Per-Block-Convex Data Modeling by Accelerated Stochastic Approximation

Konstantinos Slavakis  Georgios B. Giannakis†

Abstract

Applications involving dictionary learning, non-negative matrix factorization, subspace clustering, and parallel factor tensor decomposition tasks motivate well algorithms for per-block-convex and non-smooth optimization problems. By leveraging the stochastic approximation paradigm and first-order acceleration schemes, this paper develops an online and modular learning algorithm for a large class of non-convex data models, where convexity is manifested only per-block of variables whenever the rest of them are held fixed. The advocated algorithm incurs computational complexity that scales linearly with the number of unknowns. Under minimal assumptions on the cost functions of the composite optimization task, without bounding constraints on the optimization variables, or any explicit information on bounds of Lipschitz coefficients, the expected cost evaluated online at the resultant iterates is provably convergent with quadratic rate to an accumulation point of the (per-block) minima, while subgradients of the expected cost asymptotically vanish in the mean-squared sense. The merits of the general approach are demonstrated in two online learning setups: (i) Robust linear regression using a sparsity-cognizant total least-squares criterion; and (ii) semi-supervised dictionary learning for network-wide link load tracking and imputation with missing entries. Numerical tests on synthetic and real data highlight the potential of the proposed framework for streaming data analytics by demonstrating superior performance over block coordinate descent, and reduced complexity relative to the popular alternating-direction method of multipliers.

1 Introduction

Aiming at succinct representations of large-scale data, models relying on non-convex functions have emerged as a prominent tool to learn low-dimensional structure from (possibly high-dimensional) data. Areas of interest span signal processing and machine learning applications including dictionary learning (DL) [1, 3], non-negative matrix factorization (NMF) [4], subspace clustering (SSC) [5], parallel factor (PARAFAC) decomposition of multi-way tensors [6], and total least-squares (TLS) [7], to name a few.

Consider DL for specificity, where a given \( M \times 1 \) vector \( \mathbf{y}_t \) is modeled as the product of an unknown over-complete dictionary \( \mathbf{D} := [\mathbf{d}_1, \ldots, \mathbf{d}_Q] \), \( Q \geq M \), times an unknown sparse coefficient vector \( \mathbf{s}_t \) [2, 3]. With \( \mathbf{Y}_t := [\mathbf{y}_1, \ldots, \mathbf{y}_t] \), and likewise for \( \mathbf{S} \), DL solves

\[
\min_{(\mathbf{S}, \mathbf{D})} \frac{1}{2t} \| \mathbf{Y}_t - \mathbf{DS} \|_F^2 + \lambda \| \mathbf{S} \|_1 + \nu_{\mathcal{B}}(\mathbf{D})
\]  

(1)
where \( \| \cdot \|_F \) denotes the Frobenius norm; the scale \( \lambda_s > 0 \) controls the sparsity effected by the \( \ell_1 \)-norm \( \| S \|_1 := \sum_{q,t} |s_{q,t}| \); the set indicator is \( \nu_{\mathcal{D}}(D) = 0 \) if \( D \in \mathcal{D} \), and \( \nu_{\mathcal{D}}(D) = +\infty \) otherwise, where the set \( \mathcal{D} := \{ D \in \mathbb{R}^{M \times Q} \mid \| d_q \| \leq 1, q \in \{1, \ldots, Q\} \} \) confines the dictionary to have bounded-norm columns. This constraint fixes the inherent scale ambiguity of the bilinear fit \( DS \), and also ensures that the solution of (1) remains bounded. Sparsity on the other hand, renders DL representations identifiable even when \( y_t \) has missing entries [8], due to e.g., malfunctioning, privacy reasons, or, high cost of data gathering.

Due to the bilinear term \( DS \), the three-summand cost \( F_t(S, D; \theta_t) := f_t(S, D; \theta_t) + g_1(S) + g_2(D) \) in (1) is non-convex (set \( \theta_t := \{ y_1, \ldots, y_t \} \) collects observations up to \( t \)) However, \( F_t \) is clearly “per-block-convex,” as it is convex in either \( S \) or \( D \), if the other one is fixed. Related multilinear forms emerge also with NMF, SSC, PARAFAC, and TLS models. Mainly for offline optimization, block coordinate descent methods (BCDMs) are popular largely because they exploit efficiently the per-block-convexity of the cost functions involved [3][9][19]. For online DL, BCDMs alternate between two iterations to update current estimates \((S_{t-1} := [s_1, \ldots, s_{t-1}], D_{t-1})\) as follows [3]

\[
\begin{align*}
  s_t &\in \arg \min_{s} F_t([S_{t-1}, s], D_{t-1}; \theta_t) & \text{(2a)} \\
  D_t &\in \arg \min_{D} F_t(S_t, D; \theta_t). & \text{(2b)}
\end{align*}
\]

Given \( \theta_t \), each step in (2) is a convex optimization task: Basis pursuit [20] in (2a), and constrained least-squares (LS) in (2b). However, the per-block minimizations in BCD may not be affordable by today’s big data applications, where the sheer volume and dimensionality of \( \theta_t \) strain computing resources [21][22]. Further, as data are streaming, analytics must often be performed in real time, without a chance to revisit past entries – a feature common to stochastic approximation (SA) setups [23].

In the spirit of SA, the present paper deals with minimizing expected value costs of the form

\[ \min_{x} \mathbb{E}_{\theta} \{ F_t(x; \theta_t) \} \]  

(3)

where \( x \) comprises all blocks of variables, e.g., \( x := (S, D) \) in (1), and expectation \( \mathbb{E} \) is over the random \( \theta_t \), whose probability density function (pdf) is unknown. The goal is to develop a modular algorithmic framework for solving (3) that: i) Leverages per-block-convexity of \( F_t \) as in BCDMs; ii) it operates online with streaming data \( \theta_t \) of unknown pdf; iii) relies only on first-order (sub)gradient information of \( F_t \), bypassing the need for (almost) exact minimizers per block as in (2); iv) it incurs affordable complexity per iteration, at most linear with respect to (w.r.t.) the number of unknowns; and v) iterations converge quadratically to a solution of (3), which is optimal among first-order methods in the sense of [24].

To place our contributions i)-v) in context, related first-order online BCDMs include the proximal stochastic (sub)gradient iterations [23][25][28], whose convergence tends to be slow even for convex problems, on top of being challenged by step-size choices. A relevant stochastic algorithm is the SA-based alternating-direction method of multipliers (ADMM) [29][32], that is known to be sublinearly convergent for convex costs [33][36], but no similar results are available for per-block-convex functions. On the other hand, accelerated first-order quadratically convergent iterations are available for off-line convex optimization [24][37][40]; see also [11] for related SA-based minimizers of convex costs. Even though [42] deals with non-convex costs, it requires bounds on the (primal) variables, knowledge of a
bound on the Lipschitz coefficient of the gradient operator for the algorithm to operate, and does not exploit the modularity offered by per-block-convexity.

Our work markedly broadens the offline acceleration technique introduced for convex costs in [39], to per-block-convex and to online SA setups. Unless the per-block optimization task can be solved in simple closed-form, there is no need for exact minimizers per block. Without knowing the data pdf and by relying only on first-order information of the instantaneous cost $F_i$, at linear complexity per iteration, we prove that the expected cost converges quadratically. Neither bounds on the block variables nor knowledge of bounds on Lipschitz coefficients are required. Under minimal assumptions and without imposing any block-wise strong convexity on the cost, performance analysis is carried out both for the cost values and the block variables of task [43]. Specifically, we establish that the expected limit cost is an accumulation point of (per-block) minima, and that subgradients of the expected cost asymptotically vanish in the mean-squared sense. The analytical results are tested on two instances of broad practical interest: (i) Online, robust and sparsity-aware linear TLS regression, using synthetic data; and (ii) online semi-supervised DL for network-wide link load tracking and imputation of real data. Numerical tests corroborate our analytical claims, and demonstrate that under a linear computational complexity footprint the proposed algorithm outperforms BCDMs and the computationally heavier ADMM-based alternatives [8].

The rest of the manuscript is organized as follows. Preliminaries are given in Sec. 2, while the proposed algorithm is developed in Sec. 3. Performance analysis is the subject of Sec. 4, with proofs delegated to Appendix A. Two examples of principal practical interest are provided in Sec. 5. Numerical tests both on synthetic and real data are presented in Sec. 6, while the manuscript is concluded in Sec. 7. Preliminary results were presented in [43], and outlined in [21].

2 Preliminaries

A first-order algorithm for the off-line minimization of a convex cost $\varphi(x) := f(x) + g(x)$, $x \in \mathcal{M}$, was studied in [39] (presented for convenience in Table A) where $\mathcal{M}$ is a linear vector space; $f$ is convex as well as $L$-Lipschitz continuously differentiable; and $g$ is convex but possibly non-smooth, e.g., the $\ell_1$-norm. Auxiliary variables $\{\psi_i, \zeta_i\} \subset \mathcal{M}$ are utilized to generate a sequence $(\varphi(x_i))_{i \in \mathbb{Z}_+}$ that converges as $i \to +\infty$ to the (global) minimum of $\varphi$ with quadratic rate. The engine under the hood is the forward-backward (FB) [44] or proximal-gradient iteration of line 5 where the proximal mapping is defined as $\text{Prox}_{\beta_i g} : \mathcal{M} \to \mathcal{M} : x \mapsto \arg\min_{\xi \in \mathcal{M}} \|x - \xi\|^2/2 + \beta_i g(\xi)$ for any $\beta_i \in \mathbb{R}_{>0}$ [44]. The FB iteration splits operation on $\varphi$ into two concatenated stages: Firstly on the differentiable $f$ through the classical steepest-descent operator $(\text{Id} - \beta_i \nabla f)$, and secondly on $g$ via $\text{Prox}_{\beta_i g}$, which usually obtains closed-forms for the majority of regularizers $g$, e.g., $\text{Prox}_{\|\cdot\|_1}$ boils down to the soft-thresholding operator [45]. If the FB iteration were performed with $\psi_{i+1}$ taking the place of $\zeta_i$ in line 5 then $(\psi_i)_{i \in \mathbb{Z}_+}$ would converge to a minimizer of $\varphi$ [44], but with no claims on quadratic rate of convergence. Towards establishing such claims, $\zeta_i$ of line 5 is convexly combined with $x_{i-1}$ to form $(1 - \lambda_i)x_{i-1} + \lambda_i \zeta_i$, which, together with line 3 guarantee that the values of $\varphi$ are monotonically non-increasing: $\varphi(x_i) \leq \varphi(x_{i-1})$. Parameters $\{\eta_{i+1}, \lambda_{i+1}\}$ in line 2 are used to define stepsize $\beta_{i+1}$ through line 3 offering the flexibility of a variable stepsize from the interval $[(1 - \sqrt{1 - \eta_{i+1} \lambda_{i+1} L})/L, (1 + \sqrt{1 - \eta_{i+1} \lambda_{i+1} L})/L]$ per iteration, as opposed to the rigid $\beta_{i+1} = 1/L$ in [37, 38]. Instrumental in establishing quadratic rate of convergence
Table 1: Minimizing the convex cost $\varphi := f + g$ \[39\]

**Require:** $\lambda, \eta \in \mathbb{R}_{>0}; \mu_1 := \lambda_1 \in [\lambda, 1]$; Lipschitz coeff. $L$ of $f$; arbitrary $(x_0, y_1) \in M^2$

1: for $i = 1, \ldots, +\infty$ do

2: Choose $(\lambda_{i+1}, \eta_{i+1})$ s.t. $0 < \lambda \leq \lambda_{i+1} \leq 1, 0 < \eta \leq \eta_{i+1} \leq \eta_i, \eta_{i+1} \lambda_{i+1} L \leq 1$.

3: Stepsize $\beta_{i+1} \in \{\beta \in \mathbb{R} | L^2 - 2\beta + \eta_{i+1} \lambda_{i+1} \leq 0\}$

   \[\frac{1 - \sqrt{1 - \eta_{i+1} \lambda_{i+1} L}}{L}, \frac{1 + \sqrt{1 - \eta_{i+1} \lambda_{i+1} L}}{L}\]

4: $\mu_{i+1} := \sqrt{(4\mu_i^2 + \lambda_i^2 \lambda_{i+1}^2 + \lambda_i \lambda_{i+1})/2}$

5: $\zeta_i := \text{Prox}_{\beta_i g}(\psi - \beta_i \nabla f(\psi))$  \hspace{1cm} $\triangleright$ Forward-backward iteration

6: $x_i := \begin{cases} (1 - \lambda_i) x_{i-1} + \lambda_i \zeta_i, & \text{if } \varphi((1 - \lambda_i) x_{i-1} + \lambda_i \zeta_i) \leq \varphi(x_{i-1}) \\ x_{i-1}, & \text{otherwise} \end{cases}$

7: $\psi_{i+1} := x_i + \frac{\mu_i}{\mu_{i+1}}(\zeta_i - x_i) + \frac{\mu_i - \lambda_i}{\mu_{i+1}}(x_i - x_i) + (\beta_i \eta_i \lambda_i) \frac{(\beta_i - \eta_i \lambda_i)}{\beta_i}(\psi_i - \zeta_i)$

8: end for

is the sequence of positive coefficients $(\mu_i)_{i \in \mathbb{Z}_{>0}}$ of line \[39\] (cf. Fact \[1\] in Appendix \[3, 4\], where $\lim_{i \to \infty} \mu_i = +\infty$). Finally, line \[7\] links variables $\{x_{i-1}, x_i, \psi_i, \psi_{i+1}, \zeta_i\}$ together, and, as it will be shown later on, it facilitates performance analysis via helpful telescoping terms. Under proper parameter selection $(\eta_i = \beta_i = 1/L, \lambda_i = 1)$, the algorithm of Table \[4\] boils down to \[38\]. Moreover, with its guaranteed monotonically non-increasing behavior of cost values through line \[6\] and the flexibility offered by the variable step-sizes $(\beta_i)_{i \in \mathbb{Z}_{>0}}$ in line \[3\] the algorithm in Table \[4\] has merits over \[37, 38\]. Notwithstanding, neither \[37, 38\] nor \[39\] can offer guarantees on the convergence of the (primal) variables $\{x_i, \psi_i, \zeta_i\}$. The purpose of this study is to extend the merits of the algorithm in Table \[4\] to the much more general setting where not only the underlying cost is time-varying, but it is per-block-convex, and several of its parameters are of stochastic nature.

To this end, and with reference to \[3\], consider a sequence of functions $F_i$ of composite structure $F_i(x; \Theta_i) := f_i(x; \Theta_i) + \sum_{b=1}^B g_b(x^{(b)})$, where $x$ gathers all unknowns, split in $B$ blocks of variables $x := (x^{(1)}, \ldots, x^{(B)})$, with $x^{(b)}$ belonging to a finite-dimensional linear space $M_b$ with inner-product $\langle \cdot, \cdot \rangle_{M_b}$; $f_i$ exhibits per-block-convexity, i.e., $f_i$ is convex w.r.t. each of the blocks $x^{(b)}$ whenever the rest of them are fixed; and $g_b$ is a convex function used to regularize and account for prior information on each $x^{(b)}$. Symbol $M$ stands for the Cartesian product $M := \times_{b=1}^B M_b$. An inner product on $M$ is defined as $\langle x_1, x_2 \rangle_M := \sum_{b=1}^B \langle x_1^{(b)}, x_2^{(b)} \rangle_{M_b}$, $\forall(x_1, x_2) \in M^2$. Whenever clear from the context, subscripts $M_b$ will be dropped from the inner-product symbols for notational convenience.

For any function $\Phi$ on $M$, notation $\Phi(x^{(b)} | x^{(-b)})$ stresses the dependence of $\Phi$ onto the $b$th block of variables $x^{(b)}$, whenever the rest of them are fixed, with $x^{(-b)}$ denoting all but the $b$th blocks contained in $x$. Term $\Phi(x | x^{(-b)})$ serves also the previous purpose, but with superscript $(b)$ dropped from $x^{(b)}$ to avoid overloading notation.

Let $\Gamma_0(M_b)$ denote all proper, convex, and lower-semicontinuous (l.s.c.) functions defined on $M_b$ with values in $\mathbb{R} \cup \{+\infty\}$ \[41\]. For any $\varphi \in \Gamma_0(M_b)$, the subdifferential $\partial \varphi(x^{(b)})$ is defined as the set of all subgradients $\varphi'(x^{(b)})$ of $\varphi$ at $x^{(b)}$: $\partial \varphi(x^{(b)}) := \{\varphi'(x^{(b)}) \in M_b | \langle \varphi'(x^{(b)}) \rangle_{M_b} \xi - x^{(b)} \rangle + \varphi(x^{(b)}) \leq \varphi(\xi), \forall \xi \in M_b\}$.
If $\varphi$ is differentiable at $x^{(b)}$, $\partial \varphi(x^{(b)}) = \{\nabla_b \varphi(x^{(b)})\}$, with $\nabla_b$ denoting the gradient operator w.r.t. the $b$th block in $x$. In this context, per realization of the r.v.s in $\Theta_t$, $f_t(\cdot \mid x^{(-b)}; \Theta_t) \in \Gamma_0(M_b)$ is assumed $[L_t^{(b)} := L_t^{(b)}(x^{(-b)}, \Theta_t)]$-Lipschitz continuously differentiable. Moreover, $g_b \in \Gamma_0(M_b)$.

Data $\Theta_t$ are considered to be r.v.s. defined on a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ with $\mathbb{E}\{\cdot\} := \mathbb{E}_{\mathfrak{A}}\{\cdot\}$ denoting expectation w.r.t. the $\sigma$-algebra $\mathfrak{A}$ \cite{44}. If $\mathcal{D}$ is the $\sigma$-subalgebra of $\mathfrak{A}$ including all events related to data $\{\Theta_t\}_{t \in \mathbb{Z}_{>0}}$, $\mathbb{E}_\mathcal{D}\{\cdot\}$ stands for expectation w.r.t. $\mathcal{D}$. Whenever block $x^{(b)}$ is viewed as an r.v., it is assumed to have finite second-order moment. In this context, $\mathcal{H}_b := \{x : \Omega \rightarrow M_b \mid \mathbb{E}\{\|x\|_2^2\} < +\infty\}$ turns out to be a Hilbert space with inner-product $\mathbb{E}\{(x_1 \mid x_2)_{M_b}\}$, $\forall (x_1, x_2) \in \mathcal{H}_b^2$ \cite{44} Example 2.6. To generalize, $\mathcal{H} := \bigotimes_{b=1}^B \mathcal{H}_b$ is also a Hilbert space with inner-product $\mathbb{E}\{(x_1 \mid x_2)_{\mathfrak{M}}\}$, $\forall (x_1, x_2) \in \mathcal{H}^2$. If $\mathcal{X}$ denotes a $\sigma$-subalgebra of $\mathfrak{A}$ which includes all events related to all blocks of variables, then $\mathbb{E}_{\mathcal{X}}\{\cdot\}$ stands for expectation w.r.t. $\mathcal{X}$ \cite{16}. Moreover, $\mathbb{E}_{\mathcal{D} \cup \mathcal{X}}\{\cdot\}$ denotes conditional expectation w.r.t. $\mathcal{D}$, conditioned on $\mathcal{X}$ \cite{16}. Hereafter, it is assumed that $\mathcal{X} \cup \mathcal{D} = \mathfrak{A}$, so that $\mathbb{E}\{\cdot\} = \mathbb{E}_{\mathfrak{A}}\{\cdot\} = \mathbb{E}_{\mathcal{X} \cup \mathcal{D}}\{\cdot\}$.

### 3 Algorithm

The algorithm of this section incorporates the acceleration module of Table I into the online learning setup of \cite{5}. Given that variables are split in blocks $x = (x^{(1)}, \ldots, x^{(B)})$, the proposed algorithm takes advantage of the per-block convexity of the cost $F_t$ and visits blocks of variables in a Gauss-Seidel or successive fashion. The basic principles of this modular algorithm are depicted in the block diagram of Fig. 1. Per iteration (time slot) $t$ and given observed data $\Theta_t$, the algorithm visits all blocks of variables successively to solve the per-block $b$ convex minimization task $\min_{x^{(b)} \in \mathfrak{M}_b} f_t(x^{(b)} \mid x^{(-b)}_t) + g_b(x^{(b)})$. Symbol $\Theta_t$ is dropped from $f_t(x^{(b)} \mid x^{(-b)}_t; \Theta_t)$ for notational convenience, and $x^{(-b)}_t := (x^{(1)}_t, \ldots, x^{(b-1)}_t, x^{(b+1)}_t, \ldots, x^{(B)}_t)$ comprises all updated blocks up to the $(b-1)$st one, as well as the $\{b+1, \ldots, B\}$ unvisited ones. If the solution to the previous minimization task
is affordable both w.r.t. time and computational resources, then block $b$ is updated by the obtained minimizer; otherwise, the acceleration module of Table 1 is run only for a finite number of iterations $R_b$, i.e., $i \in \{1, \ldots, R_b\}$ in the context of Table 1 and not indefinitely often ($i \rightarrow +\infty$) as in the batch and off-line mode of [39]. Having the $b$th block updated, effort is put on the next $(b+1)$st one. Once all blocks have been updated, the previous procedure is repeated for the $(t+1)$st time instant, and so on.

Since the acceleration module of Table 1 is allowed to be employed for $R_b$ times per $(b, t)$, the time index $\tau^{(b)} := (t-1)R_b + r_b$, with $r_b \in \{1, \ldots, R_b\}$, is introduced here to account for this “overclocking” or finer-time-scaling of the original $t$-axis. Because data are originally observed according to the “$t$-clicks,” to abide by the $\tau$-click notation, define $\Theta_{\tau^{(b)}} := \Theta_{(t-1)R_b + r_b} := \Theta_t$, as well as the induced functions $F_{\tau^{(b)}} := F_{(t-1)R_b + r_b} := F_t$ and $f_{\tau^{(b)}} := f_{(t-1)R_b + r_b} := f_t$.

A more detailed version of the block diagram of Fig. 1 equipped with the $\tau$-time notation, is given in Table 2. The acceleration module of Table 1 called here Accel($\cdot$), has been revised in Table 2b to abide by the notation which pertains to per-block $b$ operation. Overclocking and acceleration can be
seen in lines 7–13 of Table 2a. Symbol $x^{(-b)}_\tau$ in line 9 (the $b$ superscript from $\tau^{(b)}$ is omitted for notational convenience) stands for all but the $b$th blocks of variables, where only blocks $\{1, \ldots, b-1\}$ have been updated. On the other hand, $x^{(b)}_\tau$ in line 12 collects $x^{(-b)}_\tau$ and the recently updated $b$th block of variables. Symbol $\mathcal{J}^{(b)}_\tau$ stands for the input arguments of the acceleration module Accel($\mathcal{J}^{(b)}_\tau$) in line 11, which is expanded in Table 2b. These input arguments consist of only those parts of the cost which are affected by the $b$th block update; the value of $x^{(b)}_\tau$ at the previous time instant $\tau^{(b)} - 1$ is used as well as the Accel($\cdot$)'s intrinsic variables $\{\psi^{(b)}_\tau, \eta^{(b)}_\tau\}$. Finally, $x_t$ in line 10 collects all $B$ updated blocks at time $t$.

Implementing only the first-order information of $f_t(\cdot | x^{(-b)}_\tau)$ in (M5) equips the algorithm in Table 2 with a low computational footprint which scales linearly w.r.t. the number of unknown parameters. For specificity, the computational complexities on two practical examples will be provided in Sec. 5.

4 Main Result

The following assumptions will be instrumental in the subsequent discussion.

[As0] (Stationarity.) Expectation $F(x) := \mathbb{E}_\mathcal{O}\{F_t(x; \mathcal{O}_t)\}$ is time-invariant.

[As1] (Boundedness from below.) Functions $F_t(x; \mathcal{O}_t)$ are bounded from below almost surely (a.s.).

[As2] (Coercivity.) If $\lim_{k \to \infty} \mathbb{E}\{\|\xi_k\|^2\} = +\infty$ for any $(\xi_k)_{k \in \mathbb{Z}_{\geq 0}} \subset \mathcal{H}$, then $\lim_{k \to \infty} \mathbb{E}\{F(\xi_k)\} = +\infty$ [43].

[As3] With $(x^{(b)}_\tau)_{\tau \in \mathbb{Z}_{\geq 0}}$ standing for the sequence of estimates of the algorithm in Table 2 and $F_*$ denoting the limit which appears in Thm. 11, there exist $x^{(b)}_* \in \mathcal{H}$ s.t. $\limsup_{\tau \to \infty} \mathbb{E}\{F(x^{(b)}_* | x^{(-b)}_\tau)\} \leq F_*$. 

[As4] Given the sequence of Lipschitz coefficients $(L^{(b)}_\tau)_{\tau \in \mathbb{Z}_{\geq 0}}$ produced by the algorithm in Table 2, there exists $\hat{L} \in \mathbb{R}_{\geq 0}$ s.t. $L^{(b)}_\tau \leq \hat{L}$ a.s. $\forall (\tau, b)$. Moreover, there exist $\tau^{(b)}_* \in \mathbb{Z}_{\geq 0}$ and a sufficiently small $\delta^{(b)} \in \mathbb{R}_{\geq 0}$ s.t. $\eta^{(b)}_\tau \lambda^{(b)}_\tau \hat{L} \leq 1 - \delta^{(b)}$, $\forall \tau \geq \tau^{(b)}_*$.

[As5] If $w^{(b)}_\tau \in \arg \min_{x^{(b)} \in \mathcal{H}} \mathbb{E}\{F(x^{(b)} | x^{(-b)}_\tau)\}$, then $(\mathbb{E}\{\|w^{(b)}_\tau\|^2\})_{\tau \in \mathbb{Z}_{\geq 0}}$ is bounded.

Comments on the previous assumptions are in order. First, As0 can be recognized as one of the principal hypotheses in SA. As1 will be used to prevent cost values from sinking to $-\infty$, and it is usually met in practice, e.g., any quadratic data-fit term as well as any vector-norm satisfy As1 due to non-negativity. As2 will be used to prevent the proposed algorithm from generating unbounded sequences of estimates, without any a-priori enforcement of hard bounds on the variables, as in [12]. Examples where coercivity is introduced via $\{g_b\}_{b=1}^B$ will be given shortly in Sec. 5. With regard to As3, it will be shown in Lemma 3 that it is a necessary condition for properties related to (weak/strong sequential) cluster points of $(x^{(b)}_\tau)_{\tau \in \mathbb{Z}_{\geq 0}}$, as well as to the boundedness of (sub)gradients of the expected cost. The existence of a sufficiently large $\hat{L}$, which upper-bounds the data-dependent Lipschitz coefficients in As4, is well-motivated by the coercivity assumption As2 that promotes bounded sequences of iterates (cf. Thm. 11). The clarification of the previous statement will be given through concrete examples in Remark 2 where
the boundedness of the resultant iterates, as well as an assumption on the boundedness of the moments of the observed data, justify the existence of \( \tilde{L} \). \textbf{As4} and the related performance analysis suggest also strategies for selecting \( \{ \eta^b_\tau, \lambda^b_\tau \} \) (cf. Remark \[1\]). It is important to stress here that the stepsize \( \beta^b_\tau \) of the forward-backward iteration relies on the “local” Lipschitz coefficient \( L^b_\tau \) and not on \( \check{L} \), e.g., \( \beta^b_\tau := 1/L^b_\tau \). Finally, \textbf{As5} imposes a uniform bound, across time, on the second-order moment of (per-block) minimizers of the expected cost. This is the case if moments of the observed data \( (\Theta_t)_{t \in \mathbb{Z}^+} \) and block variables are bounded. In this sense, \textbf{As5} is necessary here since the present framework does not enforce hard bounds on block variables and follows the more relaxed coercivity assumption of \textbf{As2}.

It is also important to stress here that \textbf{As3} is a condition on existence; there is no need of constructing such minimizers for the algorithm to operate.

The following lemma gathers a few helpful properties on the function \( \mathbb{E}\{F(\cdot \mid \mathbf{x}^{(b)})\} \).

\textbf{Lemma 1.} Per block \( b \) and for any realization of \( \mathbf{x}^{(b)} \), function \( x^{(b)} \mapsto \mathbb{E}\{F(x^{(b)} \mid \mathbf{x}^{(b)})\} \) is convex on \( \mathcal{H}_b \). Moreover, \( \mathbb{E}_{\mathcal{D}:\mathbf{x}} \{F'(x^{(b)}; \Theta_t \mid \mathbf{x}^{(b)})\} \in \partial \mathbb{E}\{F(x^{(b)} \mid \mathbf{x}^{(b)})\} \). In other words, \( \mathbb{E}_{\mathcal{D}:\mathbf{x}} \{F'(x^{(b)}; \Theta_t \mid \mathbf{x}^{(b)})\} \) is a subgradient of \( \mathbb{E}\{F(\cdot \mid \mathbf{x}^{(b)})\} \) at \( x^{(b)} \). Further, under \textbf{As0} and \textbf{As1} \( \mathbb{E}\{F(\cdot \mid \mathbf{x}^{(b)})\} \) is bounded from below on \( \mathcal{H} \).

\textbf{Proof.} See Appendix A.1. \hfill \Box

The following lemma sheds light on the connection between the selection of \( \{ \eta^b_\tau, \lambda^b_\tau \} \) and the negativity of the quadratic polynomial in (M3).

\textbf{Lemma 2.} Under \textbf{As4} there exists a sequence of stepsizes \( \{\beta^b_\tau\}_{\tau \in \mathbb{Z}^+} \) and a \( \check{\delta}^b \in \mathbb{R}^b \) s.t.

\[
L^b_\tau \beta^b_\tau \leq 2\beta^b_\tau + \eta^b_\tau \lambda^b_\tau \leq -\check{\delta}^b, \quad \forall \tau \geq \tau_\ast.
\]

\textbf{Proof.} See Appendix A.2. \hfill \Box

The main results of this paper are summarized in the following theorem.

\textbf{Theorem 1.}

1. Under \textbf{As0} and \textbf{As1} there exists \( F_* \) to which \( (\mathbb{E}\{F(x_t)\})_{t \in \mathbb{Z}^+} \) converges. In other words, \( \exists F_* \in \mathbb{R} \) s.t. \( F_* = \lim_{t \to \infty} \mathbb{E}\{F(x_t)\} = \lim_{\tau \to \infty} \mathbb{E}\{F(x^{(b)}_\tau)\}, \forall b \in \{1, \ldots, B\} \), where \( x^{(b)}_\tau \) is defined in line 12 of Table 2a.

2. Under \textbf{As0} \textbf{As1} and \textbf{As2} sequences \( (x_t)_{t \in \mathbb{Z}^+} \) and \( (x^{(b)}_\tau)_{\tau \in \mathbb{Z}^+} \) in \( \mathcal{H} \), as well as \( (x^{(b)}_{\tau(b)})_{\tau(b) \in \mathbb{Z}^+} \) in \( \mathcal{H}_b \) are bounded. Consequently, the sets of weak sequential cluster points \( \mathcal{W}\{(x_t)_{t \in \mathbb{Z}^+}\}, \mathcal{W}\{(x^{(b)}_\tau)_{\tau \in \mathbb{Z}^+}\}, \) and \( \mathcal{W}\{(x^{(b)}_{\tau(b)})_{\tau(b) \in \mathbb{Z}^+}\} \) are non-empty \[14\] Lem. 2.37. Moreover, according to (M6) define

\[
\mathcal{F}^b_\tau := \left\{ \tau(b) \in \mathbb{Z}^+ \mid x^{(b)}_{\tau(b)} \neq x^{(b)}_{\tau(b)-1} \right\}.
\]

Then, sequence \( \{\zeta^{(b)}_{\tau(b)}\}_{\tau(b) \in \mathcal{F}^b_\tau} \subset \mathcal{H}_b \) is bounded and \( \mathcal{W}\{(\zeta^{(b)}_{\tau(b)})_{\tau(b) \in \mathcal{F}^b_\tau}\} \neq \emptyset \).
3. Under \(\text{As0} \text{ As3}\), \(\mathbb{E}\{F(x_{(b)}^{(b)})\}\) enjoys a quadratic rate of convergence to \(F_x\). More precisely, for any arbitrarily fixed \(\epsilon \in \mathbb{R}_{>0}\), there exists \(\tau'_0 \in \mathbb{Z}_{>0}\) s.t. \(\forall \tau' > \tau'_0\),
\[
\mathbb{E}\{F(x_{(b)}^{(b)})\} - F_x \\
\leq \frac{4}{\lambda_1^{(b)}2\bar{\lambda}_0^{(b)}2\bar{\eta}_0^{(b)}(1 + \tau)^2} \left[ \eta_0^{(b)} \mu_0^{(b)}2(\mathbb{E}\{F(x_{(b)}^{(b)})\} - F_x) \right] \\
+ \frac{1}{2} \mathbb{E}\left\{ \| \eta_0^{(b)} \lambda_0^{(b)} \mu_0^{(b)}(\eta_0^{(b)} + \nu_0^{(b)} - \lambda_1^{(b)} x_{(b)}^{(b)} \| \right\}^2 + \epsilon .
\]

4. Under \(\text{As0} \text{ As1} \text{ and As4}\) \(\sum_{\tau=0}^{\infty}\mathbb{E}\{\|\zeta_\tau^{(b)} - \psi_\tau^{(b)}\|^2\} < +\infty\). Necessarily, \(\lim_{\tau \to \infty}\mathbb{E}\{\|\zeta_\tau^{(b)} - \psi_\tau^{(b)}\|^2\} = 0\).

5. Under \(\text{As0} \text{ As1} \text{ and As4}\)
\[
\lim_{\tau \to \infty}\mathbb{E}\{\|\mathbb{E}_{\mathcal{D}|\mathcal{X}}\{F'_\tau(\zeta_\tau | x_\tau^{(b)}; \mathcal{D}_\tau)\}\|^2\} = 0.
\]
In other words, according to Lemma \(\text{I}\) the subgradient \(\mathbb{E}_{\mathcal{D}|\mathcal{X}}\{F'_\tau(\zeta_\tau | x_\tau^{(b)}; \mathcal{D}_\tau)\}\) of \(\mathbb{E}\{F(\cdot | x_\tau^{(b)})\}\) at \(\zeta_\tau^{(b)}\) converges to \(0\) in the mean-squared sense.

6. Under \(\text{As0} \text{ As2} \text{ and As4 As5}\)
\[
F_x = \lim_{\tau \in \bar{\zeta}_\tau^{(b)}} \mathbb{E}\{F(\zeta_\tau | x_\tau^{(b)})\} \\
= \lim_{\tau \to \infty} \min_{x_\tau^{(b)} \in \mathcal{H}_0} \mathbb{E}\{F(x_\tau^{(b)} | x_\tau^{(b)})\}.
\]
The last equation implies that there exists a subsequence \((x_{\tau_k}^{(b)})_{k \in \mathbb{Z}_{\geq 0}}\) that satisfies the following property: For any arbitrarily small \(\epsilon \in \mathbb{R}_{>0}\), there exists a \(k_0\) s.t. \(\forall k \geq k_0\),
\[
\mathbb{E}\{F(x_{\tau_k}^{(b)} | x_{\tau_k}^{(b)})\} - \min_{x_\tau^{(b)} \in \mathcal{H}_0} \mathbb{E}\{F(x_\tau^{(b)} | x_\tau^{(b)})\} \leq \epsilon .
\]
In other words, there exists \((x_{\tau_k}^{(b)})_{k \in \mathbb{Z}_{\geq 0}}\) onto which \(\mathbb{E}\{F(x_{\tau_k}^{(b)} | x_{\tau_k}^{(b)})\}\) approximate arbitrarily close the per-block minima \(\min_{x_{\tau_k}^{(b)} \in \mathcal{H}_0} \mathbb{E}\{F(x_{\tau_k}^{(b)} | x_{\tau_k}^{(b)})\}\).

\textit{Proof.} The proof is given in Appendix \(\text{A.3}\) \(\square\)

To show that \(\text{As3}\) is a rather weak assumption, the following Lemma \(\text{3}\) demonstrates that \(\text{As3}\) is necessary to the more “conventional” \(\text{As6a}\) and \(\text{As6b}\) on weak and strong sequential cluster points of the sequence of r.vs. \((x_\tau)_{\tau \in \mathbb{Z}_{\geq 0}}\). More specifically, \(\text{As6a}\) assumes existence of a strong sequential cluster point and bounds a sequence of subgradients of the expected cost, similarly to the bound on gradients introduced in \(\text{[47]}\).

\textbf{Lemma 3.} Under \(\text{As0 As2}\) if \(\mathbb{E}\{F(\cdot)\}\) is also l.s.c. on \(\mathcal{H}\), then \(\text{As3}\) is necessary to \(\text{As6a}\). Moreover, under \(\text{As0 As2 As3}\) is also necessary to \(\text{As6b}\).

\footnote{Both \(\text{As6a}\) and \(\text{As6b}\) are placed in Appendix \(\text{A.3}\) for not disrupting the flow of the present discussion.}
Two concrete examples of practical interest follow.

Remark 2. Setting \( \lambda^{(b)}_r := 1 \) for simplicity, \( \text{As}4 \) suggests that \( \eta^{(b)}_r \) should be sufficiently small for \( \eta^{(b)}_r \leq (1 - \delta^{(b)})/\hat{L} \), given \( \delta^{(b)} \in (0, 1) \) and assuming that \( \hat{L} \) is available. Without having knowledge of \( \hat{L} \), selection rules for \( \eta^{(b)}_r \) are (i) \( \eta^{(b)}_r := \hat{\eta}^{(b)} \), and (ii) \( \eta^{(b)}_r := \hat{\eta}^{(b)} + 1/\tau \), after choosing a sufficiently small \( \hat{\eta}^{(b)} > 0 \) [cf. \( \text{M}2 \)]. Notice that the previous rules abide by the monotonicity of \( \eta^{(b)}_r \), i.e., \( \eta^{(b)}_{r+1} \leq \eta^{(b)}_r \), in \( \text{M}2 \). In practice, and in the numerical tests of Sec. 6 the following rule is adopted:

\[
\lambda^{(b)}_r := 1, \quad \eta^{(b)}_r := \min \left\{ \eta^{(b)}_{r-1}, \frac{1 - \delta^{(b)}}{\hat{L}^{(b)}}, \hat{\eta}^{(b)} + \frac{1}{\tau} \right\}
\]

(4)

It is important to stress here that the selection of stepsize \( \beta_r^{(b)} \), in the forward-backward iteration of (M3), is based on the “local” coefficient \( L^{(b)}_r \) and not on \( \hat{L} \). Further discussion on \( \hat{L} \) is provided in Remark 2.

5 Examples

Two concrete examples of practical interest follow.

5.1 Total least-squares

Data \((y_t)_t \in \mathbb{Z}_{>0}\) are generated by \( y_t = u_t^T s_t + v_t \), where \((u_t, s_t) \in \mathbb{R}^Q \times \mathbb{R}^Q \); the unknown \( s_t \) is sparse; \( \top \) denotes transposition; and \( v_t \) stands for noise. Observed data are \( \mathcal{O}_t := \{y_t, u_t\} \), where \( u_t := u_t - e_t \) is a noisy version of \( u_t \), and no statistical information on the process \((e_t)_t \in \mathbb{Z}_{>0}\) is available. Motivated by the TLS criterion and the resultant errors-in-variables (EIV) modeling approach [7, 16], the following sequence of per-block-convex costs is considered:

\[
F_t(s, e; \mathcal{O}_t) := \frac{1}{2} \left[ y_t - (u_t + e) \right] \left[ y_t - (u_t + e) \right]^T + \frac{\lambda s_1}{2} \| s \|_1^2 + \frac{\lambda e_1}{2} \| e \|_1^2
\]

\[
=: g_1(s)
\]

\[
=: g_2(e)
\]

where the first quadratic term in (5) quantifies fitness to the observed data, with \( e \) modeling EIV; \( \| s \|_1 \) promotes sparsity on \( s \); \( \| e \|_1 \) penalizes large entries of \( e \); and \( \| s \|_1 \) is used to regularize the cost in (5) by imposing coercivity (cf. \( \text{As}2 \)). To draw connections with Sec. 2, \( \mathcal{M}_1 := \mathcal{M}_2 := \mathbb{R}^Q \), and \( x^{(1)} := s \), \( x^{(2)} := e \).

If \( \{y_t, u_t\} \) are (jointly) wide sense stationary, \( \text{As}0 \) holds with \( F(s, e) := \mathbb{E}_\mathcal{O} \{F_t(s, e; \mathcal{O}_t)\} \). Due to the non-negativity of all terms in (5), it can be readily verified that \( \text{As}1 \) is also satisfied.

In the case where both \( s \) and \( e \) are considered as r.v.s., \( x := (s, e) \in \mathcal{X} \), \( \lambda s \neq 0 \), \( \lambda e \neq 0 \), and (5) suggest that \( F_t(x; \mathcal{O}_t) \geq \lambda s_2 \| s \|_2^2/2 + \lambda e_2 \| e \|_2^2/2 \), and under \( \text{As}1 \) \( \mathbb{E} \{F(x)\} = \mathbb{E}_x \{\mathbb{E}_\mathcal{O}[F_t(x; \mathcal{O}_t)]\} \geq \lambda s_2 \mathbb{E}\{\| s \|_2^2\}/2 + \lambda e_2 \mathbb{E}\{\| e \|_2^2\}/2 \); hence, \( \text{As}2 \) holds.

Moreover, it can be verified by standard algebraic manipulations that a Lipschitz coefficient of \( \nabla_x f_t \) is \( L^{(s)}_t(e, u_t) = \| (u_t + e)(u_t + e)^\top + \lambda s_2 I_Q \| + \lambda e_2 \| I_Q \|_F = \| (u_t + e)(u_t + e)^\top \|_F + \lambda s_2 \| I_Q \|_F = \| u_t + e \|_2^2 + \lambda s_2 \sqrt{Q} \),
Following [8], consider an undirected graph \( G = (V,E) \), where \( V \) denotes the set of all vertices or nodes, with cardinality \( |V| \). Connectivity and edge strengths of \( G \) are described by the adjacency matrix \( W \in \mathbb{R}^{V \times V} \), where \( W_{ij} > 0 \) if nodes \( v_i \) and \( v_j \) are connected, while \( W_{ij} = 0 \) otherwise. Per t and node \( v_t \), r.v. \( \chi_{tv} : \Omega \to \mathbb{R} \) describes a network-wide dynamical process of interest, e.g., traffic load. All r.v.s. are collected in \( \chi_t := [\chi_{tv}]_{v=1}^{|V|} \). A succinct representation of the process over \( G \) models \( \chi_t \) as a superposition of “few” atoms in a dictionary \( D \in \mathbb{R}^{V \times Q} \), \( Q \geq V \): \( \chi_t = Ds_t \), where \( s_t \in \mathbb{R}^Q \) is sparse. Further, only a few entries of \( \chi_t \) are observed. Such a missing-entries scenario is conceivable in cases where not all of \( \{ \chi_{tv} \}_{v=1}^{|V|} \) are observable due to privacy constraints, severely corrupted measurements, node failures, or, data collection costs. To this end, let the random masking matrix \( M_t \in \mathbb{R}^{M \times V} \), \( M < V \), whose \( m \)th row is the transpose of a canonical basis vector for \( \mathbb{R}^V \); in other words, \( M_t \chi_t \) selects \( M \) out of \( V \) entries of \( \chi_t \). To summarize, \( y_t = M_t Ds_t + v_t \), with observed data \( \theta_t := \{ y_t, M_t \} \) and \( v_t \) denoting noise.

To enable imputation of missing entries, the topology of \( G \) is utilized. Spatial correlation of the network is captured by the Laplacian matrix \( L := \text{diag}(W1_V) - W \), where \( \text{diag}(a) \) defines the diagonal matrix whose main diagonal entries are those of vector \( a \), and \( 1_V \in \mathbb{R}^V \) is the all-one vector. Given a “forgetting factor” \( \delta \in (0,1] \) to gradually diminish the effect of past data, define the per-block-convex cost

\[
F_t(s, D) = \sum_{\tau=1}^t \frac{\delta^{t-\tau} \| y_{\tau} - M_{\tau} Ds_{\tau} \|^2}{2\Delta_t} + \frac{\lambda_L}{2} s^\top L D s + \frac{\lambda_s}{2} \| s \|^2
\]

where \( \Delta_t := \sum_{\tau=1}^t \delta^{t-\tau} \); \( \| s \|^2 \) and \( \| s \|_1 \) are as in [8], while the term including \( L \) quantifies prior knowledge on the topology of \( G \), promotes “smooth” solutions over strongly connected nodes of \( G \), and is instrumental in imputing missing entries [8].

To establish links with the introductory discussion, \( \mathcal{M}_1 := \mathbb{R}^Q (x^{(1)} := s) \), with \( \langle \cdot, \cdot \rangle_{\mathcal{M}_1} \) being the dot-vector product, and \( \mathcal{M}_2 := \mathbb{R}^{V \times Q} (x^{(2)} := D) \), with \( \langle D_1, D_2 \rangle_{\mathcal{M}_2} := \text{trace}(D_1^\top D_2), \forall (D_1, D_2) \in \mathcal{M}_2^2 \). If expectations in (6) are invariant w.r.t. \( t \), then \( \mathbb{E} \{ F(s, D) \} \) holds with \( \mathbb{E} \{ F(s, D; \theta_t) \} \). The non-negativity of all terms in (6) guarantees that \( \mathbb{E} \{ F(\cdot) \} \) is also satisfied. Further, by following a similar argument as in Sec. 5.1 for any \( \lambda_s \neq 0 \), \( \mathbb{E} \{ F(\cdot) \} \) satisfies Az2.

Standard algebra suggests that a Lipschitz coefficient of \( \nabla_s f_t \) is \( L_t^{(s)}(D, A_t) = \| D^\top A_t D + \lambda_s I_Q \| \leq \| D^\top A_t D + \lambda_s I_Q \|_F \leq \| D \|^2_F \| A_t \|_F + \lambda_s \sqrt{Q} \), where \( A_t := \sum_{\tau=1}^t \delta^{t-\tau} M_{\tau}^\top M_{\tau}/\Delta_t + \lambda_t L \). By \( \nabla_D f_t(D) = A_t D s s^\top - \sum_{\tau=1}^t \delta^{t-\tau} M_{\tau}^\top y_{\tau} s^\top /\Delta_t \), it can be also verified that a Lipschitz constant of \( \nabla_D f_t \) is \( L_t^{(D)}(s, A_t) = \| s \|^2 \| A_t \|_F \).
The computational complexities of the algorithm in Table 2 on examples of Secs. 5.1 and 5.2 including computations of Lipschitz constants and function evaluations, are linear w.r.t. to the number of unknown variables, and more specifically, in the order of $\Theta[(R_1 + R_2)Q]$ and $\Theta[(R_1 + R_2)(Q + V)V]$ per $t$, respectively.

Remark 2. With regard to the selection of $\hat{L}$ in (4), recall that Markov’s inequality dictates that $\Pr(L^v_t \geq \hat{L}) \leq \mathbb{E}\{L^v_t\}/\hat{L}$, for any $\hat{L}$. Provided that there exists $\hat{\Lambda}$ s.t. $\mathbb{E}\{L^v_t\} < \hat{\Lambda}$, one can arbitrarily decrease the measure of the event $\{\omega \in \Omega | L^v_t(\omega) \geq \hat{L}\}$ by choosing a sufficiently large $\hat{L}$. Examples for which $\mathbb{E}\{L^v_t\} < \hat{\Lambda}$ can be found in this section; regarding Sec. 5.1 it is straightforward to verify that $\mathbb{E}\{L^{v(s)}(e_t, u_t)\} \leq 2\mathbb{E}\{\|e_t\|^2\} + 2\mathbb{E}\{\|e_t\|^2\} + \lambda_s\sqrt{Q}$. Hence, under Thm. 1.2 and the assumption that $\mathbb{E}\{\|e_t\|^2\} < +\infty$, there exists $\hat{\Lambda} < \infty$ s.t. $\mathbb{E}\{L^{v(s)}(e_t, u_t)\} \leq \hat{\Lambda}$, $\forall t$. It can be also verified that the previous discussion carries over to the Lipschitz coefficients of Sec. 5.2 in a similar way.

6 Numerical Tests

6.1 Synthetic data

To validate the algorithm of Table 2 on the example of Sec. 5.1 entries of $(u_{st} := u_s, s_s)$ are drawn independently from a zero-mean, unit-variance Gaussian r.v. To make $s_s$ sparse, its nonzero entries are placed randomly in $s_s$ following a uniform distribution under two scenarios, a low-dimensional one corresponding to $(Q, \|s_s\|_0) = (100, 10)$, and a high-dimensional one with $(Q, \|s_s\|_0) = (10^3, 100)$, tagged “low-d” and “high-d” in Figs. 2a, 2b, and 2c, respectively. Noise $v_t$ is considered to be zero-mean and i.i.d. Gaussian, with variance $10^{-2}$. Regressor vectors $(u_t)_{t \in \mathbb{Z}^+}$ are observed after an i.i.d. zero-mean Gaussian process with variance $10^{-4}$ is added to $u_s$. Parameters $(\lambda_{s_1}, \lambda_{s_2}, \lambda_c) = (10^{-5}, 10^{-1}, 1)$ are used in (5), common to all employed methods. Minimization w.r.t. block $e$ accepts a closed-form solution; given $s$, the minimizer of (5) w.r.t. $e$ is $\hat{e} = (y_t - u_t^s)s (ss^T + \lambda_c I_Q)^{-1}$. It is worth noticing here that $R_1 = 3$ for the inner loop in Table 2 (lines 7-13). Parameters $(\eta^{(1)}, \lambda^{(1)})$ follow (7).

The algorithm in Table 2 is tested against a block-version of the classical online (sub)gradient descent method [23], tagged as BOGD in Fig. 2. BOGD adopts the Gauss-Seidel strategy of visiting blocks; per $t$, the standard subgradient descent step is applied first w.r.t. $s$ with constant step size $10^{-3}$, followed by a minimization step w.r.t. block $e$ which is given also here by the closed-form solution employed in the proposed scheme. Moreover, a block coordinate descent (BCD) strategy is validated, where (5) is “maximally” separated in scalar-valued blocks w.r.t. $s$. More specifically, following the Gauss-Seidel scheme and letting index $b \in \{1, \ldots, Q\}$ visit the $b$th entry of $s$ successively, having fixed $e$ and the entries $\{s_j\}_{j \neq b}$ of $s$, minimization of (5) w.r.t. $s_b$ amounts to the scalar-valued optimization task $\hat{s}_b := \arg\min_{s_b} |y_t - \sum_{j \neq b}(u_{tj} + e_j)s_j - (u_{tb} + e_b)s_b|^2/2 + \lambda_{s_2}s_b^2 + \lambda_{s_1}|s_b|$, which can be solved in closed form using the soft-thresholding operator as

$$\hat{s}_b = \begin{cases} \frac{(u_{tb} + e_b)s_b - \text{sgn}(u_{tb} + e_b)s_b}{(u_{tb} + e_b)^2 + \lambda_{s_2}}, & |(u_{tb} + e_b)s_b - \lambda_{s_1}| > 0 \\ 0, & \text{otherwise}. \end{cases}$$

Notice that $\theta_b := y_t - \sum_{j \neq b}(u_{tj} + e_j)s_j$, while $u_{tj}$ denotes the $j$th entry of $u_t$. After all entries of $s$ are updated, the closed form solution $\hat{e}$, leveraged in the proposed and BOGD techniques, updates block
Figure 2: Numerical results for the synthetic [(a), (b), (c)] and real data [(d)] of Secs. 6.1 and 6.2, respectively.

e to conclude step $t$ of BCD. For fairness, both BOGD and BCD run three consecutive iterations per $t$ to meet the computational load of the proposed scheme where $R_1 = 3$.

Figs. 2a, 2b, and 2c illustrate the performance of the employed methods. Fig. 2a depicts the cost function values (5) across time; Fig. 2b shows the per entry deviation $\left( \sum_{j \in \text{supp}(s_*)} (s_{tj} - s_*^j)^2 \right)^{1/2} / (Q - \|s_*\|_0)$ across $t$, where $\text{supp}(s_*)$ stands for the support of $s_*$, i.e., all those indexes $j$ s.t. $s_*^j \neq 0$; and Fig. 2c plots the time-variations of the subgradient norm of the cost. Curves in Figs. 2a, 2b, and 2c are obtained after averaging uniformly 100 realizations, and are illustrated in log-log scale for easily identifying the rate of convergence. Numerical results corroborate Thm. 1.3 which states that there exists a time instant after which the proposed algorithm converges with quadratic rate to an arbitrarily small neighborhood around $F_*$. The behavior of BOGD confirms the fact that (sub)gradient techniques are in general slow convergent.

6.2 Real data

In the context of Sec. 5.2, the advocated algorithm is validated on estimating and tracking network-wide link loads taken from the Internet2 measurement archive [48]. Analyzing the Internet2 backbone network yields a graph $\mathcal{G}$ with $V = 54$ number of vertices. Using the network topology and routing information, network-wide link loads $(\chi_t)_{t=1}^{30,000} \subset \mathbb{R}^V$ become available (in Gbps). Per time slot $t$, only $M = 30$ of the $\chi_t$ components, chosen randomly via $M_t \in \mathbb{R}^{M \times V}$, are observed in $y_t \in \mathbb{R}^M$. The cardinality of
the time-varying dictionaries is set constant to $Q = 80$. To cope with pronounced temporal variations of the Internet2 link loads, the forgetting factor $\delta$ in Sec. 5.2 is set equal to 0.95. Initial values for both $(s, D)$ are randomly drawn from the feasibility regions seen in Sec. 5.2. Parameters in (6) are defined as $(\lambda_L, \lambda_{s1}, \lambda_{s2}) = (10^{-3}, 10^{-3}, 10^{-5})$. Moreover, as in Sec. 6.1, $\{\eta_{\tau}^{(b)}, \lambda_{\tau}^{(b)}\}$ follow (4).

The advocated algorithm is tested against the state-of-the-art scheme in [8] which relies on a Gauss-Seidel alternating minimization scheme: (i) ADMM [29,30] is employed to minimize a cost closely related to (3) w.r.t. $s$, with the same parameters $(\lambda_L, \lambda_{s1}, \lambda_{s2})$ as in (6), and (ii) BCD iterations requiring matrix inversions are leveraged to optimize the associated loss w.r.t. $D$. Fig. 2d depicts the normalized squared estimation error between the true $\chi_t$ and the inferred $\hat{\chi}_t$, namely $\|\chi_t - \hat{\chi}_t\|^2/\|\chi_t\|^2$, versus time $t$ for a randomly chosen network link. For visualization reasons, only a small portion of the data is shown in Fig. 2d. To obtain computationally light recursions, the number of inner loops in Table 2 w.r.t. $s$ is set equal to $R_1 = 2$, while $R_2 = 5$ w.r.t. $D$. It is worth noticing here that ADMM in [8] requires multiple iterations to achieve a prescribed estimation accuracy, and that no matrix inversion was incorporated in the realization of Table 2. The proposed method and [8] perform similarly, scoring mean (normalized) estimation errors of 0.1166 and 0.1161 on the entire dataset of cardinality 30,000, respectively.

7 Conclusions

This manuscript presented a modular online learning algorithm which extended arguments, originally developed for accelerating first-order methods in batch convex optimization tasks, to the per-block-convex and stochastic approximation context. The proposed framework showed a computational complexity that scales linearly w.r.t. the number of unknowns. Assuming no knowledge of the underlying data statistics, the convergence rate of the expected loss on the resultant iterates was proved to be quadratic. Rigorous theoretical analysis was performed in the Hilbert space of r.v.s. of finite second-order moments. The framework was tested on two instances of broad practical interest: (i) Sparsity-aware regression based on the TLS criterion; and (ii) semi-supervised DL for network-wide link load tracking and imputation. Numerical tests on synthetic and real data demonstrated that the proposed algorithm performs better than BCDMs and comparably to state-of-the-art but computationally heavier ADMM-based methods. Future directions include the extension of the proposed framework from Gauss-Seidel strategies of visiting blocks of variables to parallel and random ones. Moreover, to study the effect of data non-stationarities, a regret analysis on the per-block convex loss will be presented in a future submission.

A Appendices

A.1 Proof of Lemma 1

Since $x^{(b)} \leftrightarrow F_t(x^{(b)} \mid x^{(-b)}; \mathcal{O}_t)$ is convex on $\mathcal{M}_b$ for any realizations of $x^{(-b)}$ and $\mathcal{O}_t$, then $\forall \lambda \in [0, 1]$, and $\forall (x_1^{(b)}, x_2^{(b)}) \in \mathcal{M}_b^2$, $F_t(\lambda x_1^{(b)} + (1 - \lambda)x_2^{(b)} \mid x^{(-b)}; \mathcal{O}_t) \leq \lambda F_t(x_1^{(b)} \mid x^{(-b)}; \mathcal{O}_t) + (1 - \lambda) F_t(x_2^{(b)} \mid x^{(-b)}; \mathcal{O}_t)$. Applying $\mathbb{E}_x \{\mathbb{E}_{\mathcal{O}}[\{\cdot\}]\}$ to both sides of the previous inequality yields the convexity of $\mathbb{E}\{F_t(\cdot \mid x^{(-b)})\}$ on $\mathcal{M}_b$.

The convexity of $F_t(\cdot \mid x^{(-b)}; \mathcal{O}_t)$ implies that for any $x^{(b)} \in \mathcal{M}_b$, $F_t(x^{(b)} \mid x^{(-b)}; \mathcal{O}_t) + (F_t(x^{(b)} \mid x^{(-b)}; \mathcal{O}_t) \mid \xi - x^{(b)}) \leq F_t(\xi \mid x^{(-b)}; \mathcal{O}_t)$, $\forall \xi \in \mathcal{M}_b$. Application of $\mathbb{E}_{\mathcal{O}}[\mathbb{E}_x[\{\cdot\}]]$ to both sides of
Moreover, by Table 2a (line 11),

\[ F_t(x^{(b)} | x^{(-b)}) + (\mathbb{E}_d | x^{(-b)}) \{ F_t'(x^{(b)} | x^{(-b)}; \mathcal{O}_t) \} \mid x^{(b)} \leq F_t(x^{(b)} | x^{(-b)}) \]

An additional application of \( \mathbb{E}\{\cdot\} \) yields \( \mathbb{E}\{ F_t(x^{(b)} | x^{(-b)}) \} + (\mathbb{E}_d | x^{(-b)}) \{ F_t'(x^{(b)} | x^{(-b)}; \mathcal{O}_t) \} \mid x^{(b)} \leq \mathbb{E}\{ F(x^{(b)} | x^{(-b)}) \} \) or, \( \mathbb{E}_d | x^{(-b)}) \{ F_t'(x^{(b)} | x^{(-b)}; \mathcal{O}_t) \} \in \partial \mathbb{E}(F_t(x^{(b)} | x^{(-b)})) \). Notice finally that it can be trivially verified by [As1] that \( \mathbb{E}(F(\cdot)) \) is bounded from below on \( \mathcal{H} \).

### A.2 Proof of Lemma 2

It can be verified that \( \min_{\beta \in \mathbb{R}} \{ L^{(b)}_2 - 2\beta + \eta^{(b)}_r \lambda^{(b)}_r \} = (\eta^{(b)}_r \lambda^{(b)}_r L^{(b)}_r - 1)/L^{(b)}_r \), which is attained at \( \beta^* = 1/L^{(b)}_r \). Due to \( \text{As1} \),

\[ (\eta^{(b)}_r \lambda^{(b)}_r L^{(b)}_r - 1)/L^{(b)}_r \leq (\eta^{(b)}_r \lambda^{(b)}_r \tilde{L} - 1)/L^{(b)}_r \leq (\eta^{(b)}_r \lambda^{(b)}_r \tilde{L} - 1)/\tilde{L} \leq -\delta^{(b)}/\tilde{L} \]

Hence, the choices of \( \beta^{(b)}_r := 1/L^{(b)}_r \) and \( \delta^{(b)} := \delta^{(b)}/\tilde{L} \) are sufficient to establish the claim of Lemma 2. Moreover, due to the continuity of the function \( L^{(b)}_r \beta^2 - 2\beta + \eta^{(b)}_r \lambda^{(b)}_r \) w.r.t. \( \beta \), one can always find a neighborhood of \( 1/L^{(b)}_r \) onto which the claim of Lemma 2 also holds for, let’s say, \( \delta^{(b)}/2 \).

### A.3 Proof of Theorem 1

First, a fact is in order.

**Fact 1** ([39] Lem. 1). 

1. Let \( \varphi := f + g \), where \( (f, g) \in \Gamma_0(\mathcal{M})^2 \), and \( f \) is \( L_f \)-Lipschitz continuously differentiable. For any \( \beta \in \mathbb{R}_{>0} \), define \( \varphi := \text{Prox}_{\beta \gamma} [\varphi - \varphi \nabla \lambda \psi] \), \( \forall \psi \in \mathcal{M} \). Assume that \( \exists \varphi, \varphi, \psi \in \mathcal{M} \), \( \exists \beta \in \mathbb{R}_{>0} \), and \( \exists \lambda \in [0, 1] \) s.t. \( \varphi(\varphi) \leq \varphi((1 - \gamma)w + \lambda \varphi) \). Then, \( \forall x \in \mathcal{M} \), \( \forall L \geq L_f \),

\[
\varphi(x) \leq (1 - \beta)\varphi(w) + \lambda \varphi(x) - \frac{\lambda}{\beta} \langle \varphi - \psi \mid \psi \rangle \\
+ \lambda \left( \frac{L}{2} - \frac{1}{\beta} \right) \| \varphi - \psi \|^2.
\]

2. Given \( (\lambda_i)_{i \in \mathbb{Z}_{>0}} \subset (0, 1] \), the sequence \( (\mu_i)_{i \in \mathbb{Z}_{>0}} \) defined in line 4 of Table 1 satisfies \( \mu_{i+1} > \mu_i \) and \( \mu_i \geq \lambda_i(1 + \sum_{i=1}^{\infty} \lambda_i) / 2 \), \( \forall i \in \mathbb{Z}_{>0} \).

**Remark 3.** Fact 1 holds for any over-estimate \( L \) of the Lipschitz coefficient \( L_f \). This flexibility is inherited by the subsequent performance analysis, and it facilitates computations in Table 2 in cases where computing \( L_f \) requires considerable effort; cf. Sec. 5 where the smallest Lipschitz coefficients require computation of the spectral norm of a matrix, while an over-estimate is provided by the manageable Frobenius norm. Under these considerations, it will be assumed that in all of the subsequent discussion there exists a sufficiently small \( \tilde{L} \in \mathbb{R}_{>0} \) that stands as a lower bound on all employed Lipschitz coefficients.

The proof of Thm. 1 now follows. Symbol \( \mathcal{O}_t \) is omitted to avoid overloading notations. For the same reason, superscript \( (b) \) is often omitted from \( (b) \).

1) By (M6), and the definition of \( x^{(-b)}_\tau \) given in Table 2a (line 9),

\[
F_t(x^{(b)} | x^{(-b)}) \leq F_t(x^{(-b)}_\tau | x^{(-b)}) = F_t(x^{(-b)} \mid x^{(-b)}_\tau).
\]

For \( \tau \in \{ (t-1)R_b + 2, \ldots, tR_b \} \), the previous inequality yields \( F_t(x^{(b)}_R | x^{(-b)}_R) \leq F_t(x^{(b)}_R | x^{(-b)}_R) \leq F_t(x^{(b)}_R | x^{(-b)}_R) \).

Moreover, by Table 2a (line 11), \( F_t(x^{(b)}_R | x^{(-b)}_R) \leq F_t(x^{(b)}_R | x^{(-b)}_R) \). Notice now by
the definition in Table 2a (line 12), that if \( \{x^{(b)}_{(t-1)R_b}\} \) is combined with \( x^{(b)}_{tR_b-1} \), then \( x^{(b)}_{tR_b-1} \) is obtained. Hence, the previous arguments summarize to \( F_t(x^{(b)}_{tR_b}) = F_t(x^{(b)}_{tR_b}) \) \( x^{(b)}_{tR_b-1} \) \( F_t(x^{(b)}_{tR_b}) \). If \( b \) assumes all consecutive values in \( \{1, \ldots, B\} \), then the previous inequality yields \( F_t(x_t) = F_t(x^{(b)}_{tR_b}) \). 

Hence, the previous arguments summarize to \( F_t(x^{(b)}_{tR_b}) = F_t(x_{t-1}) \), where \( x^{(b)}_{tR_b} := x_{t-1} \) was used in the last equality. Consequently, by \(|A_{s0}|E\{F(x_t)\} = E_X E_{D|X}(F_t(x_t)) \leq E_X E_{D|X}(F_t(x_{t-1})) = E\{F(x_{t-1})\} \). This suggests that the bounded-from-below sequence \( \{E\{F(x_t)\}\}_{t \in \mathbb{Z}_0} \) is non-increasing; thus convergent. In other words, \( \exists F_* \in \mathbb{R} \) s.t. \( F_* = \lim_{t \to \infty} E\{F(x_t)\} \).

With reference to (7) and the preceding arguments, \( \forall \tau \in \{(t - 1)R_b + r_b \mid r_b \in \{1, \ldots, R_b\}\} \)

\[
\mathbb{E}\{F(x_t)\} \leq \mathbb{E}\{F(x^{(b)}_{\tau}) \mid x^{(b)}_{\tau-1}\} = \mathbb{E}\{F(x^{(b)}_{\tau})\}
\]

\[
\leq \mathbb{E}\{F(x^{(b)}_{\tau-1}) \mid x^{(b)}_{\tau-1}\} = \mathbb{E}\{F(x^{(b)}_{\tau-1})\}
\]

\[
\leq \mathbb{E}\{F(x_{\tau-1})\}.
\]

Hence, since \( \{E\{F(x_t)\}\}_{t \in \mathbb{Z}_0} \) converges to \( F_* \), so does also the non-increasing \( \{E\{F(x^{(b)}_{\tau})\}\}_{\tau \in \mathbb{Z}_0} \). 

2) Due to \(|A_{s2}|\) and the existence of \( F_* \) by Thm. \( \text{[II]} \) the sequences \( \{x_t\}_{t \in \mathbb{Z}_0} \) and \( \{x^{(b)}_{\tau}\}_{\tau} \) are necessarily bounded. Moreover, due to the definition of \( \mathcal{H} \), boundedness of \( \{x^{(b)}_{\tau}\}_{\tau} \) implies also boundedness of the block-sequence \( \{x^{(b)}_{\tau}\}_{\tau} \subset \mathcal{H}_b, \forall i \). Now, \( \forall \tau \in \mathcal{H}_b \),

\[
\mathbb{E}\{\|\zeta^{(b)}_{\tau}\|^2\}
\]

\[
= \mathbb{E}\left\{\left\| \frac{1}{\lambda^{(b)}_{\tau}} (x^{(b)}_{\tau} - (1 - \lambda^{(b)}_{\tau})x^{(b)}_{\tau-1}) \right\|^2 \right\}
\]

\[
\leq \frac{2}{(\lambda^{(b)}_{\tau})^2} \mathbb{E}\{\|x^{(b)}_{\tau}\|^2\} + \frac{2(1 - \lambda^{(b)}_{\tau})^2}{(\lambda^{(b)}_{\tau})^2} \mathbb{E}\{\|x^{(b)}_{\tau-1}\|^2\}
\]

\[
\leq \frac{2\Delta}{(\lambda^{(b)}_{\tau})^2} \leq \frac{2\Delta}{(\lambda^{(b)})^2},
\]

where \( \Delta \geq \sup_{\tau \in \mathbb{Z}_0} \mathbb{E}\{\|x^{(b)}_{\tau}\|^2\} \). As such, \( \{\zeta^{(b)}_{\tau}\}_{\tau \in \mathcal{H}_b} \) is bounded, and, consequently, \( \mathcal{W}\{\zeta^{(b)}_{\tau}\}_{\tau \in \mathcal{H}_b} \neq \emptyset \) \( \text{[II]} \) Lem. 2.37).

3) The following proof is based on the one developed in \( \text{[9]} \) for the off-line, convex analytic case. The subsequent one offers a generalization in the context of the present online/stochastic setup.

By \(|M_6|\) in Table 2, \( F_{\tau+1}(x_{\tau+1} \mid x^{(b)}_{\tau+1}) \leq F_{\tau+1}((1 - \lambda_{\tau+1})x_{\tau} + \lambda_{\tau+1}\zeta_{\tau+1} \mid x^{(b)}_{\tau+1}) \). Fact \( \text{[II]} \) will be applied here with the convex \( F_{\tau+1}(\cdot \mid x^{(b)}_{\tau+1}) \) taking the place of \( \varphi \), and \( \{\psi_{\tau+1}, x_{\tau+1}, x_{\tau}, \zeta_{\tau+1}\} \) that of \( \{\psi, \xi, w, \zeta_{\psi}\} \). Let also \( x := x_{\tau} \) and an arbitrarily fixed \( \bar{x}^{(b)} \in M_b \) in Fact \( \text{[II]} \) to obtain

\[
F_{\tau+1}(x_{\tau+1} \mid x^{(b)}_{\tau+1}) - F_{\tau+1}(\bar{x}^{(b)} \mid x^{(b)}_{\tau+1})
\]

\[
\leq F_{\tau+1}(x_{\tau} \mid x^{(b)}_{\tau+1}) - F_{\tau+1}(\bar{x}^{(b)} \mid x^{(b)}_{\tau+1})
\]

\[
- \frac{\lambda_{\tau+1}}{\beta_{\tau+1}} (\zeta_{\tau+1} - \psi_{\tau+1} \mid \psi_{\tau+1} - x_{\tau})
\]

\[
+ \lambda_{\tau+1}\left(\frac{L_{\tau+1}}{2} - \frac{1}{\beta_{\tau+1}}\right) \|\zeta_{\tau+1} - \psi_{\tau+1}\|^2.
\]

(8)
Another application of Fact [I] with \( x = \bar{x}^{(b)} \) yields

\[
F_{r+1}(x_{r+1} | \bar{x}^{(b)}_{r+1}) - F_{r+1}(\bar{x}^{(b)} | \bar{x}^{(b)}_{r+1}) \\
\leq (1 - \lambda_{r+1}) \left[ F_{r+1}(x_{r} | \bar{x}^{(b)}_{r+1}) - F_{r+1}(\bar{x}^{(b)} | \bar{x}^{(b)}_{r+1}) \right] \\
- \frac{\lambda_{r+1}}{\beta_{r+1}} \langle \zeta_{r+1} - \psi_{r+1} | \psi_{r+1} - \bar{x}^{(b)} \rangle \\
+ \lambda_{r+1} \left( \frac{L_{r+1}}{2} - \frac{1}{\beta_{r+1}} \right) \| \zeta_{r+1} - y_{r+1} \| ^2.
\]  

(9)

Multiplying (8) by \( \mu_{r+1}(\mu_{r+1} - \lambda_1) \geq 0 \) and (10) by \( \mu_{r+1}\lambda_1 \geq 0 \), and adding the resultant inequalities,

\[
\mu_{r+1}^2 \left[ F_{r+1}(x_{r+1} | \bar{x}^{(b)}_{r+1}) - F_{r+1}(\bar{x}^{(b)} | \bar{x}^{(b)}_{r+1}) \right] \\
\leq \mu_{r+1} (\mu_{r+1} - \lambda_1 \lambda_{r+1}) \\
\times \left[ F_{r+1}(x_{r} | \bar{x}^{(b)}_{r+1}) - F_{r+1}(\bar{x}^{(b)} | \bar{x}^{(b)}_{r+1}) \right] \\
- \frac{\lambda_{r+1} \mu_{r+1}}{\beta_{r+1}} \langle \zeta_{r+1} - \psi_{r+1} \rangle \\
= : a_r \\
+ \mu_{r+1}^2 \lambda_{r+1} \left( \frac{L_{r+1}}{2} - \frac{1}{\beta_{r+1}} \right) \| \zeta_{r+1} - \psi_{r+1} \| ^2.
\]

Application of \( \langle a_r | b_r \rangle = (\| \eta a_r + b_r \| ^2 - \| b_r \| ^2 - \| \eta a_r \| ^2) / (2\eta) \), \( \forall \eta \in \mathbb{R}_{>0} \), to the previous inequality and \( v_r := \mu_r (1 - \eta_r \lambda_r/\beta_r) \psi_r - (\mu_r - \lambda_1) x_{r-1} \) yield

\[
\mu_{r+1}^2 \left[ F_{r+1}(x_{r+1} | \bar{x}^{(b)}_{r+1}) - F_{r+1}(\bar{x}^{(b)} | \bar{x}^{(b)}_{r+1}) \right] \\
\leq \mu_{r+1} (\mu_{r+1} - \lambda_1 \lambda_{r+1}) \\
\times \left[ F_{r+1}(x_{r} | \bar{x}^{(b)}_{r+1}) - F_{r+1}(\bar{x}^{(b)} | \bar{x}^{(b)}_{r+1}) \right] \\
- \frac{1}{2\eta_{r+1}} \left\| \eta_{r+1} \mu_{r+1} \bar{x}^{(b)} \right\| ^2 \\
+ \frac{1}{2\eta_{r+1}} \| \mu_{r+1} \psi_{r+1} - (\mu_{r+1} - \lambda_1) x_r - \lambda_1 \bar{x}^{(b)} \| ^2 \\
+ \frac{1}{2\eta_{r+1}} \left\| \eta_{r+1} \mu_{r+1} \bar{x}^{(b)} \right\| ^2 \\
+ \mu_{r+1}^2 \lambda_{r+1} \left( \frac{L_{r+1}}{2} - \frac{1}{\beta_{r+1}} \right) \| \zeta_{r+1} - \psi_{r+1} \| ^2.
\]  

(10)

It can be verified by (M7) that \( \mu_{r+1} \psi_{r+1} - (\mu_{r+1} - \lambda_1) x_r = \eta_r \lambda_r \mu_r \zeta_r / \beta_r + \mu_r (1 - \eta_r \lambda_r / \beta_r) \psi_r - (\mu_r - \lambda_1) x_{r-1} \). Notice also that (M4) is equivalent to \( \mu_r^2 = \mu_{r+1}(\mu_{r+1} - \lambda_1 \lambda_{r+1}) \). Incorporating these
arguments and (M3) into (11),

\[
\eta_{r+1}\mu_{\tau+1}^2 \left[ F_{r+1}(x_{r+1} \mid x_{\tau+1}) - F_{r+1} (\tilde{x}(b) \mid x_{\tau+1}) \right] \\
\leq \eta_{r+1}\mu_{\tau}^2 \left[ F_{r+1}(x_r \mid x_r) - F_{r+1} (\tilde{x}(b) \mid x_{\tau+1}) \right] \\
- \frac{1}{2} \left\| \frac{\eta_{r+1}\lambda_{r+1}\mu_{r+1}}{\beta_{r+1}} \zeta_{r+1} + v_{r+1} - \lambda_1\tilde{x}(b) \right\|^2 \\
+ \frac{1}{2} \left\| \frac{\eta_{r+1}\lambda_\mu_{r}}{\beta_r} \zeta_r + v_r - \lambda_1\tilde{x}(b) \right\|^2 \\
+ \frac{\eta_{r+1}\mu_{\tau}^2\lambda_{r+1}}{2\beta_{r+1}} (L_{r+1}\beta_\tau^2 - 2\beta_{r+1} + \eta_{r+1}\lambda_{r+1}) \\
\times \left\| \zeta_{r+1} - \psi_{r+1} \right\|^2 \\
\leq \eta_{r+1}\mu_{\tau}^2 \left[ F_{r+1}(x_r \mid x_r) - F_{r+1} (\tilde{x}(b) \mid x_{\tau+1}) \right] \\
- \frac{1}{2} \left\| \frac{\eta_{r+1}\lambda_{r+1}\mu_{r+1}}{\beta_{r+1}} \zeta_{r+1} + v_{r+1} - \lambda_1\tilde{x}(b) \right\|^2 \\
+ \frac{1}{2} \left\| \frac{\eta_{r+1}\lambda_\mu_{r}}{\beta_r} \zeta_r + v_r - \lambda_1\tilde{x}(b) \right\|^2 .
\]

The previous inequality suggests that

\[
\eta_{r+1}\mu_{\tau+1}^2 \left[ E_D \left\{ F_{r+1}(x_{r+1} \mid x_{\tau+1}) \right\} \\
- E_D \left\{ F_{r+1} (\tilde{x}(b) \mid x_{\tau+1}) \right\} \right] \\
+ \frac{1}{2} E_D \left\{ \left\| \frac{\eta_{r+1}\lambda_{r+1}\mu_{r+1}}{\beta_{r+1}} \zeta_{r+1} \\
+ v_{r+1} - \lambda_1\tilde{x}(b) \right\|^2 \right\} \\
\leq \eta_{r+1}\mu_{\tau}^2 \left[ E_D \left\{ F_{r+1}(x_r \mid x_r) \right\} \\
- E_D \left\{ F_{r+1} (\tilde{x}(b) \mid x_{\tau+1}) \right\} \right] \\
+ \frac{1}{2} E_D \left\{ \left\| \frac{\eta_{r+1}\lambda_\mu_{r}}{\beta_r} \zeta_r + v_r - \lambda_1\tilde{x}(b) \right\|^2 \right\} 
\]

which, according to [As0] results in

\[
\eta_{r+1}\mu_{\tau+1}^2 \left[ F(x_{r+1} \mid x_{\tau+1}) - F_* + F_* - F (\tilde{x}(b) \mid x_{\tau+1}) \right] \\
+ \frac{1}{2} E_D \left\{ \left\| \frac{\eta_{r+1}\lambda_{r+1}\mu_{r+1}}{\beta_{r+1}} \zeta_{r+1} \\
+ v_{r+1} - \lambda_1\tilde{x}(b) \right\|^2 \right\} \\
\leq \eta_{r+1}\mu_{\tau}^2 \left[ F(x_r \mid x_r) - F_* + F_* - F (\tilde{x}(b) \mid x_{\tau+1}) \right]
\]
After some elementary algebra,

\[
\eta_{\tau+1}\mu_{\tau+1}^2 \left[ F(x_{\tau+1} \mid x_{\tau+1}^{(b)}) - F_* \right] \\
+ \frac{1}{2} \mathbb{E}_{\mathcal{D}|x} \left\{ \left\| \frac{\eta_{\tau} \lambda_{\tau} \mu_{\tau}}{\beta_{\tau}} \zeta_{\tau} + v_{\tau} - \lambda_1 \bar{x}(b) \right\|^2 \right\}.
\]

Consequently,

\[
\eta_{\tau+1}\mu_{\tau+1}^2 \left[ \mathbb{E}\{ F(x_{\tau+1}^{(b)}) \} - F_* \right] \\
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_{\tau} \lambda_{\tau} \mu_{\tau}}{\beta_{\tau}} \zeta_{\tau} + v_{\tau} - \lambda_1 \bar{x}(b) \right\|^2 \right\} \\
\leq \eta_{\tau+1}\mu_{\tau+1}^2 \left[ \mathbb{E}\{ F(x_{\tau+1}) \} - F_* \right] \\
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_{\tau} \lambda_{\tau} \mu_{\tau}}{\beta_{\tau}} \zeta_{\tau} + v_{\tau} - \lambda_1 \bar{x}(b) \right\|^2 \right\} \\
+ \eta_0 (\mu_{\tau+1}^2 - \mu_{\tau}^2) \left[ \mathbb{E}\{ F(x_{\tau+1}^{(b)}) \} - F_* \right].
\]

(12a)

\[
\eta_{\tau+1}\mu_{\tau+1}^2 \left[ \mathbb{E}\{ F(x_{\tau+1}^{(b-1)}) \} - F_* \right] \\
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_{\tau} \lambda_{\tau} \mu_{\tau}}{\beta_{\tau}} \zeta_{\tau} + v_{\tau} - \lambda_1 \bar{x}(b) \right\|^2 \right\} \\
+ \eta_0 (\mu_{\tau+1}^2 - \mu_{\tau}^2) \left[ \mathbb{E}\{ F(x_{\tau+1}^{(b)}) \} - F_* \right].
\]

(12b)

\[
\leq \eta_{\tau+1}\mu_{\tau}^2 \left[ \mathbb{E}\{ F(x_{\tau}^{(b)}) \} - F_* \right] \\
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_{\tau} \lambda_{\tau} \mu_{\tau}}{\beta_{\tau}} \zeta_{\tau} + v_{\tau} - \lambda_1 \bar{x}(b) \right\|^2 \right\} \\
+ \eta_0 (\mu_{\tau+1}^2 - \mu_{\tau}^2) \left[ \mathbb{E}\{ F(x_{\tau+1}^{(b)}) \} - F_* \right].
\]

\[
\leq \eta_{\tau}\mu_{\tau}^2 \left[ \mathbb{E}\{ F(x_{\tau}^{(b)}) \} - F_* \right] \\
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_{\tau} \lambda_{\tau} \mu_{\tau}}{\beta_{\tau}} \zeta_{\tau} + v_{\tau} - \lambda_1 \bar{x}(b) \right\|^2 \right\} \\
+ \eta_0 (\mu_{\tau}^2 - \mu_{\tau}^2) \left[ \mathbb{E}\{ F(x_{\tau+1}^{(b)}) \} - F_* \right].
\]
where Table 2a (line 12) and the monotonicity in Thm. 1.1 were used in (12a) and (12b), while (M2) for any \( \bar{\tau} \) there exists \( \hat{\eta}/\hat{\eta} \) such that: Let now \( \vartheta(x_{*}^{(b)}) := \hat{\epsilon} \), by \( \bar{\eta}/\bar{\eta} \) \( \epsilon_{0} \in \limsup_{\tau \to 0} (\vartheta(x_{*}^{(b)}) - \mu_{*}^{2}(b) \sigma_{*}^{2} \epsilon_{0}^{2} + \mu_{*}^{2}(b) \sigma_{*}^{2} \epsilon_{0}^{2}) \), where \( \epsilon:= \epsilon \eta/\eta_{0} \). Hence, by (4) and (4)1:

\[
0 \leq \eta_{\tau+1} \mu_{\tau+1}^{2} \left[ E \{ F(x_{*+1}^{(b)}) \} - F_{*} \right] + \frac{1}{2} \mathbb{E} \left\{ \left\| \eta_{\tau+1} \lambda_{\tau+1}^{(b)} \mu_{\tau+1} - \lambda_{1} \bar{x}^{(b)} \right\|^{2} \right\} \leq \eta_{\tau+1} \mu_{\tau+1}^{2} \left[ E \{ F(x_{*}^{(b)}) \} - F_{*} \right] + \frac{1}{2} \mathbb{E} \left\{ \left\| \eta_{\tau+1} \lambda_{\tau+1}^{(b)} \mu_{\tau+1} - \lambda_{1} \bar{x}^{(b)} \right\|^{2} \right\} + \eta_{0} \left( \mu_{\tau+1}^{2} - \mu_{*}^{2}(b) \right) \epsilon^{2},
\]

and

\[
0 \leq \mathbb{E} \{ F(x_{*+1}^{(b)}) \} - F_{*} \leq \frac{1}{\eta_{\tau+1} \mu_{\tau+1}^{2}} \left[ \eta_{\tau+1} \mu_{\tau+1}^{2} \left( E \{ F(x_{*}^{(b)}) \} - F_{*} \right) \right] \leq \frac{1}{\eta_{\tau+1} \mu_{\tau+1}^{2}} \left[ \eta_{\tau+1} \mu_{\tau+1}^{2} \left( E \{ F(x_{*}^{(b)}) \} - F_{*} \right) \right].
\]
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_0}{\gamma} \frac{\lambda}{\tau} \zeta_{\tau+1} + v_{\tau+1} - \lambda_1 x_{\tau+1} \xrightarrow{\text{b}} \Gamma \right\| \right\}
+ \eta_0 \epsilon' \left(1 - \frac{\mu_{\tau+1}^2}{\mu_{\tau+1}^2} \right)
\leq \frac{4 \lambda_1^2 \mu_{\tau+1}^2}{\lambda^2 \tau^2 \eta(1 + \lambda(1 + \tau))^2} \left[ \eta_0 \mu_{\tau+1}^2 \left( \mathbb{E} \{ F(x_{\tau+1}) \} - F_\ast \right) \right.
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_0}{\gamma} \frac{\lambda}{\tau} \zeta_{\tau+1} + v_{\tau+1} - \lambda_1 x_{\tau+1} \xrightarrow{\text{b}} \Gamma \right\| \right\}
+ \frac{\eta_0}{\gamma} \epsilon'
\leq \frac{4 \lambda_1^2 \mu_{\tau+1}^2}{\lambda^2 \tau^2 \eta(1 + \lambda(1 + \tau))^2} \left[ \eta_0 \mu_{\tau+1}^2 \left( \mathbb{E} \{ F(x_{\tau+1}) \} - F_\ast \right) \right.
+ \frac{1}{2} \mathbb{E} \left\{ \left\| \frac{\eta_0}{\gamma} \frac{\lambda}{\tau} \zeta_{\tau+1} + v_{\tau+1} - \lambda_1 x_{\tau+1} \xrightarrow{\text{b}} \Gamma \right\| \right\}
+ \frac{\eta_0}{\gamma} \epsilon',
\end{aligned}
\end{equation}

where Fact 12 was utilized in (13a). The previous inequality establishes the claim of Thm. 13.

4) By (M3),

\begin{equation}
\frac{1 - \sqrt{1 - \eta_{\tau+1} L_{\tau+1}^b}}{L_{\tau+1}^b} \leq \beta_{\tau+1} \leq \frac{1 + \sqrt{1 - \eta_{\tau+1} L_{\tau+1}^b}}{L_{\tau+1}^b}.
\end{equation}

Moreover, since the preceding discussion holds for any over-estimate $L_{\tau+1}^b$ of the underlying Lipschitz constants (cf. Remark 3), one can always set a sufficiently small $L \in \mathbb{R}_{>0}$ as a lower-bound on all $L_{\tau+1}^b$, i.e., $L_{\tau+1}^b \geq L$, $\forall (\tau, i)$. Accordingly, the right-hand side of (14) suggests that $\beta_{\tau+1} \leq 2/L$. Given also that $\eta_{\tau+1} \geq \eta$, $\lambda_{\tau+1} \geq \lambda$, and Ass. then there exists $\delta > 0$ s.t. $-\eta_{\tau+1} L_{\tau+1}^b \beta_{\tau+1} - \eta_{\tau+1} \lambda_{\tau+1} (2\beta_{\tau+1} + \eta_{\tau+1} \lambda_{\tau+1}) \geq \delta$ s.t. $\forall \tau$. As a result, (13) implies that for any $x_{\tau+1}^{(b)}$,

\begin{align*}
\mu_{\tau+1}^2 \delta \| \zeta_{\tau+1} - \psi_{\tau+1} \|^2 \\
\leq \frac{\eta_{\tau+1} \mu_{\tau+1}^2 \lambda_{\tau+1} \lambda_{\tau+1}^2}{2 \beta_{\tau+1}^2} \left( \left( L_{\tau+1}^b \right)^2 \beta_{\tau+1}^2 + \eta_{\tau+1} \lambda_{\tau+1} \right) \\
\times \| \zeta_{\tau+1} - \psi_{\tau+1} \|^2 \\
\leq \eta_{\tau+1} \mu_{\tau+1}^2 \left[ F_{\tau+1}^{(b)} \left( x_{\tau+1} | x_{\tau+1}^{(b)} \right) - F_{\tau+1} \left( x_{\tau+1}^{(b)} | x_{\tau+1}^{(b)} \right) \right] \\
+ \eta_{\tau+1} \mu_{\tau+1}^2 \left[ F_{\tau+1}^{(b)} \left( x_{\tau+1} | x_{\tau+1}^{(b)} \right) - F_{\tau+1} \left( x_{\tau+1}^{(b)} | x_{\tau+1}^{(b)} \right) \right] \\
\leq \frac{1}{2} \left\| \frac{\eta_{\tau+1} \lambda_{\tau+1} \beta_{\tau+1}}{\beta_{\tau+1}^2} \zeta_{\tau+1} + v_{\tau+1} - \lambda_1 x_{\tau+1}^{(b)} \right\|^2.
\end{align*}
\[ + \frac{1}{2} \left\| \frac{\eta_r \lambda_r \mu_r}{\beta_r} \zeta_r + v_r - \lambda_1 \bar{x}(b) \right\|^2, \]

and

\[ \| \zeta_{t+1} - \psi_{t+1} \|^2 \]

\[ \leq \frac{\eta_{t+1} \mu_r}{\mu_r} \left[ F_{t+1}(x_r | x_{t+1}^{(-b)}) - F_{t+1}(\bar{x}(b) | x_{t+1}^{(-b)}) \right] \]

\[ - \frac{\eta_{t+1}}{\delta} \left[ F_{t+1}(x_r | x_{t+1}^{(-b)}) - F_{t+1}(\bar{x}(b) | x_{t+1}^{(-b)}) \right] \]

\[ + \frac{1}{2 \delta \mu_r^2} \left\| \frac{\eta_{t+1} \lambda_r + \mu_r}{\beta_r} \zeta_{t+1} + v_{t+1} - \lambda_1 \bar{x}(b) \right\|^2 \]

Applying expectations to the previous inequality yields

\[ \mathbb{E} \left\{ \| \zeta_{t+1} - \psi_{t+1} \|^2 \right\} \]

\[ \leq \frac{\eta_0}{\delta} \left[ \mathbb{E} \left\{ F(x_{t+1}^{(b-1)}) \right\} - \mathbb{E} \left\{ F(x_{t+1}^{(b)}) \right\} \right] \]

\[ - \frac{1}{2 \delta \mu_r^2} \mathbb{E} \left\{ \left\| \frac{\eta_{t+1} \lambda_r + \mu_r}{\beta_r} \zeta_{t+1} + v_{t+1} - \lambda_1 \bar{x}(b) \right\|^2 \right\} \]

\[ + \frac{1}{2 \delta \mu_r^2} \mathbb{E} \left\{ \left\| \frac{\eta_r \lambda_r \mu_r}{\beta_r} \zeta_r + v_r - \lambda_1 \bar{x}(b) \right\|^2 \right\}. \]
By definition, $f_{\tau}$ is the unique minimizer of $	ext{Prox}_{\beta_{\tau}g_{\tau}}(h) = \arg\min_{\zeta}[\|h - \xi\|^2/2 + \beta_{\tau}g_{\tau}(\xi)]$, there exists $g_{\tau}'(\text{Prox}_{\beta_{\tau}g_{\tau}}(h)) \in \partial g_{\tau}(\text{Prox}_{\beta_{\tau}g_{\tau}}(h))$ s.t. $\text{Prox}_{\beta_{\tau}g_{\tau}}(h) = h + \beta_{\tau}g_{\tau}'(\text{Prox}_{\beta_{\tau}g_{\tau}}(h)) = 0$. Following the notation of (15), a rearrangement of the previous equality yields that there exists $g_{\tau}'(\zeta_{\tau}) \in \partial g_{\tau}(\zeta_{\tau})$ s.t.

$$\psi_{\tau} - \zeta_{\tau} = \beta_{\tau}[\nabla_{b}f_{\tau}(\psi_{\tau} \mid x_{\tau}^{(b)}) + g_{\tau}'(\zeta_{\tau})].$$

By definition, $f_{\tau}(\cdot \mid x_{\tau}^{(b)})$’s range is included in $\mathbb{R}$; hence, the relative interior [41] p. 91 of its domain $f_{\tau}(\cdot \mid x_{\tau}^{(b)}) = \mathcal{M}_{b}$. Consequently, the definition $F_{\tau} = f_{\tau} + \sum_{b=1}^{B}g_{b}$, and the elementary ri dom $f_{\tau}(\cdot \mid x_{\tau}^{(b)}) \cap \text{ri dom } g_{b}(\cdot) = \text{ri dom } g_{b}(\cdot) \neq \emptyset$, suggest that $\partial_{b}F_{\tau}(\cdot \mid x_{\tau}^{(b)}) = \nabla_{b}f_{\tau}(\cdot \mid x_{\tau}^{(b)}) + \partial g_{b}(\cdot)$ [41 Cor. 16.38.iv]. As a result, the existence of $g_{\tau}'(\zeta_{\tau})$ and (15) establish that for any $\tau$, $b$, there exists $F_{\tau}'(\zeta_{\tau} \mid x_{\tau}^{(b)}) \in \partial_{b}F_{\tau}(\zeta_{\tau} \mid x_{\tau}^{(b)})$ s.t. $\psi_{\tau} - \zeta_{\tau} = \beta_{\tau}[F_{\tau}'(\zeta_{\tau} \mid x_{\tau}^{(b)}) + \nabla_{b}f_{\tau}(\psi_{\tau} \mid x_{\tau}^{(b)}) - \nabla_{b}f_{\tau}(\zeta_{\tau} \mid x_{\tau}^{(b)})].$

Hence,

$$\left\| F_{\tau}'(\zeta_{\tau} \mid x_{\tau}^{(b)}) \right\|^2 \leq \frac{2}{\beta_{\tau}^2} \| \psi_{\tau} - \zeta_{\tau} \|^2 + 2 \left\| \nabla_{b}f_{\tau}(\zeta_{\tau} \mid x_{\tau}^{(b)}) - \nabla_{b}f_{\tau}(\psi_{\tau} \mid x_{\tau}^{(b)}) \right\|^2.$$
where (16) utilized Assumption 1 and the fact that due to (14), \( \beta_t \geq (1 - (1 - \eta_t \lambda_t \rho_t)^{1/2})/L_t \geq 1/L_t \geq 1/\bar{L} \). Assuming that \( F'_\tau(\zeta | x_{\tau-1}^{(b)}) \in \mathcal{H}_t \) and applying expectations to (15) yields \( \mathbb{E}\{\|F'_\tau(\zeta | x_{\tau-1}^{(b)})\|^2\} \leq 4\bar{L}\mathbb{E}\{\|\psi_t - \zeta_t\|^2\} \), which together with Thm. 1.4 establish

\[
\lim_{\tau \to \infty} \mathbb{E}\left\{\|F'_\tau(\zeta | x_{\tau-1}^{(b)})\|^2\right\} = 0. \tag{17}
\]

Further, due to the convexity of \( \| \cdot \|_2 \), Jensen’s inequality for conditional expectations [46, § IV.3] suggests that

\[
\mathbb{E}\left\{\|F'_\tau(\zeta | x_{\tau-1}^{(b)})\|^2\right\} = \mathbb{E}_X \left\{\mathbb{E}_D | X \left\{\|F'_\tau(\zeta | x_{\tau-1}^{(b)})\|^2\right\}\right\} \geq \mathbb{E}_X \left\{\mathbb{E}_D | X \left\{F'_\tau(\zeta | x_{\tau-1}^{(b)})\right\}\right\}.
\]

Accordingly, the previous inequality together with (17) establish Thm. 1.3.

6) By (16), \( \forall \tau \in \mathcal{T}_\zeta \) (b), where \( \mathcal{T}_\zeta \) (b) is defined in Thm. 1.2, \( F_t(x_{\tau} | x_{\tau-1}^{(b)}) \leq \lambda_t F_t(\zeta | x_{\tau-1}^{(b)}) + (1 - \lambda_t) F_t(x_{\tau-1} | x_{\tau-1}^{(b)}) \). Hence, \( \forall \tau \in \mathcal{T}_\zeta \) (b),

\[
\mathbb{E}\{F(\zeta | x_{\tau-1}^{(b)})\} = \mathbb{E}\{F_t(\zeta | x_{\tau-1}^{(b)})\}
\geq \frac{1}{\lambda_t} \mathbb{E}\{F_t(x_{\tau} | x_{\tau-1}^{(b)})\} - \frac{1 - \lambda_t}{\lambda_t} \mathbb{E}\{F_t(x_{\tau-1} | x_{\tau-1}^{(b)})\}
= \frac{1}{\lambda_t} \mathbb{E}\{F(x_{\tau}^{(b)})\} - \frac{1 - \lambda_t}{\lambda_t} \mathbb{E}\{F(x_{\tau-1}^{(b)})\}
\geq \frac{1}{\lambda_t} \mathbb{E}\{F(x_{\tau}^{(b)})\} - \frac{1 - \lambda_t}{\lambda_t} \mathbb{E}\{F(x_{\tau}^{(b-1)})\}
= \frac{1}{\lambda_t} \left[\mathbb{E}\{F(x_{\tau}^{(b)})\} - \mathbb{E}\{F(x_{\tau}^{(b-1)})\}\right] + \mathbb{E}\{F(x_{\tau}^{(b-1)})\}
\geq \frac{1}{\lambda} \left[\mathbb{E}\{F(x_{\tau}^{(b)})\} - \mathbb{E}\{F(x_{\tau}^{(b-1)})\}\right] + \mathbb{E}\{F(x_{\tau}^{(b-1)})\}\tag{18a}
\]

\[
\geq \frac{1}{\lambda} \left[\mathbb{E}\{F(x_{\tau}^{(b)})\} - \mathbb{E}\{F(x_{\tau}^{(b-1)})\}\right] + F_*, \tag{18b}
\]

where \( \bar{\lambda} \leq \lambda_t, \forall \tau \), and the monotonicity properties \( \mathbb{E}\{F(x_{\tau}^{(b-1)})\} \leq \mathbb{E}\{F(x_{\tau-1} | x_{\tau-1}^{(b)})\} \), as well as \( F_* \leq \mathbb{E}\{F(x_{\tau}^{(b)})\} \leq \mathbb{E}\{F(x_{\tau-1}^{(b)})\} \), established in Thm. 1.1 were used in (18a), (18b), and (18c). Since \( \lim_{\tau \to \infty} [\mathbb{E}\{F(x_{\tau}^{(b)})\} - \mathbb{E}\{F(x_{\tau-1}^{(b)})\}] = 0 \) by Thm. 1.1 given any \( \epsilon \in \mathbb{R}_{>0} \), there exists \( \tau_0' \in \mathbb{Z}_{\geq 0} \) s.t. \( \forall \tau \in \mathcal{T}_\zeta \) (b) \( \cap [\tau_0', +\infty) \)

\[
\mathbb{E}\{F(\zeta | x_{\tau-1}^{(b)})\} \geq \frac{1 - \lambda_t}{\lambda_t} + F_* = F_* - \frac{\epsilon}{2}. \tag{19}
\]
Due to the elementary $\mathbb{E} \{\|\zeta^b_\tau - w^b_\tau\|^2\} \leq 2 \mathbb{E} \{\|\zeta^b_\tau\|^2\} + 2 \mathbb{E} \{\|w^b_\tau\|^2\}$, Thm. 1.2 and As5 suggest that there exists $\Delta$ s.t. $\mathbb{E} \{\|\zeta^b_\tau - w^b_\tau\|^2\} \leq \Delta^2$.

Now, according to Lemma 11 the subgradient inequality for function $\mathbb{E}\{F(\cdot | x^{(-b)}_\tau)\}$ yields $\forall \tau$,

$$
\mathbb{E}\{F(\zeta_\tau | x^{(-b)}_\tau)\} = \mathbb{E}\{F'_\tau(\zeta_\tau | x^{(-b)}_\tau)\}
\leq \mathbb{E}\{F'_\tau(w_\tau | x^{(-b)}_\tau)\}
+ \mathbb{E}\left\{\left(\mathbb{E}_{x|\tau}\{F'_\tau(\zeta_\tau | x^{(-b)}_\tau)\} | \zeta_\tau - w_\tau\right)\right\}_{x \in \mathcal{H}_b}
\leq \mathbb{E}\{F'_\tau(w_\tau | x^{(-b)}_\tau)\} + \Delta \left[\mathbb{E}\left\{\|\mathbb{E}_{x|\tau}\{F'_\tau(\zeta_\tau | x^{(-b)}_\tau)\}\|^2\right\}\right]^{1/2},
$$

(20)

where $\Delta$ bounds $\mathbb{E}\{\|\zeta^b_\tau - w^b_\tau\|^2\}_{\tau \in \mathcal{Z}_{>0}}$.

Since (20) holds for any subgradient, it holds also for the specific $\mathbb{E}_{\mathcal{O}|\mathcal{X}}\{F'_\tau(\zeta_\tau | x^{(-b)}_\tau)\}$ involved in Thm. 15 Moreover, by the definition of lim sup, $\exists x'_0 \in \mathcal{Z}_{>0} \cap [\tau_0', +\infty)$ s.t. $\forall \tau \in \mathcal{T}_\zeta^{(b)} \cap [\tau_0', +\infty)$,

$$
\mathbb{E}\{F(\zeta_\tau | x^{(-b)}_\tau)\} \leq \limsup_{\tau \to \infty} \mathbb{E}\{F(\tau | x_{\tau}^{(-b)})\} = \limsup_{\tau \to \infty} \min_{x \in \mathcal{H}_b} \mathbb{E}\{F(x | x_{\tau}^{(-b)})\} =: \bar{F}_b
\leq \limsup_{\tau \to \infty} \mathbb{E}\{F(x_\tau^{(b)} | x_{\tau}^{(-b)})\} + \frac{\epsilon}{2} = F_* + \frac{\epsilon}{2},
$$

(21)

To summarize, (19) and (21) suggest that $\forall \tau \in \mathcal{T}_\zeta^{(b)} \cap [\tau_0', +\infty)$, $F_* \leq \mathbb{E}\{F(\zeta_\tau | x^{(-b)}_\tau)\} + \epsilon/2 \leq \bar{F}_b + \epsilon \leq F_* + \epsilon$, and since $\epsilon$ was chosen arbitrarily, $F_* = \bar{F}_b$.

The previous result together with the definition of lim sup suggest that there exists a subsequence $(x^{(b)}_{\tau_k})_{k \in \mathcal{Z}_{>0}}$ s.t. $\lim_{k \to \infty} \min_{x \in \mathcal{H}_b} \mathbb{E}\{F(x | x_{\tau_k}^{(-b)})\} = F_*$. In other words, given any arbitrarily small $\epsilon \in \mathbb{R}_{>0}$, there exists $k_0'$ s.t. $\forall k \geq k_0'$, $|F_* - \min_{x \in \mathcal{H}_b} \mathbb{E}\{F(x | x_{\tau_k}^{(-b)})\}| \leq \epsilon/2$. Moreover, Thm. 11 has already demonstrated that $F_* = \lim_{\tau \to \infty} \mathbb{E}\{F(x^{(b)}_{\tau_k} | x^{(-b)}_{\tau_k})\}$, and there exists $k''_0$ s.t. $\forall k \geq k''_0$, $|F_* - \mathbb{E}\{F(x^{(b)}_{\tau_k} | x^{(-b)}_{\tau_k})\}| \leq \epsilon/2$. Putting everything together, by choosing any $k \geq \max\{k'_0, k''_0\}$, it can be readily deduced that $\forall k \geq k_0$, $\mathbb{E}\{F(x^{(b)}_{\tau_k} | x^{(-b)}_{\tau_k})\} - \min_{x \in \mathcal{H}_b} \mathbb{E}\{F(x | x^{(-b)}_{\tau_k})\} \leq \epsilon$.

A.4 Proof of Lemma 3

By Thm. 12 consider $x_*^{(b)} \in \mathcal{W}\{(x^{(b)}_{\tau})_{\tau \in \mathcal{Z}_{>0}}\} \neq \emptyset$. In other words, there exists a subsequence $(\tau_k)_{k \in \mathcal{Z}_{>0}}$ s.t. $\text{w-lim}_{k \to \infty} x^{(b)}_{\tau_k} = x^*_b$, where w-lim denotes the limit in the weak topology of $\mathcal{W}_b$ [44, §2.4].
Fix now arbitrarily $\epsilon \in \mathbb{R}_{>0}$. By the definition of $\limsup$, there exists $\tau'_0 \in \mathbb{Z}_{\geq 0}$ s.t. $\forall \tau \geq \tau'_0$, $\limsup_{\tau \to \infty} \mathbb{E}\{ F(x^{(b)}_\tau | x^{(-b)}_\tau) \} \leq \mathbb{E}\{ F(x^{(b)}_{\tau'_0} | x^{(-b)}_{\tau'_0}) \} + \epsilon/2$. By assumption, $\mathbb{E}\{ F(\cdot | x^{(-b)}_\tau) \}$ is l.s.c. on $\mathcal{H}_b$, and, consequently, it is also weakly sequentially l.s.c. [44 Thm. 9.1]. Hence, there exists a neighborhood $\mathcal{V}_\tau(x^{(b)}_\tau)$ of $x^{(b)}_\tau$ in the weak topology of $\mathcal{H}_b$ s.t. $\forall x \in \mathcal{V}_\tau(x^{(b)}_\tau), \mathbb{E}\{ F(x^{(b)}_\tau | x^{(-b)}_{\tau'_0}) \} \leq \mathbb{E}\{ F(x^{(b)}_{\tau'_0} | x^{(-b)}_{\tau'_0}) \} + \epsilon/2$ [44 Def. 1.21].

[As6a] There exists an open neighborhood $\mathcal{V}(x^{(b)}_\tau)$ of $x^{(b)}_\tau$ in the weak topology of $\mathcal{H}_b$ and a $\tau_{#} \in \mathbb{Z}_{\geq 0}$ s.t. $\mathcal{V}(x^{(b)}_\tau) \subset \cap_{\tau = \tau_{#}}^{+\infty} \mathcal{V}_\tau(x^{(b)}_\tau)$.

Since $w\text{-}\lim_{k \to \infty} x^{(b)}_{\tau_k} = x^{(b)}_{\tau'_0}$, there exists $k'_0 \in \mathbb{Z}_{\geq 0}$ s.t. $\forall k \geq k'_0$, $x^{(b)}_{\tau_k} \in \mathcal{V}(x^{(b)}_{\tau'_0})$. The previous arguments and [As6a] suggest that there exists a sufficiently large $k''_0 \in \mathbb{Z}_{\geq 0}$ s.t. $\tau_{k''_0} \geq \max\{ \tau_{#}, \tau'_0 \}$, $k''_0 \geq k'_0$, and consequently $\forall k \geq k''_0$,

$$\limsup_{\tau \to \infty} \mathbb{E}\{ F(x^{(b)}_\tau | x^{(-b)}_\tau) \}$$

$$\leq \mathbb{E}\{ F(x^{(b)}_{\tau''_0} | x^{(-b)}_{\tau''_0}) \} + \frac{\epsilon}{2}$$

$$\leq \mathbb{E}\{ F(x^{(b)}_{\tau''_0} | x^{(-b)}_{\tau''_0}) \} + \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

$$= \mathbb{E}\{ F(x^{(b)}_{\tau''_0}) \} + \epsilon.$$

Therefore,

$$\limsup_{\tau \to \infty} \mathbb{E}\{ F(x^{(b)}_\tau | x^{(-b)}_\tau) \} = \lim_{k \to \infty} \limsup_{\tau \to \infty} \mathbb{E}\{ F(x^{(b)}_\tau | x^{(-b)}_\tau) \} \leq \mathbb{E}\{ F(x^{(b)}_{\tau''_0}) \} + \epsilon = F_* + \epsilon.$$

The previous inequality holds for any $\epsilon$; hence $\limsup_{\tau \to \infty} \mathbb{E}\{ F(x^{(b)}_\tau | x^{(-b)}_\tau) \} \leq F_*$

[As6b] There exists a strong sequential cluster point $x^{(b)}_\tau \in \mathcal{C}\{ (x^{(b)}_\tau)_{\tau \in \mathbb{Z}_{\geq 0}} \}$, a $\Delta \in \mathbb{R}_{>0}$, and a sequence of subgradients $(\mathbb{E}_{\mathcal{H}}|_{\mathcal{X}} \{ F'_\tau(x^{(b)}_\tau | x^{(-b)}_\tau) \})_{\tau \in \mathbb{Z}_{>0}} \subset \mathcal{H}_b$ (cf. Lemma [I]) s.t.

$$\limsup_{\tau \to \infty} \mathbb{E}\left\{ \left\| \mathbb{E}_{\mathcal{H}}|_{\mathcal{X}} \{ F'_\tau(x^{(b)}_\tau | x^{(-b)}_\tau) \} \right\|^2 \right\} \leq \Delta.$$
Hence, $\forall \tau \geq \tau'_0$,

$$\limsup_{\tau \to \infty} \mathbb{E}\left\{ F\left(x_{*}^{(b)} | x_{*}^{(b)} \right) \right\}$$

$$\leq \mathbb{E}\left\{ F\left(x_{*}^{(b)} | x_{*}^{(b)} \right) \right\} + \epsilon$$

$$\leq \mathbb{E}\left\{ F\left(x_{*}^{(b)} | x_{*}^{(b)} \right) \right\} + \epsilon$$

$$+ \mathbb{E}\left\{ \langle \mathbb{E}_{O|X}\left\{ F_t(x_{*}^{(b)} | x_{*}^{(b)} \right) \right| x_{*}^{(b)} - x_{*}^{(b)} \rangle \right\}$$

$$\leq \mathbb{E}\left\{ F\left(x_{*}^{(b)} \right) \right\} + \epsilon$$

$$+ \left[ \mathbb{E}\left\{ \left\| x_{*}^{(b)} - x_{*}^{(b)} \right\|^2 \right\} \right]^\frac{1}{2} \times$$

$$\left[ \mathbb{E}\left\{ \left\| \mathbb{E}_{O|X}\left\{ F_t(x_{*}^{(b)} | x_{*}^{(b)} \right) \right| x_{*}^{(b)} - x_{*}^{(b)} \right\} \left\| ^2 \right\} \right]^\frac{1}{2}$$

$$\leq \mathbb{E}\left\{ F\left(x_{*}^{(b)} \right) \right\} + \sqrt{\Delta} \left[ \mathbb{E}\left\{ \left\| x_{*}^{(b)} - x_{*}^{(b)} \right\|^2 \right\} \right]^{\frac{1}{2}} + \epsilon ,$$

and $\forall k \geq k''_0$, for a sufficiently large $k''_0 \in \mathbb{Z}_{>0}$,

$$\limsup_{\tau \to \infty} \mathbb{E}\left\{ F\left(x_{*}^{(b)} | x_{*}^{(b)} \right) \right\}$$

$$\leq \mathbb{E}\left\{ F\left(x_{*}^{(b)} \right) \right\} + \sqrt{\Delta} \left[ \mathbb{E}\left\{ \left\| x_{*}^{(b)} - x_{*}^{(b)} \right\|^2 \right\} \right]^{\frac{1}{2}} + \epsilon .$$

By applying $\lim_{k \to \infty}$ to both sides of the previous inequality, it can be verified that

$\limsup_{\tau \to \infty} \mathbb{E}\left\{ F\left(x_{*}^{(b)} | x_{*}^{(b)} \right) \right\} \leq F_{*} + \epsilon$, and since $\epsilon$ was chosen arbitrarily, $\limsup_{\tau \to \infty} \mathbb{E}\left\{ F\left(x_{*}^{(b)} | x_{*}^{(b)} \right) \right\} \leq F_{*}$. 

**Acknowledgments**

The authors would like to thank Dr. P. Forero of SPAWAR for providing the code of the method in [8].

**References**

[1] B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?” *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[2] I. Tošić and P. Frossard, “Dictionary learning,” *IEEE Signal Process. Magaz.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[3] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *J. Machine Learn. Research*, vol. 11, pp. 19–60, Mar. 2010.

[4] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[5] R. Vidal, “Subspace clustering,” *IEEE Signal Process. Magaz.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[6] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.
[7] S. Van Huffel and P. Lemmerling, *Total Least Squares and Errors-in-Variables Modeling*. Springer, 2002.

[8] P. Forero, K. Rajawat, and G. B. Giannakis, “Prediction of partially observed dynamical processes over networks via dictionary learning,” *IEEE Trans. Signal Process.*, vol. 62, no. 13, pp. 3305–3320, July 2014.

[9] P. Tseng, “Convergence of block coordinate decent method for nondifferentiable minimization,” *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, June 2001.

[10] P. Tseng and S. Yun, “A coordinate gradient descent method for nonsmooth separable minimization,” *Math. Program.*, Ser. B, vol. 117, pp. 387–423, 2009.

[11] S. J. Wright, “Accelerated block-coordinate relaxation for regularized optimization,” *SIAM J. Optim.*, vol. 22, no. 1, pp. 159–186, 2012.

[12] Y. Nesterov, “Efficiency of coordinate descent methods on huge-scale optimization problems,” *SIAM J. Optim.*, vol. 22, no. 2, pp. 341–362, 2012.

[13] Y. Xu and W. Yin, “A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion,” *SIAM J. Imaging*, vol. 6, no. 3, pp. 1758–1789, 2013.

[14] P. Richtárik and M. Takáč, “Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function,” *Math. Program.*, Ser. A, pp. 1–38, Dec. 2012.

[15] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, “Sparse reconstruction by separable approximation,” *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, July 2009.

[16] H. Zhu, G. Leus, and G. B. Giannakis, “Sparsity-cognizant total least-squares for perturbed compressive sampling,” *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2002–2016, May 2011.

[17] M. Razaviyayn, M. Hong, and Z.-Q. Luo, “A unified convergence analysis of block successive minimization methods for nonsmooth optimization,” *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[18] Q. Tao, K. Kong, D. Chu, and G. Wu, “Stochastic coordinate descent methods for regularized smooth and nonsmooth losses;” in *Lecture Notes in Computer Science*, ser. Machine Learning and Knowledge Discovery in Databases, P. A. Flach, T. de Bie, and N. Cristianini, Eds., vol. 7523. Springer, 2012, pp. 537–552.

[19] M. Mardani, G. Mateos, and G. B. Giannakis, “Decentralized sparsity-regularized rank minimization: Algorithms and applications,” *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 5374–5388, Nov. 2013.

[20] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, 2006.

[21] K. Slavakis, G. B. Giannakis, and G. Mateos, “Modeling and optimization for big data analytics,” *IEEE Signal Process. Magaz.*, vol. 31, no. 5, pp. 18–31, Sept. 2014.

[22] 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.

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

[24] Y. Nesterov, “A method for solving the convex programming problem with convergence rate $O(1/k^2)$,” *Dokl. Akad. Nauk SSSR*, vol. 269, pp. 543–547, 1983, in Russian.

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

[26] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *J. Machine Learn. Research*, vol. 12, pp. 2121–2159, 2011.

[27] S. Ghadimi, G. Lan, and H. Zhang, “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization,” Aug. 2013, [arXiv:1308.6594].
[28] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, “A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks,” Jul. 2013, arXiv:1307.4457.

[29] 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.

[30] 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.

[31] I. D. Schizas, G. Mateos, and G. B. Giannakis, “Distributed LMS for consensus-based in-network adaptive processing,” IEEE Trans. Signal Process., vol. 57, no. 6, pp. 2365–2381, June 2009.

[32] G. Mateos, I. D. Schizas, and G. B. Giannakis, “Performance analysis of the consensus-based distributed LMS algorithm,” EURASIP Journal on Advances in Signal Processing, vol. 2009, Dec. 2009.

[33] H. Ouyang, N. He, L. Q. Tran, and A. Gray, “Stochastic alternating direction method of multipliers,” in Proc. ICML, Atlanta, Georgia: USA, June 2013.

[34] B. He and X. Yuan, “On the $O(1/t)$ convergence rate of the Douglas-Rachford alternating direction method,” SIAM J. Numerical Analysis, vol. 50, no. 2, pp. 700–709, 2012.

[35] W. Deng, M.-J. Lai, and W. Yin, “On the $o(1/k)$ convergence and parallelization of the alternating direction method of multipliers,” 2013, arXiv:1312.3040.

[36] M. Hong and Z.-Q. Luo, “On the linear convergence of the alternating direction method of multipliers,” 2012, arXiv:1208.3922.

[37] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM J. Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.

[38] ———, “Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems,” IEEE Trans. Image Process., vol. 18, pp. 2419–2439, 2009.

[39] M. Yamagishi and I. Yamada, “Over-relaxation of the fast iterative shrinkage-thresholding algorithm with variable stepsize,” Inverse Problems, vol. 27, no. 10, 2011.

[40] M. Yamagishi, M. Yukawa, and I. Yamada, “Acceleration of adaptive proximal forward-backward splitting method and its application to sparse system identification,” in Proc. ICASSP, Prague: Czech Republic, May 22–27 2011, pp. 4296–4299.

[41] C. Hu, J. T. Kwok, and W. Pan, “Accelerated gradient methods for stochastic optimization and online learning,” in Proc. NIPS, 2009.

[42] S. Ghadimi and G. Lan, “Accelerated gradient methods for nonconvex nonlinear and stochastic programming,” Oct. 2013, arXiv:1310.3787.

[43] K. Slavakis and G. B. Giannakis, “Online dictionary learning from big data using accelerated stochastic approximation algorithms,” in Proc. ICASSP, Florence: Italy, May 2014.

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

[45] P. L. Combettes and J.-C. Pesquet, “Proximal splitting methods in signal processing,” in Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer-Verlag, 2011, pp. 185–212.

[46] J. Neveu, Mathematical Foundations of the Calculus of Probability. San Francisco: Holden-Day, 1965.

[47] A. Nemirovski, A. Juditski, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” SIAM J. Optim., vol. 19, no. 4, pp. 1574–1609, 2009.

[48] [Online]. Available: http://www.internet2.edu/observatory/