



HAL
open science

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

Alberto Bietti, Julien Mairal

► **To cite this version:**

Alberto Bietti, Julien Mairal. Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure. 2017. hal-01375816v2

HAL Id: hal-01375816

<https://inria.hal.science/hal-01375816v2>

Preprint submitted on 10 Jan 2017 (v2), last revised 15 Nov 2017 (v6)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

January 10, 2017

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. Extensions of the algorithm to composite objectives and non-uniform sampling are also studied.

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, 17]. However, in the case of finite datasets, incremental methods based on variance reduction techniques (e.g., [7, 10, 11, 19, 21]) have proven to be very successful at solving the finite-sum problem

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

where h is a simple convex penalty that need not be differentiable, e.g., the ℓ_1 norm. 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 [26] and in practice [22, 25]. Examples of such procedures include random transformations of images in classification problems (e.g., [22]), and Dropout [23]. The objective describing these scenarios, which we consider in this paper, is the following:

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

*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.

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 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 when $h = 0$ 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 achieves 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. To our knowledge, our method is the first hybrid algorithm that naturally interpolates between the incremental algorithms for finite sums (when there are no perturbations), for which we recover a linear rate of convergence, and the stochastic approximation algorithms (when $n = 1$). In the stochastic composite case with $n = 1$, we obtain a novel algorithm with the same convergence properties as SGD.

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., [7, 10, 11, 19, 21]). 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 [10] and SAGA [7] 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 [8, 11, 14] to the case where each function in the finite sum can only be accessed via a first-order stochastic oracle.

Despite its relevance in machine learning (e.g., [25, 26]), the generic problem of minimizing (1) is not well studied in the optimization literature. Most related to our work, recent methods that use clustering information to improve the convergence of variance reduction techniques [2, 9] 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 [9] 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 relies on computing a full (exact) gradient at every pass, which is not available in our setting, and 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.

We present our algorithm in Section 2 for non-composite objectives and study its convergence properties in Section 3. We introduce an extension of the algorithm to non-smooth (composite) objectives and non-uniform sampling in Section 4. In Section 5, 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 and movie review data with Dropout, demonstrating the applicability of our approach to significantly different contexts.

2 The Stochastic MISO Algorithm for Non-Composite Objectives

In this section, we introduce the *stochastic MISO* approach for $h = 0$, which relies on the following assumptions:

Algorithm 1 S-MISO for smooth objectives

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:

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

$$x_t = \frac{1}{n} \sum_{i=1}^n z_i^t = x_{t-1} + \frac{1}{n} (z_{i_t}^t - z_{i_t}^{t-1}). \quad (3)$$

end for

- (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:**

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

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

Note that we will relax the smoothness assumption (A2) in Section 4 by supporting composite objective with non-smooth regularizers, and by exploiting different smoothness parameters L_i on each example, a setting where non-uniform sampling of the training points is typically helpful to accelerate the convergence of incremental methods (e.g., [19, 20]). 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 similar assumption for the SGD algorithm on the objective (1) would take the form $\mathbb{E}_{i, \rho} [\|\nabla \tilde{f}_i(x^*, \rho)\|^2] \leq \sigma_{tot}^2$ for all x (see Proposition B.1). The quantity σ_{tot}^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. In Section 3, we will show 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, 17], but with σ^2 replacing the quantity σ_{tot}^2 , leading to a much faster rate than SGD if $\sigma^2 \ll \sigma_{tot}^2$, something which we observed in our experiments (see Section 5).

Our method is given in Algorithm 1. Without the perturbations and with a constant step-size, the algorithm resembles the MISO/Finito algorithms [8, 11, 14] which may be seen as primal variants of SDCA [20, 21]. 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; for instance, this strategy is commonly adopted in bundle methods or in the EM algorithm and its incremental variants [16]. 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

$$d_i^t(x) = c_{i,1}^t + \frac{\mu}{2} \|x - z_i^t\|^2 = c_{i,2}^t - \mu \langle x, z_i^t \rangle + \frac{\mu}{2} \|x\|^2. \quad (4)$$

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

$$d_i^t(x) = \begin{cases} (1 - \alpha_t)d_i^{t-1}(x) + \alpha_t(f_i(x_{t-1}) + \langle \nabla f_i(x_{t-1}), x - x_{t-1} \rangle + \frac{\mu}{2} \|x - x_{t-1}\|^2), & \text{if } i = i_t \\ d_i^{t-1}(x), & \text{otherwise,} \end{cases} \quad (5)$$

which corresponds to updating z_i^t as

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

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 $\alpha_t = 1$ under a “big data” condition on the sample size n [8, 14], while the theory was later extended in [11] to relax this condition by supporting smaller constant steps $\alpha_t = \alpha$, 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 [5, 13], although they rely on upper bounds instead of lower bounds in MISO/Finito. The stochastic majorization-minimization scheme in [13] updates approximate upper bounds using decreasing step-sizes, and S-MISO uses a similar strategy to obtain an approximate lower bounds to each f_i , leading to the update (2). However, in contrast to [13], our analysis does not use these bounds explicitly.

Separately, SDCA [21] 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 [20] gives an analysis of SDCA in the primal, aka. “without duality”, for smooth finite sums, and our work extends this reasoning to the stochastic approximation and composite settings.

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

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

where

$$v_t := x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^{t-1}. \quad (6)$$

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 [10] and later in other variance reduction algorithms [7, 20]. It is also worth noting that when $n = 1$, we have $z_{i_t}^{t-1} = x_{t-1}$, hence our method becomes identical to “vanilla” SGD, up to a redefinition of step-sizes.

Memory requirements and handling of sparse datasets. 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 [26]. If we further assume the typical setting where the μ -strong convexity comes from an ℓ_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

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

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 of S-MISO for Smooth Objectives

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

$$C_t = \frac{1}{2} \|x_t - x^*\|^2 + \frac{\alpha_t}{n^2} \sum_{i=1}^n \|z_i^t - z_i^*\|^2. \quad (7)$$

Our main result is a recursion on C_t , given in Proposition 2, which is obtained by bounding separately each term in (7), 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 [20], 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., [10]) 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

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

The first inequality comes from the simple 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 classical relation

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

which is known to hold for any convex and L -smooth function g (see, e.g., [18, Theorem 2.1.5]). 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). \square

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

$$\alpha_1 \leq \min \left\{ \frac{1}{2}, \frac{n}{2(2\kappa - 1)} \right\}, \quad (8)$$

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

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right) \mathbb{E}[C_{t-1}] + 2 \left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma^2}{\mu^2}. \quad (9)$$

Proof. Define the quantities

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

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

Recursion on A_t . We have

$$\begin{aligned}
A_t - A_{t-1} &= \frac{1}{n} (\|z_{i_t}^t - z_{i_t}^*\|^2 - \|z_{i_t}^{t-1} - z_{i_t}^*\|^2) \\
&= \frac{1}{n} \left(\left\| (1 - \alpha_t)(z_{i_t}^{t-1} - z_{i_t}^*) + \alpha_t \left(x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^* \right) \right\|^2 - \|z_{i_t}^{t-1} - z_{i_t}^*\|^2 \right) \\
&= \frac{1}{n} \left(-\alpha_t \|z_{i_t}^{t-1} - z_{i_t}^*\|^2 + \alpha_t \left\| x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^* \right\|^2 - \alpha_t (1 - \alpha_t) \|v_t\|^2 \right), \quad (10)
\end{aligned}$$

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 (6). A similar relation is derived in the proof of SDCA without duality [20]. Using the definition of z_i^* , the second term can be expanded as follows

$$\begin{aligned}
\left\| x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^* \right\|^2 &= \left\| x_{t-1} - x^* - \frac{1}{\mu} (\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - \nabla f_{i_t}(x^*)) \right\|^2 \\
&= \|x_{t-1} - x^*\|^2 - \frac{2}{\mu} \langle x_{t-1} - x^*, \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - \nabla f_{i_t}(x^*) \rangle \\
&\quad + \frac{1}{\mu^2} \|\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - \nabla f_{i_t}(x^*)\|^2. \quad (11)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[A_t - A_{t-1}] &= -\frac{\alpha_t}{n} A_{t-1} + \frac{\alpha_t}{n} \mathbb{E} \left[\left\| x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^* \right\|^2 \right] - \frac{\alpha_t(1 - \alpha_t)}{n} \mathbb{E}[\|v_t\|^2] \\
&\leq -\frac{\alpha_t}{n} A_{t-1} + \frac{2\alpha_t(2\kappa - 1)}{n\mu} (f(x_{t-1}) - f(x^*)) - \frac{\alpha_t(1 - \alpha_t)}{n} \mathbb{E}[\|v_t\|^2] + \frac{2\alpha_t\sigma^2}{n\mu^2}. \quad (12)
\end{aligned}$$

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

$$\mathbb{E}[B_t - B_{t-1}] \leq -\frac{\alpha_t}{n}B_{t-1} - \frac{\alpha_t}{n\mu}(f(x_{t-1}) - f(x^*)) + \frac{1}{2}\left(\frac{\alpha_t}{n}\right)^2 \mathbb{E}[\|v_t\|^2]. \quad (13)$$

Recursion on C_t . If we consider $C_t = p_t A_t + B_t$ and $C'_{t-1} = p_t A_{t-1} + B_{t-1}$, combining (12) and (13) yields

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

If we take $p_t = \frac{\alpha_t}{n}$, and if $(\alpha_t)_{t \geq 1}$ is a decreasing sequence satisfying (8), 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. \square

Comparison with SGD. A simple analysis of SGD with step-sizes $(\eta_t)_{t \geq 0}$ gives the following recursion (see Proposition B.1 in the appendix for a proof) on $B_t := \frac{1}{2} \mathbb{E}[\|x_t - x^*\|^2]$:

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

where we assume $\mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_i(x^*, \rho)\|^2] \leq \sigma_{tot}^2$. Thus, after forgetting the initial condition C_0 , S-MISO minimizes $B_t \leq C_t$ at a faster rate if $2\sigma^2 \leq \sigma_{tot}^2$. In particular, if the gradient variance *across examples* (bounded by σ_{tot}^2 at the optimum) 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 has a much faster convergence rate. As shown in the experimental section, σ_{tot}^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:¹

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{L}{2} \mathbb{E}[\|x_t - x^*\|^2] \leq L \mathbb{E}[C_t]. \quad (14)$$

Theorem 3 (Convergence of Lyapunov function). *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

$$\nu := \max\left\{\frac{2\beta^2 \sigma^2}{\mu^2(\beta - 1)}, (\gamma + 1)C_0\right\}. \quad (15)$$

¹Note that the constant L is an upper bound of the smoothness constant of each function $\tilde{f}_i(\cdot, \rho)$; it can be replaced here by the global smoothness constant of f , which may be smaller than L .

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

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

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

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 (15). 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. (9) 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 (15) 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 (8), the final work complexity of the algorithm for reaching $\mathbb{E}[\|x_t - x^*\|^2] \leq \epsilon$ is

$$O\left((n + \kappa) \log \frac{C_0}{\bar{\epsilon}}\right) + O\left(\frac{\sigma^2}{\mu^2 \epsilon}\right). \quad (16)$$

We can use (14) in order to obtain the complexity for reaching $\mathbb{E}[f(x_t) - f(x^*)] \leq \epsilon$, and the second term becomes $O(L\sigma^2/\mu^2\epsilon)$. Note that following this heuristic for choosing the step-size was found to be very effective in practice (see Section 5).

4 Extension to Composite Objectives and Non-Uniform Sampling

In this section, we study extensions of the S-MISO algorithm to different situations where our previous smoothness assumption (A2) is not suitable, either because of a non-smooth term h in the objective or because it ignores additional useful knowledge about each f_i , e.g. the norm of each example.

In the presence of non-smooth regularizers (e.g. an ℓ_1 regularization term $h(x) = \lambda\|x\|_1$ in (1)), the objective is no longer smooth, but we can leverage its composite structure by using proximal operators. To this end, we assume that one can easily compute the proximal operator of h , defined by

$$\text{prox}_h(z) := \arg \min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|x - z\|^2 + h(x) \right\}.$$

Algorithm 2 S-MISO for composite objectives, with non-uniform sampling.

Input: step-size sequence $(\alpha_t)_{t \geq 1}$, sampling distribution q ;

Initialize $x_0 = \text{prox}_{h/\mu}(\bar{z}_0)$ with $\bar{z}_0 = \frac{1}{n} \sum_i z_i^0$ for some $(z_i^0)_{i=1, \dots, n}$ that satisfies (20);

for $t = 1, \dots$ **do**

 Sample an index $i_t \sim q$, a perturbation $\rho_t \sim \Gamma$, and update:

$$z_i^t = \begin{cases} (1 - \frac{\alpha_t}{q_i n}) z_i^{t-1} + \frac{\alpha_t}{q_i n} (x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_i(x_{t-1}, \rho_t)), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise.} \end{cases} \quad (17)$$

$$\begin{aligned} \bar{z}_t &= \frac{1}{n} \sum_{i=1}^n z_i^t = \bar{z}_{t-1} + \frac{1}{n} (z_{i_t}^t - z_{i_t}^{t-1}) \\ x_t &= \text{prox}_{h/\mu}(\bar{z}_t). \end{aligned} \quad (18)$$

end for

When the smoothness constants L_i vary significantly across different examples (typically through the norm of the feature vectors), the uniform upper bound $L = L_{\max} = \max_i L_i$ can be restrictive. It has been noticed (see, e.g., [19]) that when the L_i are known, one can achieve better convergence rates—typically depending on the average smoothness constant $\bar{L} = \frac{1}{n} \sum_i L_i$ rather than L_{\max} —by sampling examples in a non-uniform way. For that purpose, we make the following assumptions in this section:

- (A4) **strong convexity:** $\tilde{f}_i(\cdot, \rho)$ is μ -strongly convex for all i and ρ ;
- (A5) **smoothness:** $\tilde{f}_i(\cdot, \rho)$ is L_i -smooth for all i and ρ ;
- (A6) **small variance from perturbations at x^* :** $\mathbb{E}_\rho [\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2] \leq \sigma_i^2$ for all i .

Note that our proof relies on a slightly stronger assumption (A4) than the global strong convexity assumption (A1) made above, which holds in the common situation where strong convexity comes from an ℓ_2 regularization term. In order to exploit the different smoothness constants, we allow the algorithm to sample indices i non-uniformly, from any distribution q such that $q_i \geq 0$ for all i and $\sum_i q_i = 1$.

The extension of S-MISO to this new setting is given in Algorithm 2. Note that the step-sizes vary depending on the example, with larger step-sizes for examples that are sampled less frequently (typically “easier” examples with smaller L_i). Note that when $h = 0$, the update directions are unbiased estimates of the gradient: we have $\mathbb{E}[x_t - x_{t-1} | \mathcal{F}_{t-1}] = -\frac{\alpha_t}{n\mu} \nabla f(x_{t-1})$ as in the uniform case. However, in the composite case, the algorithm cannot be written in a proximal stochastic gradient form like Prox-SVRG [28] or SAGA [7].

Interpretation as an extension of RDA. In the $n = 1$ case, our algorithm has updates that are similar to Regularized Dual Averaging (RDA) [27] with strongly convex regularizers. In particular, if $\tilde{f}_1(x, \rho) = \phi(x^\top \xi(\rho)) + \mu/2 \|x\|^2$, the updates are the same when taking $\alpha_t = 1/t$, since

$$x_t = \arg \min_x \left\{ \frac{\mu}{2} \|x - \bar{z}_t\|^2 + h(x) \right\} = \arg \min_x \left\{ \langle -\mu \bar{z}_t, x \rangle + \frac{\mu}{2} \|x\|^2 + h(x) \right\},$$

and $-\mu \bar{z}_t$ is equal to the average of the gradients of the loss term up to t , which appears in the same way in the RDA updates [27, Section 2.2]. However, unlike RDA, our analysis does not rely on averaging—which is often avoided in practice since it hurts the sparsity of the iterates—and supports arbitrary decreasing step-sizes, in particular keeping the step-size constant, which as we have seen in Section 3, and as explained in [3], can be beneficial in the initial iterations.

We can again view the algorithm as iteratively updating approximate lower bounds on the objective F of

the form $D_t(x) = \frac{1}{n} \sum_i d_i^t(x) + h(x)$ in a similar way to (5):

$$d_i^t(x) = \begin{cases} (1 - \frac{\alpha_t}{q_i n}) d_i^{t-1}(x) + \frac{\alpha_t}{q_i n} (\tilde{f}_i(x_{t-1}, \rho_t) + \langle \nabla \tilde{f}_i(x_{t-1}, \rho_t), x - x_{t-1} \rangle + \frac{\mu}{2} \|x - x_{t-1}\|^2), & \text{if } i = i_t \\ d_i^{t-1}(x), & \text{otherwise,} \end{cases} \quad (19)$$

which is done in (17), and minimizing the new D_t in (18). Similarly to MISO-Prox, we require that d_i^0 is initialized with a μ -strongly convex quadratic such that $\tilde{f}_i(x, \tilde{\rho}_i) \geq d_i^0(x)$ with $\tilde{\rho}_i \sim \Gamma$. Given the form of d_i^t (4), it suffices to choose z_i^0 that satisfies

$$\tilde{f}_i(x, \tilde{\rho}_i) \geq \frac{\mu}{2} \|x - z_i^0\|^2 + c, \quad (20)$$

for some constant c . In the common case of an ℓ_2 regularizer with a non-negative loss, one can simply choose $z_i^0 = 0$ for all i , otherwise, z_i^0 can be obtained by considering a strong convexity lower bound on $\tilde{f}_i(\cdot, \tilde{\rho}_i)$. Our new analysis relies on the minimum $D_t(x_t)$ of the lower bounds D_t through the following Lyapunov function:

$$C_t^q = F(x^*) - D_t(x_t) + \frac{\mu \alpha_t}{n^2} \sum_{i=1}^n \frac{1}{q_i n} \|z_i^t - z_i^*\|^2. \quad (21)$$

The convergence of the iterates x_t is controlled by the convergence in C_t^q thanks to the following lemma:

Lemma 4 (Bound on the iterates). *For all t , we have*

$$\frac{\mu}{2} \mathbb{E}[\|x_t - x^*\|^2] \leq \mathbb{E}[F(x^*) - D_t(x_t)]. \quad (22)$$

Proof. Let $F_t(x) := \frac{1}{n} \sum_{i=1}^n f_i^t(x) + h(x)$, where $f_i^0(x) = \tilde{f}_i(x, \tilde{\rho}_i)$ (where $\tilde{\rho}_i$ is used in (20)), and f_i^t is updated analogously to d_i^t as follows:

$$f_i^t(x) = \begin{cases} (1 - \frac{\alpha_t}{q_i n}) f_i^{t-1}(x) + \frac{\alpha_t}{q_i n} \tilde{f}_i(x, \rho_t), & \text{if } i = i_t \\ f_i^{t-1}(x), & \text{otherwise.} \end{cases}$$

By induction, we have

$$F_t(x^*) \geq D_t(x^*) \geq D_t(x_t) + \frac{\mu}{2} \|x_t - x^*\|^2, \quad (23)$$

where the last inequality follows from the μ -strong convexity of D_t and the fact that x_t is its minimizer.

Again by induction, we now show that $\mathbb{E}[F_t(x^*)] = F(x^*)$. Indeed, we have $\mathbb{E}[F_0(x^*)] = F(x^*)$ by construction, then

$$\begin{aligned} F_t(x^*) &= F_{t-1}(x^*) + \frac{\alpha_t}{q_i n^2} (\tilde{f}_{i_t}(x^*, \rho_t) - f_{i_t}^{t-1}(x^*)) \\ \mathbb{E}[F_t(x^*) | \mathcal{F}_{t-1}] &= F_{t-1}(x^*) + \frac{\alpha_t}{n} (f(x^*) - \frac{1}{n} \sum_{i=1}^n f_i^{t-1}(x^*)) \\ &= F_{t-1}(x^*) + \frac{\alpha_t}{n} (F(x^*) - F_{t-1}(x^*)), \end{aligned}$$

After taking total expectations and using the induction hypothesis, we obtain $\mathbb{E}[F_t(x^*)] = F(x^*)$, and the result follows from (23). \square

The following proposition gives a recursion on C_t^q similar to Proposition 2.

Proposition 5 (Recursion on Lyapunov function C_t^q). *If $(\alpha_t)_{t \geq 1}$ is a positive and non-increasing sequence of step-sizes satisfying*

$$\alpha_1 \leq \min \left\{ \frac{nq_{\min}}{2}, \frac{n\mu}{4L_q} \right\}, \quad (24)$$

with $q_{\min} = \min_i q_i$ and $L_q = \max_i \frac{L_i - \mu}{q_i n}$, then C_t^q obeys the recursion

$$\mathbb{E}[C_t^q] \leq \left(1 - \frac{\alpha_t}{n}\right) \mathbb{E}[C_{t-1}^q] + 2 \left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma_q^2}{\mu}, \quad (25)$$

with $\sigma_q^2 = \frac{1}{n} \sum_i \frac{\sigma_i^2}{q_i n}$.

Note that if we consider the quantity $\mathbb{E}[C_t^q/\mu]$, which is an upper bound on $\frac{1}{2} \mathbb{E}[\|x_t - x^*\|^2]$ by Lemma 4, we have the same recursion as (2), and thus can apply Theorem 3 with the new condition (24). If we take

$$q_i = \frac{1}{2n} + \frac{L_i - \mu}{2 \sum_i (L_i - \mu)}, \quad (26)$$

we have $q_{\min} \geq 1/2n$ and $L_q \leq 2(\bar{L} - \mu)$, where $\bar{L} = \frac{1}{n} \sum_i L_i$. Then, taking $\alpha_1 = \min(1/4, n\mu/8(\bar{L} - \mu))$ satisfies (24), and using similar arguments to Section 3, the complexity for reaching $\mathbb{E}[\|x_t - x^*\|^2] \leq \epsilon$ is

$$O\left(\left(n + \frac{\bar{L}}{\mu}\right) \log \frac{C_0^q}{\bar{\epsilon}}\right) + O\left(\frac{\sigma_q^2}{\mu^2 \epsilon}\right),$$

where $\bar{\epsilon} = 4\bar{\alpha}\sigma_q^2/n\mu$, and $\bar{\alpha}$ is the initial constant step-size (in the smooth case, we can get a similar complexity on objective values with an additional L factor in the second term using (14)). Note that with our choice of q , we have $\sigma_q^2 \leq \frac{2}{n} \sum_i \sigma_i^2 = 2\bar{\sigma}^2$, for general perturbations, where $\bar{\sigma}^2 = \frac{1}{n} \sum_i \sigma_i^2$ is the variance in the uniform case. Additionally, it is often reasonable to assume that the variance from perturbations increases with the norm of examples, for instance Dropout perturbations get larger when coordinates have larger magnitudes. Based on this observation, if we make the assumption that $\sigma_i^2 \propto L_i - \mu$, that is $\sigma_i^2 = \bar{\sigma}^2 \frac{L_i - \mu}{\bar{L} - \mu}$, then for both $q_i = 1/n$ (uniform case) and $q_i = (L_i - \mu)/n(\bar{L} - \mu)$, we have $\sigma_q^2 = \bar{\sigma}^2$, and thus we have $\sigma_q^2 \leq \bar{\sigma}^2$ for the choice of q given in (26), since σ_q^2 is convex in q . Thus, we can expect that the $O(1/t)$ convergence phase will behave similarly or better than for uniform sampling, and this is confirmed by our experiments, which show similar behavior (see Section 5).

5 Experiments

We present experiments comparing S-MISO with SGD and N-SAGA [9] on four different scenarios, in order to demonstrate the wide applicability of our method: we consider an image classification dataset with two different types of image representations and random transformations, and two different classification tasks with Dropout regularization (the perturbation sets randomly a fraction of the data entries to zero), one on breast cancer data, and one on (sparse) text data. Figure 1 shows the curves we obtain for a Monte-Carlo estimate of the training objective using 5 sampled perturbations per example (the curves we obtained were similar when using more samples). The plots are shown on a logarithmic scale, and the values are compared to the best value obtained among the different approaches in 500 epochs. In all cases, the strong convexity constant μ is the regularization parameter.

Choices of step-sizes. For both S-MISO and SGD, 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 large enough to match the previous constant step-size. For N-SAGA, we maintain a constant step-size throughout the optimization, as suggested in the original paper [9]. The factor η shown in Figure 1 controls the initial step-size and is such that $\eta = 1$ corresponds to the same linear rate $(1 - \mu/L)$ for all algorithms (or $(1 - \mu/\bar{L})$ for non-uniform sampling) in the ill-conditioned regime. Note that $\eta = 1$ corresponds to a $1/L$ step-size for N-SAGA, which is larger than allowed by the theory [9] and indeed leads to instabilities in contrast to S-MISO and SGD; however, the choice $\eta = 0.1$ is justified by the theory.

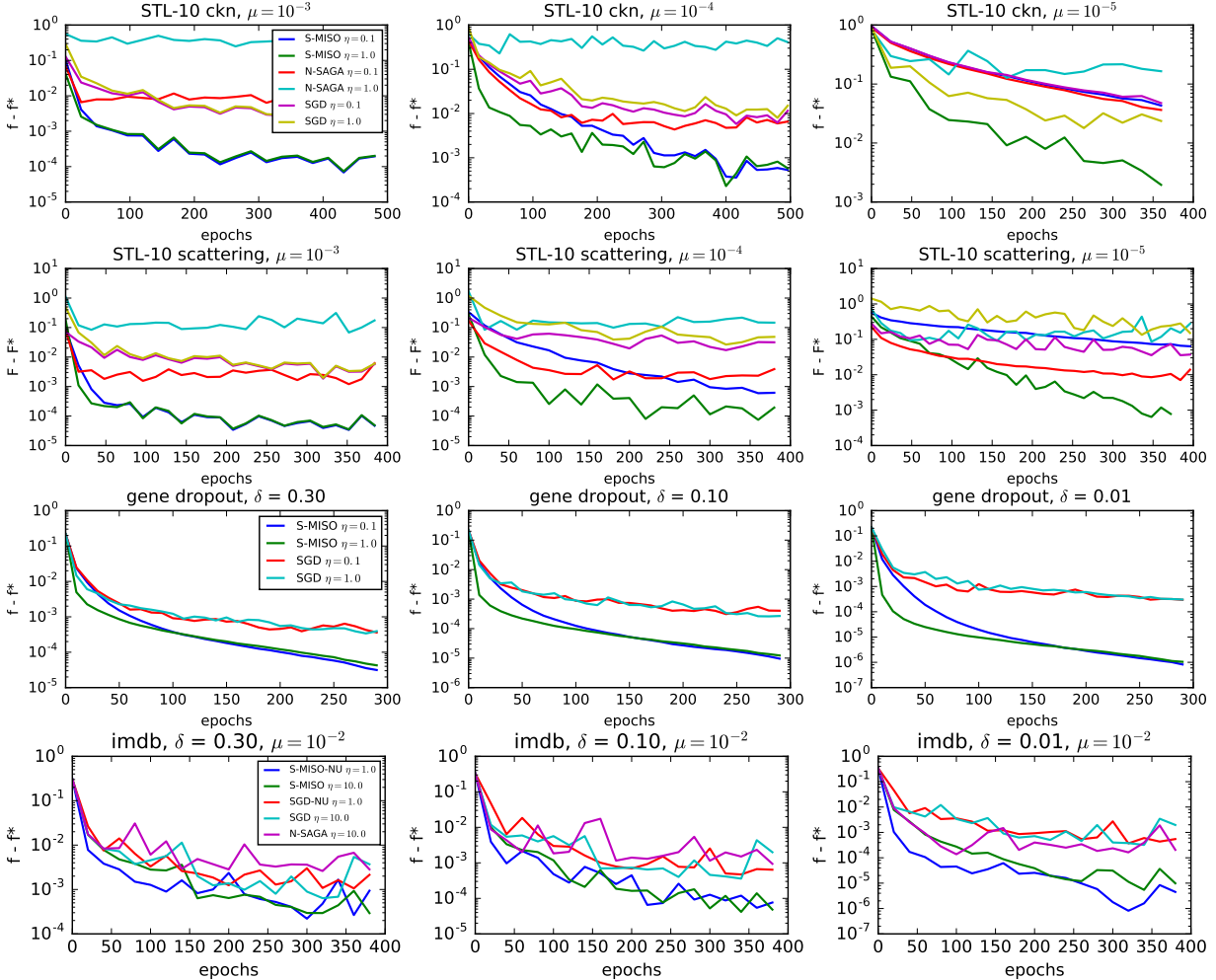


Figure 1: Comparison of S-MISO with SGD and N-SAGA [9]. Values of the training loss are shown in logarithmic scale (1 unit = factor 10). Top two rows: STL-10 dataset with different values of μ (the best value given by cross validation is around 10^{-4}), with convolutional kernel network and scattering representations. Third row: breast cancer dataset with Dropout, for different values of δ and $\mu = 0.003$ (selected by 5-fold cross validation with no Dropout). Bottom row: sentiment classification on IMDB dataset with Dropout (NU = non-uniform sampling).

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, or pre-trained on a different task (e.g., on the ImageNet dataset). We follow this approach on the STL-10 dataset [6], which contains 5 000 training images from 10 classes and 100 000 unlabeled images, using a 2-layer unsupervised convolutional kernel network [15], giving representations of dimension 9 216. The perturbation consists of randomly cropping and scaling 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. We also present results on the same dataset using the scattering representation [4], giving vectors of dimension 21 696, with random gamma corrections (raising all pixels to the power γ , where γ is chosen randomly around 1). For this representation, we add

an ℓ_1 regularization term and use the composite variant of S-MISO presented in Section 4.

Figure 1 (top two rows) 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. N-SAGA initially reaches a smaller suboptimality than SGD, but quickly gets stuck due to the bias in the algorithm, as predicted by the theory [9], while S-MISO and SGD continue to converge to the optimum thanks to the decreasing step-sizes. The best validation accuracy for both representations is obtained for $\mu \approx 10^{-4}$ (middle column in Figure 1), and we observed relative gains of up to 1% from the use of data augmentation. 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% of the total variance across multiple images.

Dropout on gene expression data. We trained a binary logistic regression model on the breast cancer dataset of Van de Vijver et al. [24], 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. The dataset consists of 8 141-dimensional vectors of gene expression data for 295 different subjects, which we normalize to unit ℓ_2 norm. Figure 1 (third row) 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.

Dropout on movie review sentiment analysis data. We trained a binary classifier with a squared hinge loss on the IMDB dataset [12] with different Dropout rates δ . We use the labeled part of the IMDB dataset, which consists of 25 000 training and 250 000 testing movie reviews, represented as 89 527-dimensional sparse bag-of-words vectors. In contrast to the previous experiments, we do not normalize the sparse representations, which have great variability in their norms, in particular the uniform smoothness constant L is roughly 100 times larger than the average constant \bar{L} . Figure 1 (bottom row) compares non-uniform sampling versions of S-MISO and SGD with their uniform sampling counterparts as well as N-SAGA. Note that we use a larger step-size than $\eta = 1$ for the uniform sampling algorithms, since $\eta = 1$ was significantly slower, likely due to outliers in the dataset. We thus allowed ourselves to tune the step-size, up to $\eta = 10$, above which the algorithms started diverging. In contrast, the non-uniform sampling algorithms required no tuning and just use $\eta = 1$. The curves clearly show that S-MISO-NU has a much faster convergence in the initial phase, thanks to the larger step-size allowed by non-uniform sampling, and later converges similarly to S-MISO, *i.e.* at a much faster rate than SGD when the perturbations are small. The value of μ used in the experiments was chosen by cross-validation, and the use of Dropout gave improvements in test accuracy from 88.51% with no dropout to $88.68 \pm 0.03\%$ with $\delta = 0.1$ and $88.86 \pm 0.11\%$ with $\delta = 0.3$ (based on 10 different runs of S-MISO-NU after 400 epochs).

6 Conclusion

In this paper, we introduced the S-MISO method, a hybrid stochastic/incremental optimization algorithm, which is able to exploit underlying finite-sum structures in stochastic optimization problems. Our approach uses variance reduction in settings where random perturbations of training examples in a finite dataset are considered during learning, thereby making the dataset infinite, and thus unsuitable for standard incremental methods. The algorithm naturally interpolates between stochastic approximation (when $n = 1$) and the variance reduction algorithm MISO-Prox for finite sums (when there are no perturbations). Our method supports composite objectives, non-uniform sampling, and gives convergence guarantees that are similar to

SGD, but with a much smaller constant term that depends on the variance of gradient estimates induced by the perturbations on a single example, rather than across the entire data.

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] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [11] H. Lin, J. Mairal, and Z. Harchaoui. A Universal Catalyst for First-Order Optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 142–150. Association for Computational Linguistics, 2011.
- [13] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [14] J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.
- [15] J. Mairal. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [16] 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.
- [17] 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.

- [18] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2004.
- [19] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2016.
- [20] S. Shalev-Shwartz. SDCA without Duality, Regularization, and Individual Convexity. In *International Conference on Machine Learning (ICML)*, 2016.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] L. van der Maaten, M. Chen, S. Tyree, and K. Q. Weinberger. Learning with marginalized corrupted features. In *International Conference on Machine Learning (ICML)*, 2013.
- [26] 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.
- [27] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
- [28] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

A Proof of Proposition 5 (Composite Objectives and Non-Uniform Sampling)

We begin by presenting a lemma that plays a similar role to Lemma 1 in our proof, but considers the composite objective and takes into account the new strong convexity and non-uniformity assumptions.

Lemma A.1. *Let $i \sim q$, where q is the sampling distribution, and $\rho \sim \Gamma$. Under assumptions (A4), (A5) and (A6) on the functions $\tilde{f}_1, \dots, \tilde{f}_n$ and their expectations f_1, \dots, f_n , we have, for all $x \in \mathbb{R}^p$,*

$$\mathbb{E}_{i,\rho} \left[\frac{1}{(q_i n)^2} \|\nabla \tilde{f}_i(x, \rho) - \mu x - (\nabla f_i(x^*) - \mu x^*)\|^2 \right] \leq 4L_q(F(x) - F(x^*)) + 2\sigma_q^2,$$

with $L_q = \max_i \frac{L_i - \mu}{q_i n}$ and $\sigma_q^2 = \frac{1}{n} \sum_i \frac{\sigma_i^2}{q_i n}$.

Proof. Since $\tilde{f}_i(\cdot, \rho)$ is μ -strongly convex and L_i -smooth, we have that $\tilde{f}_i(\cdot, \rho) - \frac{\mu}{2} \|\cdot\|^2$ is convex and $(L_i - \mu)$ -smooth (this is a straightforward consequence of [18, Eq. (2.1.9) and (2.1.22)]). Then, we have

$$\begin{aligned} & \|\nabla \tilde{f}_i(x, \rho) - \mu x - (\nabla f_i(x^*) - \mu x^*)\|^2 \\ & \leq 2\|\nabla \tilde{f}_i(x, \rho) - \mu x - (\nabla \tilde{f}_i(x^*, \rho) - \mu x^*)\|^2 + 2\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2 \\ & \leq 4(L_i - \mu) \left(\tilde{f}_i(x, \rho) - \frac{\mu}{2} \|x\|^2 - \tilde{f}_i(x^*, \rho) + \frac{\mu}{2} \|x^*\|^2 - \langle \nabla \tilde{f}_i(x^*, \rho) - \mu x^*, x - x^* \rangle \right) + 2\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2 \\ & = 4(L_i - \mu) \left(\tilde{f}_i(x, \rho) - \tilde{f}_i(x^*, \rho) - \langle \nabla \tilde{f}_i(x^*, \rho), x - x^* \rangle - \frac{\mu}{2} \|x - x^*\|^2 \right) + 2\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2 \\ & \leq 4(L_i - \mu) \left(\tilde{f}_i(x, \rho) - \tilde{f}_i(x^*, \rho) - \langle \nabla \tilde{f}_i(x^*, \rho), x - x^* \rangle \right) + 2\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2. \end{aligned}$$

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 convexity and $(L_i - \mu)$ -smoothness of $\tilde{f}_i(\cdot, \rho) - \frac{\mu}{2} \|\cdot\|^2$. Dividing by $(q_i n)^2$ and taking expectations yields

$$\begin{aligned} & \mathbb{E}_{i,\rho} \left[\frac{1}{(q_i n)^2} \|\nabla \tilde{f}_i(x, \rho) - \mu x - (\nabla f_i(x^*) - \mu x^*)\|^2 \right] \\ & \leq 4 \sum_{i=1}^n \frac{q_i(L_i - \mu)}{(q_i n)^2} (f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle) + 2 \sum_{i=1}^n \frac{q_i}{(q_i n)^2} \sigma_i^2 \\ & = 4 \frac{1}{n} \sum_{i=1}^n \frac{L_i - \mu}{q_i n} (f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle) + 2 \frac{1}{n} \sum_{i=1}^n \frac{\sigma_i^2}{q_i n} \\ & \leq 4L_q(f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle) + 2\sigma_q^2 \\ & \leq 4L_q(f(x) - f(x^*) + h(x) - h(x^*)) + 2\sigma_q^2 = 4L_q(F(x) - F(x^*)) + 2\sigma_q^2, \end{aligned}$$

where the last inequality follows from the optimality of x^* , which implies that $-\nabla f(x^*) \in \partial h(x^*)$, and in turn implies $\langle -\nabla f(x^*), x - x^* \rangle \leq h(x) - h(x^*)$ by convexity of h . \square

We can now proceed with the proof of Proposition 5.

Proof. Define the quantities

$$A_t = \frac{1}{n} \sum_{i=1}^n \frac{1}{q_i n} \|z_i^t - z_i^*\|^2$$

and $B_t = F(x^*) - D_t(x_t)$.

The proof successively describes recursions on A_t , B_t , and eventually C_t (we drop the superscript in C_t^q for simplicity), using the same approach as for the proof of Proposition 2.

Recursion on A_t . Using similar techniques as in the proof of Proposition 2, we have

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

where $v_i^t := x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_i(x_{t-1}, \rho_t) - z_i^*$. Taking conditional expectations w.r.t. \mathcal{F}_{t-1} and using Lemma A.1 to bound the second term yields

$$\mathbb{E}[A_t - A_{t-1}] \leq -\frac{\alpha_t}{n} A_{t-1} + \frac{4\alpha_t L_q}{n\mu^2} (F(x_{t-1}) - F(x^*)) + \frac{2\alpha_t \sigma_q^2}{n\mu^2} - \frac{1}{n} \sum_{i=1}^n \left(\frac{\alpha_t}{n} \frac{1}{q_i n} \left(1 - \frac{\alpha_t}{q_i n} \right) \|v_i^t\|^2 \right) \quad (27)$$

Recursion on B_t . We start by using a lemma from the proof of MISO-Prox [11, Lemma D.4], which only relies on the form of D_t and the fact that x_t minimizes it, and thus holds in our setting:

$$\begin{aligned}
D_t(x_t) &\geq D_t(x_{t-1}) - \frac{\mu}{2} \|\bar{z}_t - \bar{z}_{t-1}\|^2 \\
&= D_t(x_{t-1}) - \frac{\mu}{2(q_{i_t} n)^2} \left(\frac{\alpha_t}{n} \right)^2 \|v_{i_t}^t\|^2
\end{aligned} \quad (28)$$

We then expand $D_t(x_{t-1})$ using (19) as follows:

$$\begin{aligned}
D_t(x_{t-1}) &= D_{t-1}(x_{t-1}) + \frac{\alpha_t}{n} \frac{1}{q_{i_t} n} (\tilde{f}_{i_t}(x_{t-1}, \rho_t) - d_{i_t}^{t-1}(x_{t-1})) \\
&= D_{t-1}(x_{t-1}) + \frac{\alpha_t}{n} \frac{1}{q_{i_t} n} (\tilde{f}_{i_t}(x_{t-1}, \rho_t) + h(x_{t-1}) - d_{i_t}^{t-1}(x_{t-1}) - h(x_{t-1})).
\end{aligned}$$

After taking conditional expectations w.r.t. \mathcal{F}_{t-1} , (28) becomes

$$\mathbb{E}[D_t(x_t)] \geq D_{t-1}(x_{t-1}) + \frac{\alpha_t}{n} (F(x_{t-1}) - D_{t-1}(x_{t-1})) - \frac{\mu}{2n} \sum_{i=1}^n \left(\frac{\alpha_t}{n} \right)^2 \frac{1}{q_i n} \|v_i^t\|^2.$$

Subtracting $F(x^*)$ and rearranging yields

$$\mathbb{E}[B_t - B_{t-1}] \leq -\frac{\alpha_t}{n} B_{t-1} - \frac{\alpha_t}{n} (F(x_{t-1}) - F(x^*)) + \frac{\mu}{2n} \sum_{i=1}^n \left(\frac{\alpha_t}{n} \right)^2 \frac{1}{q_i n} \|v_i^t\|^2. \quad (29)$$

Recursion on C_t . If we consider $C_t = \mu p_t A_t + B_t$ and $C'_{t-1} = \mu p_t A_{t-1} + B_{t-1}$, combining (27) and (29) yields

$$\mathbb{E}[C_t - C'_{t-1}] \leq -\frac{\alpha_t}{n} C'_{t-1} + \frac{2\alpha_t}{n} (2p_t L_q / \mu - \frac{1}{2}) (F(x_{t-1}) - F(x^*)) + \frac{\mu \alpha_t}{n^2} \sum_{i=1}^n \frac{\delta_i^t}{q_i n} \|v_i^t\|^2 + \frac{2\alpha_t p_t \sigma_q^2}{n\mu},$$

with

$$\delta_i^t = \frac{\alpha_t}{2n} - p_t \left(1 - \frac{\alpha_t}{q_i n} \right).$$

If we take $p_t = \frac{\alpha_t}{n}$, and if $(\alpha_t)_{t \geq 1}$ is a decreasing sequence satisfying (24), then we obtain the desired recursion after noticing that $C'_{t-1} \leq C_{t-1}$ and taking total expectations on \mathcal{F}_{t-1} . \square

B SGD Recursions

Proposition B.1 (Simple SGD recursion with variance at optimum). *If $\eta_t \leq 1/2L$, then the SGD recursion $x_t := x_{t-1} - \eta_t \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)$ satisfies*

$$B_t \leq (1 - \mu\eta_t)B_{t-1} + \eta_t^2 \sigma_{tot}^2,$$

where $B_t := \frac{1}{2} \mathbb{E}[\|x_t - x^*\|^2]$ and σ_{tot} is such that

$$\mathbb{E}_{i,\rho} [\|\nabla \tilde{f}_i(x^*, \rho)\|^2] \leq \sigma_{tot}^2.$$

Proof. We have

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

where inequality (*) follows from the strong convexity of f and $\mathbb{E}_{i,\rho}[\|\tilde{f}_{i_t}(x_{t-1}, \rho_t) - \tilde{f}_{i_t}(x^*, \rho_t)\|^2]$ is bounded as in Lemma 1. When $\eta_t \leq 1/2L$, the second term is non-positive and we obtain the desired result after taking total expectations. \square