Random extrapolation for primal-dual coordinate descent

Ahmet Alacaoglu [1]   Olivier Fercoq [2]   Volkan Cevher [1]

Abstract

We introduce a randomly extrapolated primal-dual coordinate descent method that adapts to sparsity of the data matrix and the favorable structures of the objective function. Our method updates only a subset of primal and dual variables with sparse data, and it uses large step sizes with dense data, retaining the benefits of the specific methods designed for each case. In addition to adapting to sparsity, our method attains fast convergence guarantees in favorable cases *without any modifications*. In particular, we prove linear convergence under metric subregularity, which applies to strongly convex-strongly concave problems and piecewise linear quadratic functions. We show almost sure convergence of the sequence and optimal sublinear convergence rates for the primal-dual gap and objective values, in the general convex-concave case. Numerical evidence demonstrates the state-of-the-art empirical performance of our method in sparse and dense settings, matching and improving the existing methods.

1. Introduction

In this paper, we consider the problem

$$\min_{x \in \mathcal{X}} f(x) + g(x) + h(Ax), \qquad (1)$$

where $f, g \colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $h \colon \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are proper, lower semicontinuous, convex functions, $A \colon \mathcal{X} \to \mathcal{Y}$ is a linear operator. $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces such that $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$, and $\mathcal{Y} = \prod_{j=1}^{m} \mathcal{Y}_j$. Moreover, f is assumed to have coordinatewise Lipschitz continuous gradients and g, h admit easily computable proximal operators.

Problem (1) is a general template that covers many problems in different fields, such as regularized empirical risk minimization (Shalev-Shwartz & Zhang, 2013; Zhang & Xiao,

2017), optimization with large number of constraints (Patrascu & Necoara, 2017; Fercoq et al., 2019), and total variation (TV) regularized problems (Chambolle et al., 2018; Fercoq & Bianchi, 2019).

The classic choice for solving problem (1) is to use primal-dual methods (Chambolle & Pock, 2011; Vũ, 2013; Condat, 2013). These methods utilize the proximal operators for g, h^* and gradient of the differentiable component f. Randomized versions that we refer to as primal dual coordinate descent (PDCD), are proposed in several works (Zhang & Xiao, 2017; Dang & Lan, 2014; Gao et al., 2019; Fercoq & Bianchi, 2019; Chambolle et al., 2018; Latafat et al., 2019).

First advantage of coordinate-based methods is that they access to blocks of A and update a subset of variables, resulting in cheap per iteration costs. Moreover, they utilize larger step sizes depending on the properties of the problem in selected blocks.

Existing PDCD methods fail to retain both these advantages, as sparsity of A varies. In particular, methods that have cheap per-iteration costs with sparse A (Fercoq & Bianchi, 2019; Latafat et al., 2019), are restricted to use small step sizes with dense A. On the other hand, methods that can use large step sizes with dense A (Chambolle et al., 2018), have high per-iteration costs with sparse A.

Contributions. In this paper, we identify random extrapolation as the key to design a method that combines the benefits of the methods in two camps and propose the primal-dual method with random extrapolation and coordinate descent (PURE-CD). PURE-CD exhibits the advantages of (Fercoq & Bianchi, 2019; Latafat et al., 2019) in the sparse setting and the advantages of (Chambolle et al., 2018) in the dense setting simultaneously, achieving the best of both worlds. As PURE-CD has the favorable properties in both ends of the spectrum, it has the best performance in the regime in between: moderately sparse data. Table 1 compiles a summary for the comparison of PURE-CD and previous methods.

In addition to adapting to the sparsity of A, we prove that PURE-CD also adapts to unknown structures in the problem, and obtains linear rate of convergence, without any modifications in the step sizes. Our linear convergence results apply to strongly convex-strongly concave problems, linear programs, and problems with piecewise linear quadratic

[1]EPFL, Switzerland [2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France. Correspondence to: Ahmet Alacaoglu <ahmet.alacaoglu@epfl.ch>.

	Step sizes with dense data	per iteration cost	block-wise Lipschitz	probability law	Efficient implementation
(Chambolle et al., 2018)	$n\tau_i\sigma\|A_i\|^2 < 1$	m	N/A	arbitrary	direct[†]
(Fercoq & Bianchi, 2019)	$n^2\tau_i\sigma\|A_i\|^2 < 1$	$|J(i)|^*$	Yes	uniform	direct or duplication
(Latafat et al., 2019)	$n^2\tau_i\sigma\|A_i\|^2 < 1$	$|J(i)|^*$	No	arbitrary	duplication[‡]
PURE-CD	$n\tau_i\sigma\|A_i\|^2 < 1$	$|J(i)|^*$	Yes	arbitrary	direct

Table 1. Comparison of primal-dual coordinate descent methods. Note that we only compare here the most related methods to ours and include a comprehensive review of other existing methods with comparison to PURE-CD in Section 5. In the last column, we refer to the way one needs to implement the algorithm, for it to be efficient in both sparse and dense settings. $^*J(i)$ is defined in (2). †SPDHG only has implementation for dense setting and not for sparse. ‡The concept of duplication for PDCD is described in (Fercoq & Bianchi, 2019).

functions, involving Lasso, support vector machines and linearly constrained problems with piecewise linear-quadratic objectives. In the general convex case, we prove that the iterates of PURE-CD converges almost surely to a solution of problem (1). Moreover, we show that in this case, the ergodic sequence obtains the optimal $\mathcal{O}(1/k)$ sublinear rate of convergence.

2. Preliminaries

2.1. Notation

For a positive definite matrix V, we denote $\|x\|_V^2 = \langle x, Vx \rangle$. We define the distance of a point x from a set $\mathcal{X}$ as $\mathrm{dist}(x, \mathcal{X}) = \min_{u \in \mathcal{X}} \|u - x\|^2$. Given an index $i \in \{1, \ldots, n\}$, the corresponding coordinate of the gradient vector is $\nabla_i f(x)$ and the corresponding coordinate of a vector $x \in \mathcal{X}$ is x^i. Graph of mapping F is denoted by gra F. Recall that $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$, and $\mathcal{Y} = \prod_{j=1}^m \mathcal{Y}_j$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For $u \in \mathcal{X}_i$, $U_i(u) \in \mathcal{X}$ is such that each element of $U_i(u)$ is 0, except the block i which contains u. We denote the indicator function of a set $\mathcal{X}$ as $\delta_{\mathcal{X}}$.

Proximal operator with a positive definite V is defined as

$$\mathrm{prox}_{V,g}(x) = \arg\min_u g(u) + \frac{1}{2}\|u - x\|_{V^{-1}}^2.$$

We will need the following notation for the sparse setting,

$$\begin{aligned} J(i) &= \{j \in \{1, \ldots, m\} \colon A_{j,i} \neq 0\} \\ I(j) &= \{i \in \{1, \ldots, n\} \colon A_{j,i} \neq 0\}. \end{aligned} \quad (2)$$

Given a matrix A and $i \in \{1, \ldots, n\}$, $J(i)$ denotes the row indices that correspond to nonzero values in the column indexed by i. Similarly, with $j \in \{1, \ldots, m\}$, $I(j)$ gives the column indices corresponding to nonzero values in the row indexed by j.

Moreover, given positive probabilities $(p_i)_{1 \leq i \leq n}$, we define

$$\pi_j = \sum_{i \in I(j)} p_i. \quad (3)$$

In the simple case of $p_i = 1/n$, it is easy to see that $n\pi_j$ corresponds to number of nonzeros in the row indexed by j.

At iteration k, the algorithm randomly picks an index $i_{k+1} \in \{1, \ldots, n\}$. To govern the selection rule, we define the probability matrix $P = \mathrm{diag}(p_1, \ldots, p_n)$, where $p_i = \mathbb{P}(i_{k+1} = i)$, and $\underline{p} = \min_i p_i$. We define as $\mathcal{F}_k$ the filtration generated by the random indices $\{i_1, \ldots, i_k\}$.

Denoting $z = (x, y)$, we define the functions

$$\begin{aligned} D_p(x_{k+1}; z) = f(x_{k+1}) + g(x_{k+1}) - f(x) - g(x) \\ + \langle A^\top y, x_{k+1} - x \rangle, \\ D_d(\bar{y}_{k+1}; z) = h^*(\bar{y}_{k+1}) - h^*(y) - \langle Ax, \bar{y}_{k+1} - y \rangle. \end{aligned}$$

2.2. Optimality

Problem (1) has the following saddle point formulation

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x) + g(x) + \langle Ax, y \rangle - h^*(y).$$

Karush-Kuhn-Tucker (KKT) conditions state that the vector $z_\star = (x_\star, y_\star)$ is a primal-dual solution of the problem when

$$0 \in \begin{bmatrix} \nabla f(x_\star) + \partial g(x_\star) + A^\top y_\star \\ Ax_\star - \partial h^*(y_\star) \end{bmatrix} =: F(z_\star). \quad (4)$$

We call $\mathcal{Z}_\star$ the set of such solutions.

2.3. Metric subregularity

We utilize the metric subregularity assumption for proving linear convergence. This assumption has been used in primal-dual optimization literature for both deterministic (Liang et al., 2016) and randomized algorithms (Latafat et al., 2019; Alacaoglu et al., 2019).

Definition 1. *A set valued mapping $F\colon \mathcal{X} \rightrightarrows \mathcal{Y}$ is metrically subregular at $\bar{x}$ for $\bar{y}$, with $(\bar{x}, \bar{y}) \in$ gra F, if there exists $\eta > 0$ with a neighborhood of regularity $\mathcal{N}(\bar{x})$ such that*

$$\mathrm{dist}(x, F^{-1}\bar{y}) \leq \eta \, \mathrm{dist}(\bar{y}, Fx), \quad \forall x \in \mathcal{N}(\bar{x}).$$

We will be interested in the metric subregularity of KKT operator F (see (4)) for 0. Intuitively speaking, as $0 \in F(z_\star), \forall z^\star \in \mathcal{Z}_\star$, metric subregularity of F for 0 essentially

gives us a way to characterize the behavior of the iterates around the solution set.

Even though Definition 1 looks daunting, fortunately, one does not need to check it for a given problem. Metric sub-regularity is well-studied in the literature and it is known to be satisfied in the following cases:

Example 1.
▷ *If $f + g$ and h^* are strongly convex, Definition 1 holds with $\mathcal{N}(\bar{x}) = \mathbb{R}^d$ (Latafat et al., 2019, Lemma IV.2).*
▷ *If f, g, h are piecewise linear quadratic (PLQ) functions, Definition 1 holds with any bounded neighborhood $\mathcal{N}(\bar{x})$ (Latafat et al., 2019, Lemma IV.4).*

PLQ functions include ℓ_1 norm, hinge loss, indicator of polyhedral sets. Thus, second bullet point apply to Lasso, support vector machines, elastic net, and linearly constrained problems with PLQ loss functions (Latafat et al., 2019).

We now state our main assumptions which are standard in the literature (Fercoq & Bianchi, 2019; Chambolle et al., 2018; Latafat et al., 2019; Bauschke & Combettes, 2011):

Assumption 1.
▷ f, g and h are proper, lower semicontinuous, convex.
▷ g is separable, i.e., $g(x) = \sum_{i=1}^{n} g_i(x^i)$, and f has coordinatewise Lipschitz gradients such that $\forall x \in \mathcal{X}, \forall u \in \mathcal{X}_i$,

$$f(x + U_i(u)) \leq f(x) + \langle \nabla_i f(x), u \rangle + \frac{\beta_i}{2} \|u\|^2. \quad (5)$$

▷ Set of solutions to problem (1), defined in (4) is nonempty.
▷ Slater's condition holds, which states that $0 \in \mathrm{ri}(\mathrm{dom}\, h - A\, \mathrm{dom}\, g)$ where ri denotes the relative interior.

3. Algorithm

In this section, we will sketch the main ideas behind our algorithm. Primal-dual method[1], due to (Chambolle & Pock, 2011; Condat, 2013; Vũ, 2013) reads as

$$\begin{aligned}
\bar{x}_{k+1} &= \mathrm{prox}_{\tau,g}\left(\bar{x}_k - \tau\left(\nabla f(\bar{x}_k) + A^\top \bar{y}_k\right)\right) \\
\bar{y}_{k+1} &= \mathrm{prox}_{\sigma,h^*}\left(\bar{y}_k + \sigma A(2\bar{x}_{k+1} - \bar{x}_k)\right).
\end{aligned} \quad (6)$$

The main intuition behind PDCD methods proposed by (Zhang & Xiao, 2017; Fercoq & Bianchi, 2019; Chambolle et al., 2018) is to incorporate coordinate based updates. Among these methods, (Zhang & Xiao, 2017) specializes in strongly convex-strongly concave problems, whereas the other other ones focus on more general classes of problems.

A closely related approach concentrated on the following interpretation of primal-dual method (6) which is named as

---

[1]This method is also known as Vũ-Condat algorithm.

TriPD in (Latafat et al., 2019, Algorithm 1)

$$\begin{aligned}
\bar{y}_{k+1} &= \mathrm{prox}_{\sigma,h^*}\left(\hat{y}_k + \sigma A\bar{x}_k\right) \\
\bar{x}_{k+1} &= \mathrm{prox}_{\tau,g}\left(\bar{x}_k - \tau\left(\nabla f(\bar{x}_k) + A^\top \bar{y}_{k+1}\right)\right) \quad (7) \\
\hat{y}_{k+1} &= \bar{y}_{k+1} + \sigma A(\bar{x}_{k+1} - \bar{x}_k).
\end{aligned}$$

We notice that by moving the $\bar{y}_{k+1}$ update in TriPD to take place after $\hat{y}_{k+1}$ update, one obtains (6).

As observed in (Latafat et al., 2019), this particular interpretation of primal-dual method is useful for randomization. TriPD-BC as proposed in (Latafat et al., 2019) iterates as

$$\bar{y}_{k+1} = \mathrm{prox}_{\sigma,h^*}\left(y_k + \sigma A x_k\right)$$
$$\bar{x}_{k+1} = \mathrm{prox}_{\tau,g}\left(x_k - \tau\left(\nabla f(x_k) + A^\top \bar{y}_{k+1}\right)\right)$$
$$\hat{y}_{k+1} = \bar{y}_{k+1} + \sigma A(\bar{x}_{k+1} - \bar{x}_k)$$

Draw an index $i_{k+1} \in \{1, \dots, n\}$ randomly.

$$x_{k+1}^{i_{k+1}} = \bar{x}_{k+1}^{i_{k+1}}, \quad x_{k+1}^j = x_k^j, \forall j \neq i_{k+1}$$
$$y_{k+1}^j = \hat{y}_{k+1}^j, \forall j \in J(i_{k+1}), \quad y_{k+1}^j = y_k^j, \forall j \notin J(i_{k+1}).$$

One immediate limitation of TriPD-BC is that to update y_{k+1}, one needs to know $\bar{x}_{k+1}$, whereas only $\bar{x}_{k+1}^{i_{k+1}}$ is needed to update x_{k+1}. As also discussed in (Latafat et al., 2019), this scheme is suitable when A has special structure such as sparsity. When A is dense, one needs to update all elements of y_{k+1} and $\hat{y}_{k+1}$, in which case one needs to compute both $\bar{y}_{k+1}$ and $\bar{x}_{k+1}$ which has the same cost as a deterministic algorithm.

In the dense setting, for an efficient implementation, one can use duplication of dual variables as described in (Fercoq & Bianchi, 2019). However, in this case one is restricted to use small step sizes as discussed in (Fercoq & Bianchi, 2019). Compared to SPDHG in (Chambolle et al., 2018), the step sizes can be n times worse, deteriorating the performance of the method considerably in the dense setting.

On the other hand, the drawback of SPDHG is that it needs to update all dual variables at every iteration, whereas the methods in (Fercoq & Bianchi, 2019; Latafat et al., 2019) update only a subset of dual variables depending on the sparsity of A. When the dual dimension is high, per iteration cost of (Chambolle et al., 2018) becomes prohibitive.

Our idea, inspired by (Chambolle et al., 2018), to make TriPD-BC efficient for dense setting is to use x_{k+1} rather than $\bar{x}_{k+1}$ in the update of $\hat{y}_{k+1}$. Although simple to state, this modification makes $\hat{y}_{k+1}$ random, rendering the analysis of (Latafat et al., 2019) and other analyses based on monotone operator theory not applicable.

This leads to our algorithm, primal-dual method with random extrapolation and coordinate descent (PURE-CD). Our method uses large step sizes as in (Chambolle et al., 2018) in the dense setting, while staying efficient in terms of per

iteration costs in the sparse setting as in (Fercoq & Bianchi, 2019; Latafat et al., 2019); leading to the first general PDCD algorithm that obtains favorable properties in both sparse and dense settings.

Algorithm 1 Primal-dual method with random extrapolation and coordinate descent (PURE-CD)

1: **Input:** Diagonal matrices $\theta, \tau, \sigma > 0$, chosen according to (8), (9).
2: **for** $k = 0, 1 \ldots$ **do**
3: $\quad \bar{y}_{k+1} = \mathrm{prox}_{\sigma, h^*} (y_k + \sigma A x_k)$
4: $\quad \bar{x}_{k+1} = \mathrm{prox}_{\tau, g} \left(x_k - \tau \left(\nabla f(x_k) + A^\top \bar{y}_{k+1} \right) \right)$
5: $\quad$ Draw $i_{k+1} \in \{1, \ldots, n\}$ with $\mathbb{P}(i_{k+1} = i) = p_i$
6: $\quad x_{k+1}^{i_{k+1}} = \bar{x}_{k+1}^{i_{k+1}}$
7: $\quad x_{k+1}^j = x_k^j, \forall j \neq i_{k+1}$
8: $\quad y_{k+1}^j = \bar{y}_{k+1}^j + \sigma_j \theta_j (A(x_{k+1} - x_k))_j, \forall j \in J(i_{k+1}),$
$\quad y_{k+1}^j = y_k^j, \forall j \notin J(i_{k+1})$
9: **end for**

---

4. Convergence Analysis

In this section, we analyze the convergence behavior of Algorithm 1 under various assumptions. We first start with a lemma analyzing one iteration of the algorithm.

Lemma 1. *Let Assumption 1 hold. Recall the definitions of D_p and D_d from Section 2.1 and let $\theta = \mathrm{diag}(\theta_1, \ldots, \theta_m)$ and $\pi = \mathrm{diag}(\pi_1, \ldots, \pi_m)$ be chosen as*

$$\theta_j = \frac{\pi_j}{\underline{p}}, \text{ where } \pi_j = \sum_{i \in I(j)} p_i, \text{ and } \underline{p} = \min_i p_i. \quad (8)$$

We define the functions, given z,

$$V(z) = \frac{\underline{p}}{2} \|x\|_{\tau^{-1} P^{-1}}^2 + \frac{\underline{p}}{2} \|y\|_{\sigma^{-1} \pi^{-1}}^2,$$
$$\tilde{V}(z) = \frac{\underline{p}}{2} \|x\|_{C(\tau)}^2 + \frac{\underline{p}}{2} \|y\|_{\sigma^{-1}}^2,$$

where $C(\tau)_i = \frac{2p_i}{\underline{p}\tau_i} - \frac{1}{\tau_i} - p_i \sum_{j=1}^m \pi_j^{-1} \sigma_j \theta_j^2 A_{j,i}^2 - \frac{\beta_i p_i}{\underline{p}}$.

Then, for the iterates of Algorithm 1, $\forall z \in \mathcal{Z}$, it holds that:

$$\mathbb{E}_k \left[D_p(x_{k+1}; z) \right] + \underline{p} D_d(\bar{y}_{k+1}; z) + \mathbb{E}_k \left[V(z_{k+1} - z) \right]$$
$$\leq (1 - \underline{p}) D_p(x_k; z) + V(z_k - z) - \tilde{V}(\bar{z}_{k+1} - z_k).$$

The main technical challenge in the proof of the lemma, compared to the corresponding results in (Latafat et al., 2019) and (Chambolle et al., 2018) is handling stochasticity in both variables x_{k+1}, y_{k+1} (and also $\hat{y}_{k+1}$ for (Latafat et al., 2019)). Using coordinatewise Lipschitz constants of f with arbitrary sampling also requires an intricate analysis.

The result of Lemma 1 is promising for deriving convergence results for Algorithm 1. When $z = z_\star$ in Lemma 1,

as $D_p(x_{k+1}; z_\star) \geq 0$, $D_d(\bar{y}_{k+1}; z_\star) \geq 0$ and when step sizes are chosen such that $\tilde{V}$ is a squared norm, Lemma 1 describes a stochastic monotonicity property similar to (Fercoq & Bianchi, 2019). In particular, it shows that $D_p(x_{k+1}; z_\star) + V(z_{k+1} - z_\star)$ which measures the distance to solution in a Bregman distance sense, is monotonically nonincreasing in expectation.

4.1. Almost sure convergence

Almost sure convergence is a fundamental property for randomized methods describing the limiting behavior of the iterates in different realization of the algorithm. The following theorem states that the iterates of Algorithm 1 converge almost surely to a point in the solution set.

Theorem 1. *Let Assumption 1 hold and let θ, π be as in Lemma 1. Choose step sizes τ, σ such that*

$$\tau_i < \frac{2p_i - \underline{p}}{\beta_i p_i + \underline{p}^{-1} p_i \sum_{j=1}^m \pi_j \sigma_j A_{j,i}^2}. \quad (9)$$

The iterates z_k are produced by Algorithm 1. Then, almost surely, there exist $z_\star \in \mathcal{Z}_\star$ such that $z_k \to z_\star$.

We analyze the step size rule (9) in Theorem 1 and compare with existing efficient methods in dense and sparse settings.

Remark 1.
▷ Let A be dense, with all its elements being nonzero, $p_i = 1/n$ and $f(\cdot) = 0$, then the step size rule reduces to

$$\tau_i < \frac{1}{n\sigma \|A_i\|^2},$$

which is the step size rule of SPDHG (Chambolle et al., 2018; Alacaoglu et al., 2019), which is shown to be favorable in the dense setting. In contrast, step size rules of (Fercoq & Bianchi, 2019; Latafat et al., 2019) are n times worse due to duplication, in this case.

▷ Let A be diagonal, and we use $p_i = \frac{1}{n}$, which results in $\pi_j = \frac{1}{n}$. Then,

$$\tau_i < \frac{1}{\beta_i + \sum_{j=1}^m \sigma_j A_{j,i}^2},$$

which is the step size rule of Vu-Condat-CD (Fercoq & Bianchi, 2019), upon using the definition of $J(i)$ from (2). Similarly, Algorithm 1 updates 1 dual coordinate and 1 primal coordinate, in this case. In contrast, SPDHG (Chambolle et al., 2018) updates m dual coordinates, resulting in m times higher per iteration cost.

We note that the step sizes of TriPD-BC (Latafat et al., 2019) depend on global Lipschitz constant of f rather than the coordinatewise ones. Using coordinatewise Lipschitz constants in practice is very important for the success of

coordinate descent, as they give larger step sizes (Nesterov, 2012; Richtárik & Takáč, 2014; Fercoq & Richtárik, 2015).

The takeaway from Remark 1 is that Algorithm 1 recovers the characteristics of the best performing methods in fully dense and fully sparse settings. Moreover, as it is the only method with the desirable dependencies in both cases, it has the best properties in the moderate sparse cases. We validate this observation with numerical experiments in Section 6.

4.2. Linear convergence

Linear convergence of primal-dual methods in practice is a widely observed phenomenon (Chambolle & Pock, 2011; Liang et al., 2016). We show that Algorithm 1 also shares this property and obtains linear convergence under metric subregularity, without any modification on the algorithm.

We define the Bregman-type projection onto the solution set

$$z_k^\star = \arg\min_{u \in \mathcal{Z}_\star} D_p(x_k; u) + V(z_k - u). \quad (10)$$

We now show that $z_k^\star$ is well-defined under our assumptions. First, the solution set is convex and closed. Second, $D_p(x_k; u) \geq 0$ for all $u \in \mathcal{Z}_\star$ and it is also lower semicontinuous. Third, we remark that $V(z_k - u)$ is a squared norm (see Lemma 1), thus coercive, therefore the sum is coercive and lower semicontinuous over $\mathcal{Z}_\star$. Hence, $z_k^\star$ exists.

The definition of $z_k^\star$ in (10) is more involved compared to the corresponding quantity in (Latafat et al., 2019). This is in fact due to us using coordinatewise Lipschitz constants in our step sizes, rather than the global Lipschitz constant in (Latafat et al., 2019).

Assumption 2.
KKT operator F is metrically subregular at all $z_\star \in \mathcal{Z}_\star$ for 0, and $\bar{z}_k \in \mathcal{N}(z_\star), \forall z_\star, \forall k$.

Theorem 2. *Let Assumptions 1 and 2 hold. Let θ and the step sizes τ, σ be chosen according to (8) and (9), respectively. Moreover, $z_k^\star = (x_k^\star, y_k^\star)$ is as defined in (10). Then, for z_k generated by Algorithm 1, it follows that*

$$\mathbb{E}\left[\frac{p}{2}\|x_k - x_k^\star\|_{\tau^{-1}P^{-1}}^2 + \frac{p}{2}\|y_k - y_k^\star\|_{\sigma^{-1}\pi^{-1}}^2\right]$$
$$\leq (1 - \rho)^k \Delta_0,$$

where $\rho = \min\left(p, \frac{C_{2,\tilde{V}}}{C_{V,2}((2+2c)+(1+c)(\eta\|H-M\|+\bar{\beta}))^2}\right)$, $\Delta_0 = D_p(x_0; z_0^\star) + V(z_0 - z_0^\star)$, $\bar{\beta}$ *is the global Lipschitz constant of* f,
$C_{2,\tilde{V}} = \frac{p}{2}\min\{\min_i C(\tau)_i, \min_j \sigma_j^{-1}\}$,
$C_{V,2} = \frac{1}{2}\max\left\{\max_i \frac{1}{\tau_i}, \max_j \frac{1}{\sigma_j}\right\}$, $c = C_{2,V}\sqrt{\|A\|/2}$,
$C_{2,V} = \sqrt{\frac{2}{\underline{p}\min\{\min_i \tau_i^{-1}p_i^{-1}, \min_j \sigma_j^{-1}\pi_j^{-1}\}}}$, *and*

$$H = \begin{bmatrix} \tau^{-1} & A^\top \\ 0 & \sigma^{-1} \end{bmatrix}, \quad M = \begin{bmatrix} 0 & A^\top \\ -A & 0 \end{bmatrix}.$$

The first remark about Theorem 2 is that since metric subregularity constant η is not required in the algorithm, the step sizes to achieve linear convergence are the same step sizes as (9). Therefore, PURE-CD adapts to structures on the problem, without any need to modify the algorithm, and attains linear rate of convergence. This supports the well-known observation that primal-dual algorithms converge linearly on most problems, with standard step sizes in (9).

In particular, a direct corollary of our theorem is that for problems listed in Example 1, PURE-CD obtains linear rate of convergence. For the first two cases in Example 1, our result applies directly since the neighborhood of subregularity $\mathcal{N}(z_\star)$ is the whole space. For the third case, we have to assume additionally that $\bar{z}_k$ is contained in a compact set, since $\mathcal{N}(z_\star)$ is not the whole space, and is bounded. A sufficient assumption for this is when the domains of g and h^* are compact. We note that compactness is only required for this result in our paper. This is common to other results for PDCD methods with metric subregularity (Latafat et al., 2019; Alacaoglu et al., 2019). The issue, as explained in (Alacaoglu et al., 2019), stems from a fundamental limitation of the existing analyses of PDCD methods.

Many results in the literature for linear convergence only applies to the first case in Example 1, when g, h^* are strongly convex (Zhang & Xiao, 2017; Chambolle et al., 2018). Moreover, these results require setting step sizes depending on strong convexity constants of g, h^*, therefore not applicable when strong convexity is absent. Our result applies to more general problems and it uses step sizes independent of these constants. Our algorithm can be directly applied to any problem satisfying Assumption 1 and fast convergence will occur provably, if the selected problem is in Example 1.

Compared with the linear convergence rate in (Latafat et al., 2019) for TriPD-BC, our result have a similar contraction factor, however, due to larger step sizes (see Remark 1), the rate comes with a better constant.

4.3. Ergodic rates

In this section, we study Algorithm 1 in the general case, under Assumption 1, and show the optimal $\mathcal{O}(1/k)$ convergence rate on the ergodic sequence. The quantity of interest is the primal-dual gap function (Chambolle & Pock, 2011)

$$G(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{Z}} f(\bar{x}) + g(\bar{x}) + \langle A\bar{x}, y\rangle - h^*(y)$$
$$- f(x) - g(x) - \langle Ax, \bar{y}\rangle + h^*(\bar{y}). \quad (11)$$

A related quantity is the restricted gap function (Chambolle & Pock, 2011), which, for any set $\mathcal{C} \subset \mathcal{Z}$ is defined as

$$G_\mathcal{C}(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{C}} f(\bar{x}) + g(\bar{x}) + \langle A\bar{x}, y\rangle - h^*(y)$$
$$- f(x) - g(x) - \langle Ax, \bar{y}\rangle + h^*(\bar{y}). \quad (12)$$

Due to randomization in PDCD, we are interested in the expected primal-dual gap, denoted as $\mathbb{E}\left[G_{\mathcal{C}}(\bar{x},\bar{y})\right]$. As noted by Dang & Lan (2014), it is technically challenging to prove rates for this quantity as it is the expectation of supremum. Recently, (Alacaoglu et al., 2019) used a technique to show convergence of expected primal-dual gap for SPDHG of (Chambolle et al., 2018). This rate is for ergodic sequence averaging x_k and the full dual variable $\bar{y}_k$. We can use this technique for our analysis. However, there remains another technical challenge as full dual variable is not computed in PURE-CD. Thus, averaging $\bar{y}_k$ is not feasible in our case.

In addition to Assumption 1, in this section we will assume separability of h, to be able to do an efficient averaging with the dual iterate.

Due to the asymmetric nature of Algorithm 1, there are fundamental difficulties for proving a rate with averaging y_{k+1}. On this front, we propose a new type of analysis for the dual variable. To start with, we define the following iterate which has the same cost to compute as y_{k+1} each iteration. Let $\breve{y}_1 = y_1 = \bar{y}_1$,

$$
\begin{aligned}
\breve{y}_{k+1}^j &= \bar{y}_{k+1}^j, \quad \forall j \in J(i_{k+1}), \\
\breve{y}_{k+1}^j &= \breve{y}_k^j, \qquad \forall j \notin J(i_{k+1}).
\end{aligned}
\tag{13}
$$

We note that $\breve{y}_k$ is $\mathcal{F}_k$-measurable and more useful properties of $\breve{y}_k$ for analysis are given in Lemma 5 in Appendix B.4.

Due to the definition of $\breve{y}_k$, it is now feasible to compute and average this iterate. We can show the convergence of expected primal-dual gap by averaging $\breve{y}_k$ and x_k. We remark that we use some coarse inequalities to give simple constants for Theorem 3 and Theorem 4. Therefore, the bounds are not optimized with respect to dimension dependence. In Appendix B, we give these theorems with their original, tighter bounds and we show how we transform the tighter bounds into the constants we give in this section.

Theorem 3. *Let Assumption 1 hold and θ, τ, σ are chosen as in (8), (9). Moreover, let h be separable.*

We define $x_K^{av} = \frac{1}{K}\sum_{k=1}^K x_k$ and $y_K^{av} = \frac{1}{K}\sum_{k=1}^K \breve{y}_k$, where $\breve{y}_k$ is defined in (13), then it holds that for any bounded set $\mathcal{C} = \mathcal{C}_x \times \mathcal{C}_y \subset \mathcal{Z}$

$$
\mathbb{E}\left[G_{\mathcal{C}}(x_K^{av}, y_K^{av})\right] \le \frac{C_g}{\underline{p}K},
$$

where $C_g = \sum_{i=1}^4 C_{g,i}$, $C_{\tau,\tilde{V}} = \min_i C(\tau)_i \tau_i$,
$C_{g,1} = \sup_{z\in\mathcal{C}}\left\{2\underline{p}\|x_0-x\|_{\tau^{-1}P^{-1}}^2 + 2\underline{p}\|y_0-y\|_{\sigma^{-1}\pi^{-1}}^2\right\} + 4\sqrt{\Delta_0\underline{p}^{-1}}\|A\|\sup_{y\in\mathcal{C}_y}\|y\|_{\tau P}$
$+ 2\sqrt{\Delta_0(\underline{p}^{-1}+2\underline{p}^{-3}C_{\tau,\tilde{V}}^{-1})}\|A\|\sup_{x\in\mathcal{C}_x}\|x\|_{\sigma\pi}$,
$\sum_{i=2}^4 C_{g,i} = \Delta_0\left(5 + 9\underline{p}^{-1} + C_{\tau,\tilde{V}}^{-1}\left(1 + 10\underline{p}^{-1} + 14\underline{p}^{-2}\right)\right)$
$+(1-\underline{p})(f(x_0)+g(x_0)-f(x_\star)-g(x_\star))+h^*(y_0)-h^*(y_\star)$
$+\underline{p}\|Ax_\star\|_{\sigma\pi^{-1}}^2 + \|A^\top y_\star\|_{\tau P}^2.$

Remark 2. When implementing averaging of x_k, and $\breve{y}_k$, one should use a technique similar to (Dang & Lan, 2015). The main idea is to only update the averaged vector at the coordinates where an update occurred. For this, one needs to remember for each coordinate, the last time it is updated, wait until a coordinate is selected again and update the averaged vector using this information.

The result in Theorem 3 would give a rate for primal-dual gap when $\mathcal{C} = \mathcal{Z}$. However, in general such a rate is not desirable as taking a supremum over $\mathcal{Z}$ might result in an infinite bound. This rate would be meaningful when both primal and dual domains are bounded in which case one would take the supremum in $C_{g,1}$ over the bounded domains.

Alternatively, in the following theorem, we show that for two important special cases, we can extend this result to show guarantees without bounded domains. Namely, we show the same rate for the case when $h(\cdot) = \delta_{\{b\}}(\cdot)$, $b \in \mathbb{R}^m$ to cover linearly constrained problems. Moreover, we show the result for the case when h is Lipschitz continuous.

Theorem 4. *Let Assumption 1 hold. We use the same parameters θ, τ, σ and the definitions for x_K^{av} and y_K^{av} as Theorem 3. We consider two cases separately:*
▷ *If $h(\cdot) = \delta_{\{b\}}(\cdot)$, we obtain*

$$
\mathbb{E}[f(x_K^{av}) + g(x_K^{av}) - f(x_\star) - g(x_\star)] \le \frac{C_o}{\underline{p}K}.
$$

$$
\mathbb{E}[\|Ax_K^{av} - b\|] \le \frac{C_f}{\underline{p}K}.
$$

▷ *If h is L_h-Lipschitz continuous, we obtain*

$$
\begin{aligned}
\mathbb{E}[f(x_K^{av}) &+ g(x_K^{av}) + h(Ax_K^{av}) \\
&- f(x_\star) - g(x_\star) - h(Ax_\star)] \le \frac{C_l}{\underline{p}K},
\end{aligned}
$$

where $C_f = 3c_1\|x_\star - x_0\|_{\tau^{-1}P^{-1}} + 2\sqrt{c_1 C_s} + 4c_1\|y_\star - y_0\|_{\sigma^{-1}\pi^{-1}}$,
$C_o = C_s + \|y_\star\|_{\sigma^{-1}\pi^{-1}}C_f + 2c_1\underline{p}^{-1}V(z_0 - z_\star)$,
$C_l = C_s + c_1\|x_\star - x_0\|_{\tau^{-1}P^{-1}}^2 + 4c_1 L_h^2$,
$C_s = C_{g,2} + C_{g,5} + C_{g,6}$, *with $c_1 = 2\underline{p} + 2$, $C_{g,2}$ as defined in the statement of Theorem 3 in Appendix B.4 and $C_{g,5}, C_{g,6}$ are defined in the proof in (79), (80).*

5. Related works

In the deterministic setting, many primal-dual methods are proposed (Chambolle & Pock, 2011; Vũ, 2013; Condat, 2013; Tan et al., 2020; Latafat et al., 2019). The standard results in these papers include linear convergence when g, h^* are strongly convex, with step sizes selected by using strong convexity constants. In addition, these papers also show sublinear $\mathcal{O}(1/k)$ rate with general convexity, which

is known to be optimal (Nesterov, 2005). Moreover, linear rates under metric subregularity is shown for deterministic methods in (Liang et al., 2016; Latafat et al., 2019).

Randomized coordinate descent is proposed in (Nesterov, 2012) and improved by a large body of subsequent papers (Richtárik & Takáč, 2014; Fercoq & Richtárik, 2015). Primal randomized coordinate descent requires full separability on the nonsmooth parts of the objective function. Nonsmooth and nonseparable functions are handled by primal-dual coordinate descent methods (Fercoq & Bianchi, 2019).

One of the first primal-dual coordinate descent (PDCD) methods is SPDC, which is proposed in (Zhang & Xiao, 2017), that solves a special case of problem (1) with $f = 0$. SPDC has linear convergence when g, h^* are strongly convex and the step sizes are selected according to strong convexity constants. In the general convex case, SPDC has perturbation-based analysis, which needs to set an ϵ, requires knowing $\|x_\star\|^2$, and shows ϵ-based iteration complexity results, and not anytime convergence rates. Almost sure convergence of the iterates of SPDC is not proven in the general convex case. Moreover, the step sizes of SPDC are scalar and they depend on the maximum block norm of A. It is shown in (Zhang & Xiao, 2017) that in the specific cases when $g(x) = \|x\|^2$ or $g(x) = \|x\|_1 + \|x\|^2$, one can use a special implementation for efficiency with sparse data.

Tan et al. (2020) proposed a new method similar to SPDC with the same type of guarantees as (Zhang & Xiao, 2017). Due to similar analysis techniques, this method inherits the abovementioned drawbacks of SPDC. For this method, Tan et al. (2020) showed a new implementation technique for sparse data, that can be used with any separable $g(x)$.

For solving the specific case of empirical risk minimization problems, stochastic dual coordinate ascent (SDCA) is proposed in (Shalev-Shwartz & Zhang, 2013; 2014). SDCA uses strong convexity constant to set step sizes and attain linear convergence. A limitation of SDCA is to require strong convexity in the primal, to ensure smoothness of the dual objective, which is essential in the design of the method.

Another early PDCD method is by (Dang & Lan, 2014) where the authors focused on showing sublinear convergence rates. The authors showed guarantees for a relaxed version of expected primal-dual gap function in (11).

Building on (Dang & Lan, 2014), block-coordinate variants of alternating direction method of multipliers (ADMM) are proposed in (Gao et al., 2019; Xu & Zhang, 2018). These papers focus on linearly constrained problems and show ergodic sublinear convergence rates. Moreover, (Xu & Zhang, 2018) showed that under strong convexity assumption and special decomposition of the blocks, the method achieves linear convergence. This linear convergence result, similar to (Zhang & Xiao, 2017) requires knowing the strong

convexity constants to set the algorithmic parameters. Moreover, these results generally set step sizes depending on global Lipschitz constants and norm of whole matrix A.

PDCD variants are also proposed in (Combettes & Pesquet, 2015; 2019; Pesquet & Repetti, 2015) and analyzed under the general setting of monotone operators. These methods use global constants of the problem such as global Lipschitz constant of smooth part and $\|A\|$, rather than blockwise constants, resulting in worse practical performance, as illustrated in the experiments of (Chambolle et al., 2018).

Another early PDCD variant to solve problem (1) in its full generality, where f, g, h are all nonseparable, is by Fercoq & Bianchi (2019). This method uses coordinatewise Lipschitz constants of the smooth part and it is designed to exploit sparsity of A. This method has almost sure convergence guarantees as well as linear convergence when g, h^* are strongly convex. As opposed to most results in this nature, it is not required to know strong convexity constants to set the step sizes. In the general convex case, the method has $\mathcal{O}(1/\sqrt{k})$ rate for a randomly selected iterate. As argued in Section 4.1, main limitation of (Fercoq & Bianchi, 2019) is that small step sizes are required when matrix A is dense. Moreover, the results in this paper are restricted to uniform probability law for selecting coordinates.

One of the most related works to ours, and a building block of PURE-CD is TriPD-BC from (Latafat et al., 2019). The authors showed almost sure convergence of the iterates and linear convergence under metric subregularity, by using global Lipschitz constants of f for the step sizes. This paper did not have any sublinear convergence rates in the general convex case. Similar to (Fercoq & Bianchi, 2019), TriPD-BC is designed for sparse setting and a naive implementation in the dense setting requires the same per iteration cost as the deterministic algorithm. An efficient implementation is by duplication of dual variables, which as explained in (Fercoq & Bianchi, 2019) results in small step sizes (see Section 4.1).

Another building block of PURE-CD is SPDHG by (Chambolle et al., 2018), to solve (1) when $f = 0$. Linear convergence result of SPDHG by (Chambolle et al., 2018) is similar to (Zhang & Xiao, 2017) and requires setting step sizes with strong convexity constants. In the general convex case and partially strongly convex case, (Chambolle et al., 2018) proved optimal sublinear rates. Recently, (Alacaoglu et al., 2019) analyzed SPDHG and proved additional theoretical results. In particular, this work showed almost sure convergence of the iterates of SPDHG and linear convergence under metric subregularity. Even though it is fast in the dense setting, the main limitation of SPDHG, as discussed in Section 4.1 is that it needs to update all the dual coordinates, resulting in high per iterations costs in the sparse setting.
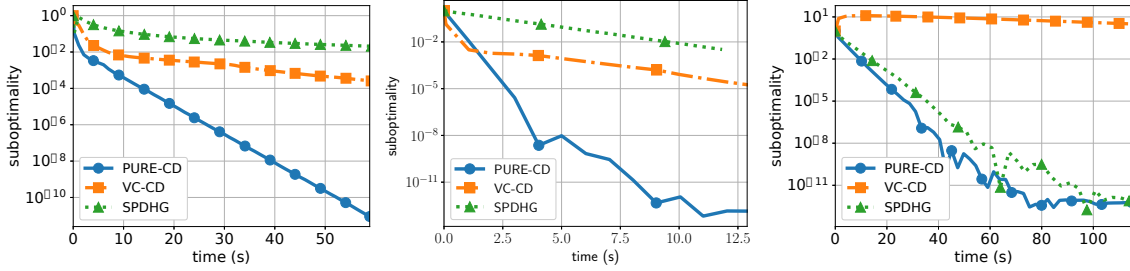
Figure 1. Lasso: Left: rcv1, $n = 20,242, m = 47,236$, density $= 0.16\%$, $\lambda = 10$; Middle: w8a, $n = 49,749, m = 300$, density $= 3.9\%$, $\lambda = 10^{-1}$; Right: covtype, $n = 581,012$, $m = 54$, density $= 22.1\%$, $\lambda = 10$.
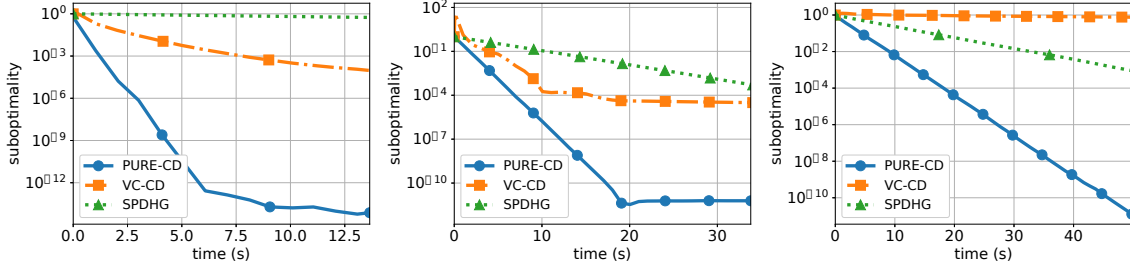

Figure 2. Ridge regression: Left: sector, $n = 6,412, m = 55,197$, density $= 0.3\%$, $\lambda = 0.1$; Middle: a9a, $n = 32,561, m = 123$, density $= 11.3\%$, $\lambda = 0.1$; Right: mnist, $n = 60,000$, $m = 780$, density $= 19.2\%$, $\lambda = 1$.

Other PDCD methods are proposed in (Luke & Malitsky, 2018; Alacaoglu et al., 2017) where the authors focused on linearly constrained problems and proved sublinear rates.

6. Numerical experiments

6.1. Effect of sparsity

As explained in Section 4.1, and Remark 1, PURE-CD brings together the benefits of different methods that are designed for dense and sparse cases. We will now compare the empirical performance of PURE-CD with Vu-Condat-CD from (Fercoq & Bianchi, 2019) which has desirable properties with sparse data and SPDHG from (Chambolle et al., 2018) which has desirable properties with dense data.

We select uniform sampling, $p_i = 1/n$, so (9) simplifies to

$$\tau_i < \frac{1}{\sum_{j=1}^{m} \theta_j \sigma_j A_{j,i}^2}. \tag{14}$$

We provide a step size policy inspired by the step size rules chosen in (Chambolle et al., 2018) and (Fercoq & Bianchi, 2019). We use the following step sizes, for $\gamma < 1$

$$\sigma_j = \frac{1}{\theta_j \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \max_{i'} \|A_{i'}\|}{\|A_i\|^2}.$$

We note that in contrast to (Chambolle et al., 2018), step sizes are both diagonal. In our case, it is important to utilize diagonal step sizes for both primal and dual variables since we perform coordinate-wise updates for both primal and

dual variables and the step sizes need to be set appropriately to obtain good practical performance. For SPDHG and Vu-Condat-CD, we use step sizes suggested in the papers.

In the edge cases (one nonzero element per row or fully dense), it is easy to see that our step size policy reduces to the suggested step sizes of (Chambolle et al., 2018) and (Fercoq & Bianchi, 2019).

For experiments, we used the generic coordinate descent solver, implemented in Cython, by Fercoq (2019), which includes an implementation of Vu-Condat-CD with duplication and we implemented SPDHG and PURE-CD. We solve Lasso and ridge regression, where we let $g(x) = \lambda\|x\|_1$, $h(Ax) = \frac{1}{2}\|Ax - b\|^2$, $f = 0$, and $g(x) = \frac{\lambda}{2}\|x\|^2$, $h(Ax) = \frac{1}{2}\|Ax - b\|^2$, $f = 0$, respectively, in our template (1). Then, we apply all the methods to the dual problems of these, to access data by rows.

We use datasets from LIBSVM with different sparsity levels (Chang & Lin, 2011). The properties of each data matrix are given in the caption of the corresponding figures. For preprocessing, we removed all-zero rows and all-zero columns of A and we performed row normalization. The results are compiled in Figures 1 and 2.

We observe the behavior predicted by theory. With sparse data such as rcv1, where density level is 0.16%, SPDHG makes very little progress in the given time window. The reason is that the per iteration cost of SPDHG in this case is updating $47,236$ dual variables, whereas for PURE-CD and Vu-Condat-CD, the cost is updating 75 dual variables.
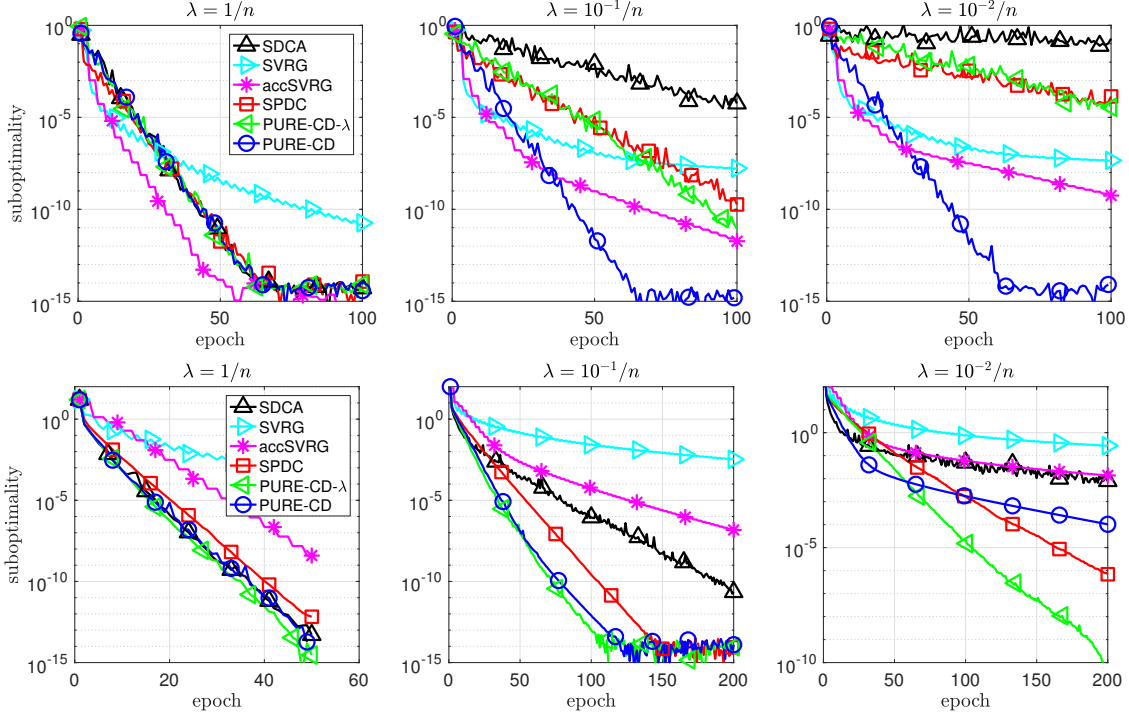
Figure 3. top: a9a, $n = 32,561$, $m = 123$, bottom: sector, $n = 6,412$, $m = 55,197$.

We note that PURE-CD is faster than Vu-Condat-CD due to better step sizes. On the other hand, with moderate sparsity, SPDHG and Vu-Condat-CD is comparable, whereas PURE-CD exhibits the best performance. For denser data, SPDHG and PURE-CD exhibit similar behavior where Vu-Condat-CD is slower than both due to smaller step sizes.

6.2. Comparison with specialized methods

In this section, we compare the practical performance of PURE-CD with state-of-the-art algorithms that are designed for strongly convex-strongly concave problems. Due to space constraints, we defer some of the plots and more details about experiments to the appendix. We focus on the problem $\min_x \frac{1}{n} \sum_{i=1}^n h_i(A_i x) + \frac{\lambda}{2} \|x\|^2$, where $h_i(x) = (x - b_i)^2$. Each h_i is smooth with Lipschitz constants $L_i = 2$ and the second component is strongly convex, which results in strong convexity in both primal and dual problems.

In this case, the algorithms SDCA (Shalev-Shwartz & Zhang, 2013), ProxSVRG (Xiao & Zhang, 2014), Accelerated SVRG (Zhou et al., 2018), SPDC (Zhang & Xiao, 2017) are all designed to use the strong convexity to obtain linear convergence. These algorithms use the strong convexity constant λ for setting the algorithmic parameters. Moreover, as all the abovementioned algorithms have special implementations to exploit sparsity, we make the comparison with respect to number of passes of the data, rather than time. The results are compiled for two datasets

in Figure 3 and more datasets are included in Appendix A. We use theoretical step sizes for all the algorithms, given in the respective papers.

• PURE-CD-λ: This variant uses the non-agnostic step sizes, using λ, which still satisfy the theoretical requirement (14).

$$\sigma_j = \frac{n}{\theta_j \sqrt{n\lambda} \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \sqrt{n\lambda} \max_{i'} \|A_{i'}\|}{n \|A_i\|^2}.$$

• PURE-CD: This variant is with the standard agnostic step sizes.

$$\sigma_j = \frac{n}{\theta_j \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \max_{i'} \|A_{i'}\|}{n \|A_i\|^2}.$$

We observe that PURE-CD has a consistent linear convergence behavior as predicted by theory. In most of the datasets (see Appendix A), it has the fastest convergence behavior. However, in some datasets, as λ gets smaller, we observed that the linear rate of PURE-CD slowed down, which motivated us to try PURE-CD-λ, which incorporates the knowledge of λ as the other methods. It seems to show favorable behavior when PURE-CD slows down.

The takeaway message is that PURE-CD, which is designed for a general problem, adapts to strong convexity well with agnostic step sizes in most cases. However, in some cases, it does not perform as good as the algorithms which are designed to exploit strong convexity. In those cases however, one can choose separating step sizes of PURE-CD accordingly, and use PURE-CD-λ to get better performance.

Acknowledgements

References

Alacaoglu, A., Dinh, Q. T., Fercoq, O., and Cevher, V. Smooth primal-dual coordinate descent algorithms for nonsmooth convex optimization. In *Advances in Neural Information Processing Systems*, pp. 5852–5861, 2017.

Alacaoglu, A., Fercoq, O., and Cevher, V. On the convergence of stochastic primal-dual hybrid gradient. *arXiv preprint arXiv:1911.00799*, 2019.

Arrow, K. J., Azawa, H., Hurwicz, L., and Uzawa, H. *Studies in linear and non-linear programming*, volume 2. Stanford University Press, 1958.

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schonlieb, C.-B. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Combettes, P. L. and Pesquet, J.-C. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

Combettes, P. L. and Pesquet, J.-C. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping ii: mean-square and linear convergence. *Mathematical Programming*, 174(1-2):433–451, 2019.

Condat, L. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

Dang, C. and Lan, G. Randomized methods for saddle point computation. *arXiv preprint arXiv:1409.8625*, 3(4), 2014.

Dang, C. D. and Lan, G. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

Fercoq, O. A generic coordinate descent solver for nonsmooth convex optimisation. *Optimization Methods and Software*, pp. 1–21, 2019.

Fercoq, O. and Bianchi, P. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM Journal on Optimization*, 29(1):100–134, 2019.

Fercoq, O. and Richtárik, P. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Fercoq, O., Alacaoglu, A., Necoara, I., and Cevher, V. Almost surely constrained convex optimization. In *International Conference on Machine Learning*, pp. 1910–1919, 2019.

Gao, X., Xu, Y.-Y., and Zhang, S.-Z. Randomized primal–dual proximal block coordinate updates. *Journal of the Operations Research Society of China*, 7(2):205–250, 2019.

Iutzeler, F., Bianchi, P., Ciblat, P., and Hachem, W. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE conference on decision and control*, pp. 3671–3676. IEEE, 2013.

Latafat, P., Freris, N. M., and Patrinos, P. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *arXiv preprint arXiv:1706.02882v4*, 2019.

Liang, J., Fadili, J., and Peyré, G. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1-2):403–434, 2016.

Luke, D. R. and Malitsky, Y. Block-coordinate primal-dual method for nonsmooth minimization over linear constraints. In *Large-Scale and Distributed Optimization*, pp. 121–147. Springer, 2018.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Patrascu, A. and Necoara, I. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18: 198–1, 2017.

Pesquet, J.-C. and Repetti, A. A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis*, 16(12):2453–2490, 2015.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2):1–38, 2014.

Shalev-Shwartz, S. and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International Conference on Machine Learning*, pp. 64–72, 2014.

Tan, C., Qian, Y., Ma, S., and Zhang, T. Accelerated dual-averaging primal-dual method for composite convex minimization. *Optimization Methods and Software*, 0(0):1–26, 2020. doi: 10.1080/10556788.2020.1713779.

Tran-Dinh, Q., Fercoq, O., and Cevher, V. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28 (1):96–134, 2018.

Vũ, B. C. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xu, Y. and Zhang, S. Accelerated primal–dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70(1):91–128, 2018.

Zhang, Y. and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *International Conference on Machine Learning*, pp. 5975–5984, 2018.

A. More experimental results

In this section, we compare the practical performance of PURE-CD with state-of-the-art algorithms that are designed to exploit problem structures. In particular, we focus on the problem

$$\min_x \frac{1}{n} \sum_{i=1}^n h_i(A_i x) + \frac{\lambda}{2} \|x\|^2, \tag{15}$$

where $h_i(x) = (x - b_i)^2$. Each h_i is smooth with Lipschitz constants $L_i = 2$ and the second component is strongly convex. This is equivalent to strong convexity in both primal and dual problems.

In this case, the algorithms SDCA, SVRG, Accelerated SVRG/Katyusha, SPDC are all designed to use the strong convexity to obtain linear convergence. These algorithms use the strong convexity constant λ for setting the algorithmic parameters (with the exception of SVRG which theoretically needs it to set number of inner loop iterations). Moreover, as all the abovementioned algorithms have special implementations to exploit sparsity, for fairness, as all algorithms have different structures, we did not try to implement them in the most efficient manner, therefore we make the comparison with respect to number of passes of the data, rather than time.

We use PURE-CD with the agnostic step size and also with a non-agnostic step size that uses λ. Both step size rules are supported by theory. Moreover, similar to Section 6, we apply PURE-CD to the dual problem of (15) to access the data row-wise as other methods.

The details of parameters for each algorithm:
• SVRG: We use the theoretical step size given in (Xiao & Zhang, 2014, Theorem 3.1)

• Accelerated SVRG/Katyusha: We use the theoretically suggested step size parameter and acceleration parameter (Zhou et al., 2018, Theorem 1, Table 2)

• SDCA: We use directly the specialization of SDCA for ridge regression, as decribed in (Shalev-Shwartz & Zhang, 2013, Section 6.2)

• SPDC: We use the step sizes from (Zhang & Xiao, 2017, Theorem 1)

• PURE-CD-λ: This variant uses the non-agnostic step sizes, using λ. We note that the step sizes satisfy the theoretical requirement (14).

$$\sigma_j = \frac{n}{\theta_j \sqrt{n\lambda} \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \sqrt{n\lambda} \max_{i'} \|A_{i'}\|}{n \|A_i\|^2}.$$

• PURE-CD: This variant is with the standard agnostic step sizes, as in Section 6. We note that the step sizes are scaled by n since the problem is scaled by $1/n$, compared to Section 6.

$$\sigma_j = \frac{n}{\theta_j \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \max_{i'} \|A_{i'}\|}{n \|A_i\|^2}.$$

We use datasets from LIBSVM, and try three different regularization parameters $\frac{1}{n}$, $\frac{10^{-1}}{n}$, and $\frac{10^{-2}}{n}$. We performed preprocessing by removing the all-zero rows and columns from the data matrix, then we normalized row norms of A. We chose the parameters as described above and did not perform any tuning for any algorithm.

We observe that PURE-CD has a consistent linear convergence behavior as predicted by theory. In most of the datasets, it has the fastest convergence behavior. However, in some datasets, as λ gets smaller, we observed that the linear rate of PURE-CD slowed down, which motivated us to try PURE-CD-λ, which incorporates the knowledge of λ as the other methods. It seems to show favorable behavior when PURE-CD slows down.

The takeaway message is that PURE-CD adapts very well with agnostic step sizes in most cases. However, in some cases, it does not perform as good as the algorithms which are designed to exploit structure. In those cases however, one can choose separating step sizes of PURE-CD accordingly, and use PURE-CD-λ to get better performance.
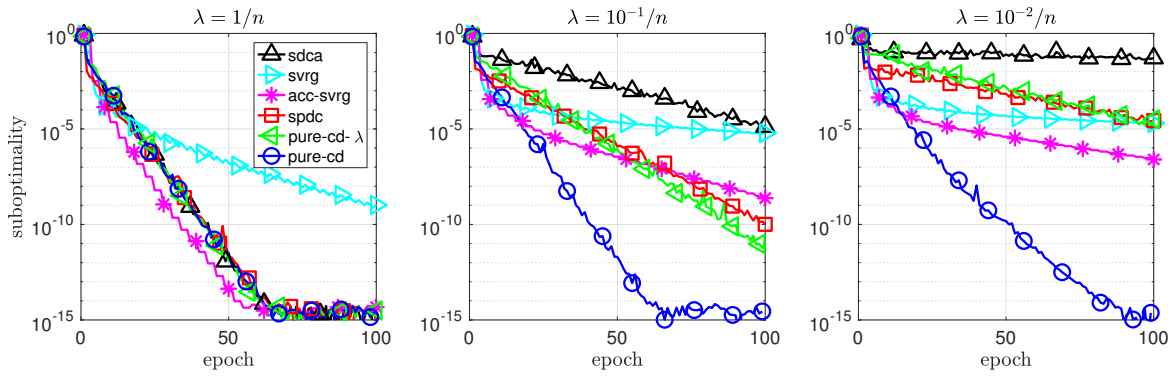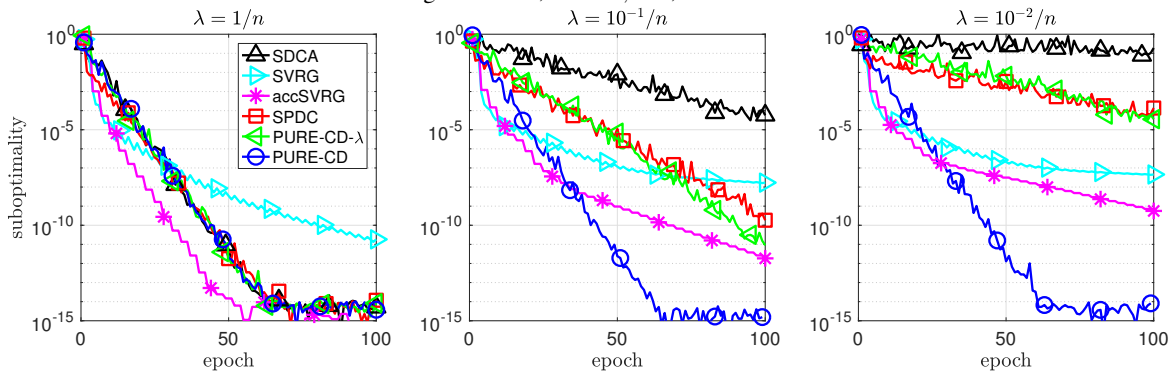
Figure 4. w8a, $n = 49,749$, $m = 300$.

Figure 5. a9a, $n = 32,561$, $m = 123$.
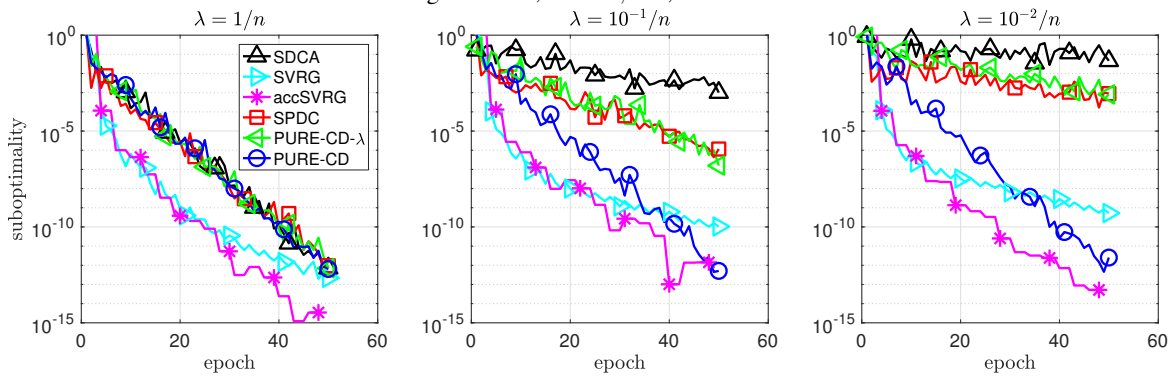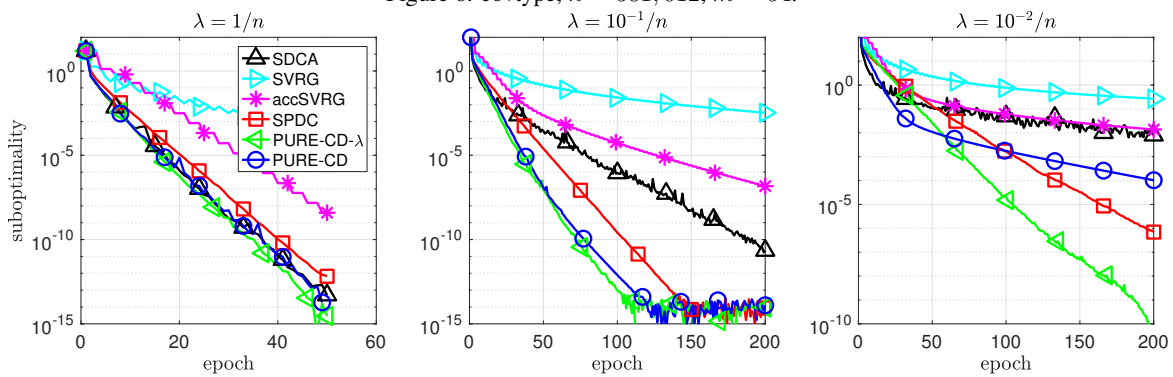

Figure 6. covtype, $n = 581,012$, $m = 54$.
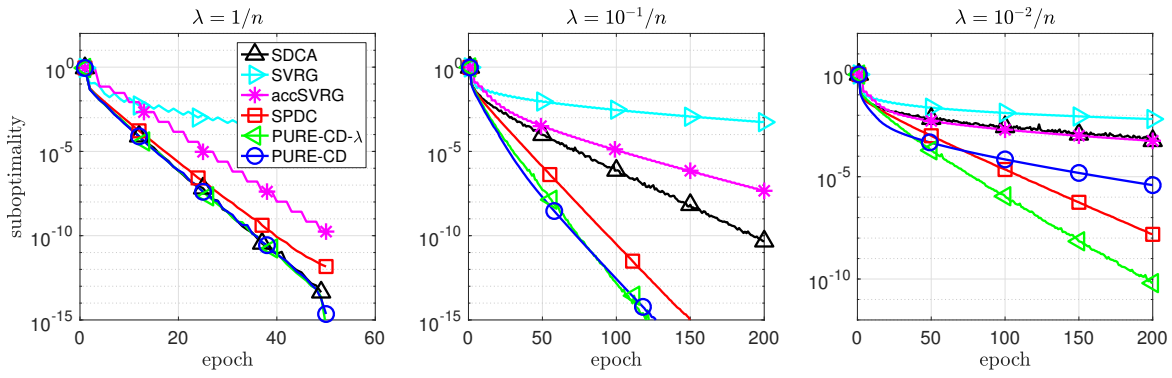

Figure 7. sector, $n = 6,412$, $m = 55,197$.
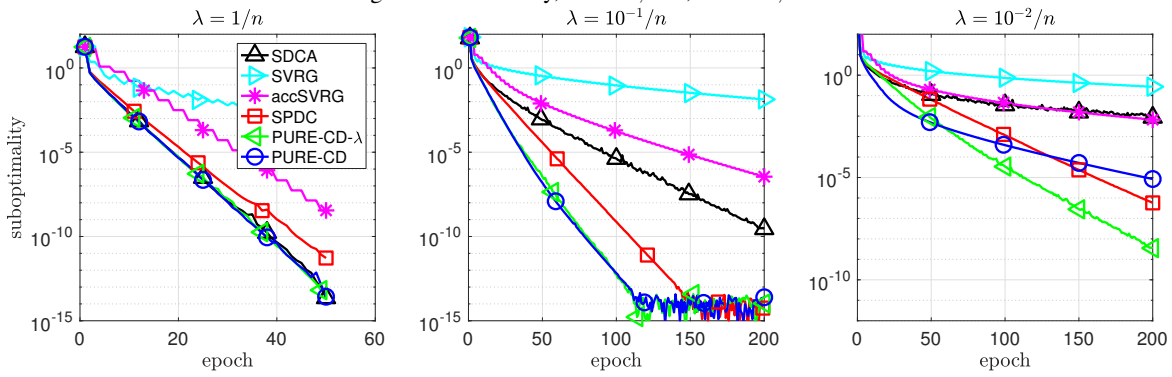
Figure 8. rcv1.binary, $n = 20,242$, $m = 47,236$.

Figure 9. news20, $n = 15,935$, $m = 62,061$.
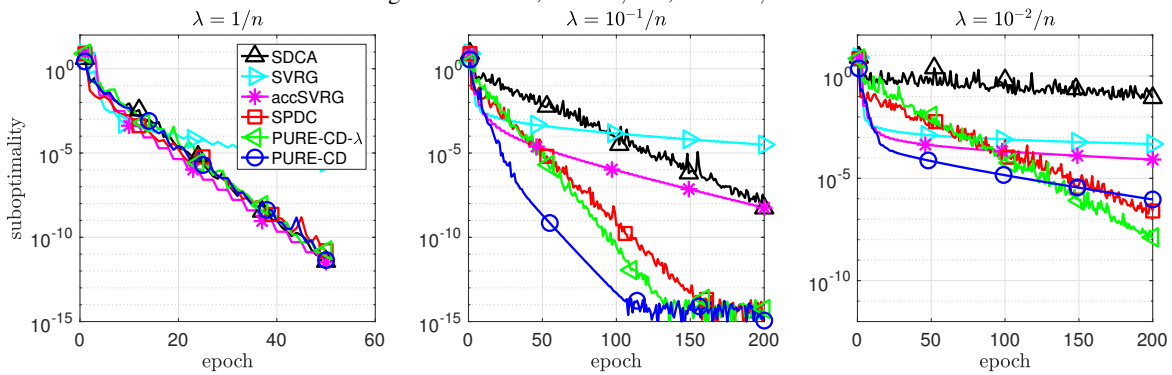
Figure 10. mnist, $n = 60,000$, $m = 780$.
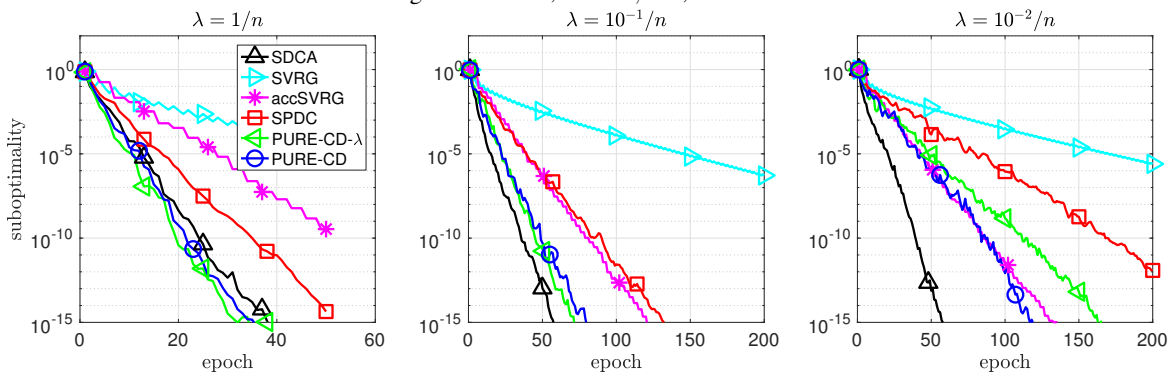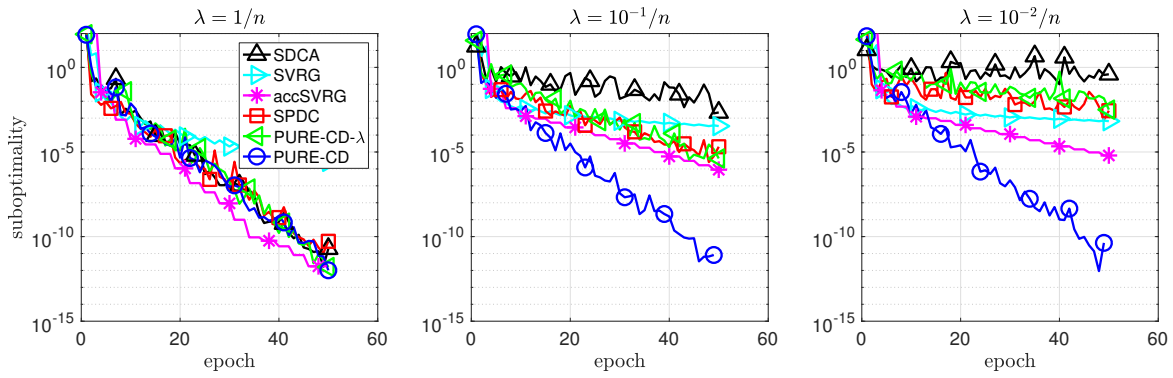

Figure 11. leukemia, $n = 38$, $m = 7,129$.

Figure 12. YearPredictionMSD, $n = 463,715$, $m = 90$.

A.1. Further details about experiments of Section 6

The experiments are done on a computer with Intel Core i7 CPUs at 2.9 GHz.

B. Proofs

B.1. Proofs for one iteration result

We start with some technical lemmas. Our first result characterizes the conditional expectation of y_{k+1}.

Lemma 2. *Let y_{k+1} as defined in Algorithm 1, and recall that $\pi = \mathrm{diag}(\pi_1, \ldots, \pi_m)$, such that $\pi_j = \sum_{i \in I(j)} p_i, \forall j \in \{1, \ldots, m\}$ and $P = \mathrm{diag}(p_1, \ldots, p_n)$. Then it holds that for any $\mathcal{F}_k$-measurable Y and $\forall \gamma = \{\gamma_1, \ldots, \gamma_m\}$, with $\gamma_i > 0$,*

$$\mathbb{E}_k \left[\|y_{k+1} - Y\|_\gamma^2 \right] = \|\bar{y}_{k+1} - Y\|_{\gamma\pi}^2 - \|y_k - Y\|_{\gamma\pi}^2 + \|y_k - Y\|_\gamma^2 + 2\langle \bar{y}_{k+1} - Y, \gamma\sigma\theta AP(\bar{x}_{k+1} - x_k) \rangle$$
$$+ \sum_{i=1}^n \sum_{j=1}^m p_i \gamma_j \sigma_j^2 \theta_j^2 A_{j,i}^2 (\bar{x}_{k+1}^i - x_k^i)^2.$$

Proof. First, we note that for $\mathcal{F}_k$-measurable Y, it follows that

$$\mathbb{E}_k \left[\|y_{k+1} - Y\|_\gamma^2 \right] = \mathbb{E}_k \left[\sum_{j=1}^m \gamma_j \left(y_{k+1}^j - Y^j \right)^2 \right]$$

$$= \mathbb{E}_k \left[\sum_{j \in J(i_{k+1})} \gamma_j (\bar{y}_{k+1}^j + \sigma_j \theta_j (A(x_{k+1} - x_k))_j - Y^j)^2 + \sum_{j \notin J(i_{k+1})} \gamma_j (y_k^j - Y^j)^2 \right]$$

$$= \sum_{i=1}^n p_i \left[\sum_{j \in J(i)} \gamma_j \left(\bar{y}_{k+1}^j + \sigma_j \theta_j A_{j,i}(\bar{x}_{k+1}^i - x_k^i) - Y^j \right)^2 + \sum_{j \notin J(i)} \gamma_j \left(y_k^j - Y^j \right)^2 \right]$$

$$= \sum_{i=1}^n \sum_{j \in J(i)} p_i \gamma_j (\bar{y}_{k+1}^j - Y^j)^2 + 2 \sum_{i=1}^n \sum_{j \in J(i)} p_i \gamma_j \sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i)(\bar{y}_{k+1}^j - Y^j)$$

$$+ \sum_{i=1}^n \sum_{j \in J(i)} p_i \gamma_j \left(\sigma_j \theta_j A_{j,i}(\bar{x}_{k+1}^i - x_k^i) \right)^2 + \sum_{i=1}^n \sum_{j \notin J(i)} p_i \gamma_j (y_k^j - Y^j)^2 \qquad (16)$$

where for the third equality, we used the fact that x_{k+1} is different from x_k only on the coordinate i_{k+1} and $\bar{x}_{k+1}^{i_{k+1}} = x_{k+1}^{i_{k+1}}$, which gives

$$(A(x_{k+1} - x_k))_j = (A((x_{k+1}^{i_{k+1}} - x_k^{i_{k+1}})e_{i_{k+1}}))_j = A_{j,i_{k+1}}(\bar{x}_{k+1}^{i_{k+1}} - x_k^{i_{k+1}}).$$

We focus on the last term on the right hand side of (16)

$$\sum_{i=1}^n \sum_{j \notin J(i)} p_i \gamma_j (y_k^j - Y^j)^2 = \sum_{i=1}^n \sum_{j=1}^n p_i \gamma_j (y_k^j - Y^j)^2 - \sum_{i=1}^n \sum_{j \in J(i)} p_i \gamma_j (y_k^j - Y^j)^2$$

$$= \|y_k - Y\|_\gamma^2 - \sum_{j=1}^m \sum_{i \in I(j)} p_i \gamma_j (y_k^j - Y^j)^2 = \|y_k - Y\|_\gamma^2 - \sum_{j=1}^m \pi_j \gamma_j (y_k^j - Y^j)^2$$

$$= \|y_k - Y\|_\gamma^2 - \|y_k - Y\|_{\gamma\pi}^2 \qquad (17)$$

where we use the fact that $\sum_{i=1}^n \sum_{j \in J(i)} \gamma'_{j,i} = \sum_{j=1}^m \sum_{i \in I(j)} \gamma'_{j,i}$, for any γ', due to the definition of $J(i)$, $I(j)$ and $\pi_j = \sum_{i \in I(j)} p_i$.

We estimate the first term of (16), similar to (17)

$$\sum_{i=1}^n \sum_{j \in J(i)} p_i \gamma_j (\bar{y}_{k+1}^j - Y^j)^2 = \sum_{j=1}^m \sum_{i \in I(j)} p_i \gamma_j (\bar{y}_{k+1}^j - Y^j)^2 = \|\bar{y}_{k+1} - Y\|_{\gamma\pi}^2. \qquad (18)$$

We lastly estimate the second and third terms of (16). We use the fact that $A_{j,i} = 0$, if $j \notin J(i)$ to obtain

$$2 \sum_{i=1}^{n} \sum_{j \in J(i)} p_i \gamma_j \sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i)(\bar{y}_{k+1}^j - Y^j) + \sum_{i=1}^{n} \sum_{j \in J(i)} p_i \gamma_j \left(\sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i) \right)^2$$

$$= 2 \sum_{i=1}^{n} \sum_{j=1}^{m} p_i \gamma_j \sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i)(\bar{y}_{k+1}^j - Y^j) + \sum_{i=1}^{n} \sum_{j=1}^{m} p_i \gamma_j \left(\sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i) \right)^2$$

$$= 2 \langle \bar{y}_{k+1} - Y, \gamma \sigma \theta A P (\bar{x}_{k+1} - x_k) \rangle + \sum_{i=1}^{n} \sum_{j=1}^{m} p_i \gamma_j \left(\sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i) \right)^2. \quad (19)$$

We use (17), (18), and (19) in (16) to obtain the final result. $\qquad \square$

We continue with the following lemma which handles necessary manipulations for the terms involving the primal variable, to handle arbitrary probabilities.

Lemma 3. *We recall that $P = \mathrm{diag}(p_1, \ldots, p_n)$, $\underline{p} = \min_i p_i$, and define*

$$x' = x_k + P^{-1} \underline{p}(x - x_k) = P^{-1} \underline{p} x + (1 - P^{-1} \underline{p}) x_k,$$

and

$$g_P(x) = \sum_{i=1}^{n} p_i g_i(x^i).$$

Then, for a function $g(x) = \sum_{i=1}^{n} g_i(x^i)$, the following conclusions hold:

$$g_P(x') \leq \underline{p} g(x) - \underline{p} g(x_k) + g_P(x_k),$$

$$\|x' - x_{k+1}\|_{\tau^{-1}}^2 - \|x' - x_k\|_{\tau^{-1}}^2 = \underline{p} \|x - x_{k+1}\|_{\tau^{-1}P^{-1}}^2 + \|x_{k+1} - x_k\|_{\tau^{-1}}^2 - \underline{p} \|x_{k+1} - x_k\|_{\tau^{-1}P^{-1}}^2$$
$$- \underline{p} \|x - x_k\|_{\tau^{-1}P^{-1}}^2.$$

Proof. We have that $x'^i = p_i^{-1} \underline{p} x^i + (1 - p_i^{-1} \underline{p}) x_k^i$. It follows by convexity of g_i that

$$g_P(x') = \sum_{i=1}^{n} p_i g_i(x'^i) \leq \sum_{i=1}^{n} \underline{p} g_i(x^i) + (p_i - \underline{p}) g_i(x_k^i) = \underline{p} g(x) - \underline{p} g(x_k) + g_P(x_k).$$

Moreover, since for any $0 \leq c \leq 1$ and any u, v, it is true that $\|cu + (1-c)v\|^2 = c\|u\|^2 + (1-c)\|v\|^2 - c(1-c)\|u-v\|^2$, we obtain

$$\|x' - x_{k+1}\|_{\tau^{-1}}^2 = \underline{p} \|x - x_{k+1}\|_{\tau^{-1}P^{-1}}^2 + \|x_{k+1} - x_k\|_{\tau^{-1}}^2 - \underline{p} \|x_{k+1} - x_k\|_{\tau^{-1}P^{-1}}^2$$
$$- \underline{p} \|x - x_k\|_{\tau^{-1}P^{-1}}^2 + \underline{p}^2 \|x - x_k\|_{\tau^{-1}P^{-2}}^2.$$

Lastly, plugging in $x' = x_k + P^{-1}\underline{p}(x - x_k)$ to $\|x' - x_k\|^2$ gives

$$-\|x' - x_k\|_{\tau^{-1}}^2 = -\underline{p}^2 \|x - x_k\|_{\tau^{-1}P^{-2}}^2. \qquad \square$$

We are now ready to prove Lemma 1 which describes the one iteration behavior of the algorithm. We restate the lemma and provide its proof.

Lemma 1. *Let Assumption 1 hold. Recall the definitions of D_p and D_d from Section 2.1 and let $\theta = \mathrm{diag}(\theta_1, \ldots, \theta_m)$ and $\pi = \mathrm{diag}(\pi_1, \ldots, \pi_m)$ be chosen as*

$$\theta_j = \frac{\pi_j}{\underline{p}}, \quad \text{where } \pi_j = \sum_{i \in I(j)} p_i \text{ and } \underline{p} = \min_i p_i.$$

We define the functions, given z,

$$V(z) = \frac{\underline{p}}{2} \|x\|_{\tau^{-1}P^{-1}}^2 + \frac{\underline{p}}{2} \|y\|_{\sigma^{-1}\pi^{-1}}^2,$$

$$\tilde{V}(z) = \frac{\underline{p}}{2}\|x\|^2_{C(\tau)} + \frac{\underline{p}}{2}\|y\|^2_{\sigma^{-1}},$$

where $C(\tau)_i = \frac{2p_i}{\underline{p}\tau_i} - \frac{1}{\tau_i} - p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2 - \frac{\beta_i p_i}{\underline{p}}$.

Then, for the iterates of Algorithm 1, $\forall z \in \mathcal{Z}$, it holds that:

$$\mathbb{E}_k\left[D_p(x_{k+1}; z)\right] + \underline{p}D_d(\bar{y}_{k+1}; z) + \mathbb{E}_k\left[V(z_{k+1} - z)\right] \leq (1 - \underline{p})D_p(x_k; z) + V(z_k - z) - \tilde{V}(\bar{z}_{k+1} - z_k).$$

Proof. By the definition of proximal operator and convexity, $\forall x' \in \mathcal{X}, \forall y \in \mathcal{Y}$,

$$p_i g_i(x'^i) \geq p_i g_i(\bar{x}_{k+1}^i) - p_i\langle \nabla_i f(x_k) + (A^\top \bar{y}_{k+1})_i, x'^i - \bar{x}_{k+1}^i\rangle$$
$$+ \frac{1}{2}\left(\|x_k^i - \bar{x}_{k+1}^i\|^2_{\tau_i^{-1}p_i} + \|x'^i - \bar{x}_{k+1}^i\|^2_{\tau_i^{-1}p_i} - \|x'^i - x_k^i\|^2_{\tau_i^{-1}p_i}\right)$$
$$\underline{p}h^*(y) \geq \underline{p}h^*(\bar{y}_{k+1}) + \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle + \frac{\underline{p}}{2}\left(\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \|y - \bar{y}_{k+1}\|^2_{\sigma^{-1}} - \|y - y_k\|^2_{\sigma^{-1}}\right).$$

We sum the first inequality for $i = 1$ to n, then add it to the second inequality and use the definition $g_P(x) = \sum_{i=1}^n p_i g_i(x^i)$ to derive

$$g_P(x') + \underline{p}h^*(y) \geq g_P(\bar{x}_{k+1}) + \underline{p}h^*(\bar{y}_{k+1}) - \langle \nabla f(x_k), P(x' - \bar{x}_{k+1})\rangle - \langle A^\top \bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle$$
$$+ \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle + \frac{1}{2}\left(\|x_k - \bar{x}_{k+1}\|^2_{\tau^{-1}P} + \|x' - \bar{x}_{k+1}\|^2_{\tau^{-1}P} - \|x' - x_k\|^2_{\tau^{-1}P}\right)$$
$$+ \frac{\underline{p}}{2}\left(\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \|y - \bar{y}_{k+1}\|^2_{\sigma^{-1}} - \|y - y_k\|^2_{\sigma^{-1}}\right). \tag{20}$$

First, we note that for $\mathcal{F}_k$-measurable X and any $\gamma = \text{diag}(\gamma_1, \ldots, \gamma_n)$, such that $\gamma_i > 0$, the following hold

$$\mathbb{E}_k[g(x_{k+1})] = g_P(\bar{x}_{k+1}) - g_P(x_k) + g(x_k), \tag{21}$$
$$\mathbb{E}_k[x_{k+1}] = P\bar{x}_{k+1} - Px_k + x_k,$$
$$\mathbb{E}_k\left[\|x_{k+1} - X\|^2_\gamma\right] = \|\bar{x}_{k+1} - X\|^2_{\gamma P} - \|x_k - X\|^2_{\gamma P} + \|x_k - X\|^2_\gamma. \tag{22}$$

We use (22) with $\gamma = \tau^{-1}$ and $X = x'$ to obtain

$$\frac{1}{2}\left(\|x_k - \bar{x}_{k+1}\|^2_{\tau^{-1}P} + \|x' - \bar{x}_{k+1}\|^2_{\tau^{-1}P} - \|x' - x_k\|^2_{\tau^{-1}P}\right)$$
$$= \frac{1}{2}\left(\|x_k - \bar{x}_{k+1}\|^2_{\tau^{-1}P} + \mathbb{E}_k\left[\|x' - x_{k+1}\|^2_{\tau^{-1}}\right] - \|x' - x_k\|^2_{\tau^{-1}}\right). \tag{23}$$

We use $\gamma = \pi^{-1}\sigma^{-1}$ and $Y = y$ in Lemma 2, then

$$\|\bar{y}_{k+1} - y\|^2_{\sigma^{-1}} = \mathbb{E}_k\left[\|y_{k+1} - y\|^2_{\sigma^{-1}\pi^{-1}}\right] + \|y_k - y\|^2_{\sigma^{-1}} - \|y_k - y\|^2_{\sigma^{-1}\pi^{-1}} - 2\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle$$
$$- \sum_{i=1}^n \sum_{j=1}^m p_i \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2. \tag{24}$$

We let $x'^i = p_i^{-1}\underline{p}x^i + (1 - p_i^{-1}\underline{p})x_k^i$, and use Lemma 3 to get

$$g_P(x') \leq \underline{p}g(x) - \underline{p}g(x_k) + g_P(x_k). \tag{25}$$

Combined with (21), the last inequality gives

$$g_P(\bar{x}_{k+1}) - g_P(x') \geq \mathbb{E}_k\left[g(x_{k+1})\right] - g(x_k) + \underline{p}g(x_k) - \underline{p}g(x). \tag{26}$$

We further use $x' = P^{-1}\underline{p}x + (1 - P^{-1}\underline{p})x_k = x_k + P^{-1}\underline{p}(x - x_k)$ in (20) to obtain

$$- \langle \nabla f(x_k), P(x' - \bar{x}_{k+1})\rangle = -\underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle, \tag{27}$$

and

$$-\langle A^\top \bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle = -\underline{p}\langle A^\top \bar{y}_{k+1}, x - x_k\rangle - \langle A^\top \bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle. \tag{28}$$

Moreover, by using Lemma 3 in (23), we get

$$\frac{1}{2}\mathbb{E}_k\left[\|x' - x_{k+1}\|_{\tau^{-1}}^2\right] - \frac{1}{2}\|x' - x_k\|_{\tau^{-1}}^2 = \frac{p}{2}\mathbb{E}_k\left[\|x - x_{k+1}\|_{\tau^{-1}P^{-1}}^2\right] + \frac{1}{2}\mathbb{E}_k\left[\|x_{k+1} - x_k\|_{\tau^{-1}}^2\right]$$
$$- \frac{p}{2}\mathbb{E}_k\left[\|x_{k+1} - x_k\|_{\tau^{-1}P^{-1}}^2\right] - \frac{p}{2}\|x - x_k\|_{\tau^{-1}P^{-1}}^2. \tag{29}$$

We also note that $\mathbb{E}_k\left[\|x_{k+1} - x_k\|_{\tau^{-1}P^{-1}}^2\right] = \|\bar{x}_{k+1} - x_k\|_{\tau^{-1}}^2$.

In (20), we collect eqs. (23), (24) and (26)–(29) to obtain

$$0 \geq \mathbb{E}_k[g(x_{k+1})] - g(x_k) + \underline{p}g(x_k) - \underline{p}g(x) + \underline{p}h^*(\bar{y}_{k+1}) - \underline{p}h^*(y) + \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle - \underline{p}\langle\nabla f(x_k), x - x_k\rangle$$
$$- \langle\nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle - \underline{p}\langle A^\top\bar{y}_{k+1}, x - x_k\rangle - \langle A^\top\bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle$$
$$+ \frac{p}{2}\mathbb{E}_k\left[\|x_{k+1} - x\|_{\tau^{-1}P^{-1}}^2\right] - \frac{p}{2}\|x_k - x\|_{\tau^{-1}P^{-1}}^2 + \|\bar{x}_{k+1} - x_k\|_{\tau^{-1}P}^2 - \frac{p}{2}\|\bar{x}_{k+1} - x_k\|_{\tau^{-1}}^2$$
$$+ \frac{p}{2}\|\bar{y}_{k+1} - y_k\|_{\sigma^{-1}}^2 + \frac{p}{2}\mathbb{E}_k\left[\|y_{k+1} - y\|_{\sigma^{-1}\pi^{-1}}^2\right] - \frac{p}{2}\|y_k - y\|_{\sigma^{-1}\pi^{-1}}^2 - \langle\bar{y}_{k+1} - y, \underline{p}\pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle$$
$$- \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^m \underline{p}p_i\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2. \tag{30}$$

We use coordinatewise smoothness of f to obtain

$$-\langle\nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle = -\langle\nabla f(x_k), \mathbb{E}_k[x_k - x_{k+1}]\rangle \geq \mathbb{E}_k\left[f(x_{k+1}) - f(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_\beta^2\right]$$
$$\geq \mathbb{E}_k[f(x_{k+1})] - f(x_k) - \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2. \tag{31}$$

Next, by using the definition of $\tilde{V}$ and $C(\tau)$, we note

$$\tilde{V}(\bar{z}_{k+1} - z_k) = \frac{p}{2}\|\bar{y}_{k+1} - y_k\|_{\sigma^{-1}}^2 + \|\bar{x}_{k+1} - x_k\|_{\tau^{-1}P}^2 - \frac{p}{2}\|\bar{x}_{k+1} - x_k\|_{\tau^{-1}}^2$$
$$- \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^m \underline{p}p_i\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2 - \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2. \tag{32}$$

Using definition of V, D_p, D_d (see Section 2.1) in (30) along with $-\underline{p}\langle\nabla f(x_k), x - x_k\rangle \geq -\underline{p}f(x) + \underline{p}f(x_k)$ and the last inequality yields

$$0 \geq \mathbb{E}_k\left[D_p(x_{k+1}; z)\right] - (1 - \underline{p})D_p(x_k; z) + \underline{p}D_d(\bar{y}_{k+1}; z) + \tilde{V}(\bar{z}_{k+1} - z_k) + \mathbb{E}_k\left[V(z_{k+1} - z)\right] - V(z_k - z)$$
$$+ \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle - \underline{p}\langle A^\top\bar{y}_{k+1}, x - x_k\rangle - \langle A^\top\bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle - \langle\bar{y}_{k+1} - y, \underline{p}\pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle$$
$$- \langle A^\top y, x_{k+1} - x\rangle + \underline{p}\langle Ax, \bar{y}_{k+1} - y\rangle + \langle A^\top y, x_k - x\rangle - \underline{p}\langle A^\top y, x_k - x\rangle. \tag{33}$$

We work on the bilinear terms to get

$$\langle A^\top y, x_k - x\rangle - \langle A^\top y, x_{k+1} - x\rangle - \langle A^\top\bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle$$
$$= \langle A^\top y, x_k - x_{k+1}\rangle - \langle A^\top\bar{y}_{k+1}, \mathbb{E}_k[x_k - x_{k+1}]\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta A\mathbb{E}_k[x_k - x_{k+1}]\rangle$$
$$= \mathbb{E}_k\left[\langle A^\top(y - \bar{y}_{k+1}), x_k - x_{k+1}\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta A(x_k - x_{k+1})\rangle\right] = 0, \tag{34}$$

where the last equality is due to the requirement in (8) as

$$\underline{p}\pi^{-1}\theta = I \iff \theta_j = \frac{\pi_j}{\underline{p}}, \quad \forall j \in \{1, \ldots, m\}.$$

We continue to estimate the remaining bilinear terms as

$$\underline{p}\left[\langle Ax_k, y - \bar{y}_{k+1}\rangle - \langle A^\top \bar{y}_{k+1}, x - x_k\rangle + \langle Ax, \bar{y}_{k+1} - y\rangle - \langle A^\top y, x_k - x\rangle\right] =$$
$$\underline{p}\left[\langle A^\top(y - \bar{y}_{k+1}), x - x_k\rangle + \langle y - \bar{y}_{k+1}, A(x_k - x)\rangle\right] = 0. \quad (35)$$

We use the last estimation, eqs. (34) and (35) in (33) to finish the proof. □

B.2. Proof for almost sure convergence

We include the statement and the proof of Theorem 1.

Theorem 1. *Let Assumption 1 hold and let* θ, π *be as in Lemma 1. Choose step sizes* τ, σ *such that*

$$\tau_i < \frac{2p_i - \underline{p}}{\beta_i p_i + \underline{p}^{-1} p_i \sum_{j=1}^m \pi_j \sigma_j A_{j,i}^2}.$$

Then, almost surely, there exist $z_\star \in \mathcal{Z}_\star$ *such that* $z_k \to z_\star$.

Proof. Equipped with Lemma 1, we will follow the standard arguments as in Fercoq & Bianchi (2019). We refer to (Fercoq & Bianchi, 2019, Theorem 1) for the finer details of the arguments.

We first invoke the main result of Lemma 1 with $z = z_\star = (x_\star, y_\star)$ where $z_\star \in \mathcal{Z}_\star$:

$$\mathbb{E}_k\left[S_p(x_{k+1})\right] + \underline{p}S_d(\bar{y}_{k+1}) + \mathbb{E}_k\left[V(z_{k+1} - z_\star)\right] \leq (1 - \underline{p})S_p(x_k) + V(z_k - z_\star) - \tilde{V}(\bar{z}_{k+1} - z_k), \quad (36)$$

where we have used the definitions

$$S_p(x_{k+1}) = D_p(x_{k+1}; z_\star) \geq 0, \quad S_d(\bar{y}_{k+1}) = D_d(\bar{y}_{k+1}; z_\star) \geq 0.$$

Nonnegativity of these quantities follow from the definition of $z_\star$ as in (4) and convexity.

Moreover, we recall the definitions of $V, \tilde{V}$ from Lemma 1. Then, we see that $\tilde{V}(z)$ is nonnegative by the choice of step sizes τ, σ and $\theta_j = \frac{\pi_j}{\underline{p}}$. It is also immediate that $V(z)$ is nonnegative.

Then, we can write (36) as

$$\mathbb{E}_k\left[S_p(x_{k+1}) + V(z_{k+1} - z_\star)\right] \leq S_p(x_k) + V(z_k - z_\star) - \underline{p}\left(S_p(x_k) + S_d(\bar{y}_{k+1})\right). \quad (37)$$

We use Robbins-Siegmund lemma on this inequality and nonnegativity of $S_p(x_k), S_d(\bar{y}_k), V(z), \tilde{V}(z)$ to conclude that $V(z_k - z_\star)$ converges almost surely, $\sum_k S_p(x_k) + S_d(\bar{y}_{k+1}) < \infty$, therefore $S_p(x_k)$ and $S_d(\bar{y}_k)$ converges to 0 almost surely. Then, we strengthen these conclusion by arguing as (Fercoq & Bianchi, 2019, Theorem 1), (Iutzeler et al., 2013), (Combettes & Pesquet, 2015, Proposition 2.3), to get the result: there exist Ω_1 with $\mathbb{P}(\Omega_1) = 1$, such that $\forall w \in \Omega_1$ and $\forall z_\star \in \mathcal{Z}_\star$, $V(z_k(\omega) - z_\star)$ converges.

We now take full expectation of (36), use the nonnegativity of $S_p(x_k), S_d(\bar{y}_k)$, and sum the inequality to obtain,

$$\sum_{k=0}^\infty \mathbb{E}\left[\tilde{V}(\bar{z}_{k+1} - z_k)\right] \leq S_p(x_0) + V(z_0 - z_\star) := \Delta_0 < \infty. \quad (38)$$

By Fubini-Tonelli theorem, $\mathbb{E}\left[\sum_{k=0}^\infty \tilde{V}(\bar{z}_{k+1} - z_k)\right] < \infty$. Since $\tilde{V}(\bar{z}_{k+1} - z_k)$ is nonnegative we have that $\tilde{V}(\bar{z}_{k+1} - z_k)$ converges almost surely to 0. Then, by the definition of $\tilde{V}(z)$ and the choice of step sizes, it follows that $\bar{z}_{k+1} - z_k$ converges to 0 almost surely. We say that there exist Ω_2 with $\mathbb{P}(\Omega_2) = 1$ such that $\forall \omega \in \Omega_2$, $\bar{z}_{k+1}(\omega) - z_k(\omega) \to 0$.

We define $T: \mathcal{Z} \to \mathcal{Z}$ such that

$$\bar{y} = \text{prox}_{\sigma, h^*}(y + \sigma Ax),$$
$$\bar{x} = \text{prox}_{\tau, g}\left(x - \tau(\nabla f(x) + A^\top \bar{y})\right).$$

It is easy to see that $\bar{z}_{k+1} = T(z_k)$.

We use the definition of proximal operator in the definition of T and compare with the definition of a saddle point in (4) to conclude that the fixed points of T correspond to the set of saddle points $\mathcal{Z}_\star$.

We now fix $\omega \in \Omega_1 \cap \Omega_2$ and then it follows that $z_k(\omega)$ is a bounded sequence, from what we have proved beforehand. As $z_k(\omega)$ is bounded, it converges on at least one subsequence. We denote by $\check{z}$ the cluster point of this subsequence. By the fact that $\bar{z}_{k+1}(\omega) = T(z_k(\omega))$ and $\bar{z}_{k+1}(\omega) - z_k(\omega) \to 0$, we conclude that $T(z_k(\omega)) - z_k(\omega) \to 0$. As T is continuous, by the nonexpansiveness of proximal operator, we get $\check{z}$ is a fixed point of T and that $\check{z} \in \mathcal{Z}_\star$.

Since we know that for any $\omega \in \Omega_1 \cap \Omega_2$ and for any $z_\star \in \mathcal{Z}_\star$, $V(z_k(\omega) - z_\star)$ converges, and we have proved $V(z_k(\omega) - \check{z})$ converges to 0 at least on one subsequence with $\check{z} \in \mathcal{Z}_\star$, we conclude that the sequence z_k converges to a point $\check{z} \in \mathcal{Z}_\star$, almost surely. $\qquad\square$

B.3. Proof for linear convergence

In this section, we include the statement and the proof of Theorem 2. First, we need a lemma to characterize the specific choice of projection onto the solution set, given in (10).

Lemma 4. *Let us denote*

$$z_k^\star = \arg\min_{u \in \mathcal{Z}_\star} D_p(x_k; u) + V(z_k - u).$$

$$z_k^{\star,e} = \arg\min_{u \in \mathcal{Z}_\star} V(z_k - u).$$

We have that $V^{1/2}(z_k^\star - z^{\star,e}) \le cV^{1/2}(z_k - z^{\star,e})$, where $c = C_{2,V}\sqrt{\frac{\|A\|}{2}}$ and $C_{2,V}$ is such that for any z, $\|z\| \le C_{2,V}V(z)^{1/2}$.

Proof. Let us first remark that since $u = (u_x, u_y) \in \mathcal{Z}_\star$ is a saddle point of the Lagrangian $L(x, y) = f(x) + g(x) + \langle Ax, y \rangle - h^*(y)$, we have that $D_p(x_k, u) = L(x_k, u_y) - L(u_x, u_y)$ is independent of u_x and affine in u_y. In particular, for any primal solution $x_\star$, there exists a constant, $C(x_k, x_\star) = f(x_k) + g(x_k) - f(x_\star) - g(x_\star)$ such that $D_p(x_k, u) = C(x_k, x_\star) + \langle A(x_k - x_\star), u_y \rangle$, for all $u \in \mathcal{Z}_\star$. We have also used here the fact that for two different primal solutions $x_{\star,1}, x_{\star,2}$, it follows that $L(x_{\star,1}, u_y) = L(x_{\star,2}, u_y)$, where u_y is a dual solution.

By prox inequality, we have for any $z \in \mathcal{Z}$,

$$C(x_k, x_\star) + \langle A(x_k - x_\star), y_k^\star \rangle + V(z - z_k^\star) \le C(x_k, x_\star) + \langle A(x_k - x_\star), y \rangle + V(z_k - z) - V(z_k^\star - z_k),$$
$$V(z_k^{\star,e} - z) \le V(z_k - z) - V(z_k^{\star,e} - z_k),$$

where $x_\star$ is any primal solution. For the first inequality, we plug in $z = z_k^{\star,e}$ and for the second inequality, we plug in $z = z_k^\star$. Summing both equalities and rearranging yields

$$2V(z_k^{\star,e} - z_k^\star) \le \langle A(x_k - x_\star), y_k^{\star,e} - y_k^\star \rangle$$

Since the inequality holds for any $x_\star \in \mathcal{X}_\star$, we can plug in $x_k^{\star,e}$ and use Cauchy-Schwarz inequality to get

$$2V(z_k^{\star,e} - z_k^\star) \le C_{2,V}^2 \|A\| V(z_k - z_k^{\star,e})^{1/2} V(z_k^\star - z_k^{\star,e})^{1/2}.$$

Lastly $C_{2,V} = \sqrt{\frac{2}{p\min\{\min_i \tau_i^{-1} p_i^{-1}, \min_j \sigma_j^{-1}\pi_j^{-1}\}}}$. $\qquad\square$

Theorem 2. *Let Assumptions 1 and 2 hold. Let θ and the step sizes τ, σ be chosen according to (8) and (9), respectively. Moreover, $z_k^\star = (x_k^\star, y_k^\star)$ is as defined in (10). Then, for z_k generated by Algorithm 1, it follows that*

$$\mathbb{E}\left[\frac{p}{2}\|x_k - x_k^\star\|_{\tau^{-1}P^{-1}}^2 + \frac{p}{2}\|y_k - y_k^\star\|_{\sigma^{-1}\pi^{-1}}^2\right] \le (1 - \rho)^k \Delta_0,$$

where $\Delta_0 = D_p(x_0; z_0^\star) + V(z_0 - z_0^\star)$, $\rho = \min\left(\underline{p}, \frac{C_{2,\tilde{V}}}{C_{V,2}((2+2c)+(1+c)(\eta\|H-M\|+\bar{\beta}))^2}\right)$, $\bar{\beta}$ is the Lipschitz constant of f,

$C_{2,\tilde{V}} = \frac{p}{2}\min\left\{\min_i C(\tau)_i, \min_j \sigma_j^{-1}\right\}$, $C_{V,2} = \frac{1}{2}\max\left\{\max_i \frac{1}{\tau_i}, \max_j \frac{1}{\sigma_j}\right\}$, $C_{2,V} = \sqrt{\frac{2}{\underline{p}\min\{\min_i \tau_i^{-1} p_i^{-1}, \min_j \sigma_j^{-1}\pi_j^{-1}\}}}$,

$c = C_{2,V} \sqrt{\|A\|/2}$, *and*

$$H = \begin{bmatrix} \tau^{-1} & A^\top \\ 0 & \sigma^{-1} \end{bmatrix}, \quad M = \begin{bmatrix} 0 & A^\top \\ -A & 0 \end{bmatrix}.$$

Proof. We note the definitions, as in (Latafat et al., 2019),

$$A\colon (x,y) \mapsto (\partial g(x) + \partial h^*(y))$$
$$M\colon (x,y) \mapsto (A^\top y, -Ax)$$
$$C\colon (x,y) \mapsto (\nabla f(x), 0)$$
$$H\colon (x,y) \mapsto (\tau^{-1}x + A^\top y, \sigma^{-1}y).$$

Under this notations, KKT operator defined in (4) can be written as

$$F = A + M + C. \tag{39}$$

Moreover, $\bar{z}_{k+1} = (H+A)^{-1}(H - M - C)z_k$, which in fact (without the term $\nabla f(x_k)$) is the well-known Arrow-Hurwicz operator (Arrow et al., 1958).

Moreover, we will use the following inequalities regarding V and $\tilde{V}$ (see Lemma 1 for the definitions)

$$\tilde{V}(z) \geq C_{2,\tilde{V}} \left(\|x\|^2 + \|y\|^2 \right) := \frac{p}{2} \min \left\{ \min_i C(\tau)_i, \min_j \sigma_j^{-1} \right\} \left(\|x\|^2 + \|y\|^2 \right), \tag{40}$$

$$V(z) \leq C_{V,2} \left(\|x\|^2 + \|y\|^2 \right) := \frac{1}{2} \max \left\{ \max_i \frac{1}{\tau_i}, \max_j \frac{1}{\sigma_j} \right\} \left(\|x\|^2 + \|y\|^2 \right). \tag{41}$$

We recall the definition of $z_k^\star$

$$z_k^\star = \arg\min_{u \in \mathcal{Z}_\star} D_p(x_k; u) + V(z_k - u).$$

We now argue that $z_k^\star$ is well-defined under our assumptions. Under Assumption 1, we know that the solution set is convex and closed; $D_p(x_k; u) \geq 0$ for all $u \in \mathcal{Z}_\star$ and it is also lower semicontinuous. Next, we remark that $V(z_k - u)$ is a squared norm, thus coercive, therefore the sum is coercive and lower semicontinuous over $\mathcal{Z}_\star$. Hence, $z_k^\star$ exists.

We use the result of Lemma 1 with $z = z_k^\star$ and $D_d(\bar{y}_{k+1}; z_k^\star) \geq 0$

$$\mathbb{E}_k \left[D_p(x_{k+1}; z_k^\star) + V(z_{k+1} - z_k^\star) \right] \leq (1 - \underline{p}) D_p(x_k; z_k^\star) + V(z_k - z_k^\star) - \tilde{V}(\bar{z}_{k+1} - z_k).$$

We use the definition of $z_{k+1}^\star$ to deduce

$$\mathbb{E}_k \left[D_p(x_{k+1}; z_{k+1}^\star) + V(z_{k+1} - z_{k+1}^\star) \right] \leq (1 - \underline{p}) D_p(x_k; z_k^\star) + V(z_k - z_k^\star) - \tilde{V}(\bar{z}_{k+1} - z_k). \tag{42}$$

In addition to Bregman projections $z_k^\star$ and $\bar{z}_{k+1}^\star$, we introduce the definitions for Euclidean projections (see Lemma 4)

$$z_k^{\star,e} = \arg\min_{u \in \mathcal{Z}_\star} V(z_k - u),$$
$$\bar{z}_{k+1}^{\star,e} = \arg\min_{u \in \mathcal{Z}_\star} V(\bar{z}_{k+1} - u).$$

Now, we use triangle inequalities and Lemma 4 to get

$$V(z_k - z_k^\star)^{1/2} \leq V(z_k - z_k^{\star,e})^{1/2} + V(z_k^{\star,e} - z_k^\star)^{1/2} \leq (1 + c) V(z_k - z_k^{\star,e})^{1/2}$$
$$\leq (1 + c)(V(z_k - \bar{z}_{k+1})^{1/2} + V(\bar{z}_{k+1} - \bar{z}_{k+1}^{\star,e})^{1/2} + V(z_k^{\star,e} - \bar{z}_{k+1}^{\star,e})^{1/2})$$

We use nonexpansiveness with metric V on the last term, to obtain

$$V(z_k - z_k^\star)^{1/2} \leq (2 + 2c) V(z_k - \bar{z}_{k+1})^{1/2} + (1 + c) V(\bar{z}_{k+1}^{\star,e} - \bar{z}_{k+1})^{1/2}.$$

We use the definition of $\bar{z}_{k+1}^{\star,e}$ to get $V(\bar{z}_{k+1} - \bar{z}_{k+1}^{\star,e}) \leq V(\bar{z}_{k+1} - P_{\mathcal{Z}_\star}(\bar{z}_{k+1}))$, where $P_{\mathcal{Z}_\star}$ is defined to be standard Euclidean projection, and then we use the relation between V and Euclidean norm from (41)

$$V(z_k - z_k^\star)^{1/2} \leq (2 + 2c)\sqrt{C_{V,2}}\|z_k - \bar{z}_{k+1}\| + (1 + c)\sqrt{C_{V,2}}\|P_{\mathcal{Z}_\star}(\bar{z}_{k+1}) - \bar{z}_{k+1}\|$$
$$= (2 + 2c)\sqrt{C_{V,2}}\|z_k - \bar{z}_{k+1}\| + (1 + c)\sqrt{C_{V,2}}\,\text{dist}(\bar{z}_{k+1}, \mathcal{Z}_\star). \tag{43}$$

We now use metric subregularity of F for 0 (see Assumption 2), and the assumption that $\bar{z}_{k+1} \in \mathcal{N}(z_\star), \forall z_\star$,

$$\text{dist}(\bar{z}_{k+1}, \mathcal{Z}_\star) \leq \eta\,\text{dist}(0, F(\bar{z}_{k+1})).$$

We now use that $(H - M - C)(z_k - \bar{z}_{k+1}) \in F(\bar{z}_{k+1})$, which can be obtained by using (39) and $\bar{z}_{k+1} = (H + A)^{-1}(H - M - C)z_k$. Therefore

$$\text{dist}(\bar{z}_{k+1}, \mathcal{Z}_\star) \leq \eta\|(H - M - C)(z_k - \bar{z}_{k+1})\| \leq \eta(\|H - M\| + \bar{\beta})\|z_k - \bar{z}_{k+1}\|,$$

where $\bar{\beta}$ is the global Lipschitz constant of f.

We plug this inequality into (43) to obtain

$$V(z_k - z_k^\star) \leq C_{V,2}((2 + 2c) + (1 + c)(\eta\|H - M\| + \bar{\beta}))^2\|\bar{z}_{k+1} - z_k\|^2.$$

Moreover, since $\tilde{V}(\bar{z}_{k+1} - z_k)$ is a squared norm, under the step size condition, it follows that $\tilde{V}(\bar{z}_{k+1} - z_k) \geq C_{2,\tilde{V}}\|\bar{z}_{k+1} - z_k\|^2$, as in (40), therefore,

$$V(z_k - z_k^\star) \leq \frac{C_{V,2}((2 + 2c) + (1 + c)(\eta\|H - M\| + \bar{\beta}))^2}{C_{2,\tilde{V}}}\tilde{V}(\bar{z}_{k+1} - z_k).$$

We use this inequality in (42) to obtain

$$\mathbb{E}_k\left[D_p(x_{k+1}; z_{k+1}^\star) + V(z_{k+1} - z_{k+1}^\star)\right] \leq (1 - \underline{p})D_p(x_k; z_k^\star)$$
$$+ \left(1 - \frac{C_{2,\tilde{V}}}{C_{V,2}((2 + 2c) + (1 + c)(\eta\|H - M\| + \bar{\beta}))^2}\right)V(z_k - z_k^\star),$$

where the constants $C_{2,\tilde{V}}, C_{V,2}$ are as defined in (40), (41).

We take full expectation and define $\rho = \min\left(\underline{p}, \frac{C_{2,\tilde{V}}}{C_{V,2}((2+2c)+(1+c)(\eta\|H-M\|+\bar{\beta}))^2}\right)$. Then, we have that

$$\mathbb{E}\left[D_p(x_{k+1}; z_{k+1}^\star) + V(z_{k+1} - z_{k+1}^\star)\right] \leq (1 - \rho)\mathbb{E}\left[D_p(x_k; z_k^\star) + V(z_k - z_k^\star)\right].$$

We have that $0 < \rho < 1$, as metric subregularity constant $\eta > 0$. Hence, linear convergence of $D_p(x_k, z_k^\star)$ and $V(z_k - z_k^\star)$ follows. We obtain the final result after using the definition of V, and the fact that $D_p(x_{k+1}; z_{k+1}^\star) \geq 0$. $\qquad\square$

B.4. Ergodic convergence rates

We introduce the following lemma, which establishes the properties of the sequence $\breve{y}_k$, defined in (13).

Lemma 5. *We recall the definition of $\breve{y}_k$ from (13): let $\breve{y}_1 = y_1 = \bar{y}_1$ and*

$$\breve{y}_{k+1}^j = \bar{y}_{k+1}^j, \quad \forall j \in J(i_{k+1})$$
$$\breve{y}_{k+1}^j = \breve{y}_k^j, \qquad \forall j \notin J(i_{k+1}),$$

where computing $\breve{y}_{k+1}$ requires the same number of operations as computing y_{k+1} every iteration, and $\breve{y}_k$ is $\mathcal{F}_k$-measurable. Moreover, if it holds for a function l that $l(y) = \sum_{j=1}^m l_j(y_j)$ and

$$l_\gamma(y) = \sum_{j=1}^m \gamma_j l_j(y_j),$$

we have the following for the sequence $\breve{y}_k$, and $\mathcal{F}_k$-measurable Y:

$$\mathbb{E}_k\left[\breve{y}_{k+1}^j - \breve{y}_k^j\right] = \pi_j\left(\bar{y}_{k+1}^j - \breve{y}_k^j\right), \forall j$$

$$\mathbb{E}_k\left[\|\breve{y}_{k+1} - Y\|_\gamma^2\right] = \|\bar{y}_{k+1} - Y\|_{\gamma\pi}^2 - \|\breve{y}_k - Y\|_{\gamma\pi}^2 + \|\breve{y}_k - Y\|_\gamma^2$$

$$\mathbb{E}_k\left[l(\breve{y}_{k+1})\right] = l_\pi(\bar{y}_{k+1}) - l_\pi(\breve{y}_k) + l(\breve{y}_k)$$

$$\mathbb{E}_k\left[\|\breve{y}_{k+1} - y_{k+1}\|_\gamma^2\right] = \|\bar{x}_{k+1} - x_k\|_{B(\gamma)}^2 + \|\breve{y}_k - y_k\|_\gamma^2 - \|\breve{y}_k - y_k\|_{\gamma\pi}^2$$

$$\sum_{k=1}^K \mathbb{E}\left[\|\breve{y}_{k+1} - \breve{y}_k\|_\gamma^2\right] \leq 2\sum_{k=1}^K \mathbb{E}\left[\|\bar{y}_{k+1} - y_k\|_{\gamma\pi}^2\right] + 2\sum_{k=1}^K \mathbb{E}\left[\|\bar{x}_{k+1} - x_k\|_{B(\gamma)}^2\right]$$

$$\mathbb{E}\left[\|\breve{y}_k - Y\|_{\gamma\pi}^2\right] \leq 2\mathbb{E}\left[\|y_k - Y\|_{\gamma\pi}^2\right] + 2\sum_{k=1}^K \mathbb{E}\left[\|\bar{x}_{k+1} - x_k\|_{B(\gamma)}^2\right],$$

where

$$B(\gamma)_i = p_i \sum_{j=1}^m \theta_j^2 \gamma_j \sigma_j^2 A_{j,i}^2, \text{ and } \pi_j = \sum_{i\in I(j)} p_i.$$

Proof. We first use the definition of $\breve{y}_k$ to get the first result. For any j,

$$\mathbb{E}_k[\breve{y}_{k+1}^j] = \mathbb{E}_k\left[\mathbb{1}_{j\in J(i_{k+1})}\bar{y}_{k+1}^j + \mathbb{1}_{j\notin J(i_{k+1})}\breve{y}_k^j\right] = \sum_{i=1}^n p_i\left[\mathbb{1}_{j\in J(i)}\left(\bar{y}_{k+1}^j\right) + \mathbb{1}_{j\notin J(i)}\breve{y}_k^j\right]$$

$$= \sum_{i\in I(j)} p_i\bar{y}_{k+1}^j + \sum_{i\notin I(j)} p_i\breve{y}_k^j = \sum_{i\in I(j)} p_i\bar{y}_{k+1}^j + \sum_{i=1}^n p_i\breve{y}_k^j - \sum_{i\in I(j)} p_i\breve{y}_k^j$$

$$= \breve{y}_k^j + \pi_j\left(\bar{y}_{k+1}^j - \breve{y}_k^j\right).$$

For the second result, we estimate similar to Lemma 2

$$\mathbb{E}_k\left[\|\breve{y}_{k+1} - Y\|_\gamma^2\right] = \mathbb{E}_k\left[\sum_{j\in J(i_{k+1})} \gamma_j(\bar{y}_{k+1}^j - Y^j)^2 + \sum_{j\notin J(i_{k+1})} \gamma_j(\breve{y}_k^j - Y^j)^2\right]$$

$$= \sum_{i=1}^n p_i\left[\sum_{j\in J(i)} \gamma_j(\bar{y}_{k+1}^j - Y^j)^2 + \sum_{j\notin J(i)} \gamma_j(\breve{y}_k^j - Y^j)^2\right]$$

$$= \sum_{j=1}^m \sum_{i\in I(j)} p_i\gamma_j(\bar{y}_{k+1}^j - Y^j)^2 + \sum_{j=1}^m \sum_{i=1}^n p_i\gamma_j(\breve{y}_k^j - Y^j)^2 - \sum_{j=1}^m \sum_{i\in I(j)} p_i\gamma_j(\breve{y}_k^j - Y^j)^2$$

$$= \|\bar{y}_{k+1} - Y\|_{\gamma\pi}^2 + \|\breve{y}_k - Y\|_\gamma^2 - \|\breve{y}_k - Y\|_{\gamma\pi}^2.$$

We derive the third result using similar estimations

$$\mathbb{E}_k\left[l(\breve{y}_{k+1})\right] = \mathbb{E}_k\left[\sum_{j\in J(i_{k+1})} l_j(\bar{y}_{k+1}^j) + \sum_{j\notin J(i_{k+1})} l_j(\breve{y}_k^j)\right]$$

$$= \sum_{i=1}^n p_i\left[\sum_{j\in J(i)} l_j(\bar{y}_{k+1}^j) + \sum_{j\notin J(i)} l_j(\breve{y}_k^j)\right]$$

$$= \sum_{j=1}^m \pi_j l_j(\bar{y}_{k+1}^j) + \sum_{j=1}^m \sum_{i=1}^n p_i l_j(\breve{y}_k^j) - \sum_{j=1}^m \pi_j l_j(\breve{y}_k^j)$$

$$= l_\pi(\bar{y}_{k+1}) - l_\pi(\breve{y}_k) + l(\breve{y}_k).$$

For the fourth result, we use the definitions of both $\breve{y}_{k+1}$ and y_{k+1} (see Algorithm 1 and (16)),

$$
\begin{aligned}
\mathbb{E}_k \left[\|\breve{y}_{k+1} - y_{k+1}\|_\gamma^2 \right] &= \mathbb{E}_k \left[\sum_{j \in J(i_{k+1})} \gamma_j \left(\bar{y}_{k+1}^j - (\bar{y}_{k+1}^j + \sigma_j \theta_j A_{j,i_{k+1}} (x_{k+1}^{i_{k+1}} - x_k^{i_{k+1}})) \right)^2 + \sum_{j \notin J(i_{k+1})} \gamma_j \left(\breve{y}_k^j - y_k^j \right)^2 \right] \\
&= \sum_{i=1}^n p_i \left[\sum_{j \in J(i)} \gamma_j \left(\sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^i - x_k^i) \right)^2 + \sum_{j \notin J(i)} \gamma_j \left(\breve{y}_k^j - y_k^j \right)^2 \right] \\
&= \sum_{i=1}^n p_i \sum_{j \in J(i)} \gamma_j \sigma_j^2 A_{j,i}^2 \theta_i^2 (\bar{x}_{k+1}^i - x_k^i)^2 + \sum_{i=1}^n p_i \sum_{j=1}^m \gamma_j (\breve{y}_k^j - y_k^j)^2 - \sum_{i=1}^n \sum_{j \in J(i)} p_i \gamma_j (\breve{y}_k^j - y_k^j)^2 \\
&= \|\bar{x}_{k+1} - x_k\|_{B(\gamma)}^2 + \|\breve{y}_k - y_k\|_\gamma^2 - \|\breve{y}_k - y_k\|_{\gamma\pi}^2,
\end{aligned}
\tag{44}
$$

where for the second equality, we used the fact that x_{k+1} is different from x_k only on the coordinate i_{k+1}, which gives

$$
(A(x_{k+1} - x_k))_j = (A((x_{k+1}^{i_{k+1}} - x_k^{i_{k+1}}) e_{i_{k+1}}))_j = A_{j,i_{k+1}} (\bar{x}_{k+1}^{i_{k+1}} - x_k^{i_{k+1}}),
$$

and for the last equality, we noted $A_{j,i} = 0, \forall j \notin J(i)$ and defined

$$
B(\gamma)_i = p_i \sum_{j=1}^m \gamma_j \sigma_j^2 \theta_j^2 A_{j,i}^2.
$$

For the fifth result, we first take full expectation and then sum the inequality (44)

$$
\sum_{k=1}^K \mathbb{E} \left[\|\breve{y}_k - y_k\|_{\gamma\pi}^2 \right] \le \sum_{k=1}^K \mathbb{E} \left[\|\bar{x}_{k+1} - x_k\|_{B(\gamma)}^2 \right] + \|y_1 - \breve{y}_1\|_\gamma^2.
\tag{45}
$$

Then, we write from the second result that

$$
\mathbb{E}_k \left[\|\breve{y}_{k+1} - \breve{y}_k\|_\gamma^2 \right] = \|\bar{y}_{k+1} - \breve{y}_k\|_{\gamma\pi}^2 \le 2\|\bar{y}_{k+1} - y_k\|_{\gamma\pi}^2 + 2\|\breve{y}_k - y_k\|_{\gamma\pi}^2.
$$

We take full expectation and sum to get

$$
\begin{aligned}
\sum_{k=1}^K \mathbb{E} \left[\|\breve{y}_{k+1} - \breve{y}_k\|_\gamma^2 \right] &\le 2 \sum_{k=1}^K \mathbb{E} \left[\|\bar{y}_{k+1} - y_k\|_{\gamma\pi}^2 \right] + 2 \sum_{k=1}^K \mathbb{E} \left[\|\breve{y}_k - y_k\|_{\gamma\pi}^2 \right] \\
&\le 2 \sum_{k=1}^K \mathbb{E} \left[\|\bar{y}_{k+1} - y_k\|_{\gamma\pi}^2 \right] + 2 \sum_{k=1}^K \mathbb{E} \left[\|\bar{x}_{k+1} - x_k\|_{B(\gamma)}^2 \right],
\end{aligned}
$$

where we have used (45) and the fact that $\breve{y}_1 = y_1$.

For the last result, we note that

$$
\mathbb{E} \left[\|\breve{y}_k - Y\|_{\gamma\pi}^2 \right] \le 2\mathbb{E} \left[\|y_k - Y\|_{\gamma\pi}^2 \right] + 2\mathbb{E} \left[\|\breve{y}_k - y_k\|_{\gamma\pi}^2 \right],
$$

and we use (45) with $\breve{y}_1 = y_1$, for the second term. $\qquad \square$

We continue with the restatement and the proof of Theorem 3. This length of the proof is due to the complications discussed earlier. First, as discussed in (Alacaoglu et al., 2019), the order of expectation and supremum requires a special proof which delays taking expectations of the estimates (which prohibits simplifications and results in long expressions). Lemma 6 thus can be seen as a version of Lemma 1 with expectation not taken.

However, this is not enough due to the special structure of our new method suited for sparse settings. In particular, we have to manipulate the terms with dual variable carefully, as we cannot average $\bar{y}_k$ (see Lemma 1). Therefore, the treatment with $\breve{y}_k$, which is characterized in Lemma 5 and Lemma 7, is an intricate part of our proof.

Lemma 6. *Let Assumption 1 hold. Given the definitions of D_p and D_d given from Lemma 1, it follows that*

$$0 \geq D_p(x_{k+1}; z) + \underline{p}D_d(\bar{y}_{k+1}; z) - (1 - \underline{p})D_p(x_k; z) + \tilde{V}(\bar{z}_{k+1} - z_k) + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2 + S_1 + S_2$$
$$+ V(z_{k+1} - z) - V(z_k - z),$$

where

$$
\begin{aligned}
S_1 &= g_P(\bar{x}_{k+1}) - g_P(x_k) - (g(x_{k+1}) - g(x_k)) \\
&\quad - f(x_{k+1}) + f(x_k) - \underline{p}f(x_k) + \underline{p}f(x) - \underline{p}\langle\nabla f(x_k), x - x_k\rangle - \langle\nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle \\
&\quad + \frac{p}{2}\|\bar{x}_{k+1}\|_{\tau^{-1}}^2 - \frac{p}{2}\|x_k\|_{\tau^{-1}}^2 - \left(\frac{p}{2}\|x_{k+1}\|_{\tau^{-1}P^{-1}}^2 - \frac{p}{2}\|x_k\|_{\tau^{-1}P^{-1}}^2\right) \\
&\quad + \frac{p}{2}\|\bar{y}_{k+1}\|_{\sigma^{-1}}^2 - \frac{p}{2}\|y_k\|_{\sigma^{-1}}^2 + \underline{p}\langle\bar{y}_{k+1}, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle + \frac{p}{2}\sum_{i=1}^{n}p_i\sum_{j=1}^{m}\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2 \\
&\quad - \frac{p}{2}\left(\|y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 - \|y_k\|_{\sigma^{-1}\pi^{-1}}^2\right), \\
S_2 &= \langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle + \underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}} \\
&\quad - \underline{p}\langle y, \pi^{-1}\sigma^{-1}(y_k - y_{k+1}) - \sigma^{-1}(y_k - \bar{y}_{k+1}) + \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle.
\end{aligned}
$$

Proof. We now follow the proof of Lemma 1 without taking conditional expectations, similar to ergodic convergence rate proof of Alacaoglu et al. (2019).

First, we have, from (20)

$$
\begin{aligned}
g_P(x') + \underline{p}h^*(y) &\geq \underbrace{g_P(\bar{x}_{k+1}) + \underline{p}h^*(\bar{y}_{k+1}) - \langle A^\top\bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle + \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle}_{T_1} \\
&\quad \underbrace{-\langle\nabla f(x_k), P(x' - \bar{x}_{k+1})\rangle}_{T_2} + \underbrace{\frac{1}{2}\left(\|x_k - \bar{x}_{k+1}\|_{\tau^{-1}P}^2 + \|x' - \bar{x}_{k+1}\|_{\tau^{-1}P}^2 - \|x' - x_k\|_{\tau^{-1}P}^2\right)}_{T_3} \\
&\quad + \underbrace{\frac{p}{2}\left(\|y_k - \bar{y}_{k+1}\|_{\sigma^{-1}}^2 + \|y - \bar{y}_{k+1}\|_{\sigma^{-1}}^2 - \|y - y_k\|_{\sigma^{-1}}^2\right)}_{T_4}.
\end{aligned}
\tag{46}
$$

We start with T_1 and add and subtract $\langle A^\top y, x_{k+1} - x\rangle - \underline{p}\langle Ax, \bar{y}_{k+1} - y\rangle - \langle A^\top y, x_k - x\rangle + \underline{p}\langle A^\top y, x_k - x\rangle + g(x_{k+1}) + g_P(x_k) - g(x_k) - \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle$

$$
\begin{aligned}
T_1 &= g(x_{k+1}) - g(x_k) + g_P(x_k) + g_P(\bar{x}_{k+1}) - g(x_{k+1}) + g(x_k) - g_P(x_k) + \underline{p}h^*(\bar{y}_{k+1}) \\
&\quad - \langle A^\top\bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle + \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle + \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle \\
&\quad + \langle A^\top y, x_{k+1} - x\rangle - \underline{p}\langle Ax, \bar{y}_{k+1} - y\rangle + \langle A^\top y, x_k - x\rangle - \underline{p}\langle A^\top y, x_k - x\rangle \\
&\quad - \langle A^\top y, x_{k+1} - x\rangle + \underline{p}\langle Ax, \bar{y}_{k+1} + y\rangle - \langle A^\top y, x_k - x\rangle + \underline{p}\langle A^\top y, x_k - x\rangle \\
&\quad - \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle.
\end{aligned}
\tag{47}
$$

We first use $x'^i = p_i^{-1}\underline{p}x^i + (1 - p_i^{-1}\underline{p})x_k^i$ as in Lemma 3 to get

$$-\langle A^\top\bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle = -\underline{p}\langle A^\top\bar{y}_{k+1}, x - x_k\rangle - \langle A^\top\bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle.$$

Next, we use that

$$
\begin{aligned}
\underline{p}\left[-\langle A^\top y, x_k - x\rangle + \langle Ax, \bar{y}_{k+1} - y\rangle + \langle Ax_k, y - \bar{y}_{k+1}\rangle - \langle A^\top\bar{y}_{k+1}, x - x_k\rangle\right] &= \\
\underline{p}\left[\langle A^\top(y - \bar{y}_{k+1}), x - x_k\rangle + \langle y - \bar{y}_{k+1}, A(x_k - x)\rangle\right] &= 0,
\end{aligned}
$$

to obtain

$$\begin{aligned}
T_1 = {}& g(x_{k+1}) - g(x_k) + g_P(x_k) + g_P(\bar{x}_{k+1}) - g(x_{k+1}) + g(x_k) - g_P(x_k) + \underline{p}h^*(\bar{y}_{k+1}) \\
& - \langle A^\top \bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle + \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta A P(x_k - \bar{x}_{k+1})\rangle \\
& + \langle A^\top y, x_{k+1} - x\rangle - \underline{p}\langle Ax, \bar{y}_{k+1} - y\rangle + \langle A^\top y, x_k - x\rangle \\
& - \langle A^\top y, x_{k+1} - x\rangle - \langle A^\top y, x_k - x\rangle + \underline{p}\langle A^\top y, x_k - x\rangle \\
& - \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta A P(x_k - \bar{x}_{k+1})\rangle.
\end{aligned} \tag{48}$$

We use $\theta_j = \frac{\pi_j}{\underline{p}}$ to deduce

$$\langle A^\top y, x_k - x_{k+1}\rangle - \langle A^\top \bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta A P(x_k - \bar{x}_{k+1})\rangle = -\langle y, A P(x_k - \bar{x}_{k+1})\rangle. \tag{49}$$

We use the last identity and recall the definitions of D_p and D_d to write T_1 as

$$\begin{aligned}
T_1 = {}& D_p(x_{k+1}; z) - f(x_{k+1}) + f(x) + \underline{p}D_d(\bar{y}_{k+1}; z) + \underline{p}h^*(y) - (1-\underline{p})D_p(x_k; z) - \underline{p}(g(x_k) - g(x)) + g_P(x_k) \\
& + g_P(\bar{x}_{k+1}) - g(x_{k+1}) + g(x_k) - g_P(x_k) + \langle y, A(x_k - x_{k+1}) - A P(x_k - \bar{x}_{k+1})\rangle \\
& + \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta A P(x_k - \bar{x}_{k+1})\rangle + (1-\underline{p})(f(x_k) - f(x)).
\end{aligned}$$

Second, for T_2, we use $x' = P^{-1}\underline{p}x + (1 - P^{-1}\underline{p})x_k = x_k + P^{-1}\underline{p}(x - x_k)$ to obtain

$$T_2 = -\langle \nabla f(x_k), P(x' - \bar{x}_{k+1})\rangle = -\underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle.$$

We now combine these two estimates

$$\begin{aligned}
T_1 + T_2 = {}& D_p(x_{k+1}; z) + \underline{p}D_d(\bar{y}_{k+1}; z) - (1-\underline{p})D_p(x_k; z) + g_P(x_k) + \underline{p}g(x) - \underline{p}g(x_k) + \underline{p}h^*(y) \\
& - f(x_{k+1}) + f(x_k) - \underline{p}f(x_k) + \underline{p}f(x) - \underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle \\
& + g_P(\bar{x}_{k+1}) - g(x_{k+1}) + g(x_k) - g_P(x_k) + \langle y, A(x_k - x_{k+1}) - A P(x_k - \bar{x}_{k+1})\rangle \\
& + \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta A P(x_k - \bar{x}_{k+1})\rangle.
\end{aligned} \tag{50}$$

We now work on T_3 in (46), in order to make terms depending on x telescope. First, we note that by Lemma 3, with the slight change of using $\bar{x}_{k+1}$ instead of x_{k+1} and $\tau^{-1}P$ instead of τ^{-1} in the metric, we get

$$\begin{aligned}
\frac{1}{2}\|x' - \bar{x}_{k+1}\|^2_{\tau^{-1}P} - \frac{1}{2}\|x' - x_k\|^2_{\tau^{-1}P} = {}& \frac{\underline{p}}{2}\|x - \bar{x}_{k+1}\|^2_{\tau^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}} \\
& + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}}.
\end{aligned}$$

Thus, on T_3, we add and subtract $\frac{\underline{p}}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}}$

$$\begin{aligned}
T_3 = {}& \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} + \frac{\underline{p}}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}} \\
& + \frac{\underline{p}}{2}\|x - \bar{x}_{k+1}\|^2_{\tau^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}} - \left(\frac{\underline{p}}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}}\right) \\
= {}& \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} + \frac{\underline{p}}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}} + \frac{\underline{p}}{2}\|\bar{x}_{k+1}\|^2_{\tau^{-1}} \\
& - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}} + \underline{p}\langle x, x_k - \bar{x}_{k+1}\rangle_{\tau^{-1}} - \left(\frac{\underline{p}}{2}\|x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}P^{-1}} + \underline{p}\langle x, (x_k - x_{k+1})\rangle_{\tau^{-1}P^{-1}}\right) \\
& + \frac{\underline{p}}{2}\|\bar{x}_{k+1}\|^2_{\tau^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}} - \left(\frac{\underline{p}}{2}\|x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}P^{-1}}\right) + \underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}}.
\end{aligned} \tag{51}$$

We estimate T_4 in (46) similarly. First note that $\theta_j = \frac{\pi_j}{\underline{p}}$ and on T_4, we add and subtract $\frac{\underline{p}}{2}\Big(\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \|y - $

$y_k\|^2_{\sigma^{-1}\pi^{-1}} + 2\langle y - \bar{y}_{k+1}, \pi^{-1}\theta A P(x_k - \bar{x}_{k+1})\rangle + \sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}\left(\bar{x}_{k+1}^i - x_k^i\right)^2\Big)$

$T_4 = \frac{\underline{p}}{2}\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \frac{\underline{p}}{2}\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \frac{\underline{p}}{2}\|y - y_k\|^2_{\sigma^{-1}\pi^{-1}}$

$$+ \frac{p}{2}\|y - \bar{y}_{k+1}\|_{\sigma^{-1}}^2 - \frac{p}{2}\|y - y_k\|_{\sigma^{-1}}^2 + \underline{p}\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle + \frac{p}{2}\sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\theta_j^2\sigma_j A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$

$$- \left(\frac{p}{2}\|y - y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 - \frac{p}{2}\|y - y_k\|_{\sigma^{-1}\pi^{-1}}^2\right)$$

$$- \underline{p}\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle - \frac{p}{2}\sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$

$$= \frac{p}{2}\|y_k - \bar{y}_{k+1}\|_{\sigma^{-1}}^2 + \frac{p}{2}\|y - y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 - \frac{p}{2}\|y - y_k\|_{\sigma^{-1}\pi^{-1}}^2$$

$$+ \frac{p}{2}\|\bar{y}_{k+1}\|_{\sigma^{-1}}^2 - \frac{p}{2}\|y_k\|_{\sigma^{-1}}^2 + \underline{p}\langle \bar{y}_{k+1}, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle + \frac{p}{2}\sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$

$$+ \underline{p}\langle y, y_k - \bar{y}_{k+1}\rangle_{\sigma^{-1}} - \underline{p}\langle y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle$$

$$- \frac{p}{2}\left(\|y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 - \|y_k\|_{\sigma^{-1}\pi^{-1}}^2\right) - \underline{p}\langle y, y_k - y_{k+1}\rangle_{\sigma^{-1}\pi^{-1}}$$

$$- \underline{p}\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle - \frac{p}{2}\sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2. \tag{52}$$

To simplify, let us introduce some more definitions to have simpler expression when we combine T_1, T_2, T_3, T_4 from eqs. (50)–(52). On a high level, S_1 will collect zero mean terms independent of (x, y) and S_2 will collect zero mean terms dependent on (x, y).

$$S_1 = g_P(\bar{x}_{k+1}) - g_P(x_k) - (g(x_{k+1}) - g(x_k))$$
$$- f(x_{k+1}) + f(x_k) - \underline{p}f(x_k) + \underline{p}f(x) - \underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle$$
$$+ \frac{p}{2}\|\bar{x}_{k+1}\|_{\tau^{-1}}^2 - \frac{p}{2}\|x_k\|_{\tau^{-1}}^2 - \left(\frac{p}{2}\|x_{k+1}\|_{\tau^{-1}P^{-1}}^2 - \frac{p}{2}\|x_k\|_{\tau^{-1}P^{-1}}^2\right)$$
$$+ \frac{p}{2}\|\bar{y}_{k+1}\|_{\sigma^{-1}}^2 - \frac{p}{2}\|y_k\|_{\sigma^{-1}}^2 + \underline{p}\langle \bar{y}_{k+1}, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle + \frac{p}{2}\sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$
$$- \frac{p}{2}\left(\|y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 - \|y_k\|_{\sigma^{-1}\pi^{-1}}^2\right) \tag{53}$$
$$S_2 = \langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle + \underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}}$$
$$- \underline{p}\langle y, \pi^{-1}\sigma^{-1}(y_k - y_{k+1}) - \sigma^{-1}(y_k - \bar{y}_{k+1}) + \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle$$

We can now collect T_1, T_2, T_3, T_4, and use the definitions of S_1, S_2, in (46)

$$g_P(x') + \underline{p}h^*(y) \geq D_p(x_{k+1}; z) + \underline{p}D_d(\bar{y}_{k+1}; z) - (1 - \underline{p})D_p(x_k; z) + g_P(x_k) + \underline{p}g(x) - \underline{p}g(x_k) + \underline{p}h^*(y)$$
$$+ \frac{p}{2}\|x - x_{k+1}\|_{\tau^{-1}P^{-1}}^2 - \frac{p}{2}\|x - x_k\|_{\tau^{-1}P^{-1}}^2 + \frac{p}{2}\|y - y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 - \frac{p}{2}\|y - y_k\|_{\sigma^{-1}\pi^{-1}}^2$$
$$+ \|\bar{x}_{k+1} - x_k\|_{\tau^{-1}P}^2 - \frac{p}{2}\|\bar{x}_{k+1} - x_k\|_{\tau^{-1}}^2 - \frac{p}{2}\sum_{i=1}^n p_i \sum_{j=1}^m \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$
$$- \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2 + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2 + \frac{p}{2}\|y_k - \bar{y}_{k+1}\|_{\sigma^{-1}}^2 + S_1 + S_2. \tag{54}$$

We make few observations on this inequality. First, by Lemma 3, as in (25)

$$g_P(x_k) + \underline{p}g(x) - \underline{p}g(x_k) - g_P(x') \geq 0.$$

Second, we have, as in (32)

$$\tilde{V}(\bar{z}_{k+1} - z_k) = \frac{p}{2}\|\bar{y}_{k+1} - y_k\|_{\sigma^{-1}}^2 + \|\bar{x}_{k+1} - x_k\|_{\tau^{-1}P}^2 - \frac{p}{2}\|\bar{x}_{k+1} - x_k\|_{\tau^{-1}}^2$$

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\underline{p}p_i\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2 - \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2$$

$$= \frac{p}{2}\|\bar{y}_{k+1} - y_k\|_{\sigma^{-1}}^2 + \frac{p}{2}\|\bar{x}_{k+1} - x_k\|_{C(\tau)}^2,$$

where

$$C(\tau)_i = \frac{2p_i}{\underline{p}\tau_i} - \frac{1}{\tau_i} - p_i\sum_{j=1}^{m}\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2 - \frac{\beta_i p_i}{\underline{p}}.$$

We use these estimates in (54) and the definition of V to conclude. $\qquad\square$

Lemma 7. *Let Assumption 1 hold and let h be separable. Given the definitions of D_p and D_d from Lemma 1, we define*

$$D_d^\gamma(\bar{y}_{k+1}; z) = \sum_{j=1}^{m}\gamma_j\left(h_j^*(\bar{y}_{k+1}^j) - h_j^*(y^j) - \langle(Ax)^j, \bar{y}_{k+1}^j - y^j\rangle\right).$$

Moreover let S_1, S_2 be as Lemma 6, and $\breve{y}_k$ as Lemma 5, it follows that

$$0 \geq \underline{p}D_p(x_{k+1}; z) + \underline{p}D_d(\breve{y}_{k+1}; z) + \tilde{V}(\bar{z}_{k+1} - z_k) + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2 + S_1 + S_2 + V(z_{k+1} - z) - V(z_k - z)$$

$$+ (1-\underline{p})D_p(x_{k+1}; z) - (1-\underline{p})D_p(x_k; z) + \underline{p}D_d^{\pi^{-1}-I}(\breve{y}_{k+1}; z) - \underline{p}D_d^{\pi^{-1}-I}(\breve{y}_k; z)$$

$$+ \underline{p}h^*(\bar{y}_{k+1}) - \underline{p}h^*(\breve{y}_k) - \underline{p}\left(h_{\pi^{-1}}^*(\breve{y}_{k+1}) - h_{\pi^{-1}}^*(\breve{y}_k)\right) + \underline{p}\langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle.$$

Proof. This lemma is an intricate part of the proof of Theorem 3. If we finish the estimations as in Alacaoglu et al. (2019), then we will end up needing to average $\bar{y}_k$. However, this is not feasible in our algorithm, since we do not update full dual vector, thus we do not compute $\bar{y}_k$ unless the data is fully dense. We will use Lemma 5 to go from $\bar{y}_k$ to $\breve{y}_k$. Let us repeat the definition of $\breve{y}_k$ from Lemma 5: Let $\breve{y}_1 = y_1 = \bar{y}_1$, and

$$\breve{y}_{k+1}^j = \bar{y}_{k+1}^j, \quad \forall j \in J(i_{k+1})$$
$$\breve{y}_{k+1}^j = \breve{y}_k^j, \qquad \forall j \notin J(i_{k+1}).$$

We now work on $D_d(\bar{y}_{k+1}; z)$ and note that h_γ^* is defined as in Lemma 5.

$$D_d(\bar{y}_{k+1}; z) = D_d(\breve{y}_{k+1}; z) - \langle Ax, \bar{y}_{k+1}\rangle + h^*(\bar{y}_{k+1}) + \langle Ax, \breve{y}_{k+1}\rangle - h^*(\breve{y}_{k+1})$$
$$= D_d(\breve{y}_{k+1}; z) - \langle Ax, \bar{y}_{k+1}\rangle + h^*(\bar{y}_{k+1}) + \langle Ax, \breve{y}_{k+1}\rangle - h^*(\breve{y}_{k+1})$$
$$+ h_{I-\pi^{-1}}^*(\breve{y}_{k+1}) - h_{I-\pi^{-1}}^*(\breve{y}_{k+1}) + h_{I-\pi^{-1}}^*(\breve{y}_k) - h_{I-\pi^{-1}}^*(\breve{y}_k)$$
$$+ \langle Ax, (I-\pi^{-1})(\breve{y}_{k+1} - \breve{y}_k)\rangle - \langle Ax, (I-\pi^{-1})(\breve{y}_{k+1} - \breve{y}_k)\rangle$$
$$= D_d(\breve{y}_{k+1}; z) + h^*(\bar{y}_{k+1}) - h^*(\breve{y}_k) - (h_{\pi^{-1}}^*(\breve{y}_{k+1}) - h_{\pi^{-1}}^*(\breve{y}_k))$$
$$+ h_{\pi^{-1}-1}^*(\breve{y}_{k+1}) - h_{\pi^{-1}-1}^*(\breve{y}_k) + \langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle$$
$$+ \langle Ax, (I-\pi^{-1})(\breve{y}_{k+1} - \breve{y}_k)\rangle$$
$$= D_d(\breve{y}_{k+1}; z) + h^*(\bar{y}_{k+1}) - h^*(\breve{y}_k) - (h_{\pi^{-1}}^*(\breve{y}_{k+1}) - h_{\pi^{-1}}^*(\breve{y}_k))$$
$$+ \langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle + D_d^{\pi^{-1}-I}(\breve{y}_{k+1}; z) - D_d^{\pi^{-1}-I}(\breve{y}_k; z)$$

We insert this estimate into the result of Lemma 6 to finish the proof. $\qquad\square$

The following lemma is similar to (Alacaoglu et al., 2019, Lemma 4.8).

Lemma 8. *Given a Euclidean space $\mathcal{S}$, a fixed diagonal matrix $\gamma \succeq 0$, let the random sequences $u_k, v_k \in \mathcal{S}$ be $\mathcal{F}_k$-measurable with*

$$u_{k+1} = v_{k+1} - \mathbb{E}_k[v_{k+1}].$$

Let $\tilde{x}_1$ be arbitrary and set for $k \geq 1$,

$$\tilde{x}_{k+1} = \tilde{x}_k + u_{k+1}.$$

Then, $\tilde{x}_k$ is $\mathcal{F}_k$-measurable and we have

$$\sum_{k=1}^{K}\langle x, u_{k+1}\rangle_\gamma \le \frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \sum_{k=1}^{K}\langle \tilde{x}_k, u_{k+1}\rangle_\gamma + \frac{1}{2}\sum_{k=1}^{K}\|v_{k+1}\|_\gamma^2,$$

with $\mathbb{E}\left[\sum_{k=1}^{K}\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right] = 0$ *and for any* $S \subset \mathcal{S}$

$$\mathbb{E}\left[\sup_{x \in S}\sum_{k=1}^{K}\langle x, u_{k+1}\rangle_\gamma\right] \le \sup_{x \in S}\frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\|v_{k+1}\|_\gamma^2\right].$$

Proof. First, by the definition of $\tilde{x}_{k+1}$, for all $x \in \mathcal{S}$

$$\frac{1}{2}\|\tilde{x}_{k+1} - x\|_\gamma^2 = \frac{1}{2}\|\tilde{x}_k - x\|_\gamma^2 + \langle \tilde{x}_k - x, u_{k+1}\rangle_\gamma + \frac{1}{2}\|u_{k+1}\|_\gamma^2.$$

Summing this inequality gives

$$\sum_{k=1}^{K}\langle x, u_{k+1}\rangle_\gamma \le \frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \sum_{k=1}^{K}\langle \tilde{x}_k, u_{k+1}\rangle_\gamma + \sum_{k=1}^{K}\frac{1}{2}\|u_{k+1}\|_\gamma^2.$$

We take supremum and expectation of both sides to get

$$\mathbb{E}\left[\sup_{x \in S}\sum_{k=1}^{K}\langle x, u_{k+1}\rangle_\gamma\right] \le \sup_{x \in S}\frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \sum_{k=1}^{K}\mathbb{E}\left[\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right] + \sum_{k=1}^{K}\frac{1}{2}\mathbb{E}\left[\|u_{k+1}\|_\gamma^2\right].$$

By the law of total expectation, $\mathcal{F}_k$-measurability of $\tilde{x}_k$ and $\mathbb{E}_k\left[u_{k+1}\right] = 0$, we have

$$\sum_{k=1}^{K}\mathbb{E}\left[\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right] = \sum_{k=1}^{K}\mathbb{E}\left[\mathbb{E}_k\left[\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right]\right] = \sum_{k=1}^{K}\mathbb{E}\left[\langle \tilde{x}_k, \mathbb{E}_k\left[u_{k+1}\right]\rangle_\gamma\right] = 0.$$

Finally, we use the definition of u_k and the inequality $\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] \le \mathbb{E}\left[\|X\|^2\right]$ which holds for any random variable X. $\qquad\square$

As mentioned in the main text, we will give the theorems in the appendix with tighter, but more complicated constants. After Theorem 4, we show how we obtained the simplified bounds in our main text.

Theorem 3. *Let Assumption 1 hold and θ, τ, σ are chosen as in* (8), (9). *Moreover, let h be separable.*

We define $x_K^{av} = \frac{1}{K}\sum_{k=1}^{K}x_k$ and $y_K^{av} = \frac{1}{K}\sum_{k=1}^{K}\breve{y}_k$, where $\breve{y}_k$ is defined in (13), *then it holds that for any bounded set $\mathcal{C} = \mathcal{C}_x \times \mathcal{C}_y \subset \mathcal{Z}$*

$$\mathbb{E}\left[G_{\mathcal{C}}(x_K^{av}, y_K^{av})\right] \le \frac{C_g}{\underline{p}K},$$

where $C_g = C_{g,1} + C_{g,2} + C_{g,3} + C_{g,4}$, $C_{\tau,\tilde{V}} = \min_i C(\tau)_i \tau_i$,

$C_{g,1} = \sup_{z \in \mathcal{C}}\left\{2\underline{p}\|x_0 - x\|_{\tau^{-1}P^{-1}}^2 + 2\underline{p}\|y_0 - y\|_{\sigma^{-1}\pi^{-1}}^2\right\} + (1 - \underline{p})4\sqrt{\Delta_0\underline{p}^{-1}}\|A\|\sup_{y \in \mathcal{C}_y}\|y\|_{\tau P}$

$+ 2\underline{p}\sqrt{\Delta_0\underline{p}^{-1} + \|2P - \underline{p}\|\Delta_0\underline{p}^{-3}C_{\tau,\tilde{V}}^{-1}}\|A\|\|\pi^{-1} - I\|\sup_{x \in \mathcal{C}_x}\|x\|_{\sigma\pi},$

$C_{g,2} = \|2P - \underline{p}\|\left(1 + \|P/\underline{p}\| + \|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2\right)\frac{2\Delta_0}{\underline{p}C_{\tau,\tilde{V}}} + \Delta_0 C_{\tau,\tilde{V}}^{-1} + (4 + 4\|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2)\Delta_0,$

$C_{g,3} = (1 - \underline{p})\left(f(x_0) + g(x_0) - f(x_\star) - g(x_\star) + \|A^\top y_\star\|_{\tau P}\sqrt{2\Delta_0\underline{p}^{-1}}\right),$

$C_{g,4} = \underline{p}h_{\pi^{-1}-I}^*(\breve{y}_0) + \underline{p}\sum_{j=1}^{m}(\pi_j^{-1} - 1)h_j^*(y_\star^j) + \frac{\underline{p}}{2}\|Ax_\star\|_{\sigma\pi^{-1}}^2 + \Delta_0 + \frac{\|2P - \underline{p}\|\Delta_0}{\underline{p}^2 C_{\tau,\tilde{V}}}.$

Proof. We start with the result of Lemma 7. First, we will manipulate the terms arising in $S_2 + \langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1}) \rangle$ (see definition of S_2 in Lemma 6).

$$- \big(S_2 + \underline{p}\langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle\big) = -\langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle$$
$$- \underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}} + \underline{p}\langle y, \pi^{-1}\sigma^{-1}(y_k - y_{k+1}) - \sigma^{-1}(y_k - \bar{y}_{k+1}) + \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle$$
$$- \underline{p}\langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle \quad (55)$$

For the four terms on the right hand side, we will apply Lemma 8. We note first, $\mathbb{E}_k\left[\pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\right] = \breve{y}_k - \bar{y}_{k+1}$ from Lemma 5, $\mathbb{E}_k\left[P^{-1}(x_k - x_{k+1}) = x_k - \bar{x}_{k+1}\right]$, by coordinate wise updates. Finally, as in the proof of Lemma 5, we can derive, as $A_{j,i} = 0, \forall i \notin I(j)$,

$$\mathbb{E}_k[y_{k+1}^j] = \sum_{i=1}^n p_i \left[\mathbb{1}_{j \in J(i)}\big(\bar{y}_{k+1}^j + \sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^i - x_k^i)\big) + \mathbb{1}_{j \notin J(i)}y_k^j\right]$$

$$= y_k^j + \sum_{i \in I(j)} p_i\big(\bar{y}_{k+1}^j - y_k^j\big) + \sum_{i=1}^n p_i\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^i - x_k^i) = y_k^j + \pi_j(\bar{y}_{k+1}^j - y_k^j) + \sigma_j\theta_j(AP(\bar{x}_{k+1} - x_k))_j$$

$$\mathbb{E}_k[y_{k+1}] = y_k + \pi(\bar{y}_{k+1} - y_k) + \sigma\theta AP(\bar{x}_{k+1} - x_k) \iff \mathbb{E}_k\left[y_k - y_{k+1}\right] = \pi(y_k - \bar{y}_{k+1}) - \sigma\theta AP(\bar{x}_{k+1} - x_k).$$

In particular, for (55), we set in Lemma 8

$$u_{k+1} = -\underline{p}^{-1}\sigma\pi A(x_k - x_{k+1}) + \underline{p}^{-1}\sigma\pi AP(x_k - \bar{x}_{k+1}), \quad \gamma = \sigma^{-1}\pi^{-1}\underline{p}, \quad \mathcal{S} = \mathcal{Y}, \quad \tilde{x}_1 = y_1.$$
$$u_{k+1} = (x_k - x_{k+1}) - P(x_k - \bar{x}_{k+1}), \quad \gamma = \tau^{-1}P^{-1}, \quad \mathcal{S} = \mathcal{X}, \quad \tilde{x}_1 = x_1,$$
$$u_{k+1} = (y_k - y_{k+1}) - \pi(y_k - \bar{y}_{k+1}) + \sigma\theta AP(\bar{x}_{k+1} - x_k), \quad \gamma = \sigma^{-1}\pi^{-1}, \quad \mathcal{S} = \mathcal{Y}, \quad \tilde{x}_1 = y_1,$$
$$u_{k+1} = \tau PA^\top\left(\pi^{-1}(\breve{y}_k - \breve{y}_{k+1}) - (\breve{y}_k - \bar{y}_{k+1})\right), \quad \gamma = \tau^{-1}P^{-1}, \quad \mathcal{S} = \mathcal{X}, \quad \tilde{x}_1 = x_1,$$

Then, we can apply Lemma 8 for these cases to bound (55) as

$$\mathbb{E}\left[\sup_{z \in \mathcal{C}}(55)\right] \leq \sup_{z \in \mathcal{C}}\left\{\underline{p}\|x - x_1\|_{\tau^{-1}P^{-1}}^2 + \underline{p}\|y - y_1\|_{\sigma^{-1}\pi^{-1}}^2\right\} + \sum_{k=1}^K \frac{p}{2}\mathbb{E}\left[\|x_k - x_{k+1}\|_{\tau^{-1}P^{-1}}^2 + \|y_k - y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2\right]$$

$$+ \sum_{k=1}^K \frac{1}{2\underline{p}}\mathbb{E}\left[\|\sigma\pi A(x_k - x_{k+1})\|_{\sigma^{-1}\pi^{-1}}^2\right] + \sum_{k=1}^K \frac{p}{2}\mathbb{E}\left[\|\tau PA^\top\pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\|_{\tau^{-1}P^{-1}}^2\right]. \quad (56)$$

We now recall the definition of S_1 from (53), and use the identities (21), (22), Lemma 2, along with the law of total expectation to estimate

$$\mathbb{E}\left[S_1 + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2\right] = \mathbb{E}\left[-f(x_{k+1}) + f(x_k) - \langle\nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2\right]$$

$$+ \underline{p}\mathbb{E}\left[f(x) - f(x_k) - \langle\nabla f(x_k), x - x_k\rangle\right]$$

$$\geq \mathbb{E}\left[-f(x_{k+1}) + f(x_k) - \mathbb{E}_k\left[\langle\nabla f(x_k), x_k - x_{k+1}\rangle\right] + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2\right]$$

$$\geq \mathbb{E}\left[-\frac{1}{2}\mathbb{E}_k\left[\|x_k - x_{k+1}\|_\beta^2\right] + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2\right]$$

$$= \mathbb{E}\left[-\frac{1}{2}\|x_k - \bar{x}_{k+1}\|_{\beta P}^2 + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|_{\beta P}^2\right]$$

$$= 0, \quad (57)$$

where the first inequality is by convexity, second inequality is by coordinatewise smoothness of f.

Furthermore, for the result of Lemma 7, by Lemma 5 and the law of total expectation

$$\mathbb{E}\left[h^*(\bar{y}_{k+1}) - h^*(\breve{y}_k) - (h_{\pi^{-1}}^*(\breve{y}_{k+1}) - h_{\pi^{-1}}^*(\breve{y}_k))\right] = 0. \quad (58)$$

We rearrange and sum the result of Lemma 7, take supremum and expectation, plug in eqs. (56)–(58), and use $\tilde{V}$ is a squared norm

$$
\mathbb{E}\left[\sup_{z\in\mathcal{C}}\sum_{k=1}^{K}\underline{p}\left(D_p(x_k)+D_d(\breve{y}_k)\right)\right] \leq \sup_{z\in\mathcal{C}}\frac{3\underline{p}}{2}\left(\|x-x_0\|_{\tau^{-1}P^{-1}}^2 + \|y-y_0\|_{\sigma^{-1}\pi^{-1}}^2\right)
$$

$$
+ \mathbb{E}\left[\sup_{z\in\mathcal{C}}(1-\underline{p})\left(D_p(x_0;z)-D_p(x_K;z)\right)\right] + \mathbb{E}\left[\sup_{z\in\mathcal{C}}\underline{p}D_d^{\pi^{-1}-I}(\breve{y}_0;z)-\underline{p}D_d^{\pi^{-1}-I}(\breve{y}_K;z)\right]
$$

$$
+ \sum_{k=1}^{K}\frac{\underline{p}}{2}\mathbb{E}\left[\|x_k-x_{k+1}\|_{\tau^{-1}P^{-1}}^2 + \|y_k-y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2\right] + \sum_{k=1}^{K}\frac{1}{2\underline{p}}\mathbb{E}\left[\|\sigma\pi A(x_k-x_{k+1})\|_{\sigma^{-1}\pi^{-1}}^2\right]
$$

$$
+ \sum_{k=1}^{K}\frac{\underline{p}}{2}\mathbb{E}\left[\|\tau PA^\top\pi^{-1}(\breve{y}_k-\breve{y}_{k+1})\|_{\tau^{-1}P^{-1}}^2\right]. \tag{59}
$$

We first note by (38)

$$
\sum_{k=1}^{\infty}\mathbb{E}\left[\tilde{V}(\bar{z}_{k+1}-z_k)\right] \leq \Delta_0. \tag{60}
$$

$$
\tilde{V}(\bar{z}_{k+1}-z_k) \geq \frac{\underline{p}C_{\tau,\tilde{V}}}{2}\|\bar{x}_{k+1}-x_k\|_{\tau^{-1}}^2 + \frac{\underline{p}}{2}\|\bar{y}_{k+1}-y_k\|_{\sigma^{-1}}^2, \tag{61}
$$

with $C_{\tau,\tilde{V}} = \min_i C(\tau)_i\tau_i$, where we used the definition of $\tilde{V}$ from Lemma 1.

We have

$$
\mathbb{E}\left[\sup_{z\in\mathcal{C}}D_p(x_0;z)-D_p(x_K;z)\right] = \mathbb{E}\left[\sup_{z\in\mathcal{C}_y}f(x_0)+g(x_0)-f(x_K)-g(x_K)+\langle A^\top y, x_0-x_K\rangle\right]
$$

$$
\leq \mathbb{E}\left[\sup_{z\in\mathcal{C}_y}f(x_0)+g(x_0)-f(x_K)-g(x_K)+\|A\|\|y\|_{\tau P}\|x_0-x_K\|_{\tau^{-1}P^{-1}}\right]. \tag{62}
$$

Then, we use the optimality conditions, convexity, and (36)

$$
\mathbb{E}\left[f(x_K)+g(x_K)\right] \geq \mathbb{E}\left[f(x_\star)+g(x_\star)-\langle A^\top y_\star, x_K-x_\star\rangle\right]
$$

$$
\geq \mathbb{E}\left[f(x_\star)+g(x_\star)-\|A^\top y_\star\|_{\tau P}\|x_K-x_\star\|_{\tau^{-1}P^{-1}}\right]
$$

$$
\geq f(x_\star)+g(x_\star)-\|A^\top y_\star\|_{\tau P}\sqrt{\frac{2\Delta_0}{\underline{p}}}. \tag{63}
$$

to obtain for this estimation

$$
\mathbb{E}\left[\sup_{z\in C}D_p(x_0;z)-D_p(x_K;z)\right] \leq f(x_0)+g(x_0)-f(x_\star)-g(x_\star)+\|A^\top y_\star\|_{\tau P}\sqrt{2\Delta_0\underline{p}^{-1}}
$$

$$
+ 4\sqrt{2\Delta_0\underline{p}^{-1}}\|A\|\sup_{y\in C_y}\|y\|_{\tau P}. \tag{64}
$$

We estimate similarly to obtain

$$
\mathbb{E}\left[\sup_{z\in\mathcal{C}}\underline{p}D_d^{\pi^{-1}-I}(\breve{y}_0;z)-\underline{p}D_d^{\pi^{-1}-I}(\breve{y}_K;z)\right] = \underline{p}\mathbb{E}\left[h_{\pi^{-1}-I}^*(\breve{y}_0)-h_{\pi^{-1}-I}^*(\breve{y}_K)-\langle Ax, (\pi^{-1}-I)(\breve{y}_0-\breve{y}_K)\rangle\right].
$$

$$
\leq \underline{p}\mathbb{E}\left[h_{\pi^{-1}-I}^*(\breve{y}_0)-h_{\pi^{-1}-I}^*(\breve{y}_K)+\|A\|\|\pi^{-1}-I\|\|x\|_{\sigma\pi}\|\breve{y}_0-\breve{y}_K\|_{\sigma^{-1}\pi^{-1}}\right]. \tag{65}
$$

By convexity and Lemma 5

$$
\mathbb{E}\left[h_{\pi^{-1}-I}^*(\breve{y}_K)\right] = \mathbb{E}\left[\sum_{j=1}^{m}(\pi_j^{-1}-1)h_j^*(\breve{y}_K^j) \geq \sum_{j=1}^{m}(\pi_j^{-1}-1)\left(h_j^*(y_\star^j)+\langle(Ax_\star)_j, \breve{y}_K^j-y_\star^j\rangle\right)\right]
$$

$$\geq \mathbb{E}\left[\sum_{j=1}^{m}(\pi_j^{-1}-1)h_j^*(y_\star^j) - \frac{1}{2}\|Ax_\star\|_{\sigma\pi^{-1}}^2 - \frac{1}{2}\|\breve{y}_K - y_\star\|_{\sigma^{-1}\pi^{-1}}^2\right]$$

$$\geq \mathbb{E}\left[\sum_{j=1}^{m}(\pi_j^{-1}-1)h_j^*(y_\star^j) - \frac{1}{2}\|Ax_\star\|_{\sigma\pi^{-1}}^2 - \|y_K - y_\star\|_{\sigma^{-1}\pi^{-1}}^2 - \sum_{k=1}^{K}\|\bar{x}_{k+1}-x_k\|_{B(\pi^{-2}\sigma^{-1})}^2\right]$$

$$\geq \sum_{j=1}^{m}(\pi_j^{-1}-1)h_j^*(y_\star^j) - \frac{1}{2}\|Ax_\star\|_{\sigma\pi^{-1}}^2 - \frac{\Delta_0}{\underline{p}} - \frac{\|2P-\underline{p}\|\Delta_0}{\underline{p}^3 C_{\tau,\tilde{V}}}, \tag{66}$$

where we used

$$\|x\|_{B(\pi^{-2}\sigma^{-1})}^2 \leq \frac{1}{\underline{p}^2}\|2P-\underline{p}\|\|x\|_{\tau^{-1}}^2, \tag{67}$$

which follows by using the step size rule from (9) and definition of $B(\gamma)$ from Lemma 5.

Thus, the final bound for (65), after using Lemma 5

$$\mathbb{E}\left[\sup_{z\in\mathcal{C}}\underline{p}D_d^{\pi^{-1}-I}(\breve{y}_0;z) - \underline{p}D_d^{\pi^{-1}-I}(\breve{y}_K;z)\right] \leq \underline{p}h_{\pi^{-1}-I}^*(\breve{y}_0) + \underline{p}\sum_{j=1}^{m}(\pi_j^{-1}-1)h_j^*(\breve{y}_\star^j) + \frac{\underline{p}}{2}\|Ax_\star\|_{\sigma\pi^{-1}}^2$$

$$+ \Delta_0 + \frac{\|2P-\underline{p}\|\Delta_0}{\underline{p}^2 C_{\tau,\tilde{V}}} + 2\underline{p}\sqrt{\Delta_0\underline{p}^{-1} + \|2P-\underline{p}\|\Delta_0\underline{p}^{-3}C_{\tau,\tilde{V}}^{-1}}\|A\|\|\pi^{-1}-I\|\sup_{x\in\mathcal{C}_x}\|x\|_{\sigma\pi}. \tag{68}$$

We continue to estimate, by Lemma 2, the definition of $B(\gamma)$ from Lemma 5, and the definition of π_j

$$\mathbb{E}_k\left[\|y_{k+1}-y_k\|_{\sigma^{-1}\pi^{-1}}^2\right] = \|\bar{y}_{k+1}-y_k\|_{\sigma^{-1}}^2 + 2\langle\bar{y}_{k+1}-y_k, \pi^{-1}\theta AP(\bar{x}_{k+1}-x_k)\rangle$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{m}p_i\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$

$$\leq \|\bar{y}_{k+1}-y_k\|_{\sigma^{-1}}^2 + \|\bar{y}_{k+1}-y_k\|_{\sigma^{-1}}^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}p_i^2\pi_j^{-2}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{m}p_i\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^i - x_k^i)^2$$

$$\leq 2\|\bar{y}_{k+1}-y_k\|_{\sigma^{-1}}^2 + \left(1 + \|P/\underline{p}\|\right)\|\bar{x}_{k+1}-x_k\|_{B(\pi^{-1}\sigma^{-1})}^2, \tag{69}$$

$$\sum_{k=1}^{K}\frac{\underline{p}}{2}\mathbb{E}_k\left[\|x_{k+1}-x_k\|_{\tau^{-1}P^{-1}}^2\right] = \frac{\underline{p}}{2}\sum_{k=1}^{K}\|\bar{x}_{k+1}-x_k\|_{\tau^{-1}}^2 \leq \frac{\Delta_0}{C_{\tau,\tilde{V}}}. \tag{70}$$

We will continue with estimating the last two terms of (59):

$$\frac{1}{2\underline{p}}\mathbb{E}_k\left[\|\sigma\pi A(x_k - x_{k+1})\|_{\sigma^{-1}\pi^{-1}}^2\right] = \mathbb{E}_k\left[\frac{1}{2\underline{p}}\sum_{j=1}^{m}\sigma_j\pi_j(A(x_k-x_{k+1}))_j^2\right]$$

$$= \mathbb{E}_k\left[\frac{1}{2\underline{p}}\sum_{j=1}^{m}\sigma_j\pi_j A_{j,i_{k+1}}^2(x_k^{i_{k+1}} - \bar{x}_{k+1}^{i_{k+1}})^2\right]$$

$$= \frac{1}{2\underline{p}}\sum_{j=1}^{m}\sum_{i=1}^{n}p_i\sigma_j\pi_j A_{j,i}^2(x_k^i - \bar{x}_{k+1}^i)^2$$

$$= \frac{\underline{p}}{2}\|\bar{x}_{k+1}-x_k\|_{B(\pi^{-1}\sigma^{-1})}^2. \tag{71}$$

Finally,

$$\sum_{k=1}^{K} \mathbb{E}\Big[\|\tau P A^\top \pi^{-1}(\breve{y}_{k+1} - \breve{y}_k)\|_{\tau^{-1}P^{-1}}^2\Big] \le \|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2 \sum_{k=1}^{K} \mathbb{E}\left[\|\breve{y}_{k+1} - \breve{y}_k\|_{\sigma^{-1}\pi^{-1}}^2\right]$$

$$\le \|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2 \sum_{k=1}^{K} \mathbb{E}\left[2\|\bar{y}_{k+1} - y_k\|_{\sigma^{-1}}^2 + 2\|\bar{x}_{k+1} - x_k\|_{B(\pi^{-1}\sigma^{-1})}^2\right], \tag{72}$$

where we use Lemma 5 for the last inequality.

By, eqs. (69)–(72), we finalized the bound for the last three terms in (59)

$$\sum_{k=1}^{K} \frac{p}{2}\mathbb{E}\left[\|x_k - x_{k+1}\|_{\tau^{-1}P^{-1}}^2 + \|y_k - y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2\right] + \sum_{k=1}^{K} \frac{1}{2\underline{p}}\mathbb{E}\left[\|\sigma\pi A(x_k - x_{k+1})\|_{\sigma^{-1}\pi^{-1}}^2\right]$$

$$+ \sum_{k=1}^{K} \frac{p}{2}\mathbb{E}\left[\|\tau P A^\top\pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\|_{\tau^{-1}P^{-1}}^2\right] \le C_{g,2}, \tag{73}$$

where

$$C_{g,2} = \|2P - \underline{p}\|\Big(1 + \|P/\underline{p}\| + \|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2\Big)\frac{2\Delta_0}{\underline{p}C_{\tau,\tilde{V}}} + \frac{\Delta_0}{C_{\tau,\tilde{V}}} + (4 + 4\|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2)\Delta_0.$$

$C_{g,2}$ basically collects the bounds in eqs. (69)–(72) and uses the estimates from eqs. (60) and (61) on this bound.

Then, on (59), we use eqs. (60), (61), (64) and (67)–(72), definition of primal-dual gap function in (11), and Jensen's inequality to conclude. $\qquad\square$

Theorem 4. *Let Assumption 1 hold. We use the same parameters θ, τ, σ and the definitions for x_K^{av} and y_K^{av} as Theorem 3. We consider two cases separately:*
▷ *If $h(\cdot) = \delta_{\{b\}}(\cdot)$, we obtain*

$$\mathbb{E}\left[f(x_K^{av}) + g(x_K^{av}) - f(x_\star) - g(x_\star)\right] \le \frac{C_o}{\underline{p}K}.$$

$$\mathbb{E}\left[\|Ax_K^{av} - b\|\right] \le \frac{C_f}{\underline{p}K}.$$

▷ *If h is L_h-Lipschitz continuous, we obtain*

$$\mathbb{E}\left[f(x_K^{av}) + g(x_K^{av}) + h(Ax_K^{av}) - f(x_\star) - g(x_\star) - h(Ax_\star)\right] \le \frac{C_l}{\underline{p}K},$$

where $C_f = 2c_2\sqrt{\|y_\star - y_0\|_{\sigma^{-1}\pi^{-1}}^2 + C_s c_2^{-1} + 2c_1 c_2^{-1}\|x_0 - x_\star\|_{\tau^{-1}P^{-1}}^2} + 2c_2\|y_\star - y_0\|_{\sigma^{-1}\pi^{-1}}$,
$C_o = C_s + \|y_\star\|_{\sigma^{-1}\pi^{-1}}C_f + c_1\|x_0 - x_\star\|_{\tau^{-1}P^{-1}}^2 + c_2\|y_\star - y_0\|_{\sigma^{-1}\pi^{-1}}^2$,
$C_l = C_s + c_1\|x_\star - x_0\|_{\tau^{-1}P^{-1}}^2 + 4c_2 L_h^2$,
$c_1 = \frac{3p}{2} + \underline{p}\|(2P - \underline{p})^{1/2}\|\|\pi^{-1} - I\|$,
$c_2 = \frac{3p}{2} + (1 - \underline{p})\|(2P - \underline{p})^{1/2}\|$, $C_s = C_{g,2} + C_{g,5} + C_{g,6}$, with $C_{g,2}$ as defined in Theorem 3 and $C_{g,5}, C_{g,6}$ are defined in the proof in (79), (80).

Proof. We will use the first result of Lemma 8 on the result of Lemma 7, similar to (59). The difference is that we take supremum and expectation after using Lemma 8 and we process the terms $(1 - \underline{p})\left(D_p(x_k; z) - D_p(x_{k+1}; z)\right) + \underline{p}\left(D_d^{\pi^{-1}-I}(\breve{y}_k; z) - D_d^{\pi^{-1}-I}(\breve{y}_{k+1}; z)\right)$ with small differences. In particular,

$$\sum_{k=0}^{K-1} (1 - \underline{p})\left(D_p(x_k; z) - D_p(x_{k+1}; z)\right) = (1 - \underline{p})\left(D_p(x_0; z) - D_p(x_K; z)\right)$$

$$= (1 - \underline{p}) \left(f(x_0) + g(x_0) - f(x_K) - g(x_K) + \langle A^\top y, x_0 - x_K \rangle \right).$$

For the final term, we estimate using the step size rule (9)

$$
\langle A^\top y, x_0 - x_K \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{j,i} y^j (x_0^i - x_K^i) \leq \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} A_{j,i}^2 (x_0^i - x_K^i)^2 \sigma_j \pi_j} \sqrt{\sum_{j=1}^{m} (y^j)^2 \sigma_j^{-1} \pi_j^{-1}}
$$

$$
\leq \sum_{i=1}^{n} \sqrt{\underline{p}(2p_i - \underline{p})(x_0^i - x_K^i)^2 \tau_i^{-1} p_i^{-1}} \sqrt{\sum_{j=1}^{m} (y^j)^2 \sigma_j^{-1} \pi_j^{-1}}
$$

$$
\leq \sum_{i=1}^{n} \left(\frac{1}{2} \sqrt{2p_i - \underline{p}}(x_0^i - x_K^i)^2 \tau_i^{-1} p_i^{-1} + \frac{\underline{p}\sqrt{2p_i - \underline{p}}}{2} \|y\|_{\sigma^{-1}\pi^{-1}}^2 \right)
$$

$$
\leq \frac{\|(2P - \underline{p})^{1/2}\|}{2} \|x_0 - x_K\|_{\tau^{-1}P^{-1}}^2 + \frac{\|(2P - \underline{p})^{1/2}\|}{2} \|y\|_{\sigma^{-1}\pi^{-1}}^2
$$

$$
\leq \frac{\|(2P - \underline{p})^{1/2}\|}{2} \|x_0 - x_K\|_{\tau^{-1}P^{-1}}^2 + \|(2P - \underline{p})^{1/2}\| \left(\|y - y_0\|_{\sigma^{-1}\pi^{-1}}^2 + \|y_0\|_{\sigma^{-1}\pi^{-1}}^2 \right). \quad (74)
$$

We estimate similarly to obtain

$$
\sum_{k=0}^{K-1} \underline{p} \left(D_d^{\pi^{-1}-I}(\breve{y}_k; z) - D_d^{\pi^{-1}-I}(\breve{y}_{k+1}; z) \right) = \underline{p} \left(h_{\pi^{-1}-I}^*(\breve{y}_0) - h_{\pi^{-1}-I}^*(\breve{y}_K) - \langle Ax, (\pi^{-1} - I)(\breve{y}_0 - \breve{y}_K) \rangle \right),
$$

and

$$
-\langle Ax, (\pi^{-1} - I)(\breve{y}_0 - \breve{y}_K) \rangle \leq \frac{1}{2} \|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\| \left(\|x\|_{\tau^{-1}P^{-1}}^2 + \|\breve{y}_0 - \breve{y}_K\|_{\sigma^{-1}\pi^{-1}}^2 \right)
$$

$$
\leq \|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\| \left(\|x - x_0\|_{\tau^{-1}P^{-1}}^2 + \|x_0\|_{\tau^{-1}P^{-1}}^2 + \frac{1}{2} \|\breve{y}_0 - \breve{y}_K\|_{\sigma^{-1}\pi^{-1}}^2 \right). \quad (75)
$$

We sum the result of Lemma 7, use eqs. (74) and (75), use the first result of Lemma 8 (see (55)), move the terms depending on (x, y) to LHS, take the supremum over $\mathcal{Z}$ and expectation. With these steps, instead of (59), we get

$$
\mathbb{E} \left[\sup_{z \in \mathcal{Z}} \sum_{k=1}^{K} \underline{p} \left(D_p(x_k) + D_d(\breve{y}_k) \right) - c_1 \|x_0 - x\|_{\tau^{-1}P^{-1}}^2 - c_2 \|y_0 - y\|_{\sigma^{-1}\pi^{-1}}^2 \right] \leq \underline{p} \left(h_{\pi^{-1}-I}^*(\breve{y}_0) - h_{\pi^{-1}-I}^*(\breve{y}_K) \right)
$$

$$
+ (1 - \underline{p}) \left(f(x_0) + g(x_0) - f(x_K) - g(x_K) \right) + \sum_{k=1}^{K} \frac{\underline{p}}{2} \mathbb{E} \left[\|x_k - x_{k+1}\|_{\tau^{-1}P^{-1}}^2 + \|y_k - y_{k+1}\|_{\sigma^{-1}\pi^{-1}}^2 \right]
$$

$$
+ \sum_{k=1}^{K} \frac{1}{2\underline{p}} \mathbb{E} \left[\|\sigma \pi A(x_k - x_{k+1})\|_{\sigma^{-1}\pi^{-1}}^2 \right] + \sum_{k=1}^{K} \frac{\underline{p}}{2} \mathbb{E} \left[\|\tau P A^\top \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\|_{\tau^{-1}P^{-1}}^2 \right],
$$

$$
+ \|(2P - \underline{p})^{1/2}\| \left(\frac{1}{2} \|x_0 - x_K\|_{\tau^{-1}P^{-1}}^2 + \|y_0\|_{\sigma^{-1}\pi^{-1}}^2 \right)
$$

$$
+ \|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\| \left(\|x_0\|_{\tau^{-1}P^{-1}}^2 + \frac{1}{2} \|\breve{y}_0 - \breve{y}_K\|_{\sigma^{-1}\pi^{-1}}^2 \right), \quad (76)
$$

where $c_1 = \frac{3\underline{p}}{2} + \underline{p}\|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\|$, $c_2 = \frac{3\underline{p}}{2} + (1 - \underline{p})\|(2P - \underline{p})^{1/2}\|$.

We divide both sides by $\underline{p}$ and use Jensen's inequality to obtain the smoothed gap function (Tran-Dinh et al., 2018)

$$
\mathcal{G}_{\frac{2c_1}{\underline{p}K}, \frac{2c_2}{\underline{p}K}} (x_K^{av}, y_K^{av}; x_0, y_0) = \sup_{z \in \mathcal{Z}} D_p(x_K^{av}; z) + D_d(y_K^{av}; z) - \frac{c_1}{\underline{p}K} \|x - x_0\|_{\tau^{-1}P^{-1}}^2 - \frac{c_2}{\underline{p}K} \|y - y_0\|_{\sigma^{-1}\pi^{-1}}^2.
$$

Then, we have, as in the proof of Theorem 3 that (see eqs. (59) and (73))

$$
\underline{p}K \mathbb{E} \left[\mathcal{G}_{\frac{2c_1}{\underline{p}K}, \frac{2c_2}{\underline{p}K}} (x_K^{av}, y_K^{av}; x_0, y_0) \right] \leq C_{g,2} + \underline{p} \left(h_{\pi^{-1}-I}^*(\breve{y}_0) - h_{\pi^{-1}-I}^*(\breve{y}_K) \right)
$$

$$+ (1 - \underline{p}) \left(f(x_0) + g(x_0) - f(x_K) - g(x_K) \right)$$

$$+ (1 - \underline{p}) \| (2P - \underline{p})^{1/2} \| \left(\frac{1}{2} \| x_0 - x_K \|^2_{\tau^{-1}P^{-1}} + \| y_0 \|^2_{\sigma^{-1}\pi^{-1}} \right)$$

$$+ \underline{p} \| (2P - \underline{p})^{1/2} \| \| \pi^{-1} - I \| \left(\| x_0 \|^2_{\tau^{-1}P^{-1}} + \frac{1}{2} \| \breve{y}_0 - \breve{y}_K \|^2_{\sigma^{-1}\pi^{-1}} \right) \qquad (77)$$

By using (63) and (66) we obtain the bound

$$(1 - \underline{p}) \left(f(x_0) + g(x_0) - f(x_K) - g(x_K) \right) + \underline{p} \left(h^*_{\pi^{-1}-I}(\breve{y}_0) - h^*_{\pi^{-1}-I}(\breve{y}_K) \right) \le C_{g,5}, \qquad (78)$$

where

$$C_{g,5} = (1 - \underline{p}) \left(f(x_0) + g(x_0) - f(x_\star) - g(x_\star) + \| A^\top y_\star \|_{\tau P} \sqrt{\frac{2\Delta_0}{\underline{p}}} \right)$$

$$+ \underline{p} \left(h^*_{\pi^{-1}-I}(y_0) - \sum_{j=1}^{m} (\pi_j^{-1} - 1) h^*_j(y^j_\star) + \frac{1}{2} \| Ax_\star \|^2_{\sigma\pi^{-1}} + \frac{\Delta_0}{\underline{p}} + \frac{\Delta_0 \| 2P - \underline{p} \|}{\underline{p}^2 C_{\tau,\tilde{V}}} \right). \qquad (79)$$

Next, we bound the last two terms in (77) using $\mathbb{E} \left[V(z_k - z_\star) \right] \le \Delta_0$ from (37) and we denote the bound as $C_{g,6}$

$$C_{g,6} = (1 - \underline{p}) \| (2P - \underline{p})^{1/2} \| \left(\frac{4\Delta_0}{\underline{p}} + 2\| y_\star \|^2_{\sigma^{-1}\pi^{-1}} \right)$$

$$+ \underline{p} \| (2P - \underline{p})^{1/2} \| \| \pi^{-1} - I \| \left(\frac{4\Delta_0}{\underline{p}} + 2\| x_\star \|^2_{\tau^{-1}P^{-1}} \right), \qquad (80)$$

so that RHS of (77) is upper bounded by $C_{g,2} + C_{g,5} + C_{g,6}$.

We consider two cases:

• If h is L_h Lipschitz continuous in norm $\| \cdot \|_{\sigma\pi}$, then $\| y - y_0 \|^2_{\sigma^{-1}\pi^{-1}} \le 4L_h^2$. Then, we argue as in (Fercoq & Bianchi, 2019, Theorem 11) to get

$$\mathbb{E} \left[f(x_K^{av}) + g(x_K^{av}) + h(Ax_K^{av}) - f(x_\star) - g(x_\star) - h(Ax_\star) \right] \le \mathbb{E} \left[\mathcal{G}_{\frac{2c_1}{\underline{p}K}, \frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0) \right]$$

$$+ \frac{c_1}{\underline{p}K} \| x_\star - x_0 \|^2_{\tau^{-1}P^{-1}} + \frac{4c_2}{\underline{p}K} L_h^2.$$

• If $h(\cdot) = \delta_b(\cdot)$, we use (Tran-Dinh et al., 2018, Lemma 1) to obtain

$$\mathbb{E} \left[f(x_K^{av}) + g(x_K^{av}) - f(x_\star) - g(x_\star) \right] \le \mathbb{E} \left[\mathcal{G}_{\frac{2c_1}{\underline{p}K}, \frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0) \right] + \frac{c_1}{\underline{p}K} \| x_0 - x_\star \|^2_{\tau^{-1}P^{-1}}$$

$$+ \frac{c_2}{\underline{p}K} \| y_\star - y_0 \|^2_{\sigma^{-1}\pi^{-1}} + \mathbb{E} \left[\| y_\star \|_{\sigma^{-1}\pi^{-1}} \| Ax_K^{av} - b \|_{\sigma\pi} \right]$$

$$\mathbb{E} \left[\| Ax - b \|_{\sigma\pi} \right] \le \frac{2c_2}{\underline{p}K} \| y_\star - y_0 \|_{\sigma^{-1}\pi^{-1}}$$

$$+ \frac{2c_2}{\underline{p}K} \sqrt{ \| y_\star - y_0 \|^2_{\sigma^{-1}\pi^{-1}} + \frac{\underline{p}K}{c_2} \left(\mathbb{E} \left[\mathcal{G}_{\frac{2c_1}{\underline{p}K}, \frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0) \right] + \frac{c_1}{\underline{p}K} \| x_0 - x_\star \|^2_{\tau^{-1}P^{-1}} \right) }.$$

We plug in the bound of $\mathbb{E} \left[\mathcal{G}_{\frac{2c_1}{\underline{p}K}, \frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0) \right]$ to obtain the final results. $\qquad \square$

Simplification of the constants. As mentioned before Theorem 3, we now give the inequalities we use to obtain the bounds we have in the main text for Theorem 3 and Theorem 4 compared to the ones we have in the appendix. It is easy to see by using coarse inequalities, we first $p_i \underline{p}^{-1} \le \underline{p}^{-1}$, second, $2p_i - \underline{p} \le 2$, third, $\pi_j^{-1} - 1 \le \underline{p}^{-1}$ as $\pi_j \ge \underline{p}$. Finally, by the definition of τ_i in (9), we can derive $\| \tau^{1/2} P^{1/2} A^\top \pi^{-1/2} \sigma^{1/2} \|^2 \le 2\underline{p}^{-1}$. By using these constants in the bounds of Theorem 3 and Theorem 4 in the appendix, we arrive at the bounds given for these theorems in the main text.