Sampling methods for scaling up empirical risk minimization

Guillaume Papa

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0005

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « "Signal et Images" »

présentée et soutenue publiquement par

Guillaume Papa

le 31 Janvier 2018

Méthode d'échantillonnage pour la minimisation du risque empirique

Directeur de thèse : **Stéphan CLEMENCON**
Co-encadrement de la thèse : **Pascal BIANCHI**

Jury

M. Gabor LUGOSI, Professeur, Université Pompeu Fabra	Rapporteur
M. Joachim BUHMANN, Professeur, ETH Zurich	Rapporteur
M. Olivier TEYTAUD, Research Scientist, Google	Examinateur
M. Gérard BIAU, Professeur, Université Pierre et Marie Curie	Examinateur
M. Patrice BERTAIL, Professeur, Université Paris Nanterre	Examinateur
M. Aurélien BELLET, Enseignant-chercheur, INRIA Lille	Invité

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

École Doctorale ED130 “Informatique, télécommunications et électronique de Paris”

Sampling Methods for Scaling Up Empirical Risk Minimization

—

Méthodes d’Echantillonnage pour la Minimisation du Risque Empirique

Thèse pour obtenir le grade de docteur délivré par

TELECOM PARISTECH

Spécialité “Signal et Images”

présentée et soutenue publiquement par

Guillaume PAPA

le 31 Janvier 2018

Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

---

Jury :

Stéphan **Clémençon**	Professeur, Télécom ParisTech	Directeur
Pascal **Bianchi**	Professeur, Télécom ParisTech	Co-directeur
Gabor **Lugosi**	Professeur, Universiété Pompeu Fabra	Rapporteur
Joachim **Buhmann**	Professeur, ETH Zurich	Rapporteur
Olivier **Teytaud**	Directeur de recherches, INRIA Saclay	Examinateur
Gérard **Biau**	Professeur, LSTA Université Pierre et Marie Curie	Examinateur
Patrice **Bertail**	Professeur, Université paris Nanterre	Examinateur
Aurélien **Bellet**	Enseignant-chercheur, INRIA Lille	Invité

List of Contributions

Journal

- Optimal survey schemes for stochastic gradient descent with applications to M-estimation. (submitted)
  Authors: Clémençon, Bertail, Chautru and Papa.

Conferences

- Adaptive sampling for Incremental Optimization using stochastic gradient descent. (ALT 2015).
  Authors: Papa, Bianchi and Clémençon.
- Scalability of Stochastic Gradient Descent based on "Smart" Sampling Techniques. (INNS Big Data 2015).
  Authors: Clémençon, Bellet, Jelassi and Papa.
- SGD Algorithms based on incomplete U-Statistics: large scale minimization of empirical risk. (NIPS 2015).
  Authors: Papa, Bellet and Clémençon.
- On graph reconstructions via empirical risk minimization: fast learning rates and scalability. (NIPS 2016).
  Authors: Papa, Bellet and Clémençon.
- Learning from survey training samples: rate bounds for Horvitz-Thompson risk minimizers. (ACML 2016).
  Authors: Clémençon, Bertail and Papa.

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

1.1 Motivation

In this manuscript, we present and study sampling strategies for statistical learning related problems. The goal is to deal with issues typically arising in a big data context when the number of observations and their dimensionality impose limitations on the learning process. We thus propose to tackle this issue by employing two sampling strategies:

- Speed-up the learning process by sampling the most useful observations.
- Scale-up the problem by discarding some observations to reduce the complexity and the size of the problem.

To introduce the problem we deal with, we first give a quick reminder on Empirical Risk Minimization (ERM) in the context of binary classification. The binary classification problem is considered a running example all along this manuscript. Because it can be easily formulated, it is undeniably the most documented statistical learning problem in the literature and many of its results extend to more general frameworks (*e.g.*, multi-class classification, regression, ranking). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and (X, Y) a random pair defined on $(\Omega, \mathcal{A}, \mathbb{P})$, taking its values in some measurable product space $\mathcal{X} \times \{-1, +1\}$, with common distribution $P(dx, dy)$: the r.v. X represents some observation, hopefully useful for predicting the binary label Y. The distribution P can also be described by the pair (F, η) where $F(dx)$ denotes the marginal distribution of the input variable X and $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$, is the *posterior distribution*. The objective is to build, based on the training dataset at disposal, a measurable mapping $g : \mathcal{X} \mapsto \{-1, +1\}$, called a *classifier*, with minimum risk:

$$L(g) \stackrel{def}{=} \mathbb{P}\{g(X) \neq Y\}. \tag{1.1}$$

It is well-known that the *Bayes classifier* $g^*(x) = 2\mathbb{I}\{\eta(x) \geq 1/2\} - 1$ is a solution of the risk minimization problem $\inf_g L(g)$, where the infinimum is taken over the collection of all classifiers defined on the input space $\mathcal{X}$. The minimum risk is denoted by $L^* = L(g^*)$. Since the distribution P of the data is unknown, one substitutes the true risk with its empirical estimate

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}, \tag{1.2}$$

based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent copies of the generic random pair (X, Y). The true risk minimization is then replaced by the empirical risk minimization

$$\min_{g \in \mathcal{G}} \widehat{L}_n(g), \tag{1.3}$$

where the minimum is taken over a class $\mathcal{G}$ of classifier candidates, supposed rich enough to include the naive Bayes classifier (or a reasonable approximation of the latter). Considering a solution $\widehat{g}_n$ of (1.3), a major problem in statistical learning theory is to establish upper confidence bounds on the *excess of risk* $L(\widehat{g}_n) - L^*$ in absence of any distributional assumptions and taking into account the complexity of the class $\mathcal{G}$ (*e.g.*, described by geometric or combinatorial features such as the VC dimension) and some measure of accuracy of approximation of P by its empirical counterpart $P_n = (1/n)\sum_{i=1}^n \delta_{(X_i,Y_i)}$ over the class $\mathcal{G}$. Indeed, the excess of risk of the empirical risk minimizers is typically bounded as follows

$$L(\widehat{g}_n) - L^* \leq 2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right). \tag{1.4}$$

The second term on the right hand side is referred to as the *bias* and depends on the richness of the class $\mathcal{G}$, while the first term, called the *stochastic error*, is controlled by means of results in empirical process theory, see Boucheron et al. (2005a). Unfortunately, one of the thing typically overlooked in this kind of analysis is how to efficiently solve the ERM problem, i.e how to find $\widehat{g}_n$. It is usually approximated by some incremental optimization algorithm, iteratively computing estimator of the gradient of the empirical risk. We investigate efficient ways to scale-up the learning process and introduce sampling-based approachs to build approximations of $\widehat{g}_n$. We do so in two different fashions:

- We replace the empirical risk $\widehat{L}_n(g)$ by an approximation based on fewer terms $\tilde{L}_n(g)$. It naturally makes the learning problem easier. Let $\tilde{g}_n$ be a minimizer of $\tilde{L}_n(g)$, then (1.4) becomes:

  $$L(\tilde{g}_n) - L^* \leq 2 \sup_{g \in \mathcal{G}} |\tilde{L}_n(g) - \widehat{L}_n(g)| + 2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right).$$

  We focus on appropriately controlling $2 \sup_{g \in \mathcal{G}} |\tilde{L}_n(g) - \widehat{L}_n(g)|$ which we typically do conditionally upon the observations. Such strategy is discussed and implemented in chapters 2 and 6 for two different problems.

- When computing estimator of the gradient, most incremental algorithms uniformly and independently sample observations within the dataset. We propose to use non uniform sampling methods to compute an estimator of the gradient of $\widehat{L}_n$ with smaller variance. For the algorithms we propose, if we denote by $g_n(T)$ the classifier obtained after T iterations of the optimization algorithm, then following the lines of Bottou & Bousquet (2007), inequality (1.4) becomes:

  $$L(g_n(T)) - L^* \leq \underbrace{\widehat{L}_n(g_n(T)) - \widehat{L}_n(\widehat{g}_n)}_{(1)} + \underbrace{2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)|}_{(2)} + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right).$$

  where (1) corresponds to the optimization error and (2) corresponds to the stochastic error. This decomposition illustrates the well known fact (see Bottou & Bousquet (2007)) that when solving the Empirical Risk Minimization problem, we have to take into account the randomness induced by the observations so that the optimization accuracy is of the same order. We particularly pay attention to this fact and illustrate it both theoretically and empirically in chapter 3, 4 and 5.

The rest of this chapter is devoted to overview our different contributions. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$ and the variance of any square integrable r.v. Z by $\sigma^2(Z)$.

1.2 Learning from Survey Training Samples

This subsection is a summary of chapter 2. We place ourselves in the context of binary classification, when the observations used to train a classifier are drawn by means of a sampling/survey scheme and exhibit a complex dependence structure. We consider, $(X_1, Y_1), \ldots, (X_n, Y_n)$ a sample of independent copies of (X, Y) observed on a finite population on $\mathcal{I}_n := \{1, \ldots, n\}$. We call a *survey sample* of (possibly random) size $N \leq n$ of the population $\mathcal{I}_n$, any subset $s := \{i_1, \ldots, i_{n(s)}\} \in \mathcal{P}(\mathcal{I}_n)$ with cardinality $N =: N(s)$ less than n. A sampling scheme is defined by a probability distribution R_n on the set of all possible samples $s \in \mathcal{P}(\mathcal{I}_n)$ conditionally on the observations $\mathcal{D}_n = \{(X_i, Y_i) : \ i \in \mathcal{I}_n\}$. The probability that the unit i belongs to a random sample S drawn from the conditional distribution $\mathcal{R}_n$ is called first order *inclusion probability* and is denoted by $\pi_i = \mathbb{P}_{R_n}\{i \in S\}$. We set $\boldsymbol{\pi}_n = (\pi_1, \ldots, \pi_n)$. Given an observed sample S, it is fully determined by the r.v. $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$, where $\epsilon_i = \mathbb{I}\{i \in S\}$ for $1 \leq i \leq n$.

Most available results (see Boucheron et al. (2005c) for instance) deal with the case where the dataset $\mathcal{D}_n$ is at disposal. However this is not the case here as we only observe a subset of observations. Therefore, these results are not directly applicable to our problem, in particular because of the dependence structure induced by the sampling scheme. Nevertheless, we show that the theory of ERM can be extended to the case where statistical learning is based on observations obtained via survey samples.

Horvitz-Thompson risk. As defined in Horvitz & Thompson (1951), for any classifier candidate g, the (not available) empirical risk $\widehat{L}_n(g) = n^{-1} \sum_{1 \leq i \leq n} \mathbb{I}\{Y_i \neq g(X_i)\}$ is replaced by its Horvitz-Thompson version :

$$\overline{L}_{\boldsymbol{\epsilon}_n}(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{\epsilon_i}{\pi_i} \mathbb{I}\{g(X_i) \neq Y_i\}, \tag{1.5}$$

where $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ denotes the vector in correspondence with the sample drawn.

While many sampling plans are of interest for our problem, we particularly pay attention to *Rejective sampling*, a sampling design $\mathcal{R}_n$ of fixed size $N \leq n$, that generalizes the *simple random sampling without replacement* (where all samples with cardinality N are equally likely to be chosen). This sampling plan is more difficult to analyse because the ϵ_i are dependent r.v.. Therefore, when statistical learning is based on the observation of a sample drawn by means of a rejective scheme, classical results from learning theory do not immediately apply. Nevertheless, we show that similar results can be established for minimizer of (1.5) in the rejective case. it is due to the fact that this scheme form a collection of *negatively associated* (see Brändén & Jonasson (2012), Kramer et al. (2011)) random variables, a rather tractable type of dependence structure. Using the negative association property, we show that for a given rejective sampling scheme $\boldsymbol{\epsilon}_n$ with first order inclusion probabilities $\boldsymbol{\pi}_n$ and with $\kappa_n = N/(n \times \min_{i \leq n} \pi_i)$ we have for any solution $\bar{g}_n$ of the minimization problem $\inf_{g \in \mathcal{G}} \overline{L}_{\boldsymbol{\epsilon}_n}(g)$, an upper-bound on the stochastic error risk of order $O_{\mathbb{P}}((\kappa_n(\log n)/N)^{1/2})$.

The property of negative association being shared by many other sampling schemes, the same argument can be thus naturally applied to carry out a similar rate analysis for training data produced by such plans. However, this analysis cannot be extended to all sampling scheme. We circumvent this difficulty using the results established for rejective plan and relying on coupling argument. Consider a complex sampling scheme $\mathcal{R}_n^*$ with first order inclusion probabilities $\boldsymbol{\pi}_n^* = (\pi_1^*, \ \ldots, \ \pi_n^*)$ described by the vector $\boldsymbol{\epsilon}_n^* = (\epsilon_1^*, \ \ldots, \ \epsilon_n^*)$ (with not necessarily negatively associated r.v). Let $\bar{g}_n^*$ be a minimizer of the HT empirical risk $\bar{L}_{\boldsymbol{\epsilon}_n^*}(g) = (1/n)\sum_{i=1}^n (\epsilon_i^*/\pi_i^*)\mathbb{I}\{Y_i \neq g(X_i)\}$ over a class $\mathcal{G}$. Since we already established results in the rejective case, we introduce a rejective sampling scheme $\mathcal{R}_n$ described by the r.v. $\boldsymbol{\epsilon}_n$, with first order inclusion probabilities $\boldsymbol{\pi}_n = (\pi_1, \ \ldots, \ \pi_n)$ as well as the following quantity:

$$\check{L}_{\boldsymbol{\epsilon}_n}(g) = \frac{1}{n}\sum_{i=1}^n \frac{\epsilon_i}{\pi_i^*}\mathbb{I}\{Y_i \neq g(X_i)\}, \tag{1.6}$$

for any classifier g. Observe that (1.6) is different from the HT empirical risk $\bar{L}_{\boldsymbol{\epsilon}_n}(g)$ related to the rejective sampling scheme $\boldsymbol{\epsilon}_n$ because it involves the π_i^*'s instead of the π_i's. The excess of risk of the HT empirical risk minimizer can be expressed as follows:

$$\begin{aligned} L(\bar{g}_n^*) - \inf_{g\in\mathcal{G}} L(g) \leq 2\sup_{g\in\mathcal{G}}\left|L(g) - \widehat{L}_n(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\widehat{L}_n(g) - \bar{L}_{\boldsymbol{\epsilon}_n}(g)\right| \\ + 2\sup_{g\in\mathcal{G}}\left|\bar{L}_{\boldsymbol{\epsilon}_n}(g) - \check{L}_{\boldsymbol{\epsilon}_n}(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\check{L}_{\boldsymbol{\epsilon}_n}(g) - \bar{L}_{\boldsymbol{\epsilon}_n^*}(g)\right|. \end{aligned} \tag{1.7}$$

We controlled the first term on the right hand side of (1.7) by using Vapnik-Chervonenkis and McDiarmid inequalities (see *e.g.* Vapnik (2001) and chapter A), assertion (i) of Proposition 2.4 established in the rejective case provides a control of the second term. The third term is bounded by means of the *coupling* argument while the last term is controlled by assumptions related to the closeness between the first order inclusion probabilities $\boldsymbol{\pi}_n^*$ and $\boldsymbol{\pi}_n$. More precisely, the assumptions required in the subsequent analysis involve the total variation distance between the sampling plans $\mathcal{R}_n$ and $\mathcal{R}_n^*$:

$$d_{TV}(R_n, \ R_n^*) \stackrel{def}{=} \frac{1}{2}\sum_{s\in\mathcal{P}(\mathcal{I}_n)} |\mathcal{R}_n(s) - \mathcal{R}_n^*(s)|.$$

With $\kappa_n^* = (N/n)\min_{i\leq n}\pi_i^*$ and $\kappa_n = (N/n \times \min_{i\leq n}\pi_i)$, we show that $L(\bar{g}_n^*) - \inf_{g\in\mathcal{G}} L(g)$ is of the order of $O_{\mathbb{P}}((\kappa_n(\log n)/N)^{1/2}) + 2(\kappa_n^* + \kappa_n)(n/N)\inf_{\mathcal{R}_n} d_{TV}(\mathcal{R}_n, \ \mathcal{R}_n^*)$, where the infinimum is taken over the set of rejective schemes $\mathcal{R}_n$ with first order inclusion probabilities $\boldsymbol{\pi}_n = (\pi_1, \ \ldots, \ \pi_n)$.

The rate bound obtained depends on the minimum error made when approximating the sampling plan by a rejective sampling plan in terms of total variation distance. It is of the same order as in case where observations are sampled uniformly up to a multiplicative term and show that learning with survey sample is possible when taking into account the first order inclusion probabilities.

1.3 Sampling Strategies for Stochastic Gradient Descent (SGD)

We present in this section a summary of the results establish in chapter 3, 4 and 5, in which we introduce the problem of non uniform sampling strategy for stochastic gradient descent (SGD in short). The Empirical Risk Minimization problem previously introduced is of the utmost importance and implementing efficient algorithms to solve this problem is a question that we

tried to answer. Here we consider a more general framework than the binary classification one, and consider optimization problems of the form :

$$\min_{\theta\in\Theta} \widehat{L}_n(\theta) = \min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(Z_i,\theta), \tag{1.8}$$

where Θ is a Euclidean space, typically $\mathbb{R}^d$ with $d \geq 1$, and $l(Z_1,.), \ \ldots, \ l(Z_n,.)$ form a collection of real-valued convex continuously differentiable functions on Θ. Indeed, such an optimization problem typically arises in a broad variety of statistical learning problems, in particular supervised tasks, where the goal pursued is to learn a predictive model, fully determined by a parameter θ. The performance of the predictive function defined by θ, is typically measured by the expectation $L(\theta) = \mathbb{E}[\ell((X,Y),\theta)]$, referred to as the *risk*, where ℓ is a *loss function* assumed convex w.r.t. θ. It is usually assessed via its empirical counterpart

$$\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell((X_i,Y_i),\theta), \tag{1.9}$$

based on $n \geq 1$ supposedly available independent training examples (X_1,Y_1), ..., (X_n,Y_n), copies of the random pair $Z = (X,Y)$. The minimization problem (1.8) can be solved incrementally, by means of variants of the *stochastic approximation* method originally introduced in the seminal contribution of Robbins & Monro (1951). This one consists in computing successive estimates of a minimizer of (1.2) using the recursive equation

$$\theta_{t+1} = \theta_t - \gamma_t \widehat{r}_t(\theta_t) \tag{1.10}$$

from a preliminarily picked initial value $\theta_0 \in \Theta$, where $\widehat{r}_t$ denotes an estimator of the gradient $\nabla\widehat{L}_n$ and γ_t is the *learning rate* or *step-size*. The implementation of SGD is quite straightforward for the minimization of standard mean statistics, as it is usually performed by sampling uniformly at random (with or without replacement) a subsample of observations before computing an estimator of the gradient. In contrast to the *batch approach*, where all the data are used to estimate the gradient at each iteration (*i.e.* $\widehat{r}_t(\theta) = \nabla\widehat{L}_n(\theta)$ for all $t \geq 0$ and $\theta \in \Theta$), only subsets of the data sample are involved in the gradient estimation steps of *sampled incremental algorithms*, with the aim to reduce computational cost when n is large. In the most commonly used implementation of the *stochastic gradient descent* (SGD) algorithm, the gradient estimator is computed from a subset of reduced size $S \leq n$ uniformly drawn without replacement among all possible subsets of the dataset of size S at each step $t \geq 0$. In practice, the main limitation of this incremental optimization technique is due to the stochastic noise induced by the choice at random of the data involved in the gradient estimator computation at each iteration. In particular, most of the theoretical justification of the SGD are established in a very general framework (see Robbins & Monro (1951) or Bach & Moulines (2011a) for instance) that encompasses the ERM case. We propose to introduce non uniform sampling strategy as well as a novel analysis highlighting the benefit of using non uniform sampling strategy for the ERM problem. We first introduce in chapter 3 a novel implementation of the SGD algorithm, where the data subset used at a given step is not picked uniformly at random among all possible subsets but drawn from a specific adaptive sampling scheme, depending on the past iterations. We then propose a general framework to extend these results using survey sampling theory in chapter 4 in which we also take into account the distribution of the observation in our final analysis. We conclude this section by considering the specific case where the empirical risk takes the form of a U-statistic and propose and efficient implementation of the SGD Algorithm in this case. Here and throughout, the gradient and Hessian operators with respect to θ are denoted by ∇ and ∇^2 respectively. By convention, ∇^0 corresponds to the identity operator and gradient values are represented as column vectors. For any vector $V \in \mathbb{R}^d$,

we denote by $||V||$ its euclidian norm and for any matrix A we denote by A^T its transpose.

1.3.1 A Non Uniform Sampling Strategy for SGD

In order to speed up the learning process, we introduce a specific variant of the SGD algorithm with an *adaptive* sampling scheme, in the sense that it may vary with t, depending on the past iterations. We consider non uniform sampling with replacement. We first start by identifying a good sampling distribution by choosing the one minimizing the variance of the estimator. When drawing a sample $\mathcal{S}$ of size S with first order inclusion p_i independently with replacement, the quantity

$$\frac{1}{S}\sum_{i\in\mathcal{S}}\frac{\nabla l(Z_i,\theta)}{p_i} \tag{1.11}$$

is an unbiased estimator of $\nabla\widehat{L}_n(\theta)$ with corresponding variance equal to :

$$\frac{1}{S}\sum_{i=1}^{n}\frac{\|\nabla l(Z_i\theta)\|^2}{p_i} - \frac{||\nabla\widehat{L}_n(\theta)||^2}{S}. \tag{1.12}$$

To achieve the best estimation of the gradient (*i.e* minimizing the variance) at parameter θ conditionally upon the observations, it is therefore natural to sample observation Z_i with probability: $p_i^*(\theta) = \|\nabla l(Z_i,\theta)\|/\sum_{j=1}^n\|\nabla l(Z_j,\theta)\|$. Unfortunately, practical implementation of the above sampling scheme is not pertinent because it requires to evaluate all gradients to calculate the norms $\|\nabla l(Z_1,\theta_t)\|,\ldots,\|\nabla l(Z_n,\theta_t)\|$ at each iteration, which is precisely what we are try to avoid when using SGD. We therefore propose a sampling scheme approximating $\mathbf{p}_t^* := (p_i^*(\theta_t))_{i=1}^n$ without requiring any additional gradient evaluations. We use some old values of the gradient in our approximation. More precisely, the main idea is to replace each unknown gradient norm $\|\nabla l(Z_i,\theta_t)\|$ by a (possibly outdated) norm $g_{t,i} = \|\nabla l(Z_i,\theta_k)\|$ at some former instant $k = k(i,t)$ corresponding to the last time $k \le t$ when the Z_i was picked. More formally, we define the random sequence g_t as

$$g_{t+1,i} = \begin{cases} \|\nabla l(Z_i,\theta_{t+1})\| & \text{if } i \in \{i_{t+1}^{(1)},\ldots,i_{t+1}^{(S)}\} \\ g_{t,i} & \text{otherwise.} \end{cases} \tag{1.13}$$

Then, a natural way to approximate $\mathbf{p}_t^*$ is to construct $\bar{\mathbf{p}}_t = (\bar{p}_{t,i})_{i=1}^n$ where we set for each i

$$\bar{p}_{t,i} = \frac{g_{t,i}}{\sum_{j=1}^n g_{t,j}}\,. \tag{1.14}$$

It turns out that convergence cannot be guaranteed with this choice, because a certain component $\bar{p}_{t,i}$ can get arbitrarily close to zero. A possible remedy is to enforce a greedy sampling scheme:

$$\forall i \in \{1,\ \ldots,\ n\},\ \ p_{t,i} = \rho\nu_i + (1-\rho)\bar{p}_{t,i}, \tag{1.15}$$

where $\nu = (\nu_1,\ldots,\nu_n)$ is a probability distribution with $\nu_i > 0$ for $1 \le i \le n$, and $0 < \rho \le 1$. This condition has the following interpretation : p_t is a mixture between two laws of probability and one of this law (ν) is independent from the past. Now that we have defined our sampling strategy, the algorithm we propose is simply to compute at each iteration t an estimator of the gradient based on equation (1.11) by sampling observation according to $\mathbf{p}_t := (p_{t,i})_{i=1}^n$. This sampling strategy can also be efficiently implemented in practice and we show that sampling under this strategy only has an additional cost of $O(\log(n))$. Theoretical results are then established by means of asymptotic argument where we show that with this sampling strategy, the asymptotic behaviour of θ_t is optimal up to an error proportional to ρ. Note that

all the result obtained in chapter 3 hold true conditionally upon the observations and therefore do no take into account the statistical nature of our problem (i.e we solve ERM because we do not know the true risk). We deal with this issue in the next section where we discuss of a similar problem (Non uniform sampling strategy for SGD) in the context of M-estimation. More specifically we use the framework of survey sampling introduced previously to extend our results .

1.3.2 Horvitz Thompson Gradient Descent (HTSGD) and Applications to M-Estimation

The previous sections strongly suggests to use sampling technique to scale up learning from datasets. We now show how to incorporate efficiently survey schemes into such iterative techniques for M-estimation. More precisely, we propose a specific estimator of the gradient, that is referred to as the *Horvitz-Thompson gradient estimator* (HTGD estimator in short). For the estimator thus produced, consistency and asymptotic normality results describing its statistical performance are established. The framework we consider is the same than in section 1.3.1. We define the *Horvitz-Thompson estimator* of the gradient $\ell_n(\theta)$ based on a survey sample S drawn from a design $\mathcal{S}_n$ with (first order) inclusion probabilities $\{\pi_i\}_{1\leq i\leq n}$ and inclusion vector $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ as

$$\ell_{\mathcal{S}_n}(\theta) = \frac{1}{n}\sum_{i\in S}\frac{1}{\pi_i}\nabla l(Z_i,\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\epsilon_i}{\pi_i}\nabla l(Z_i,\theta). \tag{1.16}$$

Equipped with this notation, we study the property of a SGD algorithm when the estimator of the gradient is computed by sampling observations within some dataset under some sampling plan $\mathcal{S}_n$ (possibly depending upon t and the current value of the parameter). We denote by $\theta_n(T)$ the value of the parameter at time T. Conditioned upon the data $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$, we study the asymptotic properties of the M-estimator produced by the HTGD algorithm. The limit results stated below essentially rely on the fact that the HT estimator (1.16) of the gradient of the empirical risk is unbiased. Reduction of the asymptotic variance of $\widehat{\theta}_n(T)$ (of $\widehat{L}_n(\widehat{\theta}_n(T))$, respectively) is investigated later in the Poisson case (i.e when the ϵ_i are independent r.v.). We show that under some appropriate assumptions, the sequence $\{\widehat{\theta}_n(t)\}_{t\geq 0}$ converges to a minimizer of the empirical risk and

$$\sqrt{1/\gamma_t}\left(\widehat{\theta}_n(t) - \theta_n^*\right) \Rightarrow \mathcal{N}(0, \Sigma_{\pi,n}),$$

where the asymptotic covariance matrix $\Sigma_{\pi,n}$, is the unique solution of a Lyapunov equation involving the π_i's and defined in chapter 4. By a direct application of the second order Delta method, we then characterize the behaviour of $\widehat{L}_n(\widehat{\theta}_n(t)) - \widehat{L}_n(\theta_n^*)$ where θ_n^* is an empirical risk minimizer. We then discuss how one should choose π_i in the Poisson case (case where the ϵ_i are independent) and recall the result of chapter 3 by showing that sampling with π_i proportional to $\|\nabla l(Z_i,\theta)\|$ yields optimal results when trying to minimize the variance of $\widehat{L}_n(\widehat{\theta}_n(t)) - \widehat{L}_n(\theta_n^*)$. Denoting by N the expected size of a sample, our analysis is then finally completed by studying the behaviour of $\widehat{\theta}_n(t)$ as n, N tend to $+\infty$ at appropriate rates. This is quite different from what we did in chapter 3 because we limited our analysis to the case where n is set in advance and all our results were obtained conditionally upon the observations. Doing so allow us to illustrate the well-known trade-off between (asymptotic) generalization and optimization errors, ruled by the limit behaviour of $n\gamma_t/N$ (see Bottou & Bousquet (2008) for instance). More precisely we show that under supplementary assumptions, if $\lim n\gamma_t/N =$

$c > 0$, then we have the convergence in distribution:

$$\lim_{n,\, N\to\infty} \left\{ \lim_{t\to\infty} \sqrt{n} \left(\widehat{\theta}_n(t) - \theta^* \right) \right\} = \mathcal{N}(0, \Lambda^* + c\Sigma^*).$$

where $\lim_{n,\, N\to\infty} N\Gamma_n^* = \Gamma^*$ and Λ^* is the asymptotic covariance matrix involved in the TCL for M-estimator applied to $\theta_n^* - \theta^*$. With γ_t typically or order $O(1/t)$, the condition $\lim n\gamma_t/N = c > 0$ gives an idea of how the number of iteration should be tuned according to the number of observation and the batch size to yield optimal results. Numerical experiments are then displayed in section 4.6.

1.3.3 Stochastic Gradient Descent Algorithms based on *Incomplete U-Statistic*

Here we discuss of the implementation of SGD for *U-Statistics*. We briefly introduce the problem and notations and explain the difference with the problems of section 1.3.1 and 1.3.2.

Generalized U-statistics are extensions of standard sample mean statistics. In machine learning, they are used as performance criteria in many problems, Metric Learning and AUC in particular are two examples that we consider in our experiments. It is defined as follows:

Definition 1.1. Let $K \geq 1$ and $(d_1, \ \ldots, \ d_K) \in \mathbb{N}^{*K}$. Let $\mathbf{X}_{\{1,\, \ldots,\, n_k\}} = (X_1^{(k)}, \ \ldots, \ X_{n_k}^{(k)})$, $1 \leq k \leq K$, be K independent samples of sizes $n_k \geq d_k$ and composed of i.i.d. random variables taking their values in some measurable space $\mathcal{X}_k$ with distribution $F_k(dx)$ respectively. Let $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$ be a measurable function, square integrable with respect to the probability distribution $\mu = F_1^{\otimes d_1} \otimes \cdots \otimes F_K^{\otimes d_K}$. Assume in addition (without loss of generality) that $H(\mathbf{x}^{(1)}, \ \ldots, \ \mathbf{x}^{(K)})$ is symmetric within each block of arguments $\mathbf{x}^{(k)}$ (valued in $\mathcal{X}_k^{d_k}$), $1 \leq k \leq K$. The generalized (or K-sample) U-statistic of degrees $(d_1, \ \ldots, \ d_K)$ with kernel H, is then defined as

$$U_{\mathbf{n}}(H) = \frac{1}{\prod_{k=1}^K \binom{n_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} H\left(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)} \right), \tag{1.17}$$

where $\mathbf{n} = (n_1, \ \ldots, \ n_K)$, the symbol $\sum_{I_1} \cdots \sum_{I_K}$ refers to summation over all elements of Λ, the set of the $\prod_{k=1}^K \binom{n_k}{d_k}$ index vectors $(I_1, \ \ldots, \ I_K)$, I_k being a set of d_k indexes $1 \leq i_1 < \ldots < i_{d_k} \leq n_k$ and $\mathbf{X}_{I_k}^{(k)} = (X_{i_1}^{(k)}, \ \ldots, \ X_{i_{d_k}}^{(k)})$ for $1 \leq k \leq K$.

Subsection 1.3.1 and 1.3.2 advocates the use of SGD to deal with the number of terms involved in (1.8). Note that when the empirical risk takes the form of a generalized U-statistics, the number of terms involved in the sum is of order $O(n^{d_1 + \ldots d_K})$ making the problem extremely difficult to solve. Nevertheless we show how to implement the SGD in this case.
We place ourselves in the parametrized setting. Still denoting by Θ the parameter space, with $H : \prod_{k=1}^K \mathcal{X}_k^{d_k} \times \Theta \to \mathbb{R}$ a convex loss function, we denote the empirical version of the risk by $\theta \in \Theta \mapsto \widehat{L}_n(\theta) = U_n(H(.;\ \theta))$. As we have mentioned before, the implementation of SGD is quite straightforward for the minimization of standard mean statistics, as it is usually performed by sampling uniformly at random (with or without replacement) a batch of observations before computing an estimator of the gradient. When the empirical risk takes the form of a U-statistic, the SGD algorithm could be implemented this way. It would lead to estimator of the gradient equal to:

$$\tilde{g}_{n'}(\theta) = \frac{1}{\prod_{k=1}^K \binom{n_k'}{d_k}} \sum_{I_1} \cdots \sum_{I_K} \nabla H(\mathbf{X}_{I_1}^{\prime(1)}; \mathbf{X}_{I_2}^{\prime(2)}; \ldots; \mathbf{X}_{I_K}^{\prime(K)};\ \theta), \tag{1.18}$$

where $\sum_{I_k}$ refers to summation over all $\binom{n'_k}{d_k}$ subsets $\mathbf{X}'^{(k)}_{I_k} = (X'^{(k)}_{i_1}, \ldots, X'^{(k)}_{i_{d_k}})$ related to a set I_k of d_k indexes $1 \leq i_1 < \ldots < i_{d_k} \leq n'_k$ and $\mathbf{n}' = (n'_1, \ldots, n'_K)$. In the case of U-Statistic, we prove that this strategy (which we later refer to as "computing a *complete* U-statistic") is not efficient. We instead propose to proceed differently by drawing independently with replacement among the set of index vectors Λ, giving a gradient estimator in the form of a so-called *incomplete* U-statistic (see Lee (1990a)):

$$\bar{g}_B(\theta) = \frac{1}{B} \sum_{(I_1, \ldots, I_K) \in \mathcal{D}_B} \nabla H(\mathbf{X}^{(1)}_{I_1}, \ldots, \mathbf{X}^{(K)}_{I_K}; \theta), \tag{1.19}$$

where $\mathcal{D}_B$ is built by sampling B times with replacement in the set Λ. The parameter B is the number of terms averaged. For the same computational cost (*i.e.*, taking $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$) and implementing SGD with (1.19) rather than (1.18) yields more accurate results, essentially because (1.19) has smaller variance w.r.t. to $\nabla L(\theta)$ (except in the case where $K = 1 = d_1$). Intuitively, sampling an *incomplete* U-statistic is better because the number of observations involved is greater than the number of observation involved in the *complete* estimator.

This is highlighted when we compare the performance of the SGD methods described above conditionally upon the observed data samples by studying both the asymptotic and non asymptotic behaviour of the SGD algorithm for both of the implementations. As we have done earlier in this chapter, we propose generalization bounds where (see Bottou & Bousquet (2007)), we decompose the stochastic error as follows:

$$\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq \underbrace{2\mathbb{E}\left[\sup_{\theta \in \Theta} |\widehat{L}_n(\theta) - L(\theta)|\right]}_{\mathcal{E}_1} + \underbrace{\mathbb{E}\left[\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta^*_n)\right]}_{\mathcal{E}_2}. \tag{1.20}$$

where $\theta^* = \operatorname{argmin}_{\theta \in \Theta} L(\theta)$. The generalization bound show the advantage of using an incomplete U-statistic (1.19) as the gradient estimator while highlighting the well-known fact that when using some optimization method to solve the Empirical Risk Minimization problem, we have to take into account the generalization bounds so that the optimization accuracy is of the same order (see subsection 1.3.2). Numerical experiments on Metric Learning problem and AUC optimization are then displayed.

1.4 Fast Learning Rates for Graph Reconstruction

This section is a summary of chapter 6 in which we present a brief overview of the *graph reconstruction* problem. We first set our context before describing the problem of interest. Let $G = (V, E)$ be an undirected random graph with a set $V = \{1, \ldots, n\}$ of $n \geq 2$ vertices and a set $E = \{e_{i,j} : 1 \leq i \neq j \leq n\} \in \{0, 1\}^{n(n-1)}$ describing its edges: for all $i \neq j$, we have $e_{i,j} = e_{j,i} = +1$ if the vertices i and j are connected by an edge and $e_{i,j} = e_{j,i} = 0$ otherwise. We also assume that for all $i \in V$, a continuous r.v. X_i is associated to vertex i. The X_i's are i.i.d. and for any $i \neq j$, the random pair (X_i, X_j) gives some information useful to predict the occurrence of an edge connecting the vertices i and j. Conditioned upon the features $(X_1, \ldots, X_n)$, any binary variables $e_{i,j}$ and $e_{k,l}$ are independent only if $\{i, j\} \cap \{k, l\} = \emptyset$. In particular, the conditional distribution of $e_{i,j}$, $i \neq j$, is supposed to depend on (X_i, X_j) solely and is described by:

$$\eta(X_i, X_j) = \mathbb{P}\{e_{i,j} = +1 \mid (X_i, X_j)\}. \tag{1.21}$$

The learning problem introduced by Biau & Bleakley (2006), referred to as *graph reconstruction*, consists in building a symmetric *reconstruction rule* $g : \mathcal{X}^2 \to \{0, 1\}$, from a training graph G, with nearly minimum *reconstruction risk*

$$\mathcal{R}(g) = \mathbb{P}\left\{g(\mathbf{X}_1, \mathbf{X}_2) \neq \mathbf{e}_{1,2}\right\}, \tag{1.22}$$

thus achieving a comparable performance to that of the Bayes rule $g^*(x_1, x_2) = \mathbb{I}\{\eta(x_1, x_2) > 1/2\}$, whose risk is given by $\mathcal{R}^* = \mathbb{E}[\min\{\eta(\mathbf{X}_1, \mathbf{X}_2),\ 1 - \eta(\mathbf{X}_1, \mathbf{X}_2)\}] = \inf_g \mathcal{R}(g)$.

The reconstruction risk (1.22) is replaced by its empirical version based on the labelled sample $\mathbb{D}_n = \{(X_i, X_j, e_{i,j}) :\ 1 \leq i < j \leq n\}$ related to G:

$$\widehat{\mathcal{R}}_n(g) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}. \tag{1.23}$$

Let $\widehat{g}_n$ be an empirical risk minimizer : $\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_n(g)$, where $\mathcal{G}$ is a class of reconstruction rules. As we have done before, the performance of $\widehat{g}_n$ is then measured by $\mathcal{R}(\widehat{g}_n) - \inf_{g \in \mathcal{G}} \mathcal{R}(g)$, which can be bounded if we can derive probability inequalities for the maximal deviation

$$\sup_{g \in \mathcal{G}} |\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g)|. \tag{1.24}$$

Biau & Bleakley (2006) establish rate bounds of the order $O_{\mathbb{P}}(1/\sqrt{n})$ for the reconstruction risk of $\widehat{g}_n$ under appropriate complexity assumptions (namely, $\mathcal{G}$ is of finite VC-dimension). We prove that rates of order $O_{\mathbb{P}}(\log n/n)$ are always attained by the minimizers of the empirical reconstruction risk (1.23) with no additional assumptions. To establish this result, we rely on some arguments used in the fast rate analysis for empirical minimization of U-statistics (Clémençon et al., 2008a), although these results only hold true under restrictive distributional assumptions. Whereas the quantity (1.23) is not a U-statistic, we rewrite the difference between the excess of reconstruction risk of any candidate rule $g \in \mathcal{G}$ and its empirical counterpart as the sum of its conditional expectation given the X_i's, which is a U-statistic, plus a residual term. Denoting by $\Lambda(g) = \mathcal{R}(g) - \mathcal{R}^*$ the excess reconstruction risk with respect to the Bayes rule, its empirical estimator is given by

$$\Lambda_n(g) = \widehat{\mathcal{R}}_n(g) - \widehat{\mathcal{R}}_n(g^*).$$

For all $g \in \mathcal{G}$, one may write:

$$\Lambda_n(g) - \Lambda(g) = U_n(g) + \widehat{W}_n(g), \tag{1.25}$$

where

$$U_n(g) = \mathbb{E}\left[\Lambda_n(g) - \Lambda(g) \mid X_1,\ \ldots,\ X_n\right]$$

is a one sample U-statistic of degree 2.

Under a certain "low-noise" condition, the analysis carried out by Clémençon et al. (2008a) shows that the small variance property of U-statistics lead to fast learning rates for empirical risk minimizers. We show that this condition is always fulfilled for the specific U-statistic $U_n(g)$ involved in the decomposition (1.25). This result is due to the fact that the empirical reconstruction risk is not an average over all pairs (*i.e.*, a U-statistic of order 2) but an average over *randomly* selected pairs (random selection being ruled by the function η). We then conclude the proof of the results by establishing that the remaining term $\widehat{W}_n(g)$ is also of order $O_{\mathbb{P}}(1/n)$.

We finally conclude our analysis by scaling up the learning process. Similarly to chapter 5, for large training graphs, the complexity of merely computing $\widehat{\mathcal{R}}_n(g)$ is prohibitive as the number of terms involved in the summation is $O(n^2)$. Just like we did in section B.3.3, we introduce a sampling-based approach to build approximations of the reconstruction risk with much fewer terms than $O(n^2)$. Instead of the empirical reconstruction risk (1.23), we consider an incomplete approximation obtained by sampling *pairs of vertices* (and not vertices) with replacement. A parallel can easily be drawn with the results obtained in chapter 5 where we recommended to implement the SGD algorithm with incomplete U-Statistic, which corresponds in this case to sampling edges instead of nodes. Formally, we define the *incomplete graph reconstruction risk* based on $B \geq 1$ pairs of vertices as

$$\widetilde{\mathcal{R}}_B(g) = \frac{1}{B} \sum_{(i,j)\in\mathcal{P}_B} \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}, \tag{1.26}$$

where $\mathcal{P}_B$ is a set of cardinality B built by sampling with replacement in the set $\Theta_n = \{(i,j) : 1 \leq i < j \leq n\}$. For any $b \in \{1, \ldots, B\}$ and all $(i,j) \in \Theta_n$, denote by $\epsilon_b(i,j)$ the variable indicating whether the pair (i,j) has been picked at the b-th draw. The incomplete risk is then represented by:

$$\widetilde{\mathcal{R}}_B(g) = \frac{1}{B} \sum_{b=1}^{B} \sum_{(i,j)\in\Theta_n} \epsilon_b(i,j) \cdot \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}. \tag{1.27}$$

(B.38) is an unbiased estimator of the true risk (1.22) and given the X_i's, its conditional expectation is equal to (4.2). When taking $B = o(n^2)$, computational costs are significantly reduced, at the price of a slightly increased variance. We characterize the performance of solutions $\widetilde{g}_B$ to the computationally simpler problem $\min_{g\in\mathcal{G}} \widetilde{\mathcal{R}}_B(g)$ and show that with only $B = O(n)$ pairs, the rate is of the same order (up to a $\log$ factor) as the one obtained by Biau & Bleakley (2006) for the maximal deviation $\sup_{g\in\mathcal{G}} |\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$ related to the complete reconstruction risk $\widehat{\mathcal{R}}_n(g)$ with $O(n^2)$ pairs. As expected, we show that the number $B \geq 1$ of pairs of vertices plays the role of a tuning parameter, ruling a trade-off between statistical accuracy (taking $B(n) = O(n^2)$ fully preserves the convergence rate) and computational complexity. Numerical experiments are displayed in section 6.5 to illustrate our different results.

PART I

Learning from Survey Training Samples

CHAPTER 2
Learning from Survey Training Samples

Abstract The generalization ability of minimizers of the empirical risk in the context of binary classification has been investigated under a wide variety of complexity assumptions for the collection of classifiers over which optimization is performed. In contrast, the vast majority of the works dedicated to this issue stipulate that the training dataset used to compute the empirical risk functional is composed of i.i.d. observations and involve sharp control of uniform deviation of i.i.d. averages from their expectation. Beyond the cases where training data are drawn uniformly without replacement among a large i.i.d. sample or modelled as a realization of a weakly dependent sequence of r.v.'s, statistical guarantees when the data used to train a classifier are drawn by means of a more general sampling/survey scheme and exhibit a complex dependence structure have not been documented in the literature yet. In this chapter, we show that the theory of empirical risk minimization can be extended to situations where statistical learning is based on survey samples and knowledge of the related inclusion probabilities. We prove that minimizing a weighted version of the empirical risk, referred to as the Horvitz-Thompson risk (HT risk), over a class of controlled complexity lead to a rate for the excess risk of the order $O((\kappa_n(\log n)/N)^{1/2})$ with $\kappa_n = (N/n)/\min_{i\leq n}\pi_i$, when data are sampled by means of a rejective scheme of (deterministic) size N within a statistical population of cardinality $N \leq n$ with probability weights $\pi_i > 0$. Extension to other sampling schemes are then established by a coupling argument. Beyond theoretical results, numerical experiments are displayed in order to show the relevance of HT risk minimization and that ignoring the sampling scheme used to generate the training dataset may completely jeopardize the learning procedure.

2.1 Introduction

Whereas statistical learning techniques crucially exploit data that can serve as examples to train a decision rule, they may also make use of weights individually assigned to the observations, resulting from survey sampling. Such weights could correspond either to true inclusion probabilities or else to calibrated or post-stratification weights, minimizing some discrepancy under certain margin constraints for the inclusion probabilities. In the context of statistical inference based on survey data, the asymptotic properties of specific statistics such as Horvitz-Thompson estimators (*cf* Horvitz & Thompson (1951)), whose computation involves not only the observations but also the weights, have been investigated, in particular, mean estimation and regression have been the subject of much attention, refer to Hajek (1964), Rosen (1972), Robinson (1982), Deville & Särndal (1992), Berger (1998) for instance, and a comprehensive functional limit theory for distribution function estimation is progressively documented in the statistical literature, see Gill et al. (1988), Breslow & Wellner (2007), Breslow & Wellner (2008), Breslow et al. (2009), Saegusa & Wellner (2011). At the same time, the last decades have witnessed a rapid development of the field of machine-learning. Revitalized by different breakout algorithms (*e.g.* SVM, boosting methods), its practice is now supported by a sound probabilistic theory based on recent non asymptotic results in the study of empirical processes,

see Devroye et al. (1996a), Koltchinskii (2006), Boucheron et al. (2005a). However, most papers dedicated to theoretical results grounding the *Empirical Risk Minimization* approach, the main paradigm of statistical learning, assume that the training of a decision rule is based on a dataset formed of independent replications of a generic random vector Z, a collection of $n \geq 1$ i.i.d. observations $Z_1, \ldots, Z_n$ namely. In contrast, few results are available in situations where the training dataset is generated by a more complex sampling scheme. One may refer to Bardenet & Maillard (2015) for concentration inequalities permitting to study the generalization ability of empirical risk minimizers when the training data are obtained by standard *sampling without replacement* (SWOR in abbreviated form) or to Steinwart et al. (2009) in the case where the decision rule is learnt from a path of a *weakly dependent stochastic* process.

In this chapter, we extend the ERM theory to situations where the training dataset is generated by means of a more general sampling scheme, with possibly unequal probability weights. We first consider the case of *rejective* sampling (sometimes referred to as *conditional Poisson* sampling), an important generalization of basic SWOR. The rate bound results obtained by means of properties of so-termed *negatively related random variables* in this case are next shown to extend to a class of more general sampling schemes by a coupling argument. In addition, numerical experiments are carried out in order to provide empirical evidence of the approach developed. They show in particular that statistical accuracy of the ERM paradigm may fail if the sampling scheme underlying the training dataset is ignored.

The chapter is organized as follows. In section 2.2, the probabilistic framework of the present study is described and basic results of the probabilistic theory of classification are briefly recalled, together with some important notions of survey theory. The main theoretical results are stated in section 2.3, while illustrative numerical experiments are presented in section 2.4.

2.2 Background and Preliminaries

As a first go, we start with recalling key concepts pertaining to the theory of empirical risk minimization in binary classification, the flagship problem in statistical learning. A few notions related to survey theory are next described, which will be involved in the subsequent analysis. Throughout the chapter, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point a by δ_a, the power set of any set E by$\mathcal{P}(E)$, the cardinality of any finite set A by $\#A$.

2.2.1 Binary Classification - Empirical Risk Minimization Theory

The binary classification problem is undeniably the most documented statistical learning problem in the literature and certain results extend to other general frameworks (*e.g.* multi-class classification, regression,ranking). With $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and (X, Y) a random pair defined on $(\Omega, \mathcal{A}, \mathbb{P})$, taking its values in some measurable product space $\mathcal{X} \times \{-1, +1\}$, with common distribution $P(dx, dy)$: the r.v. X models some observation, hopefully useful for predicting the binary label Y. The distribution P can also be described by the pair (F, η) where $F(dx)$ denotes the marginal distribution of the input variable X and $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$, is the *posterior distribution*. The objective of binary classification is to build, based on the training dataset at disposal, a measurable mapping $g : \mathcal{X} \mapsto \{-1, +1\}$, called a *classifier*, with minimum risk:

$$L(g) \stackrel{def}{=} \mathbb{P}\{g(X) \neq Y\}. \tag{2.1}$$

IThe *Bayes classifier* $g^*(x) = 2\mathbb{I}\{\eta(x) \geq 1/2\} - 1$ is a solution of the risk minimization problem $\inf_g L(g)$, where the infinimum is taken over the collection of all classifiers defined on the input space $\mathcal{X}$. The minimum risk is denoted by $L^* = L(g^*)$. Since the distribution P of the data is unknown, one substitutes the true risk with its empirical estimate

$$\widehat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}, \tag{2.2}$$

based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent copies of the generic random pair (X, Y). The true risk minimization is then replaced by the empirical risk minimization

$$\min_{g\in\mathcal{G}} \widehat{L}_n(g), \tag{2.3}$$

where the minimum is taken over a class $\mathcal{G}$ of classifier candidates, supposed rich enough to include the naive Bayes classifier (or a reasonable approximation of the latter). Considering a solution $\widehat{g}_n$ of (2.3), a major problem in statistical learning theory is to establish upper confidence bounds on the *excess of risk* $L(\widehat{g}_n) - L^*$ in absence of any distributional assumptions and taking into account the complexity of the class $\mathcal{G}$ (*e.g.* described by geometric or combinatorial features such as the VCdimension) and some measure of accuracy of approximation of P by its empirical counterpart $P_n = (1/n)\sum_{i=1}^n \delta_{(X_i,Y_i)}$ over the class $\mathcal{G}$. Indeed, one typically bounds the excess of risk of the empirical risk minimizers as follows

$$L(\widehat{g}_n) - L^* \leq 2\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left(\inf_{g\in\mathcal{G}} L(g) - L^*\right).$$

The second term on the right hand side is referred to as the *bias* and depends on the richness of the class $\mathcal{G}$, while the first term, called the *stochastic error*, is controlled by means of results in empirical process theory, see chapter A and Boucheron et al. (2005a).

Remark 2.1. (ON RISK SURROGATES) Although its study is of major interest from a theoretical perspective, the problem (2.3) is generally NP-hard. For this reason, the cost function $\mathbb{I}\{-Yg(X) > 0\}$ is replaced in practice by a non-negative convex cost function $\phi(Yg(X))$, turning empirical risk minimization to a tractable convex optimization problem. Typical choices include the exponential cost $\phi(u) = \exp(u)$ used in boosting algorithms, the hinge loss $\phi(u) = (1+u)_+$ in the case of support vector machines and the *logit* cost $\phi(u) = \log(1 + \exp(u))$ for Neural networks, see Bartlett et al. (2006) and the references therein.

In this chapter, we consider the situation where the training data used to compute of the empirical risk (2.2) is not an i.i.d. sample but the product of a more general sampling plan of fixed size $N \geq 1$.

2.2.2 Sampling Schemes and Horvitz-Thompson Estimation

Let $n \geq 1$. In the standard *superpopulation* framework we consider, $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a sample of independent copies of (X, Y) observed on a finite population $\mathcal{I}_n := \{1, \ldots, n\}$. We call a *survey sample* of (possibly random) size $N \leq n$ of the population $\mathcal{I}_n$, any subset $s := \{i_1, \ldots, i_{n(s)}\} \in \mathcal{P}(\mathcal{I}_n)$ with cardinality $N =: N(s)$ less than n. A sampling design is determined by a conditional probability distribution R_n on the set of all possible samples $s \in \mathcal{P}(\mathcal{I}_n)$ given the original data $\mathcal{D}_n = \{(X_i, Y_i) : i \in \mathcal{I}_n\}$. For any $i \in \{1, \ldots, n\}$, the first order *inclusion probability*, $\pi_i = \mathbb{P}_{R_n}\{i \in S\}$ is the probability that the unit i belongs to a random sample S drawn from the conditional distribution R_n. We set $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$.

The second order inclusion probabilities are denoted by $\pi_{i,j} = \mathbb{P}_{R_n}\{(i,j) \in S^2\}$, for any $i \neq j$ in $\{1,\ldots,n\}^2$. The information related to the observed sample $S \subset \{1,\ldots,n\}$ is fully enclosed in the r.v. $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$, where $\epsilon_i = \mathbb{I}\{i \in S\}$ for $1 \leq i \leq n$. The 1-d marginal conditional distributions of the sampling scheme $\boldsymbol{\epsilon}_n$ given $\mathcal{D}_n$ are the Bernoulli distributions $\mathcal{B}(\pi_i) = \pi_i\delta_1 + (1-\pi_i)\delta_0$, $1 \leq i \leq n$, and the covariance matrix Γ_n of the r.v. $\boldsymbol{\epsilon}_n$ has entries given by $\Gamma_n(i,j) = \pi_{i,j} - \pi_i\pi_j$, with $\pi_{i,i} = \pi_i$ by convention, for $1 \leq i,j \leq n$. Observe that, equipped with the notations above, $\sum_{1\leq i\leq n} \epsilon_i = n(S)$. One may refer to Cochran (1977), Deville (1987) for accounts of survey sampling techniques. Notice also that, in many applications, the inclusion probabilities are built using some extra information, typically by means of *auxiliary random variables* $W_1, \ldots, W_n$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking their values in some measurable space $\mathcal{W}$: $\forall i \in \{1, \ldots, n\}$, $\pi_i = Nh(W_i)/\sum_{1\leq j\leq n} h(W_j)$, where $N \max_{1\leq i\leq n} h(W_i) \leq \sum_{1\leq i\leq n} h(W_i)$ almost-surely and $h : \mathcal{W} \to]0, +\infty[$ is a measurable *link function*. The (X_i, Y_i, W_i)'s are generally supposed to be i.i.d. copies of a generic r.v. (X, Y, W). See Särndall & B. Swensson (2003) for more details. For simplicity, the π_i's are supposed to be deterministic in the subsequent analysis, which boils down to carrying out the study conditionally upon the W_i's in the example aforementioned.

Horvitz-Thompson risk. As defined in Horvitz & Thompson (1951), the Horvitz-Thompson version of the (not available) empirical risk $\widehat{L}_n(g) = n^{-1}\sum_{1\leq i\leq n} \mathbb{I}\{Y_i \neq g(X_i)\}$ of any classifier candidate g based on the sampled data $\{(X_i, Y_i) : i \in S\}$ with $S \sim \mathcal{R}_n$ is given by:

$$\overline{L}_{\boldsymbol{\epsilon}_n}(g) = \frac{1}{n}\sum_{i\in S} \frac{1}{\pi_i}\mathbb{I}\{g(X_i) \neq Y_i\} = \frac{1}{n}\sum_{i=1}^{n} \frac{\epsilon_i}{\pi_i}\,\mathbb{I}\{g(X_i) \neq Y_i\} \tag{2.4}$$

with the convention that $0/0 = 0$ and where $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ denotes the vector in correspondence with the sample S. Observe that, conditionally upon the (X_i, Y_i)'s, the quantity (2.4), that shall be referred to as the *empirical Horvitz-Thompson risk* (empirical HT risk in short) throughout the chapter, is an unbiased estimate of the empirical risk $\widehat{L}_n(g)$. Its (pointwise) consistency and asymptotic normality are established in Robinson (1982) and Berger (1998) for a variety of sampling schemes. Limit results of functional nature are established in Gill et al. (1988) for specific biased sampling models, refer also to Breslow & Wellner (2007), Breslow & Wellner (2008), Saegusa & Wellner (2011), Bertail et al. (2013).

We investigate the statistical performance of minimizers $\bar{g}_n$ of the HT risk (2.4) over the class $\mathcal{G}$ under adequate assumptions for the sampling scheme $\mathcal{R}_n$ used to generate the training dataset. We point out that such an analysis is far from straightforward due to the possible dependence structure of the terms involved in the summation (2.4): except in the Poisson case (recalled below), concentration results for empirical processes cannot be directly applied to control maximal deviations of the type

$$\sup_{g\in\mathcal{G}} |\overline{L}_{\boldsymbol{\epsilon}_n}(g) - L(g)|.$$

Conditional Poisson sampling. One of the simplest sampling plan is undeniably the *Poisson survey scheme* (without replacement), a generalization of *Bernoulli sampling* originally proposed in Goodman (1949) for the case of unequal weights: the ϵ_i's are independent and the sampling distribution is thus entirely determined by the first order inclusion probabilities $\mathbf{p}_n = (p_1, \ldots, p_n) \in]0,1[^n$:

$$\forall s \in \mathcal{P}(\mathcal{I}_n), \quad P_n(s) = \prod_{i\in S} p_i \prod_{i\notin S}(1-p_i). \tag{2.5}$$

Observe in addition that the behaviour of the quantity (2.4) can be then investigated by means of results established for sums of independent random variables. However, the major drawback of this sampling plan lies in the random nature of the corresponding sample size, impacting significantly the variability of (2.4). The variance of the Poisson sample size is given by $d_n = \sum_{i=1}^{n} p_i(1-p_i)$, while the conditional variance of (2.4) is in this case: $\sum_{i=1}^{n}((1-p_i)/p_i)\mathbb{I}\{g(X_i) \neq Y_i\}$. For this reason, *rejective sampling*, a sampling design $\mathcal{R}_n$ of fixed size $N \leq n$, is often preferred in practice. It generalizes the *simple random sampling without replacement* (where all samples with cardinality N are equally likely to be chosen, with probability $(n-N)!/n!$, all the corresponding first and second order probabilities being thus equal to N/n and $N(N-1)/(n(n-1))$ respectively). Denoting by $\boldsymbol{\pi}_n = (\pi_1, \ldots, \pi_N)$ its first order inclusion probabilities and by $\mathcal{S}_n = \{s \in \mathcal{P}(\mathcal{I}_n) : \#s = N\}$ the subset of all possible samples of size N, it is defined by:

$$\forall s \in \mathcal{S}_n, \quad \mathcal{R}_n(s) = C \prod_{i \in s} p_i \prod_{i \notin s} (1-p_i), \tag{2.6}$$

where $C = 1/\sum_{s \in \mathcal{S}_n} \prod_{i \in s} p_i \prod_{i \notin s}(1-p_i)$ and the vector $\mathbf{p}_n = (p_1, \ldots, p_n) \in]0,1[^n$ yields first order inclusion probabilities equal to the π_i's and is such that $\sum_{i \leq n} p_i = n$. Under this latter additional condition, such a vector $\mathbf{p}_n$ exists and is unique (see Dupacova (1979)) and the related representation (2.6) is then said to be *canonical*[1]. Comparing (2.6) and (2.5) reveals that rejective $\mathcal{R}_n$ sampling of fixed size N can be viewed as Poisson sampling given that the sample size is equal to N. It is for this reason that rejective sampling is usually referred to as *conditional Poisson sampling*. One must pay attention not to get the π_i's and the p_i's mixed up: the latter are the first order inclusion probabilities of P_n, whereas the former are those of its conditional version $\mathcal{R}_n$. However they can be related by means of the results stated in Hajek (1964) (see Theorem 5.1 therein): $\forall i \in \{1, \ldots, n\}$,

$$\pi_i(1-p_i) = p_i(1-\pi_i) \times \left(1 - (\tilde{\pi} - \pi_i)/d_n^* + o(1/d_n^*)\right), \tag{2.7}$$

$$p_i(1-\pi_i) = \pi_i(1-p_i) \times \left(1 - (\tilde{p} - p_i)/d_n + o(1/d_n)\right), \tag{2.8}$$

where $d_n^* = \sum_{i=1}^{n} \pi_i(1-\pi_i)$, $\tilde{\pi} = (1/d_n^*)\sum_{i=1}^{n} \pi_i^2(1-\pi_i)$ and $\tilde{p} = (1/d_n)\sum_{i=1}^{n}(p_i)^2(1-p_i)$.

More examples of sampling schemes with fixed size are given in section 2.8.

2.3 Main Results

We first consider the case where statistical learning is based on the observation of a sample drawn by means of a rejective scheme. As shall be seen below, the main argument underlying the results obtained relies on the fact that the related scheme form a collection of *negatively associated* (binary) random variables, a rather tractable type of dependence structure. This property being shared by many other sampling schemes of deterministic size, the same argument can be thus naturally applied to carry out a similar rate analysis for training data produced by such plans. Extensions of these results to more general sampling schemes are also considered by means of a *coupling* technique.

[1]Notice that any vector $\mathbf{p}'_n \in]0,1[^n$ such that $p_i/(1-p_i) = cp'_i/(1-p'_i)$ for all $i \in \{1, \ldots, n\}$ for some constant $c > 0$ can be used to write a representation of $\mathcal{R}_n$ of the same type as (2.6)

2.3.1 Horvitz-Thompson Empirical Risk Minimization in the Rejective Case

For clarity, we first recall the definition of *negatively associated random variables*, see Joag-Dev & Proschan (1983).

Definition 2.2. Let $Z_1, \ldots, Z_n$ be random variables defined on the same probability space, valued in a measurable space $(E, \mathcal{E})$. They are said to be negatively associated iff for any pair of disjoint subsets A_1 and A_2 of the index set $\{1, \ldots, n\}$

$$Cov\left(f((Z_i)_{i\in A_1}),\ g((Z_j)_{j\in A_2})\right) \leq 0, \tag{2.9}$$

for any real valued measurable functions $f : E^{\#A_1} \to \mathbb{R}$ and $g : E^{\#A_2} \to \mathbb{R}$ that are both increasing in each variable.

The theorem stated below reveals that any rejective scheme $\boldsymbol{\epsilon}_n$ forms a collection of negatively associated r.v.'s. The proof is given in Appendix 2.6.

Theorem 2.3. *Let $n \geq 1$ and $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_N)$ be the vector of indicator variables related to a rejective plan on $\mathcal{I}_n$. Then, the binary random variables $\epsilon_1, \ldots, \epsilon_n$ are negatively associated.*

The result above permits to handle the dependence of the terms involved in the summation (2.4). It is the key argument for proving the following proposition, which extends results for training datasets generated by basic sampling without replacement (*i.e.* in the case of all equal weights: $\pi_i = N/n$ for $i = 1, \ldots, n$), refer to Bardenet & Maillard (2015) (see also Serfling (1974)).

Proposition 2.4. *Suppose that the sampling scheme $\boldsymbol{\epsilon}_n$ is rejective with first order inclusion probabilities $\boldsymbol{\pi}_n$ and that the class $\mathcal{G}$ is of finite* VC *dimension $V < +\infty$. Set $\kappa_n = N/(n \times \min_{i\leq n} \pi_i)$. Then, the following assertions hold true.*

(i) For any $\delta \in (0,1)$, with probability larger than $1-\delta$, we have: $\forall N \leq n$,

$$\sup_{g\in\mathcal{G}} \left| \bar{L}_{\boldsymbol{\epsilon}_n}(g) - \widehat{L}_n(g) \right| \leq \sqrt{2\kappa_n \frac{\log(\frac{2}{\delta}) + V\log(n+1)}{N}} + 2\kappa_n \frac{\log(\frac{2}{\delta}) + V\log(n+1)}{3N}. \tag{2.10}$$

(ii) For any solution $\bar{g}_n$ of the minimization problem $\inf_{g\in\mathcal{G}} \overline{L}_{\boldsymbol{\epsilon}_n}(g)$ is such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have: $\forall n \geq 1$,

$$L(\bar{g}_n) - L^* \leq 2\sqrt{2\kappa_N \frac{\log(\frac{4}{\delta}) + V\log(n+1)}{N}} + 4\kappa_n \frac{\log(\frac{4}{\delta}) + V\log(n+1)}{3N} + C\sqrt{\frac{V}{n}} + 2\sqrt{\frac{2\log(\frac{2}{\delta})}{n}} + \inf_{g\in\mathcal{G}} L(g) - L^*.$$

The factor κ_N involved in the bounds above reflects the influence of the sampling scheme (notice incidentally that $\kappa_n \geq 1$ since $\sum_{i\leq n} \pi_i = N$). In the SWOR case, *i.e.* when $\pi_i = N/n$ for all $i \in \{1, \ldots, n\}$, it is then minimum, equal to 1. More generally, when $N = o(n)$ as $n \to +\infty$, as soon as the weights cannot vanish faster than N/n, the rate achieved by minimizers of the HT risk is of the order $O(\sqrt{(\log n)/N})$. Many sampling schemes (*e.g.* Rao-Sampford sampling, Pareto sampling, Srinivasan sampling) of fixed size are actually described

by random vectors $\boldsymbol{\epsilon}_n$ with negatively associated components, see Brändén & Jonasson (2012) or Kramer et al. (2011). Hence, Proposition 2.4's proof shows that the bounds stated above immediately extend to these cases. See section 2.8 for more details and references. Before showing how the rate bounds established can be extended to even more general sampling schemes, a few remarks are in order.

Remark 2.5. (COMPLEXITY ASSUMPTIONS) We point out that the results stated can be established, essentially by means of the same argument as that developed in the section 2.6, under complexity assumptions of different nature, involving metric entropy conditions for instance (see *e.g.* van der Vaart & Wellner (1996)).

Remark 2.6. (MODEL SELECTION) A slight modification of the argument involved in Proposition 2.4 straightforwardly leads to bounds on the expected excess risk $\mathbb{E}[L(\bar{g}_{\boldsymbol{\epsilon}_n})]-\inf_{g\in\mathcal{G}} L(g)$. Following the *Structural Risk Minimization* principle (see Vapnik (2001)), such VC bounds can be next used as complexity regularization terms to penalize additively the HT risk (2.4) and, for a sequence of model classes $\mathcal{G}_k$ with $k \geq 1$ of finite VC dimension, select the classifier among the minimizers $\{\arg\min_{g\in\mathcal{G}_k} \bar{L}_{\boldsymbol{\epsilon}_n}(g),\ k \geq 1\}$, which has approximately minimal risk.

Remark 2.7. (BIASED HT RISK) As recalled in section 2.7, the canonical parameters $\mathbf{p}_n$ are practically used to build a rejective sampling scheme $\boldsymbol{\epsilon}_n$ rather than its vector of first order inclusion probabilities $(\pi_1,\ \ldots,\ \pi_N)$, whose explicit computation based on the p_i's is a difficult task, refer to Chen et al. (1994) for dedicated algorithms. For this reason, one could be naturally tempted to minimize the alternative risk estimate $\widetilde{L}_{\boldsymbol{\epsilon}_n}(g) = (1/n)\sum_{i\leq n}(\epsilon_i/p_i)\mathbb{I}\{Y_i \neq g(X_i)\}$. As proved in section 2.7, refinements of Eq. (2.7)-(2.8) show that

$$\sup_{g\in\mathcal{G}} |\widetilde{L}_{\boldsymbol{\epsilon}_n}(g) - \bar{L}_{\boldsymbol{\epsilon}_n}(g)| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{p_i} - \frac{1}{\pi_i}\right| \leq 6N\kappa_N/(nd_N), \tag{2.11}$$

one may directly derive a rate bound for solutions of $\inf_{g\in\mathcal{G}} \widetilde{L}_{\boldsymbol{\epsilon}_n}(g)$ from bound (ii) in Proposition 2.4. In particular, the learning rate achieved by $\bar{g}_n$ is preserved when $1/\sqrt{N} = O(\min_{i\leq n}\pi_i)$ as $n,\ N \to +\infty$.

2.3.2 Extensions to More General Sampling Schemes

We now extend the rate bound analysis carried out in the previous subsection to more complex sampling schemes (described by a random vector ϵ_n^* possibly exhibiting a very complex dependence structure). In order to give an insight into the arguments which the extension is based on, additional notations are required. In this section, we consider a general sampling design $\mathcal{R}_n^*$ with first order inclusion probabilities $\boldsymbol{\pi}_n^* = (\pi_1^*,\ \ldots,\ \pi_n^*)$ described by the vector $\boldsymbol{\epsilon}_n^* = (\epsilon_1^*,\ \ldots,\ \epsilon_n^*)$ and investigate the performance of minimizers $\bar{g}_n^*$ of the HT empirical risk $\bar{L}_{\boldsymbol{\epsilon}_n^*}(g) = (1/n)\sum_{i=1}^n(\epsilon_i^*/\pi_i^*)\mathbb{I}\{Y_i \neq g(X_i)\}$ over a class $\mathcal{G}$. We also consider a rejective sampling scheme $\mathcal{R}_n$ described by the r.v. $\boldsymbol{\epsilon}_n$, with first order inclusion probabilities $\boldsymbol{\pi}_n = (\pi_1,\ \ldots,\ \pi_n)$ defined on the same probability space, as well as the following quantity:

$$\check{L}_{\boldsymbol{\epsilon}_n}(g) = \frac{1}{n}\sum_{i=1}^{n}\frac{\epsilon_i}{\pi_i^*}\mathbb{I}\{Y_i \neq g(X_i)\} \tag{2.12}$$

for any classifier g. Observe that (2.12) differs from the HT empirical risk $\bar{L}_{\boldsymbol{\epsilon}_n}(g)$ related to the rejective sampling scheme $\boldsymbol{\epsilon}_n$ in the weights it involves, the π_i^*'s instead of the π_i's namely. Equipped with this notation, the excess of risk of the HT empirical risk minimizer can

be bounded as follows:

$$L(\bar{g}_n^*) - \inf_{g\in\mathcal{G}} L(g) \leq 2\sup_{g\in\mathcal{G}}\left|L(g) - \widehat{L}_n(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\widehat{L}_n(g) - \bar{L}_{\boldsymbol{\epsilon}_n}(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\bar{L}_{\boldsymbol{\epsilon}_n}(g) - \check{L}_{\boldsymbol{\epsilon}_n}(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\check{L}_{\boldsymbol{\epsilon}_n}(g) - \bar{L}_{\boldsymbol{\epsilon}_n^*}(g)\right|. \quad (2.13)$$

Whereas the first term on the right hand side of (2.13) can be classically controlled using Vapnik-Chervonenkis and McDiarmid inequalities (see *e.g.* Vapnik (2001)), assertion (i) of Proposition 2.4 provides a control of the second term. Following in the footsteps of Hajek (1964), the third term shall be bounded by means of a *coupling* argument, *i.e.* a specific choice of the joint distribution of $(\boldsymbol{\epsilon}_n^*,\ \boldsymbol{\epsilon}_n)$ satisfying the distributional margin constraints, while the second term is controlled by assumptions related to the closeness between the first order inclusion probabilities $\boldsymbol{\pi}_n^*$ and $\boldsymbol{\pi}_n$. More precisely, the assumptions required in the subsequent analysis involve the total variation distance between the sampling plans $\mathcal{R}_n$ and $\mathcal{R}_n^*$:

$$d_{TV}(R_n,\ R_n^*) \stackrel{def}{=} \frac{1}{2}\sum_{s\in\mathcal{P}(\mathcal{I}_n)} |\mathcal{R}_n(s) - \mathcal{R}_n^*(s)|.$$

Theorem 2.8. *Suppose that Proposition 2.4's assumptions are fulfilled. Set* $\kappa_N^* = (N/n)\min_{i\leq n}\pi_i^*$ *and* $\kappa_N = (N/n)/\min_{i\leq n}\pi_i$. *Then, there exists a universal constant* $C < +\infty$ *such that we have,* $\forall n \geq 1$,

$$\mathbb{E}\left[L(\bar{g}_n^*) - \inf_{g\in\mathcal{G}} L(g)\right] \leq 2\sqrt{2\kappa_N\frac{\log(\frac{4}{\delta}) + V\log(n+1)}{N}} + 4\kappa_N\frac{\log(\frac{4}{\delta}) + V\log(n+1)}{3N} + C\sqrt{\frac{V}{n}} + 2\sqrt{\frac{2\log(\frac{2}{\delta})}{n}} + 2(\kappa_N^* + \kappa_N)(n/N)d_{TV}(\mathcal{R}_n,\ \mathcal{R}_n^*), \quad (2.14)$$

where the infinimum is taken over the set of rejective schemes $\mathcal{R}_n$ *with first order inclusion probabilities* $\boldsymbol{\pi}_n = (\pi_1,\ \ldots,\ \pi_N)$.

The proof is given in section 2.6. The rate bound obtained depends on the minimum error made when approximating the sampling plan by a rejective sampling plan in terms of total variation distance. In practice, following in the footsteps of Hajek (1964) or Berger (1998), it can be controlled by exhibiting a specific coupling $(\boldsymbol{\epsilon}_n^*,\ \boldsymbol{\epsilon}_n)$. One may refer to Berger (1998) for many coupling results of this nature, in particular when the approximating scheme $\boldsymbol{\epsilon}_n$ is of rejective type.

2.4 Illustrative Numerical Experiments

In this section we display numerical experiments to illustrate the relevance of HT risk minimization. We first consider the case where $g(X) = sign(k(X)^T\theta + b)$, where k is some mapping function, T denotes the transposition operator, θ, b are some parameters. As mentioned in 2.1, we consider the hinge loss as a convex surrogate of the $0-1$ loss and add some l_2 regularization term. This leads to the "Weighted SVM" formulation below:

$$\min_{\theta,b}\frac{1}{n}\sum_{i\in S}\frac{1}{\pi_i}\max(0, 1 - Y_i(k(X_i)^T\theta - b)) + \lambda\|\theta\|^2.$$

We use the gaussian r.b.f kernel and perform cross validation to appropriately choose the value of λ. We then consider the task of learning classification trees using the CART algorithm. These classifiers are trained using the scikit-learn library Pedregosa et al. (2011) and, we account for the randomness of our experiments by shuffling our datasets and repeating the experiments 50 times.

We first generate a two class dataset $\mathcal{D}$ in $\mathbb{R}^{10}$ of size 20000 by sampling independent observations from two multivariate normal distribution. A similar dataset $\mathcal{D}_{test}$ of size 2000 is generated to test our classifiers. Denoting by I_d the identity matrix in $\mathbb{R}^d$, the positive class has mean $(0, \ldots, 0)$ and covariance matrix equal to I_{10}, the negative class has mean $(1, \ldots, 1)$ and covariance matrix equal to $10 \times I_{10}$. We then build a dataset $\widetilde{\mathcal{D}}$ of size 1100 via a rejective sampling scheme applied to $\mathcal{D}$. Observations from the negative class being more noisy we assign them first order probability equal to 0.1, and assign first order probability equal to 0.01 to observation from the positive class. To allow for a fair comparison, we also build a dataset $\widehat{\mathcal{D}}$ of size 1100 by sampling without replacement within $\mathcal{D}$. We then learn the different classifiers on $\widetilde{\mathcal{D}}$ and $\widehat{\mathcal{D}}$, and display the results in Table2.1.

	Mean	*Standard deviation*
Weighted SVM on $\widetilde{\mathcal{D}}$	0.02	0.005
Unweighted SVM on $\widetilde{\mathcal{D}}$	0.18	0.02
SVM on $\widehat{\mathcal{D}}$	0.04	0.005
Weighted CART on $\widetilde{\mathcal{D}}$	0.06	0.01
Unweighted CART on $\widetilde{\mathcal{D}}$	0.11	0.03
CART on $\widehat{\mathcal{D}}$	0.08	0.01

TABLE 2.1: Average over 50 runs of the prediction error on $\mathcal{D}_{test}$ and its standard deviation.

Overall, taking into account the inclusion probability allows to consider a training set of reduced size and therefore reduce the computational complexity of the learning procedure without damaging the quality of the prediction.

The same conclusions can be drawn from the analysis of the following datasets which were obtained via a stratified sampling design. We point out that this sampling scheme involves *negatively associated* (binary) random variables so that the theoretical results obtained in this chapter apply to training data sampled by means of this scheme as well.

	incaIndiv	*GJB*	*privacy3*	*privacy4*
N	4079	2001	316	301
Number of features	326	130	95	124

TABLE 2.2: Number of observations and features for our different datasets

The dataset *incaIndiv* [2] contain informations on the food consumption of the french population. The dataset *GJB*[3] contains questions about job seeking and the internet, workforce automation, online dating and smartphone use among Americans. The datasets *privacy3*[4] and

[2] https://https://www.data.gouv.fr/fr/datasets/
[3] http://www.pewinternet.org/datasets/june-10-july-12-2015-gaming-jobs-and-broadband/
[4] http://www.pewinternet.org/datasets/nov-26-2014-jan-3-2015-privacy-panel-3/

privacy4[5] contain questions about privacy and information sharing. On the datasets *incaIndiv* and *incaCompl* we try to predict whether or not someone is an adult, on the dataset *GJB* we will try to learn to predict the gender, and on the datasets *privacy3* and *privacy4* we will predict an answer to some questions among 5 possibilities.

We perform our experiments by randomly splitting the datasets *incaIndiv*, *incaCompl*,*GJB* into a training set (roughly 70 percent of the initial dataset) and a test set. The size of *privacy3* and *privacy4* being much smaller we perform 10-fold cross-validation.

	incaIndiv	*GJB*	*privacy3*	*privacy4*
Weighted SVM	0.16	0.36	0.46	0.48
Unweighted SVM	0.19	0.43	0.50	0.52
Weighted CART	0.04	0.41	0.49	0.54
Unweighted CART	0.05	0.43	0.52	0.57

TABLE 2.3: Average over 50 runs of the prediction error

2.5 Conclusion

Most theoretical studies providing a statistical explanation for the success of learning algorithms based on the ERM paradigm fully ignore the possible impact of the sampling scheme producing the training data and stipulate that observations are independent replications of a generic r.v. or are uniformly sampled without replacement in a larger dataset. Through the generalizable example of rejective sampling, this chapter shows that such studies can be extended to situations where training data are obtained by more general sampling schemes and possibly exhibit a complex dependence structure, provided that related probability weights are appropriately incorporated in the risk functional.

2.6 Technical Proofs

2.6.1 Proof of Theorem 2.3

Considering the usual representation of the distribution of $(\epsilon_1, \ldots, \epsilon_N)$ as the conditional distribution of a sample of independent Bernoulli variables $(\epsilon_1^*, \ldots, \epsilon_n^*)$ conditioned upon the event $\sum_{i=1}^n \epsilon_i^* = N$ (see subsection 2.2.2), the result is a consequence of Theorem 2.8 in Joag-Dev & Proschan (1983).

2.6.2 Bernstein's Inequality for Sums of Negatively Associated Random Variables

For simplicity, we first establish the following tail bound for negatively associated random variables, which extends the usual Bernstein inequality in the i.i.d. setting, see Bernstein (1964) and section A.2.2. Proofs of Proposition 2.4 and Theorem 2.8 are then deduced from

[5] http://www.pewinternet.org/datasets/jan-27-feb-16-2015-privacy-panel-4/

Theorem 2.3 and Theorem 2.9 (see section 2.6). Althought it is not done here, Hoeffding inequality for sums of negatively associated random variables could also be easily derived and we refer to the proofs of A.2.1.2 to establish this result.

Theorem 2.9. *Let $Z_1, \ldots, Z_N$ be negatively associated real valued random variables such that $|Z_i| \leq c < +\infty$ a.s. $\mathbb{E}[Z_i] = 0$ and $\mathbb{E}[Z_i^2] = \sigma_i^2$ for $1 \leq i \leq n$. Then, for all $t > 0$, we have: $\forall n \geq 1$,*

$$\mathbb{P}\left\{\sum_{i=1}^{N} Z_i \geq t\right\} \leq \exp\left(-\frac{t^2}{\frac{2}{3}ct + 2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Before detailing the proof, observe that a similar bound holds true for the tail probability $\mathbb{P}\left(\sum_{i=1}^{N} Z_i \leq -t\right)$ (and for $\mathbb{P}\left(|\sum_{i=1}^{N} Z_i| \geq t\right)$ as well, up to a multiplicative factor 2). Refer also to Theorem 4 in Janson (1994) for a similar result in a more restrictive setting (*i.e.* for tail bounds related to sums of negatively associated r.v.'s).

Proof. Similarly to what we did in section chapter A, the proof starts off with the usual Chernoff method: for all $\lambda > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{N} Z_i \geq t\right\} \leq \exp\left(-t\lambda + \log\mathbb{E}\left[e^{t\sum_{i=1}^{n} Z_i}\right]\right). \tag{2.15}$$

Next, observe that, for all $t > 0$, we have

$$\mathbb{E}\left[e^{t\sum_{i=1}^{n} Z_i}\right] = \mathbb{E}\left[e^{tZ_n} e^{t\sum_{i=1}^{n-1} Z_i}\right] \leq \mathbb{E}\left[e^{tZ_n}\right]\mathbb{E}\left[e^{t\sum_{i=1}^{n-1} Z_i}\right] \leq \prod_{i=1}^{n}\mathbb{E}\left[e^{tZ_i}\right], \tag{2.16}$$

using the property (2.9) combined with a descending recurrence on i. The proof is finished by plugging (2.16) into (2.15), using an adequate control of the log-Laplace transform of the Z_i's throught Hoeffding's Lemma A.7 and optimizing finally the resulting bound w.r.t. $\lambda > 0$, just like in the proof of the classic Bernstein inequality, see chapter A and Bernstein (1964). □

2.7 On Biased HT Risk Minimization

Eq. (2.11) directly results from the following lemma.

Lemma 2.10. *We have, for $p_i's$ such Suppose that $d_n \geq 1$. We have, for all $i \in \{1, \ldots, n\}$,*

$$|1/\pi_i - 1/p_i| \leq \frac{6}{d_n} \times (1-\pi_i)/\pi_i.$$

Proof. The proof follows from the representation (5.14) on p1509 in Hajek (1964). Denote by P_N a Poisson sampling distribution on $\mathcal{I}_n$ with inclusion probabilities $p_1, \ldots, p_n$, the canonical parameters of $\mathcal{R}_n$. For all $i \in \{1, \ldots, n\}$, we have:

$$\begin{aligned}\frac{\pi_i}{p_i}\frac{1-p_i}{1-\pi_i} &= \left(\sum_{s\in\mathcal{P}(\mathcal{I}_n):\, i\in\mathcal{I}_n\setminus\{s\}} P(s)\right)^{-1} \sum_{s\in\mathcal{P}(\mathcal{I}_n):\, i\in\mathcal{I}_n\setminus\{s\}} P(s)\sum_{h\in s}\frac{1-p_h}{\sum_{j\in s}(1-p_j)+(p_h-p_i)} \\ &= \left(\sum_{s:\, i\in\mathcal{I}_n\setminus\{s\}} P_N(s)\right)^{-1} \sum_{s:\, i\in\mathcal{I}_n\setminus\{s\}} P_N(s)\sum_{h\in s}\frac{1-p_h}{\sum_{j\in s}(1-p_j)\left(1+\frac{(p_h-p_i)}{\sum_{j\in s}(1-p_j)}\right)}\end{aligned}$$

Now recall that for any $x \in]-1, 1[$, we have:

$$1 - x \leq \frac{1}{1+x} \leq 1 - x + x^2.$$

It follows that

$$\frac{\pi_i}{p_i}\frac{1-p_i}{1-\pi_i} \leq 1 - \left(\sum_{s:\, i\in\mathcal{I}_n\setminus\{s\}} P(s)\right)^{-1} \sum_{s:\, i\in\mathcal{I}_n\setminus\{s\}} P(s) \sum_{h\in s} \frac{(1-p_h)(p_h-p_i)}{\left(\sum_{j\in s}(1-p_j)\right)^2} + \left(\sum_{s:\, i\in\mathcal{I}_n\setminus\{s\}} P(s)\right)^{-1} \sum_{s:\, i\in\mathcal{I}_n\setminus\{s\}} P(s) \sum_{h\in s} \frac{(1-p_h)(p_h-p_i)^2}{\left(\sum_{j\in s}(1-p_j)\right)^3}$$

Following now line by line the proof on p. 1510 in Hajek (1964) and noticing that $\sum_{j\in s}(1-p_j) \geq 1/2d_n$ (see Lemma 2.2 in Hajek (1964)), we have

$$\left|\sum_{h\in s} \frac{(1-p_h)(p_h-p_i)}{\left(\sum_{j\in s}(1-p_j)\right)^2}\right| \leq \frac{1}{\left(\sum_{j\in s}(1-p_j)\right)} \leq \frac{2}{d_n}$$

and similarly

$$\sum_{h\in s} \frac{(1-p_h)(p_h-p_i)^2}{\left(\sum_{j\in s}(1-p_j)\right)^3} \leq \frac{1}{\left(\sum_{j\in s}(1-p_j)\right)^2} \leq \frac{4}{d_n^2}.$$

This yieds: $\forall i \in \{1, \ldots, n\}$,

$$1 - \frac{2}{d_n} \leq \frac{\pi_i}{p_i}\frac{1-p_i}{1-\pi_i} \leq 1 + \frac{2}{d_n} + \frac{4}{d_n^2}$$

and

$$p_i(1-\pi_i)(1-\frac{2}{d_n}) \leq \pi_i(1-p_i) \leq p_i(1-\pi_i)(1+\frac{2}{d_n}+\frac{4}{d_n^2}),$$

leading then to

$$-\frac{2}{d_n}(1-\pi_i)p_i \leq \pi_i - p_i \leq p_i(1-\pi_i)(\frac{2}{d_n}+\frac{4}{d_n^2})$$

and finally to

$$-\frac{(1-\pi_i)}{\pi_i}\frac{2}{d_n} \leq \frac{1}{p_i} - \frac{1}{\pi_i} \leq \frac{(1-\pi_i)}{\pi_i}(\frac{2}{d_n}+\frac{4}{d_n^2}).$$

Since $1/d_n^2 \leq 1/d_n$ as soon as $d_n \geq 1$, the lemma is proved. □

2.8 Sampling Training Data - Technical Details

2.8.1 Further Details on the Rejective Scheme

Let $N \leq n$ and consider a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ of first order inclusion probabilities. Further define $\mathcal{S}_n := \{s \in \mathcal{P}(\mathcal{I}_n) : \#s = N\}$, the set of all samples in population $\mathcal{I}_n$ with cardinality N. The rejective sampling Hajek (1964); Berger (1998), sometimes called conditional Poisson sampling, exponential design without replacement or maximum entropy

design, is the sampling design R_n that selects samples of fixed size $N(s) = N$ so as to maximize the entropy measure $H(R_n) = -\sum_{s \in \mathcal{S}_n} R_n(s) \log R_n(s)$, subject to the constraint that its vector of first order inclusion probabilities coincides with $\boldsymbol{\pi}$. It is easily implemented in two steps:

1. Draw a sample S according to a Poisson plan P_n, with properly chosen first order inclusion probabilities $\mathbf{p}_n = (p_1, \ldots, p_n)$. The representation is called canonical if $\sum_{i=1}^n p_i = n$. In that case, relationships between each p_i and π_i, $1 \leq i \leq n$, are established in Hajek (1964).

2. If $n(S) \neq n$, then reject sample S and go back to step one, otherwise stop.

Vector $\mathbf{p}$ must be chosen in a way that the resulting first order inclusion probabilities coincide with $\boldsymbol{\pi}$, by means of a dedicated optimization algorithm Tillé (2006). The corresponding probability distribution is given for all $s \in \mathcal{P}(\mathcal{I}_n)$ by $R_n(s) = \frac{P_n(s)\,\mathbb{I}\{\#s=N\}}{\sum_{s' \in \mathcal{S}_n} P_n(s')} \propto \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i) \times \mathbb{I}\{\#s = N\}$, where $\propto$ denotes the proportionality.

2.8.2 Examples of Sampling Plan with Negatively Associated Random Variables

Below we list two examples of sampling scheme involving negatively associated r.v..

2.8.2.1 Stratified Sampling

A stratified sampling design permits to draw a sample S of fixed size $N(S) = N \leq n$ within a population $\mathcal{I}_n$ that can be partitioned into $K \geq 1$ distinct strata $\mathcal{I}_{n_1}, \ldots, \mathcal{I}_{n_K}$ (known a priori) of respective sizes $N_1, \ldots, N_K$ adding up to n. Let $N_1, \ldots, N_K$ be non-negative integers such that $N_1 + \cdots + N_K = N$, then the drawing procedure is implemented in K steps: within each stratum $\mathcal{I}_{n_k}$, $k \in \{1, \ldots, K\}$, perform a SWOR of size $N_k \leq n_k$ yielding a sample S_k. The final sample is obtained by assembling these sub-samples: $S = \bigcup_{k=1}^K S_k$. The probability of drawing a specific sample s by means of this survey design is $R_n^{\text{str}}(s) = \sum_{k=1}^K \binom{N_k}{n_k}^{-1}$. Naturally, first and second order inclusion probabilities depend on the stratum to which each unit belong: for all $i \neq j$ in $\mathcal{U}_n$, $\pi_i(R_n^{\text{str}}) = \sum_{k=1}^K \frac{n_k}{N_k} \mathbb{I}\{i \in \mathcal{U}_{N_k}\}$ and $\pi_{i,j}(R_n^{\text{str}}) = \sum_{k=1}^K \frac{n_k(n_k-1)}{N_k(N_k-1)} \mathbb{I}\{(i,j) \in \mathcal{U}_{N_k}^2\}$.

2.8.2.2 Rao-Sampford Sampling

The Rao-Sampford sampling design generates samples $s \in \mathcal{P}(\mathcal{I}_n)$ of fixed size $N(s) = N$ with respect to some given first order inclusion probabilities $\boldsymbol{\pi}^{RS} := (\pi_1^{RS}, \ldots, \pi_n^{RS})$, fulfilling the condition $\sum_{i=1}^n \pi_i^{RS} = n$, with probability

$$R_n^{RS}(s) = \eta \sum_{i \in s} \pi_i^{RS} \prod_{j \notin s} \frac{\pi_j^{RS}}{1 - \pi_j^{RS}}.$$

Here, $\eta > 0$ is chosen such that $\sum_{s \in \mathcal{P}(\mathcal{I}_n)} R_n^{RS}(s) = 1$. In practice, the following algorithm is often used to implement such a design Berger (1998):

1. select the first unit i with probability π_i^{RS}/n,
2. select the remaining $n-1$ units j with drawing probabilities proportional to $\pi_j^{RS}/(1-\pi_j^{RS})$, $j = 1, \ldots, n$,
3. accept the sample if the units drawn are all distinct, otherwise reject it and go back to step one.

PART II

Sampling strategies for Stochastic Gradient Descent

CHAPTER 3
Adaptive Sampling Scheme for Incremental Optimization using Stochastic Gradient Descent Algorithm

Abstract A wide collection of popular statistical learning methods, ranging from K-means to Support Vector Machines through Neural Networks, can be formulated as a stochastic gradient descent (SGD) algorithm in a specific setup. In practice, the main limitation of this incremental optimization technique is due to the stochastic noise induced by the choice at random of the data involved in the gradient estimator computed at each iteration. In this chapter, we introduce a novel implementation of the SGD algorithm, where the data subset used at a given step is not picked uniformly at random among all possible subsets but drawn from a specific adaptive sampling scheme, depending on the past iterations in a Markovian manner, in order to refine the current statistical estimation of the gradient. Beyond an algorithmic description of the approach we propose, rate bounds are established and illustrative numerical results are displayed in order to provide theoretical and empirical evidence of its statistical performance, compared to more "naive" SGD implementations. Computational issues are also discussed, revealing the practical advantages of the method promoted.

3.1 Introduction

In this chapter, we consider the generic minimization problem

$$\min_{\theta\in\Theta} \widehat{L}_n(\theta) = \min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(Z_i,\theta), \tag{3.1}$$

where Θ is a Euclidean space, typically $\mathbb{R}^d$ with $d \geq 1$, and $l(Z_1,.),\ \ldots,\ l(Z_n,.)$ form a collection of real-valued convex continuously differentiable functions on Θ. Such an optimization problem typically arises in a broad variety of statistical learning problems, in particular supervised tasks, where the goal pursued is to learn a predictive model, fully determined by a parameter θ, in order to predict a random variable Y (the response/output) from an input observation X taking its values in a feature space $\mathcal{X}$. The performance of the predictive function defined by θ is measured by the expectation $L(\theta) = \mathbb{E}[l(\theta;(X,Y))]$, referred to as the *risk*, where l is a *loss function* assumed convex w.r.t. θ. As mentionned before, the distribution (X,Y) being unknown in practice, the risk functional is replaced by its statistical counterpart, the *empirical risk* namely, given by

$$\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l((X_i,Y_i),\theta), \tag{3.2}$$

based on $n \geq 1$ supposedly available independent training examples (X_1,Y_1), ..., (X_n,Y_n), copies of the random pair (X,Y). This supervised problem shall serve as a running example throughout the chapter. The minimization problem (3.2) can be solved incrementally, by

means of variants of the *stochastic approximation* method originally introduced in the seminal contribution of Robbins & Monro (1951). This consists in computing successive estimates of a minimizer of (3.2) using the recursive equation

$$\theta_{t+1} = \theta_t - \gamma_t \widehat{r}_t(\theta_t) \tag{3.3}$$

from a preliminarily picked initial value $\theta_0 \in \Theta$, where $\widehat{r}_t$ denotes an estimate of the gradient $\nabla \widehat{L}_n$ and γ_t is the *learning rate* or *stepsize*. In contrast to the *batch approach*, where all the data are used to estimate the gradient at each iteration (*i.e.* $\widehat{r}_t(\theta) = \nabla \widehat{L}_n(\theta)$ for all $t \geq 0$ and $\theta \in \Theta$), subsets of the data sample only are involved in the gradient estimation steps of *sampled incremental algorithms*, with the aim to reduce computational cost when n is large. In the most commonly used implementation of the *stochastic gradient descent* (SGD) algorithm, the gradient estimate is computed from a subset of reduced size $S \leq n$ uniformly drawn without replacement among all possible subsets of the dataset of size S at each step $t \geq 0$.

In the present chapter, we introduce a specific variant of the SGD algorithm with an *adaptive* sampling scheme, in the sense that it may possibly be different from *sampling without replacement* (SWOR) and vary with t, depending on the past iterations. Rate bounds and limit theorems guaranteeing the theoretical validity of the methodology we propose are established. In addition, the Markovian dynamics governing the evolution of the instrumental sampling distribution is shown to offer crucial advantages regarding computational efficiency. Finally, very encouraging experimental results are displayed, supporting the relevance of our method, in comparison to the usual mini-batch SGD implementation or alternative SGD techniques standing for natural competitors.

The chapter is structured as follows. A short review of the SGD methods documented in the literature, those based on non SWOR sampling schemes in particular, can be found in section 3.2. A description of the specific variant we propose in this chapter is given in section 3.3, together with a detailed discussion about the computational cost inherent to its implementation. The analysis assessing the validity of the estimate output by the algorithm proposed is carried out in section 3.4, whereas illustrative experiments are presented in section 3.5. Finally, some concluding remarks are collected in section 3.6.

3.2 Non Uniform Sampling (NUS) - State of the Art

We start off with a brief review of sampled incremental optimization algorithms, whose archetype is the celebrated SGD algorithm (Robbins & Monro (1951)). Although relevant references in this area are much too numerous to be listed in an exhaustive manner, we point out that significant advances have been recently made in the design of efficient incremental methods, see for instance Mairal (2014), Mairal (2013), Johnson & Zhang (2013a), Shalev-Shwartz & Zhang (2012), Schmidt et al. (2013) or Defazio et al. (2014) that achieve better performances than the traditional SGD method (for instance, by having the variance of the estimator going to 0 as in Johnson & Zhang (2013a), Defazio et al. (2014)). In order to study adaptative sampling scheme, we only considered the classical framework of SGD and did not compare to theses methods. In the original implementation of the SGD algorithms, a single observation (*i.e.* $S = 1$), indexed by i_{t+1} say, is chosen uniformly at random in $\{1, \ldots, n\}$ at each iteration $t + 1$ to form the gradient estimate $\nabla l(Z_{i_{t+1}}, \theta_t)$, its convergence following then from basic stochastic approximation theory. More recently, several papers have shown the possible gain from the use of non-uniform sampling, that is, choosing i_{t+1} according to a non-trivial distribution p well-suited to the specific optimization problem considered: this approach boils down to finding an optimal distribution, in the sense that it minimizes an upper

bound on the convergence rate of the estimator Zhao & Zhang (2014), D.Needell et al. (2014), Clemencon et al. (2014). As far as NUS is concerned, it is natural to ask what relevant choice of the probability distribution must be chosen in order to achieve the smallest expected risk: for instance, the sampling scheme may depend on the Lipschitz constant of the gradient as proposed in D.Needell et al. (2014) or on upper bounds for the norm of the gradient, see Zhao & Zhang (2014). Despite these recent contributions, some questions remain open. 1) In general, usual analyses of NUS algorithms do not fully grasp the impact of the sampling scheme on the performance. For instance, Defazio et al. (2014) establish performance bounds which prove the convergence of NUS scheme for SAGA. Although the attractivity of NUS is demonstrated in the simulations, the bound itself does not fully reveal the performance gain w.r.t. uniform sampling. 2) As opposed to uniform sampling, the choice of an index i_{t+1} according to a non-trivial distribution on $\{1, \ldots, n\}$ is obviously more demanding in terms of computational time. The question of an efficient sampling implementation remains posed. In particular, it is important to quantify the increased complexity caused by NUS. 3) Proposed rules for choosing the sampling distribution p depend on global properties of the functions $l(Z_i, \theta)$ (namely the Lipschitz constants L_i). They do not build upon the amount of information gathered on the optimization problem as the algorithm proceeds. An alternative is to use adaptive sampling, updating the sampling distribution $p = p_t$ at each iteration t in a Markovian fashion, as in the algorithm described below.

As shall be seen in the subsequent analysis, the NUS approach proposed in this chapter has theoretical and practical advantages regarding all these aspects. Whereas rate bound analysis by means of standard tools is poorly informative in general in the present setting, the asymptotic viewpoint developed in this chapter clearly highlights the benefit to using the specific NUS method we promote.

3.3 Adaptive Sampling SGD (AS-SGD)

We now turn to the description of the variant of SGD method considered in this chapter. The main novelty arises from the use of a specific instrumental sampling distribution evolving at each iteration. We also provide some insight into the gain one may expect from such a method and discuss the computational issues related to its implementation. Here and throughout, if $I = (i_1, \ldots, i_S)$ is a S-uplet on $\{1, \ \ldots, \ n\}$ and h is a function on $\{1, \ \ldots, \ n\}$, we use the (slightly abusive) notation $\sum_{i \in I} h(i)$ to represent the sum $\sum_{n=1}^{S} h(i_n)$. For the rest of this chapter, all expectations are taken conditionally upon the observations (i.e the randomness only lies in the sampling strategy) and for any $\theta \in \Theta$, $||\theta||$ denotes its euclidean norm.

3.3.1 The Algorithmic Principle

Let $S \leq n$ be fixed. At each iteration $t \geq 1$, the generic AS-SGD is implemented in three steps as follows:

1. Compute the instrumental probability distribution p_t on $\{1, \ \ldots, \ n\}$ from the information available at iteration t.
2. Form a random sequence of S indexes $I_{t+1} = (i_{t+1}^{(1)}, \ldots, i_{t+1}^{(S)})$ by sampling independently S times according to distribution p_t.

3. Update the estimate using the equation

$$\theta_{t+1} = \Pi_{\mathcal{K}}(\theta_t - \frac{\gamma_t}{S} \sum_{i \in I_{t+1}} \frac{\nabla l(Z_i \theta_t)}{n p_{t,i}}) \,. \tag{3.4}$$

where $\mathcal{K}$ is a compact convex set and $\Pi_{\mathcal{K}}$ is the orthogonal projection on $\mathcal{K}$. Let $\mathcal{F}_t = \sigma(I_1, \ldots, I_t)$ be the σ-algebra generated by the past variables up to time t. Conditioned upon $\mathcal{F}_t$, we have: $i_t^{(1)}, \ldots, i_t^{(S)} \overset{i.i.d.}{\sim} p_t$. Set $p_{t,j} = \mathbb{P}(i_{t+1}^{(1)} = j | \mathcal{F}_t)$ for $j = 1, \ldots, n$. Equipped with these notations, observe that the original SGD algorithm corresponds to the case where $S = 1$ and $p_{t,j} \equiv 1/n$. When $S > 1$, notice that, in contrast to the mini-batch SGD (based on the SWOR scheme), one samples with replacement here and, due to the fact that the $p_{t,j}$'s are not equal in general, it is necessary to normalize the individual gradients by $N p_{t,i}$ in order to guarantee the unbiasedness condition of the increment, which is classically required to ensure proper convergence of the algorithm, see Robbins & Monro (1951), D.Needell et al. (2014), Zhao & Zhang (2014), Clemencon et al. (2014):

$$\mathbb{E}\left(\frac{1}{S} \sum_{i \in I_{t+1}} \frac{\nabla l(Z_i \theta_t)}{n p_{t,i}} \,\middle|\, \mathcal{F}_t \right) = \sum_{j=1}^{n} p_{t,j} \frac{\nabla l(Z_j, \theta_t)}{n p_{t,j}} = \nabla \widehat{L}_n(\theta_t) \,. \tag{3.5}$$

3.3.2 Ideal Sampling Distribution

In order to provide some insight into the specific dynamics we propose to build the successive sampling distributions p_t, we first evaluate a bound on the amount of decrease of the functional. We assume that hypothesis below is fulfilled.

Assumption 1. For all $i \in \{1, \ldots, n\}$, the function $\theta \to l(Z_i, \theta)$ is convex, continuously differentiable and its gradient $\nabla l(Z_i, \theta)$ is L_i-Lipschitz continuous with $L_i < +\infty$.

Let θ_t be the sequence defined by (3.4) and $\mathcal{K} = \Theta$. By virtue of Assumption 1 and Theorem 2.1.5 in Nesterov & Nesterov (2004), we have $\widehat{L}_n(\theta_{t+1}) \leq \widehat{L}_n(\theta_t) + \langle \nabla \widehat{L}_n(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\bar{L}}{2} \|\theta_{t+1} - \theta_t\|^2$ where $\bar{L} = \frac{1}{n} \sum_i L_i$. Taking the conditional expectation, we obtain that

$$\mathbb{E}[\widehat{L}_n(\theta_{t+1}) | \mathcal{F}_t] \leq \widehat{L}_n(\theta_t) - \gamma_t \|\nabla \widehat{L}_n(\theta_t)\|^2 + \frac{\bar{L} \gamma_t^2 \Delta(\theta_t, p_t)}{2n^2},$$

where

$$\begin{aligned} \Delta(\theta_t, p_t) &= \mathbb{E}\left(\left\| \frac{1}{S} \sum_{i \in I_{t+1}} \frac{\nabla l(Z_i, \theta_t)}{p_{t,i}} \right\|^2 \,\middle|\, \mathcal{F}_t \right) \\ &= \frac{1}{S} \sum_{i=1}^{n} \frac{\|\nabla l(Z_i, \theta_t)\|^2}{p_{t,i}} \frac{n^2 (S-1) \|\nabla \widehat{L}_n(\theta_t)\|^2}{S}. \end{aligned}$$

At iteration t, the probability $p_t^* = p_t^*(\theta_t)$ ensuring the steepest descent on the above bound is clearly given by $p_t^* = \arg\min_p \Delta(\theta_t, p)$ or equivalently by $p_{t,i}^* = \|\nabla l(Z_i, \theta_t)\| / \sum_{j=1}^n \|\nabla l(Z_j, \theta_t)\|$ for $i \in \{1, \ldots, n\}$, as mentioned in Zhao & Zhang (2014) and Clemencon et al. (2014). Unfortunately, practical implementation of the above sampling scheme is prohibitively complex, as it would require to evaluate all gradients to calculate the norms $\|\nabla l(Z_1, \theta_t)\|, \ldots, \|\nabla l(Z_n, \theta_t)\|$ at each iteration which is precisely what we try to avoid. The crucial point is therefore to propose a sampling scheme approximating p_t^* without requiring any additional gradient evaluations.

3.3.3 A Practical Sampling Distribution - Our Proposal

The main idea is to replace each unknown gradient norm $\|\nabla l(Z_i, \theta_t)\|$ by a (possibly outdated) norm $g_{t,i} = \|\nabla l(Z_i, \theta_k)\|$ at some former instant $k = k(i,t)$ corresponding to the last time $k \leq t$ when the i-th component was picked. More formally, we define the random sequence g_t as

$$g_{t+1,i} = \begin{cases} \|\nabla l(Z_i, \theta_{t+1})\| & \text{if } i \in \{i_{t+1}^{(1)}, \ldots, i_{t+1}^{(S)}\} \\ g_{t,i} & \text{otherwise.} \end{cases} \tag{3.6}$$

Then, a natural way to approximate p_t^* is to set for each i

$$\bar{p}_{t,i} = \frac{g_{t,i}}{\sum_{j=1}^{n} g_{t,j}}\,. \tag{3.7}$$

It turns out that convergence cannot be guaranteed with the choice (3.7), because a certain component $\bar{p}_{t,i}$ can get arbitrarily close to zero, so that the i-th index is too rarely, or even never, picked[1]. A possible remedy is to enforce a greedy sampling scheme or, as we will refer to it, a Doeblin-like condition on the transition kernel of the underlying Markov chain, see S.Meyn & Tweedie (2009):

$$\forall i \in \{1,\ \ldots,\ n\},\ \ p_{t,i} = \rho\nu_i + (1-\rho)\bar{p}_{t,i}, \tag{3.8}$$

where $\nu = (\nu_1, \ldots, \nu_n)$ is an arbitrary probability distribution satisfying $\nu_i > 0$ for $1 \leq i \leq n$, and $0 < \rho \leq 1$. This condition has the following interpretation: p_t is a mixture between two laws of probability and one of this law is independent from the past. The AS-SGD is summarized in Algorithm 1 below.

Algorithm 1 AS-SGD

Input: θ_0, ρ, T, S, $(\gamma_t)_{t=0}^{T-1}$, ν
Initialization:
for $i = 1$ **to** n **do**
  Set $g_{0,t} = \|\nabla l(Z_i, \theta_0)\|$
end for
$\mathcal{A}_0$ = `buildtree`(g_0)
for $t = 0$ **to** $T-1$ **do**
  Define $\bar{p}_t$ and p_t according to (3.7) and (3.8)
  $I_{t+1} =$ `sample`$(\bar{p}_t, S, \rho, \nu, \mathcal{A}_t)$
  $\theta_{t+1} = \Pi_{\mathcal{K}}(\theta_t - \frac{\gamma_t}{nS}\sum_{i \in I_{t+1}} \frac{\nabla l(Z_i, \theta_t)}{p_{t,i}})$
  Update g_{t+1} according to (3.6)
  $\mathcal{A}_{t+1} =$ `updatetree`$(\mathcal{A}_t, I_{t+1}, g_{t+1})$
end for
Return θ_T

3.3.4 Computationally Efficient Sampling

We point out that there is an additional computational price to pay when implementing NUS instead of uniform sampling. Given a non uniform distribution $p = (p_1, ..., p_n)$ on $\{1,\ \ldots,\ n\}$, the simulation time needed to generate a r.v. with distribution p is larger than

[1] Consider for instance the case $n = 2$, $\mathcal{X} = \mathbb{R}$, $l(Z_1, \theta) = \theta^2$, $l(Z_2, \theta) = (\theta - 1)^2$ and $\theta_0 = 0$.

in the case of uniform distribution. Indeed, one may resort to the inversion method (see Devroye (1986) for instance), which boils down to inserting an element in the sorted vector $\tilde{p} = (p_1, p_1 + p_2, ..., p_1 + ...p_n)$ and requires $\lceil \log_2(n) \rceil$ operations. Unfortunately, changing the i-th component of p (just like in the algorithm we propose) changes $n - i$ components in the vector $\tilde{p}$. Our approach based on the notion of *binary research tree* is inspired from Devroye (1986) and overcome this difficulty.

Building / updating a tree. For simplicity, assume that n is even. Define a tree $\mathcal{A}_t$ with n terminal leaves in correspondence with the indexes in $\{1, \ldots, n\}$ the weight $g_{t,i}$ is assigned to the leaf No. i. Each pair $(2k+1, 2(k+1))$ of adjacent terminal leaves, $k \in \{0, \ldots, n/2\}$, admits a common ancestor, which the weight $g_{t,2k+1} + g_{t,2(k+1)}$ is assigned to. Continuing this way in a "bottom-up" fashion, the weight $S_t = \sum_i g_{t,i}$ is assigned to the root node. The function `buildtree` used in Algorithm 1 generates such a tree from scratch and is used at the initialization $t = 0$. At step $t+1$, g_{t+1} is essentially identical to g_t except for a few S elements which have been updated. The tree $\mathcal{A}_{t+1}$ being close to $\mathcal{A}_t$, it does *not* have to be rebuilt from scratch. The routine `updatetree` given in section 3.7 provides a computationally efficient way to update the tree $\mathcal{A}_{t+1}$ from $\mathcal{A}_t$.

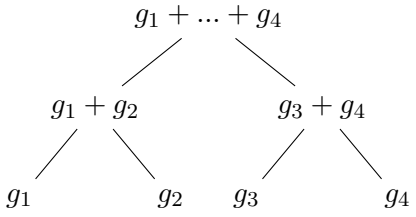


FIGURE 3.1: A binary tree for $n = 4$.

Simulating a random index. Suppose that we seek to generate a r.v., say $i_{t+1}^{(1)}$, according to distribution $\bar{p}_t$. Using a r.v. U generated according to the uniform distribution on $[0, 1]$, a path is generated from the root to one of the leaves by comparing U to successive thresholds. The generated variable $i_{t+1}^{(1)}$ is defined as the index of the obtained leaf. The procedure is detailed in the subroutines `sample` and `sample_tree` (see section 3.7). Therefore updating an element i of the distribution is equivalent to update the path from i to the root of the tree and takes $\lceil \log_2(n) \rceil$ operations, while sampling from our distribution is equivalent to follow a path from the root of the tree to one of its leaves. Table 3.1 summarizes the iteration complexity of the proposed method.

	SGD	$AS-SGD$
Complexity	Sc	$S(c + (2-\rho)\lceil \log_2(n) \rceil)$

TABLE 3.1: Comparison of iteration complexities of AS-SGD and SGD with uniform sampling: c = complexity of pointwise gradient computation, S = sample size.

3.4 Performance Analysis

We recall the transpose of a matrix A is denoted by A^T. We start off by giving a rate bound analysis of the algorithm we proposed and then study the asymptotic behavior of the estimator θ_t produced by our algorithms.

3.4.1 Preliminary results

In our analysis, we assume that

Assumption 2. $i)$ The function $\theta \to \widehat{L}_n(\theta)$ is α-strongly convex, $ii)$ The minimizer θ_n^* of $\widehat{L}_n$ belongs to the interior of $\mathcal{K}$.

The lemma stated below provides a bound on the Mean Square Error $a_t = \mathbb{E}(\|\theta_t - \theta_n^*\|^2)$, where θ_t is generated by Algorithm 1. Its proof is strongly inspired by Bach & Moulines (2011a) and A.Nemirovski et al. (2009), where similar bounds are provided.

Lemma 3.1. *Let Assumptions 1 and 2 hold true. Set $\gamma_t = \gamma_1 t^{-\beta}$ where $\beta \in (0, 1]$ and assume $\gamma_1 > \beta/(2\alpha)$. For each $t \in \mathbb{N}^*$,*

$$a_t \leq C\gamma_t/\rho, \tag{3.9}$$

where $C = \max(\frac{2B_\nu^2\gamma_1}{2\alpha\gamma_1 - 1}, \frac{a_1}{\gamma_1})$ when $\beta = 1$ and $C = \max(\frac{B_\nu^2\gamma_1}{2\alpha}, \frac{a_1}{\gamma_1})$ otherwise, with

$$B_\nu = \frac{1}{Sn^2}\sum_{i=1}^{n} \nu_i^{-1} \sup_{\theta\in\mathcal{K}} \|\nabla l(Z_i, \theta)\|^2 .$$

Remark 3.2. One mights easily check that choosing ν so as to minimize B_ν would lead to take $\nu_i \propto \sup_{\theta\in\mathcal{K}} \|\nabla l(Z_i, \theta)\|$ which is the sampling distribution proposed in Zhao & Zhang (2014).

We emphasize the fact that sharper bounds on a_t can be obtained. One could for instance easily generalize the approach of Bach & Moulines (2011a) which provides bounds on a_t which are not only tighter but also valid under weaker assumptions on the step size. This would come at the price of a rather tedious bound in (3.9). As explained below, such an involved bound would actually be unnecessary for our purpose, and Lemma 3.1 is in fact sufficient to derive the main result of the next paragraph. It also admits a simpler proof provided in section 3.8.1. Before skipping to the main result, we first discuss the bound in (3.9). Lemma 3.1 establishes that the adaptive sampling scheme preserves the convergence rate in $O(t^{-\beta})$ obtained in the uniform sampling case. Nevertheless, Lemma 3.1 is merely a sanity check, because unfortunately, the bound does **not** suggest that adaptive non-uniform sampling generates a performance improvement: the minimum of the right hand side in (3.9) is attained for $\rho = 1$ which boils down to select a constant sampling probability. This is in contradiction with numerical results, which suggest on the opposite that strong benefits can be obtained by adaptively selecting the sampling probability. In order to obtain results that fully grasp the benefits of our adaptive sampling strategy, we investigate from now on the asymptotic regime $t \to \infty$. The following Lemma provides an estimate of the value $b_{t,i} = \mathbb{E}[\|g_t^i - \nabla l(Z_i, \theta_{t-1})\|^2]$, which quantifies the mean square gap between the current (unobserved) gradients and the outdated gradients used to generate the next samples.

Lemma 3.3. *Suppose that the assumptions of Lemma 3.1 hold true. For any $t \in \mathbb{N}^*$,*

$$b_{t,i} \leqslant \frac{(2L_i)^2 2^\beta}{1 - (1 - \rho\nu_i)^S} \frac{C}{\rho t^\beta} + o(t^{-\beta}) . \tag{3.10}$$

Lemma 3.3 upper-bounds the gap between our approximation of the gradient and its true value. The bound obtained depend on the batch-size and the step-size. We now state in the following section a Theorem Central Limit (TCL) on $\theta_t - \theta_n^*$. The covariance matrix involved in the TCL depends on the sampling distribution and we show that its norm is minimal with the non uniform sampling strategy we propose.

3.4.2 Main results

We now turn to the analysis of the asymptotic behaviour and prove how our algorithm improve on SGD.

Assumption 3. The function $\widehat{L}_n$ is twice differentiable in a neighborhood of θ_n^*.

We introduce the sampling probability given by $\pi^* = \rho\nu + (1-\rho)\bar{\pi}^*$, where,

$$\bar{\pi}_i^* = \frac{\|\nabla l(Z_i, \theta_n^*)\|}{\sum_{j=1}^n \|\nabla l(Z_j, \theta_n^*)\|} \tag{3.11}$$

for any $i = 1, \ldots, n$. We define $Q^* = \sum_{i=1}^n \nabla l(Z_i, \theta_n^*) \nabla l(Z_i, \theta_n^*)^T / (Sn^2 \pi_i^*)$ and denote by $H = \nabla^2 \widehat{L}_n(\theta_n^*)$ the Hessian at point θ^*.

Theorem 3.4. *Suppose that Assumptions 1, 2 and 3 hold true and that the stepsize satisfies the condition of Lemma 3.1. Then the sequence $(\theta_t - \theta_n^*)/\sqrt{\gamma_t}$ converges in distribution to a zero-mean Gaussian variable whose covariance matrix $\Sigma = \Sigma(\rho, \nu)$ is the solution to the following Lyapunov equation*

$$\begin{aligned} &\Sigma H + H\Sigma = Q^* \quad (\textit{if } \beta < 1) \\ &\Sigma(I_d + 2\gamma_1 H) + (I_d + 2\gamma_1 H)\Sigma = 2\gamma_1 Q^* \quad (\textit{if } \beta = 1)\,. \end{aligned}$$

The proof is provided in section 3.8.3. The following Corollary is directly obtained by use of the second order delta-method Pelletier (1998). We denote by $\mathrm{Tr}(A)$ the trace of any square matrix A.

Corollary 3.5. *Under the assumptions of Theorem 3.4, $\gamma_t^{-1}(\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta^*))$ converges in distribution to the r.v. $V = (1/2) Z^T \Sigma(\rho,\nu)^{1/2} H \Sigma(\rho,\nu)^{1/2} Z$ where Z is a Gaussian vector $\mathcal{N}(0, I_d)$. In addition, we have $\mathbb{E}(V) = \mathrm{tr}(H\Sigma(\rho,\nu))/2$.*

We now use Corollary 3.5 to compare our method with the best possible *fixed* choice of a sampling distribution. Note that the search for an optimal fixed distribution is also discussed in Clemencon et al. (2014).

When the distribution is fixed, say to p, the asymptotic covariance of the normalized error is given by $\Sigma(1, p)$ as defined in Theorem 3.4. Motivated by Corollary 3.5, we refer to the optimal fixed sampling distribution as the distribution p minimizing $\mathrm{tr}(H\Sigma(1,p))$. It is straightforward to show that

$$\arg\min_p \mathrm{tr}(H\Sigma(1,p)) = \bar{\pi}^*,$$

where $\bar{\pi}^*$ is defined in (3.11). We also set $\sigma_*^2 = \mathrm{tr}(H\Sigma(1,\bar{\pi}^*))$. The following proposition follows from standard algebra and its proof is omitted due to the lack of space.

Proposition 1. Let $\Sigma(\rho,\nu)$ be the asymptotic covariance matrix defined in Theorem 3.4. Then,

$$\sigma_*^2 \leq \mathrm{tr}(H\Sigma(\rho,\nu)) \leq \sigma_*^2(1 + S\rho/(1-\rho)).$$

Proposition 1 implies that the asymptotic performance of the proposed AS-SGA can be made arbitrarily closed to the one associated with the best sampling distribution provided that ρ is chosen closed to zero. It is of course tempting to set $\rho = 0$ in (3.8) however in this case, the statement of Theorem 3.4 would be no longer valid.

3.5 Numerical experiments

We consider the l_2-regularized logistic regression problem. Denoting by n the number of observations and by d the number of features, the optimization problem can be formulated as follows:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f(y_i, x_i, \theta) + \frac{\lambda}{2} \|\theta\|^2, \tag{3.12}$$

where $f(x, y, \theta) = \log(1 + \exp(-yx^T\theta)$ the $(y_i)_{i=1}^n$ are in $\{-1, +1\}$, the $(x_i)_{i=1}^n$ are in $\mathbb{R}^d$ and $\lambda > 0$ is a scalar. Note that for this problem one mights easily have access to the quantities L_i and $B_i = \sup_{\theta \in \mathbb{R}^d} \|\nabla l(Z_i, \theta)\|$ .We used the benchmark dataset *covtype* with $n = 581012$, $d = 54$, $\lambda = \frac{1}{\sqrt{n}}$ and $\gamma_t = \frac{\gamma_1}{1+\gamma_1 \lambda t}$ as proposed in L.Bottou (2012) , where γ_1 is determined using a small sample of the training set. We considered the cases $\nu_i = \frac{1}{n}$ (ASGD), $\nu_i \sim L_i$ (ASGD-Lip) and $\nu_i \sim B_i$ (ASG-B), ran the algorithm for different values of the parameter ρ and compared it to the usual stochastic gradient descent with uniform sampling (SGD), lipschitz sampling (SGD-Lip) and upper-bound sampling (SGD-B) for the same parameters. In this scenario, $\mathcal{C} \sim d$, the computational times related to the SGD and the AS-SGD are comparable (see Table 1).

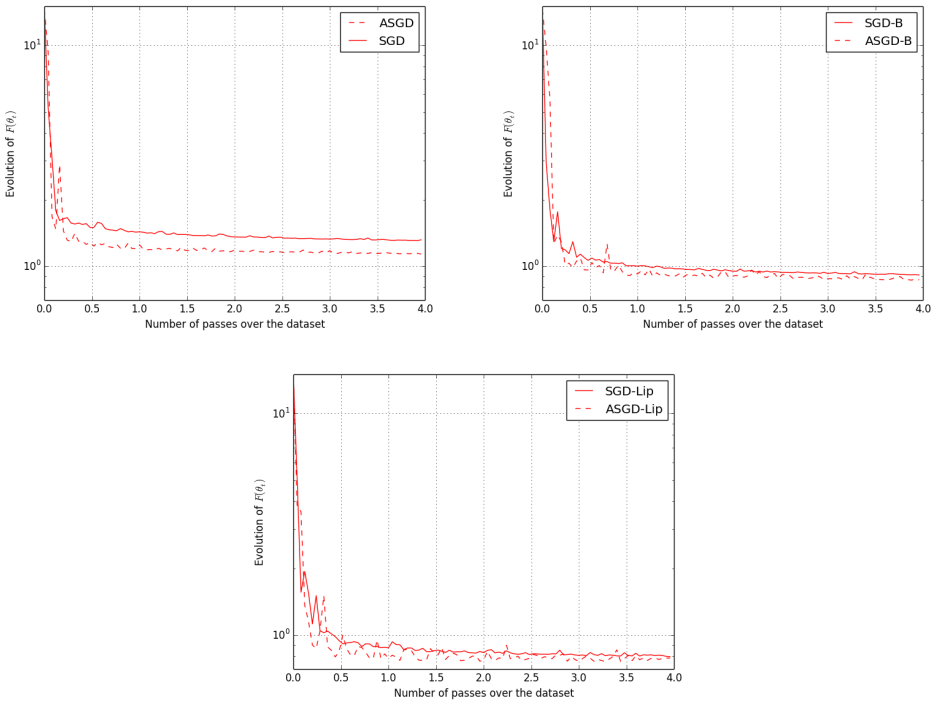


FIGURE 3.2: Evolution of $\widehat{L}_n(\theta_t)$ with $S = 10$ and $\rho = 0.7$

Experiments suggest that choosing $\nu_i \sim L_i$ leads to better performances and that using a small value of ρ leads to poor (respectively good) performance when θ_1 is far (respectively close) from θ_n^*. This suggests that a strategy could consist in running a classical SGD to get closer to θ_n^* and then run AS-SGD. One could also use the AS-SGD with a decreasing step-size policy. We dit not study these policies due to space limitations.

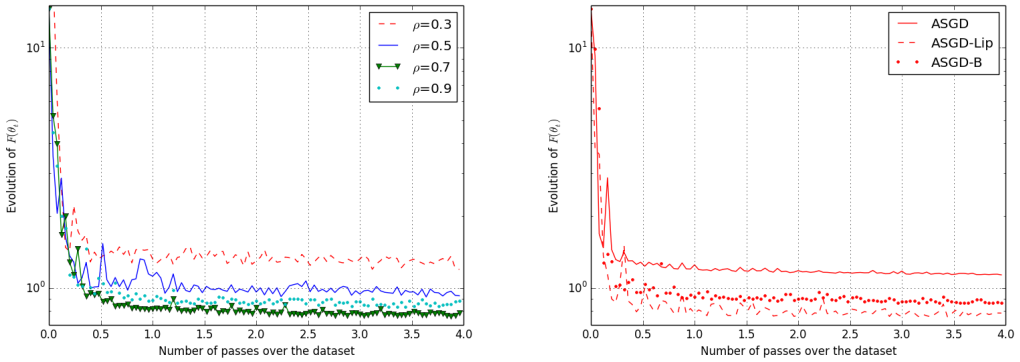


FIGURE 3.3: Evolution of $\widehat{L}_n(\theta_t)$ with different values of ρ ($\nu_i \sim L_i$, $S = 10$) (left) and different sampling strategies ($\rho = 0.7, S = 10$) (right)

3.6 Conclusion

Motivated by recent work on SGD with non uniform probability, we introduced a novel implementation of the SGD algorithm with an adaptative sampling. We proposed a specific adaptive sampling scheme, depending on the past iterations in a Markovian manner that achieves low simulation cost. We also proposed a rigorous analysis to justify our approach and gave sufficient conditions to obtain theoretical guarantees.

3.7 Algorithms for Efficient NUS

In this section, we provide the main procedures called by Algorithm 1 to efficiently generate a collection of S i.i.d random indexes on $\{1, \dots, n\}$. If $\mathcal{A}$ is a tree with n leaves and e is a node, we denote by $w(e)$ the weight of node e. The root of $\mathcal{A}$ is denoted by root(e). Father(e) is the father of a node e (and the empty set if $e = root(\mathcal{A})$. Son(e) is the list of sons of e (and the emptyset if e is a leaf) and the elements of the list are referred to as son$(e)[1]$, son$(e)[2]$. The functions isroot(e) and isleaf(e) return boolean values equal to one if e is the root or a leaf respectively. Finally, if e is a leaf, label(e) returns the index of the leaf e in $\{1, \dots, n\}$. The procedure `buildtree` is omitted but discussed in section 3.3.4. The algorithm `Sample` simply consists in writing the probability distribution as a mixture and is also omitted.

Algorithm 2 Sample_tree

Input: $\mathcal{A}$
$e =$root$(\mathcal{A})$
Draw $U \sim$ Uniform$([0, w(e)])$
repeat
  if $U < w($son$(e)[1])$ **then**
    $e \leftarrow$ son$(e)[1]$)
  else
    $U \leftarrow U - w($son$(e)[1])$
    $e \leftarrow$ son$(e)[2]$
  end if
until isleaf(e)
Return label(e)

Algorithm 3 Update_tree

Input: $\mathcal{A}$, I, g
for $i \in I$ **do**
  $e =$leaf$(\mathcal{A}, i)$
  $\delta = g_i - w(e)$
  $w(e) \leftarrow g_i$
  repeat
    $e \leftarrow$father(e)
    $w(e) \leftarrow w(e) + \delta$
  until isroot(e)
end for
Return $\mathcal{A}$

3.8 Technical Proofs

3.8.1 Proof of Lemma 3.1

Using the non expansiveness of the projection, the strong convexity and the definition of B_ν, we get $a_{t+1} \leqslant (1-2\alpha\gamma_t)a_t + \gamma_t^2 \frac{B_\nu^2}{\rho}$. We will now prove the lemma by induction. The property is checked for $t = 1$ by definition of C. Assume the result holds true for a_t then, we have $a_{t+1} \leqslant ((1-2\alpha\gamma_t)\gamma_t C + \gamma_t^2 B^2)/\rho$ and it is sufficient to show that $(1-2\alpha\gamma_t)\gamma_t C + \gamma_t^2 B^2 \leqslant \gamma_{t+1} C$ which is equivalent to $\gamma_t^2 B_\nu^2 \leqslant C(\gamma_{t+1} - \gamma_t + 2\alpha\gamma_t^2)$. If $\beta = 1$, then, using $2\alpha\gamma_1 > 1$, we get: $\gamma_{t+1} - \gamma_t + 2\alpha\gamma_t^2 \geqslant \frac{(2\alpha\gamma_1 - 1)\gamma_1}{t(t+1))} > 0$ and therefore

$$\frac{B_\nu^2 \gamma_t^2}{\gamma_{t+1} - \gamma_t + 2\alpha\gamma_t^2} \leqslant \frac{B_\nu^2 \gamma_1}{t^2} \frac{t(t+1)}{(2\alpha\gamma_1 - 1)} \leqslant \frac{2B_\nu^2 \gamma_1}{2\alpha\gamma_1 - 1}, \tag{3.13}$$

which gives the result. If $0 < \beta < 1$,

$$\begin{aligned}\gamma_{t+1} - \gamma_t + 2\alpha\gamma_t^2 &= \gamma_1 \int_t^{t+1} \frac{-\beta}{u^{1+\beta}} + 2\alpha\frac{\gamma_1^2}{t^{2\beta}} \geqslant 2\alpha\frac{\gamma_1^2}{t^{2\beta}} - \frac{\gamma_1\beta}{t^{\beta+1}} \\ &\geqslant \frac{\gamma_1(2\alpha\gamma_1 t^{1-\beta} - \beta)}{t^{\beta+1}} > 0\end{aligned}$$

since $2\alpha\gamma_1 > \beta$. We get

$$\frac{B^2\gamma_t^2}{\gamma_{t+1} - \gamma_t + 2\alpha\gamma_t^2} \leqslant \frac{B^2\gamma_1}{t^{2\beta}} \frac{t^{\beta+1}}{(2\alpha\gamma_1 t^{1-\beta} - \beta))} < \frac{B^2}{2\alpha},$$

which concludes the proof.

3.8.2 Proof of Lemma 3.3

We consider a stepsize γ_t such that

$$a_t \leqslant \frac{C\gamma_{t+1}}{\rho}, \tag{3.14}$$

and we denote by $\mathbb{I}_A$ the indicator function of any event A *i.e.*, the r.v. equal to one on this event and to 0 elsewhere. Consider any index i and instant t and let $A_{k,t}^i$ the event *"the index i has not been picked since instant k"*. Using the Doeblin Condition we have:

$$\begin{aligned}\mathbb{P}(A_{k,t}^i) &= \mathbb{E}[\mathbb{I}_{\epsilon_t^i=0}...\mathbb{I}_{\epsilon_{k+1}^i=0}\mathbb{I}_{\epsilon_k^i\neq 0}] \leqslant \mathbb{E}[\mathbb{E}[\mathbb{I}_{\epsilon_t^i=0}|\mathcal{F}_{t-1}]...\mathbb{I}_{\epsilon_{k+1}^i=0}] \\ &\leqslant (1-\rho\nu_i)^S \mathbb{E}[\mathbb{I}_{\epsilon_{t-1}^i=0}...\mathbb{I}_{\epsilon_{k+1}^i=0}] \leqslant (1-\rho\nu_i)^{S(t-k)}\end{aligned}$$

Conditionally upon $A_{k,t}^i$, we have $g_{t,i} = \nabla l(Z_i, \theta_{k-1})$. Since $(A_{k,t}^i)_{k\leqslant t}$ is a partition of the state space, the law of total probability combined with Assumption 1 and the independence induced by the Doeblin Condition yields: $\mathbb{E}[\|g_{t,i} - \nabla l(Z_i, \theta_{t-1})\|^2] \leqslant L_i^2 \sum_{k=1}^t \mathbb{E}[\|\theta_{k-1} -$

$\theta_{t-1}\|^2](1-\rho\nu_i)^{S(t-k)}$. Using $\mathbb{E}[\|\theta_{k-1}-\theta_{t-1}\|^2] \leqslant 2a_{t-1}+2a_{k-1} \leqslant 4C\gamma_k/\rho$ leads to:

$$\begin{aligned}\mathbb{E}[\|g_{t,i}-\nabla l(Z_i,\theta_{t-1})\|^2] &\leqslant 4(L_i)^2\frac{C}{\rho}\sum_{k=1}^{t}\gamma_k(1-\rho\nu_i)^{S(t-k)} \\ &= 4(L_i)^2\frac{C}{\rho}\sum_{k=0}^{t-1}\gamma_{t-k}(1-\rho\nu_i)^{Sk}.\end{aligned}$$

For all $1 < t_0 < t$, we have by splitting the sum in two terms:

$$\sum_{k=0}^{t-1}\gamma_{t-k}(1-\rho\nu_i)^{Sk} \leqslant \frac{\gamma_{t-t_0}}{1-(1-\rho\nu_i)^S} + (1-\rho\nu_i)^{St_0}\sum_{k=t_0}^{t-1}\gamma_{t-k}.$$

Taking $2t_0 \sim t$ for instance and using the classical integral test for convergence gives the result.

3.8.3 Proof of Theorem 3.4

The proof is prefaced by the following Lemma, whose proof is given at the end of this section.

Lemma 3.6. *Under the Assumptions of Theorem 3.4, the sequence* (θ_t, p_t) *converges to* (θ_n^*, π^*) *with probability one.*

We now prove the main result. We use the decomposition $\theta_{t+1} = \theta_t - \gamma_t\nabla\widehat{L}_n(\theta_t) + \gamma_t e_{t+1} + \gamma_t\eta_{t+1}$, where we set $D_{t+1} = (1/S)\sum_{i\in I_{t+1}}\nabla l(Z_i,\theta_t)/(np_{t,i})$, $e_{t+1} = \nabla\widehat{L}_n(\theta_t) - D_{t+1}$, $\eta_{t+1} = (\Pi_{\mathcal{K}}(\phi_{t+1}) - \phi_{t+1})/\gamma_t$, $\phi_{t+1} = \theta_t - \gamma_t D_{t+1}$.

We next check Conditions **C1** to **C4** in G.Fort (2014). Conditions **C1** and **C4** are immediate consequences of Assumptions 2 and 3. We check that e_t satisfies Condition **C2**. First, by virtue of(3.5), (e_t) is a martingale increment sequence adapted to $\mathcal{F}_t$ *i.e.*, $\mathbb{E}(e_{t+1}|\mathcal{F}_t) = 0$. Second, for all $t \in \mathbb{N}^*$ and $i = 1,\ldots,n$, $p_{t,i} \geq \rho\nu_i > 0$ with probability one. Therefore, it is straightforward to check that $\|e_t\| \leq M$ a.s. for some constant M, which only depends on ρ, ν and $B_1,\ldots,B_N$. Third, we analyze the asymptotic behaviour of the conditional covariance $Q_t = \mathbb{E}(e_{t+1}e_{t+1}^T|\mathcal{F}_t)$. After some algebra, we obtain that $Q_t = (1/n^2)\sum_{i=1}^n(1/(Sp_{t,i}))(\nabla l(Z_i,\theta_t) - \nabla\widehat{L}_n(\theta_t))(\nabla l(Z_i,\theta_t) - \nabla\widehat{L}_n(\theta_t))^T$. Using Lemma 3.6 along with the continuity of $\nabla l(Z_i,\theta)$ for each i, we directly obtain that Q_t tends to Q^* with probability one. Condition **C2** in G.Fort (2014) is thus fulfilled. Turning to Condition **C3**, we shall prove that $\gamma_t^{-1/2}\eta_{t+1}$ converges to 0 in L_1. Using Cauchy-Schwarz inequality:

$$\begin{aligned}\mathbb{E}(\|\gamma_t^{-1/2}\eta_{t+1}\|) &= \mathbb{E}(\|\gamma_t^{-1/2}\eta_{t+1}\|\mathbb{I}_{\eta_{t+1}\neq 0}) \\ &\leq \mathbb{E}(\|\eta_{t+1}\|^2)^{1/2}\left(\gamma_t^{-1}\mathbb{P}(\eta_{t+1}\neq 0)\right)^{1/2}.\end{aligned}$$

Defining $u_t = \mathbb{E}(\|\eta_{t+1}\|^2)$ and $v_t = \mathbb{P}(\eta_{t+1}\neq 0)$, the above inequality reads

$$\mathbb{E}(\|\gamma_t^{-1/2}\eta_{t+1}\|) \leq \sqrt{u_t}\sqrt{\frac{v_t}{\gamma_t}}. \tag{3.15}$$

We first analyze u_t. Observe that

$$\begin{aligned}\|\eta_{t+1}\|^2 &= \gamma_t^{-2}\|\Pi_{\mathcal{K}}(\phi_{t+1}) - \phi_{t+1}\|^2 \leq 2\gamma_t^{-2}(\|\Pi_{\mathcal{K}}(\phi_{t+1}) - \theta_t\|^2 + \|\phi_{t+1}-\theta_t\|^2) \\ &\leq 4\gamma_t^{-2}\|\phi_{t+1}-\theta_t\|^2 = 4\|D_{t+1}\|^2,\end{aligned}$$

where the last inequality is due to the non-expansiveness of the projection operator. Therefore Since $p_{t,i}$ admits a fixed deterministic lower-bound for each i and since the gradients $\nabla l(Z_i, \theta)$ are bounded on $\mathcal{K}$, there exists a deterministic constant M' such that $\|D_{t+1}\|^2 \leq M'$. In particular, the sequence $\|\eta_{t+1}\|^2$ is uniformly integrable. We now prove that $\eta_{t+1} \to 0$ almost surely. Consider an $\epsilon > 0$ such that the ball $B(\theta^*, 2\epsilon)$ of center θ^* and radius ϵ is contained in $\mathcal{K}$. By Lemma 3.6, for all ω on a set of probability one, there exists $N(\omega)$ such that $\|\theta_{t+1}(\omega) - \theta^*\| \leq \epsilon$ for all $t \geq N(\omega)$. Using again that $\|D_{t+1}(\omega)\|$ is a bounded sequence and $\gamma_t \to 0$, it is clear that $\gamma_t \|D_{t+1}(\omega)\| < \epsilon$ for t large enough. Thus, for t large enough, $\phi_{t+1}(\omega) \in \mathcal{K}$ which implies that $\eta_{t+1}(\omega) = 0$. Almost surely, the sequence η_{t+1} converges to zero. Putting all pieces together, $\|\eta_{t+1}\|^2$ is a uniformly integrable sequence which tends a.s. to zero. As a consequence, $u_t = \mathbb{E}(\|\eta_{t+1}\|^2)$ tends to zero as $t \to \infty$. We now analyze $v_t = \mathbb{P}(\eta_{t+1} \neq 0)$. For $\epsilon > 0$ be defined as above, note that the event $\{\eta_{t+1} \neq 0\}$ is included in the event $\{\|\phi_{t+1} - \theta^*\| \geq \epsilon\}$. By Markov inequality,

$$\begin{aligned} v_t &\leq \epsilon^{-2}\mathbb{E}(\|\phi_{t+1} - \theta^*\|^2) \leq 2\epsilon^{-2}(\mathbb{E}(\|\phi_{t+1} - \theta_t\|^2) + \mathbb{E}(\|\theta_t - \theta^*\|^2)) \\ &\leq 2\epsilon^{-2}\left(\gamma_t^2 \mathbb{E}(\|D_{t+1}\|^2) + \frac{C\gamma_t}{\rho}\right), \end{aligned}$$

where we used Lemma 3.1 to obtain the last inequality. Recalling that D_{t+1} is bounded, it is clear that v_t/γ_t is a bounded sequence. Finally, by inequality (3.15), we conclude that $\mathbb{E}(\|\gamma_t^{-1/2}\eta_{t+1}\|)$ tends to zero. Condition **C3** in G.Fort (2014) is satisfied. This completes the proof of Theorem 3.4.

Proof of Lemma 3.6

Almost sure convergence of θ_t to θ_n^* directly follows from the Robbins-Siegmund Lemma and follows the same line of reasoning than L.Bottou (1998): it is therefore omitted. It remains to show that for each $j = 1, \ldots, n$, $\bar{p}_{t,j} \to \bar{\pi}_j^*$ a.s., where $\bar{\pi}_j^* = \|\nabla l(Z_j, \theta_n^*)\| / \sum_j \|\nabla l(Z_j, \theta_n^*)\|$. Let A_j denote the event that index j is picked infinitely often (*i.e.*, there exists an infinite sequence $(t_k, n_k)_{k\in\mathbb{N}}$ on $\mathbb{N}^* \times \{1, \ldots, S\}$ such that $i_{t_k}^{(n_k)} = j$). As can be easily checked, the Doeblin condition $p_{t,j} \geq \rho\nu_i$ ensures that A_j has probability one. For a fixed $\omega \in A_j$ and using the continuity of $\theta \to \nabla l(Z_j, \theta)$, we obtain that $g_{t,j}(\omega)$ converges to $\|\nabla l(Z_j, \theta_n^*)\|$. As a consequence, $\bar{p}_{t,j}(\omega) \to \bar{\pi}_j^*(\omega)$ and the result follows.

CHAPTER 4
Horvitz Thompson Stochastic Gradient Descent : Application to M-estimation

Abstract Building upon the results of chapter 2 and 3, we propose to incorporate survey schemes into the SGD Algorithm. In the M-estimation context, we establish asymptotic results for the estimator produced, highlighting the trade-off between statistical and optimization accuracy (see Bottou & Bousquet (2008)) in large scale learning. This chapter extends the result establish in the previous chapter by studying the limit behaviour of the estimator produced as the number of observation goes to infinity and by taking into account the distribution of the observations.

4.1 Introduction

In many situations, data are collected by means of a survey technique and the related weights (the true inclusion probabilities of the individual units forming the statistical population of interest) must be used by the statistician to compute unbiased statistics (see chapter 2). The availability of massive information in the Big Data era, which statistical procedures could theoretically now rely on, has motivated the recent development of *parallelized/distributed* variants of certain inference techniques or statistical learning algorithms, see Bekkerman et al. (2011), Mateos et al. (2010), Navia-Vazquez et al. (2006) or Bianchi et al. (2013) among others. It also strongly suggests to use sampling techniques as we did in the previous chapter, as a remedy to the apparent intractability of learning from datasets of explosive size, in order to break the current computational barriers, see Clémençon et al. (2013) or Clémençon et al. (2016). It is the purpose of the present chapter to explore this approach further, by showing how to incorporate efficiently survey schemes into the SGD algorithm and highlight how it affects the statistical performance of the estimator produced. More precisely, the variant of the SGD method we propose involves a specific estimator of the gradient, that shall be referred to as the *Horvitz-Thompson gradient estimator* (HTGD estimator in short) throughout the chapter and accounts for the sampling design used to select the subsample for gradient evaluation at each iteration. For the estimator thus produced, consistency and asymptotic normality results describing its statistical performance are established under adequate assumptions on the first and second order inclusion probabilities. They reveal that accuracy may significantly increase, *i.e.* the asymptotic variance of the estimator produced by the HTGD procedure may be drastically reduced, when the inclusion probabilities of the survey design are picked adequately, depending on some supposedly available extra information, compared to a naive implementation with equal inclusion probabilities. This is thoroughly discussed in the particular case of the Poisson survey scheme. Although it is one of the simplest sampling designs, many more general survey schemes may be expressed as Poisson schemes conditioned upon specific events, see *e.g.* Berger (2011). These theoretical results are also supported by strong empirical evidence. Many variants of the SGD technique, far too numerous to be listed here, have been

introduced these last few years in order to improve its scalability/speed,; attention should be paid to the fact that the analysis presented here only aims at sheding light on the impact of survey sampling on this technique, in its most generic form.

The rest of the chapter is structured as follows. Basics in M-estimation and SGD techniques together with key notions in survey sampling theory are briefly recalled in section 4.2. Section 4.3 first describes the Horvitz-Thompson variant of the SGD in the context of a general M-estimation problem. In section 4.4, limit results are established in a general framework, revealing the possible significant gain in terms of asymptotic variance resulting from sampling with unequal probabilities in presence of extra information. They are next discussed in more depth in the specific case of Poisson surveys. Illustrative numerical experiments, consisting in fitting a logistic regression model (respectively, a semi-parametric shift model) with extra information, are displayed in section 4.6.

4.2 Theoretical Background and Preliminaries

As a first go, we start off with describing the mathematical setup and recalling key concepts in survey theory involved in the subsequent analysis. We recall that the indicator function of an event $\mathcal{B}$ is written $\mathbb{I}\{\mathcal{B}\}$. The square root of a symmetric semi-definite positive matrix B by $B^{1/2}$.

4.2.1 Iterative M-Estimation and SGD Methods

Set two positive integers d and q. Let Z be an $\mathbb{R}^d$-valued random vector (r.v.) with unknown distribution $\mathbb{P}_Z$ and Θ a compact subspace of $\mathbb{R}^q$ equipped with the euclidean norm $\|.\|$. Consider a certain smooth loss function $l : \mathbb{R}^d \times \Theta \to \mathbb{R}$ that is square $\mathbb{P}_Z$-integrable for any $\theta \in \Theta$. Given this theoretical framework, we are interested in solving the *risk minimization* problem

$$\min_{\theta\in\Theta} L(\theta), \tag{4.1}$$

where $L : \theta \in \Theta \mapsto \mathbb{E}[l(Z,\theta)] \in \mathbb{R}$ is the risk function with its empirical counterpart

$$\widehat{L}_n : \theta \in \Theta \mapsto \frac{1}{n}\sum_{i=1}^{n} l(Z_i,\theta), \tag{4.2}$$

based on the observation of $n \geq 1$ independent copies $Z_1, \ldots, Z_N$ of Z (see Examples 4.1 and 4.2 below). As $n \to +\infty$, asymptotic properties of M-estimators, *i.e.* minimizers of $\widehat{L}_n(\theta)$, have been extensively investigated, see van de Geer (2000) for instance.

Gradient descent. The popular approach to solve empirical risk minimization, introduced in chapter 3, consists in implementing variants of the standard gradient descent method, following the iterations

$$\theta_n(t+1) = \theta_n(t) - \gamma_t \nabla \widehat{L}_n(\theta_t), \quad t \geq 1, \tag{4.3}$$

with an initial value $\theta_n(0)$ arbitrarily chosen and a non-negative learning rate (step size or gain) γ_t. The latter is taken such that $\sum_{t=1}^{+\infty} \gamma_t = +\infty$ and $\sum_{t=1}^{+\infty} \gamma^2(t) < +\infty$, see *e.g.* Bertsekas (2003). The true gradient is replaced by a counterpart computed from a subsample $S \subset \{1, \ldots, n\}$ of reduced size $N \ll n$, so as to fulfill the computational constraints, and

drawn at random (uniformly) among all possible subsets of same size at each iteration:

$$\ell_N : \theta \in \Theta \mapsto \frac{1}{N} \sum_{i \in S} \nabla l(Z_i, \theta). \tag{4.4}$$

The variant of the SGD that we introduced in chapter 3 proposed to use more complex sampling schemes to speed-up the learning process. Unfortunately the results established were true conditionally upon the observations and only related to the empirical risk. We extend this analysis in the present chapter to the true risk by characterizing the asymptotic behaviour of $L(\theta_n(t)) - L(\theta^*)$. This analysis can be applied to the two following examples.

Example 4.1. (BINARY CLASSIFICATION) *In the usual binary classification framework introduced in chapter 2, Y is a binary random output, taking its values in $\{-1, +1\}$ say, and X is an input random vector valued in a high-dimensional space $\mathcal{X}$, modeling some (hopefully) useful observation for predicting Y. Based on training data $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, the goal is to build a prediction rule $\text{sign}(h(X))$, where $h : \mathcal{X} \to \mathbb{R}$ is some measurable function that minimizes the risk*

$$L_\varphi(h) = \mathbb{E}\left[\varphi(-Y h(X))\right].$$

Here, the expectation is taken over the unknown distribution of the random vector (X, Y) and $\varphi : \mathbb{R} \to [0, +\infty)$ denotes a cost function, i.e. a measurable function such that $\varphi(u) \geq \mathbb{I}\{u \geq 0\}$ for any $u \in \mathbb{R}$. For example, when φ is chosen as the convex function $u \in \mathbb{R} \mapsto (u+1)^2/2 \in \mathbb{R}_+$, then the optimal decision function is given by $h^* : x \in \mathcal{X} \mapsto 2\,\mathbb{P}\{Y = +1 \mid X = x\} - 1 \in [-1, 1]$ and the classification rule $H^* : x \in \mathbf{X} \mapsto \text{sign}(h^*(x)) \in \{-1, +1\}$ coincides with the naive Bayes classifier. For simplicity, assume that φ is differentiable and that the decision function candidates $h(x)$ belong to the parametric set $\{h(., \theta) : \theta \in \Theta\}$ of square integrable functions (with respect to the distribution of X) indexed by $\Theta \subset \mathbb{R}^q$, $q \geq 1$, such that $\theta \mapsto h(., \theta)$ is differentiable. Finding the prediction rule with minimum risk amounts to solving (4.1) with $Z = (X, Y)$ and $l(Z, \theta) = \varphi(-Y\, h(X, \theta))$ for all $\theta \in \Theta$. In the ideal case where a standard gradient descent could be applied, a sequence $\theta_t = (\theta_1(t), \cdots, \theta_q(t))$, $t \geq 1$, would be iteratively generated using the update equation

$$\theta(t+1) = \theta_t + \gamma_t\, \mathbb{E}\left[Y\, \nabla h(X, \theta_t)\, \varphi'(-Y\, h(X, \theta_t))\right]$$

with learning rate $\gamma_t > 0$. Naturally, as the distribution of (X, Y) is unknown, the expectation involved in the t-th iteration cannot be computed and must be replaced by a statistical version:

$$\frac{1}{n} \sum_{i=1}^{n} Y_i\, \nabla h(X_i, \theta_t)\, \varphi'(-Y_i\, h(X_i, \theta_t)),$$

in accordance with the Empirical Risk Minimization paradigm.

Example 4.2. (LOGISTIC REGRESSION) *Consider the same probabilistic model as above, except that the goal pursued is to find $\theta \in \Theta$ so as to minimize $\widehat{L}_n(\theta)$ in (4.2) with $Z_i = (X_i, Y_i)$ and $l(Z_i, \theta)$ defined as*

$$- \quad \left\{ \frac{Y_i + 1}{2} \log\left(\frac{\exp(h(X_i, \theta))}{1 + \exp(h(X_i, \theta))} \right) + \frac{1 - Y_i}{2} \log\left(\frac{1}{1 + \exp(h(X_i, \theta))} \right) \right\}$$

for all $i \in \{1, \ldots, n\}$ and $\theta \in \Theta$. This is equivalent to maximizing the conditional log-likelihood given the X_i's related to the parametric logistic regression model:

$$\mathbb{P}_\theta\{Y = +1 \mid X\} = \exp(h(X, \theta))/(1 + \exp(h(X, \theta))), \quad \theta \in \Theta.$$

4.2.2 Survey Sampling and Horvitz-Thompson Estimation

We recall a few notations introduced in chapter 2 and 4. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and n a positive integer. In the framework we consider throughout the chapter, it is assumed that $Z_1, \ldots, Z_n$ is a sample of i.i.d. random vectors defined on $(\Omega, \mathcal{A}, \mathbf{P})$ and taking their values in $\mathbb{R}^d$. They are interpreted as independent copies of a generic r.v. Z observed on a finite population $\mathcal{U}_n = \{1, \ldots, n\}$. A *survey sample* of the population is defined as a non-empty subset $S \subset \mathcal{U}_n$ with cardinality $N = N(S)$ less that n, selected at random according to a probability distribution R_n on $\mathcal{P}(\mathcal{U}_n)$, the power set of $\mathcal{U}_n$. The latter is called a *sampling scheme/design/plan* without replacement. We shall consider R_n as a conditional distribution given the statistical population $\mathcal{U}_n$ and the possible observations assigned to each of its units. In this setting, for any $i \in \mathcal{U}_n$, the probability that the unit i belongs to a random sample S drawn from such a R_n is called the (first order) *inclusion probability*:

$$\pi_i(R_n) = \mathbb{P}_{R_n}\{i \in S\}.$$

We set $\boldsymbol{\pi}(R_n) = (\pi_1(R_n), \ldots, \pi_N(R_n))$. The second order inclusion probabilities are

$$\pi_{i,j}(R_n) = \mathbb{P}_{R_n}\{i \in S, j \in S\},$$

for any (i, j) in $\mathcal{U}_n^2$. In particular, $\pi_{i,i}(R_n) = \pi_i(R_n)$. When no confusion is possible, we shall omit to mention the dependence in R_n when writing the first/second order probabilities of inclusion. The information related to the random sample $S \subset \mathcal{U}_n$ is fully enclosed in the random vector $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ with components $\epsilon_i = \mathbb{I}\{i \in S\}$, $i \in \mathcal{U}_n$. Given the statistical population, the conditional 1-d marginal distributions of the sampling scheme $\boldsymbol{\epsilon}_n$ are the Bernoulli distributions $\mathcal{B}(\pi_i) = \pi_i\delta_1 + (1 - \pi_i)\delta_0$, $i \in \mathcal{U}_n$, with δ_x the Dirac mass at point $x \in \mathbb{R}$. The conditional covariance matrix of the r.v. $\boldsymbol{\epsilon}_n$ is given by $\Gamma_n = \{\pi_{i,j} - \pi_i\pi_j\}_{1 \leq i,j \leq n}$. Observe that $\sum_{i=1}^n \epsilon_i = N(S)$, which can be fixed or random depending on R_n.

Poisson schemes. One of the simplest survey designs is the Poisson scheme (without replacement), denoted by p_n. For such a plan, conditioned upon the statistical population of interest, the ϵ_i's are independent Bernoulli random variables with parameters $p_1, \ldots, p_n$ in $(0, 1]$. Thus, the first order inclusion probabilities $\pi_i(p_n) = p_i$, $i \in \mathcal{U}_n$, fully characterize p_n. The size $N(S)$ of a sample S generated this way is random with mean $\sum_{i=1}^n p_i$ and goes to infinity as $n \to +\infty$ with probability one, provided that $\min_{1 \leq i \leq n} p_i$ remains bounded away from zero. In addition to its simplicity (regarding the procedure to select a sample thus distributed), the Poisson design plays a crucial role in sampling theory, insofar as it can be used to build a wide range of survey plans by conditioning arguments Hajek (1964). For instance, a *rejective sampling plan* of fixed size $N \leq n$ corresponds to the distribution of a Poisson scheme $\boldsymbol{\epsilon}_n$ conditioned upon the event $\{\sum_{i=1}^n \epsilon_i = N\}$. One may refer to Cochran (1977), Deville (1987) for accounts of survey sampling techniques and examples of designs to which the subsequent analysis applies.

Horvitz-Thompson estimators. We recall a few definitions and give the variance of Horvitz-Thompson estimators in the Poisson case. Suppose that independent random vectors $Q_1, \ldots, Q_n$ are observed on the population $\mathcal{U}_n$. They are viewed as copies of a generic r.v. Q taking its values in $\mathbb{R}^q$. A natural approach to estimate the total $\mathbf{Q}_n = \sum_{i=1}^n Q_i$ based on a sample $S \subset \mathcal{U}_n$ generated from a survey design R_n with *positive* (first order) inclusion probabilities $\{\pi_i\}_{1 \leq i \leq n}$ consists in computing the *Horvitz-Thompson estimator* (HT estimator

in abbreviated form)

$$\mathbf{Q}_{R_n} = \sum_{i \in S} \frac{1}{\pi_i} Q_i = \sum_{i=1}^{n} \frac{\epsilon_i}{\pi_i} Q_i. \tag{4.5}$$

Given the whole statistical population $Q_1, \ldots, Q_N$, the HT estimator is an unbiased estimate of the total: $\mathbb{E}\left(\mathbf{Q}_{R_n} \mid Q_1, \ldots, Q_n\right) = \mathbf{Q}_N$ almost-surely. Its conditional variance is given by

$$\mathbb{V}\left(\mathbf{Q}_{R_n} \mid Q_1, \ldots, Q_n\right) = \sum_{i,j=1}^{n} \frac{\pi_{i,j} - \pi_i\, \pi_j}{\pi_i\, \pi_j}\, Q_i\, Q_j^T.$$

In particular, when the survey design is a Poisson plan P_n with positive probabilities $p_1, \ldots, p_n$, this turns into

$$\mathbb{V}\left(\mathbf{Q}_{P_n} \mid Q_1, \ldots, Q_n\right) = \sum_{i=1}^{n} \frac{1 - p_i}{p_i}\, Q_i\, Q_i^T. \tag{4.6}$$

Remark 4.1. (AUXILIARY INFORMATION) In practice, the first order inclusion probabilities are defined as a function of an *auxiliary variable*, say W taking its values in $\mathbb{R}^{d'}$, $d' \geq 1$, which is observed on the entire population. Specifically, a link function $\pi : \mathbb{R}^{d'} \to (0, 1]$ is chosen so that $\pi_i = \pi(W_i)$ for all $i \in \mathcal{U}_n$. When $\pi(W)$ and Q are dependent, proceeding this way may help us select more informative samples and consequently yield estimators with reduced variance. A more detailed discussion on the use of auxiliary information in the present context can be found in subsection 4.4.1.

Going back to the SGD problem, the *Horvitz-Thompson estimator* of the gradient $\nabla \widehat{L}_n(\theta)$ based on a survey sample S drawn within the population $\mathcal{U}_n = \{1, \ldots, n\}$ from a design R_n with vector of (first order) inclusion probabilities $\boldsymbol{\pi}_n = (\pi_1, \ldots, \pi_N)$ and inclusion vector $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ is

$$\ell_{R_n}(\theta) = \frac{1}{n} \sum_{i \in S} \frac{1}{\pi_i} \nabla l(Z_i, \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\epsilon_i}{\pi_i} \nabla l(Z_i, \theta), \quad \theta \in \Theta. \tag{4.7}$$

As pointed out in Remark 4.1 and in section 3.3.2, this estimator would be most efficient if each π_i was strongly correlated with the corresponding $\nabla l(Z_i, \theta)$, $i \in \mathcal{U}_n$. This suggests to devise a procedure where the survey design used to estimate the gradient may change at each step, as in the HTGD algorithm described in the next section. For instance, one could stipulate the availability of extra information $W_1, \ldots, W_N$ and assume the existence of a link function $\pi : \mathcal{W} \times \Theta \to (0, 1]$ such that $\pi_i = \pi(W_i, \theta)$ for all $i \in \mathcal{U}_n$.

Of course, such an approach would be beneficial only if the cost of the computation of the weight $\pi(W_i, \theta)$ is smaller than that of the gradient $\nabla l(Z_i, \theta)$. As shall be seen in section 4.6, this happens to be the case in many situations encountered in practice.

4.3 The Horvitz-Thompson Gradient Descent

This section presents, in full generality, the variant of the SGD method we propose. It can be implemented in particular when some extra information about the target (the gradient vector field in the present case) is available, allowing hopefully for picking a sample yielding a more accurate estimation of the (true) gradient than that obtained by means of a sample chosen completely at random. Several tuning parameters must be picked by the user, including the

parameter N which controls the number of terms involved in the empirical gradient estimation at each iteration.

HORVITZ-THOMPSON GRADIENT DESCENT ALGORITHM (HTGD)

(INPUT.) Datasets $\{Z_1, \ldots, Z_n\}$ and $\{W_1, \ldots, W_n\}$. Maximum (expected) sample size $N \leq n$. Collection of sampling plans $R_n(\theta)$ with positive first order inclusion probabilities $\pi_i(\theta)$ for $1 \leq i \leq n$, indexed by $\theta \in \Theta$ with (expected) sample sizes less than N. Learning rate $\gamma_t > 0$. Number of iterations $T \geq 1$.

1. (INITIALIZATION.) Choose $\theta_n(0)$ in Θ.
2. (ITERATIONS.) For $t = 0, \ldots, T$
   (a) Draw a survey sample from $\mathcal{U}_n = \{1, \ldots, n\}$, described by the inclusion vector $\boldsymbol{\epsilon}_n^{(t)} = \left(\epsilon_1^{(t)}, \ldots, \epsilon_n^{(t)}\right)$, according to $R_n = R_n(\theta_n(t))$ with inclusion probabilities $\pi_i(\theta_n(t))$ for $i \in \mathcal{U}_n$.
   (b) Compute the HT gradient estimate at $\theta_n(t)$

$$\ell_{R_n}(\theta_n(t)) := \frac{1}{n} \sum_{i=1}^{n} \frac{\epsilon_i^{(t)}}{\pi_i(\theta_n(t))} \nabla l(Z_i, \theta_n(t)).$$

   (c) Update the estimator

$$\theta_n(t+1) = \theta_n(t) - \gamma_t \, \ell_{R_n}(\theta_n(t)).$$

(OUTPUT.) The HTGD estimator $\theta_n(T)$.

Conditioned upon the data $(Z_1, W_1), \ldots, (Z_n, W_n)$, the asymptotic accuracy of the estimator or decision rule produced by the algorithm above as the number of iterations T tends to infinity is investigated in the next section under specific assumptions. Beyond consistency, special attention is paid to the issue of choosing properly the sampling plans $R_n(\theta)$ so as to minimize the asymptotic variance of the estimator $\theta_n(T)$ or that of its empirical risk.

4.4 Conditional Asymptotic Analysis - Main Results

This section is dedicated to the analysis of the performance of the HTGD method, conditioned upon the observed population and under adequate constraints related to the (expected) size of the survey samples. We first discuss the case of Poisson survey schemes and next investigate how to establish limit results in a more general framework. They are similar to the results established in section 3.4.2 and serve as a preliminary for our main results. All expectations are taken conditionally upon the observations in this section.

4.4.1 Poisson Schemes with Unequal Inclusion Probabilities

Here we will recall a result establish in chapter 3. Fix $\theta \in \Theta$ and $N \leq n$. Given $Z_1, \ldots, Z_n$, consider a Poisson scheme P_n on the population $\mathcal{U}_n = \{1, \ldots, n\}$ with positive parameter

$\mathbf{p}_n = (p_1, \ldots, p_n)$. Then, Eq. (4.6) implies

$$\mathbb{E}\Big[\Big\|\ell_{P_n}(\theta) - \nabla\widehat{L}_n(\theta)\Big\|^2 \mid Z_1, \ldots, Z_n\Big] = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1-p_i}{p_i}\,\|\nabla l(Z_i,\theta)\|^2\,.$$

Searching for the parameter $\widetilde{\mathbf{p}}_n$ such that the L_2 distance between the empirical gradient evaluated at θ and the HT version given $Z_1, \ldots, Z_n$ is minimum under the constraint that the expected sample size is equal to $N \leq n$ yields the optimization problem

$$\min_{\mathbf{p}_n \in (0,1]^n} \sum_{i=1}^{n}\frac{1-p_i}{p_i}\,\|\nabla l(Z_i,\theta)\|^2 \quad \text{subject to} \quad \sum_{i=1}^{n} p_i = N. \tag{4.8}$$

Suppose that $\mathbb{P}\{\nabla l(Z,\theta) = 0\} = 0$ for all $\theta \in \Theta$; this is true in particular when the set $\{z \in \mathbb{R}^d : \nabla l(z,\theta) = 0\}$ has finite cardinality and the distribution of Z is absolutely continuous with respect to the Lebesgue measure. Then we have $\|\nabla l(Z_i,\theta)\| > 0$ with probability one for all $i \in \mathcal{U}_n$ and $\theta \in \Theta$. As can be shown by means of the Lagrange multipliers method, in this setting the solution corresponds to weights being proportional to the values taken by the norm of the gradient just like in section 3.3.2.

$$\widetilde{p}_i(\theta) := N\,\frac{\|\nabla l(Z_i,\theta)\|}{\sum_{j=1}^{n}\|\nabla l(Z_j,\theta)\|},$$

provided that the following condition is fulfilled:

$$\widetilde{p}_i(\theta) \leq 1 \quad \text{for all} \quad i \in \mathcal{U}_n. \tag{4.9}$$

A straightforward application of Hoeffding's inequality shows that if

$$\varepsilon := \frac{\mathbb{E}\left[\|\nabla l(Z,\theta)\|\right]}{\sup_{z\in\mathbb{R}^d}\|\nabla l(z,\theta)\|} - \frac{N}{n} \in \left(0, \frac{\mathbb{E}\left[\|\nabla l(Z,\theta)\|\right]}{\sup_{z\in\mathbb{R}^d}\|\nabla l(z,\theta)\|}\right),$$

then condition (4.9) is satisfied with probability larger than $1 - \exp\left(-2n\varepsilon^2\right)$.

Remark 4.2. (ON THE SATURATION OF THE LINEAR CONSTRAINTS) When the latter condition is not satisfied, some of the conditions $\widetilde{p}_i(\theta) \leq 1$ are saturated and the solution of (4.8) is given by the Karush-Kuhn-Tucker method. Since the objective function is strictly convex and the constraints are affine, the following conditions, related to the Lagrangian

$$\sum_{i=1}^{n}\frac{1-p_i}{p_i}\,\|\nabla l(Z_i,\theta)\|^2 + \lambda\left(\sum_{i=1}^{n} p_i - N\right) + \sum_{i=1}^{n}\mu_i(p_i - 1),$$

are necessary and sufficient: (i) $\displaystyle\sum_{i=1}^{n} p_i = N$ and for all $i \in \mathcal{U}_n$ (ii) $0 < p_i \leq 1$,

$$(iii)\ \frac{\|\nabla l(Z_i,\theta)\|^2}{p_i^2} = \lambda + \mu_i, \quad (iv)\ \mu_i \geq 0, \quad (v)\ \mu_i(p_i - 1) = 0.$$

Denoting by $m < N$ the number of components of the solution $\widetilde{\mathbf{p}}_n$ that are equal to 1 and by σ a permutation of $\mathcal{U}_n$ such that $\|\nabla l(Z_{\sigma(1)},\theta)\| \leq \ldots \leq \|\nabla l(Z_{\sigma(n)},\theta)\|$, the constraint (i) can be rewritten as $N = m + \sum_{i=1}^{n-m}\|\nabla l(Z_{\sigma(i)},\theta)\|/\sqrt{\lambda}$, so that $p_{\sigma(i)} = (N-m)\|\nabla l(Z_{\sigma(i)},\theta)\|/\sum_{j=1}^{n-m}\|\nabla l(Z_{\sigma(j)},\theta)\|$ for $i \leq n-m$ and $p_{\sigma(i)} = 1$ for $i \geq n-m+1$.

However, selecting a sample distributed this way requires to know the full statistical population $\nabla l(Z_i, \theta)$. In practice, one may consider situations where the weights are defined by means of a link function $\pi(W, \theta)$ and auxiliary variables $W_1, \ldots, W_n$ such that the inclusion probabilities are correlated with their corresponding gradient, as suggested previously. Observe in addition that the goal pursued here is not to estimate the gradient but to implement a stochastic gradient descent involving an expected number of terms fixed in advance, while yielding results close to those that would be obtained by means of a gradient descent algorithm with mean field $(1/n)\sum_{i=1}^{n} \nabla l(Z_i, \theta)$ based on the whole dataset. However, as shall be seen in the subsequent analysis (see Proposition 4.6), in general these two problems do not share the same solution from the angle embraced in this article.

In the next subsection, assumptions on the survey design under which the HTGD method yields accurate asymptotic results, surpassing those obtained with all equal inclusion probabilities (*i.e.* $\pi_i = N/n$ for all $i \in \mathcal{U}_n$), are exhibited.

4.4.2 Limit Theorems - Conditional Consistency and Asymptotic Normality

We now consider a collection of general (*i.e.* not necessarily Poisson) sampling schemes $\{R_n(\theta)\}_{\theta\in\Theta}$ with positive first order inclusion probabilities $\{\boldsymbol{\pi}_n(\theta)\}_{\theta\in\Theta}$. Conditioned upon the data $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ (or $\mathcal{D}_n = \{(Z_1, W_1), \ldots, (Z_n, W_n)\}$ in the presence of extra variables, *cf* Remark 4.1) available in the population $\mathcal{U}_n = \{1, \ldots, n\}$, we study the asymptotic properties of the M-estimator produced by the HTGD algorithm. The limit results stated below essentially rely on the fact that the HT estimator (4.7) of the gradient of the empirical risk is unbiased. Reduction of the asymptotic variance of $\theta_n(T)$ and $\widehat{L}_n(\theta_n(T))$ will be investigated in the Poisson case in the next subsection. The asymptotic analysis also involves the *regularity conditions* listed below, which are classically required in stochastic approximation.

Assumption 4. The conditions below hold true.

1. For any $z \in \mathbb{R}^d$, the mapping $\theta \in \Theta \mapsto l(z, \theta)$ is of class $\mathcal{C}^1$.
2. For any compact set $\mathcal{K} \subset \Theta$, we have with probability one: $\forall i \in \mathcal{U}_n$,
$$\sup_{\theta\in\mathcal{K}} \frac{\|\nabla l(Z_i, \theta)\|}{\pi_i(\theta)} < +\infty.$$
3. The set of stationary points $\mathcal{L}_n = \{\theta \in \Theta : \ell_N(\theta) = 0\}$ is of finite cardinality.

Theorem 4.3. (CONDITIONAL CONSISTENCY) *Suppose that Assumption 4 is fulfilled and that*

- *the learning rate decays to* 0 *so that* $\sum_{t\geq 1} \gamma_t = +\infty$ *and* $\sum_{t\geq 0} \gamma^2(t) < +\infty$,
- *the* HTGD *algorithm is stable, i.e. there exists a compact set* $\mathcal{K}_0 \subset \mathbb{R}^q$ *such that* $\theta_n(t) \in \mathcal{K}_0$ *for all* $t \geq 0$.

Then, conditioned upon the data $\mathcal{D}_n$, *the sequence* $\{\theta_n(t)\}_{t\geq 0}$ *converges to an element of the set* $\mathcal{L}_n$ *with probability one as* $t \to +\infty$.

The stability condition is generally difficult to check. In practice, one may guarantee it by confining the sequence to a compact set fixed in advance and using a *projected* version of

the algorithm above. For simplicity, the present study is restricted to the simplest framework for stochastic gradient descent and we refer to Kushner & Yin (2010) or Borkar (2008) (see section 5.4 therein) for further details.

Consider a stationary point $\theta_n^* \in \mathcal{L}_n$. The following *local* assumptions are also required to establish asymptotic normality results conditioned upon the event

$$\mathcal{E}(\theta_n^*) = \left\{ \lim_{t \to +\infty} \theta_n(t) = \theta_n^* \right\}.$$

Assumption 5. The conditions below hold true.

1. There exists a neighbourhood $\mathcal{V}$ of θ_n^* such that for all $z \in \mathbb{R}^d$, the mapping $\theta \in \Theta \mapsto l(z, \theta)$ is of class $\mathcal{C}^2$ on $\mathcal{V}$.
2. The Hessian matrix $H_n = \nabla^2 \widehat{L}_n(\theta_n^*)$ is a stable $q \times q$ positive-definite matrix, *i.e.* its smallest eigenvalue l is positive.
3. For all $(i, j) \in \mathcal{U}_n^2$, the mapping $\theta \in \mathcal{V} \mapsto \pi_{i,j}(\theta)$ is continuous.

Under these assumptions, and similarly to the results establish in section 3.4.2 we have the following TCL.

Theorem 4.4. (CONDITIONAL CENTRAL LIMIT THEOREM) *Suppose that Assumptions 4–5 are fulfilled and that $\gamma_t = \gamma_0\, t^{-\alpha}$ for some constants $\alpha \in (1/2, 1]$ and $\gamma_0 > 0$. When $\alpha = 1$, take $\gamma_0 > 1/(2l)$ and set $\eta := 1/(2\gamma_0)$; set $\eta := 0$ otherwise. Given the observations $\mathcal{D}_n$ and conditioned upon the event $\mathcal{E}(\theta_n^*)$, we have the convergence in distribution as $t \to +\infty$*

$$\sqrt{1/\gamma_t}\,(\theta_n(t) - \theta_n^*) \Rightarrow \mathcal{N}(0, \Sigma_{\boldsymbol{\pi}_n}),$$

where the asymptotic covariance matrix $\Sigma_{\boldsymbol{\pi}_n}$ is the unique solution of the Lyapunov equation

$$H_n \Sigma + \Sigma H_n + 2\eta\Sigma = \Gamma_n^*, \tag{4.10}$$

with $\Gamma_n^* = \Gamma_n(\theta_n^*)$ and, for all $\theta \in \Theta$,

$$\Gamma_n(\theta) = \frac{1}{n^2} \sum_{i,j=1}^{n} \left(\frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\pi_j(\theta)} - 1 \right) \nabla l(Z_i, \theta)\, \nabla l(Z_j, \theta)^T. \tag{4.11}$$

The result stated below provides the asymptotic conditional distribution of the error. Because it is a direct application of the second order delta method, the proof is omitted.

Corollary 4.5. (ERROR RATE) *Under the hypotheses of Theorem 4.4, given the observations $\mathcal{D}_n$ and conditioned upon the event $\mathcal{E}(\theta_n^*)$, as $t \to +\infty$ we have the convergence in distribution towards a non-central chi-square distribution:*

$$1/\gamma_t \left(\widehat{L}_n(\theta_n(t)) - \widehat{L}_n(\theta_n^*) \right) \Rightarrow \frac{1}{2}\, U^T\, \Sigma_{\boldsymbol{\pi}_n}^{1/2}\, H_n\, \Sigma_{\boldsymbol{\pi}_n}^{1/2}\, U,$$

where U is a q-dimensional standard Gaussian random vector.

4.4.3 Asymptotic Covariance Optimization in the Poisson Case

Now that the limit behavior of the solution produced by the HTGD algorithm has been described for general collections of survey designs $\mathcal{R} = \{R_n(\theta)\}_{\theta \in \Theta}$ of fixed expected sample

size, we turn to the problem of finding survey plans yielding estimates with best accuracy. Formulating this objective in a quantitative manner, this boils down to finding $\mathcal{R}$ so as to minimize the asymptotic covariance matrix summary $\|\Sigma_{\boldsymbol{\pi}_n}^{1/2}\|$, for an appropriately chosen norm $\|.\|$ on the space $\mathcal{M}_q(\mathbb{R})$ of $q \times q$ matrices with real entries for instance, when it comes to estimate θ_n^*. In order to get a natural summary of the asymptotic variability, we consider here the Frobenius (Hilbert-Schmidt) norm, *i.e.* $\|A\| = \sqrt{Tr(A^T A)} = (\sum_{i,j} a_{i,j}^2)^{1/2}$ for any $A = (a_{i,j}) \in \mathcal{M}_q(\mathbb{R})$. For simplicity's sake, we focus on Poisson schemes and consider the case where $\eta = 0$ in Theorem 4.4. Notice that the cross terms ($i \neq j$) in Eq. (4.11), *i.e.* the U-statistic part of the conditional asymptotic variance, vanish in the Poisson case. The following result exhibits an optimal collection of Poisson schemes among those with N as expected sizes, in the sense that it yields an HTGD estimator with an asymptotic covariance of square root with minimum Frobenius norm. We point out that it is generally different from that considered in subsection 4.4.1, revealing the difference between the issue of estimating the empirical gradient accurately by means of a Poisson Scheme and that of optimizing the HTGD procedure.

Proposition 4.6. (OPTIMALITY) *Consider the same assumptions as in Theorem 5 in the case where $\eta = 0$ and suppose that*

$$N \leq \inf_{\theta \in \Theta} \frac{\sum_{i=1}^n \|G_n \nabla l(Z_i, \theta)\|}{\max_{1 \leq i \leq n} \|G_n \nabla l(Z_i, \theta)\|}, \tag{4.12}$$

with $G_n := H_n^{-1/2}$. Then, the collection of Poisson schemes with positive inclusion probabilities $\{\mathbf{p}_n^*(\theta)\}_{\theta \in \Theta}$ defined for all $\theta \in \Theta$ and $i \in \mathcal{U}_n$ by

$$p_i^*(\theta) = N \frac{\|G_n \nabla l(Z_i, \theta)\|}{\sum_{j=1}^n \|G_n \nabla l(Z_j, \theta)\|}$$

is a solution of the minimization problem

$$\min_{\mathbf{p}_n = \{\mathbf{p}_n(\theta)\}_{\theta \in \Theta}} \left\| \Sigma_{\mathbf{p}_n}^{1/2} \right\| \text{ subject to } \sum_{i=1}^n p_i(\theta) = N.$$

In addition, we have

$$\left\| \Sigma_{\mathbf{p}_n^*}^{1/2} \right\|^2 = \frac{1}{2} \left\{ \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n \|G_n \nabla l(Z_i, \theta_n^*)\| \right)^2 - \frac{1}{n^2} \sum_{i=1}^n \|G_n \nabla l(Z_i, \theta_n^*)\|^2 \right\}.$$

Of course, the optimal solution exhibited in the result stated above is completely useless from a practical perspective, since the matrix H_n is unknown in general and the computation of the values taken by the gradient at each point Z_i is precisely what we are trying to avoid in order to reduce the computational cost of the GD procedure. However, we show in the next section that choosing inclusion probabilities positively correlated with the $p_i^*(\theta)$'s is actually sufficient to reduce asymptotic variability (compared to the situation where equal inclusion probabilities are used). In addition, as illustrated by the two easily generalizable examples described in section 4.6, such a sampling strategy can be implemented in many situations.

Notice finally that, if we consider the asymptotic excess of empirical risk of the estimate $\widehat{L}_n(\theta_n(T)) - \widehat{L}_n(\theta_n^*)$ rather than the asymptotic variance of the estimate itself, the survey

design $\mathcal{R}$ must be picked in order to minimize the quantity

$$\begin{aligned}\mathbb{E}\left[U^T \Sigma_{\boldsymbol{\pi}_n}^{1/2} H_n \Sigma_{\boldsymbol{\pi}_n}^{1/2} U \mid \mathcal{D}_n\right] &= \mathbb{E}\left[\left(H_n^{1/2} \Sigma_{\boldsymbol{\pi}_n}^{1/2} U\right)^T \left(H_n^{1/2} \Sigma_{\boldsymbol{\pi}_n}^{1/2} U\right) \mid \mathcal{D}_n\right] \\ &= Tr\left(H_n \Sigma_{\boldsymbol{\pi}_n}\right),\end{aligned}$$

using the fact that $U \sim \mathcal{N}(0, I_q)$ is chosen independent from $\mathcal{D}_n$ here. Observing that $H_n \Sigma_{\boldsymbol{\pi}_n} + \Sigma_{\boldsymbol{\pi}_n} H_n = \Gamma_n^*$ in the case $\eta = 0$, we have

$$Tr\left(H_n \Sigma_{\boldsymbol{\pi}_n}\right) = \frac{1}{2} Tr(\Gamma_n^*) = \mathbb{E}[||\ell_{P_n}(\theta_n^*)||^2]. \tag{4.13}$$

Now, since we have in the Poisson case

$$Tr(\Gamma_n^*) = \frac{1}{n^2} \sum_{i=1}^n \left(\frac{1}{p_i(\theta_n^*)} - 1\right) \|\nabla l(Z_i, \theta_n^*)\|^2,$$

the optimal Poisson scheme regarding this alternative criterion generally differs from that involved in Proposition 4.6 and boils down to that introduced in subsection 4.4.1 for optimal Horvitz-Thompson estimation of the gradient.

4.4.4 Extensions to More General Poisson Survey Designs

In this subsection, we still consider Poisson schemes and the case $\eta = 0$ for simplicity and now place ourselves in the situation where the information at disposal consists of a collection of i.i.d. random pairs $(Z_1, W_1), \ldots, (Z_n, W_n)$ valued in $\mathbb{R}^d \times \mathbb{R}^{d'}$. Take a link function $p : \mathbb{R}^{d'} \times \Theta \to (0, 1]$ such that $\theta \in \Theta \mapsto p(w, \theta)$ is continuous for all $w \in \mathbb{R}^{d'}$, then choose an expected sample size $N \in \{1, \ldots, n\}$ that satisfies

$$N \leq \inf_{\theta \in \Theta} \frac{\sum_{i=1}^n p(W_i, \theta)}{\max\limits_{1 \leq i \leq n} p(W_i, \theta)}$$

and define

$$p_i(\theta) = N \frac{p(W_i, \theta)}{\sum_{j=1}^n p(W_j, \theta)}, \qquad \text{for all } (i, \theta) \in \mathcal{U}_n \times \Theta. \tag{4.14}$$

Observe that for all $\theta \in \Theta$ we have $\sum_{i=1}^n p_i(\theta) = N$ and $p_i(\theta) \in (0, 1]$ for all $i \in \mathcal{U}_n$. The computational cost of the inclusion probability $p(W_i, \theta)$ is assumed to be much smaller than that of $\nabla l(Z_i, \theta)$ (see the examples in section 4.6) for all $(i, \theta) \in \mathcal{U}_n \times \Theta$. The assumption introduced below involves the empirical covariance $c_n(\theta)$ between $\|G_n \nabla l(Z, \theta)\|^2 / p(W, \theta)$ and $p(W, \theta)$, for $\theta \in \Theta$:

$$c_n(\theta) = \frac{1}{n} \sum_{i=1}^n \|G_n \nabla l(Z_i, \theta)\|^2 \left(1 - \frac{1}{n} \frac{\sum_{j=1}^n p(W_j, \theta)}{p(W_i, \theta)}\right).$$

Assumption 6. The link function $p(w, \theta)$ fulfills the following condition:

$$c_n(\theta_n^*) > 0.$$

The result stated below reveals to which extent sampling with inclusion probabilities defined by some appropriate link function may improve upon sampling with equal inclusion probabilities, $\widetilde{p}_i = N/n$ for $1 \leq i \leq N$, when implementing stochastic gradient descent. Namely,

the accuracy of the HTGD gets closer and closer to the optimum, as the empirical covariance $c_n(\theta^*)$ increases to its maximum. Notice that in the case where inclusion probabilities are all equal, we have $c_n \equiv 0$.

Proposition 4.7. *Let N be fixed. Suppose that the collection of Poisson designs $\mathbf{p}$ with expected sizes N is defined by a link function $p(w, \theta)$ satisfying Assumption 6. Then, when Theorem 4.4 applies, we have*

$$\left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|_{?\mathsf{F}} < \left\|\Sigma_{\check{\mathbf{p}}_n}^{1/2}\right\|_{?\mathsf{F}},$$

as well as

$$0 \leq \left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|_{?\mathsf{F}}^2 - \left\|\Sigma_{\mathbf{p}_n^*}^{1/2}\right\|_{?\mathsf{F}}^2 = \frac{1}{2N}\left\{\sigma_n^2(\theta_n^*) - c_n(\theta_n^*)\right\},$$

where

$$\sigma_n^2(\theta) = \frac{1}{n}\sum_{i=1}^{n} \|G_n\, \nabla\psi(Z_i, \theta)\|^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \|G_n\, \nabla\psi(Z_i, \theta)\|\right)^2$$

denotes the empirical variance of the r.v. $\left\|\mathbb{E}\left[\nabla^2\psi(Z,\theta)\right]^{-1/2}\, \nabla\psi(Z,\theta)\right\|$, $\theta \in \Theta$.

As illustrated by the easily generalizable examples provided in the next section, one may generally find link functions fulfilling Assumption 6 without great effort, permitting to gain in accuracy from the implementation of the HTGD algorithm.

4.5 Unconditional Asymptotic Analysis

Building upon the results of the previous section, we now investigate the behaviour of the HTGD estimator as n, N and t simultaneously tend to $+\infty$ at appropriate rates. For the sake of simplicity we assume in this section that the minimizer θ^* over the supposedly compact parameter space Θ is unique, as well as the empirical minimizer θ_n^* with probability one. All the results stated in this section can be directly extended to more general cases.

The assumption below, related to the asymptotic behavior of (4.11), is involved in the subsequent unconditional analysis.

Assumption 7. As both n and N tend to ∞, $N\Gamma_n^*$ converges in probability toward a positive semi-definite matrix Γ^*.

Although this condition may seem strong at first glance, it is actually fulfilled in several important situations. In particular, the following proposition shows it holds true in the Poisson case under weak conditions.

Proposition 4.8. *Suppose that the survey schemes are of Poisson type with link functions $p(., \theta) : \mathbb{R}^{d'} \to (0, 1)$, $\theta \in \Theta$, based on the auxiliary information W observed on the statistical population. Assume also that the following conditions are fulfilled.*

(i) We have $\theta_n^* \to \theta^*$ with probability one, as $n \to +\infty$.

(ii) The expected size N tend to infinity as $n \to \infty$, so that $\dfrac{N}{n} \to c_0 \in [0, 1]$.

(iii) For all $\theta \in \Theta$:

$$\mathbb{E}\left[p(W, \theta)\right] < +\infty \text{ and } \mathbb{E}\left[\frac{1}{p(W,\theta)}\nabla l(Z,\theta)\nabla\theta l(Z,\theta)^T\right] < +\infty.$$

(iv) The essential supremum below is finite:

$$\sup_{\theta\in\Theta}\sup\left\{x>0:\mathbb{P}\left\{\frac{\|\nabla l(Z,\theta)\|}{p(W,\theta)}>x\right\}=0\right\}<+\infty.$$

(v) We have: $\underline{p}=\inf_{w,\theta}p(w,\theta)>0$ *and* $\bar{p}=\sup_{w,\theta}p(w,\theta)<\infty$.

Then, the quantities (4.14) *define the inclusion probabilities of a Poisson scheme with probability one, as soon as* $N\leq n\underline{p}/\bar{p}$. *In addition, assumption 7 is fulfilled with*

$$\Gamma^*=\mathbb{E}\left[p(W,\theta^*)\right]\mathbb{E}\left[\frac{1}{p(W,\theta^*)}\nabla l(Z,\theta^*)\nabla l(Z,\theta^*)^T\right].$$

We are now ready to state the main result of this section, which illustrates the trade-off between (asymptotic) generalization and optimization errors, ruled by the limit behavior of $n\gamma_t/N$.

Theorem 4.9. *Suppose that Assumptions 4, 5, 7 are fulfilled and that the rate* γ_t *satisfies the condition of Theorem 4.4 with* $\alpha<1$ *(and thus* $\eta=0$*). Assume that the symmetric positive semi-definite matrix* $H^*=\mathbb{E}[\nabla^2 l(Z,\theta^*)]$ *is invertible, set*

$$\Lambda^*=(H^*)^{-1}\mathbb{E}[\nabla l(Z,\theta^*)\nabla l(Z,\theta^*)^T](H^*)^{-1}$$

and denote by Σ^* *the unique solution of the Lyapunov equation:* $H^*\Sigma+\Sigma H^*=\Gamma^*$. *The assertions below hold true.*

(i) If $\lim_{n,N,t\to+\infty}\frac{n}{N}\gamma_t=+\infty$, *then we have the convergence in distribution:*

$$\lim_{n,N\to\infty}\left\{\lim_{t\to\infty}\sqrt{N/\gamma_t}\left(\theta_n(t)-\theta^*\right)\right\}=\mathcal{N}(0,\Sigma^*).$$

(ii) If $\lim_{n,N,t\to+\infty}\frac{n}{N}\gamma_t=0$, *then we have the convergence in distribution:*

$$\lim_{n,N,t\to+\infty}\sqrt{n}\left(\theta_n(t)-\theta^*\right)=\mathcal{N}(0,\Lambda^*).$$

(iii) If $\lim_{n,N,t\to+\infty}\frac{n}{N}\gamma_t=c>0$, *then we have the convergence in distribution:*

$$\lim_{n,N\to\infty}\left\{\lim_{t\to\infty}\sqrt{n}\left(\theta_n(t)-\theta^*\right)\right\}=\mathcal{N}(0,\Lambda^*+c\Sigma^*).$$

We first point out that, in contrast to case (ii) where they can be swapped, the limits involved in cases (i) and (iii) must be taken sequentially: assertion (i) (respectively, assertion (ii)) describes the asymptotic regime for large values of n and N, the number t of HTGD iterations is such that $1/\gamma_t\ll n/N$ (respectively, such that $1/\gamma_t\sim n/N$). In the asymptotic regime (i), corresponding to the 'Big Data' setup, the optimization error rules the limit behavior of the HTGD estimator, whereas the estimation error determines the asymptotic covariance structure in case (ii). Case (iii) corresponds to the situation where both terms impact the limit distribution. Just like in Corollary 4.5 for the conditional analysis, the asymptotic distribution of the error can be straightforwardly deduced from the Central Limit Theorem above by means of the delta method.

Corollary 4.10. *Suppose that the assumptions of Theorem 4.9 are fulfilled. Let U be a d-dimensional Gaussian centered r.v. with the identity as covariance matrix. The assertions below hold true.*

(i) If $\lim_{n,N,t\to+\infty} \frac{n}{N}\gamma_t = +\infty$*, then we have the convergence in distribution:*

$$\lim_{n,N\to\infty} \left\{ \lim_{t\to\infty} \frac{N}{\gamma_t} \left(L(\theta_n(t)) - L(\theta^*) \right) \right\} = \frac{1}{2} U^T \Sigma^{*1/2} H^* \Sigma^{*1/2} U.$$

(ii) If $\lim_{n,N,t\to+\infty} \frac{n}{N}\gamma_t = 0$*, then we have the convergence in distribution:*

$$\lim_{n,N,t\to+\infty} n \left(L(\theta_n(t)) - L(\theta^*) \right) = \frac{1}{2} U^T \Lambda^{*1/2} H^* \Lambda^{*1/2} U.$$

(iii) If $\lim_{n,t\to+\infty} \frac{n}{N}\gamma_t = c > 0$*, then we have the convergence in distribution:*

$$\lim_{n,N\to\infty} \left\{ \lim_{t\to\infty} n \left(L(\theta_n(t)) - L(\theta^*) \right) \right\} = \frac{1}{2} U^T (\Lambda^* + c\Sigma^*)^{1/2} H^* (\Lambda^* + c\Sigma^*)^{1/2} U.$$

4.6 Illustrative Numerical Experiments

For illustration purpose, this section shows how the results previously established apply to two problems by means of simulation experiments. For both examples, the performance of the HTGD algorithm is compared with that of a basic SGD strategy with the same (mean) sample size.

4.6.1 Linear logistic regression

Consider the linear logistic regression model corresponding to Example 4.2 with $\theta = (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^d$ and $h(x, \theta) = \alpha + \beta^T x$ for all $x \in \mathbb{R}^d$. Let X' be a low dimensional marginal vector of the input r.v. X, of dimension $d' \ll d$ say, so that one may write $X = (X', X'')$ as well as $\beta = (\beta', \beta'')$ in a similar manner. The problem of fitting the parameter θ through conditional MLE corresponds to the case

$$l((x, y), \theta) = -\log \left(\frac{e^{\alpha+\beta^T x}(y+1)/2 + (1-y)/2}{1 + e^{\alpha+\beta^T x}} \right).$$

We propose to implement the HTGD with $\widetilde{p}((x', y), \theta) \propto \|\nabla l'((X, Y), \theta)\|$ as link function, where

$$l'((x, y), \theta) = -\log \left(\frac{e^{\alpha+\beta'^T x'}(y+1)/2 + (1-y)/2}{1 + e^{\alpha+\beta'^T x'}} \right).$$

In order to illustrate the advantages of the HTGD technique for logistic regression, we considered the toy numerical model with parameters $d = 11$ and $\theta = (\alpha, \beta_1, \ldots, \beta_{10}) = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)$, the 10 input variables being independent, uniformly distributed on $(0, 1)$. The maximum likelihood estimators of θ were computed using the HTGD and SGD (mini-batch) . In order to compare them, the same number of iterations was chosen in each situation and a learning rate proportionnal to $1/\sqrt{t}$ was considered. As a first go, we

drew a single sample of size $n = 5000$ on which the two algorithms were performed for 2000 iterations. Two sub-sample sizes were considered : $N = 10$ and $N = 100$. As can be seen on Fig. 4.2, while virtually equivalent in terms of computation time, thus taking a larger sample improves the efficiency of the HTGD. It also appears to reach a better level of precision in less steps than both competitors, a phenomenon that is consistent on all 11 coordinates of θ.

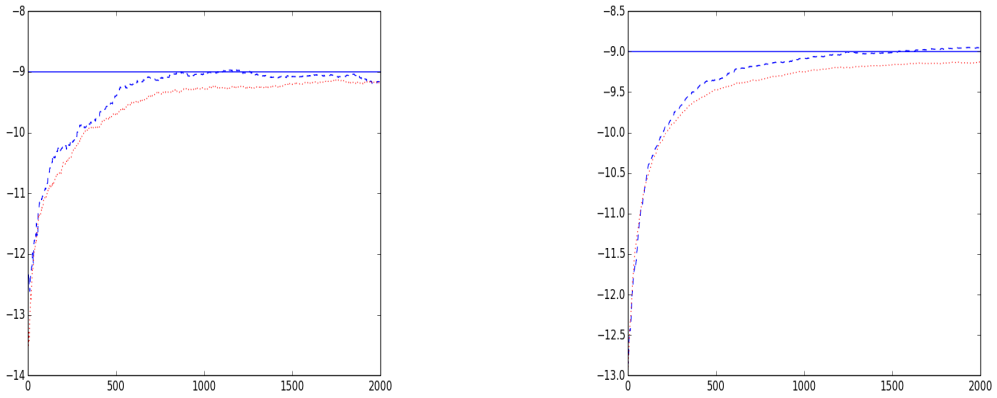


FIGURE 4.1: Evolution of the estimator of β_5 with the number of iterations in the HTGD (solid), mini-batch SGD (dotted) and GD (dashed) algorithms with $N = 10$ (left) and $N = 100$ (right).

So as to account for the randomness due to the data, we then simulated 50 samples according to the model for two population sizes, $n = 500$ and $n = 1000$. For both the HTGD and the mini-batch SGD algorithms, a sub-sample size of 20 was chosen. As shown in Table 4.1, the HTGD seems to be more robust to data randomness than SGD and GD. It is not surprising, since the sampling phase selects the most informative observations relative to the gradient descent, which makes HTGD less sensitive to the possible noise. It also provides more precise estimates, as illustrated by the results in Table 4.3.

	$n = 500$	$n = 1000$
HTGD	1.52	1.45
SGD	2.21	2.09

TABLE 4.1: Mean standard deviations of the final estimates of $\theta(= -9)$ across the 50 simulations

	Min.	Median	Max.	Mean	S.D.
HTGD					
θ_5	-9.5	-8.7	-7.8	-8.6	1.45
θ_6	13.3	14.6	15.9	14.5	1.52
SGD					
θ_5	-9.9	-8.2	-7.4	-8.2	2.09
θ_6	12.7	13.9	16.6	15.2	2.21

TABLE 4.2: Statistics on the global behavior of the final estimates of β_5 and β_6 across the 50 simulations

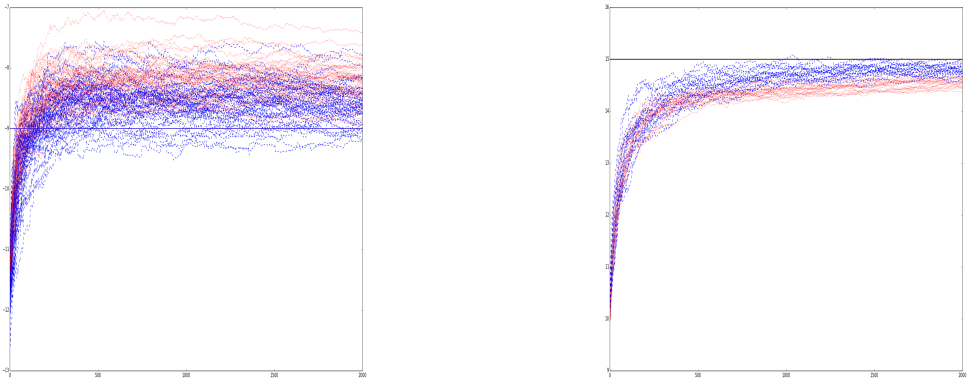


FIGURE 4.2: 50 trajectories of the estimator of β_5 with the number of iterations in the HTGD (solid), mini-batch SGD (dotted) over 50 populations (left) and of θ_6 over 1 populations (right).

4.6.2 The Symmetric Model

Consider now an i.i.d. sample $(X_1, X_2, \ldots, X_n)$ drawn from an unknown probability distribution on $\mathbb{R}^d$, supposed to belong to the semi-parametric collection $\{P_{\theta,f}, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^d$, dominated by some σ-finite measure λ. The related densities are denoted by $f(x-\theta)$, where $\theta \in \Theta$ is a location parameter and a $f(x)$ a (twice differentiable) density, symmetric about 0, *i.e.*$f(x) = f(-x)$. The density f is unknown in practice and may be multimodal. For simplicity, we assume here that $\Theta \subset \mathbb{R}$ but similar arguments can be developed when $d > 1$. For such a general semi-parametric model, it is well-known that neither the sample mean nor the median (if, for instance, the distribution does not weight the singleton $\{0\}$) are good candidates for estimating the location parameter θ. In the semiparametric literature this model is referred to as the *symmetric model*, see Bickel et al. (1993). It is known that the tangent space (*i.e.* the set of scores) with respect to the parameter of interest θ and that with respect to the nuisance parameter are orthogonal. The global tangent space at $P_{\theta,f}$ is given by

$$T_L\left[P_{\theta,f}, \mathbb{P}\right] = \left\{ c\frac{f'(x-\theta)}{f(x-\theta)} + h(x-\theta); c \in \mathbb{R},\ h \in \dot{\mathbb{P}}_2 \right\},$$

where $\dot{\mathbb{P}}_2$ is the tangent space with respect to the nuisance parameter:

$$\dot{\mathbb{P}}_2 = \left\{ h : \mathbb{E}_{P_{\theta,f}}[h(X)] = 0, h(x) = h(-x) \text{ and } \mathbb{E}_{P_{\theta,f}}[h^2(X)] < \infty \right\}.$$

Orthogonality simply results from the fact that $f'(x)$ is an odd function and implies that the parameter θ can be adaptively estimated, as if the density $f(x)$ was known, refer to Bickel et al. (1993) for more details. In practice $f(x)$ is estimated by means of some symmetrized kernel density estimator. Given a Parzen-Rosenblatt kernel $K(x)$ (*e.g.* a Gaussian kernel) for instance, consider the estimate

$$\widetilde{f}_{\theta,n}(x) = \frac{1}{Nh_n} \sum_{i=1}^{n} K\left(\frac{x - (X_i - \theta)}{h_n} \right),$$

where $h_N > 0$ is the smoothing bandwidth, and form its symetrized version (which is an even function)

$$\widehat{f}_{\theta,n}(x) = \frac{1}{2}\left(\widetilde{f}_{\theta,n}(x) + \widetilde{f}_{\theta,n}(-x) \right).$$

The related score is given by

$$\widehat{s}_n(x,\theta) = \frac{d}{d\theta}\widehat{f}_{\theta,n}(x)/\widehat{f}_{\theta,n}(x).$$

In order to perform maximum likelihood estimation approximately, one can try to implement a gradient descent method to get an efficient estimator of θ. For instance, for a reasonable sample size n, it is possible to show that, starting for instance from the empirical median θ_0 with an adequate choice of the rate γ_t, the sequence

$$\theta_n(t) = \widehat{\theta}(t-1) + \gamma_t \frac{1}{n}\sum_{j=1}^{n} \widehat{s}_n(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1))$$

converges to the true MLE. The complexity of this algorithm is typically of order $2T \times n^2$ if $T \geq 1$ is the number of iterations, due the tedious computations to evaluate the kernel density estimator (and its derivatives) at all points $X_i - \widehat{\theta}(t-1)$. It is thus relevant in this case to try to reduce it by means of (Poisson) survey sampling. The iterations of such an algorithm would be then of the form

$$\begin{aligned} \theta_n(t) &= \widehat{\theta}(t-1) + \gamma_t \frac{1}{n}\sum_{j=1}^{n} \frac{\varepsilon_j}{p_j}\widehat{s}_n(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1)), \\ \sum_{j=1}^{n} p_j &= N. \end{aligned}$$

As shown in section 4.4.3, the optimal choice would be to choose p_j proportional to $|\widehat{s}_n(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1))|$ at the t-th iteration:

$$p_j^*\left(\widehat{\theta}(t-1)\right) = \frac{N|\widehat{s}_n(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1))|}{\sum_{i=1}^{n}|\widehat{s}_n(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1))|}. \tag{4.15}$$

Unfortunately this is not possible because s is unknown and replacing $s(x-\theta)$ by $\widehat{s}_n(x-\widehat{\theta}(t-1), \widehat{\theta}(t-1))$ in (4.15) yields obvious computational difficulties. For this reason, we suggest to use the (much simpler) Poisson weights:

$$p_j(\theta) = N|X_j - \theta|/\sum_{j=1}^{n}|X_j - \theta|.$$

Fig. 4.4 depicts the performance of the HTGD algorithm when $\theta = 0$ and $f(x)$ is a balanced mixture of two Gaussian densities with means 4 and -4 respectively and same variance $\sigma^2 = 1$, compared to that of the usual SGD method. Based on a population sample of size $n = 1000$, the HTGD and SGD methods have been implemented with $N = 10$ and $T = 3000$ iterations, whereas 30 iterations have been made for the basic GD procedure (with $N = n = 1000$) so that the number of gradient computations is of the same order for each method. For each instance of the algorithms we took θ_0 equal to the median of the population.

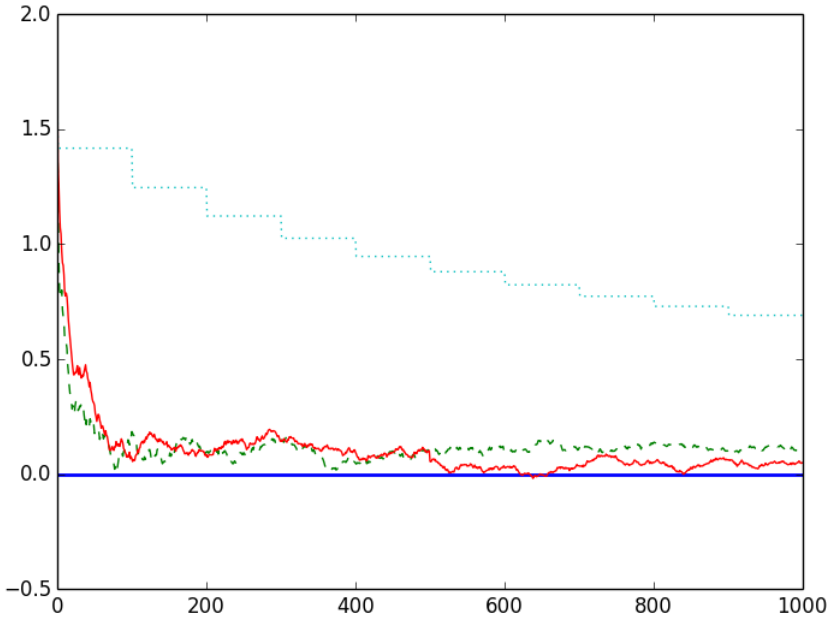


FIGURE 4.3: Evolution of the estimator of the location parameter $\theta = 0$ of the balanced Gaussian mixture with the number of iterations in the HTGD (solid red), mini-batch SGD (dashed green) and GD (dotted blue) algorithms

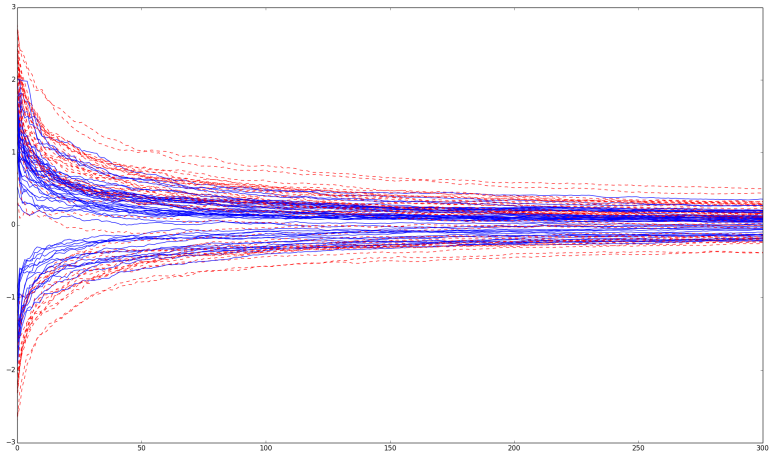


FIGURE 4.4: Evolution of the estimator of the location parameter $\theta = 0$ of the balanced Gaussian mixture with the number of iterations in the HTGD (solid blue) and mini-batch SGD (dashed red) algorithms over 50 populations

	Min.	Median	Max.	Mean	S.D.
HTGD					
θ	-0.35	0.006	0.29	0.014	0.16
SGD					
θ	-0.38	-0.036	0.42	0.025	0.22
GD					
θ	-0.52	-0.162	0.70	0.20	0.45

TABLE 4.3: Statistics on the global behavior of the final estimates of the location parameter across the 50 simulations

4.7 Conclusion

In this article, we have shown how survey sampling can be used in order to improve the accuracy of the stochastic gradient descent method in M-estimation, while preserving the complexity of the procedure. Beyond theoretical limit results, the approach we promote is illustrated by promising numerical experiments. Whereas massively parallelized/distributed approaches combined with random data splitting are now receiving much attention in the Big Data context, the present chapter explores a possible alternative way of scaling up statistical learning methods, based on gradient descent techniques.

4.8 Technical Proofs

4.1.1 Proof of Theorem 4.3

The conditional consistency of the HTGD algorithm described in Section 4.3 is obtained by applying Theorem 13 in Delyon (2000) (or Theorem 2.2 in Chapter 5 of Kushner & Yin (2010) among other references). Specifically, it states that if the following conditions are fulfilled, then $\theta_n(t)$ converges as $t \to +\infty$ to some $\theta_n^* \in \mathcal{L}_n$ with probability 1:

- $\sum_{t\geq 1} \gamma_t = +\infty$ and $\sum_{t\geq 0} \gamma^2(t) = +\infty$, which was assumed,
- $\theta_n(t)$ remains in a compact subset of Θ for all $t \geq 0$, which was also assumed,
- $\theta \in \Theta \mapsto -L_n(\theta)$ and $\theta \in \Theta \mapsto \nabla \widehat{L}_n(\theta)$ are continuous, which is guaranteed by Assumption 4-(i),
- $\mathcal{L}_n$ is finite, which corresponds to Assumption 4-(iii),
- for any compact subset $\mathcal{K} \subset \Theta$ we have that $\sup_{\theta\in\mathcal{K}} \mathbb{E}\left(\|\ell_{R_n}(\theta)\|^2 \mid \mathcal{D}_n\right) < +\infty$ with probability 1, wich we shall now check.

Let $\mathcal{K}$ be a compact subset of Θ, then

$$\sup_{\theta\in\mathcal{K}} \mathbb{E}\left(\|\ell_{R_n}(\theta)\|^2 \mid \mathcal{D}_n\right) = \sup_{\theta\in\mathcal{K}} \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\,\pi_j(\theta)} \nabla l(Z_i,\theta)^T \nabla l(Z_j,\theta)$$
$$\leq \left(\frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in\mathcal{K}} \frac{\|\nabla l(Z_i,\theta)\|}{\pi_i(\theta)}\right)^2,$$

which is finite with probability 1 by virtue of Assumption 4-(ii). □

4.1.2 Proof of Theorem 4.4

Our conditional Central Limit Theorem results from Theorem 1 in Pelletier (1998), the applicability of which needs to be checked.

First of all, rewrite the algorithm sequence as

$$\theta_n(t+1) = \theta_n(t) - \gamma_t \nabla\widehat{L}_n(\theta_t) + \gamma_t\,\xi_n(t+1),$$

where $\xi_n(t+1) := \nabla\widehat{L}_n(\theta_t) - \ell_{R_n}(\theta_n(t))$. This way, $-\nabla\widehat{L}_n(\theta_t)$ appears as the *mean field* of the algorithm and $\xi_n(t+1)$ as a *noise term*. Now consider the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 1}$ where for each $t \geq 1$, $\mathcal{F}_t$ is the σ-field generated by the indicator vectors $\boldsymbol{\epsilon}_n^{(1)} \ldots, \boldsymbol{\epsilon}_n^{(t-1)}$ and by $\mathcal{D}_n$. Then Assumption 1-(ii) guarantees that $\{\xi_n(t)\}_{t\geq 1}$ is a sequence of increments of a q-dimensional square integrable martingale adapted to the filtration $\mathcal{F}$: for all $t \geq 1$ we have both $\mathbb{E}[\xi_n(t+1) \mid \mathcal{F}_t] = 0$ and

$$\mathbb{E}\left[\|\xi_n(t+1)\|^2 \mid \mathcal{F}_t\right] =$$
$$\frac{1}{n^2}\sum_{i,j=1}^{n}\left(\frac{\pi_{i,j}(\theta_n(t))}{\pi_i(\theta_n(t))\,\pi_j(\theta_n(t))} - 1\right) \nabla l(Z_i,\theta_n(t))^T \nabla l(Z_j,\theta_n(t))$$
$$\leq \left(\frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in\mathcal{K}} \frac{\|\nabla l(Z_i,\theta)\|}{\pi_i(\theta)}\right)^2 < +\infty.$$

Given this representation, our result is assured by Theorem 1 in Pelletier (1998) provided that the following conditions hold true:

- $\nabla\widehat{L}_n(\theta_n^*) = 0$, which was assumed,
- on a neighborhood $\mathcal{V}$ of θ_n^* we have $\nabla\widehat{L}_n(\theta) = H_n\,(\theta - \theta_n^*) + O(\|\theta - \theta_n^*\|^2)$, which results from a simple Taylor expansion made possible by Assumption 5-(i),
- $-H_n$ is a stable $q \times q$ matrix (the largest real part of its eigenvalues is negative), which corresponds to Assumption 5-(ii),
- γ is regularly varying with index $-\alpha \in (-1, 0]$ or $\gamma_t = \gamma_0/t$ with $\gamma_0 > 1/(2l)$ for all $t \geq 1$, which was also assumed,
- (A) $\sup_{t\geq 0} \mathbb{E}\left(\|\xi_n(t+1)\|^b \mid \mathcal{F}_t\right) \mathbb{I}\{\theta_n(t) \in \mathcal{V}\} < +\infty$ almost-surely for any $b > 2$, which we shall verify,

- (B) $\mathbb{E}\left(\xi_n(t+1)\,\xi_n(t+1)^T \mid \mathcal{F}_t\right) \to \Gamma$ almost-surely on $\mathcal{E}(\theta_n^*)$ as $t \to +\infty$, with Γ a positive-definite deterministic matrix, which also needs to be checked.

Let us start with condition (A). Since $0 \leq \|\xi_n(t+1)\| \leq \frac{1}{n}\sum_{i=1}^{n} \frac{\|\nabla l(Z_i, \theta_n(t))\|}{\pi_i(\theta_n(t))}$ almost-surely for all $t \geq 0$, then for any $b > 2$

$$\sup_{t\geq 0} \mathbb{E}\left(\|\xi_n(t+1)\|^b \mid \mathcal{F}_t\right) \mathbb{I}\{\theta_n(t) \in \mathcal{V}\} \leq \frac{1}{n}\sum_{i=1}^{n} \left(\sup_{\theta\in\mathcal{V}} \frac{\|\nabla l(Z_i,\theta)\|}{\pi_i(\theta)}\right)^b < +\infty$$

with probability 1, by Assumption 4-(ii).

Regarding condition (B), we have $\mathbb{E}[\xi_n(t+1)\,\xi_n(t+1)^T \mid \mathcal{F}_t] = \Gamma_n(\theta_n(t))$ for all $t \geq 1$, where for any $\theta \in \Theta$,

$$\Gamma_n(\theta) = \frac{1}{n^2}\sum_{i,j=1}^{n} \left(\frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\,\pi_j(\theta)} - 1\right) \nabla l(Z_i,\theta)\,\nabla l(Z_j,\theta)^T.$$

By virtue of the continuity assumptions of the inclusion probabilities (Assumption 5-(iii)) and of the gradient (Assumption 4-(i)), given the population data $\mathcal{D}_n$ we have the almost-sure convergence $\Gamma_n(\theta_n(t)) \to \Gamma_n^* = \Gamma_n(\theta_n^*)$ on the event $\mathcal{E}(\theta_n^*)$ as $t \to +\infty$. The limit matrix is, indeed, positive-definite and deterministic (given $\mathcal{D}_n$). □

4.1.3 Proof of Proposition 4.6

Let us start by calculating $\|\Sigma_{\mathbf{p}_n}^{1/2}\|^2$ for some collection of positive Poisson inclusion probabilities $\mathbf{p}_n = \{\mathbf{p}_n(\theta)\}_{\theta\in\Theta}$. In the case where $\eta = 0$, since H_n is invertible by Assumption 5-(ii), the Lyapunov equation (4.10) can be rewritten as

$$\Sigma_{\mathbf{p}_n} + H_n^{-1}\Sigma_{\mathbf{p}_n}H_n = H_n^{-1}\Gamma_n^*,$$

which implies that $\left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|^2 = \frac{1}{2}\,Tr\left(H_n^{-1}\,\Gamma_n^*\right) = \frac{1}{2}\,Tr\left(G_n\,\Gamma_n^*\,G_n^T\right)$. Now recall that θ_n^* is a stationary point, *i.e.* $\nabla\widehat{L}_n(\theta_n^*) = 0$, and that we are considering a Poisson scheme (with independent inclusion variables). Thus,

$$\Gamma_n^* = \frac{1}{n^2}\sum_{i=1}^{n} \left(\frac{1}{p_i(\theta_n^*)} - 1\right) \nabla l(Z_i,\theta_n^*)\,\nabla l(Z_i,\theta_n^*)^T$$

and then

$$\left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|^2 = \frac{1}{2}\,\frac{1}{n^2}\sum_{i=1}^{n} \left(\frac{1}{p_i(\theta_n^*)} - 1\right) \|G_n\,\nabla l(Z_i,\theta_n^*)\|^2. \tag{4.1}$$

Let us now turn to the definition of an optimal collection of Poisson plans. Using the Lagrange multiplier method like in subsection 4.4.1, we find that any $\mathbf{p}_n$ minimizing (4.1) must satisfy the equalities

$$p_i(\theta_n^*) = N\,\frac{\|G_n\,\nabla l(Z_i,\theta_n^*)\|}{\sum_{j=1}^{n}\|G_n\,\nabla l(Z_j,\theta_n^*)\|}, \qquad i \in \mathcal{U}_n.$$

This is the case, in particular, of the collection $\mathbf{p}_n^*$ defined in Proposition 4.6. Condition (4.12) and the positive-definiteness of H_n (Assumption 5-(ii)) ensure that $p_i^*(\theta)$ is almost-surely in $(0, 1]$ for all $\theta \in \Theta$ and $i \in \mathcal{U}_n$. □

4.1.4 Proof of Proposition 4.7

Let us start by proving the first assertion. Using Eq. (4.1) with the corresponding inclusion probabilities, we immediately obtain

$$\left\|\Sigma_{\widetilde{\mathbf{p}}_N}^{1/2}\right\|^2 - \left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|^2 = \frac{c_n(\theta_n^*)}{2\,N},$$

which is positive by Assumption 6.

Turning to the second assertion, observe that

$$\begin{aligned}\left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|^2 - \left\|\Sigma_{\mathbf{p}_n^*}^{1/2}\right\|^2 &= \left\|\Sigma_{\mathbf{p}_n}^{1/2}\right\|^2 - \left\|\Sigma_{\widetilde{\mathbf{p}}_N}^{1/2}\right\|^2 + \left\|\Sigma_{\widetilde{\mathbf{p}}_N}^{1/2}\right\|^2 - \left\|\Sigma_{\mathbf{p}_n^*}^{1/2}\right\|^2 \\ &= \frac{1}{2\,N}\left\{\sigma_n^2(\theta_n^*) - c_n(\theta_n^*)\right\}.\end{aligned}$$

By definition of $\mathbf{p}_n^*$ (see Proposition 4.6), this quantity is always nonnegative. □

4.1.5 Proof of Proposition 4.8

Consider a Poisson sampling plan with inclusion probabilities as in ((4.14)). We shall prove that Assumption 7 is fulfilled by establishing the asymptotic convergences (as $n, N \to +\infty$) of the three averages in brackets that appear in the following decomposition:

$$\begin{aligned}N\,\Gamma_n^* = \left[\frac{1}{n}\sum_{i=1}^n p(W_i, \theta_n^*)\right] \times \left[\frac{1}{n}\sum_{i=1}^n \frac{1}{p(W_i, \theta_n^*)}\,\nabla l(Z_i, \theta_n^*)\nabla l(Z_i, \theta_n^*)^T\right] \\ - \frac{N}{n}\left[\frac{1}{n}\sum_{i=1}^n \nabla l(Z_i, \theta_n^*)\nabla l(Z_i, \theta_n^*)^T\right]\end{aligned}$$

Recall that $(Z_1, W_1), \ldots, (Z_n, W_n)$ were taken as independent copies of some generic r.v. (Z, W), which is thus independent from θ_n^*, for any $n \in \mathbb{N}^*$. The respective distributions of Z, W and (Z, W) are denoted by $\mathbb{P}_Z$, $\mathbb{P}_W$ and $\mathbb{P}_{Z,W}$.

First average We shall verify that the first term in brackets tends to $\mathbb{E}\left[p(W, \theta^*)\right]$ almost-surely as $n \to +\infty$. For any $n \in \mathbb{N}^*$ we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n p(W_i, \theta_n^*) - \mathbb{E}[p(W, \theta^*)]\right| \leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n p(W_i, \theta) - \mathbb{E}[p(W, \theta)]\right| \\ + \left|\mathbb{E}[p(W, \theta_n^*)] - \mathbb{E}[p(W, \theta^*)]\right|.\end{aligned}$$

Thus, it suffices to check that both the supremum and the difference of expectations above (almost-surely) vanish as n grows.

The supremum can be controlled using Lemma 3.10 in van de Geer (2000). The parameter space Θ is assumed to be a compact metric space and the map $\theta \in \Theta \mapsto p(w, \theta)$ is supposed

to be continuous for all $w \in \mathbb{R}^{d'}$. In addition, since the link function p was chosen to be bounded by some finite positive constant $\bar{p}$, the envelope $w \in \mathbb{R}^{d'} \mapsto \sup_{\theta\in\Theta} |p(w,\theta)|$ is $\mathbb{P}_W$-integrable. By virtue of the aforementioned Lemma, these conditions are sufficient to obtain the uniform law of large numbers: as $n \to +\infty$, with probability one,

$$\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{i=1}^{n} p(W_i,\theta) - \mathbb{E}[p(W,\theta)] \right| \to 0.$$

Let us now turn to the difference of expectations. Fix some $w \in \mathbb{R}^{d'}$. The empirical risk minimizer θ_n^* is assumed to be strongly consistent (condition (i)) and the mapping $\theta \in \Theta \mapsto p(w,\theta)$ to be continuous. Thus, by the continuous mapping theorem, $p(w,\theta_n^*)$ converges almost-surely to $p(w,\theta^*)$, as $n \to +\infty$. Then, because p is a bounded function, by the dominated convergence theorem we also have $\mathbb{E}[p(w,\theta_n^*)] \to p(w,\theta^*)$ as $n \to +\infty$. Next, for any $n \in \mathbb{N}^*$, the independence between W and θ_n^* implies $\mathbb{E}[p(W,\theta_n^*)] = \int_{\mathbb{R}^{d'}} \mathbb{E}[p(w,\theta_n^*)]\, \mathbb{P}_W(dw)$. Applying the dominated convergence theorem to this last integral, we finally obtain that $\mathbb{E}[p(W,\theta_n^*)] \to \mathbb{E}[p(W,\theta^*)]$ as $n \to +\infty$.

Second average The second term in brackets is a $q \times q$ matrix, the convergence of which shall be established element-wise. Let $(k,h) \in \{1,\dots,q\}^2$ and define the function $\Psi_{k,h} : (z,\theta) \in \mathbb{R}^d \times \Theta \mapsto (\partial l/\partial\theta_k)(z,\theta) \times (\partial l/\partial\theta_h)(z,\theta)$. The element at the intersection of the k-th row and h-th column of this matrix is

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{p(W_i,\theta_n^*)}\, \Psi_{k,h}(Z_i,\theta_n^*).$$

Using the same reasoning as before, this quantity can be shown to converge almost-surely, as $n, N \to +\infty$, to $\mathbb{E}\left[p(W,\theta^*)^{-1}\, \Psi_{k,h}(Z,\theta^*)\right]$. We only need to check that

1. the map $\theta \in \Theta \mapsto p(w,\theta)^{-1}\, \Psi_{k,h}(z,\theta)$ is continuous for any (z,w) in $\mathbb{R}^d \times \mathbb{R}^{d'}$,
2. the envelope $(z,w) \in \mathbb{R}^d \times \mathbb{R}^{d'} \mapsto \sup_{\theta\in\Theta} |p(w,\theta)^{-1}\, \Psi_{k,h}(z,\theta)|$ is $\mathbb{P}_{Z,W}$-integrable,
3. the class of random variables $\{p(W,\theta_n^*)^{-1}\, \Psi_{k,h}(Z,\theta_n^*)\}$ is uniformly $\mathbb{P}_{Z,W}$-integrable.

Condition (a) is guaranteed by the construction of p and by Assumption 4-(i).

4.1.6 An Intermediary Result

Before beginning to prove Theorem 4.9, we first establish the following lemma. It describes the limit behavior of the solution of the Lyapunov equation ((4.10)) as $n, N \to +\infty$.

Lemma 4.11. *Suppose that the assumptions of Theorem 4.9 are fulfilled. Then as n, N tend to $+\infty$, we have:*

$$N\Sigma_{\boldsymbol{\pi}_n} \to \Sigma^* \text{ in probability.}$$

Proof. Observe first that it follows from $H_n\Sigma_{\boldsymbol{\pi}_n} + \Sigma_{\boldsymbol{\pi}_n}H_n = \Gamma_n^*$ that

$$\|\Gamma_n^*\|^2 = 2\|H_n\Sigma_{\boldsymbol{\pi}_n}\|^2 + \underbrace{2Tr(H_n\Sigma_{\boldsymbol{\pi}_n}H_n\Sigma_{\boldsymbol{\pi}_n})}_{\geqslant 0},$$

Hence, we have:

$$\|\Gamma_n^*\| \geqslant \sqrt{2}\|H_n\Sigma_{\boldsymbol{\pi}_n}\|.$$

We deduce from this inequality combined with assumptions 7 and the fact that $H_n^{-1} = O_{\mathbb{P}}(1)$ as $n \to \infty$ (this can be deduced from the LLN $H_n \to H^*$ and the hypothesis that the Hessian matrices H_n and H^* are invertible) that

$$\Sigma_{\boldsymbol{\pi}_n} = O_{\mathbb{P}}(1/N) \text{ as } n \to \infty. \tag{4.2}$$

Since $H^*\Sigma^* + \Sigma^*H^* = \Gamma^*$ and $H_n\Sigma_{\boldsymbol{\pi}_n} + \Sigma_{\boldsymbol{\pi}_n}H_n = \Gamma_n^*$, we have:

$$\begin{aligned}\Gamma^* - N\Gamma_n^* = H^*(\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n}) + (\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n})H^* + \\ + N(H_n - H^*)\Sigma_{\boldsymbol{\pi}_n} + N\Sigma_{\boldsymbol{\pi}_n}(H_n - H^*).\end{aligned} \tag{4.3}$$

Combining (4.3) with

$$\begin{aligned}\|H^*(\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n})\| &= \|(\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n})H^*\| \\ &\leqslant \frac{1}{\sqrt{2}}\|H^*(\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n}) + (\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n})H^*\|\end{aligned} \tag{4.4}$$

we easily get

$$\|H^*(\Sigma^* - N\Sigma_{\boldsymbol{\pi}_n})\| \leqslant \frac{1}{\sqrt{2}}\|\Gamma^* - N\Gamma_n^*\| + \sqrt{2}\|N\Sigma_{\boldsymbol{\pi}_n}(H_n - H^*)\|. \tag{4.5}$$

By virtue of the LLN, we have $H_n - H^* \to 0$ almost surely as $n \to \infty$. Combining this with (4.2) and assumption 7 we see that the term on the right hand side of (4.5) converges toward 0 in probability as $n \to \infty$. Combined with the invertibility of H^*, this establishes the desired result. □ □

4.1.7 Proof of Theorem 4.9

Consider the decomposition:

$$\begin{aligned}\theta_n(t) - \theta^* &= \theta_n(t) - \theta_n^* + \theta_n^* - \theta^* \\ &= \sqrt{\frac{\gamma_t}{N}}\sqrt{N/\gamma_t}\,(\theta_n(t) - \theta_n^*) + \frac{1}{\sqrt{n}}\sqrt{n}\,(\theta_n^* - \theta^*) \\ &= \sqrt{\frac{\gamma_t}{N}}\underbrace{\sqrt{N/\gamma_t}\,(\theta_n(t) - \theta_n^*)}_{(1)} \\ &+ \frac{1}{\sqrt{n}}\underbrace{\sqrt{n}\,(\theta_n^* - \theta^*)}_{(2)}.\end{aligned}$$

The term (2) above is asymptotically normal. By virtue of the classical Central Limit Theorem for M-estimators, see Theorem 5.23 in Van der Vaart (2000) for instance, we have:

$$\sqrt{n}\,(\theta_n^* - \theta^*) \Rightarrow \mathcal{N}(0, \Lambda^*) \text{ as } n \to \infty. \tag{4.6}$$

This suffices to establish assertion (ii) since the parameter space Θ is assumed to be compact here. Turning to term (1), holding n and N fixed, Theorem 4.4 claims that, in probability

along the sequence X (respectively, the sequence (X, W)):

$$\sqrt{1/\gamma_t}\Sigma_{\boldsymbol{\pi}_n}^{-1/2}\left(\theta_n(t) - \theta_n^*\right) \Rightarrow Z \text{ as } t \to \infty, \tag{4.7}$$

where Z denotes a q-dimensional centered Gaussian random vector with the identity as covariance, independent from the sequence X (from the sequence (X, W) respectively). Now it follows from Lemma 4.11 combined with the continuity of the application that maps any symmetric positive semi-definite matrix to its square root that

$$(N\Sigma_{\boldsymbol{\pi}_n})^{1/2} \to \Sigma^{*1/2} \text{ in probability, as } n, N \to \infty. \tag{4.8}$$

Given that one may write

$$\sqrt{\frac{N}{\gamma_t}}\left(\theta_n(t) - \theta_n^*\right) = (N\Sigma_{\boldsymbol{\pi}_n})^{1/2}\sqrt{1/\gamma_t}\Sigma_{\boldsymbol{\pi}_n}^{-1/2}\left(\theta_n(t) - \theta_n^*\right),$$

it results from (4.7) and (4.8) that the following convergence in distribution holds true:

$$\lim_{n,N\to\infty}\lim_{t\to\infty}\sqrt{\frac{N}{\gamma_t}}\left(\theta_n(t) - \theta_n^*\right) = \Sigma^{*1/2}Z. \tag{4.9}$$

Assertions (i) and (iii) can be then deduced from (4.9) and (4.6) in a straightforward fashion (using the independence of the limits, regarding (iii)).

4.1.8 Rate Bound Analysis

Here we establish a rate bound for the HTGD algorithm under the additional assumption that the mapping $\theta \mapsto l(z, \theta)$ is convex, referred to as Assumption 8, as we have done in chapter 4. Note that Assumptions 5 and 8 imply that θ_n^* is unique and $\widehat{L}_n$ is l strongly convex on $\mathcal{V}$. For simplicity's sake, we suppose that the strong convexity property holds true on $\mathbb{R}^d$. The following result relies on standard arguments in stochastic approximation, see Nemirovski et al. (2009a), Bach & Moulines (2011a) or Nesterov (2004a). All expectations are taken conditionally upon the observations.

Theorem 4.12. *Under Assumptions 4, 5 and 8 and for a stepsize $\gamma_t = \gamma_0 t^{-\alpha}$ with some constants $\gamma_0 > 0$ and $\alpha \in (1/2, 1]$ (when $\alpha = 1$, take $\gamma_0 > 1/(2l)$), there exists a constant $\widetilde{C}_\alpha < +\infty$ such that: $\forall t \geq 1$,*

$$\mathbb{E}[\|\theta_n(t) - \theta_n^*\|^2] \leq \frac{\widetilde{C}_\alpha}{t^\alpha}. \tag{4.1}$$

Proof. We restrict ourselves to the case $\alpha = 1$ and follow the proof of Bach & Moulines (2011a). By construction, we have

$$\|\theta_n(t+1) - \theta_n^*\|^2 = \|\theta_n(t) - \theta_n^*\|^2 - 2\gamma_t\ell_{R_n}(\theta_n(t))^T(\theta_n(t) - \theta_n^*) + \|\gamma_t\ell_{R_n}(\theta_n(t))\|^2.$$

Since

$$\mathbb{E}[\ell_{R_n}(\theta_n(t))|\mathcal{F}_t] = \nabla\widehat{L}_n(\theta_n(t)),$$

we get

$$\mathbb{E}[|\theta_n(t+1)-\theta_n^*|^2 \mid \theta_n(t)] = \|\theta_n(t)-\theta_n^*\|^2 - 2\gamma_t \nabla \widehat{L}_n(\theta_n(t))^T(\theta_n(t)-\theta_n^*) + \gamma_t^2 \mathbb{E}[\|\ell_{R_n}(\theta_n(t))\|^2 \mid \theta_n(t)].$$

The strong convexity property gives

$$\widehat{L}_n(\theta_n(t)) - \widehat{L}_n(\theta_n^*) \leq \nabla \widehat{L}_n(\theta_n(t))^T(\theta_n(t)-\theta_n^*) - \frac{l}{2}\|\theta_n(t)-\theta_n^*\|^2$$

and

$$\widehat{L}_n(\theta_n^*) - \widehat{L}_n(\theta_n(t)) \leq -\frac{l}{2}\|\theta_n(t)-\theta_n^*\|^2,$$

so that

$$\|\theta_n(t)-\theta_n^*\|^2 \leqslant \nabla \widehat{L}_n(\theta_n(t))^T(\theta_n(t)-\theta_n^*).$$

Combining this inequality with the previous one and taking the expectation, we obtain

$$\mathbb{E}[\|\theta_n(t+1)-\theta_n^*\|^2] \leq \mathbb{E}[\|\theta_n(t)-\theta_n^*\|^2](1-2\gamma_t l) + \gamma_t^2 \mathbb{E}[\|\ell_{R_n}(\theta_n(t))\|^2].$$

Under Assumption 4, we have $\mathbb{E}[\|\ell_{R_n}(\theta_n(t))\|^2] \leq D$ for some constant $D > 0$. Using this bound and iterating the recursion, we finally obtain

$$\mathbb{E}[\|\theta_n(t+1)-\theta_n^*\|^2] \leqslant \mathbb{E}[\|\widehat{\theta}(1)-\theta_n^*\|^2] \prod_{j=1}^{t}(1-2l\gamma(j)) + D\sum_{j=1}^{t}\gamma_t^2 \prod_{k=j+1}^{t}(1-2l\gamma(k))$$

with the convention $\prod_{k=t+1}^{t}(1-2l\gamma(k)) = 1$ We now substitute the expression of γ_t and, using the following classical inequalities

$$1+x \leqslant e^x$$

and

$$\log(t+1) - \log(j+1) \leqslant \sum_{k=j+1}^{t}\frac{1}{k},$$

we get

$$\mathbb{E}\|\theta_n(t+1)-\theta_n^*\|^2 \leqslant \frac{(\mathbb{E}\|\theta_n(1)-\theta_n^*\|^2 + \tilde{D}\sum_{j=1}^{t}\frac{1}{j^{2-2l\gamma_0}})}{(t+1)^{2l\gamma_0}},$$

where $\tilde{D}$ is a positive constant. Since $\gamma_0 > 1/(2l)$, we have

$$\sum_{j=1}^{t}\frac{1}{j^{2-2l\gamma_0}} \leqslant \frac{t^{2l\gamma_0-1}}{2l\gamma_0-1}$$

and we finally obtain the desired bound. □

CHAPTER 5
Stochastic Gradient Descent based on *incomplete* U-Statistics

Abstract In many learning problems, ranging from clustering to ranking through metric learning, empirical estimates of the risk functional consist of an average over tuples (*e.g.*, pairs or triplets) of observations, rather than over individual observations. In this chapter, we focus on how to best implement a stochastic approximation approach to solve such risk minimization problems. We argue that in the large-scale setting, gradient estimates should be obtained by sampling tuples of data points with replacement (*incomplete* U-statistics) instead of sampling data points without replacement (*complete* U-statistics based on subsamples). We develop a theoretical framework accounting for the substantial impact of this strategy on the generalization ability of the prediction model returned by the Stochastic Gradient Descent (SGD) algorithm. It reveals that the method we promote achieves a much better trade-off between statistical accuracy and computational cost. Beyond the rate bound analysis, experiments on AUC maximization and metric learning provide strong empirical evidence of the superiority of the proposed approach.

5.1 Background and Problem Setup

For clarity, we start with recalling the definition of generalized U-statistics and their crucial properties and providing examples of learning problems motivated by various applications where such data functionals are naturally involved. We recall that the variance of any square integrable r.v. Z is denoted by $\sigma^2(Z)$

5.1.1 U-statistics: Definition and Examples

Generalized U-statistics are extensions of standard sample mean statistics, as defined below.

Definition 5.1. Let $K \geq 1$ and $(d_1, \ldots, d_K) \in \mathbb{N}^{*K}$. Let $\mathbf{X}_{\{1,\ldots,n_k\}} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$, $1 \leq k \leq K$, be K independent samples of sizes $n_k \geq d_k$ and composed of i.i.d. random variables taking their values in some measurable space $\mathcal{X}_k$ with distribution $F_k(dx)$ respectively. Let $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$ be a measurable function, square integrable with respect to the probability distribution $\mu = F_1^{\otimes d_1} \otimes \cdots \otimes F_K^{\otimes d_K}$. Assume in addition (without loss of generality) that $H(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)})$ is symmetric within each block of arguments $\mathbf{x}^{(k)}$ (valued in $\mathcal{X}_k^{d_k}$), $1 \leq k \leq K$. The generalized (or K-sample) U-statistic of degrees $(d_1, \ldots, d_K)$ with kernel H, is then defined as

$$U_{\mathbf{n}}(H) = \frac{1}{\prod_{k=1}^{K} \binom{n_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} H\left(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)}\right), \tag{5.1}$$

where $\mathbf{n} = (n_1, \ldots, n_K)$, the symbol $\sum_{I_1} \cdots \sum_{I_K}$ refers to summation over all elements of Λ, the set of the $\prod_{k=1}^{K} \binom{n_k}{d_k}$ index vectors $(I_1, \ldots, I_K)$, I_k being a set of d_k indexes $1 \leq i_1 < \ldots < i_{d_k} \leq n_k$ and $\mathbf{X}_{I_k}^{(k)} = (X_{i_1}^{(k)}, \ldots, X_{i_{d_k}}^{(k)})$ for $1 \leq k \leq K$.

In the above definition, standard mean statistics correspond to the case where $K = 1 = d_1$. More generally when $K = 1$, $U_n(H)$ is an average over all d_1-tuples of observations. Finally, $K \geq 2$ corresponds to the multi-sample situation where a d_k-tuple is used for each sample $k \in \{1, \ldots, K\}$.

The key property of the statistic (5.1) is that it has minimum variance among all unbiased estimates of

$$\mu(H) = \mathbb{E}\left[H\left(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_k}^{(K)}\right)\right] = \mathbb{E}\left[U_n(H)\right].$$

One may refer to Lee (1990a) for further results on the theory of U-statistics. In machine learning, generalized U-statistics are widely used for estimating properties of probability distributions (*e.g.* variance, Gini mean difference), for statistical hypothesis testing (*i.e.* Kendall tau, Mann-Whitney statistic) and used as performance criteria in various problems, such as those listed below.

Clustering. Given a distance $D : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}_+$, the quality of a partition $\mathcal{P}$ of $\mathcal{X}_1$ with respect to the clustering of an i.i.d. sample $X_1, \ldots, X_n$ drawn from $F_1(dx)$ can be assessed through the *within cluster point scatter*:

$$\widehat{W}_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{i<j} D(X_i, X_j) \cdot \sum_{\mathcal{C} \in \mathcal{P}} \mathbb{I}\left\{(X_i, X_j) \in \mathcal{C}^2\right\}. \tag{5.2}$$

It is a one sample U-statistic of degree 2 with kernel $H_{\mathcal{P}}(x, x') = D(x, x') \cdot \sum_{\mathcal{C} \in \mathcal{P}} \mathbb{I}\{(x, x') \in \mathcal{C}^2\}$.

Multi-partite ranking. Suppose that K independent i.i.d. samples $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ with $n_k \geq 1$ and $1 \leq k \leq K$ on $\mathcal{X}_1 \subset \mathbb{R}^p$ have been observed. The accuracy of a scoring function $s : \mathcal{X}_1 \to \mathbb{R}$ with respect to the K-partite ranking is empirically estimated by the rate of concordant K-tuples (sometimes referred to as the *Volume Under the ROC Surface*):

$$\widehat{\mathrm{VUS}}_n(s) = \frac{1}{n_1 \times \cdots \times n_K} \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} \mathbb{I}\left\{s(X_{i_1}^{(1)}) < \cdots < s(X_{i_K}^{(K)})\right\}.$$

The quantity above is a K-sample U-statistic with degrees $d_1 = \ldots = d_K = 1$ and kernel $\bar{H}_s(x_1, \ldots, x_K) = \mathbb{I}\{s(x_1) < \cdots < s(x_K)\}$.

Metric learning. Based on an i.i.d. sample of labelled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ on $\mathcal{X}_1 = \mathbb{R}^p \times \{1, \ldots, J\}$, the empirical pairwise classification performance of a distance $D : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}_+$ can be evaluated by:

$$\widehat{R}_n(D) = \frac{6}{n(n-1)(n-2)} \sum_{i<j<k} \mathbb{I}\left\{D(X_i, X_j) < D(X_i, X_k),\ Y_i = Y_j \neq Y_k\right\} \tag{5.3}$$

which is a one sample U-statistic of degree three with kernel $\tilde{H}_D((x, y), (x', y'), (x'', y'')) = \mathbb{I}\{D(x, x') < D(x, x''),\ y = y' \neq y''\}$.

- **Pairwise classification.** Based on an i.i.d. sample of labelled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ on $\mathcal{X}_1 = \mathbb{R}^d \times \mathbb{R}$, the empirical ranking performance of any antisymmetric ranking rule $r : \mathbb{R}^d \times \mathbb{R}^d \to \{-1, +1\}$ (the quantity $r(x, x')$ being expected to be $+1$ when the label assigned to x is larger than that assigned to x') is evaluated through the *pairwise classification risk*

$$\widehat{P}_n(r) = \frac{2}{n(n-1)} \sum_{i \neq j} \mathbb{I}\left\{r(X_i, X_j) \cdot (Y_i - Y_j) < 0\right\}.$$

  It is a one sample U-statistic of degree 2 with kernel $\tilde{H}_r((x, y), (x', y')) = \mathbb{I}\{r(x, x') \cdot (y - y') < 0\}$.

- **Metric learning.** Suppose that K independent i.i.d. samples $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ with $n_k \geq 1$ and $1 \leq k \leq K$ on $\mathcal{X}_1 \subset \mathbb{R}^d$ are available. The empirical performance of a similarity measure $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ regarding its capacity to discriminate among the K populations when combined with a nearest-neighbor can be evaluated by:

$$\widehat{R}_n(D) = 2 \sum_{k=1}^{K} \frac{1}{n_k(n_k-1)} \sum_{1 \leq i < j \leq n_k} \Big(1 + D(X_i^{(k)}, X_j^{(k)}) - b\Big)_+ - 2 \sum_{1 \leq k < l \leq K} \frac{1}{n_k n_l} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \Big(1 + D(X_i^{(k)}, X_j^{(l)}) - b\Big)_+,$$

  where $b > 0$ is a tuning threshold parameter and z_+ denotes the positive part of any real number z. It is a K sample U-statistic of degrees $d_1 = \ldots = d_K = 2$.

5.2 SGD Implementation based on Incomplete U-Statistics

Let $\Theta \subset \mathbb{R}^q$ with $q \geq 1$ be some parameter space, we consider the risk minimization problem $\min_{\theta \in \Theta} L(\theta)$ with

$$L(\theta) = \mathbb{E}[H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}; \theta)] = \mu(H(.; \theta)),$$

where $H : \prod_{k=1}^{K} \mathcal{X}_k^{d_k} \times \Theta \to \mathbb{R}$ is a convex loss function, the $(X_1^{(k)}, \ldots, X_{d_k}^{(k)})$'s, $1 \leq k \leq K$, are K independent random variables with distribution $F_k^{\otimes d_k}(dx)$ on $\mathcal{X}_k^{d_k}$ respectively so that H is square integrable for any $\theta \in \Theta$. Based on K independent i.i.d. samples $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ with $1 \leq k \leq K$, the empirical version of the risk function is $\theta \in \Theta \mapsto \widehat{L}_n(\theta) = U_n(H(.; \theta))$. So far in this manuscript, we emphasized how important are stochastic method to solve the ERM problem, this is even truer in this case where the sample sizes $n_1, \ldots, n_K$ of the training datasets are such that computing the empirical gradient

$$\widehat{g}_n(\theta) \overset{def}{=} \nabla \widehat{L}_n(\theta) = \left(1 / \prod_{k=1}^{K} \binom{n_k}{d_k}\right) \sum_{I_1} \ldots \sum_{I_K} \nabla H(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)}; \theta) \tag{5.4}$$

just once is intractable due to the number $\#\Lambda = \prod_{k=1}^{K} \binom{n_k}{d_k}$ of terms to be averaged. Note also that trying to adopt a similar approach to what we have done in chapter 3 and 4 is impossible as we would have to maintain a probability distribution over all the $\prod_{k=1}^{K} \binom{n_k}{d_k}$ couples of observations. A possible approach could consist in replacing (5.4) by a (complete) U-statistic computed from subsamples of reduced sizes $n_k' << n_k$, $\{(X_1'^{(k)}, \ldots, X_{n_k'}'^{(k)}) : k = 1, \ldots, K\}$

say, drawn uniformly at random without replacement among the original samples, leading to the following gradient estimator:

$$\tilde{g}_{n'}(\theta) = \frac{1}{\prod_{k=1}^{K} \binom{n'_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} \nabla H(\mathbf{X}'^{(1)}_{I_1}; \mathbf{X}'^{(2)}_{I_2}; \ldots; \mathbf{X}'^{(K)}_{I_K};\ \theta), \tag{5.5}$$

where $\sum_{I_k}$ refers to summation over all $\binom{n'_k}{d_k}$ subsets $\mathbf{X}'^{(k)}_{I_k} = (X'^{(k)}_{i_1},\ \ldots,\ X'^{(k)}_{i_{d_k}})$ related to a set I_k of d_k indexes $1 \leq i_1 < \ldots < i_{d_k} \leq n'_k$ and $\mathbf{n}' = (n'_1,\ \ldots,\ n'_K)$. Provided that the n'_k's are sufficiently small, it may be numerically tractable to average over the $\prod_{k=1}^{K} \binom{n'_k}{d_k}$ terms involved in the definition of (5.5). Although this approach is very natural, one can obtain a better estimate for the same computational cost, as shall be seen below.

5.2.1 Monte-Carlo Estimation of the Empirical Gradient

From a practical perspective, the alternative strategy we propose is of disarming simplicity. It is based on a Monte-Carlo sampling scheme that consists in drawing independently with replacement among the set of index vectors Λ, yielding a gradient estimator in the form of a so-called *incomplete* U-statistic (see Lee (1990a)):

$$\bar{g}_B(\theta) = \frac{1}{B} \sum_{(I_1,\ \ldots,\ I_K) \in \mathcal{D}_B} \nabla H(\mathbf{X}^{(1)}_{I_1},\ \ldots,\ \mathbf{X}^{(K)}_{I_K};\ \theta), \tag{5.6}$$

where $\mathcal{D}_B$ is built by sampling B times with replacement in the set Λ. We point out that the conditional expectation of (5.6) given the K observed data samples is equal to $\widehat{g}_n(\theta)$. The parameter B, corresponding to the number of terms to be averaged, controls the computational complexity of the SGD implementation. Observe incidentally that an incomplete U-statistic is not a U-statistic in general. Hence, as an unbiased estimator of the gradient of the statistical risk $L(\theta)$, (5.6) is of course less accurate than the full empirical gradient (5.4) (*i.e.*, it has larger variance), but this slight increase in variance leads to a large reduction in computational cost. In our subsequent analysis, we will show that for the same computational cost (*i.e.*, taking $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$), implementing SGD with (5.6) rather than (5.5) leads to much more accurate results. We will rely on the fact that (5.6) has smaller variance w.r.t. to $\nabla L(\theta)$ (except in the case where $K = 1 = d_1$), as shown in the proposition below.

Proposition 5.2. *Set $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$. There exists a universal constant $c > 0$, such that we have:*

$$\sigma^2\left(\tilde{g}_{n'}(\theta)\right) \leq c \cdot \sigma_\theta^2 / \sum_{k=1}^{K} n'_k \quad \text{and} \quad \sigma^2\left(\bar{g}_B(\theta)\right) \leq c \cdot \sigma_\theta^2 / \prod_{k=1}^{K} \binom{n'_k}{d_k},$$

for all $\mathbf{n} \in \mathbb{N}^{*K}$, with $\sigma_\theta^2 = \sigma^2(\nabla H(X^{(1)}_1,\ \ldots,\ X^{(K)}_{d_K}; \theta))$. Explicit but lengthy expressions of the variances are given in Lee (1990a).

Remark 5.3. The results of this chapter can be extended to other sampling schemes to approximate (5.4), such as Bernoulli sampling or sampling without replacement in Λ, following the proposal of Janson (1984). For clarity, we focus on sampling with replacement, which is computationally more efficient.

5.2.1.1 A Conditional Performance Analysis

As a first go, we investigate and compare the performance of the SGD methods described above conditionally upon the observed data samples. In this section, all expectations are taken conditionally upon the observations. Given a matrix M, we recall that we defined M^T to be the transpose of M and $\|M\| := \sqrt{Tr(MM^T)}$ to be its Hilbert-Schmidt norm. We assume that the loss function H is l-smooth in θ, i.e its gradient is l-Lipschitz, with $l > 0$. We also restrict ourselves to the case where $\widehat{L}_n$ is α-strongly convex for some deterministic constant α:

$$\widehat{L}_n(\theta_1) - \widehat{L}_n(\theta_2) \leqslant \nabla \widehat{L}_n(\theta_1)^T(x-y) - \frac{\alpha}{2}\|\theta_1 - \theta_2\|^2 \tag{5.7}$$

and we denote by θ_n^* its unique minimizer. We point out that the present analysis can be extended to the smooth but non-strongly convex case, see Bach & Moulines (2011b). A classical argument based on convex analysis and stochastic optimization (see Bach & Moulines (2011b); Nemirovski et al. (2009b) for instance) shows precisely how the conditional variance of the gradient estimator impacts the empirical performance of the solution produced by the corresponding SGD method and thus strongly advocates the use of the SGD variant proposed in Section 5.2.1.

Proposition 5.4. *Consider the recursion $\theta_{t+1} = \theta_t - \gamma_t g(\theta_t)$ where $\mathbb{E}[g(\theta_t)|\theta_t] = \nabla \widehat{L}_n(\theta_t)$, and denote by $\sigma_n^2(g(\theta))$ the conditional variance of $g(\theta)$. With $\gamma_t = \gamma_1/t^\beta$, the following holds.*

1. *If $\frac{1}{2} < \beta < 1$, then:*

$$\mathbb{E}[\widehat{L}_n(\theta_{t+1}) - \widehat{L}_n(\theta_n^*)] \leqslant \frac{\sigma_n^2(g(\theta_n^*))}{t^\beta} \underbrace{\gamma_1 l 2^{\beta-1}\left(\frac{1}{2\alpha} + \frac{l\gamma_1^2}{2\beta-1}\right) + o\left(\frac{1}{t^\beta}\right)}_{C_1}.$$

   and with probability at least $1-\delta$:

$$\widehat{L}_n(\theta_{t+1}) - \widehat{L}_n(\theta_n^*) \leqslant \frac{\sigma_n^2(\theta_n^*)}{t^\beta} C_1 + \sqrt{\frac{D_\beta \log(L/\delta)}{t^\beta}}$$

2. *If $\beta = 1$ and $\gamma_1 > \frac{1}{2\alpha}$, then:*

$$\mathbb{E}[\widehat{L}_n(\theta_{t+1}) - \widehat{L}_n(\theta_n^*)] \leqslant \frac{\sigma_{\mathbf{n}}^2(g(\theta_{\mathbf{n}}^*))}{t+1} \underbrace{\frac{2^{\alpha\gamma_1} l \exp(2\alpha l \gamma_1^2)\gamma_1^2}{(2\alpha\gamma_1 - 1)} + o\left(\frac{1}{t}\right)}_{C_2}.$$

   and with probability at least $1-\delta$:

$$\widehat{L}_n(\theta_{t+1}) - \widehat{L}_n(\theta_n^*) \leqslant \frac{\sigma_n^2(\theta_n^*)}{t} C_2 + \sqrt{\frac{D \log(L/\delta)}{t}}$$

for some constants D and D_β depending on the parameters L, α, γ_1, a_1.

Proposition 5.4 is another illustration of the fact that the convergence rate of SGD is dominated by the variance term and corroborate our approach of chapter 3 and 4 (see Zhao & Zhang (2015); Johnson & Zhang (2013b); Defazio et al. (2014) for instance).

As we have done in theorem 4.4, we can also give the asymptotic behaviour of the algorithm (when $t \to +\infty$), under the following assumptions:

$\mathbf{A_1}$ The function $\widehat{L}_n(\theta)$ is twice differentiable on a neighborhood of θ_n^*.

$\mathbf{A_2}$ The function $\nabla \widehat{L}_n(\theta)$ is bounded.

Let us set $\Gamma = \nabla^2 \widehat{L}_n(\theta_n^*)$, then we have under the conditional probability:

Theorem 5.5. *Let the covariance matrix Σ_n^* be the unique solution of the Lyapunov equation:*

$$\Gamma \Sigma_n^* + \Sigma_n^* \Gamma - \eta \Sigma_n^* = \Sigma_n(\theta_n^*), \tag{5.8}$$

where $\Sigma_n(\theta_n^*) = \mathbb{E}[g(\theta_n^*)g(\theta_n^*)^T]$ and $\eta = \gamma_1 > \frac{1}{2\alpha}$ if $\beta = 1$, 0 if not. Then, under Assumptions $\mathbf{A_1} - \mathbf{A_2}$, we have:

$$1/\gamma_t \left(\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta^*) \right) \Rightarrow \frac{1}{2} U^T (\Sigma_n^*)^{1/2} \Gamma (\Sigma_n^*)^{1/2} U,$$

where $U \sim \mathcal{N}(0, I_q)$. In addition, in the case $\eta = 0$, we have:

$$\|(\Sigma_{\mathbf{n}}^* \Gamma)^{1/2}\|_{HS}^2 = \mathbb{E}[U^T (\Sigma_{\mathbf{n}}^*)^{1/2} \Gamma (\Sigma_{\mathbf{n}}^*)^{1/2} U] = \frac{1}{2} \sigma_{\mathbf{n}}^2(g(\theta_{\mathbf{n}}^*)). \tag{5.9}$$

Theorem 5.5 reveals that the conditional variance term again plays a key role in the asymptotic performance of the algorithm. In particular, it is the dominating term in the precision of the solution. In the next section, we build on these results to derive a generalization bound in the spirit of Bottou & Bousquet (2007) which explicitly depend on the true variance of the gradient estimator. We also point out that the result established in section 4.5 (i.e an unconditional TCL) could also be established in a straightforward fashion, and comparison of norms of the asymptotic variance would also be straightforward thanks to equation (4.13).

5.3 Generalization Bounds

All expectations are now taken w.r.t the sampling procedure and the distribution of the observations. We introduce the following notations. Let $\theta^* = \operatorname{argmin}_{\theta \in \Theta} L(\theta)$ be the minimizer of the true risk. As proposed in Bottou & Bousquet (2007), the mean excess risk can be decomposed as follows: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq \underbrace{2\mathbb{E}\left[\sup_{\theta \in \Theta} |\widehat{L}_n(\theta) - L(\theta)|\right]}_{\mathcal{E}_1} + \underbrace{\mathbb{E}\left[\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta_{\mathbf{n}}^*)\right]}_{\mathcal{E}_2}. \tag{5.10}$$

Beyond the optimization error (the second term on the right hand side of (5.10)), the analysis of the generalization ability of the learning method previously described requires to control the estimation error (the first term). This can be achieved by means of the result stated below, which extends Corollary 3 in Clémençon et al. (2008b) to the K-sample situation.

Proposition 5.6. *Let $\mathcal{H}$ be a collection of bounded symmetric kernels on $\prod_{k=1}^K \mathcal{X}_k^{d_k}$ such that $\mathcal{M}_{\mathcal{H}} = \sup_{(H,x) \in \mathcal{H} \times \mathcal{X}} |H(x)| < +\infty$. Suppose also that $\mathcal{H}$ is a* VC *major class of functions with finite Vapnik-Chervonenkis dimension $V < +\infty$. Let $\kappa = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$. Then, for any $\mathbf{n} \in \mathbb{N}^{*K}$*

$$\mathbb{E}\left[\sup_{H \in \mathcal{H}} |U_n(H) - \mu(H)|\right] \leq \mathcal{M}_{\mathcal{H}} \left\{ 2\sqrt{\frac{2V \log(1+\kappa)}{\kappa}} \right\}. \tag{5.11}$$

for all $\delta \in (0,1)$, we have with probability at least $1-\delta$,

$$\sup_{H\in\mathcal{H}} |U_n(H) - \mu(H)| \leq \mathcal{M}_{\mathcal{H}} \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(1/\delta)}{\kappa}} \right\}. \tag{5.12}$$

We are now ready to derive our main result.

Theorem 5.7. *Let θ_t be the sequence generated by SGD using the incomplete statistic gradient estimator (5.6) with $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$ terms for some $n'_1, \ldots, n'_K$. Assume that $\{L(.;\ \theta) :\ \theta \in \Theta\}$ is a* VC *major class class of finite* VC *dimension V s.t.*

$$\mathcal{M}_\Theta = \sup_{\theta\in\Theta,\ (\mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(K)})\in\prod_{k=1}^{K}\mathcal{X}_k^{d_k}} |H(\mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(K)};\ \theta)| < +\infty, \tag{5.13}$$

and $\mathcal{N}_\Theta = \sup_{\theta\in\Theta}\sigma_\theta^2 < +\infty$. If the step size satisfies the condition of Proposition 5.4, we have:

$$\forall \mathbf{n} \in \mathbb{N}^{*K}, \quad \mathbb{E}[|L(\theta_t) - L(\theta^*)|] \leqslant \frac{C\mathcal{N}_\Theta}{Bt^\beta} + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} \right\}.$$

For any $\delta \in (0,1)$, we also have with probability at least $1-\delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$|L(\theta_t) - L(\theta^*)| \leqslant \left(\frac{C\mathcal{N}_\Theta}{Bt^\beta} + \sqrt{\frac{D_\beta \log(2/\delta)}{t^\beta}} \right) + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(4/\delta)}{\kappa}} \right\}. \tag{5.14}$$

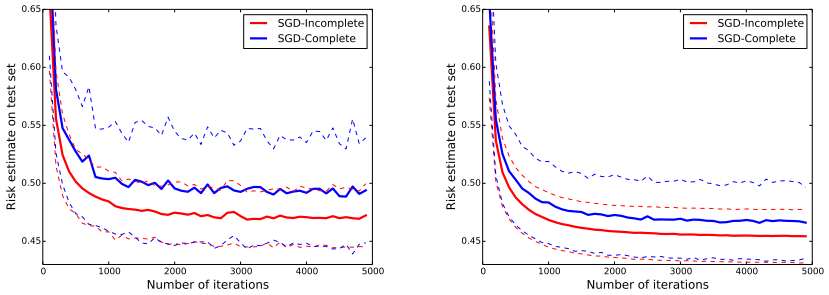
for some constants C and D_β depending on the parameters l, α, γ_1, a_1.

The generalization bound provided by Theorem B.15 shows the advantage of using an incomplete U-statistic (5.6) as the gradient estimator. In particular, we can obtain results of the same form as Theorem B.15 for the complete U-statistic estimator (5.5), but $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$ is then replaced by $\sum_{k=1}^{K} n'_k$ (following Proposition 5.2), leading to greatly damaged bounds. Using an incomplete U-statistic, we thus achieve better performance on the test set while reducing the number of iterations (and therefore the numbers of gradient computations) required to converge to a accurate solution. To the best of our knowledge, this is the first result of this type for empirical minimization of U-statistics. In the next section, we provide experiments showing that these gains are very significant in practice.

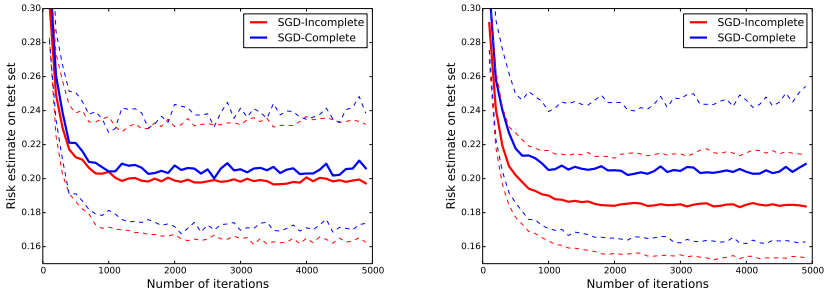
5.4 Numerical Experiments

In this section, we provide numerical experiments to compare the incomplete and complete U-statistic gradient estimators (5.5) and (5.6) when they rely on the same number of terms B. The datasets we use are available online.[1] In all experiments, we randomly split the data into 80% training set and 20% test set and sample 100K pairs from the test set to estimate the test performance. For SGD, we used a step size of the form $\gamma_t = \gamma_1/t$, where γ_1 is the initial value. The results below are with respect to the number of SGD iterations. Computational time comparisons can be found in the supplementary material.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) Covtype, Batch size $= 9$, $\gamma_1 = 1$ (b) Covtype, Batch size $= 400$, $\gamma_1 = 1$



(c) Ijcnn1, Batch size $= 9$, $\gamma_1 = 2$ (d) Ijcnn1, Batch size $= 25$, $\gamma_1 = 2$

FIGURE 5.1: Average over 50 runs of the risk estimate with the number of iterations (solid lines) +/- their standard deviation (dashed lines)

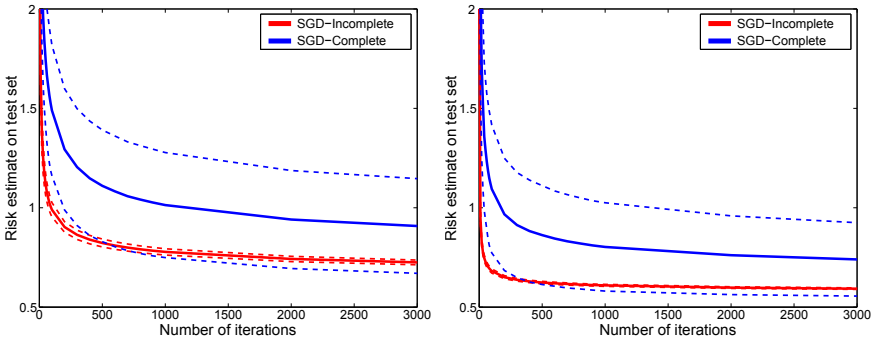
AUC Optimization We address the problem of learning a binary classifier by optimizing the Area Under the Curve, which corresponds to the VUS criterion (Eq. 5.2) when $K = 2$. Given a sequence of i.i.d observations $Z_i = (X_i, Y_i)$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$, we denote by $X^+ = \{X_i; Y_i = 1\}$, $X^- = \{X_i; Y_i = -1\}$ and $n = |X^+||X^-|$. As done in Zhao et al. (2011); Herschtal & Raskutti (2004), we take a linear scoring rule $s_\theta(x) = \theta^T x$ where $\theta \in \mathbb{R}^p$ is the parameter to learn, and use the logistic loss as a smooth convex function upper bounding the Heaviside function, leading to the following ERM problem:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{X_i^+ \in X^+} \sum_{X_j^- \in X^-} \log(1 + \exp(s_\theta(X_i^-) - s_\theta(X_i^+))).$$

We use two datasets: IJCNN1 ($\sim$200K examples, 22 features) and covtype ($\sim$600K examples, 54 features). We try different values for the initial step size γ_1 and the batch size B. Some of the results, averaged over 50 runs of SGD, are displayed in Figure 5.1. As predicted by our theoretical findings, the incomplete U-statistic estimator consistently outperforms its complete variant on average.[2] We also observe a smaller variance between SGD runs when using the incomplete version.

Metric Learning We now turn to a metric learning formulation, where we are given a sample of n i.i.d observations $Z_i = (X_i, Y_i)$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \{1, \dots, c\}$. Following the existing literature Bellet et al. (2013), we focus on (pseudo) distances of the form $D_M(x, x') = (x - x')^T M (x - x')$ where M is a $p \times p$ symmetric positive semi-definite matrix. We again

[2] Of course, the step size must be in an appropriate range. If it is unnecessarily small, both methods have comparable performance, while if it is too large, they both diverge.



(a) SUSY, Batch size $= 120$, $\gamma_1 = 0.5$ (b) SUSY, Batch size $= 455$, $\gamma_1 = 1$

FIGURE 5.2: Average over 50 runs of the error test with the number of iterations (solid lines) +/- their standard deviation (dashed lines)

use the logistic loss to obtain a convex and smooth surrogate for (5.3). The ERM problem is as follows:

$$\min_M \frac{6}{n(n-1)(n-2)} \sum_{i<j<k} \mathbb{I}\{Y_i = Y_j \neq Y_k\} \log(1 + \exp(D_M(X_i, X_j) - D_M(X_i, X_k))).$$

We use the binary classification dataset SUSY (5M examples, 18 features). Figure 5.2 shows that the performance gap between the two strategies is much larger on this problem than in Figure 5.1. This is consistent with the theory: one can see from Proposition 5.2 that the variance gap between the incomplete and the complete approximations is much wider for a one-sample U-statistic of degree 3 (metric learning) than for a two-sample U-statistic of degree 1 (AUC optimization).

5.5 Conclusion and Perspectives

In this chapter, we have studied a specific implementation of the SGD algorithm when the natural empirical estimates of the objective function are of the form of generalized U-statistics. This situation covers a wide variety of statistical learning problems such as multi-partite ranking, pairwise clustering and metric learning. The gradient estimator we propose in this context is based on an incomplete U-statistic obtained by sampling tuples with replacement. First, we have provided asymptotic and non-asymptotic rates for the convergence of this SGD-based learning method to the empirical risk minimizer. Our main result is a thorough analysis of the generalization ability of the predictive rules produced by this algorithm involving both the optimization and the estimation error in the spirit of Bottou & Bousquet (2007). These results show that the SGD variant we propose far surpasses a more naive implementation (of same computational cost) based on subsampling the data points with replacement. Furthermore, we have shown that these performance gains are very significant in practice when dealing with large-scale datasets. Note all the results established can be extended to case where we sample tuples without replacement like we did in chapter 2 (in the general framework of rejective sampling) and would lead to upper bound of the same order.

5.6 Technical Proofs

5.6.1 Proof of Proposition 5.4

We follow the proof of Bach & Moulines (2011b) to derive bounds. We highlight the fact that since the loss function H is l-smooth, $\widehat{g}_n(\theta_t)$ and $\tilde{g}_{n'}(\theta_t)$ are l-Lipschitz. We introduce the sequence $\tilde{\gamma}_t = \gamma_t(1 - l\gamma_t)$. In all generality we will denote by $g_t(\theta)$ an unbiased estimator of the gradient at iteration t, l-Lipschitz in θ. We study the recursion $\theta_{t+1} = \theta_t - \gamma_t g_t(\theta_t)$.
We will make use of the two following classical inequalities from convex analysis (see Nesterov (2004b)):

$$\widehat{L}_n(\theta_1) - \widehat{L}_n(\theta_2) \leqslant \nabla \widehat{L}_n(\theta_1)^T(x - y) - \frac{\alpha}{2}\|\theta_1 - \theta_2\|^2 \tag{5.15}$$

$$\frac{1}{l}\|g_t(\theta_1) - g_t(\theta_2)\|^2 \leqslant (g_t(\theta_1) - g_t(\theta_1))^T(\theta_1 - \theta 2) \tag{5.16}$$

As mentioned previously the analysis we proposed can easily be extended to a more general setting as in Bach & Moulines (2011b). We now begin the proof of the proposition:

$$\|\theta_{t+1} - \theta_n^*\|^2 = \|\theta_t - \theta_n^*\|^2 - 2\gamma_t g_t(\theta_t)^T(\theta_t - \theta_n^*) + \gamma_t^2\|g_t(\theta_t)\|^2$$

Using (5.16) we get

$$\begin{aligned}\|g_t(\theta_t)\|^2 &\leqslant 2(\|g_t(\theta_t) - g_t(\theta_n^*)\|^2 + \|g_t(\theta_n^*)\|^2) \\ &\leqslant 2l(g_t(\theta_t) - g_t(\theta_n^*))^T(\theta_t - \theta_n^*) + 2\|g_t(\theta_n^*)\|^2)\end{aligned} \tag{5.17}$$

which together with $\mathbb{E}[g_t(\theta_t)|\theta_t] = \widehat{g}_n(\theta_t)$ gives

$$\mathbb{E}[\|\theta_{t+1} - \theta_n^*\|^2|\theta_t] \leqslant \|\theta_t - \theta_n^*\|^2 - 2\tilde{\gamma}_t\widehat{g}_n(\theta_t)^T(\theta_t - \theta) + 2\gamma_t^2\|g_t(\theta_n^*)\|^2$$

For the sake of simplicity we assume $(1 - l\gamma_t) > 0\ \forall t$ (which is eventually true since the sequence $(\gamma_t)_{t\geqslant 0}$ goes to 0 as t goes to infinity). Let $a_t = \mathbb{E}[\|\theta_t - \theta_n^*\|^2]$, $\sigma_n^2(\theta_n^*)$ the variance (conditionally upon the data) of $g_t(\theta_n^*)$. Using (5.15) and taking the expectation we get the following recursion:

$$\begin{aligned}a_{t+1} &\leqslant a_t\,(1 - 2\alpha\tilde{\gamma}_t) + 2\gamma_t^2\sigma_n^2(\theta_n^*) \\ &\leqslant a_1\prod_{j=1}^{t}(1 - 2\alpha\tilde{\gamma}_j) + 2\sigma_n^2(\theta_n^*)\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\end{aligned} \tag{5.18}$$

with the convention $\prod_{k=t+1}^{t}(1 - 2\alpha\tilde{\gamma}_k) = 1$. Using $1 + x \leqslant e^x$ we get the following upper bound:

$$\prod_{j=1}^{t}(1 - 2\alpha\tilde{\gamma}_j) \leqslant \exp(-2\alpha\sum_{j=1}^{t}\gamma_j)\exp(2\alpha l\sum_{j=1}^{t}\gamma_j^2)$$

We now need to distinguish two cases:

5.6.1.1 Case $\beta = 1$

If $\beta = 1$ we have:

1. $\log(t+1) - \log(j+1) \leqslant \sum_{k=j+1}^{t} \frac{1}{k}$
2. $\exp(2\alpha l \sum_{k=1}^{t} \frac{1}{k^2}) \leqslant \exp(4\alpha l)$
3. $\exp(2\alpha l \sum_{k=j+1}^{t} \frac{1}{k^2}) \leqslant \exp(\frac{2\alpha l}{j}) \leqslant \exp(2\alpha l)$

Under the assumption $2\alpha\gamma_1 > 1$:

$$\begin{aligned} a_{t+1} &\leqslant \frac{a_1}{(t+1)^{2\alpha\gamma_1}} \exp(4\alpha l\gamma_1^2) + 2\sigma_n^2(\theta_n^*) \exp(2\alpha l\gamma_1^2)\gamma_1^2 \sum_{j=1}^{t} \frac{(j+1)^{2\alpha\gamma_1}}{j^2} \frac{1}{(t+1)^{2\alpha\gamma_1}} \\ &\leqslant \frac{a_1}{(t+1)^{2\alpha\gamma_1}} \exp(4\alpha l\gamma_1^2) + \frac{2^{\alpha\gamma_1} 2\sigma_n^2(\theta_n^*) \exp(2\alpha l\gamma_1^2)\gamma_1^2}{(2\alpha\gamma_1 - 1)(t+1)} \end{aligned}$$

which gives the result.

5.6.1.2 Case $\beta < 1$

If $\beta < 1$, let t_0 be a positive index, by splitting the sum in two parts we get:

$$\begin{aligned} \sum_{j=1}^{t} \gamma_j^2 \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) &= \sum_{j=1}^{t_0} \gamma_j^2 \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) + \sum_{j=t_0+1}^{t} \gamma_j^2 \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) \\ &\leqslant \prod_{k=t_0+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) \sum_{j=1}^{t_0} \gamma_j^2 + \gamma_{t_0} \sum_{j=t_0+1}^{t} \gamma_j \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) \end{aligned}$$

where we used that the sequence $(\gamma_t)_{t\geqslant 1}$ is decreasing. Since $\gamma_j = \frac{1-(1-2\alpha\tilde{\gamma}_j)}{2\alpha} + l\gamma_j^2$ we have:

$$\begin{aligned} \sum_{j=t_0+1}^{t} \gamma_j \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) &= \frac{1}{2\alpha} \sum_{j=t_0+1}^{t} \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) - \prod_{k=j}^{t} (1 - 2\alpha\tilde{\gamma}_k) \\ &\quad + l \sum_{j=t_0+1}^{t} \gamma_j^2 \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) \leqslant \frac{1}{2\alpha} + l \sum_{j=t_0+1}^{t} \gamma_j^2 \end{aligned}$$

which leads to

$$\begin{aligned} \sum_{j=1}^{t} \gamma_j^2 \prod_{k=j+1}^{t} (1 - 2\alpha\tilde{\gamma}_k) &\leqslant \exp(-2\alpha \sum_{j=t_0+1}^{t} \gamma_j) \exp(2\alpha l \sum_{j=t_0+1}^{t} \gamma_j^2) \sum_{j=1}^{t_0} \gamma_j^2 \\ &\quad + \frac{\gamma_{t_0}}{2\alpha} + \gamma_{t_0} l \sum_{j=t_0+1}^{t} \gamma_j^2 \end{aligned}$$

Taking $t_0 \sim \frac{t}{2}$ and using the integral test for convergence:

1. $\sum_{j=t_0+1}^{t} \gamma_j = \gamma_1 \sum_{j=t_0+1}^{t} \frac{1}{j^\beta} \geqslant \gamma_1 \frac{(t+1)^{1-\beta}-(t_0+1)^{1-\beta}}{1-\beta} \geqslant \gamma_1 \frac{(t+1)^{1-\beta}}{2(1-\beta)}$
2. $\sum_{j=t_0+1}^{t} \gamma_j^2 = \gamma_1^2 \sum_{j=t_0+1}^{t} \frac{1}{j^{2\beta}} \leqslant \gamma_1^2 \sum_{j=2}^{+\infty} \frac{1}{j^{2\beta}} \leqslant \frac{\gamma_1^2}{2\beta-1}$
3. $\sum_{j=1}^{t_0} \gamma_j^2 \leqslant \gamma_1^2(1+\frac{1}{2\beta-1}) = \frac{2\beta}{2\beta-1}\gamma_1^2$

gives the following bound:

$$\begin{aligned} a_{t+1} &\leqslant a_1 \exp(-2\alpha\gamma_1 \frac{(t+1)^{1-\beta}}{2(1-\beta)}) \exp(2\alpha l \frac{\gamma_1^2}{2\beta-1}) \\ &+ 2\sigma_n^2(\theta_n^*)(\exp(-2\alpha \frac{(t+1)^{1-\beta}}{2(1-\beta)}) \exp(2\alpha l \frac{\gamma_1^2}{2\beta-1}) \frac{2\beta}{2\beta-1}\gamma_1^2 \\ &+ \frac{2^\beta \gamma_1}{2\alpha t^\beta} + \frac{\gamma_1 2^\beta}{t^\beta} \frac{2l\beta}{2\beta-1}\gamma_1^2) = \sigma_n^2(\theta_n^*) \frac{\gamma_1 2^\beta}{t^\beta}(\frac{1}{2\alpha} + \frac{l\gamma_1^2}{2\beta-1}) + o(\frac{1}{t^\beta}) \end{aligned}$$

which concludes the proof.

5.6.2 Proof of Theorem 5.5

We recall that $\Gamma = \nabla^2 \widehat{L}_n(\theta_n^*)$, $\Sigma_n(\theta_n^*) = \mathbb{E}[g_t(\theta_n^*) g_t(\theta_n^*)^T]$ and Σ_n^* is the solution of Lyapunov's equation:

$$\Gamma\Sigma_n^* + \Sigma_n^*\Gamma - \eta\Sigma_n^* = \Sigma_n(\theta_n^*), \tag{5.19}$$

Using classical results from stochastic approximation theory (see Delyon (2000); G.Fort (2014); Pelletier (1998) for instance) , we first show that under our assumptions:

$$\sqrt{1/\gamma_t}\,(\theta_t - \theta_n^*) \Rightarrow \mathcal{N}(0, \Sigma_n^*),$$

The asymptotic behavior of $1/\gamma_t \left(\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta^*)\right)$ is therefore a consequence of the second order delta method. We now turn to the second part of Theorem 5.5 and similarly to chapter 4:

$$\begin{aligned} \mathbb{E}[U^T(\Sigma_n^*)^{1/2}\Gamma(\Sigma_n^*)^{1/2}U] &= \mathbb{E}[Tr(\Gamma^{1/2}(\Sigma_n^*)^{1/2}UU^T(\Sigma_n^*)^{1/2}\Gamma^{1/2}] \\ &= Tr(\Gamma^{1/2}(\Sigma_n^*)^{1/2}\mathbb{E}[UU^T](\Sigma_n^*)^{1/2}\Gamma^{1/2}) \\ &= Tr(\Gamma^{1/2}(\Sigma_n^*)\Gamma^{1/2}) = Tr(\Gamma\Sigma_n^*) \\ &= \frac{1}{2}Tr(\Sigma_n(\theta_n^*)) = \frac{1}{2}\sigma_n^2(\theta_n^*) \end{aligned}$$

where we used the linearity of the trace, the linearity of the expectation, the dominated convergence theorem (to arrange the different terms) and Lyapunov's equation to conclude.

5.6.3 Proof of Theorem B.15

We prove a more general result and apply it to our specific setting. We consider the recursion defined in the proof of Proposition 2 and keep the same notations.

Theorem 5.8. *Let θ_t be the sequence generated by SGD and define $\sigma^2 = \mathbb{E}[\sigma_n^2(g(\theta_n^*))]$. Assume that $\{L(.;\,\theta) :\, \theta \in \Theta\}$ is a* VC *major class class of finite* VC *dimension V s.t*

$$\mathcal{M}_\Theta = \sup_{\theta\in\Theta,\, (\mathbf{x}^{(1)},\,...,\,\mathbf{x}^{(K)})\in\prod_{k=1}^K \mathcal{X}_k^{d_k}} |H(\mathbf{x}^{(1)},\,\ldots,\,\mathbf{x}^{(K)};\,\theta)| < +\infty, \tag{5.20}$$

If the step size satisfies the condition of Proposition 2, we have:

$$\forall \mathbf{n} \in \mathbb{N}^{*K}, \quad \mathbb{E}[|L(\theta_t) - L(\theta^*)|] \leqslant \frac{C\sigma^2}{t^\beta} + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} \right\}.$$

For any $\delta \in (0,1)$, we also have with probability at least $1-\delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$|L(\theta_t) - L(\theta^*)| \leqslant \left(\frac{C\sigma^2}{t^\beta} + \sqrt{\frac{D_\beta \log(2/\delta)}{t^\beta}} \right) + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(4/\delta)}{\kappa}} \right\}. \tag{5.21}$$

for some constant D_β depending on the parameters l, α, γ_1, a_1 and where $C = C_1$ if $\beta < 1$ and C_2 otherwise.

Proof. For the sake of simplicity, we place ourselves in the special case where Θ is compact, but tedious calculations would lead to similar results under less restrictive assumptions. We therefore introduce the quantities M and M_1 that satisfy $\|g_t(\theta_n^*)\|^2 \leqslant M^2$ and $\|\theta_t - \theta_n^*\| \leqslant M_1^2$. We now turn to the proof of the result:

$$L(\theta_t) - L(\theta^*) \leq 2 \sup_{\theta\in\Theta} |\widehat{L}_n(\theta) - L(\theta)| + \widehat{L}_n(\theta_t) - \widehat{L}_n(\theta_{\mathbf{n}}^*).$$

Taking a union bound we directly get:

$$\mathbb{P}\left(L(\theta_t) - L(\theta^*) \geqslant \frac{l}{2}\frac{\sigma^2}{t^\beta}C + \epsilon \right) \leqslant \underbrace{\mathbb{P}\left(|\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta_n^*)| \geqslant \frac{l}{2}\frac{\sigma^2}{t^\beta}C + \frac{\epsilon}{2}) \right)}_{\mathcal{P}_1}$$
$$+ \underbrace{\mathbb{P}(\sup_{\theta\in\Theta} |\widehat{L}_n(\theta) - L(\theta)| \geqslant \frac{\epsilon}{4})}_{\mathcal{P}_2}$$

The analysis of $\mathcal{P}_2$ is classical and we refer to Clémençon et al. (2008b) to obtain that for all $\delta \in (0,1)$, we have with probability at least $1-\delta$,

$$\sup_{\theta\in\Theta} |\widehat{L}_n(\theta) - L(\theta)| \leq \mathcal{M}_{\mathcal{H}} \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(1/\delta)}{\kappa}} \right\}. \tag{5.22}$$

We now focus on the second term.
Using $\widehat{L}_n(\theta) - \widehat{L}_n(\theta_n^*) \leqslant \frac{l}{2}\|\theta - \theta_n^*\|^2$ (see Nesterov & Nesterov (2004) for instance) we have:

$$\mathbb{P}\left(|\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta_n^*)| - \frac{l}{2}\frac{\sigma^2}{t^\beta}C \geqslant \frac{\epsilon}{2} \right) \leqslant \mathbb{P}\left(\|\theta_t - \theta_n^*\|^2 - \frac{\sigma^2}{t^\beta}C \geqslant \frac{\epsilon}{l} \right)$$

Applying the recursion we get:

$$\begin{aligned}\|\theta_{t+1} - \theta_n^*\|^2 &= \|\theta_t - \theta_n^*\|^2 - 2\gamma_t g_t(\theta_t)^T(\theta_t - \theta_n^*) + \gamma_t^2\|g_t(\theta_t)\|^2 \\ &= \|\theta_t - \theta_n^*\|^2 - 2\gamma_t(g_t(\theta_t) - \nabla\widehat{L}_n(\theta_t))^T(\theta_t - \theta_n^*) - 2\gamma_t\nabla\widehat{L}_n(\theta_t)^T(\theta_t - \theta_n^*) + \gamma_t^2\|g_t(\theta_t)\|^2\end{aligned}$$

and using (5.17):

$$\begin{aligned}\|g_t(\theta_t)\|^2 &\leqslant 2l(g_t(\theta_t) - g_t(\theta_n^*))^T(\theta_t - \theta_n^*) + 2\|g_t(\theta_n^*)\|^2) \\ &= 2l(g_t(\theta_t) - \nabla\widehat{L}_n(\theta_t) - g_t(\theta_n^*))^T(\theta_t - \theta_n^*) + 2l(\nabla\widehat{L}_n(\theta_t))^T(\theta_t - \theta_n^*) + 2\|g_t(\theta_n^*)\|^2\end{aligned}$$

which with $\tilde{a}_t := \|\theta_{t+1} - \theta_n^*\|^2$ gives

$$\begin{aligned}\tilde{a}_{t+1} \leqslant \tilde{a}_t(1 - 2\alpha\tilde{\gamma}_t) + 2\gamma_t^2\sigma_n^2(\theta_n^*) - 2\tilde{\gamma}_t(g_t(\theta_t) - \nabla\widehat{L}_n(\theta_t))^T(\theta_t - \theta_n^*) \\ - 2\gamma_t^2 l(g_t(\theta_n^*))^T(\theta_t - \theta_n^*) + 2\gamma_t^2(\|g_t(\theta_n^*)\|^2 - \sigma_n^2(\theta_n^*))).\end{aligned}$$

An immediate recursion leads to

$$\begin{aligned}\tilde{a}_{t+1} \leqslant \tilde{a}_1 \prod_{j=1}^{t}(1 - 2\alpha\tilde{\gamma}_j) + 2\sigma_n^2(\theta_n^*)\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k) \\ + 2\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\left(\|g_t(\theta_n^*)\|^2 - \sigma_n^2(\theta_n^*)\right) \\ - 2\sum_{j=1}^{t}\tilde{\gamma}_j\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\left(g_t(\theta_t) - \nabla\widehat{L}_{\mathbf{n}}(\theta_t)\right)^T(\theta_t - \theta_n^*) \\ - 2l\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\, g_t(\theta_n^*)^T(\theta_t - \theta_n^*)\end{aligned}$$

The first two terms are analyzed in 5.6.1. We turn now to the tree remaining terms and we introduce the following quantities:

1. $S_{1,t} = 2\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\left(\|g_t(\theta_n^*)\|^2 - \sigma_n^2(\theta_n^*)\right)$
2. $S_{2,t} = 2\sum_{j=1}^{t}\tilde{\gamma}_j\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\left(g_t(\theta_t) - \nabla\widehat{l}_{\mathbf{n}}(\theta_t)\right)^T(\theta_t - \theta_n^*)$
3. $S_{3,t} = 2l\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)\, g_t(\theta_n^*)^T(\theta_t - \theta_n^*)$

Placing ourself under the conditional probability and applying a union bound yields

$$\begin{aligned}\mathbb{P}_{\mathbf{n}}\left(\|\theta_{t+1} - \theta_n^*\|^2 \geqslant \tilde{a}_1\prod_{j=1}^{t}(1 - 2\alpha\tilde{\gamma}_j) + 2\sigma_n^2(\theta_n^*)\sum_{j=1}^{t}\gamma_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k) + \epsilon\right) \\ \leqslant \mathbb{P}_{\mathbf{n}}(S_{1,t} \geqslant \frac{\epsilon}{3}) + \mathbb{P}_{\mathbf{n}}(S_{2,t} \geqslant \frac{\epsilon}{3}) + \mathbb{P}_{\mathbf{n}}(S_{3,t} \geqslant \frac{\epsilon}{3})\end{aligned}$$

Under our assumptions, we have $\|g_t(\theta_n^*)\|^2 \leqslant M^2$, $|g_t(\theta_n^*)^T(\theta_t - \theta_n^*)| \leqslant MM_1$ and $|(g_t(\theta_t) - \nabla\widehat{L}_{\mathbf{n}}(\theta_t))^T(\theta_t - \theta_n^*)| \leqslant (2lM_1 + M)M_1$. Applying Azuma's A.9 inequality yields the following bounds (conditionally upon the observations):

$$\begin{aligned}\mathbb{P}(S_{1,t} \geqslant \epsilon) &\leqslant \exp\left(\frac{-\epsilon^2}{4M^4\sum_{j=1}^{t}\gamma_j^4\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)^2}\right) \\ \mathbb{P}(S_{2,t} \geqslant \epsilon) &\leqslant \exp\left(\frac{-\epsilon^2}{8M_1^2(2lM_1 + M)^2\sum_{j=1}^{t}\tilde{\gamma}_j^2\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)^2}\right) \\ \mathbb{P}(S_{3,t} \geqslant \epsilon) &\leqslant \exp\left(\frac{-\epsilon^2}{8M^2M_1^2\sum_{j=1}^{t}\gamma_j^4\prod_{k=j+1}^{t}(1 - 2\alpha\tilde{\gamma}_k)^2}\right)\end{aligned}$$

We thus need to bound the sums $\sum_{j=1}^{t} \tilde{\gamma}_j^2 \prod_{k=j+1}^{t} (1-2\alpha\tilde{\gamma}_k)^2$ and $\sum_{j=1}^{t} \tilde{\gamma}_j^2 \prod_{k=j+1}^{t} (1-2\alpha\tilde{\gamma}_k)^2$.

5.6.3.1 Case $\beta < 1$

We have for $\beta < 1$:

$$\begin{aligned}
\sum_{t=1}^{T} \gamma_j^4 \prod_{k=j+1}^{T} (1-2\alpha\tilde{\gamma}_k)^2 &\leqslant \prod_{j=t_0+1}^{T} (1-2\alpha\tilde{\gamma}_j)^2 \sum_{j=1}^{t_0} \gamma_j^4 + \gamma_{t_0}^2 \sum_{j=t_0+1}^{T} \gamma_j^2 \prod_{k=j+1}^{t} (1-2\alpha\tilde{\gamma}_k)^2 \\
&\leqslant \exp(-4\alpha \sum_{j=t_0+1}^{T} \gamma_j) \exp(4\alpha l \sum_{j=t_0+1}^{T} \gamma_j^2) \sum_{j=1}^{t_0} \gamma_j^4 \\
&+ \gamma_{t_0}^2 (\frac{1}{2\alpha} + l \sum_{j=t_0+1}^{T} \gamma_j^2)^2 \\
&\leqslant \exp(-4\alpha\gamma_1 \frac{(T+1)^{(1-\beta)}}{2(1-\beta)}) \exp(\frac{4\alpha l \gamma_1^2}{2\beta-1}) \frac{4\beta}{4\beta-1} \gamma_1^4 \\
&+ \frac{\gamma_1^2 2^{2\beta}}{T^{2\beta}} (\frac{1}{2\alpha} + \frac{2l\beta}{2\beta-1})^2 \\
&= \frac{\gamma_1^2 2^{2\beta}}{T^{2\beta}} (\frac{1}{2\alpha} + \frac{2l\beta}{2\beta-1})^2 + o(\frac{1}{T^{2\beta}})
\end{aligned}$$

where we used $\sum_{j=1}^{T} x_j^2 \leqslant (\sum_{j=1}^{T} x_j)^2$ for $x_j = \gamma_j \prod_{k=j+1}^{t} (1-2\alpha\tilde{\gamma}_k) \geqslant 0$ for the last inequality.
We now analyze the second sum:

$$\begin{aligned}
\sum_{j=1}^{T} \tilde{\gamma}_j^2 \prod_{k=j+1}^{T} (1-2\alpha\tilde{\gamma}_k)^2 &\leqslant \prod_{j=t_0+1}^{T} (1-2\alpha\tilde{\gamma}_j)^2 \sum_{j=1}^{t_0} \tilde{\gamma}_j^2 + \tilde{\gamma}_{t_0} \sum_{j=t_0+1}^{T} \tilde{\gamma}_j \prod_{k=j+1}^{T} (1-2\alpha\tilde{\gamma}_k)^2 \\
&\leqslant \exp(-4\alpha \sum_{j=t_0+1}^{T} \gamma_j) \exp(4\alpha l \sum_{j=t_0+1}^{T} \gamma_j^2) \sum_{j=1}^{t_0} \tilde{\gamma}_j^2 \\
&+ \tilde{\gamma}_{t_0} \sum_{j=t_0+1}^{T} \frac{1-(1-2\alpha\tilde{\gamma}_j)}{2\alpha} \prod_{k=j+1}^{T} (1-2\alpha\tilde{\gamma}_k) \\
&\leqslant \exp(-4\alpha \sum_{j=t_0+1}^{T} \gamma_j) \exp(4\alpha l \sum_{j=t_0+1}^{T} \gamma_j^2) \sum_{j=1}^{t_0} \tilde{\gamma}_j^2 + \frac{\tilde{\gamma}_{t_0}}{2\alpha} \\
&\leqslant \exp(-4\alpha\gamma_1 \frac{(T+1)^{(1-\beta)}}{2(1-\beta)}) \exp(\frac{4\alpha l \gamma_1^2}{2\beta-1}) \frac{8\beta}{2\beta-1} \gamma_1^2 + \frac{\gamma_1 2^\beta}{\alpha T^\beta} \\
&= \frac{\gamma_1 2^\beta}{\alpha T^\beta} + o(\frac{1}{T^\beta})
\end{aligned}$$

which concludes this case.

5.6.3.2 Case $\beta = 1$

We have:

$$\begin{aligned}\sum_{t=1}^{T} \gamma_j^4 \prod_{k=j+1}^{T} (1 - 2\alpha\tilde{\gamma}_k)^2 &\leqslant \sum_{t=1}^{T} \gamma_j^4 \exp(-4\alpha \sum_{j=t+1}^{T} \gamma_j) \exp(4\alpha l \sum_{j=t+1}^{T} \gamma_j^2) \\ &\leqslant \gamma_1^4 \exp(8\alpha l \gamma_1^2) \sum_{t=1}^{T} \frac{(j+1)^{4\alpha\gamma_1}}{j^4 (T+1)^{4\alpha\gamma_1}} \\ &\leqslant \frac{\gamma_1^4 \exp(8\alpha l \gamma_1^2) 2^{4\alpha\gamma_1}}{(T+1)^{4\alpha\gamma_1}} \sum_{t=1}^{T} \frac{1}{j^{4-4\alpha\gamma_1}}\end{aligned}$$

We thus need to distinguish several cases:

1. If $2 < 4\alpha\gamma_1 < 3$ then

$$\sum_{t=1}^{T} \frac{1}{j^{4-4\alpha\gamma_1}} \leqslant 1 + \frac{1}{3 - 4\alpha\gamma_1} = \frac{4 - 4\alpha\gamma_1}{3 - 4\alpha\gamma_1}$$

so

$$\sum_{t=1}^{T} \gamma_j^4 \prod_{k=j+1}^{T} (1 - 2\alpha\tilde{\gamma}_k)^2 \leqslant \frac{\gamma_1^4 \exp(8\alpha l \gamma_1^2) 2^{4\alpha\gamma_1}}{(T+1)^{4\alpha\gamma_1}} \frac{4 - 4\alpha\gamma_1}{3 - 4\alpha\gamma_1}$$

2. $4\alpha\gamma_1 = 3$ then

$$\sum_{t=1}^{T} \frac{1}{j^{4-4\alpha\gamma_1}} \leqslant 1 + \log(T)$$

so

$$\sum_{t=1}^{T} \gamma_j^4 \prod_{k=j+1}^{T} (1 - 2\alpha\tilde{\gamma}_k)^2 \leqslant \gamma_1^4 \exp(8\alpha l \gamma_1^2) 2^3 \frac{1 + \log(T)}{(T+1)^3}$$

3. $4\alpha\gamma_1 > 3$ then

$$\sum_{t=1}^{T} \frac{1}{j^{4-4\alpha\gamma_1}} \leqslant \frac{(T+1)^{4\alpha\gamma_1 - 3}}{4\alpha\gamma_1 - 3}$$

so

$$\sum_{t=1}^{T} \gamma_j^4 \prod_{k=j+1}^{T} (1 - 2\alpha\tilde{\gamma}_k)^2 \leqslant \frac{\gamma_1^4 \exp(8\alpha l \gamma_1^2) 2^{4\alpha\gamma_1}}{(4\alpha\gamma_1 - 3)(T+1)^3}$$

Consider $\gamma_1 < \frac{1}{2L}$, then $\tilde{\gamma_j} < \frac{1}{2}\gamma_j$, and

$$\begin{aligned}\sum_{j=1}^{t} \tilde{\gamma}_j^2 \prod_{k=j+1}^{T} (1 - 2\alpha\tilde{\gamma}_k)^2 &\leqslant 4\gamma_1^2 \sum_{j=1}^{t} \frac{1}{j^2} \exp(-4\alpha \sum_{j=t+1}^{T} \gamma_j) \exp(4\alpha l \sum_{j=t+1}^{T} \gamma_j^2) \\ &\leqslant \frac{4\gamma_1^2 \exp(8\alpha l \gamma_1^2) 2^{4\alpha\gamma_1}}{(T+1)^{4\alpha\gamma_1}} \sum_{t=1}^{T} \frac{1}{j^{2-4\alpha\gamma_1}} \\ &\leqslant \frac{4\gamma_1^2 \exp(8\alpha l \gamma_1^2) 2^{4\alpha\gamma_1}}{(4\alpha\gamma_1 - 1)(T+1)}\end{aligned}$$

bringing all the pieces back together and substituting the corresponding bounds give the result under the conditional probability . Taking the expectation over the distribution of the observations gives the result in terms of $\mathbb{E}[\sigma_n^2(\theta_n^*)]$. Since $\mathbb{E}[g(\theta_n^*] = 0$, we have $\mathbb{E}[\sigma_n^2(\theta_n^*)] = \mathbb{E}[\|g(\theta_n^*)\|^2] = \sigma^2(g(\theta_n^*))$ and we get the final result.

□

PART III

Fast Learning Rates for Graph Reconstruction

CHAPTER 6
On Graph Reconstruction via Empirical Risk Minimization

Abstract The problem of predicting connections between a set of data points finds many applications, in systems biology and social network analysis among others. This chapter focuses on the *graph reconstruction* problem, where the prediction rule is obtained by minimizing the average error over all $n(n-1)/2$ possible pairs of the n nodes of a training graph. Our first contribution is to derive learning rates of order $O(\log n/n)$ for this problem, significantly improving upon the rates of order $O(1/\sqrt{n})$ established in the seminal work of Biau & Bleakley (2006). Strikingly, these fast rates are *universal*, in contrast to similar results known for other statistical learning problems (*e.g.*, classification, density level set estimation, ranking, clustering) which require strong assumptions on the distribution of the data. Motivated by applications to large graphs, our second contribution deals with the computational complexity of graph reconstruction. Specifically, we investigate to which extent the learning rates can be preserved when replacing the empirical reconstruction risk by a computationally cheaper Monte-Carlo version, obtained by sampling with replacement $B \ll n^2$ pairs of nodes. Finally, we illustrate our theoretical results by numerical experiments on synthetic and real graphs.

6.1 Introduction

Although statistical learning theory mainly focuses on establishing *universal* rate bounds (*i.e.*, which hold for any distribution of the data) for the accuracy of a decision rule based on training observations, refined concentration inequalities have recently helped understanding conditions on the data distribution under which learning paradigms lead to faster rates. In binary classification, *i.e.*, the problem of learning to predict a random binary label $Y \in \{-1, +1\}$ from on an input random variable X based on independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of the pair (X, Y), rates faster than $1/\sqrt{n}$ are achieved when little mass in the vicinity of $1/2$ is assigned by the distribution of the random variable $\eta(X) = \mathbb{P}\{Y = +1 \mid X\}$. This condition and its generalizations are referred to as the *Mammen-Tsybakov noise conditions* (see Mammen et al., 1999; Tsybakov, 2004; Massart & Nédélec, 2006). It has been shown that a similar phenomenon occurs for various other statistical learning problems. Indeed, specific conditions under which fast rate results hold have been exhibited for density level set estimation (Rigollet & Vert, 2009), (bipartite) ranking (Clémençon et al., 2008a; Clémençon & Robbiano, 2011; Agarwal, 2014), clustering (Antos et al., 2005; Clémençon, 2014) and composite hypothesis testing (Clémençon & Vayatis, 2010).

In this chapter, we consider the supervised learning problem on graphs referred to as *graph reconstruction*, rigorously formulated by Biau & Bleakley (2006). The objective of graph reconstruction is to predict the possible occurrence of connections between a set of objects/individuals known to form the nodes of an undirected graph. Precisely, each node is described by a random vector X which defines a form of *conditional preferential attachment*:

one predicts whether two nodes are connected based on their features X and X'. This statistical learning problem is motivated by a variety of applications such as systems biology (*e.g.*, inferring protein-protein interactions or metabolic networks, see Jansen et al., 2003; Kanehisa, 2001) and social network analysis (*e.g.*, predicting future connections between users, see Liben-Nowell & Kleinberg, 2003). It has recently been the subject of a good deal of attention in the machine learning literature (see Vert & Yamanishi, 2004; Biau & Bleakley, 2006; Shaw et al., 2011), and is also known as *supervised link prediction* (Lichtenwalter et al., 2010; Cukierski et al., 2011). The learning task is formulated as the minimization of a *reconstruction risk*, whose natural empirical version is the average prediction error over the $n(n-1)/2$ pairs of nodes in a training graph of size n. Under standard complexity assumptions on the set of candidate prediction rules, excess risk bounds of the order $O(1/\sqrt{n})$ for the empirical risk minimizers have been established by Biau & Bleakley (2006) based on a representation of the objective functional very similar to the *first Hoeffding decomposition* for second-order U-statistics (see Hoeffding, 1948). However, the computational complexity of finding an empirical risk minimizer, which scales at least as $O(n^2)$ since the empirical graph reconstruction risk involves summing up over $n(n-1)/2$ terms was ignored. This makes the approach impractical when dealing with large graphs commonly found in many applications.

Building up on the above work, our contributions to statistical graph reconstruction are twofold:

Universal fast rates. We prove that a fast rate of order $O(\log n/n)$ is always achieved by empirical reconstruction risk minimizers, in absence of any restrictive condition imposed on the data distribution. This is much faster than the $O(1/\sqrt{n})$ rate established by Biau & Bleakley (2006). Our analysis is based on a different decomposition of the excess of reconstruction risk of any decision rule candidate, involving the *second Hoeffding representation* of a U-statistic approximating it, as well as appropriate maximal/concentration inequalities.

Scaling-up ERM. We investigate the performance of minimizers of computationally cheaper Monte-Carlo estimates of the empirical reconstruction risk, built by averaging over $B \ll n^2$ pairs of vertices drawn with replacement. The rate bounds we obtain highlight that B plays the role of a tuning parameter to achieve an effective trade-off between statistical accuracy and computational cost. Numerical results based on simulated graphs and real-world networks are presented in order to support these theoretical findings.

The chapter is organized as follows. In Section 6.2, we present the probabilistic setting for graph reconstruction and recall state-of-the-art results. Section 6.3 provides our fast rate bound analysis, while Section 6.4 deals with the problem of scaling-up reconstruction risk minimization to large graphs. Numerical experiments are displayed in Section 6.5, and a few concluding remarks are collected in Section 6.6.

6.2 Background and Preliminaries

We start by describing at length the probabilistic framework we consider for statistical inference on graphs, as introduced by Biau & Bleakley (2006). We then briefly recall the related theoretical results documented in the literature.

6.2.1 A Probabilistic Setup for Preferential Attachment

In this chapter, $G = (V,\ E)$ is an undirected random graph with a set $V = \{1,\ \ldots,\ n\}$ of $n \geq 2$ vertices and a set $E = \{e_{i,j} :\ 1 \leq i \neq j \leq n\} \in \{0,1\}^{n(n-1)}$ describing its edges: for all $i \neq j$, we have $e_{i,j} = e_{j,i} = +1$ if the vertices i and j are connected by an edge and $e_{i,j} = e_{j,i} = 0$ otherwise. We assume that G is a *Bernoulli graph*, *i.e.* the random variables $e_{i,j}$, $1 \leq i < j \leq n$, are independent labels drawn from a Bernoulli distribution $Ber(p)$ with parameter $p = \mathbb{P}\{e_{i,j} = +1\}$, the probability that two vertices of G are connected by an edge. The degree of each vertex is thus distributed as a binomial with parameters n and p, which can be classically approximated by a Poisson distribution of parameter $\lambda > 0$ in the limit of large n, when $np \to \lambda$.

Whereas the marginal distribution of the graph G is that of a Bernoulli graph (also sometimes abusively referred to as a *random graph*), a form of *conditional preferential attachment* is also specified in the framework considered here. Precisely, we assume that, for all $i \in V$, a continuous r.v. X_i, taking its values in a separable Banach space $\mathcal{X}$, describes some features related to vertex i. The X_i's are i.i.d. with common distribution $\mu(dx)$ and, for any $i \neq j$, the random pair (X_i, X_j) models some information useful for predicting the occurrence of an edge connecting the vertices i and j. Conditioned upon the features $(X_1,\ \ldots,\ X_n)$, any binary variables $e_{i,j}$ and $e_{k,l}$ are independent only if $\{i,j\} \cap \{k,l\} = \emptyset$. The conditional distribution of $e_{i,j}$, $i \neq j$, is supposed to depend on (X_i, X_j) solely, described by the *posterior preferential attachment probability*:

$$\eta\left(X_i, X_j\right) = \mathbb{P}\left\{e_{i,j} = +1 \mid (X_i, X_j)\right\}. \tag{6.1}$$

For instance, $\forall (x_1, x_2) \in \mathcal{X}^2$, $\eta\left(x_1, x_2\right)$ can be a certain function of a specific distance or similarity measure between x_1 and x_2, as in the synthetic graphs described in Section 6.5.

The conditional average degree of the vertex $i \in V$ given X_i (respectively, given $(X_1,\ \ldots,\ X_n)$) is thus $(n-1)\int_{x \in \mathcal{X}} \eta(X_i, x)\mu(dx)$ (respectively, $\sum_{j \neq i} \eta(X_i, X_j)$). Observe incidentally that, equipped with these notations, $p = \int_{(x,x') \in \mathcal{X}^2} \eta(x, x')\mu(dx)\mu(dx')$. Hence, the 3-tuples $(X_i, X_j, e_{i,j})$, $1 \leq i < j \leq n$, are *non-i.i.d.* copies of a generic random vector $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{e}_{1,2})$ whose distribution $\mathcal{L}$ is given by the tensorial product $\mu(dx_1) \otimes \mu(dx_2) \otimes Ber(\eta(x_1, x_2))$, which is fully described by the pair (μ, η). Observe also that the function η is symmetric by construction: $\forall (x_1, x_2) \in \mathcal{X}^2$, $\eta(x_1, x_2) = \eta(x_2, x_1)$.

In this framework, the learning problem introduced by Biau & Bleakley (2006), referred to as *graph reconstruction*, consists in building a symmetric *reconstruction rule* $g : \mathcal{X}^2 \to \{0,1\}$, from a training graph G, with nearly minimum *reconstruction risk*

$$\mathcal{R}(g) = \mathbb{P}\left\{g(\mathbf{X}_1, \mathbf{X}_2) \neq \mathbf{e}_{1,2}\right\}, \tag{6.2}$$

thus achieving a comparable performance to that of the Bayes rule $g^*(x_1, x_2) = \mathbb{I}\{\eta(x_1, x_2) > 1/2\}$, whose risk is given by $\mathcal{R}^* = \mathbb{E}[\min\{\eta(\mathbf{X}_1, \mathbf{X}_2),\ 1 - \eta(\mathbf{X}_1, \mathbf{X}_2)\}] = \inf_g \mathcal{R}(g)$.

6.2.2 Related Results on Empirical Risk Minimization

Based on the labeled sample $\mathbb{D}_n = \{(X_i, X_j, e_{i,j}) : \ 1 \leq i < j \leq n\}$ related to G, (6.2) is replaced by its empirical version[1]:

$$\widehat{\mathcal{R}}_n(g) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}. \tag{6.3}$$

We denote by $\widehat{g}_n$, an empirical risk minimizer $\widehat{g}_n$, i.e a solution of the optimization problem $\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_n(g)$, where $\mathcal{G}$ is a class of reconstruction rules of controlled complexity, hopefully rich enough to yield a small bias $\inf_{g \in \mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*$. The performance of $\widehat{g}_n$ is measured by its excess risk $\mathcal{R}(\widehat{g}_n) - \inf_{g \in \mathcal{G}} \mathcal{R}(g)$, which can be bounded if we can derive probability inequalities for the maximal deviation

$$\sup_{g \in \mathcal{G}} |\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g)|. \tag{6.4}$$

In the framework of classification, the flagship problem of statistical learning theory, the empirical risk is of the form of an average of i.i.d. r.v.'s, so that results pertaining to empirical process theory can be readily used to obtain bounds for the performance of empirical error minimization. Unfortunately, the empirical risk (6.3) is a sum of *dependent* variables. Following in the footsteps of Clémençon et al. (2008a), the work of Biau & Bleakley (2006) circumvents this difficulty by means of a representation of $\widehat{\mathcal{R}}_n(g)$ as an average of sums of i.i.d. r.v.'s, namely

$$\frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathbb{I}\{g(X_{\sigma(i)}, X_{\sigma(i+\lfloor \frac{n}{2} \rfloor)}) \neq e_{\sigma(i),\sigma(i+\lfloor \frac{n}{2} \rfloor)}\},$$

where the sum is taken over all permutations of $\mathfrak{S}_n$, the symmetric group of order n, and $\lfloor u \rfloor$ denotes the integer part of any $u \in \mathbb{R}$. Very similar to the first Hoeffding decomposition for U-statistics (see Lee, 1990b), this representation reduces the *first order* analysis of the concentration properties of (6.4) to the study of a basic empirical process (see Biau & Bleakley, 2006, Lemma 3.1). Biau & Bleakley (2006) thereby establish rate bounds of the order $O(1/\sqrt{n})$ for the excess of reconstruction risk of $\widehat{g}_n$ under appropriate complexity assumptions (namely, $\mathcal{G}$ is of finite VC-dimension). Note incidentally that (6.3) is a U-statistic only when the variable $\eta(\mathbf{X}_1, \mathbf{X}_2)$ is almost-surely constant (see Janson & Nowicki, 1991, for an asymptotic study of graph reconstruction in this restrictive context).

We finally point out that instead of estimating the reconstruction risk by $\widehat{\mathcal{R}}_n(g)$, one could split the training dataset into two halves and consider the unbiased estimate of $\mathcal{R}(g)$ given by

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{I}\{g(X_i, X_{i+\lfloor n/2 \rfloor}) \neq e_{i,i+\lfloor n/2 \rfloor}\}. \tag{6.5}$$

Since only independent r.v.'s are involved in the sum (6.5), the analysis of its generalization ability is much simpler. In particular, fast rates can be obtained under the following classical noise condition (see Tsybakov, 2004; Boucheron et al., 2005c; Massart & Nédélec, 2006).

[1] A classical Lehmann-Scheffé argument shows that (6.3) is the estimator of (6.2) with smallest variance among all unbiased estimators.

Assumption 8. There exists $\beta > 0$ and $\alpha \in [0, 1]$ such that for all $t > 0$:

$$\mathbb{P}\left(\left|\eta(\mathbf{X}_1, \mathbf{X}_2) - \frac{1}{2}\right| \leqslant t\right) \leqslant \beta t^{\alpha/(1-\alpha)}.$$

One can then show that minimizers of (6.5) achieve a learning rate of order $O((\frac{\log n}{n})^{1/(2-\alpha)})$. We make the following observations:

- Assumption 8 is always satisfied for $\alpha = 0$ and corresponds to the classical learning rate of $O(\sqrt{\log(n)/n})$ obtained by Biau & Bleakley (2006).
- Fast learning rates of the same order as the one we obtained for the minimizer of $\widehat{\mathcal{R}}_n(g)$ are achieved if and only if Assumption 8 is satisfied with $\alpha = 1$. This corresponds to the case where the posterior preferential attachment probability η stays bounded away from $1/2$ with probability one (*cf* Massart & Nédélec, 2006).

In fact, the assumption $\alpha = 1$ is very restrictive. We illustrate this using the following toy example. Let $N_0 \in \mathbb{N}^*$. For each node $1 \leqslant i \leqslant n$, we observe $X_i = (X_i^1, X_i^2)$, where X_i^1 and X_i^2 are two distinct elements uniformly drawn from $\mathcal{P}(\{1, ..., N_0\})$. Consider now the case where two nodes are likely to be connected if they share common preferences, say $e_{i,j} \sim Ber(\#(X_i \cap X_j)/\#(X_i \cup X_j))$. One can easily check that $\mathbb{P}(|\eta(\mathbf{X}_1, \mathbf{X}_2) - \frac{1}{2}| = 0) > 0$, so fast learning rates cannot be obtained for minimizers of (6.5). In contrast, the fast rates of Theorem 6.1 always hold for minimizers of $\widehat{\mathcal{R}}_n(g)$, as shall be seen in next section.

6.3 Empirical Reconstruction is Always Fast!

In this section, we show that the rate bounds established by Biau & Bleakley (2006) can be largely improved *without* any additional assumptions. Precisely, we prove that fast learning rates of order $O(\log n/n)$ are always attained by the minimizers of the empirical reconstruction risk (6.3), as revealed by the following theorem. For simplicity, our results are stated for the case of the classic 0-1 loss $\mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\}$, but we point out that the same arguments directly apply to other loss functions (see Section 5.4 of Boucheron et al., 2005b, for instance). This includes the following two practical examples:

- In order to solve the ERM problem with efficient optimization methods (*e.g.* gradient-based), one typically considers a convex surrogate of the 0-1 loss (*e.g.*, hinge loss, logistic loss), as done in our experiments.
- Real-world networks can be very sparse (*i.e.*, they have very few edges), leading to a highly imbalanced prediction problem. One may then consider a *weighted* loss in order to assign a higher cost to errors on edges than to errors on non-edges.

We now state our main theorem:

Theorem 6.1. (FAST RATES) *Let $\widehat{g}_n$ be any minimizer of the empirical reconstruction risk (6.3) over a class $\mathcal{G}$ of finite* VC*-dimension $V < +\infty$. For all $\delta \in (0, 1)$, we have w.p. at least $1 - \delta$: $\forall n \geq 2$,*

$$\mathcal{R}(\widehat{g}_n) - \mathcal{R}^* \leq 2\left(\inf_{g \in \mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*\right) + C \times \frac{V \log(n/\delta)}{n},$$

where $C < +\infty$ is a universal constant.[2]

Remark 6.2. (ON THE BIAS TERM) Apart from its remarkable universality, Theorem 6.1 takes the same form as in the case of empirical minimization of U-statistics (Clémençon et al., 2008a, Corollary 6), with the same constant 2 in front of the bias term $\inf_{g\in\mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*$. As can be seen from the proof, this constant has no special meaning and can be replaced by any constant strictly larger than 1 at the cost of increasing the constant C. Note that the $O(1/\sqrt{n})$ rate obtained by Biau & Bleakley (2006) has a factor 1 in front of the bias term. Therefore, Theorem 6.1 provides a significant improvement unless the bias overly dominates the second term of the bound (*i.e.*, the complexity of $\mathcal{G}$ is too small).

Remark 6.3. (ON COMPLEXITY ASSUMPTIONS) We point out that a similar result can be established under weaker complexity assumptions involving Rademacher averages . As may be seen by carefully examining the proof of Theorem 6.1, this would require to use the moment inequality for degenerate U-processes stated in (Clémençon et al., 2008a, Theorem 11) instead of that proved by Arcones & Giné (1994).

In the rest of this section, we outline the main ideas used to obtain this result. We rely on some arguments used in the fast rate analysis for empirical minimization of U-statistics (Clémençon et al., 2008a), although these results only hold true under restrictive distributional assumptions. Whereas the quantity (6.3) is not a U-statistic, one may decompose the difference between the excess of reconstruction risk of any candidate rule $g \in \mathcal{G}$ and its empirical counterpart as the sum of its conditional expectation given the X_i's, which is a U-statistic, plus a residual term. In order to explain the main argument underlying the present analysis, additional notation is required. Set

$$\begin{aligned} H_g(x_1, x_2, e_{1,2}) &= \mathbb{I}\{g(x_1, x_2) \neq e_{1,2}\} \\ q_g(x_1, x_2, e_{1,2}) &= H_g(x_1, x_2, e_{1,2}) - H_{g^*}(x_1, x_2, e_{1,2}) \end{aligned}$$

for any $(x_1, x_2, e_{1,2}) \in \mathcal{X} \times \mathcal{X} \times \{0, 1\}$. Denoting by $\Lambda(g) = \mathcal{R}(g) - \mathcal{R}^* = \mathbb{E}[q_g(\mathbf{X}_1, \mathbf{X}_2, \mathbf{e}_{1,2})]$ the excess reconstruction risk with respect to the Bayes rule, its empirical estimate is given by

$$\Lambda_n(g) = \widehat{\mathcal{R}}_n(g) - \widehat{\mathcal{R}}_n(g^*) = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} q_g(X_i, X_j, e_{i,j}).$$

For all $g \in \mathcal{G}$, one may write:

$$\Lambda_n(g) - \Lambda(g) = U_n(g) + \widehat{W}_n(g), \tag{6.6}$$

where

$$U_n(g) = \mathbb{E}\left[\Lambda_n(g) - \Lambda(g) \mid X_1, \ldots, X_n\right] = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \widetilde{q}_g(X_i, X_j) - \Lambda(g)$$

is a U-statistic of degree 2 with symmetric kernel $\widetilde{q}_g(\mathbf{X}_1, \mathbf{X}_2) - \Lambda(g)$, where we denote $\widetilde{q}_g(\mathbf{X}_1, \mathbf{X}_2) = \mathbb{E}[q_g(\mathbf{X}_1, \mathbf{X}_2, \mathbf{e}_{1,2}) \mid \mathbf{X}_1, \mathbf{X}_2]$, and $\widehat{W}_n(g) = \frac{2}{n(n-1)} \sum_{i<j}\{q_g(X_i, X_j, e_{i,j}) - \widetilde{q}_g(X_i, X_j)\}$.

Equipped with this notation, we can now sketch the main steps of the proof of the fast rate bound stated in Theorem 6.1. It is based on Eq. (6.6) combined with two intermediary results, each providing a control of one of the terms involved in it. The second order analysis carried

[2] Note that, throughout the chapter, the constant C is not necessarily the same at each appearance.

out by Clémençon et al. (2008a) shows that the small variance property of U-statistics may yield fast learning rates for empirical risk minimizers when U-statistics are used to estimate the risk, under a certain "low-noise" condition. One of our main findings is that this condition is always fulfilled for the specific U-statistic $U_n(g)$ involved in the decomposition (6.6) of the excess of reconstruction risk of any rule candidate g, as shown by the following lemma.

Lemma 6.4. (VARIANCE CONTROL) *For any distribution $\mathcal{L}$ and any reconstruction rule g, we have*

$$\mathrm{Var}\left(\mathbb{E}\left[q_g(\mathbf{X}_1, \mathbf{X}_2, \mathbf{e}_{1,2}) \mid \mathbf{X}_1\right]\right) \leq \Lambda(g).$$

The fundamental reason for the universal character of this result lies in the fact that the empirical reconstruction risk is not an average over all pairs (*i.e.*, a U-statistic of order 2) but an average over *randomly* selected pairs (random selection being ruled by the function η). The resulting smoothness is the key ingredient allowing us to establish the desired property.

Empirical reconstruction risk minimization over a class $\mathcal{G}$ being equivalent to minimization of $\Lambda_n(g) - \Lambda(g)$, the result below, combined with (6.6), proves that it also boils down to minimizing $U_n(g)$ under appropriate conditions on $\mathcal{G}$, so that the fast rate analysis of Clémençon et al. (2008a) can be extended to graph reconstruction.

Lemma 6.5. (UNIFORM APPROXIMATION) *Under the same assumptions as in Theorem 6.1, for any $\delta \in (0,1)$, we have with probability larger than $1-\delta$: $\forall n \geq 2$,*

$$\sup_{g \in \mathcal{G}} \left| \widehat{W}_n(g) \right| \leq C \times \frac{V \log(n/\delta)}{n},$$

where $C < +\infty$ is a universal constant.

The proof relies on classical symmetrization and randomization tricks combined with the *decoupling method*, in order to cope with the dependence structure of the variables and apply maximal/concentration inequalities for sums of independent random variables (see De la Pena & Giné, 1999).

Based on the above results, Theorem 6.1 can then be derived by relying on the *second Hoeffding decomposition* (see Hoeffding, 1948). This allows us to write $U_n(g)$ as a leading term taking the form of a sum of i.i.d r.v.'s with variance $4Var(\mathbb{E}[q_g(\mathbf{X}_1, \mathbf{X}_2, \mathbf{e}_{1,2}) \mid \mathbf{X}_1])$, plus a degenerate U-statistic (*i.e.*, a U-statistic of symmetric kernel $h(\mathbf{x}_1, \mathbf{x}_2)$ such that $\mathbb{E}[h(\mathbf{x}_1, \mathbf{X}_2)] = 0$ for all $\mathbf{x}_1 \in \mathcal{X}$). The latter can be shown to be of order $O(1/n)$ uniformly over the class $\mathcal{G}$ by means of concentration results for degenerate U-processes.

6.4 Scaling-up Empirical Risk Minimization

The results of the previous section, as well as those of Biau & Bleakley (2006), characterize the excess risk achieved by minimizers of the empirical reconstruction risk $\widehat{\mathcal{R}}_n(g)$ but do not consider the computational complexity of finding such minimizers. For large training graphs, the complexity of merely computing $\widehat{\mathcal{R}}_n(g)$ is prohibitive as the number of terms involved in the summation is $O(n^2)$. In this section, we introduce a sampling-based approach to build approximations of the reconstruction risk with much fewer terms than $O(n^2)$, so as to scale-up risk minimization to large graphs.

The strategy we propose, inspired from the notion of *incomplete U-statistic* (see Lee, 1990b), is of disarming simplicity: instead of the empirical reconstruction risk (6.3), we will consider

an incomplete approximation obtained by sampling *pairs of vertices* (and not vertices) with replacement. Formally, we define the *incomplete graph reconstruction risk* based on $B \geq 1$ pairs of vertices as

$$\widetilde{\mathcal{R}}_B(g) = \frac{1}{B} \sum_{(i,j)\in\mathcal{P}_B} \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}, \tag{6.7}$$

where $\mathcal{P}_B$ is a set of cardinality B built by sampling with replacement in the set $\Theta_n = \{(i,j) : 1 \leq i < j \leq n\}$ of all pairs of vertices of the training graph G. For any $b \in \{1, \ldots, B\}$ and all $(i,j) \in \Theta_n$, denote by $\epsilon_b(i,j)$ the variable indicating whether the pair (i,j) has been picked at the b-th draw ($\epsilon_b(i,j) = +1$) or not ($\epsilon_b(i,j) = +0$). The (multinomial) random vectors $\epsilon_b = (\epsilon_b(i,j))_{(i,j)\in\Theta_n}$ are i.i.d. (notice that $\sum_{(i,j)\in\Theta_n} \epsilon_b(i,j) = +1$ for $1 \leq b \leq B$) and the incomplete risk can be then rewritten as

$$\widetilde{\mathcal{R}}_B(g) = \frac{1}{B} \sum_{b=1}^{B} \sum_{(i,j)\in\Theta_n} \epsilon_b(i,j) \cdot \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}. \tag{6.8}$$

Observe that the statistic (6.7) is an unbiased estimate of the true risk (6.2) and that, given the X_i's, its conditional expectation is equal to (6.3). Considering (6.7) with $B = o(n^2)$ as our empirical risk estimate significantly reduces the computational cost, at the price of a slightly increased variance:

$$\mathrm{Var}\left(\widetilde{\mathcal{R}}_B(g)\right) = \mathrm{Var}\left(\widehat{\mathcal{R}}_n(g)\right) + \frac{1}{B}\left(\mathrm{Var}\left(\widehat{\mathcal{R}}_1(g)\right) - \mathrm{Var}\left(\widehat{\mathcal{R}}_n(g)\right)\right),$$

for any reconstruction rule g. Note in particular that the above variance is smaller than that of the complete reconstruction risk based on a subsample of $\lfloor\sqrt{B}\rfloor$ vertices drawn at random (thus involving $O(B)$ pairs as well). To characterize $\mathrm{Var}(\widetilde{\mathcal{R}}_B(g))$, we need to derive an explicit expression for $Var(\widehat{\mathcal{R}}_n(g))$. This is done by relying on the *second Hoeffding decomposition* (see Hoeffding, 1948) of $\widehat{\mathcal{R}}_n(g)$. For all $1 \leq i < j \leq n$, let us define

- $K_1(X_i) = \mathbb{E}[\mathbb{I}\{g(X_1, X_i) \neq e_{1,i}\}|X_i]$,
- $K_2(X_i, X_j) = \mathcal{R}(g) - K_1(X_i) - K_1(X_j) + \mathbb{E}[\mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\}|X_i, X_j]$,
- $K_3(X_i, X_j, e_{i,j}) = \mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\} - \mathbb{E}[\mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\}|X_i, X_j]$.

We have the following "orthogonal" decomposition:

$$\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g) = \frac{2}{n}\sum_{i=1}^{n} K_1(X_i) + \frac{2}{n(n-1)}\sum_{i<j} K_2(X_i, X_j) + \frac{2}{n(n-1)}\sum_{i<j} K_3(X_i, X_j, e_{i,j}).$$

Introducing the following variance terms:

- $\sigma_1^2 = \mathrm{Var}(K_1(X_1))$,
- $\sigma_2^2 = \mathrm{Var}(K_2(X_1, X_2))$,
- $\sigma_3^2 = \mathrm{Var}(K_3(X_1, X_2, e_{1,2}))$,

one easily gets

$$\mathrm{Var}(\widehat{\mathcal{R}}_n(g)) = \frac{4}{n}\sigma_1^2 + \frac{4}{n(n-1)}\left(\sigma_2^2 + \sigma_3^2\right).$$

Substituting the expression of $\mathrm{Var}(\widehat{\mathcal{R}}_n(g))$ gives

$$\mathrm{Var}\left(\widetilde{\mathcal{R}}_B(g)\right) = O\left(\max\left(\frac{1}{B}, \frac{1}{n}\right)\right).$$

This shows that if $B = O(n)$, $\mathrm{Var}(\widetilde{\mathcal{R}}_B(g))$ is of the same order as $\mathrm{Var}(\widehat{\mathcal{R}}_n(g))$ while $\widetilde{\mathcal{R}}_B(g)$ is computationally much cheaper than $\widehat{\mathcal{R}}_n(g)$ as it consists of only $O(n)$ terms. In contrast to (6.7), the estimator obtained by sampling m nodes has a larger variance: it is equal to $\mathrm{Var}(\widehat{\mathcal{R}}_m(g))$, which is of order $1/m = O(1/\sqrt{B})$. We are thus interested in characterizing the performance of solutions $\widetilde{g}_B$ to the computationally simpler problem $\min_{g\in\mathcal{G}} \widetilde{\mathcal{R}}_B(g)$. The following theorem shows that, when the class $\mathcal{G}$ is of finite VC-dimension, the concentration properties of the *incomplete reconstruction risk process* $\{\widetilde{\mathcal{R}}_B(g)\}_{g\in\mathcal{G}}$ can be deduced from those of the complete version $\{\widehat{\mathcal{R}}_n(g)\}_{g\in\mathcal{G}}$.

Theorem 6.6. (UNIFORM DEVIATIONS) *Suppose that the class $\mathcal{G}$ is of finite* VC*-dimension $V < +\infty$. For all $\delta > 0$, $n \geq 1$ and $B \geq 1$, we have with probability at least $1 - \delta$:,*

$$\sup_{g\in\mathcal{G}} |\widetilde{\mathcal{R}}_B(g) - \widehat{\mathcal{R}}_n(g)| \leq \sqrt{\frac{\log 2 + V \log\left((1 + n(n-1)/2)/\delta\right)}{2B}}.$$

The finite VC-dimension hypothesis can be relaxed and a bound of the same order can be proved to hold true under weaker complexity assumptions involving Rademacher averages (see Remark 6.3). Remarkably, with only $B = O(n)$ pairs, the rate in Theorem 6.6 is of the same order (up to a log factor) as that obtained by Biau & Bleakley (2006) for the maximal deviation $\sup_{g\in\mathcal{G}} |\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$ related to the complete reconstruction risk $\widehat{\mathcal{R}}_n(g)$ with $O(n^2)$ pairs. From Theorem 6.6, one can get a learning rate of order $O(1/\sqrt{n})$ for the minimizer of the incomplete risk involving only $O(n)$ pairs.

Unfortunately, such an analysis does not exploit the relationship between conditional variance and expectation formulated in Lemma 6.4, and is thus not sufficient to show that reconstruction rules minimizing the incomplete risk (6.7) can achieve learning rates comparable to those stated in Theorem 6.1. In contrast, the next theorem provides sharper statistical guarantees.

Theorem 6.7. *Let $\widetilde{g}_B$ be any minimizer of the incomplete reconstruction risk* (6.7) *over a class $\mathcal{G}$ of finite* VC*-dimension $V < +\infty$. Then, for all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$: $\forall n \geq 2$,*

$$\mathcal{R}(\widetilde{g}_B) - \mathcal{R}^* \leq 2\left(\inf_{g\in\mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*\right) + CV\log(n/\delta) \times \left(\frac{1}{n} + \frac{1}{\sqrt{B}}\right),$$

where $C < +\infty$ is a universal constant.

This bound reveals that the number $B \geq 1$ of pairs of vertices plays the role of a tuning parameter, ruling a trade-off between statistical accuracy (taking $B(n) = O(n^2)$ fully preserves the convergence rate) and computational complexity. This will be confirmed numerically in Section 6.5.

The above results can be extended to other sampling techniques, such as Bernoulli sampling and sampling without replacement.

6.4.1 Extensions to Alternative Sampling Schemes

Although the Monte-Carlo scheme previously described is very appealing from a computational perspective, In this section, we show that the results of the previous section can be extended to other sampling techniques, such as Bernoulli sampling and sampling without replacement. Borrowing the terminology of survey theory, the (finite) population under study is the collection Θ_n of all pairs of vertices of the graph G. Its cardinality is $\#\Theta_n = n(n-1)/2$. In this context, a *survey sample* is any subset S of Θ_n with (possibly random) cardinality $m \leq n(n-1)/2$, referred to as its *size*. A survey scheme without replacement is thus any conditional probability distribution $\mathcal{D}$ on the power set of Θ_n, $\mathcal{P}(\Theta_n)$, given the data $\mathbb{D}_n = \{(X_i, X_j, e_{i,j}) : \ (i,j) \in \Theta_n\}$. The probability that the pair $(i,j) \in \Theta_n$ belongs to the sample S drawn from $\mathcal{D}$, conditioned upon $\mathbb{D}_n$, is denoted by $\pi_{(i,j)} = \mathbb{P}_{\mathcal{D}}\{(i,j) \in S\}$ and termed a first order inclusion probability. The quantities $\pi_{(i,j),(k,l)} = \mathbb{P}_{\mathcal{D}}\{((i,j),\ (k,l)) \in S^2\}$, for $(i,j) \neq (k,l)$, are referred to as second order inclusion probabilities. Equipped with these notations, the Horvitz-Thompson version (Horvitz & Thompson, 1951) of the empirical reconstruction risk of a rule $g \in \mathcal{G}$ based on a sample $S \sim \mathcal{D}$ is then given by:

$$\widetilde{\mathcal{R}}^{(\mathcal{D})}(g) = \frac{2}{n(n-1)} \sum_{(i,j)\in\Lambda} \frac{\epsilon_{i,j}}{\pi_{i,j}} \mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\}, \tag{6.9}$$

where $\epsilon_{i,j} = \mathbb{I}\{(i,j) \in S\}$ for all $(i,j) \in \Theta_n$ and $0/0 = 0$ by convention. Provided that the $\pi_{(i,j)}$'s are all strictly positive, (6.9) is an unbiased estimate of (6.3) and, when the size $B \leq n(n-1)/2$ of the survey scheme is deterministic, its conditional variance given the training graph is $Var(\widetilde{\mathcal{R}}^{(\mathcal{D})}(g) \mid \mathbb{D}_n) = 4/(n(n-1))^2 \times \sum_{(i,j)\neq(k,l)} \sigma^2_{(i,j),(k,l)}$, where $\sigma^2_{(i,j),(k,l)}$ is given by

$$\left(\frac{\mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\}}{\pi_{(i,j)}} - \frac{\mathbb{I}\{g(X_k, X_l) \neq e_{k,l}\}}{\pi_{(k,l)}} \right)^2 \times \left(\pi_{(i,j),(k,l)} - \pi_{(i,j)}\pi_{(k,l)} \right)$$

for all $(i,j) \neq (k,l)$. Two specific sampling techniques can naturally be considered.

Bernoulli sampling. Let $B \leq n(n-1)/2$. This sampling plan corresponds to the situation where the $\epsilon_{(i,j)}$'s are i.i.d. Bernoulli r.v.'s with parameter $2B/(n(n-1))$. In this case, the (random) size is a binomial variable of size $n(n-1)/2$ with B as expected value. Incidentally, we mention that Bernoulli sampling is a particular case of Poisson sampling (relaxing the assumption that the $\epsilon_{i,j}$'s are identically distributed), widely used for the purpose of graph sparsification (see *e.g.* Spielman, 2005, Section 6).

Sampling without replacement (SWOR). Fixing in advance $B \leq n(n-1)/2$, one may uniformly draw a sample S among all samples of size B (there are $\binom{n(n-1)/2}{B}$ such samples). In this case, we have $\pi_{(i,j)} = 2B/(n(n-1))$ and $\pi_{(i,j),(k,l)} = 4B(B-1)/(n(n-1)^2(n-2))$ for all $(i,j) \neq (k,l)$ in Θ_n. This is a special case of *rejective sampling*, corresponding to the situations where the $\pi_{(i,j)}$'s are all equal.

The following proposition reveals that, just like (6.7), the Horvitz-Thompson reconstruction risk (6.9), when based on SWOR or Bernoulli schemes, estimates the empirical reconstruction risk of rules in $\mathcal{G}$ uniformly well (provided that $\mathcal{G}$ is of finite VC dimension).

Proposition 6.8. (UNIFORM DEVIATIONS) *Suppose that $\mathcal{G}$ is of finite* VC *dimension $V < +\infty$. For all $\delta \in (0,1)$, we have with probability larger than $1-\delta$: for all $n \geq 1$, $B \leq$*

$n(n-1)/2$,

$$\sup_{g\in\mathcal{G}} |\widetilde{\mathcal{R}}^{(\mathcal{D})}(g) - \widehat{\mathcal{R}}_n(g)| \leq \sqrt{\frac{2\log\left(\frac{2(1+n(n-1)/2)^V}{\delta}\right)}{B}},$$

if $\mathcal{D}$ is a Bernoulli plan of expected size $B \leq n(n-1)/2$,

$$\sup_{g\in\mathcal{G}} |\widetilde{\mathcal{R}}^{(\mathcal{D})}(g) - \widehat{\mathcal{R}}_n(g)| \leq \frac{2\log\left(\frac{2(1+n(n-1)/2)^V}{\delta}\right)}{B} + \sqrt{\frac{2\log\left(\frac{2(1+n(n-1)/2)^V}{\delta}\right)}{B}},$$

when $\mathcal{D}$ is a SWOR plan of size $B \leq n(n-1)/2$.

Performance of minimizer of (6.9) are given by the following proposition.

Proposition 6.9. *Let $\widetilde{g}_B^{\mathcal{D}}$ be any minimizer of (6.9) over a class $\mathcal{G}$ of finite* VC*-dimension $V < +\infty$. When $\mathcal{D}$ is a Bernoulli plan or a SWOR plan of size B, then, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\forall n \geq 2$,*

$$\mathcal{R}(\widetilde{g}_B^{\mathcal{D}}) - \mathcal{R}^* \leq 2\left(\inf_{g\in\mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*\right) + CV\log(n/\delta) \times \left(\frac{1}{n} + \frac{1}{\sqrt{B}}\right),$$

where $C < +\infty$ is a universal constant.

6.5 Numerical Experiments

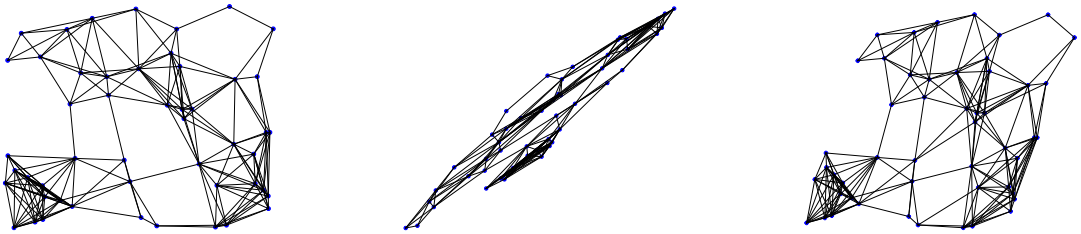
In this section, we present some numerical experiments on large-scale graph reconstruction to illustrate the practical relevance of the idea of incomplete risk introduced in Section 6.4. Following a well-established line of work (Vert & Yamanishi, 2004; Vert et al., 2007; Shaw et al., 2011), we formulate graph reconstruction as a distance metric learning problem (Bellet et al., 2015): we learn a distance function such that we predict an edge between two nodes if the distance between their features is smaller than some threshold. Assuming $\mathcal{X} \subseteq \mathbb{R}^q$, let $\mathbb{S}_+^q$ be the cone of symmetric PSD $q \times q$ real-valued matrices. The reconstruction rules we consider are parametrized by $M \in \mathbb{S}_+^q$ and have the form

$$g_M(x_1, x_2) = \mathbb{I}\left\{D_M(x_1, x_2) \leq 1\right\},$$

where $D_M(x_1, x_2) = (x_1 - x_2)^T M (x_1 - x_2)$ is a (pseudo) distance equivalent to the Euclidean distance after a linear transformation $L \in \mathbb{R}^{q\times q}$, with $M = L^T L$. Note that $g_M(x_1, x_2)$ can be seen as a linear separator operating on the pairwise representation $\mathrm{vec}((x_1 - x_2)(x_1 - x_2')^T) \in \mathbb{R}^{q^2}$, hence the class of learning rules we consider has VC-dimension bounded by $q^2 + 1$. We define the reconstruction risk as:

$$\widehat{\mathcal{S}}_n(g_M) = \frac{2}{n(n-1)} \sum_{i<j} [(2e_{i,j} - 1)(D_M(X_i, X_j) - 1)]_+,$$

where $[\cdot]_+ = \max(0, \cdot)$ is a convex surrogate for the 0-1 loss. In earlier work, ERM has only been applied to graphs with at most a few hundred or thousand nodes due to scalability issues. Thanks to our results, we are able to scale it up to much larger networks by sampling pairs of nodes and solve the resulting simpler optimization problem. In all experiments, we used the generic convex optimization method Adagrad (implemented in Python) and stopped the optimization when the training error had not been improving for the last 10 epochs.



(a) True graph (b) Graph with scrambled features (c) Reconstructed graph

FIGURE 6.1: Illustrative experiment with $n = 50$, $q = 2$, $\tau = 0.27$ and $p = 0$. Figure 6.1(a) shows the training graph, where the position of each node is given by its 2D feature vector. Figure 6.1(b) depicts the same graph after applying a random transformation R to the features. On this graph, the Euclidean distance with optimal threshold achieves a reconstruction error of 0.1311. In contrast, the reconstruction rule learned from $B = 100$ pairs of nodes (out of 1225 possible pairs) successfully inverts R and accurately recovers the original graph (Figure 6.1(c)). Its reconstruction error is 0.008 on the training graph and 0.009 on a held-out graph generated with the same parameters.

TABLE 6.1: Results (averaged over 10 runs) on synthetic graph with $n = 1,000,000$, $q = 100$, $p = 0.05$.

	$\mathbf{B = 0.01n}$	$\mathbf{B = 0.1n}$	$\mathbf{B = n}$	$\mathbf{B = 5n}$	$\mathbf{B = 10n}$
Reconstruction error	0.2272	0.1543	0.1276	0.1185	0.1159
Relative improvement	–	32%	17%	7%	2%
Training time (seconds)	21	398	5,705	20,815	42,574

6.5.1 Synthetic Graph

We create a synthetic graph with n nodes as follows. Each node i has features $X_i^{true} \in \mathbb{R}^q$ sampled uniformly over $[0, 1]$. We then add an edge between nodes that are at Euclidean distance smaller than some threshold τ, and introduce some noise by flipping the value of $e_{i,j}$ for each pair of nodes (i, j) independently with probability p. We then apply a random linear transformation $R \in \mathbb{R}^{q \times q}$ to each node to generate a "scrambled" version $X_i = RX_i^{true}$ of the nodes' features. The learning algorithm is only allowed to observe the scrambled features and must find a rule which accurately recovers the graph by solving the ERM problem above. Note that, denoting $D_{ij} = \|R^{-1}X_i - R^{-1}X_j\|_2$, the posterior preferential attachment probability is given by

$$\eta\left(X_i, X_j\right) = (1 - p) \cdot \mathbb{I}\{D_{ij} \leq \tau\} + p \cdot \mathbb{I}\{D_{ij} > \tau\}.$$

The process is illustrated in Figure 6.1. Using this procedure, we generate a training graph with $n = 1,000,000$ and $q = 100$. We set the threshold τ such that there is an edge between about 20% of the node pairs, and set $p = 0.05$. We also generate a test graph using the same parameters. We then sample uniformly with replacement B pairs of nodes from the training graph to construct our incomplete reconstruction risk. The reconstruction error of the resulting empirical risk minimizer is estimated on 1,000,000 pairs of nodes drawn from the test graph. Table 6.1 shows the test error (averaged over 10 runs) as well as the training time for several values of B. Consistently with our theoretical findings, B implements a trade-off between statistical accuracy and computational cost. For this dataset, sampling $B = 5,000,000$ pairs (out of 10^{12} possible pairs!) is sufficient to find an accurate reconstruction rule. A larger B would result in increased training time for negligible gains in reconstruction error.

TABLE 6.2: Reconstruction error on synthetic graph with parameters $n = 1,000,000$, $q = 100$, $p = 0.05$.

	$\mathbf{B = 0.01n}$	$\mathbf{B = 0.1n}$	$\mathbf{B = n}$	$\mathbf{B = 5n}$	$\mathbf{B = 10n}$
Sampling nodes	0.2552	0.1847	0.1411	0.1279	0.1233
Sampling pairs of nodes	0.2272	0.1543	0.1276	0.1185	0.1159

We also compare the above scheme with an alternative strategy based on sampling *nodes*. Specifically, we randomly sample m nodes and use the reconstruction risk evaluated on the resulting sub-graph as an approximation to the risk on the full graph. To allow a fair comparison, we set m such that $B \simeq m(m-1)/2$ in order to get approximately the same computational complexity for both approaches. Table 6.2 compares the reconstruction error of both strategies for various values of B, averaged over 10 runs. Sampling nodes leads to significantly larger error than sampling pairs, as it only leverages information from a subset of training nodes. We have noticed experimentally that the error gap between these two strategies widens as the complexity of the class of reconstruction rules gets larger (for instance, if we increase the data dimension q).

Figure 6.2 summarizes these results by displaying both the test error and the training time with respect to the number of terms in the risk estimate. For completeness, we also show the performance of the "dataset splitting" strategy (see Eq. 7 of the main text). It exhibits a good statistical/runtime trade-off but leads to suboptimal test error.

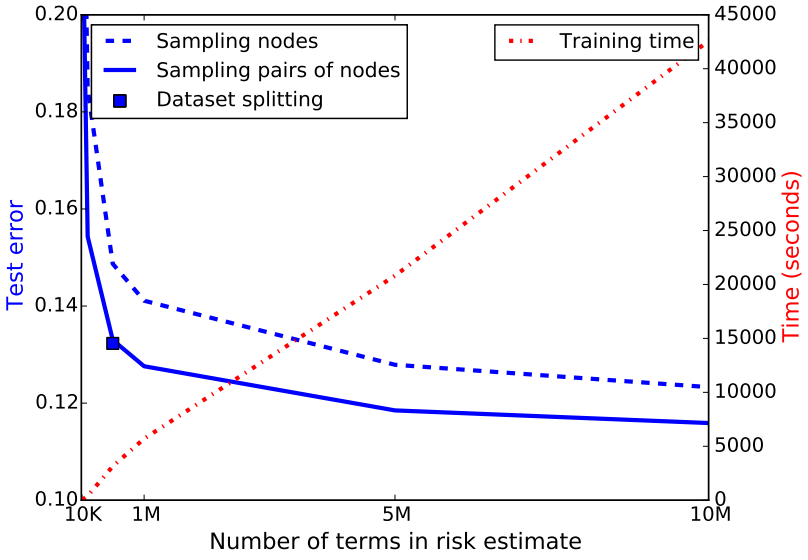


FIGURE 6.2: Summary of results on the synthetic graph.

6.5.2 Real Network

Finally, we also validate our approach on Cit-HepTh, the high-energy physics theory citation network extracted from arXiv.[3] The graph has 27,770 nodes representing research chapters and

[3] http://snap.stanford.edu/data/cit-HepTh.html

TABLE 6.3: Reconstruction error (averaged over 10 runs) on the Cit-HepTh graph.

	$\mathbf{B = 10K}$	$\mathbf{B = 100K}$	$\mathbf{B = 1M}$	$\mathbf{B = 5M}$
Balanced reconstruction error	0.3080	0.2629	0.2484	0.2464
Relative improvement	–	15%	6%	<1%
Training time (seconds)	418	1,675	4,481	18,895

352,807 edges corresponding to a citation between two papers. We generate simple features based on the paper abstracts as follows. We first remove stop words and those with less than 4 characters, then apply a tokenizer and stemmer from the NLTK library[4] and keep only the 300 most frequent words among all abstracts. Finally, we build a 300-dimensional bag-of-words feature vector for each paper by counting the number of occurrences of these words in its abstract and applying an L_1-norm normalization. We randomly split the nodes into a training set (80%) and a test set (20%). Note that the graph is very sparse: there is an edge between about 0.1% of the node pairs. Since classification error is not meaningful in such an imbalanced regime, we optimize a balanced error rate by sampling active edges with higher probability (this is equivalent to optimizing a weighted version of the reconstruction risk, see the remark at the beginning of Section 6.7). Note that the Euclidean distance (with threshold tuned on the training set) achieves a balanced test error of about 0.37.

Table 6.3 shows the balanced test error (averaged over 10 runs) as well as the training time for several values of B. Despite the higher dimensional and sparse nature of the features, we are able to significantly improve over the Euclidean baseline using few training pairs. Furthermore, sampling $B = 1$M pairs is sufficient to get very close to the best performance: going from 1M to 5M pairs brings less than 1% relative improvement in test error at the expense of a 4 times increase in training time.

6.6 Conclusion

In this chapter, we proved that the learning rates for ERM in the graph reconstruction problem are always of order $O(\log n/n)$. We also showed how sampling schemes applied to the population of edges (not nodes) can be used to scale-up such ERM-based predictive methods to very large graphs by means of a detailed rate bound analysis, further supported by empirical results. A first possible extension of this work would naturally consist in considering more general sampling designs, such as Poisson sampling (which generalizes Bernoulli sampling) used in graph sparsification (cf Spielman, 2005), and investigating the properties of minimizers of Horvitz-Thompson versions of the reconstruction risk (see Horvitz & Thompson, 1951). Another challenging line of future research is to extend the results of this chapter to more complex unconditional graph structures in order to account for properties shared by some real-world graphs (*e.g.*, graphs with a power law degree distribution).

[4] http://www.nltk.org

6.7 Technical Proofs

6.7.1 Proof of Lemma 6.4

For any reconstruction rule g, observe first that with probability one:

$$\begin{aligned}\mathbb{E}[q_g(\mathbf{X}_1,\mathbf{X}_2,\mathbf{e}_{1,2}) \mid \mathbf{X}_1] &= \mathbb{E}[\mathbb{E}[q_g(\mathbf{X}_1,\mathbf{X}_2,\mathbf{e}_{1,2}) \mid \mathbf{X}_1,\mathbf{X}_2] \mid \mathbf{X}_1] \\ &= \mathbb{E}_{\mathbf{X}_2}[|1-2\eta(\mathbf{X}_1,\mathbf{X}_2)|\mathbb{I}\{g(\mathbf{X}_1,\mathbf{X}_2) \neq g^*(\mathbf{X}_1,\mathbf{X}_2)\}]]\end{aligned}$$

Observing that we have

$$|1-2\eta(\mathbf{X}_1,\mathbf{X}_2)|^2 \leq |1-2\eta(\mathbf{X}_1,\mathbf{X}_2)|$$

almost surely, and combining with Jensen inequality, we have

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[q_g(\mathbf{X}_1,\mathbf{X}_2,\mathbf{e}_{1,2}) \mid \mathbf{X}_1]) &\leq \mathbb{E}_{\mathbf{X}_1}[(\mathbb{E}_{\mathbf{X}_2}[|1-2\eta(\mathbf{X}_1,\mathbf{X}_2)|\mathbb{I}\{g(\mathbf{X}_1,\mathbf{X}_2) \neq g^*(\mathbf{X}_1,\mathbf{X}_2)\}])^2] \\ &\leq \mathbb{E}[|1-2\eta(\mathbf{X}_1,\mathbf{X}_2)|\mathbb{I}\{g(\mathbf{X}_1,\mathbf{X}_2) \neq g^*(\mathbf{X}_1,\mathbf{X}_2)\}] \\ &= \Lambda(g).\end{aligned}$$

6.7.2 Proof of Lemma 6.5

By definition, for all g, we have: $\forall n \geq 2$,

$$\widehat{W}_n(g) = \frac{2}{n(n-1)} \sum_{i<j} \left\{ q_g(X_i,X_j,e_{i,j}) - \widetilde{q}_g(X_i,X_j) \right\}.$$

The proof relies on the key property: for all $i \neq j$,

$$\mathbb{E}[q_g(X_i,X_j,e_{i,j}) - \widetilde{q}_g(X_i,X_j)|X_i] = \mathbb{E}[q_g(X_i,X_j,e_{i,j}) - \widetilde{q}_g(X_i,X_j)|X_j] = 0$$

almost surely. This basically implies that the process $\{\widehat{W}_n(g)\}_{g\in\mathcal{G}}$ "behaves" as a second order Rademacher Chaos. Mimicking the techniques introduced in De la Pena & Giné (1999), this can be deduced from the following two technical lemmas, which we prove separately in Section 6.7.3 and Section 6.7.4 for clarity.

Lemma 6.10. (DECOUPLING) *Let $(X_i')_{i=1}^n$ be an independent copy of the sequence $(X_i)_{i=1}^n$. Consider r.v.'s valued in $\{0,1\}$, $\{\tilde{e}_{i,j}, i<j\}$, conditionally independent given the X_i's and the X_j''s and such that $\mathbb{P}(\tilde{e}_{i,j}=1|X_i,X_j') = \eta(X_i,X_j')$. Then, for all $q \geq 1$, we have:*

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} q_g(X_i,X_j,e_{i,j}) - \widetilde{q}_g(X_i,X_j)|^q] \leqslant 4^q \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} q_g(X_i,X_j',\tilde{e}_{i,j}) - \widetilde{q}_g(X_i,X_j')|^q].$$

Thanks to the decoupling argument above, one can next introduce the following randomization.

Lemma 6.11. *Let $(\sigma_i)_{i=1}^n$ and $(\sigma_i')_{i=1}^n$ be two independent sequences of i.i.d. Rademacher variables, independent from the $(X_i,X_i',e_{i,j},\tilde{e}_{i,j})$'s. Then, for all $q \geq 1$, we have:*

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} q_g(X_i,X_j',\tilde{e}_{i,j}) - \widetilde{q}_g(X_i,X_j')|^q] \leqslant 4^q \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} \sigma_i\sigma_j'(q_g(X_i,X_j',\tilde{e}_{i,j}) - \widetilde{q}_g(X_i,X_j'))|^q].$$

Consider the conditional Rademacher average

$$\mathbb{E}_{\sigma,\sigma'}[\sup_{g\in\mathcal{G}}|\sum_{i<j}\sigma_i\sigma'_j(q_g(X_i,X'_j,\tilde{e}_{i,j})-\widetilde{q}_g(X_i,X'_j))|^q],$$

where $\mathbb{E}_{\sigma,\sigma'}$ denotes the expectation taken w.r.t. the (σ_i,σ'_i)'s. Following Clémençon et al. (2008a), we can derive an exponential inequality using Markov's inequality and show that, w.p. at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}|\sum_{i<j}\sigma_i\sigma'_j\hat{h}_g(X_i,X'_j,\tilde{e}_{i,j})|\leqslant C\times\frac{V\log(n/\delta)}{n}.$$

6.7.3 Proof of Lemma 6.10

For any random variable ξ, we denote by $\mathcal{L}(\xi)$ its distribution. Let $(X'_i)_{i=1}^n$ be an independent copy of $(X_i)_{i=1}^n$, and $\mathcal{F}$ (respectively $\mathcal{F}'$) be the sigma-field generated by $\{X_1,...,X_n\}$ (respectively $\{X'_1,...,X'_n\}$). Let $\left\{e'_{i,j},1\leqslant i<j\leqslant n\right\}$ be Bernoulli random variables such that $\mathbb{P}(e'_{i,j}=1|\mathcal{F},\mathcal{F}')=\eta(X'_i,X'_j)$ (*i.e*, the conditional distribution of $e'_{i,j}$ depends on (X'_i,X'_j) only). As in De la Pena & Giné (1999), let $(\sigma_i)_{i=1}^n$ be independent Rademacher variables and define:

$$\begin{aligned}Z_i&=X_i\text{ if }\sigma_i=1\text{ and }X'_i\text{ otherwise,}\\ Z'_i&=X'_i\text{ if }\sigma_i=1\text{ and }X_i\text{ otherwise.}\end{aligned}$$

Conditionally upon the X_i and X'_i, the random vector (Z_i,Z'_i) takes the values (X_i,X'_i) or (X'_i,X_i), each with probability 1/2. In particular, we have (see De la Pena & Giné, 1999):

$$\mathcal{L}(X_1,...,X_n,X'_1,...,X'_n)=\mathcal{L}(Z_1,...,Z_n,Z'_1,...,Z'_n). \tag{6.10}$$

Let $\{\tilde{e}_{i,j},1\leqslant i<j\leqslant n\}$ be Bernoulli random variables such that $\mathbb{P}(\tilde{e}_{i,j}=1|\mathcal{F},\mathcal{F}')=\eta(X_i,X'_j)$ and define for $i<j$:

$$\hat{e}_{i,j}=\begin{cases}e_{i,j} & \text{if }\sigma_i=1\text{ and }\sigma_j=-1\\ e'_{i,j} & \text{if }\sigma_i=-1\text{ and }\sigma_j=1\\ \tilde{e}_{i,j} & \text{if }\sigma_i=1\text{ and }\sigma_j=1\\ \tilde{e}_{j,i} & \text{if }\sigma_i=-1\text{ and }\sigma_j=1.\end{cases}$$

We also recall the following notations:

$$\begin{aligned}H_g(x_1,x_2,e_{1,2}) &= \mathbb{I}\{g(x_1,x_2)\neq e_{1,2}\},\\ q_g(x_1,x_2,e_{1,2}) &= H_g(x_1,x_2,e_{1,2})-H_{g^*}(x_1,x_2,e_{1,2})\\ \widetilde{q}_g(X_1,X_2) &= \mathbb{E}[q_g(X_1,X_2,e_{1,2})|X_1,X_2]\end{aligned}$$

Let $\hat{h}_g=q_g-\widetilde{q}_g$ and notice that for all $i<j$:

$$\mathbb{E}_\sigma[\hat{h}_g(Z_i,Z'_j,\hat{e}_{i,j})]=\frac{1}{4}\Big(\hat{h}_g(X_i,X_j,e_{i,j})+\hat{h}_g(X'_i,X'_j,e'_{i,j})+\hat{h}_g(X_i,X'_j,\tilde{e}_{i,j})+\hat{h}_g(X'_i,X_j,\tilde{e}_{j,i})\Big),$$

where $\mathbb{E}_\sigma$ denotes the expectation taken with respect to $\sigma_1, ..., \sigma_n$. Moreover, using

$$\mathbb{E}[\hat{h}_g(X_i', X_j', e_{i,j}')|\mathcal{F}] = \mathbb{E}[\hat{h}_g(X_i', X_j', e_{i,j}')] = 0$$

and

$$\mathbb{E}[\hat{h}_g(X_i, X_j', \tilde{e}_{i,j})|\mathcal{F}] = \mathbb{E}[q_g(X_i, X_j', \tilde{e}_{i,j})|X_i] - \mathbb{E}[\mathbb{E}[q_g(X_i, X_j', \tilde{e}_{i,j})|X_i, X_j']|X_i] = 0,$$

we easily get

$$\hat{h}_g(X_i, X_j, e_{i,j}) = 4\mathbb{E}[\hat{h}_g(Z_i, Z_j', \hat{e}_{i,j})]|\mathcal{F}].$$

For all $q > 1$, we therefore have:

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} \hat{h}_g(X_i, X_j, e_{i,j})|^q] \leqslant 4^q \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} \hat{h}_g(Z_i, Z_j', \hat{e}_{i,j})|^q].$$

We now use (6.10) combined with the fact that by construction, the law of $\hat{e}_{i,j}$ only depends on the realizations Z_i, Z_j', *i.e.*, $\mathbb{P}(\hat{e}_{i,j} = 1|Z_i, Z_j') = \eta(Z_i, Z_j')$, to obtain

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} \hat{h}_g(Z_i, Z_j', \hat{e}_{i,j})|^q] = \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} \hat{h}_g(X_i, X_j', \tilde{e}_{i,j})|^q],$$

which concludes the proof.

6.7.4 Proof of Lemma 6.11

In this section, we find it more convenient to work with sums over $\{1 \leqslant i \neq j \leqslant n\}$ than sums over $\{1 \leqslant i < j \leqslant n\}$, so that for $i < j$ and any random variables $a_{i,j}$, we set $a_{j,i} = a_{i,j}$. Using the symmetry of our problem we have:

$$2^q \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i<j} \hat{h}_g(X_i, X_j', \tilde{e}_{i,j})|^q] = \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i\neq j} \hat{h}_g(X_i, X_j', \tilde{e}_{i,j})|^q],$$

where $\hat{h}_g$ is defined as in Section 6.7.3. Re-using the notations used in Section 6.7.3, we further introduce $(X_i'')_{i=1}^n$, a copy of $(X_i')_{i=1}^n$, independent from $\mathcal{F}$, $\mathcal{F}'$, and denote by $\mathcal{F}''$ its sigma-field. Let $\left\{\tilde{e}_{i,j}'', 1 \leqslant i < j \leqslant n\right\}$ Bernoulli random variables such that $\mathbb{P}(\tilde{e}_{i,j}'' = 1|\mathcal{F}, \mathcal{F}', \mathcal{F}'') = \eta(X_i, X_j'')$. We now use classical randomization techniques and introduce our "ghost" sample:

$$\begin{aligned}
\mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i\neq j} \hat{h}_g(X_i, X_j', \tilde{e}_{i,j})|^q] &= \mathbb{E}[\sup_{g\in\mathcal{G}} |\sum_{i\neq j} \hat{h}_g(X_i, X_j', \tilde{e}_{i,j}) - \mathbb{E}_{\mathcal{F}''}[\hat{h}_g(X_i, X_j'', \tilde{e}_{i,j}'')]|^q] \\
&\leqslant \mathbb{E}\left\{\sup_{g\in\mathcal{G}} \left|\sum_{j=1}^n \sum_{i\neq j} \hat{h}_g(X_i, X_j', \tilde{e}_{i,j}) - \hat{h}_g(X_i, X_j'', \tilde{e}_{i,j}'')\right|^q\right\}.
\end{aligned}$$

where $\mathbb{E}_{\mathcal{F}''}$ denotes expectation with respect to the $(X_i'')_{i=1}^n$. Let $(\sigma_i)_{i=1}^n$ be independent Rademacher variables, independent of $\mathcal{F}$, $\mathcal{F}'$ and $\mathcal{F}''$, then we have:

$$\begin{aligned}
&\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})-\hat{h}_g(X_i,X_j'',\tilde{e}_{i,j}'')|^q\mathcal{F}] \\
&\qquad=\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sigma_j\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})-\hat{h}_g(X_i,X_j'',\tilde{e}_{i,j}'')|^q|\mathcal{F}] \\
&\qquad\leqslant 2^q\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sigma_j\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q|\mathcal{F}],
\end{aligned}$$

and get:

$$\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q]\leqslant 2^q\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sigma_j\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q].$$

We now repeat the same argument but for the $(X_i)_{i=1}^n$. Let $(X_i''')_{i=1}^n$ be a copy of $(X_i)_{i=1}^n$, independent of $\mathcal{F}$, $\mathcal{F}'$, and denote by $\mathcal{F}'''$ its sigma-field. Let $\left\{\tilde{e}_{i,j}''', 1\leqslant i<j\leqslant n\right\}$ be Bernoulli random variables such that $\mathbb{P}(\tilde{e}_{i,j}'''=1|\mathcal{F},\mathcal{F}',\mathcal{F}''')=\eta(X_i''',X_j')$. Then:

$$\begin{aligned}
\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sigma_j\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q] &= \mathbb{E}\Big[\sup_{g\in\mathcal{G}}|\sum_{i=1}^n\sum_{j\neq i}\sigma_j\hat{h}_g(X_i,X_j',\tilde{e}_{i,j}) \\
&\qquad -\mathbb{E}_{\mathcal{F}'''}[\sigma_j\hat{h}_g(X_i''',X_j',\tilde{e}_{i,j}'')]|^q\Big] \\
&\leqslant \mathbb{E}\Big[\sup_{g\in\mathcal{G}}|\sum_{i=1}^n\sum_{j\neq i}\sigma_j\hat{h}_g(X_i,X_j',\tilde{e}_{i,j}) \\
&\qquad -\sigma_j\hat{h}_g(X_i''',X_j',\tilde{e}_{i,j}'')|^q\Big].
\end{aligned}$$

Let $(\sigma_i')_{i=1}^n$ be a copy of $(\sigma_i)_{i=1}^n$, independent of $(\sigma_i)_{i=1}^n$, $\mathcal{F}$, $\mathcal{F}'$, $\mathcal{F}'''$, we have

$$\begin{aligned}
&\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sum_{i\neq j}\sigma_j\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})-\sigma_j\hat{h}_g(X_i''',X_j',\tilde{e}_{i,j}'')|^q|\mathcal{F}',(\sigma_i)_{i=1}^n] \\
&\qquad=\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{i=1}^n\sigma_i'\sum_{j\neq i}\sigma_j\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})-\sigma_j\hat{h}_g(X_i,X_j'',\tilde{e}_{i,j}'')|^q|\mathcal{F}',(\sigma_i)_{i=1}^n] \\
&\qquad\leqslant 2^q\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{i=1}^n\sigma_i'\sum_{j\neq i}\sigma_j\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q|\mathcal{F}',(\sigma_i)_{i=1}^n].
\end{aligned}$$

Finally, we get

$$\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{j=1}^n\sum_{i\neq j}\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q]\leqslant 4^q\mathbb{E}[\sup_{g\in\mathcal{G}}|\sum_{i\neq j}\sigma_i\sigma_j'\hat{h}_g(X_i,X_j',\tilde{e}_{i,j})|^q].$$

6.7.5 Proof of Theorem 6.1

We prove a more general version of Theorem 1.

Theorem 6.12. *For any $\theta > 0$, with probability at least $1 - \delta$, the empirical risk minimizer $\widehat{g}_n$ satisfies:*

$$\mathcal{R}(\widehat{g}_n) - \mathcal{R}(g') \leqslant \theta(\mathcal{R}(g') - \mathcal{R}(g^*)) + (1 + \theta + \frac{1}{\theta})DV\frac{\log(n/\delta)}{n}$$

for some universal constant D.

The version of the main text is obtained by taking $\theta = 1$ and adding $\mathcal{R}(g') - \mathcal{R}(g^*)$ on both sides of the inequality.

Proof. Following the analysis of Clémençon et al. (2008a), for all $g \in \mathcal{G}$ we rewrite:

$$\Lambda_n(g) - \Lambda(g) = 2(T_n(g) - \Lambda(g)) + W_n(g) + \widehat{W}_n(g).$$

where

$$T_n(g) = \frac{1}{n}\sum_{i=1}^{n} h_g(X_i)$$

is a sum of i.i.d random variables with $h_g(X_i) = \mathbb{E}[q_g(X_i, X_j, e_{i,j})|X_i]$,

$$W_n(g) = \frac{1}{n(n-1)}\sum_{i \neq j} \widetilde{h}_g(X_i, X_j)$$

is a degenerate U-statistic with symmetric kernel

$$\widetilde{h}_g(X_i, X_j) = \mathbb{E}[q_g(X_i, X_j, e_{i,j})|X_i, X_j] + \Lambda(g) - h_g(X_i) - h_g(X_j)$$

and

$$\widehat{W}_n(g) = \frac{1}{n(n-1)}\sum_{i \neq j} \hat{h}_g(X_i, X_j, e_{i,j})$$

with

$$\hat{h}_g(X_i, X_j, e_{i,j}) = q_g(X_i, X_j, e_{i,j}) - \mathbb{E}[q_g(X_i, X_j, e_{i,j})|X_i, X_j].$$

We also recall that we proved in Lemma 2 that

$$\mathrm{Var}\left(\mathbb{E}\left[q_g(\mathbf{X}_1, \mathbf{X}_2, \mathbf{e}_{1,2}) \mid \mathbf{X}_1\right]\right) \leq \Lambda(g).$$

As mentioned before, the fact that we can upper-bound the variance of $h_g(X)$ by its expectation is the key property that will allow us to derive fast rates. We now follow the analysis of Boucheron et al. (2005b) and introduce the following quantities:

- $\mathcal{H} = \{h_g : g \in \mathcal{G}\}$.
- $\mathcal{H}^* = \{\alpha h_g : g \in \mathcal{G}, \alpha \in [0, 1]\}$.
- $\psi(r) = \mathbb{E}R_n\left\{f \in \mathcal{H}^* : \sqrt{\mathbb{E}[f(X)^2]} \leqslant r\right\}$, where R_n denotes the Rademacher chaos taken over the observation X_i.
- For $r > 0$, we note $\mathcal{G}_r = \left\{\frac{rh_g}{\max(r, \sqrt{\Lambda(g)})} : g \in \mathcal{G}\right\}$.

For all $h \in \mathcal{G}_r$, $\mathbb{E}[h] - h \leqslant 2$ and $\mathrm{Var}(h) \leqslant r$ so that applying Bousquet's Inequality for the Supremum of Empirical Processes Bousquet (2002); Boucheron et al. (2005b) to the class $\mathcal{G}_r$ gives that with probability at least $1 - \delta/6$, for any $g \in \mathcal{G}$:

$$\Lambda(g) - T_n(g) \leqslant \frac{\max(r, \sqrt{\Lambda(g)})}{r}\Big(2\mathbb{E}[\sup_{h_g \in \mathcal{G}_r} \Lambda(g) - T_n(g)] + r\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{8\log(1/\delta)}{3n}\Big).$$

Since $\mathbb{E}[\sup_{h_g \in \mathcal{G}_r}(\Lambda(g) - T_n(g))] \leqslant 2\mathbb{E}R_n[G_r] \leqslant 2\psi(r)$ we get:

$$\Lambda(g) - T_n(g) \leqslant \frac{\max(r, \sqrt{\Lambda(g)})}{r}\Big(4\psi(r) + r\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{8\log(6/\delta)}{3n}\Big).$$

We now apply Bernstein's Inequality to h'_g and using the fact that $\mathrm{Var}(h'_g) \leqslant \Lambda(g') \leqslant \Lambda(g)$ for any $g \in \mathcal{G}$, we get that with probability at least $1 - \delta/6$:

$$T_n(g') - \Lambda(g') \leqslant \max(r, \sqrt{\Lambda(g)})\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{8\log(6/\delta)}{3n}.$$

Summing the two inequality and taking a union bound we get that with probability at least $1 - \delta/3$, for all $g \in \mathcal{G}$:

$$\Lambda(g) - T_n(g) + T_n(g') - \Lambda(g') \leqslant \frac{\max(r, \sqrt{\Lambda(g)})}{r}\Big(4\psi(r) + 2r\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{16\log(6/\delta)}{3n}\Big).$$

We now rewrite $T_n(g)$ as:

$$T_n(g) = \frac{1}{2}(\Lambda(g) + \Lambda_n(g) - W_n(g) - \widehat{W}_n(g))$$

which we substitute in the previous inequality and obtain:

$$\begin{aligned}\frac{3}{2}(\Lambda(g) - \Lambda(g')) \leqslant \frac{1}{2}(\Lambda_n(g) - \Lambda_n(g')) + W_n(g') - W_n(g) + \widehat{W}_n(g') - \widehat{W}_n(g) \\ + \frac{\max(r, \sqrt{\Lambda(g)})}{r}\Big(4\psi(r) + 2r\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{16\log(6/\delta)}{3n}\Big).\end{aligned}$$

We take $g = \widehat{g}_n$ so that $\Lambda_n(\widehat{g}_n) - \Lambda_n(g') \leqslant 0$ and use Lemma 3 together with a result from Clémençon et al. (2008a) to obtain that with probability at least $1 - 2\delta/3$:

$$W_n(g') - W_n(g) + \widehat{W}_n(g') - \widehat{W}_n(g) \leqslant 2\sup_{g \in \mathcal{G}}|W_n(g)| + 2\sup_{g \in \mathcal{G}}|\widehat{W}_n(g)| \leqslant \frac{4CV\log(3n/\delta)}{n}.$$

We finally get that with probability at least $1 - \delta$:

$$\Lambda(\widehat{g}_n) - \Lambda(g') \leqslant \frac{4CV\log(3n/\delta)}{n} + \frac{\max(r, \sqrt{\Lambda(\widehat{g}_n)})}{r} \times \Big(4\psi(r) + 2r\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{16\log(6/\delta)}{3n}\Big).$$

Now, we either have $\Lambda(\widehat{g}_n) \leqslant r^2$, in which case we have in particular $\Lambda(\widehat{g}_n) - \Lambda(g') \leqslant r^2$, or $\Lambda(\widehat{g}_n) \geqslant r^2$. Under the latter hypothesis:

$$\Lambda(\widehat{g}_n) - \Lambda(g') \leqslant \frac{4CV\log(3n/\delta)}{n} + \frac{\sqrt{\Lambda(\widehat{g}_n)}}{r}\Big(4\psi(r) + 2r\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{16\log(6/\delta)}{3n}\Big).$$

For $\delta \in [0,1]$, we finally introduce $r^*(\delta)$ as solution of

$$r = 4\psi(\sqrt{r}) + 2\sqrt{r}\sqrt{\frac{2\log(6/\delta)}{n}} + \frac{16\log(6/\delta)}{3n}.$$

Substituting $r^*(\delta)^2$ for r and using its definition in the previous bound gives:

$$\Lambda(\widehat{g}_n) - \Lambda(g') \leqslant \frac{4CV\log(3n/\delta)}{n} + \sqrt{\Lambda(\widehat{g}_n)r^*(\delta)}.$$

Now, using for all $\theta > 0$:

$$\sqrt{\Lambda(\widehat{g}_n)r^*(\delta)} \leqslant \frac{1}{2}\left(\frac{2\theta}{1+\theta}\Lambda(\widehat{g}_n) + \frac{1+\theta}{2\theta}r^*(\delta)\right)$$

gives that with probability at least $1-\delta$:

$$\mathcal{R}(\widehat{g}_n) - \mathcal{R}(g') \leqslant \theta(\mathcal{R}(g') - \mathcal{R}(g^*)) + \frac{(\theta+1)^2}{4\theta}r^*(\delta) + (1+\theta)4CV\log(3n/\delta).$$

Putting all the pieces back together, we have shown

$$\mathcal{R}(\widehat{g}_n) - \mathcal{R}(g') \leqslant \max(r^*(\delta)^2, \theta(\mathcal{R}(g') - \mathcal{R}(g^*)) + \frac{(\theta+1)^2}{4\theta}r^*(\delta) + (1+\theta)4CV\log(3n/\delta)).$$

Convenient upper bound for $r^*(\delta)$ can be found in Boucheron et al. (2005b):

$$r^*(\delta) \leqslant CV\frac{log(n/\delta)}{n},$$

for some universal constant C. This concludes the proof. □

6.7.6 Proof of Theorem 6.6

One may write for all $g \in \mathcal{G}$, $n \geq 2$ and $B \geq 1$,

$$\widetilde{\mathcal{R}}_B(g) - \widehat{\mathcal{R}}_n(g) = \frac{1}{B}\sum_{b=1}^{B}\mathcal{Z}_b(g),$$

where

$$\mathcal{Z}_b(g) = \sum_{i<j}\left(\epsilon_b(i,j) - \frac{2}{n(n-1)}\right)\mathbb{I}\{g(X_i,X_j) \neq e_{i,j}\}$$

for all $(g,b) \in \mathcal{G} \times \{1, \ldots, B\}$. Conditioned upon the $(X_i, X_j, e_{i,j})$'s, for all $g \in \mathcal{G}$, the $\mathcal{Z}_b(g)$'s are i.i.d. centered random variables, bounded by 1. In addition, the collection $\mathcal{G}$ being of finite VC-dimension V, Sauer's lemma yields:

$$\#\{\{\mathbb{I}\{g(X_i,X_j) \neq e_{i,j}\} : g \in \mathcal{G}\} \leq (1+n(n-1)/2)^V.$$

Applying Hoeffding's inequality to the $\mathcal{Z}_b(g)$'s conditioned upon the $(X_i, X_j, e_{i,j})$'s and the union bound leads to: $\forall \zeta > 0$,

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{B}\sum_{b=1}^{B}\mathcal{Z}_b(g)\right| > \zeta \mid \{(X_i,X_j,e_{i,j})\}_{(i,j)\in\Lambda}\right\} \leq 2(1+n(n-1)/2)^V\exp\left(-2B\zeta^2\right).$$

Taking the expectation w.r.t. the $(X_i, X_j, e_{i,j})$'s yields the desired bound.

6.7.7 Proof of Theorem 6.7

As done for Theorem 1, we prove the following generalization of Theorem 5.

Theorem 6.5. *For any $\theta > 0$, with probability at least $1-\delta$, the minimizer $\widetilde{g}_B$ of the incomplete risk satisfies:*

$$\mathcal{R}(\widetilde{g}_B) - \mathcal{R}(g') \leqslant \theta(\mathcal{R}(g') - \mathcal{R}(g^*)) + (1 + \theta + \frac{1}{\theta})D_1 V\Big(\frac{\log(n/\delta)}{n} + \sqrt{\frac{\log(n/\delta)}{B}}\Big).$$

for some universal constant D_1.

Proof. We proceed in a similar fashion than for the proof of Theorem 6.12 and first start by recalling that we have with probability at least $1 - \delta/4$,

$$\sup_{g\in\mathcal{G}} |W_n(g)| \leqslant \frac{CV\log(4n/\delta)}{n}$$

and

$$\sup_{g\in\mathcal{G}} |\widehat{W}_n(g)| \leqslant \frac{CV\log(4n/\delta)}{n}.$$

We also recall that Theorem 4 gives that with probability at least $1 - \delta/4$:

$$\sup_{g\in\mathcal{G}} |\widetilde{\Lambda}_B(g) - \widehat{\Lambda}_n(g)| \leq \sqrt{\frac{\log 2 + V\log\left(\frac{4(1+n(n-1)/2)}{\delta}\right)}{2B}} := \sqrt{\frac{C_1 V\log(4n/\delta)}{B}}.$$

We follow the proof of Theorem 6.12. For all $r > 0$ with probability at least $1 - \delta/4$:

$$\Lambda(g) - T_n(g) + T_n(g') - \Lambda(g') \leqslant \frac{\max(r, \sqrt{\Lambda(g)})}{r}\Big(4\psi(r) + \frac{16\log(8/\delta)}{3n} + 2r\sqrt{\frac{2\log(8/\delta)}{n}}\Big).$$

For any $g \in \mathcal{G}$, let $\widetilde{\Lambda}_B(g) = \widetilde{\mathcal{R}}_B(g) - \widetilde{\mathcal{R}}_B(\widetilde{g}_B)$ be the incomplete excess risk of g. We rewrite:

$$\begin{aligned} T_n(g') - T_n(g) = \frac{1}{2}(\Lambda(g') - \Lambda(g) + \Lambda_n(g') - \widetilde{\Lambda}_B(g') + \widetilde{\Lambda}_B(g') - \widetilde{\Lambda}_B(g) + \widetilde{\Lambda}_B(g) - \Lambda_n(g) \\ + W_n(g') - W_n(g) + \widehat{W}_n(g') - \widehat{W}_n(g)), \quad (6.11) \end{aligned}$$

which we substitute in the previous bound, take $g = \widetilde{g}_B$ so that $\widetilde{\Lambda}_B(\widetilde{g}_B) - \widetilde{\Lambda}_B(g') \leqslant 0$ and get that with probability at least $1 - \delta$:

$$\begin{aligned} \Lambda(\widetilde{g}_B) - \Lambda(g') \leqslant \frac{\max(r, \sqrt{\Lambda(g)})}{r}\Big(4\psi(r) + 2r\sqrt{\frac{2\log(8/\delta)}{n}} + \frac{16\log(8/\delta)}{3n}\Big) \\ + 4CV\log(4n/\delta) + \sqrt{\frac{C_1 V\log(4n/\delta)}{B}}. \end{aligned}$$

Let $r_1^*(\delta)$ be solution of

$$r = 4\psi(r) + 2r\sqrt{\frac{2\log(8/\delta)}{n}} + \frac{16\log(8/\delta)}{3n}.$$

Then we either have $\Lambda(\widetilde{g}_B) \leqslant r_1^*(\delta)^2$ or:

$$\mathcal{R}(\widetilde{g}_B) - \mathcal{R}(g') = \Lambda(\widetilde{g}_B) - \Lambda(g') \leqslant \sqrt{\Lambda(\widetilde{g}_B) r_1^*(\delta)} + 4CV\log(4n/\delta) + \sqrt{\frac{C_1 V \log(4n/\delta)}{B}}.$$

In the latter case we easily get that for all $\theta > 0$,

$$\mathcal{R}(\widetilde{g}_B) - \mathcal{R}(g') \leqslant \theta(\mathcal{R}(g') - \mathcal{R}(g^*)) + \frac{(\theta+1)^2}{4\theta} r_1^*(\delta) + (1+\theta)\Big(4CV\log(4n/\delta) + \sqrt{\frac{C_1 V \log(4n/\delta)}{B}}\Big).$$

Upper-bounding $r_1^*(\delta)$ as in the proof of Theorem 6.12 gives the result. □

6.7.8 Proof of Proposition 6.8

Proof. We first establish the following preliminary result.

Lemma 6.6. *Suppose that the hypotheses of Proposition 6.8 are fulfilled. Then, we have:* $\forall g \in \mathcal{G}$,

$$\mathbb{E}\left[\left(\widetilde{\mathcal{R}}^{(\mathcal{D})}(g) - \widehat{\mathcal{R}}_n(g)\right)^2 \mid \mathbb{D}_n\right] \leq \frac{2}{B}. \tag{6.12}$$

Proof. It suffices to notice that, for both sampling plans, we have for all $(i,j) \neq (k,l)$ in Λ,

$$\mathbb{E}\left[\left(\epsilon_{i,j} - \frac{2B}{n(n-1)}\right)^2\right] \leq \frac{2B}{n(n-1)}, \tag{6.13}$$

as well as

$$\mathbb{E}\left[\left(\epsilon_{i,j} - \frac{2B}{n(n-1)}\right)\left(\epsilon_{k,l} - \frac{2B}{n(n-1)}\right)\right] \leq \frac{4B}{n^2(n-1)^2},$$

and to the fact that the collection $\{\epsilon_{i,j} : (i,j) \in \Lambda\}$ is independent from the training data $\mathbb{D}_n$ by assumption. □

For the Bernoulli case, we apply Bernstein inequality to the sum $Z(g)$ of the r.v.'s

$$\left(\epsilon_{i,j} - \frac{2B}{n(n-1)}\right)\mathbb{I}\{g(X_i, X_j) \neq e_{i,j}\}$$

conditioned upon the graph $\mathbb{D}_n$, which are bounded by 1 and have conditional variance less than $2B/(n(n-1))$. We obtain: $\forall g \in \mathcal{G}$, $\forall \zeta > 0$,

$$\mathbb{P}\left\{|Z(g)| > \zeta \mid \mathbb{D}_n\right\} \leq 2\exp\left(-\frac{\zeta^2}{2B + 2\zeta/3}\right).$$

Using the union bound, one gets: $\forall \zeta > 0$,

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}} |Z(g)| > B\zeta \mid \mathbb{D}_n\right\} \leq 2\exp\left(-\frac{B\zeta^2}{2 + 2\zeta/3}\right).$$

Solving $\delta = 2(1 + n(n-1)/2)^d \exp(-B\zeta^2/(2 + 2\zeta/3))$ leads to the first bound.

Turning next to the second bound, the exponential inequality tailored to the SWOR case (see Serfling, 1974, Corollary 1.1) yields:

$$\mathbb{P}\left\{\frac{1}{B}\left|Z(g)\right| > \zeta \mid \mathbb{D}_n\right\} \leq 2\exp\left(-\frac{B\zeta^2}{2}\right),$$

for all $g \in \mathcal{G}$, $\zeta > 0$. Using the union bound and then solving $\delta = 2(1 + n(n-1)/2)^d \exp(-B\zeta^2/2)$ completes the proof. □

Note that following Section 6.7.7, one can easily derive a version of Theorem 5 for the minimizer of $\widetilde{\mathcal{R}}^{\mathcal{D}}$ by replacing $\widetilde{\Lambda}_B(g)$ with $\widetilde{\Lambda}_B^{\mathcal{D}}(g)$ (the incomplete excess risk corresponding to the sampling plan $\mathcal{D}$) in the decomposition (6.11) and making the appropriate modifications.

CHAPTER 7
Conclusion, Limitations & Perspectives

In this manuscript we study and implement sampling strategies for statistical learning related problems. By establishing theoretical results and displaying numerical experiments, we show that the different strategies that we propose are an efficient way to deal with scaling issues. In particular, we show how sampling strategies are useful to reduce the complexity induced by the size of the dataset, for problems typically arising in a Big Data context. We highlight the impact of such strategies by establishing upper bounds and asymptotic limit illustrating the trade-off between accuracy and statistical error. We also show how such strategies can be used to speed-up the learning process by providing an analysis in the spirit of Bottou & Bousquet (2008), bridging the gap between results from statistics and optimization.

For the first problem we consider in Chapter 2 (*i.e* learning from survey training samples), we show that learning is possible when taking into account the sampling procedure used to sample the observations. Our analysis focuses on the binary Classification problem but can be extended to more general frameworks. However, in Chapter 6, we introduce fast learning rates and the type of hypothesis on the data distribution required to prove them. It is therefore natural to investigate to which extent fast learning rates results can be established when observation are drawn from a general survey sampling scheme. The analysis we develop in chapter 2 would fail because of two main reasons:

- Fast learning rates are established because of small variance property of minimizer of the empirical risk. Unfortunately our analysis is based on upper bounding $\sup_{g\in\mathcal{G}} \left|\bar{L}_{\boldsymbol{\epsilon}_n}(g) - \widehat{L}_n(g)\right|$ and we can not expect to establish fast rates for this quantity. However, this difficulty is easily overcome by controlling the risk as we do in chapter 6 (or in Boucheron et al. (2005a)).
- The second issue is that we separate the randomness coming from the observations with the randomness coming from the sampling procedure when decomposing the risk. We then dealt with each term separately by using well known results and working conditionally upon the observations. The fast rates being established under assumptions on the data distribution, our analysis would therefore fail.

An other point of interest to study is what lower bounds can we expect to have on the quantity $\sup_{g\in\mathcal{G}} \left|\bar{L}_{\boldsymbol{\epsilon}_n}(g) - \widehat{L}_n(g)\right|$. In the rejective case (and more generally for negatively associated r.v.) we expect it to be of order $O(1/\sqrt{N})$ because negatively associated r.v. "behaves at least as well" as independent r.v. for which lower bound exists (see Lecué & Mendelson (2010) in the i.i.d case).

The second problem we consider in Chapter 3, 4 and 5 was the implementation of sampling strategies for SGD. A very simple extension to our results would be to see how

they can be extended in presence of some regularization term. When the regularization term is not differentiable, one typically use the stochastic proximal gradient descent (see Atchade et al. (2014) for instance) which essentially boils down to perform one step of SGD on the data-fitting term before performing a proximal step on the regularization term. Incorporating non uniform sampling strategies is therefore a problem that would be very interesting to study for stochastic proximal gradient descent. Finally and as mentioned in chapter 3, significant advances have been recently made in the design of efficient incremental methods(see for instance Mairal (2014), Mairal (2013), Johnson & Zhang (2013a), Shalev-Shwartz & Zhang (2012), Schmidt et al. (2013) or Defazio et al. (2014)) achieving better performances than the traditional SGD method, some of these methods do work when observation are sampled non uniformly but they do not shed light on how observations should be sampled or what would be the potential gain on the risk, which would be very interesting to study.

For the problem we considered in chapter 6, we prove that the learning rates for ERM in the graph reconstruction problem are always of order $O(\log n/n)$, it would be interesting to see if we can establish a lower bound that matches this upper bound. Another challenging line of future research is to see what happens when the noise condition of assumption 8 is satisfied. It would straightforwardly lead to fast rate for minimizer of (6.5) but what it would change for minimizers of (6.3) is an open question. Extension of the results of this chapter to more complex dependence structures can also be investigated. For instance, the probability that two nodes are connected could be modelled as a function of their features, as in the present framework, but also on those of their neighbours in the graph,(*i.e.* relaxing the independence assumption for the $e_{i,j}$'s) in order to account for properties shared by some real-world graphs (*e.g.*, graphs with a power law degree distribution).

APPENDIX A
Concentration Inequalities and Applications to Empirical Risk Minimization

We introduce well-known results as well as the methodology typically applied to derive concentration inequalities. The ideas and results introduced in this chapter are of the utmost importance and are used all along this manuscript. We refer to McDiarmid (1998) and Janson (2002) for good references on this subject, and Massart (2007); Boucheron et al. (2013) for a very complete review on concentration inequalities. References on classification and statistical learning theory include Vapnik & Chervonenkis (1974); Devroye et al. (1996b); Bousquet et al. (2004); Boucheron et al. (2005a); Bishop (2006); Friedman et al. (2001); Vapnik (2013). We keep the notations introduced for the binary classification problem of section 1.1. We recall that $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and $Z = (X, Y)$ a random pair defined on $(\Omega, \mathcal{A}, \mathbb{P})$, taking its values in some measurable product space $\mathcal{X} \times \{-1, +1\}$ and for some classifier g in $\mathcal{G}$, $L(g) = \mathbb{P}\{g(X) \neq Y\}$ denotes its risk and $\widehat{L}_n(g)$ its empirical counterpart. We first present in section A.1 some classical result pertaining to the Vapnik-Chervonenkis theory as well as the methodology used to establish this result. In section A.2 we give traditional concentration inequalities that we use in this manuscript.

A.1 Vapnik-Chervonenkis's Inequality and The Method of Bounded Difference

This section is a step by step guide on how to establish Vapnik-Chervonenkis's inequality A.4 for the binary classification problem. Its purpose is to serve as a reference for the different problems we consider in this manuscript. Let $\widehat{g}_n$ be an empirical risk minimizer and $\tilde{g}^* \in \arg\min_{g \in \mathcal{G}} L(g)$. As mentioned in chapter 1, the ERM paradigm requires to establish upper bound for the quantity $L(\widehat{g}_n) - L^*$. This is typically upper bounded by introducing the empirical counterpart of the risk and rewriting this quantity as:

$$L(\widehat{g}_n) - L^* = L(\widehat{g}_n) - \widehat{L}_n(\widehat{g}_n) + \widehat{L}_n(\widehat{g}_n) - \widehat{L}_n(\tilde{g}^*) + \widehat{L}_n(\tilde{g}^*) - L(\tilde{g}^*) + L(\tilde{g}^*) - L^*.$$

Since $\widehat{g}_n$ is an empirical risk minimizer, the quantity $\widehat{L}_n(\widehat{g}_n) - \widehat{L}_n(\tilde{g}^*)$ is non positive. The quantities $L(\widehat{g}_n) - \widehat{L}_n(\widehat{g}_n)$ and $\widehat{L}_n(\tilde{g}^*) - L(\tilde{g}^*)$ are upper bounded by $\sup_{g \in \mathcal{G}} |L(g) - \widehat{L}_n(g)|$ which overall leads to the bound (1.4):

$$L(\widehat{g}_n) - L^* \leq 2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right). \tag{A.1}$$

The class $\mathcal{G}$ is supposed rich enough to make the second term on the right hand side small, so that we only have to check that $2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)|$ is small to be sure that the empirical risk minimizer $\widehat{g}_n$ achieves a similar performance to the one achieved by $\tilde{g}^*$. This result is

typically established by showing that the r.v. $\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-L(g)|$ does not vary much from its expectation by using the bounded difference inequality McDiarmid (1998):

Proposition A.1. *(*MCDIARMID INEQUALITY, OR 'INDEPENDENT BOUNDED DIFFERENCES INEQUALITY'*) Let $Z=(Z_1,\dots,Z_n)$ where the Z_i's are independent r.v. with values in $\mathcal{X}\times\{-1,+1\}$. Let $f:(\mathcal{X}\times\{-1,+1\})^n\to\mathbb{R}$ verifying the following Lipschitz condition.*

$$\text{For any } z,\ z'\in(\mathcal{X}\times\{-1,+1\})^n,\quad |f(z)-f(z')|\leq c_k \quad \text{if} \quad z_j=z'_j,\ \text{for } j\neq k,\ 1\leq j\leq n. \tag{A.2}$$

Let $\mu=\mathbb{E}\left[f(Z)\right]$. Then, for any $t\geq 0$,

$$\mathbb{P}\left[f(Z)-\mu\geq t\right]\leq e^{-2t^2/\sum c_k^2}\,.$$

The same inequality holds true when replacing $f(X)-\mu$ by $\mu-f(X)$.

It is easy to check that Proposition A.1 can be applied to the r.v. $\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-L(g)|$ with $c_k=\frac{2}{n}$. In particular, this gives that with probability at least $1-\delta$

$$\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-L(g)|\leqslant\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-L(g)|]+\sqrt{\frac{2\log(1/\delta)}{n}}.$$

The next step is to upper bound $\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-L(g)|]$ by distribution free constant depending on the number of observations n and the richness of the class $\mathcal{G}$. A first step toward this direction is to introduce a ghost sample $z'_1,\dots,z'_n$ independent of the z_i with same distribution to symmetrize our quantity. We denote by $\widehat{L}'_n$ the empirical risk based on the ghost sample. Then we have:

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-L(g)|]&=\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-\mathbb{E}[\widehat{L}'_n(g)|Z_1,\dots,Z_n]|]\\&\leqslant\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-\widehat{L}'_n(g)|].\end{aligned}$$

We then randomize this inequality by introducing rademacher r.v. $\sigma_1,\dots,\sigma_n$ with $\mathbb{P}(\sigma_i=1)=\mathbb{P}(\sigma_i=-1)=\frac{1}{2}$ independent from our samples $Z_1,\dots,Z_n$ and $Z'_1,\dots,Z'_n$. We have:

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{L}_n(g)-\widehat{L}'_n(g)|]&=\mathbb{E}\left\{\sup_{g\in\mathcal{G}}|\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{g(X_i)\neq Y_i\}-\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{g(X'_i)\neq Y'_i\}|\right\}\\&=\mathbb{E}\left\{\sup_{g\in\mathcal{G}}|\frac{1}{n}\sum_{i=1}^n\sigma_i(\mathbb{I}\{g(X_i)\neq Y_i\}-\mathbb{I}\{g(X'_i)\neq Y'_i\})|\right\}\\&\leqslant 2\mathbb{E}\left\{\sup_{g\in\mathcal{G}}|\frac{1}{n}\sum_{i=1}^n\sigma_i\mathbb{I}\{g(X_i)\neq Y_i\}|\right\}.\end{aligned}$$

The quantity $\mathbb{E}[\sup_{g\in\mathcal{G}}|\frac{1}{n}\sum_{i=1}^n\sigma_i\mathbb{I}\{g(X_i)\neq Y_i\}|]$ is very interesting as it can be expressed as:

$$\mathbb{E}[\sup_{g\in\mathcal{G}}|\frac{1}{n}\sum_{i=1}^n\sigma_i\mathbb{I}\{g(X_i)\neq Y_i\}|]=\mathbb{E}[\mathbb{E}[\sup_{g\in\mathcal{G}}|\frac{1}{n}\sum_{i=1}^n\sigma_i\mathbb{I}\{g(X_i)\neq Y_i\}|]|Z_1,\dots,Z_n], \tag{A.3}$$

and is the first step toward a distribution-free bound. Indeed, if for any bounded set of vectors $\mathcal{V}$ in $\mathbb{R}^n$ we define $R_n(\mathcal{V})=\mathbb{E}[\sup_{v\in\mathcal{V}}|\frac{1}{n}\sum_{i=1}^n\sigma_iv_i|]$ as the *Rademacher Average* of order n

associated to $\mathcal{V}$, then

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - \widehat{L}'_n(g)|] \leqslant 2\mathbb{E}[R_n(\mathcal{G}(\mathcal{Z}^n))] \tag{A.4}$$

where $\mathcal{G}(\mathcal{Z}^n)$ denotes the class of vector $(\mathbb{I}\{g(X_1) \neq Y_1\}, \ldots, \mathbb{I}\{g(X_n) \neq Y_n\})$ for any g in $\mathcal{G}$. It is very easy to establish upper bound for $R_n(\mathcal{V})$ when the set $\mathcal{V}$ is finite (see Boucheron et al. (2005a) for instance).

Proposition A.2. *Assume $\mathcal{V}$ has cardinal N_0 and let $V_0 \in \mathbb{R}$ such that for any $v \in \mathcal{V}$, $||v|| \leqslant V_0$, then:*

$$R_n(\mathcal{V}) \leqslant V_0 \frac{\sqrt{2\log(N_0)}}{n}.$$

The proof of this proposition is given remark A.8. For our problem, $\mathcal{G}(\mathcal{Z}^n)$ is a subset of $\{0,+1\}^n$ so its elements have norm smaller than $\sqrt{n}$ and its cardinality is upper bounded by 2^n. Unfortunately this is not satisfactory when plugged into proposition A.2 because th bound obtained would be a $O(1)$. Note that if we could get an upper bound on the cardinal of $\mathcal{G}(\mathcal{Z}^n)$ that would be polynomial in n instead of exponential, then proposition A.2 would give an upper bound of order $O(\frac{\log(n)}{n})$. This is essentially what happens what when we assume that the class of indicator functions indexed by $\mathcal{G}$ has finite VC-dimension. It is defined for any set $\mathcal{V}$ as:

Definition A.3. Let $\mathcal{V}$ a subset of $\{0,+1\}^n$, we call VC dimension of $\mathcal{V}$ the size V of the largest set of indices $\{i_1, \ldots, i_V\} \subset \{1, \ldots, n\}$ such that for each binary V-vector $b = (b_1, \ldots, b_V)$ there exists $v = (v_1, \ldots, v_n)$ in $\mathcal{V}$ such that $(v_{i_1}, \ldots, v_{i_V}) = b$.

An other very important result is Sauer's Lemma (see Sauer (1972) for instance) which states that a subset $\mathcal{V}$ with VC dimension V has its cardinality bounded by $(n+1)^V$.

Going back to our problem, the cardinal of $\mathcal{G}(\mathcal{Z}^n)$ is called the VC shatter coefficient and let V_n be its VC-dimension. The logarithm of the shatter coefficient is upper bounded by $V_n \log(n+1)$ thanks to Sauer's Lemma. If we assume that we have finite VC-dimension, then there exists $V \leq \infty$ such that $\sup_{n\in\mathbb{N}} V_n \leqslant V$. Applying proposition A.2 this leads to distribution-free upper bound known as Vapnik-Chervonenkis's inequality:

Theorem A.4. (VAPNIK-CHERVONENKIS) *For any distribution, we have with probability at least $1-\delta$:*

$$L(\widehat{g}_n) - L^* \leqslant L(\tilde{g}^*) - L^* + 2\sqrt{\frac{2V\log(n+1)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{A.5}$$

The VC-dimension is an important combinatorial parameter of the class, in particular, when $\mathcal{G}$ is a d-dimensional vector space of real-valued functions, the class of indicator functions $\{\mathbb{I}\{g(.) \geq 0\}\}$ has finite VC-dimension $V \leqslant d$. The rest of this chapter is devoted to present other useful results for statistical learning.

A.2 Concentration Inequalities for Empirical Risk Minimization

In this section we present popular inequalities often used in statistical learning. We first deal with bounded r.v. and establish McDiarmid's inequality and Azuma-Hoeffding's inequality.

We then focus on r.v. whose variance is bounded and establish Bernstein's inequality. We first set a few definitions :

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be a random variable on this space and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$.

Notation 1. X is a real r.v. in $L^\infty(\Omega)$. The conditional essential supremum $\sup(X|\mathcal{G})$ is the unique real r.v. $f\tilde{X} : \Omega \to \mathbb{R}$ satisfying:

- $\tilde{X}$ is $\mathcal{G}$-measurable
- $X \leq \tilde{X}$ a.s.
- If $\tilde{Y} : \Omega \to \mathbb{R}$ verifies the two previous conditions then $\tilde{X} \leq \tilde{Y}$ a.s.

It is straightforward to check that $\sup(X|\mathcal{G}) \geq \mathbb{E}(X|\mathcal{G})$ and $\sup(X|\mathcal{G}_1) \geq \sup(X|\mathcal{G}_2)$ when $\mathcal{G}_1 \subset \mathcal{G}_2$. Equipped with these notations we can state our main results in the following section.

A.2.1 Inequality for Bounded Random Variables(Azuma-Hoeffding and McDiarmid)

In this section we work with bounded r.v.. The theorem that we give now is strong enough to derive the fundamental inequality of Azuma-Hoeffding (A.9) and McDiarmid 's bounded difference inequality (A.10). We make use of this inequalities later in chapter 2,5 and 6. It involves the following notations:

Notation 2. Let X be a bounded r.v.. Let $(\mathcal{F}_k)_{0\leq k\leq n}$ be a filtration of $\mathcal{F}$ such that X is $\mathcal{F}_n$-measurable. We denote $X_1, \ldots, X_n$ the martingale $X_k = \mathbb{E}(X|\mathcal{F}_k)$ and $Y_k = X_k - X_{k-1}$ the associated martingale difference. The r.v. $\mathbf{ran}(X|\mathcal{G}) := \sup(X|\mathcal{G}) + \sup(-X|\mathcal{G})$ is the conditional range of X w.r.t $\mathcal{G}$. We also denote:

- $\mathbf{ran_k} = \mathbf{ran}(Y_k|\mathcal{F}_{k-1}) = \mathbf{ran}(X_k|\mathcal{F}_{k-1})$ the conditional range,
- $\mathbf{R^2} = \sum_1^n \mathbf{ran_k}^2$ the sum of squared conditional ranges, and $\hat{\mathbf{r}}^\mathbf{2} = \operatorname{ess\,sup}(R^2)$ the maximum sum of squared conditional ranges.

A.2.1.1 A Preliminary Theorem

Here we establish the following theorem and give its proof. The methodology used to establish this result is classical and treated as an example. In particular its proof will be adapted to derive concentration inequalities in chapter 2.

Theorem A.5. *(McDiarmid, 1998) Let X be a bounded r.v. with $\mathbb{E}(X) = \mu$, and $(\mathcal{F}_k)_{0\leq k\leq n}$ a filtration of $\mathcal{F}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and such that X is $\mathcal{F}_n$-measurable. Then for any $t \geq 0$,*

$$\mathbb{P}(X - \mu \geq t) \leq e^{-2t^2/\hat{r}^2},$$

and more generally

$$\forall r^2 \geq 0, \quad \mathbb{P}((X - \mu \geq t) \cap (R^2 \leq r^2)) \leq e^{-2t^2/r^2}.$$

The main idea used to derive this theorem is to express $X-\mu$ as a sum of martingale increment: $X-\mu=\sum_{i=1}^{n} Y_k$ and control its Laplace transform using the following lemma:

Lemma A.6. *Let $(\mathcal{F}_k)_{0\leq k\leq n}$ be a filtration of $\mathcal{F}$ with $\mathcal{F}_0=\{\emptyset,\Omega\}$, and $(Y_k)_{1\leq k\leq n}$ be a martingale difference for this filtration such that each Y_k is bounded. Let Z be any random variable. Then*

$$\mathbb{E}(Ze^{h\sum Y_k})\leq \sup(Z\prod_{k=1}^{n}\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1})).$$

Proof. This result can be easily proved by induction.

$$\begin{aligned}
\mathbb{E}\left[Ze^{h\sum Y_k}\right] &= \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[Ze^{h\sum_2^n Y_k}\ \middle|\ \mathcal{F}_1\right]\right]\\
&= \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[e^{hY_2}\dots\ \mathbb{E}\left[\mathbb{E}\left[Z\ \middle|\ \mathcal{F}_n\right]e^{hY_n}\ \middle|\ \mathcal{F}_{n-1}\right]\dots\ \middle|\ \mathcal{F}_1\right]\right]\\
&\leq \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[e^{hY_2}\dots\ \mathbb{E}\left[\sup\left[Z\mid\mathcal{F}_n\right]e^{hY_n}\ \middle|\ \mathcal{F}_{n-1}\right]\dots\ \middle|\ \mathcal{F}_1\right]\right]\\
&= \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[e^{hY_2}\dots\ \sup\left[Z\mathbb{E}\left[e^{hY_n}\mid\mathcal{F}_{n-1}\right]\mid\mathcal{F}_n\right]\dots\ \middle|\ \mathcal{F}_1\right]\right]\\
&= \sup\left[Z\prod_k\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1})\ \middle|\ \mathcal{F}_n\right]\\
&\leq \sup\left[Z\prod_k\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1})\right]\qquad(\text{since }\ \mathcal{F}_0\subset\mathcal{F}_n).
\end{aligned}$$

□

With this lemma, we decompose the expectation of a product into a product of expectations. It can be interpreted as doing 'almost as if' $\sum Y_k$ was a sum of independent variables. The next step to prove Theorem A.5 is to control the Laplace transform of bounded r.v., it can be done using Hoeffding's Lemma:

Lemma A.7. *(*HOEFFDING'S LEMMA*) Let X be a centered random variable and $(a,b)\in\mathbb{R}^2$ such that $a\leq X\leq b$, then for any $h>0$, we have $\mathbb{E}(e^{hX})\leq e^{\frac{1}{8}h^2(b-a)^2}$.*

Equipped with these intermediate results, the proof of Theorem A.5 is quite straightforward:

Proof of Theorem A.5. We start with Chernoff method then apply Lemma A.6, before using Lemma A.7, and finally choose the value of parameter h.

Let $X_k=\mathbb{E}(X|\mathcal{F}_{k-1})$ and $Y_k=X_k-X_{k-1}$ the associated martingale difference. Define the r.v. Z as $Z=\mathbb{1}_{R^2\leq r^2}$. Exponential Markov inequality yields, for any $h>0$,

$$\begin{aligned}
\mathbb{P}((X-\mu\geq t)\cap(R^2\leq r^2)) &= \mathbb{P}(Ze^{h(X-\mu)}\geq e^{ht})\\
&\leq e^{-ht}\mathbb{E}(Ze^{h(X-\mu)})\\
&\leq e^{-ht}\mathbb{E}(Ze^{h(\sum Y_k)}).
\end{aligned}$$

From Lemma A.7, $\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1}) \leq e^{\frac{1}{8}h^2r_k^2}$ so that using Lemma A.6,

$$\begin{aligned}
\mathbb{E}(Ze^{h\sum Y_k}) &\leq \sup(Z\prod \mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1})),\\
&\leq \sup(Z\prod e^{\frac{1}{8}h^2r_k^2}),\\
&= \sup(Ze^{\frac{1}{8}h^2R^2})\\
&\leq e^{\frac{1}{8}\sup(ZR^2)}\\
&\leq e^{\frac{1}{8}h^2r^2}.
\end{aligned}$$

We take $h = 4t/r^2$, we finally obtain

$$\mathbb{P}((X-\mu \geq t) \cap (R^2 \leq r^2)) \leq e^{-ht+\frac{1}{8}h^2r^2} \leq e^{-2t^2/r^2}.$$

□

Remark A.8. Using Lemma A.7, we can prove proposition A.2. Indeed for any $\lambda \geq 0$,

$$\begin{aligned}
\exp(\lambda R_n(\mathcal{V})) = \exp(\lambda\mathbb{E}[\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^n\sigma_i v_i]) &\leqslant \mathbb{E}[\exp(\lambda\sup_{v\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^n\sigma_i v_i]\\
&\leqslant \sum_{v\in\mathcal{V}}\mathbb{E}[\exp(\lambda\frac{1}{n}\sum_{i=1}^n\sigma_i v_i)]\\
&\leqslant \sum_{v\in\mathcal{V}}\prod_{i=1}^n\mathbb{E}[\exp(\lambda\sigma_i v_i)],
\end{aligned}$$

where we use the convexity of the exponential for the first inequality. Hoeffding's Lemma A.7 gives $\mathbb{E}[\exp(\lambda\sigma_i v_i)] \leqslant \exp(\lambda^2 v_i^2/2n^2)$ so that

$$\begin{aligned}
\exp(\lambda R_n(\mathcal{V})) &\leqslant \sum_{v\in\mathcal{V}}\exp(\frac{\lambda^2||v||^2}{2n^2})\\
&\leqslant N_0\exp(\frac{\lambda^2V_0^2}{2n^2}).
\end{aligned}$$

Taking $\lambda = \sqrt{\log(N_0)2n^2/V_0^2}$ gives Theorem A.4.

A.2.1.2 Azuma-Hoeffding's Inequality and McDiarmid's Inequality

In this section, we apply Theorem A.5 and derive classical concentration inequalities as corollaries . The first result that we give is Azuma-Hoeffding's inequality that deals with sum of random variables.

Proposition A.9. *(AZUMA-HOEFFDING INEQUALITY) Let $(\mathcal{F}_k)_{0\leq k\leq n}$ be a filtration of $\mathcal{F}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$, Z a martingale and Y the associated martingale difference. If for every k, $|Y_k| \leq c_k$, then we have*

$$\mathbb{P}(\sum_{k=1}^n Y_k \geq t) \leq e^{-\frac{t^2}{2\sum_{k=1}^n c_k^2}}.$$

The same inequality holds true when replacing $\sum_{k=1}^n Y_k$ by $-\sum_{k=1}^n Y_k$.

Proof. Apply Theorem A.5 with $X = \sum_1^n Y_k$, $\mathcal{F}_k = \sigma(Y_1, \dots, Y_k)$ and $X_k = \mathbb{E}(X|\mathcal{F}_k)$. Thus, $\mu = 0$, $X_k = \sum_1^k Y_i$ because Z is a martingale, and $Y_i = X_i - X_{i-1}$. Therefore, $\mathbf{ran_k} = \mathbf{ran}(Y_k|\mathcal{F}_k) \leq 2c_k$, hence $R^2 \leq 4\sum c_k^2$ and $\hat{r}^2 \leq 4\sum c_k^2$. By Theorem A.5, $\mathbb{P}(X \geq t) \leq e^{\frac{-2t^2}{\hat{r}^2}} \leq e^{-\frac{t^2}{2\sum c_k^2}}$. Applying this inequality to $-X$, we obtain the desired result. □

The second Theorem that can be derived from Theorem A.5 is McDiarmid inequality used previously in section A.1.

Proposition A.10. *(*MCDIARMID INEQUALITY, OR 'INDEPENDENT BOUNDED DIFFERENCES INEQUALITY'*) Let $X = (X_1, \dots, X_n)$ where the X_i's are independent r.v. with respected values in A_i. Let $f : \prod A_k \to \mathbb{R}$ verifying the following Lipschitz condition.*

$$\text{For any } x,\ x' \in \prod_1^n A_k, \quad |f(x) - f(x')| \leq c_k \quad \text{if} \quad x_j = x'_j, \ \text{for} \ j \neq k, \ 1 \leq j \leq n. \tag{A.6}$$

Let us denote $\mu = \mathbb{E}[f(X)]$. Then, for any $t \geq 0$,

$$\mathbb{P}[f(X) - \mu \geq t] \leq e^{-2t^2/\sum c_k^2}.$$

The same inequality holds true when replacing $f(X) - \mu$ by $\mu - f(X)$.

Proof. Lipschitz condition (A.6) implies that f is bounded, thus from Theorem A.5 we have

$$\mathbb{P}[f(X) - \mu \geq t] \leq e^{-2t^2/\hat{r}^2},$$

where $\hat{r}^2$ is defined with the filtration $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ and $X = f(X_1, \dots, X_n)$. We can easily establish an upper bound on $\hat{r}^2$: $\mathbf{ran_k} = \mathbf{ran}(\mathbb{E}(f(X)|\mathcal{F}_k) - \mathbb{E}[f(X)|\mathcal{F}_{k-1}]\,|\mathcal{F}_{k-1}) \leq c_k$. This concludes the proof □

The following inequality is similar to Proposition A.9 and can be applied to sum of independent r.v.. It provides a tighter bound than the one in Proposition A.9.

Proposition A.11. *(*HOEFFDING INEQUALITY*) Let $X_1, \dots, X_n$ be n independent random variables such that $a_i \leq X_i \leq b_i$, $1 \leq i \leq n$. Define $S_n = \sum X_k$ and $\mu = \mathbb{E}(S_n)$. Then,*

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-2t^2/\sum (b_k - a_k)^2}.$$

The same inequality holds true when replacing $S_n - \mu$ by $\mu - S_n$.

Proof. This is a immediate consequence of previous McDiarmid inequality (Proposition A.10) with $A_k = [a_k, b_k]$, $f(x) = \sum x_k$ and $c_k = b_k - a_k$. Within this setting, $\hat{r}^2 \leq b_k - a_k$. □

A.2.2 Bernstein-type Inequality (with Variance Term)

Now, we remove our boundedness hypothesis and instead make variance assumption. The theorem stated below will help us derive the popular Bernstein inequality, that is used to derive fast rate of convergence under appropriate variance control (see Tsybakov (2004)). We make use of a general version of this inequality in chapter 2 and 6. We first set a few notations:

Notation 3. The r.v. $\mathbf{var}(X|\mathcal{G}) := \mathbb{E}((X - \mathbb{E}(X|\mathcal{G}))^2|\mathcal{G})$ is called the conditional variance of X *w.r.t.* $\mathcal{G}$ and we set:

- $\mathbf{var_k} = \mathbf{var}(Y_k|\mathcal{F}_{k-1}) = \mathbf{var}(X_k|\mathcal{F}_{k-1})$ the conditional variance,
- $\mathbf{V} = \sum_1^n \mathbf{var_k}$ the sum of conditional variances and $\hat{\nu} = \operatorname{ess\,sup}(V)$ the maximum sum of conditional variances,
- $\mathbf{dev_k^+} = \sup(Y_k|\mathcal{F}_{k-1})$ the conditional positive deviation,
- $\mathbf{maxdev}^+ = \operatorname{ess\,sup}(\max_{0\leq k\leq n} dev_k^+)$ the maximum conditional positive deviation.

We now give the following theorem that will help us establish Bernstein-type inequalities.

Theorem A.12. *(McDiarmid, 1998) Let X be a r.v. with $\mathbb{E}(X) = \mu$ and $(\mathcal{F}_k)_{0\leq k\leq n}$ a filtration of $\mathcal{F}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and such that X is $\mathcal{F}_n$-measurable. Let $b = \mathbf{maxdev}^+$ the maximum conditional deviation assumed to be finite, and $\hat{\nu} = \operatorname{ess\,sup} V$ the maximum sum of conditional variances also assumed to be finite. Then, for any $t \geq 0$,*

$$\mathbb{P}(X - \mu \geq t) \leq e^{-\frac{t^2}{2(\hat{\nu}+bt/3)}},$$

and more generally, for any $v \geq 0$,

$$\mathbb{P}((X - \mu \geq t) \cap (V \leq v)) \leq e^{-\frac{t^2}{2(v+bt/3)}}.$$

In spirit, the proof of Theorem A.12 is similar to the proof of Theorem A.5 but require a different control of the Laplace transform to make use of variance assumptions. This is achieved by the following Lemma which combined with Hoeffding's lemma A.7 gives the result.

Lemma A.13. *Let g defined for $x \neq 0$ by $g(x) = \frac{e^x-1-x}{x^2}$, and X a centered r.v. satisfting for some $b \geq 0$, $X \leq b$. Then $\mathbb{E}(e^X) \leq e^{g(b)var(X)}$.*

Proof. g is non decreasing so by taylor expansion of the exponential we have $e^x \leq 1 + x + x^2 g(b)$ for $x \leq b$. Taking expectation yields $\mathbb{E}(e^X) \leq 1 + g(b)var(X) \leq e^{g(b)var(X)}$. □

Proof of Theorem A.12. The proof follows the same classical lines as the ones of Theorem A.5. Let $Y_1, \ldots, Y_n$ be the martingale differences associated to X and $(\mathcal{F}_k)$, and $Z = \mathbb{1}_{V\leq v}$. Exponential Markov inequality yields, for every $h > 0$,

$$\begin{aligned}
\mathbb{P}((X - \mu \geq t) \cap (V \leq v)) &= \mathbb{P}(Ze^{h(X-\mu)} \geq e^{ht}) \\
&\leq e^{-ht}\mathbb{E}(Ze^{h(X-\mu)}) \\
&\leq e^{-ht}\mathbb{E}(Ze^{h(\sum Y_k)})
\end{aligned}$$

From Lemma A.13, $\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1}) \leq e^{h^2 g(hdev_k^+)var_k} \leq e^{h^2 g(hb)var_k}$ so that from Lemma A.6 we obtain,

$$\begin{aligned}
\mathbb{E}(Ze^{h\sum Y_k}) &\leq \sup(Z\prod \mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1})) \\
&\leq \sup(Z\prod e^{h^2 g(hb)var_k}) \\
&= \sup(Ze^{h^2 g(hb)V}) \\
&\leq e^{h^2 g(hb)\sup(ZV)} \\
&\leq e^{h^2 g(hb)v}.
\end{aligned}$$

By setting $h = \frac{1}{b}ln(1+\frac{bt}{v})$ and using the fact that for every positive x, we have $(1+x)\ln(1+x) - x \geq 3x^2/(6+2x)$, we finally get

$$\begin{aligned}
\mathbb{P}((X-\mu \geq t) \cap (R^2 \leq r^2)) &\leq e^{-ht+h^2 g(hb)v} \\
&\leq e^{-\frac{t^2}{2(v+bt/3)}}.
\end{aligned}$$

□

We finally state Bernstein's inequality as a corollary of Theorem A.12.

Proposition A.14. *(*BERNSTEIN INEQUALITY*) Let $X_1, \ldots, X_n$ be n independent random variables with $X_k - \mathbb{E}(X_k) \leq b$. We consider their sum $S_n = \sum X_k$, the sum variance $V = var(S_n)$ as well as the sum expectation $\mathbb{E}(S_n) = \mu$. Then, for any $t \geq 0$,*

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-\frac{t^2}{2(V+bt/3)}},$$

and more generally,

$$\mathbb{P}((S_n - \mu \geq t) \cap (V \leq v)) \leq e^{-\frac{t^2}{2(v+bt/3)}}.$$

Remark A.15. (GAIN WITH RESPECT TO INEQUALITIES WITHOUT VARIANCE TERM) Assume that $0 \leq X_i \leq 1$ and consider renormalized quantities, namely $\tilde{S}_n := S_n/n$, $\tilde{\mu} := \mu/n$, $\tilde{V} = V/n^2$. Then,

$$\begin{aligned}
&\mathbb{P}(\tilde{S}_n - \tilde{\mu} \geq t) \leq e^{-2nt^2} && \text{(Hoeffding)} \\
\text{and} \quad &\mathbb{P}(\tilde{S}_n - \tilde{\mu} \geq t) \leq e^{-\frac{nt^2}{2(\tilde{V}+t/3)}} && \text{(Bernstein),}
\end{aligned}$$

with t typically of order between $1/n$ and $1/\sqrt{n}$. Thus, if the variance $\tilde{V}$ is small enough, Bernstein inequality 'almost' allows to have rates in e^{-nt} instead of e^{-nt^2}. In other words, Bernstein-type inequality may give high probability bounds in $\frac{1}{n}\log\frac{1}{\delta}$ instead of $\sqrt{\frac{1}{n}\log\frac{1}{\delta}}$. This is of the utmost importance in the fast rate analysis we provide in chapter 6.

Proof. Let $F_k = \sigma(X_1, \ldots, X_n)$, $X = \sum(X_k - \mathbb{E}X_k) = S_n - \mu$, $\tilde{X}_k = \mathbb{E}(X|\mathcal{F}_k) = \sum_1^k (X_i - \mathbb{E}X_i)$ and $Y_k = \tilde{X}_k - \tilde{X}_{k-1}$. Then $Y_k = X_k - \mathbb{E}X_k$, hence $dev_k^+ \leq b$, $maxdev^+ \leq b$ and $var_k = var(Y_k|\mathcal{F}_{k-1}) = \mathbb{E}((Y_k - \mathbb{E}(Y_k|\mathcal{F}_{k-1}))^2|\mathcal{F}_{k-1}) = \mathbb{E}((Y_k - \mathbb{E}Y_k)^2) = var(Y_k)$.

Therefore $\hat{\nu} = \operatorname{ess\,sup}(\sum var_k) = \operatorname{ess\,sup}(V) = V$. Theorem A.12 applies and yields,

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-\frac{t^2}{2(V+bt/3)}},$$
$$\mathbb{P}((S_n - \mu \geq t) \cap (V \leq v)) \leq e^{-\frac{t^2}{2(v+bt/3)}}.$$

□

PART IV

Résumé des contributions en Français

APPENDIX B
Résumé des contributions en français

B.1 Motivation

Dans ce manuscrit, nous présentons et étudions des stratégies d'échantillonnage appliquées à problèmes liés à l'apprentissage statistique. L'objectif est de traiter les problèmes qui surviennent généralement dans un contexte de données volumineuses lorsque le nombre d'observations et leur dimensionnalité contraignent le processus d'apprentissage. Nous proposons donc d'aborder ce problème en utilisant deux stratégies d'échantillonnage:

- Accélérer le processus d'apprentissage en échantillonnant les observations les plus utiles.
- Simplifier le problème en écartant certaines observations pour réduire la complexité et la taille du problème.

Pour introduire le problème que nous traitons, nous introduisons très rapidement la minimisation du risque empirique (ERM) dans le contexte de la classification binaire. Le problème de classification binaire est considéré comme un exemple récurrent tout au long de ce manuscrit. Parce qu'il peut être facilement formulé, il est indéniablement le problème d'apprentissage statistique le plus documenté dans la littérature et beaucoup de ses résultats s'étendent à des cadres plus généraux (*e.g.*, Classification multi-classe, régression, ranking). Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et (X, Y) une paire de variables aléatoires définie sur $(\Omega, \mathcal{A}, \mathbb{P})$, prenant ses valeurs dans un espace mesurable $\mathcal{X} \times \{-1, +1\}$, avec distribution jointe $P(dx, dy)$: la variable aléatoire X représente une observation, utile pour prédire le label Y. La distribution P peut aussi être décrite par la paire (F, η) où $F(dx)$ désigne la distribution marginale de la variable d'entrée X et $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$, est la distribution *postérieure*. L'objectif est de construire, à partir de l'ensemble de données d'apprentissage à disposition, un mapping mesurable $g : \mathcal{X} \mapsto \{-1, +1\}$, appelé *classifier*, avec riqsue minimum:

$$L(g) \stackrel{def}{=} \mathbb{P}\{g(X) \neq Y\}. \tag{B.1}$$

Il est bien connu que le *Bayes classifier* $g^*(x) = 2\mathbb{I}\{\eta(x) \geq 1/2\} - 1$ est une solution du problème de minimisation du risque $\inf_g L(g)$, où l'infinimum est pris sur la collection de tous les classifieurs définis sur l'espace $\mathcal{X}$. Le risque minimum est noté $L^* = L(g^*)$. Puisque la distribution P des données est inconnue, le vrai risque est remplacé par son estimation empirique:

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}, \tag{B.2}$$

basé sur l'échantillon d'exemples $(X_1, Y_1), \ \ldots, \ (X_n, Y_n)$ de copies indépendantes de (X, Y). La véritable minimisation du risque est alors remplacée par la minimisation empirique

du risque

$$\min_{g \in \mathcal{G}} \widehat{L}_n(g), \tag{B.3}$$

où le minimum est pris sur une classe $\mathcal{G}$ de classifieurs, supposés assez riches pour inclure le classifieur naïf de Bayes (ou une approximation raisonnable de ce dernier). Considérant une solution $\widehat{g}_n$ de (B.3), un problème majeur dans la théorie de l'apprentissage statistique est d'établir des bornes de confiance sur l' *excès de risque* $L(\widehat{g}_n) - L^*$ en l'absence de toute hypothèse distributionnelle mais en prenant en compte de la complexité de la classe $\mathcal{G}$ (*eg*, décrite par des caractéristiques géométriques ou combinatoires telles que la dimension de Vapnick-Chervnonenkis VC), et un controle de l'approximation de P par sa contrepartie empirique $P_n = (1/n) \sum_{i=1}^n \delta_{(X_i,Y_i)}$ sur la classe $\mathcal{G}$. En effet, l'excès de risque des minimiseurs du risque empirique est typiquement borné de ma façon suivante:

$$L(\widehat{g}_n) - L^* \leq 2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right). \tag{B.4}$$

Le second terme du côté droit est appelé *bias* et dépend de la richesse de la classe $\mathcal{G}$, tandis que le premier terme, appelé *erreur stochastique*, est contrôlé au moyen de résultats de la théorie des processus empiriques, voir Boucheron et al. (2005a). Malheureusement, l'une des choses généralement négligées dans ce type d'analyse est de savoir comment résoudre efficacement le problème de minimisation du ridque empirique, c'est-à-dire comment trouver $\widehat{g}_n$. Ce problème est habituellement résolu par un algorithme d'optimisation incrémentale, calculant itérativement l'estimateur du gradient du risque empirique. Nous étudions des moyens efficaces d'améliorer le processus d'apprentissage et comment introduire des approches basées sur l'échantillonnage pour construire des approximations de $\widehat{g}_n$. Nous le faisons de deux manières différentes:

- Nous remplaçons le risque empirique $\widehat{L}_n(g)$ par une approximation basée sur moins de termes $\tilde{L}_n(g)$. Cela facilite naturellement le problème d'apprentissage. Soit $\tilde{g}_n$ un minimiseur de $\tilde{L}_n(g)$, alors (B.4) devient:

$$L(\tilde{g}_n) - L^* \leq 2 \sup_{g \in \mathcal{G}} |\tilde{L}_n(g) - \widehat{L}_n(g)| + 2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right).$$

  Nous nous concentrons sur le contrôle approprié $2 \sup_{g \in \mathcal{G}} |\tilde{L}_n(g) - \widehat{L}_n(g)|$ que nous faisons généralement conditionnellement sur les observations. Une telle stratégie est discutée et implémentée dans les chapitres **??** et 6 pour deux problèmes différents.

- Lors du calcul de l'estimateur du gradient, la plupart des algorithmes incrémentaux échantillonnent uniformément et indépendamment les observations dans l'ensemble de données. Nous proposons d'utiliser des méthodes d'échantillonnage non uniformes pour calculer un estimateur du gradient de $\widehat{L}_n$ avec une variance plus faible. Pour les algorithmes que nous proposons, si nous notons $g_n(T)$ le classificateur obtenu après T itérations de l'algorithme d'optimisation, alors suivant le raisonnement introduit dans de Bottou & Bousquet (2007), l'inégalité (B.4) peut être bornée par:

$$L(g_n(T)) - L^* \leq \underbrace{\widehat{L}_n(g_n(T)) - \widehat{L}_n(\widehat{g}_n)}_{(1)} + \underbrace{2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)|}_{(2)} + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right).$$

où (1) correspond à l'erreur d'optimisation et (2) correspond à l'erreur stochastique. Cette décomposition illustre le fait bien connu (voir Bottou & Bousquet (2007)) que lorsque nous résolvons le problème de minimisation des risques empiriques, nous devons prendre en compte le caractère aléatoire induit par les observations afin que l'erreur d'optimisation soit du même ordre que le terme de généralisation. Nous portons particulièrement attention à ce fait et l'illustrons théoriquement et empiriquement dans les chapitres 3, 4 et 5.

Le reste de ce chapitre est consacré à la présentation de nos différentes contributions. Ici et dans le reste de ce chapitre, la fonction indicateur de tout événement $\mathcal{E}$ est notée par $\mathbb{I}\{\mathcal{E}\}$ et la variance de toute variable aléatoire de carré intégrable Z par $\sigma^2(Z)$.

B.2 Apprendre de données de sondages

Cette sous-section est un résumé du chapitre 2. Nous nous plaçons dans le contexte de la classification binaire, lorsque les observations utilisées pour former un classificateur sont issues d'un schéma d'échantillonnage/sondage et présentent une structure de dépendance complexe. Nous considérons, $(X_1, Y_1),\ \ldots,\ (X_n, Y_n)$ un échantillon de copies indépendantes de (X, Y) observées sur une population finie $\mathcal{I}_n := \{1,\ \ldots,\ n\}$. Nous appelons un *survey sample* de taille (éventuellement aléatoire) $N \leq n$ de la population $\mathcal{I}_n$, tout sous-ensemble $s := \{i_1, \ldots, i_{n(s)}\} \in \mathcal{P}(\mathcal{I}_n)$ avec cardinalité $N =: N(s)$ inférieur à n. Un schéma d'échantillonnage est défini par une distribution de probabilité R_n sur l'ensemble de tous les échantillons possibles $s \in \mathcal{P}(\mathcal{I}_n)$ conditionnellement aux observations $\mathcal{D}_n = \{(X_i, Y_i) :\ i \in \mathcal{I}_n\}$. La probabilité que l'unité i appartienne à

Theorem B.1. *Let θ_t be the sequence generated by SGD using the incomplete statistic gradient estimator (5.6) with $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$ terms for some $n'_1, \ldots, n'_K$. Assume that $\{L(.;\ \theta) :\ \theta \in \Theta\}$ is a* VC *major class class of finite* VC *dimension V s.t.*

$$\mathcal{M}_\Theta = \sup_{\theta\in\Theta,\ (\mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(K)})\in\prod_{k=1}^{K}\mathcal{X}_k^{d_k}} |H(\mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(K)};\ \theta)| < +\infty, \tag{B.5}$$

and $\mathcal{N}_\Theta = \sup_{\theta\in\Theta} \sigma_\theta^2 < +\infty$. If the step size satisfies the condition of Proposition 5.4, we have:

$$\forall \mathbf{n} \in \mathbb{N}^{*K}, \quad \mathbb{E}[|L(\theta_t) - L(\theta^*)|] \leqslant \frac{C\mathcal{N}_\Theta}{Bt^\beta} + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} \right\}.$$

For any $\delta \in (0, 1)$, we also have with probability at least $1 - \delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$|L(\theta_t) - L(\theta^*)| \leqslant \left(\frac{C\mathcal{N}_\Theta}{Bt^\beta} + \sqrt{\frac{D_\beta \log(2/\delta)}{t^\beta}} \right) + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(4/\delta)}{\kappa}} \right\}. \tag{B.6}$$

for some constants C and D_β depending on the parameters l, α, γ_1, a_1.

échantillon aléatoire S tiré de la distribution conditionnelle $\mathcal{R}_n$ est appelée *probabilité d'inclusion* du premier ordre et est notée $\pi_i = \mathbb{P}_{R_n}\{i \in S\}$. Nous définissons $\boldsymbol{\pi}_n = (\pi_1,\ \ldots,\ \pi_n)$. Étant donné un échantillon observé S, il est entièrement déterminé par les variables aléatoires $\boldsymbol{\epsilon}_n = (\epsilon_1,\ \ldots,\ \epsilon_n)$, où $\epsilon_i = \mathbb{I}\{i \in S\}$ pour $1 \leq i \leq n$. La plupart des résultats disponibles dans la littérature (voir Boucheron et al. (2005c) par exemple) traitent du cas où l'ensemble de données $\mathcal{D}_n$ est à disposition. Cependant, ce n'est pas le cas ici car nous n'observons qu'un sous-ensemble d'observations issue du schéma de sondage. Par conséquent, ces résultats ne sont pas directement applicables à notre problème, essentiellment à cause de la structure de dépendance induite par le schéma d'échantillonnage. Néanmoins, nous montrons que la théorie de l'ERM peut être étendue au cas où l'apprentissage statistique est basé sur des observations obtenues par sondages. Nous prouvons que, en minimisant une version pondérée du risque empirique, que nous nommons le risque d'Horvitz-Thompson , l'erreur stochastique peut être bornée par $O_{\mathbb{P}}((\kappa_n(\log n)/n)^{1/2})$ (où κ_n est défini plus tard) lorsque les données sont échantillonnés au moyen d'un schéma rejectif. Nous étendons ensuite ces résultats à d'autres schémas d'échantillonnage par un argument de couplage.

Risque de Horvitz-Thompson .

Comme défini dans Horvitz & Thompson (1951), pour tout candidat classificateur g, le risque empirique (non disponible) $\widehat{L}_n(g) = n^{-1} \sum_{1 \leq i \leq n} \mathbb{I}\{Y_i \neq g(X_i)\}$ est remplacée par sa version de Horvitz-Thompson:

$$\overline{L}_{\boldsymbol{\epsilon}_n}(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{\epsilon_i}{\pi_i} \mathbb{I}\left\{g(X_i) \neq Y_i\right\}, \tag{B.7}$$

où $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ indique le vecteur en correspondance avec l'échantillon tiré.

Alors que de nombreux plans d'échantillonnage pourraient etre considérés pour le problème que nous considérons, nous faisons particulièrement attention au schéma rejectif, un plan d'échantillonnage $\mathcal{R}_n$ de taille fixe $N \leq n$, qui généralise le *tirage aléatoire simple sans remise* (où tous les échantillons de cardinalité N sont également susceptibles d'être choisis). Ce plan d'échantillonnage est plus difficile à analyser car les ϵ_i sont des variables aléatoires dépendantes. Par conséquent, lorsque l'apprentissage statistique est basé sur des observations obtenues au moyen d'un schéma rejectif, les résultats classiques de minimisation du risque empirique ne peuvent être appliqués. Néanmoins, nous montrons que des résultats similaires à ceux classiquement établis peuvent être établis pour le minimiseur de (B.7) dans le cas du schéma rejectif. Pour établir ces résulatats, nous montrons et utilisons une propriéré du schéma rejectif: ce schéma forme une collection de variables aléatoires *négativement associées* (voir Brändén & Jonasson (2012), Kramer et al. (2011)), un type de structure de dépendance particulièrement utile pour l'téblissement de bornes de concentration. En utilisant la propriété d'association négative, nous montrons que pour un schéma d'échantillonnage rejeté $\boldsymbol{\epsilon}_n$ avec des probabilités d'inclusion de premier ordre $\boldsymbol{\pi}_n$ et avec $\kappa_n = N/(n \times \min_{i \leq n} \pi_i)$ nous avons pour toute solution $\bar{g}_n$ du problème de minimisation $\inf_{g \in \mathcal{G}} \overline{L}_{\boldsymbol{\epsilon}_n}(g)$, une limite supérieure du risque d'erreur stochastique de l'ordre $O_{\mathbb{P}}((\kappa_n(\log n)/N)^{1/2})$.

Proposition B.2. *Supposons que le schéma d'échantillonage $\boldsymbol{\epsilon}_n$ soit rejectif avec probabilté d'inclusion du premier ordre $\boldsymbol{\pi}_n$ et que la classe de fonctions $\mathcal{G}$ a dimension* VC $V < +\infty$. *Soit $\kappa_n = N/(n \times \min_{i \leq n} \pi_i)$. Alors les affirmations suivantes sont vraies:*

(i) Pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$, nous havons: $\forall N \leq n$,

$$\sup_{g \in \mathcal{G}} \left| \bar{L}_{\boldsymbol{\epsilon}_n}(g) - \widehat{L}_n(g) \right| \leq \sqrt{2\kappa_n \frac{\log(\frac{2}{\delta}) + V \log(n+1)}{N}} + 2\kappa_n \frac{\log(\frac{2}{\delta}) + V \log(n+1)}{3N}. \tag{B.8}$$

(ii) Pour toute solution $\bar{g}_n$ du problème de minimisation $\inf_{g \in \mathcal{G}} \overline{L}_{\boldsymbol{\epsilon}_n}(g)$, nous avons que pour tout $\delta \in (0,1)$, avec probabilité au moins $1-\delta$, nous havons: $\forall n \geq 1$,

$$L(\bar{g}_n) - L^* \leq 2\sqrt{2\kappa_N \frac{\log(\frac{4}{\delta}) + V \log(n+1)}{N}} + 4\kappa_n \frac{\log(\frac{4}{\delta}) + V \log(n+1)}{3N} + C\sqrt{\frac{V}{n}} + 2\sqrt{\frac{2\log(\frac{2}{\delta})}{n}} + \inf_{g \in \mathcal{G}} L(g) - L^*.$$

Le facteur κ_N présent dans les bornes reflète l'influence du schéma d'échantillonnage, (remarquons en particulier que $\kappa_n \geq 1$ puisque $\sum_{i \leq n} \pi_i = N$). Dans le cas du tirage sans remise, *i.e.* quand $\pi_i = N/n$ pour tout $i \in \{1, \ldots, n\}$, le coefficient est alors minimum et est égal à 1. Plus généralement quand $N = o(n)$ avec $n \to +\infty$, dès que les poids ne tendent pas vers 0 plus rapidement que N/n, le taux de convergence atteint par les minimiseurs du risque empirique sont de l'ordre de $O(\sqrt{(\log n)/N})$. Il existe de nombreux schéma d'échantillonnage (*par exemple* Rao-Sampford sampling, Pareto sampling, Srinivasan sampling) de taille fixe

étant décrits par des vecteurs aléatoires $\boldsymbol{\epsilon}_n$ avec des variables aléatoires négativement associés, voir par exemple Brändén & Jonasson (2012) ou Kramer et al. (2011). Ainsi, la preuve de la proposition B.2 permet d'étendre le résulyat à tout schéma négativement associé. Voir la section 2.8 pour plus de détails et de référence. Avant de montrer comment ce résultat peut être étendu à des schéma d'échantillonnage génériques, nous faisons les remarques suivantes:

Remark B.3. (SUR L'HYPOTHÈSE DE COMPLÉXITÉ) Nous faisons remarquer qu'il est possible d'établir le résultat précédent avec des hypothèses plus faibles, en ayant recours au mêmes arguments que ceux développés dans la section 2.6, sous des hypothèses de compléxité différentes, avec des conditions sur l'entropie de la classe de reconstruction $\mathcal{G}$ (voir *par exemple* van der Vaart & Wellner (1996)).

Remark B.4. (MODEL SELECTION) Une légère modification de la Proposition B.2 mène à des bornes sur l'éxcès de risque $\mathbb{E}[L(\bar{g}_{\boldsymbol{\epsilon}_n})] - \inf_{g\in\mathcal{G}} L(g)$. En suivant le principe de *Structural Risk Minimization* (voir Vapnik (2001)), de telles VC bornes peuvent ensuite être utilisés en tant que termes de régularisation pour pénaliser le risque de HT (2.4) et, pour une suite de classes de modèlisation $\mathcal{G}_k$ avec $k \geq 1$ de VC dimension fini, permet d'obtenit de nouvelles bornes quand on choisit le classifieur ayant le minimum de risque $\{\arg\min_{g\in\mathcal{G}_k} \bar{L}_{\boldsymbol{\epsilon}_n}(g),\ k \geq 1\}$ sur la suite de classe.

La propriété de l'association négative étant partagée par de nombreux autres schémas d'échantillonnage, le même argument peut donc être naturellement appliqué pour effectuer une analyse et établir des vitesses de convergence similaire pour les données d'apprentissage produites par de tels plans. Cependant, cette analyse ne peut pas être étendue à tous les schémas d'échantillonnage. Nous contournons cette difficulté en utilisant les résultats établis pour le plan rejectif et en nous appuyant sur un argument de couplage. Considérons un schéma d'échantillonnage abec une structure de dépendance a priori complexe $\mathcal{R}_n^*$ avec des probabilités d'inclusion du premier ordre $\boldsymbol{\pi}_n^* = (\pi_1^*,\ \ldots,\ \pi_n^*)$ représenté par le vecteur $\boldsymbol{\epsilon}_n^* = (\epsilon_1^*,\ \ldots,\ \epsilon_n^*)$ (avec des variables aléatoires non nécessairement négativement associés). Soit $\bar{g}_n^*$ un minimiseur du risque empirique HT $\bar{L}_{\boldsymbol{\epsilon}_n^*}(g) = (1/n)\sum_{i=1}^n (\epsilon_i^*/\pi_i^*)\mathbb{I}\{Y_i \neq g(X_i)\}$ sur une classe $\mathcal{G}$. Puisque nous avons déjà établi des résultats dans le cas du schéma rejectif, nous introduisons un schéma d'échantillonnage rejectif $\mathcal{R}_n$ décrit par les variables aléatoires $\boldsymbol{\epsilon}_n$, avec probabilités d'inclusion du premier ordre $\boldsymbol{\pi}_n = (\pi_1,\ \ldots,\ \pi_n)$ ainsi que la quantité suivante:

$$\check{L}_{\boldsymbol{\epsilon}_n}(g) = \frac{1}{n}\sum_{i=1}^{n} \frac{\epsilon_i}{\pi_i^*}\mathbb{I}\{Y_i \neq g(X_i)\}, \tag{B.9}$$

pour tout classifieur g. Observez que (B.9) est différent du risque empirique de Horvitz Thompson $\bar{L}_{\boldsymbol{\epsilon}_n}(g)$ lié au schéma d'échantillonnage réjectif $\boldsymbol{\epsilon}_n$ car elle fait intervenir les π_i^* au lieu des π_i. L'excès de risque du minimiseur du risque empirique de Horvitz Thompson peut être borné comme suit:

$$\begin{aligned} L(\bar{g}_n^*) - \inf_{g\in\mathcal{G}} L(g) \leq\ & 2\sup_{g\in\mathcal{G}}\left|L(g) - \widehat{L}_n(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\widehat{L}_n(g) - \bar{L}_{\boldsymbol{\epsilon}_n}(g)\right| \\ & + 2\sup_{g\in\mathcal{G}}\left|\bar{L}_{\boldsymbol{\epsilon}_n}(g) - \check{L}_{\boldsymbol{\epsilon}_n}(g)\right| + 2\sup_{g\in\mathcal{G}}\left|\check{L}_{\boldsymbol{\epsilon}_n}(g) - \bar{L}_{\boldsymbol{\epsilon}_n^*}(g)\right|. \end{aligned} \tag{B.10}$$

Nous avons contrôlé le premier terme du côté droit de (B.10) en utilisant les inégalités Vapnik-Chervonenkis et McDiarmid (voir par exemple Vapnik (2001) et le chapitre A), assertion (i) de la proposition B.2 établie dans le cas du schéma rejectif permet d'obtenir un contrôle du second terme. Le troisième terme est borné au moyen d'un argument de *couplage* alors que le dernier terme est contrôlé par des hypothèses liées à la proximité entre les probabilités d'inclusion du premier ordre $\boldsymbol{\pi}_n^*$ et $\boldsymbol{\pi}_n$. Plus précisément, les hypothèses requises dans l'analyse qui suit fait

appel à la distance de variation entre les plans d'échantillonnage $\mathcal{R}_n$ et $\mathcal{R}_n^*$ définie par:

$$d_{TV}(R_n,\ R_n^*) \stackrel{def}{=} \frac{1}{2} \sum_{s \in \mathcal{P}(\mathcal{I}_n)} |\mathcal{R}_n(s) - \mathcal{R}_n^*(s)|.$$

Avec $\kappa_n^* = (N/n) \min_{i \leq n} \pi_i^*$ et $\kappa_n = (N/n \times \min_{i \leq n} \pi_i)$, nous établissons que $L(\bar{g}_n^*) - \inf_{g \in \mathcal{G}} L(g)$ est de l'ordre de $O_{\mathbb{P}}((\kappa_n(\log n)/N)^{1/2}) + 2(\kappa_n^* + \kappa_n)(n/N) \inf_{\mathcal{R}_n} d_{TV}(\mathcal{R}_n,\ \mathcal{R}_n^*)$, où le minimmum est pris sur l'ensemble des plans d'échantillonnage rejectif $\mathcal{R}_n$ ave probabilité d'inclusion du premier ordre $\boldsymbol{\pi}_n = (\pi_1,\ \ldots,\ \pi_n)$.

Theorem B.5. *Supposons que les hypothèses de la proposition B.2 sont satisfaites. Soit $\kappa_N^* = (N/n) \min_{i \leq n} \pi_i^*$ et $\kappa_N = (N/n)/\min_{i \leq n} \pi_i$. Alors il existe une constante $C < +\infty$ tel que, $\forall n \geq 1$,*

$$\mathbb{E}\left[L(\bar{g}_n^*) - \inf_{g \in \mathcal{G}} L(g)\right] \leq 2\sqrt{2\kappa_N \frac{\log(\frac{4}{\delta}) + V \log(n+1)}{N}} + 4\kappa_N \frac{\log(\frac{4}{\delta}) + V \log(n+1)}{3N}$$
$$+ C\sqrt{\frac{V}{n}} + 2\sqrt{\frac{2\log(\frac{2}{\delta})}{n}} + 2(\kappa_N^* + \kappa_N)(n/N) d_{TV}(\mathcal{R}_n,\ \mathcal{R}_n^*), \quad \text{(B.11)}$$

où le minimum est pris sur l'ensemble des schémas d'échantillonnage rejectif $\mathcal{R}_n$ ayant pour probabilité d'inclusion du premier ordre les $\boldsymbol{\pi}_n = (\pi_1,\ \ldots,\ \pi_N)$.

Le taux de convergence obtenu dépend de l'erreur minimale faite lors de l'approximation du plan d'échantillonnage par un plan d'échantillonnage réjectif en termes de distance de variation totale. Il est du même ordre dans le cas où les observations sont échantillonnées uniformément à un terme multiplicatif près et montrent que l'apprentissage avec un échantillon de sondage est possible en tenant compte des probabilités d'inclusion du premier ordre.

B.3 Stratégie d'échantillonnage pour l'algorithme du gradient stochastique

Nous présentons dans cette section un résumé des résultats établis dans les chapitres 3, 4 et 5, dans lequel nous présentons le problème de la stratégie d'échantillonnage non uniforme pour la descente de gradient stochastique (SGD en abrégé). Le problème de minimisation du risque empirique précédemment introduit est de la plus haute importance et la mise en œuvre d'algorithmes efficaces pour résoudre ce problème est une question à laquelle nous avons tenté de répondre. Nous considérons ici un cadre plus général que celui de la classification binaire, et considérons les problèmes d'optimisation de la forme:

$$\min_{\theta\in\Theta} \widehat{L}_n(\theta) = \min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(Z_i,\theta), \tag{B.12}$$

où Θ est un espace euclidien, typiquement $\mathbb{R}^d$ avec $d \geq 1$, et $l(Z_1,.),\ \ldots,\ l(Z_n,.)$ forme une collection de fonctions convexes continuellement dérivables valables sur Θ. En effet, un tel problème d'optimisation se pose typiquement dans une grande variété de problèmes d'apprentissage statistique, en particulier de tâches supervisées, où l'objectif poursuivi est d'apprendre un modèle prédictif, entièrement déterminé par un paramètre θ. La performance de la fonction prédictive définie par θ est typiquement mesurée par l'espérance $L(\theta) = \mathbb{E}[\ell((X,Y),\theta)]$, appelée le *risque*, où ℓ est un *fonction de perte* supposé convexe θ. Il est généralement évalué via son homologue empirique

$$\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell((X_i,Y_i),\theta), \tag{B.13}$$

basé sur $n \geq 1$ exemples d'entraînement indépendants disponibles (X_1,Y_1), ..., (X_n,Y_n), copies de la paire aléatoire $Z = (X,Y)$. Le problème de minimisation (B.12) peut être résolu de façon incrémentale, au moyen de variantes de la méthode *approximation stochastique* initialement introduite dans la contribution séminale de Robbins & Monro (1951). Celui-ci consiste à calculer les estimations successives d'un minimiseur de (B.2) en utilisant l'équation récursive

$$\theta_{t+1} = \theta_t - \gamma_t \widehat{r}_t(\theta_t) \tag{B.14}$$

à partir d'une valeur initiale $\theta_0 \in \Theta$, où $\widehat{r}_t$ dénote un estimateur du gradient $\nabla\widehat{L}_n$ et γ_t est le *taux d'apprentissage* où *step-size*. La mise en œuvre de SGD est assez simple pour la minimisation d'un risque empirique prenant la forme d'une moyennes standard, car elle est généralement effectuée en échantillonnant uniformément au hasard (avec ou sans remplacement) un sous-échantillon d'observations avant de calculer un estimateur du gradient. Contrairement à l'approche *batch*, où toutes les données sont utilisées pour estimer le gradient à chaque itération (*ie* $\widehat{r}_t(\theta) = \nabla\widehat{L}_n(\theta)$ pour tout $t \geq 0$ et $\theta \in \Theta$), seuls les sous-ensembles de l'échantillon de données sont impliqués dans les étapes d'estimation de gradient de dans ces *algorithmes incrémentaux*, dans le but de réduire le coût de calcul lorsque n est grand. Dans l'implémentation la plus couramment utilisée de l'algorithme de la *descente de gradient stochastique* (SGD), l'estimateur de gradient est calculé à partir d'un sous-ensemble de taille réduite $S \leq n$ uniformément dessiné sans remplacement parmi tous les sous-ensembles possibles de taille S à chaque étape $t \geq 0$. En pratique, la limitation principale de cette technique d'optimisation incrémentale est due au bruit stochastique induit par le choix aléatoire des données impliquées dans le calcul de l'estimateur de gradient à chaque itération. En particulier, la plupart des justifications théoriques du SGD sont établies dans un cadre très général (voir Robbins & Monro (1951) ou Bach & Moulines (2011a) par exemple) qui englobe le cas de la minimisation du

risque empirique. Nous proposons d'introduire une stratégie d'échantillonnage non uniforme ainsi qu'une nouvelle analyse soulignant l'avantage d'utiliser une stratégie d'échantillonnage non uniforme pour le problème de la minimisation du risque empirique. Nous introduisons d'abord dans le chapitre 3 une nouvelle implémentation de l'algorithme SGD, où le sous-ensemble de données utilisé à un pas donné n'est pas choisi au hasard parmi tous les sous-ensembles possibles mais simulé à l'aide d'un schéma d'échantillonnage adaptatif spécifique, construit à partir des itérations passées. Nous proposons ensuite un cadre général pour étendre ces résultats en utilisant la théorie des sondages dans le chapitre 4 dans lequel nous prenons également en compte la distribution de l'observation dans notre analyse finale. Nous concluons cette section en considérant le cas spécifique où le risque empirique prend la forme d'une U-statistique et proposons une implémentation efficace de l'algorithme SGD dans ce cas. Ici dans le reste de cette sections, les opérateurs du gradient et de la hessienne par rapport à une variable θ sont notés ∇ et ∇^2 respectivement. Par convention, ∇^0 correspond à l'opérateur d'identité et les valeurs de gradient sont représentées comme des vecteurs de colonne. Pour tout vecteur $V \in \mathbb{R}^d$, on note $||V||$ sa norme euclidienne et pour toute matrice A on note A^T sa transposée.

B.3.1 Un stratégie d'échantillonnage non uniforme pour le SGD

Afin d'accélérer le processus d'apprentissage, nous introduisons une variante spécifique de l'algorithme SGD avec un schéma d'échantillonnage *adaptatif*, en ce sens qu'il peut varier avec t, en fonction des itérations passées. Nous considérons un échantillonnage non uniforme avec remplacement. Nous commençons par identifier une bonne distribution d'échantillonnage en choisissant celle qui minimise la variance de l'estimateur. Lorsque l'on tire un échantillon $\mathcal{S}$ de taille S avec probébilités d'inclusion du premier ordre p_i indépendamment avec remplacement, la quantité

$$\frac{1}{S}\sum_{i\in\mathcal{S}}\frac{\nabla l(Z_i,\theta)}{p_i} \tag{B.15}$$

est un estimateur sans biais de $\nabla\widehat{L}_n(\theta)$ avec une variance égale à:

$$\frac{1}{S}\sum_{i=1}^{n}\frac{\|\nabla l(Z_i\theta)\|^2}{p_i} - \frac{||\nabla\widehat{L}_n(\theta)||^2}{S}. \tag{B.16}$$

Pour obtenir la meilleure estimation du gradient (*i.e* en minimisant la variance) au paramètre θ conditionnellement aux observations, il est donc naturel d'échantillonner l'observation Z_i avec probabilité: $p_i^*(\theta) = \|\nabla l(Z_i,\theta)\|/\sum_{j=1}^n\|\nabla l(Z_j,\theta)\|$ car ce choix minimise la valeur de la variance. Malheureusement, la mise en oeuvre pratique du schéma d'échantillonnage ci-dessus n'est pas pertinente car elle nécessite d'évaluer tous les gradients pour calculer les normes $\|\nabla l(Z_1,\theta_t)\|,\dots,\|\nabla l(Z_n,\theta_t)\|$ à chaque itération, ce que nous essayons précisément d'éviter en utilisant l'algorithme du gradient stochastique. Nous proposons donc un schéma d'échantillonnage approximant $\mathbf{p}_t^* := (p_i^*(\theta_t))_{i=1}^n$ sans nécessiter d'évaluation de gradient supplémentaire. Nous utilisons quelques anciennes valeurs du gradient dans notre approximation. Plus précisément, l'idée principale est de remplacer chaque norme de gradient inconnue $\|\nabla l(Z_i,\theta_t)\|$ par une norme précédemment calculée $g_{t,i} = \|\nabla l(Z_i,\theta_k)\|$ à un ancien instant $k = k(i,t)$ correspondant au dernier instant $k \le t$ où Z_i a été tiré. Plus formellement, nous définissons la suite aléatoire g_t comme étant égale à:

$$g_{t+1,i} = \begin{cases} \|\nabla l(Z_i,\theta_{t+1})\| & \text{if } i \in \{i_{t+1}^{(1)},\dots,i_{t+1}^{(S)}\} \\ g_{t,i} & \text{otherwise.} \end{cases} \tag{B.17}$$

Alors, une façon naturelle d'approximer $\mathbf{p}_t^*$ est de construire $\bar{\mathbf{p}}_t = (\bar{p}_{t,i})_{i=1}^n$ où nous définissons pour chaque i

$$\bar{p}_{t,i} = \frac{g_{t,i}}{\sum_{j=1}^n g_{t,j}} \,. \tag{B.18}$$

Il s'avère que la convergence ne peut pas être garantie avec ce choix, car un certain composant $\bar{p}_{t,i}$ peut être arbitrairement proche de zéro. Un remède possible consiste à appliquer un plan d'échantillonnage greedy:

$$\forall i \in \{1, \ldots, n\}, \;\; p_{t,i} = \rho\nu_i + (1-\rho)\bar{p}_{t,i}, \tag{B.19}$$

où $\nu = (\nu_1, \ldots, \nu_n)$ est une distribution de probabilité avec $\nu_i > 0$ pour $1 \leq i \leq n$, et $0 < \rho \leq 1$. Cette condition a l'interprétation suivante: p_t est un mélange entre deux lois de probabilité et une de cette loi (ν) est indépendante du passé. Maintenant que nous avons défini notre stratégie d'échantillonnage, l'algorithme que nous proposons est simplement de calculer à chaque itération t un estimateur du gradient basé sur l'équation (1.11) par l'observation d'échantillonnage selon $\mathbf{p}_t := (p_{t,i})_{i=1}^n$. Cette stratégie d'échantillonnage peut également être mise en œuvre efficacement dans la pratique et nous montrons que l'échantillonnage dans le cadre de cette stratégie n'a qu'un coût additionnel de $O(\log(n))$. Les résultats théoriques sont ensuite établis au moyen d'un argument asymptotique où nous montrons qu'avec cette stratégie d'échantillonnage, le comportement asymptotique de θ_t est optimal jusqu'à une erreur proportionnelle à ρ. Sous les hypothèses suivantes nous montrons d'abord la convergence de l'agorithme proposé:

Assumption 9. Pour tout $i \in \{1, \ldots, n\}$, la fonction $\theta \to l(Z_i, \theta)$ est convexe, dérivable et son gradient $\nabla l(Z_i, \theta)$ est L_i-Lipschitz continue avec $L_i < +\infty$.

Assumption 10. i) La fonction $\theta \to \widehat{L}_n(\theta)$ est α fortement convexe, ii) Le minimiseur θ_n^* de $\widehat{L}_n$ est dans l'intérieur de $\mathcal{K}$.

Le lemme ci-dessous permet de contrôler l'erreur quadratique moyenne $a_t = \mathbb{E}(\|\theta_t - \theta_n^*\|^2)$, où θ_t est généré par l'algorithme 1. La preuve de ce résultat est fortement inspirée de Bach & Moulines (2011a) et A.Nemirovski et al. (2009), où des bornes similaires sont établies.

Lemma B.6. *Sous les hypothèses 9 et 10, soit $\gamma_t = \gamma_1 t^{-\beta}$ où $\beta \in (0, 1]$ et supposons que $\gamma_1 > \beta/(2\alpha)$. Pour tout $t \in \mathbb{N}^*$,*

$$a_t \leq C\gamma_t/\rho, \tag{B.20}$$

où $C = \max(\frac{2B_\nu^2\gamma_1}{2\alpha\gamma_1 - 1}, \frac{a_1}{\gamma_1})$ quand $\beta = 1$ et $C = \max(\frac{B_\nu^2\gamma_1}{2\alpha}, \frac{a_1}{\gamma_1})$ sinon, avec

$$B_\nu = \frac{1}{Sn^2} \sum_{i=1}^n \nu_i^{-1} \sup_{\theta \in \mathcal{K}} \|\nabla l(Z_i, \theta)\|^2 \,.$$

Ce résultat bien que classique quand l'on cherche à prouver la convergence d'un algorithme du gradient stochastique ne met pas en évidence l'avantage que l'on aurait d'utiliser l'algorithme que nous proposons. Nous contournons cette difficulté en établissant le premier lemme suivant, caractèrisant l'écart moyen entre les gradients utlisés dans le schéma d'approximation et les vrais gradients:

Lemma B.7. *Sous les hypothèses du lemme 3.1, Pour tout $t \in \mathbb{N}^*$,*

$$b_{t,i} \leqslant \frac{(2L_i)^2 2^\beta}{1 - (1 - \rho\nu_i)^S} \frac{C}{\rho t^\beta} + o(t^{-\beta}) \,. \tag{B.21}$$

Ce lemme est le premier résultat intermédiaire requis pour montrer le meilleur comportement asymptotique de l'estimateur que nous proposons, et caractérisé par le résultat suivant, établie sous l'hypothèse de la double différenciabilité au voisinage de θ_n^* de la fonction θ_n^*. Nous introduisons la distribution de probabilité $\pi^* = \rho\nu + (1-\rho)\bar{\pi}^*$, où

$$\bar{\pi}_i^* = \frac{\|\nabla l(Z_i, \theta_n^*)\|}{\sum_{j=1}^n \|\nabla l(Z_j, \theta_n^*)\|} \tag{B.22}$$

porut tou $i = 1, \ldots, n$. Nous définisssons $Q^* = \sum_{i=1}^n \nabla l(Z_i, \theta_n^*) \nabla l(Z_i, \theta_n^*)^T / (Sn^2 \pi_i^*)$ et notons $H = \nabla^2 \widehat{L}_n(\theta_n^*)$ la hessienne au point θ^*.

Theorem B.8. *Sous les hypothèses 9, 10 and 3, et avec un stepsize satisfaisant les conditions énoncés dans le Lemme 3.1. Alors la suite $(\theta_t - \theta_n^*)/\sqrt{\gamma_t}$ converge eb distribution vers variable aléatoire gaussience centrée avec matrice de covariance $\Sigma = \Sigma(\rho, \nu)$ est solution de l'équation de Lyapunov suivante:*

$$\begin{aligned} &\Sigma H + H\Sigma = Q^* \quad (\textit{if } \beta < 1) \\ &\Sigma(I_d + 2\gamma_1 H) + (I_d + 2\gamma_1 H)\Sigma = 2\gamma_1 Q^* \quad (\textit{if } \beta = 1)\,. \end{aligned}$$

Le corollaire suivant est directement obtenu par la méthode delta du second ordre Pelletier (1998). Nous notons $\mathrm{Tr}(A)$ la trace d'une matrice carrée A.

Corollary B.9. *Sous les hypothèses que théorème B.8, $\gamma_t^{-1}(\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta^*))$ converge en distribution vers la variable aléatoire $V = (1/2) Z^T \Sigma(\rho, \nu)^{1/2} H \Sigma(\rho, \nu)^{1/2} Z$ où Z es un vecteur gaussien $\mathcal{N}(0, I_d)$. De plus, nous havons $\mathbb{E}(V) = \mathrm{tr}(H\Sigma(\rho, \nu))/2$.*

Nous utilisons maintenat le Corollaire 3.5 afin de comparer notre algorithme avec et son choix de schéma d'échantillonage avec le meilleur schéma d'échantillonnage possible fixé. Notez aussi que la recherche d'un meilleur d'un schéma optimal est aussi discuté dans Clemencon et al. (2014).

Quand la distribution est fixée égale à p, la covariance asymptotique de l'erreur est donnée par by $\Sigma(1, p)$ tel qu'il est définie dans Theorem 3.4. Motivé par le Corollaire 3.5, nous faisons référence à la meilleur distribution de probabilité comme étant le minimiseur p de $\mathrm{tr}(H\Sigma(1, p))$. Il est facile de montrer que

$$\arg\min_p \mathrm{tr}(H\Sigma(1, p)) = \bar{\pi}^*,$$

où $\bar{\pi}^*$ est définie dans (B.22). Nous définissons aussi $\sigma_*^2 = \mathrm{tr}(H\Sigma(1, \bar{\pi}^*))$. La proposition suivante est facilement établie à l'aide de résultats d'algèbre linéaire classique.

Proposition 2. Soit $\Sigma(\rho, \nu)$ la matrice de covariance asymptotique définie dans le Théorème B.8. alors,

$$\sigma_*^2 \leq \mathrm{tr}(H\Sigma(\rho, \nu)) \leq \sigma_*^2(1 + S\rho/(1-\rho)).$$

La Proposition 2 établit que la performance de l'algorithme que nous proposons peut être rendu aussi proche que possible de la performance associée à l'algorithme employant les probabilités optimales, pour peu que ρ soit proche de zero. Il est bien entendu tentant de choisir $\rho = 0$ in (3.8), cependant dans ce cas le Théorème 3.4 n'est plus vrai.

Notez que tout les résultats obtenus dans le chapitre 3 restent vrais conditionnellement aux observations et ne tient donc pas compte de la nature statistique de notre problème (nous résolvons ERM car nous ne connaissons pas le vrai risque). Nous abordons ce problème dans la section suivante où nous discutons d'un problème similaire (Stratégie d'échantillonnage non uniforme pour SGD) dans le contexte de la M-estimation. Plus précisément, nous utilisons le cadre des sondages introduit précédemment pour étendre nos résultats.

B.3.2 (HTSGD) et applications à la M-estimation

Les sections précédentes suggèrent fortement d'utiliser les techniques de sondage pour améliorer l'algorithme d'apprentissage. Nous montrons maintenant comment incorporer des schémas de sondage efficaces dans de telles techniques itératives pour la M-estimation. Plus précisément, nous proposons un estimateur spécifique du gradient, que l'on appelle l'estimateur *estimation du gradient de Horvitz-Thompson* (HTGD en bref). Pour l'estimateur ainsi produit, les résultats de normalité asymptotique décrivant sa performance statistique sont établis. Le framework que nous considérons est le même que celui de la section 1.3.1. Nous définissons l' *estimateur de Horvitz-Thompson* du gradient $\ell_n(\theta)$ basé sur un échantillon de sondage S tiré par un design $\mathcal{S}_n$ avec probabilités d'inclusion du premier ordre $\{\pi_i\}_{1\leq i\leq n}$ et le vecteur d'inclusion $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)$ as

$$\ell_{\mathcal{S}_n}(\theta) = \frac{1}{n}\sum_{i\in S}\frac{1}{\pi_i}\nabla l(Z_i,\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\epsilon_i}{\pi_i}\nabla l(Z_i,\theta). \tag{B.23}$$

Avec ces notations, nous étudions la convergence d'un algorithme SGD lorsque l'estimateur du gradient est calculé en échantillonnant des observations dans un ensemble de données sous un plan d'échantillonnage $\mathcal{S}_n$ (éventuellement en fonction de déoendant de t et la valeur actuelle du paramètre). On note $\theta_n(T)$ la valeur du paramètre à l'instant T. Conditionné sur les données $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$, nous étudions les propriétés asymptotiques de l'estimateur M produit par l'algorithme HTGD. Les résultats limites indiqués ci-dessous reposent essentiellement sur le fait que l'estimateur HT (B.23) du gradient du risque empirique est non biaisé et sous certaines hypothèses listées dans la section précédente.

Theorem B.10. (CONDITIONAL CENTRAL LIMIT THEOREM) *Supposons que $\gamma_t = \gamma_0\, t^{-\alpha}$ avec $\alpha \in (1/2, 1]$ et $\gamma_0 > 0$. Quand $\alpha = 1$, $\gamma_0 > 1/(2l)$ et $\eta := 1/(2\gamma_0)$; $\eta := 0$ sinon. Conditionnellement à $\mathcal{D}_n$ nous avons la convergence en distribution quand $t \to +\infty$*

$$\sqrt{1/\gamma_t}\,(\theta_n(t) - \theta_n^*) \Rightarrow \mathcal{N}(0, \Sigma_{\boldsymbol{\pi}_n}),$$

où la matrice de covariance asymptotique $\Sigma_{\boldsymbol{\pi}_n}$ est l'unique solution de

$$H_n\Sigma + \Sigma H_n + 2\eta\Sigma = \Gamma_n^*, \tag{B.24}$$

avec $\Gamma_n^* = \Gamma_n(\theta_n^*)$ and, for all $\theta \in \Theta$,

$$\Gamma_n(\theta) = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(\frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\pi_j(\theta)} - 1\right)\nabla l(Z_i,\theta)\,\nabla l(Z_j,\theta)^T. \tag{B.25}$$

La réduction de la variance asymptotique de $\widehat{\theta}_n(T)$ (de $\widehat{L}_n(\widehat{\theta}_n(T))$, respectivement) est étudiée plus tard dans le cas de Poisson (c'est-à-dire quand les ϵ_i sont indépendants). Par une application directe de la méthode Delta du second ordre, nous caractérisons alors le comportement de $\widehat{L}_n(\widehat{\theta}_n(t)) - \widehat{L}_n(\theta_n^*)$ où θ_n^* est un minimiseur empirique des risques.

Corollary B.11. (ERROR RATE) *Sous les hypothèses du theorème B.10, conditionnellement à $\mathcal{D}_n$ quand $t \to +\infty$ nous avons la convergence en distribution :*

$$1/\gamma_t \left(\widehat{L}_n(\theta_n(t)) - \widehat{L}_n(\theta_n^*) \right) \Rightarrow \frac{1}{2} U^T \Sigma_{\boldsymbol{\pi}_n}^{1/2} H_n \Sigma_{\boldsymbol{\pi}_n}^{1/2} U,$$

où U est une q-dimensionnelle variable aléatoire gaussienne centrée réduite.

Nous discutons ensuite comment choisir π_i dans le cas où les variables aléatoires sont de type Poisson (cas où les ϵ_i sont indépendants) et retrouvons le résultat du chapitre 3 en montrant que l'échantillonnage avec π_i proportionnel à $\|\nabla l(Z_i, \theta)\|$ donne des résultats optimaux en essayant de minimiser la variance de $\widehat{L}_n(\widehat{\theta}_n(t)) - \widehat{L}_n(\theta_n^*)$.

Proposition B.12. (OPTIMALITY) *Sous les hypothèses du théorème 5 quand $\eta = 0$ avec*

$$N \leq \inf_{\theta \in \Theta} \frac{\sum_{i=1}^n \|G_n \nabla l(Z_i, \theta)\|}{\max_{1 \leq i \leq n} \|G_n \nabla l(Z_i, \theta)\|}, \tag{B.26}$$

et $G_n := H_n^{-1/2}$. Alors le sondage de Poisson avec probabilité d'inclusion $\{\mathbf{p}_n^*(\theta)\}_{\theta \in \Theta}$ defined for all $\theta \in \Theta$ and $i \in \mathcal{U}_n$ by

$$p_i^*(\theta) = N \frac{\|G_n \nabla l(Z_i, \theta)\|}{\sum_{j=1}^n \|G_n \nabla l(Z_j, \theta)\|}$$

est solution du problème de minimisation

$$\min_{\mathbf{p}_n = \{\mathbf{p}_n(\theta)\}_{\theta \in \Theta}} \left\| \Sigma_{\mathbf{p}_n}^{1/2} \right\| \text{ subject to } \sum_{i=1}^n p_i(\theta) = N.$$

De plus, nous havons

$$\left\| \Sigma_{\mathbf{p}_n^*}^{1/2} \right\|^2 = \frac{1}{2} \left\{ \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n \|G_n \nabla l(Z_i, \theta_n^*)\| \right)^2 - \frac{1}{n^2} \sum_{i=1}^n \|G_n \nabla l(Z_i, \theta_n^*)\|^2 \right\}.$$

Nous établissons ensuite des résultats similaires dans le cas général où le sondage n'est pas de Poisson. Le chapitre est ensuite conclu par une analyse non conditionnelle par rapport aux données. En dénotant par N la taille moyenne d'un échantillon obtenue par un sondage, notre analyse est finalement complétée en étudiant le comportement de $\widehat{\theta}_n(t)$ quand n, N tend vers $+\infty$. Ceci est très différent de ce que nous avons fait dans le chapitre 3 parce que notre analyse s'était limitée au cas où n est fixé à l'avance nos résultats étaient obtenus conditionnellement aux observations. Cela nous permet d'illustrer le compromis bien connu entre la généralisation (asymptotique) et les erreurs d'optimisation, régies par le comportement limite de $n\gamma_t/N$ (voir Bottou & Bousquet (2008) par exemple). Nous montrons que sous des hypothèses supplémentaires, si $\lim n\gamma_t/N = c > 0$, alors nous avons la convergence dans la distribution:

$$\lim_{n,\, N \to \infty} \left\{ \lim_{t \to \infty} \sqrt{n} \left(\widehat{\theta}_n(t) - \theta^* \right) \right\} = \mathcal{N}(0, \Lambda^* + c\Sigma^*).$$

où $\lim_{n,\, N \to \infty} N\Gamma_n^* = \Gamma^*$ et Λ^* est la matrice de covariance asymptotique impliquée dans le TCL pour la M estimation appliquée à $\theta_n^* - \theta^*$. L'énoncé complet du théorème requiert l'hypothèses suivante:

Assumption 11. Quand n et N tendent vers ∞, $N\Gamma_n^*$ converge en probabilité vers une matrice matrix Γ^* semi-définie positive.

Bien que cette condition ait l'air forte au premier abord, elle est en général assez souvent satisfaite. En particulier elle est vrai dans le cas poissonien sous de faibles hypothèses.

Theorem B.13. *Sous les hypothèses 4, 5, 7 et si γ_t satisfait les conditions du Théorème B.10 avec $\alpha < 1$ (et donc $\eta = 0$). Si $H^* = \mathbb{E}[\nabla^2 l(Z, \theta^*)]$ soit*

$$\Lambda^* = (H^*)^{-1}\mathbb{E}[\nabla l(Z, \theta^*)\nabla l(Z, \theta^*)^T](H^*)^{-1}$$

et Σ^* la solution unique de: $H^*\Sigma + \Sigma H^* = \Gamma^*$. alors:

(i) If $\lim_{n,N,t\to+\infty} \frac{n}{N}\gamma_t = +\infty$, alors nous avons la convergence en distribution:

$$\lim_{n,N\to\infty} \left\{ \lim_{t\to\infty} \sqrt{N/\gamma_t}\,(\theta_n(t) - \theta^*) \right\} = \mathcal{N}(0, \Sigma^*).$$

(ii) If $\lim_{n,N,t\to+\infty} \frac{n}{N}\gamma_t = 0$, alors nous avons la convergence en distribution:

$$\lim_{n,N,t\to+\infty} \sqrt{n}\,(\theta_n(t) - \theta^*) = \mathcal{N}(0, \Lambda^*).$$

(iii) If $\lim_{n,N,t\to+\infty} \frac{n}{N}\gamma_t = c > 0$, alors nous avons la convergence en distribution:

$$\lim_{n,N\to\infty} \left\{ \lim_{t\to\infty} \sqrt{n}\,(\theta_n(t) - \theta^*) \right\} = \mathcal{N}(0, \Lambda^* + c\Sigma^*).$$

Avec γ_t typiquement de l'ordre $O(1/t)$, la condition $\lim n\gamma_t/N = c > 0$ donne une idée de la façon dont le nombre d'itérations doit être réglé en fonction du nombre d'observations et la taille du dataset à disposition pour obtenir des résultats optimaux.

B.3.3 SGD pour la minimisation de U-Statistic

Nous discutons ici de l'implémentation de SGD dans le cas où le risque empirique prend la forme d'une *U-Statistique*. Nous présentons brièvement le problème et les notations et expliquons la différence avec les problèmes de la section B.3.1 et B.3.2.

Les U-statistiques sont des extensions des moyennes. En apprentissage automatique, elles sont utilisées comme critères de performance dans de nombreux problèmes, Metric Learning et AUC en particulier sont deux exemples que nous considérons dans nos expériences. Elles sont définies comme suit:

Definition B.14. Soit $K \geq 1$ et $(d_1, \ldots, d_K) \in \mathbb{N}^{*K}$. Soit $\mathbf{X}_{\{1,\ldots,n_k\}} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$, $1 \leq k \leq K$, soit K échantillons indépendants de tailles $n_k \geq d_k$ et composés de variables aléatoires i.i.d prenant leurs valeurs dans un espace mesurable $\mathcal{X}_k$ avec distribution $F_k(dx)$ respectivement. Soit $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$ une fonction mesurable, de carré intégrable par rapport à $\mu = F_1^{\otimes d_1} \otimes \cdots \otimes F_K^{\otimes d_K}$. Supposons en plus (en toute généralité) que $H(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)})$ soit symétrique en ses arguments $\mathbf{x}^{(k)}$ (à valeurs dans $\mathcal{X}_k^{d_k}$), $1 \leq k \leq K$. La K U-statistique généralisée de degrés $(d_1, \ldots, d_K)$ avec noyau H, est alors défini comme suit:

$$U_{\mathbf{n}}(H) = \frac{1}{\prod_{k=1}^K \binom{n_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} H\left(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)}\right), \tag{B.27}$$

où $\mathbf{n} = (n_1, \ldots, n_K)$, le symbole $\sum_{I_1} \cdots \sum_{I_K}$ fait référence à la sommation sur tous les éléments de Λ, l'ensemble des vecteurs d'index $\prod_{k=1}^{K} \binom{n_k}{d_k}$ $(I_1, \ldots, I_K)$, I_k étant un ensemble de d_k index $1 \leq i_1 < \ldots < i_{d_k} \leq n_k$ et $\mathbf{X}_{I_k}^{(k)} = (X_{i_1}^{(k)}, \ldots, X_{i_{d_k}}^{(k)})$ pour $1 \leq k \leq K$.

La sous-section B.3.1 et B.3.2 préconise l'utilisation de SGD pour gérer le nombre de termes polynomial dans (1.8). Notons que lorsque le risque empirique prend la forme d'une statistique U généralisée, le nombre de termes impliqués dans la somme est de l'ordre $O(n^{d_1+\ldots d_K})$ rendant le problème extrêmement difficile à résoudre. Néanmoins, nous montrons comment implémenter le SGD dans ce cas.
Nous nous plaçons dans le cadre paramétré. Notant encore Θ l'espace des paramètres, avec $H : \prod_{k=1}^{K} \mathcal{X}_k^{d_k} \times \Theta \to \mathbb{R}$ une fonction de perte convexe , nous notons la version empirique du risque par $\theta \in \Theta \mapsto \widehat{L}_n(\theta) = U_n(H(.;\, \theta))$. Comme nous l'avons mentionné précédemment, la mise en œuvre de SGD est assez simple pour la minimisation des statistiques standard, car elle est généralement réalisée en échantillonnant uniformément au hasard (avec ou sans remplacement) un sous-ensemble d'observations avant de calculer un estimateur du gradient. Lorsque le risque empirique prend la forme d'une U statistique, l'algorithme SGD pourrait être implémenté de cette façon. Cela conduirait à un estimateur du gradient égal à:

$$\tilde{g}_{n'}(\theta) = \frac{1}{\prod_{k=1}^{K} \binom{n'_k}{d_k}} \sum_{I_1} \ldots \sum_{I_K} \nabla H(\mathbf{X}_{I_1}^{\prime(1)}; \mathbf{X}_{I_2}^{\prime(2)}; \ldots; \mathbf{X}_{I_K}^{\prime(K)};\, \theta), \tag{B.28}$$

où $\sum_{I_k}$ fait référence à la somme des $\binom{n'_k}{d_k}$ sous-ensembles $\mathbf{X}_{I_k}^{\prime(k)} = (X_{i_1}^{\prime(k)}, \ldots, X_{i_{d_k}}^{\prime(k)})$ liés à un ensemble I_k de d_k indexes $1 \leq i_1 < \ldots < i_{d_k} \leq n'_k$ et $\mathbf{n}' = (n'_1, \ldots, n'_K)$. Dans le cas de U-Statistiques,, nous montrons que cette stratégie (que nous appellerons plus tard "*complète* U -statistique") n'est pas efficace. Nous proposons plutôt de procéder différemment en tirant indépendamment avec remise parmi l'ensemble des vecteurs d'index Λ, obtenant un estimateur du gradient sous la forme d'une *incomplète* U -statistique (voir Lee (1990a)):

$$\bar{g}_B(\theta) = \frac{1}{B} \sum_{(I_1, \ldots, I_K) \in \mathcal{D}_B} \nabla H(\mathbf{X}_{I_1}^{(1)}, \ldots, \mathbf{X}_{I_K}^{(K)};\, \theta), \tag{B.29}$$

où $\mathcal{D}_B$ est construit en échantillonnant B fois avec remplacement dans l'ensemble Λ. Le paramètre B est le nombre de termes moyennés. Pour le même coût de calcul (*ie*, prendre $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$) et implémenter SGD avec (B.29) plutôt que (B.28) donne de "meilleurs" solutions, essentiellement parce que (B.29) est un estimteur avec une plus petite variance (sauf dans le cas où $K = 1 = d_1$). Intuitivement, l'échantillonnage d'une *incomplète* U-statistique est préférable car le nombre d'observations présent dans cet estimateur est supérieur au nombre d'observations présent dans l'estimateur *complet*.

Ceci est mis en évidence lorsque l'on compare conditionnellement aux observations les performances des méthodes SGD décrites ci-dessus en étudiant à la fois le comportement asymptotique et non asymptotique de l'algorithme SGD pour les deux implémentations. Comme nous l'avons déjà fait dans les chapitres précédents, nous proposons des bornes de généralisation similaire à (voir Bottou & Bousquet (2007)), nous décomposons l'erreur stochastique comme suit:

$$\mathbb{E}[L(\theta_t) - L(\theta^*)] \leq \underbrace{2\mathbb{E}\left[\sup_{\theta \in \Theta} |\widehat{L}_n(\theta) - L(\theta)|\right]}_{\mathcal{E}_1} + \underbrace{\mathbb{E}\left[\widehat{L}_n(\theta_t) - \widehat{L}_n(\theta_n^*)\right]}_{\mathcal{E}_2}. \tag{B.30}$$

où $\theta^* = \operatorname{argmin}_{\theta\in\Theta} L(\theta)$ et montrons le téorème suivant:

Theorem B.15. *Soi θ_t la séquence obtenue par l'algorithme du gradient stochastique avec la U-statistique incomplète, (5.6) avec $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$ termes avec des $n'_1, \ldots, n'_K$. Suppose que $\{L(.;\ \theta) :\ \theta \in \Theta\}$ est une* VC *classe avec dimension* VC *dimension V tel que*

$$\mathcal{M}_\Theta = \sup_{\theta\in\Theta,\ (\mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(K)})\in\prod_{k=1}^{K}\mathcal{X}_k^{d_k}} |H(\mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(K)};\ \theta)| < +\infty, \tag{B.31}$$

et $\mathcal{N}_\Theta = \sup_{\theta\in\Theta} \sigma_\theta^2 < +\infty$. Si le stepsize satisait les conditions de 5.4, on a:

$$\forall \mathbf{n} \in \mathbb{N}^{*K}, \quad \mathbb{E}[|L(\theta_t) - L(\theta^*)|] \leqslant \frac{C\mathcal{N}_\Theta}{Bt^\beta} + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} \right\}.$$

Pour tout $\delta \in (0,1)$, on a aussi avec probabilité au moins $1-\delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$|L(\theta_t) - L(\theta^*)| \leqslant \left(\frac{C\mathcal{N}_\Theta}{Bt^\beta} + \sqrt{\frac{D_\beta \log(2/\delta)}{t^\beta}} \right) + 2\mathcal{M}_\Theta \left\{ 2\sqrt{\frac{2V\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(4/\delta)}{\kappa}} \right\}. \tag{B.32}$$

pour des constantes C et D_β dépendant des paramètres l, α, γ_1, a_1.

La limite de généralisation montre l'avantage d'utiliser la méthode incomplète pour obtenir un estimateur de gradient tout en mettant en lumière le fait bien connu que lorsque nous utilisons une méthode d'optimisation pour résoudre le problème de minimisation du risque empirique, nous devons prendre en compte les bornes de généralisation pour que l'erreur d'optimisation soit du même ordre (voir la sous-section B.3.2).

B.4 Vitesse rapide pour la reconstruction de graphe

Cette section est un résumé du chapitre 6 dans lequel nous présentons un bref aperçu du problème de *reconstruction de graphes*. Nous introduisons dans un premier temps le contexte avant de décrire le problème d'intérêt. Soit $G = (V,\ E)$ un graphe aléatoire non orienté avec un ensemble $V = \{1,\ \ldots,\ n\}$ de $n \geq 2$ sommets et un ensemble $E = \{e_{i,j} :\ 1 \leq i \neq j \leq n\} \in \{0,1\}^{n(n-1)}$ décrivant ses arêtes: pour tout $i \neq j$, on a $e_{i,j} = e_{j,i} = +1$ si les sommets i et j sont reliés par une arête et $e_{i,j} = e_{j,i} = 0$ sinon. Nous supposons également que pour tout $i \in V$, une variable aléatoire continue X_i est associé au sommet i. Les X_i sont i.i.d. et pour tout $i \neq j$, la paire aléatoire (X_i, X_j) représente un ensemble d'informations utiles pour prédire l'occurrence d'une arête reliant les sommets i et j. Conditionnellement à $(X_1,\ \ldots,\ X_n)$, les variables aléatoires$e_{i,j}$ et $e_{k,l}$ ne sont indépendantes que si $\{i,j\} \cap \{k,l\} = \emptyset$. En particulier, la distribution conditionnelle de $e_{i,j}$, $i \neq j$, est supposée dépendre uniquement de (X_i, X_j) et est décrite par:

$$\eta\left(X_i, X_j\right) = \mathbb{P}\left\{e_{i,j} = +1 \mid (X_i, X_j)\right\}. \tag{B.33}$$

Le problème d'apprentissage introduit par Biau & Bleakley (2006), appelé *graph reconstruction*, consiste à construire une règle symétrique de *reconstruction* $g : \mathcal{X}^2 \to \{0,1\}$, à partir d'un graphe d'apprentissage G, avec un risque de reconstruction le plus petit possible:

$$\mathcal{R}(g) = \mathbb{P}\left\{g(\mathbf{X}_1, \mathbf{X}_2) \neq \mathbf{e}_{1,2}\right\}, \tag{B.34}$$

obtenant ainsi dans le meilleur des cas une performance comparable à celle de la règle de Bayes $g^*(x_1, x_2) = \mathbb{I}\{\eta(x_1, x_2) > 1/2\}$, dont le risque est donné par $\mathcal{R}^* = \mathbb{E}[\min\{\eta(\mathbf{X}_1, \mathbf{X}_2),\ 1 - \eta(\mathbf{X}_1, \mathbf{X}_2)\}] = \inf_g \mathcal{R}(g)$.

Le risque de reconstruction (B.34) n'étant pas disponible, il est remplacé par sa version empirique basée sur l'ensemble labélisé $\mathbb{D}_n = \{(X_i, X_j, e_{i,j}) :\ 1 \leq i < j \leq n\}$ associé à G:

$$\widehat{\mathcal{R}}_n(g) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}. \tag{B.35}$$

Notez que (B.35) présente a prioir une structure de dépendance complexe car ce n'est pas une somme de variables aléatoires indépendantes. Soit $\widehat{g}_n$ un minimiseur du risque empirique: $\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_n(g)$, où $\mathcal{G}$ est une classe de reconstruction. Comme nous l'avons déjà fait précédemment, les performances de $\widehat{g}_n$ sont alors mesurées par $\mathcal{R}(\widehat{g}_n) - \inf_{g \in \mathcal{G}} \mathcal{R}(g)$, qui peut être borné si nous pouvons obtenir des inégalités de concentration sur l'écart maximal

$$\sup_{g \in \mathcal{G}} |\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g)|. \tag{B.36}$$

Biau & Bleakley (2006) établissent des bornes de l'ordre $O_{\mathbb{P}}(1/\sqrt{n})$ pour le risque de reconstruction de $\widehat{g}_n$ sous des hypothèses de complexité appropriées (à savoir que $\mathcal{G}$ a sa VC-dimension finie). Nous prouvons que les taux d'ordre $O_{\mathbb{P}}(\log n/n)$ sont toujours atteints par les minimiseurs du risque de reconstruction empirique (B.35) sans autres hypothèses. Pour établir ce résultat, nous nous appuyons sur certains arguments utilisés dans l'analyse de taux rapide pour la minimisation empirique de U-statistiques (Clémençon et al., 2008a), bien que ces résultats soient établis sous des hyôthèses resctrictives. Alors que la quantité (1.23) n'est pas une U -statistique, nous réécrivons la différence entre l'excès de risque de reconstruction de toute règle candidate $g \in \mathcal{G}$ et sa contrepartie empirique comme somme de son espérance

conditionnelle étant donné les X_i, qui sont des U-statistiques, plus un terme résiduel. Dénotant par $\Lambda(g) = \mathcal{R}(g) - \mathcal{R}^*$ l'excès de risque de reconstruction par rapport à la règle de Bayes, son estimateur empirique est donné par

$$\Lambda_n(g) = \widehat{\mathcal{R}}_n(g) - \widehat{\mathcal{R}}_n(g^*).$$

Pour tout $g \in \mathcal{G}$, nous avons la décomposition suivante:

$$\Lambda_n(g) - \Lambda(g) = U_n(g) + \widehat{W}_n(g), \tag{B.37}$$

où

$$U_n(g) = \mathbb{E}\left[\Lambda_n(g) - \Lambda(g) \mid X_1, \ldots, X_n\right]$$

est une U-statistique de degree 2.

Sous une certaine condition de «faible bruit», l'analyse effectuée par Clémençon et al. (2008a) montre qu'une "small variance property" des U-statistiques conduit à des vitesses d'apprentissage rapides pour les minimiseurs du risque empirique. Nous montrons que cette condition est toujours remplie pour la U-statistique spécifique $U_n(g)$, apparaissant dans la décomposition (B.37). Ce résultat est dû au fait que le risque de reconstruction empirique n'est pas une moyenne sur toutes les paires (*ie*, une statistique U d'ordre 2) mais une moyenne sur des paires sélectionnées au hasard *aléatoirement* par la fonction η). Nous concluons ensuite la preuve des résultats en établissant que le terme restant $\widehat{W}_n(g)$ est également d'ordre $O_{\mathbb{P}}(1/n)$.

Theorem B.16. (VITESSE RAPIDES) *Soit $\widehat{g}_n$ un minimiseur du risque empirique* (**??**) *sur une classe $\mathcal{G}$ avec* VC*-dimension $V < +\infty$. Pour tout $\delta \in (0,1)$, nous avons avec probabilité au moins $1-\delta$: $\forall n \geq 2$,*

$$\mathcal{R}(\widehat{g}_n) - \mathcal{R}^* \leq 2\left(\inf_{g \in \mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*\right) + C \times \frac{V \log(n/\delta)}{n},$$

où $C < +\infty$ est une constante universelle.

Remark B.17. (SUR LE TERME DE BIAIS) Au delà de son universalité, Theorem B.16 a la même forme que dans le cas des U-Statistiques (Clémençon et al., 2008a, Corollary 6), avec la même constante devant ls biais $\inf_{g \in \mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*$. Néanmoins, la preuve du theorème montre que cette constante n'a pas de signification particulière et peut être remplacée par n'importe quelle constance plus grande que 1 mais en augmentant alors la valeur de la constante C. Notez aussi que la vitesse $O(1/\sqrt{n})$ obtenu par Biau & Bleakley (2006) a un facteur 1 en face du terme de biais. Donc, Théorème B.16 est un meilleur résultat à moins que le terme de biais ne domine complètement le second terme de la borne (*i.e.*, la compléxité de $\mathcal{G}$ est trop faible).

Nous concluons enfin notre analyse en passant le processus d'apprentissage à l'échelle. Comme pour le chapitre 5, pour les graphes d'entraînement, la complexité de simplement calculer $\widehat{\mathcal{R}}_n(g)$ est prohibitive car le nombre de termes impliqués dans la sommation est de l'ordre de $O(n^2)$. Tout comme nous l'avons fait dans la section B.3.3, nous introduisons une approche basée sur l'échantillonnage pour construire des approximations du risque de reconstruction avec beaucoup moins de termes que $O(n^2)$. Au lieu du risque de reconstruction empirique (B.35), nous considérons une approximation incomplète obtenue en échantillonnant des *paires de sommets* (et non pas sommets) avec remise. Un parallèle peut facilement être tiré avec les résultats obtenus dans le chapitre 5 où nous avons recommandé d'implémenter l'algorithme SGD avec des U -Statistiques incomplètes, ce qui correspond dans ce cas à tirer des arètes plutôt que de noeuds. Formellement, nous définissons le *risque de reconstruction*

de graphe incomplet basé sur $B \geq 1$ paires de sommets comme

$$\widetilde{\mathcal{R}}_B(g) = \frac{1}{B} \sum_{(i,j)\in\mathcal{P}_B} \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}, \tag{B.38}$$

où $\mathcal{P}_B$ est un ensemble de cardinalité B construit en échantillonnant avec remise dans l'ensemble $\Theta_n = \{(i,j) : \ 1 \leq i < j \leq n\}$. Pour tout $b \in \{1, \ \ldots, \ B\}$ et tout $(i,j) \in \Theta_n$, notons $\epsilon_b(i,j)$ la variable aléatoire indiquant si la paire (i,j) a été sélectionnée au b-ième tirage. Le risque incomplet est alors représenté par:

$$\widetilde{\mathcal{R}}_B(g) = \frac{1}{B} \sum_{b=1}^{B} \sum_{(i,j)\in\Theta_n} \epsilon_b(i,j) \cdot \mathbb{I}\left\{g(X_i, X_j) \neq e_{i,j}\right\}. \tag{B.39}$$

et étant donné les X_i, son espérance conditionnelle est égale à (4.2). En prenant $B = o(n^2)$, les coûts de calcul sont considérablement réduits, au prix d'une variance légèrement accrue. Nous caractérisons la performance des solutions $\widetilde{g}_B$ pour le problème de calcul plus simple $\min_{g\in\mathcal{G}} \widetilde{\mathcal{R}}_B(g)$ et montrons qu'avec seulement $B = O(n)$ paires, le taux est du même ordre (jusqu'à un facteur log) que celui obtenu par Biau & Bleakley (2006) pour l'écart maximal $\sup_{g\in\mathcal{G}} |\widehat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$ lié au risque de reconstruction complet $\widehat{\mathcal{R}}_n(g)$ avec des paires $O(n^2)$.

Theorem B.18. (UNIFORM DEVIATIONS) *Supposons que la class $\mathcal{G}$ a* VC*-dimension $V < +\infty$. Pour tout $\delta > 0$, $n \geq 1$ et $B \geq 1$, nous avons avec probabilité au moins $1 - \delta$:,*

$$\sup_{g\in\mathcal{G}} |\widetilde{\mathcal{R}}_B(g) - \widehat{\mathcal{R}}_n(g)| \leq \sqrt{\frac{\log 2 + V \log\left((1 + n(n-1)/2)/\delta\right)}{2B}}.$$

Comme attendu, nous montrons que le nombre $B \geq 1$ de paires de sommets joue le rôle d'un paramètre de réglage, arbitrant un compromis entre précision statistique (prenant $B(n) = O(n^2)$ entièrement préserve le taux de convergence) et la complexité de calcul.

Theorem B.19. *Soit $\widetilde{g}_B$ un minimiseur de (6.7) sur une classe $\mathcal{G}$ avec* VC*-dimension $V < +\infty$. Alors pour tout $\delta \in (0,1)$, nous avons avec probabilité au moins $1 - \delta$: $\forall n \geq 2$,*

$$\mathcal{R}(\widetilde{g}_B) - \mathcal{R}^* \leq 2\left(\inf_{g\in\mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*\right) + CV \log(n/\delta) \times \left(\frac{1}{n} + \frac{1}{\sqrt{B}}\right),$$

où $C < +\infty$ est une constance universelle.

des expériences numériques illustrant et justifiant les différents résultats théoriques proposés dans cette sous-section sont aussi réalisés.

Bibliography

S. Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *JMLR*, 15:1653–1674, 2014.

A.Nemirovski, A. Juditsky, G.Lan, and A.Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J.OPTIM*, 2009.

A. Antos, , L. Györfi, and A. György. Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inf. Theory*, 51(11):4013–4023, 2005.

M.A. Arcones and E. Giné. U-processes indexed by Vapnik-Chervonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *Stochastic Processes and their Applications*, 52:17–38, 1994.

YVES F Atchade, GERSENDE Fort, and ERIC Moulines. On stochastic proximal gradient algorithms. *arXiv preprint arXiv:1402.2365*, 23, 2014.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 23*, 2011a.

F. R. Bach and E. Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *NIPS*, 2011b.

R Bardenet and O.A. Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

R. Bekkerman, M. Bilenko, and J. Langford. *Scaling Up Machine Learning*. Cambridge, 2011.

A. Bellet, A. Habrard, and M. Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data. Technical report, arXiv:1306.6709, June 2013.

A. Bellet, A. Habrard, and M. Sebban. *Metric Learning*. Morgan & Claypool Publishers, 2015.

Y.G. Berger. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Stat. Plan. Inf*, 67(2):209–226, 1998.

Y.G. Berger. Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pak. J. Statist*, 27(4):407–426, 2011.

S. N. Bernstein. On a modification of chebyshev's inequality and on the error in laplace formula. *Collected Works, Izd-vo 'Nauka', Moscow (in Russian)*, 4:71–80, 1964.

P. Bertail, E. Chautru, and S. Clémençon. Empirical processes in survey sampling. *Submitted to the Scandinavian Journal of Statistics*, 2013.

D. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

P. Bianchi, S. Clémençon, J. Jakubowicz, and G. Moral-Adell. On-Line Learning Gossip Algorithm in Multi-Agent Systems with Local Decision Rules. In *Proceedings of the IEEE International Conference on Big Data*, 2013.

G. Biau and L. Bleakley. Statistical Inference on Graphs. *Statistics & Decisions*, 24:209–232, 2006.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993.

C. M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

V. Borkar. *Stochastic Approximation: a Dynamical Systems Viewpoint*. Cambridge, 2008.

L. Bottou and O. Bousquet. The Tradeoffs of Large Scale Learning. In *NIPS*, 2007.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20*, 2008.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005a.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005b.

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Stat.*, 33(2):514–560, 03 2005c.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 2002.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004.

P. Brändén and J. Jonasson. Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39 (4):830–838, 2012.

N.E. Breslow, T. Lumley, C. Ballantyne, L. Chambless, and M. Kulich. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Stat. Biosc.*, 1:32–49, 2009.

N.E. Breslow and J.A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 35:186–192, 2007.

N.E. Breslow and J.A. Wellner. A Z-theorem with estimated nuisance parameters and correction note for "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression". *Scandinavian Journal of Statistics*, 35:186–192, 2008.

X.H. Chen, Dempster A.P., and J.S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3):457–469, 1994.

S. Clémençon. A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42 – 56, 2014.

S. Clémençon, A. Bellet, and I. Colin. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *Journal of Machine Learning Research*, 17:1–36, 2016.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008a.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *Ann. Statist.*, 36, 2008b.

S. Clémençon and S. Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *ICML*, 2011.

S. Clémençon, S. Robbiano, and J. Tressou. Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling. In *Proceedings of the SIAM International Conference on Data-Mining*, 2013.

S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, 32:619–648, 2010.

S. Clemencon, P. Bertail, E. Chautru, and G. Papa. Scaling up m-estimation via sampling designs: The horvitz-thompson stochastic gradient descent. In *Big Data (Big Data), 2014 IEEE International Conference on*, 2014.

W.G. Cochran. *Sampling techniques*. Wiley, NY, 1977.

W. Cukierski, B. Hamner, and B. Yang. Graph-based features for supervised link prediction. In *IJCNN*, 2011.

V. De la Pena and E. Giné. *Decoupling : from dependence to independence*. Springer, 1999.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *ArXiv e-prints*, 2014.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *NIPS*, 2014.

B. Delyon. Stochastic Approximation with Decreasing Gain: Convergence and Asymptotic Theory. 2000.

J.C. Deville. *Réplications d'échantillons, demi-échantillons, Jackknife, bootstrap dans les sondages*. Economica, Ed. Droesbeke, Tassi, Fichet, 1987.

J.C. Deville and C.E. Särndal. Calibration estimators in survey sampling. *JASA*, 87:376–382, 1992.

L. Devroye. Non-uniform random variate generation, 1986.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996a.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of mathematics : stochastic modelling and applied probability. U.S. Government Printing Office, 1996b.

D.Needell, N.Srebro, and R.Ward. Stochastic gradient descent, weighted sampling and the randomized kaczmarz algorithm. 2014.

J. Dupacova. A note on rejective sampling. *Contribution to Statistics (J. Hajek memorial volume) Academia Prague*, pages 71–78, 1979.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001.

G.Fort. Central limit theorems for stochastic approximation with controlled Markov Chain. *EsaimPS*, 2014.

R.D. Gill, Y. Vardi, and J.A. Wellner. Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, 16(3):1069–1112, 1988.

L.A. Goodman. On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*, 20:572–579, 1949.

J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.*, 35(4):1491–1523, 1964.

A. Herschtal and B. Raskutti. Optimising area under the ROC curve using gradient descent. In *ICML*, page 49, 2004.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, 19: 293–325, 1948.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *JASA*, 47:663–685, 1951.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

S. Janson. The asymptotic distributions of Incomplete U-statistics. *Z. Wahrsch. verw. Gebiete*, 66: 495–505, 1984.

S. Janson. Large deviation inequalities for sums of indicator variables. 1994.

S. Janson. On concentration of probability. *Contemporary combinatorics*, 11, 2002.

S. Janson and K. Nowicki. The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probability Theory and Related Fields*, 90:341–375, 1991.

K. Joag-Dev and F. Proschan. Negative Association of Random Variables with Applications. *The Annals of Statistics*, 11(1):286–295, 1983.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323. 2013a.

R. Johnson and T. Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *NIPS*, pages 315–323, 2013b.

M. Kanehisa. Prediction of higher order functional networks from genomic data. *Pharmacogenomics*, 2(4):373–385, 2001.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.

J.B. Kramer, J. Cutler, and A.J. Radcliffe. Negative dependence and srinivasan's sampling process. *Combinatorics, Probability and Computing*, 20(3):347–361, 2011.

H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2010.

L.Bottou. Online algorithms and stochastic approximations. In *Online Learning and Neural Networks*. 1998.

L.Bottou. Stochastic gradient tricks. In *Neural Networks, Tricks of the Trade, Reloaded*. 2012.

Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, 16(3):605–613, 08 2010.

A. J. Lee. *U-Statistics: Theory and Practice*. 1990a.

A.J. Lee. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990b.

D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

R. Lichtenwalter, J. Lussier, and N. Chawla. New perspectives and methods in link prediction. In *KDD*, 2010.

J. Mairal. Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization. *ArXiv e-prints*, 2013.

J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *ArXiv e-prints*, 2014.

Enno Mammen, Alexandre B Tsybakov, et al. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIV, volume 1896 of Lecture Notes in Mathematics.* Springer-Verlag, 2007.

P. Massart and E. Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5), 2006.

G. Mateos, J.A. Bazerque, and G.B. Giannakis. Distributed sparse linear regression. *Signal Processing, IEEE Transactions on*, 58(10):5262–5276, 2010.

Colin McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, Algorithms and Combinatorics. Springer, 1998.

A. Navia-Vazquez, D. Gutierrez-Gonzalez, E. Parrado-Hernandez, and JJ Navarro-Abellan. Distributed support vector machines. *Neural Networks, IEEE Transactions on*, 17(4):1091–1097, 2006.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009a. ISSN 1052-6234.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009b.

Y. Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004a. ISBN 1-4020-7553-7.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004b.

Y. Nesterov and I.U.E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Prob.*, 8(1):10–44, 1998.

P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 14(4):1154–1178, 2009.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22, 1951.

P.M. Robinson. On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24(2):234–238, 1982.

P. Rosen. Asymptotic theory for successive sampling. *AMS*, 43:373–397, 1972.

T. Saegusa and J.A. Wellner. Weighted likelihood estimation under two-phase sampling. *Preprint available at http://arxiv.org/abs/1112.4951v1*, 2011.

C.E. Särndall and J. Wretman B. Swensson. *Model assisted survey sampling*. Springer-Verlag, NY, 2003.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972.

M. W. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *CoRR*, 2013.

R. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

S. Shalev-Shwartz and T. Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *ArXiv e-prints*, 2012.

B. Shaw, B. Huang, and T. Jebara. Learning a Distance Metric from a Network. In *NIPS*, 2011.

S.Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. 2009.

D. Spielman. Fast Randomized Algorithms for Partitioning, Sparsification, and Solving Linear Systems. *Lecture notes from IPCO Summer School 2005*, 2005.

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

Y. Tillé. *Sampling algorithms*. Springer Series in Statistics, 2006.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, 32(1):135–166, 02 2004.

S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.

A.W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New-York, 2001.

V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

J.-P. Vert, J. Qiu, and W. S. Noble. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8(10), 2007.

J-P. Vert and Y. Yamanishi. Supervised graph inference. In *NIPS*, pages 1433–1440, 2004.

P. Zhao, S. Hoi, R. Jin, and T. Yang. AUC Maximization. In *ICML*, pages 233–240, 2011.

P. Zhao and T. Zhang. Stochastic Optimization with Importance Sampling. *ArXiv e-prints*, 2014.

P. Zhao and T. Zhang. Stochastic Optimization with Importance Sampling for Regularized Loss Minimization. In *ICML*, 2015.