The Stochastic Fejér-Monotone Hybrid Steepest Descent Method and the Hierarchical RLS

Konstantinos Slavakis, *Senior Member, IEEE*

***Abstract*—This paper introduces the stochastic Fejér-monotone hybrid steepest descent method (S-FM-HSDM) to solve affinely constrained and composite convex minimization tasks. The minimization task is not known exactly; noise contaminates the information about the composite loss function and the affine constraints. S-FM-HSDM generates sequences of random variables that, under certain conditions and with respect to a probability space, converge point-wise to solutions of the noiseless minimization task. S-FM-HSDM enjoys desirable attributes of optimization techniques such as splitting of variables and constant step size (learning rate). Furthermore, it provides a novel way of exploiting the information about the affine constraints via fixed-point sets of appropriate nonexpansive mappings. Among the offsprings of S-FM-HSDM, the hierarchical recursive least squares (HRLS) takes advantage of S-FM-HSDM's versatility toward affine constraints and offers a novel twist to LS by generating sequences of estimates that converge to solutions of a hierarchical optimization task: minimize a convex loss over the set of minimizers of the ensemble LS loss. Numerical tests on a sparsity-aware LS task show that HRLS compares favorably to several state-of-the-art convex, as well as non-convex, stochastic-approximation, and online-learning counterparts.**

***Index Terms*—Stochastic approximation, online learning, convex, composite, RLS.**

I. Introduction

A. Problem Statement

THE following problem is considered: With a stochastic oracle providing estimates f_n (or even ∇f_n), h_n and $\mathcal{A}_n$ per n (n denotes discrete time *and* iteration index; $n \in \mathbb{Z}_{>0} := \{1, 2, \ldots\}$) of the generally unknown convex functions f, h and the affine set $\mathcal{A}$, respectively, solve

$$\min_{x \in \mathcal{A} \subset \mathcal{X}} f(x) + h(x) + g(x), \tag{P}$$

where $\mathcal{X}$ is a finite-dimensional real Hilbert space. Only the convex (regularizing) function g is assumed to be known exactly. The goal is to construct a sequence of estimates $(x_n)_n := (x_n)_{n \in \mathbb{Z}_{\geq 0}} \subset \mathcal{X}$ by exploiting the information about $(f_n)_n$, or $(\nabla f_n)_n$, $(h_n, \mathcal{A}_n)_n$ as well as g, and to identify the conditions which ensure, despite the uncertainty about f, h and $\mathcal{A}$, the point-wise convergence of $(x_n)_n$ to a solution of (P) with respect to (w.r.t.) a probability space.

Instances of (P) appear in adaptive filtering (AF) [2]–[4]; in particular, in linear equalization, channel estimation, beamforming, tracking of fading channels, line and acoustic echo cancellation and active noise control [2]. Special cases of (P) appear also in stochastic approximation (SA) [4], [5] and online learning (OL) [4], [6] as in training artificial neural networks, learning optimal strategies in Markov decision processes, recursive games, sequential-decision tasks in economics [5], online classification and multi-armed bandit problems [6]. (An outline of the strong ties and distinct differences between SA and online learning is provided in [7].)

Each one of the three loss terms in (P) plays a distinct role: f is smooth and generally unknown, h can be non-smooth and unknown, while g comprises all *known* and possibly non-smooth regularizing losses. The affine constraint $\mathcal{A}$ renders (P) a versatile framework that encompasses a large variety of problems. For example, given the finite-dimensional Hilbert spaces $\{\mathsf{X}_k\}_{k=0}^{I_h+J_g}$, with $I_h, J_g \in \mathbb{Z}_{>0}$, the convex functions $\mathfrak{f} : \mathsf{X}_0 \to \mathbb{R}$, $\mathfrak{h}^{(i)} : \mathsf{X}_i \to \mathbb{R} \cup \{+\infty\}$, $\mathfrak{g}^{(j)} : \mathsf{X}_{j+I_h} \to \mathbb{R} \cup \{+\infty\}$, the linear mappings $H^{(i)} : \mathsf{X}_0 \to \mathsf{X}_i, G^{(j)} : \mathsf{X}_0 \to \mathsf{X}_{j+I_h}, \mathsf{p}^{(i)} \in \mathsf{X}_i$ and $\mathsf{q}^{(j)} \in \mathsf{X}_{j+I_h}$, for $i \in \{1, \ldots, I_h\}$ and $j \in \{1, \ldots, J_g\}$, then the highly structured composite problem

$$\min_{\mathsf{x}^{(0)} \in \mathsf{X}_0} \mathfrak{f}(\mathsf{x}^{(0)}) + \sum_{i=1}^{I_h} \mathfrak{h}^{(i)}(H^{(i)}\mathsf{x}^{(0)} - \mathsf{p}^{(i)}) + \sum_{j=1}^{J_g} \mathfrak{g}^{(j)}(G^{(j)}\mathsf{x}^{(0)} - \mathsf{q}^{(j)}) \tag{1}$$

can be recast as (P) if $\mathcal{X} := \mathsf{X}_0 \times \mathsf{X}_1 \times \cdots \times \mathsf{X}_{I_h+J_g} = \{x := (\mathsf{x}^{(0)}, \ldots, \mathsf{x}^{(I_h+J_g)}) | \mathsf{x}^{(k)} \in \mathsf{X}_k, \forall k \in \{0, \ldots, I_h + J_g\}\}$, $f(x) := \mathfrak{f}(\mathsf{x}^{(0)})$, $h(x) := \sum_{i=1}^{I_h} \mathfrak{h}^{(i)}(\mathsf{x}^{(i)})$, $g(x) := \sum_{j=1}^{J_g} \mathfrak{g}^{(j)}(\mathsf{x}^{(j+I_h)})$, and the closed affine set $\mathcal{A} := \{x \in \mathcal{X} | \mathsf{x}^{(0)} \in \mathsf{X}_0, \mathsf{x}^{(i)} = H^{(i)}\mathsf{x}^{(0)} - \mathsf{p}^{(i)}, \mathsf{x}^{(j+I_h)} = G^{(j)}\mathsf{x}^{(0)} - \mathsf{q}^{(j)}, i \in \{1, \ldots, I_h\}, j \in \{1, \ldots, J_g\}\}$. The splitting of variables via Cartesian-product spaces facilitates processing; e.g., (6). The (P) formulation can also accommodate any closed convex (not necessarily affine) constraint $\mathcal{C}$ as follows: Consider (1) and let one of the $\{\mathfrak{h}^{(i)}\}_i$ or $\{\mathfrak{g}^{(j)}\}_j$, depending on whether $\mathcal{C}$ bears stochasticity or not, take the form of the indicator function $\iota_{\mathcal{C}}$ (see Appendix A for the definition). More importantly, (P) allows for cases where the information about $\mathcal{A}$ is not known exactly, introduces thus stochasticity into $\mathcal{A}$ and opens the door to new problem formulations and novel algorithmic developments, e.g., (HLS) and Algorithm 2.

Manuscript received August 12, 2018; revised January 17, 2019; accepted March 10, 2019. Date of publication March 25, 2019; date of current version April 26, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Athanasios A. Rontogiannis. This work was supported by the NSF under Grants 1525194 and 1718796. This paper was presented in part at the 52nd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, October 2018 [1].

The author is with the Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260-2500 USA (e-mail: kslavaki@buffalo.edu).

Digital Object Identifier 10.1109/TSP.2019.2907257

Algorithm 1: S-FM-HSDM.

Stochastic oracle's input : $(\nabla f_n, h_n, \mathcal{A}_n)_{n\in\mathbb{Z}_{>0}}$, $L_{\nabla f}$, g.

User's input : $\alpha \in [0.5, 1)$, $\lambda \in (0, 2(1-\alpha)/L_{\nabla f})$, T_0, ∇f_0, x_0, and $(T_n)_{n\in\mathbb{Z}_{\geq 0}}$.

Output : Sequence $(x_n)_{n\in\mathbb{Z}_{\geq 0}}$.

Initialization

$x_{1/2} \coloneqq T_0^{(\alpha)} x_0 - \lambda\nabla f_0(x_0)$.

$x_1 \coloneqq \mathrm{Prox}_{\lambda(h_0+g)}(x_{1/2})$.

for $n = 1$ **to** $+\infty$ **do**

$x_{n+1/2} \coloneqq x_{n-1/2} - [T_{n-1}^{(\alpha)} x_{n-1} - \lambda\nabla f_{n-1}(x_{n-1})] + [T_n x_n - \lambda\nabla f_n(x_n)]$.

$x_{n+1} \coloneqq \mathrm{Prox}_{\lambda(h_n+g)}(x_{n+1/2})$.

Algorithm 2: HRLSa.

Stochastic oracle's input : $(\mathbf{a}_n, b_n)_{n\in\mathbb{Z}_{>0}}$.

User's input : $\alpha \in [0.5, 1)$, $\lambda \in \mathbb{R}_{>0}$, $\mathbf{R}_0$, $\mathbf{r}_0$, $\mathbf{x}_0$, and $\varpi_0 \geq \|\mathbf{R}_0\|$.

Output : Sequence $(\mathbf{x}_n)_{n\in\mathbb{Z}_{\geq 0}}$.

Initialization

$\mathbf{x}_{1/2} \coloneqq \mathbf{x}_0 - \alpha\frac{1}{\varpi_0}(\mathbf{R}_0\mathbf{x}_0 - \mathbf{r}_0)$.

For any $d \in \{1, \ldots, D\}$, $[\mathbf{x}_1]_d \coloneqq [\mathbf{x}_{1/2}]_d \cdot (1 - \lambda/\max\{\lambda, |[\mathbf{x}_{1/2}]_d|\})$.

for $n = 1$ **to** $+\infty$ **do**

Set $\varpi_n \geq \|\mathbf{R}_n\|$.

$\mathbf{x}_{n+1/2} \coloneqq \mathbf{x}_n + \mathbf{x}_{n-1/2} - \mathbf{x}_{n-1} + \alpha\frac{1}{\varpi_{n-1}}(\mathbf{R}_{n-1}\mathbf{x}_{n-1} - \mathbf{r}_{n-1}) - \frac{1}{\varpi_n}(\mathbf{R}_n\mathbf{x}_n - \mathbf{r}_n)$.

For any $d \in \{1, \ldots, D\}$, $[\mathbf{x}_{n+1}]_d \coloneqq [\mathbf{x}_{n+1/2}]_d \cdot (1 - \lambda/\max\{\lambda, |[\mathbf{x}_{n+1/2}]_d|\})$.

B. Case Study: Sparsity-Aware Least Squares

To highlight the versatility of (P) and to unfold all features of the proposed algorithmic solution, coined stochastic Fejér-monotone hybrid steepest descent method (S-FM-HSDM), it is instructive to build the discussion around specific instances of (P). To this end, let $\mathcal{X}$ be the Euclidean $\mathbb{R}^D$. Bold-faced symbols indicate that $\mathcal{X} = \mathbb{R}^D$; in particular, lowercase bold-faced symbols denote vectors in $\mathbb{R}^D$. Consider a sparse system $\boldsymbol{\theta}_* \in \mathcal{X}$ and the classical *linear-regression model:* $b_n = \mathbf{a}_n^\intercal\boldsymbol{\theta}_* + \eta_n$, almost surely (a.s.), $\forall n \in \mathbb{Z}_{>0}$, with input-output data pair $(\mathbf{a}_n, b_n) \in \mathcal{X}\times\mathbb{R}$, the noise process $(\eta_n)_n$ is assumed to be zero-mean and independent of $(\mathbf{a}_n)_n$, and $\intercal$ denotes vector/matrix transposition. Typical stationarity assumptions on $(\mathbf{a}_n, b_n)_n$ are adopted also here: $\mathbf{R} := \mathbb{E}(\mathbf{a}_n\mathbf{a}_n^\intercal)$, $\mathbf{r} := \mathbb{E}(b_n\mathbf{a}_n)$, and $\mathbb{E}(b_n^2)$ stay constant $\forall n$, where $\mathbb{E}(\cdot)$ denotes expectation. It is well-known that $\boldsymbol{\theta}_*$ satisfies the normal equations $\boldsymbol{\theta}_* \in \{\mathbf{x}\in\mathcal{X} | \mathbf{R}\mathbf{x} = \mathbf{r}\}$ [2, (3.9)]. This section deals with the system-identification problem of estimating the sparse $\boldsymbol{\theta}_*$ without knowing $(\mathbf{R}, \mathbf{r})$ but relying only on the information $(\mathbf{a}_n, b_n)_n$ provided by the stochastic oracle.

Motivated by the celebrated (Lagrangian form of the) least absolute shrinkage and selection operator (LASSO) [8, (3.52)], designed to solve sparse system-identification problems, the first instance of (P) is the convexly regularized least squares: $\forall n \in \mathbb{Z}_{>0}$,

$$\min_{\mathbf{x}\in\mathbb{R}^D} \overbrace{\frac{1}{2}\mathbf{x}^\intercal\mathbf{R}\mathbf{x} - \mathbf{r}^\intercal\mathbf{x} + \frac{1}{2}\,\mathbb{E}(b_n^2)}^{l(\mathbf{x})} + \overbrace{\rho\|\mathbf{x}\|_1}^{g(\mathbf{x})}$$

$$= \min_{\mathbf{x}\in\mathbb{R}^D} \mathbb{E}\left[\underbrace{\frac{1}{2n}\sum\nolimits_{\nu=1}^{n}(\mathbf{a}_\nu^\intercal\mathbf{x} - b_\nu)^2}_{l_n(\mathbf{x})}\right] + \rho\|\mathbf{x}\|_1, \quad \text{(CRegLS)}$$

where the ℓ_1-norm regularizer promotes sparse solutions. (CRegLS) becomes a special case of (P), if $\mathcal{A} := \mathcal{X} = \mathbb{R}^D =: \mathcal{A}_n$, $(f, f_n) := (l, l_n)$, or, $(h, h_n) := (l, l_n)$ a.s.

The second instance of (P) exploits the fact that even the information about $\mathcal{A}$ may be inexact, and takes the form of a *hierarchical* (H)LS estimation task, which appears to be new in the AF, SA and OL literature: $\forall n$,

$$\min_{\mathbf{x}\in\mathbb{R}^D} \ [\, \|\mathbf{x}\|_1 =: g(\mathbf{x}) \,]$$

$$\text{s.to } \mathbf{x} \in \underbrace{\arg\min_{\mathbf{x}'\in\mathbb{R}^D} \mathbb{E}\left[\sum\nolimits_{\nu=1}^{n}(\mathbf{a}_\nu^\intercal\mathbf{x}' - b_\nu)^2\right]}_{\mathcal{A}}, \quad \text{(HLS)}$$

i.e., the convex loss $g(\cdot)$, here $\|\cdot\|_1$, is minimized over the set of minimizers of the classical (ensemble) LS loss. Recall that $\mathcal{A}$ in (HLS) comprises all vectors, including $\boldsymbol{\theta}_*$, that satisfy the normal equations. In the case of $g(\cdot) := \|\cdot\|_1$, (HLS) can be also viewed as an SA extension of (the *deterministic*) basis pursuit [9]. The mainstream approach, e.g., [4], [10], [11], to deal with (HLS) is to employ the indicator function $\iota_{\mathcal{A}}$ in the place of one of the $\mathfrak{h}^{(i)}$ and $\mathfrak{g}^{(j)}$ in (1). Such a path restricts the means of treating $\mathcal{A}$ to the projection mapping $P_{\mathcal{A}}$ [recall that $P_{\mathcal{A}}$ is the proximal mapping of $\iota_{\mathcal{A}}$; *cf.* (9)]. Since $(\mathbf{R}, \mathbf{r})$ are generally unknown, $\mathcal{A}$ is also unknown to the user. Still, the goal is to solve (HLS). If $(f, f_n) := (0, 0) =: (h, h_n)$, and $\mathcal{A}_n$ is defined as an estimate of $\mathcal{A}$, then (HLS) turns out to be a special instance of (P). This paper provides a novel way of using the available estimates $(\mathcal{A}_n)_n$ of $\mathcal{A}$ via fixed-point sets of appropriate nonexpansive mappings (*cf.* Section II). This new viewpoint pays off in the computationally efficient HRLSa (*cf.* Algorithm 2), which solves (HLS) under certain conditions, despite the uncertainty in the estimates $(\mathcal{A}_n)_n$, while scoring the lowest estimation error across a variety of numerical-test scenarios versus several state-of-the-art schemes (*cf.* Section IV).

C. Prior Art

In most cases, OL and SA algorithms have their origins in deterministic optimization schemes. For example, the OL scheme [12] draws inspiration from the forward-backward (a.k.a. proximal-gradient) algorithm [13, §27.3] and incorporates variance-reduction arguments [14] into its iterations to effect convergence speed-ups in solving a special case of (P), which appears to be of primary importance in machine learning: $f := (1/M)\sum_{m=1}^{M}\mathfrak{f}^{(m)}$, where $\{\mathfrak{f}^{(m)}\}_{m=1}^{M}$ are convex and smooth, $M \in \mathbb{Z}_{>0}$ is very large, $h := 0$ and $\mathcal{A} := \mathcal{X}$. Driven by the need to avoid the cumbersome computation of ∇f, stochasticity is introduced by selecting randomly only a small subset of $\{\mathfrak{f}^{(m)}\}_{m=1}^{M}$, per time/iteration index n, to form an estimate of ∇f. Recent SA schemes, motivated by the forward-backward algorithm and formulated in the more general setting of monotone-operator inclusions, can be found in [10], [15], [16]. An SA extension of primal-dual methods, where stochasticity is introduced via general sampling techniques to deal with massive data, is reported in [17]. An SA extension of the Douglas-Rachford algorithm [13, §25.2, §27.2] is reported in [10]. Study [18] extends the celebrated alternating direction method of multipliers (ADMM) to the OL setting, and blends it with variance-reduction arguments to solve a problem similar to that of [12], but with a non-trivial, yet deterministic affine constraint $\mathcal{A} \subsetneq \mathcal{X}$. Furthermore, [19] explores the dual-averaging scheme of [20] in the SA context offering linear-convergence guarantees for a quadratic f in (P), while $\mathcal{X}$ is a closed convex set with non-empty interior. Moreover, the SA schemes [11], [21] are motivated by the deterministic acceleration method of [22]; in particular, [21] uses specific step sizes (*cf.* [21, (33)]) to effect convergence acceleration in the case where $h := 0$, g is (Lipschitz) continuous and a deterministic convex compact constraint takes the place of $\mathcal{A}$ in (P).

With regards to the specific setting of Section I-B, the state-of-the-art AF schemes [23]–[25] are built around a variation of (CRegLS), where the regularizing coefficient ρ_n converges to zero as $n \to \infty$. A Bayesian approach to the LS sparse system-identification problem appears in [26], and a greedy RLS approach based on the orthogonal-matching-pursuit algorithm is reported in [27]. A majorization-minimization approach, which includes also non-convex regularizers, is studied in [28]. Basis pursuit [9] is used in [29] to provide an interpretation of the estimate-update equation per iteration n of several proportionate-type AF schemes; however, an ensemble-based viewpoint, such as (HLS), and a performance analysis are not provided.

D. Contributions and Structure of the Manuscript

Similarly to [10], [12], [15], [16], [18], [19], [21], the proposed S-FM-HSDM (Algorithm 1) springs from the deterministic FM-HSDM [30], which belongs to the HSDM family [31] and solves (P) in infinite-dimensional Hilbert spaces with no stochasticity involved. In [30], the information about the affine constraint $\mathcal{A}$ is incorporated into FM-HSDM via an affine nonexpansive mapping $T : \mathcal{X} \to \mathcal{X}$ whose fixed-point set is $\mathcal{A} = \operatorname{Fix} T := \{x \in \mathcal{X} | Tx = x\}$. For example, the (metric) projection mapping $P_{\mathcal{A}}$ onto $\mathcal{A}$ (*cf.* Appendix A) may serve as T [30, Prop. 2.11]. Interestingly, the versatile [30] allows for numerous choices of T other than the mainstream $P_{\mathcal{A}}$ [*cf.* (3)].

S-FM-HSDM extends FM-HSDM to the stochastic setting. With a stochastic oracle providing a sequence of affine constraints $(\mathcal{A}_n)_n$ as estimates of the generally unknown $\mathcal{A}$, a mapping T_n is chosen per time index n, with $\mathcal{A}_n = \operatorname{Fix} T_n$, to serve as an estimate of T. There are numerous choices of T_n other than the obvious $P_{\mathcal{A}_n}$. Furthermore, f and h are not required to be known exactly and only estimates $(f_n)_n$ [or even $(\nabla f_n)_n$] and $(h_n)_n$ are provided to the user by the stochastic oracle. The versatility of S-FM-HSDM is demonstrated in the system-identification context of Section I-B, where S-FM-HSDM solves (HLS) in Section II, with its specific form coined *hierarchical (H)RLS*. Mappings $(T_n)_n$ drive the HRLS iterates asymptotically to a vector in $\mathcal{A}$, and HRLS solves (HLS) *without* employing any sub-routines for identifying $\mathcal{A}$ prior to minimizing g over $\mathcal{A}$. It is worth recalling here that identifying $\mathcal{A}$ requires the computation of $\mathbb{E}(\cdot)$ which is a usually intractable task for the user. A specific choice of T_n [*cf.* (5a)] yields the computationally efficient HRLSa flavor of S-FM-HSDM (*cf.* Section II).

Many SA methods, such as the classical [32] and its convex-analytic extension [33], rely on diminishing step sizes (learning rates) to ensure a.s. convergence of their iterates. Nevertheless, constant step-size schemes, e.g., [10], [16], are highly desirable in signal processing and machine learning since they appear to **(i)** reach the neighborhood of solutions in a fewer number of iterations than the diminishing step-size methods [16]; and **(ii)** adapt quickly to changes of non-stationary environments and track dynamically changing sets of solutions (*cf.* Figure 3 and [2, Ch. 21]). S-FM-HSDM operates with a constant step size $\forall n$. The performance analysis of Section III identifies those conditions which ensure that S-FM-HSDM converges a.s. to a solution of (P). For clarity, those conditions are exemplified in the context of Section I-B.

To validate the theoretical developments of this work, extensive numerical tests on synthetic data, within the context of Section I-B, are reported in Section IV. Flavors HRLSa and HRLSb of S-FM-HSDM appear to be the most consistent methods in achieving the lowest estimation error across a variety of scenarios versus several state-of-the-art AF, SA and OL schemes.

To improve readability, S-FM-HSDM, its specific flavors within the context of Section I-B and their main theoretical results are presented first in Section II. The performance analysis and the accompanying assumptions are detailed in Section III, while the necessary mathematical preliminaries and proofs are deferred to the appendices.

II. The S-FM-HSDM Family and Its Properties

A. The User-Defined Mappings $(T_n)_n$

To utilize the information about $\mathcal{A}$, this work follows [30] and considers a mapping T s.t. $\operatorname{Fix}(T) = \mathcal{A}$. An obvious choice for T would be the (metric) projection mapping $P_{\mathcal{A}}$ onto $\mathcal{A}$ [30, Prop. 2.11]. Nevertheless, this study revolves around less obvious cases. In the context of Section I-B, such examples are:

$$T = \begin{cases} (\mathbf{I} - \frac{\mu}{\varpi}\mathbf{R}) + \frac{\mu}{\varpi}\mathbf{r}, \quad \varpi \geq \|\mathbf{R}\|,\ \mu \in (0,1], & \text{(2a)} \\ (\mathbf{I} + \kappa\mathbf{R})^{-1} + \kappa(\mathbf{I} + \kappa\mathbf{R})^{-1}\mathbf{r}, \quad \kappa \in \mathbb{R}_{>0}, & \text{(2b)} \end{cases}$$

where $\mathbf{I}$ is the identity matrix and the spectral norm $\|\mathbf{R}\|$ is equal to the maximum eigenvalue of $\mathbf{R}$. In fact, any mapping which belongs to the following family of mappings may serve as a candidate for T [30, Prop. 2.11]:

$$\mathfrak{T}_{\mathcal{A}} := \left\{ T : \mathcal{X} \to \mathcal{X} \,\middle|\, \begin{array}{l} \operatorname{Fix} T = \mathcal{A}; T = Q + \pi; \\ Q \text{ is positive}; \|Q\| \leq 1; \pi \in \mathcal{X} \end{array} \right\}. \quad (3)$$

Any $T \in \mathfrak{T}_{\mathcal{A}}$ is affine, i.e., there exists a linear mapping $Q : \mathcal{X} \to \mathcal{X}$ and a $\pi \in \mathcal{X}$ s.t. $Tx = Qx + \pi$, $\forall x \in \mathcal{X}$; in short, $T = Q + \pi$. For the linear Q, $\|Q\| := \sup_{\{x | \|x\| \leq 1\}} \langle x \mid Qx \rangle$. Mapping $Q : \mathcal{X} \to \mathcal{X}$ is called positive if it is linear, bounded, self-adjoint and $\langle x \mid Qx \rangle \geq 0$, $\forall x \in \mathcal{X}$ [34, §9.3]. Since $\|Q\| \leq 1$, every mapping $T \in \mathfrak{T}_{\mathcal{A}}$ turns out to be nonexpansive [13]: $\forall (x, x') \in \mathcal{X}^2$, $\|Tx - Tx'\| = \|Qx - Qx'\| = \|Q(x - x')\| \leq \|Q\| \, \|x - x'\| \leq \|x - x'\|$. It is also worth noticing here that $\mathfrak{T}_{\mathcal{A}}$ is closed under any convex combination and certain compositions of its members [30, Prop. 2.10]. Notice also that $P_{\mathcal{A}} \in \mathfrak{T}_{\mathcal{A}}$ [30, Prop. 2.11]. Further information on $\mathfrak{T}_{\mathcal{A}}$ is deferred to Appendix A.

Notwithstanding, $\mathcal{A}$ is in general unknown to the user; hence, so is T as well. With the stochastic oracle providing estimates $(\mathcal{A}_n)_n$ of $\mathcal{A}$, the user needs to construct mappings $(T_n)_n$ that serve as estimates of the unknown T. In the context of Section I-B, for example, instead of the unknown $\mathbf{R}$ and $\mathbf{r}$, their classical running-average estimates [2], $\forall n \in \mathbb{Z}_{>0}$,

$$\mathbf{R}_n := \frac{1}{n}\sum_{\nu=1}^{n} \mathbf{a}_\nu \mathbf{a}_\nu^{\intercal}, \quad \mathbf{r}_n := \frac{1}{n}\sum_{\nu=1}^{n} b_\nu \mathbf{a}_\nu, \tag{4}$$

can be used to define

$$T_n := \begin{cases} (\mathbf{I} - \frac{\mu}{\varpi_n}\mathbf{R}_n) + \frac{\mu}{\varpi_n}\mathbf{r}_n, & \varpi_n \geq \|\mathbf{R}_n\|, \\ & \mu \in (0,1], \quad (5a) \\ (\mathbf{I} + \kappa\mathbf{R}_n)^{-1} + \kappa(\mathbf{I} + \kappa\mathbf{R}_n)^{-1}\mathbf{r}_n, & \kappa \in \mathbb{R}_{>0}. \quad (5b) \end{cases}$$

Lemma 1: For the affine set $\mathcal{A} := \{\mathbf{x} | \mathbf{R}\mathbf{x} = \mathbf{r}\}$ in Section I-B, mappings (2) belong to $\mathfrak{T}_{\mathcal{A}}$. Moreover, mappings (5) take the form $T_n = Q_n + \pi_n$, where $Q_n : \mathcal{X} \to \mathcal{X}$ is positive, with $\|Q_n\| \leq 1$, and $\pi_n \in \mathcal{X}$, a.s., $\forall n$.

Proof: See Appendix B. ■

B. The S-FM-HSDM Family

With mapping T_n available, and with the *averaged* mapping $T_n^{(\alpha)}$ defined as $T_n^{(\alpha)} := \alpha T_n + (1-\alpha)\,\mathrm{Id}$, $\forall n$, where $\mathrm{Id} : \mathcal{X} \to \mathcal{X}$ stands for the identity operator, S-FM-HSDM is presented in Algorithm 1. Prox in lines 3 and 6 of Algorithm 1 denotes the proximal mapping [*cf.* (9)]. In the case where $L_{\nabla f}$ is not available or cannot be estimated, S-FM-HSDM offers the option of setting $(f, f_n) := (0, 0)$, where $L_{\nabla f}$ can be set equal to any positive real-valued number (*cf.* Section IV), and any estimate of f can be transferred to the loss h_n, since assumptions on h and h_n are weaker than those on f and f_n (*cf.* Section III). Strategies for estimating $L_{\nabla f}$, in the case it is unknown, will be reported elsewhere. Line 5 requires only the computation of the current first-order information $\nabla f_n(x_n)$, whereas $\nabla f_{n-1}(x_{n-1})$, which was computed at the previous time instance, can be pulled from a buffer that stores information.

In the context of (HLS), if (5a) with $\mu := 1$ is adopted, S-FM-HSDM takes the flavor of Algorithm 2, coined HRLSa. Since $g(\cdot) = \|\cdot\|_1$, $\mathrm{Prox}_{\lambda g}(\cdot)$ in lines 3 and 7 of Algorithm 2 boils down to the popular soft-thresholding operation. Following (5a), line 5 of Algorithm 2 introduces an overestimate ϖ_n of the maximum eigenvalue $\lambda_{\max}(\mathbf{R}_n) = \|\mathbf{R}_n\|$. To this end, motivated by the celebrated power iteration [35], and for an arbitrarily fixed initial vector $\mathbf{p}_0 \in \mathcal{X}$, the following iterative procedure, run over all $n \in \mathbb{Z}_{>0}$, is used in Section IV to generate $(\varpi_n)_n$: **(i)** $\mathbf{q}_n := \mathbf{R}_n\mathbf{p}_{n-1}$; **(ii)** $\mathbf{p}_n := \mathbf{q}_n/\|\mathbf{q}_n\|$; **(iii)** $\varpi_n := \mathbf{p}_n^{\intercal}\mathbf{R}_n\mathbf{p}_n + \epsilon_\varpi$, for a user-defined $\epsilon_\varpi \in \mathbb{R}_{>0}$. If (5b) is used as T_n in Algorithm 1, the flavor of S-FM-HSDM is coined HRLSb. Due to space limitations, the detailed pseudo-code description of HRLSb is omitted. Other options for T_n will be explored elsewhere. Between HRLSa and HRLSb, HRLSa exhibits the lowest computational complexity, of order $\mathcal{O}(D^2)$ per n. HRLSb requires the matrix inversion $(\mathbf{I} + \lambda\mathbf{R}_n)^{-1}$ for the running average $\mathbf{R}_n$ in (4).

In the context of (CRegLS), Algorithm 1 yields Algorithm 3, tagged S-FM-HSDM(CRegLS). To verify that Algorithm 3 is indeed a by-product of Algorithm 1, notice that (CRegLS) can

Algorithm 3: S-FM-HSDM(CRegLS).

Stochastic oracle's input : $(\mathbf{a}_n, b_n)_{n\in\mathbb{Z}_{\geq 0}}$.
User's input : $\alpha \in [0.5, 1)$, $\lambda \in \mathbb{R}_{>0}$, $\mathbf{R}_0$, $\mathbf{r}_0$, $(\mathsf{x}_0^{(1)}, \mathsf{x}_0^{(2)})$.
Output : Sequence $(\mathbf{x}_n)_{n\in\mathbb{Z}_{\geq 0}}$.

1 **Initialization**
2 $\quad \mathsf{x}_0 := \frac{1}{2}(\mathsf{x}_0^{(1)} + \mathsf{x}_0^{(2)})$.
3 $\quad \mathsf{x}_{1/2}^{(i)} := \alpha\mathsf{x}_0 + (1-\alpha)\mathsf{x}_0^{(i)}$, $i \in \{1, 2\}$.
4 $\quad \mathsf{x}_1^{(1)} := (\mathbf{I} + \lambda\mathbf{R}_0)^{-1}(\mathsf{x}_{1/2}^{(1)} + \lambda\mathbf{r}_0)$.
5 $\quad [\mathsf{x}_1^{(2)}]_d := [\mathsf{x}_{1/2}^{(2)}]_d \cdot (1 - \lambda\rho/\max\{\lambda\rho, |[\mathsf{x}_{1/2}^{(2)}]_d|\})$, $\forall d \in \{1, \ldots, D\}$.
6 $\quad \mathsf{x}_1 := \frac{1}{2}(\mathsf{x}_1^{(1)} + \mathsf{x}_1^{(2)})$.
7 **for** $n = 1$ **to** $+\infty$ **do**
8 $\quad \mathsf{x}_{n+1/2}^{(i)} := \mathsf{x}_{n-1/2}^{(i)} - \alpha\mathsf{x}_{n-1} - (1-\alpha)\mathsf{x}_{n-1}^{(i)} + \mathsf{x}_n$, $i \in \{1, 2\}$.
9 $\quad \mathsf{x}_{n+1}^{(1)} := (\mathbf{I} + \lambda\mathbf{R}_n)^{-1}(\mathsf{x}_{n+1/2}^{(1)} + \lambda\mathbf{r}_n)$.
10 $\quad [\mathsf{x}_{n+1}^{(2)}]_d := [\mathsf{x}_{n+1/2}^{(2)}]_d \cdot (1 - \lambda\rho/\max\{\lambda\rho, |[\mathsf{x}_{n+1/2}^{(2)}]_d|\})$, $\forall d \in \{1, \ldots, D\}$.
11 $\quad \mathsf{x}_{n+1} := \frac{1}{2}(\mathsf{x}_{n+1}^{(1)} + \mathsf{x}_{n+1}^{(2)})$.

be seen, via variable splitting in the spirit of (1), as

$$\min_{(\mathsf{x}^{(1)},\mathsf{x}^{(2)})\in\mathbb{R}^D\times\mathbb{R}^D} \mathbb{E}\left[\overbrace{\frac{1}{2n}\sum\nolimits_{\nu=1}^{n}\left(\mathbf{a}_\nu^{\intercal}\mathsf{x}^{(1)} - b_\nu\right)^2}^{h_n(\mathsf{x}^{(1)})}\right] + \underbrace{\rho\|\mathsf{x}^{(2)}\|_1}_{g(\mathsf{x}^{(2)})} \text{ s.to } \mathsf{x}^{(1)} = \mathsf{x}^{(2)}, \tag{6}$$

so that space $\mathcal{X}$ is set to be $\mathbb{R}^D \times \mathbb{R}^D$, with inner product $\langle(\mathsf{x}^{(1)}, \mathsf{x}^{(2)}) \mid (\mathsf{x}'^{(1)}, \mathsf{x}'^{(2)})\rangle := \langle\mathsf{x}^{(1)} \mid \mathsf{x}'^{(1)}\rangle + \langle\mathsf{x}^{(2)} \mid \mathsf{x}'^{(2)}\rangle$. Moreover, $\mathcal{A}$ is the linear subspace $\{(\mathsf{x}^{(1)}, \mathsf{x}^{(2)}) \in \mathcal{X} | \mathsf{x}^{(1)} = \mathsf{x}^{(2)}\}$, with (orthogonal) projection mapping given by $P_{\mathcal{A}}[(\mathsf{x}^{(1)}, \mathsf{x}^{(2)})] = ((\mathsf{x}^{(1)} + \mathsf{x}^{(2)})/2, (\mathsf{x}^{(1)} + \mathsf{x}^{(2)})/2)$ and $T_n := T := P_{\mathcal{A}}$ in Algorithm 1. Moreover, $\mathrm{Prox}_{\lambda(h_n+g)}[(\mathsf{x}^{(1)}, \mathsf{x}^{(2)})] = (\mathrm{Prox}_{\lambda h_n}(\mathsf{x}^{(1)}), \mathrm{Prox}_{\lambda g}(\mathsf{x}^{(2)}))$. Lines 9 and 10 in Algorithm 3 correspond to $\mathrm{Prox}_{\lambda h_n}(\cdot)$ and $\mathrm{Prox}_{\lambda g}(\cdot)$, respectively. Due to the fact that $\mathbf{R}_n$ is obtained by $\mathbf{R}_{n-1}$ via a rank-one modification, i.e., $\mathbf{R}_n = (n-1)\mathbf{R}_{n-1}/n + \mathbf{a}_n\mathbf{a}_n^{\intercal}/n$, a way to compute $(\mathbf{I} + \lambda\mathbf{R}_n)^{-1}$ efficiently via modifications of the eigen-decomposition of $(\mathbf{I} + \lambda\mathbf{R}_{n-1})^{-1}$ can be deduced, for example, via arguments found in [36], [37]; details are omitted. Quantities $\mathbf{R}_0$ and $\mathbf{r}_0$ are arbitrarily fixed and used in Line 4 of Algorithm 3 to initialize the iterative process.

C. Main Theoretical Properties

The main properties of S-FM-HSDM (Algorithm 1) are summarized in the following Theorems 1 and 2. To improve the readability of the manuscript, the detailed description of the necessary assumptions is deferred to Section III. Nevertheless, as a high-level description, Assumption 1 gathers all the assumptions about the sequence of the user-defined nonexpansive mappings $(T_n)_n$, such as asymptotic consistency, while

Assumption 3 refers to the loss functions $(f_n, h_n)_n$. The typical SA presupposition of asymptotic unbiasedness is introduced in Assumption 4. The technical Assumption 6 imposes a summability constraint on the random variables (RVs) defined via (19b) and (19c). A typical SA boundedness constraint of variances is introduced by Assumption 7, while Assumption 8 imposes the weak condition on loss functions that bounded estimates imply bounded subgradients. Assumptions 2 and 5 refer to the special cases of (CRegLS) and (HLS).

Theorem 1: Under Assumptions 1, 3, 4, 6, 7, and 8 (see Section III), the set of cluster points $\mathfrak{C}[(x_n)_n]$ of the S-FM-HSDM sequence $(x_n)_n$ (Algorithm 1) is nonempty a.s. Furthermore, every point in the nonempty $\mathfrak{C}[(x_n)_n]$ is a solution of (P) a.s.

Proof: See Appendix C. ∎

Theorem 2: Consider the case where T is known exactly, i.e., $T = T_n, \forall n$. Then, under the same setting as in Theorem 1, but without Assumptions 1, 7(ii), 8(ii), 8(iii), the sequence $(x_n)_n$ generated by Algorithm 1 converges a.s. to a solution of (P).

Proof: See Appendix D. ∎

It is worth mentioning here that the qualifier "FM" in S-FM-HSDM comes from the deterministic predecessor FM-HSDM [30] and the Fejér-monotonicity property of (22) and (24) in Appendix C.

Since HRLS and S-FM-HSDM(CRegLS) are offsprings of S-FM-HSDM, assertions about their convergence properties can be deduced from Theorems 1 and 2 and can take various forms. This study avoids to provide an exhaustive list of all such assertions with their forms, but brings only a couple of examples in the form of the following corollaries.

Corollary 1: Let Assumptions 2, 6, and 7(ii) hold true. Assume also that the stochastic process $(\mathbf{a}_n)_n$ possesses a non-singular $\mathbf{R}$, and that $\exists \varpi \in \mathbb{R}_{>0}$ s.t. $\varpi_n := \varpi \geq \max\{\|\mathbf{R}\|, \|\mathbf{R}_n\|\}$, a.s., $\forall n$. Then, the set of cluster points of the sequence $(\mathbf{x}_n)_n$, generated by either HRLSa or HRLSb, is non-empty, and any of its cluster points is a solution of (HLS) a.s.

Proof: See Appendix E. ∎

As a postscript to Corollary 1, recall that the matrix $\mathbf{R}$ of *any* regular process $(\mathbf{a}_n)_n$, i.e., a process with non-zero innovation [38, §2.6], is non-singular [38, Prob. 2.2].

Corollary 2: Let Assumptions 2, 4(i), and 6 hold true. Then, the sequence $(\mathbf{x}_n)_n$ generated by Algorithm 3 converges a.s. to a solution of (CRegLS).

Proof: See Appendix F. ∎

III. Performance Analysis of S-FM-HSDM

Rather than simply listing all assumptions needed for Theorems 1 and 2, as well as for Corollaries 1 and 2, this section follows a more instructive route by exemplifying the assumptions in the context of (CRegLS) and (HLS). With symbol $\xrightarrow{\text{a.s.}}_n$ introduced in Appendix A, the following assumptions are imposed on the mappings T and $(T_n)_n$.

Assumption 1 (Mappings T and T_n):

i) $T \in \mathfrak{T}_{\mathcal{A}}$.

ii) $T_n := Q_n + \pi_n$, where mapping $Q_n : \mathcal{X} \to \mathcal{X}$ is positive, with $\|Q_n\| \leq 1$, and $\pi_n \in \mathcal{X}$, a.s., $\forall n$.

iii) $(T - T_n) \xrightarrow{\text{a.s.}}_n 0$, i.e., $(T - T_n)x \xrightarrow{\text{a.s.}}_n 0$, $\forall x \in \mathcal{X}$, or, equivalently, $(Q - Q_n) \xrightarrow{\text{a.s.}}_n 0$ and $(\pi - \pi_n) \xrightarrow{\text{a.s.}}_n 0$.

iv) Define $\forall n$, $t_n := \mathbb{E}_{|\mathcal{F}_n}[\sum_{\nu=1}^{n}(T - T_\nu)x_\nu]$. All cluster points of any bounded subsequence of $(\mathbb{E}(t_n))_n$ belong to $\operatorname{ran}(\mathrm{Id} - Q)$.

To underline the generality of Assumption 1, the following popular Assumption 2, placed in the context of Section I-B and (4), provides a special case of Assumption 1, as Lemma 2 demonstrates.

Assumption 2 (Pointwise ergodicity): $\boldsymbol{\mathcal{E}}_n^R := \mathbf{R} - \mathbf{R}_n \xrightarrow{\text{a.s.}}_n \mathbf{0}$ and $\boldsymbol{\varepsilon}_n^r := \mathbf{r} - \mathbf{r}_n \xrightarrow{\text{a.s.}}_n \mathbf{0}$.

To save space, a discussion on conditions which suffice to guarantee Assumption 2, such as statistical independency or mixing conditions [39], [40], via the strong law of large numbers, is omitted. Notice that due to Assumption 2, $(\mathbf{R}_n)_n$ is bounded a.s.; hence, $\exists \varpi := \varpi(\omega) \geq \max\{\|\mathbf{R}\|, \|\mathbf{R}_n\|\}$, a.s., $\forall n$ (symbol ω is introduced in Appendix A).

Lemma 2: Assume that $\exists \varpi \in \mathbb{R}_{>0}$ s.t. $\varpi_n := \varpi \geq \max\{\|\mathbf{R}\|, \|\mathbf{R}_n\|\}$, a.s., $\forall n$, and that the matrix $\mathbf{R}$ of the stochastic process $(\mathbf{a}_n)_n$ is non-singular. Then, under also Assumption 2, mappings (5) satisfy Assumptions 1(ii) and 1(iii). Moreover, for any $T \in \mathfrak{T}_{\mathcal{A}}$ and any of its estimates $(T_n)_n$, Assumption 1(iv) holds true.

Proof: See Appendix G. ∎

Assumption 3 (Loss functions):

i) $f, h, g : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ belong to the class $\Gamma_0(\mathcal{X})$ of proper, lower semicontinuous (l.s.c.), convex functions [13].

ii) f is everywhere (Fréchet) differentiable, with $L_{\nabla f}$-Lipschitz continuous ∇f: $\|\nabla f(x) - \nabla f(x')\| \leq L_{\nabla f} \|x - x'\|$, $\forall (x, x') \in \mathcal{X} \times \mathcal{X}$, for some $L_{\nabla f} \in \mathbb{R}_{>0}$. Moreover, for any sub-$\sigma$-algebra $\mathcal{G}$ of Σ (*cf.* Appendix A) and $\forall x \in \mathrm{m}\mathcal{G}$, $\nabla f(x) \in \mathrm{m}\mathcal{G}$.

iii) $f_n, h_n \in \Gamma_0(\mathcal{X})$ a.s., $\forall n$.

iv) f_n is everywhere (Fréchet) differentiable, with L_n-Lipschitz continuous ∇f_n a.s., $\forall n$.

v) There exist $n_\# \in \mathbb{Z}_{\geq 0}$ and a $C_{\text{Lip}} \in \mathbb{R}_{>0}$, which is constant over all $\omega \in \Omega$, s.t. $L_n \leq C_{\text{Lip}}$ a.s., $\forall n \geq n_\#$.

vi) $(\nabla f - \nabla f_n) \xrightarrow{\text{a.s.}}_n 0$.

It can be readily verified in the context of (CRegLS) that Assumptions 3(i), 3(ii), 3(iii), and 3(iv) are satisfied by either $(f, f_n) := (l, l_n)$, or, $(h, h_n) := (l, l_n)$. If $f_n := l_n$, then $\forall \mathbf{x} \in \mathcal{X}$,

$$\nabla f_n(\mathbf{x}) = \frac{1}{n} \sum_{\nu=1}^{n} \mathbf{a}_\nu (\mathbf{a}_\nu^\intercal \mathbf{x} - b_\nu) = \mathbf{R}_n \mathbf{x} - \mathbf{r}_n . \tag{7}$$

Scalar $L_n := \|\mathbf{R}_n\|$ can be considered as ∇f_n's Lipschitz coefficient. Hence, if $(\mathbf{R}_n)_n$ is uniformly bounded over Ω, i.e., $\exists C_{\text{Lip}}$ s.t. $\|\mathbf{R}_n\| \leq C_{\text{Lip}}$, a.s., $\forall n$, then Assumption 3(v) holds true. On the other hand, if the uniform boundedness of $(\mathbf{R}_n)_n$ cannot be guaranteed, and since the current framework places no requirements on the uniform boundedness of the subgradients of $(h_n)_n$, (P) offers the flexibility to set $h_n := l_n$ and $f_n := 0$, for which Assumption 3(v) holds trivially true. Given that $\nabla f(\mathbf{x}) = \mathbf{R}\mathbf{x} - \mathbf{r}, \forall \mathbf{x} \in \mathcal{X}$, whenever $f = l$, it can be verified via Assumption 2 and (7) that Assumption 3(vi) holds true.

Assumption 4 (Asymptotic unbiasedness):

i) For any $x \in \mathcal{X}$, $\mathbb{E}_{|\mathcal{F}_n}[(h - h_n)(x)] =: \varepsilon_n^h(x) \xrightarrow{\text{a.s.}}_n 0$ and $\varepsilon_n^h(x_n) \xrightarrow{\text{a.s.}}_n 0$.

ii) $\mathbb{E}_{|\mathcal{F}_n}[(\nabla f - \nabla f_n)(x_n)] =: \varepsilon_n^f(x_n) \xrightarrow{\text{a.s.}}_n 0$.

Asymptotic unbiasedness appears often in SA, e.g., [5, p. 132, Thm. 2.3]. Lemma 3 presents cases where Assumption 4 holds true. Several of the results of Lemma 3 serve also as intermediate steps that justify the introduction of Assumption 6; more precisely, (8) suffice for Assumption 6 to hold true. To prove the claims of Lemma 3, the following assumption is needed.

Assumption 5: Motivated by [5, p. 162], the approximation errors $\boldsymbol{\mathcal{E}}_n^R$ and ε_n^r (*cf.* Assumption 2) are assumed here to be *exogenous* w.r.t. $\mathcal{F}_n := \sigma(\{x_\nu\}_{\nu=0}^n)$ for any n. In other words, $\boldsymbol{\mathcal{E}}_n^R, \varepsilon_n^r$, which are provided by the stochastic oracle, are considered to be independent of the past history $\mathcal{F}_n$ of the iterates: $\forall n$ and a.s., $\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\mathcal{E}}_n^R) = \mathbb{E}(\boldsymbol{\mathcal{E}}_n^R)$, $\mathbb{E}_{|\mathcal{F}_n}(\varepsilon_n^r) = \mathbb{E}(\varepsilon_n^r)$. Moreover, the stochastic process $(\mathbf{a}_n)_n$ is assumed to be independent and identically distributed (IID) and $\mathcal{F}_n$ is considered to be conditionally independent with $\sigma(\{\mathbf{a}_\nu\}_{\nu=1}^n)$ given $\sigma(\mathbf{R}_n)$.

It can be verified via the stationarity conditions of Section I-B that Assumption 5 implies $\mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_n) = \mathbb{E}(\mathbf{R}_n) = \mathbf{R}$ and $\mathbb{E}_{|\mathcal{F}_n}(\mathbf{r}_n) = \mathbb{E}(\mathbf{r}_n) = \mathbf{r}$. Thus, $\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\mathcal{E}}_n^R) = \mathbf{0}$ and $\mathbb{E}_{|\mathcal{F}_n}(\varepsilon_n^r) = \mathbf{0}$.

Lemma 3:

i) Consider T of (2a), T_n of (5a), and let Assumption 5 hold true. Moreover, assume the existence of $\varpi \in \mathbb{R}_{>0}$ s.t. $\varpi_n := \varpi \geq \max\{\|\mathbf{R}\|, \|\mathbf{R}_n\|\}$, a.s. and $\forall n$. Then,

$$\mathbb{E}_{|\mathcal{F}_n}[(T - T_n)\mathbf{x}_n] = \mathbf{0}, \tag{8a}$$

$$\mathbb{E}_{|\mathcal{F}_n}[(Q - Q_n)(\mathbf{x}_n - \mathbf{x}_{n-1})] = \mathbf{0}, \tag{8b}$$

$$\mathbf{t}_n = \mathbf{0}, \tag{8c}$$

$$\mathbb{E}_{|\mathcal{F}_n}[(T_n - T_{n-1})\mathbf{x}_{n-1}] = \mathbf{0}\,. \tag{8d}$$

Clearly, $\mathbb{E}(\mathbf{t}_n) = \mathbf{0}$, and thus, the only cluster point $\lim_{n\to\infty} \mathbb{E}(\mathbf{t}_n) = \mathbf{0}$ of sequence $(\mathbf{t}_n)_n$ belongs trivially to $\operatorname{ran}(\mathrm{Id} - Q)$; that is, Assumption 1(iv) holds true.

ii) Consider $(h, h_n) := (l, l_n)$ in (CRegLS), and let Assumption 5 hold true. Then, Assumption 4(i) holds true with $\varepsilon_n^h(\mathbf{x}) = 0 = \varepsilon_n^h(\mathbf{x}_n)$, a.s., $\forall n, \forall \mathbf{x} \in \mathcal{X}$.

iii) Consider ∇f_n in (7), $f := l$, and let Assumption 5 hold true. Then, $\forall n$,

$$\mathbb{E}_{|\mathcal{F}_n}[(\nabla f - \nabla f_n)\mathbf{x}_n] = \mathbf{0}, \tag{8e}$$

$$\mathbb{E}_{|\mathcal{F}_n}[(\nabla f_n - \nabla f_{n-1})\mathbf{x}_{n-1}] = \mathbf{0}\,. \tag{8f}$$

iv) Let Assumption 5 hold true. Let also $g := \|\cdot\|_1$ and either $(h, h_n) := (0, 0)$ or $(h, h_n) := (l, l_n)$. Consider also the sequence $(\boldsymbol{\xi}_n \in \partial(h_{n-1} + g)(\mathbf{x}_n))_n$ of subgradients defined in (11). Then, $\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\xi}_n) \in \partial(h + g)(\mathbf{x}_n)$, $\forall n$. More generally,

$$\exists(\epsilon_n)_n \subset (\mathrm{m}\Sigma)^+ \text{ with } \sum_n \mathbb{E}(\epsilon_n) < +\infty$$

$$\text{s.t. } \mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\xi}_n) \in \partial_{\epsilon_n}(h + g)(\mathbf{x}_n), \forall n\,. \tag{8g}$$

Proof: See Appendix H. ■

Results (8) can be relaxed as follows: Appendix C [*cf.* (23)] demonstrates that (8) suffice to establish Assumption 6.

Assumption 6 (Dominated $(\vartheta_n)_{n\in\mathbb{Z}_{\geq 0}}$): Consider the sequence $(\vartheta_n)_{n\in\mathbb{Z}_{\geq 0}}$ of RVs defined by the expression which starts from (19b) and ends at (19c). There exists $\psi \in (\mathrm{m}\Sigma)^+$ with $\mathbb{E}(\psi) < +\infty$ s.t. $\sum_n \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ \leq \psi$ a.s., where $\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ := \max\{0, \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)\}$.

Assumption 7 (Bounded variances):

i) Given $z \in \mathcal{X}$, there exists $C_{\nabla f} := C_{\nabla f}(z) \in \mathbb{R}_{>0}$ s.t. $\mathbb{E}[\|(\nabla f - \nabla f_n)z\|^2] \leq C_{\nabla f}$, $\forall n$.

ii) There exists $C_\pi \in \mathbb{R}_{>0}$ s.t. $\mathbb{E}(\|\pi_n - \pi\|^2) \leq C_\pi$, $\forall n$.

Bounded-variance assumptions appear often in SA, e.g., [5, p. 126, (A2.1)].

Assumption 8 (Bounded estimates yield bounded subgradients):

i) For any a.s. bounded $(z_n)_n$, there exist a sequence $(\tau_n)_n$ and $C_\partial := C_\partial(\omega) \in \mathbb{R}_{>0}$ s.t. $\tau_n \in \partial(h_n + g)(z_n)$ and $\mathbb{E}_{|\mathcal{F}_n}(\|\tau_n\|) \leq C_\partial$, $\forall n$, a.s.

ii) Consider the sequence $(\xi_n)_n$ of subgradients defined in (11). If $(x_n)_n$ is bounded a.s., then $(\xi_n)_n$ is bounded a.s.

iii) If $(\mathbb{E}(\|x_n\|^2))_n$ is bounded, then $(\mathbb{E}(\|\xi_n\|^2))_n$ is bounded.

Lemma 4: Let Assumption 2 hold true. If $(h, h_n) := (0, 0)$ and g is defined as a scalar multiple of $\|\cdot\|_1$, then Assumption 8 holds true. If $(h, h_n) := (l, l_n)$ and $(f, f_n) := (0, 0)$ in (CRegLS), then the following claims can be established.

i) Assumptions 8(i) and 8(ii) hold true.

ii) If there exist also $\varpi, \varpi' \in \mathbb{R}_{>0}$, fixed over the probability space, s.t. $\|\mathbf{R}_n\| \leq \varpi$ and $\|\mathbf{r}_n\| \leq \varpi'$, $\forall n$ and a.s., then Assumption 8(iii) holds true.

Proof: See Appendix I. ■

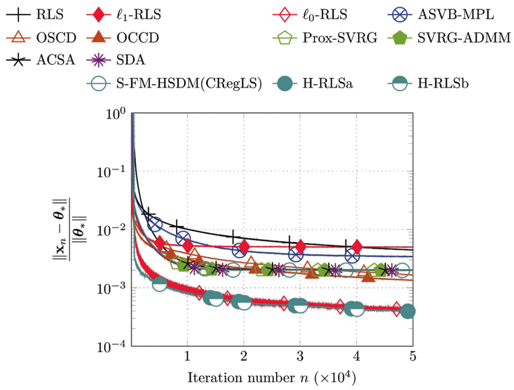
IV. Numerical Tests

The proposed framework is validated within the setting of Section I-B where S-FM-HSDM(CRegLS), HRLSa and HRLSb are compared with the following OL and SA schemes:

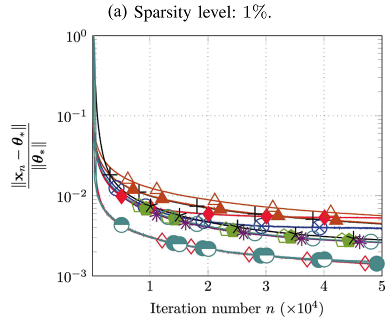
1) The classical RLS [2, §30.2];
2) the ℓ_1-norm regularized (ℓ_1-)RLS [25], and its extension, the ℓ_0-norm (ℓ_0-)RLS [25], where a *non-convex* regularizing function is used instead of $\|\cdot\|_1$;
3) the LASSO-motivated online selective coordinate descent (OSCD) and online cyclic coordinate descent (OCCD) methods [23], where, according to [23, Sec. V], the power of the additive noise in the linear-regression model is assumed to be known and incorporated in the regularizing coefficient ρ_n in (CRegLS) s.t. $\rho_n \to_n 0$;
4) the proximal stochastic variance-reduced gradient (Prox-SVRG) method [12], applied to the setting of the ever-growing data regime $f := (1/n)\sum_{\nu=1}^n \mathfrak{f}_\nu$ in (CRegLS), with $\mathfrak{f}_\nu(\mathbf{x}) := (1/2)(\mathbf{a}_\nu^\intercal \mathbf{x} - b_\nu)^2$;
5) SVRG-ADMM [18], where f is identical to that of the Prox-SVRG case;
6) the accelerated stochastic approximation (ACSA) with the step sizes of [21, (33)];
7) the adaptive sparse variational Bayes multi-parameter Laplace prior (ASVB-MPL) method [26]; and
8) the stochastic dual-averaging (SDA) scheme with linear-convergence-rate guarantees [19].

It is worth stressing here that all of [12], [18], [19], [21], [23], [25] are built around the mainstream (CRegLS). As explained in Sections I-A and I-B, any attempt to pass $\mathcal{A}$ of (HLS) to the objective function via the indicator function $\iota_{\mathcal{A}}$ entails the use of the projection mapping $P_{\mathcal{A}}$ and, thus, the eigen-decompositions of $(\mathbf{R}_n)_n$ via the (Moore-Penrose-)pseudoinverse operation. Recall that this is not the case for the computationally "light" HRLSa.

In all tests, the dimension of the Euclidean space $\mathcal{X} = \mathbb{R}^D$ is set to be $D := 100$. The sparse system $\boldsymbol{\theta}_*$ is created by placing ± 1 s at randomly selected entries of the $D \times 1$ all-zero vector. The "sparsity level" of $\boldsymbol{\theta}_*$ is defined as the percentage of the number of non-zero entries of $\boldsymbol{\theta}_*$ over D. All of the methods were tested in several scenarios detailed below. Since focus is

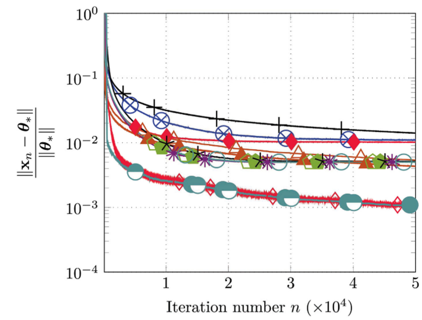


(a) Sparsity level: 1%.

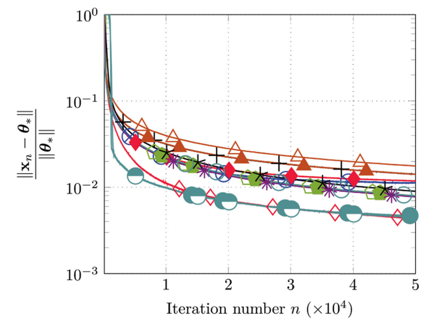


(b) Sparsity level: 10%.

Fig. 1. IID $(\mathbf{a}_n)_n$; SNR $= 20$ dB.



(a) Sparsity level: 1%.



(b) Sparsity level: 10%.

Fig. 2. IID $(\mathbf{a}_n)_n$; SNR $= 10$ dB.

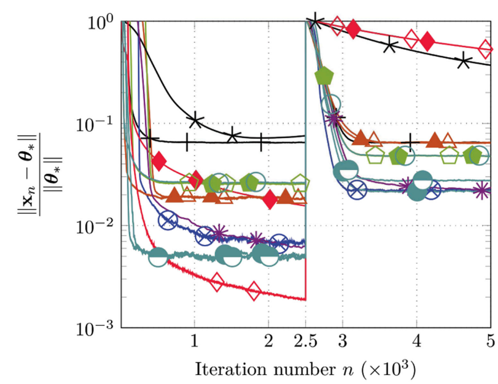


Fig. 3. IID $(\mathbf{a}_n)_n$; SNR $= 20$ dB; the sparsity level of $\boldsymbol{\theta}_*$ changes at time $n = 2,500$ from 1% to 10%.

placed on the system-identification problem of Section I-B, the criterion of performance is the normalized-root-mean-square-deviation loss $\|\mathbf{x}_n - \boldsymbol{\theta}_*\|/\|\boldsymbol{\theta}_*\|$. Each curve in the figures is the uniform average of 500 independently performed tests.

To ensure fair comparisons, the parameters of every method were carefully tuned to reach optimal performance per given scenario. Due to space limitations, lists of all parameters for each test are omitted. However, few things can be stated here about the parameters α and λ of Algorithm 1. With $\alpha \in [0.5, 1)$ in Algorithm 1, the general trend is that the fastest convergence speed of S-FM-HSDM is achieved for $\alpha = 0.5$. Moreover, with $\lambda \in (0, 2(1-\alpha)/L_{\nabla f})$, the fastest convergence speed was observed for values of λ close to $2(1-\alpha)/L_{\nabla f}$. In the case where $L_{\nabla f}$ is unknown, e.g., the case of $f := 0$, the values of $L_{\nabla f}$ used in the following tests were drawn from the interval $[10^{-3}, 10^{-1}]$.

A. $(\mathbf{a}_n)_n$ *is an IID Process*

With regards to the linear-regression model of Section I-B, process $(\mathbf{a}_n)_n$ is considered to be IID Gaussian. Independency is also assumed among the entries $([\mathbf{a}_n]_d)_{d=1}^D$ of each vector $\mathbf{a}_n$, $\forall n$. Given a value for the signal-to-noise ratio (SNR) in dB, the "power" of the additive noise $\mathbb{E}(\eta_n^2) = 10^{-\mathrm{SNR(dB)}/10}\|\boldsymbol{\theta}_*\|^2$ $\mathbb{E}([\mathbf{a}_n]_d^2)$. The SNR values $\{10, 20\}$ dB were examined and results are illustrated in Figures 1 and 2. Remarkably, the (HLS) formulation seems to be more appropriate than (CRegLS) for the sparse system-identification problem: The best performance among all methods is achieved by the proposed HRLSa, HRLSb and the non-convex ℓ_0-RLS.

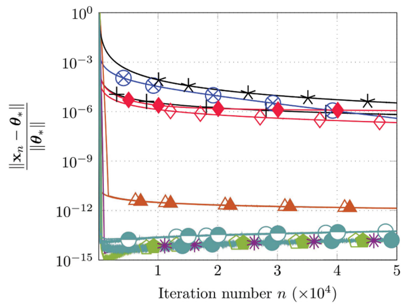
1) Time-Varying System: To test the ability of the methods to adapt to dynamic system changes, a typical AF test is considered here [2]: The sparsity level of the estimandum $\boldsymbol{\theta}_*$ changes abruptly at the time instance 2,500 from 1% to 10%, where the non-zero entries of $\boldsymbol{\theta}_*$ are re-allocated randomly.

As in the classical exponentially-weighted RLS [2, §30.6], (CRegLS) is modified to

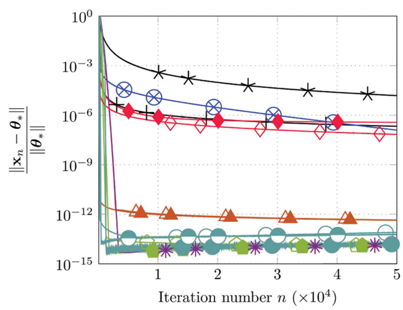
$$\min_{\mathbf{x}\in\mathbb{R}^D} \mathbb{E}\left[\frac{1}{2\Gamma_{\mathrm{f},n}}\sum_{\nu=1}^{n}\gamma_{\mathrm{f}}^{n-\nu}\left(\mathbf{a}_\nu^\intercal\mathbf{x} - b_\nu\right)^2\right] + \rho\|\mathbf{x}\|_1,$$

where $\Gamma_{\mathrm{f},n} := \sum_{\nu=1}^{n}\gamma_{\mathrm{f}}^{n-\nu}$ and $\gamma_{\mathrm{f}} \in (0, 1]$ is a "forgetting coefficient" that enforces a "short-memory" effect, via the exponential rule $\gamma_{\mathrm{f}}^{n-\nu}$, to account for the non-stationaries of the input-output data statistics. Results are illustrated in Figure 3. HRLSa, HRLSb and the Bayesian ASVB-MPL seem to be both agile and accurate in their estimation task.

2) No Additive Noise: Here, $\eta_n = 0$, or, SNR $= +\infty$, in the linear-regression model of Section I-B. Results are illustrated in Figure 4. The best performance is achieved by S-FM-HSDM(CRegLS), HRLSa, HRLSb, SDA, Prox-SVRG and SVRG-ADMM.

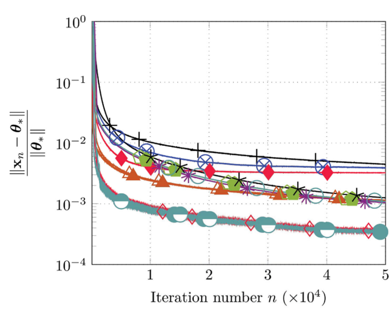


(a) IID $(\mathbf{a}_n)_n$.

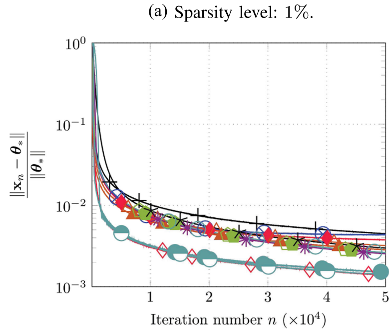


(b) AR $(\mathbf{a}_n)_n$.

Fig. 4. SNR $= +\infty$ (no additive noise); sparsity level: 10%; $\rho = 10^{-20}$.



(a) Sparsity level: 1%.

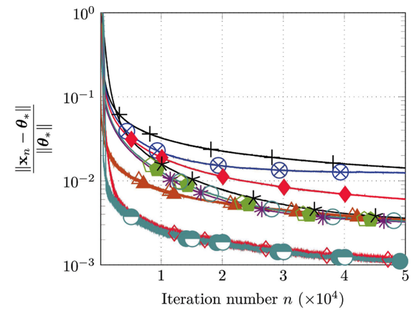


(b) Sparsity level: 10%.

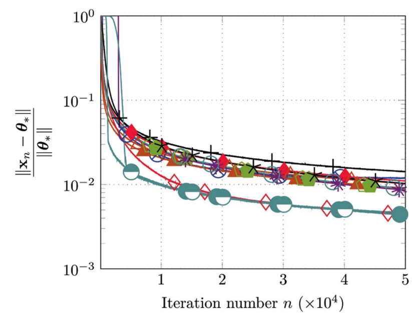
Fig. 5. AR $(\mathbf{a}_n)_n$; SNR $=$ 20 dB.

B. $(\mathbf{a}_n)_n$ *is an Auto-Regressive (AR) Process*

A first-order auto-regressive [AR(1)] process $(\mathbf{a}_n)_n$ is considered: $\forall n \in \mathbb{Z}_{\geq 0}$, $\mathbf{a}_n := \delta_{\text{AR}}\mathbf{a}_{n-1} + \boldsymbol{\upsilon}_n$, with $\delta_{\text{AR}} \in \mathbb{R}$ and $|\delta_{\text{AR}}| < 1$, $(\boldsymbol{\upsilon}_n)_{n=-1}^{+\infty}$ is a zero-mean Gaussian IID process, where independency is also assumed among the entries $([\boldsymbol{\upsilon}_n]_d)_{d=1}^D$ of each vector $\boldsymbol{\upsilon}_n$, and $\mathbf{a}_{-1} := \boldsymbol{\upsilon}_{-1}$. Recall here that $\mathbb{E}([\mathbf{a}_n]_d^2) = \mathbb{E}([\boldsymbol{\upsilon}_n]_d^2)/(1-\delta_{\text{AR}}^2)$ [38, (2.12.7)]. In all tests, the ratio $\mathbb{E}([\mathbf{a}_n]_d^2)/\,\mathbb{E}([\boldsymbol{\upsilon}_n]_d^2)$ is set to 5dB. Results are illustrated in Figures 5 and 6. The best performance is achieved by HRLSa, HRLSb and ℓ_0-RLS.



(a) Sparsity level: 1%.



(b) Sparsity level: 10%.

Fig. 6. AR $(\mathbf{a}_n)_n$; SNR $=$ 10 dB.

V. Conclusions

This paper presented a novel stochastic-approximation tool, namely the stochastic Fejér-monotone hybrid steepest descent method (S-FM-HSDM), to solve convex and affinely constrained composite minimization tasks. Noise contaminates the information about the task, affecting not only the loss terms but also the affine constraints. S-FM-HSDM provides a novel way of dealing with stochastic affine constraints via fixed-point sets of appropriate mappings, while retaining several desirable properties such as splitting of variables and constant step size. A performance analysis is also provided to identify the conditions under which the sequence of random variables, generated by S-FM-HSDM, converges a.s. to solutions of the latent noiseless minimization task. Several offsprings of S-FM-HSDM were presented in the context of a well-studied convexly regularized least-squares task. The versatility of S-FM-HSDM toward affine constraints opens the door to computationally efficient novel designs, called hierarchical recursive least squares, which, according to extensive numerical tests on synthetic data, appear to score the lowest estimation error across a variety of scenarios versus several state-of-the-art adaptive-filtering, stochastic-approximation and online-learning schemes. Due to space limitations, rates of convergence, other theoretical contributions and additional applications of S-FM-HSDM will be presented elsewhere.

Appendix A
Mathematical Preliminaries

Symbol $\mathcal{X}$ denotes a finite-dimensional Hilbert space, with inner product $\langle \cdot \mid \cdot \rangle$ and induced norm $\|\cdot\| := \langle \cdot \mid \cdot \rangle^{1/2}$. Given a linear operator $U : \mathcal{X} \to \mathcal{X}$, $\operatorname{ran} U$ and $\ker U$ denote the range and kernel spaces of U, respectively. Whenever $\mathcal{X} = \mathbb{R}^D$, the inner product of $\mathcal{X}$ is the standard dot-vector one: $\langle \mathbf{x} \mid \mathbf{x}' \rangle := \mathbf{x}^\intercal \mathbf{x}'$, $\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$.

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, with $\mathbb{E}(\cdot)$ denoting expectation [41]. Given a sub-σ-algebra $\mathcal{G}$ of Σ, the class of (non-negative) $\mathcal{G}$-measurable functions is denoted by $((\mathrm{m}\mathcal{G})^+)$ $\mathrm{m}\mathcal{G}$ [41]. Given an orthonormal basis $\{b_i\}_{i=1}^{\dim\mathcal{X}}$ of $\mathcal{X}$, $x : \Omega \to \mathcal{X}$ is called a random variable (RV) if there exist RVs $\{\gamma^i : \Omega \to \mathbb{R}\}_{i=1}^{\dim\mathcal{X}}$ s.t. $x = \sum_{i=1}^{\dim\mathcal{X}} \gamma^i b_i$. To avoid congestion in notations, a lowercase symbol x denotes both a RV, i.e., $x : \Omega \to \mathcal{X}$ with $x \in \mathrm{m}\Sigma$, and its realization $x(\omega)$, $\omega \in \Omega$. The abbreviation a.s. stands for either "almost surely," or, "almost sure" with respect to (w.r.t.) Ω, depending on the syntax of the sentence. A.s. convergence of the RV-sequence $(x_n)_n$ to $\bar{x}$ is denoted by $x_n \xrightarrow{\text{a.s.}}_n \bar{x}$. For an RV $\gamma : \Omega \to \mathbb{R}$, let $\mathbb{E}_{|\mathcal{G}}(\gamma)$ denote the conditional expectation of γ, conditioned on $\mathcal{G}$. The conditional expectation $\mathbb{E}_{|\mathcal{G}}(x)$ is defined by $\mathbb{E}_{|\mathcal{G}}(x) := \sum_i \mathbb{E}_{|\mathcal{G}}(\gamma^i) b_i$. Notice that $\mathbb{E}_{|\mathcal{G}}(x) \in \mathrm{m}\mathcal{G}$ [41]. Moreover, $\sigma(x)$ denotes the sub-σ-algebra of Σ generated by x [41, §3.8]. For a "random" linear mapping $Q : \mathcal{X} \to \mathcal{X}$ s.t. $Qx \in \mathrm{m}\Sigma$, $\forall x \in \mathcal{X}$, let $\mathbb{E}_{|\mathcal{G}}(Q)$ denote the linear mapping $\mathbb{E}_{|\mathcal{G}}(Q) : \mathcal{X} \to \mathcal{X} : x \mapsto \mathbb{E}_{|\mathcal{G}}(Q)x := \sum_i \gamma^i\, \mathbb{E}_{|\mathcal{G}}(Qb_i)$. Further, for each n, define the filtration $\mathcal{F}_n := \sigma(\{x_\nu\}_{\nu=0}^n)$, i.e., the sub-$\sigma$-algebra generated by the RVs $\{x_\nu\}_{\nu=0}^n$ [41, §10.1].

Given $\varphi \in \Gamma_0(\mathcal{X})$ [*cf.* Assumption 3(i)] and $\epsilon \in \mathbb{R}_{>0}$, the ϵ-subdifferential $\partial_\epsilon\varphi$ is the set-valued mapping which maps to any $z \in \mathcal{X}$ all ϵ-subgradients of φ at z: $\partial_\epsilon\varphi(z) := \{\xi \in \mathcal{X} | \varphi(z) + \langle \xi \mid x - z\rangle - \epsilon \leq \varphi(x), \forall x \in \mathcal{X}\}$. The graph of $\partial_\epsilon\varphi$ is defined as $\operatorname{gph}\partial_\epsilon\varphi := \{(z, \xi) \in \mathcal{X} \times \mathcal{X} | \xi \in \partial_\epsilon\varphi(z)\}$. Symbol $\partial\varphi$ stands for the subdifferential mapping, defined as $\partial\varphi(z) := \cap_{\epsilon\in\mathbb{R}_{>0}} \partial_\epsilon\varphi(z)$. Moreover, given $\lambda \in \mathbb{R}_{>0}$, the proximal mapping $\operatorname{Prox}_{\lambda\varphi} : \mathcal{X} \to \mathcal{X}$ is defined as [13]

$$z = \operatorname{Prox}_{\lambda\varphi}(x) \Leftrightarrow z = \arg\min_{a\in\mathcal{X}} \frac{1}{2}\|a - x\|^2 + \lambda\varphi(a)$$
$$\Leftrightarrow \exists \xi \in \partial\varphi(z) \text{ s.t. } z + \lambda\xi = x\,. \tag{9}$$

In the case where φ is the indicator function $\iota_{\mathcal{C}}$ for a closed convex set $\mathcal{C}$, i.e., $\iota_{\mathcal{C}}(x) := 0$, if $x \in \mathcal{C}$, and $\iota_{\mathcal{C}}(x) := +\infty$, if $x \notin \mathcal{C}$ [13], then $\operatorname{Prox}_{\lambda\varphi}$, for any $\lambda \in \mathbb{R}_{>0}$, is nothing but the (metric) projection mapping $P_{\mathcal{C}}$ onto $\mathcal{C}$. The following holds true for any member T of the family $\mathfrak{T}_{\mathcal{A}}$ in (3).

Fact 1 ([30, Prop. 2.12]): The affine constraint $\mathcal{A} = \operatorname{Fix} T = \ker(\mathrm{Id} - Q) + a = \ker U + a$, where $a \in \mathcal{A}$ and U stands for the square root of the positive $\mathrm{Id} - Q$, i.e., the (unique) positive mapping s.t. $U^2 = \mathrm{Id} - Q$ [34, Thm. 9.4-2]. Moreover, $\|\mathrm{Id} - Q\| \leq 1$ [30, (7)], and hence, $\|U\| \leq 1$.

Fact 2 ([30, Prop. 2.15]): For any $\lambda \in \mathbb{R}_{>0}$, define

$$\mathcal{A}_* := \{x \in \mathcal{A} | [\nabla f(x) + \partial(h + g)(x)] \cap \operatorname{ran} U \neq \emptyset\}\,,$$
$$\Upsilon_*^{(\lambda)} := \left\{(x, v) \in \mathcal{A} \times \mathcal{X} | \frac{-Uv}{\lambda} \in \nabla f(x) + \partial(h+g)(x)\right\}\,.$$

Then, x_* solves (P) $\Leftrightarrow x_* \in \mathcal{A}_* \Leftrightarrow \exists v_* \in \mathcal{X}$ s.t. $(x_*, v_*) \in \Upsilon_*^{(\lambda)}$.

The following lemma is used repeatedly in the sequel.

Lemma 5: For any sub-σ-algebra $\mathcal{G} \subset \Sigma$, $\forall (x, x') \in \mathrm{m}\mathcal{G} \times \mathrm{m}\Sigma$, $\mathbb{E}_{|\mathcal{G}}\langle x \mid x'\rangle = \langle x \mid \mathbb{E}_{|\mathcal{G}}(x')\rangle$ a.s. Given a linear mapping $Q : \mathcal{X} \to \mathcal{X}$, then, $\mathbb{E}_{|\mathcal{G}}(Qx') = Q\,\mathbb{E}_{|\mathcal{G}}(x')$. Further, if Q is "random," in the sense described earlier in this appendix, then $\mathbb{E}_{|\mathcal{G}}(Qx) = \mathbb{E}_{|\mathcal{G}}(Q)x$.

Proof: First, expectations are assumed to exist. Given an orthonormal basis $\{b_i\}_{i=1}^{\dim\mathcal{X}}$ of $\mathcal{X}$, there exist $\mathbb{R}$-valued RVs $\{\gamma^i, \gamma'^i\}_{i=1}^{\dim\mathcal{X}}$ s.t. $x = \sum_i \gamma^i b_i$ and $x' = \sum_i \gamma'^i b_i$ a.s. Hence, according to basic properties of conditional expectation [41, §9.7(c)(j)], $\mathbb{E}_{|\mathcal{G}}\langle x \mid x'\rangle = \sum_{i,i'} \gamma^i\, \mathbb{E}_{|\mathcal{G}}(\gamma'^{i'})\langle b_i \mid b_{i'}\rangle = \langle \sum_i \gamma^i b_i \mid \sum_{i'} \mathbb{E}_{|\mathcal{G}}(\gamma'^{i'}) b_{i'}\rangle = \langle x \mid \mathbb{E}_{|\mathcal{G}}(x')\rangle$ a.s. Further, $\mathbb{E}_{|\mathcal{G}}(Qx') = \mathbb{E}_{|\mathcal{G}}(\sum_i \gamma'^i Qb_i) = \sum_i \mathbb{E}_{|\mathcal{G}}(\gamma'^i) Qb_i = Q\,\mathbb{E}_{|\mathcal{G}}(x')$. Similar arguments can lead to the final claim of Lemma 5. ■

Appendix B
Proof of Lemma 1

First, notice that $\mathcal{A} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{R}^{1/2}\mathbf{x} - \mathbf{R}^{\dagger/2}\mathbf{r}\|^2$, where $\mathbf{R}^{1/2}$ is the square root of $\mathbf{R}$ and $\dagger$ stands for the Moore-Penrose pseudoinverse operation. The previous equality can be established by observing that $\mathbf{R}^{1/2}\mathbf{R}^{\dagger/2}$ is the orthogonal projection mapping onto the range space $\operatorname{ran}\mathbf{R}^{1/2} = \operatorname{ran}\mathbf{R}$, and that $\mathbf{r} \in \operatorname{ran}\mathbf{R}$ due to the normal equations. The claim that mappings (2) belong to $\mathfrak{T}_{\mathcal{A}}$ follows then directly from [30, (70a) and (70d)]. The claim of Lemma 1 with regards to the mappings (5) can be also established in a similar way; details are omitted.

Appendix C
Proof of Theorem 1

Line 5 of Algorithm 1 yields

$$x_{n+3/2} - x_{n+1/2} = T_{n+1}x_{n+1} - T_n^{(\alpha)}x_n$$
$$- \lambda\nabla f_{n+1}(x_{n+1}) + \lambda\nabla f_n(x_n)\,. \tag{10}$$

By line 6 of Algorithm 1 and (9), $\exists \xi_n \in \partial(h_{n-1} + g)(x_n)$, or, equivalently

$$(x_n, \xi_n) \in \operatorname{gph}\partial(h_{n-1} + g), \tag{11}$$

s.t. $x_{n-1/2} = x_n + \lambda\xi_n$, $\forall n \in \mathbb{Z}_{>0}$. Moreover, lines 2 and 3 of Algorithm 1, as well as (10) and (11) suggest

$$x_1 = T_0^{(\alpha)}x_0 - \lambda[\nabla f_0(x_0) + \xi_1]\,, \tag{12a}$$

$$x_{n+2} - x_{n+1} = T_{n+1}x_{n+1} - T_n^{(\alpha)}x_n$$
$$- \lambda[\nabla f_{n+1}(x_{n+1}) + \xi_{n+2}]$$
$$+ \lambda[\nabla f_n(x_n) + \xi_{n+1}]\,. \tag{12b}$$

By telescoping (12),

$$x_{n+1} = T_n x_n - \sum_{\nu=1}^{n-1}(T_\nu^{(\alpha)} - T_\nu)x_\nu - \lambda[\nabla f_n(x_n) + \xi_{n+1}]$$
$$= 2T_{n+1}^{(\alpha)}x_{n+1} - T_{n+1}x_{n+1} + (T_n^{(\alpha)}x_n - T_{n+1}^{(\alpha)}x_{n+1})$$
$$- \sum_{\nu=1}^{n+1}(T_\nu^{(\alpha)} - T_\nu)x_\nu - \lambda[\nabla f_n(x_n) + \xi_{n+1}]\,,$$

or, equivalently, via $T_\nu^{(\alpha)} - T_\nu = (1 - \alpha)(\mathrm{Id} - T_\nu)$,

$$(\mathrm{Id} + T_{n+1} - 2T_{n+1}^{(\alpha)})x_{n+1} + (T_{n+1}^{(\alpha)}x_{n+1} - T_n^{(\alpha)}x_n)$$
$$= -(1 - \alpha)\sum_{\nu=1}^{n+1}(\mathrm{Id} - T_\nu)x_\nu - \lambda[\nabla f_n(x_n) + \xi_{n+1}]\,, \tag{13}$$

where (13) holds true $\forall n \in \mathbb{Z}_{\geq 0}$. Furthermore,

$$(1 - 2\alpha)(T_{n+1} - \mathrm{Id})x_{n+1} + Q_{n+1}^{(\alpha)}(x_{n+1} - x_n)$$
$$+ \alpha(T_{n+1} - T_n)x_n$$

$$= (1-2\alpha)(T_{n+1} - \mathrm{Id})x_{n+1} + (T_{n+1}^{(\alpha)}x_{n+1} - T_{n+1}^{(\alpha)}x_n)$$
$$+ (T_{n+1}^{(\alpha)} - T_n^{(\alpha)})x_n$$
$$= (\mathrm{Id} + T_{n+1} - 2T_{n+1}^{(\alpha)})x_{n+1} + (T_{n+1}^{(\alpha)}x_{n+1} - T_n^{(\alpha)}x_n)$$
$$\overset{(13)}{=} -w_{n+1} - \lambda\left[\nabla f_n(x_n) + \xi_{n+1}\right], \tag{14}$$

where $\forall n \in \mathbb{Z}_{>0}$,

$$w_{n+1} := (1-\alpha)\sum_{\nu=1}^{n+1}(\mathrm{Id} - T_\nu)x_\nu\,. \tag{15}$$

Moreover, given $x_* \in \mathcal{A}$, define $\forall n \in \mathbb{Z}_{>0}$,

$$v_{n+1} := (1-\alpha)\sum_{\nu=1}^{n+1}U(x_\nu - x_*)\,. \tag{16}$$

where U is defined in Fact 1. Also, let $v_0 := 0 =: w_0$. Notice that v_{n+1} does not depend on the choice of $x_* \in \mathcal{A}$, since $\forall x_*' \in \mathcal{A}$, with $x_*' \neq x_*$, Fact 1 yields $x_*' - x_* \in \ker U$, and

$$v_{n+1} = (1-\alpha)\sum_{\nu=1}^{n+1}U(x_\nu - x_*' + x_*' - x_*)$$
$$= (1-\alpha)\sum_{\nu=1}^{n+1}U(x_\nu - x_*')\,.$$

Notice again by Fact 1 that $x_* \in \mathcal{A} \Leftrightarrow (\mathrm{Id} - T)x_* = 0$. Then, by (15) and (16), $\forall n \in \mathbb{Z}_{>0}$,

$$w_{n+1} = (1-\alpha)\sum_{\nu=1}^{n+1}(T - T_\nu)x_\nu$$
$$+ (1-\alpha)\sum_{\nu=1}^{n+1}\left[(\mathrm{Id} - T)x_\nu - (\mathrm{Id} - T)x_*\right]$$
$$= (1-\alpha)\sum_{\nu=1}^{n+1}(T - T_\nu)x_\nu$$
$$+ (1-\alpha)\sum_{\nu=1}^{n+1}(\mathrm{Id} - Q)(x_\nu - x_*)$$
$$= (1-\alpha)\sum_{\nu=1}^{n+1}(T - T_\nu)x_\nu + Uv_{n+1}\,. \tag{17}$$

Arbitrarily fix, now, $(x_*, v_*) \in \Upsilon_*^{(\lambda)}$ of Fact 2: $(\mathrm{Id} - T)x_* = 0$ and $\exists \xi_* \in \partial(h+g)(x_*)$ s.t. $Uv_* + \lambda[\nabla f(x_*) + \xi_*] = 0$. Then, by (14), (17),

$$(1-2\alpha)(T_{n+1} - T)x_{n+1} + (1-2\alpha)(T - \mathrm{Id})x_{n+1}$$
$$+ \alpha(Q_{n+1} - Q)(x_{n+1} - x_n)$$
$$+ Q^{(\alpha)}(x_{n+1} - x_n) + \alpha(T_{n+1} - T_n)x_n$$
$$= -(1-\alpha)\sum_{\nu=1}^{n+1}(T - T_\nu)x_\nu$$
$$- U(v_{n+1} - v_*) - \lambda\left[\nabla f_n(x_n) - \nabla f(x_*)\right]$$
$$- \lambda(\xi_{n+1} - \xi_*)$$
$$\Leftrightarrow (1-2\alpha)(\mathrm{Id} - T)x_{n+1} + Q^{(\alpha)}(x_n - x_{n+1})$$
$$+ U(v_* - v_{n+1}) + (1-2\alpha)(T - T_{n+1})x_{n+1}$$
$$+ \alpha(Q - Q_{n+1})(x_{n+1} - x_n)$$
$$+ \alpha(T_n - T_{n+1})x_n$$
$$+ (1-\alpha)\sum_{\nu=1}^{n+1}(T_\nu - T)x_\nu$$
$$= \lambda\left[\nabla f_n(x_n) - \nabla f(x_*) + \xi_{n+1} - \xi_*\right]\,. \tag{18}$$

Based on Assumption 3(ii), the application of the Baillon-Haddad theorem [13, Cor. 18.16] to f suggests that

$$\frac{2\lambda}{L_{\nabla f}}\|\nabla f(x_n) - \nabla f(x_*)\|^2$$
$$\leq 2\lambda\langle x_n - x_* \mid \nabla f(x_n) - \nabla f(x_*)\rangle$$
$$= 2\lambda\langle x_{n+1} - x_* \mid \nabla f_n(x_n) - \nabla f(x_*) + \xi_{n+1} - \xi_*\rangle$$
$$+ 2\lambda\langle x_{n+1} - x_* \mid (\nabla f - \nabla f_n)x_n\rangle$$
$$+ 2\lambda\langle x_n - x_{n+1} \mid \nabla f(x_n) - \nabla f(x_*)\rangle$$
$$+ 2\lambda\langle x_* - x_{n+1} \mid \xi_{n+1} - \xi_*\rangle$$
$$\overset{(18)}{=} 2(1-2\alpha)\langle x_{n+1} - x_* \mid (\mathrm{Id} - T)x_{n+1}\rangle$$
$$+ 2\langle x_{n+1} - x_* \mid Q^{(\alpha)}(x_n - x_{n+1})\rangle$$
$$+ 2\langle x_{n+1} - x_* \mid U(v_* - v_{n+1})\rangle$$
$$+ 2\lambda\langle x_n - x_{n+1} \mid \nabla f(x_n) - \nabla f(x_*)\rangle$$
$$+ 2(1-2\alpha)\langle x_{n+1} - x_* \mid (T - T_{n+1})x_{n+1}\rangle$$
$$+ 2\alpha\langle x_{n+1} - x_* \mid (Q - Q_{n+1})(x_{n+1} - x_n)\rangle$$
$$+ 2\alpha\langle x_{n+1} - x_* \mid (T_n - T_{n+1})x_n\rangle$$
$$+ 2(1-\alpha)\left\langle x_{n+1} - x_* \mid \sum_{\nu=1}^{n+1}(T_\nu - T)x_\nu\right\rangle$$
$$+ 2\lambda\langle x_{n+1} - x_* \mid (\nabla f - \nabla f_n)x_n\rangle$$
$$+ 2\lambda\langle x_* - x_{n+1} \mid \xi_{n+1} - \xi_*\rangle$$
$$\leq 2(1-2\alpha)\langle x_{n+1} - x_* \mid (\mathrm{Id} - T)x_{n+1} - (\mathrm{Id} - T)x_*\rangle$$
$$+ 2\langle x_{n+1} - x_* \mid Q^{(\alpha)}(x_n - x_{n+1})\rangle$$
$$+ 2\langle x_* - x_{n+1} \mid U(v_{n+1} - v_*)\rangle$$
$$+ \frac{\lambda L_{\nabla f}}{2}\|x_n - x_{n+1}\|^2 + \frac{2\lambda}{L_{\nabla f}}\|\nabla f(x_n) - \nabla f(x_*)\|^2 \tag{19a}$$
$$+ 2(1-2\alpha)\langle x_{n+1} - x_* \mid (T - T_{n+1})x_{n+1}\rangle \tag{19b}$$
$$+ 2\alpha\langle x_{n+1} - x_* \mid (Q - Q_{n+1})(x_{n+1} - x_n)\rangle$$
$$+ 2\alpha\langle x_{n+1} - x_* \mid (T_n - T_{n+1})x_n\rangle$$
$$+ 2(1-\alpha)\left\langle x_{n+1} - x_* \mid \sum_{\nu=1}^{n+1}(T_\nu - T)x_\nu\right\rangle$$
$$+ 2\lambda\langle x_{n+1} - x_* \mid (\nabla f - \nabla f_n)x_n\rangle$$
$$+ 2\lambda\langle x_* - x_{n+1} \mid \xi_{n+1} - \xi_*\rangle, \tag{19c}$$

where $2\langle\sqrt{\beta}a \mid b/\sqrt{\beta}\rangle \leq \beta\|a\|^2 + \|b\|^2/\beta$, $a := x_n - x_{n+1}$, $b := \nabla f(x_n) - \nabla f(x_*)$ and $\beta := L_{\nabla f}/2$, were used in (19a). Let ϑ_n be the RV defined by the expression which starts from (19b) and ends at (19c).

Let $y := (x, v)$ denote an element of the finite-dimensional Hilbert space $(\mathcal{X}^2, \langle\cdot \mid \cdot\rangle_{\mathcal{X}^2})$, where the inner product is defined as $\langle(x,v) \mid (x',v')\rangle_{\mathcal{X}^2} := \langle x \mid x'\rangle + \langle v \mid v'\rangle$, for any $(x,v),(x',v') \in \mathcal{X}^2$. Let also the bounded linear and self-adjoint operator $\Theta : \mathcal{X}^2 \to \mathcal{X}^2 : (x,v) \mapsto (Q^{(\alpha)}x, v/(1-\alpha))$. By virtue of the positivity of Q, $\forall x$,

$$\begin{aligned}\langle Q^{(\alpha)}x \mid x\rangle &= \alpha\langle Qx \mid x\rangle + (1-\alpha)\|x\|^2 \\ &\geq (1-\alpha)\|x\|^2, \end{aligned} \tag{20}$$

which renders Θ strongly positive (recall $\alpha < 1$). Operator Θ induces thus the Hilbert space $\mathcal{X}^2_\Theta := (\mathcal{X}^2, \langle\cdot \mid \cdot\rangle_{\mathcal{X}^2_\Theta})$ with inner product $\langle\cdot \mid \cdot\rangle_{\mathcal{X}^2_\Theta} := \langle\Theta(\cdot) \mid \cdot\rangle_{\mathcal{X}^2}$. Then, upon defining $y_* := (x_*, v_*)$, (19c) yields

$$\begin{aligned}
0 &\leq 2(1-2\alpha)\langle x_{n+1} - x_* \mid (\mathrm{Id} - Q)(x_{n+1} - x_*)\rangle \\
&\quad + 2\langle Q^{(\alpha)}(x_n - x_{n+1}) \mid x_{n+1} - x_*\rangle \\
&\quad + 2\langle U(x_* - x_{n+1}) \mid v_{n+1} - v_*\rangle \\
&\quad + \frac{\lambda L_{\nabla f}}{2}\|x_n - x_{n+1}\|^2 + \vartheta_n \\
&\leq 2\langle Q^{(\alpha)}(x_n - x_{n+1}) \mid x_{n+1} - x_*\rangle + \frac{\lambda L_{\nabla f}}{2}\|x_n - x_{n+1}\|^2 \\
&\quad + \frac{2}{1-\alpha}\langle v_n - v_{n+1} \mid v_{n+1} - v_*\rangle + \vartheta_n \qquad (21a)\\
&= 2\langle\Theta(y_n - y_{n+1}) \mid y_{n+1} - y_*\rangle_{\mathcal{X}^2} \\
&\quad + \frac{\lambda L_{\nabla f}}{2}\|x_n - x_{n+1}\|^2 + \vartheta_n \\
&= \|y_n - y_*\|^2_{\mathcal{X}^2_\Theta} - \|y_{n+1} - y_*\|^2_{\mathcal{X}^2_\Theta} - \|y_{n+1} - y_n\|^2_{\mathcal{X}^2_\Theta} \\
&\quad + \frac{\lambda L_{\nabla f}}{2}\|x_n - x_{n+1}\|^2 + \vartheta_n \\
&\leq \|y_n - y_*\|^2_{\mathcal{X}^2_\Theta} - \|y_{n+1} - y_*\|^2_{\mathcal{X}^2_\Theta} \\
&\quad - (1-\zeta)\|y_{n+1} - y_n\|^2_{\mathcal{X}^2_\Theta} + \vartheta_n, \qquad (21b)
\end{aligned}$$

where the positivity of $\mathrm{Id} - Q$ from Fact 1, $\alpha \geq 1/2$, and $v_n - v_{n+1} = (1-\alpha)U(x_* - x_{n+1})$, from (16), were used in (21a). Any $\zeta \in (\lambda L_{\nabla f}/[2(1-\alpha)], 1)$ justifies (21b), since for $\lambda < 2(1-\alpha)/L_{\nabla f}$ and $\forall y := (x,v) \in \mathcal{X}^2$, (20) suggests

$$\begin{aligned}\frac{\lambda L_{\nabla f}}{2}\|x\|^2 &< \zeta(1-\alpha)\|x\|^2 \leq \zeta\langle x \mid Q^{(\alpha)}x\rangle \\ &\leq \zeta\langle x \mid Q^{(\alpha)}x\rangle + \zeta\frac{1}{1-\alpha}\|v\|^2 = \zeta\|y\|^2_{\mathcal{X}^2_\Theta}.\end{aligned}$$

Notice by (16) that $v_n \in \mathrm{m}\mathcal{F}_n$. Hence, $y_n = (x_n, v_n) \in \mathrm{m}\mathcal{F}_n$. Applying $\mathbb{E}_{|\mathcal{F}_n}(\cdot)$ to (21b) yields

$$\begin{aligned}&\mathbb{E}_{|\mathcal{F}_n}\|y_{n+1} - y_*\|^2_{\mathcal{X}^2_\Theta} + (1-\zeta)\,\mathbb{E}_{|\mathcal{F}_n}\|y_{n+1} - y_n\|^2_{\mathcal{X}^2_\Theta} \\ &\leq \|y_n - y_*\|^2_{\mathcal{X}^2_\Theta} + \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ \quad \text{a.s.,}\end{aligned} \tag{22}$$

where $\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ := \max\{0, \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)\}$. Since (x_*, v_*) was arbitrarily chosen from $\Upsilon^{(\lambda)}_*$ (*cf.* Fact 2), Assumption 6 and [10, Prop. 2.3] render $(y_n)_n$ stochastic quasi-Fejér monotonous w.r.t. $\Upsilon^{(\lambda)}_*$; thus, bounded a.s. Due to $y_n = (x_n, v_n)$, sequences $(x_n)_n$ and $(v_n)_n$ are also bounded a.s.

This paragraph proves the claim that (8) suffice for Assumption 6 to hold true. Via (8g), a.s.,

$$\begin{aligned}&(h+g)(\mathbf{x}_*) \\ &\geq (h+g)(\mathbf{x}_{n+1}) + \langle\mathbf{x}_* - \mathbf{x}_{n+1} \mid \mathbb{E}_{|\mathcal{F}_{n+1}}(\boldsymbol{\xi}_{n+1})\rangle - \epsilon_{n+1}.\end{aligned}$$

Moreover, $(\mathbf{x}_*, \boldsymbol{\xi}_*) \in \operatorname{gph}\partial(h+g) \Rightarrow (h+g)(\mathbf{x}_{n+1}) \geq (h+g)(\mathbf{x}_*) + \langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid \boldsymbol{\xi}_*\rangle$. Hence, by adding the previous two inequalities, $\langle\mathbf{x}_* - \mathbf{x}_{n+1} \mid \mathbb{E}_{|\mathcal{F}_{n+1}}(\boldsymbol{\xi}_{n+1}) - \boldsymbol{\xi}_*\rangle \leq \epsilon_{n+1}$. The "tower property" of conditional probability suggests $\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n) = \mathbb{E}_{|\mathcal{F}_n}\mathbb{E}_{|\mathcal{F}_{n+1}}(\vartheta_n)$ [41, §9.7(i)]. By $\mathbf{x}_{n+1} - \mathbf{x}_* \in \mathrm{m}\mathcal{F}_{n+1}$, (8), (19c) and Lemma 5,

$$\begin{aligned}
&\mathbb{E}_{|\mathcal{F}_{n+1}}(\vartheta_n) \\
&= 2(1-2\alpha)\langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid \mathbb{E}_{|\mathcal{F}_{n+1}}[(T - T_{n+1})\mathbf{x}_{n+1}]\rangle \\
&\quad + 2\alpha\langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid \mathbb{E}_{|\mathcal{F}_{n+1}}[(Q - Q_{n+1})(\mathbf{x}_{n+1} - \mathbf{x}_n)]\rangle \\
&\quad + 2\alpha\langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid \mathbb{E}_{|\mathcal{F}_{n+1}}[(T_n - T_{n+1})\mathbf{x}_n]\rangle \\
&\quad + 2(1-\alpha)\langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid -\mathbf{t}_{n+1}\rangle \\
&\quad + 2\lambda\langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid \mathbb{E}_{|\mathcal{F}_{n+1}}[(\nabla f - \nabla f_{n+1})\mathbf{x}_n]\rangle \\
&\quad + 2\lambda\langle\mathbf{x}_{n+1} - \mathbf{x}_* \mid \mathbb{E}_{|\mathcal{F}_{n+1}}[(\nabla f_{n+1} - \nabla f_n)\mathbf{x}_n]\rangle \\
&\quad + 2\lambda\langle\mathbf{x}_* - \mathbf{x}_{n+1} \mid \mathbb{E}_{|\mathcal{F}_{n+1}}(\boldsymbol{\xi}_{n+1}) - \boldsymbol{\xi}_*\rangle \leq 2\lambda\epsilon_{n+1}. \qquad (23)
\end{aligned}$$

Thus, $\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n) \leq 2\lambda\,\mathbb{E}_{|\mathcal{F}_n}(\epsilon_{n+1}) \Rightarrow \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ \leq 2\lambda\,\mathbb{E}_{|\mathcal{F}_n}(\epsilon_{n+1}) \Rightarrow \sum_n \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ \leq 2\lambda\sum_n \mathbb{E}_{|\mathcal{F}_n}(\epsilon_{n+1})$ a.s. By (8g), $\sum_n \mathbb{E}[\mathbb{E}_{|\mathcal{F}_n}(\epsilon_{n+1})] = \sum_n \mathbb{E}(\epsilon_{n+1}) < +\infty$, and $\psi := 2\lambda\sum_n \mathbb{E}_{|\mathcal{F}_n}(\epsilon_{n+1})$, a.s., satisfies Assumption 6.

Going back to the general setting, define now space $\mathfrak{X} := L^2[(\Omega, \Sigma, \mathbb{P}), \mathcal{X}]$ of (equivalent classes of Borel) measurable functions, or, RVs $x : \Omega \to \mathcal{X}$ s.t. $\int_\Omega \|x(\omega)\|^2\,\mathbb{P}(d\omega) < +\infty$. This RV-space $\mathfrak{X}$ turns out to be a real Hilbert one with inner product $\langle x \mid x'\rangle_{\mathfrak{X}} := \mathbb{E}(\langle x \mid x'\rangle) := \int_\Omega \langle x(\omega) \mid x'(\omega)\rangle\,\mathbb{P}(d\omega)$, $\forall(x,x') \in \mathfrak{X} \times \mathfrak{X}$ [13, Ex. 2.5, p. 28]. Hilbert space $\mathfrak{X}^2_\Theta := L^2[(\Omega, \Sigma, \mathbb{P}), \mathcal{X}^2_\Theta]$ is similarly defined, with inner product $\langle y \mid y'\rangle_{\mathfrak{X}} := \mathbb{E}(\langle y \mid y'\rangle_{\mathcal{X}^2_\Theta})$.

Application of $\mathbb{E}(\cdot)$ to (22), under the light of $\mathbb{E}(\cdot) = \mathbb{E}[\mathbb{E}_{|\mathcal{F}_n}(\cdot)]$ [41, §9.7(a)], yields

$$\begin{aligned}&\|y_{n+1} - y_*\|^2_{\mathfrak{X}^2_\Theta} + (1-\zeta)\|y_{n+1} - y_n\|^2_{\mathfrak{X}^2_\Theta} \\ &\leq \|y_n - y_*\|^2_{\mathfrak{X}^2_\Theta} + \mathbb{E}[\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+].\end{aligned} \tag{24}$$

The monotone-convergence theorem [41, §5.3] and Assumption 6 imply that $\sum_n \mathbb{E}[\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+] \leq \mathbb{E}(\psi) < +\infty$. As such, (24) renders $(y_n)_n$ quasi-Fejér (of type III) w.r.t. $\Upsilon^{(\lambda)}_*$ and, thus, bounded within $\mathfrak{X}^2_\Theta$ [42, Prop. 3.3]. Hence, both $(x_n)_n$ and $(v_n)_n$ are bounded within $\mathfrak{X}$. Moreover, by telescoping (24), $\forall n$, $\sum_{\nu=0}^n \|y_{\nu+1} - y_\nu\|^2_{\mathfrak{X}^2_\Theta} \leq (1-\zeta)^{-1}[\|y_0 - y_*\|^2_{\mathfrak{X}^2_\Theta} + \mathbb{E}(\psi)]$. Hence, $\sum_{n=0}^{+\infty}\mathbb{E}(\|y_{n+1} - y_n\|^2_{\mathcal{X}^2_\Theta}) < +\infty$. By [41, §6.5], $\|y_{n+1} - y_n\|^2_{\mathcal{X}^2_\Theta} \xrightarrow{\text{a.s.}}_n 0$, and thus, $(y_{n+1} - y_n) \xrightarrow{\text{a.s.}}_n 0$ by virtue of the strong positivity of Θ. Consequently,

$$(x_{n+1} - x_n) \xrightarrow{\text{a.s.}}_n 0, \quad (v_{n+1} - v_n) \xrightarrow{\text{a.s.}}_n 0; \tag{25}$$

hence, both $(x_{n+1} - x_n)_n$ and $(v_{n+1} - v_n)_n$ are bounded a.s.

By (17),

$$\begin{aligned} & w_{n+1} - w_n \\ & = (1-\alpha)(T - T_{n+1})x_{n+1} + U(v_{n+1} - v_n) \\ & = (1-\alpha)(Tx_{n+1} - Tx_n) + (1-\alpha)(T - T_{n+1})x_n \\ & \quad + (1-\alpha)(T_{n+1}x_n - T_{n+1}x_{n+1}) + U(v_{n+1} - v_n) \\ & = (1-\alpha)Q(x_{n+1} - x_n) + (1-\alpha)(Q - Q_{n+1})x_n \\ & \quad + (1-\alpha)(\pi - \pi_{n+1}) + (1-\alpha)Q_{n+1}(x_n - x_{n+1}) \\ & \quad + U(v_{n+1} - v_n)\,. \end{aligned}$$

Since $(x_n)_n$ is bounded a.s., there exists $C_1 := C_1(\omega) \in \mathbb{R}_{>0}$ s.t. $\|x_n\| \leq C_1$, $\forall n$, a.s. Consequently, by Fact 1, Assumptions 1(i) and 1(ii),

$$\begin{aligned} & \|w_{n+1} - w_n\| \\ & \leq (1-\alpha)\|Q\|\,\|x_{n+1} - x_n\| + (1-\alpha)\|Q - Q_{n+1}\|\,\|x_n\| \\ & \quad + (1-\alpha)\|\pi - \pi_{n+1}\| + (1-\alpha)\|Q_{n+1}\|\,\|x_n - x_{n+1}\| \\ & \quad + \|U\|\,\|v_{n+1} - v_n\| \\ & \leq \|x_{n+1} - x_n\| + C_1\|Q - Q_{n+1}\| + \|\pi - \pi_{n+1}\| \\ & \quad + \|x_n - x_{n+1}\| + \|v_{n+1} - v_n\|, \end{aligned} \tag{26}$$

Via Assumption 1(iii) and (25), (26) yields

$$w_{n+1} - w_n \xrightarrow{\text{a.s.}}_n 0\,. \tag{27}$$

Hence, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $n_{\#} := n_{\#}(\omega) \in \mathbb{Z}_{\geq 0}$ s.t. $\forall n \geq n_{\#}$, $\|w_{n+1} - w_n\| \leq \epsilon$ a.s. Notice also via Jensen's inequality [41, §9.7(h)] that $\|\mathbb{E}_{|\mathcal{F}_n}(w_{n+1}) - \mathbb{E}_{|\mathcal{F}_n}(w_n)\| = \|\mathbb{E}_{|\mathcal{F}_n}(w_{n+1} - w_n)\| \leq \mathbb{E}_{|\mathcal{F}_n}(\|w_{n+1} - w_n\|) \leq \epsilon$, and thus, $\limsup_n \|\mathbb{E}_{|\mathcal{F}_n}(w_{n+1}) - \mathbb{E}_{|\mathcal{F}_n}(w_n)\| \leq \epsilon$ a.s. Since ϵ is chosen arbitrarily,

$$\mathbb{E}_{|\mathcal{F}_n}(w_{n+1}) - \mathbb{E}_{|\mathcal{F}_n}(w_n) \xrightarrow{\text{a.s.}}_n 0\,. \tag{28}$$

Furthermore, by (15), $(\mathrm{Id} - T_{n+1})x_{n+1} = (w_{n+1} - w_n)/(1 - \alpha)$, which, together with (27), yields

$$(\mathrm{Id} - T_n)x_n \xrightarrow{\text{a.s.}}_n 0\,. \tag{29}$$

As such, $((\mathrm{Id} - T_n)x_n)_n$ is bounded a.s. Moreover,

$$\begin{aligned} & \|(\mathrm{Id} - T)x_n\| \\ & \leq \|[(\mathrm{Id} - T) - (\mathrm{Id} - T_n)]x_n\| + \|(\mathrm{Id} - T_n)x_n\| \\ & \leq \|Q_n - Q\|\,\|x_n\| + \|\pi_n - \pi\| + \|(\mathrm{Id} - T_n)x_n\| \\ & \leq C_1\|Q_n - Q\| + \|\pi_n - \pi\| + \|(\mathrm{Id} - T_n)x_n\|\,. \end{aligned}$$

Referring again to Assumption 1(iii), (29) and the previous inequality yield

$$(\mathrm{Id} - T)x_n \xrightarrow{\text{a.s.}}_n 0\,. \tag{30}$$

Moreover,

$$\begin{aligned} \|(T_n - T_{n+1})x_n\| & \leq \|(Q_n - Q_{n+1})x_n\| + \|\pi_n - \pi_{n+1}\| \\ & \leq \|Q_n - Q_{n+1}\|\|x_n\| + \|\pi_n - \pi_{n+1}\| \\ & \leq C_1\|Q_n - Q_{n+1}\| + \|\pi_n - \pi_{n+1}\|, \end{aligned}$$

which, according to Assumption 1(iii), leads to

$$(T_n - T_{n+1})x_n \xrightarrow{\text{a.s.}}_n 0\,. \tag{31}$$

Hence, $((T_n - T_{n+1})x_n)_n$ is bounded a.s.

Assumptions 3(iv) and 3(v) suggest that for any $z \in \mathcal{X}$,

$$\begin{aligned} & \|\nabla f_n(x_n) - \nabla f(z)\|^2 \\ & \leq 2\|\nabla f_n(x_n) - \nabla f_n(z)\|^2 + 2\|(\nabla f_n - \nabla f)z\|^2 \\ & \leq 2L_n^2\|x_n - z\|^2 + 2\|(\nabla f_n - \nabla f)z\|^2 \\ & \leq 2C_{\text{Lip}}^2\|x_n - z\|^2 + 2\|(\nabla f_n - \nabla f)z\|^2\,. \end{aligned} \tag{32}$$

The a.s. boundedness of $(x_n)_n$ implies the a.s. boundedness of $(x_n - z)_n$. Moreover, Assumption 3(vi) suggests the a.s. boundedness of $((\nabla f_n - \nabla f)z)_n$. Due also to $\|\nabla f_n(x_n)\| \leq \|\nabla f_n(x_n) - \nabla f(z)\| + \|\nabla f(z)\|$, (32) guarantees that $(\nabla f_n(x_n))_n$ is bounded a.s. Notice also by (14),

$$\begin{aligned} \xi_{n+1} + \frac{1}{\lambda}w_{n+1} = {} & \frac{1-2\alpha}{\lambda}(\mathrm{Id} - T_{n+1})x_{n+1} \\ & + \frac{1}{\lambda}Q_{n+1}^{(\alpha)}(x_n - x_{n+1}) \\ & + \frac{\alpha}{\lambda}(T_n - T_{n+1})x_n - \nabla f_n(x_n)\,. \end{aligned} \tag{33}$$

Due also to the a.s. boundedness of $((\mathrm{Id} - T_n)x_n)_n$, $(x_{n+1} - x_n)_n$, $((T_{n+1} - T_n)x_n)_n$ and $(\nabla f_n(x_n))_n$, there exists $C_2 := C_2(\omega) \in \mathbb{R}_{>0}$ s.t.

$$\begin{aligned} & \|\xi_{n+1} + \frac{1}{\lambda}w_{n+1}\| \\ & \leq \frac{2\alpha - 1}{\lambda}\|(\mathrm{Id} - T_{n+1})x_{n+1}\| + \frac{1}{\lambda}\|Q_{n+1}^{(\alpha)}(x_n - x_{n+1})\| \\ & \quad + \frac{\alpha}{\lambda}\|(T_n - T_{n+1})x_n\| + \|\nabla f_n(x_n)\| \\ & \leq \frac{2\alpha - 1}{\lambda}\|(\mathrm{Id} - T_{n+1})x_{n+1}\| + \frac{1}{\lambda}\|x_n - x_{n+1}\| \\ & \quad + \frac{\alpha}{\lambda}\|(T_n - T_{n+1})x_n\| + \|\nabla f_n(x_n)\| \leq C_2 \quad \text{a.s.} \end{aligned} \tag{34}$$

Lemma 6: The cluster-point set $\mathfrak{C}[(y_n)_n]$ of sequence $(y_n)_n$, as well as $\mathfrak{C}[(x_n)_n]$ and $\mathfrak{C}[(v_n)_n]$ are nonempty. If $\bar{y} =: (\bar{x}, \bar{v}) \in \mathfrak{C}[(y_n)_n]$, then, $\bar{x} \in \mathfrak{C}[(x_n)_n]$ and $\bar{v} \in \mathfrak{C}[(v_n)_n]$. For any $\bar{x} \in \mathfrak{C}[(x_n)_n]$, there exists $\bar{v} \in \mathfrak{C}[(v_n)_n]$ s.t. $\bar{y} := (\bar{x}, \bar{v}) \in \mathfrak{C}[(y_n)_n]$. All of the previous statements hold true a.s.

Proof: Since $(y_n)_n$ is bounded a.s. [*cf.* discussion after (22)], its set of cluster points is nonempty [13, Fact 2.26(iii) and Lem. 2.37]. Moreover, due to the boundedness of $(x_n)_n$ and $(v_n)_n$, $\mathfrak{C}[(x_n)_n]$ and $\mathfrak{C}[(v_n)_n]$ are also nonempty. For any cluster point $\bar{y} =: (\bar{x}, \bar{v}) \in \mathfrak{C}[(y_n)_n]$, there exists a subsequence $(n_k)_k$ s.t. $y_{n_k} := (x_{n_k}, v_{n_k}) \xrightarrow{\text{a.s.}}_k (\bar{x}, \bar{v})$, i.e., $\bar{x} \in \mathfrak{C}[(x_n)_n]$ and $\bar{v} \in \mathfrak{C}[(v_n)_n]$. On the other hand, given any $\bar{x} \in \mathfrak{C}[(x_n)_n]$, there exists a subsequence $(x_{n_k})_k$ s.t. $x_{n_k} \xrightarrow{\text{a.s.}}_k \bar{x}$. Since $(v_n)_n$ is bounded, passing to a subsequence of $(n_k)_k$ if necessary (avoided here to avoid notational congestion), there exists $\bar{v} \in \mathfrak{C}[(v_n)_n]$ s.t. $v_{n_k} \xrightarrow{\text{a.s.}}_k \bar{v}$, and thus, $y_{n_k} := (x_{n_k}, v_{n_k}) \xrightarrow{\text{a.s.}}_k (\bar{x}, \bar{v}) =: \bar{y} \in \mathfrak{C}[(y_n)_n]$. ■

Choose, now, arbitrarily a cluster point $\bar{y} =: (\bar{x}, \bar{v}) \in \mathfrak{C}[(y_n)_n] \neq \emptyset$. Hence, there exists a subsequence $(n_k)_k$ s.t. $y_{n_k} =: (x_{n_k}, v_{n_k}) \xrightarrow{\text{a.s.}}_k (\bar{x}, \bar{v})$. Then, by (30), applied to $(x_{n_k})_k$, and by the nonexpansivity (thus continuity) of T,

$$\bar{x} \in \operatorname{Fix} T = \mathcal{A} \quad \text{a.s.} \tag{35}$$

Setting $n = n_k$ and $z = \bar{x}$ in (32) yields $\|\nabla f_{n_k}(x_{n_k}) - \nabla f(\bar{x})\|^2 \leq 2C_{\text{Lip}}^2 \|x_{n_k} - \bar{x}\|^2 + 2\|(\nabla f_{n_k} - \nabla f)\bar{x}\|^2$, which, by Assumption 3(iv) and $x_{n_k} \xrightarrow{\text{a.s.}}_k \bar{x}$, deduces $\nabla f_{n_k}(x_{n_k}) \xrightarrow{\text{a.s.}}_k \nabla f(\bar{x})$. Moreover, by Assumption 4(ii), $\nabla f(x_{n_k}) \in \mathrm{m}\mathcal{F}_{n_k}$ and

$$\begin{aligned}\mathbb{E}_{|\mathcal{F}_{n_k}}[\nabla f_{n_k}(x_{n_k})] &= \mathbb{E}_{|\mathcal{F}_{n_k}}[\nabla f(x_{n_k})] + \varepsilon_{n_k}^f(x_{n_k}) \\ &= \nabla f(x_{n_k}) + \varepsilon_{n_k}^f(x_{n_k}) \xrightarrow{\text{a.s.}}_k \nabla f(\bar{x}).\end{aligned} \tag{36}$$

Lemma 7: The range space $\operatorname{ran} U$ is closed in the strong topology of $\mathfrak{X}$, i.e., $\operatorname{ran} U = \overline{\operatorname{ran}}\, U$, where $\overline{\operatorname{ran}}\, U$ denotes the smallest closed set containing $\operatorname{ran} U$ (notice that $\mathfrak{X} := L^2[(\Omega, \Sigma, \mathbb{P}), \mathcal{X}]$ is infinite dimensional).

Proof: Since $\operatorname{ran} U$ is finite dimensional within $\mathcal{X}$, there exists an orthonormal set $\{u_i\}_{i=1}^{\operatorname{rank} U}$ which spans $\operatorname{ran} U$. Hence, for any $z \in \mathfrak{X} \cap \operatorname{ran} U$, there exist real-valued RVs $\{\gamma^i\}_{i=1}^{\operatorname{rank} U}$ s.t. $z = \sum_i \gamma^i u_i$ a.s. Due to the orthonormality of u_is, it can be verified that $\|z\|_{\mathfrak{X}}^2 = \sum_i \mathbb{E}[(\gamma^i)^2]$. Thus, $z \in \mathfrak{X} \Rightarrow \gamma^i \in L^2[(\Omega, \Sigma, \mathbb{P}), \mathbb{R}]$, $\forall i$. Consider, now, a sequence $(z_k)_k \subset \operatorname{ran} U \cap \mathfrak{X}$, with the associated coefficients $\{\gamma_k^i | i \in \{1, \ldots, \operatorname{rank} U\}; k \in \mathbb{Z}_{\geq 0}\} \subset L^2[(\Omega, \Sigma, \mathbb{P}), \mathbb{R}]$. Let $\bar{z}$ s.t. $z_k \xrightarrow{\mathfrak{X}}_k \bar{z}$. Since $(z_k)_k$ is convergent, it is also Cauchy [13], and thus, $(\gamma_k^i)_k$ is also Cauchy, $\forall i$. By virtue of the completeness of the Hilbert space $L^2[(\Omega, \Sigma, \mathbb{P}), \mathbb{R}]$ [13], there exists $\bar{\gamma}^i$ s.t. $\gamma_k^i \xrightarrow{L^2[(\Omega,\Sigma,\mathbb{P}),\mathbb{R}]}_k \bar{\gamma}^i$, $\forall i$. In other words, $\bar{z} = \lim_{k\to\infty} \sum_i \gamma_k^i u_i = \sum_i \lim_{k\to\infty} \gamma_k^i u_i = \sum_i \bar{\gamma}^i u_i \in \operatorname{ran} U$, which establishes the claim. ∎

Since $(x_{n_k})_k$ is bounded a.s., Assumption 8(ii) suggests that $(\xi_{n_k})_k$ is also bounded a.s. There exists, thus, $\bar{\xi}$ and a subsequence of $(n_k)_k$, denoted here also by $(n_k)_k$ to avoid notational congestion, s.t. $\xi_{n_k} \xrightarrow{\text{a.s.}}_k \bar{\xi}$. Further, via $\|w_{n_k+1}\| \leq \lambda\|\xi_{n_k+1} + w_{n_k+1}/\lambda\| + \lambda\|\xi_{n_k+1}\|$ and (34), $(w_{n_k})_k$ is also bounded a.s., and hence, so is $(\mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}))_k$. Consequently, passing again to a subsequence of $(n_k)_k$ if necessary, there exists $\bar{\mathsf{w}}$ s.t. $\mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}) \xrightarrow{\text{a.s.}}_k \bar{\mathsf{w}}$.

Recall now that $(x_n)_n$ is bounded within $\mathfrak{X}$ [*cf.* discussion after (24)]. Moreover, the application of $\mathbb{E}(\cdot)$ to (32), Assumption 7(i), and by arguments similar to those after (32), it can be shown that there exists $C_3 \in \mathbb{R}_{>0}$ s.t.

$$\|\nabla f_n(x_n)\|_{\mathfrak{X}} \leq C_3, \quad \forall n. \tag{37}$$

Notice by Assumption 7(ii) that $\forall n$, $\|\pi_n - \pi_{n+1}\|_{\mathfrak{X}}^2 = \|\pi_n - \pi + \pi - \pi_{n+1}\|_{\mathfrak{X}}^2 \leq 2\|\pi_n - \pi\|_{\mathfrak{X}}^2 + 2\|\pi_{n+1} - \pi\|_{\mathfrak{X}}^2 \leq 4C_\pi$. Further, $\|\pi_n\|_{\mathfrak{X}}^2 = \|\pi_n - \pi + \pi\|_{\mathfrak{X}}^2 \leq 2\|\pi_n - \pi\|_{\mathfrak{X}}^2 + 2\|\pi\|_{\mathfrak{X}}^2 \leq 2C_\pi + 2\|\pi\|_{\mathfrak{X}}^2$; thus, $(\pi_n)_n$ is bounded. By (33), the a.s. nonexpansivity of $(Q_n)_n$ suggests that $\exists C_4 \in \mathbb{R}_{>0}$ s.t.

$$\begin{aligned}&\|\xi_{n+1} + \frac{1}{\lambda} w_{n+1}\|_{\mathfrak{X}} \\ &\leq \frac{2\alpha - 1}{\lambda}\|(\mathrm{Id} - T_{n+1})x_{n+1}\|_{\mathfrak{X}} + \frac{1}{\lambda}\|x_n - x_{n+1}\|_{\mathfrak{X}} \\ &\quad + \frac{\alpha}{\lambda}\|(T_n - T_{n+1})x_n\|_{\mathfrak{X}} + \|\nabla f_n(x_n)\|_{\mathfrak{X}} \\ &\leq \frac{2\alpha - 1}{\lambda}\|x_{n+1}\|_{\mathfrak{X}} + \frac{2\alpha - 1}{\lambda}\|Q_{n+1}x_{n+1}\|_{\mathfrak{X}} \\ &\quad + \frac{2\alpha - 1}{\lambda}\|\pi_{n+1}\|_{\mathfrak{X}} + \frac{1}{\lambda}\|x_n\|_{\mathfrak{X}} + \frac{1}{\lambda}\|x_{n+1}\|_{\mathfrak{X}} + \frac{\alpha}{\lambda}\|Q_n x_n\|_{\mathfrak{X}} \\ &\quad + \frac{\alpha}{\lambda}\|\pi_n\|_{\mathfrak{X}} + \frac{\alpha}{\lambda}\|Q_{n+1}x_n\|_{\mathfrak{X}} + \frac{\alpha}{\lambda}\|\pi_{n+1}\|_{\mathfrak{X}} + \|\nabla f_n(x_n)\|_{\mathfrak{X}} \\ &\leq \frac{4\alpha - 1}{\lambda}\|x_{n+1}\|_{\mathfrak{X}} + \frac{2\alpha + 1}{\lambda}\|x_n\|_{\mathfrak{X}} + \frac{3\alpha - 1}{\lambda}\|\pi_{n+1}\|_{\mathfrak{X}} \\ &\quad + \frac{\alpha}{\lambda}\|\pi_n\|_{\mathfrak{X}} + \|\nabla f_n(x_n)\|_{\mathfrak{X}} \leq C_4.\end{aligned}$$

Due to Assumption 8(iii), which establishes the boundedness of $(\xi_{n_k})_k$, the previous discussion renders $(w_{n_k})_k$ bounded. By Jensen's inequality [41, §9.7(h)], $(\mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}))_k$ is also bounded in $\mathfrak{X}$, and hence uniformly integrable (UI) [41, §13.3(a)]. Since $\mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}) \xrightarrow{\text{a.s.}}_k \bar{\mathsf{w}}$, then, this convergence holds also in probability [41, App. A13.2(a)]. This and the UI argument imply that $\mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}) \xrightarrow{\mathfrak{X}}_k \bar{\mathsf{w}}$ [41, App. A13.2(f)]. Going back to (17), notice by Lemma 5 and $\mathbb{E}(\cdot) = \mathbb{E}[\mathbb{E}_{|\mathcal{F}_n}(\cdot)]$ [41, §9.7(a)] that $\forall u \in \ker U \cap \mathcal{X}$,

$$\begin{aligned}&\langle u \mid \mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k})\rangle_{\mathfrak{X}} \\ &= (1 - \alpha)\langle u \mid t_{n_k}\rangle_{\mathfrak{X}} + \langle u \mid U\,\mathbb{E}_{|\mathcal{F}_{n_k}}(v_{n_k})\rangle_{\mathfrak{X}} \\ &= (1 - \alpha)\langle u \mid t_{n_k}\rangle_{\mathfrak{X}} + \langle Uu \mid v_{n_k}\rangle_{\mathfrak{X}} \\ &= (1 - \alpha)\langle u \mid t_{n_k}\rangle_{\mathfrak{X}} = (1 - \alpha)\langle u \mid \mathbb{E}(t_{n_k})\rangle.\end{aligned} \tag{38}$$

It can be also seen via (17) that $(1 - \alpha)t_{n_k} = \mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}) - U\,\mathbb{E}_{|\mathcal{F}_{n_k}}(v_{n_k})$; hence, $(1 - \alpha)\,\mathbb{E}(t_{n_k}) = \mathbb{E}(w_{n_k}) - U\,\mathbb{E}(v_{n_k})$ and $(1 - \alpha)^2\|\mathbb{E}(t_{n_k})\|^2 \leq 2\|\mathbb{E}(w_{n_k})\|^2 + 2\|U\|^2\|\mathbb{E}(v_{n_k})\|^2 \leq 2\|\mathbb{E}(w_{n_k})\|^2 + 2\|\mathbb{E}(v_{n_k})\|^2$. Furthermore, Jensen's inequality [41, §9.7(h)] yields $(1 - \alpha)^2\|\mathbb{E}(t_{n_k})\|^2 \leq 2\,\mathbb{E}(\|w_{n_k}\|^2) + 2\,\mathbb{E}(\|v_{n_k}\|^2) = 2\|w_{n_k}\|_{\mathfrak{X}}^2 + 2\|v_{n_k}\|_{\mathfrak{X}}^2$, and consequently, the boundedness of $(w_{n_k})_k$ and $(v_{n_k})_k$ in $\mathfrak{X}$ results in that $(\mathbb{E}(t_{n_k}))_k$ is also bounded in $\mathcal{X}$. According now to Assumption 1(iv), there exists a subsequence of $(n_k)_k$, denoted here again by $(n_k)_k$ to avoid clutter in notations, s.t. $\lim_k \mathbb{E}(t_{n_k}) \in \operatorname{ran}(\mathrm{Id} - Q) = \operatorname{ran} U = (\ker U)^\perp$. Thus, via (38) and the continuity of the inner product [13, Lem. 2.41(iii)], $\langle u \mid \bar{\mathsf{w}}\rangle_{\mathfrak{X}} = \lim_k \langle u \mid \mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k})\rangle_{\mathfrak{X}} = (1 - \alpha)\langle u \mid \lim_k \mathbb{E}(t_{n_k})\rangle = 0$. Hence, $\bar{\mathsf{w}} \in (\ker U)^\perp = \overline{\operatorname{ran}}\, U = \operatorname{ran} U$, according to [13, Fact 2.18(iii)] and Lemma 7.

Fix arbitrarily an $\epsilon > 0$. By the convexity of $h_{n_k} + g$, $\forall z \in \mathcal{X}$ and a.s.,

$$\begin{aligned}&(h_{n_k} + g)(z) \\ &\geq \langle z - x_{n_k+1} \mid \xi_{n_k+1}\rangle + (h_{n_k} + g)(x_{n_k+1}) \\ &\geq \langle z - x_{n_k+1} \mid \xi_{n_k+1}\rangle + \langle x_{n_k+1} - x_{n_k} \mid \tau_{n_k}\rangle \\ &\quad + (h_{n_k} + g)(x_{n_k}) \\ &= \langle z - x_{n_k} \mid \xi_{n_k+1}\rangle + (h_{n_k} + g)(x_{n_k}) \\ &\quad + \langle x_{n_k} - x_{n_k+1} \mid \xi_{n_k+1}\rangle + \langle x_{n_k+1} - x_{n_k} \mid \tau_{n_k}\rangle \\ &\geq \langle z - x_{n_k} \mid \xi_{n_k+1}\rangle + (h_{n_k} + g)(x_{n_k}) \\ &\quad - \|x_{n_k} - x_{n_k+1}\|\|\xi_{n_k+1}\| - \|x_{n_k} - x_{n_k+1}\|\|\tau_{n_k}\|,\end{aligned} \tag{39}$$

where $\tau_{n_k} \in \partial(h_{n_k} + g)(x_{n_k})$ is chosen according to Assumption 8(i), $\forall k$. Moreover, by (25) and Assumption 4, there exists an integer $k_\# := k_\#(\omega)$ s.t. $\forall k \geq k_\#$, $\|x_{n_k} - x_{n_k+1}\|$

$\leq \epsilon/[3(C_\partial + C_5)]$, $-\epsilon/3 \leq \varepsilon^h_{n_k}(x_{n_k}) \leq \epsilon/3$ and $-\epsilon/3 \leq -\varepsilon^h_{n_k}(z) \leq \epsilon/3$. By (39),

$$(h_{n_k} + g)(z) \geq \langle z - x_{n_k} \mid \xi_{n_k+1}\rangle + (h_{n_k} + g)(x_{n_k}) - \frac{\epsilon/3}{C_\partial + C_5}(\|\tau_{n_k}\| + \|\xi_{n_k+1}\|) .$$

Notice that Assumption 8(ii) implies the existence of $C_5 := C_5(\omega) \in \mathbb{R}_{>0}$ s.t. $\|\xi_n\| \leq C_5$. Applying $\mathbb{E}_{|\mathcal{F}_{n_k}}(\cdot)$ to the previous inequality and adhering to Assumptions 4(i) and 8(i), as well as Lemma 5, $\forall z \in \mathcal{X}$, $\forall k \geq k_\#$ and a.s.,

$$\begin{aligned}
&\frac{\epsilon}{3} + (h+g)(z) \\
&\geq -\varepsilon^h_{n_k}(z) + (h+g)(z) = \mathbb{E}_{|\mathcal{F}_{n_k}}(h_{n_k} + g)(z) \\
&\geq \langle z - x_{n_k} \mid \mathbb{E}_{|\mathcal{F}_{n_k}}(\xi_{n_k+1})\rangle \\
&\quad + \mathbb{E}_{|\mathcal{F}_{n_k}}[(h_{n_k} + g)(x_{n_k})] - \frac{\epsilon}{3} \\
&= \langle z - x_{n_k} \mid \mathbb{E}_{|\mathcal{F}_{n_k}}(\xi_{n_k+1})\rangle + (h+g)(x_{n_k}) \\
&\quad - \varepsilon^h_{n_k}(x_{n_k}) - \frac{\epsilon}{3} \\
&\geq \langle z - x_{n_k} \mid \mathbb{E}_{|\mathcal{F}_{n_k}}(\xi_{n_k+1})\rangle + (h+g)(x_{n_k}) - \frac{\epsilon}{3} - \frac{\epsilon}{3} .
\end{aligned} \tag{40}$$

Since $(\xi_n)_n$ is bounded a.s., so is $(\xi_{n_k})_k$ and, consequently, so is $(\mathbb{E}_{|\mathcal{F}_{n_k}}(\xi_{n_k+1}))_k$. Hence, there exists $\bar{\xi}$ s.t. $\mathbb{E}_{|\mathcal{F}_{n_k}}(\xi_{n_k+1}) \xrightarrow{\text{a.s.}}_k \bar{\xi}$ (once again, passing to a subsequence of $(n_k)_k$ is avoided). Moreover, since $h+g$ is l.s.c., $\liminf_k (h+g)(x_{n_k}) \geq (h+g)(\bar{x})$ [13]. Hence, by the application of $\liminf_k$ to (40) and the continuity of the inner product [13, Lem. 2.41(iii)], $\epsilon/3 + (h+g)(z) \geq \langle z - \bar{x} \mid \bar{\xi}\rangle + (h+g)(\bar{x}) - 2\epsilon/3$, $\forall z \in \mathcal{X}$, and $(\bar{x}, \bar{\xi}) \in \operatorname{gph} \partial_\epsilon(h+g)$ a.s. Since $\epsilon > 0$ was chosen arbitrarily,

$$(\bar{x}, \bar{\xi}) \in \cap_{\epsilon \in \mathbb{R}_{>0}} \operatorname{gph} \partial_\epsilon(h+g) = \operatorname{gph} \partial(h+g) \quad \text{a.s.} \tag{41}$$

Similarly to the way that (28) follows from (27), it can be verified that (29) yields $\mathbb{E}_{|\mathcal{F}_{n_k}}[(T_{n_k+1} - \mathrm{Id})x_{n_k+1}] \xrightarrow{\text{a.s.}}_k 0$, (31) leads to $\mathbb{E}_{|\mathcal{F}_{n_k}}[(T_{n_k+1} - T_{n_k})x_{n_k}] \xrightarrow{\text{a.s.}}_k 0$, and (25) gives, via fact $\|Q^{(\alpha)}_n\| = \|\alpha Q_n + (1-\alpha)\,\mathrm{Id}\,\| \leq \alpha\|Q_n\| + (1-\alpha) \leq 1$, $\forall n$, a.s., $\mathbb{E}_{|\mathcal{F}_{n_k}}[Q^{(\alpha)}_{n_k+1}(x_{n_k+1} - x_{n_k})] \xrightarrow{\text{a.s.}}_k 0$. Recalling (28) and (36), the application of $\lim_k \mathbb{E}_{|\mathcal{F}_{n_k}}(\cdot)$ to (14) yields

$$\begin{aligned}
&-\lim_{k\to\infty} \mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k+1}) - \lambda \lim_{k\to\infty} \mathbb{E}_{|\mathcal{F}_{n_k}}[\nabla f_{n_k}(x_{n_k})] \\
&\quad - \lambda \lim_{k\to\infty} \mathbb{E}_{|\mathcal{F}_{n_k}}(\xi_{n_k+1}) \\
&\quad = (1-2\alpha) \lim_{k\to\infty} \mathbb{E}_{|\mathcal{F}_{n_k}}[(T_{n_k+1} - \mathrm{Id})x_{n_k+1}] \\
&\qquad + \lim_{k\to\infty} \mathbb{E}_{|\mathcal{F}_{n_k}}[Q^{(\alpha)}_{n_k+1}(x_{n_k+1} - x_{n_k})] \\
&\qquad + \alpha \lim_{k\to\infty} \mathbb{E}_{|\mathcal{F}_{n_k}}[(T_{n_k+1} - T_{n_k})x_{n_k}] \\
&\quad \Leftrightarrow \quad \nabla f(\bar{x}) + \bar{\xi} = -\frac{1}{\lambda}\bar{\mathsf{w}} \in \operatorname{ran} U \quad \text{a.s.}
\end{aligned} \tag{42}$$

Since (42) holds true for any cluster point in $\mathfrak{C}[(y_n)_n]$, Fact 2 and Lemma 6 suggest that *any* $\bar{x} \in \mathfrak{C}[(x_n)_n]$ belongs to $\mathcal{A}_*$, solving thus (P) a.s.

Appendix D
Proof of Theorem 2

Assumptions 8(ii) and 8(iii) are used in Appendix C to establish the boundedness of $(w_n)_n$ a.s. and in $\mathfrak{X}$. However, in the case where T is known exactly, the boundedness of $(w_n)_n$ follows from the boundedness of $(v_n)_n$, since by (15) and (16), $w_n = Uv_n$. Moreover, by Lemma 5 and the fact that $v_n \in \mathrm{m}\mathcal{F}_n$, $\mathbb{E}_{|\mathcal{F}_n}(w_n) = U\,\mathbb{E}_{|\mathcal{F}_n}(v_n) = Uv_n$. Thus, $\bar{\mathsf{w}} \xleftarrow{\text{a.s.}}_k \mathbb{E}_{|\mathcal{F}_{n_k}}(w_{n_k}) = Uv_{n_k} \xrightarrow{\text{a.s.}}_k U\bar{v}$, and (42) becomes $\nabla f(\bar{x}) + \bar{\xi} = -(1/\lambda)U\bar{v}$ a.s. Hence, according to Fact 2, the arbitrarily chosen cluster point $\bar{y} = (\bar{x}, \bar{v}) \in \Upsilon^{(\lambda)}_*$. This result together with the stochastic Fejér monotonicity of $(y_n)_n$ w.r.t. $\Upsilon^{(\lambda)}_*$ [*cf.* (22)] suggest that $\mathfrak{C}[(y_n)_n]$ is a singleton [10, Prop. 2.3(iv)], that $\mathfrak{C}[(x_n)_n]$ is also a singleton by virtue of Lemma 6, and that $(x_n)_n$ converges a.s. to a solution of (P).

Appendix E
Proof of Corollary 1

According to Lemma 2, Assumption 1 holds true. Since $(f, f_n) := (0,0)$ and $(h, h_n) := (0,0)$ in (HLS), Assumptions 3, 4, and 7(i) hold trivially true. Moreover, Lemma 4 and $(h, h_n) := (0,0)$ suggest that Assumption 8 holds true. In the context of Theorem 1, $L_{\nabla f}$ and λ can take any values in $\mathbb{R}_{>0}$. The claim of Corollary 1 follows now from Theorem 1.

Appendix F
Proof of Corollary 2

Since $(f, f_n) := (0,0)$ in (CRegLS) and $T_n := P_{\mathcal{A}} =: T$ in Algorithm 3, Assumptions 3, 4(ii) and 7(i) hold trivially true. Moreover, according to Lemma 4, Assumption 8(i) holds also true. The claim of Corollary 2 follows now from Theorem 2.

Appendix G
Proof of Lemma 2

The proof that Assumption 1(ii) holds true follows exactly the steps of the proof of [30, (70a) and (70d)], in the case where $\delta := 1$ and $\varphi_\delta(\mathbf{x}) = \varphi_1(\mathbf{x}) := [1/(2n)]\sum_{\nu=1}^n (\mathbf{a}_\nu^\intercal \mathbf{x} - b_\nu)^2$, $\forall \mathbf{x} \in \mathcal{X}$, in [30, (73)]. Furthermore, by virtue of Assumption 2, of $(\varpi_n := \varpi,\ \forall n)$, and the continuity of the matrix-inversion operation, mappings (5) satisfy Assumption 1(iii). According to Fact 1, the normal equations suggest that for any $T \in \mathfrak{T}_{\mathcal{A}}$, $\{\mathbf{x} | \mathbf{R}\mathbf{x} = \mathbf{r}\} = \ker(\mathrm{Id} - Q) + \boldsymbol{\theta}_* \Rightarrow \ker(\mathrm{Id} - Q) = \{\mathbf{x} - \boldsymbol{\theta}_* | \mathbf{R}\mathbf{x} = \mathbf{r} = \mathbf{R}\boldsymbol{\theta}_*\} = \{\mathbf{x}' | \mathbf{R}\mathbf{x}' = \mathbf{0}\} = \ker \mathbf{R} = \{\mathbf{0}\}$, due to the non-singularity of $\mathbf{R}$. However, $\operatorname{ran}(\mathrm{Id} - Q) = [\ker(\mathrm{Id} - Q)]^\perp = \{\mathbf{0}\}^\perp = \mathcal{X}$. Hence, if $(\mathbb{E}(\mathbf{t}_n))_n$ is bounded, then *any* of its cluster points belongs trivially to $\operatorname{ran}(\mathrm{Id} - Q) = \mathcal{X}$.

Appendix H
Proof of Lemma 3

Notice first that due to the IID assumption, $\forall \nu \in \{1, \ldots, n\}$, $\mathbb{E}_{|\sigma(\mathbf{R}_n)}(\mathbf{a}_\nu \mathbf{a}_\nu^\intercal) = \mathbf{R}_n$ [41, §9.11]. Thus, by applying $\mathbb{E}_{|\sigma(\mathbf{R}_n)}(\cdot)$ to $\mathbf{R}_\nu = (n/\nu)\mathbf{R}_n - (1/\nu)\sum_{i=\nu+1}^n \mathbf{a}_i \mathbf{a}_i^\intercal$, which can be straightforwardly derived from (4), $\mathbb{E}_{|\sigma(\mathbf{R}_n)}(\mathbf{R}_\nu) = \mathbf{R}_n$ can be established. Moreover, due to the conditional-independence hypothesis $\mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_\nu) = \mathbb{E}_{|\mathcal{F}_n}\mathbb{E}_{|\mathcal{F}_n,\sigma(\mathbf{R}_n)}[(1/\nu)\sum_{i=1}^{\nu}\mathbf{a}_i\mathbf{a}_i^\intercal] = (1/\nu)\sum_{i=1}^{\nu}\mathbb{E}_{|\mathcal{F}_n}\mathbb{E}_{|\mathcal{F}_n,\sigma(\mathbf{R}_n)}(\mathbf{a}_i\mathbf{a}_i^\intercal) = (1/\nu)\sum_{i=1}^{\nu}\mathbb{E}_{|\mathcal{F}_n}\mathbb{E}_{|\sigma(\mathbf{R}_n)}(\mathbf{a}_i\mathbf{a}_i^\intercal) = \mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_n) = \mathbf{R}$. Furthermore, by the linear-regression model of Section I-B and the assumptions on the noise process $(\eta_n)_n$, $\mathbb{E}_{|\sigma(\mathbf{R}_n)}(b_\nu\mathbf{a}_\nu) = \mathbb{E}_{|\sigma(\mathbf{R}_n)}(\mathbf{a}_\nu\mathbf{a}_\nu^\intercal\boldsymbol{\theta}_*) + \mathbb{E}_{|\sigma(\mathbf{R}_n)}(\eta_n\mathbf{a}_\nu) = \mathbb{E}_{|\sigma(\mathbf{R}_n)}(\mathbf{a}_\nu\mathbf{a}_\nu^\intercal)\boldsymbol{\theta}_* + \mathbb{E}(\eta_n)\,\mathbb{E}_{|\sigma(\mathbf{R}_n)}(\mathbf{a}_\nu) = \mathbf{R}_n\boldsymbol{\theta}_*$. Furthermore, in a way similar to that of the $\mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_\nu)$ case, $\mathbb{E}_{|\mathcal{F}_n}(\mathbf{r}_\nu) = \mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_n\boldsymbol{\theta}_*) = \mathbf{R}\boldsymbol{\theta}_* = \mathbf{r}$.

i) Applying $\mathbb{E}_{|\mathcal{F}_n}(\cdot)$ to $(T - T_n)\mathbf{x}_n = -(\mu/\varpi)\boldsymbol{\mathcal{E}}_n^R\mathbf{x}_n + (\mu/\varpi)\varepsilon_n^r$ yields (8a). In a similar way, (8b) can be established. Moreover,

$$\mathbf{t}_n = \mathbb{E}_{|\mathcal{F}_n}\left[\sum_{\nu=1}^{n}(T - T_\nu)x_\nu\right] = -\frac{\mu}{\varpi}\sum_{\nu=1}^{n}\left[\mathbf{R} - \mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_\nu)\right]\mathbf{x}_\nu + \frac{\mu}{\varpi}\sum_{\nu=1}^{n}\left[\mathbf{r} - \mathbb{E}_{|\mathcal{F}_n}(\mathbf{r}_\nu)\right] = \mathbf{0}\,.$$

Furthermore, $\mathbf{R}_{n-1} - \mathbf{R}_n = \mathbf{R}_n/(n-1) - \mathbf{a}_n\mathbf{a}_n^\intercal/(n-1) = \mathbf{R}/(n-1) - \mathbf{a}_n\mathbf{a}_n^\intercal/(n-1) - \boldsymbol{\mathcal{E}}_n^R/(n-1)$ and $\mathbf{r}_{n-1} - \mathbf{r}_n = \mathbf{r}/(n-1) - b_n\mathbf{a}_n/(n-1) - \varepsilon_n^r/(n-1)$. Hence, due to $T_n - T_{n-1} = (\mu/\varpi)(\mathbf{R}_{n-1} - \mathbf{R}_n) + (\mu/\varpi)(\mathbf{r}_n - \mathbf{r}_{n-1})$,

$$\begin{aligned}&\mathbb{E}_{|\mathcal{F}_n}\left[(T_n - T_{n-1})\mathbf{x}_{n-1}\right]\\ &= \frac{\mu}{\varpi(n-1)}\mathbf{R}\mathbf{x}_{n-1} - \frac{\mu}{\varpi(n-1)}\,\mathbb{E}_{|\mathcal{F}_n}(\mathbf{a}_n\mathbf{a}_n^\intercal)\mathbf{x}_{n-1}\\ &\quad - \frac{\mu}{\varpi(n-1)}\,\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\mathcal{E}}_n^R)\mathbf{x}_{n-1} - \frac{\mu}{\varpi(n-1)}\mathbf{r}\\ &\quad + \frac{\mu}{\varpi(n-1)}\,\mathbb{E}_{|\mathcal{F}_n}(b_n\mathbf{a}_n) + \frac{\mu}{\varpi(n-1)}\,\mathbb{E}_{|\mathcal{F}_n}(\varepsilon_n^r) = \mathbf{0}\,.\end{aligned}$$

ii) By the assumption that noise η_n is independent of $\mathbf{a}_n$, and thus independent also of $\mathcal{F}_n$, as well as by $\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\mathcal{E}}_n^R) = \mathbf{0}$, it can be verified that $\mathbb{E}_{|\mathcal{F}_n}(b_n^2) = \mathbb{E}(b_n^2)$ a.s. Now, by (CRegLS) and Lemma 5, $\forall n$, $\forall \mathbf{x} \in \mathcal{X}$ and a.s.,

$$\begin{aligned}\varepsilon_n^h(\mathbf{x}) &= \mathbb{E}_{|\mathcal{F}_n}\left[(h - h_n)(\mathbf{x})\right]\\ &= \frac{1}{2}\,\mathbb{E}_{|\mathcal{F}_n}(\mathbf{x}^\intercal\boldsymbol{\mathcal{E}}_n^R\mathbf{x}) - \mathbb{E}_{|\mathcal{F}_n}(\mathbf{x}^\intercal\varepsilon_n^r)\\ &= \frac{1}{2}\mathbf{x}^\intercal\,\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\mathcal{E}}_n^R)\mathbf{x} - \mathbf{x}^\intercal\,\mathbb{E}_{|\mathcal{F}_n}(\varepsilon_n^r) = 0\,.\end{aligned}$$

The claim $\varepsilon_n^h(\mathbf{x}_n) = 0$, a.s., can be similarly verified.

iii) Notice that $\mathbb{E}_{|\mathcal{F}_n}[(\nabla f - \nabla f_n)\mathbf{x}_n] = \mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\mathcal{E}}_n^R\mathbf{x}_n) - \mathbb{E}_{|\mathcal{F}_n}(\varepsilon_n^r) = \mathbb{E}(\boldsymbol{\mathcal{E}}_n^R)\mathbf{x}_n - \mathbb{E}(\varepsilon_n^r) = \mathbf{0}$. Moreover, by following similar steps as those in the last part of Appendix H(i), $\mathbb{E}_{|\mathcal{F}_n}[(\nabla f_n - \nabla f_{n-1})\mathbf{x}_{n-1}] = \mathbf{0}$ can be also established.

iv) Here, only the case of $(h, h_n) = (l, l_n)$ is considered, since the case of $(h, h_n) = (0, 0)$ can be trivially deduced. To this end, there exists $\boldsymbol{\chi}_n \in \partial g(\mathbf{x}_n)$ s.t. $\boldsymbol{\xi}_n = \mathbf{R}_{n-1}\mathbf{x}_n - \mathbf{r}_{n-1} + \boldsymbol{\chi}_n$. Recall here that $\boldsymbol{\chi}_n \in \partial g(\mathbf{x}_n) = \partial\|\cdot\|_1(\mathbf{x}_n)$ iff

$$[\boldsymbol{\chi}_n]_d \in \begin{cases}\{\mathrm{sgn}([\mathbf{x}_n]_d)\}\,, & \text{if } [\mathbf{x}_n]_d \neq 0,\\ [-1, 1], & \text{if } [\mathbf{x}_n]_d = 0,\end{cases} \tag{43}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign of a real-valued number. Hence, $\boldsymbol{\chi}_n \in \mathrm{m}\mathcal{F}_n$. By arguments similar to those in the first part of Appendix H(i), $\mathbb{E}_{|\mathcal{F}_n}(\boldsymbol{\xi}_n) = \mathbb{E}_{|\mathcal{F}_n}(\mathbf{R}_{n-1})\mathbf{x}_n - \mathbb{E}_{|\mathcal{F}_n}(\mathbf{r}_{n-1}) + \boldsymbol{\chi}_n = \mathbf{R}\mathbf{x}_n - \mathbf{r} + \boldsymbol{\chi}_n \in \partial(h + g)(\mathbf{x}_n)$. In other words, $\epsilon_n = 0$, a.s., in (8g).

Appendix I
Proof of Lemma 4

i) According to (43), for any $(\mathbf{z}, \boldsymbol{\tau})$ s.t. $\boldsymbol{\tau} \in \partial\|\|_1(\mathbf{z})$, $|[\boldsymbol{\tau}]_d| \leq 1$, $\forall d$. Hence, $\|\boldsymbol{\tau}\| \leq \sqrt{D}$ and $\mathbb{E}_{|\mathcal{F}_n}(\|\boldsymbol{\tau}\|) \leq \sqrt{D}$ a.s. This bound renders Assumptions 8(i), 8(ii) and 8(iii) true in the case where $h_n := 0$ and $h := 0$ in (CRegLS).

The following discussion deals with the case of $(h, h_n) := (l, l_n)$ in (CRegLS), where, according to (7), $\nabla h_n(\mathbf{x}) = \mathbf{R}_n\mathbf{x} - \mathbf{r}_n$, $\forall \mathbf{x} \in \mathcal{X}$. By Assumption 2, given $\epsilon := \epsilon(\omega) \in \mathbb{R}_{>0}$, there exists $n_\# := n_\#(\omega) \in \mathbb{Z}_{\geq 0}$ s.t. $\|\mathbf{R}_n\| \leq \|\mathbf{R}\| + \epsilon$ and $\|\mathbf{r}_n\| \leq \|\mathbf{r}\| + \epsilon$, $\forall n \geq n_\#$. Define, then, $\varpi := \varpi(\omega) := \max\{\{\|\mathbf{R}_n\| | 0 \leq n < n_\# - 1\}, \|\mathbf{R}\| + \epsilon\}$ and $\varpi' := \varpi'(\omega) := \max\{\{\|\mathbf{r}_n\| | 0 \leq n < n_\# - 1\}, \|\mathbf{r}\| + \epsilon\}$. According to the hypothesis, there exists $C_z := C_z(\omega) \in \mathbb{R}_{>0}$ s.t. $\|\mathbf{z}_n\| \leq C_z$, $\forall n$. Thus, $\forall \boldsymbol{\delta}_n \in \partial\|\cdot\|_1(\mathbf{z}_n)$, $\boldsymbol{\tau}_n := \mathbf{R}_n\mathbf{z}_n - \mathbf{r}_n + \rho\boldsymbol{\delta}_n \in \partial(h_n + g)(\mathbf{z}_n)$ and $\|\boldsymbol{\tau}_n\| \leq \|\mathbf{R}_n\|\|\mathbf{z}_n\| + \|\mathbf{r}_n\| + \rho\|\boldsymbol{\delta}_n\| \leq \varpi C_z + \varpi' + \rho\sqrt{D} =: C_\partial$, $\forall n$; thus, Assumption 8(i) holds true. By using $\mathbf{x}_n$ in the place of $\mathbf{z}_n$ in the previous discussion, it can be verified that Assumption 8(ii) holds also true.

ii) Observe that $\|\boldsymbol{\xi}_n\|^2 = \|\mathbf{R}_n\mathbf{x}_n - \mathbf{r}_n + \rho\boldsymbol{\delta}_n\|^2 \leq 2\|\mathbf{R}_n\mathbf{x}_n\|^2 + 2\|\mathbf{r}_n + \rho\boldsymbol{\delta}_n\|^2 \leq 2\|\mathbf{R}_n\|^2\|\mathbf{x}_n\|^2 + 4\|\mathbf{r}_n\|^2 + 4\rho^2\|\boldsymbol{\delta}_n\|^2 \leq 2\varpi^2\|\mathbf{x}_n\|^2 + 4\varpi'^2 + 4\rho^2 D$, a.s. Hence, $\|\boldsymbol{\xi}_n\|_{\mathfrak{X}}^2 = \mathbb{E}(\|\boldsymbol{\xi}_n\|^2) \leq 2\varpi^2\,\mathbb{E}(\|\mathbf{x}_n\|^2) + 4\varpi'^2 + 4\rho^2 D = 2\varpi^2\|\mathbf{x}_n\|_{\mathfrak{X}}^2 + 4\varpi'^2 + 4\rho^2 D$, and consequently, $(\|\boldsymbol{\xi}_n\|_{\mathfrak{X}}^2)_n$ is bounded.

References

[1] K. Slavakis, “Stochastic composite convex minimization with affine constraints,” in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, USA, Oct. 2018, pp. 1871–1875.

[2] A. H. Sayed, *Adaptive Filters*. Hoboken, NJ, USA, Wiley, 2008.

[3] N. Kalouptsidis and S. Theodoridis, Eds., *Adaptive System Identification and Signal Processing Algorithms*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[4] M. Pereyra *et al.*, “A survey of stochastic simulation and optimization methods in signal processing,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 224–241, Mar. 2016.

[5] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York, NY, USA: Springer, 2003.

[6] S. Shalev-Shwartz, “Online learning and online convex optimization,” *Found. Trends Mach. Learn.*, vol. 4, no. 2, pp. 107–194, 2012.

[7] K. Slavakis, S.-J. Kim, G. Mateos, and G. B. Giannakis, “Stochastic approximation vis-à-vis online learning for big-data analytics,” *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 124–129, Nov. 2014.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
[10] P. L. Combettes and J.-C. Pesquet, "Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping," *SIAM J. Optim.*, vol. 25, no. 2, pp. 1221–1248, 2015.
[11] L. Rosasco, S. Villa, and B. C. Vũ, "A stochastic inertial forwardbackward splitting algorithm for multivariate monotone inclusions," *Optimization*, vol. 65, no. 6, pp. 1293–1314, 2016.
[12] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM J. Optim.*, vol. 24, no. 4, pp. 2057–2075, 2014.
[13] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York, NY, USA: Springer, 2011.
[14] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, vol. 26, pp. 315–323.
[15] P. L. Combettes and J.-C. Pesquet, "Stochastic approximations and perturbations in forward-backward splitting for monotone operators," *Pure Appl. Functional Anal.*, vol. 1, no. 1, pp. 13–37, 2016.
[16] P. Bianchi, W. Hachem, and A. Salim, "A constant step forward-backward algorithm involving random maximal monotone operators," *J. Convex Anal.*, vol. 26, no. 2, 2019.
[17] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications," *SIAM J. Optim.*, vol. 28, no. 4, pp. 2783–2808, 2018.
[18] S. Zheng and J. T. Kwok, "Fast-and-light stochastic ADMM," in *Proc. Int. Joint Conf. Artif. Intell.*, Las Vegas, NV, USA, Jul. 2016, pp. 2407–2413.
[19] N. Flammarion and F. Bach, "Stochastic composite least-squares regression with convergence rate $\mathcal{O}(1/n)$," in *Proc. Conf. Learn. Theory*, in Proc. Machine Learning Research, Jul. 2017, vol. 65, pp. 831–875.
[20] Y. E. Nesterov, "Primal-dual subgradient methods for convex problems," *Math. Program., Ser. B*, vol. 120, no. 1, pp. 221–259, 2009.
[21] G. Lan, "An optimal method for stochastic composite optimization," *Math. Program., Ser. A*, vol. 133, pp. 365–397, 2012.
[22] Y. E. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$," *Doklady AN SSSR*, vol. 269, pp. 543–547, 1983.
[23] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online adaptive estimation of sparse signals: Where RLS meets the ℓ_1-norm," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3436–3447, Jul. 2010.
[24] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, Aug. 2010.
[25] E. M. Eksioglu and A. K. Tanc, "RLS algorithm with convex regularization," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 470–473, Aug. 2011.
[26] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A variational Bayes framework for sparse adaptive estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4723–4736, Sep. 2014.
[27] B. Dumitrescu, A. Onose, P. Helin, and I. Tăbuş, "Greedy sparse RLS," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2194–2207, May 2012.
[28] E. Chouzenoux and J.-C. Pesquet, "A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4770–4783, Sep. 2017.
[29] J. Benesty, C. Paleologu, and S. Ciochină, "Proportionate adaptive filters from a basis-pursuit perspective," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 985–988, Dec. 2010.
[30] K. Slavakis and I. Yamada, "Fejér-monotone hybrid steepest descent method for affinely constrained and composite convex minimization tasks," *Optimization*, vol. 67, no. 11, pp. 1963–2001, 2018.
[31] I. Yamada, "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed-point sets of nonexpansive mappings," in *Inherently Parallel Algorithms for Feasibility and Optimization and Their Applications*, D. Butnariu, Y. Cencor, and S. Reich, Eds., New York, NY, USA: Elsevier, 2001, pp. 473–504.
[32] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.
[33] L. Rosasco, S. Villa, and B. C. Vũ, "Convergence of stochastic proximal gradient algorithm," 2014, arXiv:1403.5074.
[34] E. Kreyszig, *Introductory Functional Analysis and Applications* (Wiley Classics Library ed.). New York, NY, USA: Wiley, 1989.
[35] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.
[36] M. Gu and S. C. Eisenstat, "A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem," *SIAM J. Matrix Anal. Appl.*, vol. 15, no. 4, pp. 1266–1276, 1994.
[37] E. Livne and A. Brandt, "N roots of the secular equation in $\mathcal{O}(N)$ operations," *SIAM J. Matrix Anal. Appl.*, vol. 24, no. 2, pp. 439–453, 2002.
[38] B. Porat, *Digital Processing of Random Signals: Theory and Methods*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.
[39] D. W. K. Andrews, "Laws of large numbers for dependent non-identically distributed random variables," *Econometric Theory*, vol. 4, no. 3, pp. 458–467, 1988.
[40] K. Petersen, *Ergodic Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1983.
[41] D. Williams, *Probability With Martingales*. New York, NY, USA: Cambridge Univ. Press, 1991.
[42] P. L. Combettes, "Quasi-Fejérian analysis of some optimization algorithms," in *Inherently Parallel Algorithms for Feasibility and Optimization and Their Applications*, D. Butnariu, Y. Cencor, and S. Reich, Eds., Amsterdam, The Netherlands: Elsevier, 2001, pp. 115–152.

Konstantinos Slavakis (M'08–SM'12) received the M.Sc. and Ph.D. degrees in electrical and electronic engineering from Tokyo Institute of Technology, Tokyo, Japan. He joined University at Buffalo (SUNY) as an Associate Professor in 2015. His research interests are in the areas of signal processing, machine learning, and data analytics. He has been an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING (2009–2013), *EURASIP Journal on Advances in Signal Processing* (2016–2018), and *EURASIP Signal Processing* (2016–2018). He served IEEE Transactions on Signal Processing also as a Senior Associate Editor (2010–2015), and he is currently a Senior Associate Editor for *EURASIP Signal Processing*. He has also been a member of the IEEE Signal Processing Theory and Methods (SPTM) (2012–2017) and the EURASIP Signal and Data Analytics for Machine Learning (SiG-DML) (2015–2018) technical committees. He was the recipient of the 2014 IEEE Signal Processing Magazine best-paper award.