Web-based supporting materials for *"The Proximal Robbins–Monro Method"*

Panos Toulis
Booth School of Business, University of Chicago
Thibaut Horel
Department of Computer Science, Harvard University
Edoardo M. Airoldi
Fox School of Business, Temple University

1. Proofs of theorems for main method

We recall the procedure below:

$$\theta_n^+ = \theta_{n-1} - \gamma_n h(\theta_n^+), \tag{1}$$
$$\theta_n = \theta_n^+ - \gamma_n \varepsilon_n. \qquad \text{(Stochastic Proximal Point Algorithm)} \tag{2}$$

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 $b_n \le c a_n$, 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 (1) is guaranteed by the following assumption, that we make throughout the paper without further mention:

$$\text{There exists a convex potential } F \text{ such that } \nabla F = h. \tag{3}$$

This assumption is not strictly necessary but covers most applications, including settings where stochastic gradient descent is applied. In Section 6 of the paper, 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 $\theta_1, \theta_2 \in \Theta$,*

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

Assumption 3. *Function h satisfies either*

(a) $(\theta - \theta_\star)^\top h(\theta) \ge 0$, *for all* $\theta \in \Theta$;

(b) $(\theta-\theta_\star)^\top h(\theta) > 0$, *for all* $\theta \in \Theta \setminus \{\theta_\star\}$;

(c) $(\theta-\theta_\star)^\top h(\theta) \geq \mu\|\theta-\theta_\star\|^2$, *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* $\Xi_n = \mathrm{E}\left(\varepsilon_n\varepsilon_n^\top|\mathcal{F}_{n-1}\right)$, *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.*

Note about proofs. A key equation of implicit stochastic approximation is Equation (1):

$$\theta_n^+ + \gamma_n h(\theta_n^+) = \theta_{n-1}. \tag{4}$$

As this fixed-point equation has a unique solution, θ_n^+ is a *deterministic function* of θ_{n-1}.

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 (2) 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) and using Assumption 4, we have:

$$\mathrm{E}\left(\|\theta_n - \theta_\star\|^2|\mathcal{F}_{n-1}\right) \leq \|\theta_n^+ - \theta_\star\|^2 + \gamma_n^2\sigma^2\,.$$

Taking norms in (4):

$$\|\theta_n^+ - \theta_\star\|^2 = \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma_n \cdot h(\theta_n^+)^\top(\theta_n^+ - \theta_\star) - \gamma_n^2\|h(\theta_n^+\|^2 \tag{5}$$

which together with the previous inequality implies:

$$\begin{aligned}\mathrm{E}\left(\|\theta_n - \theta_\star\|^2|\mathcal{F}_{n-1}\right) &\leq \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma_n \cdot h(\theta_n^+)^\top(\theta_n^+ - \theta_\star) - \gamma_n^2\|h(\theta_n^+\|^2 + \gamma_n^2\sigma^2 \\ &\leq \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma_n \cdot h(\theta_n^+)^\top(\theta_n^+ - \theta_\star) + \gamma_n^2\sigma^2\,.\end{aligned}$$

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 $R_n = h(\theta_n^+)^\top(\theta_n^+ - \theta_\star)$ is positive by Assumption 3(b), 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 $\|\theta_n - \theta_\star\|^2 \to B > 0$ and $\sum \gamma_n R_n < \infty$, almost surely. If $B \neq 0$ then $\liminf \|\theta_n - \theta_\star\| > 0$, and thus the series $\sum_n \gamma_n R_n$ diverges sinc $\sum \gamma_i = \infty$ (Assumption 1). This is a contradiction. Thus, $B = 0$.

THEOREM 2. *Suppose that Assumptions 1, 2, 3(a), and 4 hold. Let* $\Gamma^2 = \mathrm{E}\|\theta_0 - \theta_\star\|^2 + \sigma^2\sum_{i=1}^\infty \gamma_i^2 + \gamma_1^2\sigma^2$. *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 (2) satisfies:*

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

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

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

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

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

Proof. Note that $\theta_n^+ + \gamma_n h(\theta_n^+) = \theta_{n-1}$ is equivalent to $\theta_n^+ = \arg\min_\theta\{\frac{1}{2\gamma_n}\|\theta - \theta_{n-1}\|^2 + F(\theta)\}$. Therefore, comparing the values of the expression for $\theta = \theta_n^+$ and $\theta = \theta_{n-1}$, we obtain

$$F(\theta_n^+) + \frac{1}{2\gamma_n}\|\theta_n^+ - \theta_{n-1}\|^2 \leq F(\theta_{n-1}). \tag{6}$$

Since $\theta_{n-1} - \theta_n^+ = \gamma_n h(\theta_n^+)$, Inequality (6) can be written as

$$F(\theta_{n-1}) - F(\theta_n^+) - \frac{1}{2}\gamma_n\|h(\theta_n^+)\|^2 \geq 0. \tag{7}$$

Note that $F(\theta_\star) \leq F(\theta)$, for all θ. Thus, we have:

$$\begin{aligned}
F(\theta_n^+) - F(\theta_\star) &\leq h(\theta_n^+)^\top(\theta_n^+ - \theta_\star) \qquad [\textit{by convexity Assumption 3(a)}]\\
F(\theta_n^+) - F(\theta_\star) &\leq \|h(\theta_n^+)\| \cdot \|\theta_n^+ - \theta_\star\|\\
[\mathrm{E}(F(\theta_n^+) - F(\theta_\star))]^2 &\leq [\mathrm{E}(\|h(\theta_n^+)\| \cdot \|(\theta_n^+ - \theta_\star\|)]^2\\
[\mathrm{E}(F(\theta_n^+) - F(\theta_\star))]^2 &\leq \mathrm{E}(\|h(\theta_n^+)\|^2)\mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) \qquad [\textit{by Cauchy-Schwarz inequality}].
\end{aligned} \tag{8}$$

Therefore,

$$\begin{aligned}
\mathrm{E}(\|\theta_n - \theta_\star\|^2) &= \mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) - 2\gamma_n\mathrm{E}((\theta_n^+ - \theta_\star)^\top\varepsilon_n) + \gamma_n^2\mathrm{E}(\|\varepsilon_n\|^2)\\
&= \mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) + \gamma_n^2\mathrm{E}(\|\varepsilon_n\|^2)\\
&\leq \mathrm{E}(\|\theta_{n-1} - \theta_\star\|^2) + \gamma_n^2\sigma^2. \qquad [\textit{by Inequality (5) and Assumption 4}]\\
&\leq \mathrm{E}(\|\theta_0 - \theta_\star\|^2) + \sigma^2\sum_{i=1}^n \gamma_i^2. \qquad [\textit{by induction.}]
\end{aligned} \tag{9}$$

For brevity, define $h_n = \mathrm{E}(F(\theta_n) - F(\theta_\star))$ and $h_n^+ = \mathrm{E}(F(\theta_n^+) - F(\theta_\star))$. It follows that $h_n > 0, h_n^+ > 0$, everywhere. We want to derive a bound for h_n. Since $\mathrm{E}(\varepsilon_n|\mathcal{F}_{n-1}) = 0$, it follows from Assumption 4 that $\mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) \leq \mathrm{E}(\|\theta_n - \theta_\star\|^2) + \gamma_n^2\sigma^2$. Using Inequality (9), we get

$$\mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) \leq \mathrm{E}(\|\theta_0 - \theta_\star\|^2) + \sigma^2\sum_{i=1}^\infty \gamma_i^2 + \gamma_n^2\sigma^2 \leq \Gamma^2. \tag{10}$$

From Inequality (8) and Inequality (10), we get

$$\mathrm{E}(\|h(\theta_n^+)\|^2) \geq \frac{1}{\Gamma^2}[\mathrm{E}(F(\theta_n^+) - F(\theta_\star))]^2 = \frac{1}{\Gamma^2}{h_n^+}^2. \tag{11}$$

Furthermore, by convexity of F, Assumption 3(a), and Assumption 4, we have that

$$
\begin{aligned}
F(\theta_n) &= F(\theta_n^+ - \gamma_n \varepsilon_n) \\
F(\theta_n) &\leq F(\theta_n^+) - \gamma_n h(\theta_n^+)^\top \varepsilon_n + \gamma_n^2 \frac{L}{2} \|\varepsilon_n\|^2 \quad \text{[by Lipschitz continuity]} \\
F(\theta_n) - F(\theta_\star) &\leq F(\theta_n^+) - F(\theta_\star) - \gamma_n h(\theta_n^+)^\top \varepsilon_n + \gamma_n^2 \frac{L}{2} \|\varepsilon_n\|^2 \\
h_n &\leq h_n^+ + \gamma_n^2 \frac{L\sigma^2}{2}. \quad \text{[by taking expectations.]}
\end{aligned} \tag{12}
$$

Now, in Inequality (7), we substract $F(\theta_\star)$ from the left-hand side, take expectations, and combine with Inequality (11) to obtain

$$
h_{n-1} \geq h_n^+ + \frac{1}{2\Gamma^2} \gamma_n {h_n^+}^2 \triangleq R_{\gamma_n}(h_n^+). \tag{13}
$$

Function $R_{\gamma_n}(x)$ defines a nondecreasing map, since its argument, h_n^+, is always positive. Let $R_{\gamma_n}^{-1}$ denote its inverse, which is also nondecreasing. Thus, we obtain $h_n^+ \leq R_{\gamma_n}^{-1}(h_{n-1})$. Using Equation (13), we can rewrite Inequality (12) as

$$
h_n \leq R_{\gamma_n}^{-1}(h_{n-1}) + \gamma_n^2 \frac{L\sigma^2}{2}. \tag{14}
$$

Inequality (14) 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 $b_n \geq R_{\gamma_n}^{-1}(b_{n-1}) + \gamma_n^2 \frac{L\sigma^2}{2}$. 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 (14). Then, $h_n \leq R_{\gamma_n}^{-1}(h_{n-1}) + \gamma_n^2 \frac{L\sigma^2}{2} \leq R_{\gamma_n}^{-1}(b_{n-1}) + \gamma_n^2 \frac{L\sigma^2}{2} \leq b_n$, where the first inequality follows from the monotonicity of R_{γ_n}, and the second inequality follows from definition of b_n.

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

$$
b_{n-1} \leq b_n - \gamma_n^2 \frac{L\sigma^2}{2} + \gamma_n \frac{1}{2\Gamma^2} (b_n - \gamma_n^2 \frac{L\sigma^2}{2})^2 \tag{15}
$$

Using $b_n = b_1 n^{-\beta}$ and $\gamma_n = \gamma_1 n^{-\gamma}$ (Assumption 1), we obtain

$$
b_1[(n-1)^{-\beta} - n^{-\beta}] + \frac{L\sigma^2\gamma_1^2}{2} n^{-2\gamma} + \frac{L\sigma^2\gamma_1^3 b_1}{2\Gamma^2} n^{-\beta-3\gamma} - \frac{\gamma_1 b_1^2}{2\Gamma^2} n^{-2\beta-\gamma} - \frac{L^2\sigma^4\gamma_1^5}{8\Gamma^2} n^{-5\gamma} \leq 0. \tag{16}
$$

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

$$
\frac{b_1}{1-\beta} n^{-1-\beta} + \frac{L\sigma^2\gamma_1^2}{2} n^{-2\gamma} + \frac{L\sigma^2\gamma_1^3 b_1}{2\Gamma^2} n^{-\beta-3\gamma} - \frac{\gamma_1 b_1^2}{2\Gamma^2} n^{-2\beta-\gamma} \leq 0, \tag{17}
$$

where we dropped the $n^{-5\gamma}$ term without loss of generality. The positive terms in Inequality (17) 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 β to satisfy (17), 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 (17) for all $n > n_{0,1}$, for some $n_{0,1}$, if

$$\frac{b_1}{1-\beta} < \frac{\gamma_1 b_1^2}{2\Gamma^2} \Leftrightarrow b_1 > \frac{2\Gamma^2}{\gamma\gamma_1}. \tag{18}$$

(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 (17) for all $n > n_{0,2}$, for some $n_{0,2}$, if

$$\frac{\gamma_1^2 L\sigma^2}{2} < \frac{\gamma_1 b_1^2}{2\Gamma^2} \Leftrightarrow b_1 > \Gamma\sigma\sqrt{L\gamma_1}. \tag{19}$$

(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

$$\frac{b_1}{1-\beta} + \frac{\gamma_1^2 L\sigma^2}{2} < \frac{\gamma_1 b_1^2}{2\Gamma^2}. \tag{20}$$

Because all constants are positive in Inequality (20), including b_1, it follows that

$$b_1 > \frac{3+\sqrt{9+4\gamma_1^3 L\sigma^2/\Gamma^2}}{2\gamma_1/\Gamma^2}. \tag{21}$$

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 $\frac{\gamma_1 b_1^2}{2\Gamma^2} - \frac{b_1}{1-\beta} = \epsilon > 0$. Ignoring the term $n^{-3\gamma-\beta}$ (for simplicity), Inequality (17) becomes

$$\epsilon n^{-2+\gamma} \geq \frac{L\sigma^2\gamma_1^2}{2} n^{-2\gamma} \Rightarrow n \geq (\frac{L\sigma^2\gamma_1^2}{2\epsilon})^c \equiv n_{0,1}, \tag{22}$$

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 ϵ. Similarly, we can derive expressions for $n_{0,2}$ and $n_{0,3}$.

THEOREM 3. *Suppose that Assumptions 1, 3(c), and 4 hold. Let $\zeta_n = \mathrm{E}(\|\theta_n - \theta_\star\|^2)$ 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 (2). 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

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

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*

$$\prod_{i=1}^{n} \frac{1}{1+b_i} \leq \exp(-K \sum_{i=1}^{n} b_i). \tag{23}$$

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 (23).

LEMMA 2 (TOULIS AND AIROLDI (2017)). *Consider sequences $a_n \downarrow 0, b_n \downarrow 0$, and $c_n \downarrow 0$ such that, $a_n = \mathrm{o}(b_n)$, $\sum_{i=1}^{\infty} a_i = A < \infty$, and there is n' such that $c_n/b_n < 1$ for all $n > n'$. Define,*

$$\delta_n \triangleq \frac{1}{a_n}(a_{n-1}/b_{n-1} - a_n/b_n) \text{ and } \zeta_n \triangleq \frac{c_n}{b_{n-1}} \frac{a_{n-1}}{a_n}, \tag{24}$$

and suppose that $\delta_n \downarrow 0$ and $\zeta_n \downarrow 0$. Pick a positive n_0 such that $\delta_n + \zeta_n < 1$ and $(1+c_n)/(1+b_n) < 1$, for all $n \geq n_0$.
Consider a positive sequence $y_n > 0$ that satisfies the recursive inequality,

$$y_n \leq \frac{1 + c_n}{1 + b_n} y_{n-1} + a_n. \tag{25}$$

Then, for every $n > 0$,

$$y_n \leq K_0 \frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n (1 + c_1)^{n_0} A, \tag{26}$$

where $K_0 = (1 + b_1)(1 - \delta_{n_0} - \zeta_{n_0})^{-1}$, $Q_i^n = \prod_{j=i}^n (1 + c_i)/(1 + b_i)$, and $Q_i^n = 1$ if $n < i$, by definition.

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,*

$$y_n \leq 2\frac{a_1(1 + b_1)}{b_1} n^{-\alpha+\beta} + \exp(-\log(1 + b_1)n^{1-\beta})[y_0 + (1 + b_1)^{n_0} A], \tag{27}$$

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.)
By definition,

$$\begin{aligned}
\delta_n = \frac{1}{a_n}(\frac{a_{n-1}}{b_{n-1}} - \frac{a_n}{b_n}) &= \frac{1}{a_1 n^{-\alpha}} \frac{a_1}{b_1}((n-1)^{-\alpha+\beta} - n^{-\alpha+\beta}) \\
&= \frac{1}{n^{-\alpha} b_1}[(n-1)^{-\alpha+\beta} - n^{-\alpha+\beta}] \\
&\leq \frac{1}{b_1} n^{-1+\beta}.
\end{aligned} \tag{28}$$

Also, $\zeta_n = 0$ since $c_n = 0$. We can take $n_0 = \lceil (2/b_1)^{1/(1-\beta)} \rceil$, for which $\delta_{n_0} \leq 1/2$. Therefore, $K_0 = (1+b_1)(1-\delta_{n_0})^{-1} \leq 2(1+b_1)$; we can simply take $K_0 = 2(1+b_1)$. Since $c_n = 0$, $Q_i^n = \prod_{j=i}^n (1+b_i)^{-1}$. Thus,

$$
\begin{aligned}
& Q_1^n \geq (1+b_1)^{-n}, \text{ and} \\
& Q_1^n \leq \exp(-\log(1+b_1)/b_1 \sum_{i=1}^n b_i), \qquad [\textit{by Lemma 1.}] \\
& Q_1^n \leq \exp(-\log(1+b_1) n^{1-\beta}). \qquad [\textit{because} \sum_{i=1}^n i^{-\beta} \geq n^{1-\beta}.]
\end{aligned}
\tag{29}
$$

Lemma 2 and Ineqs. (29) imply

$$
\begin{aligned}
y_n & \leq K_0 \frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n (1+c_1)^{n_0} A \qquad [\textit{by Lemma 2}] \\
& \leq 2\frac{a_1(1+b_1)}{b_1} n^{-\alpha+\beta} + Q_1^n [y_0 + (1+b_1)^{n_0} A] \qquad [\textit{by Ineqs. (29)}, c_1 = 0] \\
& \leq 2\frac{a_1(1+b_1)}{b_1} n^{-\alpha+\beta} + \exp(-\log(1+b_1) n^{1-\beta}) [y_0 + (1+b_1)^{n_0} A],
\end{aligned}
\tag{30}
$$

where the last inequality also follows from Ineqs. (29).

Proof of Theorem 3. Now we are ready to prove the main theorem. By definition, $\theta_n = \theta_n^+ - \gamma_n \varepsilon_n$, and thus, by Assumption 4,

$$
\mathrm{E}(\|\theta_n - \theta_\star\|^2) \leq \mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) + \gamma_n^2 \sigma^2. \tag{31}
$$

Also by definition we have $\gamma_n h(\theta_n^+) + \theta_n^+ = \theta_{n-1}$, and thus

$$
\|\theta_{n-1} - \theta_\star\|^2 = \|\theta_n^+ - \theta_\star\|^2 + 2\gamma_n (\theta_n^+ - \theta_\star)^\top h(\theta_n^+) + \gamma_n^2 \|h(\theta_n^+)\|^2. \tag{32}
$$

Therefore,

$$
\begin{aligned}
\|\theta_n^+ - \theta_\star\|^2 + 2\gamma_n (\theta_n^+ - \theta_\star)^\top h(\theta_n^+) & \leq \|\theta_{n-1} - \theta_\star\|^2 \\
\|\theta_n^+ - \theta_\star\|^2 + 2\gamma_n \mu \|\theta_n^+ - \theta_\star\|^2 & \leq \|\theta_{n-1} - \theta_\star\|^2 \qquad [\textit{by Assumption 3(c)}] \\
\|\theta_n^+ - \theta_\star\|^2 & \leq \frac{1}{1+2\gamma_n\mu} \|\theta_{n-1} - \theta_\star\|^2.
\end{aligned}
\tag{33}
$$

Combining Inequality (31) and Inequality (33) yields

$$
\begin{aligned}
\mathrm{E}(\|\theta_n - \theta_\star\|^2) & = \mathrm{E}(\|\theta_n^+ - \theta_\star\|^2) + \gamma_n^2 \sigma^2 \\
& \leq \frac{1}{1+2\gamma_n\mu} \mathrm{E}(\|\theta_{n-1} - \theta_\star\|^2) + \gamma_n^2 \sigma^2.
\end{aligned}
\tag{34}
$$

The final result of Theorem 3 is obtained through a direct application of Corollary 1 on recursion (34), by setting $y_n \equiv \mathrm{E}\|\theta_n - \theta_\star\|^2$, $b_n \equiv 2\gamma_n\mu$, and $a_n \equiv \gamma_n^2\sigma^2$. The case where $\gamma = 1$ only changes Inequality (29) by replacing $\sum b_i$ with $\log n$.

THEOREM 4. *Suppose that Assumptions 1,2, 3(a), 4, and 5 hold, and that $(2\gamma_1 J_h(\theta_\star) - I)$ 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 (2) is asymptotically normal:*

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

The covariance matrix Σ is the unique solution of

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

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

PROOF. Convergence of $\theta_n \rightarrow \theta_\star$ is established from Theorem 1. By definition of the stochastic proximal point algorithm in Equation (2),

$$\theta_n = \theta_{n-1} - \gamma_n(h(\theta_n^+) + \varepsilon_n), \text{ and} \tag{35}$$

$$\theta_n^+ + \gamma_n h(\theta_n^+) = \theta_{n-1}. \tag{36}$$

We use Equation (36) and expand $h(\cdot)$ to obtain

$$\begin{aligned} h(\theta_n^+) &= h(\theta_{n-1}) - \gamma_n J_h(\theta_{n-1}) h(\theta_n^+) + \epsilon_n \\ h(\theta_n^+) &= (I + \gamma_n J_h(\theta_{n-1}))^{-1} h(\theta_{n-1}) + (I + \gamma_n J_h(\theta_{n-1}))^{-1} \epsilon_n, \end{aligned} \tag{37}$$

where $\|\epsilon_n\| = \mathrm{O}(\gamma_n^2)$ by Theorem 3. By Lipschitz continuity of $h(\cdot)$ (Assumption 3(a)) and the almost sure convergence of θ_n to $\theta_\star$, it follows $h(\theta_{n-1}) = J_h(\theta_\star)(\theta_{n-1} - \theta_\star) + \mathrm{o}(1)$, where $\mathrm{o}(1)$ is a vector with vanishing norm. Therefore we can rewrite (37) as follows,

$$h(\theta_n^+) = A_n(\theta_{n-1} - \theta_\star) + \mathrm{O}(\gamma_n^2), \tag{38}$$

such that $\|A_n - J_h(\theta_\star)\| \rightarrow 0$, and $\mathrm{O}(\gamma_n^2)$ denotes a vector with norm $\mathrm{O}(\gamma_n^2)$. Thus, we can rewrite (35) as

$$\theta_n - \theta_\star = (I - \gamma_n A_n)(\theta_{n-1} - \theta_\star) - \gamma_n \varepsilon_n + \mathrm{O}(\gamma_n^2). \tag{39}$$

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

2. Proofs for approximate implementations

First, we recall our main approximate implementation:

$$\begin{aligned} w_1 &= \theta_{n-1}, \\ w_k &= w_{k-1} - a_k\big(\gamma_n H(w_{k-1}, \xi_k) + w_{k-1} - w_1\big), \quad 1 < k \leq K, \\ \theta_n &= w_k. \end{aligned} \tag{40}$$

Note about proofs. The procedures analyzed in this section involve two nested iterative processes. Throughout, we use n as the index variable of the outer iteration and k for the inner iteration. The randomness entering the kth step of the inner iteration inside the nth step of the outer iteration is denoted by ξ_k^n and $\mathcal{F}_{n,k}$ denotes the σ-algebra generated by $\{\xi_i^j\}_{1\leq i\leq K}^{1\leq j\leq n-1} \cup \{\xi_i^n\}_{1\leq i\leq k}$. We also write w_k^n instead w_k in (40) to explicitely keep track of the outer iteration index. Finally, we use $\mathcal{F}_{n-1}$ as a shorthand for $\mathcal{F}_{n-1,K}$.

Let $\chi_n(\theta)$ denote the output of the same procedure in the theoretical case where $K = \infty$. In other words, χ_n is the proximal operator that satisfies:

$$\chi_n(\theta) + \gamma_n h(\chi_n(\theta)) = \theta. \tag{41}$$

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

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

(b) $\|\chi_n(x) - x\| \leq \frac{\gamma_n L}{1+\gamma_n\mu}\|x - \theta_\star\|$.

PROOF. First note that since $h(\theta_\star) = 0$, $\theta_\star$ is a fixed point of χ_n.

(a) By definition of χ_n in Equation (41), one can write:

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

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

$$\begin{aligned} \|\chi_n(x) - \chi_n(y)\|^2 = {} & (x - y)^\top \big(\chi_n(x) - \chi_n(y)\big) \\ & - \gamma_n \left[h\big(\chi_n(x)\big) - h\big(\chi_n(y)\big)\right]^\top \big(\chi_n(x) - \chi_n(y)\big). \end{aligned} \tag{42}$$

Using 3(c), we obtain:

$$(1 + \gamma_n\mu)\|\chi_n(x) - \chi_n(y)\|^2 \leq (x - y)^\top \big(\chi_n(x) - \chi_n(y)\big),$$

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

(b) We can write $\|\chi_n(x) - x\| = \gamma_n\|h\big(\chi_n(x)\big)\|$ by definition of χ_n. Because $h\big(\chi_n(\theta_\star)\big) = 0$:

$$\begin{aligned} \|\chi_n(x) - x\| &= \gamma_n\|h\big(\chi_n(x)\big) - h\big(\chi_n(\theta_\star)\big)\| \\ &\leq \gamma_n L\|\chi_n(x) - \chi_n(\theta_\star)\| \leq \frac{\gamma_n L}{1 + \gamma_n\mu}\|x - \theta_\star\|, \end{aligned}$$

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

LEMMA 4. *Suppose that Assumptions 2, 4 and 3(a) hold. Consider the choice of parameter $a_k = a_n$, $1 \leq k \leq K$ in (40) with $a_n \leq \frac{1}{(1+\gamma_n L)^2}$, then:*

$$\mathrm{E}\left(\|\theta_n - \theta_n^+\|^2 | \mathcal{F}_{n-1}\right) \leq (1 - a_n)^K \|\theta_{n-1} - \theta_n^+\|^2 + \sigma^2\gamma_n^2 a_n.$$

PROOF. Let us write $H(w_k^n, \xi_{k+1}^n) = h(w_k^n) + \varepsilon_{k+1}^n$ and define $g(x) = \gamma_n h(x) + x - \theta_{n-1}$. We can write:

$$\begin{aligned}\|w_{k+1}^n - \theta_n^+\|^2 &= \|w_k^n - a_n\big(g(w_k^n) + \gamma_n \varepsilon_{k+1}^n\big) - \theta_n^+\|^2 \\ &= \|w_k^n - \theta_n^+\|^2 - 2a_n\big(g(w_k^n) + \gamma_n\varepsilon_{k+1}^n\big)^T\big(w_k^n - \theta_n^+\big) \\ &\quad + a_n^2\big(\|g(w_k^n)\|^2 + \gamma_n^2\|\varepsilon_{k+1}^n\|^2 + 2g(w_k^n)^T\gamma_n\varepsilon_{k+1}^n\big).\end{aligned}$$

Taking expectations on both sides conditioned on $\mathcal{F}_{n,k}$ and noting that $\mathrm{E}(\varepsilon_{k+1}|\mathcal{F}_{n,k}) = 0$ and $\mathrm{E}(\|\varepsilon_{k+1}\|^2|\mathcal{F}_{n,k}) \le \sigma^2$ by Assumption 4 we get:

$$\mathrm{E}(\|w_{k+1}^n - \theta_n^+\|^2|\mathcal{F}_{n,k}) \le \|w_k^n - \theta_n^+\|^2 - 2a_n g(w_k^n)^T\big(w_k^n - \theta_n^+\big) + a_n^2\|g(w_k^n)\|^2 + a_n^2\gamma_n^2\sigma^2,$$

It follows easily from Assumptions 2 and 3(a) that g is $(\gamma_n L + 1)$-Lipschitz continuous and that $\big(g(x) - g(y)\big)^\top(x-y) \ge \|x-y\|^2$ for al x and y in $\mathbb{R}^p$. Furthermore, since $g(\theta_n^+) = 0$ by definition:

$$\delta_{k+1}^n \le \big[1 - 2a_n + a_n^2(1+\gamma_n L)^2\big]\,\delta_k + a_n^2\gamma_n^2\sigma^2\ .$$

where we took expectations on both sides conditioned on $\mathcal{F}_{n-1}$ and write $\delta_k = \mathrm{E}\left(\|w_k^n - \theta_n^+\|^2|\mathcal{F}_{n-1}\right)$. For $a_n \le \frac{1}{(1+\gamma_n L)^2}$, the above recursion becomes:

$$\delta_{k+1}^n \le (1-a_n)\delta_k + a_n^2\gamma_n^2\sigma^2\ .$$

Note that $w_K^n = \theta_n$, and $w_1^n = \theta_{n-1}$ by definition. Therefore, we obtain:

$$\mathrm{E}\left(\|\theta_n - \theta_n^+\|^2|\mathcal{F}_{n-1}\right) \le (1-a_n)^K\|\theta_{n-1} - \theta_n^+\|^2 + \sigma^2\gamma_n^2 a_n\big(1 - (1-a_n)^K\big).$$

THEOREM 5. *Suppose that Assumptions 2, 4 and 3(c) hold, then the proximal stochastic fixed point procedure in Equation (40) with parameters $\gamma_n = \gamma$ and $a_k = 2a/K$, such that $e^{-a} < \mu/L$ and $K \ge 2a(1+\gamma L)^2$, satisfies:*

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

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

PROOF. We decompose the distance between θ_n and $\theta_\star$ as the distance between θ_n and θ_n^+, and the distance of θ_n^+ to $\theta_\star$:

$$\begin{aligned}\mathrm{E}\|\theta_n - \theta_\star\| &\le \mathrm{E}\|\theta_n - \theta_n^+\| + \mathrm{E}\|\theta_n^+ - \theta_\star\| \quad [\textit{triangle inequality}] \\ &= \mathrm{E}\|\theta_n - \theta_n^+\| + \mathrm{E}\|\chi_n(\theta_{n-1}) - \chi_n(\theta_\star)\| \quad [\textit{by definition of } \chi_n \textit{ in Equation } (41)] \\ &\le \mathrm{E}\|\theta_n - \theta_n^+\| + \frac{1}{1+\gamma\mu}\mathrm{E}\|\theta_{n-1} - \theta_\star\| \quad [\textit{by Lemma 3 } \text{(a)}] \\ &\le (1-a_n)^{K/2}\mathrm{E}\|\theta_{n-1} - \chi_n(\theta_{n-1})\| + \sigma\gamma\sqrt{a_n} + \frac{1}{1+\gamma\mu}\mathrm{E}\|\theta_{n-1} - \theta_\star\| \quad [\textit{by Lemma 4}] \\ &\le \frac{(1-a_n)^{K/2}\gamma L}{1+\gamma\mu}\mathrm{E}\|\theta_{n-1} - \theta_\star\| + \sigma\gamma\sqrt{a_n} + \frac{1}{1+\gamma\mu}\mathrm{E}\|\theta_{n-1} - \theta_\star\| \quad [\textit{by Lemma 3}\text{(b)}] \\ &= \left(\frac{1 + (1-a_n)^{K/2}\gamma L}{1+\gamma\mu}\right)\mathrm{E}\|\theta_{n-1} - \theta'_{n-1}\| + \sigma\gamma\sqrt{a_n}\,.\end{aligned}$$

We now choose a_n constant of the form $\frac{2a}{K}$ and obtain the following recursion:

$$\mathrm{E}\|\theta_n - \theta_\star\| \leq C \cdot \mathrm{E}\|\theta_{n-1} - \theta_\star\| + \sigma\gamma\frac{\sqrt{2a}}{\sqrt{K}},$$

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

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

□

For completeness, we finally present a variant of the previous procedure, also providing an approximate implementation of the proximal Robbins–Monro procedure via proximal stochastic fixed points. Compared to the procedure (40) analyzed in Theorem 5, we now perform an extra gradient step to compute θ_n from θ_{n-1} instead of simply using w_K^n. Formally:

$$\begin{aligned} &w_1^n = \theta_{n-1}, \\ &w_k^n = w_{k-1}^n - a_k\big(\gamma_n H(w_{k-1}^n, \xi_k^n) + w_{k-1}^n - w_1^n\big), \quad 1 < k \leq K, \\ &\theta_n = \theta_{n-1} - \gamma_n H(w_K^n, \xi_{K+1}^n) \end{aligned} \tag{43}$$

THEOREM 6. *Suppose that Assumptions 2, 4 and 3(c) hold, then the procedure in Equation (43) with parameters $\gamma_n = \gamma_1/n$ and $a_k = 2a/K$, where a and K are constants satisfying:*

$$e^{-a} \leq \frac{\mu}{2\gamma_1 L^2}, \quad K \geq 3a \cdot \max\left\{(1+\gamma_1 L)^2, (\gamma_1 L)^2 + e^{3a}\right\}.$$

Then:

$$\mathrm{E}\|\theta_n - \theta_\star\|^2 \leq \frac{e^{4\gamma_1^2\mu^2}}{n^{\gamma_1\mu}}\|\theta_0 - \theta_\star\|^2 + 2\gamma_1^2\sigma^2 e^{2\gamma_1^2\mu^2} e^{\gamma_1\mu} \cdot S(n),$$

where:

$$S(n) \leq \begin{cases} \frac{1}{\gamma_1\mu - 1}\frac{1}{n} & \text{if } \gamma_1\mu > 1 \\ \log(en)/n & \text{if } \gamma_1\mu = 1 \\ \frac{2}{1-\gamma_1\mu}\frac{1}{n^{\gamma_1\mu}} & \text{if } \gamma_1\mu < 1 \end{cases}$$

PROOF. We focus on a single iteration n and write $H(w_K^n, \xi_{K+1}^n) = h(w_K^n) + \epsilon_n$. We first decompose the error as usual:

$$\begin{aligned} \|\theta_n - \theta_\star\|^2 &= \|\theta_{n-1} - \gamma_n h(w_K^n) - \gamma_n\epsilon_n - \theta_\star\|^2 \\ &= \|\theta_{n-1} - \gamma_n h(w_K^n) - \theta_\star\|^2 + \gamma_n^2\|\epsilon_n\|^2 - 2\gamma_n\epsilon_n^T\big(\theta_{n-1} - \gamma_n h(w_K^n) - \theta_\star\big). \end{aligned}$$

Recall that $\mathrm{E}(\epsilon_n|\mathcal{F}_{n,K}) = 0$ and $\mathrm{E}\left(\|\epsilon_n\|^2|\mathcal{F}_{n,K}\right) \leq \sigma^2$ by Assumption 4. Hence:

$$\begin{aligned} \mathrm{E}\left(\|\theta_n - \theta_\star\|^2|\mathcal{F}_{n,K}\right) &\leq \|\theta_{n-1} - \gamma_n h(w_K^n) - \theta_\star\|^2 + \gamma_n^2\sigma^2 \\ &= \|\theta_n^+ + \gamma_n\big(h(\theta_n^+) - h(w_K^n)\big) - \theta_\star\|^2 + \gamma_n^2\sigma^2 \end{aligned}$$

where the equality uses that $\theta_{n-1} - \gamma_n h(\theta_n^+) = \theta_n^+$ by Eq. (1).

Next, using that $\|a+b\|^2 \le (1+\alpha)\|a\|^2 + (1+\alpha^{-1})\|b\|^2$ for all $\alpha > 0$ by Young's inequality:

$$\begin{aligned}
\mathrm{E}\left(\|\theta_n - \theta_\star\|^2 | \mathcal{F}_{n,K}\right) &\le (1+\alpha)\|\theta_n^+ - \theta_\star\|^2 + \gamma_n^2(1+\alpha^{-1})\|h(\theta_n^+) - h(w_K^n)\|^2 + \gamma_n^2\sigma^2 \\
&\le \frac{1+\alpha}{(1+\gamma_n\mu)^2}\|\theta_{n-1} - \theta_\star\|^2 + (1+\alpha^{-1})(\gamma_n L)^2\|\theta_n^+ - w_K^n\|^2 + \gamma_n^2\sigma^2
\end{aligned}$$

where the second inequality uses Lemma 3 *(a)* and Assumption 2.

Taking expectations conditioned on $\mathcal{F}_{n-1}$ and using Lemma 4 (our choice of parameters satisfies in particular $a_n \le 1/(1+\gamma_n L)^2$ as required by the Lemma):

$$\begin{aligned}
\mathrm{E}\left(\|\theta_n - \theta_\star\|^2 | \mathcal{F}_{n-1}\right) &\le \frac{1+\alpha}{(1+\gamma_n\mu)^2}\|\theta_{n-1} - \theta_\star\|^2 + (1+\alpha^{-1})(\gamma_n L)^2(1-a_n)^K\|\theta_n^+ - \theta_{n-1}\|^2 \\
&\quad + (1+\alpha^{-1})(\gamma_n L)^2\gamma_n^2\sigma^2 a_n + \gamma_n^2\sigma^2 \\
&\le \frac{1+\alpha+(1+\alpha^{-1})(\gamma_n L)^4(1-a_n)^K}{(1+\gamma_n\mu)^2}\|\theta_{n-1} - \theta_\star\|^2 \\
&\quad + \gamma_n^2\sigma^2\big[1 + (1+\alpha^{-1})(\gamma_n L)^2 a_n\big]\,.
\end{aligned}$$

where the second inequality uses Lemma 3 *(b)*.

We now pick $\alpha = (\gamma_n L)^2(1-a_n)^{K/2}$ and take expectations on both sides:

$$\mathrm{E}\|\theta_n - \theta_\star\|^2 \le \left(\frac{1+(\gamma_n L)^2(1-a_n)^{K/2}}{1+\gamma_n\mu}\right)^2 \mathrm{E}\|\theta_{n-1} - \theta_\star\|^2 + \gamma_n^2\sigma^2\left[1 + (\gamma_n L)^2 a_n + \frac{a_n}{(1-a_n)^{K/2}}\right].$$

Using the inequality $\exp\big(-nx/(1-x)\big) \le (1-x)^n \le \exp(-nx)$, it is easy to see that the choice of parameters in the theorem statement implies:

$$(1-a_n)^{K/2} \le e^{-a}, \quad e^{-a}(\gamma_n L)^2 \le \gamma_n\mu/2, \quad (\gamma_n L)^2 a_n + \frac{a_n}{(1-a_n)^{K/2}} \le 1\,,$$

hence the previous inequality yields:

$$\begin{aligned}
\mathrm{E}\|\theta_n - \theta_\star\|^2 &\le \left(\frac{1+\gamma_n\mu/2}{1+\gamma_n\mu}\right)^2 \mathrm{E}\|\theta_{n-1} - \theta_\star\|^2 + 2\gamma_n^2\sigma^2 \\
&\le \left(1 - \frac{\gamma_n\mu}{(1+\gamma_n\mu)^2}\right)\mathrm{E}\|\theta_{n-1} - \theta_\star\|^2 + 2\gamma_n^2\sigma^2\,.
\end{aligned}$$

Writing $y_n = \mathrm{E}\|\theta_n - \theta_\star\|^2$, $a_n = \gamma_n\mu/(1+\gamma_n\mu)^2$ and $b_n = 2\gamma_n^2\sigma^2$, the previous inequality reads $y_n \le (1-a_n)y_{n-1} + b_n$. Define $p_n = \prod_{k=1}^n(1-a_k)$, an easy induction gives:

$$y_n \le p_n y_0 + p_n\sum_{k=1}^n \frac{b_k}{p_k}\,. \tag{44}$$

We first focus on getting a lower bound and upper bound on p_n. For the lower bound, using that $(1-x) \ge \exp\big(-x/(1-x)\big)$, we obtain:

$$\begin{aligned}
p_n &\ge \exp\left(-\sum_{k=1}^n a_k\right)\exp\left(-\sum_{k=1}^n \frac{a_k^2}{1-a_k}\right) \\
&\ge \exp\left(-\sum_{k=1}^n \gamma_k\mu\right)\exp\left(-\sum_{k=1}^n \gamma_k^2\mu^2\right) \ge \frac{e^{2\gamma_1^2\mu^2 - \gamma_1\mu}}{n^{\gamma_1\mu}}\,.
\end{aligned}$$

where the second inequality uses the definition of a_k and the last inequality uses that $\gamma_n = \gamma_1/n$ an the series approximations of Lemma 5. Similarly for the upper bound, using that $(1-x) \le \exp(-x)$:

$$p_n \le \exp\left(-\sum_{k=1}^{n} a_k\right) = \exp\left(-\sum_{k=1}^{n} \gamma_k\mu\right) \exp\left(\sum_{k=1}^{n} \frac{\gamma_k^2\mu^2(2+\gamma_k\mu)}{(1+\gamma_k\mu)^2}\right)$$
$$\le \exp\left(-\sum_{k=1}^{n} \gamma_k\mu\right) \exp\left(\sum_{k=1}^{n} 2\gamma_k^2\mu^2\right) \le \frac{e^{4\gamma_1^2\mu^2}}{(n+1)^{\gamma_1\mu}}\,.$$

Plugging the previous two bounds into (44), we obtain:

$$y_n \le \frac{e^{4\gamma_1^2\mu^2}}{n^{\gamma_1\mu}} y_0 + \frac{2\gamma_1^2\sigma^2 e^{2\gamma_1^2\mu^2} e^{\gamma_1\mu}}{(n+1)^{\gamma_1\mu}} \sum_{k=1}^{n} \frac{1}{k^{2-\gamma_1\mu}}\,.$$

Finally, we conclude by defining $S(n) = (n+1)^{-\gamma_1\mu} \sum_{k=1}^{n} k^{\gamma_1\mu-2}$ and using Lemma 5 to obtain the upper bounds on $S(n)$ given in the theorem statement depending on the value of $\gamma_1\mu$.

LEMMA 5. *For any $\alpha > 0$ and $n \ge 1$:*

$$\frac{(1+n)^{1-\alpha}-1}{1-\alpha} \le \sum_{k=1}^{n} \frac{1}{k^\alpha} \le \frac{n^{1-\alpha}-\alpha}{1-\alpha} \quad \textit{and} \quad \frac{n^{1+\alpha}}{1+\alpha} \le \sum_{k=1}^{n} k^\alpha \le \frac{(n+1)^{1+\alpha}-1}{1+\alpha}\,,$$

where the first bound remains true by continuity at $\alpha = 1$: $\log(1+n) \le \sum_{k=1}^{n} \frac{1}{k} \le 1 + \log n$.

PROOF. Immediate by approximating the discrete sums from above and below by integrals.

3. Computation of implicit updates

At a first glance, the computation of the implicit procedure,

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

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_θ to simplify the implicit update.

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

$$\theta_n = \theta_{n-1} - \gamma_n s(\theta_n) U_n = \theta_{n-1} - \eta U_n,$$

for some scalar η. Thus, we have to solve:

$$\gamma_n s(\theta_n) = \eta \Leftrightarrow \gamma_n s(\theta_{n-1} - \eta U_n) = \eta.$$

The problem is now reduced to a one-dimensional fixed-point equation for ξ. 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 algorithm for generalized linear models.

References

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

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

Gladyshev, E. (1965). On stochastic approximation. *Theory of Probability & Its Applications 10*(2), 275–278.

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

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

Toulis, P., E. Airoldi, and J. Rennie (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.

Toulis, P. and E. M. Airoldi (2017, 08). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist. 45*(4), 1694–1727.