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

June 1, 2017

Abstract

Stochastic optimization algorithms with variance reduction have proven successful for minimizing large finite sums of functions. Unfortunately, these techniques are unable to deal with stochastic perturbations of input data, induced for example by data augmentation. In such cases, 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 these settings when the objective is composite and strongly convex. The convergence rate outperforms SGD with a typically much smaller constant factor, which depends on the variance of gradient estimates only due to perturbations on a *single* example.

1 Introduction

Many supervised machine learning problems can be cast as the minimization of an expected loss over a data distribution with respect to a vector x in $\mathbb{R}^p$ of model parameters. When an infinite amount of data is available, stochastic optimization methods such as SGD or stochastic mirror descent algorithms, or their variants, are typically used (see [4, 10, 23, 33]). Nevertheless, when the dataset is finite, incremental methods based on variance reduction techniques (*e.g.*, [2, 7, 14, 16, 17, 26, 28]) have proven to be significantly faster at solving the finite-sum problem

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

where the functions f_i are smooth and convex, and h is a simple convex penalty that need not be differentiable such as 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.

In this paper, we are interested in a variant of (1), where random perturbations of data are introduced, which is a fundamental concept in machine learning; then, the functions f_i are stochastic and involve an expectation over a random perturbation ρ, leading to the problem

$$\min_{x\in\mathbb{R}^p} \Big\{ F(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) + h(x) \Big\}. \quad \text{with} \quad f_i(x) = \mathbb{E}_\rho[\tilde{f}_i(x,\rho)]. \tag{2}$$

Unfortunately, variance reduction methods are not compatible with the setting (2), since evaluating a single gradient $\nabla f_i(x)$ requires computing a full expectation. Yet, dealing with random perturbations is of

*This work was supported by a grant from ANR (MACARON project under grant number ANR-14-CE23-0003-01), by the ERC grant number 714381 (SOLARIS project), and from the MSR-Inria joint centre.

utmost interest; for instance, this is a key to achieve stable feature selection [22], improving the generalization error both in theory [32] and in practice [18, 31], obtaining stable and robust predictors [35], or using complex a priori knowledge about data to generate virtually larger datasets [18, 25, 29]. Injecting noise in data is also useful to hide gradient information for privacy-aware learning [9].

Despite its importance, the optimization problem (2) has been littled studied and to the best of our knowledge, no dedicated optimization method that is able to exploit the problem structure has been developed so far. A natural way to optimize this objective when $h = 0$ is indeed SGD, but ignoring the finite-sum structure leads to gradient estimates with high variance and slow convergence. The goal of this paper is to introduce an algorithm for strongly convex objectives, called *stochastic MISO*, which exploits the underlying finite sum using variance reduction. Our method achieves a faster convergence rate than SGD, by removing the dependence on the gradient variance due to sampling the data points i in $\{1, \dots, n\}$; the dependence remains only for the variance due to random perturbations ρ.

To the best of our knowledge, our method is the first algorithm that interpolates naturally between incremental methods for finite sums (when there are no perturbations) and the stochastic approximation setting (when $n = 1$), while being able to efficiently tackle the hybrid case.

Related work. Many optimization methods dedicated to the finite-sum problem (*e.g.*, [14, 28]) have been motivated by the fact that their updates can be interpreted as SGD steps with unbiased estimates of the full gradient, but with a variance that decreases as the algorithm approaches the optimum [14]; on the other hand, vanilla SGD requires decreasing step-sizes to achieve this reduction of variance, thereby slowing down convergence. Our work aims at extending these techniques to the case where each function in the finite sum can only be accessed via a first-order stochastic oracle.

Most related to our work, recent methods that use data clustering to accelerate variance reduction techniques [3, 13] can be seen as tackling a special case of (2), where the expectations in f_i are replaced by empirical averages over points in a cluster. While N-SAGA [13] was originally not designed for the stochastic context we consider, we remark that their method can be applied to (2). Their algorithm is however asymptotically biased and does not converge to the optimum. On the other hand, ClusterSVRG [3] is not biased, but does not support infinite datasets. The method proposed in [1] uses variance reduction in a setting where gradients are computed approximately, but the algorithm computes a full gradient at every pass, which is not available in our stochastic setting.

Paper organization. In Section 2, we present our algorithm for smooth objectives, and we analyze its convergence in Section 3. We present an extension to composite objectives and non-uniform sampling in Section 4. Section 5 is devoted to empirical results.

2 The Stochastic MISO Algorithm for Smooth Objectives

In this section, we introduce the *stochastic MISO* approach for smooth objectives ($h = 0$), 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.*, with L-Lipschitz gradients).

Note that these assumptions are relaxed in Section 4 by supporting composite objectives 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 convergence (*e.g.*, [34]).

Complexity results. We now introduce the following quantity, which is essential in our analysis:

$$\sigma_p^2 := \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2, \quad \text{with} \quad \sigma_i^2 := \mathbb{E}_\rho \left[\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2 \right],$$

Table 1: Iteration complexity of different methods for solving the objective (2) in terms of number of iterations required to minimize the objective up to accuracy ε. The complexity of N-SAGA [13] matches the first term of S-MISO but is asymptotically biased. Note that we always have the perturbation noise variance σ_p^2 smaller than the total variance σ_{tot}^2 and thus S-MISO improves on SGD both in the first term (linear convergence to a smaller $\bar{\epsilon}$) and in the second (smaller constant in the asymptotic rate). In many application cases, we also have $\sigma_p^2 \ll \sigma_{\text{tot}}^2$ (see main text and Table 2).

Method	Asymptotic error	Iteration complexity
SGD	0	$O\left(\frac{L}{\mu}\log\frac{1}{\bar{\epsilon}} + \frac{\sigma_{\text{tot}}^2}{\mu\epsilon}\right)$ with $\bar{\epsilon} = O\left(\frac{\sigma_{\text{tot}}^2}{\mu}\right)$
N-SAGA [13]	$\epsilon_0 = O\left(\frac{\sigma_p^2}{\mu}\right)$	$O\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$ with $\epsilon > \epsilon_0$
S-MISO	0	$O\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\bar{\epsilon}} + \frac{\sigma_p^2}{\mu\epsilon}\right)$ with $\bar{\epsilon} = O\left(\frac{\sigma_p^2}{\mu}\right)$

Table 2: Estimated ratio $\sigma_{\text{tot}}^2/\sigma_p^2$, which corresponds to the expected acceleration of S-MISO over SGD. These numbers are based on feature vectors variance, which is closely related to the gradient variance when learning a linear model. For direct perturbations, we consider standardized data (zero mean, unit standard deviation). ResNet-50 denotes a 50 layer network [11] pre-trained on the ImageNet dataset. For image transformations, the numbers are empirically evaluated from 100 different images, with 100 random perturbations for each image. R_{tot} and R_{cluster} denote the radius of the data and the average radius of clusters, respectively, following [13]. The settings for unsupervised CKN and Scattering are described in Section 5. More details are also given in the main text.

Type of perturbation	**Application case**	**Estimated ratio $\sigma_{\text{tot}}^2/\sigma_p^2$**
Direct perturbation of linear model features	Data clustering as in [3, 13]	$\approx R_{\text{tot}}^2/R_{\text{cluster}}^2$
	Additive Gaussian noise $\mathcal{N}(0, \alpha^2 I)$	$\approx 1 + 1/\alpha^2$
	Dropout with probability δ	$\approx 1 + 1/\delta$
	Feature rescaling by s in $\mathcal{U}(1-w, 1+w)$	$\approx 1 + 3/w^2$
Random image transformations	ResNet-50 [11], color perturbation	21.9
	ResNet-50 [11], rescaling + crop	13.6
	Unsupervised CKN [21], rescaling + crop	9.6
	Scattering [5], gamma correction	9.8

where x^* is the (unique) minimizer of f. The quantity σ_p^2 represents the part of the variance of the gradients at the optimum, which is due to the perturbations ρ. In contrast, another quantity of interest is the total variance σ_{tot}^2, which also includes the randomness in the choice of the index i, defined as

$$\sigma_{\text{tot}}^2 = \mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_i(x^*, \rho)\|^2] = \sigma_p^2 + \mathbb{E}_i[\|\nabla f_i(x^*)\|^2] \qquad \text{(note that } \nabla f(x^*) = 0\text{)}.$$

The relation between σ_{tot}^2 and σ_p^2 is obtained by simple algebraic manipulations.

The goal of our paper is to exploit the potential imbalance $\sigma_p^2 \ll \sigma_{\text{tot}}^2$, occurring when perturbations on input data are small compared to the sampling noise. The assumption is reasonable: given a data point, selecting a different one should lead to larger variation than a simple perturbation. From a theoretical point of view, the approach we propose achieves the iteration complexity presented in Table 1, see also Appendix C and [4, 23] for the complexity analysis of SGD. The gain over SGD is of order $\sigma_{\text{tot}}^2/\sigma_p^2$, which is also observed in our experiments in Section 5. We also compare against the method N-SAGA; its complexity matches ours up to some unimprovable asymptotic bias.

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 uniformly at random, a perturbation ρ_t, and update

$$z_i^t = \begin{cases} (1-\alpha_t) z_i^{t-1} + \alpha_t (x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise.} \end{cases} \tag{3}$$

$$x_t = \frac{1}{n}\sum_{i=1}^n z_i^t = x_{t-1} + \frac{1}{n}(z_{i_t}^t - z_{i_t}^{t-1}). \tag{4}$$

end for

Motivation from applications cases. One clear framework of application is the data clustering scenario already investigated in [3, 13]. Nevertheless, we will focus on less-studied data augmentation settings that lead instead to true stochastic formulations such as (2). First, we consider learning a linear model when adding simple direct manipulations of feature vectors, via rescaling (multiplying each entry vector by a random scalar), Dropout, or additive Gaussian noise, in order to improve the generalization error [32], or to get more stable estimators [22]. In Table 2, we present the potential gain over SGD in these scenarios. To do that, we study the variance of perturbations applied to a feature vector ξ. Indeed, the gradient of the loss is proportional to ξ, which allows us to obtain good estimates of the ratio $\sigma_{\text{tot}}^2/\sigma_p^2$, as we observed in our empirical study of Dropout presented in Section 5. Whereas some perturbations are friendly for our method such as feature rescaling (a rescaling window of $[0.9, 1.1]$ yields for instance a huge gain factor of 300), a large Dropout rate would lead to less impressive acceleration (*e.g.*, a Dropout with $\delta = 0.5$ simply yields a factor 2).

Second, we also consider more interesting domain-driven data perturbations such as classical image transformations considered in computer vision [25, 35] including image cropping, rescaling, brightness, contrast, hue, and saturation changes. These transformations may be used to train a linear classifier on top of an unsupervised multilayer image model such as unsupervised CKNs [21] or the scattering transform [5]. It may also be used for retraining the last layer of a pre-trained deep neural network: given a new task unseen during the full network training and limited amount of training data, data augmentation may be indeed crucial to obtain good prediction and S-MISO can help accelerate learning in this setting. These scenarios are also studied in Table 2, where the experiment with ResNet-50 involving random cropping and rescaling produces 224×224 images from 256×256 ones. For these scenarios with realistic perturbations, the potential gain varies from 10 to 20.

Description of stochastic MISO. We are now in shape to present our method, described in Algorithm 1. Without perturbations and with a constant step-size, the algorithm resembles the MISO/Finito algorithms [8, 17, 20], which may be seen as primal variants of SDCA [27, 28]. Specifically, MISO is not able to deal with our stochastic objective (2), but it may address the deterministic finite-sum problem (1). It is part of a larger body of optimization methods that iteratively build a *model* of the objective function, typically a lower or upper bound on the objective that is easier to optimize; for instance, this strategy is commonly adopted in bundle methods [12, 24].

More precisely, MISO assumes that each f_i is strongly convex and builds a model using lower bounds $D_t(x) = \frac{1}{n}\sum_{i=1}^n d_i^t(x)$, where each d_i^t is a quadratic lower bound on f_i of the form

$$d_i^t(x) = c_{i,1}^t + \frac{\mu}{2}\|x - z_i^t\|^2 = c_{i,2}^t - \mu\langle x, z_i^t\rangle + \frac{\mu}{2}\|x\|^2. \tag{5}$$

These lower bounds are updated during the algorithm using strong convexity lower bounds at x_{t-1} of the

form $l_i^t(x) = f_i(x_{t-1}) + \langle \nabla f_i(x_{t-1}), x - x_{t-1} \rangle + \frac{\mu}{2}\|x - x_{t-1}\|^2 \leq f_i(x)$:

$$d_i^t(x) = \begin{cases} (1-\alpha_t) d_i^{t-1}(x) + \alpha_t l_i^t(x), & \text{if } i = i_t \\ d_i^{t-1}(x), & \text{otherwise,} \end{cases} \tag{6}$$

which corresponds to an update of the quantity z_i^t:

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

The next iterate is then computed as $x_t = \arg\min_x D_t(x)$, which is equivalent to (4). The original MISO/Finito algorithms use $\alpha_t = 1$ under a "big data" condition on the sample size n [8, 20], while the theory was later extended in [17] 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 (see [27]).

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. For this reason, we have introduced S-MISO, which can exploit *approximate* lower bounds to each f_i using gradient estimates, by letting the step-sizes α_t decrease appropriately as commonly done in stochastic approximation. This leads to update (3).

Separately, SDCA [28] 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, [27] gives an analysis of SDCA in the primal, aka. "without duality", for smooth finite sums, and our work extends this line of reasoning to the stochastic approximation and composite settings.

Relationship with SGD in the smooth case. The link between S-MISO in the non-composite setting and SGD can be seen by rewriting the update (4) 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

$$v_t := x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^{t-1}. \tag{7}$$

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 [14] and later in other variance reduction algorithms [7, 27]. It is also worth noting that in the absence of a finite-sum structure ($n=1$), we have $z_{i_t}^{t-1} = x_{t-1}$, hence our method becomes identical to SGD, up to a redefinition of step-sizes. In the composite case (see Section 4), our approach yields a new algorithm that resembles regularized dual averaging [33].

Memory requirements and handling of sparse datasets. The algorithm requires storing the vectors $(z_i^t)_{i=1,\dots,n}$, which takes the same amount of memory as the original dataset and which is therefore a reasonable requirement in many practical cases. In the case of sparse datasets, it is fair to assume that random perturbations applied to 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 [32]. 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 ξ_i^ρ is the (sparse) perturbed example and ϕ_i encodes the loss, then the update (3) 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 ξ_i^ρ throughout the algorithm (assuming the initialization $z_i^0 = 0$), making the update (3) efficient. The update (4) has the same cost since $v_t = z_{i_t}^t - z_{i_t}^{t-1}$ is also sparse.

Limitations and alternative approaches. Since our algorithm is uniformly better than SGD in terms of iteration complexity, its main limitation is in terms of memory storage when the dataset cannot fit into memory (remember that the memory cost of S-MISO is the same as the input dataset). In these huge-scale settings, SGD should be preferred; this holds true in fact for all incremental methods when one cannot afford to perform more than one (or very few) passes over the data. Our paper focuses instead on non-huge datasets, which are those benefiting most from data augmentation.

We note that a different approach to variance reduction like SVRG [14] is able to trade off storage requirements for additional full gradient computations, which would be desirable in some situations. However, we were not able to obtain any decreasing step-size strategy that works for these methods, both in theory and practice, leaving us with constant step-size approaches as in [1, 13] that either maintain a non-zero asymptotic error, or require dynamically reducing the variance of gradient estimates. One possible way to explain this difficulty is that SVRG and SAGA [7] "forget" past gradients for a given example i, while S-MISO averages them in (3), which seems to be a technical key to make it suitable to stochastic approximation. Nevertheless, the question of whether it is possible to trade-off storage with computation in a setting like ours is open and of utmost interest.

3 Convergence Analysis of S-MISO

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

$$C_t = \frac{1}{2}\|x_t - x^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^{n}\|z_i^t - z_i^*\|^2. \tag{8}$$

Proposition 1 gives a recursion on C_t, obtained by upper-bounding separately its two terms, and finding coefficients to cancel out other appearing quantities when relating C_t to C_{t-1}. To this end, we borrow elements of the convergence proof of SDCA without duality [27]; our technical contribution is to extend their result to the stochastic approximation and composite (see Section 4) cases.

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

$$\alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{2(2\kappa - 1)}\right\}, \tag{9}$$

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

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}] + 2\left(\frac{\alpha_t}{n}\right)^2\frac{\sigma_p^2}{\mu^2}. \tag{10}$$

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 [4] for SGD. Note that convergence of objective function values is directly related to that of the Lyapunov function C_t via smoothness:[1]

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{L}{2}\,\mathbb{E}\left[\|x_t - x^*\|^2\right] \leq L\,\mathbb{E}[C_t]. \tag{11}$$

Theorem 2 (Convergence of Lyapunov function)**.** *Let the sequence of step-sizes $(\alpha_t)_{t\geq 1}$ be defined by $\alpha_t = \frac{\beta n}{\gamma + t}$ with $\beta > 1$ and $\gamma \geq 0$ such that α_1 satisfies (9). For all $t \geq 0$, it holds that*

$$\mathbb{E}[C_t] \leq \frac{\nu}{\gamma + t + 1} \quad \text{where} \quad \nu := \max\left\{\frac{2\beta^2\sigma_p^2}{\mu^2(\beta - 1)}, (\gamma + 1)C_0\right\}. \tag{12}$$

[1] Note 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.

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 (12). 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 [4] 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_p, and then start decaying the step-size. It is easy to see that by using a constant step-size $\bar{\alpha}$, C_t converges near a value $\bar{C} := 2\bar{\alpha}\sigma_p^2/n\mu^2$. Indeed, Eq. (10) with $\alpha_t = \bar{\alpha}$ yields

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

Thus, we can reach a precision 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. Then, if we start decaying step-sizes as in Theorem 2 with $\beta = 2$ and γ large enough so that $\alpha_1 = \bar{\alpha}$, we have

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

making both terms in (12) smaller than or equal to $\nu = 8\sigma_p^2/\mu^2$. Considering these two phases, with an initial step-size $\bar{\alpha}$ given by (9), the final work complexity for reaching $\mathbb{E}[\|x_t - x^*\|^2] \leq \epsilon$ is

$$O\left(\left(n + \frac{L}{\mu}\right) \log \frac{C_0}{\bar{\epsilon}}\right) + O\left(\frac{\sigma_p^2}{\mu^2 \epsilon}\right). \tag{13}$$

We can then use (11) in order to obtain the complexity for reaching $\mathbb{E}[f(x_t) - f(x^*)] \leq \epsilon$. Note that following this step-size strategy was found to be very effective in practice (see Section 5).

Acceleration by iterate averaging. When one is interested in the convergence in function values, the complexity (13) combined with (11) yields $O(L\sigma_p^2/\mu^2\epsilon)$, which can be problematic for ill-conditioned problems (large condition number L/μ). The following theorem presents an iterate averaging scheme which brings the complexity term down to $O(\sigma_p^2/\mu\epsilon)$, which appeared in Table 1.

Theorem 3 (Convergence under iterate averaging)**.** *Let the step-size sequence $(\alpha_t)_{t\geq 1}$ be defined by*

$$\alpha_t = \frac{2n}{\gamma + t} \quad \text{for } \gamma \geq 1 \text{ s.t. } \alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{4(2\kappa - 1)}\right\}.$$

We have

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{2\mu\gamma(\gamma - 1)C_0}{T(2\gamma + T - 1)} + \frac{16\sigma_p^2}{\mu(2\gamma + T - 1)},$$

where

$$\bar{x}_T := \frac{2}{T(2\gamma + T - 1)} \sum_{t=0}^{T-1} (\gamma + t) x_t.$$

The proof uses a similar telescoping sum technique to [15]. Note that if $T \gg \gamma$, the first term, which depends on the initial condition C_0, decays as $1/T^2$ and is thus dominated by the second term. Moreover, if we start averaging after an initial phase with constant step-size $\bar{\alpha}$, we can consider $C_0 \approx 4\bar{\alpha}\sigma_p^2/n\mu^2$. In the ill-conditioned regime, taking $\bar{\alpha} = \alpha_1 = 2n/(\gamma + 1)$ as large as allowed we have γ of the order of $\kappa = L/\mu \gg 1$. The full convergence rate then becomes

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq O\left(\frac{\sigma_p^2}{\mu(\gamma + T)}\left(1 + \frac{\gamma}{T}\right)\right).$$

When T is large enough compared to γ, this becomes $O(\sigma_p^2/\mu T)$, leading to a complexity $O(\sigma_p^2/\mu\epsilon)$.

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

Input: step-sizes $(\alpha_t)_{t\geq 1}$, sampling distribution q;
Initialize $x_0 = \mathrm{prox}_{h/\mu}(\bar{z}_0)$ with $\bar{z}_0 = \frac{1}{n}\sum_i z_i^0$ for some $(z_i^0)_{i=1,\dots,n}$ that safisfies (16);
for $t = 1, \dots$ **do**
Sample an index $i_t \sim q$, a perturbation ρ_t, and update (with $\alpha_t^i = \frac{\alpha_t}{q_i n}$):

$$z_i^t = \begin{cases} (1-\alpha_t^i) z_i^{t-1} + \alpha_t^i (x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise} \end{cases} \tag{14}$$

$$\bar{z}_t = \frac{1}{n}\sum_{i=1}^n z_i^t = \bar{z}_{t-1} + \frac{1}{n}(z_{i_t}^t - z_{i_t}^{t-1})$$

$$x_t = \mathrm{prox}_{h/\mu}(\bar{z}_t). \tag{15}$$

end for

4 Extension to Composite Objectives and Non-Uniform Sampling

In this section, we study extensions of S-MISO 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 such as the norm of each example.

In the presence of non-smooth regularizers such as the ℓ_1-norm, 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

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

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.*, [26, 34]) 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 now make the following assumptions:

- (A3) **strong convexity**: $\tilde{f}_i(\cdot, \rho)$ is μ-strongly convex for all i, ρ;
- (A4) **smoothness**: $\tilde{f}_i(\cdot, \rho)$ is L_i-smooth for all i, ρ;

Note that our proof relies on a slightly stronger assumption (A3) than the global strong convexity assumption (A1) made above, which holds in the 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 setting is given in Algorithm 2. Note that the step-sizes vary depending on the example, with larger steps 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 [34] or SAGA [7].

Relationship with RDA. When $n = 1$, our algorithm performs similar updates to Regularized Dual Averaging (RDA) [33] 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

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

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 [33, Section 2.2]. However, unlike RDA, our method supports arbitrary decreasing step-sizes, in particular keeping the step-size constant, which can lead to faster convergence in the initial iterations (see Section 3).

Lower-bound model and convergence analysis. Again, we can view the algorithm as iteratively updating approximate lower bounds on the objective F of the form $D_t(x) = \frac{1}{n}\sum_i d_i^t(x) + h(x)$ analogously to (6), and minimizing the new D_t in (15). Similar 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 the random perturbation $\tilde{\rho}_i$. Given the form of d_i^t in (5), it suffices to choose z_i^0 that satisfies

$$\tilde{f}_i(x, \tilde{\rho}_i) \geq \frac{\mu}{2}\|x - z_i^0\| + c, \tag{16}$$

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:

$$C_t^q = F(x^*) - D_t(x_t) + \frac{\mu\alpha_t}{n^2}\sum_{i=1}^n \frac{1}{q_i n}\|z_i^t - z_i^*\|^2. \tag{17}$$

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

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

$$\frac{\mu}{2}\,\mathbb{E}[\|x_t - x^*\|^2] \leq \mathbb{E}[F(x^*) - D_t(x_t)]. \tag{18}$$

The following proposition gives a recursion on C_t^q similar to Proposition 1.

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

$$\alpha_1 \leq \min\left\{\frac{n q_{\min}}{2}, \frac{n\mu}{4L_q}\right\}, \tag{19}$$

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

$$\mathbb{E}[C_t^q] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}^q] + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma_q^2}{\mu}, \tag{20}$$

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 (10), and thus can apply Theorem 2 with the new condition (19). If we choose

$$q_i = \frac{1}{2n} + \frac{L_i - \mu}{2\sum_i (L_i - \mu)}, \tag{21}$$

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 (19), 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. For the complexity in function suboptimality, the second term becomes $O(\sigma_q^2/\mu\epsilon)$ by using the same averaging scheme presented in Theorem 3 and

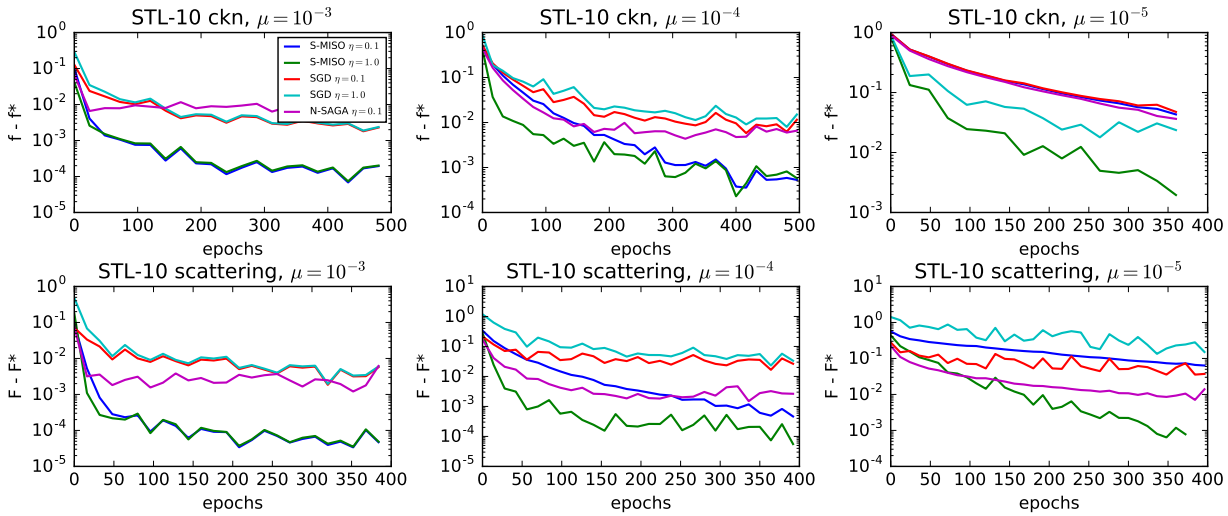

Figure 1: Impact of conditioning for data augmentation on STL-10 (controlled by μ, where $\mu = 10^{-4}$ gives the best accuracy). Values of the loss are shown in **logarithmic scale** (1 unit = factor 10). $\eta = 0.1$ satisfies the theory for all methods, and we include curves for larger step-sizes $\eta = 1$. We omit N-SAGA for $\eta = 1$ because it remains far from the optimum. For the scattering representation, the problem we study is ℓ_1-regularized, and we use the composite algorithm of Section 4.

adapting the proof. Note that with our choice of q, we have $\sigma_q^2 \leq \frac{2}{n}\sum_i \sigma_i^2 = 2\bar{\sigma_p}^2$, for general perturbations, where $\bar{\sigma_p}^2 = \frac{1}{n}\sum_i \sigma_i^2$ 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 $\sigma_i^2 = \bar{\sigma_p}^2 \frac{L_i - \mu}{\bar{L} - \mu}$, then for both $q_i = 1/n$ (uniform case) and $q_i = (L_i - \mu)/n(\bar{L} - \mu)$, we have $\sigma_q^2 = \bar{\sigma_p}^2$, and thus we have $\sigma_q^2 \leq \bar{\sigma_p}^2$ for the choice of q given in (21), since σ_q^2 is convex in q. Thus, we can expect that the $O(1/t)$ convergence phase behaves similarly or better than for uniform sampling, which is confirmed by our experiments (see Section 5).

5 Experiments

We present experiments comparing S-MISO with SGD and N-SAGA [13] on four different scenarios, in order to demonstrate the wide applicability of our method: we consider an image classification dataset with two different image representations and random transformations, and two classification tasks with Dropout regularization, one on genetic data, and one on (sparse) text data. Figures 1 and 2 show the curves for an estimate of the training objective using 5 sampled perturbations per example. The plots are shown on a logarithmic scale, and the values are compared to the best value obtained among the different methods in 500 epochs. The strong convexity constant μ is the regularization parameter. For all methods, we consider step-sizes supported by the theory as well as larger step-sizes that may work better in practice.

Choices of step-sizes. For both S-MISO and SGD, we use the step-size strategy mentioned in Section 3 and advised by [4], 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 the figures) 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 [13]. The factor η shown in the figures is such that $\eta = 1$ corresponds to an initial step-size $n\mu/(L - \mu)$ for S-MISO (from (19) in the uniform case) and $1/L$ for SGD

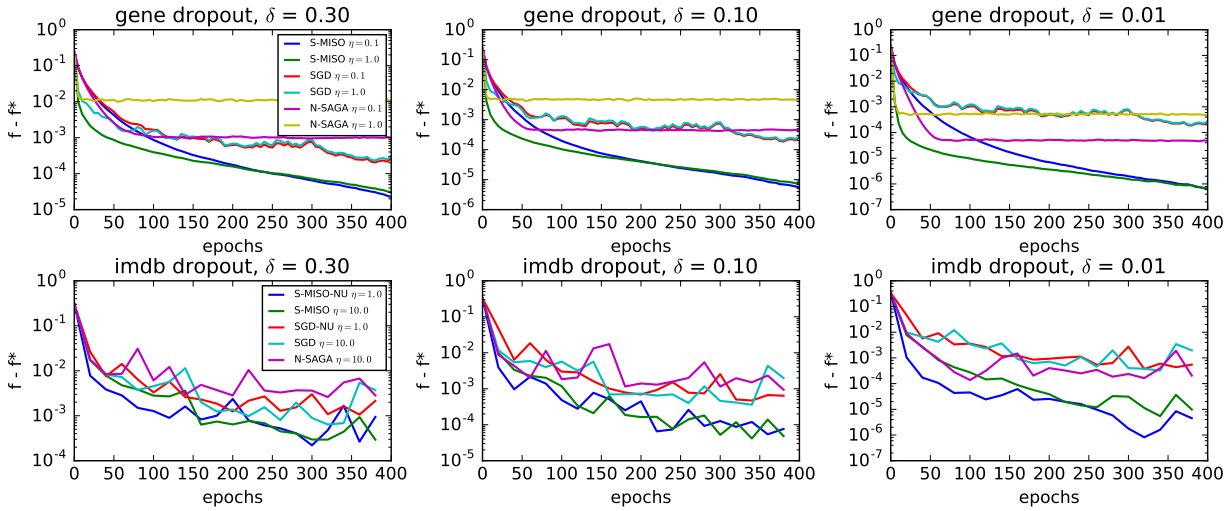

Figure 2: Impact of perturbations controlled by the Dropout rate δ. The gene data is ℓ_2-normalized; hence, we consider similar step-sizes as Figure 1. The IMDB dataset is highly heterogeneous; thus, we also include non-uniform (NU) sampling variants of Section 4. For uniform sampling, theoretical step-sizes perform poorly for all methods; thus, we show a larger tuned step-size $\eta = 10$.

and N-SAGA (with $\bar{L}$ instead of L in the non-uniform case when using the variant of Section 4).

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 5K training images from 10 classes and 100K unlabeled images, using a 2-layer unsupervised convolutional kernel network [21], giving representations of dimension 9 216. The perturbation consists of randomly cropping and scaling the input images. We use the squared hinge loss 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 a scattering representation [5] 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 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 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 [13], 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), and we observed relative gains of up to 1% from using 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.

Dropout on gene expression data. We trained a binary logistic regression model on the breast cancer dataset of [30], 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 295 vectors of dimension 8 141 of gene expression data, which we normalize in ℓ_2 norm. Figure 2 (top)

compares S-MISO with SGD and N-SAGA 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_p^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 [19] with different Dropout rates δ. We use the labeled part of the IMDB dataset, which consists of 25K training and 250K testing movie reviews, represented as 89 527-dimensional sparse bag-of-words vectors. In contrast to the previous experiments, we do not normalize the representations, which have great variability in their norms, in particular, the maximum Lipschitz constant across training points is roughly 100 times larger than the average one. Figure 2 (bottom) compares non-uniform sampling versions of S-MISO (see Section 4) and SGD with their uniform sampling counterparts as well as N-SAGA. Note that we use a large step-size $\eta = 10$ for the uniform sampling algorithms, since $\eta = 1$ was significantly slower for all methods, likely due to outliers in the dataset. 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).

Finally, we also study the effect of the iterate averaging scheme of Theorem 3 in Appendix D.

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. Katyusha: The first direct acceleration of stochastic gradient methods. *arXiv:1603.05953*, 2016.

[3] 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.

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

[5] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, 35(8):1872–1886, 2013.

[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] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[10] J. C. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research (JMLR)*, 10:2899–2934, 2009.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*. Springer science & business media, 1993.

[13] 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.

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

[15] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv:1212.2002*, 2012.

[16] G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *arXiv:1507.02000*, 2015.

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

[18] G. Loosli, S. Canu, and L. Bottou. Training invariant support vector machines using selective sampling. In *Large Scale Kernel Machines*, pages 301–320. MIT Press, Cambridge, MA., 2007.

[19] 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.

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

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

[22] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[23] 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.

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

[25] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid. Transformation pursuit for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[26] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research (JMLR)*, 11:2543–2596, 2010.

[34] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[35] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

A Proofs for the Smooth Case (Section 3)

A.1 Proof of Proposition 1 (Recursion on Lyapunov function C_t)

We begin by stating the following lemma, which extends a key result of variance reduction methods (see, *e.g.*, [14]) to the situation considered in this paper, where one only has access to noisy estimates of the gradients of each f_i.

Lemma A.1. *Let i be uniformly distributed in $\{1,\dots,n\}$ and ρ according to a perturbation distribution Γ. Under assumption (A2) on the functions $\tilde{f}_1,\dots,\tilde{f}_n$ and their expectations $f_1,\dots,f_n$, we have, for all $x\in\mathbb{R}^p$,*

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

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.*, [24, 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 the definition of σ_p^2. □

We now proceed with the proof of Proposition 1.

Proof. Define the quantities

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

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

Recursion on A_t. We have

$$\begin{aligned}A_t-A_{t-1}&=\frac{1}{n}(\|z_{i_t}^t-z_{i_t}^*\|^2-\|z_{i_t}^{t-1}-z_{i_t}^*\|^2)\\&=\frac{1}{n}\left(\left\|(1-\alpha_t)(z_{i_t}^{t-1}-z_{i_t}^*)+\alpha_t\left(x_{t-1}-\frac{1}{\mu}\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t)-z_{i_t}^*\right)\right\|^2-\|z_{i_t}^{t-1}-z_{i_t}^*\|^2\right)\\&=\frac{1}{n}\left(-\alpha_t\|z_{i_t}^{t-1}-z_{i_t}^*\|^2+\alpha_t\left\|x_{t-1}-\frac{1}{\mu}\nabla\tilde{f}_{i_t}(x_{t-1},\rho_t)-z_{i_t}^*\right\|^2-\alpha_t(1-\alpha_t)\|v_t\|^2\right),\end{aligned}\tag{22}$$

where we first use the definition of z_i^t in (3), 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 (7). A similar relation is derived in the proof of SDCA

without duality [27]. Using the definition of z_i^*, the second term can be expanded as

$$\begin{aligned}\left\|x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^*\right\|^2 &= \left\|x_{t-1} - x^* - \frac{1}{\mu}(\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - \nabla f_{i_t}(x^*))\right\|^2 \\ &= \|x_{t-1} - x^*\|^2 - \frac{2}{\mu}\langle x_{t-1} - x^*, \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - \nabla f_{i_t}(x^*)\rangle \\ &\quad + \frac{1}{\mu^2}\left\|\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - \nabla f_{i_t}(x^*))\right\|^2. \end{aligned} \tag{23}$$

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.

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

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

$$\begin{aligned}\mathbb{E}[A_t - A_{t-1}] &= -\frac{\alpha_t}{n}A_{t-1} + \frac{\alpha_t}{n}\,\mathbb{E}\left[\left\|x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t) - z_{i_t}^*\right\|^2\right] - \frac{\alpha_t(1-\alpha_t)}{n}\,\mathbb{E}[\|v_t\|^2] \\ &\leq -\frac{\alpha_t}{n}A_{t-1} + \frac{2\alpha_t(2\kappa - 1)}{n\mu}(f(x_{t-1}) - f(x^*)) - \frac{\alpha_t(1-\alpha_t)}{n}\,\mathbb{E}[\|v_t\|^2] + \frac{2\alpha_t\sigma_p^2}{n\mu^2}. \end{aligned} \tag{24}$$

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

$$\mathbb{E}[B_t - B_{t-1}] \leq -\frac{\alpha_t}{n}B_{t-1} - \frac{\alpha_t}{n\mu}(f(x_{t-1}) - f(x^*)) + \frac{1}{2}\left(\frac{\alpha_t}{n}\right)^2\mathbb{E}[\|v_t\|^2]. \tag{25}$$

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 (24) and (25) yields

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

If we take $p_t = \frac{\alpha_t}{n}$, and if $(\alpha_t)_{t\geq 1}$ is a decreasing sequence satisfying (9), then the factors in front of $f(x_{t-1}) - f(x^*)$ and $\mathbb{E}[\|v_t\|^2]$ are non-positive and we get

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

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.

□

A.2 Proof of Theorem 2 (Convergence of C_t under decreasing step-sizes)

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

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

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

A.3 Proof of Theorem 3 (Convergence in function values under iterate averaging)

Proof. From the proof of Proposition 1, we have

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}] + \frac{2\alpha_t}{n\mu}\left(\frac{\alpha_t}{n}(2\kappa - 1) - \frac{1}{2}\right)\mathbb{E}[f(x_{t-1}) - f(x^*)] + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma_p^2}{\mu^2}.$$

The result holds because the choice of step-sizes $(\alpha_t)_{t\geq 1}$ safisfies the assumptions of Proposition 1. With our new choice of step-sizes, we have the stronger bound

$$\frac{\alpha_t}{n}(2\kappa - 1) - \frac{1}{2} \leq -\frac{1}{4}.$$

After rearranging, we obtain

$$\frac{\alpha_t}{2n\mu}\mathbb{E}[f(x_{t-1}) - f(x^*)] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}] - \mathbb{E}[C_t] + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma_p^2}{\mu^2}. \tag{26}$$

Dividing by $\frac{\alpha_t}{2n\mu}$ gives

$$\begin{aligned}\mathbb{E}[f(x_{t-1}) - f(x^*)] &\leq 2\mu\left[\left(\frac{n}{\alpha_t} - 1\right)\mathbb{E}[C_{t-1}] - \frac{n}{\alpha_t}\mathbb{E}[C_t]\right] + 4\frac{\alpha_t}{n}\frac{\sigma_p^2}{\mu} \\ &= \mu\left((\gamma + t - 2)\,\mathbb{E}[C_{t-1}] - (\gamma + t)\,\mathbb{E}[C_t]\right) + \frac{8}{\gamma + t}\frac{\sigma_p^2}{\mu}.\end{aligned}$$

Multiplying by $(\gamma + t - 1)$ yields

$$\begin{aligned}&(\gamma + t - 1)\,\mathbb{E}[f(x_{t-1}) - f(x^*)] \\ &\qquad \leq \mu\left((\gamma + t - 1)(\gamma + t - 2)\,\mathbb{E}[C_{t-1}] - (\gamma + t)(\gamma + t - 1)\,\mathbb{E}[C_t]\right) + \frac{8(\gamma + t - 1)}{\gamma + t}\frac{\sigma_p^2}{\mu} \\ &\qquad \leq \mu\left((\gamma + t - 1)(\gamma + t - 2)\,\mathbb{E}[C_{t-1}] - (\gamma + t)(\gamma + t - 1)\,\mathbb{E}[C_t]\right) + \frac{8\sigma_p^2}{\mu}.\end{aligned}$$

By summing the above inequality from $t = 1$ to $t = T$, we have a telescoping sum that simplifies as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}(\gamma + t - 1)(f(x_{t-1}) - f(x^*))\right] &\leq \mu\left(\gamma(\gamma - 1)C_0 - (\gamma + T)(\gamma + T - 1)\,\mathbb{E}[C_T]\right) + \frac{8T\sigma_p^2}{\mu} \\ &\leq \mu\gamma(\gamma - 1)C_0 + \frac{8T\sigma_p^2}{\mu}.\end{aligned}$$

Dividing by $\sum_{t=1}^{T}(\gamma + t - 1) = (2T\gamma + T(T-1))/2$ and using Jensen's inequality on $f(\bar{x}_T)$ gives the desired result. □

B Proofs for Composite Objectives and Non-Uniform Sampling (Section 4)

We recall here the updates to the lower bounds d_i^t in the setting of this section, which are analogous to (6) but with non-uniform weights and stochastic perturbations,: for $i = i_t$, we have

$$d_i^t(x) = \Big(1 - \frac{\alpha_t}{q_i n}\Big)d_i^{t-1}(x) + \frac{\alpha_t}{q_i n}\Big(\tilde{f}_i(x_{t-1}, \rho_t) + \langle \nabla \tilde{f}_i(x_{t-1}, \rho_t), x - x_{t-1}\rangle + \frac{\mu}{2}\|x - x_{t-1}\|^2\Big), \tag{27}$$

and $d_i^t(x) = d_i^{t-1}(x)$ otherwise.

B.1 Proof of Lemma 4 (Bound on the iterates)

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 (16)), 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

$$F_t(x^*) \geq D_t(x^*) \geq D_t(x_t) + \frac{\mu}{2}\|x_t - x^*\|^2, \tag{28}$$

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 (28). □

B.2 Proof of Proposition 5 (Recursion on Lyapunov function C_t^q)

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

Lemma B.1. *Let $i \sim q$, where q is the sampling distribution, and ρ be a random perturbation. Under assumptions (A4-5) 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 [24, Eq. 2.1.9 and 2.1.22]). Then, by denoting by F_i the quantity $2\|\nabla \tilde{f}_i(x^*,\rho) - \nabla f_i(x^*)\|^2$, we have

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

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

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

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. □

We can now proceed with the proof of Proposition 5.

Proof. Define the quantities

$$A_t = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{q_i n}\|z_i^t - z_i^*\|^2$$
$$\text{and} \quad B_t = F(x^*) - D_t(x_t).$$

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 1.

Recursion on A_t. Using similar techniques as in the proof of Proposition 1, we have

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

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

$$\begin{aligned}
&\mathbb{E}[A_t - A_{t-1}] \leq \\
&\qquad -\frac{\alpha_t}{n} A_{t-1} + \frac{4\alpha_t L_q}{n\mu^2}(F(x_{t-1}) - F(x^*)) + \frac{2\alpha_t \sigma_q^2}{n\mu^2} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\alpha_t}{n}\frac{1}{q_i n}\left(1 - \frac{\alpha_t}{q_i n}\right)\|v_i^t\|^2\right) \qquad (29)
\end{aligned}$$

Recursion on B_t. We start by using a lemma from the proof of MISO-Prox [17, 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:

$$\begin{aligned}
D_t(x_t) &\geq D_t(x_{t-1}) - \frac{\mu}{2}\|\bar{z}_t - \bar{z}_{t-1}\|^2 \\
&= D_t(x_{t-1}) - \frac{\mu}{2(q_{i_t} n)^2}\left(\frac{\alpha_t}{n}\right)^2 \|v_{i_t}^t\|^2 \qquad (30)
\end{aligned}$$

We then expand $D_t(x_{t-1})$ using (27) as follows:

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

After taking conditional expections w.r.t. $\mathcal{F}_{t-1}$, (30) 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

$$\mathbb{E}[B_t - B_{t-1}] \leq -\frac{\alpha_t}{n} B_{t-1} - \frac{\alpha_t}{n}(F(x_{t-1}) - F(x^*)) + \frac{\mu}{2n}\sum_{i=1}^{n}\left(\frac{\alpha_t}{n}\right)^2 \frac{1}{q_i n}\|v_i^t\|^2. \qquad (31)$$

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 (29) and (31) yields

$$\mathbb{E}[C_t - C'_{t-1}] \leq -\frac{\alpha_t}{n} C'_{t-1} + \frac{2\alpha_t}{n}\Big(\frac{2p_t L_q}{\mu} - \frac{1}{2}\Big)(F(x_{t-1}) - F(x^*)) + \frac{\mu\alpha_t}{n^2}\sum_{i=1}^{n}\frac{\delta_i^t}{q_i n}\|v_i^t\|^2 + \frac{2\alpha_t p_t \sigma_q^2}{n\mu}, \qquad (32)$$

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 (19), then we obtain the desired recursion after noticing that $C'_{t-1} \leq C_{t-1}$ and taking total expectations on $\mathcal{F}_{t-1}$.

□

Note that if we take

$$\alpha_1 \leq \min\left\{\frac{n q_{\min}}{2}, \frac{n\mu}{8L_q}\right\},$$

then (32) yields

$$\mathbb{E}\left[\frac{C_t^q}{\mu}\right] \leq \Big(1 - \frac{\alpha_t}{n}\Big)\,\mathbb{E}\left[\frac{C_{t-1}^q}{\mu}\right] - \frac{\alpha_t}{2n\mu}(F(x_{t-1}) - F(x^*)) + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma_q^2}{\mu^2}.$$

This relation takes the same form as Eq. (26), hence it is straightforward to adapt the proof of Theorem 3 to this setting, and the same iterate averaging scheme applies.

C Complexity Analysis of SGD

In this section, we provide a proof of a simple result for SGD in the smooth case, giving a recursion that depends on a variance condition at the optimum (in contrast to [4, 23] where this condition needs to hold everywhere), for a more natural comparison with S-MISO.

Proposition C.1 (Simple SGD recursion with variance at optimum)**.** *Under assumptions (A1) and (A2), 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

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

Proof. We have

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

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

Note that when $\eta_t \leq 1/4L$, we have

$$\mathbb{E}\left[\|x_t - x^*\|^2\right] \leq (1-\mu\eta_t)\,\mathbb{E}\left[\|x_{t-1} - x^*\|^2\right] - \eta_t(f(x_{t-1}) - f(x^*)) + 2\eta_t^2\sigma_{\rm tot}^2.$$

This takes a similar form to Eq. (26), and one can use the same iterate averaging scheme as Theorem 3 with step-sizes $\eta_t = 2/\mu(\gamma+t)$ by adapting the proof.

We now give a similar recursion for the proximal SGD algorithm (see, *e.g.*, [10]). This allows us to apply the results of Theorem 2 and the step-size strategy mentioned in Section 3.

Proposition C.2 (Simple recursion for proximal SGD with variance at optimum)**.** *Under assumptions (A1) and (A2), if $\eta_t \leq 1/2L$, then the proximal SGD recursion*

$$x_t := \operatorname{prox}_{\eta_t h}(x_{t-1} - \eta_t \nabla \tilde{f}_{i_t}(x_{t-1},\rho_t))$$

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

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

Proof. We have

$$\begin{aligned}
&\|x_t - x^*\|^2 \\
&= \|\operatorname{prox}_{\eta_t h}(x_{t-1} - \eta_t \nabla \tilde{f}_{i_t}(x_{t-1},\rho_t)) - \operatorname{prox}_{\eta_t h}(x^* - \eta_t \nabla f(x^*))\|^2 \\
&\leq \|x_{t-1} - \eta_t \nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - x^* + \eta_t \nabla f(x^*)\|^2 \\
&= \|x_{t-1} - x^*\|^2 - 2\eta_t\langle \nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla f(x^*), x_{t-1} - x^*\rangle + \eta_t^2\|\nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla f(x^*)\|^2 \\
&\leq \|x_{t-1} - x^*\|^2 - 2\eta_t\langle \nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla f(x^*), x_{t-1} - x^*\rangle \\
&\quad + 2\eta_t^2\|\nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla \tilde{f}_{i_t}(x^*,\rho_t)\|^2 + 2\eta_t^2\|\nabla \tilde{f}_{i_t}(x^*,\rho_t) - \nabla f(x^*)\|^2,
\end{aligned}$$

where the first equality follows from the optimality of x^* and the following inequality follows from the non-expansiveness of proximal operators. Taking conditional expectations on $\mathcal{F}_{t-1}$ yields

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

where inequality $(*)$ follows from the μ-strong convexity of f and $\mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_{i_t}(x_{t-1},\rho_t) - \nabla \tilde{f}_{i_t}(x^*,\rho_t)\|^2]$ is bounded by $2L(f(x_{t-1}) - f(x^*) - \langle \nabla f(x^*), x_{t-1} - x^*\rangle)$ as in the proof of Lemma B.1. By convexity of f, we have $f(x_{t-1}) - f(x^*) - \langle \nabla f(x^*), x_{t-1} - x^*\rangle \geq 0$, hence the second term is non-positive when $\eta_t \leq 1/2L$. We conclude by taking total expectations. $\square$

We note that Propositions C.1 and C.2 can be easily adapted to non-uniform sampling with sampling distribution q and step-sizes $\eta_t/q_{i_t}n$, leading to step-size conditions $\eta_t \leq 1/2L_q$, with $L_q = \max_i \frac{L_i}{q_i n}$ and variance $\sigma_{q,tot}^2 = \mathbb{E}_{i,\rho}[\frac{1}{(q_i n)^2}\|\nabla \tilde{f}_i(x^*,\rho) - \nabla f(x^*)\|^2]$.

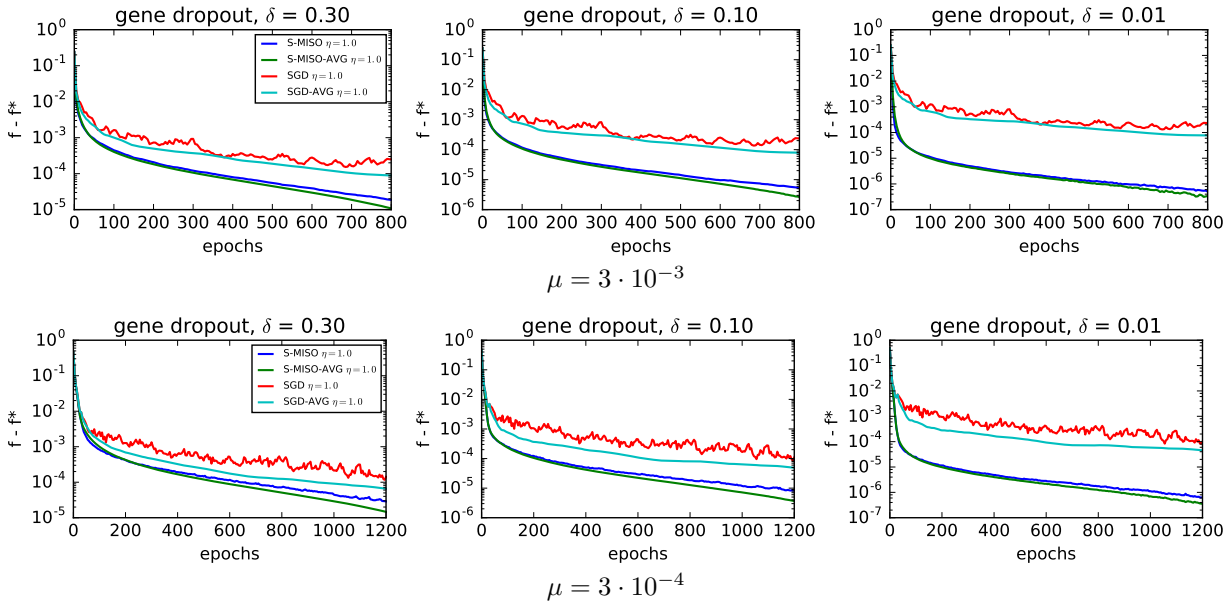

Figure 3: Comparison of S-MISO and SGD with averaging, for different condition numbers (controlled by μ) and different Dropout rates δ. We begin step-size decay and averaging at epoch 3 (top) and 30 (bottom).

D Experiments with Averaging Scheme

Figure 3 shows a comparison of S-MISO and SGD with the averaging scheme proposed in Theorem 3 (see Appendix C for comments on how it applies to SGD), on the breast cancer dataset presented in Section 5, for different values of the regularization μ (and thus of the condition number $\kappa = L/\mu$), and Dropout rates δ. We can see that the averaging scheme gives some small improvements for both methods, and that the improvements are more significant when the problem is more ill-conditioned (Figure 3, bottom). We note that the time at which we start averaging can have significant impact on the convergence, in particular, starting too early can significantly slow down the initial convergence, as commonly noticed for stochastic gradient methods (see, *e.g.*, [23]).