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

Abstract

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

1 Introduction

Many supervised machine learning problems can be cast into the problem of minimizing an expected loss over a data distribution $\mathcal{D}$ with respect to a vector x 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 are typically used [3]. However, in the case of finite datasets, incremental methods based on variance reduction techniques (e.g., [5, 8, 9, 13, 15]) have proven to be very successful at solving the finite sum problem

$$\min_x \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}.$$

A classical setting is $f_i(x) = \ell(y_i, x^\top \xi_i) + \mu/2\|x\|^2$, where (ξ_i, y_i) is an example-label pair, ℓ is a convex loss function, and μ is a regularization parameter. However, in many situations, augmenting the finite training set with well-chosen random perturbations of each example can lead to a smaller test error in theory [19] and in practice [16]. Examples of such procedures include random transformations of images in classification problems (e.g., [16]), and Dropout [17]. The objective describing these scenarios, which we consider in this paper, is the following:

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

where ρ parametrizes the random perturbation. Because each function f_i is an expectation, computing the exact gradient is intractable in general, and standard variance reduction methods cannot be used. A straightforward way to optimize this objective is to use SGD by choosing an index i_t randomly in $\{1, \ldots, n\}$ at iteration t and sampling a perturbation $\rho_t \sim \Gamma$. 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, while guaranteeing convergence behavior similar to that of stochastic approximation, with a smaller variance only due to the random perturbations on a single example.

This work was supported by a grant from ANR (MACARON project ANR-14-CE23-0003-01) and from the MSR-Inria joint centre. A longer version of this paper is available at https://arxiv.org/abs/1610.00970.

Related work. The problem of minimizing (1) is not well studied in the optimization and machine learning literature. Most related to our work, recent methods that use clustering information to improve the convergence of variance reduction techniques [2, 7] 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 [7] can be seen as a variance condition on stochastic gradients as in our case, their algorithm is asymptotically biased and does not converge to the optimum. On the other hand, ClusterSVRG [2] is not biased, but requires a finite sum structure and hence does not support infinite datasets. The method proposed in [1] also bears similarity with ours, since it uses variance reduction in a setting where gradients are computed approximately, but the algorithm requires reducing the approximation variance by dynamically increasing the number of MCMC samples used in order to reach the optimum, while our algorithm overcomes this requirement by supporting decreasing step-sizes.

2 The Stochastic MISO Algorithm

In this section, we introduce the *stochastic MISO* approach, given in Algorithm 1, which relies on the following assumptions:

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

```
Algorithm 1: S-MISO
Input: step-size sequence $(\alpha_t)_{t \geq 1}$;
initialize $x_0 = \frac{1}{n}\sum_i z_i^0$ for some $(z_i^0)_{i=1,\dots,n}$;
for $t = 1, \dots$ do
    Sample $i_t$ randomly in $\{1, \dots, n\}$, $\rho_t \sim \Gamma$, and update:
        $z_i^t = \begin{cases} (1-\alpha_t) z_i^{t-1} + \alpha_t (x_{t-1} - \frac{1}{\mu}\nabla \tilde{f}_i(x_{t-1}, \rho_t)), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise.} \end{cases}$
        $x_t = \frac{1}{n}\sum_{i=1}^n z_i^t.$        (2)
end
```

Without the perturbations and with a constant step-size, the algorithm resembles the MISO/Finito algorithms [6, 9, 10] which may be seen as primal variants of SDCA [14, 15]. Specifically, MISO/Finito assumes that each f_i is strongly convex, and builds a model of the objective using lower bounds of the form $f_i(x) \geq c_i + \frac{\mu}{2}\|x - z_i\|^2$. Note that when f_i is an expectation, it is hard to obtain such bounds since exact gradients are not available in closed form. Separately, SDCA [15] considers the Fenchel conjugates of f_i, which usually are not available in closed form either when f_i is an expectation, and in fact exploiting stochastic gradient estimates is difficult in the duality framework. In contrast, Shalev-Shwartz [14] gives an analysis of SDCA in the primal, aka. "without duality", for finite sums, and our work extends this reasoning to the stochastic approximation setting.

The link between S-MISO and SGD can be seen by rewriting the update (2) as

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

where

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

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 [8] and later in other variance reduction algorithms [5, 14].

3 Convergence Analysis

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

$$C_t = \frac{1}{2}\|x_t - x^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^n \|z_i^t - z_i^*\|^2, \tag{3}$$

which allows us to state our main result:

Proposition 1 (Recursion on C_t)**.** *If $(\alpha_t)_{t\geq 1}$ is a positive and non-increasing sequence of step-sizes with $\alpha_1 \leq \min(\frac{1}{2}, \frac{n}{2(2\kappa-1)})$, then C_t obeys the following recursion*

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}] + 2\left(\frac{\alpha_t}{n}\right)^2 \frac{\sigma^2}{\mu^2}. \tag{4}$$

This result is obtained by bounding separately each term in (3), 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 [14], while taking into account the stochastic perturbations. Ultimately, the variance σ in (4) only depends on the amount of data perturbation.

Comparison with SGD. A classical analysis of SGD with step-sizes $(\eta_t)_{t\geq 0}$ gives the following recursion (see, e.g., [12]) on $B_t := \frac{1}{2}\mathbb{E}[\|x_t - x^*\|^2]$:

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

where we assume $\mathbb{E}_{i,\rho}[\|\nabla\tilde{f}_i(x,\rho)\|^2] \leq M^2$ for all x. Thus, after forgetting the initial condition B_0, S-MISO minimizes $B_t \leq C_t$ at a faster rate if $2\sigma^2 \leq M^2/2$. In particular, if the gradient variance *across examples* (bounded by M here) is much smaller than the gradient variance due to the data perturbation only $\rho \sim \Gamma$ (bounded by σ^2 at the optimum), then our algorithm will have a much faster convergence rate. As shown in the experiments presented in the next section, M 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, Theorem 4.7] for SGD. Note that convergence of objective function values is directly related to that of the Lyapunov function C_t via smoothness:

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

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

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

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

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

where

$$\nu := \max\left\{\frac{2\beta^2\sigma^2}{\mu^2(\beta - 1)}, (\gamma + 1)C_0\right\}. \tag{5}$$

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 (5). 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], 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 $\bar{\alpha}$, we can reach a suboptimality $\bar{\epsilon} := \frac{2\bar{\alpha}\sigma^2}{n\mu^2}$ in $O(\frac{n}{\bar{\alpha}}\log C_0/\bar{\epsilon})$ iterations. If we then set $\beta = 2$ and γ large enough so that $\alpha_1 = \frac{\beta n}{\gamma+1} \approx \bar{\alpha}$, we will have $\nu = 8\sigma^2/\mu^2$. Considering these two phases, the final work complexity of the algorithm is

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

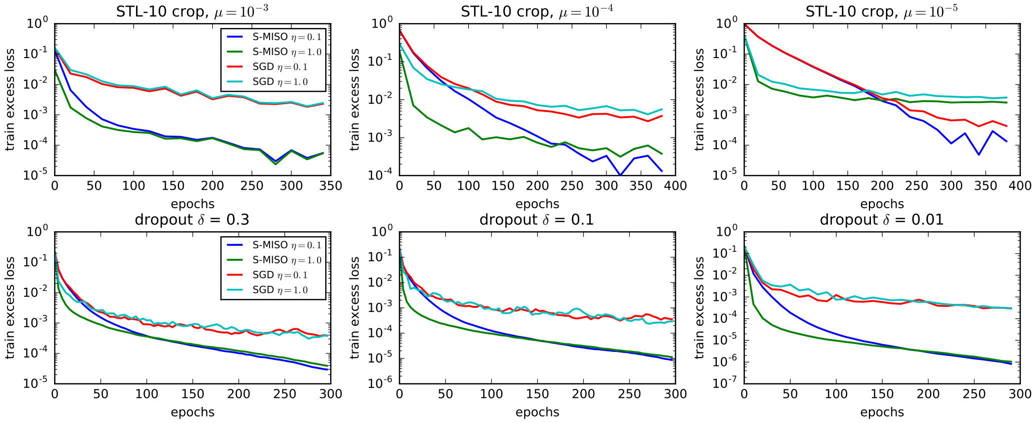

Figure 1: Comparison of S-MISO with SGD. Training loss are shown in logarithmic scale (1 unit = factor 10). (Top) STL-10 dataset with different values of μ (the best value given by cross validation is around 10^{-4}); (bottom) breast cancer dataset with Dropout, for different values of δ and $\mu = 0.003$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

4 Experiments

We present preliminary experiments on two significantly different scenarios: we consider an image classification dataset with random transformations and a classification task on breast cancer data with Dropout (the perturbation sets randomly a fraction of the data entries to zero). For both algorithms, we use the step-size strategy mentioned in section 3 and advised by Bottou et al. [3], which we have found to be most effective among many strategies 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. Figure 1 shows the curves we obtain for a Monte-Carlo estimate of the training objective. The plots are shown on a logarithmic scale, and the values are compared to the best value obtained in 400 epochs. In both cases, the strong convexity constant μ is the regularization parameter.

Image classification with "data augmentation". The success of deep neural networks is often limited by the availability of large amounts of labeled images. When there are many unlabeled images but few labeled ones, a common approach is to train a linear classifier on top of a deep network learned in an unsupervised manner. We follow this approach on the STL-10 dataset [4], which contains 5000 training images from 10 classes and 100000 unlabeled images, using a 2-layer unsupervised convolutional kernel network [11], giving representations of dimension 102400. The perturbation consists of randomly cropping the input images. The loss function is the squared hinge loss used in a one-versus-all setting. The vector representations are ℓ_2-normalized such that $L = 1$.

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

Dropout on gene expression data. We trained a binary logistic regression model on the breast cancer gene expression dataset of Van de Vijver et al. [18] with different dropout rates δ, i.e. where at every iteration, each coordinate ξ_j of a feature vector ξ is set to zero independently with probability δ and to $\xi_j/(1-\delta)$ otherwise. Figure 1 (bottom) compares S-MISO with SGD for three values of δ, as a way to control the variance of the perturbations. We include a dropout rate of 0.01 to illustrate the impact of δ on the algorithms, 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.

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] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

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

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

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.