

ASYMPTOTIC AND FINITE-SAMPLE PROPERTIES OF ESTIMATORS BASED ON STOCHASTIC GRADIENTS¹

BY PANOS TOULIS AND EDOARDO M. AIROLDI

University of Chicago and Harvard University

Stochastic gradient descent procedures have gained popularity for parameter estimation from large data sets. However, their statistical properties are not well understood, in theory. And in practice, avoiding numerical instability requires careful tuning of key parameters. Here, we introduce implicit stochastic gradient descent procedures, which involve parameter updates that are implicitly defined. Intuitively, implicit updates shrink standard stochastic gradient descent updates. The amount of shrinkage depends on the observed Fisher information matrix, which does not need to be explicitly computed; thus, implicit procedures increase stability without increasing the computational burden. Our theoretical analysis provides the first full characterization of the asymptotic behavior of both standard and implicit stochastic gradient descent-based estimators, including finite-sample error bounds. Importantly, analytical expressions for the variances of these stochastic gradient-based estimators reveal their exact loss of efficiency. We also develop new algorithms to compute implicit stochastic gradient descent-based estimators for generalized linear models, Cox proportional hazards, M-estimators, in practice, and perform extensive experiments. Our results suggest that implicit stochastic gradient descent procedures are poised to become a workhorse for approximate inference from large data sets.

1. Introduction. Parameter estimation by optimization of an objective function is a fundamental idea in statistics and machine learning [Fisher (1922), Hastie, Tibshirani and Friedman (2009), Lehmann and Casella (1998)]. However, classical procedures, such as Fisher scoring, the EM algorithm or iteratively reweighted least squares [Dempster, Laird and Rubin (1977), Fisher (1925), Green (1984)], do not scale to modern data sets with millions of data points and hundreds or thousands of parameters [National Research Council (2013)].

In particular, suppose we want to estimate the true parameter $\theta_\star \in \mathbb{R}^p$ of a distribution f from N i.i.d. data points (X_i, Y_i) such that conditional on covariate $X_i \in \mathbb{R}^p$ outcome $Y_i \in \mathbb{R}^d$ is distributed according to $f(Y_i; X_i, \theta_\star)$. Such estimation problems often reduce to optimization problems. For instance, the maximum

Received September 2015; revised August 2016.

¹Supported in part by NSF CAREER award IIS-1149662, ARO MURI award W911NF-11-1-0036, ONR YIP award N00014-14-1-0485, and the 2012 Google US/Canada Ph.D. Fellowship in Statistics. *MSC2010 subject classifications.* 62L20, 62F10, 62L12, 62F12, 62F35.

Key words and phrases. Stochastic approximation, implicit updates, asymptotic variance, generalized linear models, Cox proportional hazards, M-estimation, maximum likelihood, exponential family, statistical efficiency, numerical stability.

likelihood estimator (MLE) is obtained by solving $\theta_N^{\text{mle}} = \arg \max_{\theta} \sum_{i=1}^N \log f(Y_i; X_i, \theta)$. Classical optimization procedures, such as Newton–Raphson or Fisher scoring, have a runtime complexity that ranges between $O(Np^{1+\varepsilon})$ and $O(Np^{2+\varepsilon})$, in the best case and worst case, respectively [Lange (2010)]. Quasi-Newton (QN) procedures are the only viable alternative in practice because they have $O(Np^2)$ complexity per iteration, or $O(Np^{1+\varepsilon})$ in certain favorable cases [Hennig and Kiefel (2013)]. However, estimation from large data sets requires an even better runtime complexity that is roughly $O(Np^{1-\varepsilon})$, that is, linear in data size N but sublinear in parameter dimension p . The first requirement on N is generally unavoidable because all data points carry information from the i.i.d. assumption. Sublinearity in p is therefore critical.

Such requirements have recently generated interest in stochastic optimization procedures, especially those only relying on first-order information, that is, gradients. Perhaps the most widely popular procedure in this family is *stochastic gradient descent* (SGD), defined for $n = 1, 2, \dots$, as

$$(1) \quad \theta_n^{\text{sgd}} = \theta_{n-1}^{\text{sgd}} + \gamma_n C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{sgd}}),$$

where $\gamma_n > 0$ is the learning rate sequence, typically defined as $\gamma_n = \gamma_1 n^{-\gamma}$, $\gamma_1 > 0$ is the learning rate parameter, $\gamma \in (0.5, 1]$, and C_n are $p \times p$ positive-definite matrices, also known as condition matrices.

Stochastic optimization procedures of this kind are special cases of stochastic approximation [Robbins and Monro (1951)], where the estimation problem is not formulated as an optimization problem but more generally as a characteristic equation. Early research considered a streaming data setting—akin to a superpopulation setting—where the characteristic equation is

$$(2) \quad \mathbb{E}(\nabla \log f(Y; X, \theta_{\star}) \mid X) = 0,$$

with the expectation being over the true conditional distribution of outcome Y given covariate X . More recent research, largely in computer science and optimization, considers a finite N setting with characteristic equation

$$(3) \quad \mathbb{E}(\nabla \log f(Y; X, \theta_N^{\text{mle}})) = 0,$$

where the expectation is over the empirical distribution of (X, Y) in the finite data set.² In both settings, SGD of equation (1) is well defined: in the finite population setting of equation (3) the data point (X_n, Y_n) is a random sample with replacement from the finite data set; in the infinite population setting of equation (2) the data point (X_n, Y_n) is simply the n th data point in the stream.

From a computational perspective, SGD in equation (1) is appealing because it avoids expensive matrix inversions, as in Newton–Raphson, and the log-likelihood

²If regularization is used, then equation (3) could approximate the maximum a posteriori estimate (MAP) instead of the MLE.

is evaluated at a single data point (X_n, Y_n) and not on the entire data set. From a theoretical perspective, SGD in equation (1) converges, under suitable conditions, to $\theta_\infty^{\text{sgd}}$ where $\mathbb{E}(\log f(Y; X, \theta_\infty^{\text{sgd}}) \mid X) = 0$ [Benveniste, Métivier and Priouret (1990), Ljung, Pflug and Walk (1992), Borkar (2008)]. This condition can satisfy both equation (2) and equation (3), implying that SGD can be used on both finite and infinite population settings. For the rest of this paper, we assume an infinite population setting, as it is the most natural setting for stochastic approximations. The main difference between the data setting studied in the computer science and optimization literature and the infinite population setting we consider here is that we do not condition on the observed ordering of data points, but we condition on a random ordering instead. Moreover, most of the theoretical results presented in this paper for the infinite population setting can be applied to the finite population setting, where instead of estimating θ_\star we estimate, say, the MLE, or the MAP estimate if there is regularization.

In this paper, we introduce *implicit* stochastic gradient descent procedures—*implicit SGD* for short—defined as

$$(4) \quad \theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n C_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}),$$

where γ_n, C_n are defined as in standard SGD equation (1). Furthermore, we provide a theoretical analysis of estimators based on stochastic gradients, for both implicit and standard procedures. To distinguish the two procedures, we will refer to standard SGD in equation (1) as SGD with *explicit updates*, or *explicit SGD* for short, because the next iterate θ_n^{sgd} can be immediately computed given $\theta_{n-1}^{\text{sgd}}$ and the data point (X_n, Y_n) . In contrast, the update in equation (4) is *implicit* because the next iterate θ_n^{im} appears on both sides of the equation, where the iterate was typed in boldface to emphasize the fact.

1.1. Illustrative example. Here, we motivate the main results of this paper on the comparison between implicit and explicit SGD. Let $\theta_\star \in \mathbb{R}$ be the true parameter of a normal model with i.i.d. observations $Y_i \mid X_i \sim \mathcal{N}(X_i \theta_\star, \sigma^2)$, where the variance σ^2 is assumed known for simplicity. The log-likelihood is $\log f(Y_i; X_i, \theta) = -\frac{1}{2\sigma^2}(Y_i - X_i \theta)^2$, and the score function (i.e., gradient of log-likelihood) is given by $\nabla \log f(Y_i; X_i, \theta) = \frac{1}{\sigma^2}(Y_i - X_i \theta)X_i$. Let X_i be distributed according to some unknown distribution with bounded second moment. Assume $\gamma_n = \gamma_1/n$, for some $\gamma_1 > 0$ as the learning rate. Then the explicit SGD procedure in equation (1) is

$$(5) \quad \begin{aligned} \theta_n^{\text{sgd}} &= \theta_{n-1}^{\text{sgd}} + \gamma_n (Y_n - \theta_{n-1}^{\text{sgd}} X_n) X_n \\ &= (1 - \gamma_n X_n^2) \theta_{n-1}^{\text{sgd}} + \gamma_n Y_n X_n. \end{aligned}$$

Procedure (5) is the least mean squares filter (LMS) in signal processing, also known as the Widrow–Hoff algorithm [Widrow and Hoff (1960)]. The implicit

SGD procedure can be derived in closed form in this problem using update in equation (4) as

$$\begin{aligned}
 (6) \quad \theta_n^{\text{im}} &= \theta_{n-1}^{\text{im}} + \gamma_n(Y_n - X_n\theta_n^{\text{im}})X_n \\
 &= \frac{1}{1 + \gamma_n X_n^2} \theta_{n-1}^{\text{im}} + \frac{\gamma_n}{1 + \gamma_n X_n^2} Y_n X_n.
 \end{aligned}$$

The procedure defined by equation (6) is known as the normalized least mean squares filter (NLMS) in signal processing [Nagumo and Noda (1967)].

From equation (5), we see that it is crucial for explicit SGD to have a well-specified learning rate parameter γ_1 . For instance, if $\gamma_1 X_1^2 \gg 1$ then θ_n^{sgd} will diverge to a value at the order of $2^{\gamma_1}/\sqrt{\gamma_1}$, before converging to the true value θ_* (see Section 2.5, Lemma 2.1). In contrast, implicit SGD is more stable to misspecification of the learning rate parameter γ_1 . For example, a very large γ_1 will not cause divergence as in explicit SGD, but it will simply put more weight on the n th observation $Y_n X_n$ than the previous iterate θ_{n-1}^{im} . Assuming for simplicity $\theta_{n-1}^{\text{sgd}} = \theta_{n-1}^{\text{im}} = 0$, it also holds $\theta_n^{\text{im}} = \frac{1}{1 + \gamma_n X_n^2} \theta_n^{\text{sgd}}$, showing that implicit SGD iterates are shrunk versions of explicit ones (see also Section 5).

Let $v^2 = \mathbb{E}(X^2)$, then by Theorem 2.2 the asymptotic variance of θ_n^{im} (and of θ_n^{sgd}) satisfies $n \text{Var}(\theta_n^{\text{im}}) \rightarrow \gamma_1^2 \sigma^2 v^2 / (2\gamma_1 v^2 - 1)$ if $2\gamma_1 v^2 - 1 > 0$. Since $\gamma_1^2 / (2\gamma_1 v^2 - 1) \geq 1/v^2$, it is best to set $\gamma_1 = 1/v^2$. In this case $n \text{Var}(\theta_n^{\text{im}}) \rightarrow \sigma^2/v^2$. Implicit SGD can thus be optimal by setting $\gamma_n = (\sum_{i=1}^n X_i^2)^{-1}$ in which case θ_n^{sgd} is exactly the OLS estimator, and θ_n^{im} is an approximate but more stable version of the OLS estimator. Thus, the implicit SGD estimator θ_n^{im} in equation (6) inherits the efficiency properties of θ_n^{sgd} , with the added benefit of being stable over a wide range of learning rates. Overall, implicit SGD is a superior form of SGD.

1.2. Related work. Historically, the duo of explicit–implicit updates originate from the numerical methods introduced by Euler (ca. 1770) for solving ordinary differential equations [Hoffman and Frankel (2001)]. The explicit SGD procedure was first proposed by Sakrison (1965) as a recursive statistical estimation method and it is theoretically based on the stochastic approximation method of Robbins and Monro (1951). Statistical estimation with explicit SGD is a straightforward generalization of Sakrison’s method and has recently attracted attention in the machine learning community as a fast learning method for large-scale problems [Zhang (2004), Bottou (2010), Toulis and Airolidi (2015a)]. Applications of explicit SGD procedures in massive data problems can be found in many diverse areas such as large-scale machine learning [Zhang (2004)], online EM algorithm [Cappé and Moulines (2009)], image analysis and deep learning [Dean et al. (2012)] and MCMC sampling [Welling and Teh (2011)].

The implicit SGD procedure is less known and not well understood. In optimization, implicit methods have recently attracted attention under the guise of proximal

methods, such as mirror-descent [Nemirovsky and Yudin (1983)]. In fact, the implicit SGD update in equation (4) can be expressed as a proximal update:

$$(7) \quad \theta_n^{\text{im}} = \arg \max_{\theta} \left\{ -\frac{1}{2\gamma_n} \|\theta - \theta_{n-1}^{\text{im}}\|^2 + \log f(Y_n; X_n, \theta) \right\}.$$

From a Bayesian perspective, θ_n^{im} is the posterior mode of a model with the standard multivariate normal $\mathcal{N}(\theta_{n-1}^{\text{im}}, \gamma_n I)$ as the prior, and $\log f(Y_n; X_n, \theta)$ as the log-likelihood of θ for observation (X_n, Y_n) . Arguably, the normalized least mean squares (NLMS) filter [Nagumo and Noda (1967)], introduced in equation (6), was the first statistical model that used an implicit update as in equation (4), and was shown to be consistent and robust under excessive input noise [Slock (1993)]. From an optimization perspective, the update in equation (7) corresponds to a stochastic version of the proximal point algorithm by Rockafellar (1976), which has been generalized through the idea of splitting algorithms [Lions and Mercier (1979), Beck and Teboulle (2009), Singer and Duchi (2009)]; see also the comprehensive review of proximal methods in optimization by Parikh and Boyd (2013). Additional intuition of implicit methods has been provided by Krzysztow et al. (2007) and Nemirovski et al. (2008), who have argued that proximal methods can fit better in the geometry of the parameter space. Bertsekas (2011) derived the convergence rate of an implicit procedure similar to equation (4) on a fixed data set, and compared the rates between procedures that randomly sampled data (X_n, Y_n) or simply cycled through them. Toulis, Airolidi and Rennie (2014) derived the asymptotic variance of θ_n^{im} as estimator of θ_* in the family of generalized linear models, and provided an algorithm to efficiently compute the implicit update of equation (4) in such models and in the simplified setting where $C_n = I$.

1.3. Contributions. Prior work on procedures similar to implicit SGD has considered mostly an optimization setting, in which the focus is on speed of convergence [for example, Bertsekas (2011)]. Instead, we focus on statistical efficiency, that is, the sampling variability of the estimator implied by implicit and explicit SGD procedures—the relevant analysis and the results of Theorem 2.1 and Theorem 2.2 are novel. Furthermore, our procedure, which we generalized in Toulis and Airolidi (2015b), is different than typical stochastic proximal gradient procedures [see, e.g., Duchi and Singer (2009), Rosasco, Villa and Vü (2014)]. In such procedures, the parameter updates are obtained by combining a stochastic explicit update and a deterministic implicit update. In implicit SGD, there is a single stochastic implicit update, which prevents numerical instability.

With regard to theoretical contributions, the asymptotic statistical efficiency of SGD procedures (both explicit and implicit) derived in Theorem 2.2 is a key contribution of our work. Our analysis is in fact general enough that allowed us to derive the asymptotic efficiency of other popular stochastic optimization procedures, notably of AdaGrad [Duchi, Hazan and Singer (2011)] in equation (13) of our paper.

The asymptotic normality of implicit SGD in Theorem 2.4 is new and enables a novel comparison of explicit SGD and implicit SGD in terms of the normality of their iterates, which is also a clear point of departure from the typical optimization literature. The results in Section 2.5 are also new, and formalize the advantages of implicit SGD over explicit SGD in terms of numerical stability.

With regard to practical contributions, Algorithm 1 and its variants presented in the paper are a significant extension of our earlier work beyond first-order GLMs [Toulis, Airolidi and Rennie (2014), Algorithm 1]. The key contribution here is that these new algorithms make implicit SGD as simple to implement as standard explicit SGD, whenever the fixed-point computation of the implicit update is feasible. We provide extensive applications in Section 3 and experiments in Section 4 of implicit SGD compared to explicit SGD. Importantly, we developed a concrete implementation of implicit SGD through the R package `sgd` [Tran, Toulis and Airolidi (2015)] available at <https://cran.r-project.org/web/packages/sgd/index.html> to compare implicit SGD with state-of-art procedures, including R's `glm()` function (Fisher scoring), `biglm` package, the elastic net [Friedman, Hastie and Tibshirani (2010), `glmnet`], AdaGrad [Duchi, Hazan and Singer (2011)], Prox-SVRG [Xiao and Zhang (2014)], and Prox-SAG [Schmidt, Le Roux and Bach (2013)].

2. Theory. The norm $\|\cdot\|$ denotes the L_2 norm. If a positive scalar sequence a_n is nonincreasing and $a_n \rightarrow 0$, we write $a_n \downarrow 0$. For two positive scalar sequences a_n, b_n , equation $b_n = O(a_n)$ denotes that b_n is bounded above by a_n , that is, there exists a fixed $c > 0$ such that $b_n \leq ca_n$, for all n . Furthermore, $b_n = o(a_n)$ denotes that $b_n/a_n \rightarrow 0$. Similarly, for a sequence of vectors (or matrices) X_n , we write $X_n = O(a_n)$ to denote $\|X_n\| = O(a_n)$, and write $X_n = o(a_n)$ to denote $\|X_n\| = o(a_n)$. For two positive definite matrices A, B , we write $A \prec B$ to express that $B - A$ is positive definite. The set of eigenvalues of a matrix A is denoted by $\text{eig}(A)$; thus, $A \succ 0$ if and only if $\lambda > 0$ for every $\lambda \in \text{eig}(A)$.

ASSUMPTION 2.1. The explicit SGD procedure in equation (1) and the implicit SGD procedure in equation (4) 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 $\log f(Y; X, \theta) \equiv \ell(X^\top \theta; Y)$, which depends on θ only through the natural parameter $X^\top \theta$.
- (c) Function ℓ is concave, twice differentiable almost surely w.r.t. natural parameter $X^\top \theta$ and Lipschitz with constant L_0 w.r.t. θ .
- (d) The observed Fisher information matrix $\hat{\mathcal{I}}_n(\theta) = -\nabla^2 \ell(X_n^\top \theta; Y_n)$ has nonvanishing trace, that is, there exists constant $b > 0$ such that $\text{trace}(\hat{\mathcal{I}}_n(\theta)) \geq b$

- almost surely, for all θ . The Fisher information matrix, $\mathcal{I}(\theta_\star) = \mathbb{E}(\hat{\mathcal{I}}_n(\theta_\star))$, has minimum eigenvalue $\lambda_f > 0$ and maximum eigenvalue $\bar{\lambda}_f < \infty$. Typical regularity conditions hold [Lehmann and Casella (1998), Theorem 5.1, page 463].
- (e) Every condition matrix C_n is a fixed positive-definite matrix, such that $C_n = C + O(\gamma_n)$, where $\|C\| = 1$, $C \succ 0$ and symmetric, and C commutes with $\mathcal{I}(\theta_\star)$. For every C_n , $\min \text{eig}(C_n) = \underline{\lambda}_c > 0$, and $\max \text{eig}(C_n) = \bar{\lambda}_c < \infty$.
- (f) Let $\Xi_n = \mathbb{E}(\nabla \log f(Y_n; X_n, \theta_\star) \nabla \log f(Y_n; X_n, \theta_\star)^\top \mid \mathcal{F}_{n-1})$, then $\|\Xi_n - \Xi\| = O(1)$ for all n , and $\|\Xi_n - \Xi\| \rightarrow 0$, for a symmetric positive-definite Ξ . Let $\sigma_{n,s}^2 = \mathbb{E}(\|\xi_n(\theta_\star)\|^2_{\geq s/\gamma_n} \mid \mathcal{F}_n)$, then for all $s > 0$, $\sum_{i=1}^n \sigma_{i,s}^2 = o(n)$ if $\gamma = 1$, and $\sigma_{n,s}^2 = o(1)$ otherwise.

REMARKS. Assumption 2.1(a) is typical in stochastic approximation as it implies that $\sum_i \gamma_i = \infty$ and $\sum_i \gamma_i^2 < \infty$, as posited by Robbins and Monro (1951). Assumption 2.1(b) narrows our focus to models for which the likelihood depends on parameters θ through the linear combination $X^\top \theta$. This family of models is large and includes generalized linear models, Cox proportional hazards models, and M-estimation. Furthermore, in Section 5 we discuss a significant relaxation of Assumption 2.1(b). Assumption 2.1(c) puts a Lipschitz condition on the log-likelihood but it is used only for deriving finite-sample error bounds in Theorem 2.1—it is possible that this condition can be relaxed. Assumption 2.1(d) is equivalent to assuming strong convexity for the negative log-likelihood, which is typical for proving convergence in probability. The assumption on the observed Fisher information is less standard and, intuitively, it posits that a minimum of statistical information is received from any data point, at least for certain model parameters. Making this assumption allows us to forgo boundedness assumptions on the errors of stochastic gradients that were originally used by Robbins and Monro (1951), and have since been standard. Finally, Assumption 2.1(f) posits the typical Lindeberg conditions that are necessary to invoke the central limit theorem and prove asymptotic normality; this assumption follows the conditions defined by Fabian (1968) for the normality of explicit SGD procedures.

2.1. *Finite-sample error bounds.* Here, we derive bounds for the errors $\mathbb{E}(\|\theta_n^{\text{im}} - \theta_\star\|^2)$ on a finite sample of fixed size n .

THEOREM 2.1. *Let $\delta_n = \mathbb{E}(\|\theta_n^{\text{im}} - \theta_\star\|^2)$. 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*

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

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

Not surprisingly, implicit SGD in equation (4) matches the asymptotic rate of explicit SGD in equation (1). In particular, the iterates θ_n^{im} have squared error with rate $O(n^{-\gamma})$, as seen in Theorem 2.1, which is identical to the rate of error for the explicit iterates θ_n^{sgd} [Benveniste, Métivier and Priouret (1990), Theorem 22, page 244]. One way to explain intuitively this similarity in convergence rates is to assume that both explicit and implicit SGD are at the same estimate θ_0 . Then, using definitions in equation (1) and in equation (4), a Taylor approximation of the gradient $\nabla \log f(Y_n; X_n, \theta_n^{\text{im}})$ yields

$$(8) \quad \Delta \theta_n^{\text{im}} \approx [I + \gamma_n \hat{\mathcal{I}}_n(\theta_0)]^{-1} \Delta \theta_n^{\text{sgd}},$$

where $\Delta \theta_n^{\text{im}} = \theta_n^{\text{im}} - \theta_0$ and $\Delta \theta_n^{\text{sgd}} = \theta_n^{\text{sgd}} - \theta_0$. Therefore, as $n \rightarrow \infty$, we have $\Delta \theta_n^{\text{im}} \approx \Delta \theta_n^{\text{sgd}}$, and the two procedures coincide.

Despite the similarity in convergence rates, the critical advantage of implicit SGD—more generally of implicit procedures—is their robustness to initial conditions and excess noise. This can be seen in Theorem 2.1 where the implicit procedure discounts the initial conditions $\mathbb{E}(\|\theta_0^{\text{im}} - \theta_\star\|^2)$ at an exponential rate through the term $\exp(-\log \kappa \cdot \phi_\gamma(n))$. Importantly, the discounting of initial conditions happens regardless of the specification of the learning rate. In fact, large values of γ_1 can lead to faster discounting, and thus possibly to faster convergence, however, at the expense of increased variance as implied by Theorem 2.2, which is presented in the following section. The implicit iterates are therefore *unconditionally stable*, that is, virtually any specification of the learning rate will lead to a stable discounting of the initial conditions.

In contrast, explicit SGD is known to be very sensitive to the learning rate, and can numerically diverge if the rate is misspecified. For example, Moulines and Bach (2011), Theorem 1, showed that there exists a term $\exp(L^2 \gamma_1^2 n^{1-2\gamma})$, where L is a Lipschitz constant for the gradient of the log-likelihood, amplifying the initial conditions $\mathbb{E}(\|\theta_0^{\text{sgd}} - \theta_\star\|^2)$ of explicit SGD, which can be catastrophic if the learning rate parameter γ_1 is misspecified.³ Thus, although implicit and explicit SGD have identical asymptotic performance, they are crucially different in their stability properties. This is investigated further in Section 2.5 and in the experiments of Section 4.

2.2. Asymptotic variance and optimal learning rates. In the previous section, we showed that $\theta_n^{\text{im}} \rightarrow \theta_\star$ in quadratic mean, that is, the implicit SGD iterates converge to the true model parameters θ_\star , similar to classical results for the explicit SGD iterates θ_n^{sgd} . Thus, θ_n^{im} and θ_n^{sgd} are consistent estimators of θ_\star . In the

³The Lipschitz conditions are different in the two works; however, this does not affect our conclusions. Our result remains effectively unchanged if we assume Lipschitz continuity of the gradient $\nabla \ell$ instead of the log-likelihood ℓ , similar to Moulines and Bach (2011); see comment after the proof of Theorem 2.1.

following theorem, we show that both SGD estimators have the same asymptotic variance.

THEOREM 2.2. *Consider SGD procedures in equation (1) and in equation (4), 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 in equation (1) satisfies*

$$n \operatorname{Var}(\theta_n^{\text{sgd}}) \rightarrow \gamma_1^2 (2\gamma_1 C\mathcal{I}(\theta_\star) - I)^{-1} C\mathcal{I}(\theta_\star) C.$$

The asymptotic variance of the implicit SGD estimator in equation (4) satisfies

$$n \operatorname{Var}(\theta_n^{\text{im}}) \rightarrow \gamma_1^2 (2\gamma_1 C\mathcal{I}(\theta_\star) - I)^{-1} C\mathcal{I}(\theta_\star) C.$$

REMARKS. Although the implicit SGD estimator θ_n^{im} is significantly more stable than the explicit estimator θ_n^{sgd} (Theorem 2.1), both estimators have the same asymptotic efficiency in the limit according to Theorem 2.2. This implies that implicit SGD is a superior form of SGD, and should be preferred when the calculation of implicit updates in equation (4) is computationally feasible. In Section 3, we show that this is possible in a large family of statistical models, and illustrate with several numerical experiments in Section 4.1.

Asymptotic variance results in stochastic approximation similar to Theorem 2.2 were first obtained by Chung (1954), Sacks (1958), and followed by Fabian (1968), Polyak and Tsykin (1979), and several other authors [see also Ljung, Pflug and Walk (1992), Parts I, II]. We contribute to this literature in two important ways. First, our asymptotic variance result includes implicit SGD, which is a stochastic approximation procedure with implicitly defined updates, whereas other works consider only explicit updates. Second, in our setting we estimate recursively the true parameters θ_\star of a statistical model, and thus we can exploit the typical regularity conditions of Assumption 2.1(d) to derive the asymptotic variance of θ_n^{im} (and θ_n^{sgd}) in a simplified closed-form. We illustrate the asymptotic variance results of Theorem 2.2 in Section 4.1.1.

2.2.1. Optimal learning rates. Crucially, the asymptotic variance formula of Theorem 2.2 depends on the limit of the sequence C_n used in the SGD procedures of equation (1) and equation (4). We distinguish two classes of procedures, one where $C_n = I$, known as *first-order procedures*, and a second class where C_n is not trivial, known as *second-order procedures*.

In first-order procedures, only gradients are used in the SGD procedures. Inevitably, no matter how we set the learning rate parameter γ_1 , first-order SGD procedures will lose statistical efficiency. We can immediately verify this by comparing the asymptotic variance in Theorem 2.2 with the asymptotic variance of the maximum likelihood estimator (MLE), denoted by θ_N^{mle} , on a data set with

N data points $\{(X_n, Y_n)\}$, $n = 1, 2, \dots, N$. Under the regularity conditions of Assumption 2.1(d), the MLE is the asymptotically optimal unbiased estimator and $N \text{Var}(\theta_N^{\text{mle}} - \theta_\star) \rightarrow \mathcal{I}(\theta_\star)^{-1}$. By Theorem 2.2 and convergence of implicit SGD, it holds $N \text{Var}(\theta_N^{\text{im}} - \theta_\star) \rightarrow \gamma_1^2(2\gamma_1\mathcal{I}(\theta_\star) - I)^{-1}\mathcal{I}(\theta_\star)$, which also holds for θ_N^{sgd} . For any $\gamma_1 > 0$, we have as an identity that

$$(9) \quad \gamma_1^2(2\gamma_1\mathcal{I}(\theta_\star) - I)^{-1}\mathcal{I}(\theta_\star) \geq \mathcal{I}(\theta_\star)^{-1}.$$

The proof is rather quick if we consider $\lambda_i \in \text{eig}(\mathcal{I}(\theta_\star))$ and note that $\gamma_1^2\lambda_i/(2\gamma_1\lambda_i - 1)$ is the corresponding eigenvalue of the left-hand matrix in inequality (9) and $1/\lambda_i$ is the eigenvalue of $\mathcal{I}(\theta_\star)^{-1}$, and that $(2\gamma_1\underline{\lambda}_f - 1) > 0$ implies that

$$\gamma_1^2\lambda_i/(2\gamma_1\lambda_i - 1) \geq 1/\lambda_i,$$

for every $\lambda_i \in \text{eig}(\mathcal{I}(\theta_\star))$. Therefore, both SGD estimators lose information and this loss can be quantified exactly by inequality (9). This inequality can also be used to find the optimal choice for γ_1 given an appropriate objective. As demonstrated in the experiments in Section 4, this often suffices to achieve estimates that are comparable with MLE in statistical efficiency but with substantial computational gains. One reasonable objective is to minimize the trace of the asymptotic variance matrix, that is, to set γ_1 equal to

$$(10) \quad \gamma_1^\star = \arg \min_{x > 1/(2\underline{\lambda}_f)} \sum_i x^2\lambda_i/(2x\lambda_i - 1).$$

Equation (10) is defined under the constraint $x > 1/(2\underline{\lambda}_f)$ because Theorem 2.2 requires $2\gamma_1\mathcal{I}(\theta_\star) - I$ to be positive definite.

Of course, the eigenvalues λ_i are unknown in practice and need to be estimated from the data. This problem has received significant attention recently and several methods exist [see El Karoui (2008), and references within]. We will use equation (10) extensively in our experiments (Section 4) in order to tune the SGD procedures. However, we note that in first-order SGD procedures, knowing the eigenvalues λ_i of $\mathcal{I}(\theta_\star)$ does not necessarily achieve statistical efficiency because of the spectral gap of $\mathcal{I}(\theta_\star)$, that is, the ratio between its maximum eigenvalue $\overline{\lambda}_f$ and minimum eigenvalue $\underline{\lambda}_f$; for instance, if $\underline{\lambda}_f = \overline{\lambda}_f$, then the choice of learning rate parameter according to equation (10) leads to statistically efficient first-order SGD procedures. However, this case is not typical in practice, especially in many dimensions.

In second-order procedures, we assume nontrivial condition matrices C_n . Such procedures are called second-order because they usually leverage curvature information from the Fisher information matrix (or the Hessian of the log-likelihood). They are also known as *adaptive* procedures because they adapt their hyperparameters, that is, learning rates γ_n or condition matrices C_n , according to observed

data. For instance, let $C_n \equiv \mathcal{I}(\theta_\star)^{-1}$ and $\gamma_1 = 1$. Plugging in $C_n = \mathcal{I}(\theta_\star)^{-1}$ in Theorem 2.2, the normalized asymptotic variance of the SGD estimators is

$$\gamma_1^2 (2\gamma_1 \mathcal{I}(\theta_\star)^{-1} \mathcal{I}(\theta_\star) - I)^{-1} \mathcal{I}(\theta_\star)^{-1} \mathcal{I}(\theta_\star) \mathcal{I}(\theta_\star)^{-1} = \mathcal{I}(\theta_\star)^{-1},$$

which is the theoretically optimal asymptotic variance of the MLE, that is, the Cramér–Rao lower bound.

Therefore, to achieve asymptotic efficiency, second-order procedures need to estimate the Fisher information matrix at θ_\star . Because θ_\star is unknown one can simply use $C_n = \mathcal{I}(\theta_n^{\text{im}})^{-1}$ [or $C_n = \mathcal{I}(\theta_{n-1}^{\text{sgd}})^{-1}$] as an iterative estimate of $\mathcal{I}(\theta_\star)$, and the same optimality result holds. This approach in second-order explicit SGD was first studied by [Sakrisson \(1965\)](#), and later by [Nevelson and Khasminskii \(1973\)](#), Chapter 8, Theorem 5.4. It was later extended by [Fabian \(1978\)](#) and several other authors. Notably, [Amari \(1998\)](#) refers to the direction $\mathcal{I}(\theta_{n-1}^{\text{sgd}})^{-1} \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{sgd}})$ as the “natural gradient” and uses information geometry arguments to prove statistical optimality.

An alternative way to implement second-order procedures is to use stochastic approximation to estimate $\mathcal{I}(\theta_\star)$, in addition to the approximation procedure estimating θ_\star . For example, [Amari, Park and Fukumizu \(2000\)](#) proposed the following second-order procedure:

$$(11) \quad \begin{aligned} C_n^{-1} &= (1 - a_n) C_{n-1}^{-1} + a_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{am}}) \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{am}})^\top, \\ \theta_n^{\text{am}} &= \theta_{n-1}^{\text{am}} + \gamma_n C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{am}}), \end{aligned}$$

where $a_n = a_1/n$ is a learning rate sequence, separate from γ_n . By standard stochastic approximation, C_n^{-1} converges to $\mathcal{I}(\theta_\star)$, and thus the procedure in equation (11) is asymptotically optimal. However, there are two important problems with this procedure. First, it is computationally costly because of matrix inversions. A faster way is to apply quasi-Newton ideas. SGD-QN developed by [Bordes, Bottou and Gallinari \(2009\)](#) is such a procedure where the first expensive matrix computations are substituted by the secant condition. Second, the stochastic approximation of $\mathcal{I}(\theta_\star)$ is usually very noisy in high-dimensional problems and this affects the main approximation for θ_\star . Recently, more robust variants of SGD-QN have been proposed [[Byrd et al. \(2016\)](#)].

Another notable adaptive procedure is AdaGrad [[Duchi, Hazan and Singer \(2011\)](#)], which is defined as

$$(12) \quad \begin{aligned} C_n^{-1} &= C_{n-1}^{-1} + \text{diag}(\nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ada}}) \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ada}})^\top), \\ \theta_n^{\text{ada}} &= \theta_{n-1}^{\text{ada}} + \gamma_1 C_n^{1/2} \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ada}}), \end{aligned}$$

where $\text{diag}(\cdot)$ takes the diagonal matrix of its matrix argument, and the learning rate is set constant to $\gamma_n \equiv \gamma_1$. AdaGrad can be considered a second-order procedure because it tries to approximate the Fisher information matrix; however, it only

uses gradient information so technically it is first-order. Under appropriate conditions, $C_n^{-1} \rightarrow \text{diag}(\mathcal{I}(\theta_\star))$ and a simple modification in the proof of Theorem 2.2 can show that the asymptotic variance of the AdaGrad estimate is given by

$$(13) \quad \sqrt{n} \text{Var}(\theta_n^{\text{ada}}) \rightarrow \frac{\gamma_1}{2} \text{diag}(\mathcal{I}(\theta_\star))^{-1/2}.$$

This result reveals an interesting trade-off achieved by AdaGrad and a subtle contrast to first-order SGD procedures. The asymptotic variance of AdaGrad is $O(1/\sqrt{n})$, which indicates significant loss of information. However, this rate is attained *regardless* of the specification of the learning rate parameter γ_1 .⁴ In contrast, as shown in Theorem 2.2, first-order SGD procedures require $2\gamma_1 \mathcal{I}(\theta_\star) - I > 0$ in order to achieve the $O(1/n)$ rate, and the rate is significantly worse if this condition is not met. For instance, Nemirovski et al. (2008) give an example of misspecification of γ_1 where the rate of first-order explicit SGD is $O(n^{-\varepsilon})$, and ε can be arbitrarily small. The variance result in equation (13) is illustrated in the numerical experiments of Section 4.1.1.

2.3. Optimality with averaging. As shown in Section 2.2.1, Theorem 2.2 implies that first-order SGD procedures can be statistically inefficient, especially in many dimensions. One surprisingly simple idea to achieve statistical efficiency is to combine larger learning rates with averaging of the iterates. In particular, we consider the procedure

$$(14) \quad \begin{aligned} \theta_n^{\text{im}} &= \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}), \\ \overline{\theta}_n^{\text{im}} &= \frac{1}{n} \sum_{i=1}^n \theta_i^{\text{im}}, \end{aligned}$$

where θ_n^{im} are the typical implicit SGD iterates in equation (4), and $\gamma_n = \gamma_1 n^{-\gamma}$, $\gamma \in [0.5, 1)$. Under suitable conditions, the iterates $\overline{\theta}_n^{\text{im}}$ are asymptotically efficient. This is formalized in the following theorem.

THEOREM 2.3. *Consider the SGD procedure defined in equation (14) and suppose Assumptions 2.1(a), (c), (d) and (e) hold, where $\gamma \in [0.5, 1)$. Then $\overline{\theta}_n^{\text{im}}$ converges to θ_\star in probability and is asymptotically efficient, that is,*

$$n \text{Var}(\overline{\theta}_n^{\text{im}}) \rightarrow \mathcal{I}(\theta_\star)^{-1}.$$

⁴This follows from a property of recursions [Toulis and Airoldi (2017), Lemma 2.4]. On a high-level, the term γ_{n-1}/γ_n is important for the variance rates of AdaGrad and SGD. When $\gamma_n \propto 1/n$, as in Theorem 2.2, it holds that $\gamma_{n-1}/\gamma_n = 1 + \gamma_n/\gamma_1 + O(\gamma_n^2)$, which explains the quantity $2\mathcal{I}(\theta_\star) - I/\gamma_1$ in first-order SGD. The rate $O(1/n)$ is attained only if $2\mathcal{I}(\theta_\star) - I/\gamma_1 > 0$. When $\gamma_n \propto 1/\sqrt{n}$, as in AdaGrad, it holds that $\gamma_{n-1}/\gamma_n = 1 + o(\gamma_n)$ and the rate $O(1/\sqrt{n})$ is attained without any additional requirements.

REMARKS. In the context of explicit stochastic approximations, averaging was first proposed and analyzed by [Ruppert \(1988\)](#) and [Bather \(1989\)](#). [Ruppert \(1988\)](#) argued that larger learning rates in stochastic approximation uncorrelates the iterates allowing averaging to improve efficiency. [Polyak and Juditsky \(1992\)](#) expanded the scope of averaging by proving asymptotic optimality in more general explicit stochastic approximations that operate under suitable conditions similar to Theorem 2.3. [Polyak and Juditsky \(1992\)](#) thus proved that slowly converging stochastic approximations can be improved by using larger learning rates and averaging of the iterates. Recent work has analyzed explicit updates with averaging [[Zhang \(2004\)](#), [Xu \(2011\)](#), [Bach and Moulines \(2013\)](#), [Shamir and Zhang \(2012\)](#)], and has shown their superiority in numerous learning tasks. More recently, [Toulis, Tran and Airolidi \(2016\)](#) derived the finite-sample error bounds of the averaged implicit SGD estimator.

2.4. *Asymptotic normality.* Asymptotic distributions, or more generally invariance principles, are well studied in classical stochastic approximation [[Ljung, Pflug and Walk \(1992\)](#), Chapter II.8]. In this section, we leverage Fabian's theorem [[Fabian \(1968\)](#)] to show that iterates from implicit SGD are asymptotically normal.

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

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

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

REMARKS. The combined results of Theorems 2.1, 2.2 and 2.4 indicate that implicit SGD is numerically stable and has known asymptotic variance and distribution. Therefore, contrary to explicit SGD that has severe stability issues, implicit SGD emerges as a stable estimation procedure with known standard errors, which enables typical statistical tasks, such as confidence intervals, hypothesis testing and model checking. We show empirical evidence supporting this claim in Section 4.1.2.

2.5. *Stability.* To illustrate the stability, or lack thereof, of both SGD estimators in small-to-moderate samples, we simplify the SGD procedures and inspect the size of the biases $\mathbb{E}(\theta_n^{\text{sgd}} - \theta_\star)$ and $\mathbb{E}(\theta_n^{\text{im}} - \theta_\star)$. In particular, based on Theorem 2.1, we simply assume the Taylor expansion $\nabla \log f(Y_n; X_n, \theta_n) = -\mathcal{I}(\theta_\star)(\theta_n - \theta_\star) + O(\gamma_n)$; to simplify further we ignore the remainder term $O(\gamma_n)$.

Under this simplification, the SGD procedures in equation (1) and in equation (4) can be written as follows:

$$(15) \quad \mathbb{E}(\theta_n^{\text{sgd}} - \theta_\star) = (I - \gamma_n \mathcal{I}(\theta_\star))\mathbb{E}(\theta_{n-1}^{\text{sgd}} - \theta_\star) = P_1^n b_0,$$

$$(16) \quad \mathbb{E}(\theta_n^{\text{im}} - \theta_\star) = (I + \gamma_n \mathcal{I}(\theta_\star))^{-1}\mathbb{E}(\theta_{n-1}^{\text{im}} - \theta_\star) = Q_1^n b_0,$$

where $P_1^n = \prod_{i=1}^n (I - \gamma_i \mathcal{I}(\theta_\star))$, $Q_1^n = \prod_{i=1}^n (I + \gamma_i \mathcal{I}(\theta_\star))^{-1}$, and b_0 denotes the initial bias of the two procedures from a common starting point θ_0 . Thus, the matrices P_1^n and Q_1^n describe how fast the initial bias decays for the explicit and implicit SGD, respectively. In the limit, $P_1^n \rightarrow 0$ and $Q_1^n \rightarrow 0$ [Toulis and Airoldi (2017), proof of Lemma 2.4], and thus both methods are *asymptotically stable*.

However, the explicit procedure has significant stability issues in small-to-moderate samples. By inspection of equation (15), the magnitude of P_1^n is dominated by $\overline{\lambda}_f$, the maximum eigenvalue of $\mathcal{I}(\theta_\star)$. Furthermore, the rate of convergence is dominated by λ_f , the minimum eigenvalue of $\mathcal{I}(\theta_\star)$.⁵ For stability, it is desirable $|1 - \gamma_1 \lambda_i| < 1$, for all eigenvalues $\lambda_i \in \text{eig}(\mathcal{I}(\theta_\star))$. This implies the requirement $\gamma_1 < 2/\overline{\lambda}_f$ for stability. Furthermore, Theorem 2.2 implies the requirement $\gamma_1 > 1/2\lambda_f$ for fast convergence. This is problematic in high-dimensional settings because $\overline{\lambda}_f$ is typically orders of magnitude larger than λ_f . Thus, the requirements for stability and speed of convergence are in conflict in explicit procedures: to ensure stability we need a small learning rate parameter γ_1 , thus paying a high price in convergence which will be at the order of $O(n^{-\gamma_1 \lambda_f})$, and vice versa.

In contrast, the implicit procedure is *unconditionally stable*. The eigenvalues of Q_1^n are $\lambda'_i = \prod_{j=1}^n 1/(1 + \gamma_1 \lambda_i/j) = O(n^{-\gamma_1 \lambda_i})$. Critically, it is no longer required to have a small γ_1 for stability because the eigenvalues of Q_1^n are always less than one. We summarize these findings in the following lemma.

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

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

For the implicit method,

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

REMARKS. Lemma 2.1 shows that in the explicit SGD procedure the effect from the initial bias can be amplified in an arbitrarily large way before fading out, if the learning rate is misspecified (i.e., if $\gamma_1 \gg 1/\overline{\lambda}_f$). This sensitivity of explicit SGD is well known and requires problem-specific considerations to be avoided in practice, for example, preprocessing, small-sample tests, projections, truncation [Chen, Lei and Gao (1988)]. In fact, there exists voluminous work, which is still ongoing, in designing learning rates to stabilize explicit SGD; see, for example, a review by George and Powell (2006). Implicit procedures render such ad-hoc designs obsolete because they remain stable regardless of learning rate design, and still maintain the asymptotic convergence and efficiency properties of explicit SGD procedures.

⁵To see this, note that the eigenvalues of P_1^n are $\lambda'_i = \prod_j (1 - \gamma_1 \lambda_i/j) = O(n^{-\gamma_1 \lambda_i})$ if $0 < \gamma_1 \lambda_i < 1$. See also the proof of Lemma 2.1.

3. Applications. Here, we show how to apply implicit SGD in equation (4) for estimation in generalized linear models, Cox proportional hazards, and more general M-estimation problems. We start by developing an algorithm that efficiently computes the implicit update in equation (4), and is applicable to all aforementioned models.

3.1. Efficient computation of implicit updates. The main difficulty in applying implicit SGD is the solution of the multidimensional fixed-point equation (4). In a large family of models where the likelihood depends on the parameter θ_\star only through the natural parameter $X_n^\top \theta_\star$, the solution of the fixed-point equation is feasible and computationally efficient. We prove the general result in Theorem 3.1.

For the rest of this section, we will treat $\ell(X^\top \theta; Y)$ as a function of the natural parameter $X^\top \theta$ for a fixed outcome Y . Thus, $\ell'(X^\top \theta; Y)$ will refer to the first derivative of ℓ with respect to $X^\top \theta$ with fixed Y .

THEOREM 3.1. *Suppose Assumption 2.1(b) holds, then the gradient of the log-likelihood is a scaled version of covariate X , that is, 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 equation (4) is a scaled version of the gradient calculated at the previous iterate, that is,

$$(17) \quad \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) = \lambda_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}),$$

where the scalar λ_n satisfies

$$(18) \quad \lambda_n \ell'(X_n^\top \theta_{n-1}^{\text{im}}; Y_n) = \ell'(X_n^\top \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n \ell'(X_n^\top \theta_{n-1}^{\text{im}}; Y_n) X_n^\top C_n X_n; Y_n).$$

REMARKS. Theorem 3.1 implies that computing the implicit update in equation (4) reduces to numerically solving the one-dimensional fixed-point equation for λ_n —this idea is implemented in Algorithm 1. As shown in the proof of Theorem 3.1, this implementation is fast because λ_n lies on an interval B_n of size $O(\gamma_n)$. We also note that Theorem 3.1 can be readily extended to cases with linearly separable regularizers, for instance, regularizers using the L_1 norm $\|\theta\| = \sum_i |\theta_i|$. In such cases, there are additional fixed-point equations as in Step 9 of Algorithm 1 that involve the components of the regularizer. More generally, for families of models that do not satisfy Assumption 2.1(b) there are methods to *approximately* perform the implicit update—we discuss one such method in Section 3.3.

3.2. Generalized linear models. In this section, we apply implicit SGD to estimate generalized linear models (GLMs). In such models, Y_n follows an exponential distribution conditional on X_n , and $\mathbb{E}(Y_n | X_n) = h(X_n^\top \theta_\star)$, where h is the

Algorithm 1: Efficient implementation of implicit SGD in equation (4)

```

1: for all  $n \in \{1, 2, \dots\}$  do
2:   # compute search bounds  $B_n$ 
3:    $r_n \leftarrow \gamma_n \ell'(X_n^\top \theta_{n-1}^{\text{im}}; Y_n)$ 
4:    $B_n \leftarrow [0, r_n]$ 
5:   if  $r_n \leq 0$  then
6:      $B_n \leftarrow [r_n, 0]$ 
7:   end if
8:   # solve fixed-point equation by a root-finding method
9:    $\xi = \gamma_n \ell'(X_n^\top \theta_{n-1}^{\text{im}} + \xi X_n^\top C_n X_n; Y_n), \xi \in B_n$ 
10:   $\lambda_n \leftarrow \xi / r_n$ 
11:  # following update is equivalent to update in equation (4)
12:   $\theta_n^{\text{im}} \leftarrow \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}})$ 
13: end for
    
```

transfer function of the GLM model [Nelder and Wedderburn (1972)]. Furthermore, the gradient of the GLM log-likelihood for parameter value θ at data point (X_n, Y_n) is given by

$$(19) \quad \nabla \log f(Y_n; X_n, \theta) = [Y_n - h(X_n^\top \theta)] X_n.$$

The conditional variance of Y_n is $\text{Var}(Y_n | X_n) = h'(X_n^\top \theta_\star) X_n X_n^\top$, and thus the Fisher information matrix is $\mathcal{I}(\theta) = \mathbb{E}(h'(X_n^\top \theta) X_n X_n^\top)$. Thus, the SGD procedures in equation (1) and in equation (4) can be written as

$$(20) \quad \theta_n^{\text{sgd}} = \theta_{n-1}^{\text{sgd}} + \gamma_n C_n [Y_n - h(X_n^\top \theta_{n-1}^{\text{sgd}})] X_n,$$

$$(21) \quad \theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n C_n [Y_n - h(X_n^\top \theta_n^{\text{im}})] X_n.$$

Implementation of explicit SGD is straightforward. Implicit SGD can be implemented through Algorithm 1. In particular, $\log f(Y; X, \theta) \equiv \ell(X^\top \theta; Y)$ with $\ell(\eta; Y) = Y - h(\eta)$. In typical GLMs, h is twice-differentiable and also $h'(\eta) \geq 0$ because it is proportional to the conditional variance of Y given X , thus fulfilling Assumption 2.1(b). In the simplified case where $C_n = I$, the identity matrix, for all n , Algorithm 1 simplifies to Algorithm 2, which was first derived by Toulis, Airolidi and Rennie (2014). We make extensive experiments using Algorithm 2 in Section 4.2.

3.3. Cox proportional hazards model. Here, we apply SGD to estimate a Cox proportional hazards model, which is a popular model in survival analysis [Cox (1972), Klein and Moeschberger (2003)]. Multiple variations of the model exist but for simplicity we will analyze one simple variation that is popular in practice [Davison (2003)]. Consider N individuals, indexed by i , with observed survival

Algorithm 2: Estimation of GLMs with first-order implicit SGD ($C_n = I$)

```

1: for all  $n \in \{1, 2, \dots\}$  do
2:    $r_n \leftarrow \gamma_n[Y_n - h(X_n^\top \theta_{n-1}^{\text{im}})]$ 
3:    $B_n \leftarrow [0, r_n]$ 
4:   if  $r_n \leq 0$  then
5:      $B_n \leftarrow [r_n, 0]$ 
6:   end if
7:    $\xi = \gamma_n[Y_n - h(X_n^\top \theta_{n-1}^{\text{im}} + \xi \|X_n\|^2)], \xi \in B_n$ 
8:    $\theta_n^{\text{im}} \leftarrow \theta_{n-1}^{\text{im}} + \xi X_n$ 
9: end for

```

times Y_i , failure indicators d_i , and covariates X_i . The survival times can be assumed ordered, $Y_1 < Y_2 < \dots < Y_N$, whereas $d_i = 1$ denotes failure (e.g., death) and $d_i = 0$ indicates censoring (e.g., patient dropped out of study). Given a failure for unit i ($d_i = 1$) at time Y_i , the *risk set* \mathcal{R}_i is defined as the set of individuals that could possibly fail at Y_i , that is, all individuals except those who failed or were censored before Y_i . In our simplified model, $\mathcal{R}_i = \{i, i + 1, \dots, N\}$. Define $\eta_i(\theta) = \exp(X_i^\top \theta)$, then the log-likelihood ℓ for θ is given by Davison (2003), Chapter 10,

$$(22) \quad \ell(\theta; X, Y) = \sum_{i=1}^N [d_i - H_i(\theta) \eta_i(\theta)] X_i,$$

where $H_i(\theta) = \sum_{j:i \in \mathcal{R}_j} d_j (\sum_{k \in \mathcal{R}_j} \eta_k(\theta))^{-1}$. In an online setting, where N is infinite and data points (X_i, Y_i) are observed one at a time, future observations affect the likelihood of previous ones, as can be seen by inspection of equation (22). Therefore, we apply SGD assuming fixed N to estimate the MLE θ_N^{MLE} . As mentioned in Section 1, our theory in Section 2 can be applied unchanged if we only substitute θ_* , the true parameter, with the MLE θ_N^{MLE} .

A straightforward implementation of explicit SGD in equation (1) for the Cox model is shown in Algorithm 3. For implicit SGD in equation (4), we have the

Algorithm 3: Explicit SGD for Cox proportional hazards model

```

1 for  $n = 1, 2, \dots$  do
2    $i \leftarrow \text{sample}(1, N)$ 
3    $\hat{H}_i \leftarrow \sum_{j:i \in \mathcal{R}_j} \frac{d_j}{\sum_{k \in \mathcal{R}_j} \eta_k(\theta_{n-1}^{\text{sgd}})}$ 
4    $w_{n-1} \leftarrow [d_i - \hat{H}_i \eta_i(\theta_{n-1}^{\text{sgd}})]$ 
5    $\theta_n^{\text{sgd}} = \theta_{n-1}^{\text{sgd}} + \gamma_n w_{n-1} C_n X_i$ 

```

Algorithm 4: Implicit SGD for Cox proportional hazards model

```

1 for  $n = 1, 2, \dots$  do
2    $i \leftarrow \text{sample}(1, N)$ 
3    $\hat{H}_i \leftarrow \sum_{j:i \in \mathcal{R}_j} \frac{d_j}{\sum_{k \in \mathcal{R}_j} \eta_k(\theta_{n-1}^{\text{im}})}$ 
4    $w(\theta) = d_i - \hat{H}_i \eta_i(\theta)$ 
5    $W_n \leftarrow w(\theta_{n-1}^{\text{im}}) C_n X_i$ 
6    $\lambda_n w(\theta_{n-1}^{\text{im}}) = w(\theta_{n-1}^{\text{im}} + \gamma_n \lambda_n W_n)$ 
7    $\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n W_n$ 
    
```

update

$$(23) \quad \theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n [d_i - H_i(\theta_n^{\text{im}}) \eta_i(\theta_n^{\text{im}})] X_i,$$

which is similar to the implicit procedure for GLMs in equation (21). However, the log-likelihood term $d_i - H_i(\theta_n^{\text{im}}) \eta_i(\theta_n^{\text{im}})$ does not satisfy the conditions of Assumption 2.1(b) because $H_i(\theta)$ may be increasing or decreasing since it depends on terms $X_j^\top \theta$, $j \neq i$. Thus, Theorem 3.1 cannot be applied.

One idea to circumvent this problem is to simply compute $H_i(\cdot)$ on the previous update θ_{n-1}^{im} instead of the current θ_n^{im} . Then update (23) becomes

$$(24) \quad \theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n [d_i - H_i(\theta_{n-1}^{\text{im}}) \eta_i(\theta_n^{\text{im}})] X_i,$$

which now satisfies Assumption 2.1(b) since $H_i(\theta_{n-1}^{\text{im}})$ is constant with respect to θ_n^{im} . This idea is implemented in Algorithm 4, but can be more generally applied in models that don't satisfy Assumption 2.1(b); see Section 5 for a discussion.

3.4. M -estimation. Given N observed data points (X_i, Y_i) and a convex function $\rho : \mathbb{R} \rightarrow \mathbb{R}^+$, the M -estimator is defined as

$$(25) \quad \hat{\theta}^m = \arg \min_{\theta} \sum_{i=1}^N \rho(Y_i - X_i^\top \theta),$$

where it is assumed $Y_i = X_i^\top \theta_\star + \varepsilon_i$, and ε_i are i.i.d. zero mean-valued noise. M -estimators are especially useful in robust statistics [Huber (1964)] because appropriate choice of ρ can reduce the influence of outliers in data. Typically, ρ is twice-differentiable around zero. In this case,

$$(26) \quad \mathbb{E}(\rho'(Y - X^\top \hat{\theta}^m) X) = 0,$$

where the expectation is over the empirical data distribution. Thus, according to Section 1, SGD procedures can be applied to approximate the M -estimator $\hat{\theta}^m$. There has been increased interest in the literature for fast approximation of

Algorithm 5: Implicit SGD for M-estimation

```

1 for  $n = 1, 2, \dots$  do
2    $i \leftarrow \text{sample}(1, N)$ 
3    $w(\theta) = \rho'(Y_i - X_i^\top \theta)$ 
4    $\lambda_n w(\theta_{n-1}^{\text{im}}) \leftarrow w(\theta_{n-1}^{\text{im}} + \gamma_n \lambda_n w(\theta_{n-1}^{\text{im}}) C_n X_i)$    # implicit update
5    $\theta_n^{\text{im}} \leftarrow \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n w(\theta_{n-1}^{\text{im}}) C_n X_i$ 

```

M-estimators due to their robustness [Donoho and Montanari (2016), Jain, Tewari and Kar (2014)]. The implicit SGD procedure for approximating M-estimators is defined in Algorithm 5, and is a simple adaptation of Algorithm 1.

Importantly, the conditions of Assumption 2.1(b) are met because ρ is convex, and thus $\rho'' \geq 0$. Thus, Step 4 of Algorithm 5 is a straightforward application of Algorithm 1 by simply setting $\ell'(X_n' \theta_{n-1}^{\text{im}}; Y_n) \equiv \rho'(Y_n - X_n^\top \theta_{n-1}^{\text{im}})$. The asymptotic variance of θ_n^{im} is also easy to derive. If $S = \mathbb{E}(X_n X_n^\top)$, $C_n \rightarrow C > 0$ such that S and C commute, $\psi^2 = \mathbb{E}(\rho'(\varepsilon_i)^2)$, and $v(z) = \mathbb{E}(\rho'(\varepsilon_i + z))$, Theorem 2.2 can be leveraged to show that

$$(27) \quad n \text{Var}(\theta_n^{\text{im}}) \rightarrow \psi^2 (2v'(0)CS - I)^{-1} CSC.$$

Historically, one of the first applications of explicit stochastic approximation procedures in robust estimation was due to Martin and Masreliez (1975). The asymptotic variance (27) was first derived, only for the explicit SGD case, by Poljak and Tsyppkin (1980) using stochastic approximation theory from Nevelson and Khasminskiĭ (1973).

4. Simulation and data analysis. Here, we demonstrate the computational and statistical advantages of the SGD estimation procedures in equation (1) and in equation (4). For our experiments we developed a new R package, namely `sgd`, which has been published on CRAN. All experiments were conducted on a single laptop running Linux Ubuntu 13.x with 8 cores@2.4 GHz, 16 Gb of RAM memory and 256 Gb of physical storage with SSD technology. A separate set of experiments, which is presented in the supplemental article [Toulis and Airolidi (2017), Section 3], focuses on comparisons of implicit SGD with popular machine learning methods on typical estimation tasks.

4.1. Numerical results. In this section, we aim to illustrate the theoretical results of Section 2, namely the result on asymptotic variance (Theorem 2.2) and asymptotic normality (Theorem 2.4) of SGD procedures.

4.1.1. Asymptotic variance. In this experiment, we use a normal linear model following Xu (2011). The procedures we test are explicit SGD in equation (1),

implicit SGD in equation (4), and AdaGrad in equation (12). For simplicity, we use first-order SGD, that is, $C_n = I$. In the experiment we calculate the empirical variance of said procedures for 25 values of their common learning rate parameter γ_1 in the interval $[1.2, 10]$. For every value of γ_1 , we calculate the empirical variances through the following process, repeated for 150 times. First, we set $\theta_\star = (1, 1, \dots, 1)^\top \in \mathbb{R}^{20}$ as the true parameter value. For iterations $n = 1, 2, \dots, 1500$, we sample covariates as $X_n \sim \mathcal{N}_p(0, S)$, where S is diagonal with elements uniformly on $[0.5, 5]$. The outcome Y_n is then sampled as $Y_n|X_n \sim \mathcal{N}(X_n^\top \theta_\star, 1)$. In every repetition, we store the iterate θ_{1500} for every tested procedure and then calculate the empirical variance of stored iterates over all 150 repetitions.

For any fixed learning rate parameter γ_1 , we set $\gamma_n = \gamma_1/n$ for implicit SGD and $\gamma_n = \gamma_1$ for AdaGrad. For explicit SGD, we set $\gamma_n = \min(0.3, \gamma_1/(n + \|X_n\|^2))$ in order to stabilize its updates. This trick is necessary by the analysis of Section 2.5. In particular, the Fisher information matrix here is $\mathcal{I}(\theta_\star) = \mathbb{E}(X_n X_n^\top) = S$, and thus the minimum eigenvalue is $\lambda_f = 0.5$ and the maximum is $\bar{\lambda}_f = 5$. Therefore, for stability we require $\gamma_1 < 2/\bar{\lambda}_f = 0.4$ and for fast convergence we require $\gamma_1 > 1/(2\lambda_f) = 1$. The two requirements are incompatible, which indicates that explicit SGD can have serious stability issues.

For given $\gamma_1 > 1$, the asymptotic variance of SGD procedures after n iterations is $(1/n)\gamma_1^2(2\gamma_1 S - I)^{-1}S$, by Theorem 2.2. The asymptotic variance of AdaGrad after n iterations is equal to $(\gamma_1/2\sqrt{n})S^{-1/2}$ by equation (13). The log traces of the empirical variance of the SGD procedures and AdaGrad in this experiment are shown in Figure 1. The x -axis corresponds to different values of the learning rate parameter γ_1 , and the y -axis corresponds to the log trace of the empirical variance of the iterates for all three different procedures. We also include curves for the theoretical values of the empirical variances.

We see that our theory predicts well the empirical variances of all methods. Explicit SGD performs on par with implicit SGD for moderate values of γ_1 , however, it required a modification in its learning rate to make it work. Furthermore, explicit SGD quickly becomes unstable at larger values of γ_1 (see, e.g., its empirical variance for $\gamma_1 = 10$), and in several instances, not considered in Figure 1, it numerically diverged. On the other hand, AdaGrad is stable to the specification of γ_1 and tracks its theoretical variance well. However, it gives inefficient estimators because their variance has order $O(1/\sqrt{n})$. Implicit SGD effectively combines stability and good statistical efficiency. First, it remains very stable to the entire range of the learning rate parameter γ_1 . Second, its empirical variance is $O(1/n)$ and it tracks closely the theoretical value predicted by Theorem 2.2 for all γ_1 .

4.1.2. Asymptotic normality. In this experiment, we use the normal linear model in the setup of Section 4.1.1 to check the asymptotic normality result of Theorem 2.4. For simplicity, we only test first-order implicit SGD in equation (4) and first-order explicit SGD.

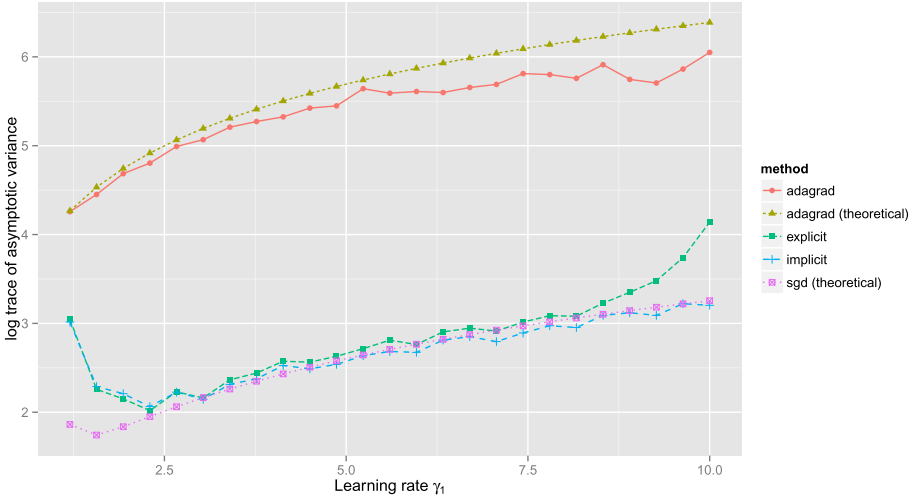


FIG. 1. *Simulation with normal model. The x-axis corresponds to learning rate parameter γ_1 ; the y-axis curves corresponds to log trace of the empirical variance of tested procedures (explicit/implicit SGD, AdaGrad). Theoretical asymptotic variances of SGD and AdaGrad are plotted as well. Implicit SGD is stable and its empirical variance is very close to its asymptotic value. Explicit SGD becomes unstable at large γ_1 . AdaGrad is statistically inefficient but remains stable to large learning rates.*

In the experiment, we define a set of learning rates $(0.5, 1, 3, 5, 6, 7)$. For every learning rate, we take 400 samples of $N(\theta_N - \theta_\star)^\top \Sigma^{-1}(\theta_N - \theta_\star)$, where $N = 1200$ and θ_N denotes either θ_N^{sgd} or θ_N^{im} . The matrix Σ is the asymptotic variance matrix in Theorem 2.4, and $\theta_\star = 10 \exp(-2 \cdot (1, 2, \dots, p))$, is the true parameter value. We use the ground-truth values both for Σ and θ_\star , as we are only interested to test normality of the iterates in this experiment. We also tried $p = 5, 10, 100$ as the parameter dimension. Because the explicit SGD procedure was very unstable across experiments we only report results for $p = 5$. Results on the implicit procedure for larger p are given in the supplemental article [Toulis and Airolidi (2017)], where we also include results for a logistic regression model.

By Theorem 2.4 for implicit SGD, and by classical normality results for explicit SGD [Fabian (1968), Ljung, Pflug and Walk (1992)], the quadratic form $N(\theta_N - \theta_\star)^\top \Sigma^{-1}(\theta_N - \theta_\star)$ is a chi-squared random variable with p degrees of freedom. Thus, for every procedure we plot this quantity against independent samples from a χ_p^2 distribution and visually check for deviations. As before, we tried to stabilize explicit SGD as much as possible by setting $\gamma_n = \min(0.3, \gamma_1/(n + \|X_n\|^2))$. This worked in many iterations, but not for all. Iterations for which explicit SGD diverged were not considered. For implicit SGD, we simply set $\gamma_n = \gamma_1/n$ without additional tuning.

The results of this experiment are shown in Figure 2. The vertical axis on the grid corresponds to different values of the learning rate parameter γ_1 , and the hori-

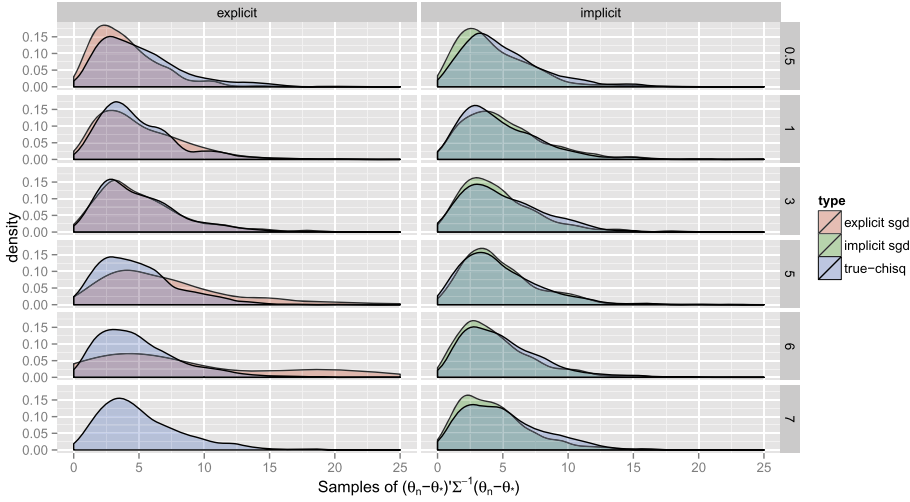


FIG. 2. Simulation with normal model. The x -axis corresponds to the SGD procedure (explicit or implicit) for various values of the learning rate parameter, $\gamma_1 \in \{0.5, 1, 3, 5, 7\}$. The histograms (x -axis) for the SGD procedures are 500 replications of SGD where at each replication we only store the quantity $N(\theta_N - \theta_\star)^\top \Sigma^{-1}(\theta_N - \theta_\star)$, for every method ($N = 1200$); the theoretical covariance matrix Σ is different for every learning rate and is given in Theorem 2.2. The data generative model is the same as in Section 4.1.1. We observe that implicit SGD is stable and follows the nominal chi-squared distribution. Explicit SGD becomes unstable at larger γ_1 and its distribution does not follow the nominal one well. In particular, the distribution of $N(\theta_N^{\text{sgd}} - \theta_\star)^\top \Sigma^{-1}(\theta_N^{\text{sgd}} - \theta_\star)$ becomes increasingly heavy-tailed as the learning rate parameter gets larger, and eventually diverges for $\gamma_1 \geq 7$.

zontal axis has histograms of $N(\theta_N - \theta_\star)^\top \Sigma^{-1}(\theta_N - \theta_\star)$, and also includes samples from a χ^2_5 distribution for visual comparison.

We see that the distribution $N(\theta_N^{\text{im}} - \theta_\star)^\top \Sigma^{-1}(\theta_N^{\text{im}} - \theta_\star)$ of the implicit iterates follows the nominal chi-squared distribution. This also seems to be unaffected by the learning rate parameter. However, the distribution of $N(\theta_N^{\text{sgd}} - \theta_\star)^\top \Sigma^{-1}(\theta_N^{\text{sgd}} - \theta_\star)$ does not follow a chi-squared distribution, except for small learning rate parameter values. For example, as the learning rate parameter increases, the distribution becomes more heavy-tailed (e.g., for $\gamma_1 = 6$), indicating that explicit SGD becomes unstable. Particularly for $\gamma_1 = 7$ explicit SGD diverged in almost all replications, and thus a histogram could not be constructed.

4.2. Comparative performance analysis. In this section, we aim to illustrate the performance of implicit SGD estimation against deterministic estimation procedures that are optimal. The goal is to investigate the extent to which implicit SGD can be as fast as deterministic methods, and to quantify how much statistical efficiency needs be sacrificed to accomplish that.

4.2.1. *Experiments with `glm()` function.* The built-in function `glm()` in R performs deterministic maximum-likelihood estimation through iterative re-weighted least squares. In this experiment, we wish to compare computing time and MSE between first-order implicit SGD and `glm()`. Our simulated data set is a simple normal linear model constructed as follows. First, we sample a binary $p \times p$ design matrix $X = (x_{ij})$ such that $x_{i1} = 1$ (intercept) and $P(x_{ij} = 1) = s$ i.i.d., where $s \in (0, 1)$ determines the sparsity of X . We set $s = 0.08$ indicating that roughly 8% of the X matrix will be nonzero. We generate θ_\star by sampling p elements from $(-1, -0.35, 0, 0.35, 1)$ with replacement. The outcomes are $Y_i = X_i^\top \theta_\star + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$ i.i.d., and $X_i = (x_{ij})$ is the $p \times 1$ vector of i 's covariates. By GLM properties,

$$\mathcal{I}(\theta_\star) = \mathbb{E}(h'(X_i^\top \theta_\star) X_i X_i^\top) = \begin{pmatrix} 1 & s & s & \cdots & s \\ s & s & s^2 & \cdots & s^2 \\ s & s^2 & s & s^2 & \cdots \\ \cdots & s^2 & \cdots & s & \cdots \\ s & s^2 & \cdots & \cdots & s \end{pmatrix}.$$

Slightly tedious algebra can show that the eigenvalues of $\mathcal{I}(\theta_\star)$ are $s(1 - s)$ with multiplicity $(p - 2)$ and the two solutions of $x^2 - A(s)x + B(s) = 0$, where $A(s) = 1 + s + s^2(p - 2)$ and $B(s) = s(1 - s)$. It is thus possible to use the analysis of Section 2.2 and equation (10) to derive a theoretically optimal learning rate. We sample 200 pairs (p, N) for the problem size, uniformly in the ranges $p \sim [10, 500]$ and $N \sim [500, 50,000]$, and obtain running times and MSE of the estimates from implicit SGD and `glm()`. Finally, we then run a regression of computing time and MSE against the problem size (N, p) .

The results are shown in Table 1. We observe that implicit SGD scales better in both sample size N , and especially in the model size p . We also observe that this significant computational gain does not come with much efficiency loss. In fact, averaged over all samples, the MSE of the implicit SGD is 10% higher than the MSE of `glm()`, with a standard error of ± 0.005 . Furthermore, the memory

TABLE 1
Parameters from regressing computation time and MSE against (N, p) in log-scale for `glm()` and implicit GLM. Computation time for `glm()` is roughly $O(p^{1.47}N)$ and for implicit SGD, it is $O(p^{0.2}N^{0.9})$. Implicit SGD scales better in parameter dimension p , whereas MSE for both methods are comparable, at the order of $O(\sqrt{p/N})$

Method	Time (sec)		MSE	
	log p (se)	log N (se)	log p (se)	log N (se)
<code>glm()</code> function	1.46 (0.019)	1.03 (0.02)	0.52 (0.007)	-0.52 (0.006)
implicit SGD	0.19 (0.012)	0.9 (0.01)	0.58 (0.007)	-0.53 (0.006)

TABLE 2

Comparison of implicit SGD with `biglm`. MSE is defined as $\|\theta_N - \theta_\star\|/\|\theta_0 - \theta_\star\|$. Values “*” indicate out-of-memory errors. `biglm` was run in combination with the `ffdf` package to map big data files to memory. Implicit SGD used a similar but slower ad-hoc method. The table reports computation times excluding file access

p	N	Size (GB)	Procedure			
			<code>biglm</code>		Implicit SGD	
			Time (secs)	MSE	Time (secs)	MSE
1e2	1e5	0.021	2.32	0.028	2.4	0.028
1e2	5e5	0.103	8.32	0.012	7.1	0.012
1e2	1e6	0.206	16	0.008	14.7	0.009
1e2	1e7	2.1	232	0.002	127.9	0.002
1e2	1e8	20.6	*	*	1397	0.00
1e3	1e6	2.0	*	*	31.38	0.153
1e4	1e5	2.0	*	*	25.05	0.160

requirements (not reported in Table 1) are roughly $O(Np^2)$ for `glm()` and only $O(p)$ for implicit SGD.

4.2.2. *Experiments with `biglm`.* The package `biglm` is a popular choice for fitting GLMs with data sets where N is large but p is small.⁶ It works in an iterative way by splitting the data set in many parts, and by updating the model parameters using incremental QR decomposition [Miller (1992)], which results in only $O(p^2)$ memory requirement. In this experiment, we compare implicit SGD with `biglm` on larger data sets of Section 4.2.1. with small p and large N such that Np remains roughly constant.

The results are shown in Table 2. We observe that implicit SGD is significantly faster at a very small efficiency loss. The difference is more dramatic at large p ; for example, when $p = 10^3$ or $p = 10^4$, `biglm` quickly runs out of memory, whereas implicit SGD works without problems.

4.2.3. *Experiments with `glmnet`.* The `glmnet` package in R [Friedman, Hastie and Tibshirani (2010)] is a deterministic optimization algorithm for generalized linear models that uses the elastic net. It performs a component-wise update of the parameter vector, utilizing thresholding from the regularization penalties for more computationally efficient updates. One update over all parameters costs roughly $O(Np)$ operations. Additional computational gains are achieved when the design matrix is sparse because fewer components are updated per each iteration.

⁶See <http://cran.r-project.org/web/packages/biglm/index.html> for the `biglm` package. `biglm` is part of the High-Performance Computing (HPC) task view of the CRAN project here <http://cran.r-project.org/web/views/HighPerformanceComputing.html>.

TABLE 3

Comparing implicit SGD with `glmnet`. Table reports running times (in secs.) and MSE for both procedures. The MSE of `glmnet` is calculated as the median MSE over the 100 grid values of regularization parameter computed by default [Friedman, Hastie and Tibshirani (2010)]

Method	Metric	Correlation (ρ)			
		0	0.2	0.6	0.9
$N = 1000, p = 10$					
glmnet	time (sec)	0.005	0.005	0.008	0.022
	mse	0.083	0.085	0.099	0.163
sgd	time (sec)	0.011	0.011	0.011	0.011
	mse	0.042	0.042	0.049	0.053
$N = 5000, p = 50$					
glmnet		0.058	0.067	0.119	0.273
		0.044	0.046	0.057	0.09
sgd		0.059	0.056	0.057	0.057
		0.019	0.02	0.023	0.031
$N = 100,000, p = 200$					
glmnet		2.775	3.017	4.009	10.827
		0.017	0.017	0.021	0.033
sgd		1.475	1.464	1.474	1.446
		0.004	0.004	0.004	0.006

In this experiment, we compare implicit SGD with `glmnet` on a subset of experiments in the original package release [Friedman, Hastie and Tibshirani (2010)]. In particular, we implement the experiment of Section 5.1 in that paper, as follows. First, we sample the design matrix $X \sim \mathcal{N}_p(0, \Sigma)$, where $\Sigma = b^2 U + I$ and U is the $p \times p$ matrix of ones. The parameter $b = \sqrt{\rho/(1 - \rho)}$, where ρ is the target correlation of columns of X , is controlled in the experiments. The outcomes are $Y = X\theta_\star + \sigma^2 \varepsilon$, where $\theta_j^\star = (-1)^j \exp(-2(j - 1)/20)$, and ε is a standard p -variate normal. The parameter σ is tuned to achieve a predefined signal-noise ratio. We report average computation times in Table 3 over 10 replications, which expands Table 1 of Friedman, Hastie and Tibshirani (2010).

First, we observe that implicit SGD is consistently faster than the `glmnet` method. In particular, the SGD method scales better at larger p following a sub-linear growth as noted in Section 4.2.1. Interestingly, it is also not affected by covariate correlation, whereas `glmnet` gets slower as more components need to be updated at every iteration. For example, with correlation $\rho = 0.9$ and $N = 1e5$, $p = 200$, the SGD method is almost $10\times$ faster.

Second, to compare `glmnet` with implicit SGD in terms of MSE we picked the median MSE produced by the grid of regularization parameters computed by `glmnet`. We picked the median because `glmnet` is a deterministic method and so at the best regularization value its MSE will be lower than the MSE of implicit

SGD. However, implicit SGD seems to perform better against the median performance of `glmnet`. Furthermore, Table 3 indicates a clear trend where, for bigger dimensions p and higher correlation ρ , implicit SGD is performing better than `glmnet` in terms of efficiency as well. We obtain similar results in a comparison on a logistic regression model, which we present in Section 3 of the supplemental article [Toulis and Airolidi (2017)].

4.2.4. Cox proportional hazards. In this experiment, we test the performance of implicit SGD on estimating the parameters of a Cox proportional hazards model in a setup that is similar to the numerical example of Simon et al. (2011), Section 3.

We consider $N = 1000$ units with covariates $X \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = 0.2U + I$, and U is the matrix of ones. We sample times as $Y_i \sim \text{Expo}(\eta_i(\theta_\star))$, where $\eta_i(\theta) = \exp(X_i^\top \theta)$, and $\theta_\star = (\theta_{\star,k})$ is a vector with $p = 20$ elements defined as $\theta_{\star,k} = 2(-1)^{-k} \exp(-0.1k)$. Time Y_i is censored, and thus $d_i = 0$, according to probability $(1 + \exp(-a(Y_i - q))^{-1})$, where q is a quantile of choice (set here as $q = 0.8$), and a is set such that $\min\{Y_i\}$ is censored with a prespecified probability (set here as 0.1%). We replicate 50 times the following process. First, we run implicit SGD for $2N$ iterations, and then measure $\text{MSE } \|\theta_n^{\text{im}} - \theta_\star\|^2$, for all $n = 1, 2, \dots, 2N$. To set the learning rates, we use equation (10), where the Fisher matrix is diagonally approximated, through the AdaGrad procedure (12). We then take the 5%, 50% and 95% quantiles of MSE across all repetitions and plot them against iteration number n .

The results are shown in Figure 3 (left panel). In the figure, we also plot (horizontal dashed lines) the 5% and 95% quantiles of the MSE of the MLE, assumed to be the best MSE achievable for SGD. We observe that implicit SGD performs well compared to MLE in this small-sized problem. In particular, implicit SGD, under the aforementioned generic tuning of learning rates, converges to the region of optimal MLE in a few thousands of iterations. In experiments with explicit SGD,

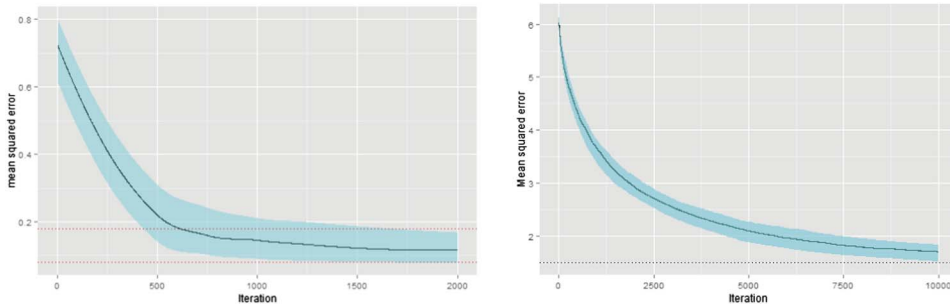


FIG. 3. Left panel: 5%–95% quantile band of implicit SGD estimates (in cyan) against 5%–95% band of the MLE (dashed lines) for a Cox proportional hazards model (50 replications); Right panel: 5%–95% quantile band of implicit SGD estimates (in cyan) against median MLE (dashed line) on an M-estimation task (100 replications).

we were not able to replicate this performance because of numerical instability. We note that there are no standard implementations of explicit SGD for estimating Cox proportional hazards models, to our best knowledge.

4.2.5. *M-estimation.* In this experiment, we test the performance of implicit SGD, in particular Algorithm 5, on a M-estimation problem in a setup that is similar to the simulation example of [Donoho and Montanari \(2016\)](#), Example 2.4.

We set $N = 1000$ data points and $p = 200$ as the parameter dimension. We sample θ_\star as a random vector with norm $\|\theta_\star\| = 6\sqrt{p}$, and sample the design matrix as $X \sim \mathcal{N}(0, (1/N)I)$. The outcomes are sampled i.i.d. from a contaminated normal distribution, that is, with probability 95%, $Y_n \sim \mathcal{N}(X_n^\top \theta_\star, 1)$ and $Y_n = 10$ with probability 5%.

The results over 2000 iterations of implicit SGD are shown in Figure 3 (right panel). In the figure, we plot the 5% and 95% quantiles of MSE of implicit SGD over 100 replications of the experiment. We also plot (horizontal dashed line) the median MSE of the MLE estimator, computed using the `coxph` built-in command of R. We observe that SGD converges steadily to the best possible MSE. Similar behavior was observed under various modifications of the simulation parameters.

4.3. *National morbidity-mortality air pollution (NMMAPS) study.* The NMMAPS study [[Samet et al. \(2000\)](#), [Dominici et al. \(2002\)](#)] analyzed the risks of air pollution to public health. Several cities (108 in the US) are included in the study with daily measurements covering more than 13 years (roughly 5000 days) including air pollution data (e.g., concentration of CO in the atmosphere) together with health outcome variables such as number of respiratory-related deaths.

The original study fitted a Poisson generalized additive model (GAM), separately for each city due to data set size. Recent research [[Wood, Goude and Shaw \(2015\)](#)] has developed procedures similar to `biglm`'s iterative QR decomposition to fit all cities simultaneously on the full data set with approximately $N = 1.2$ million observations and $p = 802$ covariates (7 Gb in size). In this experiment, we construct a GAM model using data from all cities in the NMMAPS study in a process that is very similar (but not identical) to the data set of [Wood, Goude and Shaw \(2015\)](#).

Our final data set has $N = 1,426,806$ observations and $p = 794$ covariates including all cities in the NMMAPS study (8.6 GB in size), and is fit using a simple first-order implicit SGD procedure with $C_n = I$ and $\gamma_1 = 1$. The runtime for implicit SGD was roughly 120 seconds, which is 6x faster than the 12 minutes reported by [Wood, Goude and Shaw \(2015\)](#) on a similar computer. We cannot directly compare the estimates from the two procedures because the datasets used were different. However, we can compare the estimates of our model with the estimates of `glm()` on a random small subset of the data. For that purpose, we subsampled $N = 50,000$ observations and $p = 50$ covariates (19.5 MB in size) and fit the smaller data set using implicit SGD and `glm()`. A scatter plot of the

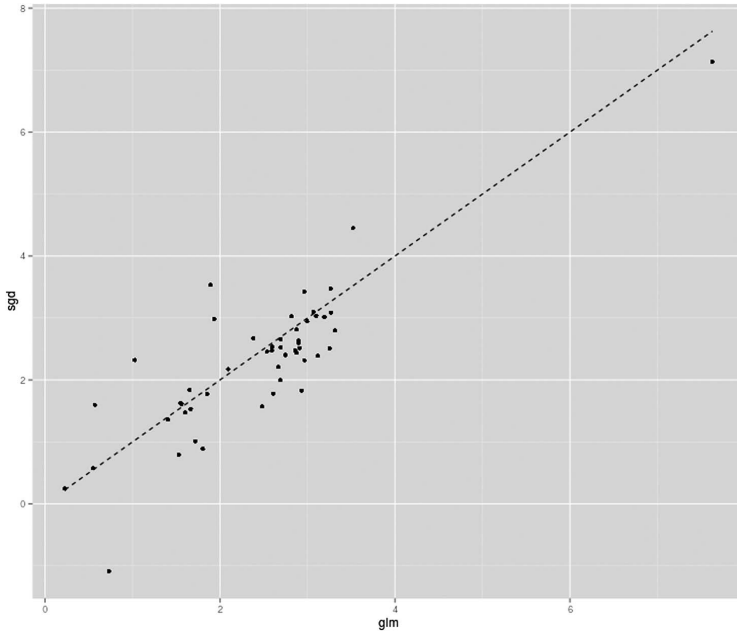


FIG. 4. Estimates of implicit SGD (y-axis) and $\text{glm}()$ (x-axis) on a subset of the NMMAPS data set with $N = 50,000$ observations and $p = 50$ covariates, which is roughly 5% of the entire data set.

estimates is shown in Figure 4. The estimates of the implicit of the SGD procedure are very close to MLE, while further replications of the aforementioned testing process revealed the same pattern indicating that implicit SGD converged on all replications.

5. Discussion. The theory in Section 2 suggests that implicit SGD is numerically stable and has known asymptotic variance and asymptotic distribution. The experiments in Section 4 show that the empirical properties of SGD are well predicted by theory. In contrast, explicit SGD is unstable and cannot work well without problem-specific tuning. Thus, we conclude that implicit SGD is a principled estimation procedure and is superior to widely-used explicit SGD procedures.

Intuitively, implicit SGD leverages second-order information at every iteration, although second-order quantities do not need to be computed in equation (4). To demonstrate this, we build upon the argument that was first introduced in Section 1 and equation (8), which is repeated below for convenience:

$$\Delta\theta_n^{\text{im}} \approx [I + \gamma_n \hat{\mathcal{I}}(\theta_0; X_n, Y_n)]^{-1} \Delta\theta_n^{\text{sgd}};$$

here, $\Delta\theta_n^{\text{im}} = \theta_n^{\text{im}} - \theta_0$ and $\Delta\theta_n^{\text{sgd}} = \theta_n^{\text{sgd}} - \theta_0$, and the matrix $\hat{\mathcal{I}}(\theta_0; X_n, Y_n) = -\nabla^2 \log f(Y_n; X_n, \theta)|_{\theta=\theta_0}$ is the observed Fisher information at θ_0 . In other words, the implicit procedure is a *shrunked* version of the explicit one, where the shrinkage factor depends on the observed information.

Naturally, the implicit SGD iterate θ_n^{im} has also a Bayesian interpretation. In particular, θ_n^{im} is the posterior mode of a Bayesian model defined as

$$(28) \quad \begin{aligned} \theta | \theta_{n-1}^{\text{im}} &\sim \mathcal{N}(\theta_{n-1}^{\text{im}}, \gamma_n C_n), \\ Y_n | X_n, \theta &\sim f(\cdot; X_n, \theta). \end{aligned}$$

The explicit SGD update θ_n^{sgd} can be written as in equation (28); however, f needs to be substituted with its linear approximation around $\theta_{n-1}^{\text{sgd}}$. Thus, equation (28) provides an alternative explanation why implicit SGD is more principled than explicit SGD. Furthermore, it indicates possible improvements for implicit SGD. For example, the prior in equation (28) could be chosen to fit better the parameter space (e.g., θ_\star being on the simplex). Krzysztow et al. (2007) and Nemirovski et al. (2008) have argued that appropriate implicit updates can fit better in the geometry of the parameter space, and thus converge faster. Setting up the parameters of the prior is also crucial. Whereas in explicit SGD there is no statistical intuition behind learning rates γ_n , equation (28) reveals that in implicit SGD the terms $(\gamma_n C_n)^{-1}$ encode the statistical information up to iteration n . It follows immediately that it is optimal, in general, to set $\gamma_n C_n = \mathcal{I}(\theta_\star)^{-1}/n$, which is a special case of Theorem 2.2.

The Bayesian formulation of equation (28) also explains the stability of implicit SGD. In Theorem 2.1, we showed that the initial conditions are discounted at an exponential rate, regardless of misspecification of the learning rates. This stability of implicit SGD allows several ideas for improvements. For example, constant learning rates could be used in implicit SGD to speed up convergence toward a region around θ_\star . A sequential hypothesis test could decide on whether θ_n^{im} has reached that region or not, and switch to the theoretically optimal $1/n$ rate accordingly. Alternatively, we could run implicit SGD with AdaGrad learning rates and switch to $1/n$ rates when the theoretical $O(1/\sqrt{n})$ variance of AdaGrad becomes larger than the $O(1/n)$ variance of implicit SGD. Such schemes using constant rates with explicit SGD are very hard to do in practice because of instability.

Regarding statistical efficiency, a key technical result in this paper is that the asymptotic variance of implicit SGD can be calculated exactly using Theorem 2.2. Optimal learning rates were suggested in equation (10) that depend on the eigenvalues of the unknown Fisher matrix $\mathcal{I}(\theta_\star)$. In this paper, we used second-order procedures of Section 2.2.1 to iteratively estimate the eigenvalues, however better methods are certainly possible and could improve the performance of implicit SGD. For example, it is known that typical iterative methods usually overestimate the largest eigenvalue and underestimate the smallest eigenvalue, in small-to-moderate samples. This crucially affects the behavior of stochastic approximations with learning rates that depend on sample eigenvalues. Empirical Bayes methods have been shown to be superior in iterative estimation of eigenvalues of large matrices [Mestre (2008)], and it would be interesting to apply such methods to design the learning rates of implicit SGD procedures.

Regarding computational efficiency, we developed Algorithm 1 which implements implicit SGD on a large family of statistical models. However, the trick used in fitting the Cox proportional hazards model in Section 3.3 can be more generally applied to models outside this family. For example, assume a log-likelihood gradient of the form $s(X^\top \theta; Y)G(\theta; X, Y)$, where both its scale $s(\cdot)$ and direction $G(\cdot)$ depend on model parameters θ ; this violates conditions of Assumption 2.1(b). The implicit update in equation (4)—where $C_n = I$ for simplicity—would be $\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n s(X_n^\top \theta_n^{\text{im}}; Y_n)G(\theta_n^{\text{im}}; X_n, Y_n)$, which cannot be computed by Algorithm 1. One way to circumvent this problem is to use an implicit update only on the scale and use an explicit update on the direction, that is, $\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n s(X_n^\top \theta_n^{\text{im}}; Y_n)G(\theta_{n-1}^{\text{im}}; X_n, Y_n)$. This form of updates expands the applicability of implicit SGD.

Finally, hypothesis testing and construction of confidence intervals using SGD estimates is an important issue that has remained unexplored. In experiments of Section 4.1.2, we showed that implicit SGD is indeed asymptotically normal in several simulation scenarios. However, as SGD procedures are iterative, there needs to be a rigorous and general method to decide whether SGD iterates have converged to the asymptotic regime. Several methods, such as bootstrapping the data set, could be used for that. Furthermore, conservative confidence intervals could be constructed through multivariate Chebyshev inequalities or other strategies [Marshall and Olkin (1960)].

5.1. Concluding remarks. In this paper, we introduced a new stochastic gradient descent procedure that uses implicit updates at every iteration, which we termed implicit SGD. Equation (8) shows, intuitively, that the iterates of implicit SGD are a shrunk version of the standard iterates, where the shrinkage factor depends on the observed Fisher information matrix. Thus, implicit SGD combines the computational efficiency of first-order methods with the numerical stability of second-order methods.

In a theoretical analysis, we derived nonasymptotic upper bounds for the mean-squared errors of implicit SGD iterates, and the asymptotic variance of both explicit and implicit SGD iterates. Our analysis quantifies the efficiency loss of SGD procedures, and suggests principled strategies to calibrate a hyperparameter that is common to both explicit and implicit SGD procedures, known as the learning rate. We illustrated the use of implicit SGD for statistical estimation in generalized linear models, Cox proportional hazards model, and general M-estimation problems.

Viewed as statistical estimation procedures, our results suggest that implicit SGD has the same asymptotic efficiency to explicit SGD. However, the implicit procedure is significantly more stable than the explicit one with respect to misspecification of the learning rate. In general, explicit SGD procedures are sensitive to outliers and to misspecification of the learning rates, making it impossible to apply without problem-specific tuning. In theory and in extensive experiments, implicit procedures emerge as principled iterative estimation methods because they

are numerically stable, they are robust to tuning of hyperparameters, and their standard errors are well predicted by theory. Thus, implicit stochastic gradient descent is poised to become a workhorse of estimation from large data sets in statistical practice.

Acknowledgments. The authors wish to thank Joe Blitzstein, Leon Bottou, Bob Carpenter, David Dunson, Andrew Gelman, Brian Kulis, Xiao-Li Meng, Natesh Pillai and Neil Shephard for useful comments and discussion. We are grateful to four reviewers, the Associate Editor, and the Editor for offering suggestions that improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to “Asymptotic and finite-sample properties of estimators based on stochastic gradients” (DOI: [10.1214/16-AOS1506SUPP](https://doi.org/10.1214/16-AOS1506SUPP); .pdf). The proofs of all technical results are provided in an online supplement [Toulis and Airolidi (2017)]. There, we also provide numerical results that extend the results in Section 4 of this article—referred to as the “main paper” in the supplement.

REFERENCES

- AMARI, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Comput.* **10** 251–276.
- AMARI, S.-I., PARK, H. and FUKUMIZU, K. (2000). Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Comput.* **12** 1399–1409.
- BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems* 773–781.
- BATHER, J. A. (1989). Stochastic approximation: A generalisation of the Robbins–Monro procedure. In *Proceedings of the Fourth Prague Symposium on Asymptotic Statistics (Prague, 1988)* 13–27. Charles Univ., Prague. [MR1051424](#)
- BECK, A. and TEOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. [MR2486527](#)
- BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer, Berlin. [MR1082341](#)
- BERTSEKAS, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Math. Program.* **129** 163–195. [MR2837879](#)
- BORDES, A., BOTTOU, L. and GALLINARI, P. (2009). SGD-QN: Careful quasi-Newton stochastic gradient descent. *J. Mach. Learn. Res.* **10** 1737–1754. [MR2534877](#)
- BORKAR, V. S. (2008). *Stochastic Approximation*. Cambridge Univ. Press, Cambridge. [MR2442439](#)
- BOTTOU, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT*2010* 177–186. Springer, Heidelberg. [MR3362066](#)
- BYRD, R. H., HANSEN, S. L., NOCEDAL, J. and SINGER, Y. (2016). A stochastic quasi-Newton method for large-scale optimization. *SIAM J. Optim.* **26** 1008–1031. [MR3485979](#)
- CAPPÉ, O. and MOULINES, E. (2009). On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 593–613. [MR2749909](#)
- CHEN, H. F., LEI, G. and GAO, A. J. (1988). Convergence and robustness of the Robbins–Monro algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.* **27** 217–231. [MR0931029](#)

- CHUNG, K. L. (1954). On a stochastic approximation method. *Ann. Math. Stat.* **25** 463–483. [MR0064365](#)
- COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. [MR0341758](#)
- DAVISON, A. C. (2003). *Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. [MR1998913](#)
- DEAN, J., CORRADO, G., MONGA, R., CHEN, K., DEVIN, M., MAO, M., SENIOR, A., TUCKER, P., YANG, K., LE, Q. V. et al. (2012). Large scale distributed deep networks. In *Advances in Neural Information Processing Systems* 1223–1231.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. [MR0501537](#)
- DOMINICI, F., DANIELS, M., ZEGER, S. L. and SAMET, J. M. (2002). Air pollution and mortality: Estimating regional and national dose-response relationships. *J. Amer. Statist. Assoc.* **97** 100–111. [MR1963390](#)
- DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. [MR3568043](#)
- DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159. [MR2825422](#)
- DUCHI, J. and SINGER, Y. (2009). Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.* **10** 2899–2934. [MR2579916](#)
- EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. [MR2485012](#)
- FABIAN, V. (1968). On asymptotic normality in stochastic approximation. *Ann. Math. Stat.* **39** 1327–1332. [MR0231429](#)
- FABIAN, V. (1978). On asymptotically efficient recursive estimation. *Ann. Statist.* **6** 854–866. [MR0478506](#)
- FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222** 309–368.
- FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- GEORGE, A. P. and POWELL, W. B. (2006). Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming. *Mach. Learn.* **65** 167–198.
- GREEN, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B* **46** 149–192. [MR0781879](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. [MR2722294](#)
- HENNIG, P. and KIEFEL, M. (2013). Quasi-Newton methods: A new direction. *J. Mach. Learn. Res.* **14** 843–865. [MR3049491](#)
- HOFFMAN, J. D. and FRANKEL, S. (2001). *Numerical Methods for Engineers and Scientists*. CRC press, Boca Raton, FL.
- HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. [MR0161415](#)
- JAIN, P., TEWARI, A. and KAR, P. (2014). On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems* 685–693.
- KLEIN, J. P. and MOESCHBERGER, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media New York.

- KRAKOWSKI, K. A., MAHONY, R. E., WILLIAMSON, R. C. and WARMUTH, M. K. (2007). A geometric view of non-linear on-line stochastic gradient descent. Available at <https://users.soe.ucsc.edu/~manfred/pubs/T3.pdf>.
- LANGE, K. (2010). *Numerical Analysis for Statisticians*, 2nd ed. Springer, New York. [MR2655999](#)
- LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics* **31**. Springer, New York. [MR1639875](#)
- LIONS, P.-L. and MERCIER, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* **16** 964–979. [MR0551319](#)
- LJUNG, L., PFLUG, G. and WALK, H. (1992). *Stochastic Approximation and Optimization of Random Systems*. *DMV Seminar* **17**. Birkhäuser, Basel. [MR1162311](#)
- MARSHALL, A. W. and OLKIN, I. (1960). Multivariate Chebyshev inequalities. *Ann. Math. Stat.* **31** 1001–1014. [MR0119234](#)
- MARTIN, R. D. and MASRELIEZ, C. J. (1975). Robust estimation via stochastic approximation. *IEEE Trans. Inform. Theory* **IT-21** 263–271. [MR0395111](#)
- MESTRE, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory* **54** 5113–5129. [MR2589886](#)
- MILLER, A. J. (1992). Algorithm as 274: Least squares routines to supplement those of gentleman. *Applied Statistics* 458–478.
- MOULINES, E. and BACH, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems* 451–459.
- NAGUMO, J.-I. and NODA, A. (1967). A learning method for system identification. *IEEE Trans. Automat. Control* **12** 282–287.
- NATIONAL RESEARCH COUNCIL (2013). *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, DC.
- NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.
- NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. [MR2486041](#)
- NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York. [MR0702836](#)
- NEVELSON, M. B. and KHASMINSKIĬ, R. Z. (1973). *Stochastic Approximation and Recursive Estimation* **47**. Amer. Math. Society, Washington.
- PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.
- POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. [MR1167814](#)
- POLJAK, B. T. and TSYPKIN, JA. Z. (1980). Robust identification. *Automatica J. IFAC* **16** 53–63. [MR0571554](#)
- POLYAK, B. T. and TSYPKIN, YA. Z. (1979). Adaptive estimation algorithms (convergence, optimality, stability). *Avtomat. i Telemekh.* **3** 71–84. [MR0544876](#)
- ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. [MR0042668](#)
- ROCKAFELLAR, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14** 877–898. [MR0410483](#)
- ROSASCO, L., VILLA, S. and VÜ, B. C. (2014). Convergence of stochastic proximal gradient algorithm. Preprint. Available at [arXiv:1403.5074](#).
- RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process. Technical report, Dept. Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY.
- SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Stat.* **29** 373–405. [MR0098427](#)
- SAKRISON, D. J. (1965). Efficient recursive estimation; application to estimating the parameters of a covariance function. *Internat. J. Engrg. Sci.* **3** 461–483. [MR0182082](#)

- SAMET, J. M., ZEGER, S. L., DOMINICI, F., CURRIERO, F., COURSAK, I., DOCKERY, D. W., SCHWARTZ, J. and ZANOBETTI, A. (2000). The national morbidity, mortality, and air pollution study. Part II: Morbidity and mortality from air pollution in the United States. *Res. Rep. Health Eff. Inst.* **94** 5–79.
- SCHMIDT, M., LE ROUX, N. and BACH, F. (2013). Minimizing finite sums with the stochastic average gradient. Technical report, HAL 00860051.
- SHAMIR, O. and ZHANG, T. (2012). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. Preprint. Available at [arXiv:1212.1824](https://arxiv.org/abs/1212.1824).
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox’s proportional Hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13.
- SINGER, Y. and DUCHI, J. C. (2009). Efficient learning using forward-backward splitting. In *Advances in Neural Information Processing Systems* 495–503.
- SLOCK, D. T. M. (1993). On the convergence behavior of the lms and the normalized lms algorithms. *IEEE Trans. Signal Process.* **41** 2811–2825.
- TOULIS, P. and AIROLDI, E. M. (2015a). Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Stat. Comput.* **25** 781–795. [MR3360492](https://arxiv.org/abs/1510.00967)
- TOULIS, P. and AIROLDI, E. M. (2015b). Implicit stochastic approximation. Preprint. Available at [arXiv:1510.00967](https://arxiv.org/abs/1510.00967).
- TOULIS, P. and AIROLDI, E. M. (2017). Supplement to “Asymptotic and finite-sample properties of estimators based on stochastic gradients.” DOI:[10.1214/16-AOS1506SUPP](https://doi.org/10.1214/16-AOS1506SUPP).
- TOULIS, P., AIROLDI, E. M. and RENNIE, J. (2014). Statistical analysis of stochastic gradient methods for generalized linear models. *J. Mach. Learn. Res. W&CP* **32 (ICML)** 667–675.
- TOULIS, P., TRAN, D. and AIROLDI, E. M. (2016). Towards stability and optimality in stochastic gradient descent. *J. Mach. Learn. Res. W&CP* **51 (AISTATS)**.
- TRAN, D., TOULIS, P. and AIROLDI, E. M. (2015). Stochastic gradient descent methods for estimation with large data sets. Preprint. Available at [arXiv:1509.06459](https://arxiv.org/abs/1509.06459).
- WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* 681–688.
- WIDROW, B. and HOFF, M. E. (1960). Adaptive switching circuits. *Defense Technical Information Center*.
- WOOD, S. N., GOUDE, Y. and SHAW, S. (2015). Generalized additive models for large data sets. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 139–155. [MR3293922](https://arxiv.org/abs/1309.6213)
- XIAO, L. and ZHANG, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.* **24** 2057–2075. [MR3285905](https://arxiv.org/abs/1309.6213)
- XU, W. (2011). Towards optimal one pass large scale learning with averaged stochastic gradient descent. Preprint. Available at [arXiv:1107.2490](https://arxiv.org/abs/1107.2490).
- ZHANG, T. (2004). Solving large scale linear prediction problems using gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning* 116. ACM, New York.

ECONOMETRICS AND STATISTICS
BOOTH SCHOOL OF BUSINESS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: panos.toulis@chicagobooth.edu

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: airoldi@fas.harvard.edu

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 Airolidi, 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 $\log f(Y; X, \theta) \equiv \ell(X^\top \theta; Y)$, which depends on θ only through the natural parameter $X^\top \theta$.*
- (c) *Function ℓ is concave, twice differentiable almost surely wrt natural parameter $X^\top \theta$ and Lipschitz with constant L_0 wrt θ .*
- (d) *The observed Fisher information matrix $\hat{\mathcal{I}}_n(\theta) = -\nabla^2 \ell(X_n^\top \theta; Y_n)$ has non-vanishing trace, i.e., there exists constant $b > 0$ such that $\text{trace}(\hat{\mathcal{I}}_n(\theta)) \geq b$ almost surely, for all θ . The Fisher information matrix $\mathcal{I}(\theta_\star) = \mathbb{E}(\hat{\mathcal{I}}_n(\theta_\star))$ 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 $C_n = C + O(\gamma_n)$, where $\|C\| = 1$, $C \succ 0$ and symmetric, and C commutes with $\mathcal{I}(\theta_\star)$. For every C_n , $\min \text{eig}(C_n) = \underline{\lambda}_c > 0$, and $\max \text{eig}(C_n) = \overline{\lambda}_c < \infty$.*
- (f) *Let $\Xi_n = \mathbb{E}(\nabla \log f(Y_n; X_n, \theta_\star) \nabla \log f(Y_n; X_n, \theta_\star)^\top \mid \mathcal{F}_{n-1})$, then $\|\Xi_n - \Xi\| = O(1)$ for all n , and $\|\Xi_n - \Xi\| \rightarrow 0$, for a symmetric positive-definite Ξ . Let $\sigma_{n,s}^2 = \mathbb{E}(\mathbb{I}_{\|\xi_n(\theta_\star)\|^2 \geq s/\gamma_n} \|\xi_n(\theta_\star)\|^2)$, then for all $s > 0$, $\sum_{i=1}^n \sigma_{i,s}^2 = o(n)$ if $\gamma = 1$, and $\sigma_{n,s}^2 = o(1)$ 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*

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

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). \square

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

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

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,

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

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

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

where $Q_i^n = \prod_{j=i}^n (1+c_j)/(1+b_j)$, 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,

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

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

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

This inequality also holds for $n = n_0$.

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

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

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

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

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

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

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

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

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

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 $n \geq n_0$,*

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

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

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

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

By definition of δ_n and Ineq. (12),

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

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 $Q_i^n = \prod_{j=i}^n (1+b_j)^{-1}$. Thus, for a lower bound,

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

and for an upper bound,

$$(16) \quad \begin{aligned} Q_1^n &\leq \exp(-\log(1+b_1)/b_1 \sum_{i=1}^n b_i), \quad [\text{by Lemma 2.1}] \\ Q_1^n &\leq \exp(-\log(1+b_1)\phi_\beta(n)). \quad [\text{by Ineq. (13)}] \end{aligned}$$

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

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

where the last inequality follows from Ineq. (16). \square

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

$$(18) \quad \lambda_n \leq \frac{1}{1 + \gamma_n \lambda_c \underline{b}},$$

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

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

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

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

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

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

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

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

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

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

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

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

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

□

Finite-sample analysis.

THEOREM 2.1. *Let $\delta_n = \mathbb{E}(\|\theta_n^{\text{im}} - \theta_\star\|^2)$. 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,*

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

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

$$\begin{aligned} \theta_n^{\text{im}} - \theta_\star &= \theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n C_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) \\ \theta_n^{\text{im}} - \theta_\star &= \theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n \lambda_n C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) \quad [\text{By Theorem (3.1)}] \\ \|\theta_n^{\text{im}} - \theta_\star\|^2 &= \|\theta_{n-1}^{\text{im}} - \theta_\star\|^2 \\ &\quad + 2\gamma_n \lambda_n (\theta_{n-1}^{\text{im}} - \theta_\star)^\top C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) \\ (24) \quad &\quad + \gamma_n^2 \|C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}})\|^2. \end{aligned}$$

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

$$(25) \quad \|C_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}})\|^2 \leq \bar{\lambda}_c^2 \|\theta_n^{\text{im}} - \theta_{n-1}^{\text{im}}\|^2 \leq 4L_0^2 \bar{\lambda}_c^2 \gamma_n^2,$$

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

$$\begin{aligned} & \mathbb{E}(2\gamma_n \lambda_n (\theta_{n-1}^{\text{im}} - \theta_\star)^\top C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}})) \\ & \leq \frac{2\gamma_n}{1 + \gamma_n \underline{\lambda}_c b} \mathbb{E}((\theta_{n-1}^{\text{im}} - \theta_\star)^\top C_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}})) \quad [\text{by Lemma 2.3}] \\ & \leq \frac{2\gamma_n}{1 + \gamma_n \underline{\lambda}_c b} \mathbb{E}((\theta_{n-1}^{\text{im}} - \theta_\star)^\top C_n \nabla h(\theta_{n-1}^{\text{im}})) \quad [\text{where } \nabla h(\theta_{n-1}^{\text{im}}) = \mathbb{E}(\nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) | \mathcal{F}_{n-1})] \\ (26) \quad & \leq -\frac{2\gamma_n \underline{\lambda}_f \underline{\lambda}_c}{1 + \gamma_n \underline{\lambda}_c b} \|\theta_{n-1}^{\text{im}} - \theta_\star\|^2 \quad [\text{by strong convexity, Assumption 2.1(d).}] \end{aligned}$$

Taking expectations in Eq. (24) and substituting Ineqs. (25) and (26) into Eq. (24) yields the recursion,

$$(27) \quad \mathbb{E}(\|\theta_n^{\text{im}} - \theta_\star\|^2) \leq (1 - \frac{2\gamma_n \underline{\lambda}_f \underline{\lambda}_c}{1 + \gamma_n \underline{\lambda}_c b}) \mathbb{E}(\|\theta_{n-1}^{\text{im}} - \theta_\star\|^2) + 4L_0^2 \bar{\lambda}_c^2 \gamma_n^2.$$

Define $\mu_1 = 2\underline{\lambda}_f$, $\mu_2 = \max\{\gamma_1 \underline{\lambda}_c \mu_1 (b - \mu_1), 0\}$ and $\mu = \mu_1 / (\mu_1 + \mu_2)$; note that $\mu \in (0, 1]$, and $\mu = 1$ only when $\mu_2 = 0$, i.e., $2\underline{\lambda}_f \geq b$. Through simple algebra we obtain

$$(28) \quad (1 - \frac{2\gamma_n \underline{\lambda}_f \underline{\lambda}_c}{1 + \gamma_n \underline{\lambda}_c b}) \leq \frac{1}{1 + 2\gamma_n \mu \underline{\lambda}_f \underline{\lambda}_c},$$

for all $n > 0$. Therefore we can write recursion (27) as

$$(29) \quad \mathbb{E}(\|\theta_n^{\text{im}} - \theta_\star\|^2) \leq \frac{1}{1 + 2\gamma_n \mu \underline{\lambda}_f \underline{\lambda}_c} \mathbb{E}(\|\theta_{n-1}^{\text{im}} - \theta_\star\|^2) + 4L_0^2 \bar{\lambda}_c^2 \gamma_n^2.$$

We can now apply Corollary 2.1 with $a_n = 4L_0^2 \bar{\lambda}_c^2 \gamma_n^2$ and $b_n = 2\gamma_n \mu \underline{\lambda}_f \underline{\lambda}_c$. \square

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

$\mathbb{E}(\|\nabla \log f(Y_n; X_n, \theta_*)\|^2) \leq \sigma^2$, as it is typical in the literature, would simply add a term $\gamma_n^2 L^2 \mathbb{E}(\|\theta_n^{\text{im}} - \theta_*\|^2) + \gamma_n^2 \sigma^2$ in the right-hand side of Eq. (24). 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 $\gamma_n^2 L^2 \|\theta_{n-1}^{\text{sgd}} - \theta_*\|^2$ increases the upper bound and can make $\|\theta_n^{\text{sgd}} - \theta_*\|^2$ 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 $\|\mathbb{L}_B\| = O(\|B\|)$.

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

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

$$(31) \quad Y_n = \mathbb{L}_{I + \gamma_n B_n}^{-1} \{X_{n-1} + \gamma_n (C + D_n)\},$$

such that

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

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

$$(32) \quad \|X_n B + B X_n - 2C\| \rightarrow 0 \text{ and } \|Y_n B + B Y_n - 2C\| \rightarrow 0.$$

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

PROOF. We make the following definitions.

$$(33) \quad \Gamma_n = I - \gamma_n B_n,$$

$$(34) \quad P_i^n = \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \circ \mathbb{L}_{\Gamma_i},$$

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

$$(35) \quad \|P_i^n\| = O\left(\prod_{j=i}^n \|I - \gamma_j B_j\|\right) \leq K_0 e^{-K_1 \sum_{j=i}^n \gamma_j},$$

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), $\Gamma(n) \rightarrow \infty$ and thus $P_i^n \rightarrow \mathbb{L}_0$ as $n \rightarrow \infty$ 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

$$(36) \quad \begin{aligned} X_n &= \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \circ \mathbb{L}_{\Gamma_1} \{X_0\} + \gamma_n C + \gamma_n D_n \\ &\quad + a_{n-1} \mathbb{L}_{\Gamma_n} \{C\} + a_{n-1} \mathbb{L}_{\Gamma_n} \{D_{n-1}\} \\ &\quad + \cdots + \\ &\quad + a_1 \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \circ \mathbb{L}_{\Gamma_2} \{C\} + a_1 \mathbb{L}_{\Gamma_n} \circ \mathbb{L}_{\Gamma_{n-1}} \circ \cdots \circ \mathbb{L}_{\Gamma_2} \{D_1\} \\ &= P_1^n \{X_0\} + S_n \{C\} + \tilde{D}_n, \end{aligned}$$

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

$$(37) \quad \sum_{i=1}^n \gamma_i P_{i+1}^n = \mathbb{L}_{B_n}^{-1} + \sum_{i=2}^n P_i^n (\mathbb{L}_{B_{i-1}}^{-1} - \mathbb{L}_{B_i}^{-1}) - P_1^n \mathbb{L}_{B_1}^{-1}.$$

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

$$(38) \quad \gamma_n \mathbb{I} = \mathbb{L}_{I - \Gamma_n} \circ \mathbb{L}_{B_n}^{-1}.$$

Therefore, if we collect the coefficients of the terms $\mathbb{L}_{B_n}^{-1}$ in the right-hand side of (37), we get

$$\begin{aligned} &\mathbb{L}_{B_n}^{-1} + \sum_{i=2}^n P_i^n (\mathbb{L}_{B_{i-1}}^{-1} - \mathbb{L}_{B_i}^{-1}) - P_1^n \mathbb{L}_{B_1}^{-1} \\ &= (P_2^n - P_1^n) \mathbb{L}_{B_1}^{-1} + (P_3^n - P_2^n) \mathbb{L}_{B_2}^{-1} + \cdots + (P_{n+1}^n - P_n^n) \mathbb{L}_{B_n}^{-1} \\ &= P_2^n \circ \mathbb{L}_{I - \Gamma_1} \circ \mathbb{L}_{B_1}^{-1} + P_3^n \circ \mathbb{L}_{I - \Gamma_2} \circ \mathbb{L}_{B_2}^{-1} + \cdots + P_{n+1}^n \circ \mathbb{L}_{I - \Gamma_n} \circ \mathbb{L}_{B_n}^{-1} \\ &= P_2^n (\gamma_1 \mathbb{I}) + P_3^n (\gamma_2 \mathbb{I}) + \cdots + P_{n+1}^n (\gamma_n \mathbb{I}) \quad [\text{by Eq. (38)}] \\ &= \sum_{i=1}^n \gamma_i P_{i+1}^n, \end{aligned}$$

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,

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

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

$$(39) \quad \sum_{i=2}^n P_i^n \circ (\mathbb{L}_{B_{i-1}}^{-1} - \mathbb{L}_{B_i}^{-1}) \rightarrow \mathbb{L}_0,$$

and since $P_1^n \rightarrow \mathbb{L}_0$, we conclude from Equation (38) that

$$(40) \quad \lim_{n \rightarrow \infty} \sum_{i=1}^n \gamma_i P_{i+1}^n = \lim_{n \rightarrow \infty} \mathbb{L}_{B_n}^{-1} = \mathbb{L}_B^{-1}.$$

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

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

Since $\|D_n\| \rightarrow 0$ it follows that $\|\mathbb{L}_{B_n}^{-1} \{D_n\}\| \rightarrow 0$ and $\|(\mathbb{L}_{B_{i-1}}^{-1} \{D_{i-1}\} - \mathbb{L}_{B_i}^{-1} \{D_i\})\| = O(\gamma_i^2)$. Recall that $P_1^n \rightarrow \mathbb{L}_0$, and thus $\tilde{D}_n \rightarrow \mathbf{0}$. Finally, we substitute this result in Equation (38) to get $X_n \rightarrow \mathbb{L}_B^{-1} \{C\}$.

For the second recursion of the lemma,

$$(41) \quad Y_n = \mathbb{L}_{I+\gamma_n B_n}^{-1} \{Y_{n-1} + \gamma_n (C + D_n)\},$$

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

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

As before, $Q_i^n \rightarrow \mathbb{L}_0$. Solving the recursion (41) yields

$$(42) \quad Y_n = Q_1^n \{Y_0\} + S_n \{C\} + \tilde{D}_n,$$

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

$$(43) \quad \mathbb{L}_{B_n}^{-1} \circ (\mathbb{I} - \mathbb{L}_{\Gamma_n}^{-1}) = \gamma_n \mathbb{L}_{\Gamma_n}^{-1},$$

$$(44) \quad \mathbb{L}_{B_n}^{-1} \mathbb{L}_{\Gamma_n}^{-1} = \mathbb{L}_{\Gamma_n}^{-1} \mathbb{L}_{B_n}^{-1}.$$

It holds,

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

where the first line is obtained by Eq. (43) and the second line by Eq. (44). Thus, similar to the previously analyzed recursion, $S_n \rightarrow \mathbb{L}_B^{-1}$ and $\tilde{D}_n \rightarrow 0$. Therefore, $Y_n \rightarrow \mathbb{L}_B^{-1} \{C\}$.

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. \square

COROLLARY 2.2. *Consider the matrix recursions*

$$(45) \quad X_n = \mathbb{L}_{I - \gamma_n B_n} \{X_{n-1}\} + \gamma_n^2 (C + D_n),$$

$$(46) \quad Y_n = \mathbb{L}_{I + \gamma_n B_n}^{-1} \{Y_{n-1} + \gamma_n^2 (C + D_n)\},$$

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

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

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

PROOF. Both $X_n, Y_n \rightarrow 0$ by direct application of Lemma (2.4). Let $\tilde{X}_n = (1/\gamma_n)X_n$. First, divide (45) by γ_n to obtain

$$(47) \quad \tilde{X}_n = \mathbb{L}_{I - \gamma_n B_n} \left\{ \tilde{X}_{n-1} \right\} \frac{\gamma_{n-1}}{\gamma_n} + \gamma_n (C + D_n).$$

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

$$(48) \quad \mathbb{L}_{I - \gamma_n B_n} \left\{ \tilde{X}_{n-1} \right\} \frac{\gamma_{n-1}}{\gamma_n} = \mathbb{L}_{I - \gamma_n B_n} \left\{ \tilde{X}_{n-1} \right\} + \gamma_n \tilde{X}_{n-1} + O(\gamma_n^2).$$

Therefore, we can rewrite Eq. (47) as

$$(49) \quad \tilde{X}_n = \mathbb{L}_{I-\gamma_n\Gamma_n} \left\{ \tilde{X}_{n-1} \right\} + \gamma_n(C + D_n),$$

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

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*

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

The asymptotic variance of the implicit SGD estimator satisfies

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

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

$$\begin{aligned} \mathbb{E}(S_n(\theta_\star) \mid X_n) &= 0 \\ \text{Var}(S_n(\theta_\star)) &= \mathbb{E}(\text{Var}(S_n(\theta_\star) \mid X_n)) = \mathcal{I}(\theta_\star) \\ J_h(\theta) &= -\mathcal{I}(\theta), \quad [\text{under regularity conditions}] \\ h(\theta_n^{\text{im}}) &= -\mathcal{I}(\theta_\star)(\theta_n^{\text{im}} - \theta_\star) + O(\gamma_n) \quad [\text{by Theorem 2.1}], \\ (50) \quad \|\text{Var}(S_n(\theta) - S_n(\theta_\star))\| &\leq \mathbb{E}(\|S_n(\theta) - S_n(\theta_\star)\|^2) \leq L_0^2 \mathbb{E}(\|\theta - \theta_\star\|^2). \end{aligned}$$

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

$$(51) \quad \theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n C_n S_n(\theta_n^{\text{im}}) = \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n C_n S_n(\theta_{n-1}^{\text{im}}),$$

where λ_n is defined in Theorem 3.1 and $\lambda_n = 1 - O(\gamma_n)$ by Eq. (18). Then, taking variances on both sides of Eq. (51) yields

$$(52) \quad V_n = V_{n-1} + \gamma_n^2 C_n \text{Var}(S_n(\theta_n^{\text{im}})) C_n^\top + \gamma_n \text{Cov}(\theta_{n-1}^{\text{im}}, S_n(\theta_n^{\text{im}})) C_n^\top + \gamma_n C_n \text{Cov}(S_n(\theta_n^{\text{im}}), \theta_{n-1}^{\text{im}}).$$

We can simplify all variance/covariance terms in Eq. (52) as follows.

$$\begin{aligned}
 C_n \text{Var} (S_n(\theta_n^{\text{im}})) C_n^\top &= C_n \text{Var} (S_n(\theta_\star) + [S_n(\theta_n^{\text{im}}) - S_n(\theta_\star)]) C_n^\top \\
 &= C\mathcal{I}(\theta_\star)C^\top + o(1), \quad [\text{by Eqs. (50), Theorem (2.1), and Assumption 2.1(e)}] \\
 \text{Cov} (\theta_{n-1}^{\text{im}}, S_n(\theta_n^{\text{im}})) &= \text{Cov} (\theta_{n-1}^{\text{im}}, S_n(\theta_{n-1}^{\text{im}})) + \text{Cov} (\theta_{n-1}^{\text{im}}, (\lambda_n - 1)S_n(\theta_{n-1}^{\text{im}})) \\
 &= \text{Cov} (\theta_{n-1}^{\text{im}}, h(\theta_{n-1}^{\text{im}})) + O(\gamma_n) \\
 &= V_{n-1}\mathcal{I}(\theta_\star) + O(\gamma_n). \quad [\text{by Eq. (50), Theorem (2.1), Eq. (18)}].
 \end{aligned}$$

Similarly, $\text{Cov} (h(\theta_n^{\text{im}}), \theta_{n-1}^{\text{im}}) = V_{n-1}\mathcal{I}(\theta_\star) + O(\gamma_n)$. We can now rewrite Eq. (52) as

$$(53) \quad V_n = \mathbb{L}_{I-\gamma_n B_n} \{V_{n-1}\} + \gamma_n^2 [C\mathcal{I}(\theta_\star)C^\top + o(1)],$$

where $B_n = 2C_n\mathcal{I}(\theta_\star)$ and $B_n \rightarrow 2C\mathcal{I}(\theta_\star)$. Corollary 2.2 on recursion (53) yields the following closed-form, since B and C commute and C is symmetric:

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

The regularity conditions (50) and the convergence rates of Theorem 2.1 that are crucial for this proof also hold for the explicit procedure. \square

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 θ_\star in probability and is asymptotically efficient, i.e.,*

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

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

$$(54) \quad \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) = \nabla \log f(Y_n; X_n, \theta_\star) - \mathcal{I}(\theta_\star)(\theta_n^{\text{im}} - \theta_\star) + O(\gamma_n).$$

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

$$(55) \quad \theta_n^{\text{im}} - \theta_\star = (I + \gamma_n F)^{-1}(\theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n \varepsilon_n + O(\gamma_n^2)).$$

We make the following definitions.

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

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

$$(57) \quad \overline{\theta}_n^{\text{im}} - \theta_\star = (1/n)D_0^n(\overline{\theta}_n^{\text{im}} - \theta_\star) + (1/n)\sum_i^{n-1} D_i^n e_i.$$

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 $1 + \frac{1}{1+\gamma_1\lambda} + \frac{1}{1+\gamma_1\lambda} \frac{1}{1+\gamma_2\lambda} + \dots \leq \sum_{i=0}^n \exp(-K\lambda \sum_{k=1}^i \gamma_k)$, where the last inequality is obtained by Lemma 2.1. Because $\sum \gamma_i \rightarrow \infty$, the summands are $o(1)$, and thus D_0^n is $o(n)$. \square

LEMMA 2.6. *Suppose Assumption 2.1(a) and Eq. (54) hold. Then,*

$$(58) \quad \gamma_i D_i^n (I + \gamma_i F)^{-1} = \Omega_i^n + F^{-1},$$

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 $q_i^j = O(\exp(-K\lambda s_i^j))$, 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

$$(59) \quad \lambda' = \frac{\gamma_i}{1 + \gamma_i \lambda} (q_{i+1}^i + q_{i+1}^{i+1} + \dots + q_{i+1}^{n-1}).$$

Thus,

$$(60) \quad \lambda' (1 + \gamma_i \lambda) = \sum_{k=i}^{n-1} \gamma_i q_{i+1}^k.$$

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

$$(61) \quad \begin{aligned} \gamma_{i+1} \lambda q_{i+1}^{i+1} + q_{i+1}^{i+1} &= 1 \\ \gamma_{i+2} \lambda q_{i+1}^{i+2} + q_{i+1}^{i+2} &= q_{i+1}^{i+1} \\ &\dots\dots\dots \\ \gamma_{n-2} \lambda q_{i+1}^{n-2} + q_{i+1}^{n-2} &= q_{i+1}^{n-3} \\ \gamma_{n-1} \lambda q_{i+1}^{n-1} + q_{i+1}^{n-1} &= q_{i+1}^{n-2}. \end{aligned}$$

By summing over the terms we obtain:

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

If we combine with (59) we obtain

$$(63) \quad \lambda \sum_{k=i}^{n-1} \gamma_i q_{i+1}^k + \lambda \sum_{k=i}^{n-1} (\gamma_k - \gamma_i) q_{i+1}^k + q_{i+1}^{n-1} = 1 + \gamma_i \lambda \quad \text{or}$$

$$(64) \quad (1 + \gamma_i \lambda) \lambda \lambda' + \lambda \sum_{k=i}^{n-1} (\gamma_k - \gamma_i) q_{i+1}^k + q_{i+1}^{n-1} = 1 + \gamma_i \lambda.$$

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

$$(65) \quad \begin{aligned} \lambda \sum_{k=i}^{n-1} (\gamma_k - \gamma_i) q_{i+1}^k &= \lambda \sum_{k=i}^{n-1} \left[\sum_{j=i}^k (\gamma_{j+1} - \gamma_j) \right] q_{i+1}^k = \lambda \sum_{k=i}^{n-1} \left[\sum_{j=i}^k \gamma_j o(\gamma_j) \right] q_{i+1}^k \\ &\leq \lambda o(\gamma_i) \sum_{k=i}^{n-1} s_i^k q_{i+1}^k \triangleq q_i^n. \end{aligned}$$

In Eq. (65) we used $(\gamma_{j+1} - \gamma_j)/\gamma_j = O(n^{-1-\gamma})/n^{-\gamma} = O(n^{-1}) = o(\gamma_j)$, by Assumption 2.1(a). Our goal is now to show $\sum_{i=0}^{n-1} q_i^n = o(n)$. Since $q_{i+1}^k = O(\exp(-K\lambda s_{i+1}^k))$ by (Polyak and Juditsky, 1992, p845, see A6 and A7) we obtain that $q_i^n \rightarrow 0$ for fixed i as $n \rightarrow \infty$. Therefore we can rewrite Eq. (63) as

$$(66) \quad \lambda' \lambda + q_i^n + O(q_{i+1}^n) = 1,$$

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

Our proof is now complete. By Eq. (57) and Lemmas 2.5 and 2.6 we have

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

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

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

□

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*

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

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

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

$$(67) \quad \theta_n^{\text{im}} - \theta_\star = (I - \gamma_n A_n)(\theta_{n-1}^{\text{im}} - \theta_\star) + \gamma_n \xi_n(\theta_\star) + O(\gamma_n^2),$$

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

$$(68) \quad (I + \gamma_n \hat{\mathcal{I}}_n(\theta_\star))(\theta_n^{\text{im}} - \theta_\star) = \theta_{n-1}^{\text{im}} - \theta_\star + \gamma_n S_n(\theta_\star),$$

where $\hat{\mathcal{I}}_n(\theta_\star) = -\nabla^2 S_n(\theta_\star)$; note, $\mathbb{E}(\hat{\mathcal{I}}_n(\theta_\star)) = \mathcal{I}(\theta_\star)$. Because $(I + \gamma_n \hat{\mathcal{I}}_n(\theta_\star))^{-1} = I - \gamma_n \hat{\mathcal{I}}_n(\theta_\star) + O(\gamma_n^2)$, we can rewrite Eq. (68) as

$$(69) \quad \theta_n^{\text{im}} - \theta_\star = (I - \gamma_n \hat{\mathcal{I}}_n(\theta_\star))(\theta_{n-1}^{\text{im}} - \theta_\star) + \gamma_n S_n(\theta_\star) + O(\gamma_n^2).$$

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. \square

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

LEMMA 2.1. *Let $\bar{\lambda}_f = \max \text{eig}(\mathcal{I}(\theta_\star))$, and suppose $\gamma_n = \gamma_1/n$ and $\gamma_1 \bar{\lambda}_f > 1$. Then, the maximum eigenvalue of P_1^n satisfies*

$$\max_{n>0} \max \text{eig}(P_1^n) = \Theta(2^{\gamma_1 \bar{\lambda}_f} / \sqrt{\gamma_1 \bar{\lambda}_f}).$$

For the implicit method,

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

PROOF. We will use the following intermediate result:

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

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

$$(70) \quad (b-1)(b/2-1)(b/3-1) \cdots (2-1) = \frac{1}{2} \binom{b}{b/2} = \Theta(2^b / \sqrt{2\pi b}),$$

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 $\bar{\lambda}_f$ of $\mathcal{I}(\theta_*)$. For the implicit SGD updates, we note that the eigenvalues of $(I + \gamma_n \mathcal{I}(\theta_*))^{-1}$ are less than one, for any $\gamma_n > 0$ and any Fisher matrix. \square

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.,

$$(71) \quad \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) = \lambda_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}),$$

where the scalar λ_n satisfies

$$(72) \quad \lambda_n \ell'(X_n^\top \theta_{n-1}^{\text{im}}; Y_n) = \ell'(X_n^\top \theta_{n-1}^{\text{im}} + \gamma_n \lambda_n \ell'(X_n^\top \theta_{n-1}^{\text{im}}; Y_n) X_n^\top C_n X_n; Y_n).$$

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

$$(73) \quad \begin{aligned} \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) &= \lambda_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}) \quad \text{or} \\ \ell'(X_n^\top \theta_n^{\text{im}}; Y_n) X_n &= \lambda_n \ell'(X_n^\top \theta_{n-1}^{\text{im}}; Y_n) X_n. \end{aligned}$$

We also have,

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

Substituting the expression for θ_n^{im} in Eq.(74) into Eq. (73) we obtain the desired result of the Theorem in Eq. (71).

We now prove the last claim of the theorem regarding the search bounds for λ_n . For notational convenience, define $a = X_n^T \theta_{n-1}^{\text{im}}$, $g(x) = \ell'(x; Y_n)$, and $c = X_n^T C_n X_n$, where $c > 0$ because C_n are positive definite. Also let $x_\star = \gamma_n \lambda_n g(a)$, then the fixed-point equation (72) can be written as

$$(75) \quad x_\star = \gamma_n g(a + x_\star c).$$

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 (75). 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.

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

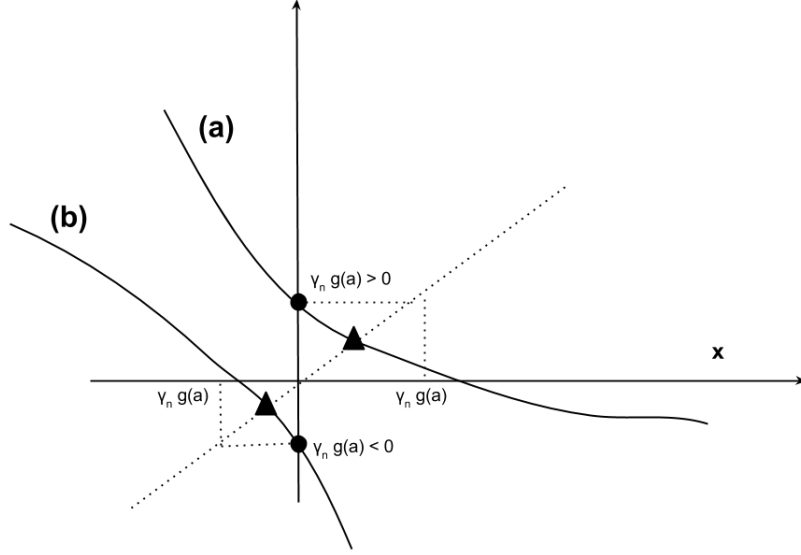


FIG 1. (Search bounds for solution of Eq. (75)) **Case $g(a) > 0$:** Corresponds to Curve (a) defined as $\gamma_n g(a + xc)$, $c > 0$. The solution x_* of fixed point equation (75) (corresponding to right triangle) is between 0 and $\gamma_n g(a)$ since Curve (a) is decreasing. **Case $g(a) < 0$:** Corresponds to Curve (b) also defined as $\gamma_n g(a + xc)$. The solution x_* of fixed point equation (75) (left triangle) is between $\gamma_n g(a)$ and 0 since Curve (b) is also decreasing.

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.

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 $(0, 0)^\top$, $(1, 0)^\top$ or $(0, 1)^\top$ with probabilities 0.6, 0.2 and 0.2 respectively. We set $\theta_* = (\theta_1, \theta_2)^\top$ for some θ_1, θ_2 , and assume $Y_n \sim \text{Poisson}(\exp(X_n^\top \theta_*))$, where the transfer function h is the exponential, i.e., $h(x) = \exp(x)$. It follows,

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

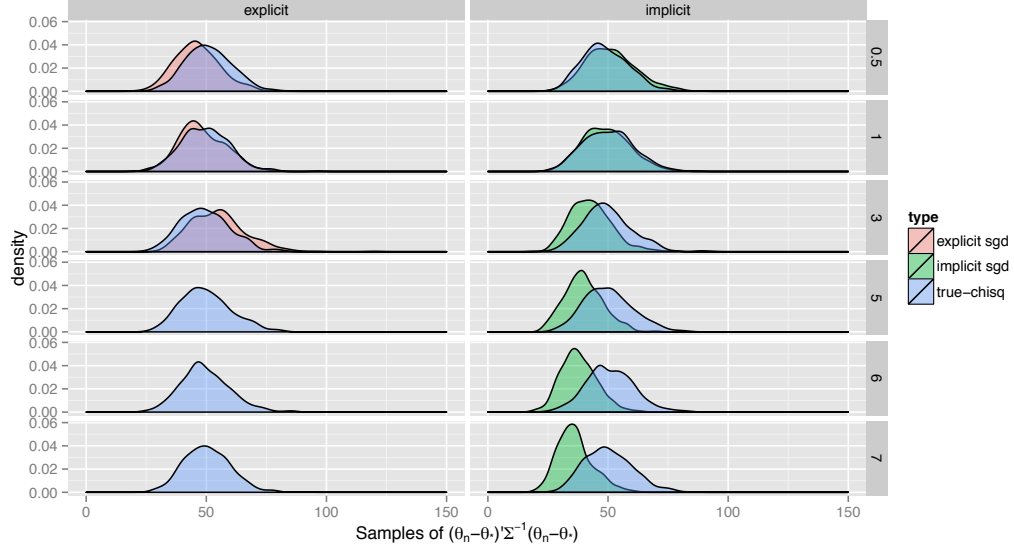


FIG 2. *Simulation with normal model for $p = 50$ parameters. Implicit SGD is stable and follows the nominal chi-squared distribution well, regardless of the particular learning rate. Explicit SGD becomes unstable at larger γ_1 and its distribution does not follow the theoretical distribution chi-squared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

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

$$(76) \quad \Sigma = \frac{2}{3} \begin{pmatrix} \frac{e^{\theta_1}}{(4/3)e^{\theta_1}-1} & 0 \\ 0 & \frac{e^{\theta_2}}{(4/3)e^{\theta_2}-1} \end{pmatrix} = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.62 \end{pmatrix}.$$

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. (76). In contrast, the standard SGD estimates are unstable and their L_2 distance to the true values θ_* are orders of magnitude larger than the implicit ones (see Table 1 for sample

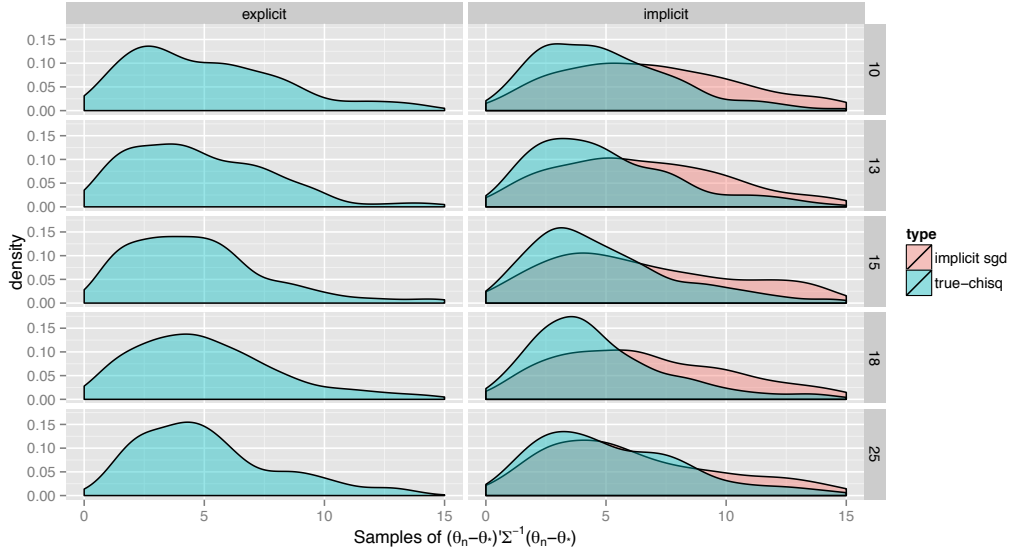


FIG 3. *Simulation with logistic regression model for $p = 5$. Learning rates are larger than in the linear normal model to ensure the asymptotic covariance matrix of Theorem (2.2) is positive definite. Implicit SGD is stable and follows the nominal chi-squared distribution regardless of the learning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

quantiles). By Lemma 2.1 in the main paper, such deviations are expected for standard SGD because the largest eigenvalue of $\mathcal{I}(\theta_*)$ 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 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

TABLE 1
Quantiles of $\|\theta_n^{\text{sgd}} - \theta_\star\|$ and $\|\theta_n^{\text{im}} - \theta_\star\|$. Values larger than $1e3$ are marked “*”.

METHOD	QUANTILES					
	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

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, $\theta_n^{\text{sgd}} = \theta_{n-1}^{\text{sgd}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{sgd}})$.
- 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

¹We 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 θ_\star .

TABLE 2

Experiments comparing implicit SGD with *glmnet*. Covariates X are sampled as normal, with cross-correlation ρ , and the outcomes are sampled as $\mathbf{y} \sim \text{Binom}(\mathbf{p})$, $\text{logit}(\mathbf{p}) = \mathcal{N}(X\theta_*, \sigma^2 I)$. 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
<hr/>					
$N = 1000, p = 10$					
<hr/>					
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
<hr/>					
$N = 5000, p = 50$					
<hr/>					
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
<hr/>					
$N = 100000, p = 200$					
<hr/>					
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

natural gradient by using a larger-dimensional learning rate. It has the added advantage of being less sensitive than first-order methods 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 COV-TYPE 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

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}

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 hyperparameters. 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 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 $N = 1\text{e}6$ as the number of observations, and $p = 20$ be the number of covariates. We also set $\theta_\star = (0, 0, \dots, 0)^\top \in \mathbb{R}^{20}$ as the true parameter value. The random variables X_n are distributed i.i.d. as $X_n \sim \mathcal{N}_p(0, H)$, where H is a randomly generated symmetric matrix with eigenvalues $1/k$, for $k = 1, \dots, p$. The outcome Y_n is sampled from a nor-

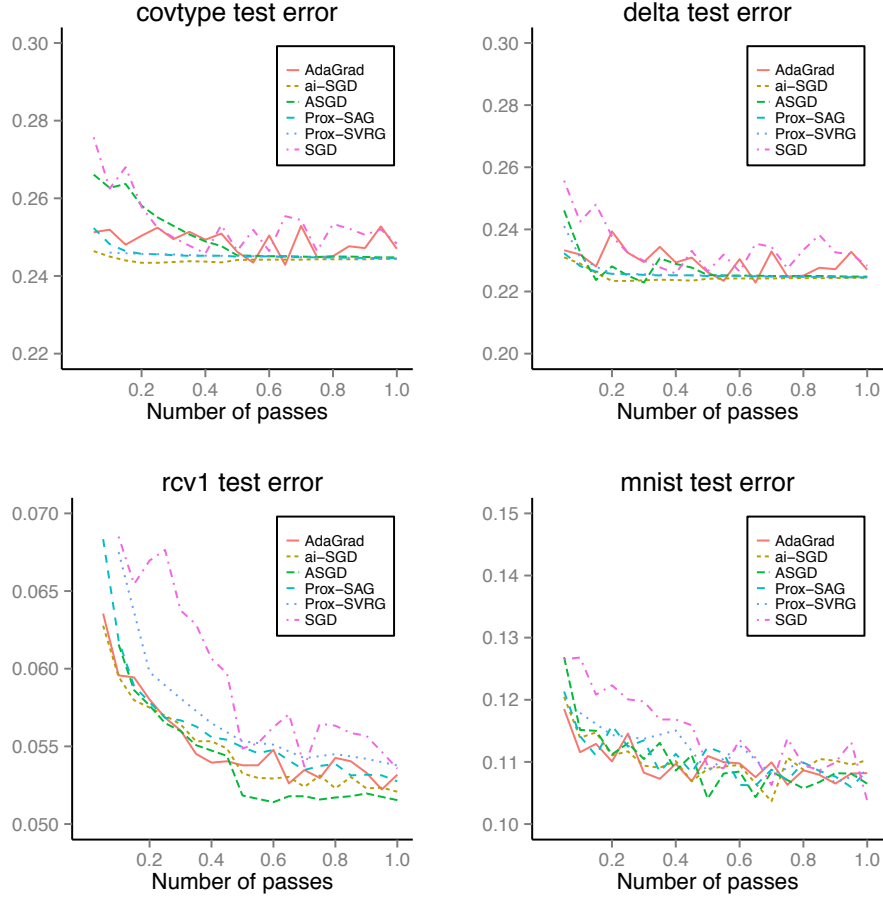


FIG 4. Large scale linear classification with log loss on four data sets. Each plot indicates the test error of various stochastic gradient methods over a single pass of the data.

mal distribution as $Y_n | X_n \sim \mathcal{N}(X_n^\top \theta_*, 1)$, for $n = 1, \dots, N$. 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 $n = 1, \dots, N$, and keep the average $\bar{\theta}_n$. In Figure 5, we plot $(\theta_n - \theta_*)^\top H(\theta_n - \theta_*)$ 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

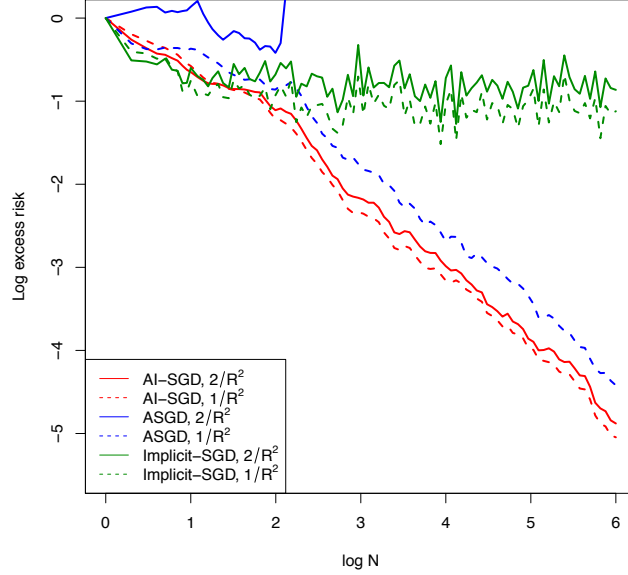


FIG 5. Loss of averaged implicit SGD, averaged explicit SGD, and plain implicit SGD in Eq. (4) ($C_n = I$), on simulated multivariate normal data with $N = 1e6$ observations $p = 20$ features. The plot shows that averaged implicit SGD is stable regardless of the specification of the learning rate γ and without sacrificing performance. In contrast, explicit averaged SGD is very sensitive to misspecification of the learning rate.

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 on-line 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.