Statistical analysis of stochastic gradient methods for generalized linear models (extended manuscript with proofs)

Abstract

We study the statistical properties of stochastic gradient descent (SGD) using explicit and *implicit* updates for fitting generalized linear models (GLMs). Initially, we develop a computationally efficient algorithm to implement implicit SGD learning of GLMs. Next, we obtain exact formulas for the bias and variance of both updates which leads to two important observations on their comparative statistical properties. First, in small samples, the estimates from the implicit procedure are more biased than the estimates from the explicit one, but their empirical variance is smaller and they are more robust to learning rate misspecification. Second, the two procedures are statistically identical in the limit: they are both unbiased, converge at the same rate and have the same asymptotic variance. Our set of experiments confirm our theory and more broadly suggest that the implicit procedure can be a competitive choice for fitting large-scale models, especially when robustness is a concern.

1. Introduction

Stochastic gradient descent (SGD) is a stochastic approximation (SA) method. Assume a random variable $y \in \mathbb{R}$ as the outcome of interest controlled by a parameter $\theta \in \mathbb{R}$ with regression function $M(\theta) = \mathbb{E}\left(y|\, \theta\right)$, and consider the problem of finding θ^* such that $M(\theta^*) = \mathbb{E}\left(y|\, \theta^*\right) = 0$. The classical SA procedure (Robbins & Monro, 1951) maintains an estimate θ_{n-1} of θ^* at each iteration n and then obtains a sample y_n (e.g. through an experiment) such that $\mathbb{E}\left(y_n\right) = M(\theta_{n-1})$. The estimate is then updated to $\theta_n = \theta_{n-1} - a_n y_n$. The scalar $a_n > 0$ is the *learning rate* and should decay to zero but not too fast in order to guarantee consistency of the method even when the analytic form of $M(\theta)$ is not known. Although initially applied in experimental design, the SA procedure was soon adapted for statistical estimation. Sakrison (Sakrison, 1965) assumed $y_n \sim f(y_n; \theta^*)$ i.e., that observations are drawn independently from a statistical model with unknown fixed parameter θ^*. Sakrison's recursive estimation procedure was defined using the update $\theta_n = \theta_{n-1} + a_n \ell'(\theta_{n-1}; y_n)$, where $\ell(\theta; y_n) = \log f(y_n; \theta)$ is the log-likelihood of θ given observation y_n. Under certain regularity and monotonicity conditions, $\mathbb{E}\left(\ell'(\theta^*; y_n)\right) = 0$ and so according to SA theory, the estimates θ_n converge to the real parameter θ^* with possibly optimal asymptotic efficiency (Anbar, 1973; Fabian, 1973). In recent years, Sakrison's recursive estimation method has become known as stochastic gradient descent (SGD). Further, note that the aforementioned SGD update is *explicit* i.e., θ_n can be calculated immediately from θ_{n-1} and the data at the n-th iteration. For the rest of this paper, we will refer to this procedure as "SGD with explicit updates" or *standard SGD* for short. Despite the theoretical guarantees, standard SGD is generally not robust to learning rate misspecification or input noise. Recursive procedures that aim to control the size of updates have thus been proposed such as,

$$\theta_n = \arg\min_{\theta} \left\{D(\theta, \theta_{n-1}) - a_n \ell(\theta; y_n)\right\} \tag{1}$$

in which $\ell(\cdot)$ is the log-likelihood as before and $D(\cdot, \cdot)$ is some distance function. Minimizing (1) yields updates of the form $\theta_n = g(\theta_n; \theta_{n-1}, y_n, a_n)$, that are called *implicit* since the future estimate θ_n appears in both sides of the equation. We will refer to procedure (1) as "SGD with implicit updates"[1] or *implicit SGD* for short.

Historically, the duo of explicit-implicit updates originate from the numerical methods invented by Euler (ca. 1770) for approximating solutions of ordinary differential equations (Hoffman & Frankel, 2001). However, the normalized least mean squares (NLMS) filter (Nagumo & Noda, 1967) was, arguably, the first statistical model that used an implicit update as in Equation (1) and was shown to be consistent and robust to excess input noise (Slock, 1993).[2]

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] Thus, we still regard procedure (1) to be a SGD procedure because the gradient is calculated at one data point y_n at a time and, hence, it is stochastic.

[2] In the NLMS algorithm (Nagumo & Noda, 1967), the multivariate update has the form $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + (a + b||\boldsymbol{x}_n||^2)^{-1}(y_n - \boldsymbol{x}_n^\intercal \theta_{n-1})\boldsymbol{x}_n, a, b > 0$ which can be written in the form of (1) for which $D(\cdot, \cdot)$ is the usual L_2 norm and the log-likelihood is that of a linear normal model.

Since then, several online learning models have been using implicit updates in various forms. For example, mirror-descent and projected subgradient algorithms (Nemirovski, 1983; Beck & Teboulle, 2003) and variants such as FOBOS (Duchi & Singer, 2009) include updates that can be written in the form of Equation (1). The regret of such algorithms using implicit updates and $D(\cdot, \cdot)$ a Bregman divergence has been shown to be comparable to standard SGD bounds (Kivinen & Warmuth, 1995; Kivinen et al., 2006; Kulis & Bartlett, 2010) and their robustness has been proven useful in a wide range of modern machine learning problems (Nemirovski et al., 2009; Kulis & Bartlett, 2010; Schuurmans & Caelli, 2007).

However, the statistical properties of SGD methods, either implicit or standard, remain not well-understood. In this paper, we perform a statistical analysis of the implicit method vis-à-vis with a standard SGD counterpart in the family of generalized linear models (GLMs). Our main contributions are the following:

(i) We adapt the classical SA procedure (Robbins & Monro, 1951) and the proof therein to formalize its implicit counterpart and show that the method is consistent in quadratic mean (Theorem 2.1). Next, we focus on the problem of online estimation of GLMs and provide a computationally efficient algorithm for applying implicit updates (Algorithm 1).

(ii) We derive formulas for the asymptotic bias of the implicit and standard SGD procedures (Theorem 4.1). We show that the implicit procedure converges slower (in general) but asymptotically at the same rate as the standard one. Furthermore, we derive exact formulas for the asymptotic variance of both procedures (Theorem 4.2) and, thus, show that they have the same asymptotic efficiency.

(iii) We show that the implicit method is unconditionally stable under any specification of the learning rate, whereas standard SGD can deviate arbitrarily when the learning rate is misspecified (Lemma 4.2).

For clarity of exposition, we omit most proofs from the current document and make them all available online in the unpublished, full version of this paper. [3]

2. Implicit stochastic approximation

We first introduce a general definition of the implicit SA procedure by adapting the original work of (Robbins & Monro, 1951). Assume a function $M : \mathbb{R} \to \mathbb{R}$ for which we wish to estimate the zero $M(\theta^*) = 0$. Starting from some $\theta_0 \in \mathbb{R}$, we update our estimates at iteration n according to observed data $y_n \in \mathbb{R}$, a learning rate $a_n \in \mathbb{R}^+$ and the following rule:

$$\theta_n = \theta_{n-1} - a_n y_n \tag{2}$$

Equation (2) defines the *implicit stochastic approximation* procedure under the following assumptions:

Assumption 2.1. $a_n > 0$, $\sum a_n^2 < \infty$, $\sum a_n = \infty$.

Assumption 2.2. *The random variable y_n has a distribution that depends on θ_n such that $\mathbb{E}\left(y_n | \theta_n\right) = M(\theta_n)$. Furthermore, it is bounded such that $P(|y_n| < C) = 1$ for some constant $C > 0$.*

Assumption 2.3. *The function $M(x)$ is non-decreasing and differentiable. Furthermore, $M'(\theta^*) > 0$.*

Note that only Assumption (2.2) differentiates between the standard SA procedure and the implicit one. Rather counter-intuitively, the observation y_n is considered a sample from the distribution of the *future* update θ_n.[4] Clearly, one may need to know the form of that distribution in order to perform the update. The following theorem establishes that the implicit SA procedure converges in quadratic mean, just like the standard SA counterpart.[5].

Theorem 2.1. *Suppose that assumptions (2.1)-(2.3) hold. Then, the implicit SA procedure (2) converges in quadratic mean i.e.,*

$$\mathbb{E}\left(\theta_n - \theta^*\right)^2 \to 0 \text{ as } n \to \infty \tag{3}$$

The proof is an adaptation of the original proof from (Robbins & Monro, 1951) and is given in the full version of this paper.

3. Preliminaries

We now focus on the family of GLMs (Nelder & Wedderburn, 1972). Let $y \in \mathbb{R}$ denote the outcome of interest, $\boldsymbol{\theta^*} \in \mathbb{R}^p$ be the vector of unknown model parameters and

[3] The full version of the paper together with the accompanying source code and documentation can be found at the following location: http://www.people.fas.harvard.edu/~ptoulis/harvard-homepage/implicit-sgd.html.

[4] Originally, the SA procedure was created for experimental design. Assume that y is a random response to a drug at a dose level θ. Assume θ^* to be the dose level in which the response will have no effects i.e., $\mathbb{E}\left(y | \theta^*\right) = 0$. Then the SA procedure describes a succession of dose levels θ_n, and thus a series of experiments with outcomes y_n, that converges to θ^*. This is not possible with the implicit procedure since y_n does not depend on θ_{n-1} but on θ_n, i.e. the future dose level!

[5] Note that the convergence of implicit update procedures for linear classifiers has already been shown in the literature (Cesa-Bianchi, 2006). However, Theorem 2.1 formalizes the implicit procedure under the general stochastic approximation framework which is broader in scope than linear classifiers.

$\boldsymbol{x} \in \mathbb{R}^p$ denote a vector of features. In a GLM, we assume that the outcome y follows some distribution in the exponential family as follows

$$y|\boldsymbol{x} \sim \exp\left(\frac{\eta y - b(\eta)}{\psi}\right) c(y, \psi) \text{ where } \eta = \boldsymbol{\theta}^{*\intercal}\boldsymbol{x} \quad (4)$$

The quantity η is the *linear predictor*, the scalar $\psi > 0$ is the *dispersion parameter* as it affects the variance of the outcome, and $b(\cdot), c(\cdot, \cdot)$ are appropriate real-valued functions. Equation (4) is known as the *canonical form* because the linear predictor appears as a coefficient of y in the density function. Furthermore, the expected value of the outcome is given by the *link function* $g(\cdot)$ of the model i.e.,

$$g\left(\mathbb{E}\left(y|\,\boldsymbol{x}\right)\right) = \boldsymbol{\theta}^{*\intercal}\boldsymbol{x} = \eta \quad (5)$$

The inverse of the link function $h = g^{-1}$ is the *transfer function* of the GLM model so that $\mathbb{E}\left(y|\,\boldsymbol{x}\right) = h(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x})$. We will assume GLMs in the canonical form with a monotone link function which we will refer to as *canonical GLMs*. This family is very broad and widely-applicable as it contains models such as the linear normal model, logistic regression, Poisson regression and so on. To illustrate our notation, in logistic regression we assume $P(y = 1|\boldsymbol{x}) = p^y(1-p)^{1-y}$ where p is a function of $\boldsymbol{x}$. This can be written in the form of Equation (4) with $\eta = \log(p/(1-p))$, $b(\eta) = \log(1 + e^{\eta})$, $\psi = 1$ and $c(y, \psi) = 1$. We know that $\mathbb{E}\left(y|\,\boldsymbol{x}\right) = p = \exp(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x})(1 + \exp(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x}))^{-1}$ and so the link function $g(\cdot)$ is the logit function $g(u) = \log(u/(1-u))$ and the transfer function is the logistic i.e., $h(u) = e^u(1+e^u)^{-1}$. Table 1 summarizes the link/transfer functions for the three aforementioned models. The results in the proposition that follows will be useful for the rest of our analysis. As these are standard results in the theory of GLMs (Nelder & Wedderburn, 1972), we just give short proofs in the full version of this paper.

model	$g(u)$	$h(u)$
Normal	u	u
Logistic	$\log(\frac{u}{1-u})$	$e^u(1+e^u)^{-1}$
Poisson	$\log(u)$	e^u

Table 1. Three well-known canonical GLMs.

Proposition 3.1. *Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be the true parameter vector of a canonical GLM model with outcome $y \in \mathbb{R}$, and $\boldsymbol{x} \in \mathbb{R}^p$ be some feature vector. Then,*

(a) $\mathbb{E}\left(y|\,\boldsymbol{x}\right) = h(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x}) = b'(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x})$

(b) $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = \psi h'(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x}) = b''(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x})$

(c) $\nabla\ell(\boldsymbol{\theta}; y, \boldsymbol{x}) = \frac{1}{\psi}\left(y - h(\boldsymbol{\theta}^{\intercal}\boldsymbol{x})\right)\boldsymbol{x}$

(d) $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -E\left(\nabla\nabla\ell(\boldsymbol{\theta}; y, \boldsymbol{x})\right) = \frac{1}{\psi}E\left(h'(\boldsymbol{\theta}^{\intercal}\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\intercal}\right)$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is an arbitrary vector.

3.1. Online learning of GLMs using SGD

In this paper, we assume the task of learning online the unknown parameter vector $\boldsymbol{\theta}^*$ of a GLM model (4). At every iteration indexed by $n = 1, 2, \cdots$ a new feature vector, denoted by $\boldsymbol{x}_n$, is sampled independently from a fixed and *known* distribution. Given $\boldsymbol{x}_n$, the outcome y_n is sampled according to a canonical GLM as in (4). Upon observing $(y_n, \boldsymbol{x}_n)$, we update our estimate of $\boldsymbol{\theta}^*$ from $\boldsymbol{\theta}_{n-1}$ to $\boldsymbol{\theta}_n$. The initial estimate $\boldsymbol{\theta}_0$ has been set a priori to a reasonable value. We now proceed to define the implicit and the standard SGD procedures for learning GLMs. The exact requirements for the learning rate a_n will be set in Assumptions (4.1).

Definition 3.1. *The standard SGD learning procedure for a canonical GLM is defined as:*

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n\left(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n)\right)\boldsymbol{x}_n \quad (6)$$

After n steps of procedure (6)*, the vector $\boldsymbol{\theta}_n^{\mathrm{sgd}}$ is the standard SGD estimator of $\boldsymbol{\theta}^*$.*

Definition 3.2. *The implicit SGD learning procedure for a canonical GLM is defined as:*

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n\left(y_n - h(\boldsymbol{\theta}_n^{\intercal}\boldsymbol{x}_n)\right)\boldsymbol{x}_n \quad (7)$$

After n steps of procedure (7)*, the vector $\boldsymbol{\theta}_n^{\mathrm{im}}$ is the implicit SGD estimator of $\boldsymbol{\theta}^*$.*

Discussion. First, note that we omit the term of $(1/\psi)$ of the log-likelihood gradient (see Proposition (3.1)-(c)) since it can be factored into a_n. Second, we clarify that Definitions 6 and 7 correspond to "baseline" definitions of the two learning procedures. Especially for standard SGD, there has been significant work in improving the performance of the procedure. These methods include averaging of the updates to speed up convergence (Polyak & Juditsky, 1992), approximating second-order information as in SGD-QN (Bordes et al., 2009), using adaptive learning rates as in AdaGrad (Duchi et al., 2011; Schaul et al., 2012) or variance reduction methods (Johnson & Zhang, 2013; Roux et al., 2012). So far, the implicit procedure has received disproportionately less attention, however it is reasonable to expect that similar methods could be employed there as well. In fact, our subsequent analysis suggests that the implicit procedure, being less sensitive to learning rate specification, is likely to be more amenable to performance improvements.

3.2. Efficient implicit updates for canonical GLMs

The implicit equation (7) cannot be solved in general because the form of $h(\cdot)$ can be arbitrary. Furthermore, in a multi-dimensional setting, this would require the solution of multiple simultaneous equations. However, it has already been noted that line search methods can be employed

Algorithm 1 Implicit learning of canonical GLMs

```
1: for all $n \in \{1, 2, \cdots\}$ do
2:     $r_n \leftarrow a_n \left(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n)\right)$
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:     $\xi_n = a_n \left[y_n - h\left(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + ||\boldsymbol{x}_n||^2\xi_n\right)\right], \xi_n \in B_n$
8:     $\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_{n-1} + \xi_n \boldsymbol{x}_n$
9: end for
```

to implement general implicit updates (Kivinen et al., 2006; Kulis & Bartlett, 2010). Here, we show that the special structure of the log-likelihood gradient of the GLMs (see Proposition 3.1-(c)) can be exploited in order to efficiently compute the implicit updates. Algorithm 1 reduces equation (7) to a one-dimensional implicit equation which can be solved efficiently, since narrow search bounds can be derived by using the monotone property of the transfer function in canonical GLMs.

Lemma 3.1. *Algorithm 1 computes estimates $\boldsymbol{\theta}_n$ that are identical to the estimates of the implicit procedure* (7)*.*

Proof. We first show that $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \xi_n \boldsymbol{x}_n$ is the correct update for the implicit procedure, where ξ_n is computed in Step 7 of Algorithm 1. We multiply with $\boldsymbol{x}_n$ on both sides of (7) to get,

$$\boldsymbol{\theta}_n^{\intercal}\boldsymbol{x}_n = \boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + a_n\left(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n)\right)\boldsymbol{x}_n^{\intercal}\boldsymbol{x}_n$$

and we apply $h(\cdot)$ on both sides to further obtain,

$$h(\boldsymbol{\theta}_n^{\intercal}\boldsymbol{x}_n) = h\left(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + a_n(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n))||\boldsymbol{x}_n||^2\right)$$

Setting $\xi_n = a_n(y_n - h(\boldsymbol{\theta}_n^{\intercal}\boldsymbol{x}_n))$, we can rewrite the above equation as

$$h(\boldsymbol{\theta}_n^{\intercal}\boldsymbol{x}_n) = h\left(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + \xi_n||\boldsymbol{x}_n||^2\right) \tag{8}$$

It also holds that $h(\boldsymbol{\theta}_n^{\intercal}\boldsymbol{x}_n) = y_n - \xi_n/a_n$ and so Equation (8) now becomes,

$$y_n - \xi_n/a_n = h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + \xi_n||\boldsymbol{x}_n||^2)$$

Solving for ξ_n we finally get the one-dimensional implicit equation,

$$\xi_n = a_n(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + \xi_n||\boldsymbol{x}_n||^2)) \tag{9}$$

By the definition of ξ_n and the implicit procedure (7), we have that $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \xi_n\boldsymbol{x}_n$.

Next we show why the bounds B_n in Algorithm 1 are correct. Let $m(u) = a_n\left(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n + u||\boldsymbol{x}_n||^2)\right)$ and let $l(u) = u$ be the straight line. We wish to find the fixed point ξ_n such that $m(\xi_n) = l(\xi_n)$. Since $m(u)$ is monotone decreasing and $l(u)$ is monotone increasing and both functions are continuous in $\mathbb{R}$, the intersection point is unique. The sign of ξ_n depends on where $m(\xi_n)$ crosses the y-axis i.e., $m(0) = a_n\left(y_n - h(\boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n)\right) \equiv r_n$. If $r_n > 0$ then $\xi_n > 0$. Furthermore, since $l(u)$ is increasing, $l(r_n) > l(\xi_n) \Rightarrow \xi_n < r_n$, and thus $[0, r_n]$ is a search interval for ξ_n. Similarly, if $r_n < 0$ then $\xi_n < 0$ and $\xi_n > r_n$.

Note that more restrictive bounds might be available. For example, if $r_n > 0$ we know that ξ_n has to be smaller than the point u_0 where $m(u)$ crosses the x-axis, i.e. $m(u_0) = 0$. Through standard algebra we can obtain that $u_0 = (g(y_n) - \boldsymbol{\theta}_{n-1}^{\intercal}\boldsymbol{x}_n)/||\boldsymbol{x}_n||^2$. In this case, a better bound for ξ_n is $[0, \min(u_0, r_n)]$, while a similar argument works also if $r_n < 0$. Significant improvements are expected in models where $g(u) = o(u)$ such as the Poisson regression model. In this case, instead of searching in an interval $[0, r_n]$, the algorithm could search in $[0, \log r_n]$. □

4. Statistical analysis

For an estimate $\boldsymbol{\theta}_n$ of $\boldsymbol{\theta}^*$, let $\boldsymbol{\mu}_n = \mathbb{E}(\boldsymbol{\theta}_n)$ and $\mathbf{V}_n = \mathrm{Var}(\boldsymbol{\theta}_n)$. Denote the bias of $\boldsymbol{\theta}_n$ with $\boldsymbol{b}_n = \boldsymbol{\mu}_n - \boldsymbol{\theta}^*$. We will use a superscript to denote the specific procedure under study, thus $\boldsymbol{\mu}_n^{\mathrm{sgd}}$ and $\boldsymbol{\mu}_n^{\mathrm{im}}$ denote the means of the estimates from the standard SGD and the implicit procedure respectively, and so on. Also, let $\boldsymbol{z}_n(\boldsymbol{\theta};\boldsymbol{x}) = h(\boldsymbol{\theta}^{\intercal}\boldsymbol{x}_n)\boldsymbol{x}_n$, $\boldsymbol{r}(\boldsymbol{\theta}) = \mathbb{E}\left(\boldsymbol{z}_n(\boldsymbol{\theta};\boldsymbol{x})|\,\boldsymbol{\theta}\right)$ and let $\mathbf{D}_r(\cdot)$ be the Jacobian of $\boldsymbol{r}(\cdot)$. If a nonnegative series $\{\gamma_n\}$ satisfies $\sum_{i=1}^{\infty} a_i\gamma_i < \infty$, we will call it *$a_n$-convergent*. We also write $\mathcal{C}(a_n)$ for an arbitrary a_n-convergent series. Last, throughout this paper the notation $||\cdot||$ for a vector or a matrix argument denotes the L_2 norm.

Assumption 4.1. *(a) Let $a_n > 0$ be a decreasing sequence of numbers such that $\sum a_n = \infty$, $\sum a_n^2 < \infty$. Furthermore, $a_{n-1}/a_n = 1 + (1/\alpha)a_n + \mathcal{O}(a_n^2)$, for some $\alpha > 0$.*

(b) For sufficiently large n, in the neighborhood of $\boldsymbol{\theta}^*$, make the approximation

$$\mathbf{r}(\boldsymbol{\theta}_n) = \mathbf{r}(\boldsymbol{\theta}^*) + \mathbf{D}_r(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*) + \boldsymbol{\epsilon}_n$$

and assume convergence $\boldsymbol{\theta}_n \to \boldsymbol{\theta}^*$ such that the series

(i) $||\boldsymbol{\epsilon}_n||$ *and* $(1/a_n)||\mathrm{Cov}(\boldsymbol{\theta}_n, \boldsymbol{\epsilon}_n)||$

(ii) $||\mathrm{Var}(\boldsymbol{z}_n(\boldsymbol{\theta}_n;\boldsymbol{x}) - \boldsymbol{z}_n(\boldsymbol{\theta}^*;\boldsymbol{x}))||$

(iii) $||\mathrm{Var}(\boldsymbol{z}_n(\boldsymbol{\theta}_n;\boldsymbol{x})) - \mathrm{Var}(\boldsymbol{z}_n(\boldsymbol{\theta}^*;\boldsymbol{x}))||$

are all a_n-convergent.

The first part of the assumption is essential for convergence of the stochastic approximation procedure. One schedule that satisfies such assumptions is $a_n = \alpha(\beta + n)^{-1}$, $\alpha, \beta > 0$. Note also that Assumption 4.1 does not cover rates of the form $(\beta + n)^{-c}$ for $c < 1$; this will be part of future work. Part (b) of our Assumption 4.1 puts weak constraints on convergence to $\boldsymbol{\theta}^*$. For example, if $a_n \propto n^{-1}$, then the series of Assumption 4.1 (i)-(iii) are allowed to be of the form $n^{-\epsilon}$ for any $\epsilon > 0$. Approximation assumptions like Assumption 4.1 are generally necessary in order to derive *exact* asymptotic variance formulas. For example, the classical derivation of the variance of the MLE relies on a second-order approximation (Fisher, 1922). Classical results on the variance and/or normality of the SA estimators also rely on linearly-bounded derivatives (Sacks, 1958; Fabian, 1968). In a similar statistical analysis of the standard SGD iterates, Murata (1998) relies on a complete second order approximation of the loss function (see (Murata, 1998), Equation 2.4). In machine learning, the assumptions are generally weaker since the goal is only to bound the regret of the online procedure. However, assumptions on *global* bounds on the loss-gradients (Zinkevich, 2003) or the distance $||\boldsymbol{\theta}_n - \boldsymbol{\theta}^*||$ (Kulis & Bartlett, 2010) are common.

Lemma 4.1. *Let the sequence a_n satisfy part (a) of Assumption 4.1 and consider the following matrix recursions,*

$$\boldsymbol{X}_n = (\boldsymbol{I} - a_n\boldsymbol{B}_n)\boldsymbol{X}_{n-1} + a_n(\boldsymbol{C} + \boldsymbol{D}_n) \tag{10}$$

$$\boldsymbol{Y}_n = (\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1}\left[\boldsymbol{Y}_{n-1} + a_n(\boldsymbol{C} + \boldsymbol{D}_n)\right] \tag{11}$$

such that,

(a) $\boldsymbol{B}_n \to \boldsymbol{B} > 0$, $||\boldsymbol{B}_n - \boldsymbol{B}_{n-1}|| = \mathcal{O}(a_n^2)$, *and*

(b) $\sum_{i=1}^{\infty} a_i||\boldsymbol{D}_i|| < \infty$ *i.e., $||D_n||$ is a_n-convergent.*

Then, both recursions approximate the matrix $\boldsymbol{B}^{-1}\boldsymbol{C}$ i.e.,

$$||\boldsymbol{X}_n - \boldsymbol{B}^{-1}\boldsymbol{C}|| \to 0 \text{ and } ||\boldsymbol{Y}_n - \boldsymbol{B}^{-1}\boldsymbol{C}|| \to 0 \tag{12}$$

Corollary 4.1. *Consider the following matrix recursions,*

$$\boldsymbol{X}_n = (\boldsymbol{I} - a_n\boldsymbol{B}_n)\boldsymbol{X}_{n-1} + a_n^2(\boldsymbol{C} + \boldsymbol{D}_n) \tag{13}$$

$$\boldsymbol{Y}_n = (\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1}\left[\boldsymbol{Y}_{n-1} + a_n^2(\boldsymbol{C} + \boldsymbol{D}_n)\right] \tag{14}$$

where $a_n, \boldsymbol{B}_n, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}_n$ satisfy the assumptions of Lemma 4.1. Then, $\boldsymbol{X}_n \to \boldsymbol{0}$ and $\boldsymbol{Y}_n \to \boldsymbol{0}$. Furthermore, if the matrix $(\boldsymbol{B} - \boldsymbol{I}/\alpha)$ is positive-definite,

$$(1/a_n)\boldsymbol{X}_n \to (\boldsymbol{B} - \boldsymbol{I}/\alpha)^{-1}\boldsymbol{C} \tag{15}$$

$$(1/a_n)\boldsymbol{Y}_n \to (\boldsymbol{B} - \boldsymbol{I}/\alpha)^{-1}\boldsymbol{C} \tag{16}$$

Proof. Both $\boldsymbol{X}_n, \boldsymbol{Y}_n \to \boldsymbol{0}$ by direction application of Lemma (4.1). Let $\tilde{\boldsymbol{X}}_n = (1/a_n)\boldsymbol{X}_n$. Divide (13) by a_n to obtain

$$\tilde{\boldsymbol{X}}_n = (\boldsymbol{I} - a_n\boldsymbol{B}_n)\,\tilde{\boldsymbol{X}}_{n-1}\frac{a_{n-1}}{a_n} + a_n(\boldsymbol{C} + \boldsymbol{D}_n) \tag{17}$$

Using Assumption (4.1) (a), we can rewrite as (18) as

$$\tilde{\boldsymbol{X}}_n = (\boldsymbol{I} - a_n\boldsymbol{\Gamma}_n)\,\tilde{\boldsymbol{X}}_{n-1} + a_n(\boldsymbol{C} + \boldsymbol{D}_n) \tag{18}$$

where $\boldsymbol{\Gamma}_n = \boldsymbol{B}_n - \boldsymbol{I}/\alpha + \mathcal{O}(a_n)$. Then, in the limit, $\boldsymbol{\Gamma}_n \to \boldsymbol{B} - \boldsymbol{I}/\alpha > 0$. Furthermore, $||\boldsymbol{\Gamma}_i - \boldsymbol{\Gamma}_{i-1}|| = ||\boldsymbol{B}_i - \boldsymbol{B}_{i-1} + \mathcal{O}(a_n^2)|| = \mathcal{O}(a_n^2)$. Thus, we can apply Lemma (4.1) to conclude that $(1/a_n)\boldsymbol{X}_n = \tilde{\boldsymbol{X}}_n \to (\boldsymbol{B} - \boldsymbol{I}/\alpha)^{-1}\boldsymbol{C}$. The proof for $\boldsymbol{Y}_n$ follows exactly the same reasoning, noting that

$$(\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1}(a_{n-1}/a_n) = (\boldsymbol{I} + a_n\boldsymbol{\Gamma}_n)^{-1}$$

where $\boldsymbol{\Gamma}_n = \boldsymbol{B}_n - \boldsymbol{I}/\alpha + \mathcal{O}(a_n^2)$.

□

4.1. Asymptotic bias

Taking expectations on both sides of updates (6) and (7), and using Assumption 4.1 and Lemma 4.1, we obtain the asymptotic unbiasedness of the SGD estimators through the following theorem. The complete proof is given in the full version of the paper.

Theorem 4.1. *Under Assumption 4.1, the asymptotic bias of the standard SGD estimator satisfies,*

$$\boldsymbol{b}_n^{\text{sgd}} = (\boldsymbol{I} - a_n\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))\,\boldsymbol{b}_{n-1}^{\text{sgd}} + \mathcal{C}(a_n) \tag{19}$$

The asymptotic bias of the implicit SGD estimator satisfies,

$$\boldsymbol{b}_n^{\text{im}} = (\boldsymbol{I} + a_n\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))^{-1}\Big[\boldsymbol{b}_{n-1}^{\text{im}} + \mathcal{C}(a_n)\Big] \tag{20}$$

Both methods are asymptotically unbiased i.e., $\boldsymbol{\mu}_n^{\text{sgd}} \to \boldsymbol{\theta}^*$ and $\boldsymbol{\mu}_n^{\text{im}} \to \boldsymbol{\theta}^*$.

Proof. First, we show that $\mathbf{D}_r(\boldsymbol{\theta}) = \psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$. Note that the (i, j) element of the Jacobian is $\frac{\partial r_i}{\partial \theta_j}$. Denote by x_{ni} the i-th element of $\boldsymbol{x}_n$, and note that $r_i(\cdot)$ is equal to $\mathbb{E}\left(h(\boldsymbol{\theta}^\intercal\boldsymbol{x}_n)x_{ni}\right)$. Assuming differentiability of expectation (holds in the general canonical GLM models), we obtain $\frac{\partial r_i}{\partial \theta_j} = \psi\mathbb{E}\left(h'(\boldsymbol{\theta}^\intercal\boldsymbol{x}_n)x_{ni}x_{nj}\right)$. Therefore, by Proposition (3.1), it holds

$$\mathbf{D}_r(\boldsymbol{\theta}) = \mathbb{E}\left(h'(\boldsymbol{\theta}^\intercal\boldsymbol{x}_n)\boldsymbol{x}_n\boldsymbol{x}_n^\intercal\right) = \psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \tag{21}$$

Recall that the SGD procedure is:

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \alpha_n(y_n - h(\boldsymbol{\theta}_{n-1}^\intercal\boldsymbol{x}_n))\boldsymbol{x}_n$$

Taking expectations on both sides, we have

$$\begin{aligned}\boldsymbol{\mu}_n^{\text{sgd}} &= \boldsymbol{\mu}_{n-1}^{\text{sgd}} + \alpha_n\mathbb{E}\left(y_n\boldsymbol{x}_n\right) - \alpha_n\mathbb{E}\left(h(\boldsymbol{\theta}_{n-1}^\intercal\boldsymbol{x}_n)\boldsymbol{x}_n\right) \\ &= \boldsymbol{\mu}_{n-1}^{\text{sgd}} + a_n\mathbb{E}\left(\boldsymbol{r}(\boldsymbol{\theta}^*) - \boldsymbol{r}(\boldsymbol{\theta}_{n-1})\right) \end{aligned} \tag{22}$$

By Assumption 4.1-(b), we have

$$\boldsymbol{r}(\boldsymbol{\theta}^*) - \boldsymbol{r}(\boldsymbol{\theta}_{n-1}) = -\mathbf{D}_r(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*) - \boldsymbol{\epsilon}_{n-1}$$

Subtracting $\boldsymbol{\theta}^*$ from both sides and using (21), gives

$$\boldsymbol{b}_n^{\mathrm{sgd}} = (\boldsymbol{I} - a_n \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))\boldsymbol{b}_{n-1}^{\mathrm{sgd}} - a_n \mathbb{E}\left(\boldsymbol{\epsilon}_{n-1}\right)$$

By Assumption (4.1)-(bi), $\sum_i a_i ||\boldsymbol{\epsilon}_n|| < \infty$, and this yields the desired result in Equation (19). Direct application of Lemma (4.1) yields $\boldsymbol{b}_n^{\mathrm{sgd}} \to \mathbf{0}$.

The implicit case is symmetrical:

$$\begin{aligned} \boldsymbol{\mu}_n^{\mathrm{im}} &= \boldsymbol{\mu}_{n-1}^{\mathrm{im}} + \alpha_n \mathbb{E}\left(\boldsymbol{r}(\boldsymbol{\theta}^*) - \boldsymbol{r}(\boldsymbol{\theta}_n)\right) \Rightarrow \\ \boldsymbol{b}_n^{\mathrm{im}} &= (\boldsymbol{I} + a_n \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))^{-1} \boldsymbol{b}_n^{\mathrm{im}} - a_n \boldsymbol{D}_n \end{aligned} \tag{23}$$

where we set $\boldsymbol{D}_n = (\boldsymbol{I} + a_n \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))^{-1} \mathbb{E}\left(\boldsymbol{\epsilon}_n\right) = \mathcal{O}(a_n)$, and so the result in (20) is established. Furthermore, by Lemma (4.1), we have $\boldsymbol{b}_n^{\mathrm{im}} \to \mathbf{0}$ as well. □

Note that Theorem 4.1 implies that the standard SGD procedure converges faster than the implicit one, since $||(\boldsymbol{I} - a_n \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))|| < ||(\boldsymbol{I} + a_n \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))^{-1}||$ for sufficiently large n. However, the rates become equal in the limit.

4.2. Asymptotic variance

Taking variances on both sides of updates (6) and (7), and using Assumption 4.1 and Corollary 4.1, we obtain the exact formula for the asymptotic variances of the SGD estimators through the following theorem. The complete proof is also given in the full version of the paper.

Theorem 4.2. *Under Assumption* (4.1) *and if the matrix* $(2\psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I}/\alpha)$ *is positive-definite, the asymptotic variance of the standard SGD estimator satisfies,*

$$(1/a_n)\mathbf{V}_n^{\mathrm{sgd}} \to \alpha \psi^2 \left(2\alpha\psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I}\right)^{-1} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \tag{24}$$

The asymptotic variance of the implicit SGD estimator satisfies,

$$(1/a_n)\mathbf{V}_n^{\mathrm{im}} \to \alpha \psi^2 \left(2\alpha\psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I}\right)^{-1} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \tag{25}$$

Therefore, both estimators have the same asymptotic efficiency.

Proof. We begin with the SGD procedure and omit the superscript for notational convenience. Start with Equation (6) and take variances on both sides:

$$\begin{aligned} \mathbf{V}_n = \mathbf{V}_{n-1} &+ a_n^2 \mathrm{Var}\left(y_n \boldsymbol{x}_n\right) + a_n^2 \mathrm{Var}\left(h(\boldsymbol{\theta}_{n-1}^{\intercal} \boldsymbol{x}_n)\boldsymbol{x}_n\right) \\ &+ 2a_n \mathrm{Cov}\left(\boldsymbol{\theta}_{n-1}, y_n \boldsymbol{x}_n\right) \\ &- 2a_n \mathrm{Cov}\left(\boldsymbol{\theta}_{n-1}, h(\boldsymbol{\theta}_{n-1}^{\intercal} \boldsymbol{x}_n)\boldsymbol{x}_n\right) \\ &- 2a_n^2 \mathrm{Cov}\left(y_n \boldsymbol{x}_n, h(\boldsymbol{\theta}_{n-1}^{\intercal} \boldsymbol{x}_n)\boldsymbol{x}_n\right) \end{aligned} \tag{26}$$

We proceed to simplify Equation (26) by computing all variance/covariance terms.

$$\begin{aligned} &\mathrm{Var}\left(y_n \boldsymbol{x}_n\right) = \psi^2 \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) + \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}^*; \boldsymbol{x})\right) \\ &\mathrm{Var}\left(h(\boldsymbol{\theta}_{n-1}^{\intercal} \boldsymbol{x}_n)\boldsymbol{x}_n\right) = \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}_{n-1}; \boldsymbol{x})\right) \\ &\mathrm{Cov}\left(\boldsymbol{\theta}_{n-1}, y_n \boldsymbol{x}_n\right) = 0 \\ &\mathrm{Cov}\left(\boldsymbol{\theta}_{n-1}, h(\boldsymbol{\theta}_{n-1}^{\intercal} \boldsymbol{x}_n)\boldsymbol{x}_n\right) = \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)\mathbf{V}_{n-1} + \boldsymbol{U}_{n-1} \\ &\mathrm{Cov}\left(y_n \boldsymbol{x}_n, h(\boldsymbol{\theta}_{n-1}^{\intercal} \boldsymbol{x}_n)\boldsymbol{x}_n\right) = \mathrm{Cov}\left(\boldsymbol{z}_n(\boldsymbol{\theta}^*; \boldsymbol{x}), \boldsymbol{z}_n(\boldsymbol{\theta}_{n-1}; \boldsymbol{x})\right) \end{aligned}$$

where we defined $\boldsymbol{U}_n = \mathrm{Cov}\left(\boldsymbol{\theta}_n, \boldsymbol{\epsilon}_n\right)$. We can now rewrite Equation (26) as

$$\mathbf{V}_n = \left(\boldsymbol{I} - a_n \boldsymbol{B}\right)\mathbf{V}_{n-1} + a_n^2(\psi^2 \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) + \boldsymbol{\Delta}_n) \tag{27}$$

where we have defined $\boldsymbol{B} = 2\psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$ and the remainder matrix

$$\begin{aligned} \boldsymbol{\Delta}_n =& \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}^*; \boldsymbol{x})\right) + \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}_{n-1}; \boldsymbol{x})\right) - \\ &- 2\mathrm{Cov}\left(\boldsymbol{z}_n(\boldsymbol{\theta}^*; \boldsymbol{x}), \boldsymbol{z}_n(\boldsymbol{\theta}_{n-1}; \boldsymbol{x})\right) - (2/a_n)\boldsymbol{U}_{n-1} \\ =& \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}_{n-1}; \boldsymbol{x}) - \boldsymbol{z}_n(\boldsymbol{\theta}^*; \boldsymbol{x})\right) - (2/a_n)\boldsymbol{U}_{n-1} \end{aligned}$$

By Assumptions (4.1)-(bi) and (bii) we establish that $\sum_i a_i ||\boldsymbol{\Delta}_n|| < \infty$. Therefore, we can directly apply Corollary (4.1) to obtain

$$(1/a_n)\mathbf{V}_n^{\mathrm{sgd}} \to \alpha \psi^2 \left(2\alpha\psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I}\right)^{-1} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$$

For the implicit updates, we follow the same approach. Taking the variance on both sides of the implicit procedure (7) yields

$$(\boldsymbol{I} + a_n \boldsymbol{B})\mathbf{V}_n = \mathbf{V}_{n-1} + a_n^2(\psi^2 \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) + \boldsymbol{\Delta}_n^{\mathrm{im}}) \tag{28}$$

where we set

$$\boldsymbol{\Delta}_n^{\mathrm{im}} = \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}^*; \boldsymbol{x})\right) - \mathrm{Var}\left(\boldsymbol{z}_n(\boldsymbol{\theta}_n; \boldsymbol{x})\right) - (2/a_n)\boldsymbol{U}_n$$

Thus, by Assumptions (4.1)-(bi) and (biii) we establish that $\sum_i a_i ||\boldsymbol{\Delta}_n^{\mathrm{im}}|| < \infty$ and using the approximation $(\boldsymbol{I} + a_n \boldsymbol{B})^{-1} = \boldsymbol{I} - a_n \boldsymbol{B} + \mathcal{O}(a_n)$, we can apply Corollary (4.1) on (28) to finally get

$$(1/a_n)\mathbf{V}_n^{\mathrm{im}} \to \alpha \psi^2 \left(2\alpha\psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I}\right)^{-1} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$$

□

Similar forms to the asymptotic variance of Theorem 4.2 have been discovered before. For example, assume a one-dimensional normal model ($p = 1$) where $\boldsymbol{x}_n = 1$, $a_n = \alpha/n$ and $y_n | \boldsymbol{x}_n \sim \mathcal{N}(\theta^*, \sigma^2)$. In our notation, $\psi = \sigma^2$, $h(u) = u$ and $\boldsymbol{\mathcal{I}}(\theta^*) = 1/\sigma^2$. Thus, the asymptotic variance in (24) becomes $nV_n^{\mathrm{sgd}} \to \alpha^2\sigma^2/(2\alpha - 1)$. This form of asymptotic variance, which by Theorem 4.2 holds for the implicit procedure as well, was first proved by Sacks (1958) and has since been rediscovered multiple times. [6]

We can also confirm that the asymptotic variance in Theorem 4.2 is larger than the variance of the maximum likelihood estimator (MLE) defined as $\boldsymbol{\theta}_n^{\text{mle}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$. Standard theory suggests that the MLE is *asymptotically optimal* as an estimator and that $\sqrt{n}\boldsymbol{\theta}_n^{\text{mle}}$ has variance $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}$ for large n. Let $c = \alpha\psi > 0$, then the variance of $\sqrt{n}\boldsymbol{\theta}_n^{\text{sgd}}$ and of $\sqrt{n}\boldsymbol{\theta}_n^{\text{im}}$ is given by $c^2(2c\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I})^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \geq \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}$ for any $c > 0$,[7] thus showing that both estimators (standard SGD and implicit) are not optimal. However, we can still utilize the variance formula to derive optimal learning rates. Note that the eigenvalues of the variance of both estimators are given by $c^2\lambda_i/(2c\lambda_i - 1)$ where λ_i are the eigenvalues of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$. This formula could then be used to pick optimal learning rates according to an appropriate criterion. For example, if we would like to minimize the trace of the SGD asymptotic variance then we should pick $\hat{c} = \arg\min_c \sum_i c^2\lambda_i/(2c\lambda_i - 1)$ and $\hat{\alpha} = \hat{c}/\psi$. Of course, the λ_i's are unknown in general and thus we would need to estimate them from data. Using our theory in order to develop optimal learning rates, especially for the implicit procedure, will be the focus of future work.

4.3. Stability

We simplify the bias recursions (19) and (20) by ignoring the remainder terms and by considering a simpler form as follows:

$$\begin{aligned} \boldsymbol{b}_n^{\text{sgd}} &= (\boldsymbol{I} - a_n\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))\boldsymbol{b}_{n-1}^{\text{sgd}} = \boldsymbol{P}_1^n\boldsymbol{b}_0 \\ \boldsymbol{b}_n^{\text{im}} &= (\boldsymbol{I} + a_n\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))^{-1}\boldsymbol{b}_{n-1}^{\text{im}} = \boldsymbol{Q}_1^n\boldsymbol{b}_0 \qquad (29) \end{aligned}$$

where $\boldsymbol{P}_1^n = \prod_{i=1}^n(\boldsymbol{I} - a_i\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))$, $\boldsymbol{Q}_1^n = \prod_{i=1}^n(\boldsymbol{I} + a_i\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))^{-1}$, and $\boldsymbol{b}_0$ denotes the initial bias from a common starting point $\boldsymbol{\theta}_0$. Thus, the simplified form actually describes the effect of the starting point $\boldsymbol{\theta}_0$ on the estimates $\boldsymbol{\theta}_n$ after n iterations. Also, let $\text{eig}(\boldsymbol{A})$ be the set of eigenvalues of a matrix $\boldsymbol{A}$, and let $\lambda_{(p)} = \max\{\text{eig}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))\}$, $\lambda_{(1)} = \min\{\text{eig}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))\}$ be the maximum and minimum eigenvalues of the Fisher information matrix respectively. Note that, $\boldsymbol{P}_1^n \to \boldsymbol{0}$ and $\boldsymbol{Q}_1^n \to \boldsymbol{0}$ (based on the proof of Lemma 4.1) and thus both methods are *asymptotically stable* i.e., both will converge, in theory, to the true parameter vector regardless of the starting point. However, we are interested in *deviations* of the standard and implicit SGD methods as captured here by the empirical bias. Based on the simplified bias equations (29), this information can be summarized by the eigenvalues of the matrices $\boldsymbol{P}_1^n$ and $\boldsymbol{Q}_1^n$ through the following lemma.

Lemma 4.2. *If $a_n = \alpha/n$ and $\alpha\psi\lambda_{(p)} > 1$, then the maximum possible eigenvalue of a matrix $\boldsymbol{P}_1^n$ is given by*

$$\max_{n>0} \max\{\text{eig}(\boldsymbol{P}_1^n)\} = \Theta(2^{\alpha\psi\lambda_{(p)}}/\sqrt{\alpha\psi\lambda_{(p)}}) \qquad (30)$$

For the implicit method,

$$\max_{n>0} \max\{\text{eig}(\boldsymbol{Q}_1^n)\} = \mathcal{O}(1) \qquad (31)$$

Lemma 4.2 shows that in the standard SGD procedure, the effect from the initial conditions can be amplified in an exponentially large way before fading out, if the learning rate is misspecified (i.e., if $\alpha > 1/\psi\lambda_{(p)}$). This sensitivity of the standard SGD procedure is well-known and requires problem-specific considerations to be avoided in practice. However, it is less well-known that the effects of the initial conditions *monotonically decrease* in the implicit method as shown in Equation (31). Rather remarkably, this robustness property of the implicit method holds under arbitrary misspecifications of the learning rate.

5. Experiments

We illustrate the different aspects of our theory on three separate sets of experiments. In Section 5.1, we work on a simple bivariate Poisson regression model and verify the variance asymptotics in Theorem 4.2, both analytically and in simulation. In Section 5.2, we compare convergence and stability of the standard and implicit SGD procedures on a larger Normal model. Last, in Section 5.3 we implement an implicit learning algorithm for SVM and compare with a standard SGD algorithm on a typical classification task.

5.1. Bivariate Poisson model

We first illustrate on a bivariate Poisson model which is simple enough to derive the relevant formulas analytically. We assume binary features such that, for any iteration n, $\boldsymbol{x}_n$ is either $(0,0)^\intercal$, $(1,0)^\intercal$ or $(0,1)^\intercal$ with probabilities 0.6, 0.2 and 0.2 respectively. We set $\boldsymbol{\theta}^* = (\theta_1, \theta_2)^\intercal$ for some θ_1, θ_2, and assume $y_n \sim \text{Poisson}(e^{\boldsymbol{\theta}^{*\intercal}\boldsymbol{x}_n})$. In our GLM notation, $p = 2$, $\psi = 1$ and $h(u) = e^u$. By Proposition 3.1 it easily follows that,

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) = \mathbb{E}\left(h'(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x}_n)\boldsymbol{x}_n\boldsymbol{x}_n^\intercal\right) = 0.2\begin{pmatrix} e^{\theta_1} & 0 \\ 0 & e^{\theta_2} \end{pmatrix}$$

We set $a_n = 10/3n$ which implies $\alpha = 10/3$ using Assumption 4.1. For notational convenience, let $c = (2\alpha)0.2 = 4/3$. Setting $\theta_1 = \log 2$ and $\theta_2 = \log 4$, the

[6] Sacks (1958) proved normality of θ_n^{sgd} with variance $(1/n)\alpha^2\sigma^2(2aM'(\theta^*) - 1)^{-1}$, under certain conditions on the regression function $M(\theta) = \mathbb{E}(y_n|\theta)$, but without requiring a normal distributional assumption of y_n. See also (Nemirovski et al., 2009), page 1578, for similar variance results but for general strongly-convex objective functions.

[7] To clarify, by $\boldsymbol{A} \geq \boldsymbol{B}$ we mean that $\boldsymbol{A} - \boldsymbol{B}$ is a nonnegative-definite matrix.

Table 2. Sample quantiles of $||\boldsymbol{\theta}^{\text{sgd}}_{20000} - \boldsymbol{\theta}^*||$ and $||\boldsymbol{\theta}^{\text{im}}_{20000} - \boldsymbol{\theta}^*||$. Values that are larger than `1e3` are marked with "*".

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

asymptotic variance $\boldsymbol{\Sigma} = \alpha(2\alpha\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I})^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$ in Theorem 4.2 is equal to,

$$\boldsymbol{\Sigma} = \frac{c}{2}\begin{pmatrix} \frac{e^{\theta_1}}{ce^{\theta_1}-1} & 0 \\ 0 & \frac{e^{\theta_2}}{ce^{\theta_2}-1} \end{pmatrix} = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.62 \end{pmatrix} \quad (32)$$

Next, we obtain 100 independent samples of $\boldsymbol{\theta}^{\text{sgd}}_N$ and $\boldsymbol{\theta}^{\text{im}}_N$ for $N = 20000$ iterations through the procedures defined in (6) and (7), and compute their empirical variances. We observe that the implicit estimates are particularly stable and have an empirical variance that satisfies,

$$(1/a_N)\widehat{\text{Var}}(\boldsymbol{\theta}^{\text{im}}_N) = \begin{pmatrix} 0.86 & -0.06 \\ -0.06 & 0.64 \end{pmatrix}$$

and is close to the theoretical value calculated in (32). In contrast, the standard SGD estimates are quite unstable and their L_2 distance to the true values $\boldsymbol{\theta}^*$ are orders of magnitude larger than the implicit ones (see Table 2 for sample quantiles). By Lemma 4.2, such deviations are expected for standard SGD because the largest eigenvalue of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$ is $\lambda_{(2)} = 0.8$ satisfying $\alpha\psi\lambda_{(2)} = 8/3 > 1$. Note however, that it is fairly straightforward to stabilize the standard SGD procedure in this problem, for example by modifying the learning rate to $a_n = \min\{0.15, 10/3n\}$. In general, when the optimization problem is well-understood, it is easy to determine the learning rate schedule that avoids out-of-band explicit updates; in practice, we are working with problems that are not so well-understood and determining the correct learning rate schedule may take substantial effort, especially in multi-dimensional settings. The implicit method eliminates this overhead: a wide range of learning rate schedules leads to convergence on all problems.

5.2. Multivariate Normal model

In our next experiment, we wish to validate our theory through a toy problem of normal linear regression following (Xu, 2011). We assume $\boldsymbol{\theta}^* = (1, 1, \cdots, 1)^\intercal \in \mathbb{R}^{20}$ to be the ground-truth (i.e., $p = 20$ parameters). At each iteration n, the feature vector $\boldsymbol{x}_n$ is sampled i.i.d. from a multivariate normal $\boldsymbol{x}_n \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{V}_x)$ for a fixed matrix $\boldsymbol{V}_x$.[8] The outcome y_n is then sampled from a normal as $y_n|\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\theta}^{*\intercal}\boldsymbol{x}_n, 1)$. For each procedure, i.e., standard and implicit SGD, we collect iterates $\boldsymbol{\theta}_n$ for $n = 1, 2, \cdots N$. We also repeat the procedure M times so that we finally have M samples of $\boldsymbol{\theta}^{\text{sgd}}_n$ and $\boldsymbol{\theta}^{\text{im}}_n$, similar to Section 5.1.

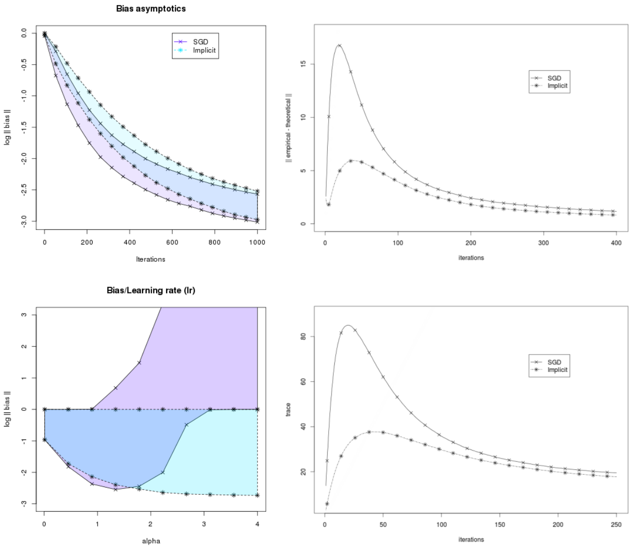

Figure 1. Standard SGD (dark shade, "x") and implicit SGD (light shade, "*") procedures on normal linear regression. The figure shows for each procedure, the (i) 2.5%/97.5% quantiles of log-bias over iterations (top-left) (ii) 2.5%/97.5% quantiles of log-bias over learning-rate scaling (bottom-left), (iii) L_2 norm of empirical minus theoretical variance over iterations (top-right), and (iv) trace of empirical variance over iterations (bottom-right).

Figure 1 shows results for a maximum $N = 1000$ iterations and $M = 2000$ samples. In the top-left subfigure, we plot the log-norm of the bias over N iterations, where for each method we plot two lines corresponding to the 2.5% (lower line) and 97.5% (upper line) over all M samples. We observe that the implicit method is slightly slower to converge but eventually obtains a similar rate of convergence to standard SGD, as predicted by Theorem 4.1. In the bottom-left figure, we plot the log-norm of the bias achieved by $\boldsymbol{\theta}_N$, over M samples for each method and for different learning rates (x-axis). In particular, we scale the baseline learning-rate up to being 3x the optimal value as calculated for the standard SGD. We observe that the bias of the standard SGD method is significantly affected by this scaling whereas the implicit method remains robust. In particular, the maximum observed bias of the implicit method remains constant whereas the minimum bias is actually improving when scaling the learning rate.

In the top-right figure, we plot for each method the L_2 norm of the empirical variance (computed over $M = 2000$ samples) subtracted from the one predicted by Theorem 4.2, and thus observe that both variances are converging to the

[8] The covariance matrix is designed to have eigenvalues almost-uniformly in the interval $[0.2, 1.0]$ and one larger at the order of $0.1p$.

Table 3. Test errors of standard and implicit SGD methods on the RCV1 dataset benchmark. Training times are roughly comparable. Best scores, for a particular loss and regularization, are bolded.

		REGULARIZATION (λ)		
LOSS		1E-5	1E-7	1E-12
HINGE	SGD	**4.65%**	**3.57%**	4.85%
	IMPLICIT	4.68%	3.6%	**3.46%**
LOG	SGD	5.23%	3.87%	5.42%
	IMPLICIT	**4.28%**	**3.69%**	**4.01%**

theoretical one. Finally, the bottom-right figure shows the trace of the variances of the iterates $\boldsymbol{\theta}_n$ for every method. This plot shows that the implicit method exhibits smaller empirical variance of the iterates, thus achieving an interesting trade-off: it gives up bias at the early stages of the iteration (see top-left) in order to compensate for more robustness (bottom-left) and smaller empirical variance (bottom-right). Asymptotically and *assuming* convergence, both methods provide identical estimators in terms of bias (top-left) and variance (top-right) as predicted by Theorems 4.1 and 4.2.

5.3. Additional experiments on SVM model

We are also interested to test the performance of the implicit procedure outside the family of GLMs. For that purpose, we implement an implicit online learning procedure for a SVM model and compare it to a standard SGD method on the RCV1 benchmark.[9]Some results using variations on the loss functions and the regularization parameter are shown in Table 3. A complete understanding of these results is still missing, however we do observe that the implicit method fares well compared to standard SGD and, at the same time, remains remarkably robust to misspecification. For example, note that in all experiments the standard SGD method degrades in performance for small or large regularization (in these experiments, the regularization parameter λ also affects the learning rate such that, larger λ means larger learning rates). However, the implicit method maintains a more stable performance accross experiments and, interestingly, it achieves best performance under minimal regularization using the hinge loss.

6. Conclusion

We study the statistical properties of explicit and implicit updates for fitting GLM models using SGD. In this model family, we derive a computationally efficient algorithm to perform the implicit updates. Furthermore, we derive formulas for the asymptotic bias and variance of both updates and show the fundamental bias/variance trade-off achieved by the implicit method. In small samples, the implicit estimates are more biased than the explicit ones but exhibit smaller empirical variance and are substantially more robust to misspecification. In the limit, both methods are statistically equivalent: they are both unbiased at the same convergence rate and enjoy the same statistical efficiency. Our theoretical results thus suggest that the implicit method could safely be the method of choice in estimating large-scale GLMs, especially when robustness is a concern. Our experiments confirm our theory and, more broadly, suggest that the implicit method can be a competitive method in large-scale machine learning tasks, requiring less tuning of learning-rate or regularization parameters. Future work will focus on the implicit method towards optimal learning rate schedules and a more detailed characterization of its robustness properties.

Acknowledgments

This work was sponsored, in part, by the 2012 Google PhD Fellowship in Statistics awarded to Panos Toulis, and grants from the NSF, NIH, ARO, ONR and the Sloan Foundation to Edoardo M. Airoldi.

Convergence of implicit RM procedure in quadratic mean.

Our proof is an adaptation of the original proof of convergence of the RM procedure (Robbins & Monro, 1951) and so we follow the same naming conventions for easy reference.

Let $b_n = \mathbb{E}\,(\theta_n - \theta^*)^2$, $d_n = \mathbb{E}\,((\theta_n - \theta^*)M(\theta_n))$, $e_n = \mathbb{E}\,(y_n^2)$ and $f_n = \mathbb{E}\,(M(\theta_n)^2)$. Using the definition of the implicit RM procedure (2) we obtain

$$b_n = b_{n-1} - 2a_n\mathbb{E}\,((\theta_{n-1} - \theta^*)y_n) + a_n^2 e_n \qquad (33)$$

Furthermore, by (2) and Assumption (2.2) we have $\mathbb{E}\,(\theta_{n-1}|\theta_n) = \theta_n + a_n M(\theta_n)$. Therefore,

$$\begin{aligned}\mathbb{E}\,((\theta_{n-1} - \theta^*)y_n) =& \mathbb{E}\,([\theta_n - \theta^* + a_n M(\theta_n)]M(\theta_n)) \\ =& d_n + a_n f_n \qquad (34)\end{aligned}$$

We substitute (34) into (33) and get

$$b_n = b_{n-1} - 2a_n d_n + a_n^2 e_n - 2a_n^2 f_n$$

The term $a_n^2 f_n$ is the only part that differentiates this proof with the original one in (Robbins & Monro, 1951). Since, $\sum a_n^2 < \infty$, essentially, this additional term has no effect, and the original analysis still carries through unaltered. We repeat the arguments here for completeness.

[9] We used Bottou's SVM SGD implementation available at http://leon.bottou.org/projects/sgd. Our implicit SVM is available at the first author's website.

Let $g_n = e_n - 2f_n$ and sum up all the terms b_n to obtain

$$b_n = b_0 - 2\sum_{j=1}^{n} a_j d_j + \sum_{j=1}^{n} a_j^2 g_j \tag{35}$$

By Assumptions (2.1)-(2.3) it holds $\sum a_j^2 d_j < \infty$ and $b_n \geq 0$ for all n. Therefore, we conclude that $\sum_{j=1}^{\infty} a_j d_j$ is finite. Thus, the series b_n is converging to a finite value i.e., $b_n \to b \geq 0$. Now, the goal is to construct a nonnegative series k_n such that

$$d_n \geq k_n b_n \text{ and } \sum^{\infty} a_j k_j = \infty \tag{36}$$

If this is possible, then $\sum a_j k_j b_j \leq \sum_j a_j d_j$ and so it is finite. Since $\sum_j a_j k_j$ diverges then it must be that $b_n \to 0$ in the limit. This construction is identical to (Robbins & Monro, 1951), page 403. In particular, by definition of (2) and Assumption (2.2), we can find a constant $K > 0$ such that $P(|\theta_n - \theta^*| \leq A_n) = 1$, where $A_n = K\sum_{j=1}^{n} a_j$. Then, k_n can be defined as

$$k_n = \inf\{\frac{M(x)}{x - \theta^*} : 0 < |x - \theta^*| \leq A_n\} \tag{37}$$

It can be shown that k_n satisfies the requirements in (36) (see pages 403-405 of (Robbins & Monro, 1951), Equation (24) and Theorem 2, in particular). Intuitively, this is because (for large enough n), $k_n \leq M'(\theta^*)$ by definition and $d_n \approx M'(\theta^*)b_n$, and thus the first requirement of (36) is fulfilled. Furthermore, by the monotonicity of $M(\cdot)$ (Assumption (2.3)), $k_n \geq \delta\frac{M'(\theta^*)}{A_n}$ for some fixed constant $\delta > 0$, which satisfies the second requirement of (36) by Assumption (2.1).

Proof for Proposition 3.1 (GLM moments)

Proof. For convenience, let $\eta = \boldsymbol{x}^\intercal\boldsymbol{\theta}$ and let $f(y; \eta, \psi)$ denote the density of the GLM model. The moment-generating function of y is given by

$$\begin{aligned} M(n) = \mathbb{E}\left(e^{ty}\right) &= \int e^{ty} f(y; \eta, \psi) dy \\ &= \int e^{\frac{b(\eta+t\psi)-b(\eta))}{\psi}} f(y; \eta + t\psi, \psi) dy \\ &= \exp\left\{\frac{b(\eta + t\psi) - b(\eta)}{\psi}\right\} \end{aligned}$$

Thus, the expected value is $\mathbb{E}\left(y|\,\boldsymbol{x}\right) = M'(0) = b'(\eta)$. Furthermore, by definition, $\mathbb{E}\left(y|\,\boldsymbol{x}\right) = h(\eta)$ and this concludes Part (a). For Part (b), note that $\mathbb{E}\left(y^2\middle|\,\boldsymbol{x}\right) = M''(0) = M'(0)b'(\eta) + M(0)b''(\eta)\psi$ and so $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = \mathbb{E}\left(y^2\middle|\,\boldsymbol{x}\right) - M'(0)^2 = \psi M(0)b''(\eta) = \psi h''(\eta)$. For Part (c) note that the log-likelihood is

$$\ell(\boldsymbol{\theta}; y, \boldsymbol{x}) = \log f(y; \eta, \psi) = (1/\psi)(\eta y - b(\eta))$$

Thus,

$$\begin{aligned} \nabla\ell(\boldsymbol{\theta}; y, \boldsymbol{x}) =&(1/\psi)(y - b'(\eta))\nabla_{\boldsymbol{\theta}}\eta \\ =&(1/\psi)\left(y - h(\mathbf{x}^\intercal\boldsymbol{\theta})\right)\boldsymbol{x} \end{aligned} \tag{38}$$

Subsequent differentiation yields Part (d). □

Proof for Lemma 4.1

Proof. For convenience, we make the following definitions. Let $\boldsymbol{\Gamma}_n = \boldsymbol{I} - a_n\boldsymbol{B}_n$ and the partial products $\boldsymbol{P}_i^n = \boldsymbol{\Gamma}_n\boldsymbol{\Gamma}_{n-1}\cdots\boldsymbol{\Gamma}_i = \prod_{k=t}^{i}\boldsymbol{\Gamma}_k$, for $i \leq t$, and $\boldsymbol{P}_n^{t+1} = \boldsymbol{I}$. Note that it holds that $||\boldsymbol{P}_i^n|| \leq Ke^{-\gamma\sum_{i=1}^{n} a_i}$ for suitable constants $K, \gamma > 0$ (Polyak & Juditsky, 1992; **?**). Let $A(n) = \gamma\sum_{i=1}^{n} a_i$ so that $||\boldsymbol{P}_i^n|| \leq Ke^{-\gamma A(n)}e^{\gamma A(i)}$. Also, $A(n)$ is increasing and $A(n) \to \infty$ and so $\boldsymbol{P}_i^n \to \boldsymbol{0}$ as $t \to \infty$, and for a fixed i.

The matrix recursion in Lemma 4.1 can now be rewritten as $\boldsymbol{X}_n = \boldsymbol{\Gamma}_n\boldsymbol{X}_{n-1} + a_n\boldsymbol{C} + a_n\boldsymbol{D}_n$ and, by performing successive multiplications we get:

$$\begin{aligned} \boldsymbol{X}_n =&(\boldsymbol{\Gamma}_n\boldsymbol{\Gamma}_{n-1}\cdots\boldsymbol{\Gamma}_1)\cdot\boldsymbol{X}_0 + a_n\boldsymbol{C} + a_n\boldsymbol{D}_n \\ &+ a_{n-1}\boldsymbol{\Gamma}_n\boldsymbol{C} + a_{n-1}\boldsymbol{\Gamma}_n\boldsymbol{D}_{n-1}\cdots \\ &+ a_1\boldsymbol{\Gamma}_n\cdots\boldsymbol{\Gamma}_2\boldsymbol{C} + a_1\boldsymbol{\Gamma}_n\cdots\boldsymbol{\Gamma}_2\boldsymbol{D}_1 \\ =&\boldsymbol{P}_1^n\boldsymbol{X}_0 + \boldsymbol{S}_n^0 + \boldsymbol{S}_n^1 \end{aligned} \tag{39}$$

where we have defined $\boldsymbol{S}_t^0 = \sum_{i=1}^{n} a_i\boldsymbol{P}_{i+1}^n\boldsymbol{C}$ and $\boldsymbol{S}_t^1 = \sum_{i=1}^{n} a_i\boldsymbol{P}_{i+1}^n D_i$. We have already established that $\boldsymbol{P}_1^n \to \boldsymbol{0}$. Out proof strategy will be to prove that $\boldsymbol{S}_t^0 \to \boldsymbol{B}^{-1}$ and that $\boldsymbol{S}_t^1 \to \boldsymbol{0}$.

By definition it holds that,

$$\sum_{i=1}^{n} a_i\boldsymbol{P}_{i+1}^n = \boldsymbol{B}_n^{-1} + \sum_{i=1}^{n}\boldsymbol{P}_i^n(\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1}) \tag{40}$$

To see this, note that $a_n\boldsymbol{I} = (\boldsymbol{I} - \boldsymbol{\Gamma}_n)\boldsymbol{B}_n^{-1}$. So, $\sum_{i=1}^{n}\boldsymbol{P}_i^n(\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1}) = (\boldsymbol{I} - \boldsymbol{\Gamma}_n)\boldsymbol{B}_n^{-1} + \boldsymbol{\Gamma}_n(\boldsymbol{I} - \boldsymbol{\Gamma}_{n-1})\boldsymbol{B}_{n-1}^{-1} + \cdots + \boldsymbol{\Gamma}_n\boldsymbol{\Gamma}_{n-1}\cdots\boldsymbol{\Gamma}_2(\boldsymbol{I} - \boldsymbol{\Gamma}_1)\boldsymbol{B}_1^{-1} = \boldsymbol{P}_n^{t+1}a_n\boldsymbol{I} + \boldsymbol{P}_n^n a_{n-1}\boldsymbol{I} + \cdots + \boldsymbol{P}_n^2 a_1\boldsymbol{I} = \sum_{i=1}^{n} a_i\boldsymbol{P}_{i+1}^n$ For $t = 1$ this reduces to $a_1\boldsymbol{P}_2^1 = \boldsymbol{B}_1^{-1} + \boldsymbol{P}_1^1\boldsymbol{E}_1 = \sum_{i=1}^{n} a_i\boldsymbol{P}_{i+1}^n$. Thus, for the last summation in Equation (40) we have, $||\sum_{i=1}\boldsymbol{P}_i^n(\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1})|| \leq Ke^{-\gamma A(n)}\sum_i e^{\gamma A(i)}O(a_i^2)$ Note that, since $\boldsymbol{B}_n$ converges, $||\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1}|| = ||\boldsymbol{B}_i^{-1}(\boldsymbol{B}_i - \boldsymbol{B}_{i-1})\boldsymbol{B}_{i-1}^{-1}|| = O(||\boldsymbol{B}_i - \boldsymbol{B}_{i-1}||) = O(a_i^2)$. Since $\sum_i O(a_i^2) < \infty$ and $e^{\gamma A(i)}$ is positive, increasing and diverging, by Kronecker's lemma we obtain $\sum_i O(a_i^2)e^{\gamma A(i)} = o(e^{A(n)})$. Thus, $\sum_{i=1}^{n}\boldsymbol{P}_i^n(\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1}) \to \boldsymbol{0}$. Therefore it holds:

$$\lim_{t\to\infty}\sum_{i=1}^{n} a_i\boldsymbol{P}_{i+1}^n = \boldsymbol{B}^{-1} \tag{41}$$

Thus, we have established that $\boldsymbol{S}_t^0 \to \boldsymbol{B}^{-1}\boldsymbol{C}$. Furthermore, since $\sum_i a_i||\boldsymbol{D}_i|| < \infty$, by applying Kronecker's lemma once more on the sum $\sum_i a_i \boldsymbol{P}_{i+1}^n \boldsymbol{D}_i$, we obtain that $\boldsymbol{S}_t^1 \to \boldsymbol{0}$. By substitution in Equation (39), we finally get that $\boldsymbol{X}_n \to \boldsymbol{B}^{-1}\boldsymbol{C}$.

For the second part and the recursion

$$\boldsymbol{Y}_n = (\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1}\boldsymbol{Y}_{n-1} + a_n(\boldsymbol{C} + \boldsymbol{D}_n) \qquad (42)$$

the proof is almost identical. For an intuition, note that for small enough a_n it holds $(\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1} = (\boldsymbol{I} - a_n\boldsymbol{B}_n) + O(a_n^2)$ and so the result should follow from the aforementioned analysis. For a complete formal proof, we just need to (re)define $\boldsymbol{\Gamma}_n = (\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1}$ and show that $a_n\boldsymbol{\Gamma}_n + a_{n-1}\boldsymbol{\Gamma}_n\boldsymbol{\Gamma}_{n-1} + \cdots + a_1\boldsymbol{\Gamma}_n\cdots\boldsymbol{\Gamma}_1 = \boldsymbol{B}_n^{-1} + \sum_i^n \boldsymbol{P}_i^n(\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1})$, similar to the case before (the difference in this recusion to the previous one is that this sum has $\boldsymbol{\Gamma}_n$ in all terms). To see why this is true, note that $\boldsymbol{I} - \boldsymbol{\Gamma}_n = a_n\boldsymbol{B}_n\boldsymbol{\Gamma}_n$ and so the right-hand side of the above equation is re-written as $\boldsymbol{B}_n^{-1} + \sum_i^n \boldsymbol{P}_i^n(\boldsymbol{B}_{i-1}^{-1} - \boldsymbol{B}_i^{-1}) = \boldsymbol{B}_n^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma}_n) + \boldsymbol{\Gamma}_n\boldsymbol{B}_{n-1}^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma}_{n-1}) + \cdots = a_n\boldsymbol{\Gamma}_n + a_{n-1}\boldsymbol{\Gamma}_n\boldsymbol{\Gamma}_{n-1}\cdots$ as needed. Noting that, $||(\boldsymbol{I} + a_n\boldsymbol{B}_n)^{-1}|| = O(||\boldsymbol{I} - a_n\boldsymbol{B}_n||)$ completes the proof.

Finally, note that the Lemma also holds $\boldsymbol{X}_n, \boldsymbol{C}, \boldsymbol{D}_n$ are vectors; the proof is identical. □

Proof for Lemma 4.2

Proof. We will use the following intermediate result:

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

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

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

where the last approximation follows from Stirling's formula. The stability result on the explicit SGD updates of Lemma 4.2 follows immediately by using the largest eigenvalue $\psi\lambda_{(p)}$ of $\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$. For the implicit SGD updates, simply note that the eigenvalues of $(\boldsymbol{I}+a_n\boldsymbol{J})^{-1}$ are less than 1, for any $a_n > 0$ and any nonnegative-definite matrix $\boldsymbol{J}$. □

References

Anbar, Dan. On optimal estimation methods using stochastic approximation procedures. *The Annals of Statistics*, pp. 1175–1184, 1973.

Beck, Amir and Teboulle, Marc. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bordes, Antoine, Bottou, Léon, and Gallinari, Patrick. Sgd-qn: Careful quasi-newton stochastic gradient descent. *The Journal of Machine Learning Research*, 10:1737–1754, 2009.

Cesa-Bianchi, Nicolo. *Prediction, learning, and games*. Cambridge University Press, 2006.

Duchi, John and Singer, Yoram. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

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

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

Fabian, Vaclav. Asymptotically efficient stochastic approximation; the rm case. *The Annals of Statistics*, pp. 486–495, 1973.

Fisher, Ronald A. 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, 1922.

Hoffman, Joe D and Frankel, Steven. *Numerical methods for engineers and scientists*. CRC press, 2001.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Kivinen, Jyrki and Warmuth, Manfred K. Additive versus exponentiated gradient updates for linear prediction. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pp. 209–218. ACM, 1995.

Kivinen, Jyrki, Warmuth, Manfred K, and Hassibi, Babak. The p-norm generalization of the lms algorithm for adaptive filtering. *Signal Processing, IEEE Transactions on*, 54(5):1782–1793, 2006.

Kulis, Brian and Bartlett, Peter L. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 575–582, 2010.

Murata, Noboru. A statistical study of on-line learning. *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK*, 1998.

Nagumo, Jin-Ichi and Noda, Atsuhiko. A learning method for system identification. *Automatic Control, IEEE Transactions on*, 12(3):282–287, 1967.

Nelder, J.A. and Wedderburn, R.W.M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pp. 370–384, 1972.

Nemirovski, Arkadi, Juditsky, Anatoli, Lan, Guanghui, and Shapiro, Alexander. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19 (4):1574–1609, 2009.

Nemirovski, Yudin, DB. *Problem complexity and method efficiency in optimization*. Wiley (Chichester and New York), 1983.

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

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Roux, Nicolas Le, Schmidt, Mark, and Bach, Francis. A stochastic gradient method with an exponential convergence rate for finite training sets. *arXiv preprint arXiv:1202.6258*, 2012.

Sacks, Jerome. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2): 373–405, 1958.

Sakrison, David J. Efficient recursive estimation; application to estimating the parameters of a covariance function. *International Journal of Engineering Science*, 3(4):461–483, 1965.

Schaul, Tom, Zhang, Sixin, and LeCun, Yann. No more pesky learning rates. *arXiv preprint arXiv:1206.1106*, 2012.

Schuurmans, Li Cheng SVN Vishwanathan Dale and Caelli, Shaojun Wang Terry. Implicit online learning with kernels. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, pp. 249. MIT Press, 2007.

Slock, Dirk TM. On the convergence behavior of the lms and the normalized lms algorithms. *Signal Processing, IEEE Transactions on*, 41(9):2811–2825, 1993.

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

Zinkevich, Martin. Online convex programming and generalized infinitesimal gradient ascent. 2003.