Journées SMAI MODE 2018

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

Alberto BIETTI
Inria

Julien MAIRAL
Inria

Résumé. 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). 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.
This work was published at the NIPS 2017 conference [1].

Mots-clefs : machine learning, stochastic optimization, variance reduction, regularization.

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 [9]). Nevertheless, when the dataset is finite, incremental methods based on variance reduction techniques (*e.g.*, [4, 10, 11]) 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. We consider the smooth case here ($h = 0$), and point the interested reader to the full paper [1] for extensions to the non-smooth (composite) case.

A hybrid stochastic/finite-sum setting. In this work, we are interested in a variant of (1) where random perturbations of data are introduced, which is a common scenario in machine learning. Then, the functions f_i involve an expectation over a random perturbation ρ, leading to the problem

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

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 utmost interest; for instance, this is a key to achieve stable feature selection [8], improving the generalization error both in theory [12] and in practice [5], obtaining stable and robust predictors [13], or using complex a priori knowledge about data to generate virtually larger datasets [5].

Going faster than SGD: the Stochastic MISO algorithm. 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 is indeed SGD, but ignoring the finite-sum structure leads to gradient estimates with high variance and slow convergence. The key quantity to characterize the gains we can hope to achieve relative to SGD is

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

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 that is due to the perturbations ρ. In contrast, the convergence behavior of SGD is controlled by the total variance, which also includes the randomness in the choice of the index i, and is given by

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

The goal of this work 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 applying a simple perturbation.

We introduce an algorithm for strongly convex objectives, called *stochastic MISO* (S-MISO), which exploits the underlying finite sum using variance reduction and achieves faster convergence rate than SGD by reducing the dependence on gradient variance from σ_{tot}^2 to σ_p^2. Our method is based on the MISO/Finito algorithm [4, 6], which incrementally updates quadratic lower bounds on each f_i (obtained from strong convexity) and minimizes the obtained minorizing surrogate on f. S-MISO follows a similar approach, but considers instead *approximate* lower bounds to each f_i, constructed using stochastic gradient estimates of the form $\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)$, where x_{t-1} is the current iterate and i_t, ρ_t denote a randomly sampled index and perturbation. Because of the stochastic nature of the setting, an appropriate decay of step-sizes is needed in order to obtain convergence.

The iteration complexity of our method is shown in Table 1. The gain over SGD is of order $\sigma_{\text{tot}}^2/\sigma_p^2$, which is also observed in our experiments. We also compare against N-SAGA [3]; its convergence rate is similar to ours but suffers from a non-zero asymptotic error.

Practical gains and results. In Table 2, we show estimates of the gains $\sigma_{\text{tot}}^2/\sigma_p^2$ of our algorithm compared to SGD in practical scenarios with perturbations, such as Dropout (which sets feature vector coordinates to zero with probability δ) or image data augmentation with a pre-trained or unsupervised deep convolutional network. Both settings are important in order to improve test error in typical machine learning tasks, such as text document classification

Table 1: Iteration complexity of different methods for solving the objective (2) in terms of number of iterations required to find x such that $\mathbb{E}[f(x) - f(x^*)] \leq \epsilon$. 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).

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 [3]	$\epsilon_0 = O\left(\frac{\sigma_p^2}{\mu}\right)$	$O\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$ 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. ResNet-50 denotes a 50 layer network [2] pre-trained on the ImageNet dataset.

Perturbation	**Application case**	**Estimated ratio** $\sigma_{\text{tot}}^2/\sigma_p^2$
Direct perturbation of features	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 [2], color perturbation	21.9
	ResNet-50 [2], rescaling + crop	13.6
	Unsupervised CKN [7], rescaling + crop	9.6

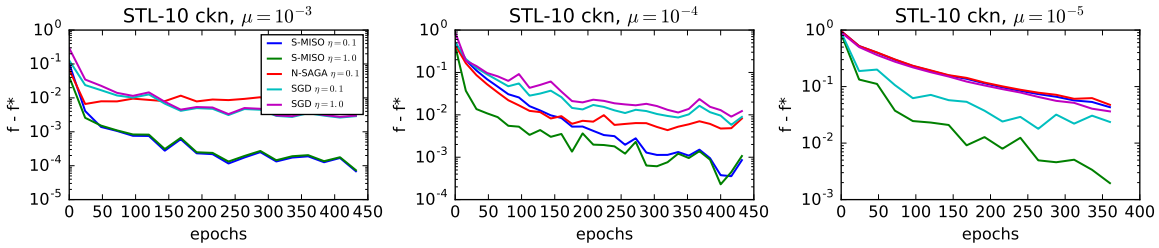

Figure 1: Impact of conditioning for data augmentation on the STL-10 dataset (controlled by μ, where $\mu = 10^{-4}$ gives the best accuracy). Values of the loss are shown on a **logarithmic scale** (1 unit = factor 10). η is a multiplier on the initial step-size.

and image classification. Figure 1 compares convergence results of S-MISO with SGD and N-SAGA, 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.

Références

[1] A. Bietti and J. Mairal. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[2] 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.

[3] 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.

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

[5] 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.

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

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

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

[9] 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.

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

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

[12] 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.

[13] 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.