Towards stability and optimality in stochastic gradient descent

Panos Toulis, Dustin Tran and Edoardo M. Airoldi

August 26, 2016

Discussion by Ikenna Odinaka
Duke University
Outline

1. Introduction
2. Theory
3. Experiments
Consider random variable $\xi \in \Xi$, convex and compact parameter space $\Theta$, and differentiable (a.s.) loss function $L : \Theta \times \Xi \to \mathbb{R}$. We want to solve the stochastic optimization problem

$$\theta^* = \arg \min_{\theta \in \Theta} \ell(\theta)$$

(1)

where

$$\ell(\theta) = \mathbb{E}(L(\theta, \xi))$$

and the expectation is wrt $\xi$. 

Toulis et al. AI-SGD
Classic SGD and Averaged SGD

- Classic stochastic gradient descent (SGD) solves problem 1 using

$$\theta_n = \theta_{n-1} - \gamma_n \nabla L(\theta_{n-1}, \xi_n), \theta_0 \in \Theta,$$

where \(\{\xi_n\}\) are i.i.d. realizations of \(\xi\) and the learning rate \(\{\gamma_n\}\) is a non-increasing sequence of positive real numbers.

- To achieve statistical efficiency, classic SGD is merged with iterate averaging in averaged SGD (ASGD)

$$\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i$$

\[\tag{3}\]
Authors’ Contribution: AI-SGD

- Implicit SGD (improves stability) and Averaging (improves statistical efficiency) of iterates
- Both aspects have been shown in a previous 2015 arXiv publication titled “Implicit stochastic gradient descen”
- Theorem 2 is new to this AISTATS 2016 submission

Main Idea: Solve problem 1 using

\[
\theta_n = \theta_{n-1} - \gamma_n \nabla L(\theta_n, \xi_n), \theta_0 \in \Theta, \quad (4)
\]

\[
\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i \quad (5)
\]
Interpretations of Implicit SGD

**Interpretations**: Implicit update 4 is equivalent to a sequence of “improved” classic SGD procedures

\[
\begin{align*}
\theta_n^{(1)} &= \theta_{n-1} - \gamma_n \nabla L(\theta_{n-1}, \xi_n), \\
\theta_n^{(2)} &= \theta_{n-1} - \gamma_n \nabla L(\theta_n^{(1)}, \xi_n), \\
\theta_n^{(3)} &= \theta_{n-1} - \gamma_n \nabla L(\theta_n^{(2)}, \xi_n), \\
&\vdots \\
\theta_n^{(\infty)} &= \theta_{n-1} - \gamma_n \nabla L(\theta_n^{(\infty)}, \xi_n)
\end{align*}
\]

Implicit update 4 is equivalent to proximal update

\[
\theta_n = \arg \min_{\theta \in \Theta} \left\{ \frac{1}{2\gamma_n} \| \theta - \theta_{n-1} \|^2 + L(\theta, \xi_n) \right\}
\] (6)
(De)Merits of Implicit SGD

**Merit:** Implicit SGD is stable. Assuming $L$ is $\mu$-strongly convex a.s., then $\|\theta_n - \theta^*\|$ is contracting a.s.

**Demerit:**
- Solving multidimensional fixed-point equation 4 for $\theta_n$ is difficult for general models.
- Easy when $L$ is a function of natural parameter $x^T \theta$
Notation

- Let $\| \cdot \|$ denote the $L_2$ norm
- $x \triangleq y$ defines $x$ as equal to known variable $y$
- $x \overset{\text{def}}{=} y$ denotes that the value of $x$ is equal to the value of $y$, by definition
- $a_n \downarrow 0$ means $a_n > 0$, $a_n$ is monotonically non-increasing, and $a_n \to 0$
- Given sequences $a_n > 0$, $b_n > 0$, $b_n = O(a_n)$ means $\exists c > 0$ s.t. $b_n \leq ca_n \forall n$
- $b_n = o(a_n)$ means $b_n/a_n \to 0$
- For a sequence of vectors or matrices $X_n$, $X_n = O(a_n)$ and $X_n = o(a_n)$ denote the equality for the scalar norm sequence $\|X_n\|$
- Given two matrices $A$ and $B$, $A \preceq B$ denotes that $B - A$ is positive semidefinite
- $\text{tr}(A)$ denotes the trace of $A$
Main Assumptions

Assumption 1

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

$$L(\theta, \xi) \equiv L(x^T \theta, y)$$

(7)

Assumption 2

The learning rate sequence $\{\gamma_n\}$ is defined as $\gamma_n = \gamma_1 n^{-\gamma}$, 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 with parameter $\lambda_0$, i.e.,

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

(b) The map $\nabla L$ is Lipschitz with parameter $\lambda_1$, 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 with parameter $\lambda_2$, 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, $\hat{I}(\theta) \triangleq \nabla^2 L(\theta, \xi)$, has non-vanishing trace, i.e., there exists $\phi > 0$ such that $\text{tr}(\hat{I}(\theta)) \geq \phi$, almost-surely, for all $\theta \in \Theta$. The expected Fisher information matrix, $I(\theta) \triangleq \mathbb{E}(\hat{I}(\theta))$, has minimum eigenvalue $0 < \lambda_f \leq \phi$, 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 $\Sigma$,

$$\mathbb{E} \left( W_{\theta*} W_{\theta*}^T \right) \preceq \Sigma$$
Remarks on Assumptions

- Assumptions 2 and 5 are standard in the stochastic approximation literature.
- Assumption 1 restricts application to models where $L$ depends on $\theta$ through $x^T \theta$.
- This excludes models with regularization. Authors claim they can easily incorporate regularization as shown in the supp. material (no such extension).
- Authors also claim there is no need for regularization, since proximal operator (equation 6) regularizes $\theta_n$ towards $\theta_{n-1}$.
- Assumptions 3(b) and (c) are used to simplify non-asymptotic analysis. These assumptions have been relaxed in classical stochastic approximation theory.
- Assumption 3(a) is not standard in stochastic approximation literature. It only shows up in implicit SGD literature.
- **Open Problem**: Establish Theorem 1 without assumption 3(a).
- Assumption 4 has two requirements:
  - $\text{tr}(\hat{I}(\theta)) \geq \phi > 0$ is a relaxed form of strong convexity on $L(\theta, \xi)$
  - Strong convexity on $\ell(\theta)$
The main technical challenge in analyzing implicit SGD (equation 4) is that $\xi_n$ is not conditionally independent of $\theta_n$.

This implies

$$\mathbb{E}(\nabla L(\theta_n, \xi_n) | \theta_n) \neq \nabla \ell(\theta_n)$$

So, convexity properties of $\ell$ cannot be used to analyze $\mathbb{E}(\|\theta_n - \theta_*\|^2)$, as is common in the classic SGD literature.

In addition to Assumption 3(a), other authors make strict assumptions of bounded gradients $\nabla L(\theta, \xi)$ almost-surely for implicit procedure.
Computational Efficiency

- Solving the fixed-point equation 4 at every iteration can be expensive.
- For a special class of problems, the multidimensional equation can be reduced to a single-variable equation

Definition 1

Suppose that Assumption 1 holds. For observation $\xi = (x, y)$, the first derivative with respect to the natural parameter $x^T\theta$ is denoted by $L'(\theta, \xi)$, and is defined as

$$L'(\theta, \xi) \triangleq \frac{\partial L(\theta, \xi)}{\partial (x^T\theta)} \equiv \frac{\partial L(x^T\theta, y)}{\partial (x^T\theta)}$$

Similarly, $L''(\theta, \xi) \triangleq \frac{\partial L'(\theta, \xi)}{\partial (x^T\theta)}$. 
Lemma 1

Suppose that Assumption 1 holds, and consider functions $L'$, $L''$ from Definition 1. Then, almost-surely,

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

the scalar $s_n$ satisfies the fixed-point equation

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

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

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

Suppose that Assumptions 1, 2, 3(a), and 4 hold. Define
\[ \delta_n \triangleq \mathbb{E} (\| \theta_n - \theta_\star \| ^2), \] and constants
\[ \Gamma^2 = 4\lambda_0^2 \sum \gamma_i^2 < \infty, \epsilon = (1 + \gamma_1 (\phi - \lambda_f))^{-1}, \text{ and } \lambda = 1 + \gamma_1 \lambda_f \epsilon. \] Also
let \( \rho_\gamma (n) = n^{1-\gamma} \) if \( \gamma \neq 1 \) and \( \rho_\gamma (n) = \log n \) if \( \gamma = 1 \). Then, there exists
constant \( n_0 > 0 \) such that, for all \( n > 0 \),
\[
\delta_n \leq (8\lambda_0^2 \gamma_1 \lambda / \lambda_f \epsilon) n^{-\gamma} + e^{-\log \lambda \cdot \rho_\gamma (n)} [\delta_0 + \lambda n_0 \Gamma^2].
\]

Remarks on convergence rate and numerical stability:

- Convergence rate of implicit SGD iterates \( \theta_n \) is \( \mathcal{O}(n^{-\gamma}) \). Same as for classic SGD
- The initial conditions \( \delta_0 \) is reduced at an exponential rate, irrespective of the learning rate
- To contrast, in classic SGD there is an exponential term \( \exp(\lambda_1^2 \gamma_1^2 n^{1-2\gamma}) \) multiplying the initial conditions
- In classic SGD, if \( \gamma_1 \) is misspecified, the iterates can diverge
Theorem 2

Consider the AI-SGD procedure 5, and suppose Assumptions 1, 2, 3(a), 3(c), 4, and 5 hold with $\gamma < 1$. Then,

$$\left( \mathbb{E} \left( \| \tilde{\theta}_n - \theta_* \|^2 \right) \right)^{1/2} \leq \left( \text{tr} \left( \nabla^2 \ell(\theta_*)^{-1} \Sigma \nabla^2 \ell(\theta_*)^{-1} \right) / n \right)^{1/2} + O \left( n^{-1+\gamma/2} \right) + O \left( n^{-\gamma} \right) + O \left( \exp \left( - \log \lambda \cdot n^{1-\gamma/2} \right) \right).$$

Remarks:

- The iterates $\tilde{\theta}_n$ attain
  - the CRLB (statistical perspective)
  - the rate $O(1/n)$, which is optimal for first-order methods (optimization perspective)
  - Same as in averaged classic SGD

- Optimal choice of $\gamma = 2/3$

- AI-SGD inherits stability properties from implicit SGD
To demonstrate the statistical efficiency and stability of AI-SGD
Let $N = 10^6$ be the number of observations
Let $p = 20$ be the number of features
Let $\theta_* = (0, 0, \ldots, 0)^T$ be the ground truth
Let $H$ be a randomly generated symmetric matrix with eigenvalues $1/k$, for $k = 1, \ldots, p$
Let $\xi_n = (x_n, y_n)$, where $x_1, \ldots, x_N \sim \mathcal{N}_p(0, H)$ are i.i.d. normal random variables
$y_n|x_n \sim \mathcal{N}(x_n^T \theta_*, 1)$, for $n = 1, \ldots, N$.
Let $L(\theta, \xi_n) = (y_n - x_n^T \theta)^2$. Thus, $
\ell(\theta) = \mathbb{E}(L(\theta, \xi)) = (\theta - \theta_*)^T H (\theta - \theta_*)$
Constant learning rate $\gamma_n \equiv \gamma_1 \propto 1/R^2$, where $R^2 = \text{tr}(H)$ is the average radius of the data.
Linear Regression: Results

![Graph showing linear regression results](image)

- AI-SGD, $2/R^2$
- AI-SGD, $1/R^2$
- ASGD, $2/R^2$
- ASGD, $1/R^2$
- Implicit-SGD, $2/R^2$
- Implicit-SGD, $1/R^2$

Toulis et al. (AI-SGD)
Classification: Description

- Compare AI-SGD to other stochastic optimization approaches using standard benchmarks.
- Benchmarks include: COVTYPE (forest cover classification), DELTA (synthetic data in PASCAL large scale challenge), RCV1 (document classification), and MNIST (handwritten digits classification).
- AI-SGD and ASGD use learning rate schedule \( \gamma_n = \eta_0 (1 + \eta_0 n)^{-3/4} \).
- \( \eta_0 \) is obtained by preprocessing on a small subset of the data.
- Other methods, grid search used for hyperparameter tuning.
Classification: Results

- Covtype test error
- Delta test error
- Rcv1 test error
- Mnist test error

Toulis et al. AI-SGD
Sensitivity Analysis

- Unfair sensitivity comparison: regularization vs algorithmic parameters
- Authors first talk about varying learning rates, but later show and talk about regularization parameters
- The trends in the text are reversed
- I’m pretty sure there is a mistake somewhere in this section.
Sensitivity Analysis Contd

![Graph showing test error versus number of passes with AI-SGD and Prox-SVRG with different parameters.](image-url)