Towards stability and optimality in stochastic gradient descent

Panos Toulis
Harvard University
ptoulis@fas.harvard.edu

Dustin Tran
Harvard University
dtran@g.harvard.edu

Edoardo M. Airoldi
Harvard University
airoldi@fas.harvard.edu

Abstract

Iterative procedures for parameter estimation based on stochastic gradient descent (SGD) allow the estimation to scale to massive data sets. However, in both theory and practice, they suffer from numerical instability. Moreover, they are statistically inefficient as estimators of the true parameter value. To address these two issues, we propose a new iterative procedure termed *averaged implicit* SGD *(*AI-SGD*)*. For statistical efficiency, AI-SGD employs averaging of the iterates, which achieves the optimal Cramér-Rao bound under strong convexity; i.e., it is an optimal unbiased estimator of the true parameter value. For numerical stability, AI-SGD employs an implicit update at each iteration, which is related to proximal operators in optimization. In practice, AI-SGD is more stable than averaging procedures that do not employ proximal operators, and is simpler to implement than procedures that do employ proximal operators but require careful tuning of several hyperparameters.

1 Introduction

Many problems in statistical estimation involve finding the parameter value $\theta_\star \in \Theta$ such that

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathbb{E}\left(L(\theta, \xi)\right), \tag{1}$$

where the expectation is with respect to the random variable $\xi \in \Xi \subseteq \mathbb{R}^d$ that represents the data, $\Theta \subseteq \mathbb{R}^p$ is the parameter space, and $L : \Theta \times \Xi \to \mathbb{R}$ is a loss function. A popular procedure for solving Eq.(1) is stochastic gradient descent (SGD) [15, 3], where a sequence θ_n approximates $\theta_\star$, and is updated iteratively, one data point at a time, through the iteration

$$\theta_n = \theta_{n-1} - \gamma_n \nabla L(\theta_{n-1}, \xi_n), \tag{2}$$

where $\{\xi_1, \xi_2, \ldots\}$ is a stream of i.i.d. realizations of ξ, and $\{\gamma_n\}$ is a non-increasing sequence of positive real numbers, known as the learning rate. While computationally efficient, the SGD procedure (2) suffers from numerical instability and statistical inefficiency. Regarding stability, SGD is sensitive to specification of the learning rate γ_n, since the mean-squared errors can diverge arbitrarily when γ_n is misspecified with the respect to problem parameters, e.g., the convexity and Lipschitz parameters of the loss function [1, 7]. Regarding statistical efficiency, SGD loses statistical information. In fact, the amount of information loss depends on the misspecification of γ_n with respect to the spectral gap of the matrix $\mathbb{E}\left(\nabla^2 L(\theta_\star, \xi)\right)$ [11, 13], also known as the Fisher information matrix if L is the negated log-likelihood. Several solutions have been proposed to resolve these two issues, e.g., using projections and gradient clipping. However, they are usually heuristic and hard to generalize.

In this paper, we aim for the ideal combination of computational efficiency, numerical stability, and statistical efficiency using the following procedure:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla L(\theta_n, \xi_n), \tag{3}$$

$$\bar{\theta}_n = (1/n)\sum_{i=1}^{n} \theta_i. \qquad \textbf{AI-SGD} \tag{4}$$

2 Preliminaries

The norm $||\cdot||$ will denote the L_2 norm. For two matrices A, B, $A \preceq B$ denotes that $B - A$ is nonnegative-definite; $\mathrm{tr}(A)$ denotes the trace of A.

Assumption 1. *The loss function $L(\theta, \xi)$ is almost-surely differentiable. The random vector ξ can be decomposed as $\xi = (x, y)$, $x \in \mathbb{R}^p$, $y \in \mathbb{R}^d$, such that*

$$L(\theta, \xi) = L(x^\intercal \theta, y). \tag{5}$$

Assumption 2. *The learning rate sequence $\{\gamma_n\}$ is $\gamma_n = \gamma_1 n^{-\gamma}$, where $\gamma_1 > 0$ and $\gamma \in (1/2, 1]$.*

Assumption 3 (Lipschitz conditions)**.** *For all $\theta_1, \theta_2 \in \Theta$, a combination of the following conditions is satisfied almost-surely:*

(a) The loss function L is Lipschitz-continuous with parameter λ_0, i.e.,

$$|L(\theta_1, \xi) - L(\theta_2, \xi)| \leq \lambda_0 ||\theta_1 - \theta_2||,$$

(b) The map ∇L is Lipschitz-continuous with parameter λ_1, i.e.,

$$||\nabla L(\theta_1, \xi) - \nabla L(\theta_2, \xi)|| \leq \lambda_1 ||\theta_1 - \theta_2||,$$

(c) The map $\nabla^2 L$ is Lipschitz-continuous with parameter λ_2, i.e.,

$$||\nabla^2 L(\theta_1, \xi) - \nabla^2 L(\theta_2, \xi)|| \leq \lambda_2 ||\theta_1 - \theta_2||.$$

Assumption 4. *The observed Fisher information matrix, $\hat{\mathcal{I}}(\theta) \triangleq \nabla^2 L(\theta, \xi)$, has non-vanishing trace, i.e., there exists $\phi > 0$ such that $\mathrm{tr}(\hat{\mathcal{I}}(\theta)) \geq \phi$, almost-surely, for all $\theta \in \Theta$. The expected Fisher information matrix, $\mathcal{I}(\theta) \triangleq \mathbb{E}\left(\hat{\mathcal{I}}(\theta)\right)$, has minimum eigenvalue $0 < \underline{\lambda_f} \leq \phi$, for all $\theta \in \Theta$.*

Assumption 5. *The zero-mean random variable $W_\theta \triangleq \nabla L(\theta, \xi) - \nabla \ell(\theta)$ is square-integrable, such that, for a fixed positive-definite Σ,*

$$\mathbb{E}\left(W_{\theta_\star} W_{\theta_\star}^\intercal\right) \preceq \Sigma.$$

Remarks. Assumption 1 is not very restrictive because the majority of machine learning models depends on parameter θ through a linear combination with features. A notable exception includes loss functions with a regularization term. Although it is easy to add regularization to AI-SGD we will not do so in this paper because AI-SGD works well without it, since AI-SGD already regularizes the estimate θ_n towards θ_{n-1}; in experiments, regularization neither improved nor worsened AI-SGD. Assumptions on Lipschitz gradients (Assumption 3(b), Assumption 3(c)) can are common [1, 7]. Assumption 3(a) is less standard in classic SGD literature but, so-far, it is standard in the limited literature on implicit SGD [2]. However, we can forgo this assumption and still maintain identical rates for the errors, although at the expense of a more complicated analysis. It is also an open problem whether a nice stability result similar to Theorem 1 can be derived under Assumption 3(b) instead of Assumption 3(a). Assumption 4 makes two claims. The first claim on the observed Fisher information matrix is a relaxed form of strong convexity for the loss $L(\theta, \xi)$. However, in contrast to strong convexity, this claim allows several eigenvalues of $\nabla^2 L$ to be zero. The second claim of Assumption 4 is equivalent to strong convexity of the expected loss $\ell(\theta)$. From a statistical perspective, strong convexity posits that there is information in the data for all elements of $\theta_\star$. This assumption is necessary to derive bounds on the errors $\mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$, and has been used to show optimality of classic SGD with averaging [9, 6, 14, 7].

Overall, our assumptions are weaker than the assumptions in the limited literature on implicit SGD. For example, Bertsekas [2, Assumptions 3.1, 3.2] assumes almost-sure bounded gradients $\nabla L(\theta, \xi)$ in addition to Assumption 3(a).

3 Theory

The main technical challenge in analyzing implicit SGD (3) is that, unlike typical analysis with classic SGD (2), the error ξ_n is not conditionally independent of θ_n. This implies that $\mathbb{E}\left(\nabla L(\theta_n, \xi_n) | \, \theta_n\right) \neq \ell(\theta_n)$, which makes it no longer possible to use the convexity properties of ℓ to analyze the errors $\mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$, as it is common in the literature. The proof strategy relies on a master lemma for the analysis of recursions that appear to be typical in implicit procedures, which is novel to our best knowledge. All proofs are given in an extended version of this paper [12].

3.1 Computational efficiency

Our first result enables efficient computation of the implicit update (3). In general, this can be expensive due to solving a fixed-point equation in many dimensions, at every iteration. We reduce this multi-dimensional equation to an equation of only one dimension. Furthermore, under almost-sure convexity of the loss function, efficient search bounds for the one-dimensional fixed-point equation are available. This result generalizes an earlier result in efficient computation of implicit updates on generalized linear models [11, Algorithm 1].

Lemma 1. *Suppose that Assumption 1 holds. For a fixed data point $\xi = (x, y)$, let $L'(\theta, \xi) = \frac{\partial L(\theta,\xi)}{\partial(x^\intercal\theta)} \stackrel{\text{def}}{=} \frac{\partial L(x^\intercal\theta,y)}{\partial(x^\intercal\theta)}$ and $L''(\theta, \xi) = \frac{\partial L'(\theta,\xi)}{\partial(x^\intercal\theta)}$. Then, almost-surely,*

$$\nabla L(\theta_n, \xi_n) = s_n \nabla L(\theta_{n-1}, \xi_n); \tag{6}$$

the scalar s_n satisfies the fixed-point equation,

$$s_n \kappa_{n-1} = L'\left(\theta_{n-1} - s_n \gamma_n \kappa_{n-1} x_n, \xi_n\right), \tag{7}$$

where $\kappa_{n-1} \triangleq L'(\theta_{n-1}, \xi_n)$. Moreover, if $L''(\theta, \xi) \geq 0$ almost-surely for all $\theta \in \Theta$, then

$$s_n \in \begin{cases} [\kappa_{n-1}, 0) & \text{if } \kappa_{n-1} < 0, \\ [0, \kappa_{n-1}] & \text{otherwise.} \end{cases}$$

Remarks. Lemma 1 has two parts. First, it shows that the implicit update can be performed by obtaining s_n from the fixed-point Eq.(6), and then using $\nabla L(\theta_n, \xi_n) = s_n \nabla L(\theta_{n-1}, \xi_n)$ in the implicit update (3). The fixed-point equation can be solved through a numerical root-finding procedure [4, 5, 11]. Second, when the loss function is convex, then narrow search bounds for s_n are available. This property holds, for example, when the loss function is the negative log-likelihood in the exponential family of models.

3.2 Non-asymptotic analysis

Theorem 1. *Suppose that Assumptions 1, 2, 3(a), and 4 hold. Define $\delta_n \triangleq \mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$, and constants $\Gamma^2 = 4\lambda_0^2 \sum \gamma_i^2 < \infty$, $\epsilon = (1 + \gamma_1(\phi - \underline{\lambda_f}))^{-1}$, and $\lambda = 1 + \gamma_1 \underline{\lambda_f} \epsilon$. Then, there exists constant $n_0 > 0$ such that, for all $n > 0$,*

$$\delta_n \leq (8\lambda_0^2 \gamma_1 \lambda / \underline{\lambda_f} \epsilon) n^{-\gamma} + e^{-\log \lambda \cdot n^{1-\gamma}} [\delta_0 + \lambda^{n_0} \Gamma^2].$$

Remarks. According to Theorem 1, the convergence rate of the implicit iterates θ_n is $\mathcal{O}(n^{-\gamma})$. This matches earlier results on rates of classic SGD [1, 7]. The most important difference, however, is that the implicit procedure discounts the initial conditions δ_0 at an exponential rate, regardless of the specification of the learning rate. As shown by Moulines and Bach [7, Theorem 1], in classic SGD there exists a term $\exp(\lambda_1^2 \gamma_1^2 n^{1-2\gamma})$ in front of the initial conditions, which can be catastrophic if the learning rate parameter γ_1 is misspecified. In contrast, the implicit iterates are unconditionally stable, i.e., any specification of the learning rate will lead to a stable discounting of the initial conditions.

Theorem 2. *Consider the AI-SGD procedure (4), and suppose that Assumptions 2, 3(a), 3(c), 4, and 5 hold, and λ is defined as in Theorem 1. Then,*

$$\begin{aligned}(\mathbb{E}\left(||\bar{\theta}_n - \theta_\star||^2\right))^{1/2} \leq & \frac{1}{\sqrt{n}} \left(\operatorname{tr}(\nabla^2 \ell(\theta_\star)^{-1} \Sigma \nabla^2 \ell(\theta_\star)^{-1})\right)^{1/2} + \mathcal{O}(n^{-1+\gamma/2}) + \mathcal{O}(n^{-\gamma}) \\ & + \mathcal{O}(\exp(-\log \lambda \cdot n^{1-\gamma}/2).\end{aligned}$$

Remarks. The full version of Theorem 2, which includes all constants, is in the extended paper [12]. Even in its shortened form, Theorem 2 delivers three main results. First, the iterates $\bar{\theta}_n$ attain the Cramér-Rao lower bound, i.e., any other unbiased estimator of $\theta_\star$ cannot have lower MSE than $\bar{\theta}_n$. From an optimization perspective, $\bar{\theta}_n$ attains the rate $\mathcal{O}(1/n)$, which is optimal for first-order methods [8]. This result matches the asymptotic optimality of averaged iterates from classic SGD procedures, which has been proven by Polyak and Juditsky [9]. Second, the remaining rates are $\mathcal{O}(n^{-2+\gamma})$

and $\mathcal{O}(n^{-2\gamma})$. This implies the optimal choice $\gamma = 2/3$ for the exponent of the learning rate. It extends the results of Ruppert [10], and more recently by Xu [14], and Moulines and Bach [7], on optimal exponents for classic SGD procedures. Third, as with non-averaged implicit iterates in Theorem 1, the averaged iterates $\bar{\theta}_n$ have a decay of the initial conditions regardless of the specification of the learning rate parameter. This stability property is inherited from the underlying implicit SGD procedure (3) that is being averaged. In contrast, averaged iterates of classic SGD procedures can have large terms amplifying arbitrarily the initial conditions [7, Theorem 3].

References

[1] Albert Benveniste, Pierre Priouret, and Michel Métivier. Adaptive algorithms and stochastic approximations. 1990.

[2] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

[3] Léon Bottou. Stochastic learning. In *Advanced lectures on machine learning*, pages 146–168. Springer, 2004.

[4] Jyrki Kivinen, Manfred K Warmuth, and Babak Hassibi. The p-norm generalization of the lms algorithm for adaptive filtering. *Signal Processing, IEEE Transactions on*, 54(5):1782–1793, 2006.

[5] Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.

[6] Lennart Ljung, Georg Ch Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*, volume 17. Springer, 1992.

[7] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[8] Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

[9] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[10] David Ruppert. Efficient estimators from a slowly convergent robbins-monro process. Technical report, School of Operations Research and Industrial Engineering, Cornell University, 1988.

[11] Panagiotis Toulis, Edoardo Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 667–675, 2014.

[12] Panos Toulis and Edoardo M. Airoldi. Implicit stochastic gradient descent. arxiv preprint:1408.2923, 2015. URL http://arxiv.org/abs/1408.2923.

[13] Panos Toulis and Edoardo M Airoldi. Scalable estimation strategies based on stochastic approximations: classical results and new insights. *Statistics and computing*, 25(4):781–795, 2015.

[14] Wei Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*, 2011.

[15] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.