STOCHASTIC COMPOSITE CONVEX MINIMIZATION WITH AFFINE CONSTRAINTS

Konstantinos Slavakis

University at Buffalo, The State University of New York, Dept. of Electrical Engineering (kslavaki@buffalo.edu)

ABSTRACT

This paper presents the basic ingredients of a novel method, the stochastic Fejér-monotone hybrid steepest descent method (S-FM-HSDM), designed 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 pointwise to solutions of the noiseless minimization task. S-FM-HSDM enjoys desirable attributes of state-of-the-art stochastic-approximation 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 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 least-squares loss. Numerical tests on a synthetic ℓ_1-norm regularized 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.

1. INTRODUCTION

1.1. 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; $n \in \mathbb{Z}_{\geq 0} := \{0, 1, 2, \ldots\}$) of the generally unknown convex functions f, h and the affine set $\mathcal{A}$, respectively, solve

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

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 pointwise 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) [16, 23]; in particular, in linear equalization, channel estimation, beamforming, tracking of fading channels, line and acoustic echo cancellation and active noise control [23]. Special cases of (P) appear also in stochastic approximation (SA) [18] and online learning [24],[1] as in training artificial neural networks, learning optimal strategies in Markov decision processes, recursive games, sequential-decision tasks in economics [18], online classification and multi-armed bandit problems [24].

1.2. Case study: Sparsity-aware least squares

To highlight the versatility of (P), it is instructive to examine specific instances of (P). To this end, let $\mathcal{X}$ be the Euclidean $\mathbb{R}^D$. Boldfaced symbols indicate that $\mathcal{X} = \mathbb{R}^D$; in particular, lowercase boldfaced symbols denote vectors in $\mathbb{R}^D$. With respect to a probability space $(\Omega, \Sigma, \mathbb{P})$, 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 $\boldsymbol{\theta}_* \in \{\mathbf{x} \in \mathcal{X} \,|\, \mathbf{R}\mathbf{x} = \mathbf{r}\}$ [23, (3.9)]. This section deals with the system-identification problem of estimating the sparse $\boldsymbol{\theta}_*$ without knowing $(\mathbf{R}, \mathbf{r})$ and 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) [14, (3.52)], designed to solve sparse system-identification problems, the first instance of (P) is the convexly regularized least squares (LS): $\forall n \in \mathbb{Z}_{>0}$,

$$\begin{aligned} &\min_{\mathbf{x} \in \mathbb{R}^D} \overbrace{\tfrac{1}{2}\mathbf{x}^\intercal \mathbf{R} \mathbf{x} - \mathbf{r}^\intercal \mathbf{x} + \tfrac{1}{2}\,\mathbb{E}(b_n^2)}^{l(\mathbf{x})} + \overbrace{\rho \|\mathbf{x}\|_1}^{g(\mathbf{x})} \\ &= \min_{\mathbf{x} \in \mathbb{R}^D} \underbrace{\mathbb{E}\Big[\, \tfrac{1}{2n} \sum\nolimits_{\nu=1}^{n} \left(\mathbf{a}_\nu^\intercal \mathbf{x} - b_\nu\right)^2 \Big]}_{l_n(\mathbf{x})} + \rho \|\mathbf{x}\|_1 \,, \end{aligned} \tag{CRegLS}$$

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

$$\begin{aligned} &\min_{\mathbf{x} \in \mathbb{R}^D} \left[\, \|\mathbf{x}\|_1 =: g(\mathbf{x}) \,\right] \\ &\text{s.to } \mathbf{x} \in \underbrace{\arg\min_{\mathbf{x}' \in \mathbb{R}^D} \mathbb{E}\Big[\sum\nolimits_{\nu=1}^{n} \left(\mathbf{a}_\nu^\intercal \mathbf{x}' - b_\nu\right)^2\Big]}_{\mathcal{A}} \,, \end{aligned} \tag{HLS}$$

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 [6]. Since

This work was supported by the NSF grants 1525194 and 1718796.

[1] [25] provides an outline of the strong ties and distinct differences between SA and online learning.

978-1-5386-9218-9/18/$31.00 ©2018 IEEE

$(\mathbf{R}, \mathbf{r})$ are generally unknown, $\mathcal{A}$ is also unknown to the user. Still, the goal is to solve (HLS). If $f := h := 0 =: h_n =: f_n$, and $\mathcal{A}_n$ is defined as an estimate of $\mathcal{A}$, a.s., 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 mappings (*cf.* Section 2). This new viewpoint pays off in the computationally efficient HRLSa (*cf.* Algorithm 2), which solves (HLS) despite the uncertainty in the estimates $(\mathcal{A}_n)_n$ and under certain conditions, while scoring the lowest estimation error across a variety of numerical-test scenarios and several state-of-the-art schemes (*cf.* Section 3).

1.3. Prior art

Online-learning and SA algorithms have their origins usually in deterministic optimization schemes. For example, the online-learning scheme [28] draws inspiration from the forward-backward (a.k.a. proximal-gradient) algorithm [3, §27.3] and incorporates variance-reduction arguments [15] 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: $f := (1/m)\sum_{i=1}^{m} \mathscr{f}_i$, where $\{\mathscr{f}_i\}_{i=1}^{m}$ are convex and smooth, m is massive, $h := 0$ 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 $\{\mathscr{f}_i\}_{i=1}^{m}$, 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 [5, 7]. An SA extension of the Douglas-Rachford algorithm [3, §25.2, §27.2] is reported in [7]. Study [29] extends the celebrated alternating direction method of multipliers (ADMM) [11, 12] to the online-learning setting, and blends it with variance-reduction arguments to solve a problem similar to that of [28], but with a non-trivial, yet deterministic affine constraint $\mathcal{A} \subsetneq \mathcal{X}$. Further, [10] explores the dual-averaging scheme of [21] in the SA context offering linear-convergence guarantees for a quadratic f and $\mathcal{A} := \mathcal{X}$ in (P), while $\mathcal{X}$ is a closed convex set with non-empty interior. Moreover, the SA schemes [19, 22] are motivated by the deterministic acceleration method of [20]; in particular, [19] uses specific step sizes (*cf.* [19, (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 1.2, the state-of-the-art AF schemes [1, 2, 9] 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 [27], and a greedy RLS approach based on the orthogonal-matching-pursuit algorithm is reported in [8]. Basis pursuit [6] is used in [4] 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.

2. STOCHASTIC FEJÉR-MONOTONE HYBRID STEEPEST DESCENT METHOD AND OFFSPRINGS

Algorithm 1: S-FM-HSDM

```
Stochastic oracle's input : $(\nabla f_n, h_n, \mathcal{A}_n)_n$, $L_{\nabla f}$, $g$.
User's input              : $\alpha$, $\lambda$, $(T_n)_n$.
Output                    : Sequence $(x_n)_n$.
1 Initialization
2     Set $\alpha \in [0.5, 1)$ and $\lambda \in (0, 2(1-\alpha)/L_{\nabla f})$.
        Arbitrarily fix $x_0$.
3     $x_{1/2} := T_0^{(\alpha)} x_0 - \lambda \nabla f_0(x_0)$.
4     $x_1 := \mathrm{Prox}_{\lambda(h_0+g)}(x_{1/2})$.
5 for $n = 0$ to $+\infty$ do
6     $x_{n+3/2} := x_{n+1/2} - [T_n^{(\alpha)} x_n - \lambda \nabla f_n(x_n)] +$
        $[T_{n+1} x_{n+1} - \lambda \nabla f_{n+1}(x_{n+1})]$.
7     $x_{n+2} := \mathrm{Prox}_{\lambda(h_{n+1}+g)}(x_{n+3/2})$.
```

With the stochastic oracle providing the estimates $(\mathcal{A}_n)_n$ of $\mathcal{A}$, the user constructs mappings $(T_n)_n$ that serve as estimates of the unknown T. Since $\mathcal{A}$ is affine, T is considered to be affine, *i.e.*, for a linear mapping $Q : \mathcal{X} \to \mathcal{X}$ and a $\pi \in \mathcal{X}$, $Tx = Qx + \pi$, $\forall x \in \mathcal{X}$; in short, $T = Q + \pi$. For an affine T, $\operatorname{Fix} T$ is an affine set. For the linear Q, $\|Q\| := \sup_{\{x \mid \|x\| \leq 1\}} \langle x \mid Qx \rangle$. Mapping $Q : \mathcal{X} \to \mathcal{X}$ is called positive if it is linear, bounded, self-adjoint and $\langle x \mid Qx \rangle \geq 0$, $\forall x \in \mathcal{X}$ [17, §9.3]. Given the affine set $\mathcal{A}$, define now the following non-empty family of mappings [26, Prop. 2.11]:

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

Since $\|Q\| \leq 1$, every mapping $T \in \mathfrak{T}_{\mathcal{A}}$ turns out to be nonexpansive [3]: $\forall (x, x') \in \mathcal{X}^2$, $\|Tx - Tx'\| = \|Qx - Qx'\| = \|Q(x - x')\| \leq \|Q\|\|x - x'\| \leq \|x - x'\|$. It is also worth noticing here that $\mathfrak{T}_{\mathcal{A}}$ is closed under any convex combination and certain compositions of its members [26, Prop. 2.10]. There are several choices for mappings $T \in \mathfrak{T}_{\mathcal{A}}$ that suit the context of Section 1.2, *e.g.*, [26, (70), and (72)]. A well-known member of $\mathfrak{T}_{\mathcal{A}}$ is the projection mapping $P_{\mathcal{A}}$ onto $\mathcal{A}$ [26, Prop. 2.11]. Nevertheless, this study revolves around less obvious cases. Examples are provided next.

Lemma 1. For $\mathcal{A} := \{\mathbf{x} \mid \mathbf{R}\mathbf{x} = \mathbf{r}\}$ in Section 1.2, mappings

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

belong to $\mathfrak{T}_{\mathcal{A}}$, where $\mathbf{I}$ is the identity matrix and the spectral norm $\|\mathbf{R}\|$ is equal to the maximum eigenvalue of $\mathbf{R}$. □

Instead of the unknown $\mathbf{R}$ and $\mathbf{r}$, the following classical running-average estimates are used [23]; $\forall n \in \mathbb{Z}_{>0}$,

$$\mathbf{R}_n := \tfrac{1}{n} \sum\nolimits_{\nu=1}^{n} \mathbf{a}_\nu \mathbf{a}_\nu^\intercal\,, \quad \mathbf{r}_n := \tfrac{1}{n} \sum\nolimits_{\nu=1}^{n} b_\nu \mathbf{a}_\nu\,. \tag{3}$$

Choices for the estimates T_n of the mapping T in (2) are

$$T_n := \begin{cases} (\mathbf{I} - \frac{\mu}{\varpi_n}\mathbf{R}_n) + \frac{\mu}{\varpi_n}\mathbf{r}_n\,, \qquad \varpi_n \geq \|\mathbf{R}_n\|\,, & \\ \qquad\qquad\qquad\qquad\qquad\quad \mu \in (0, 1]\,, & \text{(4a)} \\ (\mathbf{I} + \kappa\mathbf{R}_n)^{-1} + \kappa(\mathbf{I} + \kappa\mathbf{R}_n)^{-1}\mathbf{r}_n\,,\ \kappa \in \mathbb{R}_{>0}\,. & \text{(4b)} \end{cases}$$

S-FM-HSDM is presented in Algorithm 1. The averaged mapping $T_n^{(\alpha)}$, given parameter $\alpha \in (0, 1)$, is defined by $T_n^{(\alpha)} := \alpha T_n + (1 - \alpha)\,\mathrm{Id}$, a.s., $\forall n$. Moreover, Prox in lines 4 and 7 of Algorithm 1 denotes the celebrated proximal mapping [3]. In the case where $L_{\nabla f}$ is not available or cannot be estimated, S-FM-HSDM offers the option of setting $f := f_n := 0$, where $L_{\nabla f}$ can be set to any positive real-valued number (*cf.* Section 3), whereas 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.

In the context of Section 1.2 and if (4a) is adopted, S-FM-HSDM

Algorithm 2: HRLSa

```
Stochastic oracle's input : $(\mathbf{a}_n, b_n)_{n\in\mathbb{Z}_{>0}}$.
User's input              : $\alpha$, $\lambda$.
Output                    : Sequence $(\mathbf{x}_n)_n$.
1 Initialization
2     Set $\alpha \in [0.5, 1)$ and $\lambda \in \mathbb{R}_{>0}$. Arbitrarily fix $\mathbf{x}_0$.
3     Set $\varpi_0 \geq \|\mathbf{R}_0\|$.
4     $\mathbf{x}_{1/2} := \mathbf{x}_0 - \alpha\frac{1}{\varpi_0}(\mathbf{R}_0\mathbf{x}_0 - \mathbf{r}_0)$.
5     For any $d \in \{1, \ldots, D\}$,
        $[\mathbf{x}_1]_d := [\mathbf{x}_{1/2}]_d \cdot (1 - \lambda/\max\{\lambda, |[\mathbf{x}_{1/2}]_d|\})$.
6 for $n = 0$ to $+\infty$ do
7     Set $\varpi_{n+1} \geq \|\mathbf{R}_{n+1}\|$.
8     $\mathbf{x}_{n+3/2} := \mathbf{x}_{n+1} + \mathbf{x}_{n+1/2} - \mathbf{x}_n + \alpha\frac{1}{\varpi_n}(\mathbf{R}_n\mathbf{x}_n -$
        $\mathbf{r}_n) - \frac{1}{\varpi_{n+1}}(\mathbf{R}_{n+1}\mathbf{x}_{n+1} - \mathbf{r}_{n+1})$.
9     For any $d \in \{1, \ldots, D\}$,
        $[\mathbf{x}_{n+2}]_d := [\mathbf{x}_{n+3/2}]_d \cdot (1 - \lambda/\max\{\lambda, |[\mathbf{x}_{n+3/2}]_d|\})$.
```

takes the flavor of Algorithm 2, coined HRLSa. Following (4a), Lines 3 and 7 of Algorithm 2 introduce over-estimates $(\varpi_n)_n$ of the maximum eigenvalues $(\lambda_{\max}(\mathbf{R}_n) = \|\mathbf{R}_n\|)_n$. To this end, motivated by the celebrated power iteration [13], and for a randomly generated initial vector $\mathbf{p}_0 \in \mathcal{X}$, the following iterative procedure, run over all $n \in \mathbb{Z}_{>0}$, is used in Section 3 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 (4b) 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.

Several assumptions follow to serve as the ground onto which Theorems 1 and 2 are based. The journal version of this manuscript will also include more details and exemplifying instances of the following assumptions in the context of Section 1.2.

Assumption 1 (Pointwise ergodicity)**.** $\boldsymbol{\mathcal{E}}_n^R := \mathbf{R} - \mathbf{R}_n \xrightarrow{\text{a.s.}}_n \mathbf{0}$ and $\boldsymbol{\varepsilon}_n^r := \mathbf{r} - \mathbf{r}_n \xrightarrow{\text{a.s.}}_n \mathbf{0}$. □

Assumption 2 (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 a.s. and $\forall n$,

$$t_n := \mathbb{E}_{|\mathcal{F}_n}\left[\sum\nolimits_{\nu=1}^{n}(T - T_\nu)x_\nu\right] = \mathbb{E}_{|\mathcal{F}_n}\left\{\sum\nolimits_{\nu=1}^{n}[(\operatorname{Id} - T_\nu)x_\nu - (\operatorname{Id} - T)x_\nu]\right\} . \quad (5)$$

Consider the event

$$\Omega_t := \{\lim\nolimits_{n\to\infty}\mathbb{E}(t_n) \in \operatorname{ran}(\operatorname{Id} - Q)\} . \quad (6)$$

Let $\mathbb{P}(\Omega_t) > 0$. □

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 [3].
- **(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 $L_{\nabla f} \in \mathbb{R}_{>0}$.
- **(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)** $(\nabla f - \nabla f_n) \xrightarrow{\text{a.s.}}_n 0$. □

Assumption 4 (Asymptotic unbiasedness)**.**

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

Assumption 5 (Dominated $(\vartheta_n)_{n\in\mathbb{Z}_{\geq 0}}$)**.** With respect to a sequence of random variables $(\vartheta_n)_{n\in\mathbb{Z}_{\geq 0}}$, whose description is out of the scope of this paper and will be given in the journal version of this manuscript, define

$$\Omega_\vartheta := \begin{Bmatrix} \exists\psi \in (\mathrm{m}\Sigma)^+ \text{ with } \mathbb{E}(\psi) < +\infty \\ \text{s.t. } \sum\nolimits_n \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ \leq \psi \text{ a.s.} \end{Bmatrix} ,$$

where $\mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)^+ := \max\{0, \mathbb{E}_{|\mathcal{F}_n}(\vartheta_n)\}$. Let $\mathbb{P}(\Omega_\vartheta) > 0$. □

Assumption 6 (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. $\mathbb{E}(\|\pi_n - \pi\|^2) \leq C_\pi$, $\forall n$. □

Assumption 7 (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 associated with the sequence of estimates $(x_n)_n$ (details are deferred to the journal version of the manuscript). 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. □

The main convergence properties of S-FM-HSDM (Algorithm 1) are summarized in the following two theorems.

Theorem 1. *Under Assumptions 2–7, and by choosing $\alpha \in [0.5, 1)$ and $\lambda \in (0, 2(1-\alpha)/L_{\nabla f})$, 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. Every point in $\mathfrak{C}[(x_n)_n]$ is a solution of (P) a.s.* □

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 2, 6(ii), 7(ii) and 7(iii), the sequence $(x_n)_n$ generated by Algorithm 1 converges a.s. to a solution of (P).* □

3. NUMERICAL TESTS

The proposed framework is validated within the setting of Section 1.2 where S-FM-HSDM(CRegLS), HRLSa and HRLSb are compared with the following online-learning and SA schemes: **i)** The classical RLS [23, §30.2]; **ii)** the ℓ_1-norm regularized (ℓ_1-)RLS [9], and its extension, the ℓ_0-norm (ℓ_0-)RLS [9], where a *non-convex* regularizing function is used instead of $\|\cdot\|_1$; **iii)** the LASSO-motivated online selective coordinate descent (OSCD) and online cyclic coordinate descent (OCCD) methods [1], where, according to [1, 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. $\rho_n \to_n 0$; **iv**) the proximal stochastic variance-reduced gradient (Prox-SVRG) method [28], applied to the setting of the ever-growing data regime $f := (1/n)\sum_{\nu=1}^{n} f_\nu$ in (CRegLS), with $f_\nu(\mathbf{x}) := (1/2)(\mathbf{a}_\nu^\intercal \mathbf{x} - b_\nu)^2$; **v**) SVRG-ADMM [29], where f is identical to that of the Prox-SVRG case; **vi**) the accelerated stochastic approximation (ACSA) with the step sizes of [19, (33)]; **vii**) the adaptive sparse variational Bayes multi-parameter Laplace prior (ASVB-MPL) method [27]; and **viii**) the stochastic dual-averaging (SDA) scheme with linear-convergence-rate guarantees [10]. It is worth stressing here that all of [1, 9, 10, 19, 28, 29] are built around the mainstream (CRegLS). As explained in Sections 1.1 and 1.2, any attempt to pass $\mathcal{A}$ of (HLS) to the objective function via the indicator function $\iota_{\mathcal{A}}$ [3] 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.

In all tests, the dimension of the Euclidean space $\mathcal{X} = \mathbb{R}^D$ is set to $D := 100$. The sparse system $\boldsymbol{\theta}_*$ is created by placing ± 1s 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. Since focus is placed on the system-identification problem of Section 1.2, the criterion of performance is the normalized-root-mean-square-deviation loss $\|\mathbf{x}_n - \boldsymbol{\theta}_*\|/\|\boldsymbol{\theta}_*\|$. The parameters of every method were carefully tuned to yield optimal performance per given scenario. Each curve in the figures is the uniform average of 500 independently performed tests.

With regards to the linear-regression model of Section 1.2, 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 $\mathbb{E}(\eta_n^2) = 10^{-\text{SNR(dB)}/10}\|\boldsymbol{\theta}_*\|^2\,\mathbb{E}([\mathbf{a}_n]_d^2)$. 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.

To test the ability of the methods to adapt to dynamic system changes, a typical AF test is considered here [23]: The sparsity level of the estimandum $\boldsymbol{\theta}_*$ changes abruptly at the time instance $2.5 \cdot 10^3$ from 1% to 10%, where the non-zero entries of $\boldsymbol{\theta}_*$ are re-allocated randomly.

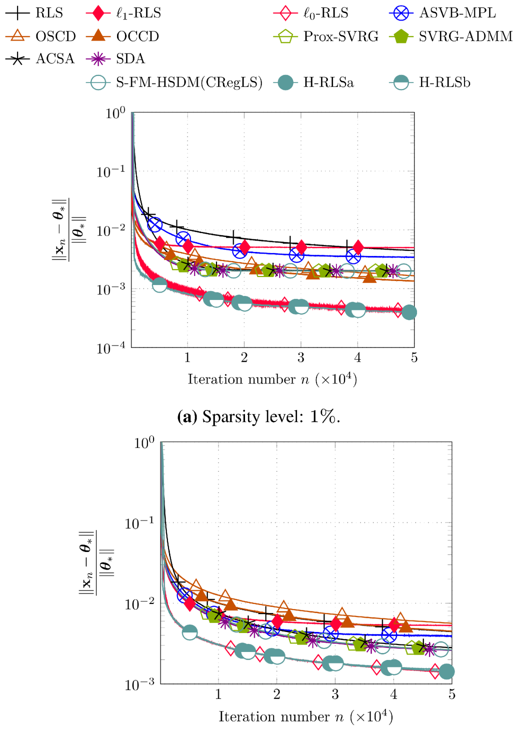
As in the classical exponentially-weighted RLS [23, §30.6], (CRegLS) is modified to

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

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.

4. CONCLUSIONS AND THE ROAD AHEAD

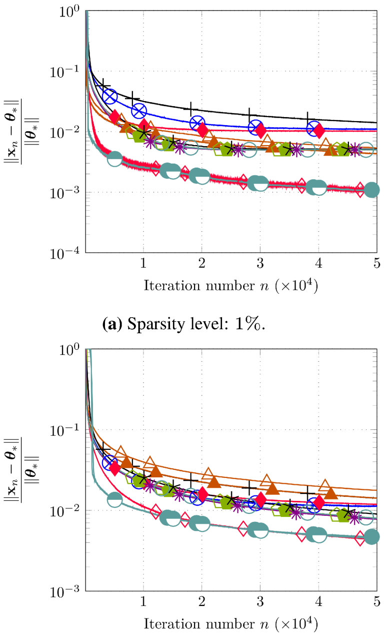
This paper presents the basic ingredients of a novel stochastic-approximation tool, namely the stochastic Fejér-monotone hybrid steepest descent method (S-FM-HSDM), designed to solve convex and affinely constrained composite minimization tasks. Noise contaminates the information about the task, affecting not only the



(a) Sparsity level: 1%.

(b) Sparsity level: 10%.

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



(a) Sparsity level: 1%.

(b) Sparsity level: 10%.

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

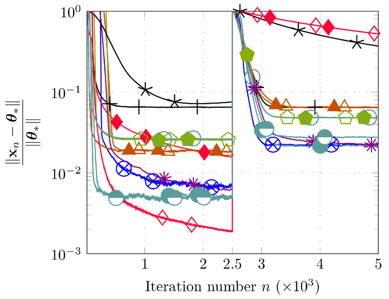


Fig. 3: IID $(\mathbf{a}_n)_n$; SNR $=$ 20dB; the sparsity level of $\boldsymbol{\theta}_*$ changes at time $n = 2.5 \cdot 10^3$ from 1% to 10%.

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 of state-of-the-art stochastic-approximation methods, such as splitting of variables and constant step size. Theorems, without proofs, are 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 are 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 for 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 and among several state-of-the-art AF, SA and online-learning schemes. Due to space limitations, the proofs of the theorems, rates of convergence, other theoretical contributions, as well as further applications of S-FM-HSDM will be presented elsewhere.

5. REFERENCES

[1] 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, July 2010.

[2] 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.

[3] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.

[4] J. Benesty, C. Paleologu, and S. Ciochină, "Proportionate adaptive filters from a basis-pursuit perspective," *IEEE Signal Process. Letters*, vol. 17, no. 12, pp. 985–988, Dec. 2010.

[5] P. Bianchi, W. Hachem, and A. Salim, "A constant step forward-backward algorithm involving random maximal monotone operators," *arXiv e-print*, 2017. [Online]. Available: arxiv.org/abs/1702.04144v3

[6] 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.

[7] 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.

[8] B. Dumitrescu, A. Onose, P. Helin, and I. Tăbuș, "Greedy sparse RLS," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2194–2207, May 2012.

[9] E. M. Eksioglu and A. K. Tanc, "RLS algorithm with convex regularization," *IEEE Signal Process. Letters*, vol. 18, no. 8, pp. 470–473, Aug. 2011.

[10] N. Flammarion and F. Bach, "Stochastic composite least-squares regression with convergence rate $\mathcal{O}(1/n)$," in *Proc. 2017 Conf. Learning Theory*, ser. Proc. Machine Learning Research, vol. 65, Jul. 2017, pp. 831–875.

[11] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite-element approximations," *Comp. Math. Appl.*, vol. 2, pp. 17–40, 1976.

[12] R. Glowinski and A. Marrocco, "Sur l'approximation par éléments finis et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Rev. Francaise d'Aut. Inf. Rech. Oper.*, vol. 9, no. 2, pp. 41–76, 1975.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 1996.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[15] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 315–323.

[16] N. Kalouptsidis and S. Theodoridis, Eds., *Adaptive System Identification and Signal Processing Algorithms*. Prentice Hall, 1993.

[17] E. Kreyszig, *Introductory Functional Analysis and Applications*, Wiley Classics Library ed. New York: Wiley, 1989.

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

[19] G. Lan, "An optimal method for stochastic composite optimization," *Math. Program., Ser. A*, vol. 133, pp. 365–397, 2012.

[20] 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.

[21] ——, "Primal-dual subgradient methods for convex problems," *Math. Program.*, vol. 120, no. 1, Ser. B, pp. 221–259, 2009.

[22] L. Rosasco, S. Villa, and B. C. Vũ, "A stochastic inertial forward–backward splitting algorithm for multivariate monotone inclusions," *Optimization*, vol. 65, no. 6, pp. 1293–1314, 2016.

[23] A. H. Sayed, *Adaptive Filters*. Hoboken: New Jersey: John Wiley & Sons, 2008.

[24] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[25] 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. Magaz.*, vol. 31, no. 6, pp. 124–129, Nov. 2014.

[26] K. Slavakis and I. Yamada, "Fejér-monotone hybrid steepest descent method for affinely constrained and composite convex minimization tasks," *Optimization*, 2018, DOI: 10.1080/02331934.2018.1505885.

[27] 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, Sept. 2014.

[28] 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.

[29] S. Zheng and J. T. Kwok, "Fast-and-light stochastic ADMM," in *Proc. Intern. Joint Conf. Artificial Intelligence*, Las Vegas: NV: USA, July 2016, pp. 2407–2413.