The Proximal Robbins–Monro Method

Panos Toulis\textsuperscript{1}, Thibaut Horel\textsuperscript{2}, and Edoardo M. Airoldi\textsuperscript{3}

\textsuperscript{1}Booth School of Business, University of Chicago
\textsuperscript{2}Department of Computer Science, Harvard University
\textsuperscript{3}Department of Statistics, Harvard University

Abstract

The need for parameter estimation with massive data has reinvigorated interest in iterative estimation procedures. Stochastic approximations, such as stochastic gradient descent, are at the forefront of this recent development because they yield estimation procedures that are simple, general, and fast. However, standard stochastic approximations are often numerically unstable, which is only partially addressed in current practice with ad hoc techniques. As a consequence, deterministic methods increasingly use proximal operators, which can induce numerical stability in a principled manner. A theoretical gap has thus emerged. While the majority of classical iterative estimation procedures are subsumed by the framework of Robbins and Monro (1951), there is no such generalization for stochastic approximations with proximal updates. In this paper, we conceptualize a general stochastic approximation method with proximal updates. The key idea is to solve a stochastic fixed point equation at each iteration. This method can be applied even in situations where the analytical form of the objective is not known, and so it generalizes many stochastic gradient procedures with proximal operators currently in use. Our theoretical analysis indicates that the proposed method has important stability benefits over the classical stochastic approximation method, while it retains the best known convergence rates of their Robbins–Monro counterparts. Exact instantiations of the proposed method are challenging, but we show that approximate instantiations lead to procedures that are easy to implement, and still dominate classical procedures by achieving numerical stability without tradeoffs. This last advantage is akin to that seen in deterministic proximal optimization, where the framework is typically impossible to instantiate exactly, but where approximate instantiations lead to new optimization procedures that dominate classical ones.

Keywords: iterative estimation; stochastic approximation; stochastic gradient descent; stochastic fixed-point equations; proximal operators; implicit updates.
## Contents

1 Introduction 3

2 The proximal Robbins–Monro method 4  
   2.1 Related work and contributions 6

3 Theory of implicit stochastic approximation 9  
   3.1 Convergence of implicit stochastic approximation 10  
   3.2 Non-asymptotic analysis 11  
   3.3 Asymptotic normality 13

4 Approximate implementation through the plug-in principle 13

5 Approximate implementation with nested Robbins–Monro 15

6 Application: iterative quantile estimation 17  
   6.1 Stability of the proximal Robbins–Monro method 18  
   6.2 Numerical evaluation 19

7 Concluding remarks 21

8 Acknowledgements 22

A Proofs of theorems for main method 26

B Proof of Theorem 5 37

C Stability of implicit SGD: iterative least-squares regression 41

D Computation of implicit updates 42
1 Introduction

In a seminal paper, Robbins and Monro (1951) considered the problem of estimating the zero $\theta_*$ of a function $h : \mathbb{R}^p \to \mathbb{R}$, where $h(\theta)$ is unknown but can be unbiasedly estimated by a random variable $W_{\theta}$, such that $\mathbb{E}(W_{\theta}) = h(\theta)$, for fixed $\theta \in \Theta \subseteq \mathbb{R}^p$. Starting from an estimate $\theta_0$, Robbins and Monro (1951) iteratively estimated $\theta_*$ as follows:

$$\theta_n = \theta_{n-1} - \gamma_n W_{\theta_{n-1}},$$

(1)

where $\gamma_n$ is known as the learning rate sequence, typically defined as $\gamma_n \propto 1/n$, for $n = 1, 2, \ldots$, so that $\sum \gamma_i^2 < \infty$ and $\sum \gamma_i = \infty$. Robbins and Monro (1951) proved convergence in quadratic mean for the procedure in Equation (1), under a monotonicity assumption for $h$ and bounded second moments for the noise, $W_{\theta} - h(\theta)$. Ljung et al. (1992); Kushner and Yin (2003); Borkar (2008) later strengthened this convergence result. Due to its remarkable simplicity and empirical performance, the Robbins–Monro method has found widespread applications across scientific fields, including statistics (Nevel’son et al., 1973; Ruppert, 1988), engineering (Benveniste et al., 1990), and optimization (Nesterov, 2004).

Recently, the Robbins–Monro method has attracted considerable interest in machine learning with large data sets (Zhang, 2004; Bottou, 2010; Moulines and Bach, 2011; Bottou et al., 2016), and in statistical inference at scale (Toulis and Airoldi, 2015; Chen et al., 2016; Su and Zhu, 2018; Li et al., 2017; Toulis and Airoldi, 2017). Given a dataset $D$, the Robbins–Monro method in Equation (1) can be applied with $W_{\theta}$ being the gradient of the negative log-likelihood of $\theta$ calculated at a single data point sampled with replacement from $D$. Standard theory then implies that $\theta_n$ converges to a point $\theta_{\infty}$ such that $h(\theta_{\infty}) = 0$—in other words, $\theta_n$ converges to the maximum-likelihood estimator (or maximum a posteriori if regularization is used) given dataset $D$. In this context, $h$ is the gradient of a convex scalar potential $H$, and the Robbins–Monro method is commonly referred to as stochastic gradient descent (SGD). While this context is widely applicable, and also helps with concreteness and interpretation, in this paper we consider the more general case where $h$ is not necessarily the gradient of a convex potential (with the exception of Theorem 2). Thus, our present work is placed in the broader context of stochastic approximation.
A well-known issue with the Robbins–Monro method is that the learning rate sequence crucially affects both its numerical stability and convergence. The procedure can actually be arbitrarily slow if $\gamma_n$ is even slightly misspecified. To illustrate, suppose that $\gamma_n = \gamma_1/n$, and there exists a scalar potential $H$, such that $\nabla H(\theta) = h(\theta)$, for all $\theta \in \Theta$. If $H$ is strongly convex with parameter $\mu$, then $\mathbb{E}(\|\theta_n - \theta^*\|^2) = O(n^{-\epsilon})$ if $\epsilon = 2\mu\gamma_1 < 1$ (Nemirovski et al., 2009, Section 1); (Moulines and Bach, 2011, Section 3.1). On the other hand, the procedure can diverge, even in the first few iterations, if the learning rate is too large, especially with non-Lipschitz likelihoods such as Poisson regression (Toulis et al., 2014). In summary, small learning rates can make the Robbins–Monro iterates converge very slowly, whereas large learning rates can make the iterates diverge numerically. Importantly, the requirements for numerical stability and fast convergence are very hard to reconcile in practice, especially in large-scale problems (Toulis and Airoldi, 2017), which renders the Robbins–Monro method, and all its derived procedures, inapplicable without extensive heuristic modifications (Bottou, 2012).

2 The proximal Robbins–Monro method

Our idea to improve the stability of the Robbins–Monro method is to transform Equation (1) into a fixed-point equation as follows:

$$\theta_n = \theta_{n-1} - \gamma_n W_{\theta^+_n},$$

where $\mathbb{E}(\theta_n|\mathcal{F}_{n-1}) = \theta^+_n$. (3)

Here, $\mathcal{F}_{n-1}$ is the natural filtration, $\sigma(\theta_0, \theta_1, \ldots, \theta_{n-1})$. The update in Equation (2) differs from the classical update in Equation (1) in calculating $W_\theta$ at an intermediate iterate $\theta^+_n$. The intermediate iterate generally differs from $\theta_{n-1}$ but depends deterministically on it, since Equation (3) is equivalent to the following equation:

$$\theta_{n-1} - \gamma_n h(\theta^+_n) = \theta^+_n.$$

This is an implicit equation because $\theta^+_n$ appears on both sides of the equation. As such, the method in Equations (2) and (3) is a form of implicit stochastic approximation.
To gain intuition, suppose that a convex scalar potential, $H$, exists such that $\nabla H = h$. Then, Equation (4) can be expressed through a proximal operator, $\text{prox}_{\gamma_n H}$, as follows:

$$\theta^+_n = \text{prox}_{\gamma_n H}(\theta_{n-1}) \overset{\text{def}}{=} \arg \min_{\theta \in \Theta} \left\{ \frac{1}{2\gamma_n} \| \theta - \theta_{n-1} \|^2 + H(\theta) \right\}. \tag{5}$$

The update in Equation (5) is the same as the updates used in the proximal point algorithm of Rockafellar (1976), which is the quintessential proximal optimization method. In recent years, interest in optimization through proximal operators has exploded since the resulting proximal procedures are stable and converge with minimal assumptions (Bauschke and Combettes, 2011; Parikh and Boyd, 2013). In addition, they can be applied to the setting where the objective function is the sum of a smooth function and a non-smooth function (as is common when using regularization in machine learning), and often lead to efficient, parallelizable algorithms.

To illustrate the stability of proximal updates, let us take norms in Equation (4):

$$\| \theta_{n-1} - \theta^* \|^2 = \| \theta^+_n - \theta^* \|^2 + 2\gamma_n h(\theta^+_n)^\top (\theta^+_n - \theta^*) + \gamma_n^2 \| h(\theta^+_n) \|^2.$$

By convexity of $H$, we have $h(\theta)^\top (\theta - \theta^*) \geq 0$ for any $\theta$, and so unless $h(\theta^+_n) = 0$ we obtain $\| \theta^+_n - \theta^* \|^2 < \| \theta_{n-1} - \theta^* \|^2$, a contraction. More generally, $\text{prox}_{\gamma_n H}$ can be shown to be a firmly non-expansive operator, and so the procedure in Equation (5) is stable. The drawback of such deterministic proximal optimization is that Equation (5) cannot be solved exactly, otherwise we could simply minimize $H(\theta)$ directly. Nevertheless, the proximal method is useful in practice since Equation (5) can be solved approximately without affecting the convergence of the resulting algorithm, provided that the approximation errors are small enough (Rockafellar, 1976).

Our proposed procedure in Equation (2) is the stochastic approximation analog of this idea. As in deterministic proximal optimization, we cannot compute $\theta^+_n$ exactly, otherwise we could simply define $\theta_n = \theta^+_n$ in Equation (2). Furthermore, we assume that we can only observe a noisy version of $h(\theta^+_n)$ through $W_{\theta}$, which is useful in situations where the analytic form of the objective is unknown. Despite these differences, our procedure behaves increasingly as a deterministic proximal optimization method, since Equation (2) can be re-written as $\theta_n = \theta^+_n - \gamma_n \varepsilon_n$, where $\varepsilon_n = W_{\theta^+_n} - h(\theta^+_n)$ is the zero-mean intermediate noise term, and $\gamma_n \to 0$. 
The key advantage of our proposed stochastic approximation in Equation (2) with respect to the classical stochastic approximation in Equation (1) is the additional numerical stability stemming from the implicit update in Equation (4), even though the implicit update only holds in expectation for $\theta^+_n$. More specifically, in Section 3 we show that the proximal Robbins–Monro procedure is more robust with respect to the specification of the learning rates, and is less sensitive to initial conditions, than the classical Robbins–Monro. Importantly, this stability carries through to approximate implementations of the proximal procedure (see Sections 4 and 5).

Remark 2.1. As mentioned above, the proximal Robbins–Monro we study in this paper is an “idealized procedure”, i.e., it is well-defined mathematically but, in general, it cannot be directly computed. We emphasize that such idealized procedures are ubiquitous in statistics and optimization. For instance, the proximal point algorithm described above, or projected gradient descent, to name a few, are typically approximated in practice even though they have a concrete mathematical definition. The study of our suggested procedure follows an outline similar to how other idealized procedures are studied: first, we explore the properties of the idealized procedure as if it were computed exactly, and then explore the extent to which these properties carry through to approximate implementations.

2.1 Related work and contributions

There is voluminous literature on classical stochastic approximation. The early mathematical work by Robbins and Monro (1951); Sacks (1958); Fabian (1968); Nevel’son et al. (1973); Robbins and Siegmund (1985); Wei (1987) established the fundamental properties, including convergence and asymptotic distributions. Subsequently, this work was pivotal in engineering applications, and particularly systems identification and tracking (Ljung et al., 1992; Benveniste et al., 1990); see also the excellent review by Lai et al. (2003). More recently, there have been important developments in studying stochastic approximations through the lens of dynamical systems theory, spearheaded by Kushner and Yin (2003) and Borkar (2008). Roughly at the same time, stochastic approximations started appearing in modern machine learning, usually in the form of stochastic gradient descent (SGD) methods, which have been irreplaceable in applications with large data sets and complex models (Zhang, 2004; Bottou, 2010).
Table 1: Work in stochastic approximations related to this paper. Modern procedures, such as SGD, are instantiations of the classical Robbins–Monro approximation method (Robbins and Monro, 1951). Our work provides a stochastic approximation method with proximal updates. Instantiations of our method include well-known existing procedures that employ proximal updates, such as implicit SGD. Additionally, it leads to novel procedures with nested stochastic approximations, which can be applied even in cases where the objective is not known analytically.

<table>
<thead>
<tr>
<th>Objective function</th>
<th>Robbins–Monro (explicit) updates</th>
<th>Proximal (implicit) updates</th>
</tr>
</thead>
<tbody>
<tr>
<td>analytically known</td>
<td>stochastic gradient descent</td>
<td>implicit stochastic gradients</td>
</tr>
<tr>
<td></td>
<td>(Coraluppi and Young, 1969); (Zhang, 2004); (Bottou, 2010); natural gradients (Amari, 1998); adaptive gradients (Duchi et al., 2011)</td>
<td>(Toulis and Airoldi, 2017); stochastic proximal gradients (Singer and Duchi, 2009); (Rosasco et al., 2014)</td>
</tr>
<tr>
<td>analytically unknown</td>
<td>quantile estimation</td>
<td>nested procedures (Section 5).</td>
</tr>
<tr>
<td></td>
<td>(Robbins and Monro, 1951)</td>
<td></td>
</tr>
</tbody>
</table>

However, classical stochastic approximations are numerically unstable, and often impossible to apply in practice without extensive heuristics. In this paper, we introduce an implicit stochastic approximation procedure, defined in Equations (2) and (3), which aims at mitigating stability problems of classical approximation through proximal updates. In the same way that the classical method of Robbins and Monro (1951) is the stochastic analog of gradient descent in deterministic optimization, implicit stochastic approximation is the stochastic analog of the proximal point algorithm (Rockafellar, 1976), which is the quintessential method in proximal optimization. This fills a crucial gap in the literature of stochastic approximations. As visualized in Table 1, the method we introduce is general enough to cover both cases where an analytic form of the objective function is known, and cases where no such form is known for the objective or its gradients.

There are mainly two lines of literature related to our work, as shown in the table. One line of work is about procedures where the proximal update is deterministic, and is performed after a classical stochastic update. For example, the forward-backward procedure of Singer and Duchi (2009) and the proximal stochastic gradient procedure studied by Rosasco et al. (2014) fall in this category. Such procedures first make the update \( \tilde{\theta}_n = \theta_{n-1} - \gamma_n W \theta_{n-1} \), and then define \( \theta_n = \text{prox}_f(\tilde{\theta}_n) \), where \( f \) is some convex regularization function. In our work, we wish to avoid making any explicit update at all in order to ensure stability. A notable exception is presented in Section 5, where an approximate implementation of our method involves multiple explicit updates within a nested procedure.
Another line of work involves procedures where implicit updates are directly used in the update equation, in contrast to our procedure where the implicit update holds in expectation. Incremental proximal procedures (Bertsekas, 2011), and implicit SGD (Toulis et al., 2014; Toulis and Airoldi, 2017) fall in this category. In our notation, the update is $\theta_n = \theta_{n-1} - \gamma_n W\theta_n$, which requires that the implicit equation is solvable. This is not uncommon, and efficient algorithms exist for several models (Toulis et al., 2014, Algorithm 1). In numerical optimization and engineering, the stochastic proximal point algorithms studied by Bianchi (2016); Ryu and Boyd (2014); Patrascu and Necoara (2017) are closely related. Interestingly, all such procedures can be viewed as the plug-in versions of the proposed proximal Robbins–Monro, since $\theta_n$ in Equation (3) is in fact an unbiased estimator of $\theta_n^+$. By plugging in this estimator in the main update of Equation (2) we get the implicit update, $\theta_n = \theta_{n-1} - \gamma_n W\theta_n$. We discuss this idea further in Section 4.

Thus, the central contribution of this paper is the introduction of the proximal Robbins–Monro procedure as the stochastic analog of the proximal point algorithm. This fills a gap in the literature that has remained open since classical stochastic approximation was introduced by Robbins and Monro (1951) as the stochastic analog of gradient descent. We provide full analysis of the theoretical properties (both asymptotic and non-asymptotic) of the new method in Section 3. The analysis shows that proximal Robbins–Monro is more stable numerically than classical Robbins–Monro, and is also less sensitive to hyperparameter tuning in achieving the best known convergence rates for non-averaged stochastic approximation.

Regarding implementation, our proximal Robbins–Monro method does present some unique challenges. Since exact implementation will generally be impossible, we analyze the following two approximate instantiations. First, in Section 4 we assume $W_\theta$ to be known analytically, and show that the plug-in principle leads to a family of well-known implicit procedures (Bertsekas, 2011; Toulis et al., 2014), which are becoming increasingly popular for their superior numerical stability compared to classical SGD procedures. They are also easy to implement in a broad family of models, and their theoretical properties are now well understood (Kulis and Bartlett, 2010; Bertsekas, 2011; Toulis and Airoldi, 2017; Ryu and Boyd, 2014; Patrascu and Necoara, 2017).

Second, in Section 5 we discuss settings where the form of $W_\theta$ is unknown, and analyze an approximate implementation that relies on stochastic fixed-point equations solved by nested classical stochastic approximation. We present a full convergence analysis of the approximate procedure,
which is particularly challenging due to its nested structure. This procedure, and its theoretical
analysis, constitute a key contribution of this paper. We are unaware of other proximal methods
that address settings where the objective is analytically unknown, and where the underlying pro-
cedure is comprised of nested stochastic fixed points. In Section 6, we show significant benefits in
numerical stability through the quantile regression example of Robbins and Monro (1951).

3 Theory of implicit stochastic approximation

In this section, we state the theoretical guarantees for implicit stochastic approximation: conver-
gence almost-surely (Section 3.1), asymptotic normality (Section 3.3), and non-asymptotic conver-
gence rate (Section 3.2). All proofs can be found in Appendix A.

Symbol \( \| \cdot \| \) denotes the \( L_2 \) vector/matrix norm. We define the error random variables at the
intermediate iterate as \( \varepsilon_n = W_{\theta_n^+} - h(\theta_n^+) \). Because \( \theta_n^+ \) is a deterministic function of \( \theta_{n-1} \) it holds
that \( \mathbb{E}(\varepsilon_n | F_{n-1}) = 0 \). The parameter space for \( \theta \) is \( \Theta \subseteq \mathbb{R}^p \), and is convex. For positive scalar
sequences \( (a_n) \) and \( (b_n) \), we write \( b_n = O(a_n) \) to express that \( b_n \leq ca_n \), for some fixed \( c > 0 \), and
every \( n = 1, 2, \ldots \); we write \( b_n = o(a_n) \) to express that \( b_n/a_n \to 0 \) in the limit where \( n \to \infty \).

Notation \( b_n \downarrow 0 \) means that \( b_n \) is positive and decreasing towards zero. Depending on which
result we state, implicit stochastic approximation operates under a combination of the following
assumptions.

Assumption 1. It holds that \( \gamma_n = \gamma_1 n^{-\gamma} \), \( \gamma_1 > 0 \) and \( \gamma \in (0, 1] \).

Assumption 2. Function \( h \) is Lipschitz with parameter \( L \), i.e., for all \( \theta_1, \theta_2 \in \Theta \),

\[
\|h(\theta_1) - h(\theta_2)\| \leq L\|\theta_1 - \theta_2\|.
\]

Assumption 3. Function \( h \) satisfies either

(a) \( (\theta - \theta_*)^T h(\theta) \geq 0 \), for all \( \theta \in \Theta \), or, for all \( n = 1, 2, \ldots , \)

(b) \( (\theta - \theta_*)^T h(\theta) > 0 \), for all \( \theta \in \Theta \setminus \{\theta_*\} \), or, for all \( n = 1, 2, \ldots , \)

(c) \( (\theta_n^+ - \theta_*)^T h(\theta_n^+) \geq \delta_n \|\theta_n^+ - \theta_*\|^2 \), where \( \delta_n = \delta_1 n^{-\delta} \), \( \delta_1 > 0 \) and \( 0 < \gamma + \delta \leq 1 \).

Assumption 4. There exists \( H : \mathbb{R}^p \to \mathbb{R} \) such that \( \nabla H(\theta) = h(\theta) \), for all \( \theta \in \Theta \).
Assumption 5. There exists fixed $\sigma^2 > 0$ such that, for all $n = 1, 2, \ldots$,

$$\mathbb{E}(\|\epsilon_n\|^2|\mathcal{F}_{n-1}) \leq \sigma^2.$$ 

Assumption 6. Let $\Xi_n = \mathbb{E}(\epsilon_n\epsilon_n^\top|\mathcal{F}_{n-1})$, then $\|\Xi_n - \Xi\| \to 0$ for fixed positive-definite matrix $\Xi$.

Furthermore, if $\sigma^2_{n,s} = \mathbb{E}(\|\epsilon_n\|^2 \geq s/\gamma_n \|\epsilon_n\|^2)$, then for all $s > 0$, $\sum_{i=1}^n \sigma^2_{i,s} = o(n)$ if $\gamma_n \propto n^{-1}$, or $\sigma^2_{n,s} = o(1)$ otherwise.

Assumption 3(a) is a typical convexity assumption. Assumption 3(b) is akin to a strict convexity assumption (in particular, it implies that $\theta_\star$ is unique). Assumption 3(c) is stronger than the convexity assumption, but weaker than strong convexity. Assumption 4 will be used later in Sections 4 and 5, where we consider concrete implementations of our procedure. Assumption 5 was introduced by Robbins and Monro (1951), and has since been standard in stochastic approximation analysis, since bounded noise is a crucial condition for convergence. Assumption 6 is the Lindeberg condition that is used to prove asymptotic normality of $\theta_n$, later in this section. Overall, our assumptions are weaker than the assumptions in classical stochastic approximation because they refer to the idealized procedure of Equation (2); compare, for example, Assumptions 1-6 with assumptions (A1)-(A4) of Borkar (2008, Section 2.1), or the assumptions by Benveniste et al. (1990, Theorem 15).

3.1 Convergence of implicit stochastic approximation

In Theorem 1, we derive a proof of almost sure convergence of implicit stochastic approximation, which relies on the supermartingale lemma of Robbins and Siegmund (1985).

Theorem 1. Suppose that Assumptions 1, 2, 3(b), and 5 hold. Then, the iterates $\theta_n$ of the implicit stochastic approximation method in Equations (2) and (3) converge almost surely to $\theta_\star$; i.e., $\theta_n \to \theta_\star$, such that $h(\theta_\star) = 0$, almost surely.

The conditions for almost sure convergence of implicit stochastic approximation are weaker than the conditions required for classical stochastic approximation. For example, in standard stochastic approximation methods it is typically assumed that the iterates $\theta_n$ are almost surely bounded (Borkar, 2008, Assumption (A4)).
3.2 Non-asymptotic analysis

In this section, we prove results on upper bounds for deviance, \( \mathbb{E}(H(\theta_n) - H(\theta^*)) \), and the mean quadratic errors, \( \mathbb{E}(\|\theta_n - \theta^*\|^2) \). This provides information on the rate of convergence, as well as the stability of implicit stochastic approximation. Theorem 2 on deviance that follows assumes non-strong convexity of \( H \), whereas Theorem 3 on squared error assumes strong convexity.

**Theorem 2.** Suppose that Assumptions 1, 2, 3(a), 4, and 5 hold. Let \( \Gamma^2 = \mathbb{E}(\|\theta_0 - \theta^*\|^2) + \sigma^2 \sum_{i=1}^{\infty} \gamma_i^2 + \gamma_i^2 \sigma^2 \). Then, if \( \gamma \in (2/3, 1] \), there exists \( n_{0,1} < \infty \) such that, for all \( n > n_{0,1} \),

\[
\mathbb{E}(H(\theta_n) - H(\theta^*)) \leq \left[ \frac{2\Gamma^2}{\gamma \gamma_1} + o(1) \right] n^{-1+\gamma}.
\]

If \( \gamma \in (1/2, 2/3) \), there exists \( n_{0,2} < \infty \) such that, for all \( n > n_{0,2} \),

\[
\mathbb{E}(H(\theta_n) - H(\theta^*)) \leq \left[ \Gamma \sigma \sqrt{L \gamma_1} + o(1) \right] n^{-\gamma/2}.
\]

Otherwise, \( \gamma = 2/3 \) and there exists \( n_{0,3} < \infty \) such that, for all \( n > n_{0,3} \),

\[
\mathbb{E}(H(\theta_n) - H(\theta^*)) \leq \left[ \frac{3 + \sqrt{9 + 4\gamma_1^2 L \sigma^2/\Gamma^2}}{2\gamma_1/\Gamma^2} + o(1) \right] n^{-1/3}.
\]

There are two main results in Theorem 2. First, the rates of convergence for the deviance are either \( O(n^{-1+\gamma}) \) or \( O(n^{-\gamma/2}) \), depending on the learning rate, \( \gamma \). Second, there is a uniform decay of expected deviance towards zero, whereas in standard stochastic approximation under non-strong convexity, there is a term of the form \( \exp(4L^2 \gamma_1^2 n^{-2\gamma}) \) (Moulines and Bach, 2011, Theorem 4), which can amplify the initial conditions arbitrarily. Thus, the proximal Robbins–Monro method has similar asymptotic properties to classical stochastic approximation, but is more stable.

**Remark 3.1.** The best rate of convergence for the proximal Robbins–Monro as shown in Theorem 2 is \( O(n^{-1/3}) \), which matches the best known rate for standard stochastic approximations with non-strong convex objective (Moulines and Bach, 2011, Theorem 4). This rate is suboptimal since it is worse than the minimax rate of \( O(n^{-1/2}) \) that is achieved through Polyak-Ruppert averaging (Ruppert, 1988). We conjecture that our proposed procedure can also achieve the minimax rate through averaging, but we leave this for future work.
Remark 3.2. The proof of Theorem 2 presents some unique technical challenges, including an implicit inequality of the form $b_n + g(b_n) \leq b_{n-1}$, with $g$ non-decreasing but complex. Our approach is to solve the reverse recursive inequality, $\tilde{b}_n(\beta) + g(\tilde{b}_n(\beta)) \geq \tilde{b}_{n-1}(\beta)$, in some parametric family such as $\tilde{b}_n(\beta) = O(n^{-\beta})$, which is more tractable. Then, $\tilde{b}_n(\beta)$ is an upper bound for $b_n$, for any $\beta$, and so a natural upper bound for $b_n$ is given by $b_n \leq \arg \min_{\beta} \tilde{b}_n(\beta)$. This solution technique is reminiscent of the majorization-minorization idea (Lange, 2010), and may be more broadly useful.

**Theorem 3.** Suppose that Assumptions 1, 3(c), and 5 hold. Let $\zeta_n = \mathbb{E}(\|\theta_n - \theta_*\|^2)$ and $\kappa = 1 + 2\gamma_1\delta_1$. Then, if $\gamma + \delta < 1$, for every $n > 1$ it holds that

$$\zeta_n \leq e^{-\log \kappa n^{1-\gamma-\delta}} \zeta_0 + \sigma^2 \frac{\gamma_1\kappa}{\delta_1} n^{-\gamma+\delta} + O(n^{-\gamma+\delta-1}).$$

Otherwise, if $\gamma = 1, \delta = 0$, it holds that

$$\zeta_n \leq e^{-\log \kappa \log n} \zeta_0 + \sigma^2 \frac{\gamma_1\kappa}{\delta_1} n^{-1} + O(n^{-2}).$$

There are two main results in Theorem 3. First, if potential function $H$ is strongly convex ($\delta = 0$), then the rate of convergence of $\mathbb{E}(\|\theta_n - \theta_*\|^2)$ is $O(n^{-\gamma})$, which matches the rate of convergence for classic stochastic approximation under strong convexity (Benveniste et al., 1990, Theorem 22). The best possible rate here is $O(1/n)$, which is also the minimax rate with strongly convex objectives. Second, there is an exponential discounting of initial conditions $\zeta_0$ regardless of the specification of the learning rate parameter $\gamma_1$ and the Lipschitz parameter $L$. In contrast, in classical stochastic approximation there exists a term $\exp(L^2 \gamma_1^2 n^{1-2\gamma})$ in front of the initial conditions $\zeta_0$, which can make the approximation diverge numerically if $\gamma_1$ is misspecified with respect to the Lipschitz parameter $L$ (Moulines and Bach, 2011, Theorem 1). Thus, as in the non-strongly convex case of Theorem 2, implicit stochastic approximation has similar asymptotic rates to classical stochastic approximation, but is also more stable.

Remark 3.3. When $\gamma = 1$ and $\delta = 0$, misspecification of the learning rate parameter can indeed lead to arbitrary slowdown to a rate $O(\max\{n^{-1}, n^{-\log \kappa}\})$. This is also true for classical stochastic approximation (Moulines and Bach, 2011, Theorem 1). The key difference between the two procedures, as described above, is numerical stability.
3.3 Asymptotic normality

Asymptotic distribution are well studied in classical stochastic approximation. Starting from Fabian (1968) there has been extensive work in identifying asymptotic distributions of stochastic approximations. In this section, we leverage this theory to show when iterates from implicit stochastic approximation are asymptotically normal. The following theorem establishes this result using Theorem 1 of Fabian (1968); see also (Ljung et al., 1992, Chapter II.8).

**Theorem 4.** Suppose that Assumptions 1, 2, 3(a), 5, and 6 hold. Suppose also that 
\[
(\gamma_1 \mathbf{J}_h(\theta^*) - I)\]
is positive-definite, where \(\mathbf{J}_h(\theta)\) is the Jacobian of \(h\) at \(\theta\), and \(I\) is the \(p \times p\) identity matrix. Then, \(\theta_n\) of implicit stochastic approximation is asymptotically normal:

\[
n^{-\gamma/2}(\theta_n - \theta^*) \rightarrow \mathcal{N}_p(0, \Sigma).
\]

The covariance matrix \(\Sigma\) is the unique solution of

\[
(\gamma_1 \mathbf{J}_h(\theta^*) - I/2)\Sigma + \Sigma(\gamma_1 \mathbf{J}_h(\theta^*) - I/2) = \Xi.
\]

Theorem 4 shows that the asymptotic distribution of \(\theta_n\) is identical to the asymptotics of the classical Robbins–Monro method, as derived by Fabian (1968). Intuitively, in the limit as \(n\) grows, we have that \(\theta_n^+ \approx \theta_{n-1} + O(\gamma_n)\) with high probability, and thus implicit stochastic approximation behaves like the classical procedure, since \(\theta_n = \theta_n^+ - \gamma_n \varepsilon_n\).

**Remark 3.4.** A closed-form solution for \(\Sigma\) is possible if \(\Xi\) commutes with \(\mathbf{J}_h(\theta^*)\), such that \(\Xi \mathbf{J}_h(\theta^*) = \mathbf{J}_h(\theta^*) \Xi\). Then \(\Sigma\) can be derived as \(\Sigma = (2\gamma_1 \mathbf{J}_h(\theta^*) - I)^{-1}\Xi\).

4 Approximate implementation through the plug-in principle

Here, we focus on implementation of the proximal Robbins–Monro method. In this section, we assume that the analytic form of \(W_\theta\) is known. This leads to well-known stochastic procedures, and so our discussion will be short. Later, in Section 5, we focus on the more challenging setting of unknown analytic form for \(W_\theta\), and analyze the resulting procedure in detail.

When the analytic form of \(W_\theta\) is known it is possible to apply the plug-in principle to implement
the proximal Robbins–Monro update. Specifically, the idea is to use $\theta_n$ instead of $\theta_n^+$ in Equation (2) since, by definition, $\theta_n$ is an unbiased estimator of $\theta_n^+$; i.e., $\mathbb{E}(\theta_n|F_{n-1}) = \theta_n^+$. This approximate update yields the following procedure:

$$\theta_n = \theta_{n-1} - \gamma_n W_{\theta_n}. \quad (6)$$

One of the most popular applications of procedure in Equation (6) is in iterative statistical estimation, where $W_\theta = -\nabla \log f(Y; X, \theta)$, and $f$ corresponds to the likelihood of a random data point $(Y, X)$ at parameter value $\theta$. For example, if in Equation (6) we use $W_{\theta_{n-1}}$ instead of $W_{\theta_n}$, this amounts to classical SGD, which is widely popular in optimization and signal processing (Coraluppi and Young, 1969), and has been fundamental in modern machine learning with large data sets (Amari, 1998; Zhang, 2004; Bottou, 2010; Bottou et al., 2016).

When we use $W_{\theta_n}$, as originally described in Equation (6), then the resulting procedure is known as incremental proximal method in optimization (Bertsekas, 2011), or as implicit stochastic gradient descent in statistics (Toulis et al., 2014). This procedure shows superior performance to standard stochastic gradient descent, both in theory and practice (Toulis and Airoldi, 2017). In particular, in accordance to the theoretical properties of their stochastic approximation counterparts, implicit SGD has identical asymptotic efficiency and convergence rate as standard SGD, but it is significantly more stable numerically. We refer readers to (Bertsekas, 2011) and (Toulis and Airoldi, 2017) for two complementary analyses of implicit SGD, including asymptotic and non-asymptotic errors. See also (Bianchi, 2016) and (Ryu and Boyd, 2014) for analysis using monotone operator theory. In Appendix C, we also present an example on a least squares model illustrating that implicit SGD is more stable than classical SGD, and that such stability is obtained essentially for free. Last, in Appendix D we discuss computational issues in calculating the implicit update.

Remark 4.1. We see that the approximate implementation of the proximal Robbins–Monro through the plug-in principle, as defined in Equation (6), leads to methods that inherit the theoretical properties of the idealized procedure derived in Section 3. This reiterates the point in Remark 2.1 regarding the importance of studying the theoretical properties of the idealized procedure despite its implementation challenges.
Approximate implementation with nested Robbins–Monro

In this section, we consider cases where the forms of neither the regression function $h$ nor the random variable $W_\theta$ are known analytically. We present an approximate implementation of the proximal Robbins–Monro based on nested stochastic approximation that can be used without any auxiliary knowledge of the estimation problem. The nested procedure is in fact a fixed-point stochastic approximation procedure (Borkar, 2008), which, however, is run only for a finite number of steps. Section 6 illustrates the benefits of the nested procedure in quantile estimation.

The idea is to approximate $\theta_n^+$ through a separate, standard stochastic approximation procedure. At every $n$-th iteration, we run a Robbins–Monro procedure, $x_k$, as follows:

\begin{equation}
\begin{aligned}
x_1 &= \theta_{n-1}, \\
x_{k+1} &= x_k - a_k(\gamma_n W_{x_k} + x_k - x_1), \quad 1 \leq k \leq K, \\
\theta_n &= x_K.
\end{aligned}
\end{equation}

At first, it may seem that this procedure is affected by the same stability issues as classical stochastic approximation. However, our convergence result will show that this is not true. For intuition, consider the case where there exists a convex potential $H$ such that $h = \nabla H$. Then, for fixed $n$ the sequence $(x_k)$ is a standard Robbins–Monro procedure applied to a different minimization problem:

\begin{equation}
\begin{aligned}
\min \limits_{\theta \in \Theta} & \left\{ \frac{1}{2\gamma_n} \| \theta - \theta_{n-1} \|^2 + H(\theta) \right\}.
\end{aligned}
\end{equation}

With this formulation, it is easy to verify that $\theta_n^+$ is the solution to this optimization problem, so that $x_k \to \theta_n^+$. What we gain compared to applying the classical Robbins–Monro method to $h$ directly, is that the objective function in Equation (8) is now strongly convex, even when $H$ is not. Therefore, the problem structure that we designed allows the application of explicit updates, without compromising numerical stability. We illustrate this point in Section 6.

We will require the following strengthening of Assumption 3(c):

**Assumption 7.** for some $\delta > 0$ and for all $\theta \in \mathbb{R}^p$ and $\theta' \in \mathbb{R}^p$, $(h(\theta) - h(\theta'))^\top (\theta - \theta') \geq \delta \| \theta - \theta' \|^2$.

Note that when $h = \nabla H$, Assumption 7 states that $H$ is $\delta$-strongly convex.
Theorem 5. Suppose that Assumptions 2, 5 and 7 hold, then the nested stochastic approximation procedure in Equation (7) with parameters \( \gamma_n = \gamma \) and \( a_k = 2a/K \), such that \( e^{-a} < \delta/L \) and \( K \geq 2a(1 + \gamma L)^2 \), satisfies:

\[
\mathbb{E}(\|\theta_n - \theta_\ast\|) \leq \left[ \frac{1}{(1 + \gamma \delta)^n} + \frac{C^n}{e^{-a}L} \right] \|\theta_0 - \theta_\ast\| + \frac{\gamma \sigma \sqrt{2a}}{(1-C)\sqrt{K}}
\]

where \( C \overset{\text{def}}{=} (1 + e^{-a}\gamma L)/(1 + \gamma \delta) \).

Theorem 5 shows two key results. First, the initial conditions are forgotten exponentially fast at a rate \((1 + \gamma \delta)^{-n}\). Second, an approximation error smaller than \( \varepsilon \) can be obtained by choosing \( n = O(\log \frac{1}{\varepsilon}) \) and \( K = O\left(\frac{1}{\varepsilon^2}\right) \), where \( K \) is the number of iterations in the inner procedure. Taken together, these choices imply a total number of gradient observations of order \( O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}\right) \).

Remark 5.1. The proof of Theorem 5 is technically challenging due to the nested nature of the procedure. This requires carefully balancing the accumulation of approximation errors from the inner iteration jointly with the rate of convergence of the idealized procedure. To the best of our knowledge, there is no prior analysis of such stochastic fixed-point procedures in the literature. So, the proof of Theorem 5 applies novel techniques, which may be of general interest.

Remark 5.2. The nested nature of the procedure described in Equation (7) is reminiscent of the Catalyst scheme of Lin et al. (2015), which is a general acceleration technique for a large class of first-order optimization methods. Similar to the Catalyst scheme, our procedure (7) approximately computes a proximal update at each iteration. The key difference is that we explicitly analyze how to perform this approximate computation whereas the Catalyst scheme assumes “oracle” access to an algorithm able to perform this update. Furthermore, the main focus of the Catalyst scheme is to achieve acceleration à la Nesterov with the use of a “momentum” term, while our focus is to analyze the gain in stability provided by proximal updates.

In the following section, we illustrate the use of the nested procedure of Equation (7) and the use of Theorem 5 through the classical quantile estimation problem of Robbins and Monro (1951).
6 Application: iterative quantile estimation

In their seminal paper, Robbins and Monro (1951) described an application of their proposed stochastic approximation in iterative quantile estimation. In this problem, \( W_\theta \) corresponds to a sample drawn from a distribution with cumulative function \( F(\theta) \). The goal is to estimate \( \theta_* \) such that \( F(\theta_*) = \alpha \), for given quantile \( \alpha \). A relevant application from toxicology is the estimation of the dose that is lethal to 50% of experimental subjects, known as LD50 (Grieve, 1996).

In more detail, consider a random variable \( Z \) with cumulative distribution function \( F \). An experimenter wants to find the point \( \theta_* \) for which \( F(\theta_*) = \alpha \), for some fixed \( \alpha \in (0, 1) \). The experimenter can draw samples of \( Z \) but has only access to the random variable \( W_\theta = \mathbb{I}\{Z \leq \theta\} - \alpha \), for any value of \( \theta \). Robbins and Monro (1951) showed that the following iterative procedure,

\[
\theta_n = \theta_{n-1} - \gamma_n W_{\theta_{n-1}},
\]

(9)

converges to \( \theta_\infty \) for which \( \mathbb{E}(W_{\theta_\infty}) = 0 \). Consequently, \( \mathbb{E}(\mathbb{I}\{Z \leq \theta_\infty\}) - \alpha = F(\theta_\infty) - \alpha = 0 \), and by monotonicity of \( F \), we obtain \( \theta_\infty = \theta_* \).

Despite theoretical convergence, the numerical stability issue of the Robbins–Monro procedure can be challenged by the following result.

**Proposition 6.1.** Assume that \( \theta_0 < \theta_* \) and that \( \theta_0 + \gamma_1 \alpha > \theta_* \), then for any \( \varepsilon > 0 \) such that \( \theta_0 + \gamma_1 \alpha > \theta_* + \varepsilon \), with probability \( 1 - F(\theta_0) \), the number of iterations \( N_\varepsilon \) of procedure (9) required to approximate \( \theta_* \) within accuracy \( \varepsilon \) is lower-bounded:

\[
\log N_\varepsilon \geq \frac{\theta_0 + \gamma_1 \alpha - \theta_* - \varepsilon}{(1 - \alpha) \gamma_1}.
\]

(10)

**Proof.** With probability \( 1 - F(\theta_0) \) the first iterate of (9) is \( \theta_1 = \theta_0 + \gamma_1 \alpha > \theta_* \), where the inequality is by assumption. Conditioned on this event, the progress in each subsequent iteration, namely \( \theta_n - \theta_{n-1} \), is upper-bounded by \( \gamma_n(1 - \alpha) \) with probability 1 as long as \( \theta_n > \theta_* \). This implies that \( \theta_n \geq \theta_0 + \gamma_1 \alpha - (1 - \alpha) \sum_{k=2}^{n} \frac{\gamma_k}{k} \geq \theta_0 + \gamma_1 \alpha - (1 - \alpha) \gamma_1 \log n \).

Proposition 6.1 shows that there are values of the learning rate parameter \( \gamma_1 \) and initial estimate \( \theta_0 \) for which the classical Robbins–Monro procedure may be stuck indefinitely. For example, let
$F$ be the standard normal distribution, and let $\alpha = 0.999$, so that $\theta_* = 3.09$ is the solution. Suppose also that $\gamma_1 = F'(\theta_*)^{-1} \approx 297$, which is the learning rate value suggested by standard theory (Nemirovski et al., 2009). Let $\theta_0 = -10$ and suppose that $W_{\theta_0} = -\alpha$. It follows that

$$\theta_1 = -10 - \gamma_1(-\alpha) = -10 + \gamma_1\alpha \approx 287 \gg \theta_*.$$  

From there, the Robbins–Monro method makes progress by at most $\gamma_i(1 - \alpha) \approx 297 \cdot 10^{-3}$ at each step. Thus, the number of iterations required to return back from $\theta_1$ to a region near $\theta_*$ is at the order of $e^{956}$. In other words, the procedure gets stuck at large values of $\theta$, where the derivative of the objective is negligible.

This numerical example illustrates that a misspecification of $\gamma_1$ can dramatically amplify the initial conditions in classical stochastic approximation, and affect convergence. It is therefore interesting to investigate whether the proximal Robbins–Monro method offers an improvement.

### 6.1 Stability of the proximal Robbins–Monro method

In the context of quantile estimation, the proximal Robbins–Monro method can be implemented through the approximate procedure in Equation (7):

$$x_1 = \theta_n - 1, \quad x_{k+1} = x_k - a_1(\gamma_n W_{x_k} + x_k - x_1), \quad 1 \leq k \leq K, $$

$$\theta_n = x_K,$$

(11)

where $W_{\theta} = \mathbb{I}\{Z \leq \theta\} - \alpha$, as before; $a_1$ and $K$ are constants. Before presenting our numerical experiments, we discuss intuitively why the nested procedure in Equation (11) improves upon the classical Robbins–Monro method in Equation (9), and also discuss how to define the constants according to Theorem 5. We address these two issues successively.

First, consider the idealized case where $K = \infty$. In this case, the iteration in Equation (11) converges to the solution of the following fixed point equation:

$$x_\infty = \theta_{n-1} - \gamma_n[F(x_\infty) - \alpha].$$
The next iterate, $\theta_n$, is simply defined as $\theta_n = x_\infty$. It is easy to verify the stability of this fixed point. For example, if $\theta_{n-1} < \theta_*$, then $\theta_n < \theta_{n-1} < \theta_*$; and, conversely, if $\theta_{n-1} > \theta_*$, then $\theta_* < \theta_{n-1} < \theta_n$. That is, the idealized procedure with $K = \infty$ always pulls back in the right direction towards $\theta_*$, and thus always makes progress towards the global solution. Convergence is also extremely fast, as shown in the proof of Theorem 5. To illustrate numerically, consider the example of the previous section where the classical Robbins–Monro method did not converge. Using the same numbers, at the second iteration the idealized procedure will calculate:

$$\theta_1 = -10 - 297[F(\theta_1) - .999],$$

which solves to $\theta_1 \approx 1.74$; if we keep iterating, the idealized procedure will be 0.01-close to $\theta_*$ by the hundredth iteration. This is a vast improvement compared to the classical Robbins–Monro method, which remains stuck virtually forever.

Second, consider the actual nested procedure in Equation (11), where $K$ is finite. Theorem 5 shows that the procedure maintains the nice convergence and stability properties of the original procedure under certain assumptions. The assumptions in this case can be greatly simplified if we consider that for the normal distribution, the probability density function is upper-bounded by $\frac{1}{\sqrt{2\pi}}$. Hence, $L \leq 1$ and Theorem 5 suggests the following choice of hyperparameters for the nested procedure:

$$\gamma_n = \gamma_1, a = \frac{1}{(1 + \gamma_1)^2}, \text{ and } K = 50.$$  \hspace{1cm} (12)

Note in particular that this choice of parameters satisfies $K \geq 2a(1 + \gamma_1L)^2$. We can define the constants in a similar manner for arbitrary distributions from an upper bound on the probability density function. Next, we evaluate numerically the proximal Robbins–Monro procedure resulting from the aforementioned choice of hyperparameters.

### 6.2 Numerical evaluation

Here, we conduct a numerical evaluation of our proposed procedure in Equation (11), using the parameter settings of Equation (12), and compare it with the classical Robbins–Monro method.
in Equation (9). For a fair comparison, we compare $N$ iterations of the classical Robbins–Monro to our iteration executed for $n = 1, 2, \ldots, N/K$. This way, the total number of random samples (gradient observations) used by our procedure is exactly $N$ as in the classical procedure.

As before, $F(\theta)$ is the cumulative distribution function of the standard normal, $\alpha = 0.999$ and $\theta_0 = -10$. The quantity to be estimated is $\theta_\ast \approx 3.09$, for which $F(\theta_\ast) = \alpha$. For different values of $\gamma_1$ we compare the Robbins–Monro procedure with $N = 100,000$ iterations to our proposed procedure in Equation (11), with $K = 50$ and $N = \frac{100,000}{K}$, as explained earlier. For each value of $\gamma_1$, the experiment is replicated 100 times and we report a boxplot of all final estimates: $\theta_N$ for Robbins–Monro, and $\theta_{N/K}$ for the nested procedure. The results of this experiment are shown in Figure 6.2.

In the left plot, we observe that the Robbins–Monro method suffers from numerical instability, as described in the previous sections. In particular, as predicted by Proposition 6.1, when $\gamma_1$ increases beyond $\frac{\theta_\ast - \theta_0}{\alpha} \simeq 13.1$, the iterates overshoot and remain virtually stuck for all subsequent iterations. This explains why the boxplots for the Robbins–Monro method look flat for large values of $\gamma_1$; for small values of $\gamma_1$ the iterates also do not vary much because their variance depends on $\gamma_1$. In fact, there is only a small range of values for $\gamma_1$ (visually similar to [11, 15]), for which $\gamma_1$ is big enough to allow convergence, yet small enough to prevent the aforementioned numerical instability. Not shown in the figure, the estimates of Robbins–Monro are negative for very small learning rates; for example, when $\gamma_1 = 0.1$ the average estimate is $-8.8$. This is close to the starting point, $\theta_1 = -10$, and indicates that the classical procedure makes little progress when the learning rate is very small. This shows that classical Robbins–Monro approximations are extremely sensitive to specification of the learning rate values.

The results for the nested implicit procedure of Equation (11) are drastically different. In the left plot of Figure 1 we see that the estimates of the nested procedure neither overshoot nor undershoot in contrast to the classical procedure. The implicit procedure maintains a remarkable numerical stability across the entire range of learning rate values. Furthermore, the procedure is statistically efficient in that the final iterates, $\theta_{N/K}$, are centered around the true value (dashed line in figure) with small variance; this is better shown in the right subplot of Figure 1 which focuses on the nested procedure. A slight bias exists for very small or very large values of the learning rate (e.g., average estimate is 2.84 when $\gamma_1 = 0.1$), but this is an artifact of finite samples; the bias goes
Figure 1: **Left:** boxplots of 100 replications of the Robbins–Monro (RM) procedure of Equation (9) and of the nested ISA procedure of Equation (11); averages are indicated as circles and triangles respectively. **Right:** Zoom in to the boxplots of the nested ISA procedure (note the different scale on the y-axis). Left plot is in log-scale; negative values (for $\gamma_1 = 0.1$ and $\gamma_1 = 0.5$) are not shown for the Robbins–Monro method. The true parameter value, $\theta_\ast$, is depicted as the dashed horizontal line at $y = \Phi(0.999) \approx 3.09$. Both procedures start from $\theta_1 = -10$, and the nested procedure is implemented following Equation (12). We see that as $\gamma_1$ increases, the classical Robbins–Monro method overshoots and essentially remains stuck, which explains the flat boxplots. In contrast, the implicit procedure remains robust, with final iterates estimating the true value well, except for a small bias at very small or very large values of the learning rate.

away if we increase $N$. We emphasize again that the nested procedure is implemented in a fully data-driven way, by choosing its parameters using Equation (12), as prescribed by Theorem 5.

7 Concluding remarks

The theoretical and empirical results presented in this paper point to a key stability advantage of the proposed proximal Robbins–Monro method, as defined in Equation (2), over the classical
method of Robbins–Monro. Our theoretical analysis showed that such stability is obtained without sacrificing convergence or efficiency. However, the proposed method is idealized because it can only be approximately implemented. There remain several open questions regarding such approximate implementations, as presented in this paper.

First, although the implicit stochastic gradient methods described in Equation (6) are easy to implement in a wide class of models (e.g., generalized linear models, M-estimation), their application to large non-convex settings, such as neural networks, has just started to emerge (Fagan and Iyengar, 2018). In this context, the stability of implicit approximations appears to be beneficial as predicted by the theory in this paper. More work needs to be done, however, to leverage the added flexibility in designing the learning rate sequence and its robustness to misspecifications.

Second, extending the scope of nested implementations of proximal Robbins–Monro, such as the implementation in Equation (7), is intriguing, particularly because it can be applied in settings where the analytic form of the objective is not known. The nested procedure in Equation (7) can operate even when samples from the objective are only available. This introduces minimal modeling assumptions, which is desirable in many settings, such as econometric models, or sequential experimentation in clinical trials. It is also an open question whether the substantive results of the quantile estimation example of Robbins–Monro presented in Section 6.1 extend to broader applications and domains. We conjecture that this holds true.

8 Acknowledgements

This work was supported, in part, by the National Science Foundation under grants CAREER IIS-1149662 and IIS-1409177, by the Office of Naval Research under grants YIP N00014-14-1-0485 and N00014-17-1-2131, and by a Shutzer Fellowship to EMA.

References


23


### A Proofs of theorems for main method

Symbol $\| \cdot \|$ denotes the $L_2$ vector/matrix norm. We define the error random variables at the intermediate iterate as $\epsilon_n = W_{\theta_n} - h(\theta_n^+)$. Because $\theta_n^+$ is a deterministic function of $\theta_{n-1}$ it holds that $E(\epsilon_n | F_{n-1}) = 0$. The parameter space for $\theta$ is $\Theta \subseteq \mathbb{R}^p$, and is convex. For positive scalar sequences $(a_n)$ and $(b_n)$, we write $b_n = O(a_n)$ to express that $b_n \leq c a_n$, for some fixed $c > 0$, and every $n = 1, 2, \ldots$; we write $b_n = o(a_n)$ to express that $b_n / a_n \to 0$ in the limit where $n \to \infty$. Notation $b_n \downarrow 0$ means that $b_n$ is positive and decreasing towards zero. Depending on which result we state, implicit stochastic approximation operates under a combination of the following assumptions.

**Assumption 1.** It holds that $\gamma_n = \gamma_1 n^{-\gamma}, \gamma_1 > 0$ and $\gamma \in (0, 1]$. 

26
Assumption 2. Function $h$ is Lipschitz with parameter $L$, i.e., for all $\theta_1, \theta_2 \in \Theta$,

$$\|h(\theta_1) - h(\theta_2)\| \leq L\|\theta_1 - \theta_2\|.$$ 

Assumption 3. Function $h$ satisfies either

(a) $(\theta - \theta_\star)^\top h(\theta) \geq 0$, for all $\theta \in \Theta$, or, for all $n = 1, 2, \ldots$,

(b) $(\theta - \theta_\star)^\top h(\theta) > 0$, for all $\theta \in \Theta \setminus \{\theta_\star\}$, or, for all $n = 1, 2, \ldots$,

(c) $(\theta_n^+ - \theta_\star)^\top h(\theta_n^+) \geq \delta_n \|\theta_n^+ - \theta_\star\|^2$, where $\delta_n = \delta_1 n^{-\delta}$, $\delta_1 > 0$ and $0 < \gamma + \delta \leq 1$.

Assumption 4. There exists $H : \mathbb{R}^p \to \mathbb{R}$ such that $\nabla H(\theta) = h(\theta)$, for all $\theta \in \Theta$.

Assumption 5. There exists fixed $\sigma^2 > 0$ such that, for all $n = 1, 2, \ldots$,

$$\mathbb{E}(\|\varepsilon_n\|^2 | \mathcal{F}_{n-1}) \leq \sigma^2.$$ 

Assumption 6. Let $\Xi_n = \mathbb{E}(\varepsilon_n \varepsilon_n^\top | \mathcal{F}_{n-1})$, then $\|\Xi_n - \Xi\| \to 0$ for fixed positive-definite matrix $\Xi$.

Furthermore, if $\sigma_{n,s}^2 = \mathbb{E}(\mathbb{I}_{\|\varepsilon_n\|^2 \geq s/\gamma_n}\|\varepsilon_n\|^2)$, then for all $s > 0$, $\sum_{i=1}^n \sigma_{i,s}^2 = o(n)$ if $\gamma_n \propto n^{-1}$, or $\sigma_{n,s}^2 = o(1)$ otherwise.

Note about proofs. A key equation of implicit stochastic approximation is Equation (4):

$$\theta_n^+ + \gamma_n h(\theta_n^+) = \theta_{n-1}. \quad (13)$$

As this fixed point equation has a unique solution, $\theta_n^+$ is a deterministic function of $\theta_{n-1}$. By assumption, $W_{\theta_n^+} = h(\theta_n^+) + \varepsilon_n$, and so $\mathbb{E}\left(W_{\theta_n^+} | \mathcal{F}_{n-1}\right) = h(\theta_n^+)$. 

Theorem 1. Suppose that Assumptions 1, 2, 3(b), and 5 hold. Then, the iterates $\theta_n$ of the implicit stochastic approximation method in Equations (2) and (3) converge almost surely to $\theta_\star$; i.e., $\theta_n \to \theta_\star$, such that $h(\theta_\star) = 0$, almost surely.

Proof. By Equation (2):

$$\|\theta_n - \theta_\star\|^2 = \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma_n(\theta_{n-1} - \theta_\star)^\top W_{\theta_n^+} + \gamma_n^2 \|W_{\theta_n^+}\|^2. \quad (14)$$
We use decomposition $(\theta_{n-1} - \theta_*) = (\theta_n^+ - \theta_*) + (\theta_n - \theta_n^+)$, and that $\theta_{n-1} - \theta_n^+ = \gamma_n h(\theta_n^+)$ in Equation (13) to obtain:

$$R_n \triangleq \mathbb{E} \left( (\theta_{n-1} - \theta_*)^T W_{\theta_n^+} | F_{n-1} \right) = (\theta_n^+ - \theta_*)^T h(\theta_n^+) + (\theta_n - \theta_n^+)^T h(\theta_n^+),$$

$$= (\theta_n^+ - \theta_*)^T h(\theta_n^+) + \gamma_n \| h(\theta_n^+) \|^2 > 0. \quad \text{[by Assumption 3(a)]} \quad (15)$$

Taking norms in Equation (13) we obtain:

$$\| \theta_{n-1} - \theta_* \|^2 = \| \theta_n^+ - \theta_* \|^2 + 2\gamma_n h(\theta_n^+)^T (\theta_n^+ - \theta_*) + \gamma_n^2 \| h(\theta_n^+) \|^2,$$

$$> \| \theta_n^+ - \theta_* \|^2. \quad \text{[by Assumption 3(a)]} \quad (16)$$

It follows that

$$\| h(\theta_n^+) \| = \| h(\theta_n^+) - h(\theta_*) \| \leq L \| \theta_n^+ - \theta_* \| \quad \text{[by Assumption 2]}$$

$$\leq L \| \theta_{n-1} - \theta_* \|. \quad \text{[by Inequality (16)]} \quad (17)$$

Furthermore,

$$\mathbb{E} \left( \| W_{\theta_n^+} \|^2 | F_{n-1} \right) = \mathbb{E} \left( \| h(\theta_n^+) + \varepsilon_n \|^2 | F_{n-1} \right)$$

$$\leq 2 \| h(\theta_n^+) \|^2 + 2 \mathbb{E} (\| \varepsilon_n \|^2 | F_{n-1} )$$

$$\leq 2L^2 \| \theta_{n-1} - \theta_* \|^2 + 2\sigma^2. \quad \text{[by Inequality (17) and Assumption 5]} \quad (18)$$

Taking expectations in Equation (14) conditional on $F_{n-1}$, and using Equation (15) and Inequality (18) we obtain

$$\mathbb{E} (\| \theta_n - \theta_* \|^2 | F_{n-1} ) \leq (1 + 2\gamma_n^2 L^2) \| \theta_{n-1} - \theta_* \|^2 - 2\gamma_n R_n + 2\gamma_n^2 \sigma^2. \quad (19)$$

We now use an argument—due to Gladyshev (1965)— that is also applicable to the classical Robbins-Monro procedure; see, for example, Benveniste et al. (1990, Section 5.2.2), or Ljung et al. (1992, Theorem 1.9). Random variable $R_n$ is positive by Inequality (15), and $\sum \gamma_i = \infty$ and
\[ \sum \gamma_i^2 < \infty \] by Assumption 1. Therefore, we can invoke the supermartingale lemma of Robbins and Siegmund (1985) to infer that \( \|\theta_n - \theta_*\|^2 \to B > 0 \) and \( \sum \gamma_n R_n < \infty \), almost surely. If \( B \neq 0 \) then \( \lim \inf \|\theta_n - \theta_*\| > 0 \), and thus the series \( \sum \gamma_n R_n \) diverges by Inequality (15) and \( \sum \gamma_i = \infty \) (Assumption 1). This is a contradiction. Thus, \( B = 0 \).

\[ \square \]

**Theorem 2.** Suppose that Assumptions 1, 2, 3(a), 4, and 5 hold. Let \( \Gamma^2 = E(\|\theta_0 - \theta_*\|^2) + \sigma^2 \sum_{i=1}^{\infty} \gamma_i^2 + \gamma_1^2 \sigma^2 \). Then, if \( \gamma \in (2/3, 1] \), there exists \( n_{0,1} < \infty \) such that, for all \( n > n_{0,1} \),

\[ E(H(\theta_n) - H(\theta_*)) \leq \left[ \frac{2\Gamma^2}{\gamma \gamma_1} + o(1) \right] n^{-1+\gamma}. \]

If \( \gamma \in (1/2, 2/3) \), there exists \( n_{0,2} < \infty \) such that, for all \( n > n_{0,2} \),

\[ E(H(\theta_n) - H(\theta_*)) \leq \left[ \Gamma \sigma \sqrt{L} \gamma_1 + o(1) \right] n^{-\gamma/2}. \]

Otherwise, \( \gamma = 2/3 \) and there exists \( n_{0,3} < \infty \) such that, for all \( n > n_{0,3} \),

\[ E(H(\theta_n) - H(\theta_*)) \leq \left[ 3 + \sqrt{3} + 4 \gamma_1^2 L \sigma^2 / \Gamma^2 \right] 2 \gamma_1 / \Gamma^2 + o(1) \right] n^{-1/3}. \]

**Proof.** By Equation (3) and Assumption 3(a), \( \theta_n^+ + \gamma_n h(\theta_n^+) = \theta_{n-1} \) is equivalent to minimization

\[ \theta_n^+ = \arg \min_{\theta} \left\{ \frac{1}{2\gamma_n} \|\theta - \theta_{n-1}\|^2 + H(\theta) \right\}. \]

Therefore, comparing the values of the expression for \( \theta = \theta_n^+ \) and \( \theta = \theta_{n-1} \), we obtain

\[ H(\theta_n^+) + \frac{1}{2\gamma_n} \|\theta_n^+ - \theta_{n-1}\|^2 \leq H(\theta_{n-1}). \] (20)

Since \( \theta_{n-1} - \theta_n^+ = \gamma_n h(\theta_n^+) \), Inequality (20) can be written as

\[ H(\theta_{n-1}) - H(\theta_n^+) - \frac{1}{2} \gamma_n \|h(\theta_n^+))\|^2 \geq 0. \] (21)
Note that \( H(\theta_*) \leq H(\theta) \), for all \( \theta \). Thus, we have:

\[
H(\theta_n^+ \! - \! H(\theta_*) \leq h(\theta_n^+)^\top (\theta_n^+ - \theta_*) \quad \text{[by convexity Assumption 3(a)]}
\]

\[
H(\theta_n^+ \! - \! H(\theta_*) \leq \|h(\theta_n^+\| \cdot \|\theta_n^+ - \theta_*\|
\]

\[
[\mathbb{E} (H(\theta_n^+) - H(\theta_*))^2 \leq [\mathbb{E} (\|h(\theta_n^+\| \cdot \|\theta_n^+ - \theta_*\|)^2)
\]

\[
[\mathbb{E} (H(\theta_n^+) - H(\theta_*))^2 \leq \mathbb{E} (\|h(\theta_n^+\|^2) \mathbb{E} (\|\theta_n^+ - \theta_*\|^2) \quad \text{[by Cauchy-Schwarz inequality]. (22)}
\]

Furthermore,

\[
\theta_n = \theta_{n-1} - \gamma_n (h(\theta_n^+) + \varepsilon_n) = \theta_n^+ - \gamma_n\varepsilon_n. \quad \text{[by Equation (3)]}
\]

(23)

Therefore,

\[
\mathbb{E} (\|\theta_n - \theta_*\|^2) = \mathbb{E} (\|\theta_n^+ - \theta_*\|^2) - 2\gamma_n\mathbb{E} ((\theta_n^+ - \theta_*)^\top \varepsilon_n) + \gamma_n^2\mathbb{E} (\|\varepsilon_n\|^2)
\]

\[
= \mathbb{E} (\|\theta_n^+ - \theta_*\|^2) + \gamma_n^2\mathbb{E} (\|\varepsilon_n\|^2)
\]

\[
\leq \mathbb{E} (\|\theta_{n-1} - \theta_*\|^2) + \gamma_n^2\sigma^2. \quad \text{[by Inequality (16) and Assumption 5]}
\]

\[
\leq \mathbb{E} (\|\theta_0 - \theta_*\|^2) + \sigma^2\sum_{i=1}^n \gamma_i^2. \quad \text{[by induction.] (24)}
\]

For notational convenience, define \( h_n = \mathbb{E} (H(\theta_n) - H(\theta_*)) \) and \( h_n^+ = \mathbb{E} (H(\theta_n^+) - H(\theta_*)) \). It follows that \( h_n > 0, h_n^+ > 0 \), everywhere. We want to derive a bound for \( h_n \). By Equation (23), \( \theta_n^+ = \theta_n + \gamma_n\varepsilon_n \). Since \( \mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1} = 0 \), it follows from Assumption 5 that \( \mathbb{E} (\|\theta_n^+ - \theta_*\|^2) \leq \mathbb{E} (\|\theta_n - \theta_*\|^2) + \gamma_n^2\sigma^2 \). Hence, using Inequality (24) we obtain

\[
\mathbb{E} (\|\theta_n^+ - \theta_*\|^2) \leq \mathbb{E} (\|\theta_0 - \theta_*\|^2) + \sigma^2 \sum_{i=1}^\infty \gamma_i^2 + \gamma_n^2\sigma^2 \leq \Gamma^2. \quad (25)
\]

From Inequality (22) and Inequality (25), we get

\[
\mathbb{E} (\|h(\theta_n^+)\|^2) \geq \frac{1}{\Gamma^2} [\mathbb{E} (H(\theta_n^+) - H(\theta_*))^2] = \frac{1}{\Gamma^2} h_n^+ \leq \frac{1}{\Gamma^2} h_n^+ \cdot \quad (26)
\]

Furthermore, by convexity of \( H \) and Lipschitz continuity of \( h \) (Assumption 3(a)), and Assumption
5, we have that

\[ H(\theta_n) = H(\theta_n^+ - \gamma_n \varepsilon_n) \]
\[ H(\theta_n) \leq H(\theta_n^+) - \gamma_n h(\theta_n^+)^\top \varepsilon_n + \gamma_n^2 \frac{L}{2} \| \varepsilon_n \|^2 \quad [\text{by Lipschitz continuity}] \]
\[ H(\theta_n) - H(\theta_*) \leq H(\theta_n^+) - H(\theta_*) - \gamma_n h(\theta_n)^\top \varepsilon_n + \gamma_n^2 \frac{L}{2} \| \varepsilon_n \|^2 \]
\[ h_n \leq h_n^* + \gamma_n^2 \frac{L \sigma^2}{2}. \quad [\text{by taking expectations}.] \] \quad (27)

Now, in Inequality (21), we subtract \( H(\theta_*) \) from the left-hand side, take expectations, and combine with Inequality (26) to obtain

\[ h_{n-1} \geq h_n^* + \frac{1}{21^2} \gamma_n h_n^* + \frac{L \sigma^2}{2} = F_{\gamma_n}(h_n^*). \] \quad (28)

Function \( F_{\gamma_n}(x) \) defines a nondecreasing map, since its argument, \( h_n^* \), is always positive. Let \( F_{\gamma_n}^{-1} \) denote its inverse, which is also nondecreasing. Thus, we obtain \( h_n^* \leq F_{\gamma_n}^{-1}(h_{n-1}) \). Using Equation (28), we can rewrite Inequality (27) as

\[ h_n \leq F_{\gamma_n}^{-1}(h_{n-1}) + \gamma_n^2 \frac{L \sigma^2}{2}. \] \quad (29)

Inequality (29) is our main recursion, since ultimately we want to upper-bound \( h_n \). Our solution strategy is as follows. We will try to find a base sequence \((b_n)\) such that \( b_n \geq F_{\gamma_n}^{-1}(b_{n-1}) + \gamma_n^2 \frac{L \sigma^2}{2} \).

Since one can take \( b_n \) to be increasing arbitrarily, we will try to find the smallest possible sequence \((b_n)\) that satisfies the recursion. To make our analysis more tractable we will search in the family of sequences \( b_n = b_1 n^{-\beta}, \) for various values \( b_1, \beta > 0 \). Then, \( b_n \) will be an upper-bound for \( h_n \). To see this inductively, assume that \( h_{n-1} \leq b_{n-1} \) and that \( h_n \) satisfies (29). Then, \( h_n \leq F_{\gamma_n}^{-1}(h_{n-1}) + \gamma_n^2 \frac{L \sigma^2}{2} \leq F_{\gamma_n}^{-1}(b_{n-1}) + \gamma_n^2 \frac{L \sigma^2}{2} \leq b_n \), where the first inequality follows from the monotonicity of \( F_{\gamma_n} \), and the second inequality follows from definition of \( b_n \).

Now, the condition for \( b_n \) can be rewritten as \( b_{n-1} \leq F_{\gamma_n}(b_n - \gamma_n^2 \frac{L \sigma^2}{2}) \), and by definition of \( F_{\gamma_n} \) we get

\[ b_{n-1} \leq b_n - \gamma_n^2 \frac{L \sigma^2}{2} + \gamma_n \frac{1}{21^2} (b_n - \gamma_n^2 \frac{L \sigma^2}{2})^2 \] \quad (30)
Using \( b_n = b_1 n^{-\beta} \) and \( \gamma_n = \gamma_1 n^{-\gamma} \) (Assumption 1), we obtain

\[
 b_1 [(n - 1)^{-\beta} - n^{-\beta}] + \frac{L \sigma_1^2}{2} \gamma_1^2 n^{-2\gamma} + \frac{L \sigma_2^2}{2} \gamma_1^3 b_1 n^{-\beta - 3\gamma} - \frac{\gamma_1 b_1^2}{2 \Gamma^2} n^{-2\beta - \gamma} - \frac{L^2 \sigma_1^4 \gamma_1^5}{8 \Gamma^2} n^{-5\gamma} \leq 0. \quad (31)
\]

We have \((n - 1)^{-\beta} - n^{-\beta} < \frac{1}{1-\beta} n^{-1-\beta} \), for \( n > 1 \). Thus, it suffices to have

\[
 b_1 \left[ (n - 1) - \beta - n - \beta \right] + \frac{L \sigma_1^2 \gamma_1^2}{2} n^{-2\gamma} + \frac{L \sigma_2^2 \gamma_1^3 b_1}{2 \Gamma^2} n^{-\beta - 3\gamma} - \frac{\gamma_1 b_1^2}{2 \Gamma^2} n^{-2\beta - \gamma} \leq 0,
\]

where we dropped the \( n^{-5\gamma} \) term without loss of generality. The positive terms in Inequality (32) are \( n^{-1-\beta}, n^{-2\gamma}, \) and \( n^{-\beta - 3\gamma} \), and the only negative term is of order \( n^{-2\beta - \gamma} \). In order to find the largest possible \( \beta \) to satisfy (32), one needs to equate the term \( n^{-2\beta - \gamma} \) with the slowest possible term with a positive coefficient, i.e., set \( 2\beta + \gamma = \min\{1 + \beta, \beta + 3\gamma, 2\gamma\} \). However, \( \beta + 3\gamma > 1 + \beta \) and \( \beta + 3\gamma > 2\gamma \), and thus \( 2\beta + \gamma = \min\{1 + \beta, 2\gamma\} \), which implies only three cases:

(a) \( 1 + \beta < 2\gamma \), and thus \( 2\beta + \gamma = 1 + \beta \), which implies \( \beta = 1 - \gamma \). Also, \( 1 + \beta < 2\gamma \Rightarrow 2 - \gamma < 2\gamma \), and thus \( \gamma \in (2/3, 1] \). In this case, \( b_1 \) will satisfy (32) for all \( n > n_{0,1} \), for some \( n_{0,1} \), if

\[
 \frac{b_1}{1 - \beta} < \gamma_1 \frac{\gamma_1^2}{2 \Gamma^2} \iff b_1 > \frac{2 \Gamma^2}{\gamma_1 \gamma_1}. \quad (33)
\]

(b) \( 2\gamma < 1 + \beta \), and thus \( 2\beta + \gamma = 2\gamma \), which implies \( \beta = \gamma / 2 \). Also, \( 1 + \beta > 2\gamma \Rightarrow 1 + \gamma / 2 > 2\gamma \), and thus \( \gamma \in (1/2, 2/3) \). In this case, \( b_1 \) will satisfy (32) for all \( n > n_{0,2} \), for some \( n_{0,2} \), if

\[
 \frac{\gamma_1^2 \sigma_2^2}{2} < \frac{\gamma_1 b_1^2}{2 \Gamma^2} \iff b_1 > \frac{1}{\Gamma \sqrt{\gamma_1}}. \quad (34)
\]

(c) \( 2\gamma = 1 + \beta \), and thus \( 2\gamma = 1 + \beta = 2\beta + \gamma \), which solves to \( \gamma = 2/3 \) and \( \beta = 1/3 \). In this case, we need

\[
 \frac{b_1}{1 - \beta} + \frac{\gamma_1^2 \sigma_2^2}{2} < \frac{\gamma_1 b_1^2}{2 \Gamma^2} \quad (35)
\]

Because all constants are positive in Inequality (35), including \( b_1 \), it follows that

\[
 b_1 > \frac{3 + \sqrt{9 + 4 \gamma_1^2 \sigma_2^2 / \Gamma^2}}{2 \gamma_1 / \Gamma^2}. \quad (36)
\]
Remarks. The constants $n_{0,1}, n_{0,2}, n_{0,3}$ depend on the problem parameters and the desired accuracy in the bounds of Theorem 2. It is straightforward to derive exact values for them. For example, consider case $(a)$ and assume we picked $b_1$ such that $\frac{n_1 \delta}{2} > b_1 = \epsilon > 0$. Ignoring the term $n^{-3\gamma-\beta}$ (for simplicity), Inequality (32) becomes

$$
\epsilon n^{-2+\gamma} \geq \frac{L\sigma^2 \gamma_1^2}{2} n^{-2\gamma} \Rightarrow n \geq \left( \frac{L\sigma^2 \gamma_1^2}{2\epsilon} \right)^c \equiv n_{0,1},
$$

(37)

where $c = 1/(3\gamma - 2) > 0$ since $\gamma \in (2/3, 1]$. Parameter $n_{0,1}$ can therefore be set according to desired accuracy $\epsilon$. Similarly, we can derive expressions for $n_{0,2}$ and $n_{0,3}$.

Theorem 3. Suppose that Assumptions 1, 3(c), and 5 hold. Let $\zeta_n = E(\|\theta_n - \theta^*\|^2)$ and $\kappa = 1 + 2\gamma_1 \delta_1$. Then, if $\gamma + \delta < 1$, for every $n > 1$ it holds that

$$
\zeta_n \leq e^{-\log \kappa \cdot n^{1-\gamma-\delta}} \zeta_0 + \sigma^2 \gamma_1^2 \kappa \delta_1 n^{-\gamma+\delta} + O(n^{-\gamma+\delta-1}).
$$

Otherwise, if $\gamma = 1, \delta = 0$, it holds that

$$
\zeta_n \leq e^{-\log \kappa \cdot \log n} \zeta_0 + \sigma^2 \gamma_1^2 \kappa n^{-1} + O(n^{-2}).
$$

Proof. First we prove two lemmas that will be useful for Theorem 3.

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

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

(38)

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, $(1 + b_n)^{-1} \leq \exp(-Kb_n)$. Successive applications of this inequality yields Inequality (38).

Lemma 2 (Toulis and Airoldi, 2017). 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,

\[
\delta_n \triangleq \frac{1}{a_n} (a_{n-1}/b_{n-1} - a_n/b_n) \quad \text{and} \quad \zeta_n \triangleq \frac{c_n}{b_{n-1}} \frac{a_{n-1}}{a_n},
\]

(39)

and suppose that \( \delta_n \downarrow 0 \) and \( \zeta_n \downarrow 0 \). Pick a positive \( n_0 \) such that \( \delta_n + \zeta_n < 1 \) for all \( n > n_0 \).

Consider a positive sequence \( y_n > 0 \) that satisfies the recursive inequality,

\[
y_n \leq \frac{1 + c_n}{1 + b_n} y_{n-1} + a_n.
\]

(40)

Then, for every \( n > 0 \),

\[
y_n \leq K_0 \frac{a_n}{b_n} + Q^n_1 y_0 + Q^n_{n+1} (1 + c_1)^{n_0} A,
\]

(41)

where \( K_0 = (1 + b_1)(1 - \delta_{n_0} - \zeta_{n_0})^{-1} \), \( Q^n_i = \prod_{j=i}^{n} (1 + c_j)/(1 + b_j) \), and \( Q^n_i = 1 \) if \( n < i \), by definition.

**Proof.** See identical Lemma in Supplement of (Toulis and Airoldi, 2017). \( \square \)

**Corollary 1.** In Lemma 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 \( 1 < \alpha < 1 + \beta \). Then,

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

(42)

where \( n_0 > 0 \) and \( A = \sum_i a_i < \infty \).

**Proof.** In this proof, we will assume, for simplicity, \( (n - 1)^{-c} - n^{-c} \leq n^{-1-c}, c \in (0, 1) \), for every \( n > 0 \). It is straightforward to derive an appropriate bound for each value of \( c \). Furthermore, we assume \( \sum_{i=1}^{n} i^{-\gamma} \geq n^{1-\gamma} \), for every \( n > 0 \). Formally, this holds for \( n \geq n' \), where \( n' \) in practice is very small (e.g., \( n' = 14 \) if \( \gamma = 0.1 \), \( n' = 5 \) if \( \gamma = 0.5 \), and \( n' = 9 \) if \( \gamma = 0.9 \), etc.)

34
By definition,
\[
\delta_n = \frac{1}{a_n b_{n-1}} \left( \frac{a_n}{b_n} - 1 \right) = \frac{1}{a_1 n^{-\alpha} b_1} \left( (n-1)^{-\alpha+\beta} - n^{-\alpha+\beta} \right)
\]
\[
= \frac{1}{n^{-\alpha} b_1} \left[ (n-1)^{-\alpha+\beta} - n^{-\alpha+\beta} \right]
\leq \frac{1}{b_1} n^{-1+\beta}.
\] (43)

Also, \( \zeta_n = 0 \) since \( c_n = 0 \). We can take \( n_0 = \lceil \frac{2}{b_1} \rceil \), for which \( \delta_{n_0} \leq 1/2 \). Therefore,
\[
K_0 = (1 + b_1)(1 - \delta_{n_0})^{-1} \leq 2(1 + b_1); \text{ we can simply take } K_0 = 2(1 + b_1). \text{ Since } c_n = 0, \]
\[
Q^n_i = \prod_{j=i}^{n} (1 + b_i) \leq 2 \prod_{j=i}^{n} (1 + b_i).
\] (44)

Lemma 2 and Ineqs. (44) imply
\[
y_n \leq K_0 \frac{a_n}{b_n} + Q^n_1 y_0 + Q^n_{n+1} (1 + c_1)^n A \quad \text{[by Lemma 2]}
\]
\[
\leq 2 \frac{a_1 (1 + b_1)}{b_1} n^{-\alpha+\beta} + Q^n_1 [y_0 + (1 + b_1)^n A] \quad \text{[by Ineqs. (44), } c_1 = 0]\]
\[
\leq 2 \frac{a_1 (1 + b_1)}{b_1} n^{-\alpha+\beta} + \exp(-\log(1 + b_1)n^{1-\beta})[y_0 + (1 + b_1)^n A],
\] (45)

where the last inequality also follows from Ineqs. (44).

**Proof of Theorem 3.** Now we are ready to prove the main theorem. By definition (2), \( \theta_n = \theta_n^+ - \gamma_n \varepsilon_n \), and thus, by Assumption 5,
\[
\mathbb{E} \left( \| \theta_n - \theta_* \|^2 \right) \leq \mathbb{E} \left( \| \theta_n^+ - \theta_* \|^2 \right) + \gamma_n^2 \sigma^2.
\] (46)

By definition (3), \( \gamma_n h(\theta_n^+) + \theta_n^+ = \theta_{n-1} \), and thus
\[
\| \theta_{n-1} - \theta_* \|^2 = \| \theta_n^+ - \theta_* \|^2 + 2 \gamma_n (\theta_n^+ - \theta_*)^\top h(\theta_n^+) + \gamma_n^2 \| h(\theta_n^+) \|^2.
\] (47)
Therefore,

\[
\|\theta_n^+ - \theta_*\|^2 + 2\gamma_n (\theta_n^+ - \theta_*)^\top h(\theta_n^+)^2 \leq \|\theta_{n-1} - \theta_*\|^2 \\
\|\theta_n^+ - \theta_*\|^2 + 2\gamma_n \delta_n \|\theta_n^+ - \theta_*\|^2 \leq \|\theta_{n-1} - \theta_*\|^2 \quad \text{[by Assumption 3(c)]} \\
\|\theta_n^+ - \theta_*\|^2 \leq \frac{1}{1 + 2\gamma_n \delta_n} \|\theta_{n-1} - \theta_*\|^2. \tag{48}
\]

Combining Inequality (46) and Inequality (48) yields

\[
\mathbb{E}(\|\theta_n - \theta_*\|^2) = \mathbb{E}(\|\theta_n^+ - \theta_*\|^2) + \gamma_n^2 \sigma^2 \\
\leq \frac{1}{1 + 2\gamma_n \delta_n} \mathbb{E}(\|\theta_{n-1} - \theta_*\|^2) + \gamma_n^2 \sigma^2. \tag{49}
\]

The final result of Theorem 3 is obtained through a direct application of Corollary 1 on recursion (49), by setting \(y_n \equiv \mathbb{E}(\|\theta_n - \theta_*\|^2)\), \(b_n \equiv 2\gamma_n \delta_n\), and \(a_n \equiv \gamma_n^2 \sigma^2\). The case where \(\gamma = 1, \delta = 0\) only changes Inequality (44) by replacing \(\sum b_i\) with \(\log n\).

**Theorem 4.** Suppose that Assumptions 1, 2, 3(a), 5, and 6 hold. Suppose also that \((2\gamma_1 J_h(\theta_*) - I)\) is positive-definite, where \(J_h(\theta)\) is the Jacobian of \(h\) at \(\theta\), and \(I\) is the \(p \times p\) identity matrix. Then, \(\theta_n\) of implicit stochastic approximation is asymptotically normal:

\[
n\gamma/2(\theta_n - \theta_*) \to \mathcal{N}_p(0, \Sigma).
\]

The covariance matrix \(\Sigma\) is the unique solution of

\[
(\gamma_1 J_h(\theta_*) - I/2)\Sigma + \Sigma(\gamma_1 J_h(\theta_*) - I/2) = \Xi.
\]

**Proof.** Convergence of \(\theta_n \to \theta_*\) is established from Theorem 1. By definition of the implicit stochastic approximation procedure (2),

\[
\theta_n = \theta_{n-1} - \gamma_n (h(\theta_n^+) + \varepsilon_n), \quad \text{and} \\
\theta_n^+ + \gamma_n h(\theta_n^+)= \theta_{n-1}. \tag{50}
\]
We use Equation (51) and expand \( h(\cdot) \) to obtain

\[
\begin{align*}
    h(\theta^+_n) &= h(\theta_{n-1}) - \gamma_n J_h(\theta_{n-1}) h(\theta^+_n) + \epsilon_n \\
    h(\theta^+_n) &= (I + \gamma_n J_h(\theta_{n-1}))^{-1} h(\theta_{n-1}) + (I + \gamma_n J_h(\theta_{n-1}))^{-1} \epsilon_n,
\end{align*}
\]

(52)

where \( \|\epsilon_n\| = O(\gamma_n^2) \) by Theorem 3. By Lipschitz continuity of \( h(\cdot) \) (Assumption 3(a)) and the almost sure convergence of \( \theta_n \) to \( \theta_* \), it follows \( h(\theta_{n-1}) = J_h(\theta_*)(\theta_{n-1} - \theta_*) + o(1) \), where \( o(1) \) is a vector with vanishing norm. Therefore we can rewrite (52) as follows,

\[
h(\theta^+_n) = A_n(\theta_{n-1} - \theta_*) + O(\gamma_n^2),
\]

(53)

such that \( \|A_n - J_h(\theta_*)\| \to 0 \), and \( O(\gamma_n^2) \) denotes a vector with norm \( O(\gamma_n^2) \). Thus, we can rewrite (50) as

\[
\theta_n - \theta_* = (I - \gamma_n A_n)(\theta_{n-1} - \theta_*) - \gamma_n \epsilon_n + O(\gamma_n^2).
\]

(54)

The conditions for Fabian’s theorem (Fabian, 1968, Theorem 1) are now satisfied, and thus \( \theta_n - \theta_* \) is asymptotically normal with mean zero, and variance that is given in the statement of Theorem 1 by Fabian (1968).

\( \square \)

**B Proof of Theorem 5**

*Note about proofs.* We repeat the definition of the operators \( \chi_n \) and \( \xi_n \) introduced in Section 5. In particular, \( \xi_n(\theta) \) will denote the output of procedure in Equation (7), which is run for \( K \) iterations (a fixed \( K \) will be implicitly assumed). Furthermore, \( \chi_n(\theta) \) will denote the output of the same procedure in the theoretical case where \( K = \infty \). In other words, \( \chi_n \) is the proximal operator that satisfies:

\[
\chi_n(\theta) + \gamma_n h(\chi_n(\theta)) = \theta.
\]

(55)

**Lemma 3.** Suppose that Assumptions 2 and 7 hold and consider \( (x, y) \in \mathbb{R}_p^2 \), two \( p \)-component
vectors. Then, for all \( n = 1, 2, \ldots \):

(a) \( \chi_n \) is a contraction: \( \| \chi_n(x) - \chi_n(y) \| \leq \frac{1}{1 + \gamma_n \delta} \| x - y \| \).

(b) \( \| \chi_n(x) - x \| \leq \frac{\gamma_n L}{1 + \gamma_n \delta} \| x - \theta^* \| \).

\textbf{Proof.} First note that since \( h(\theta^*) = 0 \), \( \theta^* \) is a fixed point of \( \chi_n \).

(a) By definition of \( \chi_n \) in Equation (55), one can write:

\[
\chi_n(x) - \chi_n(y) = x - y + \gamma_n \left[ h(\chi_n(y)) - h(\chi_n(x)) \right].
\]

Taking the inner product with \((\chi_n(x) - \chi_n(y))\):

\[
\| \chi_n(x) - \chi_n(y) \|^2 = (x - y)^\top (\chi_n(x) - \chi_n(y)) - \gamma_n [h(\chi_n(x)) - h(\chi_n(y))]^\top (\chi_n(x) - \chi_n(y)).
\]

Using Assumption 7, we obtain:

\[
(1 + \gamma_n \delta) \| \chi_n(x) - \chi_n(y) \|^2 \leq (x - y)^\top (\chi_n(x) - \chi_n(y)),
\]

and we conclude by applying the Cauchy-Schwarz inequality to the right-hand side.

(b) We can write \( \| \chi_n(x) - x \| = \gamma_n \| h(\chi_n(x)) \| \) by definition of \( \chi_n \). Because \( h(\chi_n(\theta^*)) = 0 \):

\[
\| \chi_n(x) - x \| = \gamma_n \| h(\chi_n(x)) - h(\chi_n(\theta^*)) \|
\]

\[
\leq \gamma_n L \| \chi_n(x) - \chi_n(\theta^*) \| \leq \frac{\gamma_n L}{1 + \gamma_n \delta} \| x - \theta^* \|,
\]

where the first inequality uses Assumption 2 and the second follows from (a).

\textbf{Lemma 4.} Suppose that Assumptions 2, 5 and 7 hold. Consider the choice of parameter \( a_k = a_n, 1 \leq k \leq K \) in (7) with \( a_n \leq \frac{1}{(1 + \gamma_n L)^2} \), then:

\[
\mathbb{E} \left( \| \theta_n - \theta^+ \| \left| F_{n-1} \right\right) \leq (1 - a)^{K/2} \| \theta_{n-1} - \theta^+ \| + \sigma \gamma_n \sqrt{a}.
\]

\textbf{Proof.} Let us write \( W_\theta = h(\theta) + \varepsilon \) with \( \mathbb{E} (\| \varepsilon \|^2) \leq \sigma^2 \) by Assumption 5 and define \( g(x) = \)
\(\gamma_n h(x) + x - \theta_{n-1}\). We can write:

\[
\|x_{k+1} - \chi_n(\theta_{n-1})\|^2 = \|x_k - a_k(g(x_k) + \gamma_n \varepsilon) - \chi_n(\theta_{n-1})\|^2 \\
= \|x_k - \chi_n(\theta_{n-1})\|^2 - 2a_k(g(x_k) + \gamma_n \varepsilon)^T(x_k - \chi_n(\theta_{n-1})) \\
+ a_k^2(||g(x_k)||^2 + \gamma_n^2 \|\varepsilon\|^2 + 2g(x_k)^T\gamma_n \varepsilon).
\]

Taking expectations on both sides conditioned on \(\mathcal{F}_k\)—the \(\sigma\)-field generated by \((x_1, \ldots, x_k)\)—and noting that \(\mathbb{E}(\varepsilon|\mathcal{F}_k) = 0\) and \(\mathbb{E}(\varepsilon^2|\mathcal{F}_k) \leq \sigma^2\) we get:

\[
\delta_{k+1} \leq \delta_k - 2a_k g(x_k)^T(x_k - \chi_n(\theta_{n-1})) + a_k^2 ||g(x_k)||^2 + a_k^2 \gamma_n^2 \sigma^2,
\]

where \(\delta_k = \mathbb{E}(\|x_k - \chi_n(\theta_{n-1})\|^2)\). It follows easily from Assumptions 2 and 7 that \(g\) is \((\gamma_n L + 1)\)-Lipschitz continuous and that \((g(x) - g(y))^T(x - y) \geq \|x - y\|^2\) for all \(x\) and \(y\) in \(\mathbb{R}^p\). Furthermore, since \(g(\chi_n(\theta_{n-1})) = 0\):

\[
\delta_{k+1} \leq \left[1 - 2a_k + a_k^2(1 + \gamma_n L)^2\right] \delta_k + a_k^2 \gamma_n^2 \sigma^2.
\]

For \(a_k = a\) with \(a \leq \frac{1}{(1 + \gamma_n L)^2}\), the above recursion becomes:

\[
\delta_{k+1} \leq (1-a)\delta_k + a^2 \gamma_n^2 \sigma^2.
\]

Note that \(\chi_n(\theta_{n-1}) = \theta_{n-1}^+\) and \(x_K = \theta_n\), and \(x_1 = \theta_{n-1}\). Therefore, we obtain:

\[
\mathbb{E}(\|\theta_n - \theta_n^+\|^2|\mathcal{F}_{n-1}) \leq (1-a)^K \|\theta_{n-1} - \theta_{n-1}^+\| + a^2 \gamma_n^2 a(1 - (1-a)^K).
\]

We then obtain the statement of the lemma by applying the square root on both sides and using Jensen’s inequality on the left-hand side and subadditivity of the square root on the right-hand side.

**Lemma 5.** Let \(\theta_n'\) be an idealized procedure where: \(\theta_0' = \theta_0\), and \(\theta_n' = \chi_n(\theta_{n-1}')\). Under Assump-
tion 7, the choice of $\gamma_n = \gamma$ gives:

$$||\theta'_n - \theta_*|| \leq \frac{1}{(1 + \gamma \delta)^n} ||\theta_0 - \theta_*||.$$  

**Proof.** Because $\theta_*$ is a fixed point of $\chi_n$, we can write:

$$||\theta'_n - \theta_*|| = ||\chi_n(\theta'_{n-1}) - \chi_n(\theta_*)|| \leq \frac{1}{1 + \gamma \delta} ||\theta'_{n-1} - \theta_*||,$$

where the inequality is by Lemma 3 (a). We then conclude by solving the recursion. \qed

**Theorem 5.** Suppose that Assumptions 2, 5 and 7 hold, then the nested stochastic approximation procedure in Equation (7) with parameters $\gamma_n = \gamma$ and $a_k = 2a/K$, such that $e^{-a} < \delta/L$ and $K \geq 2a(1 + \gamma L)^2$, satisfies:

$$\mathbb{E}(||\theta_n - \theta_*||) \leq \left[ \frac{1}{(1 + \gamma \delta)^n} + \frac{C^n}{e^{-aL}} \right] ||\theta_0 - \theta_*|| + \frac{\gamma \sigma \sqrt{2a}}{(1 - C)\sqrt{K}}$$

where $C \overset{\text{def}}{=} (1 + e^{-a}\gamma L)/(1 + \gamma \delta)$.

**Proof.** We will decompose the distance between $\theta_n$ and $\theta_*$ as the distance between $\theta_n$ and $\theta_n^+$, and the distance of $\theta_n^+$ to the idealized procedure $\theta'_n$ of Lemma 5.

$$\mathbb{E} (||\theta_n - \theta'_n||) \leq \mathbb{E} (||\theta_n - \theta_n^+||) + \mathbb{E} (||\theta_n^+ - \theta'_n||) \quad \text{[triangle inequality]}$$

$$= \mathbb{E} (||\theta_n - \theta_n^+||) + \mathbb{E} (||\chi_n(\theta_{n-1}) - \chi_n(\theta'_{n-1})||) \quad \text{[by definition of $\chi_n$ in Equation (55)]}$$

$$\leq \mathbb{E} (||\theta_n - \theta_n^+||) + \frac{1}{1 + \gamma \delta} \mathbb{E}(||\theta_{n-1} - \theta'_{n-1}||) \quad \text{[by Lemma 3 (a)]}$$

$$\leq (1 - a_n)^{K/2} \mathbb{E}(||\theta_{n-1} - \chi_n(\theta_{n-1})||) + \sigma \gamma \sqrt{a_n} + \frac{1}{1 + \gamma \delta} \mathbb{E}(||\theta_{n-1} - \theta'_{n-1}||) \quad \text{[by Lemma 4]}$$

$$\leq (1 - a_n)^{K/2} \frac{\gamma L}{1 + \gamma \delta} \mathbb{E}(||\theta_{n-1} - \theta_*||) + \sigma \gamma \sqrt{a_n} + \frac{1}{1 + \gamma \delta} \mathbb{E}(||\theta_{n-1} - \theta'_{n-1}||) \quad \text{[by Lemma 3(b)]}$$

$$\leq \left( \frac{1 + (1 - a_n)^{K/2} \gamma L}{1 + \gamma \delta} \right) \mathbb{E}(||\theta_{n-1} - \theta'_{n-1}||) + \sigma \gamma \sqrt{a_n} + \frac{\gamma L}{1 + \gamma \delta} \mathbb{E}(||\theta_{n-1} - \theta_*||) \quad \text{[triangle ineq.]}$$

$$\leq \left( \frac{1 + (1 - a_n)^{K/2} \gamma L}{1 + \gamma \delta} \right) \mathbb{E}(||\theta_{n-1} - \theta'_{n-1}||) + \sigma \gamma \sqrt{a_n} + \frac{\gamma L}{(1 + \gamma \delta)^n} ||\theta_0 - \theta_*|| \quad \text{[by Lemma 5].}$$
We now choose \( a_n \) of the form \( \frac{2a}{K} \) and obtain the following recursion:

\[
\mathbb{E}
\left(
\| \theta_n - \theta'_n \|ight)
\leq
C \cdot \mathbb{E}
\left(
\| \theta_{n-1} - \theta'_{n-1} \|ight) + \sigma \gamma \frac{\sqrt{2a}}{\sqrt{K}} + \frac{\gamma L}{(1 + \gamma \delta)^n} \| \theta_0 - \theta_* \|,
\]

where \( C \) is as in the theorem statement. Observe that for our choice of parameter, \( C < 1 \). This recursion solves to:

\[
\mathbb{E}
\left(
\| \theta_n - \theta' \|ight)
\leq \frac{\gamma \sigma \sqrt{2a}}{(1 - C) \sqrt{K}} + \frac{C^n}{e^{-aL}} \| \theta_0 - \theta_* \|.
\]

Finally, using the triangle inequality and Lemma 5, we obtain:

\[
\mathbb{E}
\left(
\| \theta_n - \theta_* \|\right)
\leq \frac{\gamma \sigma \sqrt{2a}}{(1 - C) \sqrt{K}} + \left( \frac{1}{(1 + \gamma \delta)^n} + \frac{C^n}{e^{-aL}} \right) \| \theta_0 - \theta_* \|.
\]

\[\square\]

C Stability of implicit SGD: iterative least-squares regression

Let \( \theta_* \in \mathbb{R} \) be the true parameter vector of a normal model producing i.i.d. observations \( Y_n | X_n \sim \mathcal{N}(X_n \theta_*, 1) \), where \( X_n, Y_n \in \mathbb{R} \). Thus, \( \log f(Y_n; X_n, \theta) = -\frac{1}{2} (Y_n - X_n \theta)^2 \), and so \( W_{\theta} = \nabla \log f(Y_n; X_n, \theta) = (Y_n - X_n \theta)X_n \). Define \( \gamma_n = \gamma_1/n \). Then, the classical SGD procedure reduces to:

\[
\theta_n = (1 - \gamma_n X_n^2) \theta_{n-1} + \gamma_n Y_n X_n. \tag{57}
\]

Procedure (57) is known as the least mean squares filter (LMS) in signal processing, or as the Widrow-Hoff algorithm (Widrow and Hoff, 1960). The implicit SGD procedure for this problem reduces to:

\[
\theta_n = \frac{1}{1 + \gamma_n X_n^2} \theta_{n-1} + \frac{\gamma_n}{1 + \gamma_n X_n^2} Y_n X_n. \tag{58}
\]

Procedure (58) is also known as the normalized least mean squares filter (NLMS) in signal processing (Nagumo and Noda, 1967). From Equation (57) we see that it is crucial for classical SGD to have a well-specified learning rate parameter \( \gamma_1 \). For instance, assume fixed \( X_n^2 = x^2 \) for simplicity, then if \( \gamma_1 x^2 \gg 1 \) the iterate \( \theta_n \) will diverge to a value \( O(2^{n x^2}/\sqrt{\gamma_1 x^2}) \). In contrast, a very large \( \gamma_1 \) will not cause divergence in implicit SGD, but it will simply put more weight on the \( n \)th observation \( Y_n X_n \). Moreover, from a statistical perspective, implicit SGD specifies a reasonable averaging of
old and new information, by weighing the estimate and observation according to the inverse of information, \((1 + \gamma_n X_n^2)\).

### D Computation of implicit updates

At first, the computation of the implicit procedure,

\[
\theta_n = \theta_{n-1} - \gamma_n W_{\theta_n},
\]

may appear to be challenging, or even impossible. However, the implementation can actually be quite straightforward in a variety of popular models and objectives. The general idea is to exploit a special structure \(W_{\theta}\) to simplify the implicit update.

Specifically, suppose that \(W_{\theta} = s(\theta)U\), where \(s(\theta) \in \mathbb{R}\) and \(U\) is a vector that does not depend on the parameter value, \(\theta\). Then, we can write the implicit update as follows:

\[
\theta_n = \theta_{n-1} - \gamma_n s(\theta_n)U_n = \theta_{n-1} - \xi U_n,
\]

for some scalar \(\xi\). Thus, we have to solve:

\[
\gamma_n s(\theta_n) = \xi \iff \gamma_n s(\theta_{n-1} - \xi U_n) = \xi.
\]

The problem is now reduced to a one-dimensional fixed point equation for \(\xi\). In many statistical models, including generalized linear models and M-estimation, this fixed point can be efficiently solved through line search due to the structure of \(s\). For instance, Algorithm 1 of Toulis et al. (2014) provides a concrete for the family of canonical generalized linear models.