Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure

Alberto Bietti, Julien Mairal

Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure*

Alberto Bietti
Inria
alberto.bietti@inria.fr

Julien Mairal
Inria
julien.mairal@inria.fr

October 3, 2016

Abstract

Stochastic optimization algorithms with variance reduction have proven successful for minimizing large finite sums of functions. However, in the context of empirical risk minimization, it is often helpful to augment the training set by considering random perturbations of input examples. In this case, the objective is no longer a finite sum, and the main candidate for optimization is the stochastic gradient descent method (SGD). In this paper, we introduce a variance reduction approach for this setting when the objective is strongly convex. After an initial linearly convergent phase, the algorithm achieves a $O(1/t)$ convergence rate in expectation like SGD, but with a constant factor that is typically much smaller, depending on the variance of gradient estimates due to perturbations on a *single* example.

1 Introduction

Many supervised machine learning problems can be cast into the problem of minimizing an expected loss over a data distribution $\mathcal{D}$ with respect to a vector x in $\mathbb{R}^p$ of model parameters: $\mathbb{E}_{\zeta\sim\mathcal{D}}[f(x,\zeta)]$. When an infinite amount of data is available, stochastic optimization methods such as the stochastic gradient descent (SGD) or stochastic mirror descent algorithms, or their variants, are typically used [3, 15]. However, in the case of finite datasets, incremental methods based on variance reduction techniques (e.g., [6, 9, 10, 17, 19]) have proven to be very successful at solving the finite-sum problem

$$\min_{x\in\mathbb{R}^p}\left\{f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)\right\}.$$

A classical setting is $f_i(x) = \ell(y_i, x^\top\xi_i) + \mu/2\|x\|^2$, where (ξ_i, y_i) is an example-label pair, ℓ is a convex loss function, and μ is a regularization parameter. However, in many situations, augmenting the finite training set with well-chosen random perturbations of each example can lead to smaller test error in theory [23] and in practice [20]. Examples of such procedures include random transformations of images in classification problems (e.g., [20]), and Dropout [21]. The objective describing these scenarios, which we consider in this paper, is the following:

$$\min_{x\in\mathbb{R}^p}\left\{f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho\sim\Gamma}[\tilde{f}_i(x,\rho)]\right\}, \tag{1}$$

where ρ parametrizes the random perturbation and $\tilde{f}_i(\cdot,\rho)$ is a convex smooth function with L-Lipschitz continuous gradients for all i and ρ. We also assume that f is μ-strongly convex. Because each function f_i is

*This work was supported by a grant from ANR (MACARON project under grant number ANR-14-CE23-0003-01) and from the MSR-Inria joint centre.

an expectation, computing a single gradient ∇f_i is intractable in general, and standard variance reduction methods such as SAG/SVRG/SDCA cannot be used. A straightforward way to optimize this objective is to use SGD by choosing an index i_t randomly in $\{1, \dots, n\}$ at iteration t, sampling a perturbation $\rho_t \sim \Gamma$, and performing the update $x_t = x_{t-1} - \eta_t \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)$, where η_t is a step-size. Note that this approach ignores the finite-sum structure in the objective and thus leads to gradient estimates with high variance. The goal of this paper is to introduce an algorithm, *stochastic MISO*, which can exploit the problem structure using variance reduction. Our method maintains a $O(1/t)$ convergence rate like SGD, but with a constant term that is much smaller in typical settings, only depending on the variance of the gradient estimates due to the random perturbations on a single example.

We present our algorithm in Section 2 and study its convergence properties in Section 3. In Section 4, we present empirical results of our method and compare it with SGD on two different problems: image classification with random transformations of the input examples and binary classification of high-dimensional gene expression data with Dropout, demonstrating the applicability of our approach to two significantly different contexts.

Related work. Our work is inspired by the recent surge of interest in stochastic optimization methods for minimizing finite sums, which arise naturally in machine learning for the minimization of empirical risk. Surprisingly, it has been shown that by exploiting the finite-sum structure in the objective, one can develop much faster optimization methods—i.e., with lower (expected) computational complexity—than previous ones that did not exploit this structure, such as SGD or full gradient descent (see, e.g., [6, 9, 10, 17, 19]). This is particularly the case when n is large, a common scenario in the current "big data" setting.

Many of these methods such as SVRG [9] and SAGA [6] have been motivated by the fact that their updates can be interpreted as stochastic gradient descent steps with unbiased estimates of the full gradient, but with a variance that decreases as the algorithm approaches the optimum (hence the name "variance reduction"); on the other hand, SGD requires decreasing step-sizes to achieve this reduction of variance, thereby slowing down convergence. Our work aims at extending these techniques, in particular the MISO/Finito algorithm [7, 10, 12] to the case where each function in the finite sum can only be accessed via a first-order stochastic oracle.

The problem of minimizing (1) is not well studied in the optimization and machine learning literature. Most related to our work, recent methods that use clustering information to improve the convergence of variance reduction techniques [2, 8] can be seen as tackling a special case of the objective (1), where the expectations in f_i are replaced by empirical averages over the points in a cluster. While the approximation assumption of SAGA with neighbors [8] can be seen as a variance condition on stochastic gradients as in our case, their algorithm is asymptotically biased and does not converge to the optimum. On the other hand, ClusterSVRG [2] is not biased, but requires a finite-sum structure and hence does not support infinite datasets. The method proposed in [1] also bears similarity with ours, since it uses variance reduction in a setting where gradients are computed approximately, but the algorithm requires reducing the approximation variance by dynamically increasing the number of MCMC samples used in order to reach the optimum, while our algorithm overcomes this requirement by supporting decreasing step-sizes.

2 The Stochastic MISO Algorithm

In this section, we introduce the *stochastic MISO* approach, which relies on the following assumptions:

- (A1) **global strong convexity**: f is μ-strongly convex;
- (A2) **smoothness**: $\tilde{f}_i(\cdot, \rho)$ is L-smooth for all i and ρ (i.e., differentiable with L-Lipschitz gradients);
- (A3) **small variance from perturbations at optimum**:

$$\mathbb{E}_\rho \left[\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2 \right] \leq \sigma^2,$$

  for all i, where x^* is the (unique) minimizer of f.

Algorithm 1 S-MISO

Input: step-size sequence $(\alpha_t)_{t\geq 1}$;
Initialize $x_0 = \frac{1}{n}\sum_i z_i^0$ for some $(z_i^0)_{i=1,\dots,n}$;
for $t = 1, \dots$ **do**
Sample an index i_t randomly in $\{1, \dots, n\}$, a perturbation $\rho_t \sim \Gamma$, and update:

$$z_i^t = \begin{cases} (1-\alpha_t) z_i^{t-1} + \alpha_t (x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_i(x_{t-1}, \rho_t)), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise.} \end{cases} \tag{2}$$

$$x_t = \frac{1}{n}\sum_{i=1}^n z_i^t. \tag{3}$$

end for

It is important to note that our variance assumption (A3) is only affected by the noise induced by the perturbations ρ and not by the randomness in the choice of the index i. In contrast, a standard assumption for the SGD algorithm on the objective (1) would take the form $\mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_i(x,\rho)\|^2] \leq M^2$ for all x. The quantity M^2 takes into account the noise induced by the random index i in addition to ρ, and can thus be much larger than σ^2, particularly if the perturbations on input data are small. We will show in Section 3 that after an initial linearly convergent phase, and under appropriate choice of step-sizes $(\alpha_t)_{t\geq 1}$, S-MISO will satisfy $\mathbb{E}[f(x_t) - f(x^*)] \leq \epsilon$ after

$$O\left(\frac{L\sigma^2}{\mu^2 \epsilon}\right)$$

iterations. This complexity is similar to that of SGD [3, 15], but with σ^2 replacing the quantity M^2, leading to a much faster rate than SGD if $\sigma^2 \ll M^2$, something which we observed clearly in our experiments.

Our method is given in Algorithm 1. Without the perturbations and with a constant step-size, the algorithm resembles the MISO/Finito algorithms [7, 10, 12] which may be seen as primal variants of SDCA [18, 19]. MISO/Finito are part of a larger body of optimization methods that iteratively build a *model* of the objective function, typically in the form of a lower or upper bound on the objective that is easier to optimize, such as bundle methods or the EM algorithm and its incremental variants [14]. Specifically, MISO/Finito assumes that each f_i is strongly convex, and builds a model of the objective using lower bounds of the form $D_t(x) = \frac{1}{n}\sum_{i=1}^n d_i^t(x)$, where each d_i^t is a lower bound on f_i and takes the form

$$d_i^t(x) = c_{i,1}^t + \frac{\mu}{2}\|x - z_i^t\|^2 = c_{i,2}^t - \mu\langle x, z_i^t\rangle + \frac{\mu}{2}\|x\|^2.$$

These lower bounds are updated during the algorithm using strong convexity lower bounds at the current iterate:

$$d_i^t(x) = \begin{cases} (1-\alpha) d_i^{t-1}(x) + \alpha(f_i(x_{t-1}) + \langle \nabla f_i(x_{t-1}), x - x_{t-1}\rangle + \frac{\mu}{2}\|x - x_{t-1}\|^2), & \text{if } i = i_t \\ d_i^{t-1}(x), & \text{otherwise,} \end{cases} \tag{4}$$

which corresponds to updating z_i^t as

$$z_i^t = \begin{cases} (1-\alpha) z_i^{t-1} + \alpha\left(x_{t-1} - \frac{1}{\mu}\nabla f_i(x_{t-1})\right), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise.} \end{cases}$$

The next iterate is then computed as $x_t = \arg\min_x D_t(x)$, which is equivalent to (3). The original MISO/Finito algorithm uses $\alpha = 1$ under a "big data" condition on the sample size n [7, 12], while the theory was later extended in [10] to relax this condition by supporting smaller steps α, leading to an algorithm that may be interpreted as a primal variant of SDCA.

Note that when f_i is an expectation, it is hard to obtain such lower bounds since the gradient $\nabla f_i(x_{t-1})$ is not available in general. Nevertheless, similar methods based on the majorization-minimization principle have been successfully applied to stochastic approximations settings [4, 11], although they rely on upper bounds instead of lower bounds in MISO/Finito. The stochastic majorization-minimization scheme in [11] updates approximate upper bounds using decreasing step-sizes, and S-MISO uses a similar strategy by allowing decreasing step-sizes in (4) to obtain an approximate lower bounds to each f_i, leading to the update (2). However, in contrast to [11], our analysis does not use these bounds explicitly.

Separately, SDCA [19] considers the Fenchel conjugates of f_i, defined by $f_i^*(y) = \sup_x x^\top y - f_i(x)$. When f_i is an expectation, f_i^* is not available in closed form in general, nor are its gradients, and in fact exploiting stochastic gradient estimates is difficult in the duality framework. In contrast, Shalev-Shwartz [18] gives an analysis of SDCA in the primal, aka. "without duality", for finite sums, and our work extends this reasoning to the stochastic approximation setting.

Relationship with SGD. The link between S-MISO and SGD can be seen by rewriting the update (3) as

$$x_t = x_{t-1} + \frac{\alpha_t}{n}(z_{i_t}^t - z_{i_t}^{t-1}) = x_{t-1} + \frac{\alpha_t}{n} v_t,$$

where

$$v_t := x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^{t-1}. \tag{5}$$

Note that $\mathbb{E}[v_t|\mathcal{F}_{t-1}] = -\frac{1}{\mu}\nabla f(x_{t-1})$, where $\mathcal{F}_{t-1}$ contains all information up to iteration t, hence the algorithm can be seen as an instance of the stochastic gradient method with unbiased gradients, which was a key motivation in SVRG [9] and later in other variance reduction algorithms [6, 18].

Notes on implementation. The algorithm requires storing the vectors $(z_i^t)_{i=1,\dots,n}$ in memory, which takes the same amount of memory as the original dataset, and is therefore a reasonable requirement in practice. In the case of sparse datasets, it is reasonable to assume that the random perturbations applied to the input data preserve the sparsity patterns of the original vectors, as is the case, e.g., when applying Dropout to text documents described with bag-of-words representations [23]. If we further assume the typical setting where the μ-strong convexity comes from an $\ell 2$ regularizer: $\tilde{f}_i(x, \rho) = \phi_i(x^\top \xi_i(\rho)) + \mu/2\|x\|^2$, where $\xi_i(\rho)$ is the (sparse) perturbed example and ϕ_i encodes the loss, then the update (2) can be written as

$$z_i^t = \begin{cases} (1-\alpha_t) z_i^{t-1} + \alpha_t \phi_i'(x_{t-1}^\top \xi_i(\rho_t))\xi_i(\rho_t), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise,} \end{cases}$$

which shows that for every index i, the vector z_i^t preserves the same sparsity pattern as the examples $\xi_i(\rho)$ throughout the algorithm (assuming the standard initialization $z_i^0 = 0$), making the update (2) efficient. The update (3) is also efficient since $v_t = z_{i_t}^t - z_{i_t}^{t-1}$ is also sparse.

3 Convergence Analysis

We now study the convergence properties of the S-MISO algorithm. We start by defining the problem-dependent quantities $z_i^* := x^* - \frac{1}{\mu}\nabla f_i(x^*)$, and then introduce the Lyapunov function

$$C_t = \frac{1}{2}\|x_t - x^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^n \|z_i^t - z_i^*\|^2. \tag{6}$$

Our main result is a recursion on C_t, given in Proposition 2, which is obtained by bounding separately each term in (6), and finding coefficients to cancel out other appearing quantities when relating C_t to C_{t-1}; this requires borrowing elements of the convergence proof of SDCA without duality [18], while taking into account

the stochastic perturbations. We begin by stating the following lemma, which extends a key result of variance reduction methods (see, e.g., [9]) to the situation considered in this paper, where one only has access to noisy estimates of the gradients of each f_i.

Lemma 1. *Let i be uniformly distributed in $\{1,\dots,n\}$ and $\rho \sim \Gamma$. Under assumptions (A2) and (A3) on the functions $\tilde{f}_1,\dots,\tilde{f}_n$ and their expectations $f_1,\dots,f_n$, we have, for all $x \in \mathbb{R}^p$,*

$$\mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_i(x,\rho) - \nabla f_i(x^*)\|^2] \leq 4L(f(x) - f(x^*)) + 2\sigma^2.$$

Proof. We have

$$\begin{aligned}\|\nabla \tilde{f}_i(x,\rho) - \nabla f_i(x^*)\|^2 &\leq 2\|\nabla \tilde{f}_i(x,\rho) - \nabla \tilde{f}_i(x^*,\rho)\|^2 + 2\|\nabla \tilde{f}_i(x^*,\rho) - \nabla f_i(x^*)\|^2 \\ &\leq 4L(\tilde{f}_i(x,\rho) - \tilde{f}_i(x^*,\rho) - \nabla \tilde{f}_i(x^*,\rho)^\top (x - x^*)) + 2\|\nabla \tilde{f}_i(x^*,\rho) - \nabla f_i(x^*)\|^2.\end{aligned}$$

The first inequality comes from the classical relation $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$. The second inequality follows from the smoothness of $\tilde{f}_i(\cdot,\rho)$, in particular we used the relation

$$g(y) \geq g(x) + \nabla g(x)^\top (y - x) + \frac{1}{2L}\|\nabla g(y) - \nabla g(x)\|^2,$$

which is known to hold for any convex and L-smooth function g (see, e.g., Theorem 2.1.5 in [16]). The result follows by taking expectations on i and ρ and noting that $\mathbb{E}_{i,\rho}[\nabla \tilde{f}_i(x^*,\rho)] = \nabla f(x^*) = 0$, as well as assumption (A3). □

Proposition 2 (Recursion on C_t)**.** *If $(\alpha_t)_{t\geq 1}$ is a positive and non-increasing sequence of step-sizes satisfying*

$$\alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{2(2\kappa - 1)}\right\}, \tag{7}$$

with $\kappa = L/\mu$, then C_t obeys the recursion

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}] + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma^2}{\mu^2}. \tag{8}$$

Proof. Define the quantities

$$\begin{aligned} A_t &= \frac{1}{n}\sum_{i=1}^n \|z_i^t - z_i^*\|^2 \\ \text{and} \quad B_t &= \|x_t - x^*\|^2. \end{aligned}$$

The proof successively describes recursions on A_t, B_t, and eventually C_t.

Recursion on A_t. We have

$$\begin{aligned} A_t - A_{t-1} &= \frac{1}{n}(\|z_{i_t}^t - z_{i_t}^*\|^2 - \|z_{i_t}^{t-1} - z_{i_t}^*\|^2) \\ &= \frac{1}{n}\left(\left\|(1-\alpha_t)(z_{i_t}^{t-1} - z_{i_t}^*) + \alpha_t\left(x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - z_{i_t}^*\right)\right\|^2 - \|z_{i_t}^{t-1} - z_{i_t}^*\|^2\right) \\ &= \frac{1}{n}\left(-\alpha_t\|z_{i_t}^{t-1} - z_{i_t}^*\|^2 + \alpha_t\left\|x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - z_{i_t}^*\right\|^2 - \alpha_t(1-\alpha_t)\|v_t\|^2\right), \end{aligned}$$

where we first use the definition of z_i^t in (2), then the relation $\|(1-\lambda)u+\lambda v\|^2 = (1-\lambda)\|u\|^2 + \lambda\|v\|^2 - \lambda(1-\lambda)\|u-v\|^2$, and the definition of v_t given in (5). A similar relation is derived in the proof of SDCA without duality [18]. Using the definition of z_i^*, the second term can be expanded as follows

$$\begin{aligned}\left\|x_{t-1} - \frac{1}{\mu}\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t) - z_{i_t}^*\right\|^2 &= \left\|x_{t-1} - x^* - \frac{1}{\mu}(\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla f_{i_t}(x^*))\right\|^2 \\ &= \|x_{t-1}-x^*\|^2 - \frac{2}{\mu}\langle x_{t-1}-x^*, \nabla\tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla f_{i_t}(x^*)\rangle \\ &\quad + \frac{1}{\mu^2}\left\|\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla f_{i_t}(x^*))\right\|^2.\end{aligned}$$

We then take conditional expectations with respect to $\mathcal{F}_{t-1}$, defined in Section 2. Unless otherwise specified, we will simply write $\mathbb{E}[\cdot]$ instead of $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ for these conditional expectations in the rest of the proof.

$$\begin{aligned}\mathbb{E}\left[\left\|x_{t-1} - \frac{1}{\mu}\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t) - z_{i_t}^*\right\|^2\right] &\leq \|x_{t-1}-x^*\|^2 - \frac{2}{\mu}\langle x_{t-1}-x^*, \nabla f(x_{t-1})\rangle \\ &\quad + \frac{4L}{\mu^2}(f(x_{t-1}) - f(x^*)) + \frac{2\sigma^2}{\mu^2} \\ &\leq \|x_{t-1}-x^*\|^2 - \frac{2}{\mu}(f(x_{t-1}) - f(x^*) + \frac{\mu}{2}\|x_{t-1}-x^*\|^2) \\ &\quad + \frac{4L}{\mu^2}(f(x_{t-1}) - f(x^*)) + \frac{2\sigma^2}{\mu^2} \\ &= \frac{2(2\kappa-1)}{\mu}(f(x_{t-1}) - f(x^*)) + \frac{2\sigma^2}{\mu^2},\end{aligned}$$

where we used $\mathbb{E}[\nabla f_{i_t}(x^*)] = \nabla f(x^*) = 0$, Lemma 1, and the μ-strong convexity of f. Taking expectations on the previous relation on A_t yields

$$\begin{aligned}\mathbb{E}[A_t - A_{t-1}] &= -\frac{\alpha_t}{n}A_{t-1} + \frac{\alpha_t}{n}\,\mathbb{E}\left[\left\|x_{t-1} - \frac{1}{\mu}\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t) - z_{i_t}^*\right\|^2\right] - \frac{\alpha_t(1-\alpha_t)}{n}\,\mathbb{E}[\|v_t\|^2] \\ &\leq -\frac{\alpha_t}{n}A_{t-1} + \frac{2\alpha_t(2\kappa-1)}{n\mu}(f(x_{t-1}) - f(x^*)) - \frac{\alpha_t(1-\alpha_t)}{n}\,\mathbb{E}[\|v_t\|^2] + \frac{2\alpha_t\sigma^2}{n\mu^2}. \qquad (9)\end{aligned}$$

Recursion on B_t. Separately, we have

$$\begin{aligned}\|x_t - x^*\|^2 &= \left\|x_{t-1} - x^* + \frac{\alpha_t}{n}v_t\right\|^2 \\ &= \|x_{t-1}-x^*\|^2 + \frac{2\alpha_t}{n}\langle x_{t-1}-x^*, v_t\rangle + \left(\frac{\alpha_t}{n}\right)^2\|v_t\|^2 \\ \mathbb{E}[\|x_t - x^*\|^2] &= \|x_{t-1}-x^*\|^2 - \frac{2\alpha_t}{n\mu}\langle x_{t-1}-x^*, \nabla f(x_{t-1})\rangle + \left(\frac{\alpha_t}{n}\right)^2\mathbb{E}[\|v_t\|^2] \\ &\leq \|x_{t-1}-x^*\|^2 - \frac{2\alpha_t}{n\mu}(f(x_{t-1}) - f(x^*) + \frac{\mu}{2}\|x_{t-1}-x^*\|^2) + \left(\frac{\alpha_t}{n}\right)^2\mathbb{E}[\|v_t\|^2],\end{aligned}$$

using that $\mathbb{E}[v_t] = -\frac{1}{\mu}\nabla f(x_{t-1})$ and the strong convexity of f. This gives

$$\mathbb{E}[B_t - B_{t-1}] \leq -\frac{\alpha_t}{n}B_{t-1} - \frac{2\alpha_t}{n\mu}(f(x_{t-1}) - f(x^*)) + \left(\frac{\alpha_t}{n}\right)^2\mathbb{E}[\|v_t\|^2]. \qquad (10)$$

Recursion on C_t. If we consider $C_t = p_t A_t + q_t B_t$ and $C'_{t-1} = p_t A_{t-1} + q_t B_{t-1}$, combining (9) and (10) yields

$$\mathbb{E}[C_t - C'_{t-1}] \leq -\frac{\alpha_t}{n} C'_{t-1} + \frac{2\alpha_t}{n\mu}(p_t(2\kappa - 1) - q_t)(f(x_{t-1}) - f(x^*)) + \frac{\alpha_t}{n}\left(\frac{q_t \alpha_t}{n} - p_t(1-\alpha_t)\right)\mathbb{E}[\|v_t\|^2] + \frac{2\alpha_t p_t \sigma^2}{n\mu^2}.$$

If we take $p_t = \frac{\alpha_t}{n}$ and $q_t = \frac{1}{2}$, and if $(\alpha_t)_{t \geq 1}$ is a decreasing sequence satisfying (7), then the factors in front of $f(x_{t-1}) - f(x^*)$ and $\mathbb{E}[\|v_t\|^2]$ are non-positive and we get

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right) C'_{t-1} + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma^2}{\mu^2}.$$

Finally, since $\alpha_t \leq \alpha_{t-1}$, we have $C'_{t-1} \leq C_{t-1}$. After taking total expectations on $\mathcal{F}_{t-1}$, we are left with the desired recursion.

□

Comparison with SGD. A classical analysis of SGD with step-sizes $(\eta_t)_{t \geq 0}$ gives the following recursion (see, e.g., [15]) on $B_t := \frac{1}{2}\mathbb{E}[\|x_t - x^*\|^2]$:

$$B_t \leq (1 - \mu\eta_t) B_{t-1} + \frac{\eta_t^2 M^2}{2} = (1 - \mu\eta_t) B_{t-1} + (\mu\eta_t)^2 \frac{M^2}{2\mu^2},$$

where we assume $\mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_i(x, \rho)\|^2] \leq M^2$ for all x. Thus, after forgetting the initial condition C_0, S-MISO minimizes $B_t \leq C_t$ at a faster rate if $2\sigma^2 \leq M^2/2$. In particular, if the gradient variance *across examples* (bounded by M^2 here) is much larger than the gradient variance due to the data perturbation only $\rho \sim \Gamma$ (bounded by σ^2 at the optimum), then our algorithm will have a much faster convergence rate. As shown in the experiments presented in the next section, M^2 may be indeed orders of magnitude larger than σ^2 in real scenarios, leading to both theoretical and practical benefits.

We now state the main convergence result, which provides the expected rate $O(1/t)$ on C_t based on decreasing step-sizes, similar to [3] for SGD. Note that convergence of objective function values is directly related to that of the Lyapunov function C_t via smoothness:

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{L}{2}\mathbb{E}\left[\|x_t - x^*\|^2\right] \leq L\,\mathbb{E}[C_t].$$

Theorem 3. *Let the sequence of step-sizes $(\alpha_t)_{t \geq 1}$ be defined by*

$$\alpha_t = \frac{\beta n}{\gamma + t} \quad \text{for } \beta > 1 \text{ and } \gamma \geq 0 \text{ s.t. } \alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{2(2\kappa - 1)}\right\}.$$

For all $t \geq 0$, it holds that

$$\mathbb{E}[C_t] \leq \frac{\nu}{\gamma + t + 1},$$

where

$$\nu := \max\left\{\frac{2\beta^2\sigma^2}{\mu^2(\beta - 1)}, (\gamma + 1) C_0\right\}. \tag{11}$$

Proof. Let us proceed by induction. We have $C_0 \leq \nu/(\gamma+1)$ by definition of ν. For $t \geq 1$,

$$\begin{aligned}
\mathbb{E}[C_t] &\leq \left(1 - \frac{\alpha_t}{n}\right) \mathbb{E}[C_{t-1}] + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma^2}{\mu^2} \\
&\leq \left(1 - \frac{\beta}{\hat{t}}\right) \frac{\nu}{\hat{t}} + \frac{2\beta^2\sigma^2}{\hat{t}^2\mu^2} \quad \text{(with } \hat{t} := \gamma + t\text{)} \\
&= \left(\frac{\hat{t}-\beta}{\hat{t}^2}\right) \nu + \frac{2\beta^2\sigma^2}{\hat{t}^2\mu^2} \\
&= \left(\frac{\hat{t}-1}{\hat{t}^2}\right) \nu - \left(\frac{\beta-1}{\hat{t}^2}\right) \nu + \frac{2\beta^2\sigma^2}{\hat{t}^2\mu^2} \\
&\leq \left(\frac{\hat{t}-1}{\hat{t}^2}\right) \nu \leq \frac{\nu}{\hat{t}+1},
\end{aligned}$$

where the last two inequalities follow from the definition of ν and from $\hat{t}^2 \geq (\hat{t}+1)(\hat{t}-1)$. □

Choice of step-sizes in practice. Naturally, we would like ν to be small, in particular independent of the initial condition C_0 and equal to the first term in the definition (11). We would like the dependence on C_0 to vanish at a faster rate than $O(1/t)$, as it is the case in variance reduction algorithms on finite sums. As advised in [3] in the context of SGD, we can initially run the algorithm with a constant step-size $\bar{\alpha}$ and exploit this linear convergence regime until we reach the level of noise given by σ, and then start decaying the step-size.

It is easy to see that by using a constant step-size $\bar{\alpha}$, C_t will converge linearly near a value $\bar{C} := \frac{2\bar{\alpha}\sigma^2}{n\mu^2}$. Indeed, Eq. (8) with $\alpha_t = \bar{\alpha}$ yields

$$\mathbb{E}[C_t - \bar{C}] \leq \left(1 - \frac{\bar{\alpha}}{n}\right) \mathbb{E}[C_{t-1} - \bar{C}].$$

We can thus reach a value C_0' with $\mathbb{E}[C_0'] \leq \bar{\epsilon} := 2\bar{C}$ in $O(\frac{n}{\bar{\alpha}} \log C_0/\bar{\epsilon})$ iterations. If we then start decaying step-sizes as in Theorem 3 with $\beta = 2$ and γ large enough so that $\alpha_1 = \frac{\beta n}{\gamma+1} = \bar{\alpha}$, we will have

$$(\gamma+1)\,\mathbb{E}[C_0'] \leq (\gamma+1)\bar{\epsilon} = 8\sigma^2/\mu^2,$$

making both terms in (11) smaller than or equal to $\nu = 8\sigma^2/\mu^2$. Considering these two phases, with an initial step-size $\bar{\alpha}$ given by the upper bound in (7), the final work complexity of the algorithm is

$$O\left((n+\kappa)\log\frac{C_0}{\bar{\epsilon}}\right) + O\left(\frac{L\sigma^2}{\mu^2\epsilon}\right).$$

Note that following this heuristic for choosing the step-size was found to be very effective in practice (see next section).

4 Experiments

We present preliminary experiments comparing S-MISO with SGD on two significantly different scenarios, in order to demonstrate the wide applicability of our method: we consider an image classification dataset with random transformations and a classification task on breast cancer data with Dropout (the perturbation sets randomly a fraction of the data entries to zero). For both algorithms, we use the step-size strategy mentioned in Section 3 and advised by Bottou et al. [3], which we have found to be most effective among many heuristics we have tried: we initially keep the step-size constant (controlled by a factor $\eta \leq 1$ in Figure 1) for 2 epochs, and then start decaying as $\alpha_t = C/(\gamma+t)$, where $C = 2n$ for S-MISO, $C = 2/\mu$ for SGD, and γ is chosen

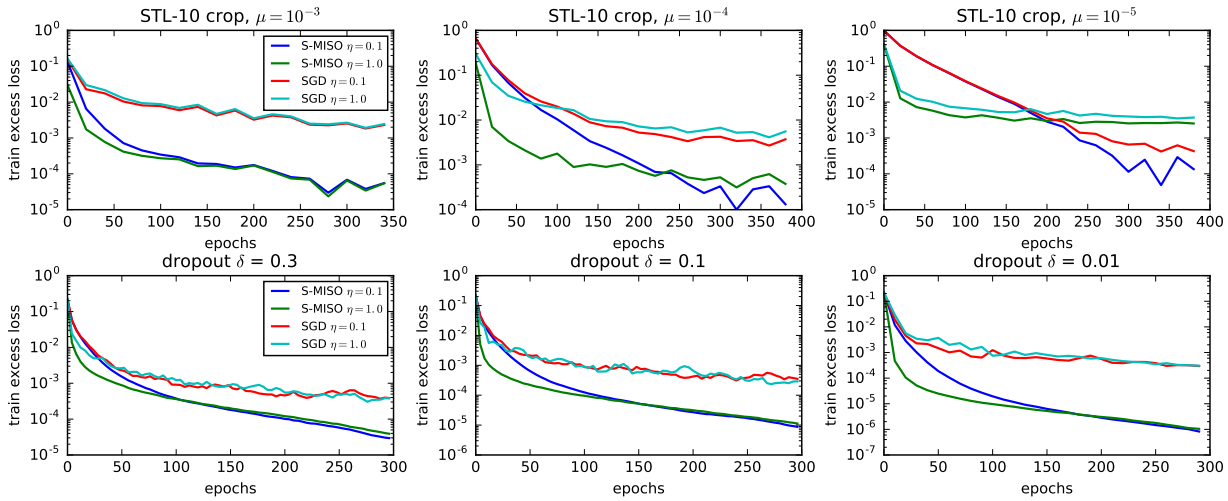

Figure 1: Comparison of S-MISO with SGD. Values of the training loss are shown in logarithmic scale (1 unit = factor 10). (Top) STL-10 dataset with different values of μ (the best value given by cross validation is around 10^{-4}); (bottom) breast cancer dataset with Dropout, for different values of δ and $\mu = 0.003$ (selected by 5-fold cross validation with no Dropout). Curves for $\eta = 10$ (not shown) are diverging.

large enough to match the previous constant step-size. Figure 1 shows the curves we obtain for a Monte-Carlo estimate of the training objective. The plots are shown on a logarithmic scale, and the values are compared to the best value obtained among the different approaches in 400 epochs. In both cases, the strong convexity constant μ is the regularization parameter.

Image classification with "data augmentation". The success of deep neural networks is often limited by the availability of large amounts of labeled images. When there are many unlabeled images but few labeled ones, a common approach is to train a linear classifier on top of a deep network learned in an unsupervised manner. We follow this approach on the STL-10 dataset [5], which contains 5 000 training images from 10 classes and 100 000 unlabeled images, using a 2-layer unsupervised convolutional kernel network [13], giving representations of dimension 102 400. The perturbation consists of randomly cropping the input images. The loss function is the squared hinge loss used in a one-versus-all setting. The vector representations are ℓ_2-normalized such that we may use the upper bound $L = 1 + \mu$ for the smoothness constant.

Figure 1 (top) shows convergence results on one training fold (500 images), for different values of μ, allowing us to study the behavior of the algorithms for different condition numbers. The low variance induced by the data transformations allows S-MISO to reach suboptimality that is orders of magnitude smaller than SGD after the same number of epochs. Note that one unit on these plots corresponds to one order of magnitude in the logarithmic scale. The best validation accuracy is obtained for $\mu \approx 10^{-4}$ (middle plot in Figure 1), giving a 0.5% accuracy improvement over the non-augmented strategy. A more aggressive augmentation strategy with resizing gave a 2% improvement. Compared to SGD, S-MISO reached the improved accuracy in less than half the number of epochs in both cases. We computed empirical variances of the image representations for these two strategies, which are closely related to the variance in gradient estimates, and observed these transformations to account for about 10% and 30% of the total variance across multiple images, respectively.

Dropout on gene expression data. We trained a binary logistic regression model on the breast cancer gene expression dataset of Van de Vijver et al. [22] with different Dropout rates δ, i.e. where at every iteration, each coordinate ξ_j of a feature vector ξ is set to zero independently with probability δ and to $\xi_j/(1-\delta)$

otherwise. Figure 1 (bottom) compares S-MISO with SGD for three values of δ, as a way to control the variance of the perturbations. We include a Dropout rate of 0.01 to illustrate the impact of δ on the algorithms and study the influence of the perturbation variance σ^2, even though this value of δ is less relevant for the task. The plots show very clearly how the variance induced by the perturbations affects the convergence of S-MISO, giving suboptimality values that may be orders of magnitude smaller than SGD. This behavior is consistent with the theoretical convergence rate established in Section 3 and shows that the practice matches the theory.

References

[1] M. Achab, A. Guilloux, S. Gaïffas, and E. Bacry. SGD with Variance Reduction beyond Empirical Risk Minimization. *arXiv:1510.04822*, 2015.

[2] Z. Allen-Zhu, Y. Yuan, and K. Sridharan. Exploiting the Structure: Stochastic Gradient Methods Using Raw Clusters. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *arXiv:1606.04838*, 2016.

[4] O. Cappé and E. Moulines. Online expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, June 2009.

[5] A. Coates, H. Lee, and A. Y. Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[6] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[7] A. Defazio, J. Domke, and T. S. Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning (ICML)*, 2014.

[8] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance Reduced Stochastic Gradient Descent with Neighbors. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[9] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[10] H. Lin, J. Mairal, and Z. Harchaoui. A Universal Catalyst for First-Order Optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[11] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[12] J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[13] J. Mairal. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[14] R. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[15] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[16] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Springer, 2004.

[17] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2016.

[18] S. Shalev-Shwartz. SDCA without Duality, Regularization, and Individual Convexity. In *International Conference on Machine Learning (ICML)*, 2016.

[19] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[20] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri. Transformation Invariance in Pattern Recognition — Tangent Distance and Tangent Propagation. In G. B. Orr and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, number 1524 in Lecture Notes in Computer Science, pages 239–274. Springer Berlin Heidelberg, 1998.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[22] M. J. van de Vijver et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, 347(25):1999–2009, Dec. 2002.

[23] S. Wager, W. Fithian, S. Wang, and P. Liang. Altitude Training: Strong Bounds for Single-layer Dropout. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.