Submitted to the Annals of Statistics
arXiv: arXiv:1408.2923

SUPPLEMENT TO "ASYMPTOTIC AND FINITE-SAMPLE PROPERTIES OF ESTIMATORS BASED ON STOCHASTIC GRADIENTS"

By Panos Toulis and Edoardo M. Airoldi

University of Chicago and Harvard University

1. R code. All experiments were run using the R package sgd, which implements explicit SGD and implicit SGD defined in Eqs. (1) and (4) of the main paper (Toulis and Airoldi, 2016). The package is published at CRAN here http://cran.r-project.org/web/packages/sgd/index.html.

2. Useful lemmas. For convenience we restate here the assumptions underlying the technical results of the main paper.

Assumption 2.1. *The explicit SGD procedure in Eq. (1) and the implicit SGD procedure in Eq. (4), both defined in the main paper, operate under a combination of the following assumptions.*

(a) *The learning rate sequence $\{\gamma_n\}$ is defined as $\gamma_n = \gamma_1 n^{-\gamma}$, where $\gamma_1 > 0$ is the learning parameter, and $\gamma \in (0.5, 1]$.*

(b) *For the log-likelihood $\log f(Y; X, \theta)$ there exists function ℓ such that $\log f(Y; X, \theta) \equiv \ell(X^\intercal\theta; Y)$, which depends on θ only through the natural parameter $X^\intercal\theta$.*

(c) *Function ℓ is concave, twice differentiable almost surely wrt natural parameter $X^\intercal\theta$ and Lipschitz with constant L_0 wrt θ.*

(d) *For observed Fisher information matrix $\hat{\mathcal{I}}_n(\theta) = -\nabla^2\ell(X_n^\intercal\theta; Y_n)$ there exists constant $F > 0$ such that $\mathrm{trace}(\hat{\mathcal{I}}_n(\theta)) \leq F$ almost surely, for all θ. The Fisher information matrix $\mathcal{I}(\theta_\star) = \mathbb{E}\left(\hat{\mathcal{I}}_n(\theta_\star)\right)$ has minimum eigenvalue $\underline{\lambda_f} > 0$ and maximum eigenvalue $\overline{\lambda_f} < \infty$. Typical regularity conditions hold (Lehmann and Casella, 1998, Theorem 5.1, p.463).*

(e) *Every condition matrix C_n is a fixed positive-definite matrix, such that $C_n = C + \mathrm{O}(\gamma_n)$, where $||C|| = 1$, $C \succ 0$ and symmetric, and C commutes with $\mathcal{I}(\theta_\star)$. For every C_n, $\min \mathrm{eig}(C_n) = \underline{\lambda_c} > 0$, and $\max \mathrm{eig}(C_n) = \overline{\lambda_c} < \infty$.*

(f) *Let $\Xi_n = \mathbb{E}\left(\nabla \log f(Y_n; X_n, \theta_\star) \nabla \log f(Y_n; X_n, \theta_\star)^\intercal \mid \mathcal{F}_{n-1}\right)$, then $||\Xi_n - \Xi|| = \mathrm{O}(1)$ for all n, and $||\Xi_n - \Xi|| \to 0$, for a symmetric positive-definite Ξ. Let $\sigma^2_{n,s} = \mathbb{E}\left(\mathbb{I}_{||\xi_n(\theta_\star)||^2 \geq s/\gamma_n} ||\xi_n(\theta_\star)||^2\right)$, then for all $s > 0$, $\sum_{i=1}^n \sigma^2_{i,s} = \mathrm{o}(n)$ if $\gamma = 1$, and $\sigma^2_{n,s} = \mathrm{o}(1)$ otherwise.*

Next, we prove lemmas on recursions that will be useful for subsequent analysis. All results are stated under a combination of Assumptions 2.1.

LEMMA 2.1. *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*

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

PROOF. The function $x \log(1+1/x)$ is increasing-concave in $(0, \infty)$. From $b_n \downarrow 0$ it follows that $\log(1+b_n)/b_n$ is non-increasing. Consider the value $K = \log(1+b_1)/b_1$. Then, $\log(1+b_1)/b_1 \geq \log(1+b_n)/b_n$ implies that $(1+b_n)^{-1} \leq \exp(-Kb_n)$. Successive applications of this inequality yields Ineq. (1). □

LEMMA 2.2. *Consider sequences $a_n \downarrow 0, b_n \downarrow 0$, and $c_n \downarrow 0$ such that, $a_n = \mathrm{o}(b_n)$, $\sum_{i=1}^{\infty} a_i = A < \infty$, and there is n' such that $c_n/b_n < 1$ for all $n > n'$. Define,*

$$\delta_n = \frac{1}{a_n}(a_{n-1}/b_{n-1} - a_n/b_n) \text{ and } \zeta_n = \frac{c_n}{b_{n-1}} \frac{a_{n-1}}{a_n}, \tag{2}$$

and suppose that $\delta_n \downarrow 0$ and $\zeta_n \downarrow 0$.

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

$$y_n \leq \frac{1+c_n}{1+b_n} y_{n-1} + a_n. \tag{3}$$

Then, for every $n > 0$, there exist constants K_0, n_0 such that

$$y_n \leq K_0 \frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n (1+c_1)^{n_0} A, \tag{4}$$

where $Q_i^n = \prod_{j=i}^{n} (1+c_i)/(1+b_i)$, with $Q_i^n = 1$ if $n < i$, by definition.

PROOF. Pick a positive n_0 such that $\delta_n + \zeta_n < 1$ and $(1+c_n)/(1+b_n) < 1$, for all $n \geq n_0$. Also, define $K_0 = (1+b_1)(1-\delta_{n_0} - \zeta_{n_0})^{-1}$. We consider two separate cases, namely, $n < n_0$ and $n \geq n_0$, and then we will combine the respective bounds.

Analysis for $n < n_0$. We first find a crude bound for Q_{i+1}^n. It holds,

$$Q_{i+1}^n \leq (1+c_{i+1})(1+c_{i+2}) \cdots (1+c_n) \leq (1+c_1)^{n_0}, \tag{5}$$

since $c_1 \geq c_n$ ($c_n \downarrow 0$ by definition) and there are no more than n_0 terms in the product. From Ineq. (3) we get

$$\begin{aligned} y_n &\leq Q_1^n y_0 + \sum_{i=1}^{n} Q_{i+1}^n a_i \quad \text{[\textit{by expanding recursive Ineq.} (3)]} \\ &\leq Q_1^n y_0 + (1+c_1)^{n_0} \sum_{i=1}^{n} a_i \quad \text{[\textit{using Ineq.} (5)]} \\ &\leq Q_1^n y_0 + (1+c_1)^{n_0} A. \end{aligned} \tag{6}$$

This inequality also holds for $n = n_0$.

Analysis for $n \geq n_0$. In this case, we have for all $n \geq n_0$,

$$\begin{aligned} (1+b_1)\left(1-\delta_n-\zeta_n\right)^{-1} &\leq K_0 \quad \text{[\textit{by definition of} } n_0, K_0\text{]} \\ K_0(\delta_n+\zeta_n)+1+b_1 &\leq K_0 \\ K_0(\delta_n+\zeta_n)+1+b_n &\leq K_0 \quad \text{[\textit{because} } b_n \leq b_1, \text{ \textit{since} } b_n \downarrow 0\text{]} \\ \frac{1}{a_n}K_0(\frac{a_{n-1}}{b_{n-1}}-\frac{a_n}{b_n})+\frac{1}{a_n}K_0\frac{c_n a_{n-1}}{b_{n-1}}+1+b_n &\leq K_0 \quad \text{[\textit{by definition of} } \delta_n, \zeta_n\text{]} \\ a_n(1+b_n) &\leq K_0 a_n - K_0\left(\frac{(1+c_n)a_{n-1}}{b_{n-1}}-\frac{a_n}{b_n}\right) \\ a_n &\leq K_0(\frac{a_n}{b_n}-\frac{1+c_n}{1+b_n}\frac{a_{n-1}}{b_{n-1}}). \end{aligned} \tag{7}$$

Now combine Ineq. (7) and Ineq. (3) to obtain

$$(y_n - K_0\frac{a_n}{b_n}) \leq \frac{1+c_n}{1+b_n}(y_{n-1}-K_0\frac{a_{n-1}}{b_{n-1}}). \tag{8}$$

Define $s_n = y_n - K_0 a_n/b_n$. Then, from Ineq. (8), $s_n \leq \frac{1+c_n}{1+b_n}s_{n-1}$, where $\frac{1+c_n}{1+b_n} < 1$ since $n \geq n_0$. Let n_1 be the smallest integer such that $n_1 \geq n_0$ and $s_{n_1} \leq 0$. If n_1 does not exist then s_n are all positive, and thus $y_n \leq K_0 a_n/b_n$, which satisfies Ineq. (3), for all $n \geq n_0$. If n_1 exists then for all $n \geq n_1$, it follows $s_n \leq 0$, and thus $y_n \leq K_0 a_n/b_n$ for all $n \geq n_1$. For $n_0 \leq n < n_1$ all s_n are positive. Using Ineq. (8), we have $s_n \leq (\prod_{i=n_0+1}^{n}\frac{1+c_i}{1+b_i})s_{n_0} = Q_{n_0+1}^n s_{n_0}$, and thus

$$\begin{aligned} y_n - K_0\frac{a_n}{b_n} &\leq Q_{n_0+1}^n s_{n_0} \quad \text{[\textit{by definition of} } s_n\text{]} \\ y_n &\leq K_0\frac{a_n}{b_n} + Q_{n_0+1}^n y_{n_0} \quad \text{[\textit{because} } s_n \leq y_n\text{]} \\ y_n &\leq K_0\frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n(1+c_1)^{n_0}A. \quad \text{[\textit{by Ineq.} (6) \textit{on} } y_{n_0}\text{]} \end{aligned} \tag{9}$$

Combining this result with Ineq. (6) and Ineq. (9), we obtain

$$y_n \leq K_0 \frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n (1+c_1)^{n_0} A, \tag{10}$$

since $Q_i^n = 1$ for $n < i$, by definition. □

Corollary 2.1. *In Lemma 2.2 assume $a_n = a_1 n^{-\alpha}$ and $b_n = b_1 n^{-\beta}$, and $c_n = 0$, where $\alpha > \beta$, and $a_1, b_1, \beta > 0$ and $\alpha > 1$. Then, there exists $n_0 > 0$ such that for all $n \geq n_0$,*

$$y_n \leq 2\frac{a_1(1+b_1)}{b_1} n^{-\alpha+\beta} + \exp(-\log(1+b_1)\phi_\beta(n))[y_0 + (1+b_1)^{n_0} A], \tag{11}$$

where $A = \sum_i a_i < \infty$, and ϕ_β is defined as in Theorem (2.1) of the main paper; i.e., $\phi_\beta(n) = n^{1-\beta}$ if $\beta \in (0.5, 1)$, and $\phi_\beta(n) = \log n$ if $\beta = 1$.

Proof. For every $n > 2$ and $\gamma \in (0.5, 1]$ it is easy to show through induction that

$$(n-1)^{-\gamma} - n^{-\gamma} \leq 2n^{-1-\gamma}, \tag{12}$$

$$\sum_{i=1}^{n} i^{-\gamma} \geq \phi_\gamma(n). \tag{13}$$

By definition of δ_n and Ineq. (12),

$$\delta_n = \frac{1}{a_n}\left(\frac{a_{n-1}}{b_{n-1}} - \frac{a_n}{b_n}\right) = \frac{1}{a_1 n^{-\alpha}} \frac{a_1}{b_1}((n-1)^{-\alpha+\beta} - n^{-\alpha+\beta}) \leq \frac{2}{b_1} n^{-1+\beta}. \tag{14}$$

Also, $\zeta_n = 0$ since $c_n = 0$. For the rest of the proof we will suppose that Ineq. (14) holds for every n since for $n = 1$ we can simply define $\delta_1 \leq 1/2$.

Next, we take $n_0 = \lceil (4/b_1)^{1/(1-\beta)} \rceil$ so that $\delta_n < 1/2$ and $\delta_n + \zeta_n < 1$ for all $n \geq n_0$. Therefore, $K_0 = (1+b_1)(1-\delta_{n_0})^{-1} \leq 2(1+b_1)$; define $K_0 = 2(1+b_1)$. Since $c_n = 0$, it follows $Q_i^n = \prod_{j=i}^n (1+b_i)^{-1}$. Thus, for a lower bound,

$$Q_1^n \geq (1+b_1)^{-n}, \tag{15}$$

and for an upper bound,

$$Q_1^n \leq \exp(-\log(1+b_1)/b_1 \sum_{i=1}^{n} b_i), \quad \text{[by Lemma 2.1]}$$

$$Q_1^n \leq \exp(-\log(1+b_1)\phi_\beta(n)). \quad \text{[by Ineq. (13)]} \tag{16}$$

Lemma 2.2, Ineq. (15) and Ineq. (16) imply that

$$
\begin{aligned}
y_n &\leq K_0 \frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n (1+c_1)^{n_0} A \quad [\textit{by Lemma } 2.2] \\
&\leq 2\frac{a_1(1+b_1)}{b_1} n^{-\alpha+\beta} + Q_1^n [y_0 + (1+b_1)^{n_0} A] \quad [\textit{by Ineq. } (15),\ c_1 = 0] \\
&\leq 2\frac{a_1(1+b_1)}{b_1} n^{-\alpha+\beta} + \exp(-\log(1+b_1)\phi_\beta(n))[y_0 + (1+b_1)^{n_0} A], \quad (17)
\end{aligned}
$$

where the last inequality follows from Ineq. (16). □

Lemma 2.3. *Suppose Assumptions 2.1(b), (c), and (d) hold. Then, almost surely it holds*

$$\lambda_n \geq \frac{1}{1 + \gamma_n \overline{\lambda_c} F}, \tag{18}$$

$$||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}||^2 \leq 4L_0^2 \gamma_n^2, \tag{19}$$

where λ_n is defined in Theorem (3.1), and θ_n^{im} is the n-th iterate of implicit SGD, defined by Eq. (4) in the main paper.

Proof. For the first part, from Theorem (3.1) we have

$$\ell'(X_n^\intercal \theta_n^{\text{im}}; Y_n) = \lambda_n \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n), \tag{20}$$

where the derivative of the log-likelihood ℓ is with respect to the natural parameter $X^\intercal \theta$. Using definition in Eq. (4),

$$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) C_n X_n. \tag{21}$$

We use this definition of θ_n^{im} into Eq.(20) and perform a Taylor approximation on ℓ' to obtain

$$\ell'(X_n^\intercal \theta_n^{\text{im}}; Y_n) = \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) + \tilde{\ell}'' \gamma_n \lambda_n \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) X_n^\intercal C_n X_n, \tag{22}$$

where $\tilde{\ell}'' = \ell''(\delta X_n^\intercal \theta_{n-1}^{\text{im}} + (1-\delta) X_n^\intercal \theta_n^{\text{im}}; Y_n) \equiv \ell''(X_n^\intercal \tilde{\theta}; Y_n)$, and $\delta \in [0,1]$. By combining Eq. (20) with Eq. (22) and cancelling out the first derivative term we get

$$
\begin{aligned}
\lambda_n &= 1 + \tilde{\ell}'' \gamma_n \lambda_n X_n^\intercal C_n X_n \\
\lambda_n &\geq 1 + \tilde{\ell}'' \gamma_n \lambda_n \overline{\lambda_c} ||X_n||^2 \quad [\textit{by Assumption } 2.1(\text{e}) \textit{ and } \ell'' < 0] \\
\lambda_n (1 - \gamma_n \overline{\lambda_c} \tilde{\ell}'' ||X_n||^2) &\geq 1 \\
\left(1 + \gamma_n \overline{\lambda_c} \text{trace}(\hat{\mathcal{I}}(\tilde{\theta}))\right) \lambda_n &\geq 1 \quad [\textit{where } \hat{\mathcal{I}} \textit{ is the observed Fisher information}] \\
(1 + \gamma_n \overline{\lambda_c} F) \lambda_n &\geq 1 \quad [\textit{by Assumption } 2.1(\text{d})].
\end{aligned}
$$

For the second part, since the log-likelihood is differentiable (Assumption 2.1(b)) we can rewrite the definition of implicit SGD in Eq. (4) (in the main paper) as

$$\theta_n^{\text{im}} = \arg\max\{-\frac{1}{2\gamma_n}||\theta - \theta_{n-1}^{\text{im}}||^2 + \ell(X_n^\intercal\theta; Y_n)\}.$$

Therefore, setting $\theta = \theta_{n-1}^{\text{im}}$ in the above equation yields

$$\begin{aligned}
-\frac{1}{2\gamma_n}||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}||^2 + \ell(X_n^\intercal\theta_n^{\text{im}}; Y_n) &\geq \ell(X_n^\intercal\theta_{n-1}^{\text{im}}; Y_n) \\
||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}||^2 &\leq 2\gamma_n\left(\ell(X_n^\intercal\theta_n^{\text{im}}; Y_n) - \ell(X_n^\intercal\theta_{n-1}^{\text{im}}; Y_n)\right) \\
||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}||^2 &\leq 2\gamma_n L_0||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}|| \qquad [\textit{By Assumption 2.1(c)}] \\
||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}|| &\leq 2L_0\gamma_n \\
||\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}||^2 &\leq 4L_0^2\gamma_n^2.
\end{aligned}$$

□

Finite-sample analysis.

THEOREM 2.1. *Let* $\delta_n = \mathbb{E}\left(||\theta_n^{\text{im}} - \theta_\star||^2\right)$. *Suppose that Assumptions 2.1(a),(b),(c), (d), and (e) hold. Then, there exist constants* $n_0 > 0$ *and* $\kappa = 1 + 2\gamma_1\mu\underline{\lambda_c}\underline{\lambda_f}$ *for some* $\mu \in (0, 1]$ *such that,*

$$\delta_n \leq \frac{4L_0^2\overline{\lambda_c}^2\gamma_1\kappa}{\mu\underline{\lambda_f}\underline{\lambda_c}}n^{-\gamma} + \exp\left(-\log\kappa \cdot \phi_\gamma(n)\right)[\delta_0 + \kappa^{n_0}\Gamma^2],$$

where $\Gamma^2 = 4L_0^2\overline{\lambda_c}^2\sum_i\gamma_i^2 < \infty$, *and* $\phi_\gamma(n) = n^{1-\gamma}$ *if* $\gamma < 1$, *and* $\phi_\gamma(n) = \log n$ *if* $\gamma = 1$.

PROOF. Starting from the procedure defined by Eq. (4) in the main paper, we have

$$\begin{aligned}
\theta_n^{\text{im}} - \theta_\star =& \theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n C_n\nabla\log f(Y_n; X_n, \theta_n^{\text{im}}) \\
\theta_n^{\text{im}} - \theta_\star =& \theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n\lambda_n C_n\nabla\log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) \qquad [\textit{By Theorem (3.1)}] \\
||\theta_n^{\text{im}} - \theta_\star||^2 =& ||\theta_{n-1}^{\text{im}} - \theta_\star||^2 \\
& + 2\gamma_n\lambda_n(\theta_{n-1}^{\text{im}} - \theta_\star)^\intercal C_n\nabla\log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) \\
& + \gamma_n^2||C_n\nabla\log f(Y_n; X_n, \theta_n^{\text{im}})||^2. \qquad (23)
\end{aligned}$$

The last term can be simply bounded since $\nabla \log f(Y_n; X_n, \theta_n^{\mathrm{im}}) = \theta_n^{\mathrm{im}} - \theta_{n-1}^{\mathrm{im}}$ by definition; thus,

$$||C_n \nabla \log f(Y_n; X_n, \theta_n^{\mathrm{im}})||^2 \leq \overline{\lambda_c}^2 ||\theta_n^{\mathrm{im}} - \theta_{n-1}^{\mathrm{im}}||^2 \leq 4L_0^2 \overline{\lambda_c}^2 \gamma_n^2, \tag{24}$$

which holds almost surely by Lemma 2.3-Eq.(19). For the second term we can bound its expectation as

$$\begin{aligned} &\mathbb{E}(2\gamma_n \lambda_n (\theta_{n-1}^{\mathrm{im}} - \theta_\star)^\intercal C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\mathrm{im}})) \\ &\leq -\frac{2\gamma_n \overline{\lambda_f}\underline{\lambda_c}}{1 + \gamma_n \overline{\lambda_c} F} ||\theta_{n-1}^{\mathrm{im}} - \theta_\star||^2 \qquad [\textit{by strong convexity, Assumption 2.1(d), Lemma 2.3}] \end{aligned} \tag{25}$$

Taking expectations in Eq. (23) and substituting Ineqs. (24) and (25) into Eq. (23) yields the recursion,

$$\mathbb{E}\left(||\theta_n^{\mathrm{im}} - \theta_\star||^2\right) \leq (1 - \frac{2\gamma_n \underline{\lambda_f}\underline{\lambda_c}}{1 + \gamma_n \overline{\lambda_c} F}) \mathbb{E}\left(||\theta_{n-1}^{\mathrm{im}} - \theta_\star||^2\right) + 4L_0^2 \overline{\lambda_c}^2 \gamma_n^2. \tag{26}$$

The following identity holds:

$$1 - \frac{1 + a\gamma_n}{1 + b\gamma_n} \leq \frac{1}{1 + c\gamma_n}, \quad c = \frac{a}{1 + \mathbb{I}\{b > a\}(b - a)\gamma_1},$$

for all $n > 0$ and $a, b > 0$. It follows that $c = \mu a$ for

$$\mu = \frac{1}{1 + \mathbb{I}\{b > a\}(b - a)\gamma_1} \in (0, 1].$$

We can use this identity to write:

$$(1 - \frac{2\gamma_n \underline{\lambda_f}\underline{\lambda_c}}{1 + \gamma_n \overline{\lambda_c} F}) \leq \frac{1}{1 + 2\gamma_n \mu \underline{\lambda_f}\underline{\lambda_c}}, \tag{27}$$

for all $n > 0$, where μ is defined by substitution as follows:

$$\mu = \frac{1}{1 + \mathbb{I}\{\overline{\lambda_c} F > 2\underline{\lambda_f}\underline{\lambda_c}\}(\overline{\lambda_c} F - 2\underline{\lambda_f}\underline{\lambda_c})\gamma_1} \in (0, 1].$$

Therefore we can write recursion (26) as

$$\mathbb{E}\left(||\theta_n^{\mathrm{im}} - \theta_\star||^2\right) \leq \frac{1}{1 + 2\gamma_n \mu \underline{\lambda_f}\underline{\lambda_c}} \mathbb{E}\left(||\theta_{n-1}^{\mathrm{im}} - \theta_\star||^2\right) + 4L_0^2 \overline{\lambda_c}^2 \gamma_n^2. \tag{28}$$

We can now apply Corollary 2.1 with $a_n = 4L_0^2 \overline{\lambda_c}^2 \gamma_n^2$ and $b_n = 2\gamma_n \mu \underline{\lambda_f}\underline{\lambda_c}$. □

Note. Assuming Lipschitz continuity of the gradient $\nabla\ell$ instead of function ℓ would not critically alter the main result of Theorem (2.1). In fact, assuming Lipschitz continuity with constant L of $\nabla\ell$ and boundedness of $\mathbb{E}\left(||\nabla \log f(Y_n; X_n, \theta_\star)||^2\right) \leq \sigma^2$, as it is typical in the literature, would simply add a term $\gamma_n^2 L^2 \mathbb{E}\left(||\theta_n^{\text{im}} - \theta_\star||^2\right) + \gamma_n^2\sigma^2$ in the right-hand side of Eq.(23). In this case the upper-bound is always satisfied for n such that $\gamma_n^2 L^2 > 1$, which also highlights a difference of implicit SGD with explicit SGD, as in explicit SGD the term $\gamma_n^2 L^2 ||\theta_{n-1}^{\text{sgd}} - \theta_\star||^2$ increases the upper bound and can make $||\theta_n^{\text{sgd}} - \theta_\star||^2$ diverge. For, $\gamma_n^2 L^2 < 1$, the discount factor for implicit SGD would be $(1 - \gamma_n^2 L^2)^{-1}(1 + 2\gamma_n \mu \underline{\lambda_f}\,\underline{\lambda_c})^{-1}$, which could then be bounded by a quantity $(1 + \gamma_n d)^{-1}$ for some constant d. This would lead to a solution that is similar to Theorem (2.1).

Asymptotic analysis. Here, we prove the main result on the asymptotic variance of implicit SGD. First, we introduce linear maps $\mathbb{L}_B\{\cdot\}$ defined as $\mathbb{L}_B\{X\} = \frac{1}{2}(BX + XB)$, where B is symmetric positive definite matrix and X is bounded. The identity map is denoted as $\mathbb{I}$ and it holds $\mathbb{I}\{X\} = X$, for all X. Also, $\mathbb{L}_0$ is the null operator for which $\mathbb{L}_0\{X\} = 0$, for all X. By the Lyapunov theorem (Lyapunov, 1992) the map $\mathbb{L}_B$ is one-to-one and thus the inverse operator $\mathbb{L}_B^{-1}\{\cdot\}$ is well-defined. Furthermore, we define the norm of a linear map as $||\mathbb{L}_B|| = \max_{||X||=1} ||\mathbb{L}_B\{X\}||$. For bounded inputs X, it holds $||\mathbb{L}_B|| = \mathrm{O}(||B||)$.

LEMMA 2.4. *Suppose that the sequence $\{\gamma_n\}$ satisfies Assumption 2.1(a). Consider the matrix recursions*

$$X_n = \mathbb{L}_{I-\gamma_n B_n}\{X_{n-1}\} + \gamma_n(C + D_n), \tag{29}$$

$$Y_n = \mathbb{L}_{I+\gamma_n B_n}^{-1}\{X_{n-1} + \gamma_n(C + D_n)\}, \tag{30}$$

such that

(a) All matrices X_n, Y_n, B_n, D_n and C are bounded,
(b) $B_n \to B$ is positive definite and $||B_n - B_{n-1}|| = \mathrm{O}(\gamma_n^2)$,
(c) C is a fixed matrix and $D_n \to 0$.

Then, both recursions approximate the matrix $\mathbb{L}_B^{-1}\{C\}$ i.e.,

$$||X_n B + BX_n - 2C|| \to 0 \text{ and } |Y_n B + BY_n - 2C|| \to 0. \tag{31}$$

If, in addition, B and C commute then $X_n \to B^{-1}C$ and $Y_n \to B^{-1}C$.

PROOF. We make the following definitions.

$$\Gamma_n = I - \gamma_n B_n, \tag{32}$$

$$P_i^n = \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \mathbb{L}_{\Gamma_i}, \tag{33}$$

where the symbol $\circ$ denotes successive application of the linear maps, and $P_i^n = \mathbb{I}$ if $n < i$, by definition. It follows,

$$||P_i^n|| = \mathrm{O}(\prod_{j=i}^{n} ||I - \gamma_i B_i||) \leq K_0 e^{-K_1 \sum_{j=i}^n \gamma_j}, \tag{34}$$

for suitable constants K_0, K_1 (see Polyak and Juditsky, 1992, Appendix, Part 3). Let $\Gamma(n) = K_1 \sum_{i=1}^n \gamma_i$. By Assumption 2.1(a), $\Gamma(n) \to \infty$ and thus $P_i^n \to \mathbb{L}_0$ as $n \to \infty$ and i is fixed. The matrix recursion in Lemma 2.4 can be rewritten as $X_n = \mathbb{L}_{\Gamma_n}\{X_{n-1}\} + \gamma_n C + \gamma_n D_n$. Solving the recursion yields

$$\begin{aligned} X_n =& \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \mathbb{L}_{\Gamma_1}\{X_0\} + \gamma_n C + \gamma_n D_n \\ &+ a_{n-1}\mathbb{L}_{\Gamma_n}\{C\} + a_{n-1}\mathbb{L}_{\Gamma_n}\{D_{n-1}\} \\ &+ \cdots + \\ &+ a_1 \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \mathbb{L}_{\Gamma_2}\{C\} + a_1 \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \mathbb{L}_{\Gamma_2}\{D_1\} \\ =& \quad P_1^n\{X_0\} + S_n\{C\} + \widetilde{D}_n, \end{aligned} \tag{35}$$

where we have defined the linear map $S_n = \sum_{i=1}^n \gamma_i P_{i+1}^n$ and the matrix $\widetilde{D}_n = \sum_{i=1}^n \gamma_i P_{i+1}^n\{D_i\}$. Since $P_1^n \to \mathbb{L}_0$, our goal is to prove that $S_n \to \mathbb{L}_B^{-1}$ and $\widetilde{D}_n \to 0$. By definition,

$$\sum_{i=1}^n \gamma_i P_{i+1}^n = \mathbb{L}_{B_n}^{-1} + \sum_{i=2}^n P_i^n (\mathbb{L}_{B_{i-1}}^{-1} - \mathbb{L}_{B_i}^{-1}) - P_1^n \mathbb{L}_{B_1}^{-1}. \tag{36}$$

To see this, first note that $\gamma_n I = (I - \Gamma_n) B_n^{-1}$ for every n, and thus

$$\gamma_n \mathbb{I} = \mathbb{L}_{I-\Gamma_n} \circ \mathbb{L}_{B_n}^{-1}. \tag{37}$$

Therefore, if we collect the coefficients of the terms $\mathbb{L}_{B_n}^{-1}$ in the right-hand side of (36), we get

$$\begin{aligned}
&\mathbb{L}_{B_n}^{-1}+\sum_{i=2}^{n} P_i^n(\mathbb{L}_{B_{i-1}}^{-1}-\mathbb{L}_{B_i}^{-1})-P_1^n\mathbb{L}_{B_1}^{-1}\\
&\quad= (P_2^n-P_1^n)\mathbb{L}_{B_1}^{-1}+(P_3^n-P_2^n)\mathbb{L}_{B_2}^{-1}+\cdots+(P_{n+1}^n-P_n^n)\mathbb{L}_{B_n}^{-1}\\
&\quad= P_2^n\circ\mathbb{L}_{I-\Gamma_1}\circ\mathbb{L}_{B_1}^{-1}+P_3^n\circ\mathbb{L}_{I-\Gamma_2}\circ\mathbb{L}_{B_2}^{-1}+\cdots+P_{n+1}^n\circ\mathbb{L}_{I-\Gamma_n}\circ\mathbb{L}_{B_n}^{-1}\\
&\quad= P_2^n(\gamma_1\mathbb{I})+P_3^n(\gamma_2\mathbb{I})+\cdots+P_{n+1}^n(\gamma_n\mathbb{I}) \qquad [\textit{by Eq. } (37)]\\
&\quad= \sum_{i=1}^{n}\gamma_i P_{i+1}^n,
\end{aligned}$$

where we used the identity $P_{i+1}^n - P_i^n = P_{i+1}^n\circ(\mathbb{I}-\mathbb{L}_{\Gamma_i}) = P_{i+1}^n\circ\mathbb{L}_{I-\Gamma_i}$. Furthermore, since B_i are bounded,

$$\begin{aligned}
||\mathbb{L}_{B_{i-1}}^{-1}-\mathbb{L}_{B_i}^{-1}|| &= |||\mathbb{L}_{B_i}^{-1}\circ(\mathbb{L}_{B_i}-\mathbb{L}_{B_{i-1}})\circ\mathbb{L}_{B_{i-1}}^{-1}|| = \mathrm{O}(||\mathbb{L}_{B_i}-\mathbb{L}_{B_{i-1}}||)\\
&= \mathrm{O}(||B_i-B_{i-1}||) = \mathrm{O}(\gamma_i^2). \qquad [\textit{By assumption of Lemma 2.4}]
\end{aligned}$$

In addition, $||\sum_{i=2}^n P_i^n\circ(\mathbb{L}_{B_{i-1}}^{-1}-\mathbb{L}_{B_i}^{-1})|| \le K_0 e^{-\Gamma(n)}\sum_{i=2}^n e^{\Gamma(i)}\mathrm{O}(\gamma_i^2)$. Since $\sum_i \mathrm{O}(\gamma_i^2)<\infty$ and $e^{\Gamma(i)}$ is positive, increasing and diverging, we can invoke Kronecker's lemma and obtain $\sum_{i=2}^n e^{\Gamma(i)}\mathrm{O}(\gamma_i^2)=o(e^{\Gamma(n)})$. Therefore

$$\sum_{i=2}^{n} P_i^n\circ(\mathbb{L}_{B_{i-1}}^{-1}-\mathbb{L}_{B_i}^{-1})\to\mathbb{L}_0, \tag{38}$$

and since $P_1^n\to\mathbb{L}_0$, we conclude from Equation (37) that

$$\lim_{n\to\infty}\sum_{i=1}^{n}\gamma_i P_{i+1}^n = \lim_{n\to\infty}\mathbb{L}_{B_n}^{-1} = \mathbb{L}_B^{-1}. \tag{39}$$

Thus, $S_n\to\mathbb{L}_B^{-1}$, as desired. For $\widetilde{D}_n$ we have

$$\begin{aligned}
\widetilde{D}_n = \sum_{i=1}^{n}\gamma_i P_{i+1}^n\{D_i\} =&\mathbb{L}_{B_n}^{-1}\{D_n\}+\sum_{i=2}^{n}P_i^n\circ(\mathbb{L}_{B_{i-1}}^{-1}\{D_{i-1}\}-\mathbb{L}_{B_i}^{-1}\{D_i\})\\
&+P_1^n\circ\mathbb{L}_{B_1}^{-1}\{D_1\}.
\end{aligned}$$

Since $||D_n||\to 0$ it follows that $||\mathbb{L}_{B_n}^{-1}\{D_n\}||\to 0$ and $||(\mathbb{L}_{B_{i-1}}^{-1}\{D_{i-1}\}-\mathbb{L}_{B_i}^{-1}\{D_i\})||=\mathrm{O}(\gamma_i^2)$. Recall that $P_1^n\to\mathbb{L}_0$, and thus $\widetilde{D}_n\to\mathbf{0}$. Finally, we substitute this result in Equation (37) to get $X_n\to\mathbb{L}_B^{-1}\{C\}$.

For the second recursion of the lemma,

$$Y_n = \mathbb{L}^{-1}_{I+\gamma_n B_n}\left\{Y_{n-1} + \gamma_n(C + D_n)\right\}, \tag{40}$$

the proof is similar. First, we make the following definitions.

$$\begin{aligned}\Gamma_n &= I + \gamma_n B_n,\\ Q_i^n &= \mathbb{L}^{-1}_{\Gamma_n} \circ \mathbb{L}^{-1}_{\Gamma_{n-1}} \circ \cdots \mathbb{L}^{-1}_{\Gamma_i}.\end{aligned}$$

As before, $Q_i^n \to \mathbb{L}_0$. Solving the recursion (40) yields

$$Y_n = \quad Q_1^n\{Y_0\} + S_n\{C\} + \widetilde{D}_n, \tag{41}$$

where we defined $S_n = \sum_{i=1}^n \gamma_i Q_i^n$ and $\widetilde{D}_n = \sum_{i=1}^n \gamma_i Q_i^n\{D_i\}$. The following identities can also be verified by the definition of the linear maps.

$$\mathbb{L}^{-1}_{B_n} \circ (\mathbb{I} - \mathbb{L}^{-1}_{\Gamma_n}) = \gamma_n \mathbb{L}^{-1}_{\Gamma_n}, \tag{42}$$

$$\mathbb{L}^{-1}_{B_n}\mathbb{L}^{-1}_{\Gamma_n} = \mathbb{L}^{-1}_{\Gamma_n}\mathbb{L}^{-1}_{B_n}. \tag{43}$$

It holds,

$$\begin{aligned}\mathbb{L}^{-1}_{B_n} + \sum_{i=1}^n Q_i^n \circ (\mathbb{L}^{-1}_{B_{i-1}} - \mathbb{L}^{-1}_{B_i}) =& \mathbb{L}^{-1}_{B_n} \circ (\mathbb{I} - \mathbb{L}^{-1}_{\Gamma_n}) + \mathbb{L}^{-1}_{\Gamma_n} \circ \mathbb{L}^{-1}_{B_{n-1}} \circ (\mathbb{I} - \mathbb{L}^{-1}_{\Gamma_n}) + \cdots\\ =& \gamma_n \mathbb{L}^{-1}_{\Gamma_n} + \gamma_{n-1}\mathbb{L}^{-1}_{\Gamma_n}\mathbb{L}^{-1}_{\Gamma_{n-1}} + \cdots = S_n,\end{aligned}$$

where the first line is obtained by Eq. (42) and the second line by Eq. (43). Thus, similar to the previously analyzed recursion, $S_n \to \mathbb{L}_B^{-1}$ and $\widetilde{D}_n \to 0$. Therefore, $Y_n \to \mathbb{L}_B^{-1}\{C\}$.

For both cases, if B, C commute then $\mathbb{L}_B^{-1}\{C\} = X$ such that $BX + XB = 2C$. Setting $X = B^{-1}C$ is a solution since $BB^{-1}C + B^{-1}CB = C + B^{-1}BC = 2C$. By the Lyapunov theorem, this solution is unique. □

Corollary 2.2. *Consider the matrix recursions*

$$X_n = \mathbb{L}_{I-\gamma_n B_n}\left\{X_{n-1}\right\} + \gamma_n^2(C + D_n), \tag{44}$$

$$Y_n = \mathbb{L}^{-1}_{I+\gamma_n B_n}\left\{Y_{n-1} + \gamma_n^2(C + D_n)\right\}, \tag{45}$$

where B_n, B, C, D_n satisfy the assumptions of Lemma 2.4. Moreover, suppose $\gamma_n = \gamma_1 n^{-1}$. If the matrix $B - I/\gamma_1$ is positive definite, then

$$(1/\gamma_n)X_n \to \mathbb{L}^{-1}_{B-I/\gamma_1}\{C\} \text{ and } (1/\gamma_n)Y_n \to \mathbb{L}^{-1}_{B-I/\gamma_1}\{C\} \text{ i.e.,}$$

both matrices $(1/\gamma_n)X_n$ and $(1/\gamma_n)Y_n$ approximate the matrix $\mathbb{L}^{-1}_{B-I/\gamma_1}\{C\}$. If, in addition, B and C commute then $(1/\gamma_n)X_n \to (B - I/\gamma_1)^{-1}C$ and $(1/\gamma_n)Y_n \to (B - I/\gamma_1)^{-1}C$.

Proof. Both $X_n, Y_n \to 0$ by direct application of Lemma (2.4). Let $\tilde{X}_n = (1/\gamma_n)X_n$. First, divide (44) by γ_n to obtain

$$\tilde{X}_n = \mathbb{L}_{I-\gamma_n B_n}\left\{\tilde{X}_{n-1}\right\}\frac{\gamma_{n-1}}{\gamma_n} + \gamma_n(C + D_n). \tag{46}$$

By Assumption 2.1(a), $\gamma_{n-1}/\gamma_n = 1 + \gamma_n/\gamma_1 + \mathrm{O}(\gamma_n^2)$. Then,

$$\mathbb{L}_{I-\gamma_n B_n}\left\{\tilde{X}_{n-1}\right\}\frac{\gamma_{n-1}}{\gamma_n} = \mathbb{L}_{I-\gamma_n B_n}\left\{\tilde{X}_{n-1}\right\} + \gamma_n\tilde{X}_{n-1} + \mathrm{O}(\gamma_n^2). \tag{47}$$

Therefore, we can rewrite Eq. (46) as

$$\tilde{X}_n = \mathbb{L}_{I-\gamma_n\Gamma_n}\left\{\tilde{X}_{n-1}\right\} + \gamma_n(C + D_n), \tag{48}$$

where $\Gamma_n = B_n - I/\gamma_1 + \mathrm{O}(\gamma_n)$. In the limit $\Gamma_n \to B - I/\gamma_1 > 0$. Furthermore, $||\Gamma_{i-1} - \Gamma_i|| = \mathrm{O}(\gamma_i^2)$ by assumptions of Corollary 2.2. Thus, we can apply Lemma 2.4 to conclude that $\tilde{X}_n = (1/\gamma_n)X_n \to \mathbb{L}^{-1}_{B-I/\gamma_1}\{C\}$. The proof for Y_n follows the same reasoning since $(I + \gamma_n B_n)^{-1}(\gamma_{n-1}/\gamma_n) = (I + \gamma_n\Gamma_n)^{-1}$, where $\Gamma_n = B_n - I/\gamma_1 + \mathrm{O}(\gamma_n)$. □

Theorem 2.2. *Consider the SGD procedures defined by Eq. (1) and by Eq. (4) in the main paper, and suppose that Assumptions 2.1(a),(c),(d),(e) hold, where $\gamma = 1$, and that $2\gamma_1 C\mathcal{I}(\theta_\star) \succ I$. The asymptotic variance of the explicit SGD estimator satisfies*

$$n\mathrm{Var}\left(\theta_n^{\mathrm{sgd}}\right) \to \gamma_1^2\left(2\gamma_1 C\mathcal{I}(\theta_\star) - I\right)^{-1} C\mathcal{I}(\theta_\star)C.$$

The asymptotic variance of the implicit SGD estimator satisfies

$$n\mathrm{Var}\left(\theta_n^{\mathrm{im}}\right) \to \gamma_1^2\left(2\gamma_1 C\mathcal{I}(\theta_\star) - I\right)^{-1} C\mathcal{I}(\theta_\star)C.$$

Proof. We begin with the implicit SGD procedure. For notational convenience we make the following definitions: $V_n = \mathrm{Var}\left(\theta_n^{\mathrm{im}}\right)$, $S_n(\theta) = \nabla \log f(Y_n; X_n, \theta)$. Denote $\mathbb{E}\left(S_n(\theta)\right) = h(\theta)$. Let J_h denote the Jacobian of function h, then, under typical regularity conditions of Assumptions 2.1(d) and by Theorem 2.1:

$$\begin{aligned}
&\mathbb{E}\left(S_n(\theta_\star) \mid X_n\right) = 0\\
&\mathrm{Var}\left(S_n(\theta_\star)\right) = \mathbb{E}\left(\mathrm{Var}\left(S_n(\theta_\star) \mid X_n\right)\right) = \mathcal{I}(\theta_\star)\\
&J_h(\theta) = -\mathcal{I}(\theta), \qquad [\textit{under regularity conditions}]\\
&h(\theta_n^{\mathrm{im}}) = -\mathcal{I}(\theta_\star)(\theta_n^{\mathrm{im}} - \theta_\star) + \mathrm{O}(\gamma_n) \qquad [\textit{by Theorem 2.1}],
\end{aligned}$$

$$||\mathrm{Var}\left(S_n(\theta) - S_n(\theta_\star)\right)|| \le \mathbb{E}\left(||S_n(\theta) - S_n(\theta_\star)||^2\right) \le L_0^2\mathbb{E}\left(||\theta - \theta_\star||^2\right). \tag{49}$$

We can now rewrite the definition of implicit SGD as follows,

$$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n C_n S_n(\theta_n^{\text{im}}) = \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n C_n S_n(\theta_{n-1}^{\text{im}}), \tag{50}$$

where λ_n is defined in Theorem 3.1 and $\lambda_n = 1 - \mathrm{O}(\gamma_n)$ by Eq. (18). Then, taking variances on both sides of Eq. (50) yields

(51)

$$V_n = V_{n-1} + \gamma_n^2 C_n \mathrm{Var}\left(S_n(\theta_n^{\text{im}}\right) C_n^\intercal + \gamma_n \mathrm{Cov}\left(\theta_{n-1}^{\text{im}}, S_n(\theta_n^{\text{im}}\right) C_n^\intercal + \gamma_n C_n \mathrm{Cov}\left(S_n(\theta_n^{\text{im}}), \theta_{n-1}^{\text{im}}\right).$$

We can simplify all variance/covariance terms in Eq. (51) as follows.

$$\begin{aligned}
C_n \mathrm{Var}\left(S_n(\theta_n^{\text{im}})\right) C_n^\intercal &= C_n \mathrm{Var}\left(S_n(\theta_\star) + [S_n(\theta_n^{\text{im}}) - S_n(\theta_\star)]\right) C_n^\intercal \\
&= C\mathcal{I}(\theta_\star)C^\intercal + \mathrm{o}(1), \quad \textit{[by Eqs. (49), Theorem (2.1), and Assumption 2.1(e)]} \\
\mathrm{Cov}\left(\theta_{n-1}^{\text{im}}, S_n(\theta_n^{\text{im}})\right) &= \mathrm{Cov}\left(\theta_{n-1}^{\text{im}}, S_n(\theta_{n-1}^{\text{im}})\right) + \mathrm{Cov}\left(\theta_{n-1}^{\text{im}}, (\lambda_n - 1) S_n(\theta_{n-1}^{\text{im}})\right) \\
&= \mathrm{Cov}\left(\theta_{n-1}^{\text{im}}, h(\theta_{n-1}^{\text{im}})\right) + \mathrm{O}(\gamma_n) \\
&= V_{n-1}\mathcal{I}(\theta_\star) + \mathrm{O}(\gamma_n). \quad \textit{[by Eq. (49), Theorem (2.1), Eq. (18)]}.
\end{aligned}$$

Similarly, $\mathrm{Cov}\left(h(\theta_n^{\text{im}}), \theta_{n-1}^{\text{im}}\right) = V_{n-1}\mathcal{I}(\theta_\star) + \mathrm{O}(\gamma_n)$. We can now rewrite Eq. (51) as

$$V_n = \mathbb{L}_{I - \gamma_n B_n}\left\{V_{n-1}\right\} + \gamma_n^2 [C\mathcal{I}(\theta_\star)C^\intercal + \mathrm{o}(1)], \tag{52}$$

where $B_n = 2C_n\mathcal{I}(\theta_\star)$ and $B_n \to 2C\mathcal{I}(\theta_\star)$. Corollary 2.2 on recursion (52) yields the following closed-form, since B and C commute and C is symmetric:

$$(1/n)V_n \to \gamma_1^2 \left(2\gamma_1 C\mathcal{I}(\theta_\star) - I\right)^{-1} C\mathcal{I}(\theta_\star)C.$$

The regularity conditions (49) and the convergence rates of Theorem 2.1 that are crucial for this proof also hold for the explicit procedure. □

THEOREM 2.3. *Consider the SGD procedure defined in Eq. (1) in the main paper, and suppose Assumptions 2.1(a),(c),(d), and (e) hold, where* $\gamma \in [0.5, 1)$. *Then, the iterate* $\overline{\theta_n^{\text{im}}}$ *converges to* $\theta_\star$ *in probability and is asymptotically efficient, i.e.,*

$$n\mathrm{Var}\left(\overline{\theta_n^{\text{im}}}\right) \to \mathcal{I}(\theta_\star)^{-1}.$$

PROOF. By Theorem 2.1 and Assumptions 2.1 (c), (d), we have

$$\nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) = \nabla \log f(Y_n; X_n, \theta_\star) - \mathcal{I}(\theta_\star)(\theta_n^{\text{im}} - \theta_\star) + \mathrm{O}(\gamma_n). \tag{53}$$

Define, for convenience $\varepsilon_n = \nabla \log f(Y_n; X_n, \theta_\star)$, $F = \mathcal{I}(\theta_\star)$. Then, the first-order implicit SGD iteration becomes

$$\theta_n^{\text{im}} - \theta_\star = (I + \gamma_n F)^{-1}(\theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n \varepsilon_n + \mathrm{O}(\gamma_n^2)). \tag{54}$$

We make the following definitions.

$$\begin{aligned} e_i &= \gamma_i (I + \gamma_i F)^{-1}(\varepsilon_i + \mathrm{O}(\gamma_i^2)), \\ B_i^j &= \prod_{k=j}^{i} (I + \gamma_k F)^{-1}, \\ D_j^n &= \prod_{k=n-1}^{i} B_{j+1}^k = I + B_{j+1}^{j+1} + B_{j+1}^{j+2} + \ldots + B_{j+1}^{n-1}. \end{aligned} \tag{55}$$

Then, we can solve the recursion for $\overline{\theta_n^{\text{im}}} - \theta_\star$ to obtain

$$\overline{\theta_n^{\text{im}}} - \theta_\star = (1/n) D_0^n (\overline{\theta_n^{\text{im}}} - \theta_\star) + (1/n) \sum_{i}^{n-1} D_i^n e_i. \tag{56}$$

Our proof is now split into proving the following two lemmas.

Lemma 2.5. *Under Assumption 2.1(a) $D_0^n = o(n)$.*

Proof. Matrix F is positive definite by Assumption 2.1(d). Thus, if λ is some eigenvalue of F then the corresponding eigenvalue of D_0^n is $1 + \frac{1}{1+\gamma_1\lambda} + \frac{1}{1+\gamma_1\lambda}\frac{1}{1+\gamma_2\lambda} + \cdots \le \sum_{i=0}^{n} \exp(-K\lambda \sum_{k=1}^{i} \gamma_k)$, where the last inequality is obtained by Lemma 2.1. Because $\sum \gamma_i \to \infty$, the summands are $o(1)$, and thus D_0^n is $o(n)$. □

Lemma 2.6. *Suppose Assumption 2.1(a) and Eq. (53) hold. Then,*

$$\gamma_i D_i^n (I + \gamma_i F)^{-1} = \Omega_i^n + F^{-1}, \tag{57}$$

such that $\sum_{i=0}^{n-1} \Omega_i^n = o(n)$.

Proof. Our goal will be to compare the eigenvalues of $\gamma_i D_i^n$ and F. Any matrix D_i^n shares the same eigenvectors with F because F is positive definite, and thus a relationship on eigenvalues will automatically establish a relationship on the matrices. For convenience, define $q_i^j = \prod_{k=i}^{j} (1 + \gamma_k \lambda)^{-1}$ for $\lambda > 0$; by convention, $q_{i-1}^i = 1$. Also let $s_i^j = \sum_{k=i}^{j} \gamma_k$ be the function of partial sums. By Lemma 2.1 $q_i^j = \mathrm{O}(\exp(-K\lambda s_i^j))$, for some $K > 0$. For an eigenvalue $\lambda > 0$ of F the corresponding eigenvalue, say λ', of matrix $\gamma_i D_i^n (I + \gamma_i F)^{-1}$ is equal to

$$\lambda' = \frac{\gamma_i}{1 + \gamma_i \lambda} (q_{i+1}^i + q_{i+1}^{i+1} + \ldots + q_{i+1}^{n-1}). \tag{58}$$

Thus,

$$\lambda'(1 + \gamma_i \lambda) = \sum_{k=i}^{n-1} \gamma_i q_{i+1}^k. \tag{59}$$

Our goal will be to derive the relationship between λ and λ'. By definition

$$\begin{aligned}
&\gamma_{i+1} \lambda q_{i+1}^{i+1} + q_{i+1}^{i+1} = 1 \\
&\gamma_{i+2} \lambda q_{i+1}^{i+2} + q_{i+1}^{i+2} = q_{i+1}^{i+1} \\
&\ldots\ldots \\
&\gamma_{n-2} \lambda q_{i+1}^{n-2} + q_{i+1}^{n-2} = q_{i+1}^{n-3} \\
&\gamma_{n-1} \lambda q_{i+1}^{n-1} + q_{i+1}^{n-1} = q_{i+1}^{n-2}.
\end{aligned} \tag{60}$$

By summing over the terms we obtain:

$$\lambda \sum_{k=i+1}^{n-1} \gamma_k q_{i+1}^k + q_{i+1}^{n-1} = 1. \tag{61}$$

If we combine with (58) we obtain

$$\lambda \sum_{k=i}^{n-1} \gamma_i q_{i+1}^k + \lambda \sum_{k=i}^{n-1} (\gamma_k - \gamma_i) q_{i+1}^k + q_{i+1}^{n-1} = 1 + \gamma_i \lambda \quad \text{or} \tag{62}$$

$$(1 + \gamma_i \lambda) \lambda \lambda' + \lambda \sum_{k=i}^{n-1} (\gamma_k - \gamma_i) q_{i+1}^k + q_{i+1}^{n-1} = 1 + \gamma_i \lambda. \tag{63}$$

We now focus on the second term. By telescoping the series we obtain

$$\begin{aligned}
\lambda \sum_{k=i}^{n-1} (\gamma_k - \gamma_i) q_{i+1}^k &= \lambda \sum_{k=i}^{n-1} \left[\sum_{j=i}^{k} (\gamma_{j+1} - \gamma_j) \right] q_{i+1}^k = \lambda \sum_{k=i}^{n-1} \left[\sum_{j=i}^{k} \gamma_j o(\gamma_j) \right] q_{i+1}^k \\
&\leq \lambda o(\gamma_i) \sum_{k=i}^{n-1} s_i^k q_{i+1}^k \triangleq q_i^n.
\end{aligned} \tag{64}$$

In Eq. (64) we used $(\gamma_{j+1} - \gamma_j)/\gamma_j = \mathrm{O}(n^{-1-\gamma})/n^{-\gamma} = \mathrm{O}(n^{-1}) = o(\gamma_j)$, by Assumption 2.1(a). Our goal is now to show $\sum_{i=0}^{n-1} q_i^n = \mathrm{o}(n)$. Since $q_{i+1}^k = \mathrm{O}(\exp(-K\lambda s_{i+1}^k))$ by (Polyak and Juditsky, 1992, p845, see A6 and A7) we obtain that $q_i^n \to 0$ for fixed i as $n \to \infty$. Therefore we can rewrite Eq. (62) as

$$\lambda'\lambda + q_i^n + \mathrm{O}(q_{i+1}^n) = 1, \tag{65}$$

where $\sum_{i=0}^{n} q_{i+1}^n = \mathrm{o}(n)$ and $\sum_{i=0}^{n-1} q_i^n = \mathrm{o}(n)$. □

Our proof is now complete. By Eq. (56) and Lemmas 2.5 and 2.6 we have

$$\overline{\theta_n^{\mathrm{im}}} - \theta_\star = F^{-1}\sum_{i=1}^{n} \varepsilon_i + (1/n)\mathrm{o}(n).$$

Because $\mathrm{Var}\,(\varepsilon_i) = \mathcal{I}(\theta_\star)$, we finally obtain

$$n\mathrm{Var}\left(\overline{\theta_n^{\mathrm{im}}} - \theta_\star\right) = \mathcal{I}(\theta_\star)^{-1}.$$

□

Theorem 2.4. *Suppose that Assumptions 2.1(a),(c),(d),(e),(f) hold. Then, the iterate θ_n^{im} of implicit SGD, defined by Eq. (4) in the main paper, is asymptotically normal, such that*

$$n^{\gamma/2}(\theta_n^{\mathrm{im}} - \theta_\star) \to \mathcal{N}_p(0, \Sigma),$$

where $\Sigma = \gamma_1^2\left(2\gamma_1 C\mathcal{I}(\theta_\star) - I\right)^{-1} C\mathcal{I}(\theta_\star)C$.

Proof. Let $S_n(\theta) = \nabla \log f(Y_n; X_n, \theta)$ as in the proof of Theorem (2.2). The conditions for Fabian's theorem—see Fabian (1968, Theorem 1)—hold also for the implicit procedure. The goal is to show that

$$\theta_n^{\mathrm{im}} - \theta_\star = (I - \gamma_n A_n)(\theta_{n-1}^{\mathrm{im}} - \theta_\star) + \gamma_n \xi_n(\theta_\star) + \mathrm{O}(\gamma_n^2), \tag{66}$$

where $A_n \to A \succeq 0$, and $\xi_n(\theta) = S_n(\theta) - h(\theta)$, and $h(\theta) = \mathbb{E}\left(S_n(\theta)\right)$; note, $\xi_n(\theta_\star) = S_n(\theta_\star)$. Indeed, by a Taylor expansion on $S_n(\theta_n^{\mathrm{im}})$ and considering that $\theta_n^{\mathrm{im}} = \theta_{n-1}^{\mathrm{im}} + \gamma_n S_n(\theta_n^{\mathrm{im}})$, by definition, we have

$$(I + \gamma_n \hat{\mathcal{I}}_n(\theta_\star))(\theta_n^{\mathrm{im}} - \theta_\star) = \theta_{n-1}^{\mathrm{im}} - \theta_\star + \gamma_n S_n(\theta_\star), \tag{67}$$

where $\hat{\mathcal{I}}_n(\theta_\star) = -\nabla^2 S_n(\theta_\star)$; note, $\mathbb{E}\left(\hat{\mathcal{I}}_n(\theta_\star)\right) = \mathcal{I}(\theta_\star)$. Because $(I+\gamma_n\hat{\mathcal{I}}_n(\theta_\star))^{-1} = I - \gamma_n\hat{\mathcal{I}}_n(\theta_\star) + \mathrm{O}(\gamma_n^2)$, we can rewrite Eq. (67) as

$$\theta_n^{\mathrm{im}} - \theta_\star = (I - \gamma_n\hat{\mathcal{I}}_n(\theta_\star))(\theta_{n-1}^{\mathrm{im}} - \theta_\star) + \gamma_n S_n(\theta_\star) + \mathrm{O}(\gamma_n^2). \tag{68}$$

We can now apply Fabian's Theorem to derive asymptotic normality of θ_n^{im}. The variance matrix of the asymptotic normal distribution is derived in Theorem 2.4 under weaker conditions. □

Stability. Here, we prove Lemma (2.1) in the main paper.

Lemma 2.1. *Let $\overline{\lambda_f} = \max\mathrm{eig}(\mathcal{I}(\theta_\star))$, and suppose $\gamma_n = \gamma_1/n$ and $\gamma_1\overline{\lambda_f} > 1$. Then, the maximum eigenvalue of P_1^n satisfies*

$$\max_{n>0}\max\mathrm{eig}(P_1^n) = \Theta(2^{\gamma_1\overline{\lambda_f}}/\sqrt{\gamma_1\overline{\lambda_f}}).$$

For the implicit method,

$$\max_{n>0}\max\mathrm{eig}(Q_1^n) = \mathrm{O}(1).$$

Proof. We will use the following intermediate result:

$$\max_{n>0}|\prod_{i=1}^{n}(1-b/i)| \approx \begin{cases} 1-b & \text{if } 0<b<1 \\ \frac{2^b}{\sqrt{2\pi b}} & \text{if } b>1 \end{cases}$$

The first case is obvious. For the second case, $b > 1$, assume without loss of generality that b is an even integer. Then the maximum is given by

$$(b-1)(b/2-1)(b/3-1)\cdots(2-1) = \frac{1}{2}\binom{b}{b/2} = \Theta(2^b/\sqrt{2\pi b}), \tag{69}$$

where the last approximation follows from Stirling's formula. The stability result on the explicit SGD updates of Lemma 2.1 follows immediately by using the largest eigenvalue $\overline{\lambda_f}$ of $\mathcal{I}(\theta_\star)$. For the implicit SGD updates, we note that the eigenvalues of $(I+\gamma_n\mathcal{I}(\theta_\star))^{-1}$ are less than one, for any $\gamma_n > 0$ and any Fisher matrix. □

Applications.

Theorem 3.1. *Suppose Assumption 2.1(b) holds, then the gradient of the log-likelihood is a scaled version of covariate X, i.e., for every $\theta \in \mathbb{R}^p$ there is a scalar $\lambda \in \mathbb{R}$ such that*

$$\nabla \log f(Y; X, \theta) = \lambda X.$$

Thus, the gradient in the implicit update in Eq. (4) (in the main paper) is a scaled version of the gradient calculated at the previous iterate, i.e.,

$$\nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) = \lambda_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}), \tag{70}$$

where the scalar λ_n satisfies

$$\lambda_n \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) = \ell' \left(X_n^\intercal \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) X_n^\intercal C_n X_n; Y_n \right). \tag{71}$$

Proof. From the chain rule $\nabla \log f(Y_n; X_n, \theta) = \ell'(X_n^\intercal \theta; Y_n) X_n$, and thus $\nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) = \ell'(X_n^\intercal \theta_n^{\text{im}}; Y_n) X_n$ and $\nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) = \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) X_n$, and thus the two gradients are colinear. Therefore there exists a scalar λ_n such that

$$\begin{aligned} \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) &= \lambda_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) \quad \text{or} \\ \ell'(X_n^\intercal \theta_n^{\text{im}}; Y_n) X_n &= \lambda_n \ell'(X_n^\intercal \theta_{n-1}^{\text{im}}; Y_n) X_n. \end{aligned} \tag{72}$$

We also have,

$$\begin{aligned} \theta_n^{\text{im}} &= \theta_{n-1}^{\text{im}} + \gamma_n C_n \log f(Y_n; X_n, \theta_n^{\text{im}}) \quad \textit{[by definition of implicit SGD in Eq. (4)]} \\ &= \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n C_n \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}). \quad \textit{[by Eq. (72)]} \end{aligned} \tag{73}$$

Substituting the expression for θ_n^{im} in Eq.(73) into Eq. (72) we obtain the desired result of the Theorem in Eq. (70).

We now prove the last claim of the theorem regarding the search bounds for λ_n. For notational convenience, define $a = X_n^\intercal \theta_{n-1}^{\text{im}}$, $g(x) = \ell'(x; Y_n)$, and $c = X_n^\intercal C_n X_n$, where $c > 0$ because C_n are positive definite. Also let $x_\star = \gamma_n \lambda_n g(a)$, then the fixed-point equation (71) can be written as

$$x_\star = \gamma_n g(a + x_\star c). \tag{74}$$

where g is decreasing by Assumption (b). If $g(a) = 0$ then $x_\star = 0$. If $g(a) > 0$ then $x_\star > 0$ and $\gamma_n g(a + xc) < \gamma_n g(a)$ for all $x > 0$, since $g(a + xc)$ is decreasing; taking $x = x_\star$ yields $\gamma_n g(a) > \gamma_n g(a + x_\star c) = x_\star$, by the fixed-point equation (74). Thus, $0 < x_\star < \gamma_n g(a)$. Similarly, if $g(a) < 0$ then $x_\star < 0$ and $\gamma_n g(a + xc) > \gamma_n g(a)$ for all $x < 0$, since $g(a + xc)$ is decreasing; taking $x = x_\star$ yields $\gamma_n g(a) < \gamma_n g(a + x_\star c) = x_\star$, by the fixed-point equation. Thus, $\gamma_n g(a) < x_\star < 0$. In both cases $0 < \lambda_n < 1$. A visual proof is given Figure 1.

□

3. Additional experiments.

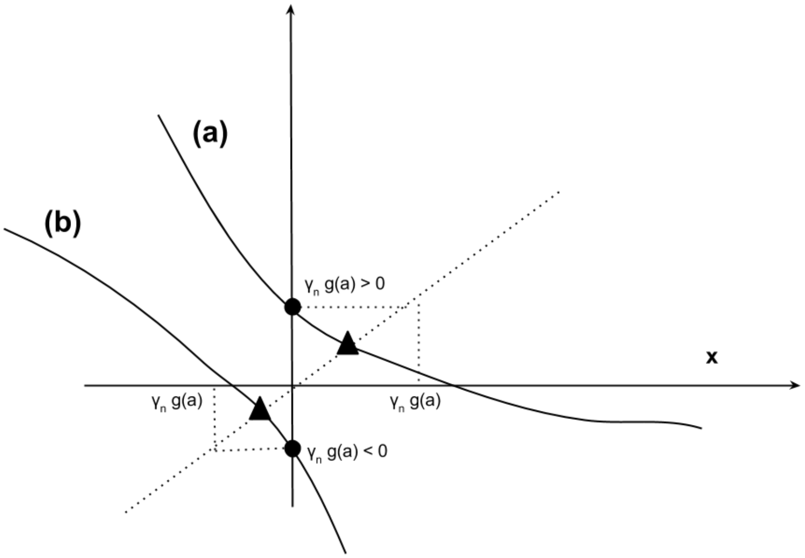

Fig 1. *(Search bounds for solution of Eq.* (74)*)* **Case** $g(a) > 0$**:** *Corresponds to Curve (a) defined as* $\gamma_n g(a + xc), c > 0$*. The solution* $x_\star$ *of fixed point equation* (74) *(corresponding to right triangle) is between 0 and* $\gamma_n g(a)$ *since Curve (a) is decreasing.* **Case** $g(a) < 0$**:** *Corresponds to Curve (b) also defined as* $\gamma_n g(a + xc)$*. The solution* $x_\star$ *of fixed point equation* (74) *(left triangle) is between* $\gamma_n g(a)$ *and 0 since Curve (b) is also decreasing.*

Normality experiments with implicit SGD. In Figure 2 we plot the experimental results of Section 4.1.2 for $p = 50$ (parameter dimension). We see that explicit SGD becomes even more unstable in more dimensions as expected. In contrast, implicit SGD remains stable and validates the theoretical normal distribution for small learning rates. In larger learning rates we observe a divergence from the asymptotic chi-squared distribution (e.g., $\gamma_1 = 6$) because when the learning rate parameter is large there is more noise in the stochastic approximations, and thus more iterations are required for convergence. In this experiment we fixed the number of iterations for each value of the learning rate, but subsequent experiments verified that implicit SGD reaches the theoretical chi-squared distribution if the number of iterations is increased. Finally, in Figure 3 we make a similar plot for a logistic regression model. In this case the learning rates need to be larger because with the same distribution of covariates for X_n, the Fisher information is smaller than in the linear normal model. In summary, in almost all experiments explicit SGD was unstable and could not converge whereas implicit SGD was stable and followed the theoretical chi-squared distribution.

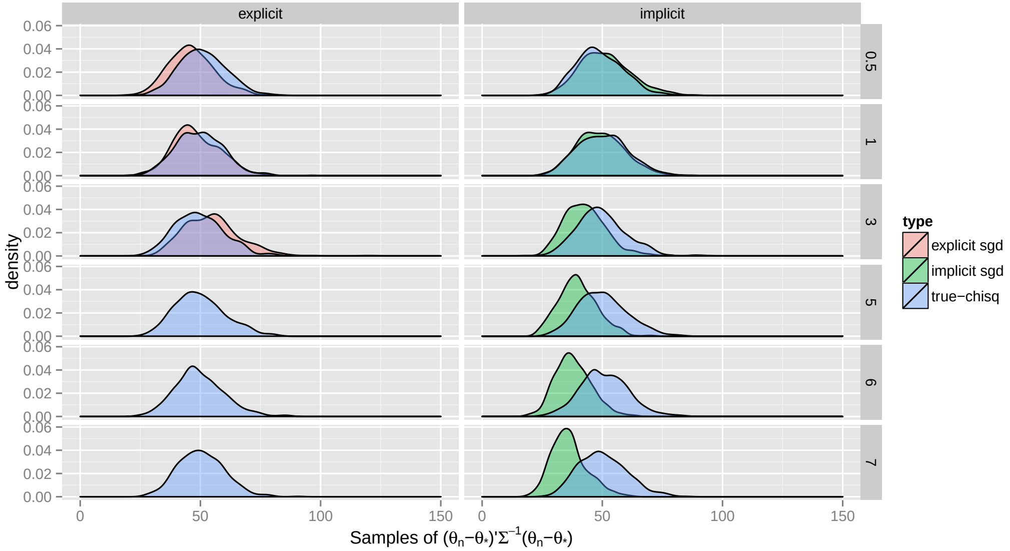

FIG 2. *Simulation with normal model for $p = 50$ parameters. Implicit SGD is stable and follows the nominal chi-squared distribution well, regardless of the particular learning rate. Explicit SGD becomes unstable at larger γ_1 and its distribution does not follow the theoretical distribution chi-squared distribution well. In particular, the distribution of $N(\theta_N^{\mathrm{sgd}} - \theta_\star)^\intercal \Sigma^{-1} (\theta_N^{\mathrm{sgd}} - \theta_\star)$ quickly becomes unstable for larger values of the learning rate parameter, and eventually diverges when $\gamma_1 > 3$.*

Poisson regression. Here, we illustrate our method on a bivariate Poisson model which is simple enough to derive the variance formula analytically. This example was first presented by Toulis et al. (2014). We assume binary features such that, for any iteration n, X_n is either $(0,0)^\intercal$, $(1,0)^\intercal$ or $(0,1)^\intercal$ with probabilities 0.6, 0.2 and 0.2 respectively. We set $\theta_\star = (\theta_1, \theta_2)^\intercal$ for some θ_1, θ_2, and assume $Y_n \sim \mathrm{Poisson}(\exp(X_n^\intercal \theta_\star))$, where the transfer function h is the exponential, i.e., $h(x) = \exp(x)$. It follows,

$$\mathcal{I}(\theta_\star) = \mathbb{E}\left(h'(X_n^\intercal \theta_\star) X_n X_n^\intercal\right) = 0.2 \begin{pmatrix} e^{\theta_1} & 0 \\ 0 & e^{\theta_2} \end{pmatrix}.$$

We set $\gamma_n = 10/3n$ and $C_n = I$. Setting $\theta_1 = \log 2$ and $\theta_2 = \log 4$, the asymptotic variance Σ in Theorem (2.2) is equal to

$$\Sigma = \frac{2}{3} \begin{pmatrix} \frac{e^{\theta_1}}{(4/3)e^{\theta_1} - 1} & 0 \\ 0 & \frac{e^{\theta_2}}{(4/3)e^{\theta_2} - 1} \end{pmatrix} = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.62 \end{pmatrix}. \tag{75}$$

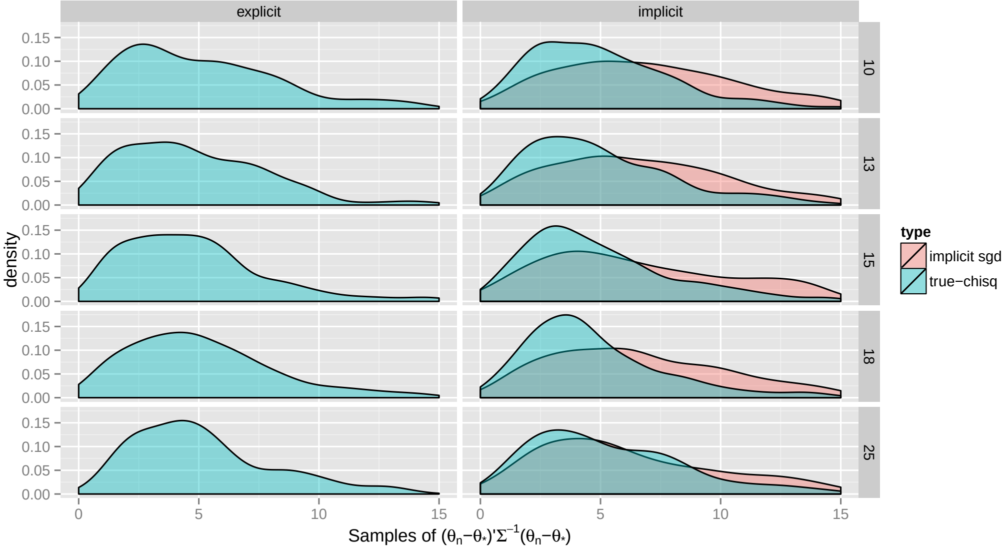

Fig 3. *Simulation with logistic regression model for* $p = 5$*. Learning rates are larger than in the linear normal model to ensure the asymptotic covariance matrix of Theorem (2.2) is positive definite. Implicit SGD is stable and follows the nominal chi-squared distribution regardless of the learning rate. Explicit SGD is unstable at virtually all replications of this experiment.*

Next, we obtain 100 independent samples of θ_n^{sgd} and θ_n^{im} for $n = 20000$ iterations of procedures in Eq. (4) and in Eq. (4), and compute their empirical variances. We observe that the implicit estimates are particularly stable and have an empirical variance satisfying

$$(1/\gamma_n)\widehat{\text{Var}}(\theta_n^{\text{im}}) = \begin{pmatrix} 0.86 & -0.06 \\ -0.06 & 0.64 \end{pmatrix},$$

and that is close to the theoretical value in Eq. (75). In contrast, the standard SGD estimates are unstable and their L_2 distance to the true values $\theta_\star$ are orders of magnitude larger than the implicit ones (see Table 1 for sample quantiles). By Lemma 2.1 in the main paper, such deviations are expected for standard SGD because the largest eigenvalue of $\mathcal{I}(\theta_\star)$ is $\lambda_{(2)} = 0.8$ satisfying $\gamma_1\lambda_{(2)} = 8/3 > 1$. Note that it is fairly straightforward to stabilize the standard SGD procedure in this problem, for example by modifying the learning rate sequence to $\gamma_n = \min\{0.15, 10/3n\}$. In general, when the optimization problem is well-understood, it is easy to determine the learning rate schedule that avoids out-of-bound explicit updates. In practice, however, we

Table 1

Quantiles of $||\theta_n^{\mathrm{sgd}} - \theta_\star||$ and $||\theta_n^{\mathrm{im}} - \theta_\star||$. Values larger than `1e3` *are marked "*".*

	QUANTILES					
METHOD	25%	50%	75%	85%	95%	100%
SGD	0.01	1.3	435.8	*	*	*
IMPLICIT	0.00	0.01	0.02	0.02	0.03	0.04

are working with problems that are not so well-understood and determining the correct learning rate parameters may take substantial effort. The implicit method eliminates this overhead.

Experiments with `glmnet`*.* In this section, we transform the outcomes in the original experiment Y through the logistic transformation and then fit a logistic regression model. The results are shown in Table 2, which replicates and expands on Table 2 of Friedman et al. (2010). The implicit SGD method maintains a stable running time over different correlations and scales sub-linearly in the model size p. In contrast, `glmnet` is affected by the model size p and covariate correlation, and remains 2x-10x slower across experiments. We note that the implicit SGD method is slower in the logistic regression example compared to the normal case (Table 3 in main paper). This is because the implicit equation of Algorithm 1 (in the main paper) needs to be solved numerically, whereas a closed-form solution is available in the normal case.

Experiments with machine learning algorithms. In this section we perform additional experiments with related methods from the machine learning literature. We focus on averaged implicit SGD defined in Eq. (14) of the main paper, which was shown to be optimal under suitable conditions, because most machine learning methods are also designed to achieve optimality in the context of maximum-likelihood (or maximum a-posteriori) computation with a finite data set. In summary, our experiments include the following procedures:

- Explicit SGD procedure in Eq. (1) of the main paper.
- Implicit SGD procedure in Eq. (4) of the main paper.
- Averaged explicit SGD: Averaged stochastic gradient descent with explicit updates of the iterates (Xu, 2011; Shamir and Zhang, 2012; Bach and Moulines, 2013). This is equivalent to the procedure in Eq.(14) of the main paper, where the implicit update is replaced by an explicit one, $\theta_n^{\mathrm{sgd}} = \theta_{n-1}^{\mathrm{sgd}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\mathrm{sgd}})$.

TABLE 2

Experiments comparing implicit SGD with `glmnet`. *Covariates X are sampled as normal, with cross-correlation ρ, and the outcomes are sampled as $\boldsymbol{y} \sim \mathrm{Binom}(\boldsymbol{p})$, $\mathrm{logit}(\boldsymbol{p}) = \mathcal{N}(X\theta_\star, \sigma^2 I)$. Running times (in secs) are reported for different values of ρ averaged over 10 repetitions.*

METHOD	METRIC	CORRELATION (ρ)			
		0	0.2	0.6	0.9
		$N = 1000, p = 10$			
GLMNET	TIME(SECS)	0.02	0.02	0.026	0.051
	MSE	0.256	0.257	0.292	0.358
SGD	TIME(SECS)	0.058	0.058	0.059	0.062
	MSE	0.214	0.215	0.237	0.27
		$N = 5000, p = 50$			
GLMNET		0.182	0.193	0.279	0.579
		0.131	0.139	0.152	0.196
SGD		0.289	0.289	0.296	0.31
		0.109	0.108	0.116	0.14
		$N = 100000, p = 200$			
GLMNET		8.129	8.524	9.921	22.042
		0.06	0.061	0.07	0.099
SGD		5.455	5.458	5.437	5.481
		0.045	0.046	0.048	0.058

- Prox-SVRG: A proximal version of the stochastic gradient descent with progressive variance reduction (SVRG) method (Xiao and Zhang, 2014).
- Prox-SAG: A proximal version of the stochastic average gradient (SAG) method (Schmidt et al., 2013). While its theory has not been formally established, Prox-SAG has shown similar convergence properties to Prox-SVRG.[1]
- Adagrad (Duchi et al., 2011) as defined in Eq. (12). We note that AdaGrad and similar adaptive methods effectively approximate the natural gradient by using a larger-dimensional learning rate. It has the added advantage of being less sensitive than first-order methods

[1] We note that the linear convergence rates for Prox-SVRG and Prox-SAG refer to convergence to the empirical minimizer (e.g., MLE), and not to ground truth $\theta_\star$.

Table 3

Summary of data sets and the L_2 regularization parameter λ used

	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}

to tuning of hyperparameters.

We test the performance of averaged implicit SGD on standard benchmarks of large-scale linear classification with real data sets against the aforementioned methods. Some of these test comparisons were recently published by Toulis et al. (2016). Our datasets are summarized in Table 3. The COVTYPE data (Blackard, 1998) consists of forest cover types in which the task is to classify class 2 among 7 forest cover types. DELTA is synthetic data offered in the PASCAL Large Scale Challenge (Sonnenburg et al., 2008) and we apply the default processing offered by the challenge organizers. The task in RCV1 is to classify documents belonging to class CCAT in the text dataset (Lewis et al., 2004), where we apply the standard preprocessing provided by Bottou (2012). In the MNIST data set (Le Cun et al., 1998) of images of handwritten digits, the task is to classify digit 9 against all others.

For averaged implicit SGD and averaged explicit SGD, we use the learning rate $\gamma_n = \eta_0(1 + \lambda\eta_0 n)^{-3/4}$ prescribed in Xu (2011), where the constant η_0 is determined using a small subset of the data. Hyperparameters for other methods are set based on a computationally intensive grid search over the entire hyperparameter space: for Prox-SVRG, this includes the step size η in the proximal update and the inner iteration count m, and for Prox-SAG, the same step size η.

The results are shown in Figure 4. We see that averaged implicit SGD achieves comparable performance with the tuned proximal methods Prox-SVRG and Prox-SAG, as well as AdaGrad. All methods have a comparable convergence rate and take roughly a single pass in order to converge. AdaGrad exhibits a larger variance in its estimate than the proximal methods, which can be explained from our theoretical results in Section 2.2.1. We also note that as averaged implicit SGD achieves comparable results to the other proximal methods, it also requires no tuning while Prox-SVRG and Prox-SAG do require careful tuning of their hyparameters. This was confirmed from separate sensitivity analyses (not reported in this paper), which indicated that aisgd is robust to fine-tuning of hyperparameters in the learning

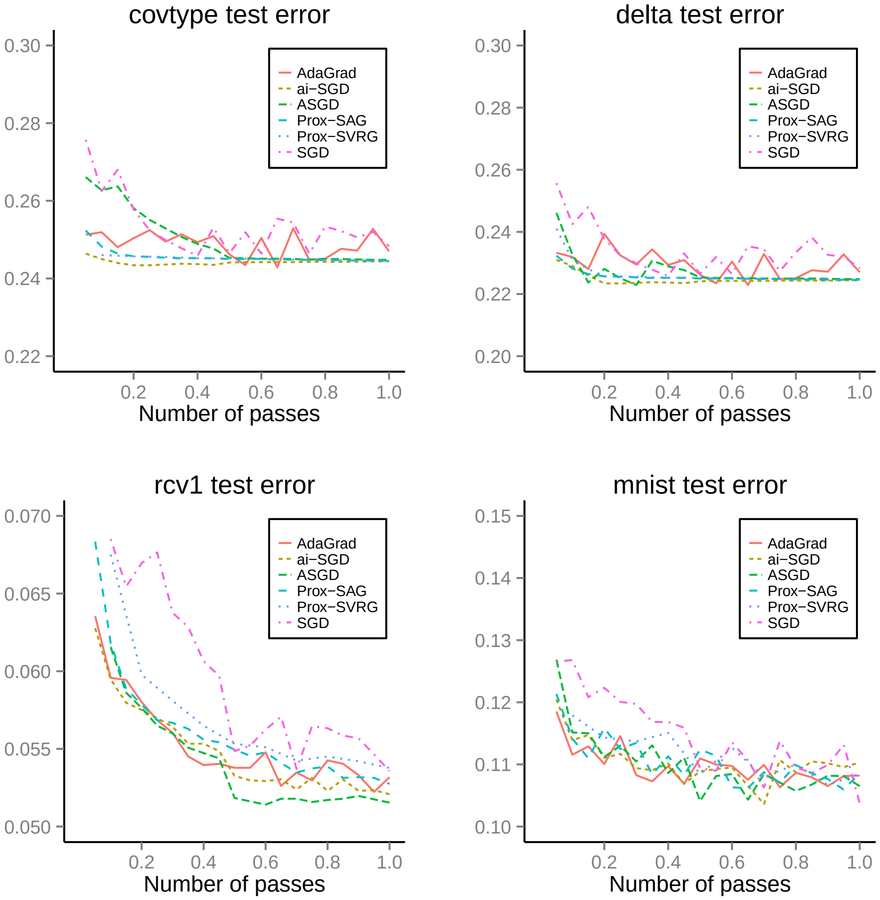

FIG 4. *Large scale linear classification with log loss on four data sets. Each plot indicates the test error of various stochastic gradient methods over a single pass of the data.*

rate, whereas small perturbations of hyperparameters in averaged explicit SGD (the learning rate), Prox-SVRG (proximal step size η and iteration m), and Prox-SAG (proximal step size η), can lead to arbitrarily bad error rates.

Averaged explicit SGD. In this experiment we validate the theory of statistical efficiency and stability of averaged implicit SGD. To do so, we follow a simple normal linear regression example from Bach and Moulines (2013). We set $N = 1e6$ as the number of observations, and $p = 20$ be the number of covariates. We also set $\theta_\star = (0, 0, \ldots, 0)^\intercal \in \mathbb{R}^{20}$ as the true parameter value. The random variables X_n are distributed i.i.d. as

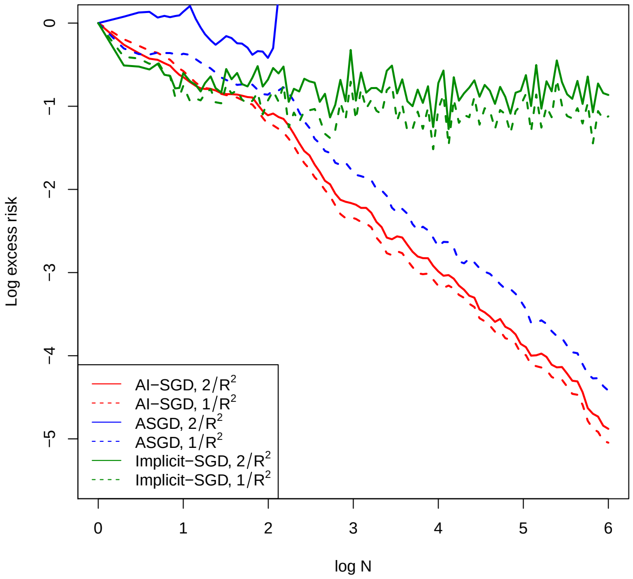

Fig 5. *Loss of averaged implicit SGD, averaged explicit SGD, and plain implicit SGD in Eq. (4) ($C_n = I$), on simulated multivariate normal data with $N = 1\text{e}6$ observations $p = 20$ features. The plot shows that averaged implicit SGD is stable regardless of the specification of the learning rate γ and without sacrificing performance. In contrast, explicit averaged SGD is very sensitive to misspecification of the learning rate.*

$X_n \sim \mathcal{N}_p(0, H)$, where H is a randomly generated symmetric matrix with eigenvalues $1/k$, for $k = 1, \ldots, p$. The outcome Y_n is sampled from a normal distribution as $Y_n \mid X_n \sim \mathcal{N}(X_n^\intercal \theta_*, 1)$, for $n = 1, \ldots, N$. We choose a constant learning rate $\gamma_n \equiv \gamma$ according to the average radius of the data $R^2 = \text{trace}(H)$, and for both averaged explicit and implicit SGD we collect iterates θ_n for $n = 1, \ldots, N$, and keep the average $\bar{\theta}_n$. In Figure 5, we plot $(\theta_n - \theta_\star)^\intercal H(\theta_n - \theta_\star)$ for each iteration for a maximum of N iterations, i.e., a full pass over the data, in log-log space.

Figure 5 shows that averaged implicit SGD performs on par with averaged explicit SGD for the rates at which averaged explicit SGD is known to be optimal. Thus, averaged implicit SGD is also optimal. However, the benefit of the implicit procedure in averaged implicit SGD becomes clear as the learning rate deviates; notably, averaged implicit SGD remains stable for learning rates that are above the theoretical threshold, i.e., $\gamma > 1/R^2$, whereas averaged explicit SGD diverges in the case of $\gamma = 2/R^2$. This stable behavior is also exhibited in implicit SGD, but it converges at a slower rate than averaged implicit SGD, and thus cannot effectively combine stability with statistical efficiency. We note that stability of averaged implicit SGD is also observed in the same experiments using decaying learning rates.

References.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

Jock Blackard. *Comparison of neural networks and discriminant analysis in predicting forest cover types.* PhD thesis, Department of Forest Sciences, Colorado State University, 1998.

Leon Bottou. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, volume 1, pages 421–436. 2012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 999999:2121–2159, 2011.

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

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Yann Le Cun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of IEEE*, 86(11):2278–2324, 1998.

Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

David Lewis, Yiming Yang, Tony Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

Aleksandr Mikhailovich Lyapunov. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992.

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

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. Technical report, HAL 00860051, 2013.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *arXiv preprint arXiv:1212.1824*, 2012.

Soeren Sonnenburg, Vojtech Franc, Elad Yom-Tov, and Michele Sebag. Pascal large scale learning challenge, 2008.

P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics*, 2016. In press.

Panos Toulis, Edoardo M. Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. *JMLR, W&CP, Volume 32 (ICML)*, pages 667–675, 2014.

Panos Toulis, Dustin Tran, and Edoardo M. Airoldi. Towards stability and optimality in stochastic gradient descent. *JMLR, W&CP, Volume 51 (AISTATS)*, 51, 2016.

Lin Xiao and Tony Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24:2057–2075, 2014.

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