Optimal survey schemes for stochastic gradient descent with applications to M-estimation

Stéphan Clémençon, Patrice Bertail, Emilie Chautru, Guillaume Papa

ESAIM: PS 23 (2019) 310–337
https://doi.org/10.1051/ps/2018021

ESAIM: Probability and Statistics
www.esaim-ps.org

OPTIMAL SURVEY SCHEMES FOR STOCHASTIC GRADIENT DESCENT WITH APPLICATIONS TO *M*-ESTIMATION

Stephan Clémençon[1,*], Patrice Bertail[2], Emilie Chautru[3] and Guillaume Papa[1]

Abstract. Iterative stochastic approximation methods are widely used to solve M-estimation problems, in the context of predictive learning in particular. In certain situations that shall be undoubtedly more and more common in the Big Data era, the datasets available are so massive that computing statistics over the full sample is hardly feasible, if not unfeasible. A natural and popular approach to gradient descent in this context consists in substituting the "full data" statistics with their counterparts based on subsamples picked at random of manageable size. It is the main purpose of this paper to investigate the impact of survey sampling with unequal inclusion probabilities on stochastic gradient descent-based M-estimation methods. Precisely, we prove that, in presence of some *a priori* information, one may significantly increase statistical accuracy in terms of limit variance, when choosing appropriate first order inclusion probabilities. These results are described by asymptotic theorems and are also supported by illustrative numerical experiments.

Mathematics Subject Classification. 62D05.

Received January 3, 2018. Accepted October 22, 2018.

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. Such quantities may also correspond either to calibrated or post-stratification weights, minimizing some measure of discrepancy under certain margin constraints for the inclusion probabilities. Since the seminal contribution of [25], asymptotic analysis of Horvitz–Thompson estimators based on survey data has received much attention, in the context of mean estimation and regression in particular, refer to *e.g.* [3, 21, 24, 33, 34] and a functional limit theory for distribution function estimation is

Keywords and phrases: Asymptotic analysis, central limit theorem, Horvitz–Thompson estimator, M-estimation, Poisson sampling, stochastic gradient descent, survey scheme.

[1] Telecom ParisTech LTCI, Université Paris Saclay, 46 Rue Barrault, Paris 75634, France.
[2] Université Paris Ouest, MODAL'X, 200 Avenue de la République, Nanterre 92000, France.
[3] Mines ParisTech, PSL University, Centre de géosciences, 35 Rue Saint Honoré, Fontainebleau 77305, France.

* Corresponding author: stephan.clemencon@telecom-paristech.fr

also progressively documented, see [5, 13–15, 23, 35] for instance. At the same time, with the design of successful algorithms such as neural networks, support vector machines or boosting methods, the practice of statistical learning has very rapidly developed and is now supported by a sound theory based on results in the study of empirical processes, see [12, 22, 26]. Nevertheless, our increasing capacity to gather data has improved much faster than our capacity to process and analyze big datasets. 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 [2, 7, 28, 29] among others. It also strongly suggests to use sampling techniques, as a remedy to the apparent intractability of learning from datasets of explosive size, in order to break the current computational barriers, see [16] or [17] and advocates in particular the use of stochastic gradient algorithms (SGD in abbreviated form, see [10]) for large-scale M-estimation problems, as discussed in [11]. It is the purpose of the present article to explore this approach further, by showing how to incorporate efficiently survey schemes into such iterative techniques. 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 paper 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.* [4]. These theoretical results are also supported by strong empirical evidence. The numerical experiments we carried out clearly show the advantages of the approach promoted in this paper. 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 shedding light on the impact of survey sampling on this technique, in its most generic form. We also point out that a very preliminary version of this work has been presented at the 2014 IEEE International Conference on Big Data, the present article offering a much more complete theoretical study, including unconditional limit results and a nonasymptotic rate bound analysis for the HTGD method, with detailed proofs and illustrative numerical examples.

The rest of the paper is structured as follows. Basics in M-estimation and SGD techniques together with key notions in survey sampling theory are briefly recalled in Section 2. Section 3 first describes the Horvitz–Thompson variant of the SGD in the context of a general M-estimation problem. In Section 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 6. Technical proofs are postponed to the Appendix section, together with a rate bound analysis of the HTGD algorithm.

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. Here and throughout, the indicator function of an event $\mathcal{B}$ is written $\mathbb{I}\{\mathcal{B}\}$. The transpose of a matrix A is denoted by $A^\intercal$ and the square root of a symmetric semi-definite positive matrix B by $B^{1/2}$.

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 $\psi : \mathbb{R}^d \times \Theta \to \mathbb{R}$ 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

$$\min_{\theta \in \Theta} L(\theta), \tag{2.1}$$

where $L : \theta \in \Theta \mapsto \mathbb{E}[\psi(Z, \theta)] \in \mathbb{R}$ is called the risk function. Because it is not directly accessible, it is typically replaced in (2.1) by its empirical counterpart

$$L_N : \theta \in \Theta \mapsto \frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta), \tag{2.2}$$

based on the observation of $N \geq 1$ independent copies $Z_1, \ldots, Z_N$ of Z (see Examples 2.1 and 2.2). As $N \to +\infty$, asymptotic properties of M-estimators, *i.e.* minimizers of $L_N(\theta)$, have been extensively investigated, see [36] for instance.

Here and throughout, the gradient and Hessian operators with respect to θ are denoted by ∇ and ∇^2, respectively, with gradient values represented as column vectors.

Gradient descent. A very popular approach to empirical risk minimization consists in implementing variants of the standard gradient descent method, following the iterations

$$\theta(t+1) = \theta(t) - \gamma(t) \nabla L_N(\theta(t)), \quad t \geq 1, \tag{2.3}$$

with an initial value $\theta(0)$ arbitrarily chosen and a non-negative learning rate (step size or gain) $\gamma(t)$. 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.* [6]. Here, we place ourselves in a large-scale setting, where the sample size N of the training dataset is so large that computing the gradient of L_N

$$\nabla L_N : \theta \in \Theta \mapsto \frac{1}{N} \sum_{i=1}^{N} \nabla \psi(Z_i, \theta), \tag{2.4}$$

at each iteration (2.3) is too demanding regarding available memory capacity. Beyond parallel and distributed implementation strategies (see [2]), a natural approach consists in replacing (2.4) 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:

$$\ell_n : \theta \in \Theta \mapsto \frac{1}{n} \sum_{i \in S} \nabla \psi(Z_i, \theta).$$

The convergence properties of variants of such a stochastic gradient descent, usually referred to as *mini-batch* SGD, have received a good deal of attention, in particular in the case $n = 1$, suited to the *on-line* situation where training data are progressively available. Results, mainly based on stochastic approximation

combined with convex minimization theory under appropriate assumptions on the decay of the step size $\gamma(t)$, are well-documented in the literature. References are much too numerous to be listed exhaustively, see [27] for instance.

Example 2.1. (Binary classification) In the usual binary classification framework, 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 $\mathrm{sign}(h(X))$, where $h:\mathcal{X}\to\mathbb{R}$ is some measurable function that minimizes the risk

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

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 $H^*:x\in\mathcal{X}\mapsto\mathrm{sign}(h^*(x))\in\{-1,+1\}$ 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 (2.1) with $Z=(X,Y)$ and $\psi(Z,\theta)=\varphi(-Y\,h(X,\theta))$ 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 tth 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 2.2. (Logistic regression) Consider the same probabilistic model as above, except that the goal pursued is to find $\theta\in\Theta$ so as to minimize $L_N(\theta)$ in (2.2) with $Z_i=(X_i,Y_i)$ and $\psi(Z_i,\theta)$ defined as

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

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:

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

2.2. Survey sampling and Horvitz–Thompson estimation

Let $(\Omega,\mathcal{A},\mathbf{P})$ be a probability space and N a positive integer. In the framework we consider throughout the article, 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. From this point forward, only sampling plans with positive first order inclusion probabilities shall be considered.

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_is 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 [24]. 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 $\left\{\sum_{i=1}^N \epsilon_i = n\right\}$. One may refer to [18, 20] for accounts of survey sampling techniques and examples of designs to which the subsequent analysis applies.

Horvitz–Thompson estimators. 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)

$$\mathbf{Q}_{R_N} = \sum_{i \in S} \frac{1}{\pi_i} Q_i = \sum_{i=1}^N \frac{\epsilon_i}{\pi_i} Q_i.$$

Given the whole statistical population $Q_1, \ldots, Q_N$, the HT estimator is an unbiased estimate of the total:

$$\mathbb{E}\left[\mathbf{Q}_{R_N} \mid Q_1, \ldots, Q_N\right] = \mathbf{Q}_N \text{ almost-surely.}$$

Its conditional variance is given by

$$\mathbb{V}\left(\mathbf{Q}_{R_N} \mid Q_1, \dots, Q_N\right) = \sum_{i,j=1}^{N} \frac{\pi_{i,j} - \pi_i\, \pi_j}{\pi_i\, \pi_j}\, Q_i\, Q_j^{\intercal}.$$

In particular, when the survey design is a Poisson plan P_N with positive probabilities $p_1, \dots, p_N$, this turns into

$$\mathbb{V}\left(\mathbf{Q}_{P_N} \mid Q_1, \dots, Q_N\right) = \sum_{i=1}^{N} \frac{1 - p_i}{p_i}\, Q_i\, Q_i^{\intercal}. \tag{2.5}$$

Remark 2.3. (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 Section 4.1.

Going back to the SGD problem, the *Horvitz–Thompson estimator* of the gradient $\nabla L_N(\theta)$ based on a survey sample S drawn within the population $\mathcal{U}_N = \{1, \dots, N\}$ from a design R_N with vector of (first order) inclusion probabilities $\boldsymbol{\pi}_N = (\pi_1, \dots, \pi_N)$ and inclusion vector $\boldsymbol{\epsilon}_N = (\epsilon_1, \dots, \epsilon_N)$ is

$$\ell_{R_N}(\theta) = \frac{1}{N} \sum_{i \in S} \frac{1}{\pi_i}\, \nabla\psi(Z_i, \theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i}\, \nabla\psi(Z_i, \theta), \quad \theta \in \Theta. \tag{2.6}$$

As pointed out in Remark 2.3, this estimator would be most efficient if each π_i was strongly correlated with the corresponding $\nabla\psi(Z_i, \theta)$, $i \in \mathcal{U}_N$. 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 $W_1, \dots, W_N$ 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 $\nabla\psi(Z_i, \theta)$. As shall be seen in Section 6, this happens to be the case in many situations encountered in practice.

3. The Horvitz–Thompson gradient descent

This section presents, in full generality, the variant of the SGD method we promote in this article. 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_0 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 $n_0 \leq N$. 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_0. 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 $\theta_N(t)$

$$\ell_{R_N}(\theta_N(t)) := \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i^{(t)}}{\pi_i(\theta_N(t))} \nabla \psi(Z_i, \theta_N(t)).$$

   (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.

Remark 3.1. (BALANCE BETWEEN ACCURACY AND COMPUTATIONAL COST) We point out that the complexity of any Poisson sampling algorithm is $O(N)$, just as in the usual case where the data involved in SGD are uniformly drawn with(out) replacement. However, even if it can be straightforwardly parallelized, the numerical computation of the inclusion probabilities at each step naturally induces a certain amount of additional latency. Hence, although HTGD may largely outperform SGD for a fixed number of iterations, this should be taken into consideration for optimizing computation time.

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.

4.1. Poisson schemes with unequal inclusion probabilities

Fix $\theta \in \Theta$ and $n_0 \leq N$. 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, equation (2.5) implies

$$\mathbb{E}\Big[\|\ell_{P_N}(\theta) - \ell_N(\theta)\|^2 \mid Z_1, \ldots, Z_N\Big] = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1-p_i}{p_i} \|\nabla\psi(Z_i, \theta)\|^2 .$$

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 $n_0 \leq N$ yields the optimization problem

$$\min_{\mathbf{p}_N \in (0,1]^N} \sum_{i=1}^{N} \frac{1-p_i}{p_i} \|\nabla\psi(Z_i, \theta)\|^2 \quad \text{subject to} \quad \sum_{i=1}^{N} p_i = n_0. \tag{4.1}$$

Suppose that $\mathbb{P}\{\nabla\psi(Z, \theta) = 0\} = 0$ for all $\theta \in \Theta$; this is true in particular when the set $\{z \in \mathbb{R}^d : \nabla\psi(z, \theta) = 0\}$ has finite cardinality and the distribution of Z is absolutely continuous with respect to the Lebesgue measure. Then we have $\|\nabla\psi(Z_i, \theta)\| > 0$ 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

$$\widetilde{p}_i(\theta) := n_0 \frac{\|\nabla\psi(Z_i, \theta)\|}{\sum_{j=1}^{N} \|\nabla\psi(Z_j, \theta)\|},$$

provided that the following condition is fulfilled:

$$\widetilde{p}_i(\theta) \leq 1 \quad \text{for all} \quad i \in \mathcal{U}_N. \tag{4.2}$$

A straightforward application of Hoeffding's inequality shows that if

$$\varepsilon := \frac{\mathbb{E}\left[\|\nabla\psi(Z, \theta)\|\right]}{\sup_{z \in \mathbb{R}^d} \|\nabla\psi(z, \theta)\|} - \frac{n_0}{N} \in \left(0, \frac{\mathbb{E}\left[\|\nabla\psi(Z, \theta)\|\right]}{\sup_{z \in \mathbb{R}^d} \|\nabla\psi(z, \theta)\|}\right),$$

then condition (4.2) is satisfied with probability larger than $1 - \exp\left(-2N\varepsilon^2\right)$.

Remark 4.1. (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 (4.1) 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

$$\sum_{i=1}^{N} \frac{1-p_i}{p_i} \|\nabla\psi(Z_i, \theta)\|^2 + \lambda \left(\sum_{i=1}^{N} p_i - n_0\right) + \sum_{i=1}^{N} \mu_i (p_i - 1),$$

are necessary and sufficient: (i) $\sum_{i=1}^{N} p_i = n_0$ and for all $i \in \mathcal{U}_N$ (ii) $0 < p_i \leq 1$,

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

Denoting by $m < n_0$ 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 $\|\nabla\psi(Z_{\sigma(1)},\theta)\| \leq \ldots \leq \|\nabla\psi(Z_{\sigma(N)},\theta)\|$, the constraint (i) can be rewritten as $n_0 = m + \sum_{i=1}^{N-m} \|\nabla\psi(Z_{\sigma(i)},\theta)\|/\sqrt{\lambda}$, so that $p_{\sigma(i)} = (n_0 - m)\|\nabla\psi(Z_{\sigma(i)},\theta)\|/\sum_{j=1}^{N-m}\|\nabla\psi(Z_{\sigma(j)},\theta)\|$ 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 $\nabla\psi(Z_i,\theta)$. 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 $(1/N)\sum_{i=1}^{N}\nabla\psi(Z_i,\theta)$ based on the whole dataset. However, as shall be seen in the subsequent analysis (see Prop. 4.8), 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 (in terms of asymptotic covariance) those obtained with all equal inclusion probabilities (*i.e.* $\pi_i = n_0/N$ for all $i \in \mathcal{U}_N$), are exhibited.

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.* Rem. 2.3) 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 (2.6) of the gradient of the empirical risk is unbiased. Reduction of the asymptotic variance of $\theta_N(T)$ and $L_N(\theta_N(T))$ 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 4.2. The conditions below hold true.

(i) For any $z \in \mathbb{R}^d$, the mapping $\theta \in \Theta \mapsto \psi(z,\theta)$ is of class $\mathcal{C}^1$.
(ii) For any compact set $\mathcal{K} \subset \Theta$, we have with probability one:

$$\forall i \in \mathcal{U}_N,\ \sup_{\theta\in\mathcal{K}} \frac{\|\nabla\psi(Z_i,\theta)\|}{\pi_i(\theta)} < +\infty.$$

(iii) The set of stationary points $\mathcal{L}_N = \{\theta \in \Theta : \ell_N(\theta) = 0\}$ is of finite cardinality.

We point out that Assumption 4.2 is essentially a mild envelope condition on the class of functions, necessary to ensure uniform convergence, that is required to establish the following theorem.

Theorem 4.3. (Conditional consistency) *Suppose that Assumption 4.2 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 $\{\theta_N(t)\}_{t\geq 0}$ almost-surely converges to an element of the set $\mathcal{L}_N$ as $t \to +\infty$.

The stability condition is automatically fulfilled when Θ is compact, which encompasses many M-estimation problems in practice, but can be generally difficult to check otherwise: we point out that the existence of a compact set $\mathcal{K}_0$ is linked to the so-called stability condition frequently used in the stochastic approximation community, see [32] for instance. 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, refer to [27] or [9] (see Sect. 5.4 therein) for further details. The goal of the subsequent analysis is to show how the choice of appropriate weights in the mini-batch sampling procedure can favourably impact the limiting variance in presence of auxiliary information. In order to avoid technicalities unnecessary to the understanding of this remarkable phenomenon, the present study is restricted to the simplest framework for stochastic gradient descent.

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 composed of outcomes such that $\theta_N(t)$ converges (in the usual sense) to θ_N^* as $t \to \infty$:

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

Assumption 4.4. The conditions below hold true.

(i) There exists a neighborhood $\mathcal{V}$ of θ_N^* such that for all $z \in \mathbb{R}^d$, the mapping $\theta \in \Theta \mapsto \psi(z,\theta)$ is of class $\mathcal{C}^2$ on $\mathcal{V}$.
(ii) The Hessian matrix $H_N = \nabla^2 L_N(\theta_N^*)$ is a stable $q \times q$ positive-definite matrix, *i.e.* its smallest eigenvalue l is positive.
(iii) For all $(i,j) \in \mathcal{U}_N^2$, the mapping $\theta \in \mathcal{V} \mapsto \pi_{i,j}(\theta)$ is continuous.

We underline that L^2 regularization is very often incorporated into the optimization problems, which automatically makes them strongly convex and also provides a lower bound on the strong convexity constant, lower bound that can be used to choose the constant γ_0 involved in the step size considered in the result stated below.

Theorem 4.5. (Conditional Central Limit Theorem) *Suppose that Assumptions 4.2–4.4 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

$$H_N \Sigma + \Sigma H_N + 2\eta\Sigma = \Gamma_N^*, \tag{4.3}$$

with $\Gamma_N^* = \Gamma_N(\theta_N^*)$ and, for all $\theta \in \Theta$,

$$\Gamma_N(\theta) = \frac{1}{N^2} \sum_{i,j=1}^{N} \left(\frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\pi_j(\theta)} - 1 \right) \nabla\psi(Z_i,\theta)\, \nabla\psi(Z_j,\theta)^\intercal. \tag{4.4}$$

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 4.6. (Error rate) *Under the hypotheses of Theorem 4.5, 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:

$$1/\gamma(t)\,(L_N(\theta_N(t)) - L_N(\theta_N^*)) \Rightarrow \frac{1}{2}\,U^\intercal\,\Sigma_{\boldsymbol{\pi}_N}^{1/2}\,H_N\,\Sigma_{\boldsymbol{\pi}_N}^{1/2}\,U,$$

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

Before showing how the results above can be used to understand how specific sampling designs may improve statistical analysis, a few comments are in order.

Remark 4.7. (Asymptotic covariance estimation) An estimate of $\Sigma_{\boldsymbol{\pi}_N}$ could be obtained by solving the equation $\Sigma H_N + H_N\Sigma + 2\eta\Sigma = \Gamma_N(\theta_N(T))$, replacing in (4.4) the (unknown) target value θ_N^* by the estimate $\theta_N(T)$ produced by the HTGD algorithm after T iterations. Alternatively, a percentile Bootstrap method could also be used for this purpose, repeating $B \geq 1$ times the HTGD algorithm based on replicates of the original sample $\mathcal{D}_N$.

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.* $\|A\|_{\circ\mathsf{F}} = \sqrt{\mathrm{Tr}(A^\intercal A)} = (\sum_{i,j} a_{i,j}^2)^{1/2}$ for any $A = (a_{i,j}) \in \mathcal{M}_q(\mathbb{R})$ where $\mathrm{Tr}(.)$ denotes the Trace operator. For simplicity's sake, we focus on Poisson schemes and consider the case where $\eta = 0$ in Theorem 4.5. Notice that the cross terms ($i \neq j$) in equation (4.4), *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_0 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 Section 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 4.8. (Optimality) *Consider the same assumptions as in Theorem 4.4 in the case where $\eta = 0$ and suppose that*

$$n_0 \leq \inf_{\theta\in\Theta} \frac{\sum_{i=1}^N \|G_N \nabla\psi(Z_i,\theta)\|}{\max_{1\leq i\leq N} \|G_N \nabla\psi(Z_i,\theta)\|}, \tag{4.5}$$

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

$$p_i^*(\theta) = n_0\,\frac{\|G_N \nabla\psi(Z_i,\theta)\|}{\sum_{j=1}^N \|G_N \nabla\psi(Z_j,\theta)\|},$$

is a solution of the minimization problem

$$\min_{\mathbf{p}_N=\{\mathbf{p}_N(\theta)\}_{\theta\in\Theta}} \left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}} \quad \textit{subject to} \quad \sum_{i=1}^N p_i(\theta) = n_0.$$

In addition, we have

$$\left\|\Sigma^{1/2}_{\mathbf{p}^*_N}\right\|^2_{\circ\mathsf{F}} = \frac{1}{2}\left\{\frac{1}{n_0}\left(\frac{1}{N}\sum_{i=1}^N \|G_N\,\nabla\psi(Z_i,\theta^*_N)\|\right)^2 - \frac{1}{N^2}\sum_{i=1}^N \|G_N\,\nabla\psi(Z_i,\theta^*_N)\|^2\right\}.$$

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 (deterministic Gradient Descent) 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 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 $L_N(\theta_N(T)) - L_N(\theta^*_N)$ rather than the asymptotic variance of the estimate itself, the survey design $\mathcal{R}$ must be picked in order to minimize the quantity

$$\mathbb{E}\left[U^\intercal\,\Sigma^{1/2}_{\boldsymbol{\pi}_N}\,H_N\,\Sigma^{1/2}_{\boldsymbol{\pi}_N}U \mid \mathcal{D}_N\right] = \mathbb{E}\left[\left(H_N^{1/2}\,\Sigma^{1/2}_{\boldsymbol{\pi}_N}U\right)^\intercal\left(H_N^{1/2}\,\Sigma^{1/2}_{\boldsymbol{\pi}_N}U\right) \mid \mathcal{D}_N\right] = \mathrm{Tr}\left(H_N\Sigma_{\boldsymbol{\pi}_N}\right),$$

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

$$\mathrm{Tr}\left(H_N\Sigma_{\boldsymbol{\pi}_N}\right) = \frac{1}{2}\,\mathrm{Tr}(\Gamma^*_N).$$

Now, since we have in the Poisson case

$$\mathrm{Tr}(\Gamma^*_N) = \frac{1}{N^2}\sum_{i=1}^N\left(\frac{1}{p_i(\theta^*_N)} - 1\right)\|\nabla\psi(Z_i,\theta^*_N)\|^2,$$

the optimal Poisson scheme regarding this alternative criterion generally differs from that involved in Proposition 4.8 and boils down to that introduced in Section 4.1 for optimal Horvitz–Thompson estimation of the gradient.

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 $p : \mathbb{R}^{d'} \times \Theta \to \mathbb{R}^*_+$ such that $\theta \in \Theta \mapsto p(w,\theta)$ is continuous for all $w \in \mathbb{R}^{d'}$, then choose an expected sample size $n_0 \in \{1, \ldots, N\}$ that satisfies

$$n_0 \leq \inf_{\theta\in\Theta}\frac{\sum_{i=1}^N p(W_i,\theta)}{\max\limits_{1\leq i\leq N} p(W_i,\theta)},$$

and define

$$p_i(\theta) = n_0\,\frac{p(W_i,\theta)}{\sum_{j=1}^N p(W_j,\theta)}, \qquad \text{for all } (i,\theta) \in \mathcal{U}_N \times \Theta. \tag{4.6}$$

Observe that for all $\theta \in \Theta$ we have $\sum_{i=1}^N p_i(\theta) = n_0$ 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 $\nabla\psi(Z_i, \theta)$ (see the examples in Sect. 6) for all $(i, \theta) \in \mathcal{U}_N \times \Theta$. The assumption introduced below involves the empirical covariance $c_N(\theta)$ between $\|G_N \nabla\psi(Z, \theta)\|^2 / p(W, \theta)$ and $p(W, \theta)$, for $\theta \in \Theta$:

$$c_N(\theta) = \frac{1}{N} \sum_{i=1}^N \|G_N \nabla\psi(Z_i, \theta)\|^2 \left(1 - \frac{1}{N} \frac{\sum_{j=1}^N p(W_j, \theta)}{p(W_i, \theta)}\right).$$

Assumption 4.9. 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, $\check{p}_i = n_0/N$ for $1 \leq i \leq n$, when implementing stochastic gradient descent (see Eq. (4.7)). Namely, the accuracy of the HTGD gets closer and closer to the optimum, as the empirical covariance $c_N(\theta^*)$ increases to its maximum (see Eq. (4.8)). Notice that in the case where inclusion probabilities are all equal, we have $c_N \equiv 0$.

Proposition 4.10. *Let n_0 be fixed. Suppose that the collection of Poisson designs $\mathbf{p}$ with expected sizes n_0 is defined by a link function $p(w, \theta)$ satisfying Assumption 4.9. Then, when Theorem 4.5 applies, we have*

$$\left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}} < \left\|\Sigma_{\check{\mathbf{p}}_N}^{1/2}\right\|_{\circ\mathsf{F}}, \tag{4.7}$$

as well as

$$0 \leq \left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 - \left\|\Sigma_{\mathbf{p}_N^*}^{1/2}\right\|_{\circ\mathsf{F}}^2 = \frac{1}{2n_0} \left\{\sigma_N^2(\theta_N^*) - c_N(\theta_N^*)\right\}, \tag{4.8}$$

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 Section 6, one may generally find link functions fulfilling Assumption 4.9 without great effort, permitting to gain in accuracy from the implementation of the HTGD algorithm.

5. Unconditional asymptotic analysis

Building upon the results of the previous section, we now investigate the behavior of the HTGD estimator as N, n_0 and t simultaneously tend to $+\infty$ at appropriate rates. For the sake of simplicity we assume in this section that the minimizer θ^* of the theoretical risk function over the supposedly compact parameter space Θ is unique (see Eq. (2.1)), 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 (4.4), is involved in the subsequent unconditional analysis.

Assumption 5.1. As both N and n_0 tend to ∞, $n_0\Gamma_N^*$ converges in probability toward a positive semi-definite matrix Γ^*.

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

Proposition 5.2. *Suppose that the survey schemes are of Poisson type with link functions $p(.,\theta) : \mathbb{R}^{d'} \to \mathbb{R}_+^*$, $\theta \in \Theta$, based on the auxiliary information W observed on the statistical population. Assume also that Assumption 4.2-(i) is fulfilled, together with the following conditions.*

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

(ii) The expected size n_0 tend to infinity as $N \to \infty$, so that $\dfrac{n_0}{N} \to c_0 \in [0,1]$.

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

$$\mathbb{E}\left[\sup_{\theta\in\Theta} \frac{1}{p(W,\theta)} \nabla\psi(Z,\theta)\nabla\psi(Z,\theta)^\intercal\right] < +\infty.$$

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

Then, the quantities (4.6) define the inclusion probabilities of a Poisson scheme with probability one, as soon as $n_0 \leq N\underline{p}/\bar{p}$. In addition, Assumption 5.1 is fulfilled with

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

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 $N\gamma(t)/n_0$.

Theorem 5.3. *Suppose that Assumptions 4.2, 4.4, 5.1 are fulfilled and that the rate $\gamma(t)$ satisfies the condition of Theorem 4.5 with $\alpha < 1$ (and thus $\eta = 0$). Assume that the symmetric positive semi-definite matrix $H^* = \mathbb{E}[\nabla_\theta^2\psi(Z,\theta^*)]$ is invertible, set*

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

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

(i) If $\lim_{N,n_0,t\to+\infty} \frac{N}{n_0}\gamma(t) = +\infty$, then we have the convergence in distribution:

$$\lim_{N,n_0\to\infty} \left\{\lim_{t\to\infty} \sqrt{n_0/\gamma(t)}\,(\theta_N(t) - \theta^*)\right\} = \mathcal{N}(0,\Sigma^*).$$

(ii) If $\lim_{N,n_0,t\to+\infty} \frac{N}{n_0}\gamma(t) = 0$, then we have the convergence in distribution:

$$\lim_{N,n_0,t\to+\infty} \sqrt{N}\,(\theta_N(t) - \theta^*) = \mathcal{N}(0,\Lambda^*).$$

(iii) If $\lim_{N,n_0,t\to+\infty} \frac{N}{n_0}\gamma(t) = c > 0$, then we have the convergence in distribution:

$$\lim_{N,n_0\to\infty} \left\{\lim_{t\to\infty} \sqrt{N}\,(\theta_N(t) - \theta^*)\right\} = \mathcal{N}(0,\Lambda^* + c\Sigma^*).$$

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 that for large values of N and n_0, the number t of HTGD iterations is such that $1/\gamma(t) \ll N/n_0$ (respectively, such that $1/\gamma(t) \sim N/n_0$). 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. The proof is given in the Appendix. Just like in Corollary 4.6 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; technical details are omitted.

Corollary 5.4. *Suppose that the assumptions of Theorem 5.3 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 $\lim_{N,n_0,t\to+\infty} \frac{N}{n_0}\gamma(t) = +\infty$, *then we have the convergence in distribution:*

$$\lim_{N,n_0\to\infty} \left\{ \lim_{t\to\infty} \frac{n_0}{\gamma(t)} \left(L(\theta_N(t)) - L(\theta^*)\right) \right\} = \frac{1}{2} U^\intercal \Sigma^{*1/2} H^* \Sigma^{*1/2} U.$$

(ii) If $\lim_{N,n_0,t\to+\infty} \frac{N}{n_0}\gamma(t) = 0$, *then we have the convergence in distribution:*

$$\lim_{N,n_0,t\to+\infty} N \left(L(\theta_N(t)) - L(\theta^*)\right) = \frac{1}{2} U^\intercal \Lambda^{*1/2} H^* \Lambda^{*1/2} U.$$

(iii) If $\lim_{N,t\to+\infty} \frac{N}{n_0}\gamma(t) = c > 0$, *then we have the convergence in distribution:*

$$\lim_{N,n_0\to\infty} \left\{ \lim_{t\to\infty} N \left(L(\theta_N(t)) - L(\theta^*)\right) \right\} = \frac{1}{2} U^\intercal (\Lambda^* + c\Sigma^*)^{1/2} H^* (\Lambda^* + c\Sigma^*)^{1/2} U.$$

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.

6.1. Linear logistic regression

Consider the linear logistic regression model corresponding to Example 2.2 with $\theta = (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^d$ and $h(x, \theta) = \alpha + \beta^\intercal x$ for all $x \in \mathbb{R}^d$. Take $K \subset \{1, \ldots, d\}$ with cardinal $d' \ll d$ and complementary set $K^c := \{1, \ldots, d\} \backslash K$. Now let X_K be a low dimensional marginal vector of the input r.v. X, so that one may write $X = (X_K, X_{K^c})$ as well as $\beta = (\beta_K, \beta_{K^c})$ in a similar manner. The problem of fitting the parameter θ through conditional MLE corresponds to the case

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

We propose to implement the HTGD using a Poisson sampling plan with link function $\check{p}((x', y), \theta) \propto \|\nabla\psi_K((X, Y), \theta)\|$, where

$$\psi_K((x, y), \theta) = -\log\left(\frac{e^{\alpha+\beta_K{}^\intercal x_K}(y+1)/2 + (1-y)/2}{1 + e^{\alpha+\beta_K{}^\intercal x_K}}\right).$$

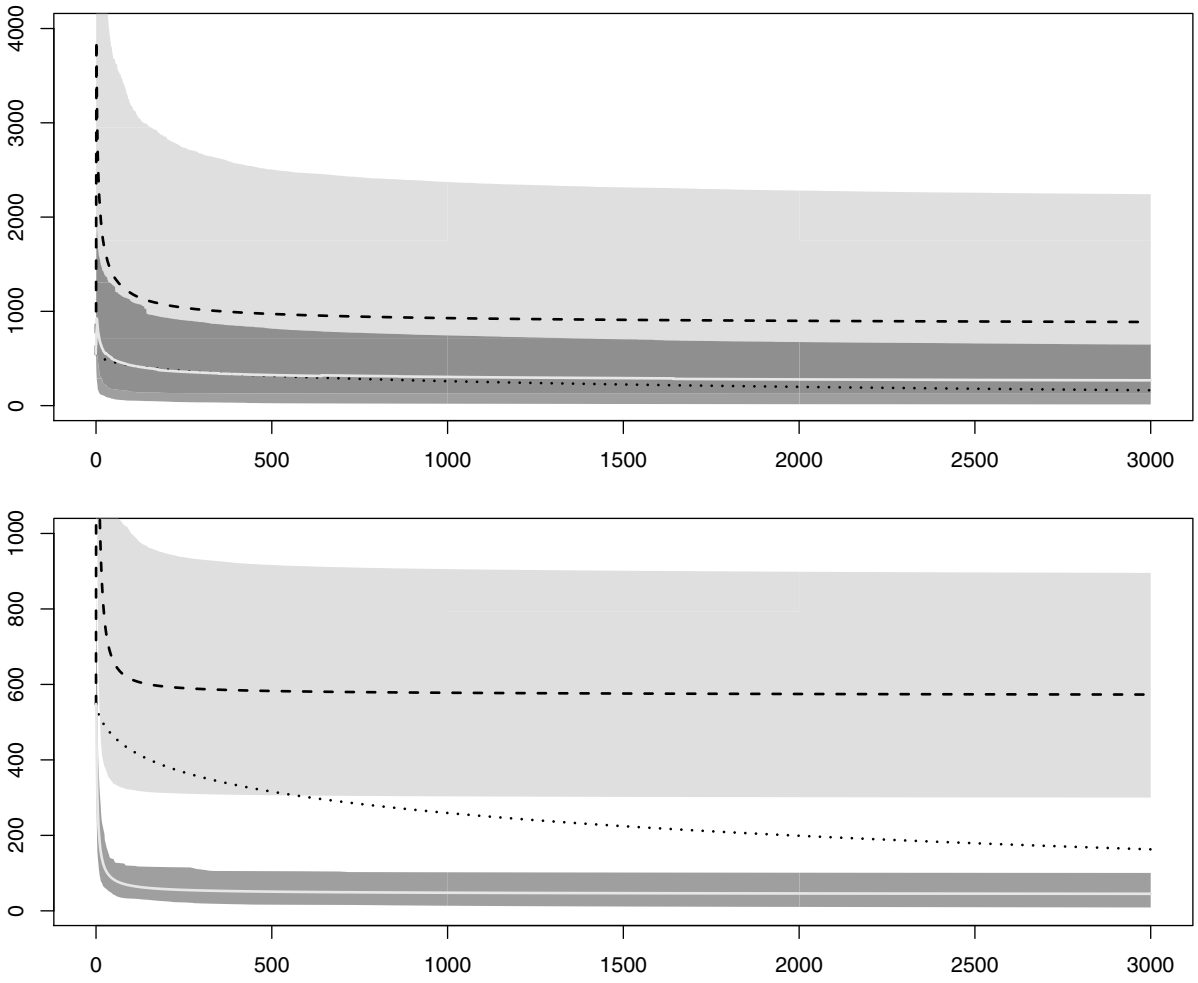

FIGURE 1. Mean squared errors of the estimators of θ with the number of iterations for $n_0 = 10$ (*left*) and $n_0 = 100$ (*right*), based on the 100 repetitions of HTGD (*solid light grey*) and SGD (*dashed black*); the corresponding 90% dispersion bands are displayed in dark and light grey, respectively, and the squared error of the GD (*dotted black*) is indicated as a point of comparison.

In order to illustrate the advantages of the HTGD technique for logistic regression, we considered the toy numerical model with parameters $d = 10$ and $\theta = (\alpha, \beta_1, \ldots, \beta_{10}) = (0, -2, 0, -9, 1, 0, -9, 11, 0, -8, -14)$, the 10 input variables being generated from a centered multivariate Gaussian distribution with all variance terms equal to 1 and all correlation coefficients set to 0.5. The maximum likelihood estimators of θ were computed using the HTGD, SGD (mini-batch) and GD algorithms. In order to compare them, the same number of iterations was chosen in each situation. The learning rate was taken as fixed for the deterministic gradient descent, but proportional to $1/t$ for both HTGD and SD.

We drew a single sample of size $N = 5000$ on which the three algorithms were performed for 3000 iterations. Both HTGD and SGD were repeated 100 times so as to account for the randomness due to their respective sampling phases. Two expected sub-sample sizes were considered : $n_0 = 10$ and $n_0 = 100$. As can be seen in Figure 1, 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 regarding the estimation than both competitors.

As clearly appears in Figure 1 and can be observed in a similar fashion with larger sample sizes in our experience (*e.g.* $N = 50\,000$), a significantly lower number of iterations is required for our algorithm to obtain satisfactory estimates of the parameters of interest. However, adapting this approach to real Big datasets requires to combine the statistical principles investigated in this paper with the use of computing infrastructures dedicated to large-scale problems. This will be the subject of a future work.

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 $\{P_{\theta,f}, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^q$, 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 $f(x)$ 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 $q > 1$. 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 [8]. 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 [8] 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

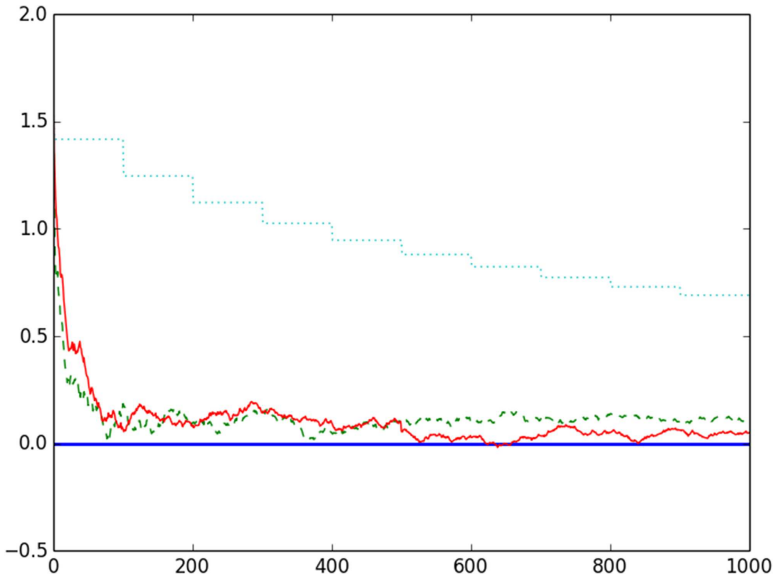

FIGURE 2. Evolution of the estimator of the location parameter $\theta = 0$ of the balanced Gaussian mixture with the number of iterations in the HTGD (*solid red*), mini-batch SGD (*dashed green*) and GD (*dotted blue*) algorithms.

that, starting for instance from the empirical median θ_0 with an adequate choice of the rate $\gamma(t)$, 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

$$\theta_N(t) = \widehat{\theta}(t-1) + \gamma(t)\frac{1}{N}\sum_{j=1}^{N} \frac{\varepsilon_j}{p_j}\widehat{s}_N(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1)),$$

$$\sum_{j=1}^{N} p_j = n_0.$$

As shown in Section 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 tth iteration:

$$p_j^*\left(\widehat{\theta}(t-1)\right) = \frac{n_0\,|\widehat{s}_N(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1))|}{\sum_{i=1}^{N} |\widehat{s}_N(X_j - \widehat{\theta}(t-1), \widehat{\theta}(t-1))|}. \tag{6.1}$$

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 (6.1) yields obvious computational difficulties. For this reason, we suggest to use the (much simpler) Poisson

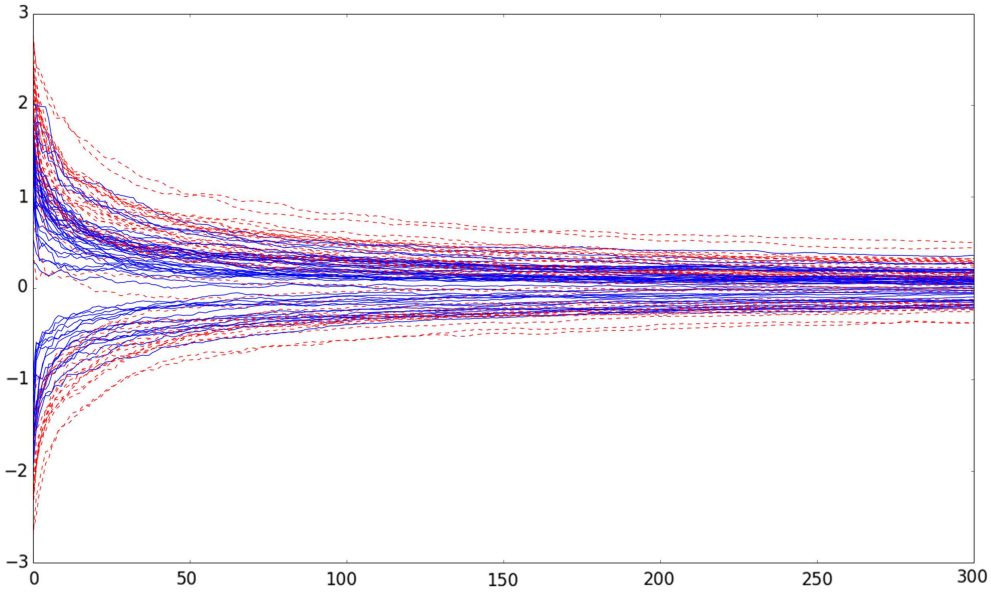

FIGURE 3. Evolution of the estimator of the location parameter $\theta = 0$ of the balanced Gaussian mixture with the number of iterations in the HTGD (*solid blue*) and mini-batch SGD (*dashed red*) algorithms over 50 populations.

TABLE 1. Statistics on the global behavior of the final estimates of the location parameter θ across the 50 simulations.

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

weights:

$$p_j(\theta) = n_0 \frac{|X_j - \theta|}{\sum_{j=1}^{N} |X_j - \theta|}.$$

Figures 2 and 3 depict 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 $n_0 = 10$ and $T = 3000$ iterations, whereas 30 iterations have been made for the basic GD procedure (with $n_0 = N = 1000$) 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, used a step-size $\gamma(t) = \frac{\gamma_0}{t}$ for the HTGD and the SGD, and a constant step-size γ_1 for the GD, see Table 1.

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 paper explores a possible alternative way of scaling up statistical learning methods, based on gradient descent techniques. It hopefully lays a first stone in efficient incorporation of survey techniques into machine-learning algorithms.

Appendix A – Technical proofs

A.1 Proof of Theorem 4.3

The conditional consistency of the HTGD algorithm described in Section 3 is obtained by applying Theorem 13 in [19] (or Thm. 2.2 in Chap. 5 of [27] among other references). Specifically, it states that if the following conditions are fulfilled, then $\theta_N(\theta)$ 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 $\theta \in \Theta \mapsto \nabla L_N(\theta)$ are continuous, which is guaranteed by Assumption 4.2-(i),
- $\mathcal{L}_N$ is finite, which corresponds to Assumption 4.2-(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, which we shall now check.

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

$$\begin{aligned}\sup_{\theta\in\mathcal{K}} \mathbb{E}\left[\|\ell_{R_N}(\theta)\|^2 \mid \mathcal{D}_N\right] &= \sup_{\theta\in\mathcal{K}} \frac{1}{N^2} \sum_{i,j=1}^{N} \frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\,\pi_j(\theta)} \nabla\psi(Z_i,\theta)^\intercal \nabla\psi(Z_j,\theta) \\ &\leq \left(\frac{1}{N}\sum_{i=1}^{N} \sup_{\theta\in\mathcal{K}} \frac{\|\nabla\psi(Z_i,\theta)\|}{\pi_i(\theta)}\right)^2 < \infty \quad a.s.\end{aligned}$$

which is finite with probability 1 by virtue of Assumption 4.2-(ii). □

A.2 Proof of Theorem 4.5

Our conditional Central Limit Theorem results from Theorem 1 in [32], the applicability of which needs to be checked.

First of all, rewrite the algorithm sequence as

$$\theta_N(t+1) = \theta_N(t) - \gamma(t)\,\nabla L_N(\theta_N(t)) + \gamma(t)\,\xi_N(t+1),$$

where $\xi_N(t+1) := \nabla L_N(\theta_N(t)) - \ell_{R_N}(\theta_N(t))$. This way, $-\nabla L_N(\theta_N(t))$ 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 4.2-(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

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

Given this representation, our result is assured by Theorem 1 in [32] provided that the following conditions hold true:

- $\nabla L_N(\theta_N^*) = 0$, which was assumed,
- on a neighborhood $\mathcal{V}$ of θ_N^* we have $\nabla L_N(\theta) = H_N\,(\theta - \theta_N^*) + O(\|\theta - \theta_N^*\|^2)$, which results from a simple Taylor expansion made possible by Assumption 4.4-(i),
- $-H_N$ is a stable $q \times q$ matrix (the largest real part of its eigenvalues is negative), which corresponds to Assumption 4.4-(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) $\mathbb{E}\left[\xi_N(t+1)\, \xi_N(t+1)^\intercal \mid \mathcal{F}_t\right] \to \Gamma$ 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 we have with probability one

$$0 \leq \|\xi_N(t+1)\| \leq \frac{1}{N} \sum_{i=1}^{N} \frac{\|\nabla\psi(Z_i, \theta_N(t))\|}{\pi_i(\theta_N(t))},$$

for all $t \geq 0$, then for any $b > 2$, we almost-surely have

$$\sup_{t \geq 0} \mathbb{E}\left[\|\xi_N(t+1)\|^b \mid \mathcal{F}_t\right] \mathbb{I}\{\theta_N(t) \in \mathcal{V}\} \leq \frac{1}{N} \sum_{i=1}^{N} \left(\sup_{\theta \in \mathcal{V}} \frac{\|\nabla\psi(Z_i, \theta)\|}{\pi_i(\theta)} \right)^b < +\infty,$$

by Assumption 4.2-(ii).

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

$$\Gamma_N(\theta) = \frac{1}{N^2} \sum_{i,j=1}^{N} \left(\frac{\pi_{i,j}(\theta)}{\pi_i(\theta)\, \pi_j(\theta)} - 1 \right) \nabla\psi(Z_i, \theta)\, \nabla\psi(Z_j, \theta)^\intercal.$$

By virtue of the continuity assumptions of the inclusion probabilities (Assumption 4.4-(iii)) and of the gradient (Assumption 4.2-(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 $\mathcal{D}_N$). □

A.3 Proof of Proposition 4.8

Let us start by calculating $\|\Sigma_{\mathbf{p}_N}^{1/2}\|_{\circ\mathsf{F}}^2$ 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 4.4-(ii), the Lyapunov equation

(4.3) 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 $\left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 = \frac{1}{2}\operatorname{Tr}\left(H_N^{-1}\,\Gamma_N^*\right) = \frac{1}{2}\operatorname{Tr}\left(G_N\,\Gamma_N^*\,G_N^\intercal\right)$. Now recall that θ_N^* is a stationary point, *i.e.* $\nabla L_N(\theta_N^*) = 0$, and that we are considering a Poisson scheme (with independent inclusion variables). Thus,

$$\Gamma_N^* = \frac{1}{N^2}\sum_{i=1}^N \left(\frac{1}{p_i(\theta_N^*)} - 1\right) \nabla\psi(Z_i,\theta_N^*)\,\nabla\psi(Z_i,\theta_N^*)^\intercal$$

and then

$$\left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 = \frac{1}{2}\,\frac{1}{N^2}\sum_{i=1}^N \left(\frac{1}{p_i(\theta_N^*)} - 1\right) \left\|G_N\,\nabla\psi(Z_i,\theta_N^*)\right\|^2. \tag{A.1}$$

Let us now turn to the definition of an optimal collection of Poisson plans. Using the Lagrange multiplier method like in Section 4.1, we find that any $\mathbf{p}_N$ minimizing (A.1) must satisfy the equalities

$$p_i(\theta_N^*) = n_0\,\frac{\left\|G_N\,\nabla\psi(Z_i,\theta_N^*)\right\|}{\sum_{j=1}^N \left\|G_N\,\nabla\psi(Z_j,\theta_N^*)\right\|}, \qquad i \in \mathcal{U}_N.$$

This is the case, in particular, of the collection $\mathbf{p}_N^*$ defined in Proposition 4.8. Condition (4.5) and the positive-definiteness of H_N (Assumption 4.4-(*ii*)) ensure that $p_i^*(\theta)$ is almost-surely in $(0,1]$ for all $\theta \in \Theta$ and $i \in \mathcal{U}_N$. □

A.4 Proof of Proposition 4.10

Let us start by proving the first assertion. Using equation (A.1) with the corresponding inclusion probabilities, we immediately obtain

$$\left\|\Sigma_{\tilde{\mathbf{p}}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 - \left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 = \frac{c_N(\theta_N^*)}{2\,n_0},$$

which is positive by Assumption 4.9.

Turning to the second assertion, observe that

$$\left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 - \left\|\Sigma_{\mathbf{p}_N^*}^{1/2}\right\|_{\circ\mathsf{F}}^2 = \left\|\Sigma_{\mathbf{p}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 - \left\|\Sigma_{\tilde{\mathbf{p}}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 + \left\|\Sigma_{\tilde{\mathbf{p}}_N}^{1/2}\right\|_{\circ\mathsf{F}}^2 - \left\|\Sigma_{\mathbf{p}_N^*}^{1/2}\right\|_{\circ\mathsf{F}}^2 = \frac{1}{2\,n_0}\left\{\sigma_N^2(\theta_N^*) - c_N(\theta_N^*)\right\}.$$

By definition of $\mathbf{p}_N^*$ (see Prop. 4.8), this quantity is always nonnegative. □

A.5 Proof of Proposition 5.2

Consider a Poisson sampling plan with inclusion probabilities as in (4.6). We shall prove that Assumption 5.1 is fulfilled by establishing the asymptotic convergences (as $N, n_0 \to +\infty$) of the three averages in brackets that

appear in the following decomposition:

$$n_0\,\Gamma_N^* = \left[\frac{1}{N}\sum_{i=1}^N p(W_i,\theta_N^*)\right] \times \left[\frac{1}{N}\sum_{i=1}^N \frac{1}{p(W_i,\theta_N^*)}\,\nabla\psi(Z_i,\theta_N^*)\nabla\psi(Z_i,\theta_N^*)^\intercal\right] - \frac{n_0}{N}\left[\frac{1}{N}\sum_{i=1}^N \nabla\psi(Z_i,\theta_N^*)\nabla\psi(Z_i,\theta_N^*)^\intercal\right].$$

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 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

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

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 [36]. 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 $(k,h)\in\{1,\ldots,q\}^2$ and define the function $\Psi_{k,h}:(z,\theta)\in\mathbb{R}^d\times\Theta\mapsto(\partial\psi/\partial\theta_k)(z,\theta)\times(\partial\psi/\partial\theta_h)(z,\theta)$. The element at the intersection of the kth row and hth 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 to $\mathbb{E}\left[p(W,\theta^*)^{-1}\,\Psi_{k,h}(Z,\theta^*)\right]$, as $N,n_0\to+\infty$. We only need to check that

(a) the map $\theta\in\Theta\mapsto p(w,\theta)^{-1}\,\Psi_{k,h}(z,\theta)$ is continuous for any (z,w) in $\mathbb{R}^d\times\mathbb{R}^{d'}$,

(b) the envelope $(z,w) \in \mathbb{R}^d \times \mathbb{R}^{d'} \mapsto \mathbb{E}[\sup_{\theta\in\Theta} |p(w,\theta)^{-1}\, \Psi_{k,h}(z,\theta)|]$ is $\mathbb{P}_{Z,W}$-integrable,

(c) 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 4.2-(i). Condition (b) and (c) are direct consequences of conditions (iii).

Third average. The convergence of the third term in bracket is easily deduced from the convergence of the second average, it is therefore omitted.

A.6 An intermediary result

Before tackling the proof of Theorem 5.3, we first establish the following lemma. It describes the limit behavior of the solution of the Lyapunov equation (4.3) as $N, n_0 \to +\infty$.

Lemma A.1. *Suppose that the assumptions of Theorem 5.3 are fulfilled. Then as N, n_0 tend to $+\infty$, we have:*

$$n_0 \Sigma_{\boldsymbol{\pi}_N} \to \Sigma^* \text{ in probability.}$$

Proof. Observe first that it follows from $H_N \Sigma_{\boldsymbol{\pi}_N} + \Sigma_{\boldsymbol{\pi}_N} H_N = \Gamma_N^*$ that

$$\|\Gamma_N^*\|_{\circ\mathsf{F}}^2 = 2\|H_N \Sigma_{\boldsymbol{\pi}_N}\|_{\circ\mathsf{F}}^2 + \underbrace{2\mathrm{Tr}(H_N \Sigma_{\boldsymbol{\pi}_N} H_N \Sigma_{\boldsymbol{\pi}_N})}_{\geqslant 0},$$

Hence, we have:

$$\|\Gamma_N^*\|_{\circ\mathsf{F}} \geqslant \sqrt{2}\|H_N \Sigma_{\boldsymbol{\pi}_N}\|_{\circ\mathsf{F}}.$$

We deduce from this inequality combined with assumptions 5.1 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

$$\Sigma_{\boldsymbol{\pi}_N} = O_{\mathbb{P}}(1/n_0) \text{ as } N \to \infty. \tag{A.2}$$

Since $H^*\Sigma^* + \Sigma^* H^* = \Gamma^*$ and $H_N \Sigma_{\boldsymbol{\pi}_N} + \Sigma_{\boldsymbol{\pi}_N} H_N = \Gamma_N^*$, we have:

$$\Gamma^* - n_0\Gamma_N^* = H^*(\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N}) + (\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N})H^* + n_0(H_N - H^*)\Sigma_{\boldsymbol{\pi}_N} + n_0\Sigma_{\boldsymbol{\pi}_N}(H_N - H^*). \tag{A.3}$$

Combining (A.3) with

$$\|H^*(\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N})\|_{\circ\mathsf{F}} = \|(\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N})H^*\|_{\circ\mathsf{F}} \leqslant \frac{1}{\sqrt{2}}\|H^*(\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N}) + (\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N})H^*\|_{\circ\mathsf{F}} \tag{A.4}$$

we easily get

$$\|H^*(\Sigma^* - n_0\Sigma_{\boldsymbol{\pi}_N})\|_{\circ\mathsf{F}} \leqslant \frac{1}{\sqrt{2}}\|\Gamma^* - n_0\Gamma_N^*\|_{\circ\mathsf{F}} + \sqrt{2}\|n_0\Sigma_{\boldsymbol{\pi}_N}(H_N - H^*)\|_{\circ\mathsf{F}}. \tag{A.5}$$

By virtue of the LLN, we have $H_N - H^* \to 0$ almost surely as $N \to \infty$. Combining this with (A.2) and Assumption 5.1 we see that the term on the right hand side of (A.5) converges toward 0 in probability as $N \to \infty$. Combined with the invertibility of H^*, this establishes the desired result. □

A.7 Proof of Theorem 5.3

Consider the decomposition:

$$\begin{aligned}\theta_N(t) - \theta^* &= \theta_N(t) - \theta_N^* + \theta_N^* - \theta^* \\ &= \sqrt{\frac{\gamma(t)}{n_0}}\sqrt{n_0/\gamma(t)}\,(\theta_N(t) - \theta_N^*) + \frac{1}{\sqrt{N}}\sqrt{N}\,(\theta_N^* - \theta^*) \\ &= \sqrt{\frac{\gamma(t)}{n_0}}\underbrace{\sqrt{n_0/\gamma(t)}\,(\theta_N(t) - \theta_N^*)}_{(1)} + \frac{1}{\sqrt{N}}\underbrace{\sqrt{N}\,(\theta_N^* - \theta^*)}_{(2)}.\end{aligned}$$

The term (2) above is asymptotically normal. By virtue of the classical Central Limit Theorem for M-estimators, see Theorem 5.23 in [37] for instance, we have:

$$\sqrt{N}\,(\theta_N^* - \theta^*) \Rightarrow \mathcal{N}(0, \Lambda^*) \text{ as } N \to \infty. \tag{A.6}$$

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_0 fixed, Theorem 4.5 claims that, in probability along the sequence X (respectively, the sequence (X, W)):

$$\sqrt{1/\gamma(t)}\Sigma_{\boldsymbol{\pi}_N}^{-1/2}\,(\theta_N(t) - \theta_N^*) \Rightarrow Z \text{ as } t \to \infty, \tag{A.7}$$

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 A.1 combined with the continuity of the application that maps any symmetric positive semi-definite matrix to its square root that

$$(n_0\Sigma_{\boldsymbol{\pi}_N})^{1/2} \to \Sigma^{*1/2} \text{ in probability, as } N, n_0 \to \infty. \tag{A.8}$$

Given that one may write

$$\sqrt{\frac{n_0}{\gamma(t)}}\,(\theta_N(t) - \theta_N^*) = (n_0\Sigma_{\boldsymbol{\pi}_N})^{1/2}\sqrt{1/\gamma(t)}\Sigma_{\boldsymbol{\pi}_N}^{-1/2}\,(\theta_N(t) - \theta_N^*)\,,$$

it results from (A.7) and (A.8) that the following convergence in distribution holds true:

$$\lim_{N, n_0 \to \infty}\lim_{t \to \infty}\sqrt{\frac{n_0}{\gamma(t)}}\,(\theta_N(t) - \theta_N^*) = \Sigma^{*1/2}Z. \tag{A.9}$$

Assertions (i) and (iii) can be then deduced from (A.9) and (A.6) in a straightforward fashion (using the independence of the limits, regarding (iii)).

Appendix B – Rate bound analysis

Here, we establish a rate bound for the HTGD algorithm under the additional assumption that the mapping $\theta \mapsto \psi(z, \theta)$ is convex, referred to as 5th assumption. Note that Assumptions 4.4 and the 5th imply that θ_N^* is unique and 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 [1, 30] or [31].

Theorem B.2. *Under the Assumptions of Theorem 4.5 & 5th assumption, 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 $\gamma_0 > 1/(2\ell)$), there exists a constant $\widetilde{C}_\alpha < +\infty$ such that:*

$$\forall t \geq 1,\ \mathbb{E}[\|\theta_N(t) - \theta_N^*\|^2] \leq \frac{\widetilde{C}_\alpha}{t^\alpha}. \tag{B.1}$$

Proof. We restrict ourselves to the case $\alpha = 1$ and follow the proof of [1]. By construction, we have

$$\|\theta_N(t+1) - \theta_N^*\|^2 = \|\theta_N(t) - \theta_N^*\|^2 - 2\gamma(t)\ell_{R_N}(\theta_N(t))^\intercal(\theta_N(t) - \theta_N^*) + \|\gamma(t)\ell_{R_N}(\theta_N(t))\|^2.$$

Since

$$\mathbb{E}[\ell_{R_N}(\theta_N(t))|\mathcal{F}_t] = \nabla L_N(\theta_N(t)),$$

we get

$$\mathbb{E}[|\theta_N(t+1) - \theta_N^*|^2 \mid \theta_N(t)] = \|\theta_N(t) - \theta_N^*\|^2 - 2\gamma(t)\nabla F(\theta_N(t))^\intercal(\theta_N(t) - \theta_N^*) + \gamma(t)^2\mathbb{E}[\|\ell_{R_N}(\theta_N(t))\|^2 \mid \theta_N(t)].$$

The strong convexity property gives

$$L_N(\theta_N(t)) - L_N(\theta_N^*) \leq \nabla L_N(\theta_N(t))^T(\theta_N(t) - \theta_N^*) - \frac{}{2}\|\theta_N(t) - \theta_N^*\|^2$$

and

$$L_N(\theta_N^*) - L_N(\theta_N(t)) \leq -\frac{}{2}\|\theta_N(t) - \theta_N^*\|^2,$$

so that

$$l\|\theta_N(t) - \theta_N^*\|^2 \leqslant \nabla L_N(\theta_N(t))^T(\theta_N(t) - \theta_N^*).$$

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 4.2, 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 $\gamma(t)$ 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. □

References

[1] F. Bach and E. Moulines, E. Moulines, and F. Bach, Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning, in *Vol. 24 of Advances in Neural Information Processing Systems*, edited by J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, K.Q. Weinberger, Curran Associates, Inc. (2011) 451–459.

[2] R. Bekkerman, M. Bilenko and J. Langford, Scaling Up Machine Learning. Cambridge University Press, Cambridge (2011).

[3] Y. Berger, Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Stat. Plan. Inference* **67** (1998) 209–226.

[4] Y. Berger, Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pak. J. Stat.* **27** (2011) 407–426.

[5] P. Bertail, E. Chautru and S. Clémençon, Empirical processes in survey sampling with (conditional) Poisson designs. *Scand. J. Stat.* **44** (2017) 97–111.

[6] D. Bertsekas, Convex Analysis and Optimization. Athena Scientific, NH (2003).

[7] P. Bianchi, S. Clémençon, J. Jakubowicz and G. Moral-Adell, On-Line Learning Gossip Algorithm in Multi-Agent Systems with Local Decision Rules, in *2013 IEEE International Conference on Big Data (BIG DATA)* (2014) 6–14.

[8] P. Bickel, C. Klaassen, Y. Ritov and J. Wellner, Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins University Press, Baltimore (1993).

[9] V. Borkar, Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press, Cambridge (2008).

[10] L. Bottou, Online Algorithms and Stochastic Approximations: Online Learning and Neural Networks. Cambridge University Press, Cambridge (1998).

[11] L. Bottou and O. Bousquet, The tradeoffs of large scale learning. *Adv. Neural Inf. Process. Syst.* **20** (2008) 161–168.

[12] S. Boucheron, O. Bousquet and G. Lugosi, Theory of classification: a survey of some recent advances. *ESAIM: PS* **9** (2005) 323–375.

[13] N. Breslow and J. Wellner, Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Stat.* **35** (2007) 186–192.

[14] N. Breslow and J. 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". *Scand. J. Stat.* **35** (2008) 186–192.

[15] N. 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. Biosci.* **1** (2009) 32–49.

[16] S. Clémençon, S. Robbiano and J. Tressou, Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling, in *Proceedings of the 2013 SIAM International Conference on Data Mining* (2013) 19–27.

[17] S. Clémençon, A. Bellet and I. Colin, Scaling-up empirical risk minimization: optimization of incomplete U-statistics. *J. Mach. Learn. Res.* **17** (2016) 1–36.

[18] W. Cochran, Sampling Techniques. Wiley, NY (1977).

[19] B. Delyon, Stochastic Approximation with Decreasing Gain: Convergence and Asymptotic Theory, 2000. Available at: http://perso.univ-rennes1.fr/bernard.delyon/.

[20] J. Deville, Réplications d'échantillons, demi-échantillons, Jackknife, bootstrap dans, Les Sondages, edited by J.-J. Droesbeke, Ph. Tassi, B. Fichet. Economica (1987).
[21] J. Deville and C. Särndal, Calibration estimators in survey sampling. *J. Acoust. Soc. Amer.* **87** (1992) 376–382.
[22] L. Devroye, L. Györfi and G. Lugosi, A Probabilistic Theory of Pattern Recognition. Springer, New York (1996).
[23] R. Gill, Y. Vardi and J. Wellner, Large sample theory of empirical distributions in biased sampling models. *Ann. Stat.* **16** (1988) 1069–1112.
[24] J. Hajek, Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Stat.* **35** (1964) 1491–1523.
[25] D. Horvitz and D. Thompson, A generalization of sampling without replacement from a finite universe. *J. Acoust. Soc. Amer.* **47** (1951) 663–685.
[26] V. Koltchinskii, Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *Ann. Stat.* **34** (2006) 2593–2706.
[27] H. Kushner and G. Yin, Stochastic Approximation and Recursive Algorithms and Applications. Springer, New York (2010).
[28] G. Mateos, J. Bazerque and G. Giannakis, Distributed sparse linear regression. *IEEE Trans. Signal Process.* **58** (2010) 5262–5276.
[29] A. Navia-Vazquez, D. Gutierrez-Gonzalez, E. Parrado-Hernandez and J. Navarro-Abellan, Distributed support vector machines. *IEEE Trans. Neural Netw.* **17** (2006) 1091–1097.
[30] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** (2009) 1574–1609.
[31] Y. Nesterov, Introductory lectures on convex optimization: a basic course, in Applied Optimization. Kluwer Academic Publ., Boston, Dordrecht, London (2004).
[32] M. Pelletier, Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Probab.* **8** (1998) 10–44.
[33] P. Robinson, On the convergence of the Horvitz-Thompson estimator. *Aust. J. Stat.* **24** (1982) 234–238.
[34] P. Rosen, Asymptotic theory for successive sampling. *AMS J.* **43** (1972) 373–397.
[35] T. Saegusa and J. Wellner, Weighted likelihood estimation under two-phase sampling. *Ann. Statist.* **41** (2013) 269–295.
[36] S. van de Geer, Empirical Processes in M-Estimation. Cambridge University Press, Cambridge (2000).
[37] A. Van der Vaart, Asymptotic Statistics. Vol. 3, Cambridge University Press, Cambridge (2000).