Statistical Perspectives of Stochastic Optimization

Matt Bonakdarpour
Department of Statistics
University of Chicago
Chicago, IL, 60637
mbonakda@uchicago.edu

Panagiotis (Panos) Toulis
Econometrics and Statistics
University of Chicago, Booth School
Chicago, IL, 60637
ptoulis@uchicago.edu

Abstract

In this technical note, we leverage theory from stochastic approximations to characterize the asymptotic variance of a general class of stochastic optimization procedures. We show that one approach to tuning parameters of such procedures is to minimize the asymptotic variance, and we illustrate such tuning in simulations. Using this result, we also obtain new statistical insights on popular methods employing mini-batches, adaptive gradients, and variance reduction.

1 Introduction

Many stochastic optimization procedures are of the following form:

$$\theta_n = \theta_{n-1} - \frac{1}{n} S(\theta_{n-1}), \tag{1}$$

where $S(\theta)$ is a random variable that depends on parameter θ such that (i) $\mathbb{E}(S(\theta_\star)) = 0$; (ii) $\mathrm{Var}(S(\theta))$ is bounded; (iii) S is convex such that $S(\theta)'(\theta - \theta_\star)$ is positive in expectation. The convergence in probability of θ_n to $\theta_\star$ follows from the seminal result of Robbins & Monro [7]. Several stochastic optimization procedures are special cases of Eq. (1), including stochastic gradient descent, AdaGrad [2], and stochastic variance reduction [4].

There are several ways to select $S(\theta)$ and still converge to $\theta_\star$. For example, for any two valid selections S_1 and S_2, their weighted average is also valid. Thus, one approach is to parameterize S and then tune the parameters to optimize some objective. Here, we choose this objective to be the limit of $n\mathrm{Var}(\theta_n)$, the asymptotic variance of θ_n as an estimator of $\theta_\star$. Better procedures have smaller asymptotic variances, so reducing such variance is a principled way to optimize them. In Section 2 we derive the main theoretical result, and subsequently use it to gain insights on popular stochastic optimization procedures.

2 Theory

Theorem 1 (Fabian [3])**.** *Let $\mathbb{E}(S(\theta)) = s(\theta)$. Let $\mathcal{J}_s(\theta)$ denote the Jacobian of s at θ and suppose that $\mathcal{J}_s(\theta) - I/2$ is positive-definite everywhere, where I is the identity matrix. Also let $V_s(\theta) = \mathrm{Var}(S(\theta))$ denote the variance of $S(\theta)$, and define the limit $\Sigma = \lim_{n\to\infty} n\mathrm{Var}(\theta_n)$. Then,*

$$(\mathcal{J}_s(\theta_\star) - I/2)\Sigma + \Sigma(\mathcal{J}_s(\theta_\star) - I/2) = V_s(\theta_\star). \tag{2}$$

Proof. Under the theorem definitions, we can write procedure (1) as follows:

$$\theta_n = \theta_{n-1} - \frac{1}{n} s(\theta_{n-1}) - \frac{1}{n} W_n,$$

The work was presented at the NIPS workshop on Optimizing the Optimizers. Barcelona, Spain, 2016.

where W_n is the stochastic component of $S(\theta_n)$ with $\mathbb{E}(W_n) = 0$ and $\mathrm{Var}(W_n) = V_s(\theta)$. Near the solution $\theta_\star$, we can write $s(\theta) = \mathcal{J}_s(\theta_\star)(\theta - \theta_\star) + o(||\theta - \theta_\star||)$ so that the procedure becomes

$$\theta_n - \theta_\star = (I - \frac{1}{n}\mathcal{J}_s(\theta_\star))(\theta_{n-1} - \theta_\star) - \frac{1}{n}(W_n + o(||\theta_{n-1} - \theta_\star||)).$$

By Fabian's theorem [3], as $n \to \infty$, $\sqrt{n}(\theta_n - \theta_\star)$ is asymptotically normal with variance Σ as defined in Eq. (2). □

Corollary 1. *In Theorem 1, suppose that $\mathcal{J}_s(\theta)$ and $V_s(\theta)$ commute. Then,*

$$\Sigma = (2\mathcal{J}_s(\theta_\star) - I)^{-1} V_s(\theta_\star). \tag{3}$$

Remarks. Eq. (3) determines the statistical efficiency of the stochastic procedure in Eq. (1), since it can be shown that the MSE of stochastic procedures can be decomposed into a $O(1/n)$ variance term and a $O(1/n^2)$ bias term, and so the variance term is more important in the limit. Moreover, Eq. (3) illustrates that the asymptotic variance depends on the Jacobian of $s(\theta) = \mathbb{E}(S(\theta))$. Intuitively, the Jacobian determines the covariance between θ_{n-1} and $S(\theta_{n-1})$ since $\mathrm{Cov}(S(\theta_{n-1}), \theta_{n-1}) \approx Cov(\mathcal{J}_s(\theta_\star)(\theta_{n-1} - \theta_\star), \theta_{n-1}) = \mathcal{J}_s(\theta_\star)\mathrm{Var}(\theta_{n-1})$. This covariance measures roughly how much information is "wasted" due to the curvature of $s(\theta)$, and so more efficient procedures will have smaller curvature – ideally, no curvature at all as in normal linear models.

3 Stochastic gradient descent

Suppose we have data $\{(X_i, Y_i)\}_{i=1}^N$ and we wish to find $\theta_\star \in \mathbb{R}^p$ to minimize a differentiable and convex loss function, typically the negative of log-likelihood of Y given X and θ:

$$\theta_\star = \arg\min_\theta -\frac{1}{N}\sum_{i=1}^N \log f(Y_i; X_i, \theta) \Rightarrow -\frac{1}{N}\sum_{i=1}^N \nabla \log f(Y_i; X_i, \theta_\star) = 0.$$

One popular approach is to find $\theta_\star$ using stochastic gradient descent (SGD):

$$\theta_n = \theta_{n-1} - \frac{\gamma}{n}\left(-\nabla \log f(Y_{i_k}; X_{i_k}, \theta_{n-1})\right), \tag{4}$$

where i_k is a random sample from $\{1, 2, \ldots, N\}$. To write Eq. (4) in the form of Eq. (1), we define Z to be a one-hot binary vector of length N, selected uniformly at random. We also let G_θ denote the $p \times N$ matrix where the jth column is $-\nabla \log f(Y_j; X_j, \theta)$. Note that the only random variable here is Z, whereas G_θ is a deterministic function of θ, conditional on the data. Then, SGD in Eq. (4) is equivalent to the procedure

$$\theta_n = \theta_{n-1} - \frac{\gamma}{n} G_{\theta_{n-1}} Z, \tag{5}$$

and thus $S(\theta) = \gamma G_\theta Z$ in the notation of Eq. (1). By definition, $E(Z) = (1/N)\mathbf{1}$, where $\mathbf{1}$ is the appropriately-sized vector of ones, and $\mathrm{Var}(Z) = (1/N)I - (1/N^2)\mathbf{1}\mathbf{1}^\intercal$. Hence,

$$s(\theta) = E(S(\theta)) = \gamma E(G_\theta Z) = \gamma G_\theta E(Z) = (\gamma/N) G_\theta \mathbf{1} = (-\gamma/N)\sum_{i=1}^N \nabla \log f(Y_i; X_i, \theta),$$

from which it follows that $s(\theta_\star) = (\gamma/N) G_{\theta_\star}\mathbf{1} = 0$, as desired. Therefore, the Jacobian is

$$\mathcal{J}_s(\theta_\star; \gamma) = (-\gamma/N)\sum_{i=1}^N \nabla^2 \log f(Y_i; X_i, \theta_\star). \tag{6}$$

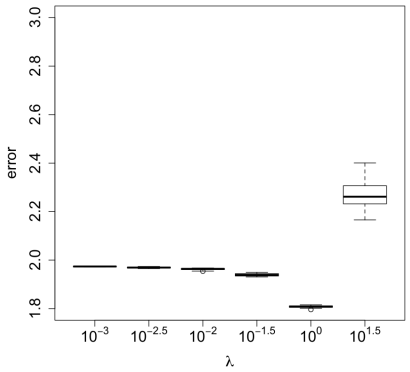
Similarly the variance is derived as follows

$$V_s(\theta_\star) = \mathrm{Var}(S(\theta_\star)) = \gamma^2 \mathrm{Var}(G_{\theta_\star} Z) = \gamma^2 G_{\theta_\star}\mathrm{Var}(Z) G_{\theta_\star}^\intercal = (\gamma^2/N) G_{\theta_\star} G_{\theta_\star}^\intercal. \tag{7}$$

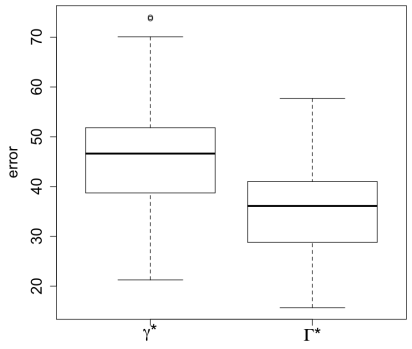
In linear models where the loss depends on θ through a linear combination with features X, i.e. where $f(Y; X, \theta) \equiv f(Y; X'\theta)$, we can show that $\mathcal{J}_s(\theta)$ and $V_s(\theta)$ commute and thus using Corollary 1 we obtain:

$$n\mathrm{Var}(\theta_n) \to (1/N)\gamma^2 (2\mathcal{J}_s(\theta_\star; \gamma) - I)^{-1} G_{\theta_\star} G_{\theta_\star}^\intercal. \tag{8}$$

We can use this formula for the asymptotic variance of SGD to tune the learning rate parameter γ in a principled way. We describe this tuning next, and subsequently describe tuning SGD with a multi-dimensional learning rate.



(a) Error for $\lambda \times \gamma^*$.



(b) Error for scalar γ^* and matrix Γ^*.

Figure 1: *(a)*: Selecting γ^* according to Eq. (9) yields smallest estimation error compared to other values of γ; *(b)*: Selecting the multi-dimensional learning rate Γ^* according to Eq. (12) yields better performance than the best scalar learning rate γ^* of Eq. (9).

3.1 Choosing γ

One approach for tuning γ is to minimize the trace of the asymptotic variance matrix in Eq. (8):

$$\gamma^* = \arg\min_{\gamma} \left\{ \text{trace}\left(\gamma^2 (2\mathcal{J}_s(\theta_\star;\gamma) - I)^{-1} G_{\theta_\star} G_{\theta_\star}^\intercal\right)\right\}. \tag{9}$$

We illustrate this approach in the following experiment. We generate a synthetic dataset $\{(X_i, Y_i)\}_{i=1}^{2000}$ from a logistic regression model. We sample and fix our covariates $X_i \sim N_{10}(0, I)$ and set $\theta_\star \in \mathbb{R}^{10}$ so that the k-th coordinate is $\theta_\star^k = 5e^{-k}$. We generate responses $Y_i \sim \text{Bern}(p_i)$ where $p_i = 1/(1 + \exp(-\theta_\star' X_i))$. We then solve Eq. (9) exactly for γ^*.

With this fixed dataset and γ^*, we perform SGD with 20e3 iterations to estimate $\theta_\star$. We perform SGD 500 separate times on the same dataset to obtain 500 different parameter estimates – variability across trials is introduced by the order in which we sample our data. Finally, we compute the mean squared error between θ_n^{SGD} and $\theta_\star$ across all 500 runs. For comparison, we also perform SGD with $\gamma = \lambda\gamma^*$ for varying levels of λ. Figure (1a) shows that $\lambda = 1$ obtains the best parameter estimation error, consistent with the theory. We observed similar results across different datasets.

3.2 Multi-dimensional learning rate

We now generalize the SGD procedure in Eq. (5) to use a $p \times p$ positive definite matrix Γ as the multi-dimensional learning rate:

$$\theta_n = \theta_{n-1} - \frac{1}{n}\Gamma G_{\theta_{n-1}} Z. \tag{10}$$

If we can set Γ to commute with $\mathcal{J}_s(\theta_\star)$ then we can show for this SGD procedure that

$$n\text{Var}(\theta_n) \to (1/N)(2\Gamma\mathcal{J}_s(\theta_\star) - I)^{-1}\Gamma G_{\theta_\star} G_{\theta_\star}^\intercal \Gamma^\intercal \tag{11}$$

This expression illustrates that we now have more freedom to reduce the asymptotic variance of SGD. Similar to the method above, we find Γ by minimizing the asymptotic variance in Eq. (11). However, to ensure that Γ and $\mathcal{J}_s(\theta_\star)$ commute, we first obtain the eigenvalue decomposition of $\mathcal{J}_s(\theta_\star) = Q\Lambda Q^T$, and then solve the following optimization problem:

$$D^* = \arg\min_{D} \left\{ \text{trace}\left((2(QDQ^T)\mathcal{J}_s(\theta_\star) - I)^{-1}(QDQ^T) G_{\theta_\star} G_{\theta_\star}^\intercal (QDQ^T)^\intercal\right)\right\} \tag{12}$$

where D is a diagonal matrix. Finally, we set $\Gamma^* = QD^*Q^T$. For these experiments, we degrade the conditioning of the design matrix X to better illustrate the improvement over the approach in Section (3.1). We show the improvement of the multi-dimensional Γ in Figure (1b).

4 Insights on other methods

4.1 Mini-batch SGD

A generalization of SGD in Eq. (1) is to use the average of m gradients selected at random. In our notation, $S(\theta) = -(\gamma/m)\gamma G_\theta Z$, where Z is now a binary vector with exactly m random components equal to 1. The expected value of S is the same as in classical SGD since $E(S(\theta)) = -(\gamma/m)G_\theta E(Z) = (-\gamma/N)G_\theta \mathbf{1}$. The variance, however, is different since $\mathrm{Var}(S(\theta)) = (\gamma^2/m^2)G_\theta \mathrm{Var}(Z) G_\theta^\intercal$, where $\mathrm{Var}(Z) = (1-\frac{1}{N})a(1-a)I - \frac{1}{N-1}a(1-a)\mathbf{1}\mathbf{1}^\intercal$, and $a = m/N$ is the mini-batch proportional to the total sample size. When $\theta = \theta_\star$, and assuming $1-a \approx 1$, we obtain $V_s(\theta_\star) = \mathrm{Var}(S(\theta_\star)) = (\gamma^2/mN)G_{\theta_\star} G_{\theta_\star}^\intercal$.

Mini-batch SGD therefore achieves an m-fold improvement in statistical efficiency relative to classical SGD. However, the computational burden increases m-fold because at every iteration we calculate and add m gradients together.

4.2 Adaptive methods

Equations (9) and (12) illustrate that for optimal selection of learning rates we need to know the true parameter values $\theta_\star$. For example, we can show that the optimal selection of Γ^* in Eq. (12) is given by $\Gamma^* = F(\theta_\star)^{-1}$, where $F(\theta) = -E(\nabla^2 \log f(Y; X, \theta)$ is the so-called Fisher information matrix. Under standard statistical theory we can also show that $\mathcal{J}_s(\theta_\star) \to F(\theta_\star)$ and $V_s(\theta_\star) \to F(\theta_\star)$, as $N \to \infty$. Thus by Eq. (11) we have

$$n\mathrm{Var}(\theta_n) \to (2F(\theta_\star)^{-1}\mathcal{J}_s(\theta_\star) - I)^{-1} F(\theta_\star)^{-1} F(\theta_\star) F(\theta_\star)^{-\intercal} = F(\theta_\star)^{-1}.$$

The quantity $F(\theta_\star)^{-1}$ is the so-called Cramer-Rao bound, and it is a fundamental estimation bound: any unbiased estimator of $\theta_\star$ cannot achieve variance that is smaller than $F(\theta_\star)^{-1}/n$ in the limit as n, the number of iid samples, increases.

It follows that if we know $F(\theta_\star)^{-1}$, we can use SGD in Eq. (10) with $\Gamma = F(\theta_\star)^{-1}$ and achieve optimal estimation efficiency. This is why most adaptive methods in fact aim to approximate $F(\theta_\star)$ along the main procedure [6, 1], e.g., using $F(\theta_n)$. AdaGrad [2] is an interesting case where $F(\theta_\star)$ is diagonally approximated but the learning rate is of order, say, $1/\sqrt{n}$. This leads to an inefficient estimator but the estimator is robust to misspecifications of the learning rate parameters, unlike SGD which is sensitive to such misspecifications [8].

4.3 Variance reduction methods

In variance reduction [5, 4] the stochastic procedure in Eq. (1) is typically modified as follows:

$$\theta_n = \theta_{n-1} - \frac{\gamma}{n}\left[-\nabla \log f(Y_{i_k}; X_{i_k}, \theta_{n-1}) + \nabla \log f(Y_{i_k}; X_{i_k}, \tilde{\theta}) + \tilde{\mu}\right], \tag{13}$$

where $\tilde{\theta}$ is some moving average of θ_n and $\tilde{\mu} = -\frac{1}{N}\sum_{i=1}^N \nabla \log f(Y_{i_k}; X_{i_k}, \tilde{\theta})$. In our notation, the above procedure is equivalent to $\theta_n = \theta_{n-1} - \frac{1}{n}S(\theta_{n-1})$, where

$$S(\theta) = \gamma[G_\theta Z - G_{\tilde{\theta}} Z + (1/N) G_{\tilde{\theta}} \mathbf{1}].$$

Note that $s(\theta) = E(S(\theta)) = \gamma[G_\theta E(Z) - G_{\tilde{\theta}} E(Z) + (1/N) G_{\tilde{\theta}} \mathbf{1}] = (\gamma/N) G_\theta \mathbf{1}$, which is the same regression function as in SGD of Eq. (4), and thus the Jacobian $\mathcal{J}_s(\theta)$ here is the same as in classical SGD. However, the variance $V_s(\theta)$ of $S(\theta)$ is different:

$$V_s(\theta) = \gamma^2 \mathrm{Var}((G_\theta - G_{\tilde{\theta}})Z) = \gamma^2 (G_\theta - G_{\tilde{\theta}}) \mathrm{Var}(Z) (G_\theta - G_{\tilde{\theta}})^\intercal = \frac{\gamma^2}{N}(G_\theta - G_{\tilde{\theta}})(G_\theta - G_{\tilde{\theta}})^\intercal + O(1/N^2).$$

Assuming Lipschitz gradients we obtain $\mathrm{Var}(S(\theta)) = O(||\theta - \tilde{\theta}||^2)$ and since $||\theta_n - \tilde{\theta}|| \to 0$, we conclude that $n\mathrm{Var}(\theta_n) \to 0$. Variance reduction methods therefore do better than the $O(1/n)$ rate of classical methods such as SGD. Our analysis illustrates that variance reduction methods essentially construct a statistic $S(\theta)$ that has the same expectation as classical SGD but with vanishing variance. This construction is possible because we allow ourselves to periodically calculate the entire gradient (parameter $\tilde{\mu}$ in Eq. (13)), which inherits the convergence rate properties of deterministic methods.

References

[1] Antoine Bordes, Léon Bottou, and Patrick Gallinari. Sgd-qn: Careful quasi-newton stochastic gradient descent. *Journal of Machine Learning Research*, 10(Jul):1737–1754, 2009.

[2] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[3] Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

[4] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[5] V Roshan Joseph. Efficient robbins–monro procedure for binary data. *Biometrika*, 91(2):461–470, 2004.

[6] Hyeyoung Park, S-I Amari, and Kenji Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755–764, 2000.

[7] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[8] Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics, forthcoming*, 2016.