Received: 8 May 2019 | Accepted: 30 September 2020

DOI: 10.1111/rssb.12405

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

The proximal Robbins–Monro method

Panos Toulis[1] | **Thibaut Horel**[2] | **Edoardo M. Airoldi**[3]

[1]Booth School of Business, University of Chicago, Chicago, IL, USA

[2]Department of Computer Science, Harvard University, Cambridge, MA, USA

[3]Fox School of Business, Temple University, Philadelphia, PA, USA

Correspondence
Panos Toulis,
Booth School of Business, University of Chicago, Chicago, IL, USA.
Email: panos.toulis@chicagobooth.edu

Funding information
National Science Foundation, Grant/Award Number: IIS-1149662 and IIS-1409177; Office of Naval Research, Grant/Award Number: N00014-14-1-0485 and N00014-17-1-2131; Shutzer Fellowship

Abstract

The need for statistical estimation with large data sets has reinvigorated interest in iterative procedures and stochastic optimization. Stochastic approximations are at the forefront of this recent development as they yield procedures that are simple, general and fast. However, standard stochastic approximations are often numerically unstable. Deterministic optimization, in contrast, increasingly uses proximal updates to achieve numerical stability in a principled manner. A theoretical gap has thus emerged. While standard stochastic approximations are subsumed by the framework Robbins and Monro (*The annals of mathematical statistics*, 1951, pp. 400–407), there is no such framework for stochastic approximations with proximal updates. In this paper, we conceptualize a proximal version of the classical Robbins–Monro procedure. Our theoretical analysis demonstrates that the proposed procedure has important stability benefits over the classical Robbins–Monro procedure, while it retains the best known convergence rates. Exact implementations of the proximal Robbins–Monro procedure are challenging, but we show that approximate implementations lead to procedures that are easy to implement, and still dominate standard procedures by achieving numerical stability, practically without trade-offs. Moreover, approximate proximal Robbins–Monro procedures can be applied even when the objective cannot be calculated analytically, and so they generalize stochastic proximal procedures currently in use.

KEYWORDS
implicit updates, iterative estimation, proximal operators, stochastic approximation, stochastic fixed-point equations, stochastic gradient descent

1 | INTRODUCTION

In a seminal paper, Robbins and Monro (1951) considered the problem of estimating the zero θ_* of a function $h:\mathbb{R}^p \to \mathbb{R}$, where $h(\theta)$ is unknown but can be unbiasedly estimated by a function H of some random variable ξ, such that $\mathrm{E}_\xi(H(\theta,\xi)) = h(\theta)$, for fixed $\theta \in \Theta \subseteq \mathbb{R}^p$. Starting from θ_0, Robbins and Monro (1951) iteratively estimated θ_* using observations $\xi_1, \xi_2, \ldots$, as follows:

$$\theta_n = \theta_{n-1} - \gamma_n H(\theta_{n-1}, \xi_n), \tag{1}$$

where $\gamma_n \propto 1/n$, 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, $H(\theta, \xi) - h(\theta)$. Blum (1954), Ljung et al. (1992), Kushner and Yin (2003), and Borkar (2008) later strengthened this convergence result. Due to its remarkable simplicity and empirical performance, the Robbins–Monro procedure 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 procedure has attracted considerable interest in machine learning with large data sets (Bottou, 2010; Bottou et al., 2016; Moulines & Bach, 2011; Zhang, 2004) and in scalable statistical inference (Chen et al., 2016; Li et al., 2017; Su & Zhu, 2018; Toulis & Airoldi, 2015; Toulis & Airoldi, 2017). In this context, given a data set D, the Robbins–Monro procedure in Equation (1) can be applied with $h(\theta)$ being the gradient of the negative log-likelihood of θ given D and $H(\theta, \xi)$ being the gradient of the negative log-likelihood of θ calculated at a single data point sampled with replacement from D. Standard theory then implies that θ_n converges to a point θ_∞ for which $h(\theta_\infty) = 0$. In other words, θ_n converges to the maximum-likelihood estimator (or maximum a posteriori if regularization is used) given data set D. In this context, h is the gradient of a convex scalar potential, and the Robbins–Monro procedure is commonly referred to as stochastic gradient descent (SGD).

A well-known issue with the Robbins–Monro procedure, however, is its sensitivity to specification of hyperparameters, especially the learning rate γ_1. For instance, the procedure can be arbitrarily slow if γ_1 is even slightly misspecified. To illustrate, suppose that $\gamma_n = \gamma_1/n$, and there exists a scalar potential F, such that $\nabla F(\theta) = h(\theta)$, for all $\theta \in \Theta$. If F is strongly convex with parameter μ, then $\mathrm{E}\|\theta_n - \theta_*\|^2 = \mathrm{O}(n^{-\epsilon})$ if $\epsilon = 2\mu\gamma_1 < 1$ (Nemirovski et al., 2009, Section 1); (Moulines & Bach, 2011, Section 3.1). In contrast, the procedure can diverge even in the first few iterations if the learning rate is too large, especially with non-Lipschitz likelihoods as in 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, which renders the Robbins–Monro method, and all its derived procedures, inapplicable without extensive heuristic modifications (Bottou, 2012).

2 | THE PROXIMAL ROBBINS–MONRO PROCEDURE: AN OVERVIEW

In this paper, our idea to improve the stability of the Robbins–Monro procedure is to leverage the proximal point algorithm of Rockafellar (1976). Assuming a known convex potential function F such that $\nabla F = h$ and a current iterate θ_{n-1}, the proximal point update is defined as follows:

$$\theta_n^+ = \text{prox}_{\gamma_n F}(\theta_{n-1}) := \arg\min_{\theta\in\Theta}\left\{\frac{1}{2\gamma_n}\|\theta-\theta_{n-1}\|^2 + F(\theta)\right\} = \theta_{n-1} - \gamma_n h(\theta_n^+). \tag{2}$$

The intuition behind the update in Equation (2) is to penalize values of θ_n^+ which are far away from θ_{n-1}: the smaller γ_n is, the stronger the penalization is. In recent years, interest in optimization through proximal operators (i.e. function $\text{prox}_{\gamma_n F}$ above) has exploded because the resulting proximal procedures are stable and converge with minimal assumptions (Bauschke & Combettes, 2011; Parikh & Boyd, 2013). In addition, they can be applied to settings where the objective function is the sum of a smooth and a non-smooth function (as is common when using regularization), and often lead to efficient, parallelizable algorithms.

To illustrate the stability of proximal updates, we can rewrite (2) as:

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

where θ_* satisfies $h(\theta_*) = 0$. Then, we take ℓ_2 norms on both sides to obtain:

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

By convexity of F, we have $h(\theta)^\top(\theta - \theta_*) \geq 0$ for any θ, and so, unless $h(\theta_n^+) = 0$, we obtain $\|\theta_n^+ - \theta_*\|^2 < \|\theta_{n-1} - \theta_*\|^2$. This shows that the proximal update in Equation (2) contracts the iterates towards the optimal solution θ_*, hence the procedure converges. In practice, however, the proximal point algorithm is infeasible since solving the minimization problem (2) is, in general, as hard as minimizing F, or even impossible if F is unknown. Fortunately, classical results show that approximate solutions to Equation (2) are still stable if the approximation errors are small enough (Rockafellar, 1976).

In this paper, we first introduce stochastic errors in the proximal point updates, and study the properties of the resulting procedure from a probabilistic viewpoint. This procedure will follow a stylized model at first, but will later serve as a template for concrete, approximate implementations. To that end, we begin with the following stylized model: an agent has an initial estimate θ_0, then an oracle calculates the proximal update θ_1^+ according to Equation (2), which the agent then observes with error ε_1, as $\theta_1 = \theta_1^+ - \gamma_1\varepsilon_1$; then, the oracle computes the proximal update θ_2^+ given θ_1, and so on. This procedure is depicted in Table 1, and is also summarized below:

$$\theta_n^+ = \theta_{n-1} - \gamma_n h(\theta_n^+), \tag{3}$$

$$\theta_n = \theta_n^+ - \gamma_n\varepsilon_n. \quad \text{(Stochastic Proximal Point Algorithm)} \tag{4}$$

An immediate concern with the *stochastic proximal point algorithm* of Equation (4) is whether it inherits the stability properties of the classical proximal point algorithm. Indeed, in Section 3, we show that when γ_n decreases at a proper rate, and ε_n is random error with uniformly bounded variance,

TABLE 1 Stylized model of the Stochastic Proximal Point Algorithm. The update from θ_n to θ_n^+ is deterministic, and from θ_n^+ to θ_{n+1} it is stochastic

Agent	θ_0				θ_1		θ_{n-1}			θ_n
		↘		↗ (error)		…		↘	↗ (error)	
Oracle			θ_1^+						θ_n^+	

then the stochastic proximal point algorithm converges, and is numerically stable. In particular, we prove results on almost sure convergence (Theorem 1), and derive error bounds for convex (Theorem 2) and strongly convex objectives (Theorem 3). We then discuss the stability of the algorithm by analysing the dependence of expected errors, $\mathrm{E}(\|\theta_n - \theta_*\|^2)$, on the initial error, $\mathrm{E}(\|\theta_0 - \theta_*\|^2)$, and the learning rate, γ_1; see Section 3.2 for details.

Towards concrete implementations of the stochastic proximal point algorithm (4), recall our key assumption that $h(\theta)$ can be estimated by some random variable $H(\theta, \xi)$, such that $\mathrm{E}_\xi(H(\theta, \xi)) = h(\theta)$. Thus, using the same notations as in Equation (1), we can define $\varepsilon_n = H(\theta_n^+, \xi_n) - h(\theta_n^+)$, whose expected value is zero conditional on $\{\xi_i\}_{1 \le i < n}$. This lets us rewrite Equations (3) and (4) as:

$$\theta_n^+ = \theta_{n-1} - \gamma_n h(\theta_n^+), \tag{5}$$

$$\theta_n = \theta_{n-1} - \gamma_n H(\theta_n^+, \xi_n). \quad \text{Proximal Robbins–Monro Procedure} \tag{6}$$

This is a form of a *proximal Robbins–Monro procedure*, since its update in Equation (5) only differs from the classical Robbins–Monro update of Equation (1) in using the proximal update θ_n^+ instead of θ_{n-1} in $H(\theta, \xi)$. As a special case of the stochastic proximal point algorithm defined above, the proximal Robbins–Monro is also numerically stable. Its exact implementation, however, remains problematic due to the presence of the proximal term θ_n^+ in Equation (5).

We thus propose and study two different ways to implement the proximal Robbins–Monro procedure, depending on whether we can observe ξ directly, or not. First, we consider settings where ξ can be observed directly. For example, in the context of stochastic gradient descent (see Section 4.1), ξ is a random datapoint from the data set, and $H(\theta, \xi)$ is the corresponding stochastic gradient calculated at parameter value θ. In such settings, we can perform an interesting—and perhaps counterintuitive—application of the 'plug-in principle'.

In particular, taking expectations in Equation (5) yields $\mathrm{E}(\theta_n | \mathcal{F}_{n-1}) = \theta_n^+$, where $\mathcal{F}_{n-1}$ is the natural filtration, $\sigma(\xi_1, \ldots, \xi_{n-1})$. Since θ_n is an unbiased estimator of θ_n^+ we can simply plug in θ_n on the right-hand side of Equation (5) to obtain:

$$\theta_n = \theta_{n-1} - \gamma_n H(\theta_n, \xi_n). \quad \text{Incremental Proximal/Implicit Methods} \tag{7}$$

Equation (7) describes a wide family of stochastic optimization methods known as incremental proximal methods, or implicit stochastic gradient descent, depending on the field of application. We discuss related work in Section 2.1, and provide details and practical examples in Section 4.

Second, we consider the more novel and challenging setting where ξ cannot be observed directly but is only observed through stochastic function H. In particular, we no longer assume that the form of the dependency of H on ξ is known, and thus solving the implicit Equation (7) becomes infeasible. This is the case, for example when $H(\theta, \xi)$ can only be sampled successively in the context of a sequential experiment (see Section 6).

To address this challenge, we first take a full expectation in Equation (5) to obtain $\mathrm{E}(\theta_n^+ - \theta_{n-1} + \gamma_n H(\theta_n^+, \xi_n)|\mathcal{F}_{n-1}) = 0$. This implies that conditional on $\mathcal{F}_{n-1}$ we can view the proximal iterate θ_n^+ as a solution to the following characteristic equation:

$$\mathrm{E}(\theta - \theta_{n-1} + \gamma_n H(\theta, \xi_n)|\mathcal{F}_{n-1}) = 0.$$

The key idea is then to apply the classical Robbins–Monro procedure directly on this characteristic equation, leading to the following stochastic fixed point procedure:

$$\begin{aligned} w_1 &= \theta_{n-1}, \\ w_k &= w_{k-1} - a_k(\gamma_n H(w_{k-1}, \xi_k) + w_{k-1} - w_1), \quad k = 1, \ldots, K, \\ \theta_n &= w_K. \quad \text{(Proximal Stochastic Fixed-Point)} \end{aligned} \tag{8}$$

At first, it may seem that the *proximal stochastic fixed-point procedure* described in Equation (8) may have the same stability issues as classical stochastic approximation, since it is using classical updates in the inner loop. To investigate this, in Section 4.2 we analyse the convergence properties of this procedure, which is particularly challenging due to its nested structure. Our analysis reveals conditions under which the procedure can be more stable than classical Robbins–Monro. In Section 6, we also show significant benefits in numerical stability using the classical quantile regression example of Robbins and Monro (1951). The proximal stochastic fixed-point procedure of Equation (8) and its theoretical analysis, therefore constitute a key contribution of this paper. We are unaware of other stochastic proximal procedures where the random components, ξ, cannot be observed directly, or where the underlying procedure is comprised of nested stochastic fixed points.

The rest of this paper is structured as follows. In Section 2.1, we discuss related work. In Section 3, we study the stochastic proximal point algorithm, and show theoretically its appealing statistical and stability properties. We then focus on the proximal Robbins–Monro procedure as a natural instance of the stochastic proximal point algorithm. As mentioned earlier, proximal Robbins–Monro is also an 'idealized procedure, that is it is well-defined mathematically but, in general, it cannot be directly computed. We thus explore the extent to which its theoretical properties carry through to the two approximate implementations outlined above. Specifically, in Section 4, we discuss the implicit procedures described in Equation (7), with concrete examples, and in Section 4.2, we analyse the stochastic fixed-point procedure of Equation (8) with a full convergence analysis. To illustrate our theory, we present empirical results on both approximate implementations in Sections 5 and 6, all showing the stability benefits of our approach.

Remark 2.1 Technically speaking, the approximate procedure in Equation (7) is not subsumed by the idealized procedure of Equation (4), since $H(\theta_n, \xi_n)$ in the former equation is not an unbiased estimator of $h(\theta_n^+)$ in the latter as θ_n and ξ_n are dependent in Equation (7). The convergence analysis of the approximate procedures in Equations (7) and (8) therefore requires special technical results that extend beyond the idealized procedure—see Sections 4.1 and 4.2, respectively. This analysis can be conceptually understood as quantifying the bias in the errors between the approximate procedures and the idealized procedure, and arguing that this bias converges to zero sufficiently quickly. As such, the idealized procedure in Equation (4) can be thought of as describing the 'limiting behaviour' of the two approximate procedures. An interesting question for future work is to identify general conditions under which convergence of (4) can be obtained when ε_n have non-zero mean.

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), and Wei (1987) established the fundamental properties, including convergence and asymptotic laws. This work was subsequently pivotal in engineering, and particularly, in systems identification and tracking (Benveniste et al., 1990; Ljung et al., 1992); see also the excellent review by Lai (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 appeared in machine learning, usually in the form of stochastic gradient descent (SGD) methods, and especially in applications with large data sets and complex models (Bottou, 2010; Zhang, 2004).

There are mainly two lines of literature that are directly related to our work, as depicted in Table 2. In one line of work, 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, 2016); Bianchi and Hachem (2016) fall into this category—see Section 3 for a related discussion. Such procedures first update $\tilde{\theta}_n = \theta_{n-1} - \gamma_n H(\theta_{n-1}, \xi_n)$, and then define $\theta_n = \text{prox}_{\gamma_n f}(\tilde{\theta}_n)$, where f is some convex regularization function. In our work, we wish to avoid making an explicit update to ensure stability. A notable exception is presented in Section 4.2, where we discuss the stochastic fixed-point procedure in Equation (8). This procedure involves multiple explicit updates within a nested procedure, which, however, do not introduce instability thanks to the problem structure.

Another line of work involves procedures as in Equation (7), where implicit updates are directly used in the update equation. Incremental proximal procedures (Bertsekas, 2011), and implicit SGD (Toulis & Airoldi, 2017; Toulis et al., 2014; Tran et al., 2016) fall into this category. The update in (7) can be solved efficiently in many statistical models (Toulis et al., 2014, Algorithm 1), including generalized linear models. In numerical optimization and engineering, the stochastic proximal point algorithms studied by Bianchi (2016); Patrascu and Necoara (2017); Patrascu and Irofti (2019) are closely related. Interestingly, all such procedures can be viewed as the plug-in versions of the proposed proximal Robbins–Monro in Equation (5).

TABLE 2 Depiction of related work. Modern procedures, such as SGD, are instantiations of the classical procedure of Robbins and Monro (1951). The proximal Robbins–Monro procedure proposed in this paper leads to well-known implicit procedures, and also to novel procedures which can work even when the random component ξ cannot be observed directly

	Solve: $\mathbf{E}_\xi(H(\theta, \xi)) = 0$.	
samples $H(\theta, \xi)$	**Classical Robbins–Monro** $\theta_n = \theta_{n-1} - \gamma_n H(\theta_{n-1}, \xi_n)$	**Proximal Robbins–Monro** $\theta_n = \theta_{n-1} - \gamma_n H(\theta_n^+, \xi_n)$
can observe ξ directly	Stochastic gradient descent (Coraluppi & Young, 1969); (Zhang, 2004); (Bottou, 2010); natural gradients (Amari, 1998); adaptive gradients (Duchi et al., 2011)	Implicit stochastic gradients (Bertsekas, 2011); (Bianchi, 2016) (Toulis & Airoldi, 2017); stochastic proximal gradients (Singer & Duchi, 2009); (Rosasco et al., 2014)
cannot observe ξ directly	Structural breaks/tracking, quantile estimation (Benveniste et al., 1990); (Robbins & Monro, 1951)	Prox-Stochastic Fixed Point (Equation (8), Section 4.2)

From a theoretical perspective, the central contribution of this paper is the introduction of the proximal Robbins–Monro procedure as the stochastic analogue of the classical proximal point algorithm. This procedure is similar to classical Robbins–Monro procedures, but differs by using the proximal point, θ_n^+, in its iterations. We provide a full convergence analysis of the new procedures in Section 3. 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 analogue of gradient descent. Our analysis shows that the proximal Robbins–Monro procedure is more stable numerically than classical Robbins–Monro, and is also less sensitive to hyperparameter tuning.

From a practical perspective, the proximal Robbins–Monro procedure is generally infeasible. We thus develop the following two approximate implementations. First, in Section 4, we discuss an implementation based on the plug-in principle, which leads to Equation (7), and a large family of well-known implicit SGD procedures (Bertsekas, 2011; Toulis et al., 2014). These procedures are becoming increasingly popular thanks to their superior numerical performance, and their theoretical properties are now well understood (Asi & Duchi, 2019; Bertsekas, 2011; Chee & Toulis, 2018; Kulis & Bartlett, 2010; Patrascu & Irofti, 2019; Patrascu & Necoara, 2017; Ryu & Boyd, 2015; Tamar et al., 2014; Toulis & Airoldi, 2017).

The key practical novelty of our work is therefore an implementation of the proximal Robbins–Monro procedure based on the fixed point procedure of Equation (8). This procedure can operate even when we cannot observe ξ directly, including settings where $h(\theta)$ is not known analytically. In Section 4.2, we present a full convergence analysis of the procedure, which is particularly challenging due to its nested structure. In Section 6, we also illustrate significant benefits in numerical stability through the classical quantile regression example of Robbins and Monro (1951). As mentioned earlier, the fixed-point procedure and its theoretical analysis constitute a key contribution of this paper. We are unaware of other proximal methods, exact or approximate, that can be applied to pure stochastic approximation settings. Moreover, while stochastic fixed point procedures exist in the classical literature (Borkar, 2008, section 10.2), they usually lack non-asymptotic error analysis (similar to Theorem 6), and so their stability properties are unknown.

3 | THEORY OF THE STOCHASTIC PROXIMAL POINT ALGORITHM

In this section, we analyse theoretically the stochastic proximal point algorithm in Equation (4). Specifically, we study convergence (Section 3.1), asymptotic normality (Section 3.3) and non-asymptotic convergence rates (Section 3.2). We emphasize that the analysis here also applies directly to the proximal Robbins–Monro procedure of Equation (5). Later, we show that the theoretical properties studied here, and especially those that relate to numerical stability, pass onto the approximate implementations of proximal Robbins–Monro. All proofs can be found in Section 1 of the supplementary material.

We need some notation first. Symbol $\|\cdot\|$ denotes the ℓ_2 vector/matrix norm. The parameter space for θ is $\Theta \subseteq \mathbb{R}^p$, and is convex. For positive scalar sequences (a_n) and (b_n), we write $b_n = \mathrm{O}(a_n)$ to express that $b_n \leq ca_n$, for some fixed $c > 0$ and every $n = 1, 2, \ldots$; we write $b_n = \mathrm{o}(a_n)$ 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.

Existence and uniqueness of θ_n^+ as a solution of Equation (3) are guaranteed by the following assumption that we make throughout the paper without further mention:

$$\text{There exists a convex potential } F \text{ such that } \nabla F = h. \tag{9}$$

This assumption is not strictly necessary but covers most applications considered in this paper, including settings where stochastic gradient descent is applied. In Section 6, for instance we study a quantile regression problem where h is scalar-valued and non-decreasing, which ensures the existence of F and θ_n^+.

Depending on which result we state, the stochastic proximal point algorithm 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

1. $(\theta - \theta_*)^\top h(\theta) \geq 0$, for all $\theta \in \Theta$;
2. $(\theta - \theta_*)^\top h(\theta) > 0$, for all $\theta \in \Theta \setminus \{\theta_*\}$;
3. $(\theta - \theta_*)^\top h(\theta) \geq \mu\|\theta - \theta_*\|^2$, for some fixed $\mu > 0$, and all $\theta \in \Theta$.

Assumption 4 There exists fixed $\sigma^2 > 0$ such that, for all $n = 1, 2, \ldots$,

$$\mathrm{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0, \text{ and } \mathrm{E}(\|\varepsilon_n\|^2 | \mathcal{F}_{n-1}) \leq \sigma^2.$$

Assumption 5 Let $\Xi_n = \mathrm{E}(\varepsilon_n \varepsilon_n^\top | \mathcal{F}_{n-1})$, then $\|\Xi_n - \Xi\| \to 0$ for fixed positive-definite matrix Ξ. Furthermore, if $\sigma_{n,s}^2 = \mathrm{E}(\mathbb{I}_{\|\varepsilon_n\|^2 \geq s/\gamma_n} \|\varepsilon_n\|^2)$, then for all $s > 0$, $\sum_{i=1}^n \sigma_{i,s}^2 = \mathrm{o}(n)$ if $\gamma_n \propto n^{-1}$, or $\sigma_{n,s}^2 = \mathrm{o}(1)$ otherwise.

Assumption 3(a) is implied by convexity of F but is weaker since monotonicity of its gradient h is only required to hold at θ_*. Assumption 3(b) states that F is strictly convex at θ_* (in particular, it implies that θ_* is unique). Assumption 3(c) is implied by strong convexity of F but is weaker since, similarly to Assumption 3(a), the quadratic lower bound is only required to hold with respect to θ_*. Assumption 4 was introduced by Robbins and Monro (1951), and has since been standard in stochastic approximation analysis. It simply states that the stochastic errors in the observations of h have zero mean and uniformly bounded variances. It could be weakened to include slowly growing errors, σ_n^2, provided that $\sum_{i=1}^\infty \sigma_i^2 \gamma_i^2 < \infty$. Assumption 5 is the typical Lindeberg condition that is used to prove asymptotic normality of θ_n, later in this section.

Overall, our assumptions are weaker than the assumptions in classical stochastic approximation because they refer to the idealized procedures of Equations (4) and (5); compare, for example Assumptions 1–5 with the assumptions (A1)–(A4) of Borkar (2008) Section 2.1) or the assumptions by Benveniste et al. (1990) Theorem 15). In comparison to forward–backward procedures (e.g. Bianchi & Hachem, 2016; Rosasco et al., 2016), we share common assumptions on Lipschitzness of the regression function h (Assumption 2) and bounded second moments for the noise term (Assumption 4). The main difference is that forward–backward procedures require certain 'fine-tuning conditions for the learning rate. For example, assumption (A2) of Rosasco et al. (2016) requires that γ_n decays sufficiently fast with respect to the noise level in ε. Our procedure does not require such assumptions because the forward–backward steps are transposed (the implicit step happens first), which adds numerical stability, as shown in Theorem 3. In some sense, our procedure is a form of 'backward-forward splitting.

3.1 | Convergence of stochastic proximal points

In Theorem 1, we derive a proof of almost sure convergence of the stochastic proximal point algorithm, which uses the supermartingale lemma of Robbins and Siegmund (1985).

Theorem 1 *Suppose that Assumptions 1, 3(b) and 4 hold with $\gamma \in (1/2, 1]$. Then, the iterates θ_n of the stochastic proximal point algorithm of Equation (4) converge almost surely to θ_*; that is $\theta_n \to \theta_*$, such that $h(\theta_*) = 0$, almost surely.*

The conditions for almost sure convergence of the stochastic proximal point are weaker than classical stochastic approximation. For example, in classical stochastic approximations, it is typically assumed that the iterates θ_n are almost surely bounded; for example Assumption (A4) of Borkar (2008).

3.2 | Non-asymptotic analysis

In this section, we derive upper bounds for deviance of the potential function, $\mathrm{E}(F(\theta_n) - F(\theta_*))$, and the mean squared errors, $\mathrm{E}\|\theta_n - \theta_*\|^2$. This provides information on the rate of convergence, as well as the stability of the stochastic proximal point algorithm. Theorem 2 on deviance assumes non-strong convexity of F, whereas Theorem 3 on squared error assumes strong convexity.

Theorem 2 *Suppose that Assumptions 1, 2, 3(a) and 4 hold. Let $\Gamma^2 = \mathrm{E}\|\theta_0 - \theta_*\|^2 + \sigma^2 \sum_{i=1}^{\infty} \gamma_i^2 + \gamma_1^2\sigma^2$. Then, if $\gamma \in (2/3, 1]$, there exists $n_{0,1} < \infty$ such that, for all $n > n_{0,1}$, the iterate θ_n of the stochastic proximal point algorithm of Equation (4) satisfies:*

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

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

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

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

$$\mathrm{E}(F(\theta_n) - F(\theta_*)) \le \left[\frac{3 + \sqrt{9 + 4\gamma_1^3 L\sigma^2/\Gamma^2}}{2\gamma_1/\Gamma^2} + \mathrm{o}(1)\right] n^{-1/3}.$$

There are two main results in Theorem 2. First, the rates of convergence for the deviance of the stochastic proximal point algorithm are either $\mathrm{O}(n^{-1+\gamma})$ or $\mathrm{O}(n^{-\gamma/2})$, depending on the learning rate, γ. 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 $\exp(4L^2\gamma_1^2 n^{1-2\gamma})$ (Moulines & Bach, 2011, Theorem 4), which can amplify the initial conditions arbitrarily. Thus, the stochastic proximal point algorithm, and consequently, the proximal Robbins–Monro procedure, have similar

asymptotic properties to classical stochastic approximation, but they are more stable numerically, and less sensitive to initial conditions or hyperparameter tuning.

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 classical stochastic approximations with non-strongly convex objective (Moulines & 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 being a non-explicit, non-decreasing function. Our strategy 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, it is easy to show that $\tilde{b}_n(\beta)$ is an upper bound for b_n, for any β. Thus, a natural upper bound for b_n is given by $b_n \leq \arg\min_\beta \tilde{b}_n(\beta)$. This solution strategy is reminiscent of the majorization–minorization idea (Lange, 2010), and may be more broadly useful.

Theorem 3 *Suppose that Assumptions 1, 3(c) and 4 hold. Let* $\zeta_n = \mathrm{E}(\|\theta_n - \theta_*\|^2)$ *and define* $\kappa = 1 + 2\gamma_1\mu$, *where the* θ_n *is the n-th iterate of the stochastic proximal point algorithm of Equation (4). If* $\gamma < 1$, *then, for every* $n > 1$, *it holds that*

$$\zeta_n \leq \exp\{-\log\kappa \cdot n^{1-\gamma}\}\zeta_0 + \sigma^2\frac{\gamma_1\kappa}{\mu}n^{-\gamma} + \mathrm{O}(n^{-\gamma-1}).$$

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

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

There are two main results presented in Theorem 3. First, the rate of convergence for the expected errors, $\mathrm{E}(\|\theta_n - \theta_*\|^2)$, is $\mathrm{O}(n^{-\gamma})$, which matches the rate of convergence for classical 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, ζ_0, regardless of the specification of the learning rate parameter γ_1 and the Lipschitz parameter L. Another way to express this is to consider the function $\omega_n = \log(\zeta_n/\zeta_0)$ under a noise-free setting ($\sigma^2 = 0$). By studying this function with respect to γ_1 (and other problem parameters, such as convexity) we can study stability. In particular, Theorem 3 shows that $\omega_n = -\log(1 + 2\gamma_1\mu)n^{1-\gamma}$. In contrast, in classical stochastic approximation, $\omega_n = L^2\gamma_1^2 n^{1-2\gamma} - O(n^{1-\gamma})$, which can make the approximation diverge numerically if γ_1 is even slightly misspecified with respect to L (Moulines & Bach, 2011, Theorem 1). Thus, as in the non-strongly convex case of Theorem 2, the stochastic proximal point algorithm has similar asymptotic rates to classical stochastic approximation and is also more stable.

Remark 3.3 When $\gamma = 1$, 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 & Bach, 2011, Theorem 1), and is generally a feature of stochastic first-order methods. The key difference between the two procedures, as described above, is numerical stability.

3.3 | Asymptotic normality

Asymptotic distributions are well studied in classical stochastic approximation. Starting from Fabian (1968), there has been extensive work in identifying asymptotic distribution laws in stochastic approximation. In this section, we leverage this theory to show when iterates from stochastic proximal point procedures can also be 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), 4 and 5 hold, and that* $(2\gamma_1 J_h(\theta_*) - I)$ *is positive-definite, where* $J_h(\theta)$ *is the Jacobian of* h *at* θ*, and* I *is the* $p \times p$ *identity matrix. Then,* θ_n *of the stochastic proximal point algorithm of Equation (4) is asymptotically normal:*

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

The covariance matrix Σ is the unique solution of

$$(\gamma_1 \mathbf{J}_h(\theta_*) - I/2)\Sigma + \Sigma(\gamma_1 \mathbf{J}_h(\theta_*) - I/2) = \Xi.$$

A closed-form solution for Σ is possible if Ξ commutes with $J_h(\theta_*)$, such that $\Xi J_h(\theta_\star) = J_h(\theta_\star)\Xi$. Then, Σ can be derived as $\Sigma = (2\gamma_1 J_h(\theta_*) - I)^{-1}\Xi$.

Theorem 4 shows that the asymptotic distribution of θ_n is identical to the asymptotics of the classical Robbins–Monro procedure (Fabian, 1968, for example). Intuitively, in the limit as n grows, we have that $\theta_n^+ \approx \theta_{n-1} + \mathrm{O}(\gamma_n)$ with high probability, and thus the stochastic proximal point behaves like the classical approximation procedure.

4 | THE PROXIMAL ROBBINS–MONRO PROCEDURE

In the following sections, we focus on the special case of the stochastic proximal point algorithm, where an unbiased estimate $H(\theta, \xi)$ of $h(\theta)$ is available, such that $\mathrm{E}_\xi(H(\theta, \xi)) = h(\theta)$. This leads to the proximal Robbins–Monro procedure introduced in Equation (5), which we repeat here:

$$\theta_n = \theta_{n-1} - \gamma_n H(\theta_n^+, \xi_n). \tag{10}$$

This procedure is still infeasible, in general, due to the proximal term, θ_n^+, and so we consider approximate implementations. Specifically, we consider two different implementations depending on whether we have direct access to samples of ξ or not. The former leads to well-known stochastic procedures, and so our discussion will be relatively short. Later, in Section 4.2, we focus on the more challenging setting where we cannot directly sample (or observe) ξ, and analyse the resulting procedures in more detail.

4.1 | Observable ξ: Approximate implementation with the plug-in principle

As mentioned earlier, when we can observe ξ directly, we can apply the plug-in principle to implement the proximal Robbins–Monro update in (10). Specifically, by definition of the proximal update

in Equation (4) and Assumption 4 we have: $\mathrm{E}(\theta_n|\mathcal{F}_{n-1}) = \theta_{n-1} - \gamma_n h(\theta_n^+) = \theta_n^+$. Plugging-in θ_n for θ_n^+ in Equation (10) yields:

$$\theta_n = \theta_{n-1} - \gamma_n H(\theta_n, \xi_n). \tag{11}$$

We see that the iterate θ_n appears on both sides of Equation (11), and the resulting implicit update can be solved, in principle, since H is known analytically. The substitution of θ_n^+ with θ_n may naturally cause concerns about whether the stability properties of the proximal Robbins–Monro procedure carry over to the approximate implementation through the implicit procedure of Equation (11). All related work, which we summarize below, generally points to the same fact: the implicit procedure indeed remains stable, and shows superior performance to classical procedures, such as stochastic gradient descent, both in theory and practice.

Specifically, one of the most popular applications of the procedure in Equation (11) is in iterative statistical estimation, where $H(\theta, \xi) = -\nabla \log \ell(Y; X, \theta)$, and ℓ corresponds to the likelihood of a random data point $\xi = (Y, X)$, at parameter value θ. For example, if in Equation (11) we use $H(\theta_{n-1}, \xi_n)$ instead of $H(\theta_n, \xi_n)$, this amounts to classical stochastic gradient descent (SGD), which is widely popular in optimization and signal processing (Coraluppi & Young, 1969), and has been fundamental in modern machine learning with large data sets (Amari, 1998; Bottou, 2010; Bottou et al., 2016; Zhang, 2004). When we use the implicit update, as originally described in Equation (11), then the resulting procedure is known as incremental proximal method in optimization (Bertsekas, 2011), or as implicit stochastic gradient descent (ISGD) in statistics and machine learning (Toulis et al., 2014). We refer readers to (Bertsekas, 2011) and (Toulis & Airoldi, 2017) for two complementary analyses of implicit SGD, including asymptotic and non-asymptotic errors; see also (Bianchi, 2016; Bianchi et al., 2018; Salim et al., 2019) for related analyses and stronger theoretical results using monotone operator theory. One way to summarize these results on ISGD is through the following 'meta-theorem'.

Theorem 5 *(Patrascu & Necoara, 2017; Ryu & Boyd, 2015; Toulis & Airoldi, 2017) Let* $\delta_n = \mathrm{E}\left(||\theta_n - \theta_*||^2\right)$ *where* θ_n *is the n-th iterate of ISGD in Equation (11), and* $\gamma_n = \gamma_1/n$ *for some constant* $\gamma_1 > 0$*. Suppose that Assumptions 2, 3(c) and 4 hold. Then,*

$$\delta_n \leq n^{-A}\delta_0 + Bn^{-1},$$

where $A = O(\gamma_1\mu)$ *and* $B = O(\gamma_1/\mu)$ *are positive constants.*

There are two main results shown in meta-Theorem 5. First, we see the inherent trade-off in first-order stochastic procedures: a larger learning rate γ_1 helps to discount faster the bias term $O(n^{-A}\delta_0)$ at the expense of a larger variance term, $O(Bn^{-1})$, and vice versa. Second, we see that ISGD can navigate this trade-off in a numerically stable way, since the initial conditions, δ_0, cannot be amplified even when the learning rate is misspecified. In contrast, classical SGD has the same bias-variance trade-off, but the term in front of δ_0 involves additional quantities growing with n. This means that the initial conditions can be amplified arbitrarily when the learning rate is even slightly misspecified, leading to numerical divergence (Moulines & Bach, 2011, Theorem 1); for more detailed results and discussion see (Toulis & Airoldi, 2017, Theorem 2.1), (Ryu & Boyd, 2015, Theorem 4), (Patrascu & Necoara, 2017, Theorem 12), and (Pătraşcu, 2020). See also (Toulis et al., 2014, section 2.5) for comparisons in statistical efficiency between ISGD, standard SGD and maximum likelihood estimators in regular models.

To illustrate these stability advantages of ISGD, we present two examples from the literature: one example is on a linear normal model where the theoretical assumptions in Section 3 of this paper hold,

and another example on a Poisson model where the assumptions do not hold because the objective is non-Lipschitz.

4.1.1 | Example: Linear normal model

Let $\theta_* \in \mathbb{R}^p$ be the true parameters of a normal model, $y|x \sim N(x^\top\theta_*, \sigma^2)$, where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$. Let $\xi = (y, x)$ denote one datapoint, and define $H(\theta, \xi) = -(y - x^\top\theta)x$ as above. Then, classical Robbins–Monro reduces to:

$$\theta_n = (I - \gamma_n x_n x_n^\top)\theta_{n-1} + \gamma_n y_n x_n. \tag{12}$$

Procedure (12) is equivalent to classical SGD on the least-squares objective. It is also known as the least mean squares (LMS) filter in signal processing or as the Widrow–Hoff algorithm (Widrow & Hoff, 1960). From Equation (11), the ISGD update for this problem can be solved in closed form:

$$\theta_n = (I + \gamma_n x_n x_n^\top)^{-1}(\theta_{n-1} + \gamma_n x_n y_n) = \left(I - \frac{\gamma_n x_n x_n^\top}{1 + \gamma_n \|x_n\|^2}\right)(\theta_{n-1} + \gamma_n x_n y_n),$$

where the second equality follows from the Sherman–Morrison formula. Simplifying, we obtain

$$\theta_n = \theta_{n-1} + \frac{\gamma_n}{1 + \gamma_n ||x_n||^2}\left(y_n - x_n^\top\theta_{n-1}\right)x_n. \tag{13}$$

Observe that Procedure (13) corresponds to Procedure (12) with the step size γ_n replaced by $\gamma_n/(1 + \gamma_n\|x_n\|^2)$. This important normalization of the step size is known as the normalized least mean squares (NLMS) filter in signal processing (Nagumo & Noda, 1967). From Equation (12), we see that it is crucial for classical SGD to have a well-specified learning rate parameter γ_1. For instance, assume fixed $\|x_n\|^2 = c^2$, for simplicity, then if $\gamma_1 c^2 \gg 1$ the iterate θ_n of classical SGD will diverge to a value $\mathrm{O}(2^{\gamma_1 c^2}/\sqrt{\gamma_1 c^2})$ (Toulis et al., 2014, for example). In contrast, a very large γ_1 will not cause divergence in ISGD, but it will simply put more weight on the n-th observation, $y_n x_n$, as can be seen in Equation (13). Intuitively, from a statistical perspective, ISGD specifies an averaging of old and new information by weighing the estimate and observation according to the inverse of information, $(1 + \gamma_n\|x_n\|^2)$.

The stability advantages of ISGD on classical SGD in the normal model are further illustrated in the numerical simulations of Section 5.

4.1.2 | Example: Poisson regression

Following the setup in Section 4.1.1, now let $y|x \sim \mathrm{Pois}(e^{x^\top\theta_*})$, where 'Pois' denotes the Poisson density. Then, the classical SGD procedure reduces to:

$$\theta_n = \theta_{n-1} - \gamma_n(y_n - e^{x_n^\top\theta_{n-1}})x_n. \tag{14}$$

The implicit SGD procedure for this problem is equivalent to:

$$\theta_n = \theta_{n-1} - \gamma_n(y_n - e^{x_n^\top\theta_n})x_n. \tag{15}$$

The implementation of such implicit update may seem intractable, but it is actually simple and efficient. The key observation is that the implicit update θ_n defined in Equation (15) is still of the form $\theta_n = \theta_{n-1} + \lambda x_n$, for some scalar λ; the goal is therefore to identify this scalar parameter. Indeed, equating this expression of the parameter update with the expression in Equation (15) implies that λ actually solves the following fixed point equation:

$$\lambda = f_n(\lambda), \text{ where } f_n(s) = \gamma_n \left[y_n - \exp\left(x_n^\top \theta_{n-1} + s\|x_n\|^2\right) \right], s \in \mathbb{R}.$$

Since f_n is non-increasing, the search bounds for its fixed point are to be found in $[0, f_n(0)]$ or $[f_n(0), 0]$ depending on whether $f_n(0) > 0$ or $f_n(0) < 0$, respectively. This can be done efficiently via a generic root finder procedure. Moreover, the idea can be extended to the family of generalized linear models, survival models based on the Cox proportional hazard and M-estimation (Toulis & Airoldi, 2017; Toulis et al., 2014).

Regarding stability, we can see that the updates in Equation (11) are extremely sensitive to specification of γ_n due to the non-Lipschitzness of the objective in the Poisson model, while implicit SGD is not as sensitive. For example, when we start at $\theta_0 = 0$ and $\|x_1\| = O(1)$, the next iterate, θ_1, will be $O(e^{\gamma_1 y_1})$ in classical SGD, which diverges arbitrarily as γ_1 increases. On the other hand, implicit SGD has a very different behaviour thanks to the implicit update in Equation (15): when γ_1 is small such that $\gamma_1 y_1 \ll 1$, then θ_1 is $O(\gamma_1 y_1)$; but when γ_1 is large, then θ_1 asymptotes to $O(\log y_1)$.

These stability advantages of ISGD over classical SGD in the Poisson model are further illustrated in the numerical experiments of Section 5.2.

4.2 | Non-observable ξ: Approximate implementation with proximal stochastic fixed points

In this section, we consider cases where we cannot observe directly the random component ξ of $H(\theta, \xi)$. As mentioned earlier, this includes cases where the analytic form of $h(\theta)$ or $H(\theta, \xi)$ is unknown, and may only be sampled through, say, a sequential experiment. We thus present an approximate implementation of the proximal Robbins–Monro procedure based on nested stochastic approximations that can be used without any auxiliary knowledge of the estimation problem. The nested procedure is in fact a proximal form of a fixed-point stochastic approximation procedure (Borkar, 2008, section 10.2), which, however, is run only for a finite number of steps. Section 6 illustrates the benefits of the nested procedure in quantile estimation.

To begin, we first take expectations in the proximal Robbins–Monro iteration:

$$\mathrm{E}(\theta_n - \theta_{n-1} + \gamma_n H(\theta_n^+, \xi_n)) = 0 \Rightarrow \mathrm{E}(\theta_n^+ - \theta_{n-1} + \gamma_n H(\theta_n^+, \xi_n)) = 0.$$

The key idea is then to treat θ_n^+ as the solution to $\mathrm{E}_\xi(\theta - \theta_{n-1} + \gamma_n H(\theta, \xi)) = 0$, and solve this characteristic equation through a separate, standard stochastic approximation procedure. At every n-th iteration, we therefore run a Robbins–Monro procedure, w_k, for K steps as follows:

$$\begin{aligned} w_1 &= \theta_{n-1}, \\ w_k &= w_{k-1} - a_k(\gamma_n H(w_{k-1}, \xi_k) + w_{k-1} - w_1), \quad 1 < k \le K, \\ \theta_n &= w_k. \end{aligned} \tag{16}$$

At first, it may seem that this procedure is affected by the same stability issues as classical stochastic approximation. However, our convergence result that follows will show that this is not true. For intuition, note that for fixed n the sequence $(w_k)_{k\geq 1}$ is a standard Robbins–Monro procedure applied to a different minimization problem:

$$\min_{\theta\in\Theta}\left\{\frac{1}{2\gamma_n}\|\theta-\theta_{n-1}\|^2+F(\theta)\right\}. \tag{17}$$

What we gain compared to applying the classical Robbins–Monro method to h directly, is that the objective function in Equation (17) is now strongly convex, even when F is not. With this formulation, it is easy to verify that θ_n^+ is the solution to this optimization problem, so that $w_k \to \theta_n^+$. Therefore, the problem structure that we designed allows the application of explicit updates, without compromising numerical stability. We illustrate this point in Section 6.

Theorem 6 *Theorem Suppose that Assumptions 2, 4 and 3(c) hold, then the proximal stochastic fixed point procedure in Equation (16) with parameters $\gamma_n = \gamma$ and $a_k = 2a/K$, such that $e^{-a} < \mu/L$ and $K \geq 2a(1+\gamma L)^2$, satisfies:*

$$\mathrm{E}x\|\theta_n-\theta_*\| \leq C^n\|\theta_0-\theta_*\|+\frac{\gamma\sigma\sqrt{2a}}{(1-C)\sqrt{K}}$$

where $C = (1+e^{-a}\gamma L)/(1+\gamma\mu)$.

Theorem 6 shows two key results. First, the initial conditions of the nested procedure are forgotten exponentially fast at a rate which can be made arbitrarily close to $(1+\gamma\mu)^{-n}$—this was also true in the idealized procedure. Second, an approximation error smaller than ε can be obtained by choosing $n = O(\log,\frac{1}{\epsilon})$, and $K = O(\frac{1}{\epsilon^2})$, where K is the number of iterations in the inner procedure. Taken together, these choices imply a total number of gradient observations of order $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$. In comparison, under the same assumptions, the stochastic proximal point algorithm (Section 3) and the standard Robbins–Monro procedure achieve an approximation error smaller than ε using $O(\frac{1}{\epsilon^2}s)$ observations. Hence, the approximate implementation of Equation (16) incurs a small (logarithmic) overhead in terms of number of observations required to achieve a given level of accuracy. Experiments in Section 6, however, show that this overhead may be negligible in practice and that the stability benefit of procedure (16) is preserved without sacrificing accuracy, even when restricted to run for the same amount of time as other procedures.

Remark 4.1 The proof of Theorem 6 is technically challenging due to the nested nature of the procedure. This requires careful balancing of 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 are no such non-asymptotic analyses of stochastic fixed-point procedures in the literature. The proof of Theorem 6 therefore applies novel techniques, which may be of general interest. We also note that convergence of the nested procedure when F is non-strongly convex is an open question, which we leave for future work.

Remark 4.2 The nested nature of the procedure described in Equation (16) is reminiscent of the Catalyst scheme of Lin et al. (2015), which is a general acceleration technique for first-order optimization methods. Similar to the Catalyst scheme, our procedure (16) approximately

computes a proximal update at each iteration. The key difference is that we analyse how to perform this approximate computation, whereas the Catalyst scheme assumes oracle access to such computation. 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 analyse the stability of proximal updates.

In the following sections, we illustrate the use of the nested procedure of Equation (16) and the use of Theorem 6 through a simulated study on the normal model of Section 4.1.1, and the classical quantile estimation problem of Robbins and Monro (1951).

5 | SIMULATED STUDIES ON STABILITY

Here, we investigate empirically the stability of plug-in implementations of the proximal Robbins–Monro procedure presented in Section 4. Specifically, we present results for the normal linear model of Section 4.1.1, and the Poisson model of Section 4.1.2. Both experiments highlight the stability benefits of proximal Robbins–Monro procedures.

5.1 | Normal linear model

Our first simulation setting with the normal linear model is defined as follows. We define parameters $\theta_* \in \mathbb{R}^p$, $p = 6$, such that $\theta_{*,j} = e^{-j}(-1)^j$, for $j = 1, \ldots, 6$; and $y|x \sim N(x^\top \theta_*, \sigma^2)$, where $x \sim N_p(0, \Sigma)$ is a p-variate normal with $\Sigma = 2I + uu^\top$, with u a column vector with p i.i.d. uniform random variables, $U(0, 1)$; we set $\sigma^2 = 4$. We estimate θ_* recursively using the procedures of SGD and ISGD as introduced in Equations (12) and (13), respectively. We also use the stochastic fixed point method of Equation (16) to estimate θ_*. To satisfy the conditions of Theorem 6, we set:

$$a = \log(L/\mu), K = 2a(1 + \gamma_1 L)^2, a_k = 2a/K. \tag{18}$$

The parameters L, μ are estimated directly from data (in the linear model, these values correspond to the maximum and minimum eigenvalue of Σ, respectively). Finally, the learning rate for all procedures is set as $\gamma_n = \gamma_1/n$ across iterations, and we vary γ_1 in the experiment to check the sensitivity of the procedures to specification of the learning rate.

As we vary γ_1, we replicate data sets of size $N = 10{,}000$ based on the above model setup, and run the three procedures above for a total wall clock time of 1 s. For every replication, we calculate the trajectory of log mean squared error (log-MSE) of all procedures, $m_{j,n} = \|\theta_{j,n} - \theta_*\|^2$, where $\theta_{j,n}$ is the n-th iterate of procedure j within the data replication. From this series, we are interested in two summary statistics. First, the 'mean-level of the log-MSE, $m_{j,n}$, which corresponds to the level around which the log-MSE 'settles'. To calculate this number, we fit an AR(1) model on the series $m_{j,n}$, and then calculate the stationary limit $b_1/(1 - b_0)$, where b_1 is the estimated slope coefficient and b_0 is the estimated intercept in the model. Second, we are interested in the maximum value, $\max_{i:t_{j,i} \le 1}\{m_{j,i}\}$, of log-MSE across iterations, where $t_{j,i}$ is the wall clock time until iteration i for method j. This acts as a proxy for the sensitivity of the procedure.

Figure 1 shows the results of this experiment. For any value of γ_1, Figure 1(top) shows the boxplot of the mean-level log-MSE for each procedure. We see that each procedure behaves differently. Across all γ_1 values, ISGD performs best, and also remains robust. The stochastic fixed-point procedure (SFP) starts

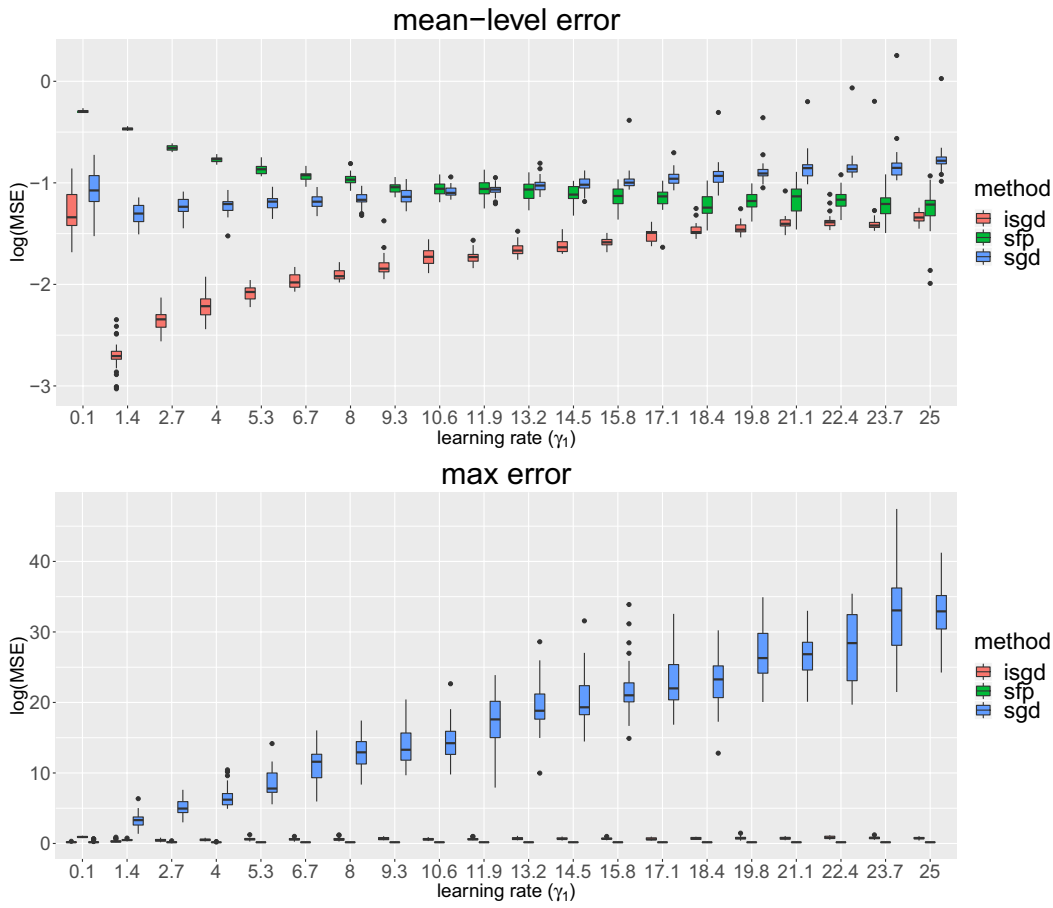

FIGURE 1 **Top:** Boxplots of mean-level log-MSE in the normal model of Section 5.1 over 50 replications of (i) the classical Robbins–Monro procedure ('sgd') of Equation (16); (ii) the nested implicit stochastic approximation procedure ('sfp') of Equation (13); and (iii) the implicit SDG procedure ('isgd') of Equation (13). **Bottom:** Boxplots of maximum log-MSE for each method. Each procedure runs for a total of 1 s of wall-clock time

from worst performance for small γ_1. However, as γ_1 increases SFP keeps improving in MSE, and its variance increases as well. This can be explained by the SFP specification in Equation (18), where larger γ_1 leads to larger K. Since we keep the computation budget (measured in wall-clock time) fixed, this means that as γ_1 increases SFP performs more inner iterations (large K) but fewer outer iterations (small n). In contrast, classical SGD is the most unstable procedure. We see that its MSE steadily increases, while the maximum MSE (Figure 1, bottom) varies widely, as γ_1 increases. This experiment illustrates a key point of our paper: classical Robbins–Monro methods are sensitive to parameter specifications, while proximal Robbins–Monro methods, and even approximate implementations of it, remain stable in a wide range of specifications.

5.2 | Poisson model

In this section, we investigate empirically the stability of plug-in implementations of proximal Robbins–Monro presented for the Poisson model of Section 4.1.1. This model has a non-Lipschitz

likelihood, and so it is not covered by our theory. This experiment is therefore meant to test the robustness of our theory and methods when our working assumptions do not hold. Our simulation setup is similar to that of the previous section. As before, we consider the true parameter vector $\theta_* \in \mathbb{R}^p$, $p = 6$, such that $\theta_{*,j} = e^{-j}$, for $j = 1, \ldots, 6$; and $y|x \sim \text{Pois}(e^{x'\theta_*})$, where x_{ij} takes values $\{0, 1, 2, 3\}$ with probabilities $\{0.4, 0.4, 0.15, 0.05\}$, respectively.

We estimate θ_* recursively using the methods of SGD and ISGD as introduced in Equations (12) and (13), respectively. We also employ the stochastic fixed point method of Equation (16) to estimate θ_*. As before we set: $a = \log(L/\mu), K = 2a(1 + \gamma_1 L)^2, a_k = 2a/K$. The parameters L, μ are estimated directly from simulated data. As expected, the estimated L is larger in the Poisson model than in the normal model, and the estimated value increases with more observations. This leads to larger values for K and fewer outer iterations for the stochastic fixed point procedure. The learning rates for all methods is set as $\gamma_n = \gamma_1/n$ across iterations. As we vary γ_1, we replicate data sets of size $N = 10{,}000$ based on the above model setup, and run each of the three procedures above for a total wall clock time of 1 s.

Figure 2 shows the results of this experiment with the Poisson model. For any value of γ_1, Figure 2 (top) shows the boxplot of the mean-level log-MSE for each method, as defined in the previous section. Since the standard SGD method is extremely unstable in this model and regularly diverges, we use Figure 2 (bottom) to zoom into the mean-level MSE values only for the proximal Robbins–Monro

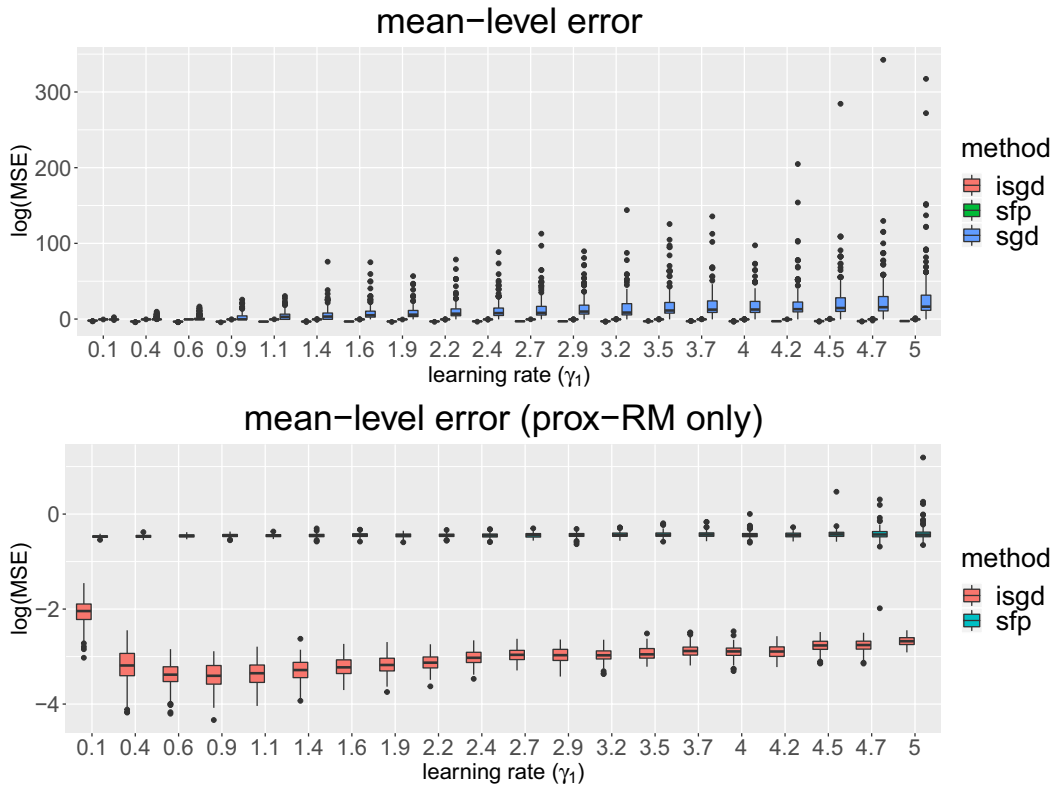

FIGURE 2 **Top:** Boxplots of mean-level log MSE over 250 replications of SGD, ISGD, and stochastic fixed point (SFP) for the Poisson model of Section 5.2. **Bottom**: 'Zoomed-in' boxplots of mean-level log-MSE only for ISGD and SFP. Each procedure runs for a total of 1 s of wall-clock time

procedures, namely, for ISGD and the stochastic fixed point procedure (SFP). We see that both ISGD and SFP are clearly more stable than classical SGD.

This experiment suggests that the proximal Robbins–Monro methods, including the approximate implementations discussed in this paper, are generally more stable than their classical counterparts, even when our theoretical assumptions do not hold. Specifically, in this example, the likelihood is non-Lipschitz, and we see that classical SGD diverges even with slight misspecifications of the learning rate. While ISGD is known to be stable in generalized linear models (Toulis et al., 2014), it is remarkable that the SFP procedure, initialized only with the theoretical values suggested by Theorem 6 and without any fine tuning, is stable in this highly non-linear model as well.

6 | APPLICATION: ITERATIVE QUANTILE ESTIMATION

In their seminal paper, Robbins and Monro (1951) applied stochastic approximations in iterative quantile estimation. In this problem, $H(\theta, \xi)$ corresponds to a sample drawn from a distribution with cumulative distribution function $Q(\theta)$. The goal is to estimate θ_* such that $Q(\theta_*) = \alpha$, for given $\alpha \in (0,1)$. A relevant application from medicine and 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 ξ with cumulative distribution function Q. An experimenter wants to find the point θ_* for which $Q(\theta_*) = \alpha$, for some fixed $\alpha \in (0, 1)$. Let $h(\theta) = Q(\theta) - \alpha$. The experimenter cannot observe ξ directly, but has only access to the outcome of an experiment, denoted by $\mathbb{I}\{\xi \leq \theta\}$, for any value of θ specified in the experiment. Robbins and Monro (1951) showed that the following iterative procedure,

$$\theta_n = \theta_{n-1} - \gamma_n H(\theta, \xi_n), \tag{19}$$

where $H(\theta, \xi) = \mathbb{I}\{\xi \leq \theta\} - \alpha$, converges to θ_∞ for which $\mathrm{E}(H(\theta_\infty, \xi)) = 0$. Consequently, it solves the characteristic equation $\mathrm{E}(\mathbb{I}\{\xi \leq \theta_\infty\}) - \alpha = Q(\theta_\infty) - \alpha = 0$. By monotonicity of Q, we obtain $\theta_\infty = \theta_*$. Despite theoretical convergence, the numerical stability of the Robbins–Monro procedure can be challenged by the following result.

Proposition 1 *Assume that $\theta_0 < \theta_*$ and that $\theta_0 + \gamma_1\alpha > \theta_*$, then for any $\epsilon > 0$ such that $\theta_0 + \gamma_1\alpha > \theta_* + \epsilon$, with probability $1 - Q(\theta_0)$, the number of iterations N_ϵ of procedure (19) required to approximate θ_* within accuracy ϵ is lower-bounded:*

$$\log N_\epsilon \geq \frac{\theta_0 + \gamma_1\alpha - \theta_* - \epsilon}{(1-\alpha)\gamma_1}. \tag{20}$$

Proof With probability $1 - Q(\theta_0)$, the first iterate of Equation (19) 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 $\theta_n \geq \theta_0 + \gamma_1\alpha - (1 - \alpha)\sum_{k=2}^{n} \frac{\gamma_1}{k} \geq \theta_0 + \gamma_1\alpha - (1 - \alpha)\gamma_1 \log n$.

Proposition 1 essentially shows that there are values of the learning rate parameter γ_1 and initial estimate θ_0 for which the classical Robbins–Monro procedure may be stuck indefinitely. For example, let Q be the standard normal distribution, and let $\alpha = 0.999$, so that $\theta_* = 3.09$ is the solution.

Suppose also that $\gamma_1 = Q'(\theta_*)^{-1} \simeq 297$, which is the learning rate value suggested by standard theory (Nemirovski et al., 2009). Let $\theta_0 = -10$ and suppose that $H(\theta_0, \xi_1) = -\alpha$. It follows that

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

From there, the Robbins–Monro procedure makes progress by at most $\gamma_i(1-\alpha) \simeq \frac{297}{i} \cdot 10^{-3}$ at each step. Thus, the number of iterations required to return back from θ_1 to a region near θ_* is at the order of e^{956}. In other words, the procedure gets stuck at large values of θ, where the derivative of the objective is negligible.

This numerical example illustrates that a misspecification of γ_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 stochastic fixed points

In the context of quantile estimation, the stochastic fixed point procedure of Equation (16) can be written as follows:

$$\begin{aligned} w_1 &= \theta_{n-1}, \\ w_k &= w_{k-1} - a_k(\gamma_n H(w_{k-1}, \xi_k) + w_k - w_1), \quad 1 < k \le K, \\ \theta_n &= w_k, \end{aligned} \tag{21}$$

where $H(\theta, \xi) = \mathbb{I}\{\xi \le \theta\} - \alpha$, $\gamma_n = \gamma_1$, $a_k = 2a/K$, and γ_1, a and K are constants to be defined.

Before presenting our numerical experiments, we discuss intuitively why the nested procedure in Equation (21) improves upon the classical Robbins–Monro method in Equation (19), and also discuss how to define the constants according to Theorem 6. We address these two issues successively. First, consider the idealized case where $K = \infty$. In this case, the iteration in Equation (21) converges to the solution of the following fixed-point equation:

$$w_\infty = \theta_{n-1} - \gamma_n(Q(w_\infty) - \alpha).$$

The next iterate, θ_n, is simply defined as $\theta_n = w_\infty$. It is easy to verify the stability of this fixed point. For example, if $\theta_{n-1} < \theta_*$, then $\theta_{n-1} < \theta_n < \theta_*$; and, conversely, if $\theta_{n-1} > \theta_*$, then $\theta_* < \theta_n < \theta_{n-1}$. That is, the idealized procedure with $K = \infty$ always pulls back in the right direction towards θ_*, and thus always makes progress towards the global solution. Convergence is also extremely fast, as shown in the proof of Theorem 6. To illustrate numerically, consider the example of the previous section where the classical Robbins–Monro procedure did not converge. Using the same numbers, at the second iteration, the idealized procedure will calculate:

$$\theta_1 = -10 - 297(Q(\theta_1) - 0.999),$$

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

Second, consider the actual nested procedure in Equation (21), where K is finite. Theorem 6 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. Hence, $L \leq 1$ and Theorem 6 suggests the following choice of hyperparameters for the nested procedure:

$$\gamma_n = \gamma_1, a = \frac{1}{(1+\gamma_1)^2}, \text{ and } K = 50. \tag{22}$$

Note in particular that this choice of parameters satisfies $K \geq 2a(1 + \gamma_1 L)^2$, as required by the theorem. 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 (approximate) proximal Robbins–Monro procedure resulting from the aforementioned choice of hyperparameters.

6.2 | Numerical evaluation

Here, we conduct a numerical evaluation of our proposed nested procedure in Equation (21), using the parameter settings of Equation (22), and compare it with the classical Robbins–Monro procedure in Equation (19). For a fair comparison, both methods run for a total of 1 s of wall-clock time. As the iteration complexity is similar for both methods, the classical Robbins–Monro procedure runs for N iterations, whereas the stochastic fixed point runs roughly for N/K outer iterations, and K inner iterations. This way, the total number of random samples (gradient observations) used by our procedure similar to those in the classical procedure.

As mentioned before, $Q(\theta)$ is here the cumulative distribution function of the standard normal, $\alpha = 0.999$ and $\theta_0 = -10$. The quantity to be estimated is $\theta_* \approx 3.09$, for which $Q(\theta_*) = \alpha$. For different values of γ_1, we compare the Robbins–Monro procedure to our proposed fixed point procedure in Equation (21), with $K = 50$. For each value of γ_1, the experiment is replicated 100 times, and we report an average of all final estimates from both procedures. The results of this experiment are shown in Figure 3.

In the left plot, we observe that the classical Robbins–Monro procedure indeed suffers from numerical instability. In particular, as predicted by Proposition 1, when γ_1 increases beyond $\frac{\theta_* - \theta_0}{\alpha} \simeq 13.1$, the iterates overshoot and remain virtually stuck for all subsequent iterations. In fact, there is only a small range of values for γ_1 (visually between values 11 and 15), for which γ_1 is big enough to allow convergence, yet small enough to prevent the aforementioned numerical instability. Not shown in the figure, the estimates from the Robbins–Monro procedure 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. Overall, these results show that classical Robbins–Monro approximations are extremely sensitive to specification of the learning rate values.

The results for proximal Robbins–Monro, as approximately implemented by the fixed point procedure of Equation (21), are drastically different. In the left subplot of Figure 3, we see that proximal Robbins–Monro neither overshoots nor undershoots in contrast to the classical procedure. We see that the proximal procedure maintains a remarkable numerical stability across the entire range of learning rate values. The procedure is also statistically efficient in that the final iterates are centred around the true value (red dashed line) with small variance around it. This is better shown in the right subplot of Figure 3, which only focuses on the estimates of the proximal Robbins–Monro. We note that a slight bias exists for very small or very large values of the learning rate. For example, the average parameter estimate is roughly 3.04 when $\gamma_1 = 0.1$. The bias goes away, however, with increased sample sizes.

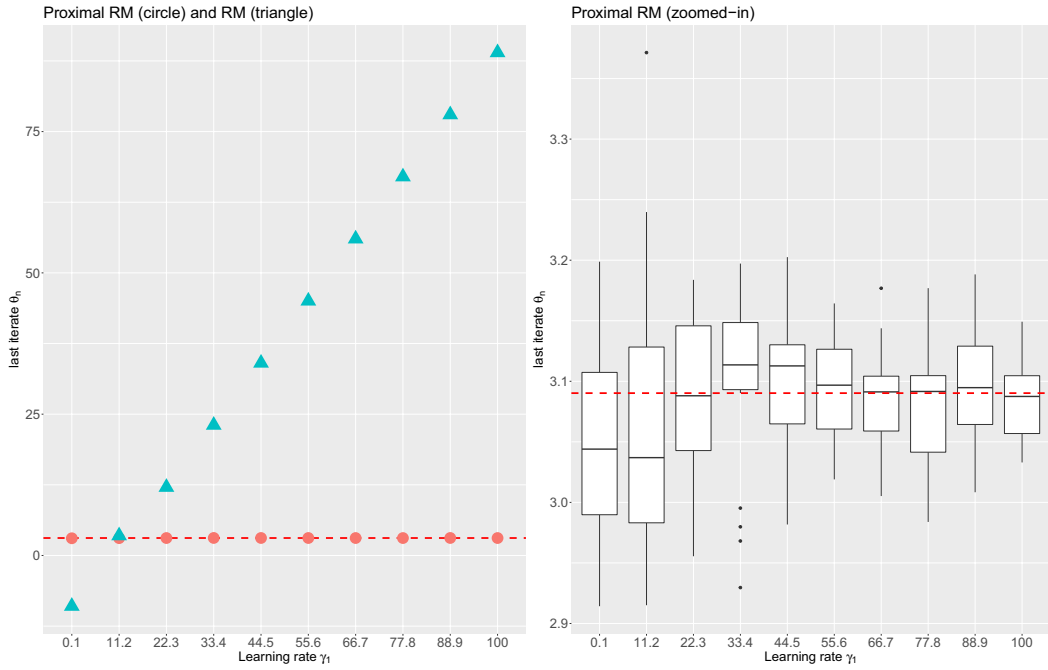

FIGURE 3 **Left:** boxplots of 100 replications of the Robbins–Monro (RM) procedure of Equation (19) and of proximal Robbins–Monro(Prox-RM), approximately implemented by Equation (21); averages of last iterates for RM and Prox-RM are indicated as triangles and circles, respectively. Each procedure runs for a total of 1 s of wall-clock time. **Right:** Zoom in to proximal RM (note the different scale on the y-axis). Left plot is in log-scale. The dashed horizontal line depicts true value, $\theta_* = 3.09$. Both procedures start from $\theta_1 = -10$, and prox-RM is implemented following Equation (22). We see that prox-RM is more stable to specification of γ_1 than classical RM [Colour figure can be viewed at wileyonlinelibrary.com]

We emphasize again that, similar to the simulation studies of Section 5, the stochastic fixed point procedure is implemented in a fully data-driven way, by choosing its parameters using Equation (22), as prescribed by Theorem 6.

7 | CONCLUDING REMARKS

The theoretical and empirical results presented in this paper point to key advantages of the proposed proximal Robbins–Monro procedure, as defined in Equation (5), over the classical procedure of Robbins–Monro. One such advantage is numerical stability. 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.

While in this paper we propose two approximate implementations that work well in general settings, there remain several open questions. First, although the implicit stochastic gradient methods described in Equation (11) are easy to implement in a wide class of models (e.g. generalized linear models, M-estimation), their application to large-scale non-convex settings, such as neural networks, has just started to emerge (Fagan & Iyengar, 2018). In this context, the stability of proximal Robbins–Monro approximations appears to be beneficial as predicted by the theory in this paper. More work needs to be done, however, to analyse these settings theoretically, and to leverage the added flexibility in designing the learning rate sequence.

Second, extending the scope of nested, fixed point implementations of proximal Robbins–Monro as in Equation (16), is interesting especially because the procedure can operate even when only samples from the objective are available. This introduces minimal modelling assumptions, which may be desirable in many settings, such as in econometric models, or in sequential experimentation of 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 provided positive empirical evidence in the simulations of Section 5, and conjecture that this holds true more generally.

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. We thank Leon Bottou, Francis Bach, Adil Salim, Pascal Bianchi, Walid Hachem and participants at the NESS conference for valuable comments and feedback. We also thank the editorial team and our two anonymous reviewers for valuable feedback and suggestions.

REFERENCES

Amari, S.-I. (1998) Natural gradient works efficiently in learning. *Neural Computation*, 10 (2), 251–276.

Asi, H. and Duchi, J.C. (2019) The importance of better models in stochastic optimization. *arXiv preprint arXiv:1903.08619*.

Bauschke, H.H. and Combettes, P.L. (2011) *Convex analysis and monotone operator theory in Hilbert spaces*. Berlin: Springer Science & Business Media.

Benveniste, A., Priouret, P. and Métivier, M. (1990) Adaptive algorithms and stochastic approximations.

Bertsekas, D.P. (2011) Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129 (2), 163–195.

Bianchi, P. (2016) Ergodic convergence of a sto chastic proximal point algorithm. *SIAM Journal on Optimization*, 26 (4), 2235–2260.

Bianchi, P. and Hachem, W. (2016) Dynamical behavior of a stochastic forward–backward algorithm using random monotone operators. *Journal of Optimization Theory and Applications*, 171(1), 90–120.

Bianchi, P., Hachem, W. and Salim, A. (2018) A constant step Forward-Backward algorithm involving random maximal monotone operators. Journal of Convex Analysis.

Blum, J.R. (1954) Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2), 382–386.

Borkar, V.S. (2008) *Stochastic approximation*. Cambridge: Cambridge Books.

Bottou, L. (2010) Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMP-STAT'2010*, pp. 177–186. Springer.

Bottou, L. (2012) Stochastic gradient tricks. *Neural Networks, Tricks of the Trade, Reloaded*, 7700, 430–445.

Bottou, L., Curtis, F.E. and Nocedal, J. (2016) Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*.

Chee, J. and Toulis, P. (2018) Convergence diagnostics for stochastic gradient descent with constant learning rate. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1476–1485.

Chen, X., Lee, J.D., Tong, X.T. and Zhang, Y. (2016) Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*.

Coraluppi, G. and Young, T.Y. (1969) Stochastic signal representation. *IEEE Transactions on Circuit Theory*, 16(2), 155–161.

Duchi, J., Hazan, E. and Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul), 2121–2159.

Fabian, V. (1968) On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4), 1327–1332.

Fagan, F. and G. Iyengar (2018) Robust implicit backpropagation. *arXiv preprint arXiv:1808.02433*.

Grieve, A.P. (1996) On likelihood and bayesian methods for interval estimation of the ld50. Statistics in Toxicology, 87–100.

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

Kushner, H.J. and Yin, G. (2003) *Stochastic approximation and recursive algorithms and applications*, Volume 35. New York: Springer Science & Business Media.

Lai, T.L. (2003) Stochastic approximation. *The Annals of Statistics*, 31 (2), 391–406.

Lange, K. (2010) *Numerical analysis for statisticians*. New York: Springer Science & Business Media.

Li, T., Liu, L., Kyrillidis, A. and Caramanis, C. (2017) Statistical inference using sgd. *arXiv preprint arXiv:1705.07477*.

Lin, H., Mairal, J. and Harchaoui, Z. (2015) A universal catalyst for first-order optimization. In: *Advances in Neural Information Processing Systems*, pp. 3384–3392.

Ljung, L., Pflug, G. and Walk, H. (1992) Stochastic approximation and optimization of random systems.

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

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

Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4), 1574–1609.

Nesterov, Y. (2004) *Introductory lectures on convex optimization*, Volume 87. Berlin: Springer Science & Business Media.

Nevel'son, M.B., Khas'minskii, R.Z. and Silver, B. (1973) *Stochastic approximation and recursive estimation*. Providence, RI: American Mathematical Society.

Parikh, N. and Boyd, S. (2013) Proximal algorithms. *Foundations and Trends in Optimization* 1 (3), 123–231.

Pătraşcu, A. (2020) New nonasymptotic convergence rates of stochastic proximal point algorithm for stochastic convex optimization. *Optimization*, 1–29. http://dx.doi.org/10.1080/02331934.2020.1761364.

Patrascu, A. and Irofti, P. (2019) Stochastic proximal splitting algorithm for stochastic composite minimization. *arXiv preprint arXiv:1912.02039*.

Patrascu, A. and Necoara, I. (2017) *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*. *arXiv preprint arXiv:1706.06297*.

Robbins, H. and Monro, S. (1951) A stochastic approximation method. *The Annals of Mathematical Statistics*, 25, 400–407.

Robbins, H. and Siegmund, D. (1985) A convergence theorem for non negative almost supermartingales and some applications. In: *Herbert Robbins Selected Papers*, pp. 111–135. Springer.

Rockafellar, R.T. (1976) Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* 14 (5), 877–898.

Rosasco, L., Villa, S. and Vũ, B.C. (2014) Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*.

Rosasco, L., Villa, S. and Vũ, B.C. (2016) A stochastic inertial forward–backward splitting algorithm for multivariate monotone inclusions. *Optimization* 65(6), 1293–1314.

Ruppert, D. (1988) Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Ryu, E.K. and Boyd, S. (2015) Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, manuscript*.

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

Salim, A., Bianchi, P. and Hachem, W. (2019) Snake: A stochastic proximal gradient algorithm for regularized problems over large graphs. *IEEE Transactions on Automatic Control*, 64, 1832–1847.

Singer, Y. and Duchi, J.C. (2009) Efficient learning using forward-backward splitting. In: *Advances in Neural Information Processing Systems*, pp. 495–503.

Su, W. and Zhu, Y. (2018) Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*.

Tamar, A., Toulis, P., Mannor, S. and Airoldi, E.M. (2014) Implicit temporal differences. *arXiv preprint arXiv:1412.6734*.

Toulis, P. and Airoldi, E.M. (2015) Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Statistics and Computing* 25(4), 781–795.

Toulis, P. and Airoldi, E.M. (2017) Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4), 1694–1727.

Toulis, P., Airoldi, E. and Rennie, J. (2014) Statistical analysis of stochastic gradient methods for generalized linear models. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 667–675.

Tran, D., Toulis, P. and Airoldi, E. (2016) Towards stability and optimality in stochastic gradient descent. In: *Artificial Intelligence and Statistics*, pp. 12901298.

Wei, C. (1987) Multivariate adaptive stochastic approximation. *The Annals of Statistics*, 15, 1115–1130.

Widrow, B. and Hoff, M.E. (1960) Adaptive switching circuits. Defense Technical Information Center.

Zhang, T. (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *Proceedings of the twenty-first international conference on Machine learning*, pp. 116. ACM.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Toulis P, Horel T, Airoldi EM. The proximal Robbins–Monro method. *J R Stat Soc Series B*. 2021;83:188–212. https://doi.org/10.1111/rssb.12405