Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions

Alexandre Défossez and **Francis Bach**
Département d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS)

Abstract

We consider the least-squares regression problem and provide a detailed asymptotic analysis of the performance of averaged constant-step-size stochastic gradient descent. In the strongly-convex case, we provide an asymptotic expansion up to explicit exponentially decaying terms. Our analysis leads to new insights into stochastic approximation algorithms: (a) it gives a tighter bound on the allowed step-size; (b) the generalization error may be divided into a variance term which is decaying as $O(1/n)$, independently of the step-size γ, and a bias term that decays as $O(1/\gamma^2 n^2)$; (c) when allowing non-uniform sampling of examples over a dataset, the choice of a good sampling density depends on the trade-off between bias and variance: when the variance term dominates, optimal sampling densities do not lead to much gain, while when the bias term dominates, we can choose larger step-sizes that lead to significant improvements.

1 Introduction

For large-scale supervised machine learning problems, optimization methods based on stochastic gradient descent (SGD) lead to computationally efficient algorithms that make a single or few passes over the data (see, e.g., Bottou and Le Cun, 2005; Bousquet and Bottou, 2008).

In recent years, for smooth problems, large step-sizes together with some form of averaging, have emerged as having optimal scaling in terms of number of examples, both with asymptotic (Polyak and Juditsky, 1992) and non-asymptotic (Bach and Moulines, 2011) results. However, these convergence rates (i.e., bounds on generalization performance) in $O(1/n)$ are only optimal in the limit of large samples, and in practice where the asymptotic regime may not be reached, notably because of the high-dimensionality of the data, other non-dominant terms may come into play, which is the main issue we are tackling in this paper.

We consider least-squares regression using constant-step-size stochastic gradient descent, often referred to as least-mean-squares in the non-averaged case (Macchi, 1995; Bach and Moulines, 2013), where the generalization error may be explicitly split into a *bias* term that characterizes how fast initial conditions are forgotten, and a *variance* term that is only impacted by the noise present in the prediction problem. In this paper, we first show that while the variance term is asymptotically dominant, the bias term may play a strong role, both in theory and in practice, that explains convergence behaviors typically seen in applications. We should emphasize that here *bias* is meant as the error we would get in a noiseless setup rather than its usual statistical meaning.

Another question that has emerged as important to improve convergence is the use of special sampling distributions of examples (Nesterov, 2012; Needell et al., 2013; Zhao and Zhang, 2014). From our theoretical results, we can optimize the first-order asymptotic terms rather than traditional upper-bounds, casting a new light on the potential gains (or lack thereof) of such different sampling distributions.

More precisely, we make the following contributions:

- We provide in Section 2 a detailed asymptotic analysis of the performance of averaged constant-step-size SGD, with all terms up to exponentially decaying ones. We also give in Section 2.1 a tighter bound on the allowed step-size γ.
- In Section 2.4, the generalization error may be divided into a variance term which is (up to first order) decaying as $O(1/n)$, independently of the step-size γ, and a bias term that decays as $O(1/\gamma^2 n^2)$.

Appearing in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

- When allowing non-uniform sampling, the choice of a good sampling density depends on the trade-off between bias and variance: as shown in Section 3, when the variance term dominates, optimal sampling densities do not lead to much gain, while when the bias term dominates, we can choose larger step-sizes that lead to significant improvements.

1.1 Problem setup

Let X be a random variable with values in $\mathbb{R}^d$ and Y another random variable with values in $\mathbb{R}$. Throughout this paper, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. We assume that $\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[X^T X\right]$ is finite and we denote by $H = \mathbb{E}\left[XX^T\right] \in \mathbb{R}^{d\times d}$ the second-order moment matrix of X. Throughout the paper, we assume that H is invertible, or equivalently, in optimization terms, that we are in the strongly convex case (see, e.g., Nesterov, 2004). We denote by μ the smallest eigenvalue of H, so that we have $\mu > 0$. Note that in our asymptotic results, the leading terms do not depend explicitly on μ.

We wish to solve the following optimization problem:

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[\left\|X^T w - Y\right\|^2\right], \tag{1.1}$$

from a stream of independent and identically distributed samples $(X_i, Y_i)_{i\geqslant 0}$.

For any given $w \in \mathbb{R}^d$, we denote by $f(w) = \mathbb{E}\left[\left\|X^T w - Y\right\|^2\right]$ the expected generalization performance; we denote by $w^* \in \mathbb{R}^d$ the optimal solution (as H is invertible, it is unique), and by $f^* = f(w^*) \in \mathbb{R}$ the value at the minimum. This set-up covers two common situations:

(a) *Single pass through the data*, where each observation is seen once and considered as an i.i.d. sample, which is the context we explicitly study in this paper; note that then, our bounds are on the *testing error*, i.e., on the expected error on unseen data.

(b) *Multiple passes through a finite dataset*, where each sample (X_i, Y_i) is selected uniformly at random from the dataset; in this situation, the *training error* is explicitly minimized, a regularizer is often added and our bound corresponds to training errors. Moreover, dedicated algorithms (Schmidt et al., 2013; Shalev-Shwartz and Zhang, 2013) have then better convergence rates than stochastic gradient.

Averaged SGD with constant step-size. From a starting point $w_0 \in \mathbb{R}^d$, at each iteration $i \geq 1$, an i.i.d. sample of (X_i, Y_i) is obtained and the following recursion is used:

$$\begin{aligned} w_i &= w_{i-1} - \gamma X_i (X_i^T w_i - Y_i), \\ \bar{w}_i &= \frac{1}{i+1}\sum_{k=0}^{i} w_k = \frac{1}{i+1} w_i + \frac{i}{i+1}\bar{w}_{i-1}, \end{aligned}$$

where $\gamma > 0$ is a user-defined step-size. We denote by $\varepsilon_i = X_i^T w^* - Y_i$ the residual. Note that by definition of w^*, $\mathbb{E}[\varepsilon_i X_i] = 0$. If the vector X includes a constant component (which is common in practice), then this implies that ε_i and X_i are *uncorrelated*. However in general they are not *independent*, unless the model with independent homoscedastic noise is well-specified (which implies in particular that $E(Y_i|X_i) = X_i^T w^*$).

We denote by $f_i = \mathbb{E}[f(\bar{w}_i)]$ the expected (with respect to the randomness of the data) value of the generalization performance f at the averaged iterate $\bar{w}_i$ (with respect to time). It will be more convenient to work with the following centered estimates $\eta_i = w_i - w^*$ and $\bar{\eta}_i = \bar{w}_i - w^*$, for which one immediately gets

$$\eta_i = (I - \gamma X_i X_i^T)\eta_{i-1} + \gamma\varepsilon_i X_i,$$

which is the recursion that we study in this paper.

1.2 Related work

Stochastic gradient methods have been heavily studied. We mention in this section some of the works which are relevant for the present paper.

Analysis of stochastic gradient algorithms. Since the work of Nemirovski and Yudin (1983), it is known that the optimal convergence rate depends in general on the presence or absence of *strong convexity*, with rates of $O(1/n\mu)$ for μ-strongly convex problems, and $O(1/\sqrt{n})$ for non-strongly convex problems. Recently, for specific smooth situations with the square or logistic loss, these rates can be improved to $O(1/n)$ in both situations (Bach and Moulines, 2013). For least-squares, this is achieved with constant-step-size SGD, hence our main focus on this algorithm.

Asymptotic analysis of stochastic gradient descent. In this paper, we focus on finding asymptotic equivalents of the generalization errors of SGD algorithms (with explicit remainder terms). For decaying step-sizes and general loss functions, this was partially considered by Fabian (1968), who provides the limiting distribution of iterates (from which the generalization performance can be derived), but only without averaging. Moreover, the traditional analysis of Polyak-Rupert averaging (Polyak and Juditsky, 1992; Ruppert, 1988) also leads to asymptotic equivalents, also for decaying step-sizes, but only the (asymptotically dominant) variance terms are considered. See also the recent analysis of Toulis et al. (2014).

Non-uniform sampling. Non-uniform sampling, that is using another distribution than the one given by (X, Y), has been already tackled from several points of views; for example, in the active learning literature, Kanamori and Shimodaira (2003) provide the optimal sampling density to optimize the generalization error (for an estimator obtained as the minimum of the empirical least-squares risk), leading to distributions that are the same than the one obtained in Section 3.3 (where the variance term dominates), for which the actual gains are limited.

Moreover, in the context of stochastic gradient descent methods, Needell et al. (2013); Zhao and Zhang (2014) show that by optimizing the sampling density, bounds on the convergence rates could be improved, but the actual gains are hard to quantify. Our focus on limits of convergence rates allows us to precisely quantify the gains and obtain extra insights (at least asymptotically).

2 Main results

We will present results about the convergence of the algorithm which are derived from the exact computation of the second-order moment matrix, which we refer to as the *covariance matrix*, $\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right]$. Since we consider a least-squares problem, we have

$$f_n - f^* = \operatorname{Tr}\big(H\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right]\big).$$

We will distinguish two terms, which can be assimilated to a variance/bias decomposition. The *variance* term Δ^{variance} can be defined as the covariance matrix we would get starting from the solution (that is, $\eta_0 = 0$). On the other hand, the *bias* term Δ^{bias} is defined as the covariance matrix we would get if the model was noiseless, that is $Y = X^T w^*$ and $\varepsilon = 0$. It thus characterizes the rate at which the initial condition is forgotten.

Each of these two terms leads to contribution to $f_n - f^*$, that is $\operatorname{Tr}\big(H\Delta^{\text{variance}}\big)$ and $\operatorname{Tr}\big(H\Delta^{\text{bias}}\big)$. Under extra assumptions that are discussed in the supplementary material, such that when X and ε are independent (i.e., well-specified model), the actual covariance matrix is exactly the sum of the bias and variance matrices, and thus

$$f_n - f^* = \operatorname{Tr}\big(H\Delta^{\text{variance}}\big) + \operatorname{Tr}\big(H\Delta^{\text{bias}}\big).$$

Moreover, even when this is not true, it has been noted by Bach and Moulines (2013) that

$$f_n - f^* \leq 2\operatorname{Tr}\big(H\Delta^{\text{variance}}\big) + 2\operatorname{Tr}\big(H\Delta^{\text{bias}}\big),$$

that is, the sum of the two terms is a factor of two away from the exact generalization error.

2.1 Improved step-size

Let us denote by $\mathcal{M}_d$ the vector space of d-by-d real matrices. For any $\gamma \geq 0$, we can define a linear operator T defined by, $\forall M \in \mathcal{M}_d$:

$$TM = HM + MH - \gamma\mathbb{E}\left[(X^T M X)XX^T\right]. \quad (2.1)$$

The operator T can be seen as a symmetric $d^2 \times d^2$ matrix and is central in our analysis, like in the work of Murata (1998), who considered a simplified version of non-averaged stochastic gradient. We also define two contraction factors,

$$\rho_T = \|I - \gamma T\|_{\text{op}} \quad \text{and} \quad \rho_H = \|I - \gamma H\|_{\text{op}}, \quad (2.2)$$

as well as $\rho = \max(\rho_T, \rho_H)$, where $\|\cdot\|_{\text{op}}$ is defined as the largest singular value. Note that in traditional analyses of gradient methods, only ρ_H is considered.

Let us define $\gamma_{\max}$ as the supremum of the set of $\gamma > 0$ verifying, $\forall A \in \mathcal{S}(\mathbb{R}^d)$ (the set of symmetric matrices of size $d \times d$):

$$2\operatorname{Tr}\left(A^T H A\right) - \gamma\mathbb{E}\left[(X^T A X)^2\right] > 0, \quad (2.3)$$

or equivalently as the supremum of $\gamma > 0$ such that T is definite positive. One can actually show that we necessarily have that

$$\gamma_{\max} \leq 2/\operatorname{Tr}(H), \quad (2.4)$$

and the following lemma (see proof in the supplementary material):

Lemma 1. *If $0 < \gamma < \gamma_{\max}$ then $\rho < 1$ so that both $I - \gamma T$ and $I - \gamma H$ are contractive.*

Note that we may rewrite $\gamma_{\max}$ as

$$\frac{2}{\gamma_{\max}} = \sup_{A \in \mathcal{S}(\mathbb{R}^d)} \frac{\mathbb{E}\left[(X^T A X)^2\right]}{\operatorname{Tr}\left(A^T H A\right)},$$

which can be computed explicitly by a generalized eigenvalue problem once all second- and fourth-order moments of X are known. This is to be contrasted with the largest step-size $\gamma_{\max}^{\text{det}}$ for deterministic gradient descent, which is such that

$$\frac{2}{\gamma_{\max}^{\text{det}}} = \sup_{a \in \mathbb{R}^d} \frac{a^T H a}{a^T a}.$$

One can observe that for any distribution on X, we necessarily have $\gamma_{\max} \leq 2/\operatorname{Tr}(H) \leq \gamma_{\max}^{\text{det}}$ so that the maximum stochastic step-size will always be smaller than the deterministic, as one would expect.

Note also that the step-size provided by $\gamma_{\max}$ is a strict improvement (see supplementary material) on the one proposed by Bach and Moulines (2013), which is equal to the supremum of the set of $\gamma > 0$ such that

$\mathbb{E}\left[XX^T\right] - \gamma\mathbb{E}\left[(X^TX)XX^T\right]$ is positive semidefinite. Our results also extend the result of Slock (1993), who considered a very special set-up for inputs while ours only relies only on fourth-order moments. We conjecture that the bound given by $\gamma_{\max}$ is tight, namely that if γ is larger than $\gamma_{\max}$ then there exists an initial condition η_0 such that the algorithm diverges.

2.2 Bias term

In this section, we provide an asymptotic expansion of the bias term Δ^{bias}. We introduce $A(\gamma)$ some linear operator over $\mathcal{M}_d$ that is given in the supplementary material, with an explicit remainder term.

Theorem 1 (Asymptotic covariance for noiseless problem). *Let $E_0 = \eta_0\eta_0^T$. If $0 < \gamma < \gamma_{\max}$ and $\forall i \geq 1, \varepsilon_i = 0$, then*

$$\Delta^{\text{bias}} = \mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] = \frac{1}{n^2\gamma^2}A(\gamma)E_0 + O\Big(\frac{\rho^n}{n}\Big). \quad (2.5)$$

Using Lemma 1, we know that $\rho < 1$ so that (2.5) converges as n^{-2}. We can thus derive the rate of convergence for $\operatorname{Tr}\big(H\Delta^{\text{bias}}\big)$, that will be of order n^{-2} as well. Although the dependency of $A(\gamma)$ is complex, one can easily derive an equivalent when γ tends to zero, and we have

$$\lim_{n\to\infty} n^2\operatorname{Tr}\big(H\Delta^{\text{bias}}\big) \underset{\gamma\to 0}{\sim} \gamma^{-2}\eta_0^T H^{-1}\eta_0. \quad (2.6)$$

2.3 Variance term

In this section, we provide an asymptotic expansion of the variance term Δ^{variance}. We introduce $B(\gamma)$ and $C(\gamma)$ some linear operators over $\mathcal{M}_d$ explicitly computed in the supplementary material.

Theorem 2 (Asymptotic covariance for problem started at the optimum). *Let $\Sigma_0 = \mathbb{E}\left[\varepsilon^2XX^T\right]$ and let assume that $\eta_0 = 0$. If $0 < \gamma < \gamma_{\max}$ then Δ^{variance} is equal to:*

$$\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] = \frac{1}{n}B(\gamma)\Sigma_0 - \frac{1}{\gamma n^2}C(\gamma)\Sigma_0 + O\Big(\frac{\rho^n}{n}\Big). \quad (2.7)$$

Unsurprisingly, the asymptotic behavior of the variance term is dominant over the bias one as it decreases only as n^{-1}, which is the overall convergence rate of this algorithm of least-mean-squares as noted by Bach and Moulines (2013).

It is also possible to get a simpler equivalent when γ goes to 0:

$$\lim_{n\to\infty} n\operatorname{Tr}\big(H\Delta^{\text{variance}}\big) \underset{\gamma\to 0}{\sim} \mathbb{E}\left[\varepsilon^2X^TH^{-1}X\right].$$

If we further assume that the noise ε is independent of X, then we recover the usual

$$\lim_{n\to\infty} n\operatorname{Tr}\big(H\Delta^{\text{variance}}\big) \underset{\gamma\sim 0}{\sim} d\sigma^2,$$

where $\sigma^2 = \mathbb{E}\left[\varepsilon^2\right]$, which is the Cramer-Rao bound for such a problem (obtained from computing the generalization performance, which depends only on the usual estimation covariance matrix). It is also interesting to notice that this is the exact same result as the one obtained by Polyak and Juditsky (1992) with a decreasing step-size.

Finally, note that if γ is small, the term in $\frac{1}{\gamma n^2}C(\gamma)\Sigma_0$ is always positive so that there is no risk of it exploding for small values of γ (unlike for the bias term).

2.4 Comparing both terms

As seen above with an asymptotic expansion around $\gamma = 0$, for n sufficiently large, the bias and variance terms are of order:

$$\begin{aligned}\operatorname{Tr}\big(H\Delta^{\text{bias}}\big) &\sim \frac{1}{\gamma^2n^2}\eta_0^TH^{-1}\eta_0\\ \operatorname{Tr}\big(H\Delta^{\text{variance}}\big) &\sim \frac{1}{n}\mathbb{E}\left[\varepsilon^2X^TH^{-1}X\right].\end{aligned}$$

The different behaviors of the bias and variance terms lead to two regimes, one in $(\gamma n)^{-2}$ and one in n^{-1} that can clearly be observed on synthetic data. On real world data, one will often observe a mixture of the two, depending on the step-size and the difficulty of the problem. Experimental results on both synthetic and real world data will be presented in Section 4.

3 Optimal sampling

Changing the sampling density for (X, Y) may be interesting in several situations, in particular (a) in presence of outliers (i.e., points with large norms) and (b) classification problems with asymmetric costs.

3.1 Impact of sampling

Using the two previous theorems, we can now try to optimize the sampling distribution to increase performance. We will sample from a distribution q instead of the given distribution p. Since we wish to keep the same objective function, we will use importance weights $c(X, Y)$, so that if we denote by $\mathbb{E}_p[A]$ the expectation of a random variable A under the probability distribution given by p over A we have

$$\mathbb{E}_p\left[|X^Tw - Y|^2\right] = \mathbb{E}_q\left[c(X,Y)|X^Tw - Y|^2\right].$$

First, one can notice that, from a practical point of view, we must restrict ourselves to q that are absolutely continuous with respect to p as one cannot invent samples. In order to be able to define c we also need p to be absolutely continuous with respect to q, so that $c = \frac{\mathrm{d}p}{\mathrm{d}q}$. Besides, c^{-1} is defined as $c^{-1} = \frac{\mathrm{d}q}{\mathrm{d}p}$.

A key consequence of using least-squares is that for a given (q, c) pair, we only have to sample using q

and scale X, Y and $\varepsilon = X^T w^* - Y$ by $\sqrt{c(X,Y)}$. Thus we can use the two previous theorems for $X' = \sqrt{c(X,Y)}X$ and $Y' = \sqrt{c(X,Y)}Y$ and sampling X, Y according to q.

As of now, we will assume that almost surely $X \neq 0$. Indeed, when $X_i = 0$, we perform no update so that we can just ignore such points.

One can notice that for any $A, B \in \{X, Y, \varepsilon\}$, taking $A' = \sqrt{c(X,Y)}A$ (same for B), $\mathbb{E}_q[A'B'] = \mathbb{E}_q[c(X,Y)AB] = \mathbb{E}_p[AB]$, and thus all second-order moments are unchanged under resampling. This is the case for the matrix $H = \mathbb{E}_p[XX^T] = \mathbb{E}_q[X'X'^T]$.

However, for terms of order 4 like T, $\mathbb{E}[(X^TX)XX^T]$ or Σ_0, an extra c appears and we have for instance

$$\mathbb{E}_q\left[(X'^TX')X'X'^T\right] = \mathbb{E}_p\left[\frac{p(X,Y)}{q(X,Y)}(X^TX)XX^T\right].$$

It means that while H will not be changed, T is impacted in non trivial ways, as is T^{-1}. This makes it tricky to truely optimize sampling for any γ. However, when assuming γ small, it is possible to optimize the limit we obtained for $\gamma \to 0$ (see Sections 3.3 and 3.4). The experiments we ran (see Section 4) seem to confirm that this is a valid assumption for values of γ as high as $\gamma_{\max}/2$.

3.2 Asymmetric binary classification

As a motivation for this work, we will present one practical application of resampling which is binary classification with highly asymmetric classes. Assume we have $Y \in \{-1, 1\}$ and that $\mathbb{P}\{Y = 1\}$ and $\mathbb{P}\{Y = -1\}$ are highly unbalanced, as it can be the case in various domains, such as ad click prediction or object detection, etc. Then, it can be useful in practice to give more weight the less frequent class (see, e.g., Perronnin et al., 2012, and references therein). This is equivalent to multiplying both X and Y by some constant $\sqrt{c_Y}$. A common choice is for instance to take $c_y = 1/\mathbb{P}\{Y = y\}$ which will give the same importance in the loss to both classes.

However, these weights will make the gradients from the less frequent class huge compared to the usual updates. This is likely to impact the convergence of the algorithm. In that case it is easy to notice that taking taking $c(x,y) = 1/c_y$ will leave the gradients unchanged but will favor sampling examples from the less frequent class.

3.3 Optimal sampling for the variance term

Let assume that we are only interested in the long term performance for our algorithm. Ultimately the variance term will be driving the performance and we need to optimize it.

Exactly optimizing the sampling for this case in uneasy as it impacts both Σ_0 and the terms $B(\gamma)$ and $C(\gamma)$ in Theorem 2, in a non trivial way. However, if we assume a small step-size γ, then we just have to minimize

$$\mathbb{E}_q\left[\varepsilon'^2 X'^T H^{-1} X'\right] = \mathbb{E}_p\left[c(X,Y)\varepsilon^2 X^T H^{-1} X\right],$$

under the constraint that $\mathbb{E}_p\left[c^{-1}(X,Y)\right] = 1$ so that q is a distribution. Using the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\mathbb{E}_p&\left[c(X,Y)\varepsilon^2 X^T H^{-1} X\right]\\ &= \mathbb{E}_p\left[c(X,Y)\varepsilon^2 X^T H^{-1} X\right]\mathbb{E}_p\left[c^{-1}(X,Y)\right]\\ &\geq \left(\mathbb{E}_p\left[|\varepsilon|\sqrt{X^T H^{-1} X}\right]\right)^2.\end{aligned}$$

When $X \neq 0$ almost surely, then this lower-bound is achieved for

$$c^{-1}(X,Y) = \frac{|\varepsilon|\sqrt{X^T H^{-1} X}}{\mathbb{E}_p\left[|\varepsilon|\sqrt{X^T H^{-1} X}\right]}.$$

Prior knowledge of H and ε is required, or just H when the noise is independent, which can be impractical.

In that case, we obtain

$$\lim_{n\to\infty} n\mathrm{Tr}\big(H\Delta^{\text{variance}}\big) = \left(\mathbb{E}\left[|\varepsilon|\sqrt{X^T H^{-1} X}\right]\right)^2.$$

One can notice that this is the exact same optimal sampling as the one obtained in the active learning set-up by Kanamori and Shimodaira (2003).

Again, it is possible to slightly simplify this expression when ε and X are independent, as we obtain

$$\lim_{n\to\infty} n\mathrm{Tr}\big(H\Delta^{\text{variance}}\big) = \sigma^2\left(\mathbb{E}\left[\sqrt{X^T H^{-1} X}\right]\right)^2,$$

with $\sigma^2 = \mathbb{E}\left[\varepsilon^2\right]$. At this point it is important to realize that the gain we have here is of the order of

$$\mathbb{E}\left[\sqrt{X^TX}\right]^2 / \ \mathbb{E}\left[X^TX\right]. \qquad (3.1)$$

During our experimentations on usual datasets, we have observed that this factor was always between 1/2 and 1 and thus there is little to be gained when optimizing the variance term.

3.4 Optimal sampling for the bias term

Although asymptotically the variance term will be the largest one, it is possible that initially the bias one is non negligible and it can be interesting to optimize for it. This is all the more possible as it depends much more on the step-size γ and if γ is too small, the bias term can stay larger than the variance term for many iterations.

If we assume γ small, then we can approximate the bias term by the expression given by (2.6), that is, proportional to $1/(\gamma^2 n^2)$. In this case, it is clear that we want to increase $\gamma_{\max}$ and, because second-order moments are not impacted by resampling, it has no effect other than changing $\gamma_{\max}$. Numerical experiments tends to show that increasing γ is beneficial even for γ close to $\frac{\gamma_{\max}}{2}$. Beyond this limit, the approximation (2.6) is no longer sustainable and besides, exponentially decreasing terms can start to grow quite large.

The maximum step-size is given by the tighter condition from Section 2.1, that is, $\forall A \in \mathcal{S}(\mathbb{R}^d)$,

$$2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathbb{E}\left[(X^T A X)^2\right] > 0, \qquad (3.2)$$

which implies that

$$\gamma_{\max} \leq 2/\mathbb{E}\left[X^T X\right], \qquad (3.3)$$

using (2.4). As this upper bound on $\gamma_{\max}$ only depends on moments of order 2, and that those moments are not changed by resampling, (3.3) is an upper bound on any $\gamma_{\max}$ for a given optimization problem, no matter how we resample. It turns out it can be achieved by the resampling given by

$$c_*^{-1}(X, Y) = X^T X / \ \mathbb{E}_p\left[X^T X\right] \text{ a.s},$$

which, unlike the variance term, does not require the knowledge of H. To show that, one can first prove that $H - \gamma \mathbb{E}\left[(X^T X) X X^T\right] \succ 0$ implies (3.2), as we already noted in Section 2.1. Then, computing $H - \gamma \mathbb{E}\left[(X'^T X') X' X'^T\right]$, one obtains

$$\begin{aligned} & H - \gamma \mathbb{E}_q\left[c_*(X,Y)^2 (X^T X) X X^T\right] \\ & = 2H - \gamma \mathbb{E}_p\left[\frac{\mathbb{E}_p\left[X^T X\right]}{X^T X}(X^T X) X X^T\right] \\ & = H(2 - \gamma \mathbb{E}_p\left[X^T X\right]), \end{aligned}$$

which is positive definite as soon as $\gamma < \frac{2}{\mathrm{Tr}(H)}$. This means that using the resampling defined by c_*, we have $\gamma_{\max} = \frac{2}{\mathbb{E}[X^T X]}$ which is thus not improvable.

If $\gamma_{\max}^{(0)}$ is the maximum step-size before resampling and $\gamma_{\max}^{(1)}$ is the maximum step-size after resampling, then the gain for $f_n - f^*$ is a factor $\left(\gamma_{\max}^{(0)} / \gamma_{\max}^{(1)}\right)^2$. It is computationally expensive to evaluate for large problems, but from our experiments where we take the largest non-diverging step-sizes (see Section 4) it was common to observe gain factors of $1/100$ or $1/400$ while the gain for the variance term was limited to $1/2$.

Unlike the variance term, the resampling in itself here has an impact only through a larger step-size. Resampling while keeping the same step-size will often lead to almost identical performances for the bias term. It is interesting to note that when $H = I$, this sampling will exactly have no impact at all on the variance term, and when $H \neq I$, it will only impact it marginally.

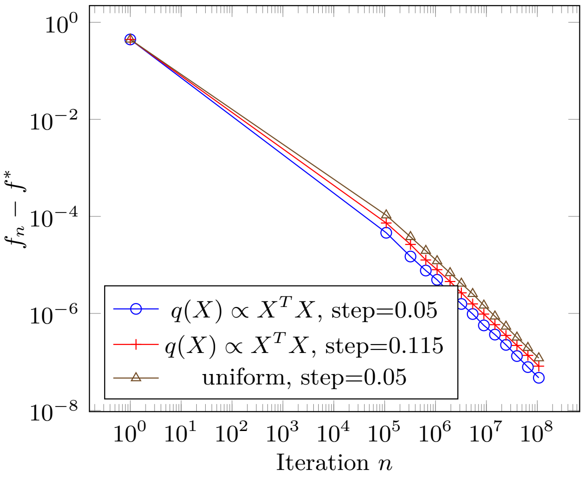

Figure 1: Convergence on *Yahoo* dataset without weights.

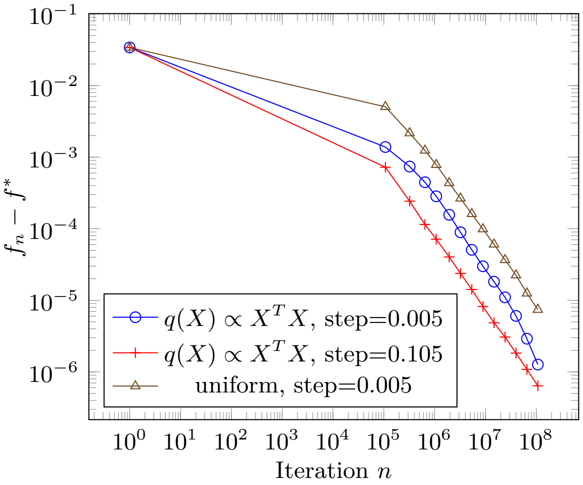

Figure 2: Convergence on *Yahoo* dataset with weights.

Link with other algorithms. When using this resampling and $\gamma = 1/\mathbb{E}\left[X^T X\right]$, the update becomes

$$w_i = w_{i-1} - \frac{X_i}{X_i^T X_i}\left(X_i^T w_{i-1} - Y_i\right), \qquad (3.4)$$

where X_i is sampled from q_*. This is very similar to normalized least mean squares (NLMS) by Bershad (1986), i.e., we first normalize X (and Y by the same factor), and then we run the usual stochastic gradient descent with a step-size of 1. However, while NLMS does not optimize the same overall objective function, we remember the norm of X in c_* and sample large ones more often and keep the same overall objective function. One can also notice some links

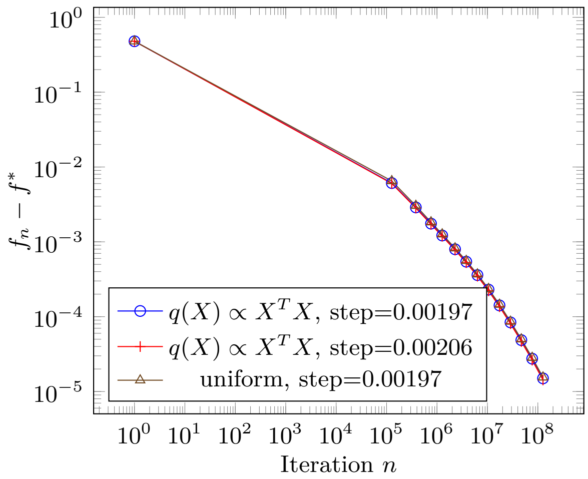

Figure 3: Convergence on *Sido* dataset without weights.

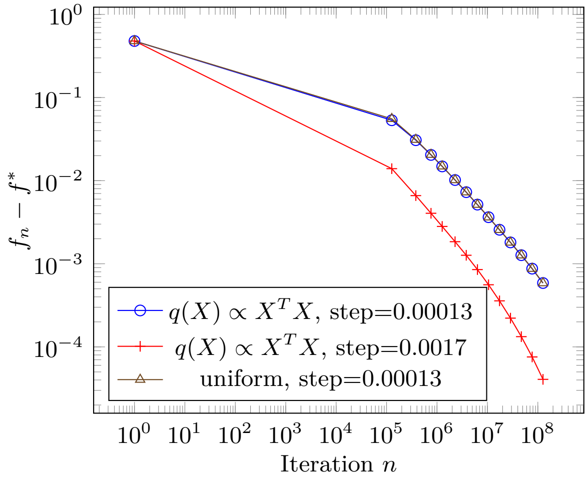

Figure 4: Convergence on *Sido* dataset with weights.

with implicit stochastic gradient descent (ISGD) by Toulis et al. (2014), where the update rule is

$$w_i = w_{i-1} - \frac{\gamma_i X_i}{1 + \gamma_i X_i^T X_i} \left(X_i^T w_{i-1} - Y_i \right), \qquad (3.5)$$

which is similar to NLMS and (3.4) when γ_i is large. As γ_i is a decreasing step-size for ISGD, during the early iterations, ISGD will behave like NLMS before switching to a regular stochastic gradient descent as γ_i goes to 0 (thus allowing convergence to the global optimum). However, the decaying step-size will prevent the fast decay of the bias terms.

4 Experiments

***Yahoo* and *Sido* datasets.** We have tried to observe evidence of our predictions into two unbalanced datasets, "yahoo" and "sido". The "yahoo" dataset[1] is composed of millions of triple (ads, context, click) from the Yahoo front page where click is 1 if the user clicked on the given ad in the given context (composed of 136 Boolean features) and 0 otherwise. It is very unbalanced, as the click rate is low. For this experiment we only looked at the rows corresponding to a specific ad (107355 rows), with a click rate of 0.03%. We used it both with and without weights in order to give the same importance to both clicks and non clicks as explained in Section 3.2. In all our experiments, we report the average performance over 50 runs (error bars are not plotted as the relative error is always less than 0.2%).

On Figures 1 and 2, we compared the incidence of the step-size and the resampling of performance. When using weights, the maximum step-size happens to be divided by ten, but when resampling proportionally to $X^T X$, we recover almost the same step-size as without weights. Moreover, without weights (Figure 1), the lowest step-size performs best, which tends to indicate that the variance term is dominant. However, with weights (Figure 2), one can observe the more pronounced dependency in γ, which show that the bias term became non negligible. Comparable results, but obtained on the "sido" dataset[2] are given in Figures 3 and 4. This dataset is composed of 12678 points with 4932 features. The less frequent class represents 3.6% of the points.

We did not plot the graphs obtained when sampling proportionally to $\sqrt{X^T H^{-1} X}$ as they are mostly the same as without resampling. If we were to achieve the regime where the variance is completely dominant, the error would be at best be divided by two for the "yahoo" dataset and would be almost the same for "sido" (divided by 1.02). These potential gains were computed directly using the expression (3.1).

Synthetic data and bias-variance decomposition. We also observed exactly the bias and variance terms on synthetic data. The data consist in an infinite stream of points X sampled from a normal distribution with covariance matrix H, so that the eigenvalues of H are $(\frac{1}{i})_{1 \leq i \leq 25}$. Y is given by $Y = X^T w_* + \varepsilon$ for a fixed $w^* \in \mathbb{R}^{25}$ and $\varepsilon \sim \mathcal{N}(0, 1)$. On Figures 5 and 6 one can see the decomposition of the error between variance and bias terms (Figure 5) as well as the sum of both the variance and bias error (Figure 6). We see that both the variance and bias curves quickly reach their asymptotic regime with a slope of -1 for the variance in log/log space and -2 for the bias as expected. We ran the algorithm with two step-sizes, one being ten times larger than the other. We observe the ratio of 100 as expected between the two bias curves and al-

[1] `webscope.sandbox.yahoo.com/`

[2] `www.causality.inf.ethz.ch/data/SIDO.html`

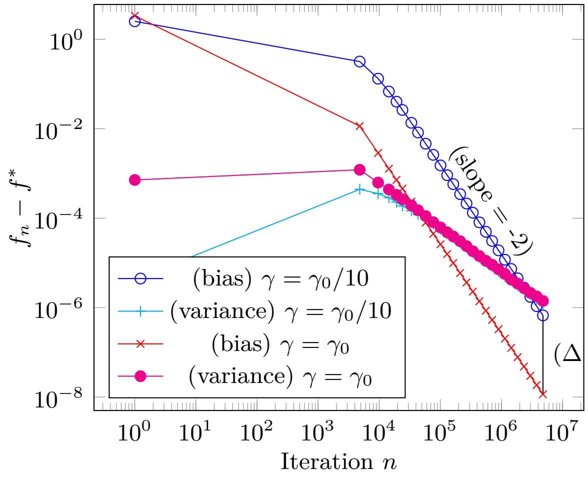

Figure 5: Convergence per term on synthetic data.

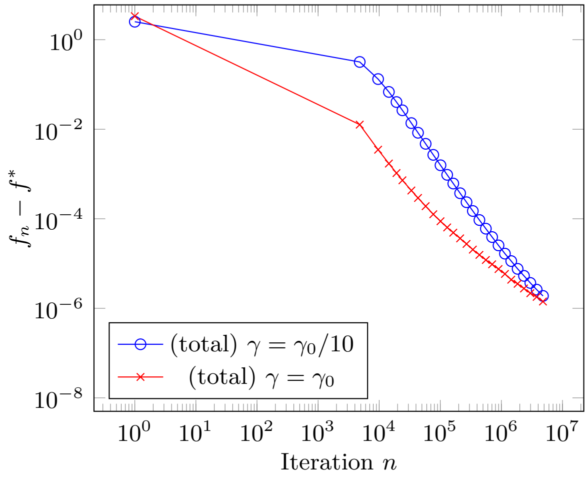

Figure 6: Convergence on synthetic data.

most no difference at all for the variance ones, except for the first iterations.

One can also see the effect of the step-size γ at a fixed number of iterations on Figure 7. Due to the symmetry between n and γ in the expression of the bias term, one can notice the resemblance at first between this curve and the one obtained when plotting $f_n - f^*$ against n. However, when the step-size is large, we get sooner into the regime where the variance dominates. At this point we observe almost no influence of the step-size on $f_n - f^*$. When getting closer to the maximum step-size, convergence becomes very slow as ρ becomes close to 1. The variance term increases at first with γ because for $\gamma = 0$, one has by definition $f_n = f_*$. The predicted flat part is only achieved for γ sufficiently large so that the exponentially decaying terms are negligible. For all the experiments on synthetic data, we used 300 independent runs to obtain statistically significant results.

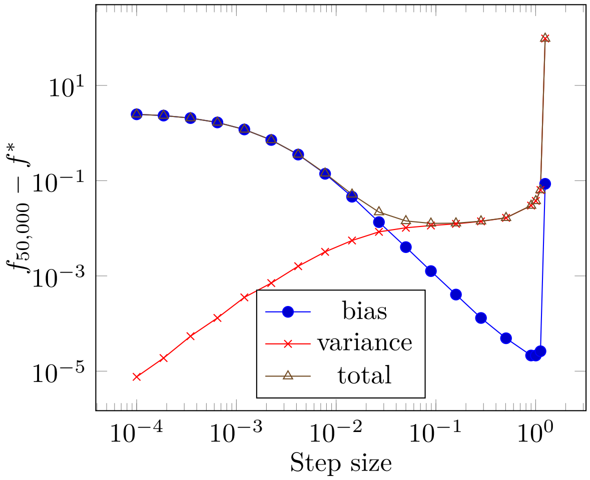

Figure 7: Impact of step-size on error, with its bias/variance decomposition.

5 Conclusion

In this paper, we have provided a tighter analysis of averaged constant-step-size SGD for least-squares, leading to a better understanding of the convergence of the algorithm at different stages, in particular regarding how the initial condition is forgotten.

We were able to deduce different sampling schemes depending on what regime we are in. Sampling proportionally to $\sqrt{X^T H^{-1} X}$ is always asymptotically the best method. The potential gain is however limited most of the time. Besides, for datasets that are more "difficult", that is with moments that increases quickly, forgetting the initial condition can happen arbitrary slow due to the strong dependency in the step-size. If this is the case, then sampling proportionally to $X^T X$ will allow us to take a much larger step-size which will then lead to a smaller error.

Our work can be extended in several ways: for simplicity we have focused on least-squares problems where the bias/variance decomposition is explicit. It would be interesting to see how these results can be extended to other smooth losses such as logistic regression (with the proper use of Fisher information matrices), where constant-step size SGD does not converge to the global optimum (Nedic and Bertsekas, 2000; Bach and Moulines, 2013). Moreover, we have only provided results in expectations and a precise study of higher-order moments would give a better understanding of additional potential effects of resampling.

Acknowledgements. This work was partially supported by a grant from the European Research Council (SIERRA project 239993). We thank Aymeric Dieuleveut and Nicolas Flammarion for interesting discussions related to this work.

References

Bach, F. and E. Moulines (2011). Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Adv. NIPS*.

Bach, F. and E. Moulines (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Adv. NIPS*.

Bershad, N. (1986). Analysis of the normalized LMS algorithm with gaussian inputs. *Speech and Signal Processing, IEEE Transactions on Acoustics 34*(4), 793–806.

Bottou, L. and Y. Le Cun (2005). On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry 21*(2), 137–151.

Bousquet, O. and L. Bottou (2008). The tradeoffs of large scale learning. In *Adv. NIPS*.

Fabian, V. (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics 39*(4), 1327–1332.

Kanamori, T. and H. Shimodaira (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of statistical planning and inference 116*(1), 149–162.

Macchi, O. (1995). *Adaptive processing: The least mean squares approach with applications in transmission.* Wiley West Sussex.

Murata, N. (1998). A statistical study of on-line learning. In *Online Learning and Neural Networks*. Cambridge University Press.

Nedic, A. and D. Bertsekas (2000). Convergence rate of incremental subgradient algorithms. *Stochastic Optimization: Algorithms and Applications*, 263–304.

Needell, D., N. Srebro, and R. Ward (2013). Stochastic gradient descent and the randomized kaczmarz algorithm. Technical Report 1310.5715, arXiv.

Nemirovski, A. S. and D. B. Yudin (1983). *Problem complexity and method efficiency in optimization.* Wiley & Sons.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: a Basic Course.* Kluwer Academic Publishers.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization 22*(2), 341–362.

Perronnin, F., Z. Akata, Z. Harchaoui, and C. Schmid (2012). Towards good practice in large-scale learning for image classification. In *Proc. CVPR*.

Polyak, B. T. and A. B. Juditsky (1992, July). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim. 30*(4), 838–855.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Schmidt, M., N. L. Roux, and F. Bach (2013). Minimizing finite sums with the stochastic average gradient. Technical Report 00860051, HAL.

Shalev-Shwartz, S. and T. Zhang (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *JMLR 14*, 567—599.

Slock, D. T. M. (1993). On the convergence behavior of the LMS and the normalized LMS algorithms. *IEEE Transactions on Signal Processing 41*(9), 2811–2825.

Toulis, P., J. Rennie, and A. M. Airoldi (2014). Statistical analysis of stochastic gradient methods for generalized linear models. In *Proc. ICML*.

Zhao, P. and T. Zhang (2014). Stochastic optimization with importance sampling. Technical Report 1401.2753, arXiv.