Randomly Activated Proximal Methods for Nonsmooth Convex Minimization

Patrick L. Combettes
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205, USA

Javier I. Madariaga
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205, USA

***Abstract*—We propose stochastic algorithms for solving large scale nonsmooth convex composite minimization problems. They activate at each iteration blocks of randomly selected proximity operators and achieve almost sure convergence of the iterates to a solution without any regularity assumptions. Numerical applications to data analysis problems are provided.**

***Index Terms*—Convex optimization, data analysis, proximal splitting, stochastic algorithm.**

I. Introduction

The objective of this paper is to propose stochastic algorithms with convergence guarantees on the sequence of iterates for solving the following general nonsmooth composite minimization problem, which is ubiquitous in signal processing, inverse problems, and machine learning applications (see Section II-A for notation).

Problem 1 H *is a separable real Hilbert space and* $\mathsf{f} \in \Gamma_0(\mathsf{H})$. *For every* $\mathsf{k} \in \{1,\dots,\mathsf{p}\}$, G_k *is a separable real Hilbert space,* $\mathsf{g}_\mathsf{k} \in \Gamma_0(\mathsf{G}_\mathsf{k})$, *and* $\mathsf{L}_\mathsf{k}\colon \mathsf{H} \to \mathsf{G}_\mathsf{k}$ *is linear and bounded. It is assumed that* $\mathsf{Z} = \operatorname{Argmin}(\mathsf{f}+\sum_{\mathsf{k}=1}^{\mathsf{p}} \mathsf{g}_\mathsf{k}\circ\mathsf{L}_\mathsf{k}) \neq \varnothing$. *The task is to*

$$\underset{\mathsf{x}\in\mathsf{H}}{\operatorname{minimize}} \;\; \mathsf{f}(\mathsf{x}) + \sum_{\mathsf{k}=1}^{\mathsf{p}} \mathsf{g}_\mathsf{k}(\mathsf{L}_\mathsf{k}\mathsf{x}). \tag{1}$$

Various deterministic proximal splitting methods are available to solve Problem 1, most of which require the activation of the proximity operators of the $\mathsf{p}+1$ functions f and $(\mathsf{g}_\mathsf{k})_{1\leqslant\mathsf{k}\leqslant\mathsf{p}}$ at each iteration [3], [5]. Our specific focus is on solving Problem 1 in instances when p is large, which makes it necessary to activate only a small number of proximity operators at each iteration. In this context, we aim at designing efficient stochastic proximal splitting algorithms with the following features:

- **F1:** They guarantee the convergence of the sequence of iterates to a solution to Problem 1 (not just objective function convergence or ergodic convergence) without any additional assumptions on the functions, the linear operators, or the underlying spaces.
- **F2:** At each iteration, they activate only a block of randomly selected proximity operators of $(\mathsf{f}, \mathsf{g}_1, \dots, \mathsf{g}_\mathsf{p})$.
- **F3:** At each iteration, more than one randomly selected proximity operators of $(\mathsf{f}, \mathsf{g}_1, \dots, \mathsf{g}_\mathsf{p})$ can be activated.
- **F4:** Knowledge of the norms of the linear operators is not required.

There is a vast literature on random activation algorithms in the special case of minimizing a sum of smooth functions $\sum_{\mathsf{k}=1}^{\mathsf{p}} \mathsf{g}_\mathsf{k}$ in $\mathsf{H} = \mathbb{R}^\mathsf{N}$ via so-called stochastic gradient descent [8]. The minimization of $\sum_{\mathsf{k}=1}^{\mathsf{p}} \mathsf{g}_\mathsf{k}$ when the functions are Lipschitzian is considered in [7], [12], [13]; they do not satisfy **F1** and **F3**. In [6] the convergence analysis requires a strong convexity assumption, and finite-dimensional spaces, which violates **F1** and **F4**. On the other hand, [9] addresses the constrained minimization of a smooth function under regularity conditions in $\mathsf{H} = \mathbb{R}^\mathsf{N}$. The work of [10], which employs the stochastic quasi-Fejér framework of [4], considers special cases of (1) and does not guarantee **F1**. The only random activation frameworks that address Problem 1 in its generality and guarantee **F1** seem to be those of [11], which is based on [4, Remark 5.10(iv)], and [2]. However, these primal-dual renorming approaches do not satisfy **F4** and [2] does not satisfy **F3**. We propose three frameworks based on results of [4] which lead to simple and efficient algorithms for solving Problem 1 that satisfy the requirements **F1–F4** above. These novel algorithms are presented in Section II and applied to support vector machine and classification problems in Section III.

II. Proposed algorithms

A. Notation

Throughout, H is a separable real Hilbert space with identity operator Id, scalar product $\langle\cdot\,|\,\cdot\rangle$, and associated norm $\|\cdot\|$. $\Gamma_0(\mathsf{H})$ denotes the class of lower semicontinuous convex functions $\mathsf{f}\colon \mathsf{H} \to\,]-\infty,+\infty]$ such that $\operatorname{dom}\mathsf{f} = \big\{\mathsf{x}\in\mathsf{H} \mid \mathsf{f}(\mathsf{x}) < +\infty\big\} \neq \varnothing$. Let C be a nonempty closed convex subset of H. Then ι_C denotes the indicator function of C and $\operatorname{proj}_\mathsf{C}$ the projection operator onto C. Let $\mathsf{f} \in \Gamma_0(\mathsf{H})$. The subdifferential of f at $\mathsf{x} \in \mathsf{H}$ is the set

This work was supported by the National Science Foundation under grant CCF-2211123.

$\partial f(x) = \big\{u \in H \mid (\forall z \in H)\ \langle z - x \,|\, u\rangle + f(x) \leqslant f(z)\big\}$ and the proximity operator of f is

$$\operatorname{prox}_f \colon H \to H \colon x \mapsto \operatorname*{argmin}_{z \in H} \left(f(z) + \frac{1}{2}\|x - z\|^2 \right). \quad (2)$$

We refer to [1] for background on convex analysis and optimization. The underlying probability space is $(\Omega, \mathcal{F}, P)$ and $\mathcal{B}_H$ denotes the Borel σ-algebra of H. An H-valued random variable is a measurable mapping $x \colon (\Omega, \mathcal{F}) \to (H, \mathcal{B}_H)$. The σ-algebra generated by a family Φ of random variables is denoted by $\sigma(\Phi)$. Given $x \colon \Omega \to H$ and $A \subset H$, we set $[x \in A] = \big\{\omega \in \Omega \mid x(\omega) \in A\big\}$.

B. General framework

Our approach consists in embedding Problem 1 into multivariate problems that have the following general form studied in [4].

Problem 2 *Let $(X_i)_{1 \leqslant i \leqslant m}$ and $(Y_k)_{1 \leqslant k \leqslant r}$ be families of separable real Hilbert spaces with direct Hilbert sums $\mathbf{X} = X_1 \oplus \cdots \oplus X_m$ and $\mathbf{Y} = Y_1 \oplus \cdots \oplus Y_r$. For every $i \in \{1, \ldots, m\}$, let $f_i \in \Gamma_0(X_i)$ and, for every $k \in \{1, \ldots, r\}$, let $h_k \in \Gamma_0(Y_k)$, and let $L_{ki} \colon X_i \to Y_k$ be linear and bounded. It is assumed that there exists $\mathbf{u} \in \mathbf{X}$ such that*

$$(\forall i \in \{1, \ldots, m\})\ 0 \in \partial f_i(u_i) + \sum_{k=1}^{r} L_{ki}^* \left(\partial h_k \left(\sum_{j=1}^{m} L_{kj} u_j \right) \right). \quad (3)$$

The task is to

$$\operatorname*{minimize}_{x \in \mathbf{X}} \ \sum_{i=1}^{m} f_i(x_i) + \sum_{k=1}^{r} h_k \left(\sum_{i=1}^{m} L_{ki} x_i \right) \quad (4)$$

and $\mathbf{Z}$ denotes the set of solutions. Further, the projection operator onto the subspace

$$\mathbf{V} = \left\{ (x, y) \in \mathbf{X} \oplus \mathbf{Y} \;\middle|\; (\forall k \in \{1, \ldots, r\})\ y_k = \sum_{i=1}^{m} L_{ki} x_i \right\} \quad (5)$$

is decomposed as $\operatorname{proj}_{\mathbf{V}} \colon (x, y) \mapsto (Q_j(x, y))_{1 \leqslant j \leqslant m+r}$, where for every $i \in \{1, \ldots, m\}$, $Q_i \colon \mathbf{X} \oplus \mathbf{Y} \to X_i$ and, for every $k \in \{1, \ldots, r\}$, $Q_{m+k} \colon \mathbf{X} \oplus \mathbf{Y} \to Y_k$.

Theorem 3 [4, Corollary 5.5] *Consider the setting of Problem 2. Set $D = \{0,1\}^{m+r} \smallsetminus \{\mathbf{0}\}$, let $\gamma \in \left]0, +\infty\right[$, let $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence in $\left]0, 2\right[$ such that $\inf_{n \in \mathbb{N}} \lambda_n > 0$ and $\sup_{n \in \mathbb{N}} \lambda_n < 2$, let x_0 and z_0 be $\mathbf{X}$-valued random variables, let y_0 and w_0 be $\mathbf{Y}$-valued random variables, and let $(\varepsilon_n)_{n \in \mathbb{N}}$ be identically distributed D-valued random variables. Iterate*

for $n = 0, 1, \ldots$
  for $i = 1, \ldots, m$
    $x_{i,n+1} = x_{i,n} + \varepsilon_{i,n}\big(Q_i(z_n, w_n) - x_{i,n}\big)$
    $z_{i,n+1} = z_{i,n} + \varepsilon_{i,n}\lambda_n\big(\operatorname{prox}_{\gamma f_i}(2x_{i,n+1} - z_{i,n}) - x_{i,n+1}\big)$
  for $k = 1, \ldots, r$
    $y_{k,n+1} = y_{k,n} + \varepsilon_{m+k,n}\big(Q_{m+k}(z_n, w_n) - y_{k,n}\big)$
    $w_{k,n+1} = w_{k,n} + \varepsilon_{m+k,n}\lambda_n\big(\operatorname{prox}_{\gamma h_k}(2y_{k,n+1} - w_{k,n}) - y_{k,n+1}\big)$.

In addition, assume that the following are satisfied:

(i) *For every $n \in \mathbb{N}$, $\sigma(\varepsilon_n)$ and $\sigma(z_j, w_j)_{0 \leqslant j \leqslant n}$ are independent.*
(ii) *For every $j \in \{1, \ldots, m + r\}$, $P[\varepsilon_{j,0} = 1] > 0$.*

Then $(x_n)_{n \in \mathbb{N}}$ converges weakly P-a.s. to a $\mathbf{Z}$-valued random variable.

Remark 4 The random variables $\varepsilon_{i,n}$ and $\varepsilon_{m+k,n}$ control which components are updated.

We now present three frameworks for solving Problem 1 which are based on specializations of Theorem 3. We define $\mathbf{G} = G_1 \oplus \cdots \oplus G_p$ and

$$\mathbf{W} = \left\{ x \in H \oplus \mathbf{G} \;\middle|\; (\forall k \in \{1, \ldots, p\})\ x_{k+1} = L_k x_1 \right\}. \quad (6)$$

C. Framework 1

We start with the following reformulation of Problem 1.

Problem 5 *Consider the setting of Problem 1. Set $f_1 = f$ and, for every $i \in \{2, \ldots, p+1\}$, $f_i = g_{i-1}$. Denote by $\mathbf{x} = (x_1, \ldots, x_{p+1})$ a generic element in $H \oplus \mathbf{G}$. The task is to*

$$\operatorname*{minimize}_{\mathbf{x} \in H \oplus \mathbf{G}} \ \sum_{i=1}^{p+1} f_i(x_i) + \iota_{\mathbf{W}}(\mathbf{x}). \quad (7)$$

We observe that Problem 5 is a special case of Problem 2 in which $m = p + 1$, $r = 1$, $X_1 = H$, $(X_i)_{2 \leqslant i \leqslant m} = (G_{i-1})_{2 \leqslant i \leqslant m}$, $Y_1 = \mathbf{X}$, $h_1 = \iota_{\mathbf{W}}$, and, for every $i \in \{1, \ldots, m\}$, $L_{1i} \colon x_i \mapsto (z_1, \ldots, z_m)$, where $z_j = 0$ if $j \neq i$, and $z_j = x_i$ if $j = i$. Altogether, Problem 1 is an instance of Problem 2 and we apply Theorem 3 to solve it as follows.

Proposition 6 *Consider the setting of Problem 1. Set $D = \{0,1\}^{p+2} \smallsetminus \{\mathbf{0}\}$, let $\gamma \in \left]0, +\infty\right[$, let $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence in $\left]0, 2\right[$ such that $\inf_{n \in \mathbb{N}} \lambda_n > 0$ and $\sup_{n \in \mathbb{N}} \lambda_n < 2$, let x_0, z_0, y_0, and w_0 be $H \oplus \mathbf{G}$-valued random variables, and let $(\varepsilon_n)_{n \in \mathbb{N}}$ be identically distributed D-valued random variables. Set $x_0 = x_{1,0}$ and iterate*

for $n = 0, 1, \ldots$
  $x_{n+1} = x_n + \varepsilon_{1,n}\big(\frac{1}{2}z_{1,n} + \frac{1}{2}w_{1,n} - x_n\big)$
  $z_{1,n+1} = z_{1,n} + \varepsilon_{1,n}\lambda_n\big(\operatorname{prox}_{\gamma f}(2x_{n+1} - z_{1,n}) - x_{n+1}\big)$
  for $k = 1, \ldots, p$
    $x_{k+1,n+1} = x_{k+1,n} + \varepsilon_{k+1,n}\big(\frac{1}{2}z_{k+1,n} + \frac{1}{2}w_{k+1,n} - x_{k+1,n}\big)$
    $z_{k+1,n+1} = z_{k+1,n} + \varepsilon_{k+1,n}\lambda_n\big(\operatorname{prox}_{\gamma g_k}(2x_{k+1,n+1} - z_{k+1,n}) - x_{k+1,n+1}\big)$
  $y_{n+1} = y_n + \varepsilon_{p+2,n}\big(\frac{1}{2}z_n + \frac{1}{2}w_n - y_n\big)$
  $q_n = (\mathrm{Id} + \sum_{k=1}^{p} L_k^* \circ L_k)^{-1}\big(2y_{1,n+1} - w_{1,n} + \sum_{k=1}^{p} L_k^*(2y_{k+1,n+1} - w_{k+1,n})\big)$
  $w_{1,n+1} = w_{1,n} + \varepsilon_{p+2,n}\lambda_n(q_n - y_{1,n+1})$
  for $k = 1, \ldots, p$
    $w_{k+1,n+1} = w_{k+1,n} + \varepsilon_{p+2,n}\lambda_n(L_k q_n - y_{k+1,n+1})$.

In addition, assume that the following are satisfied:

(i) $(\exists\, u \in H)\ 0 \in \partial f(u) + \sum_{k=1}^{p} L_k^*(\partial g_k(L_k u))$.
(ii) *For every* $n \in \mathbb{N}$, $\sigma(\varepsilon_n)$ *and* $\sigma(z_j, w_j)_{0\leqslant j\leqslant n}$ *are independent.*
(iii) *For every* $j \in \{1, \ldots, p+2\}$, $\mathrm{P}[\varepsilon_{j,0} = 1] > 0$.

Then $(x_n)_{n\in\mathbb{N}}$ *converges weakly* P-a.s. *to a* Z*-valued random variable.*

D. Framework 2

Here is an alternative reformulation of Problem 1.

Problem 7 *Consider the setting of Problem 1 and let* $(f_i)_{1\leqslant i\leqslant p+1}$ *be as in Problem 5. Let* $(K_k)_{1\leqslant k\leqslant r}$ *be separable real Hilbert spaces, set* $\mathbf{K} = \bigoplus_{k=1}^{r} K_k$*, and let*

$$\mathbf{C}\colon H \oplus \mathbf{G} \to \mathbf{K}\colon x \mapsto \left(\sum_{i=1}^{p+1} C_{ki} x_i\right)_{1\leqslant k\leqslant r} \tag{8}$$

be linear and bounded with $\ker \mathbf{C} = \mathbf{W}$*. The task is to*

$$\underset{x\in H\oplus\mathbf{G}}{\text{minimize}} \ \sum_{i=1}^{p+1} f_i(x_i) + \sum_{k=1}^{r} \iota_{\{0\}}\left(\sum_{i=1}^{p+1} C_{ki} x_i\right). \tag{9}$$

We observe that Problem 7 is the special case of Problem 2 in which $m = p+1$, $X_1 = H$, $(X_i)_{2\leqslant i\leqslant m} = (G_{i-1})_{2\leqslant i\leqslant m}$, $\mathbf{Y} = \mathbf{K}$, for every $k \in \{1, \ldots, r\}$, $h_k = \iota_{\{0\}}$, and, for every $i \in \{1, \ldots, m\}$, $L_{ki} = C_{ki}$. Thus, the subspace $\mathbf{V}$ of (5) becomes

$$\mathbf{V} = \left\{(x, y) \in \mathbf{X} \oplus \mathbf{Y} \;\middle|\; (\forall k \in \{1, \ldots, r\})\ y_k = \sum_{i=1}^{p+1} C_{ki} x_i\right\}, \tag{10}$$

which confirms that Problem 1 is an instance of Problem 7. In turn, we apply Theorem 3 to solve it as follows.

Proposition 8 *Consider the setting of Problem 1. Let* $\mathbf{K}$ *and* $\mathbf{C}$ *be as in Problem 7, set* $\mathbf{V}$ *be as in* (10)*, and decompose its projection operator as* $\mathrm{proj}_{\mathbf{V}}\colon x \mapsto (R_j x)_{1\leqslant j\leqslant p+1+r}$*, where* $R_1\colon H\oplus\mathbf{G}\oplus\mathbf{K} \to H$, $(\forall i \in \{1, \ldots, p\})\ R_{1+i}\colon H\oplus\mathbf{G}\oplus\mathbf{K} \to G_i$*, and* $(\forall k \in \{1, \ldots, r\})\ R_{p+1+k}\colon H \oplus \mathbf{G} \oplus \mathbf{K} \to K_k$*. Set* $D = \{0,1\}^{p+1+r} \smallsetminus \{\mathbf{0}\}$*, let* $\gamma \in \left]0, +\infty\right[$*, let* $(\lambda_n)_{n\in\mathbb{N}}$ *be a sequence in* $\left]0, 2\right[$ *such that* $\inf_{n\in\mathbb{N}} \lambda_n > 0$ *and* $\sup_{n\in\mathbb{N}} \lambda_n < 2$*, let* x_0 *and* z_0 *be* $H\oplus\mathbf{G}$*-valued random variables, let* y_0 *and* w_0 *be* $\mathbf{K}$*-valued random variables, and let* $(\varepsilon_n)_{n\in\mathbb{N}}$ *be identically distributed* D*-valued random variables. Set* $x_0 = x_{1,0}$ *and iterate*

for $n = 0, 1, \ldots$
  $x_{n+1} = x_n + \varepsilon_{1,n}\big(R_1(z_n, w_n) - x_n\big)$
  $z_{1,n+1} = z_{1,n} + \varepsilon_{1,n}\lambda_n\big(\mathrm{prox}_{\gamma f}\,(2x_{n+1} - z_{1,n}) - x_{n+1}\big)$
  for $k = 1, \ldots, p$
    $x_{k+1,n+1} = x_{k+1,n} + \varepsilon_{k+1,n}\big(R_{k+1}(z_n, w_n) - x_{k+1,n}\big)$
    $z_{k+1,n+1} = z_{k+1,n} +$
      $\varepsilon_{k+1,n}\lambda_n\big(\mathrm{prox}_{\gamma g_k}\,(2x_{k+1,n+1} - z_{k,n}) - x_{k+1,n+1}\big)$
  for $k = 1, \ldots, r$
    $y_{k,n+1} = y_{k,n} + \varepsilon_{p+1+k,n}\big(R_{p+1+k}(z_n, w_n) - y_{k,n}\big)$
    $w_{k,n+1} = w_{k,n} - \varepsilon_{p+1+k,n}\lambda_n y_{k,n+1}$.

In addition, assume that the following are satisfied:

(i) *There exists* $u \in \ker \mathbf{C}$ *such that* $0 \in \partial f(u_1) + \sum_{k=1}^{r} \mathrm{range}\, C_{k1}^*$ *and* $(\forall j \in \{1, \ldots, p\})\ 0 \in \partial g_j(u_{j+1}) + \sum_{k=1}^{r} \mathrm{range}\, C_{kj}^*$.
(ii) *For every* $n \in \mathbb{N}$, $\sigma(\varepsilon_n)$ *and* $\sigma(z_j, w_j)_{0\leqslant j\leqslant n}$ *are independent.*
(iii) *For every* $j \in \{1, \ldots, p+1+r\}$, $\mathrm{P}[\varepsilon_{j,0} = 1] > 0$.

Then $(x_n)_{n\in\mathbb{N}}$ *converges weakly* P-a.s. *to a* Z*-valued random variable.*

E. Framework 3

The third approach stems from the observation that Problem 1 coincides with (4) for $m = 1$, $r = p$, $X_1 = H$, $f_1 = f$, and $(\forall k \in \{1, \ldots, p\})\ Y_k = G_k$, $L_{k,1} = L_k$, and $h_k = g_k$. We therefore derive from Theorem 3 the following convergence result.

Proposition 9 *Consider the setting of Problem 1. Set* $D = \{0,1\}^{1+p} \smallsetminus \{\mathbf{0}\}$*, let* $\gamma \in \left]0, +\infty\right[$*, let* $(\lambda_n)_{n\in\mathbb{N}}$ *be a sequence in* $\left]0, 2\right[$ *such that* $\inf_{n\in\mathbb{N}} \lambda_n > 0$ *and* $\sup_{n\in\mathbb{N}} \lambda_n < 2$*, let* x_0 *and* z_0 *be* H*-valued random variables, let* y_0 *and* w_0 *be* $\mathbf{G}$*-valued random variables, and let* $(\varepsilon_n)_{n\in\mathbb{N}}$ *be identically distributed* D*-valued random variables. Iterate*

for $n = 0, 1, \ldots$
  $q_n = (\mathrm{Id} + \sum_{k=1}^{p} L_k^* \circ L_k)^{-1}\big(z_n + \sum_{k=1}^{p} L_k^* w_{k,n}\big)$
  $x_{n+1} = x_n + \varepsilon_{1,n}(q_n - x_n)$
  $z_{n+1} = z_n + \varepsilon_{1,n}\lambda_n\big(\mathrm{prox}_{\gamma f}\,(2x_{n+1} - z_n) - x_{n+1}\big)$
  for $k = 1, \ldots, p$
    $y_{k,n+1} = y_{k,n} + \varepsilon_{1+k,n}\big(L_k q_n - y_{k,n}\big)$
    $w_{k,n+1} = w_{k,n} +$
      $\varepsilon_{1+k,n}\lambda_n\big(\mathrm{prox}_{\gamma g_k}\,(2y_{k,n+1} - w_{k,n}) - y_{k,n+1}\big)$.

In addition, assume that the following are satisfied:

(i) $(\exists\, u \in H)\ 0 \in \partial f(u) + \sum_{k=1}^{p} L_k^*(\partial g_k(L_k u))$.
(ii) *For every* $n \in \mathbb{N}$, $\sigma(\varepsilon_n)$ *and* $\sigma(z_j, w_j)_{0\leqslant j\leqslant n}$ *are independent.*
(iii) *For every* $j \in \{1, \ldots, p+1\}$, $\mathrm{P}[\varepsilon_{j,0} = 1] > 0$.

Then $(x_n)_{n\in\mathbb{N}}$ *converges weakly* P-a.s. *to a* Z*-valued random variable.*

F. Examples

We provide some examples of operators arising in Propositions 6, 8, and 9.

Example 10 In Proposition 8, set $r = p$, $\mathbf{K} = \mathbf{G}$, and, for every $k \in \{1, \ldots, p\}$ and every $i \in \{1, \ldots, p+1\}$,

$$C_{ki} = \begin{cases} L_k, & \text{if } i = 1; \\ -\mathrm{Id}, & \text{if } i = k+1; \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Let $\mathbf{x} \in \mathsf{H} \oplus \mathbf{G}$, $\mathbf{y} \in \mathbf{G}$, and set $\mathsf{q} = (2\mathrm{Id} + \sum_{k=1}^{p} \mathsf{L}_k^* \circ \mathsf{L}_k)^{-1}(2\mathsf{x}_1 + \sum_{k=1}^{p} \mathsf{L}_k^*(\mathsf{x}_{k+1} + \mathsf{y}_k))$. Then, for every $i \in \{1, \ldots, p+1\}$, $\mathsf{R}_i(\mathbf{x}, \mathbf{y})$ is given by

$$\begin{cases} \mathsf{q}, & \text{if } i = 1; \\ \dfrac{1}{2}\big(\mathsf{L}_{i-1}\mathsf{q} + \mathsf{x}_i - \mathsf{y}_{i-1}\big), & \text{if } 2 \leqslant i \leqslant p+1; \\ \dfrac{1}{2}\big(\mathsf{L}_{i-p-1}\mathsf{q} - \mathsf{x}_{i-p} + \mathsf{y}_{i-p-1}\big), & \text{if } p+2 \leqslant i \leqslant 2p+1. \end{cases} \tag{12}$$

The next examples focus on the special case of Problem 1 in which, for every $k \in \{1, \ldots, p\}$, $\mathsf{G}_k = \mathsf{H}$ and $\mathsf{L}_k = \mathrm{Id}$, that is,

$$\underset{x \in \mathsf{H}}{\text{minimize}} \;\; \mathsf{f}(x) + \sum_{k=1}^{p} \mathsf{g}_k(x). \tag{13}$$

Example 11 In Example 10, for every $\mathbf{x} \in \mathsf{H}^{p+1}$ and every $\mathbf{y} \in \mathsf{H}^p$,

$$\mathsf{q} = \frac{1}{p+2}\left(2\mathsf{x}_1 + \sum_{k=1}^{p}(\mathsf{x}_{k+1} + \mathsf{y}_k)\right). \tag{14}$$

Example 12 In Proposition 8, set $\mathbf{C}$ such that, for every $k \in \{1, \ldots, p+1\}$ and every $i \in \{1, \ldots, p+1\}$,

$$\mathsf{C}_{ki} = \begin{cases} \dfrac{p}{p+1}\mathrm{Id}, & \text{if } k = i; \\ -\dfrac{1}{p+1}\mathrm{Id}, & \text{if } k \neq i. \end{cases} \tag{15}$$

Then $\ker \mathbf{C}$ is the subspace of all the vectors $\mathbf{x} \in \mathsf{H}^{p+1}$ such that, for every $i \in \{1, \ldots, p+1\}$, $\mathsf{x}_i = \frac{1}{p+1}\sum_{j=1}^{p+1} \mathsf{x}_j$. Hence, for every $i \in \{1, \ldots, 2p+2\}$, $\mathbf{x} \in \mathsf{H}^{p+1}$, and $\mathbf{y} \in \mathsf{H}^{p+1}$, $\mathsf{R}_i(\mathbf{x}, \mathbf{y})$ is given by

$$\begin{cases} \dfrac{\mathsf{x}_i + \mathsf{y}_i}{2} + \dfrac{1}{2(p+1)}\displaystyle\sum_{j=1}^{p+1}(\mathsf{x}_j - \mathsf{y}_j), & \text{if } i \leqslant p+1; \\ \dfrac{\mathsf{x}_i + \mathsf{y}_i}{2} - \dfrac{1}{2(p+1)}\displaystyle\sum_{j=1}^{p+1}(\mathsf{x}_j + \mathsf{y}_j), & \text{if } p+2 \leqslant i \leqslant 2p+2. \end{cases} \tag{16}$$

Example 13 In Propositions 6 and 9, $(\mathrm{Id} + \sum_{k=1}^{p} \mathsf{L}_k^* \circ \mathsf{L}_k)^{-1}$ is just $(p+1)^{-1}\mathrm{Id}$.

Remark 14 On the one hand, the operator $\mathbf{C}$ in Example 11 applied to $\mathbf{x} \in \mathsf{H}^{p+1}$ couples, for every $i \in \{2, \ldots, p+1\}$, x_i with x_1. On the other hand, in Example 12 the operator $\mathbf{C}$ applied to $\mathbf{x} \in \mathsf{H}^{p+1}$ couples, for every $i \in \{1, \ldots, p+1\}$, x_i with the average $\frac{1}{p+1}\sum_{j=1}^{p+1} \mathsf{x}_j$. Various alternative coupling operators $\mathbf{C}$ can be considered to enforce the condition $\mathsf{x}_1 = \cdots = \mathsf{x}_{p+1}$.

III. Numerical experiments

We present two experiments to compare the numerical behavior of the algorithms presented in Section II. The qualification condition (3) is satisfied in all cases.

A. Experiment 1: Overlapping group lasso regression

We address the overlapping group lasso regression problem of [14]. Here $\mathsf{H} = \mathbb{R}^N$ and, for every $k \in \{1, \ldots, p\}$, $\varnothing \neq \mathsf{I}_k \subset \{1, \ldots, N\}$ and

$$\mathsf{L}_k \colon \mathbb{R}^N \to \mathbb{R}^{\operatorname{card} \mathsf{I}_k} \colon x = (\xi_j)_{1 \leqslant j \leqslant N} \mapsto (\xi_j)_{j \in \mathsf{I}_k}. \tag{17}$$

Further, $\bigcup_{k=1}^{p} \mathsf{I}_k = \{1, \ldots, N\}$. The goal is to

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \frac{\alpha}{2}\|Ax - b\|^2 + \frac{1}{p}\sum_{k=1}^{p} \|\mathsf{L}_k x\|, \tag{18}$$

where $\mathsf{A} \in \mathbb{R}^{M \times N}$, $\mathsf{b} \in \mathbb{R}^M$, and $\alpha \in \left]0, +\infty\right[$. In the experiment $M = 1000$, $N = 3610$, $p = 40$, and, as in [14], $\alpha = 5/p^2$. The entries of A are i.i.d. samples from a $\mathcal{N}(0,1)$ distribution, and the entries of b are i.i.d. samples from a $\mathcal{N}(100, 100)$ distribution. Finally,

$$(\forall k \in \{1, \ldots, p\}) \;\; \mathsf{I}_k = \{90k - 89, \ldots, 90k + 10\}. \tag{19}$$

We employ the three frameworks of Sections II-C–II-E to solve (18), where Proposition 8 uses the operator $\mathbf{C}$ defined in Example 10. In each case, $\gamma = p$, the initial points $\boldsymbol{x}_0$, $\boldsymbol{z}_0$, $\boldsymbol{y}_0$, and $\boldsymbol{w}_0$ are set to $\mathbf{0}$, and, for every $n \in \mathbb{N}$, $\lambda_n = 1.9$. The random variable ε_0 activates a single index in $\{1, \ldots, p+2\}$, $\{1, \ldots, 2p+1\}$, and $\{1, \ldots, p+1\}$ respectively, and the distribution is uniform. We display in Fig. 1 the normalized error versus execution time.

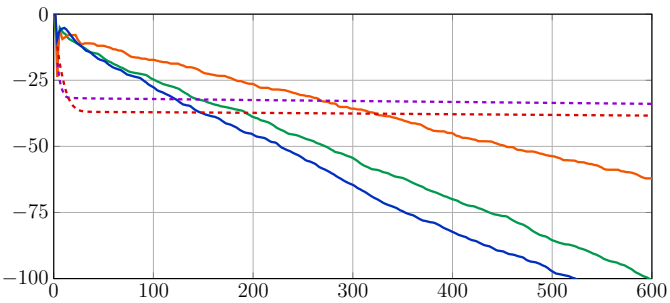


Fig. 1: Normalized error $20\log(\|x_{1,n} - x_\infty\|/\|x_{1,0} - x_\infty\|)$ (dB) versus execution time (s) in Experiment 1. Orange: Framework 1. Blue: Framework 2 with Example 12. Green: Framework 3. Dashed red: Algorithm of [11]. Dashed violet: Algorithm of [2].

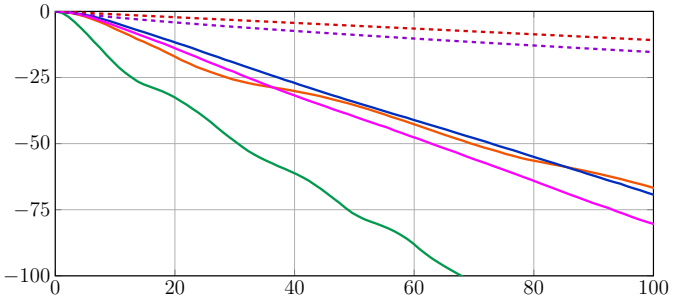
B. Experiment 2: Classification using the hinge loss

We address a binary classification problem. The training data set $((u_k, \xi_k))_{1 \leqslant k \leqslant p}$ is in $\mathbb{R}^N \times \{-1, 1\}$ and the goal is to learn a linear classifier $x \in \mathsf{H} = \mathbb{R}^N$. For this purpose, we solve the support vector machine model

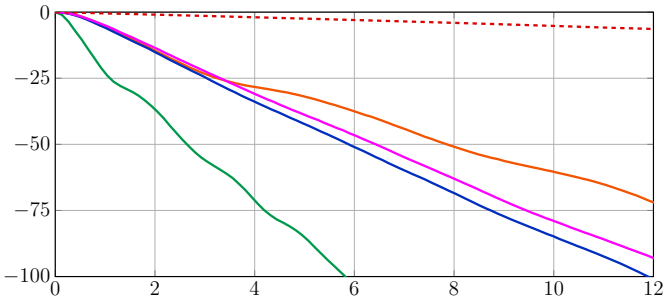
$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \frac{\alpha}{2}\|x\|^2 + \frac{1}{p}\sum_{k=1}^{p} \mathsf{g}_k(x), \tag{20}$$

where $\alpha \in \left]0, +\infty\right[$ and, for every $k \in \{1, \ldots, p\}$,

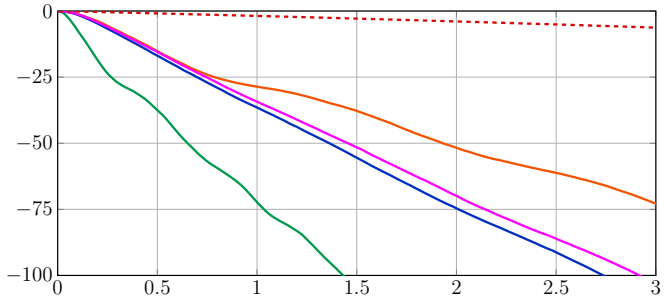
$$\mathsf{g}_k \colon x \mapsto \max\{0, 1 - \xi_k \langle x \,|\, u_k \rangle\}. \tag{21}$$



(a)



(b)



(c)

Fig. 2: Normalized error $20\log(\|x_{1,\mathsf{n}} - x_\infty\|/\|x_{1,0} - x_\infty\|)$ (dB) versus execution time (s) in Experiment 2. (a): Block size 1 with 1 core. (b): Block size 8 with 8 cores. (c): Block size 32 with 32 cores. Orange: Framework 1. Blue: Framework 2 with Example 11. Magenta: Framework 2 with Example 12. Green: Framework 3. Dashed red: Algorithm of [11]. Dashed violet in (a): Algorithm of [2].

In the experiment, $\mathsf{N} = 1000$, $\alpha = 1$, $\mathsf{p} = 500$, and, for every $\mathsf{k} \in \{1, \ldots, \mathsf{p}\}$, the entries of u_k are i.i.d. samples from a $\mathcal{N}(100, 10)$ distribution, and the entries of ξ_k are i.i.d. samples from a uniform distribution on $\{-1, 1\}$. We employ four methods to solve this problem: Framework 1, Framework 2 using the operators $\mathbf{C}$ defined in Example 11 and Example 12, and Framework 3. In each case, $\gamma = 1$, the initial points $\boldsymbol{x}_0$, $\boldsymbol{z}_0$, $\boldsymbol{y}_0$, and $\boldsymbol{w}_0$ are set to $\mathbf{0}$ and, for every $\mathsf{n} \in \mathbb{N}$, $\lambda_\mathsf{n} = 1.9$. We run three instances of the algorithms. In the first one, the random variable ε_n activates one index uniformly in $\{1, \ldots, \mathsf{p}+2\}$, $\{1, \ldots, 2\mathsf{p}+1\}$, $\{1, \ldots, 2\mathsf{p}+2\}$, and $\{1, \ldots, \mathsf{p}+1\}$ respectively. In the second, the number of activated indices is 8, and in the third it is 32. We display in Fig. 2 the normalized error versus execution time for each instances. The execution time is evaluated based on the assumption that the computation corresponding to each selected index is assigned a dedicated core and that all the cores are working in parallel.

C. Discussion

As seen in Section I, the only comparable existing algorithms are those of [2] and [11]. In Experiment 1 they display a fast initial behavior but then progress very slowly compared to the proposed Frameworks 1–3. In Experiment 2 they are consistently slower than Frameworks 1–3. Note that the algorithm of [2] is just compared in Fig. 2(a), where only one index is activated since it does not satisfy **F3**. In terms of storage, Framework 1 stores $2\mathsf{p}+1$ variables, Framework 2 stores $\mathsf{p}+\mathsf{r}+1$, and Framework 3 stores $\mathsf{p}+1$. An advantage of Framework 2 is that in the last r activations no proximal calculations are needed. In general, the execution time depends on the computational load associated with the evaluation of the proximity operators and the inversions. For instance, these are cheaper in Experiment 2, which makes Framework 3 the fastest.

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. Springer, New York, 2017.

[2] A. Chambolle, C. Delplancke, M. J. Ehrhardt, C.-B. Schönlieb, and J. Tang, Stochastic primal-dual hybrid gradient algorithm with adaptive step sizes, *J. Math. Imaging Vision*, vol. 66, pp. 294–313, 2024.

[3] P. L. Combettes, The geometry of monotone operator splitting methods, *Acta Numer.*, vol. 33, pp. 487–632, 2024.

[4] P. L. Combettes and J.-C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.*, vol. 25, pp. 1221–1248, 2015.

[5] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists, *SIAM Rev.*, vol. 65, pp. 375–435, 2023.

[6] L. Condat and P. Richtárik, Randprox: Primal-dual optimization algorithms with randomized proximal updates, *Proc. Int. Conf. Learn. Represent.* Kigali, Rwanda, May 1–5, 2023.

[7] D. Davis, Variance reduction for root-finding problems, *Math. Program.*, vol. A197, pp. 375–410, 2023.

[8] A. Dieuleveut, G. Fort, E. Moulines, and H.-T. Wai, Stochastic approximation beyond gradient for signal processing and machine learning, *IEEE Trans. Signal Process.*, vol. 71, pp. 3117–3148, 2023.

[9] A. Nedić, Random algorithms for convex minimization problems, *Math. Program.*, vol. B129, pp. 225–253, 2011.

[10] Z. Peng, Y. Xu, M. Yan, and W. Yin, ARock: An algorithmic framework for asynchronous parallel coordinate updates, *SIAM J. Sci. Comput.*, vol. 38, pp. A2851–A2879, 2016.

[11] J.-C. Pesquet and A. Repetti, A class of randomized primal-dual algorithms for distributed optimization, *J. Nonlinear Convex Anal.*, vol. 16, pp. 2453–2490, 2015.

[12] Q. Tran-Dinh, N. H. Pham, D. Phan, and L. Nguyen, FedDR–randomized Douglas–Rachford splitting algorithms for nonconvex federated composite optimization, *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 30326–30338, 2021.

[13] C. Traoré, V. Apidopoulos, S. Salzo, and S. Villa, Variance reduction techniques for stochastic proximal point algorithms, arxiv, 2023. https://arxiv.org/pdf/2308.09310

[14] Y.-L. Yu, Better approximation and faster algorithm using the proximal average, *Proc. Conf. Adv. Neural Inform. Process. Syst.*, pp. 458–466, 2013.