Stability and Optimality in Stochastic Gradient Descent

Dustin Tran, Panos Toulis, Edoardo Airoldi Harvard University, Department of Statistics

Introduction

Stochastic gradient methods have increasingly become popular for large-scale optimization. However, they are often numerically unstable and statistically inefficient because of their sensitivity to additional hyperparameters and suboptimal usage of the data's information. We propose a new learning procedure, termed *averaged implicit stochastic gradient descent* (ai-SGD), which combines stability through proximal (implicit) updates and statistical efficiency through averaging of the iterates.

We prove that ai-SGD is computationally efficient, statistically optimal, and stable. Furthermore, we demonstrate in experiments that its performance is comparable to state-of-the-art algorithms, with the added bonuses of statistical efficiency and stability.

Background

Consider a random variable $\xi \in \Xi \subseteq \mathbb{R}^d$, a parameter space $\Theta \subseteq \mathbb{R}^d$, and a loss function $\ell : \Theta \times \Xi \to \mathbb{R}$. We wish to solve the following stochastic optimization problem:

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

where the expectation is with respect to ξ. Formulation (1) encompasses a wide variety of machine learning tasks. For example, learning through least-mean squares, logistic regression or SVM, can be cast into (1) by considering ℓ as the KL-divergence between the distribution of ξ and the model family parameterized by θ.

If an empirical distribution of ξ is used, one recovers the problem of empirical loss minimization, which includes maximum likelihood estimation (MLE), or maximum a posteriori (MAP) if there are regularization terms, which are widely used in machine learning and statistics.

The standard stochastic gradient descent, which we term an *explicit* method, forms an update corresponding to information of the score function evaluated at the previous iterate, c.f., our method.

Method

We propose a stochastic approximation procedure to solve (1) defined for datapoints $n = 1, 2, \ldots$, as follows:

$$\theta_n = \theta_{n-1} - \gamma_n \partial\ell(\theta_n, \xi_n), \quad \theta_0 \in \Theta, \tag{2}$$

$$\bar{\theta}_n = (1/n)\sum_{i=1}^{n} \theta_i, \tag{3}$$

where $\{\xi_1, \xi_2, \ldots\}$ are i.i.d. realizations of ξ and assumed to be a continuous stream of data, $\partial\ell(\theta, \xi_n)$ is a subgradient of the loss function with respect to θ at realized value ξ_n, and $\{\gamma_n\}$ is a non-increasing sequence of positive real numbers.

We will refer to the procedure defined by (2) and (3) as *averaged implicit stochastic gradient descent*, or ai-SGD for short. Our approximation procedure combines two ideas, namely an implicit formulation of the updates in Eq. (2) as θ_n appears on both sides of the update, and averaging of the iterates θ_n in Eq. (3).

Theory

Theorem 1.1 (Computational efficiency) *For data $\xi = (x, y)$, and a differentiable and linear loss $\ell(\theta, \xi)$, the implicit update* (2) *of* ai-SGD *is*

$$\partial\ell(\theta_n, \xi_n) = \lambda_n \nabla\ell(\theta_{n-1}, \xi_n),$$

where the scalar $\lambda_n \in \mathbb{R}$ satisfies the fixed-point equation,

$$\lambda_n = \frac{\ell'\left(\theta_{n-1} - \lambda_n\gamma_n\ell'(\theta_{n-1}, \xi_n)x_n, \xi_n\right)}{\ell'(\theta_{n-1}, \xi_n)}, \tag{4}$$

∗ Remark: Numerical solution of Eq. (4) is straightforward.

Theorem 1.2 (Statistical efficiency) *Suppose there is a positive semi-definite $d \times d$ matrix* F, *such that*

$$\partial\ell(\theta_n, \xi) - \partial\ell(\theta_\star, \xi) = \mathrm{F}(\theta_n - \theta_\star) + r_n,$$

where $\{r_n\}$ is a sequence of random variables for which $||r_n|| = o(||\theta_n - \theta_\star||)$, almost-surely. Then,

$$\bar{\theta}_n - \theta_\star = \frac{1}{n}\mathrm{D}_0^n(\theta_0 - \theta_\star) + \mathrm{F}^{-1}\bar{\varepsilon}_n + \frac{1}{n}\sum_{i=1}^{n-1}\Omega_i^n r_i,$$

where $||\mathrm{D}_0^n|| = O(1)$, $\varepsilon_i = \nabla\ell(\theta_\star, \xi_i)$, $\bar{\varepsilon}_n = (1/n)\sum_{i=1}^{n}\varepsilon_i$, and $\sum_{i=1}^{n-1}||\Omega_i^n|| = \mathrm{o(n)}$. In particular, $(\bar{\theta}_n - \theta_\star) \to \mathrm{F}^{-1}\bar{\varepsilon}_n$, in probability.

∗ Remark: $\bar{\theta}_n$ achieves the Cramér-Rao bound.

Theorem 1.3 (Stability) *Under standard assumptions, and supposing $\mathbb{E}\left[||\nabla\ell(\theta, \xi)||^2\right] = 0$,* ai-SGD *satisfies*

$$\max_i \left\{\frac{\mathbb{E}\left[||\theta_i - \theta_\star||^2\right]}{\mathbb{E}\left[||\theta_0 - \theta_\star||^2\right]}\right\} = \mathrm{O(1)}.$$

Let $2c > 0$ be a small constant and define $\tilde{n}_c = [(1 + c)\gamma\mu]^{1/\alpha}$, then for the explicit procedure,

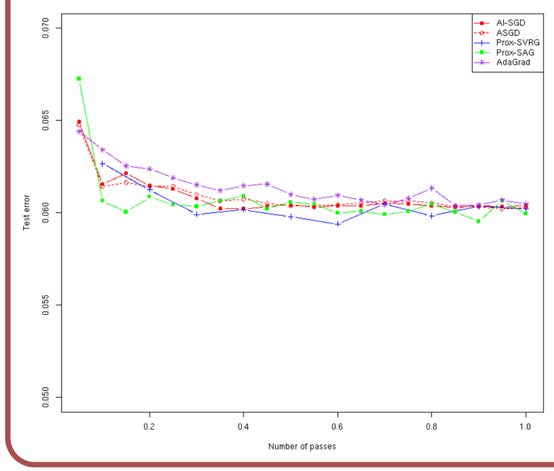
$$\max_i \left\{\frac{\mathbb{E}\left[||\theta_i - \theta_\star||^2\right]}{\mathbb{E}\left[||\theta_0 - \theta_\star||^2\right]}\right\} = \mathrm{O}((1 + 2c)^{\max\{1, \tilde{n}_c\}}).$$

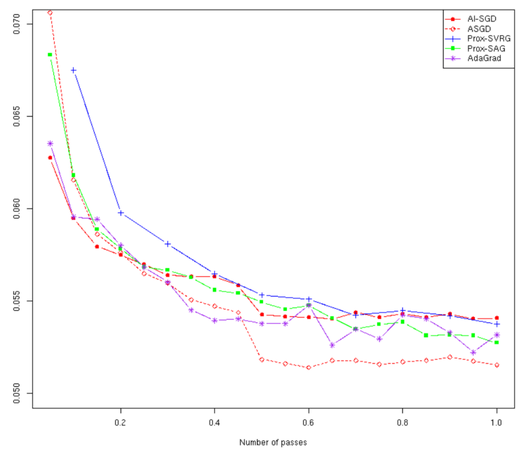
∗ Remark: Iterates of ai-SGD are unconditionally stable.

Experiments

2.1 RCV1 dataset

The task is to classify documents belonging to class CCAT, which has $d = 47,152$ features and is split into a training set of $N = 781,265$ observations and a test set of $23,149$ observations. We implement a linear SVM using hinge loss and logistic regression using log loss. We compare it to current state-of-the-art algorithms. Our results demonstrate comparable performance, measured by misclassification error, as all methods iterate over the data.





2.2 Covtype dataset

The task is to classify class 2 among 7 forest cover types, which has $d = 54$ features and is split into a training set of $N = 406,708$ observations and a test set of $174,304$ observations. Performances indicate similar behavior as in RCV1; here we perform sensitivity analysis on the regularization parameter in order to examine stability to compared to other methods. We see that while all methods achieve comparable performance for optimized hyperparameters, ai-SGD is more robust to any misspecification.