Towards stability and optimality in stochastic gradient descent

Supplementary material for AISTATS 2016

1 Note

Lemmas 1, 2, 3 and 4, and Corollary 1, were originally derived by Toulis and Airoldi (2014). These intermediate results (and Theorem 1) provide the necessary foundation to derive Lemma 5 (only in this supplement) and Theorem 2 on the asymptotic optimality of $\bar{\theta}_n$, which is the key result of the main paper. We fully state these intermediate results here for convenience but we point the reader to the aforementioned reference for the proofs and for more details on the theory of (non-averaged) implicit stochastic gradient descent (implicit SGD).

2 Introduction

Consider a random variable $\xi \in \Xi$, a parameter space Θ that is convex and compact, and a loss function $L : \Theta \times \Xi \to \mathbb{R}$. We wish to solve the following stochastic optimization problem:

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

where the expectation is with respect to ξ. Define the expected loss,

$$\ell(\theta) = \mathbb{E}\left(L(\theta, \xi)\right), \tag{2}$$

where L is differentiable almost-surely. In this work we study a stochastic approximation procedure to solve (1) defined through the iterations

$$\boldsymbol{\theta_n} = \theta_{n-1} - \gamma_n \nabla L(\boldsymbol{\theta_n}, \xi_n), \;\; \theta_0 \in \Theta, \tag{3}$$

$$\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i, \tag{4}$$

where $\{\xi_1, \xi_2, \ldots\}$ are i.i.d. realizations of ξ, and $\nabla L(\theta, \xi_n)$ is the gradient of the loss function with respect to θ given realized value ξ_n. The sequence $\{\gamma_n\}$ is a non-increasing sequence of positive real numbers. We will refer to procedure defined by (3) and (4) as *averaged implicit stochastic* *gradient descent*, or averaged implicit SGD (AI-SGD) for short. Procedure AI-SGD combines two ideas, namely an implicit update in Eq. (3) as θ_n appears on both sides of the update, and averaging of the iterates θ_n in Eq. (4).

3 Notation and assumptions

Let $\mathcal{F}_n = \{\theta_0, \xi_1, \xi_2, \ldots, \xi_n\}$ denote the filtration that process θ_n (3) is adapted to. The norm $||\cdot||$ will denote the L_2 norm. The symbol $\triangleq$ indicates a definition, and the symbol $\stackrel{\text{def}}{=}$ denotes "equal by definition". For example, $x \triangleq y$ defines x as equal to known variable y, whereas $x \stackrel{\text{def}}{=} y$ denotes that the value of x is equal to the value of y, by definition. We will not use this formalism when defining constants. For two positive sequences a_n, b_n, we write $b_n = \mathcal{O}(a_n)$ if there exists a fixed $c > 0$ such that $b_n \leq c a_n$, for all n; also, $b_n = o(a_n)$ if $b_n/a_n \to 0$. When a positive scalar sequence a_n is monotonically decreasing to zero, we write $a_n \downarrow 0$. Similarly, for a sequence X_n of vectors or matrices, $X_n = \mathcal{O}(a_n)$ denotes that $||X_n|| = \mathcal{O}(a_n)$, and $X_n = o(a_n)$ denotes that $||X_n|| = o(a_n)$. For two matrices A, B, $A \preceq B$ denotes that $B - A$ is nonnegative-definite; $\mathrm{tr}(A)$ denotes the trace of A.

We now introduce the main assumptions pertaining to the theory of this paper.

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*

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

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 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 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 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,*

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

4 Proof of Lemma 1

Definition 1. *Suppose that Assumption 1 holds. For observation $\xi = (x, y)$, the first derivative with respect to the natural parameter $x^\intercal\theta$ is denoted by $L'(\theta, \xi)$, and is defined as*

$$L'(\theta, \xi) \triangleq \frac{\partial L(\theta, \xi)}{\partial(x^\intercal\theta)} \stackrel{\text{def}}{=} \frac{\partial L(x^\intercal\theta, y)}{\partial(x^\intercal\theta)}. \tag{6}$$

Similarly, $L''(\xi, \theta) \triangleq \frac{\partial L'(\theta,\xi)}{\partial(x^\intercal\theta)}$.

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

$$\nabla L(\theta_n, \xi_n) = s_n \nabla L(\theta_{n-1}, \xi_n); \tag{7}$$

the scalar s_n satisfies the fixed-point equation,

$$s_n\kappa_{n-1} = L'\left(\theta_{n-1} - s_n\gamma_n\kappa_{n-1}x_n, \xi_n\right), \tag{8}$$

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

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

Proof. See Toulis and Airoldi (2014, Theorem 4.1). □

5 Proof of Theorem 1

5.1 Useful lemmas

In this section, we will present the intermediate lemmas on recursions that will be useful for the non-asymptotic analysis of the implicit procedures.

Lemma 2. *Consider a sequence b_n such that $b_n \downarrow 0$ and $\sum_{i=1}^{\infty} b_i = \infty$. Then, there exists a positive constant $K > 0$, such that*

$$\prod_{i=1}^{n} \frac{1}{1 + b_i} \leq \exp(-K \sum_{i=1}^{n} b_i). \tag{9}$$

Proof. See Toulis and Airoldi (2014, Lemma B.1). □

Lemma 3. *Consider scalar sequences $a_n \downarrow 0, b_n \downarrow 0$, and $c_n \downarrow 0$ such that, $a_n = o(b_n)$, and $A \triangleq \sum_{i=1}^{\infty} a_i < \infty$. Suppose there exists n' such that $c_n/b_n < 1$ for all $n > n'$. Define,*

$$\delta_n \triangleq \frac{1}{a_n}(a_{n-1}/b_{n-1} - a_n/b_n) \text{ and } \zeta_n \triangleq \frac{c_n}{b_{n-1}}\frac{a_{n-1}}{a_n}, \tag{10}$$

and suppose that $\delta_n \downarrow 0$ and $\zeta_n \downarrow 0$. Fix $n_0 > 0$ such that $\delta_n + \zeta_n < 1$ and $(1 + c_n)/(1 + b_n) < 1$, for all $n \geq n_0$.

Consider a positive sequence $y_n > 0$ that satisfies the recursive inequality,

$$y_n \leq \frac{1 + c_n}{1 + b_n} y_{n-1} + a_n. \tag{11}$$

Then, for every $n > 0$,

$$y_n \leq K_0 \frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n (1 + c_1)^{n_0} A, \tag{12}$$

where $K_0 = (1 + b_1)(1 - \delta_{n_0} - \zeta_{n_0})^{-1}$, and $Q_i^n = \prod_{j=i}^n (1 + c_i)/(1 + b_i)$, such that $Q_i^n = 1$ if $n < i$, by definition.

Proof. See Toulis and Airoldi (2014, Lemma B.2). □

Corollary 1. *In Lemma 3 assume $a_n = a_1 n^{-\alpha}$ and $b_n = b_1 n^{-\beta}$, and $c_n = 0$, where $a_1, b_1, \beta > 0$ and $\max\{\beta, 1\} < \alpha < 1 + \beta$, and $\beta \neq 1$. Then,*

$$y_n \leq 2\frac{a_1(1 + b_1)}{b_1} n^{-\alpha+\beta} + \exp(-\log(1 + b_1)n^{1-\beta})[y_0 + (1 + b_1)^{n_0} A], \tag{13}$$

where $n_0 > 0$ and $A = \sum_i a_i < \infty$. If $\beta = 1$ then the above inequality holds by replacing the term $n^{1-\beta}$ with $\log n$.

Proof. See Toulis and Airoldi (2014, Corollary B.1). □

Lemma 4. *Suppose Assumptions 1, 3(a), and 4 hold. Then, almost surely,*

$$s_n \geq \frac{1}{1 + \gamma_n \phi}, \tag{14}$$

$$||\theta_n - \theta_{n-1}||^2 \leq 4\lambda_0^2 \gamma_n^2, \tag{15}$$

where s_n is defined in Lemma 1, and θ_n is the nth iterate of implicit SGD (3).

Proof. See Toulis and Airoldi (2014, Lemma B.3). □

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 $\Gamma^2 = 4\lambda_0^2 \sum \gamma_i^2 < \infty$, $\epsilon = (1+\gamma_1(\phi - \underline{\lambda_f}))^{-1}$, and $\lambda = 1+\gamma_1\underline{\lambda_f}\epsilon$. 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 $n > 0$,*

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

Proof. See Toulis and Airoldi (2014, Theorem 3.1). □

Remarks. #1. Assuming Lipschitz continuity of the gradient ∇L instead of function L, i.e., Assumption 3(b) over Assumption 3(a) would not alter the main result of Theorem 1 about the $\mathcal{O}(n^{-\gamma})$ rate of the mean-squared error. Assuming Lipschitz continuity with constant λ_1 of ∇L and boundedness of $\mathbb{E}\left(||\nabla L(\theta_\star, \xi_n)||^2\right) \leq \sigma^2$, as it is typical in the literature, would simply add a term $\gamma_n^2 \lambda_1^2 \mathbb{E}\left(||\theta_n - \theta_\star||^2\right) + \gamma_n^2 \sigma^2$ in the corresponding recursive inequality. Specifically, by Lemma 1, $s_n \leq 1$, and thus

$$\begin{aligned}
\mathbb{E}\left(||\nabla L(\theta_n, \xi_n)||^2\right) = \mathbb{E}\left(s_n^2 ||\nabla L(\theta_{n-1}, \xi_n)||^2\right) &\leq \mathbb{E}\left(||\nabla L(\theta_{n-1}, \xi_n)||^2\right) \\
&= \mathbb{E}\left(||\nabla L(\theta_{n-1}, \xi_n) - \nabla L(\theta_\star, \xi_n) + \nabla L(\theta_\star, \xi_n)||^2\right) \\
&\leq \lambda_1^2 \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^2\right) + \gamma_n^2 \mathbb{E}\left(||\nabla L(\theta_\star, \xi_n)||^2\right) \\
&\leq \lambda_1^2 \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^2\right) + \gamma_n^2 \sigma^2. \qquad (16)
\end{aligned}$$

The recursion for the implicit errors would then be

$$\mathbb{E}\left(||\theta_n - \theta_\star||^2\right) \leq \left(\frac{1}{1 + \gamma_n \underline{\lambda_f} \epsilon} + \lambda_1^2 \gamma_n^2\right) \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^2\right) + \gamma_n^2 \sigma^2,$$

which also implies the $\mathcal{O}(n^{-\gamma})$ convergence rate. However, it is an open problem whether it is possible to derive a nice stability property for implicit SGD under Assumption 3(b) similar to the result of Theorem 1 under Assumption 3(a).

Remarks. #2. An assumption of almost-sure convexity can simplify the analysis significantly. For example, similar to the assumption of Ryu and Boyd (2014), assume that $L(\theta, \xi)$ is convex almost surely such that

$$(\theta_n - \theta_\star)^\intercal \nabla L(\theta_n, \xi_n) \geq \frac{\mu_n}{2} ||\theta_n - \theta_\star||^2, \qquad (17)$$

where $\mu_n \geq 0$ and $\mathbb{E}(\mu_n) = \mu > 0$. Then,

$$\begin{aligned}
\theta_n + 2\gamma_n \nabla L(\theta_n, \xi_n) &= \theta_{n-1} \quad \text{[\textit{by definition of implicit SGD} (3)]} \\
||\theta_n - \theta_\star||^2 + 2\gamma_n (\theta_n - \theta_\star)^\intercal \nabla L(\theta_n, \xi_n) &\leq ||\theta_{n-1} - \theta_\star||^2. \\
(1 + \gamma_n \mu_n) ||\theta_n - \theta_\star||^2 &\leq ||\theta_{n-1} - \theta_\star||^2. \\
\mathbb{E}\left(||\theta_n - \theta_\star||^2\right) &\leq \frac{1}{1 + \gamma_n \mu} \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^2\right) + \mathrm{SD}(1 + \gamma_n \mu_n) \mathrm{SD}(||\theta_n - \theta_\star||^2), \qquad (18)
\end{aligned}$$

where the last inequality follows from the identity $\mathbb{E}(XY) \geq \mathbb{E}(X)\,\mathbb{E}(Y) - \mathrm{SD}(X)\mathrm{SD}(Y)$. However, $\mathrm{SD}(1 + \gamma_n \mu_n) = \mathcal{O}(\gamma_n)$, and assuming bounded θ_n we get

$$\mathbb{E}\left(||\theta_n - \theta_\star||^2\right) \leq \frac{1}{1 + \gamma_n \mu} \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^2\right) + \mathcal{O}(\gamma_n), \qquad (19)$$

which indicates a fast convergence towards $\theta_\star$. It is also possible to work with the recursion

$$||\theta_n - \theta_\star||^2 \leq \frac{1}{1 + \gamma_n \mu_n} ||\theta_{n-1} - \theta_\star||^2, \qquad (20)$$

and then use a stochastic version of Lemma 3 although the analysis would be more complex in this case.

6 Proof of Theorem 2

In this section, we prove Theorem 2. To do so, we need bounds for $\mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$, which are available through Theorem 1, but also bounds for $\mathbb{E}\left(||\theta_n - \theta_\star||^4\right)$, which are established in the following lemma.

Lemma 5. *Suppose that Assumptions 1, 2, 3(a), and 4 hold. For a constant $K_3 > 0$, define $\zeta_n \triangleq \mathbb{E}\left(||\theta_n - \theta_\star||^4\right)$, and constants $\Delta^3 \triangleq K_3 \sum \gamma_i^3 < \infty$, $\epsilon \triangleq (1 + \gamma_1(\phi - \underline{\lambda_f}))^{-1}$, and $\lambda \triangleq 1 + \gamma_1 \underline{\lambda_f}\epsilon$. Then, there exists constant n_0 such that, for all $n > 0$,*

$$\zeta_n \leq (2K_3\gamma_1^2\lambda/\underline{\lambda_f}\epsilon)n^{-2\gamma} + e^{-\log\lambda\cdot\rho_\gamma(n)}[\zeta_0 + \lambda^{n_0}\Delta^3].$$

Proof. Define $W_n \triangleq s_n(\theta_{n-1} - \theta_\star)^\intercal \nabla L(\theta_{n-1}, \xi_n)$ for compactness, and proceed as folllows,

$$\begin{aligned}
||\theta_n - \theta_\star||^2 &= ||\theta_{n-1} - \theta_\star||^2 - 2\gamma_n s_n(\theta_{n-1} - \theta_\star)^\intercal \nabla L(\theta_{n-1}, \xi_n) + \gamma_n^2 ||\nabla L(\theta_n, \xi_n)||^2 \\
||\theta_n - \theta_\star||^2 &= ||\theta_{n-1} - \theta_\star||^2 - 2\gamma_n W_n + \gamma_n^2 ||\nabla L(\theta_n, \xi_n)||^2 \quad \textit{[by definition]} \\
||\theta_n - \theta_\star||^2 &\leq ||\theta_{n-1} - \theta_\star||^2 - 2\gamma_n W_n + 4\lambda_0^2\gamma_n^2, \\
||\theta_n - \theta_\star||^4 &\leq ||\theta_{n-1} - \theta_\star||^4 + 4\gamma_n^2 W_n^2 + 16\lambda_0^4\gamma_n^4 \\
&\quad - 2\gamma_n ||\theta_{n-1} - \theta_\star||^2 W_n + 4\lambda_0^2\gamma_n^2 ||\theta_{n-1} - \theta_\star||^2 - 8\lambda_0^2\gamma_n^3 W_n.
\end{aligned} \tag{21}$$

By Lemma 4 we have

$$\mathbb{E}\left(W_n \middle| \mathcal{F}_{n-1}\right) \geq \frac{\underline{\lambda_f}}{2(1+\gamma_n\phi)} ||\theta_{n-1} - \theta_\star||^2. \tag{22}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left(W_n^2 \middle| \mathcal{F}_{n-1}\right) &\stackrel{\text{def}}{=} \mathbb{E}\left([s_n(\theta_{n-1} - \theta_\star)^\intercal \nabla L(\theta_{n-1}, \xi_n)]^2 \middle| \mathcal{F}_{n-1}\right) \\
&\stackrel{\text{def}}{=} \mathbb{E}\left([(\theta_{n-1} - \theta_\star)^\intercal \nabla L(\theta_n, \xi_n)]^2 \middle| \mathcal{F}_{n-1}\right) \quad \textit{[by Lemma 1]} \\
&\leq ||\theta_{n-1} - \theta_\star||^2 \mathbb{E}\left(||\nabla L(\theta_n, \xi_n)||^2 \middle| \mathcal{F}_{n-1}\right) \quad \textit{[by Cauchy-Schwartz inequality]} \\
&\leq 4\lambda_0^2 ||\theta_{n-1} - \theta_\star||^2 \quad \textit{[by Lemma 4]}
\end{aligned} \tag{23}$$

Define $B_n \triangleq \mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$ for notational brevity. We use results (22) and (23) to get

$$\begin{aligned}
\mathbb{E}\left(||\theta_n - \theta_\star||^4\right) &\leq \left(1 - \frac{\gamma_n\underline{\lambda_f}}{1+\gamma_n\phi}\right) \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^4\right) + 4\lambda_0^2\gamma_n^2\left(5 - \frac{\gamma_n\underline{\lambda_f}}{1+\gamma_n\phi}\right)B_{n-1} + 16\lambda_0^4\gamma_n^4 \\
\mathbb{E}\left(||\theta_n - \theta_\star||^4\right) &\leq \left(1 - \frac{\gamma_n\underline{\lambda_f}}{1+\gamma_n\phi}\right) \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^4\right) + 20\lambda_0^2\gamma_n^2 B_{n-1} + 16\lambda_0^4\gamma_n^4 \\
\mathbb{E}\left(||\theta_n - \theta_\star||^4\right) &\leq \frac{1}{1+\gamma_n\underline{\lambda_f}\epsilon} \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^4\right) + 20\lambda_0^2\gamma_n^2 B_{n-1} + 16\lambda_0^4\gamma_n^4. \quad \textit{[by Assumption 4]} \\
\mathbb{E}\left(||\theta_n - \theta_\star||^4\right) &\leq \frac{1}{1+\gamma_n\underline{\lambda_f}\epsilon} \mathbb{E}\left(||\theta_{n-1} - \theta_\star||^4\right) + K_0\gamma_n^3 + e^{-\log\lambda\cdot n^{1-\gamma}}K_1 + K_2\gamma_n^4, \quad \textit{[by Theorem 1]}
\end{aligned} \tag{24}$$

where $\lambda = (1 + \gamma_1(\phi - \underline{\lambda_f}))^{-1}$ and $\Gamma^2 = 4\lambda_0^2 \sum \gamma_i^2$, (as in Theorem 1), $K_0 \triangleq 160\lambda_0^4\lambda/\underline{\lambda_f}$, $K_1 \triangleq 20\lambda_0^2(\mathbb{E}\,(||\theta_0 - \theta_\star||^2) + \lambda^{n_0}\Gamma^2)$, and $K_2 \triangleq 16\lambda_0^4$, and n_0 is a constant defined in the proof of Theorem 1.

Now, define

$$K_3 \triangleq K_0 + K_2\gamma_1 + \max\{\frac{e^{-\log\lambda\cdot\rho_\gamma(n)K_1}}{\gamma_n^3}\}, \tag{25}$$

which exists and is finite. Through simple algebra it is easy to verify that

$$K_0\gamma_n^3 + e^{-\log\lambda\cdot\rho_\gamma(n)}K_1 + K_2\gamma_n^4 \leq K_3\gamma_n^3, \tag{26}$$

for all n. Therefore, we can simplify Ineq. (24) as

$$\mathbb{E}\left(||\theta_n - \theta_\star||^4\right) \leq \frac{1}{1 + \gamma_n\underline{\lambda_f}\epsilon}\mathbb{E}\left(||\theta_{n-1} - \theta_\star||^4\right) + K_3\gamma_n^3. \tag{27}$$

We can now apply Corollary 1 with $a_n \equiv K_3\gamma_n^3$ and $b_n \equiv \gamma_n\underline{\lambda_f}\epsilon$ to derive the final bounds for $\mathbb{E}\,(||\theta_n - \theta_\star||^4)$. □

We now evaluate the mean squared error of the averaged iterates, $\bar{\theta}_n$.

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

$$\begin{aligned}(\mathbb{E}\left(||\bar{\theta}_n - \theta_\star||^2\right))^{1/2} \leq & \frac{1}{\sqrt{n}}\left(\mathrm{trace}(\nabla^2\ell(\theta_\star)^{-1}\Sigma\nabla^2\ell(\theta_\star)^{-1})\right)^{1/2} \\ & + \frac{2\gamma+1}{\underline{\lambda_f}^{1/2}\gamma_1}(8\lambda_0^2\gamma_1\lambda/\underline{\lambda_f}\epsilon)^{1/2}n^{-1+\gamma/2} \\ & + \frac{2\gamma+1}{\underline{\lambda_f}^{1/2}n\gamma_n}[\delta_0 + \lambda^{n_{0,1}}\Gamma^2]^{1/2}e^{-\log\lambda\cdot n^{1-\gamma}/2} \\ & + \frac{\lambda_2}{2\underline{\lambda_f}^{1/2}}(2K_3\gamma_1^2\lambda/\underline{\lambda_f}\epsilon)^{1/2}n^{-\gamma} \\ & + \frac{\lambda_2}{2n\underline{\lambda_f}^{1/2}}[\zeta_0 + \lambda^{n_{0,2}}\Delta^3]^{1/2}K_2(n).\end{aligned} \tag{28}$$

where $K_2(n) = \sum_{i=1}^n \exp\left(-\log\lambda\cdot i^{1-\gamma}/2\right)$, and constants $\lambda, \epsilon, n_{0,1}, \delta_0, \Gamma^2$ are defined in Theorem 1 (susbtituting n_0 for $n_{0,1}$), and $\zeta_0, n_{0,2}, \Delta^3$ are defined in Lemma 5, substituting (n_0 for $n_{0,2}$).

Proof. We leverage a result shown for averaged explicit stochastic gradient descent. In particular, it has been shown that the squared error for the averaged iterate satisfies:

$$\begin{aligned}(\mathbb{E}\left(||\bar{\theta}_n - \theta_\star||^2\right))^{1/2} \leq & \frac{1}{\sqrt{n}}\left(\mathrm{trace}(\nabla^2\ell(\theta_\star)^{-1}\Sigma\nabla^2\ell(\theta_\star)^{-1})\right)^{1/2} \\ & + \frac{2\gamma+1}{\underline{\lambda_f}^{1/2}n\gamma_n}(\mathbb{E}\left(||\theta_n - \theta_\star||^2\right))^{1/2} \\ & + \frac{\lambda_2}{2n\underline{\lambda_f}^{1/2}}\sum_{i=1}^n(\mathbb{E}\left(||\theta_i - \theta_\star||^4\right)^{1/2}.\end{aligned} \tag{29}$$

The proof technique for (29) was first devised by Polyak and Juditsky (1992), but was later refined by Xu (2011), and Moulines and Bach (2011). In this paper,we follow the formulation of Moulines and Bach (2011, Theorem 3, page 20); the derivation of Ineq.(29) for the implicit procedure is identical to the derivation for the explicit one, however the two procedures differ in the terms that appear in the bound (29).

All such terms in (29) have been bounded in the previous sections. In particular, we can use Theorem 1 for $\mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$; we can also use Theorem 2 and the concavity of the square-root to derive

$$\sum_{i=1}^{n} (\mathbb{E}\left(||\theta_i - \theta_\star||^4\right)^{1/2} \leq \sum_{i=1}^{n} \left((2K_3\gamma_1^2\lambda/\underline{\lambda_f}\epsilon)^{1/2} i^{-\gamma} + e^{-\log\lambda \cdot i^{1-\gamma}/2}[\zeta_0 + \lambda^{n_{0,2}}\Delta^3]^{1/2} \right)$$
$$\leq (2K_3\gamma_1^2\lambda/\underline{\lambda_f}\epsilon)^{1/2} n^{1-\gamma} + K_2(n)[\zeta_0 + \lambda^{n_{0,2}}\Delta^3]^{1/2}, \tag{30}$$

where $K_2(n) = \sum_{i=1}^{n} \exp\left(-\frac{\log\lambda}{2} i^{1-\gamma}\right)$, $\zeta_0 = \mathbb{E}\left(||\theta_0 - \theta_\star||^4\right)$, and Δ^3, $n_{0,2}$ are defined in Lemma 5, substituting n_0 for $n_{0,2}$. Similarly, using Theorem 1,

$$(\mathbb{E}\left(||\theta_n - \theta_\star||^2\right)^{1/2} \leq (8\lambda_0^2\gamma_1\lambda/\underline{\lambda_f}\epsilon)^{1/2} n^{-\gamma/2} + e^{-\log\lambda \cdot n^{1-\gamma}/2}[\delta_0 + \lambda^{n_{0,1}}\Gamma^2]^{1/2},$$

where $\delta_0 = \mathbb{E}\left(||\theta_n - \theta_\star||^2\right)$, and $n_{0,1}, \Gamma^2$ are defined in Theorem 1, substituing $n_{0,1}$ for n_0. These two bounds can be used in Ineq.(29) and thus yield the result of Theorem 2. □

7 Data sets used in experiments

	description	type	features	training set	test set	λ
covtype	forest cover type	sparse	54	464,809	116,203	10^{-6}
delta	synthetic data	dense	500	450,000	50,000	10^{-2}
rcv1	text data	sparse	47,152	781,265	23,149	10^{-5}
mnist	digit image features	dense	784	60,000	10,000	10^{-3}
sido	molecular activity	dense	4,932	10,142	2,536	10^{-3}
alpha	synthetic data	dense	500	400k	50k	10^{-5}
beta	synthetic data	dense	500	400k	50k	10^{-4}
gamma	synthetic data	dense	500	400k	50k	10^{-3}
epsilon	synthetic data	dense	2000	400k	50k	10^{-5}
zeta	synthetic data	dense	2000	400k	50k	10^{-5}
fd	character image	dense	900	1000k	470k	10^{-5}
ocr	character image	dense	1156	1000k	500k	10^{-5}
dna	DNA sequence	sparse	800	1000k	1000k	10^{-3}

Table 1: Summary of data sets and the L_2 regularization parameter λ used

Table 1 includes a full summary of all data sets considered in our experiments. The majority of regularization parameters are set according to Xu (2011).

References

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, **30**(4), 838–855.

Ryu, E. K. and Boyd, S. (2014). Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*.

Toulis, P. and Airoldi, E. M. (2014). Implicit stochastic gradient descent. *arXiv preprint arXiv:1408.2923*.

Xu, W. (2011). Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*.