On the convergence of stochastic primal-dual hybrid gradient

Ahmet Alacaoglu[§] Olivier Fercoq[†] Volkan Cevher[‡]

[§]University of Wisconsin-Madison, USA
[†]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[‡]LIONS, Ecole Polytechnique Fédérale de Lausanne, Switzerland

June 23, 2022

Abstract

In this paper, we analyze the recently proposed stochastic primal-dual hybrid gradient (SPDHG) algorithm and provide new theoretical results. In particular, we prove almost sure convergence of the iterates to a solution with convexity and linear convergence with further structure, using standard step sizes independent of strong convexity or other regularity constants. In the general convex case, we also prove the $\mathcal{O}(1/k)$ convergence rate for the ergodic sequence, on expected primal-dual gap function. Our assumption for linear convergence is metric subregularity, which is satisfied for strongly convex-strongly concave problems in addition to many nonsmooth and/or nonstrongly convex problems, such as linear programs, Lasso, and support vector machines. We also provide numerical evidence showing that SPDHG with standard step sizes shows a competitive practical performance against its specialized strongly convex variant SPDHG-μ and other state-of-the-art algorithms including variance reduction methods.

1 Introduction

Stochastic primal-dual hybrid gradient (SPDHG) algorithm is proposed by Chambolle et al. [6], for solving the optimization problem

$$\min_{x\in\mathcal{X}} \sum_{i=1}^{n} f_i(A_i x) + g(x), \tag{1.1}$$

where $f_i\colon \mathcal{Y}_i \to \mathbb{R}\cup\{+\infty\}$ and $g\colon \mathcal{X}\to\mathbb{R}\cup\{+\infty\}$ are proper, lower semicontinuous (l.s.c.), convex functions and f is defined as the separable function such that $f(y) = \sum_{i=1}^n f_i(y_i)$. $A_i\colon \mathcal{X}\to\mathcal{Y}_i$ is a linear mapping and A is defined such that $(Ax)_i = A_i x$.

The classical approaches provide numerical solutions to (1.1) via primal-dual methods. In particular, a common strategy is to have coordinate-based updates for the separable dual variable [6, 52]. These methods show competitive practical performance and are proven to converge linearly under the assumption that $f_i^*, \forall i$ and g are μ_i and μ_g-strongly convex functions, respectively. Step sizes of these methods in turn depend on μ_i, μ_g to obtain linear convergence. SPDHG belongs to this class.

Chambolle et al. provide convergence analysis for SPDHG under various assumptions on the problem template [6]. Indeed, SPDHG is a variant of celebrated primal-dual hybrid gradient (PDHG) method [7, 8] where the main difference is stochastic block updates for dual variables at each iteration. In the general convex case, [6] proved that a particular Bregman distance between the iterates of SPDHG and any primal-dual solution converges almost surely to 0 and the ergodic sequence has a $\mathcal{O}(1/k)$ rate for this quantity. Note however that this result does not imply the almost sure convergence of the sequence to a solution, in general. However, this result does not give guarantees on the expected primal-dual gap function (see (4.28), (4.21)), which is the standard optimality measure. If f_i^* and g are strongly convex functions, SPDHG-μ, which is a variant of SPDHG with step sizes depending on strong convexity constants, is proven to converge linearly [6, Theorem 6.1]. Estimation of strong convexity constants can be challenging in practice, restricting the use of SPDHG-μ.

Since its introduction, SPDHG has been popular in practice, especially in computational imaging, with implementations in different software packages [16, 26, 32, 38]. Despite the practical interest, fundamental theoretical results regarding the convergence of SPDHG remained open, including almost sure convergence, $\mathcal{O}(1/k)$ convergence rate for expected primal-dual gap and adaptive linear convergence.

In its most basic form, step sizes of SPDHG are determined using $\|A_i\|$ and probabilities of selecting coordinates [6]. It is often observed in practice that the last iterate of PDHG or SPDHG with these step sizes has competitive practical performance. Yet, only ergodic rates are known for this method with restrictive assumptions [6, 8]. In this paper, we analyze SPDHG with standard step sizes and provide new theoretical results, paving the way for explaining its fast convergence behavior in practice.

1.1 Our contributions

We prove the following results for SPDHG:

General convex case We prove that the iterates of SPDHG converge almost surely to a solution. For this purpose, we introduce a representation of SPDHG as a fixed point operator in a duplicated space. For the ergodic sequence, we show that SPDHG has $\mathcal{O}(1/k)$ rate of convergence for the expected primal-dual gap. To prove this result, we introduce a generic technique that is applicable to other stochastic primal-dual coordinate descent algorithms. Moreover, we prove the same rate for objective residual and feasibility for linearly constrained problems.

Metrically subregular case When the problem is metrically subregular (see Section 2.3), we prove that SPDHG has linear convergence with standard step sizes, depending only on A_i and probabilities for selecting coordinates. Our result shows that without any modification, basic SPDHG adapts to problem structure and attains linear rate when the assumption holds.

Practical performance We show that SPDHG shows a robust and competitive practical performance compared to SPDHG-μ of [6] and other state-of-the-art methods including variance reduction and primal-dual coordinate descent methods.

We summarize our results and compare with those of [6] in Table 2 (Page 26).

2 Preliminaries

2.1 Notation

We assume that $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces and that $\mathcal{Y} = \prod_{i=1}^n \mathcal{Y}_i$. We define $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $z = (x, y) \in \mathcal{Z}$. For positive definite Q, we use $\langle x, y\rangle_Q = \langle Qx, y\rangle$ for denoting weighted inner product and $\|x\|_Q^2 = \langle Qx, x\rangle$ for weighted Euclidean norm. We overload these notations to also write for a vector σ with $\sigma_i > 0$, $\|y\|_\sigma^2 = \langle y, \operatorname{diag}(\sigma) y\rangle$. For a set $\mathcal{C}$, and positive definite Q, distance of a point x to $\mathcal{C}$, measured in $\|\cdot\|_Q$ is defined as $\operatorname{dist}_Q^2(x, \mathcal{C}) = \min_{y\in\mathcal{C}} \|x - y\|_Q^2 = \|x - \mathcal{P}_\mathcal{C}^Q(x)\|_Q^2$, where we have defined the projection operator $\mathcal{P}$ implicitly. When $Q = I$, we drop the subscript and write $\operatorname{dist}(x, \mathcal{C})$. For $\sigma \in \mathbb{R}^n$, we use the elementwise inverse $\sigma^{-1} = (\sigma_1^{-1}, \dots, \sigma_n^{-1})$. Domain of a function h is denoted as $\operatorname{dom} h$. We encode constraints using the indicator function: $\delta_{\{b\}}(x) = 0$ if $x = b$ and $\delta_{\{b\}}(x) = +\infty$ if $x \neq b$.

Given a vector x, we access i^{th} element as x_i. We define $e(i) \in \mathcal{Y}$ such that $e(i)_j = 1$, if $j = i$ and $e(i)_j = 0$, if $j \neq i$. Moreover, we use $E(i) = e(i)e(i)^\top$. Unless used with a subscript, 1 in Kronecker products denotes $1_n \in \mathbb{R}^n$, all-ones vector.

Given a vector $x \in \mathcal{X}$, we use bold symbol $\boldsymbol{x}$ to denote the duplicated version of this vector, which consists of n *copies* of x, and the corresponding space is denoted by $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$. Similarly, the duplicated dual space is $\boldsymbol{\mathcal{Y}} = \mathcal{Y}^n$ and $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Y}}$. The copies might be the same, or different, depending on how $\boldsymbol{x}$ is set. To access i^{th} copy, we use the notation $\boldsymbol{x}(i) \in \mathcal{X}$. For the operator $T\colon \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Z}}$, and a duplicated vector $\boldsymbol{q} \in \boldsymbol{\mathcal{Z}}$, we denote the output as $T(\boldsymbol{q}) = \binom{T_x(\boldsymbol{q})}{T_y(\boldsymbol{q})}$. For example, i^{th} primal copy is denoted as $T_x(\boldsymbol{q})(i) \in \mathcal{X}$. Similarly, for the i^{th} primal copy in $\boldsymbol{q}$, we use $\boldsymbol{q}_x(i) \in \mathcal{X}$. To access i^{th} primal and dual copies, we use $\boldsymbol{q}(i) \in \mathcal{Z}$.

For example, when we pick one coordinate at a time, we can set $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^n$, which would result in the duplicated spaces $\boldsymbol{\mathcal{X}} = \mathbb{R}^{dn}$, $\boldsymbol{\mathcal{Y}} = \mathbb{R}^{n^2}$, and $\boldsymbol{\mathcal{Z}} = \mathbb{R}^{dn+n^2}$.

Probability of selecting an index $i \in \{1, \dots, n\}$ is denoted as $p_i > 0$, with $\sum_{i=1}^n p_i = 1$. We define $P = \operatorname{diag}(p_1, \dots, p_n)$ and $\underline{p} = \min_i p_i$. Notation $\mathcal{F}_k$ defines the filtration generated by randomly selected indices $\{i_1, \dots, i_{k-1}\}$. Let $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation with respect to $\mathcal{F}_k$.

The proximal operator of a function h is defined as

$$\operatorname{prox}_{\tau,h}(x) = \arg\min_{u \in \mathcal{X}} h(u) + \frac{1}{2}\|u - x\|^2_{\tau^{-1}}. \tag{2.1}$$

The Fenchel conjugate of h is defined as $h^*(y) = \sup_{z \in \mathcal{X}} \langle z, y\rangle - h(z)$.

2.2 Solution

Using Fenchel conjugate, (1.1) is cast as the saddle point problem

$$\min_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \sum_{i=1}^n \langle A_i x, y_i\rangle - f_i^*(y_i) + g(x). \tag{2.2}$$

A primal-dual solution $(x^\star, y^\star) \in \mathcal{Z}^\star$ is characterized as

$$0 \in \begin{bmatrix} A^\top y^\star + \partial g(x^\star) \\ A x^\star - \partial f^*(y^\star) \end{bmatrix} = F(x^\star, y^\star). \tag{2.3}$$

Given the functions g and f^* as in (2.2), we define

$$D_g(x; \bar{z}) = g(x) - g(\bar{x}) + \langle A^\top \bar{y}, x - \bar{x}\rangle, \tag{2.4}$$
$$D_{f^*}(y; \bar{z}) = f^*(y) - f^*(\bar{y}) - \langle A\bar{x}, y - \bar{y}\rangle. \tag{2.5}$$

When $\bar{z} = z^\star = (x^\star, y^\star)$, with $z^\star$ denoting a primal-dual solution as defined in (2.3), we have that (2.4) and (2.5) are Bregman distances generated by functions $g(x)$ and $f^*(y)$. In this case, these Bregman distances measure the distance between x and $x^\star$, and y and $y^\star$, respectively. Given z, $D_h(z; z^\star)$ is the Bregman distance generated by $h(z) = g(x) + f^*(y)$, to measure the distance between z and $z^\star$. Moreover, the primal-dual gap function can be written as $G(z) = \sup_{\bar{z} \in \mathcal{Z}} D_{f^*}(x; \bar{z}) + D_g(y; \bar{z})$.

2.3 Metric subregularity

For Euclidean spaces $\mathcal{U}, \mathcal{V}$ and a set valued mapping $F\colon \mathcal{U} \rightrightarrows \mathcal{V}$, we denote the graph of F by $\operatorname{gra} F = \{(u, v) \in \mathcal{U} \times \mathcal{V}\colon v \in Fu\}$. We say that F is metrically subregular at $\bar{u}$ for $\bar{v}$, with $(\bar{u}, \bar{v}) \in \operatorname{gra} F$, if there exists $\eta_0 > 0$ with a neighborhood of subregularity $\mathcal{N}(\bar{u})$ such that:

$$\operatorname{dist}(u, F^{-1}\bar{v}) \leq \eta_0 \operatorname{dist}(\bar{v}, Fu), \quad \forall u \in \mathcal{N}(\bar{u}). \tag{2.6}$$

If $\mathcal{N}(\bar{u}) = \mathcal{U}$, then F is globally metrically subregular [14]. Absence of metric subregularity is signaled by $\eta_0 = +\infty$. This assumption is used in the context of deterministic and stochastic primal-dual algorithms in [15, 29, 31].

In the paper we shall study how the metric subregularity of the Karush-Kuhn-Tucker (KKT) operator F in (2.3) implies linear convergence of SPDHG.

Metric subregularity of F holds in following cases:

1. f_i^* and g are strongly convex functions, since $\mathcal{N}(\bar{z}) = \mathcal{Z}$.
2. The problem (1.1) is defined with piecewise linear quadratic (PLQ) functions and $\operatorname{dom} g$ and $\operatorname{dom} f^*$ are compact sets, in which case $\mathcal{N}(\bar{z}) = \operatorname{dom} g \times \operatorname{dom} f^*$. In particular the domain of a PLQ function can be represented as the union of finitely many polyhedral sets and in each set, the function is a quadratic (see [29, Definition IV.3]). Problems with PLQ functions include Lasso, support vector machines, linear programs, etc.

Remark 2.1. In the first example above, compact domains are not needed since metric subregularity holds globally for these problems. One can also relax strong convexity in the first case to weaker conditions (see [30]). Importantly, compact domain assumption is only needed in the second example mentioned above in this paper, for PLQs. The reason, as we see in Theorem 4.5 is the lack of control on the low probability event that the trajectory makes an excursion far away. The same assumption for proving linear convergence of another primal-dual coordinate descent method is also needed in [29].

3 Algorithm

The algorithm SPDHG is given as Algorithm 1.

Algorithm 1 Stochastic PDHG (SPDHG) [6, Algorithm 1]

Input: Pick step sizes σ_i, τ by (3.1) and $x^0 \in \mathcal{X}$, $y^0 = y^1 = \bar{y}^1 \in \mathcal{Y}$. Given $P = \operatorname{diag}(p_1, \dots, p_n)$.
for $k = 1, 2, \dots$ **do**
  $x^k = \operatorname{prox}_{\tau,g}(x^{k-1} - \tau A^\top \bar{y}^k)$
  Draw $i_k \in \{1, \dots, n\}$ such that $\Pr(i_k = i) = p_i$.
  $y_{i_k}^{k+1} = \operatorname{prox}_{\sigma_{i_k}, f_{i_k}^*}(y_{i_k}^k + \sigma_{i_k} A_{i_k} x^k)$
  $y_i^{k+1} = y_i^k, \quad \forall i \neq i_k$
  $\bar{y}^{k+1} = y^{k+1} + P^{-1}(y^{k+1} - y^k)$,
end for

Remark 3.1. We use serial sampling of blocks in our analysis for the ease of notation. We can extend our results with other samplings by using expected separable overapproximation (ESO) inequality as in [6].

We use the standard step size rules for primal and dual step sizes [6]:

$$p_i^{-1} \tau \sigma_i \|A_i\|^2 \leq \gamma^2 < 1. \tag{3.1}$$

Assumption 1. We have the following assumptions concerning (1.1).

1. f_i and g are proper, lower semicontinuous (l.s.c.), convex functions.
2. The set of solutions to (1.1) is nonempty.
3. Slater's condition holds, namely $0 \in \operatorname{ri}(\operatorname{dom} f - A \operatorname{dom} g)$ where ri stands for relative interior [4].

Slater's condition is a standard sufficient assumption for strong duality, used frequently for primal-dual methods [4, 6, 7, 19, 29, 47]. Strong duality ensures that a dual solution exists in (2.2) and the set of primal-dual solutions is characterized by (2.3).

4 Convergence

We start with a lemma analyzing one iteration behavior of the algorithm. This lemma is essentially the same as [6, Lemma 4.4] up to minor modifications and is included for completeness, with its proof in Section 8.3. We first introduce some notations.

$$\begin{aligned} V(z) &= \frac{1}{2}\|x\|_{\tau^{-1}}^2 + \frac{1}{2}\|y\|_{\sigma^{-1}P^{-1}}^2 + \langle Ax, P^{-1}y\rangle, \\ V_k(x,y) &= \frac{1}{2}\|x\|_{\tau^{-1}}^2 - \langle Ax, P^{-1}(y^k - y^{k-1})\rangle + \frac{1}{2}\|y^k - y^{k-1}\|_{\sigma^{-1}P^{-1}}^2 + \frac{1}{2}\|y\|_{\sigma^{-1}P^{-1}}^2. \end{aligned} \tag{4.1}$$

We also define the full dimensional dual update

$$\hat{y}_i^{k+1} = \operatorname{prox}_{\sigma_i, f_i^*}(y_i^k + \sigma_i A_i x^k), \quad \forall i \in \{1, \dots, n\}. \tag{4.2}$$

Lemma 4.1. *Let Assumption 1 hold. It holds for SPDHG that,* $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$,

$$D_g(x^k; z) + D_{f^*}(\hat{y}^{k+1}; z) \leq V_k(x^{k-1} - x, y^k - y) - \mathbb{E}_k\left[V_{k+1}(x^k - x, y^{k+1} - y)\right] - V(z^k - z^{k-1}). \tag{4.3}$$

Moreover, with $C_1 = 1 - \gamma$, *under the step size rules in (3.1), we have*

$$V(z^k - z^{k-1}) \geq C_1 \left(\frac{1}{2}\|x^k - x^{k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}}\right), \tag{4.4}$$

$$V_k(x, y) \geq C_1 \left(\frac{1}{2}\|x\|^2_{\tau^{-1}} + \frac{1}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}}\right) + \frac{1}{2}\|y\|^2_{\sigma^{-1}P^{-1}}. \tag{4.5}$$

4.1 Almost sure convergence

In this section, we present the almost sure convergence of the iterates of SPDHG to a solution of (1.1).

We start by introducing an equivalent representation of SPDHG that is instrumental in our proofs. The motivation of this representation can be seen as similar to [22], where the focus was on PDHG. In particular, this representation shifts the primal update so that the algorithm can be written as a fixed point operator. Since $\bar{y}^{k+1}$ depends on the selected index i_k at iteration k, the operator T is defined such that all the possible values of $\bar{y}^{k+1}$ and consequently, of x^{k+1} are captured.

Lemma 4.2. *Let us define* $T\colon \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Z}}$ *that to* $(\boldsymbol{x}, \boldsymbol{y})$ *associates* $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ *such that* $\forall i \in \{1, \dots, n\}$,

$$\begin{aligned}
\hat{\boldsymbol{y}}(i) &= \operatorname{prox}_{\sigma, f^*}(\boldsymbol{y}(i) + \operatorname{diag}(\sigma) A \boldsymbol{x}(i)) \\
\bar{\boldsymbol{y}}(i) &= \boldsymbol{y}(i) + (1 + p_i^{-1})(\hat{\boldsymbol{y}}(i)_i - \boldsymbol{y}(i)_i) e(i) \\
\hat{\boldsymbol{x}}(i) &= \operatorname{prox}_{\tau, g}(\boldsymbol{x}(i) - \tau A^\top \bar{\boldsymbol{y}}(i))
\end{aligned}$$

where $\boldsymbol{x}(i) \in \mathcal{X}$, $\boldsymbol{y}(i) \in \mathcal{Y}$.

The fixed points of T *are of the form* $(\boldsymbol{x}(i), \boldsymbol{y}(i))$ *such that* $(\boldsymbol{x}(i), \boldsymbol{y}(i)) \in \mathcal{Z}^\star$, $\forall i \in \{1, \dots, n\}$. *Moreover,*

$$\left(x^{k+1}, \hat{y}^{k+1}\right) = \left(T_x(1 \otimes x^k, 1 \otimes y^k)(i_k), T_y(1 \otimes x^k, 1 \otimes y^k)(1)\right).$$

We also denote

$$\begin{aligned}
\bar{S} &= \operatorname{blkdiag}(\tau^{-1} I_{dn \times dn}, I_{n \times n} \otimes \sigma^{-1}), \\
\bar{P} &= \operatorname{blkdiag}(p_1 I_{d \times d}, \dots, p_n I_{d \times d}, p_1 I_{n \times n}, \dots, p_n I_{n \times n}).
\end{aligned}$$

We then have,

$$\|T(1 \otimes x^k, 1 \otimes y^k) - (1 \otimes x^k, 1 \otimes y^k)\|^2_{\bar{S}\bar{P}} = \mathbb{E}_k\left[\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^{k+1} - y^k\|^2_{\sigma^{-1}P^{-1}}\right].$$

Before presenting the proof of the lemma, we use an example to illustrate the notation and the main idea behind it.

Example 1. Let $d = 1$, $n = 2$, then $\boldsymbol{x} = \binom{\boldsymbol{x}(1)}{\boldsymbol{x}(2)} \in \mathbb{R}^2$, $\boldsymbol{y} = \binom{\boldsymbol{y}(1)}{\boldsymbol{y}(2)} \in \mathbb{R}^4$, and

$$\begin{aligned}
\bar{S} &= \operatorname{diag}(\tau^{-1}, \tau^{-1}, \sigma_1^{-1}, \sigma_2^{-1}, \sigma_1^{-1}, \sigma_2^{-1}) \in \mathbb{R}^{6 \times 6}, \\
\bar{P} &= \operatorname{diag}(p_1, p_2, p_1, p_1, p_2, p_2) \in \mathbb{R}^{6 \times 6}.
\end{aligned}$$

Then, we have by letting $\boldsymbol{x} = 1 \otimes x^k$, $\boldsymbol{y} = 1 \otimes y^k$,

$$\begin{aligned}
\hat{\boldsymbol{y}}(1) &= \operatorname{prox}_{\sigma, f^*}(y^k + \operatorname{diag}(\sigma) A x^k), & \hat{\boldsymbol{y}}(2) &= \operatorname{prox}_{\sigma, f^*}(y^k + \operatorname{diag}(\sigma) A x^k), \\
\bar{\boldsymbol{y}}(1) &= y^k + (1 + p_1^{-1}) \begin{bmatrix} \hat{\boldsymbol{y}}(1)_1 - y_1^k \\ 0 \end{bmatrix}, & \bar{\boldsymbol{y}}(2) &= y^k + (1 + p_2^{-1}) \begin{bmatrix} 0 \\ \hat{\boldsymbol{y}}(2)_2 - y_2^k \end{bmatrix}, \\
\hat{\boldsymbol{x}}(1) &= \operatorname{prox}_{\tau, g}(x^k - \tau A^\top \bar{\boldsymbol{y}}(1)), & \hat{\boldsymbol{x}}(2) &= \operatorname{prox}_{\tau, g}(x^k - \tau A^\top \bar{\boldsymbol{y}}(2)).
\end{aligned}$$

We have $T(1 \otimes x^k, 1 \otimes y^k) = \left(\begin{bmatrix} \hat{\boldsymbol{x}}(1) \\ \hat{\boldsymbol{x}}(2) \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{y}}(1) \\ \hat{\boldsymbol{y}}(2) \end{bmatrix}\right)$. By using the definition of $\hat{y}^{k+1}$ in Theorem 4.1, we see that $(x^{k+1}, \hat{y}^{k+1}) = (\hat{\boldsymbol{x}}(1), \hat{\boldsymbol{y}}(1))$ if $i_k = 1$ and $(x^{k+1}, \hat{y}^{k+1}) = (\hat{\boldsymbol{x}}(2), \hat{\boldsymbol{y}}(1))$ if $i_k = 2$. Note that we can take any copy of $\hat{\boldsymbol{y}}$ as $\hat{\boldsymbol{y}}(1) = \hat{\boldsymbol{y}}(2)$. Moreover, depending on i_k, one obtains y^{k+1} from $\hat{y}^{k+1}$ with a coordinate-wise update, as given in SPDHG (see Algorithm 1).

Proof of Theorem 4.2. Let $(\boldsymbol{x}, \boldsymbol{y})$ be a fixed point of T. Then it follows that $\boldsymbol{y}(i) = \text{prox}_{\sigma, f^*}(\boldsymbol{y}(i) + \text{diag}(\sigma) A \boldsymbol{x}(i))$, $\forall i$, $\bar{\boldsymbol{y}}(i) = \boldsymbol{y}(i)$, $\forall i$ and $\boldsymbol{x}(i) = \text{prox}_{\tau, g}(\boldsymbol{x}(i) - \tau A^\top \boldsymbol{y}(i))$, $\forall i$. Hence, optimality conditions for each i are the same as (2.3). Therefore fixed points of T are such that $(\boldsymbol{x}(i), \boldsymbol{y}(i)) \in \mathcal{Z}^\star, \forall i$.

The equality $(x^{k+1}, \hat{y}^{k+1}) = (T_x(1 \otimes x^k, 1 \otimes y^k)(i_k), T_y(1 \otimes x^k, 1 \otimes y^k)(1))$ is just another way to write the algorithm. Since when inputted $(1 \otimes x^k, 1 \otimes y^k)$, T outputs $(1 \otimes \hat{y}^{k+1})$ for the dual variable, we can simply take first copy for $\hat{y}^{k+1}$.

For the last result, we use $\|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}} = \mathbb{E}_k\left[\|y^{k+1} - y^k\|^2_{\sigma^{-1} P^{-1}}\right]$ to show

$$\begin{aligned}
&\|T(1 \otimes x^k, 1 \otimes y^k) - (1 \otimes x^k, 1 \otimes y^k)\|^2_{\bar{S}\bar{P}} \\
&= \sum_{i=1}^n \left(\|T_x(1 \otimes x^k, 1 \otimes y^k)(i) - x^k\|^2_{\tau^{-1}} p_i + \|T_y(1 \otimes x^k, 1 \otimes y^k)(i) - y^k\|^2_{\sigma^{-1}} p_i\right) \\
&= \sum_{i=1}^n \left(\|T_x(1 \otimes x^k, 1 \otimes y^k)(i) - x^k\|^2_{\tau^{-1}} p_i\right) + \|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}} \left(\sum_{i=1}^n p_i\right) \\
&= \mathbb{E}_k\left[\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^{k+1} - y^k\|^2_{\sigma^{-1} P^{-1}}\right],
\end{aligned}$$

where we also used that $\sum_{i=1}^n p_i = 1$. □

We proceed with the main theorem of this section. We present the main ideas and ingredients that make the proof possible in the following proof sketch. The details of the proof using classical arguments from [5, 11, 24] are deferred to Section 8.1. Let us define

$$\Delta^k = V_{k+1}(x^k - x^\star, y^{k+1} - y^\star). \tag{4.6}$$

Theorem 4.3. *Let Assumption 1 hold. Then, it holds that* $\mathbb{E}[V_k(x^{k-1} - x^\star, y^k - y^\star)] \leq \Delta^0$, $\sum_{k=1}^\infty \mathbb{E}[V(z^k - z^{k-1})] \leq \Delta^0$. *Moreover, almost surely, there exists* $(x^\star, y^\star) \in \mathcal{Z}^\star$, *such that the iterates of SPDHG satisfy* $(x^k, y^k) \to (x^\star, y^\star)$.

Proof sketch. On (4.3), we pick $(x, y) = (x^\star, y^\star)$ and by convexity, $D_g(x^k; z^\star) \geq 0$, $D_{f^*}(\hat{y}^{k+1}; z^\star) \geq 0$. Next, by using the definition of Δ^k, we write (4.3) as

$$\mathbb{E}_k\left[\Delta^k\right] \leq \Delta^{k-1} - V(z^k - z^{k-1}).$$

We apply Robbins-Siegmund lemma [43, Theorem 1] to get that almost surely, Δ^k converges to a finite valued random variable and $V(z^k - z^{k-1}) \to 0$. Consequently, by (4.4), $\|y^k - y^{k-1}\|$ converges to 0 almost surely. Since almost surely, Δ^k converges and $\|y^k - y^{k-1}\|$ converges to 0, we have that $\|z^k - z^\star\|$ converges almost surely.

Next, we denote $\boldsymbol{q}^k = (1 \otimes x^k, 1 \otimes y^k)$ and use the arguments in [11, Proposition 2.3], [19, Theorem 1] to argue that there exists a probability 1 set Ω such that for every $z^\star \in \mathcal{Z}^\star$ and for every $\omega \in \Omega$, $\|z^k(\omega) - z^\star\|$ converges and $\|T(\boldsymbol{q}^k(\omega)) - \boldsymbol{q}^k(\omega)\| \to 0$. As for every $\omega \in \Omega$, $(z^k(\omega))_k$ is bounded, we denote by $\tilde{z} = (\tilde{x}, \tilde{y})$ one of its cluster points. Then, we denote $\tilde{\boldsymbol{q}} = (1 \otimes \tilde{x}, 1 \otimes \tilde{y})$ and have that $\tilde{\boldsymbol{q}}$ is a cluster point of $(\boldsymbol{q}^k(\omega))_k$.

The key step in our proof that enables the result is the fixed point characterization of T in Theorem 4.2. With this result, we derive $\tilde{z} \in \mathcal{Z}^\star$ as $\tilde{\boldsymbol{q}}$ is a fixed point of T.

To sum up, we have shown that at least on some subsequence $z^k(\omega)$ converges to $\tilde{z} \in \mathcal{Z}^\star$. As for every $\omega \in \Omega$ and $z^\star \in \mathcal{Z}^\star$, $\|z^k(\omega) - z^\star\|$ converges, the result follows. □

4.2 Linear convergence

The standard approach for showing linear convergence with metric subregularity is to obtain a *Fejer-type inequality* of the form [29]

$$\mathbb{E}_k\left[d(z^{k+1} - z^\star)\right] \leq d(z^k - z^\star) - V(T(z^k) - z^k), \tag{4.7}$$

for suitably defined distance functions d, V and operator T. However, as evident from (4.3) and the definition of V_{k+1}, one iteration result of SPDHG does not fit into this form. When $x = x^\star, y = y^\star$, $V_{k+1}(x^k - x^\star, y^{k+1} - y^\star)$

does not only measure distance to solution, but also the distance of subsequent iterates y^{k+1} and y^k. In addition, V_{k+1} includes $x^k - x^\star$ and $y^{k+1} - y^\star$ rather than $x^{k+1} - x^\star$ and $y^{k+1} - y^\star$, which further presents a challenge due to asymmetry, for using metric subregularity. Therefore, an intricate analysis is needed to control the additional terms and handle the asymmetry in V_{k+1}. In addition, Theorem 4.2 is a necessary tool to identify T.

We need the following notation and lemma which builds on Lemma 4.2, for easier computations with metric subregularity. For the operators, we adopt the convention in [29]. Operator C is the concatenation of subdifferentials, M is the skew symmetric matrix that is formed using matrix A. Operator F is the KKT operator and $\boldsymbol{H}$ is the "metric" that helps us write the algorithm in proximal point form (see Theorem 4.2). Due to duplication in Theorem 4.2, we need duplicated versions of C and M. Consistent with the notation of Theorem 4.2 (also see Section 2.1), we use boldface to denote operators in the duplicated space.

Lemma 4.4. *Under the notations of Lemma 4.2, to write compactly the operation of T, let us define the operators*

$$\begin{aligned}
&C\colon (x,y) \mapsto (\partial g(x), \partial f^*(y)),\\
&M\colon (x,y) \mapsto (A^\top y, -Ax),\\
&\boldsymbol{C}\colon (\boldsymbol{x},\boldsymbol{y}) \mapsto (\partial g(\boldsymbol{x}(1)),\dots,\partial g(\boldsymbol{x}(n)), \partial f^*(\boldsymbol{y}(1)),\dots,\partial f^*(\boldsymbol{y}(n))),\\
&\boldsymbol{M}\colon (\boldsymbol{x},\boldsymbol{y}) \mapsto (A^\top \boldsymbol{y}(1),\dots,A^\top \boldsymbol{y}(n), -A\boldsymbol{x}(1),\dots,-A\boldsymbol{x}(n)),\\
&F = C + M,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{H}\colon (\boldsymbol{x},\boldsymbol{y}) \mapsto \big(&\tau^{-1}\boldsymbol{x}(1) + A^\top(1+p_1^{-1})E(1)\boldsymbol{y}(1),\dots,\\
&\tau^{-1}\boldsymbol{x}(n) + A^\top(1+p_n^{-1})E(n)\boldsymbol{y}(n), \sigma^{-1}\boldsymbol{y}(1),\dots,\sigma^{-1}\boldsymbol{y}(n)\big)\ .
\end{aligned}$$

Let $\boldsymbol{q}^k = (1\otimes x^k, 1\otimes y^k)$, $\hat{\boldsymbol{q}}^{k+1} = T(\boldsymbol{q}^k)$ *and* $\hat{z}^{k+1} = (x^{k+1},\hat{y}^{k+1}) = (\hat{\boldsymbol{q}}_x^{k+1}(i_k), \hat{\boldsymbol{q}}_y^{k+1}(1))$. *Then, we have* $(\boldsymbol{H}-\boldsymbol{M})\boldsymbol{q}^k \in (\boldsymbol{C}+\boldsymbol{H})\hat{\boldsymbol{q}}^{k+1}$, $(\boldsymbol{M}-\boldsymbol{H})(\hat{\boldsymbol{q}}^{k+1}-\boldsymbol{q}^k) \in (\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}^{k+1}$,

$$\mathbb{E}_k\left[\operatorname{dist}^2(0, F\hat{z}^{k+1})\right] = \mathbb{E}_k\left[\operatorname{dist}^2(0,(C+M)\hat{z}^{k+1})\right] = \operatorname{dist}^2_P(0,(\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}^{k+1}).$$

Proof. We start by the representation in Theorem 4.2 by incorporating the update of $\bar{y}^{k+1}$, and recalling the definition of $E(i) = e(i)e(i)^\top$, $\forall i \in \{1,\dots,n\}$

$$\begin{aligned}
\hat{y}(i) &= \operatorname{prox}_{\sigma,f^*}(y(i) + \operatorname{diag}(\sigma)Ax(i))\\
\hat{x}(i) &= \operatorname{prox}_{\tau,g}(x(i) - \tau A^\top\left[y(i) + (1+p_i^{-1})E(i)(\hat{y}(i)-y(i))\right])\\
&= \operatorname{prox}_{\tau,g}(x(i) - \tau A^\top(1+p_i^{-1})E(i)\hat{y}(i) + \tau A^\top(-I_{n\times n} + (1+p_i^{-1})E(i))y(i)).
\end{aligned}$$

We now use the definition of proximal operator to obtain

$$\begin{aligned}
&\sigma^{-1}y(i) + Ax(i) \in \partial f^*(\hat{y}(i)) + \sigma^{-1}\hat{y}(i)\\
&\tau^{-1}x(i) - A^\top y(i) + A^\top(1+p_i^{-1})E(i)y(i) \in \partial g(\hat{x}(i)) + \tau^{-1}\hat{x}(i) + A^\top(1+p_i^{-1})E(i)\hat{y}(i).
\end{aligned}$$

We identify

$$\boldsymbol{H}\boldsymbol{q} = \begin{bmatrix} \tau^{-1}\boldsymbol{x}(1) + A^\top(1+p_1^{-1})E(1)\boldsymbol{y}(1)\\ \vdots\\ \tau^{-1}\boldsymbol{x}(n) + A^\top(1+p_n^{-1})E(n)\boldsymbol{y}(n)\\ \sigma^{-1}\boldsymbol{y}(1)\\ \vdots\\ \sigma^{-1}\boldsymbol{y}(n)\end{bmatrix}, \boldsymbol{M}\boldsymbol{q} = \begin{bmatrix} A^\top\boldsymbol{y}(1)\\ \vdots\\ A^\top\boldsymbol{y}(n)\\ -A\boldsymbol{x}(1)\\ \vdots\\ -A\boldsymbol{x}(n)\end{bmatrix}, \boldsymbol{C}\boldsymbol{q} = \begin{bmatrix} \partial g(\boldsymbol{x}(1))\\ \vdots\\ \partial g(\boldsymbol{x}(n))\\ \partial f^*(\boldsymbol{y}(1))\\ \vdots\\ \partial f^*(\boldsymbol{y}(n))\end{bmatrix}.$$

We set $\boldsymbol{q} = \boldsymbol{q}^k$ and $\hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}^{k+1}$ and use the definition of T in Theorem 4.2 to obtain the first inclusion.

The second inclusion follows by adding to both sides $\boldsymbol{M}\hat{\boldsymbol{q}}^{k+1}$ and rearranging.
For the equality, we write

$$\begin{aligned}\mathbb{E}_k\left[\operatorname{dist}^2(0,(C+M)\hat{z}^{k+1})\right] &= \sum_{i=1}^{n}\operatorname{dist}^2(0,(C+M)\hat{\boldsymbol{q}}^{k+1}(i))p_i \\ &= \operatorname{dist}_P^2(0,(\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}^{k+1}),\end{aligned}$$

where the first equality follows by $\hat{z}^{k+1}=(x^{k+1},\hat{y}^{k+1})=(\hat{\boldsymbol{q}}_x^{k+1}(i_k),\hat{\boldsymbol{q}}_y^{k+1}(1))$ and the second equality is by the definitions of C, M, $\boldsymbol{C}$, and $\boldsymbol{M}$ and $\hat{\boldsymbol{q}}_y^{k+1}(i)=\hat{\boldsymbol{q}}_y^{k+1}(1)$, $\forall i$. □

We continue with our assumption for linear convergence (see Section 2.3).

Assumption 2. Metric subregularity holds for F (see (2.3), Section 2.3) at all $z^\star\in\mathcal{Z}^\star$ for 0 with constant $\eta>0$ using $\|\cdot\|_S$ with $S=\operatorname{diag}(\tau^{-1}1_p,\sigma_1^{-1},\dots,\sigma_n^{-1})$, and the neighborhood of regularity $\mathcal{N}(z^\star)$ contains $\hat{z}^k,\forall k$.

We present our main theoretical development in the next theorem, which states that SPDHG with step sizes in (3.1) attains linear convergence with Assumption 2. The proof idea is to utilize the term $-V(z^k-z^{k-1})$ in (4.3) to obtain contraction. For this, we have to use the results of Theorems 4.2 and 4.4 to write this term with the fixed point characterization given in Theorem 4.2, which allows using metric subregularity.
We denote

$$(x_\star^{k-1},y_\star^k)=\arg\min_{(x,y)\in\mathcal{Z}^\star}V_k(x^{k-1}-x,y^k-y),$$

which exists since V_k is a nonnegative quadratic function. We define (cf. (4.6))

$$\begin{aligned}\Delta^k &= V_{k+1}(x^k-x_\star^k,y^{k+1}-y_\star^{k+1}),\\ \Phi^k &= \Delta^k-\frac{C_1}{4\zeta}\|y^k-y_\star^k\|_{\sigma^{-1}}^2\geq 0.\end{aligned}$$

Theorem 4.5. *Let Assumptions 1 and 2 hold. Then it holds that*

$$\mathbb{E}_k\left[\Delta^k\right]\leq\Delta^{k-1}-V(z^k-z^{k-1}), \tag{4.8}$$

and

$$\mathbb{E}\left[\frac{C_1}{2}\|x^k-x_\star^k\|_{\tau^{-1}}^2+\frac{1}{2}\|y^{k+1}-y_\star^{k+1}\|_{\sigma^{-1}P^{-1}}^2\right]\leq(1-\rho)^k2\Phi^0,$$

where, $\rho=\frac{C_1p}{2\zeta}$, $\zeta=2+2\eta^2\|\boldsymbol{H}-\boldsymbol{M}\|^2$, $C_1=1-\gamma$.

Proof. Starting from the result of Theorem 4.1, we have

$$\begin{aligned}D_g(x^k;z)+D_{f^*}(\hat{y}^{k+1};z)\leq&-\mathbb{E}_k\left[V_{k+1}(x^k-x,y^{k+1}-y)\right]\\ &+V_k(x^{k-1}-x,y^k-y)-V(z^k-z^{k-1}).\end{aligned}\tag{4.9}$$

We pick $x=x_\star^{k-1}$, $y=y_\star^k$, with $z_\star^k=(x_\star^{k-1},y_\star^k)$ and use convexity to get $D_g(x^k;z_\star^k)\geq 0$ and $D_{f^*}(\hat{y}^{k+1};z_\star^k)\geq 0$. In addition, we define

$$\begin{aligned}\Delta^{k-1} &= V_k(x^{k-1}-x_\star^{k-1},y^k-y_\star^k)\\ \tilde{\Delta}^k &= V_{k+1}(x^k-x_\star^{k-1},y^{k+1}-y_\star^k).\end{aligned}$$

We use these definitions in (4.9) to write

$$\mathbb{E}_k\left[\tilde{\Delta}^k\right]\leq\Delta^{k-1}-V(z^k-z^{k-1}).$$

By definition of $(x_\star^k, y_\star^{k+1})$, we have $\Delta^k \leq \tilde{\Delta}^k$, which implies that

$$\mathbb{E}_k\left[\Delta^k\right] \leq \Delta^{k-1} - V(z^k - z^{k-1}).$$

Recursion of this inequality gives boundedness of the iterates x_k and y_k, in expectation. However, it is not possible to derive sure boundedness of the sequence. Without sure boundedness, the set that includes x_k, y_k depends on the specific trajectory of the algorithm, and it is not possible to find a set independent of these. As metric subregularity holds for PLQs with a bounded neighborhood (see Section 2.3), we cannot utilize this result and this is the main reason for the need for bounded domains in this case. This assumption would ensure sure boundedness of the sequence, which gives us a suitable set to use for using metric subregularity assumption for PLQs.

We recall $S = \operatorname{diag}(\tau^{-1}1_p, \sigma_1^{-1}, \ldots, \sigma_n^{-1})$, $\bar{S}$ and $\bar{P}$ are as defined in Theorem 4.2, and $\operatorname{dist}_S^2(z^k, \mathcal{Z}^\star) = \|z^k - \mathcal{P}_{\mathcal{Z}^\star}^S(z^k)\|_S^2 = \|x^k - \tilde{x}_\star^k\|_{\tau^{-1}}^2 + \|y^k - y_\star^k\|_{\sigma^{-1}}^2$ where $\tilde{x}_\star^k$ is the projection of x^k onto the set of solutions with respect to norm $\|\cdot\|_{\tau^{-1}}$. We now use Assumption 2 stating that $F = C + M$ is metrically subregular at $\mathcal{P}_{\mathcal{Z}^\star}^S(\hat{z}^{k+1})$ for 0. We recall, $\boldsymbol{q}^k = (1 \otimes x^k, 1 \otimes y^k)$ and $\hat{\boldsymbol{q}}^{k+1} = T(\boldsymbol{q}^k)$ and estimate as

$$\begin{aligned}\|x^k - \tilde{x}_\star^k\|_{\tau^{-1}}^2 + \|y^k - y_\star^k\|_{\sigma^{-1}}^2 = \operatorname{dist}_S^2(z^k, \mathcal{Z}^\star) &\leq \mathbb{E}_k\left[\|z^k - \mathcal{P}_{\mathcal{Z}^\star}^S(\hat{z}^{k+1})\|_S^2\right] \\ &\leq 2\mathbb{E}_k\left[\|z^k - \hat{z}^{k+1}\|_S^2\right] + 2\mathbb{E}_k\left[\|\hat{z}^{k+1} - \mathcal{P}_{\mathcal{Z}^\star}^S(\hat{z}^{k+1})\|_S^2\right], \end{aligned} \tag{4.10}$$

where the first inequality is due to the definition of $\operatorname{dist}_S^2(z^k, \mathcal{Z}^\star)$. Next, we estimate the second term on RHS

$$\begin{aligned}2\mathbb{E}_k\left[\|\hat{z}^{k+1} - \mathcal{P}_{\mathcal{Z}^\star}^S(\hat{z}^{k+1})\|_S^2\right] &\leq 2\eta^2\mathbb{E}_k\left[\operatorname{dist}_S^2(0, (C+M)\hat{z}^{k+1})\right] \\ &= 2\eta^2 \operatorname{dist}_{\bar{S}\bar{P}}^2(0, (\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}^{k+1}) \leq 2\eta^2\|\boldsymbol{M}-\boldsymbol{H}\|^2\|\hat{\boldsymbol{q}}^{k+1} - \boldsymbol{q}^k\|_{\bar{S}\bar{P}}^2, \end{aligned} \tag{4.11}$$

with the first inequality being due to metric subregularity of $C + M$ (see Remark 4.6) since $\operatorname{dist}_S^2(\hat{z}^{k+1}, \mathcal{Z}^\star) = \|\hat{z}^{k+1} - \mathcal{P}_{\mathcal{Z}^\star}^S(\hat{z}^{k+1})\|_S^2$. First equality and second inequality are by Theorem 4.4 and Cauchy-Schwarz inequality. Joining the estimates give

$$\|x^k - \tilde{x}_\star^k\|_{\tau^{-1}}^2 + \|y^k - y_\star^k\|_{\sigma^{-1}}^2 \leq 2\mathbb{E}_k\left[\|z^k - \hat{z}^{k+1}\|_S^2\right] + 2\eta^2\|\boldsymbol{M}-\boldsymbol{H}\|^2\|\hat{\boldsymbol{q}}^{k+1} - \boldsymbol{q}^k\|_{\bar{S}\bar{P}}^2. \tag{4.12}$$

First, we use $\|\hat{y}^{k+1} - y^k\|_{\sigma^{-1}}^2 = \mathbb{E}_k\left[\|y^{k+1} - y^k\|_{\sigma^{-1}P^{-1}}^2\right]$ to estimate

$$\begin{aligned}\mathbb{E}_k\left[\|z^k - \hat{z}^{k+1}\|_S^2\right] &= \mathbb{E}_k\left[\|x^{k+1} - x^k\|_{\tau^{-1}}^2\right] + \|\hat{y}^{k+1} - y^k\|_{\sigma^{-1}}^2 \\ &= \mathbb{E}_k\left[\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^{k+1} - y^k\|_{\sigma^{-1}P^{-1}}^2\right]. \end{aligned} \tag{4.13}$$

Second, we use Lemma 4.2 to obtain

$$\begin{aligned}\|\hat{\boldsymbol{q}}^{k+1} - \boldsymbol{q}^k\|_{\bar{S}\bar{P}}^2 &= \|T(1 \otimes x^k, 1 \otimes y^k) - (1 \otimes x^k, 1 \otimes y^k)\|_{\bar{S}\bar{P}}^2 \\ &= \mathbb{E}_k\left[\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^{k+1} - y^k\|_{\sigma^{-1}P^{-1}}^2\right]. \end{aligned} \tag{4.14}$$

We combine (4.13) and (4.14) in (4.10) to get

$$\begin{aligned}&\frac{1}{2}\|x^k - \tilde{x}_\star^k\|_{\tau^{-1}}^2 + \frac{1}{2}\|y^k - y_\star^k\|_{\sigma^{-1}}^2 \\ &\qquad \leq (2 + 2\eta^2\|N - H\|^2)\mathbb{E}_k\left[\frac{1}{2}\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \frac{1}{2}\|y^{k+1} - y^k\|_{\sigma^{-1}P^{-1}}^2\right]. \end{aligned} \tag{4.15}$$

Herein, we denote $\zeta = 2 + 2\eta^2\|\boldsymbol{H} - \boldsymbol{M}\|^2$.

By using (4.4), we have that, for all $\alpha \in [0, 1]$

$$\begin{aligned}\mathbb{E}_{k-1}\left[V(z^k - z^{k-1})\right] &\geq C_1\mathbb{E}_{k-1}\left[\frac{1}{2}\|x^k - x^{k-1}\|_{\tau^{-1}}^2 + \frac{1}{2}\|y^k - y^{k-1}\|_{\sigma^{-1}P^{-1}}^2\right] \\ &\geq \frac{C_1}{\zeta}\left(\frac{\alpha}{2}\|x^{k-1} - \tilde{x}_\star^{k-1}\|_{\tau^{-1}}^2 + \frac{1}{2}\|y^{k-1} - y_\star^{k-1}\|_{\sigma^{-1}}^2\right), \end{aligned} \tag{4.16}$$

where the second inequality is due to (4.15) and $\alpha \geq 1$.

We have, by definition of $x^\star_{k-1}$ that

$$\begin{aligned}\Delta^{k-1} &\leq V_k(x^{k-1} - \tilde{x}^{k-1}_\star, y^k - y^k_\star) \\ &= \frac{1}{2}\|x^{k-1} - \tilde{x}^{k-1}_\star\|^2_{\tau^{-1}} + \frac{1}{2}\|y^k - y^k_\star\|^2_{\sigma^{-1}P^{-1}} + \frac{1}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}} \\ &\quad - \langle P^{-1}A(x^{k-1} - \tilde{x}^{k-1}_\star), y^k - y^{k-1}\rangle.\end{aligned}$$

Next, by Cauchy-Schwarz and Young's inequalities with (3.1), we have

$$-\langle P^{-1}A(x^{k-1} - \tilde{x}^{k-1}_\star), y^k - y^{k-1}\rangle \leq \frac{\gamma}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}} + \frac{\gamma}{2}\|x^{k-1} - \tilde{x}^{k-1}_\star\|^2_{\tau^{-1}}.$$

Using the final estimate and adding and subtracting $\frac{1+\gamma}{2\alpha}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}}$ gives

$$\begin{aligned}\Delta^{k-1} \leq \frac{1+\gamma}{2}\|x^{k-1} - \tilde{x}^{k-1}_\star\|^2_{\tau^{-1}} + \frac{1+\gamma}{2\alpha}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}} \\ + \frac{1+\gamma}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}} - \frac{1+\gamma}{2\alpha}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}}. \quad (4.17)\end{aligned}$$

We now take conditional expectation of both sides and use (4.16) to get

$$\begin{aligned}\mathbb{E}_{k-1}\left[\Delta^{k-1}\right] \leq \frac{(1+\gamma)\zeta}{C_1\alpha}\mathbb{E}_{k-1}\left[V(z^k - z^{k-1})\right] + \frac{1+\gamma}{2}\mathbb{E}_{k-1}\left[\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}}\right] \\ + \frac{1}{2}\mathbb{E}_{k-1}\left[\|y^k - y^k_\star\|^2_{\sigma^{-1}P^{-1}}\right] - \frac{1+\gamma}{2\alpha}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}}.\end{aligned}$$

By using (4.4) and requiring that $\frac{(1+\gamma)}{C_1} \leq \frac{(1+\gamma)\zeta}{C_1\alpha}$, or equivalently $\zeta \geq \alpha$, which is not restrictive since α is finite, and one can increase η as in (2.6) to satisfy the requirement, we can combine the first two terms in the right hand side to get

$$\begin{aligned}\mathbb{E}_{k-1}\left[\Delta^{k-1}\right] \leq \frac{2(1+\gamma)\zeta}{C_1\alpha}\mathbb{E}_{k-1}\left[V(z^k - z^{k-1})\right] + \frac{1}{2}\mathbb{E}_{k-1}\left[\|y^k - y^k_\star\|^2_{\sigma^{-1}P^{-1}}\right] \\ - \frac{1+\gamma}{2\alpha}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}}.\end{aligned}$$

We now insert this inequality into (4.8) and use that $\mathbb{E}_{k-1}\left[\mathbb{E}_k\left[\Delta^k\right]\right] = \mathbb{E}_{k-1}\left[\Delta^k\right]$

$$\begin{aligned}\mathbb{E}_{k-1}\left[\Delta^k\right] \leq \mathbb{E}_{k-1}\left[\Delta^{k-1}\right] - \frac{C_1\alpha}{2(1+\gamma)\zeta}\mathbb{E}_{k-1}\left[\Delta^{k-1}\right] \\ + \frac{C_1\alpha}{4(1+\gamma)\zeta}\mathbb{E}_{k-1}\left[\|y^k - y^k_\star\|^2_{\sigma^{-1}P^{-1}}\right] - \frac{C_1}{4\zeta}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}}.\end{aligned}$$

We take full expectation and rearrange to get

$$\begin{aligned}\mathbb{E}\Big[\Delta^k - \frac{C_1\alpha}{4(1+\gamma)\zeta}\|y^k - y^k_\star\|^2_{\sigma^{-1}P^{-1}}\Big] \\ \leq \left(1 - \frac{C_1\alpha}{2(1+\gamma)\zeta}\right)\mathbb{E}\Big[\Delta^{k-1} - \frac{C_1}{4\zeta(1 - \frac{C_1\alpha}{2(1+\gamma)\zeta})}\|y^{k-1} - y^{k-1}_\star\|^2_{\sigma^{-1}}\Big]. \quad (4.18)\end{aligned}$$

We require

$$C_2 = \frac{C_1\alpha}{4\underline{p}(1+\gamma)\zeta} \leq \frac{C_1}{4\zeta} \leq \frac{C_1}{4\zeta(1 - \frac{C_1\alpha}{2(1+\gamma)\zeta})} \iff \alpha \leq (1+\gamma)\underline{p}. \quad (4.19)$$

Let us pick $\alpha = (1+\gamma)\underline{p}$ so that $C_2 = \frac{C_1}{4\zeta}$ and define

$$\Phi^k = \Delta^k - C_2\|y^k - y^k_\star\|^2_{\sigma^{-1}}.$$

We note (4.16) and (4.8) to have

$$\|y^k - y^k_\star\|^2_{\sigma^{-1}} \leq \frac{2\zeta}{C_1}\mathbb{E}_k\left[V(z^{k+1} - z^k)\right] \leq \frac{2\zeta}{C_1}\mathbb{E}_k\left[\Delta^k\right].$$

Then, we can lower bound Φ^k as

$$\mathbb{E}\left[\Phi^k\right] \geq \left(1 - C_2\frac{2\zeta}{C_1}\right)\mathbb{E}\left[\Delta^k\right] = \frac{1}{2}\mathbb{E}\left[\Delta^k\right]. \tag{4.20}$$

Therefore, it follows that $\mathbb{E}\left[\Phi^k\right]$ is nonnegative, by the definition of Δ^k and (4.5).

We can now rewrite (4.18) as

$$\mathbb{E}\left[\Phi^k\right] \leq (1-\rho)\mathbb{E}\left[\Phi^{k-1}\right],$$

where $\rho = \frac{C_1 \underline{p}}{2\zeta}$. We have shown that Φ^k converges linearly to 0 in expectation.

By (4.20), it immediately follows that Δ^k converges linearly to 0.

To conclude, we note $\Delta^k = V_{k+1}(x^k - x^k_\star, y^{k+1} - y^{k+1}_\star)$, and (4.5), from which we conclude linear convergence of $\|x^k - x^k_\star\|^2_{\tau^{-1}}$ and $\|y^{k+1} - y^{k+1}_\star\|^2_{\sigma^{-1}P^{-1}}$.

It is obvious to see that $0 < \rho$ follows by the fact that η is finite by metric subregularity and $\rho < 1$ follows since $\gamma < 1$ and $\underline{p} \leq 1$. □

One important remark about Theorem 4.5 is that the knowledge of the metric subregularity constant η is not needed for running the algorithm. Step sizes are chosen as (3.1) and linear convergence follows directly when Assumption 2 holds. Important examples where Assumption 2 holds are given in Section 2.3.

Even though Assumption 2 is more general than prior assumptions for linear convergence and our result is agnostic to the choice of the step size, we observe in practice that SPDHG can be much faster than the rate derived in Theorem 4.5. We reflect on this issue more in Section 7 and present some open questions in this context.

Remark 4.6. Strictly speaking, metric subregularity is used in Theorem 4.5 in the weighted norm, *i.e.,*

$$\operatorname{dist}_S(z, \mathcal{Z}^\star) \leq \eta \operatorname{dist}_S(0, Fz),$$

where $S = \operatorname{diag}(\tau^{-1}1_p, \sigma_1^{-1}, \ldots, \sigma_n^{-1})$. In terms of the definition in (2.6) if η_0 is the constant using the standard Euclidean norm, it is obvious that $\eta \leq \|S\|\|S^{-1}\|\eta_0$, but we use η in Theorem 4.5 since it can be smaller, resulting in a better rate.

4.3 Sublinear convergence

In this section, we prove $\mathcal{O}(1/k)$ convergence rates for the ergodic sequence with different optimality measures.

4.3.1 Convergence of expected primal-dual gap

We recall the definition of the primal-dual gap function,

$$\begin{aligned} G(\bar{x}, \bar{y}) &= \sup_{z \in \mathcal{Z}} \mathcal{H}(\bar{x}, \bar{y}; x, y) \\ &:= \sup_{z \in \mathcal{Z}} g(\bar{x}) + \langle A\bar{x}, y\rangle - f^*(y) - g(x) - \langle Ax, \bar{y}\rangle + f^*(\bar{y}). \end{aligned} \tag{4.21}$$

It is also possible to consider the restricted primal-dual gap in the sense of [6, 7], which for any set $\mathcal{B} = \mathcal{B}_x \times \mathcal{B}_y \subseteq \mathcal{Z}$ would correspond to

$$G_{\mathcal{B}}(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{B}} \mathcal{H}(\bar{x}, \bar{y}; x, y). \tag{4.22}$$

The main quantity of interest for randomized algorithms is the expected restricted primal-dual gap $\mathbb{E}\left[G_{\mathcal{B}}(\bar{x}, \bar{y})\right]$. As also mentioned in [13], showing convergence rate for this quantity is not straightforward, as the interplay of supremum and expectation can be problematic. In [13], convergence rate is shown in a weaker measure named

as perturbed gap function. We show in the sequel that obtaining the guarantee in expected primal-dual gap is also possible, however, with a more involved analysis.

The expected primal-dual gap proof in [6] has a technical issue, near the end of the proof in [6, Theorem 4.3]. Since the supremum of expectation is upper bounded by the expectation of the supremum, which is in the definition of expected primal-dual gap (4.22), the order of expectation in the proof is incorrect. As we could not find a simple way of fixing the issue using the existing techniques, we introduce a new technique and provide a proof to show that the conclusions of [6, Theorem 4.3], for the primal-dual gap, are still correct, with different constants in the bound.

Our technique in the following proof is inspired by the stochastic approximation literature of variational inequalities and saddle point problems (see [36, Lemma 3.1, Lemma 6.1] for a reference), where such an analysis is used to obtain $\mathcal{O}(1/\sqrt{k})$ rates. In the new proof, we adapt this idea by using the structure of primal-dual coordinate descent to obtain the optimal $\mathcal{O}(1/k)$ rate of convergence. Our technique uses the Euclidean structure of the dual update of SPDHG, therefore might not be directly applicable to cases where general Bregman distances are used for proximal operator, such as in [27, 28].

We start with a lemma to decouple supremum and expectation in the proof.

Lemma 4.7. *Given a point $\tilde{y}^1 \in \mathcal{Y}$, for $k \geq 1$, we define the sequences*

$$v^{k+1} = y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1}), \quad \text{and,} \quad \tilde{y}^{k+1} = \tilde{y}^k - Pv^{k+1}. \tag{4.23}$$

Then, we have for any $y \in \mathcal{Y}$,

$$\sum_{k=1}^{K} \langle \tilde{y}^k - y, v^{k+1} \rangle_{\sigma^{-1}} \leq \frac{1}{2}\|\tilde{y}^1 - y\|^2_{\sigma^{-1}P^{-1}} + \sum_{k=1}^{K} \frac{1}{2}\|v^{k+1}\|^2_{\sigma^{-1}P}, \tag{4.24}$$

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{2}\|v^{k+1}\|^2_{\sigma^{-1}P}\right] \leq \frac{1}{C_1}\Delta^0. \tag{4.25}$$

Moreover, v^k and $\tilde{y}^k$ are $\mathcal{F}_k$-measurable and $\mathbb{E}_k\left[v^{k+1}\right] = 0$.

Proof. For brevity in this proof, we denote $\Upsilon = \sigma^{-1}P^{-1}$. We have $\forall y \in \mathcal{Y}$,

$$\begin{aligned}\frac{1}{2}\|\tilde{y}^{k+1} - y\|^2_{\Upsilon} &= \frac{1}{2}\|\tilde{y}^k - y\|^2_{\Upsilon} - \langle Pv^{k+1}, \tilde{y}^k - y\rangle_{\Upsilon} + \frac{1}{2}\|Pv^{k+1}\|^2_{\Upsilon} \\ &= \frac{1}{2}\|\tilde{y}^k - y\|^2_{\sigma^{-1}P^{-1}} - \langle v^{k+1}, \tilde{y}^k - y\rangle_{\sigma^{-1}} + \frac{1}{2}\|v^{k+1}\|^2_{\sigma^{-1}P}.\end{aligned}$$

Summing this equality gives the first result.

For the second result, we use $\mathbb{E}_k\left[P^{-1}(y^k - y^{k+1})\right] = y^k - \hat{y}^{k+1}$, tower property, and the definition of variance,

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{2}\|v^{k+1}\|^2_{\sigma^{-1}P}\right] &= \sum_{k=1}^{K} \frac{1}{2}\mathbb{E}\left[\mathbb{E}_k\left[\|v^{k+1}\|^2_{\sigma^{-1}P}\right]\right] \\ &\leq \sum_{k=1}^{K} \frac{1}{2}\mathbb{E}\left[\mathbb{E}_k\left[\|P^{-1}(y^{k+1} - y^k)\|^2_{\sigma^{-1}P}\right]\right] \\ &= \sum_{k=1}^{K} \frac{1}{2}\mathbb{E}\left[\|y^{k+1} - y^k\|^2_{\sigma^{-1}P^{-1}}\right] \leq \frac{1}{C_1}\Delta^0,\end{aligned}$$

where the last inequality follows by $\sum_{k=1}^{\infty} \mathbb{E}\left[V(z^{k+1} - z^k)\right] \leq \Delta^0$ from Theorem 4.3 and $\frac{1}{2}\|y^{k+1} - y^k\|^2_{\sigma^{-1}P^{-1}} \leq \frac{1}{C_1}V(z^{k+1} - z^k)$ from Theorem 4.1.

Other results follow immediately by the definition of the sequences and the equality $\mathbb{E}_k\left[y^{k+1} - y^k\right] = P(\hat{y}^{k+1} - y^k)$. □

A direct proof of Theorem 4.1 would proceed by developing terms involving random quantities, by utilizing conditional expectations (see [6]). In this case, however, our approach is to proceed without using conditional expectation since the quantity of interest requires us to take first supremum and then the expectation of the estimates. Our proof strategy is to characterize the error term, and then utilize the results Theorem 4.7 to decouple and bound this term. First, we give the variant of Theorem 4.1 without taking expectations, with its proof given in Section 8.2.

Lemma 4.8. *We define* $f_P^*(y) = \sum_{i=1}^n p_i f_i^*(y_i)$, *and similar to* (2.5) $D_{f^*}^P(\bar{y}; z) = \sum_{i=1}^n p_i f_i^*(\bar{y}_i) - p_i f_i^*(y_i) - \langle (Ax)_i, p_i(\bar{y} - y)_i \rangle$ *and recall the definitions of* V *and* V_k *from Theorem 4.1 and* $\mathcal{H}$ *from* (4.21). *Then, it holds that*

$$\begin{aligned}\mathcal{H}(x^k, y^{k+1}; x, y) \leq V_k(x^{k-1} - x, y^k - y) - V_{k+1}(x^k - x, y^{k+1} - y) - V(z^k - z^{k-1}) \\ + \mathcal{E}^k + D_{f^*}^{P^{-1}-I}(y^k; z) - D_{f^*}^{P^{-1}-I}(y^{k+1}; z) - \langle y, v^{k+1} \rangle_{\sigma^{-1}},\end{aligned} \tag{4.26}$$

where $v^{k+1} = y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1})$ *and*

$$\begin{aligned}\mathcal{E}^k = \frac{1}{2}\left[\|y^k\|_{\sigma^{-1}}^2 - \|\hat{y}^{k+1}\|_{\sigma^{-1}}^2 - \left(\|y^k\|_{\sigma^{-1}P^{-1}}^2 - \|y^{k+1}\|_{\sigma^{-1}P^{-1}}^2\right)\right] \\ + \frac{1}{2}\|y^{k+1} - y^k\|_{\sigma^{-1}P^{-1}}^2 - \frac{1}{2}\|\hat{y}^{k+1} - y^k\|_{\sigma^{-1}}^2 + f^*(y^k) - f^*(\hat{y}^{k+1}) \\ - (f_{P^{-1}}^*(y^k) - f_{P^{-1}}^*(y^{k+1})) - \langle Ax^k, y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1}) \rangle,\end{aligned} \tag{4.27}$$

and also $\mathbb{E}_k\left[\mathcal{E}^k\right] = 0$.

With this lemma, we identify the problematic inner product for deriving the rate for expected gap: $\langle y, v^{k+1} \rangle$ (see (4.26)). This is the only term coupling the free variable z and random term v^{k+1}. In the next theorem, we use Theorem 4.7 to manipulate this inner product. For the rest, we can observe in (4.26) that the terms with V_k telescopes, $\mathcal{E}^k$ has expectation 0 and it is independent of free variable z.

Theorem 4.9. *Let Assumption 1 hold. Define the sequences* $x_{av}^K = \frac{1}{K}\sum_{k=1}^K x^k$ *and* $y_{av}^{K+1} = \frac{1}{K}\sum_{k=1}^K y^{k+1}$. *Then, for any set* $\mathcal{B} = \mathcal{B}_x \times \mathcal{B}_y \subseteq \mathcal{Z}$, *the following result holds for the expected restricted primal dual gap defined in* (4.21)

$$\mathbb{E}\left[\sup_{z \in \mathcal{B}} \mathcal{H}(x_{av}^K, y_{av}^{K+1}; x, y)\right] = \mathbb{E}\left[G_{\mathcal{B}}(x_{av}^K, y_{av}^{K+1})\right] \leq \frac{C_{\mathcal{B}}}{K}, \tag{4.28}$$

where

$$\begin{aligned}C_{\mathcal{B}} = \frac{1+2c}{2}\sup_{x \in \mathcal{B}_x}\|x^0 - x\|_{\tau^{-1}}^2 + \sup_{y \in \mathcal{B}_y}\|y^1 - y\|_{\sigma^{-1}P^{-1}}^2 + f_{P^{-1}-I}^*(y^1) - f_{P^{-1}-I}^*(y^\star) \\ + \left(\frac{1}{C_1} + 2c + c_1\right)\Delta^0 + c\|x^0\|_{\tau^{-1}}^2 + c\|y^1 - y^\star\|_{\sigma^{-1}P^{-1}}^2 + \frac{\|\sigma^{1/2}A\tau^{1/2}\|^2}{2c_1\underline{p}}\|x^\star\|_{\tau^{-1}}^2,\end{aligned}$$

where $c_1 = \|\tau^{1/2}A^\top\sigma^{1/2}P^{-1/2}\|$, $c = \|\tau^{1/2}A^\top(P^{-1} - I)\sigma^{1/2}P^{1/2}\|$, $C_1 = 1 - \gamma$.

Proof. We start from the result of Theorem 4.8. We have for the last term in (4.26)

$$-\langle y, v^{k+1} \rangle_{\sigma^{-1}} = \langle \tilde{y}^k - y, v^{k+1} \rangle_{\sigma^{-1}} - \langle \tilde{y}^k, v^{k+1} \rangle_{\sigma^{-1}}, \tag{4.29}$$

where $\tilde{y}^k$ is the random sequence defined in Theorem 4.7.

We sum (4.26) after using (4.29) and Theorem 4.1

$$\begin{aligned}\sum_{k=1}^K \mathcal{H}(x^k, y^{k+1}; x, y) \leq -V_{K+1}(x^K - x, y^{K+1} - y) + V_1(x^0 - x, y^1 - y) \\ + D_{f^*}^{P^{-1}-I}(y^1; z) - D_{f^*}^{P^{-1}-I}(y^{K+1}; z) \\ + \sum_{k=1}^K \left(\langle \tilde{y}^k - y, v^{k+1} \rangle_{\sigma^{-1}} - \langle \tilde{y}^k, v^{k+1} \rangle_{\sigma^{-1}} + \mathcal{E}^k\right),\end{aligned} \tag{4.30}$$

Next, by Young's inequality (see also (8.17))

$$-\langle A(x-x^K), P^{-1}(y^{K+1}-y^K)\rangle \le \frac{\gamma}{2}\|x-x^K\|_{\tau^{-1}}^2 + \frac{\gamma}{2}\|y^{K+1}-y^K\|_{\sigma^{-1}P^{-1}}^2. \tag{4.31}$$

On (4.30), we use (4.24) from Lem. 4.7 with $\tilde{y}^1 = y^1 = y^0$, (4.31) with the definition of $V_{K+1}(x^K - x, y^{K+1} - y)$ from Lem. 4.1 (see also (4.5), (8.17)), and $\gamma < 1$ from (3.1)

$$\begin{aligned}\sum_{k=1}^K \mathcal{H}(x^k, y^{k+1}; x, y) &\le \frac{1}{2}\|x^0 - x\|_{\tau^{-1}}^2 + \|y^1 - y\|_{\sigma^{-1}P^{-1}}^2 \\ &+ f^*_{P^{-1}-I}(y^1) - f^*_{P^{-1}-I}(y^{K+1}) + \langle Ax, (P^{-1}-I)(y^{K+1}-y^1)\rangle \\ &+ \sum_{k=1}^K \Big(\frac{1}{2}\|v^{k+1}\|_{\sigma^{-1}P}^2 - \langle \tilde{y}^k, v^{k+1}\rangle_{\sigma^{-1}} + \mathcal{E}^k\Big).\end{aligned} \tag{4.32}$$

We have $\langle Ax, (P^{-1}-I)(y^{K+1}-y^1)\rangle \le c\left(\frac{1}{2}\|x\|_{\tau^{-1}}^2 + \frac{1}{2}\|y^{K+1}-y^1\|_{\sigma^{-1}P^{-1}}^2\right)$, where $c = \|\tau^{1/2}A^\top(P^{-1}-I)\sigma^{1/2}P^{1/2}\|$ and $\frac{1}{2}\|x\|_{\tau^{-1}}^2 \le \|x - x^0\|_{\tau^{-1}}^2 + \|x^0\|_{\tau^{-1}}^2$.

We use these inequalities, arrange (4.32), and divide both sides by K

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^K \mathcal{H}(x^k, y^{k+1}; x, y) \le \frac{1}{K}\Bigg\{&\frac{1+2c}{2}\|x^0 - x\|_{\tau^{-1}}^2 + \|y^1 - y\|_{\sigma^{-1}P^{-1}}^2 + c\|x^0\|_{\tau^{-1}}^2 \\ &+ \frac{c}{2}\|y^{K+1} - y^1\|_{\sigma^{-1}P^{-1}}^2 + f^*_{P^{-1}-I}(y^1) - f^*_{P^{-1}-I}(y^{K+1}) \\ &+ \sum_{k=1}^K \left(\frac{1}{2}\|v^{k+1}\|_{\sigma^{-1}P}^2 - \langle \tilde{y}^k, v^{k+1}\rangle_{\sigma^{-1}} + \mathcal{E}^k\right)\Bigg\}.\end{aligned} \tag{4.33}$$

We now take supremum of (4.33) with respect to z, note that only the first two terms on the right hand side depend on $z = (x, y)$, and x^0, y^1 are deterministic. Then we take expectation of both sides of (4.33)

$$\begin{aligned}\mathbb{E}\left[\sup_{z\in\mathcal{B}}\frac{1}{K}\sum_{k=1}^K \mathcal{H}(x^k, y^{k+1}; x, y)\right] \le \frac{1}{K}\Bigg\{&\sup_{z\in\mathcal{B}}\left\{\frac{1+2c}{2}\|x^0 - x\|_{\tau^{-1}}^2 + \|y^1 - y\|_{\sigma^{-1}P^{-1}}^2\right\} \\ &+ \mathbb{E}\left[\frac{c}{2}\|y^{K+1} - y^1\|_{\sigma^{-1}P^{-1}}^2 + f^*_{P^{-1}-I}(y^1) - f^*_{P^{-1}-I}(y^{K+1})\right] + c\|x^0\|_{\tau^{-1}}^2 \\ &+ \sum_{k=1}^K \frac{1}{2}\mathbb{E}\left[\|v^{k+1}\|_{\sigma^{-1}P}^2\right] - \sum_{k=1}^K \mathbb{E}\left[\langle \tilde{y}^k, v^{k+1}\rangle_{\sigma^{-1}}\right] + \sum_{k=1}^K \mathbb{E}\left[\mathcal{E}^k\right]\Bigg\}.\end{aligned} \tag{4.34}$$

As $\tilde{y}^k$ is $\mathcal{F}_k$-measurable, $\mathbb{E}_k\left[v^{k+1}\right] = 0$, by Theorem 4.7, and by the tower property,

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^K \langle \tilde{y}^k, v^{k+1}\rangle_{\sigma^{-1}}\right] &= \sum_{k=1}^K \mathbb{E}\left[\mathbb{E}_k\left[\langle \tilde{y}^k, v^{k+1}\rangle_{\sigma^{-1}}\right]\right] \\ &= \sum_{k=1}^K \mathbb{E}\left[\langle \tilde{y}^k, \mathbb{E}_k[v^{k+1}]\rangle_{\sigma^{-1}}\right] = 0.\end{aligned} \tag{4.35}$$

On (4.34), we use (4.25) from Theorem 4.7, (4.35) and $\sum_{k=1}^K \mathbb{E}[\mathcal{E}^k] = \sum_{k=1}^K \mathbb{E}\left[\mathbb{E}_k\left[\mathcal{E}^k\right]\right] = 0$, which follows from Theorem 4.8 along with the tower property, to obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{z\in\mathcal{B}}\frac{1}{K}\sum_{k=1}^K \mathcal{H}(x^k, y^{k+1}; x, y)\right] \le &\sup_{z\in\mathcal{B}}\left\{\frac{1+2c}{2K}\|x^0 - x\|_{\tau^{-1}}^2 + \frac{1}{K}\|y^1 - y\|_{\sigma^{-1}P^{-1}}^2\right\} \\ &+ \frac{c}{2K}\mathbb{E}\left[\|y^{K+1} - y^1\|_{\sigma^{-1}P^{-1}}^2\right] + \frac{c}{K}\|x^0\|_{\tau^{-1}}^2 \\ &+ \frac{1}{K}\mathbb{E}\left[f^*_{P^{-1}-I}(y^1) - f^*_{P^{-1}-I}(y^{K+1})\right] + \frac{1}{C_1 K}\Delta^0.\end{aligned} \tag{4.36}$$

By Theorem 4.3 and Theorem 4.1, $\mathbb{E}\left[\|y^{K+1}-y^\star\|^2_{\sigma^{-1}P^{-1}}\right] \le 2\Delta^0$, and by Jensen's inequality, $\mathbb{E}\left[\|y^{K+1}-y^\star\|_{\sigma^{-1}P^{-1}}\right] \le \sqrt{2\Delta^0}$. With these estimations we have

$$\mathbb{E}\left[\|y^{K+1}-y^1\|^2_{\sigma^{-1}P^{-1}}\right] \le 2\|y^1-y^\star\|^2_{\sigma^{-1}P^{-1}} + 4\Delta^0. \tag{4.37}$$

As f_i is proper, l.s.c., convex, and $A_i x^\star \in \partial f_i^*(y_i^\star)$, we additionally note that

$$\begin{aligned} f_i^*(y_i^{K+1}) &\ge f_i^*(y_i^\star) + \langle A_i x^\star, y_i^{K+1}-y_i^\star\rangle \\ &\ge f_i^*(y_i^\star) - \|A_i x^\star\|_{\sigma_i}\|y_i^{K+1}-y_i^\star\|_{\sigma_i^{-1}}, \end{aligned}$$

and by substitution, Young's inequality, and defining $c_1 = \|\tau^{1/2}A^\top\sigma^{1/2}P^{-1/2}\|$

$$\begin{aligned} \mathbb{E}\big[f^*_{P^{-1}-I}(y^{K+1})\big] &= \sum_{i=1}^n \left(\frac{1}{p_i}-1\right)\mathbb{E}\left[f_i^*(y_i^{K+1})\right] \\ &\ge \sum_{i=1}^n \left(\frac{1}{p_i}-1\right)\left(f_i^*(y_i^\star) - \frac{1}{2c_1}\|A_i x^\star\|^2_{\sigma_i} - \frac{c_1}{2}\mathbb{E}\left[\|y_i^{K+1}-y_i^\star\|^2_{\sigma_i^{-1}}\right]\right) \\ &\ge \sum_{i=1}^n \left(\frac{1}{p_i}-1\right) f_i^*(y_i^\star) - \frac{1}{2c_1\underline{p}}\|Ax^\star\|^2_\sigma - \frac{c_1}{2}\mathbb{E}\|y^{K+1}-y^\star\|^2_{\sigma^{-1}P^{-1}}. \end{aligned} \tag{4.38}$$

We now use (4.37) and (4.38) in (4.36), use $\mathbb{E}\left[\|y^{K+1}-y^\star\|^2_{\sigma^{-1}P^{-1}}\right] \le 2\Delta^0$ and use definition of c_1 to obtain

$$\begin{aligned} \mathbb{E}\left[\sup_{z\in\mathcal{B}}\frac{1}{K}\sum_{k=1}^K \mathcal{H}(x^k, y^{k+1}; x, y)\right] &\le \frac{1+2c}{2K}\sup_{x\in\mathcal{B}_x}\|x^0-x\|^2_{\tau^{-1}} + \frac{1}{K}\sup_{y\in\mathcal{B}_y}\|y^1-y\|^2_{\sigma^{-1}P^{-1}} \\ &+ \frac{c}{K}\|y^1-y^\star\|^2_{\sigma^{-1}P^{-1}} + \frac{2c}{K}\Delta^0 + \frac{c}{K}\|x^0\|^2_{\tau^{-1}} + \frac{1}{K}\left(f^*_{P^{-1}-I}(y^1) - f^*_{P^{-1}-I}(y^\star)\right) \\ &+ \frac{\|\sigma^{1/2}A\tau^{1/2}\|^2}{2c_1\underline{p}K}\|x^\star\|^2_{\tau^{-1}} + \frac{c_1}{K}\Delta^0 + \frac{1}{C_1K}\Delta^0 =: \frac{C_{\mathcal{B}}}{K}. \end{aligned}$$

We define as $C_{\mathcal{B}}$ the constant of right hand side and use Jensen's inequality on the left hand side with definitions of x^K_{av} and y^{K+1}_{av} to get the result. □

Remark 4.10. In Thm. 4.9, when $p_i = \frac{1}{n}$, setting scalar step sizes $\tau = \frac{1}{n\max_i\|A_i\|}$, $\sigma = \frac{1}{\max_i\|A_i\|}$ in view of (3.1) gives $\mathcal{O}(n[\|A\| + f^*(y^1) - f^*(y^\star)]\cdot\max_{z\in\mathcal{B}}[\|x\|^2+\|y\|^2])$ as the worst case order for $C_{\mathcal{B}}$.

4.3.2 Convergence of objective values

The guarantee for the expected global primal-dual gap (see (4.28)) requires bounded primal and dual domains.

In this section, we show that $\mathcal{O}(1/k)$ rate of convergence in terms of objective values and/or feasibility can be shown with possibly unbounded primal and dual domains. The case $f(\cdot) = \delta_b(\cdot)$ is studied in [34] and a similar result was derived. The rate in [34] has a different nature in the sense that it is an almost sure rate where the constant depends on trajectory, whereas our rate is in expectation. We use the smoothed gap function introduced in [47], which, for (1.1), is defined as

$$\mathcal{G}_{\alpha,\beta}(x,y;\dot{x},\dot{y}) = \sup_{u,v} g(x) + \langle Ax, v\rangle - f^*(v) - g(u) - \langle Au, y\rangle + f^*(y) - \frac{\alpha}{2}\|u-\dot{x}\|^2 - \frac{\beta}{2}\|v-\dot{y}\|^2. \tag{4.39}$$

Theorem 4.11. *Let Assumption 1 hold. We recall* $x^K_{av} = \frac{1}{K}\sum_{k=1}^K x^k$.
- *If f is $L(f)$-Lipschitz continuous and $y^1 \in \operatorname{dom} f^*$,*

$$\mathbb{E}\left[f(Ax^K_{av}) + g(x^K_{av}) - f(Ax^\star) - g(x^\star)\right] \le \frac{C_{e,1}}{K}.$$

- *If* $f(\cdot) = \delta_{\{b\}}(\cdot)$ *with* $b \in \mathcal{Y}$,

$$\mathbb{E}\left[g(x_{av}^K) - g(x^\star)\right] \leq \frac{C_{e,2}}{K}, \quad \mathbb{E}\left[\|Ax_{av}^K - b\|_{\mathrm{diag}(\sigma)P}\right] \leq \frac{C_{e,3}}{K},$$

where (see Theorem 4.9)

$$C_e = f^*_{P^{-1}-I}(y^1) - f^*_{P^{-1}-I}(y^\star) + \left(\frac{1}{C_1} + 2c + c_1\right)\Delta^0 + c\|x^0\|^2_{\tau^{-1}} + c\|y^1 - y^\star\|^2_{\sigma^{-1}P^{-1}} + \frac{\|\sigma^{1/2}A\tau^{1/2}\|^2}{2c_1\underline{p}}\|x^\star\|^2_{\tau^{-1}},$$

$C_{e,1} = C_e + \frac{2}{\underline{p}}L(f)^2 + \frac{1+2\gamma}{2}\|x^0 - x^\star\|^2_{\tau^{-1}}$,
$C_{e,3} = \frac{1}{2}\left\{\|y^\star - y^1\|_{\sigma^{-1}P^{-1}} + \left(\|y^\star - y^1\|^2_{\sigma^{-1}P^{-1}} + 4C_e + 6\|x^\star - x^0\|_{\tau^{-1}}\right)^{1/2}\right\}$,
$C_{e,2} = C_e + \frac{1}{2}\|y^\star - y^1\|^2_{\sigma^{-1}P^{-1}} + \frac{1+2\gamma}{2}\|x^0 - x^\star\|^2_{\tau^{-1}} + \|y^\star\|_{\sigma^{-1}P^{-1}}C_{e,3}$.

Proof. For the smoothed gap (see (4.39)), from Theorem 4.9, we have

$$\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_{av}^K, y_{av}^{K+1}; x^0, y^1)\right] \leq \frac{C_e}{K}.$$

To see this, we proceed the same as in the proof of Theorem 4.9 until (4.33). Then, we move the terms $\frac{1+2\gamma}{2K}\|x^0 - x\|^2_{\tau^{-1}}$ and $\frac{1}{K}\|y^1 - y\|^2_{\sigma^{-1}P^{-1}}$ to the left hand side, take supremum, use the definition of smoothed gap, then take expectations of both sides and use the same estimations as in the first part to conclude.

- When f is Lipschitz continuous in the norm $\|\cdot\|_\sigma$, we will argue as in [19, Theorem 11]. On (4.39), with the parameters used in this theorem, we make the following observations. By [4, Corollary 17.19], when f is $L(f)$-Lipschitz continuous in the norm $\|\cdot\|_{\mathrm{diag}(\sigma)}$, it follows that $\|y^1 - y\|^2_{\sigma^{-1}} \leq 4L(f)^2$. By Lipschitzness and the definition of conjugate function, we can pick $y \in \partial f(Ax_{av}^K) \neq \emptyset$ such that $\langle Ax_{av}^K, y\rangle - f^*(y) = f(Ax_{av}^K)$. Next by Fenchel-Young inequality, $f^*(y_{av}^{K+1}) - \langle A^\top y_{av}^{K+1}, x^\star\rangle \geq -f(Ax^\star)$. We also use $\underline{p} = \min_i p_i$ to obtain (see (4.39))

$$\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{K},\frac{1}{K}}(x_{av}^K, y_{av}^{K+1}; x^0, y^1)\right] \geq \mathbb{E}\left[f(Ax_{av}^K) + g(x_{av}^K) - f(Ax^\star) - g(x^\star)\right] - \frac{2}{\underline{p}K}L(f)^2 - \frac{1+2\gamma}{2K}\|x^0 - x^\star\|^2_{\tau^{-1}},$$

where the result directly follows.

- When $f(\cdot) = \delta_b(\cdot)$, we use [47, Lemma 1], to obtain the bounds

$$\mathbb{E}\left[g(x_{av}^K) - g(x^\star)\right] \leq \mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_{av}^K, y_{av}^{K+1}; x^0, y^1)\right] + \frac{1+2\gamma}{2K}\|x^0 - x^\star\|^2_{\tau^{-1}} - \mathbb{E}\left[\langle y^\star, Ax_{av}^K - b\rangle\right] + \frac{1}{2K}\|y^\star - y^1\|^2_{\sigma^{-1}P^{-1}},$$

$$\mathbb{E}\left[\|Ax_{av}^K - b\|_{\mathrm{diag}(\sigma)P}\right] \leq \frac{1}{2K}\left\{\|y^\star - y^1\|_{\sigma^{-1}P^{-1}} + \left(\|y^\star - y^1\|^2_{\sigma^{-1}P^{-1}} + 4K\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_{av}^K, y_{av}^{K+1}; x^0, y^1)\right] + 2(1+2\gamma)\|x^0 - x^\star\|^2_{\tau^{-1}}\right)^{1/2}\right\}.$$

We use Cauchy-Schwarz inequality and the bound of $\mathbb{E}\left[\|Ax_{av}^K - b\|_{\mathrm{diag}(\sigma)P}\right]$ on $\langle y^\star, Ax_{av}^K - b\rangle$ to conclude. □

5 Related works

We summarize the comparison of the most related primal-dual coordinate descent methods (PDCD) in Table 1 at Page 26.

Primal camp. Stochastic gradient based methods (SGD) can be applied to solve (1.1) [36, 42]. SGD cannot get linear convergence except special cases [35]. Variance reduction based methods obtain linear convergence when the functions f_i are smooth and g is strongly convex; or f_i are smooth and strongly convex [2, 25, 48]. Smoothness of f_i is equivalent to strong convexity of f_i^*. Therefore, the linear convergence results of these methods require the similar assumptions as [6]. Moreover, as in [6], variance reduction based methods require knowing the constants μ_i and μ_g to set the algorithmic parameters accordingly, for obtaining linear convergence.

When $f_i(\cdot) = \delta_{\{b_i\}}(\cdot)$, SGD-type methods are proposed in [18, 39, 49]. However, these methods only obtain $\mathcal{O}(1/k)$ rate with strong convexity of g, since they focus on the general problem where the objective can be given in expectation form. Even though this rate is optimal for the given template, it is suboptimal for (1.1).

Primal-dual camp. This line of research uses coordinate descent type schemes for solving (1.1). Coordinate descent with random sampling for unconstrained optimization is proposed in [37] and later generalized and improved in [20, 41]. These methods apply coordinate descent in the primal and obtain linear convergence rates with smooth and strongly convex f_i; or smooth f_i and strongly convex g.

Another approach is to apply coordinate ascent in the dual to exploit separability of the dual in (1.1). Stochastic dual coordinate ascent (SDCA) and its accelerated variant are proposed in [44, 45]. These methods require smoothness of f_i and strong convexity of g for linear convergence and the strong convexity constants are used in the algorithms for setting the parameters.

The algorithm we analyzed in this paper is SPDHG, proposed in [6]. The authors proved linear convergence of the modified method SPDHG-μ [6, Theorem 6.1] by assuming strong convexity of f_i^*, g and special step sizes depending on strong convexity constants. Iterate convergence and ergodic $\mathcal{O}(1/k)$ rate results in [6, Theorem 4.3] are given in terms of Bregman distances which is not a valid optimality measure in general. Our analysis for SPDHG shows linear convergence with standard step sizes in (3.1) and with weaker metric subregularity assumption (see Section 2.3). Moreover, in the general convex case, we prove almost sure convergence of the iterates to a solution, which is stronger than Bregman distance based almost sure convergence in [6]. Finally, we prove $\mathcal{O}(1/k)$ rate for the ergodic sequence, with possibly unbounded domains, for optimality measures stronger than Bregman distances, such as expected primal-dual gap. The comparison of the results is also summarized in Table 2.

PDCD schemes similar to SPDHG are proposed in [13, 19, 52]. These variants assume strong convexity of f_i^*, g for linear rate of convergence. Only [19] proved linear convergence with step sizes independent of strong convexity constants, to provide a partial answer for adaptivity of PDCD methods to strong convexity. However, as detailed in Table 1, with dense A matrix and uniform sampling, this method requires step sizes n times smaller than (3.1) which can be problematic in practice (see Section 6.1). For sublinear convergence, [19] proved $\mathcal{O}(1/\sqrt{k})$ rate on a randomly selected iterate, under similar assumption to ours whereas [52] requires boundedness of the dual domain, setting a horizon and proves primal-only rates.

PDCD algorithms are also studied in [11, 12, 40]. As mentioned in [6, 19], operator theory-based proofs of these methods require using step sizes depending on global constants about the problem, causing slow performance in practice. PDCD methods for linearly constrained problems are studied in [1, 13, 34], with sublinear rates.

Latafat et al. [29] proposed TriPD-BC and proved linear convergence for this method under metric subregularity. There are two drawbacks of TriPD-BC for our setting. First, when A is not of special structure, such as block diagonal, one needs to use duplication for an efficient implementation (see [19]). Second issue is that as in [19], this method needs to use n times smaller step sizes with dense A. For the details of duplication and small step sizes, we refer to [19]. The need to use small step sizes seriously affects the practical performance of the algorithm (see Section 6.1).

Some standard references for deterministic primal-dual algorithms are in [7, 8, 17, 22, 46, 47]. As observed in [6], coordinate descent-based variants significantly increase the practical performance of these deterministic methods.

Our results imply global linear convergence for PDHG when $n = 1$, answering the question posed in [7]: *"It would be interesting to understand whether the steps can be estimated in Algorithm 1 without the a priori knowledge of μ_i, μ_g."* In the third part of Assumption 2, compact domains are not needed for this case. We highlight that such behaviour of deterministic primal-dual methods is investigated before in [29, 31].

Linear programming. A related notion to metric subregularity for linear programming is Hoffman's lemma due to classical result in [23], which is used to show linear convergence of ADMM-type methods for LPs [33, 50, 51]. The drawback of these approaches is that the knowledge of the constant η is required to run the algorithm, which is difficult to estimate. Our analysis recovers these results specific to LPs with a simpler algorithm that does not need the knowledge of η.

6 Numerical evidence

In this section, we support our theoretical findings by showing that SPDHG with step sizes in (3.1) obtains linear convergence for problems satisfying metric subregularity.

The problems we solve in this section satisfy metric subregularity (see Section 2.3). However, among these problems, only ridge regression is strongly convex-strongly concave, thus this is the only problem where existing linear convergence results from [6] apply by using the algorithm SPDHG-μ [6, Theorem 6.1]. We show that even in this case, when strong convexity constants are small, applying SPDHG can be more beneficial for some datasets. SPDHG-μ is not applicable for other problems due to lack of strong convexity or strong concavity. We also illustrate favorable behavior of SPDHG against state-of-the-art methods SVRG [25], accelerated SVRG [53] and PDCD algorithms using smaller step sizes with dense data, such as [19].

Due to limited space, we include results with one or two datasets for each problem. For SPDHG, as suggested in [6], we use uniform sampling of coordinates and the step sizes $\tau = \frac{0.99}{n \max_i \|A_i\|}$ and $\sigma_i = \frac{0.99}{\|A_i\|}$ for all problems. For the other methods, we use the suggested theoretical step sizes in the respective papers and we do not fine tune any of the methods.

6.1 Sparse recovery with basis pursuit

Basis pursuit is a fundamental problem in signal processing [10] with applications in machine learning [3, 21]:

$$\min_{x \in \mathbb{R}^d} \|x\|_1 : Ax = b. \tag{6.1}$$

Since basis pursuit is PLQ, metric subregularity holds. In this section, we aim to illustrate the importance of step sizes, as mentioned in Section 5 and Table 1 and to verify linear convergence of SPDHG. We compare SPDHG with coordinate descent version of Vu-Condat algorithm from [19], which we refer to as FB-VC-CD. Since [29] requires duplication for an efficient implementation for this problem, it uses the same step sizes as [19]. Thus, we only compare with FB-VC-CD and note that the practical performance of [29] is expected to be similar to FB-VC-CD with same step sizes.

We generate the data matrix A with $n = 500$ and $d = 1000$ and entries follow a standard normal distribution. We generate a covariance matrix $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$ and a sparse solution $x^\star$ with 100 nonzero entries. We then compute $b = Ax^\star$.

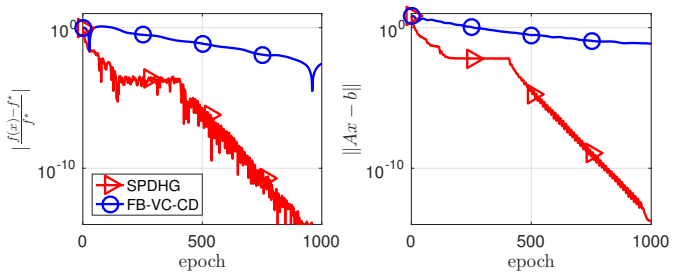

Figure 1: Linear convergence of SPDHG for basis pursuit problem.

The analysis of SPDHG by [6] shows $\mathcal{O}(1/k)$ rate on the Bregman distance to solution on the ergodic sequence whereas our analysis proves linear convergence on the last iterate. On the other hand, FB-VC-CD is proven to have $\mathcal{O}\left(1/\sqrt{k}\right)$ rate for this problem [19]. FB-VC-CD is tailored specially to exploit sparsity in the data. However, the data is dense in this problem, which causes FB-VC-CD to use n times smaller step

sizes as shown in Figure 6.1. Because of this reason, FB-VC-CD exhibits a slow rate whereas SPDHG gets fast rate as predicted by our theoretical results.

6.2 Lasso and ridge regression

In this section we solve ridge regression and Lasso problems, formulated as

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|^2 + \frac{\lambda}{2}\|x\|^2, \text{ and, } \min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1, \tag{6.2}$$

respectively. In terms of structure, (6.2) is smooth and strongly convex, or equivalently, its Lagrangian is strongly convex-strongly concave. For this problem class, [6] showed linear convergence for the method SPDHG-μ, which is a modified version of SPDHG using strong convexity and strong concavity constants for step sizes. In addition, SVRG and accelerated SVRG have linear convergence for this problem [2, 48, 53].

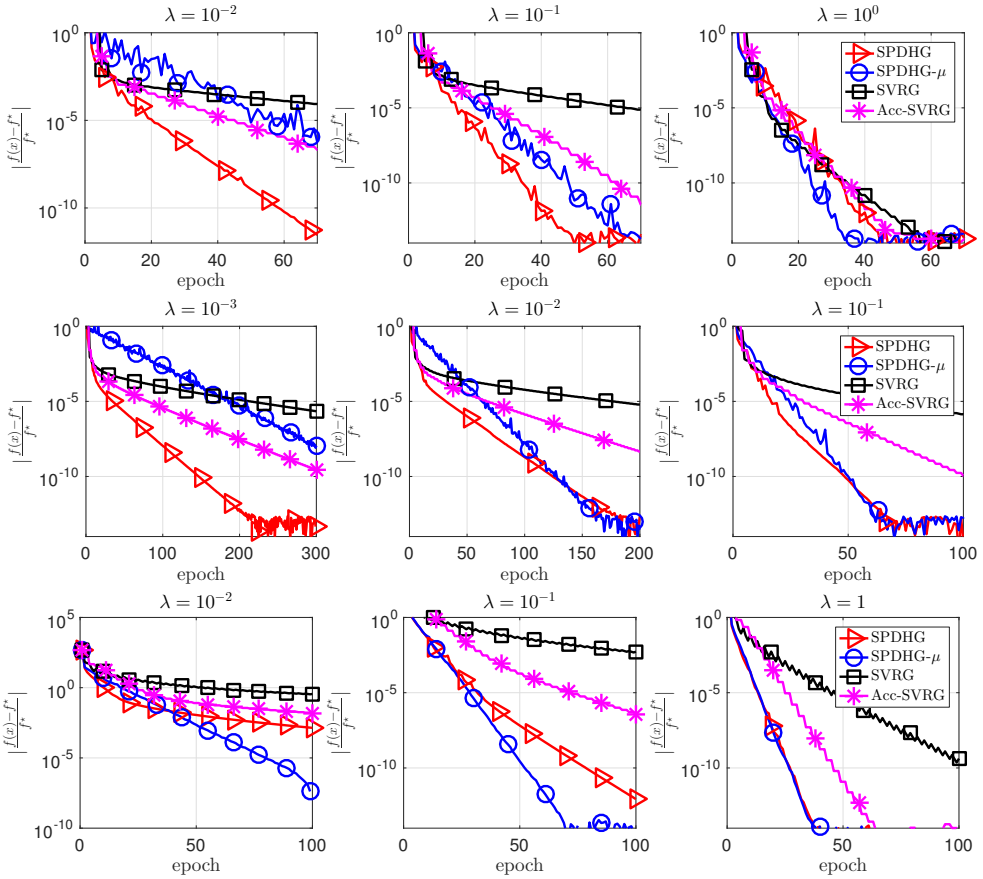

Figure 2: Ridge regression, first row: w8a, $n = 49,749, d = 300$; second row: sector, $n = 6,412$, $d = 55,197$; third row: YearPredictionMSD, $n = 463,715, d = 90$.

We use regression datasets from libsvm [9], perform row normalization, and use three different regularization parameters for each case. We compile the results in Section 6.2 along with information on datasets and regularization parameters.

The aim in this experiment is not to argue that SPDHG gets the best performance in all cases since this is a very specific instance where most algorithms can get linear convergence. Our goal is rather to show that even though our linear convergence results apply to a broad class of problems and SPDHG can apply to more general problems, it can still be competitive when compared to methods which are designed to exploit the structure of this specific setting.

When $n \geq d$, in Section 6.2, we see that for large regularization parameters, or equivalently, large strong convexity constants, SPDHG-μ is faster than SPDHG. This is expected since SPDHG-μ is designed to use strong convexity as good as possible, whereas our result holds generically without any modifications on the algorithm. Next, when strong convexity constant is small, SPDHG gets a faster linear rate than SPDHG-μ, which suggests robustness of SPDHG over SPDHG-μ in this regime. SPDHG also shows a more favorable performance than SVRG and accelerated SVRG.

When $n \leq d$, in Section 6.2, we see that SPDHG-μ shows faster convergence with small μ. This seems intuitive, since in this case the strong convexity *purely* comes from the regularization term. In this case, SPDHG-μ directly exploits this knowledge and shows a better performance.

We then solve Lasso (6.2), for which SPDHG-μ does not apply and accelerated SVRG cannot get linear rates in general. We compare with SVRG for varying regularization parameters, datasets with $n \leq d$ and $n \geq d$, and compile the results in Section 6.2. We observe that SPDHG converges linearly for this problem and exhibits a better practical performance than SVRG.

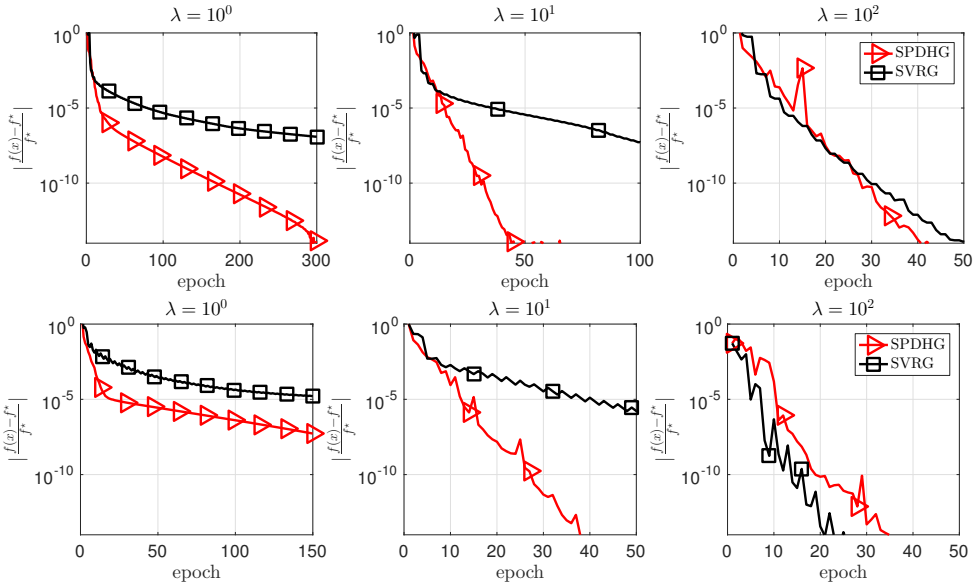

Figure 3: Lasso, top: mnist scale, $n = 60,000, d = 780$; bottom: rcv1.binary, $n = 20,242$, $d = 47,236$

7 Conclusions and open questions

In this section, we focus on the theory-practice gap mentioned in Section 4.2, before Theorem 4.6. In particular, the main aim of Section 4.2 was to show that SPDHG obtains linear rate of convergence under general assumptions that hold for a large body of problems, with an agnostic step size selection. A natural question is: How does this rate translate to practice? For this purpose, we perform a controlled experiment on a simple problem

$$\min_{x \in \mathbb{R}^d} \frac{\mu}{2}\|x\|^2 : Ax = b,$$

with $d = n = 10$. After writing the KKT conditions, we obtain $F = \begin{bmatrix} \mu I & A^\top \\ A & 0 \end{bmatrix}$ and metric subregularity constant η is the smallest eigenvalue of F in absolute value.

For simplicity, we run PDHG, which is a specific case of SPDHG, and plot the predicted rate and the empirical rate in Section 7. The resulting empirical rate is significantly faster than the worst case rate predicted by theory. We point out several possible explanations for this:

- Metric subregularity is too general to capture structures observed in practice.

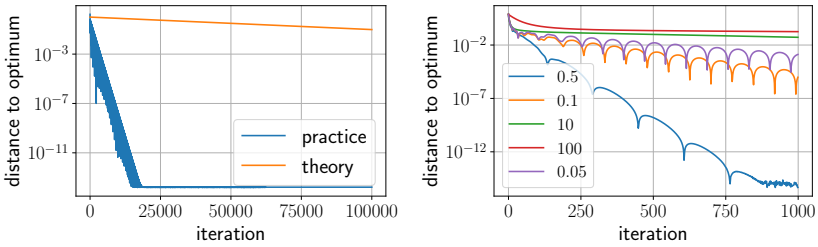

Figure 4: left: empirical and theoretical linear rates, right: empirical rates with different μ.

- Our step size choice is independent of metric subregularity constant, preventing optimizing the theoretical rate with respect to these quantities.

In fact, this phenomenon is not specific to our analysis and seems to be a common drawback of the existing analyses utilizing metric subregularity [29]. On this front, we observe that in our example, as μ increases, metric subregularity constant η degrades. However, as we see in the plot, the practical performance degrades when μ is either too big or too small (see Section 7). This observation suggests that there might exist better regularity measures beyond metric subregularity that would help us derive better rates. We believe that this is a promising future direction.

Acknowledgments

Part of the work was done while A. Alacaoglu was at EPFL. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data). This work was supported by the Swiss National Science Foundation (SNSF) under grant number 407540_167319. This project received funding from NSF Award 2023239; DOE ASCR under Subcontract 8F-30039 from Argonne National Laboratory.

We are grateful to Panayotis Mertikopoulos, Ya-Ping Hsieh, and Yura Malitsky for discussions.

8 Appendix

8.1 Proof of Theorem 4.3

Proof. On (4.3), we pick $(x, y) = (x^\star, y^\star)$ and by convexity, $D_g(x^k; z^\star) \geq 0$, $D_{f^*}(\hat{y}^{k+1}; z^\star) \geq 0$. Next, by using $\Delta^k = V_{k+1}(x^k - x^\star, y^{k+1} - y^\star)$, we write (4.3)

$$\mathbb{E}_k\left[\Delta^k\right] \leq \Delta^{k-1} - V(z^k - z^{k-1}). \tag{8.1}$$

We denote $\boldsymbol{q}^k = (1 \otimes x^k, 1 \otimes y^k)$. By taking total expectation, summing (8.1), and using Theorem 4.2, we have $\sum_{k=1}^{\infty} \mathbb{E}\left[\|T(\boldsymbol{q}^{k-1}) - \boldsymbol{q}^{k-1}\|^2_{\bar{S}\bar{P}}\right] < +\infty$. We use Fubini-Tonelli theorem to exchange the infinite sum and the expectation to obtain $\mathbb{E}\left[\sum_{k=0}^{\infty} \|T(\boldsymbol{q}^{k-1}) - \boldsymbol{q}^{k-1}\|^2_{\bar{S}\bar{P}}\right] < \infty$. Here, since $\sum_{k=0}^{\infty} \|T(\boldsymbol{q}^{k-1}) - \boldsymbol{q}^{k-1}\|^2_{\bar{S}\bar{P}}$ is nonnegative, we conclude that $\sum_{k=0}^{\infty} \|T(\boldsymbol{q}^{k-1}) - \boldsymbol{q}^{k-1}\|^2_{\bar{S}\bar{P}}$ is finite almost everywhere, which implies that $\|T(\boldsymbol{q}^{k-1}) - \boldsymbol{q}^{k-1}\|^2_{\bar{S}\bar{P}}$ converges to 0 almost surely. Thus we established: $\exists \Omega_T$ with $\mathbb{P}(\Omega_T) = 1$ such that $\forall \omega \in \Omega_T$, we have $T(\boldsymbol{q}^k(\omega)) - \boldsymbol{q}^k(\omega) \to 0$.

We apply Robbins-Siegmund lemma [43, Theorem 1] on (8.1) to get that a.s., Δ^k converges to a finite valued random variable and $V(z^k - z^{k-1}) \to 0$. Consequently, by (4.4), $\|y^k - y^{k-1}\|$ converges to 0 a.s. Since a.s., Δ^k converges and $\|y^k - y^{k-1}\|$ converges to 0, we have that $\|z^k - z^\star\|$ converges a.s.

In particular, we have shown that

$$\mathbb{P}\big(\omega \in \Omega\colon \lim_{k\to\infty} \|z_k(\omega) - z^\star\| \text{ exists.}\big) = 1. \tag{8.2}$$

The probability 1 set from which we select the trajectories is defined via $z^\star$. Let us denote the set

$$\Omega_{z^\star} = \Big\{\omega \in \Omega\colon \lim_{k\to\infty} \|z_k(\omega) - z^\star\| \text{ exists.}\Big\} \tag{8.3}$$

Thus our statement is actually: for each $z^\star \in \mathcal{Z}^\star$, there exists a set $\Omega_{z^\star}$ with probability 1, such that $\forall \omega \in \Omega_{z^\star}$, $\lim_{k\to\infty} \|z_k(\omega) - z^\star\|$ exists.

We now follow the arguments in [11, Proposition 2.3], [5, Proposition 9], [24, Theorem 2], [19, Theorem 1] to strengthen this result.

Let us pick a set $\mathcal{C}$ which is a countable subset of $\mathrm{ri}(\mathcal{Z}^\star)$ that is dense in $\mathcal{Z}^\star$. Let us denote the elements of $\mathcal{C}$ as v_i for $i \in \mathbb{N}$.

We just proved that for all $v_i \in \mathcal{Z}^\star$, $\exists \Omega_{v_i}$ with $\mathbb{P}(\Omega_{v_i}) = 1$, such that $\forall \omega \in \Omega_{v_i}$, $\lim_{k\to\infty} \|z_k(\omega) - v_i\|$ exists. Let us denote $\Omega_{\mathcal{C}} = \cap_{i\in\mathbb{N}} \Omega_{v_i}$. As $\Omega_{\mathcal{C}}$ is the intersection of a countable number of sets of probability 1, $\mathbb{P}(\Omega_{\mathcal{C}}) = 1$.

Next, we set $\tilde{z} \in \mathcal{Z}^\star$. As $\mathcal{C}$ is dense in $\mathrm{ri}(\mathcal{Z}^\star)$, there exists a subsequence $v_{\varphi(i)}$, where $\varphi\colon \mathbb{N} \to \mathbb{N}$ is an increasing function, such that $v_{\varphi(i)} \to \tilde{z}$.

We now pick $\omega \in \Omega_{\mathcal{C}}$ and study the existence of $\lim_{k\to\infty} \|z_k(\omega) - \tilde{z}\|$. By triangle inequality, $\forall i \in \mathbb{N}$,

$$\|z_k(\omega) - v_{\varphi(i)}\| - \|v_{\varphi(i)} - \tilde{z}\| \le \|z_k(\omega) - \tilde{z}\| \le \|z_k(\omega) - v_{\varphi(i)}\| + \|v_{\varphi(i)} - \tilde{z}\|.$$

Rearranging gives

$$-\|v_{\varphi(i)} - \tilde{z}\| \le \|z_k(\omega) - \tilde{z}\| - \|z_k(\omega) - v_{\varphi(i)}\| \le \|v_{\varphi(i)} - \tilde{z}\|.$$

As ω is chosen from $\Omega_{\mathcal{C}}$, and any element of $\Omega_{\mathcal{C}}$ is also an element of Ω_{v_i}, we know that $\lim_{k\to\infty} \|z_k(\omega) - v_{\varphi(i)}\|$ exists. Moreover, recall that $v_{\varphi(i)} \to \tilde{z}$.

We take limit as $k \to \infty$,

$$\begin{aligned}
-\|v_{\varphi(i)} - \tilde{z}\| &\le \liminf_{k\to\infty} \|z_k(\omega) - \tilde{z}\| - \lim_{k\to\infty} \|z_k(\omega) - v_{\varphi(i)}\| \\
&\le \limsup_{k\to\infty} \|z_k(\omega) - \tilde{z}\| - \lim_{k\to\infty} \|z_k(\omega) - v_{\varphi(i)}\| \\
&\le \|v_{\varphi(i)} - \tilde{z}\|.
\end{aligned}$$

As we take the limit along the subsequence defined by $\varphi(i)$, we have $\lim_{i\to\infty} \|v_{\varphi(i)} - \tilde{z}\| = 0$, which gives the equality of lim inf and lim sup.

Thus, $\forall \omega \in \Omega_{\mathcal{C}}$ with $\mathbb{P}(\Omega_{\mathcal{C}}) = 1$ and $\forall \tilde{z} \in \mathcal{Z}^\star$, we have that $\lim_{k\to\infty} \|z_k(\omega) - \tilde{z}\|$ exists.

We now pick $\omega \in \Omega_{\mathcal{C}} \cap \Omega_T$ and then as we have that $(z^k(\omega))_k$ is bounded, we denote by $\tilde{z} = (\tilde{x}, \tilde{y})$ one of its cluster points. Then, we denote $\tilde{\boldsymbol{q}} = (1 \otimes \tilde{x}, 1 \otimes \tilde{y})$ and say that $\tilde{\boldsymbol{q}}$ is a cluster point of $(\boldsymbol{q}^k(\omega))_k$.

As $T(\boldsymbol{q}^k(\omega)) - \boldsymbol{q}^k(\omega) \to 0$, by continuity of T we have $T(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}} \to 0$, therefore $\tilde{\boldsymbol{q}}$ is a fixed point of T. We now use Theorem 4.2 to argue that fixed points of T which we denote as $(x_f(j), y_f(j))_{j=\{1,\dots,n\}}$ are such that $(x_f(j), y_f(j)) \in \mathcal{Z}^\star, \forall j \in \{1, \dots, n\}$. Since $\tilde{\boldsymbol{q}}$ is a fixed point of T, we conclude that $\tilde{z} \in \mathcal{Z}^\star$.

To sum up, we have shown that at least on some subsequence $z^k(\omega)$ converges to $\tilde{z} \in \mathcal{Z}^\star$. Then, the result follows due to existence of the limit, proven earlier. □

8.2 Proof of Theorem 4.8

Proof. As in [6], we use (4.2) to denote full dimensional updates. By the definition of the proximal operator (2.1) along with convexity of f_i^* and g, we get, $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$ and $\forall i = \{1, \dots, n\}$

$$\begin{aligned}
g(x) &\ge g(x^k) + \langle x^k - x, A^\top \bar{y}^k \rangle + \frac{1}{2}\|x^k - x^{k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|x^k - x\|^2_{\tau^{-1}} \\
&\qquad - \frac{1}{2}\|x - x^{k-1}\|^2_{\tau^{-1}}, \\
f_i^*(y_i) &\ge f^*(\hat{y}_i^{k+1}) - \langle \hat{y}_i^{k+1} - y_i, A_i x^k \rangle + \frac{1}{2}\|\hat{y}_i^{k+1} - y_i^k\|^2_{\sigma_i^{-1}} + \frac{1}{2}\|\hat{y}_i^{k+1} - y_i\|^2_{\sigma_i^{-1}} \\
&\qquad - \frac{1}{2}\|y_i - y_i^k\|^2_{\sigma_i^{-1}}.
\end{aligned}$$

We sum the second inequality from $i = 1$ to n and add to the first inequality to obtain

$$\begin{aligned}0 \geq\ & g(x^k) - g(x) + \langle x^k - x, A^\top \bar{y}^k\rangle + f^*(\hat{y}^{k+1}) - f^*(y) - \langle \hat{y}^{k+1} - y, Ax^k\rangle \\ & + \frac{1}{2}\left(-\|x^{k-1} - x\|^2_{\tau^{-1}} + \|x^k - x\|^2_{\tau^{-1}} + \|x^k - x^{k-1}\|^2_{\tau^{-1}}\right) \\ & + \frac{1}{2}\big(- \|y^k - y\|^2_{\sigma^{-1}} + \|\hat{y}^{k+1} - y\|^2_{\sigma^{-1}} + \|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}}\big). \end{aligned} \tag{8.4}$$

We next note

$$\begin{aligned}\mathcal{H}(x^k, \hat{y}^{k+1}; x, y) &= g(x^k) + \langle Ax^k, y\rangle - f^*(y) - g(x) - \langle Ax, \hat{y}^{k+1}\rangle + f^*(\hat{y}^{k+1}), \\ \Delta_1 &= \frac{1}{2}\big(- \|x^{k-1} - x\|^2_{\tau^{-1}} + \|x^k - x\|^2_{\tau^{-1}} + \|x^k - x^{k-1}\|^2_{\tau^{-1}}\big), \\ \Delta_2 &= \frac{1}{2}\big(- \|y^k - y\|^2_{\sigma^{-1}} + \|\hat{y}^{k+1} - y\|^2_{\sigma^{-1}} + \|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}}\big).\end{aligned}$$

Then, we can write (8.4) as

$$0 \geq \mathcal{H}(x^k, \hat{y}^{k+1}; x, y) + \langle A(x - x^k), \hat{y}^{k+1} - \bar{y}^k\rangle + \Delta_1 + \Delta_2. \tag{8.5}$$

We estimate by simple manipulations

$$\begin{aligned}\mathcal{H}(x^k,\hat{y}^{k+1}; x, y) &= \mathcal{H}(x^k, y^{k+1}; x, y) + \langle Ax, y^{k+1} - \hat{y}^{k+1}\rangle + f^*(\hat{y}^{k+1}) - f^*(y^{k+1}) \\ &\quad - \left(f^*_{P^{-1}-I}(y^{k+1}) - f^*_{P^{-1}-I}(y^k)\right) + \left(f^*_{P^{-1}-I}(y^{k+1}) - f^*_{P^{-1}-I}(y^k)\right) \\ &\quad + \langle Ax, (P^{-1} - I)(y^{k+1} - y^k)\rangle - \langle Ax, (P^{-1} - I)(y^{k+1} - y^k)\rangle \\ &= \mathcal{H}(x^k, y^{k+1}; x, y) + f^*(\hat{y}^{k+1}) - f^*(y^k) - (f^*_{P^{-1}}(y^{k+1}) - f^*_{P^{-1}}(y^k)) \\ &\quad + \langle Ax, y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1})\rangle \\ &\quad + \left(f^*_{P^{-1}-I}(y^{k+1}) - f^*_{P^{-1}-I}(y^k)\right) - \langle Ax, (P^{-1} - I)(y^{k+1} - y^k)\rangle \\ &= \mathcal{H}(x^k, y^{k+1}; x, y) + f^*(\hat{y}^{k+1}) - f^*(y^k) - (f^*_{P^{-1}}(y^{k+1}) - f^*_{P^{-1}}(y^k)) \\ &\quad + \langle Ax, y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1})\rangle + D^{P^{-1}-I}_{f^*}(y_{k+1}; y) - D^{P^{-1}-I}_{f^*}(y_k; y).\end{aligned} \tag{8.6}$$

By the definition of $\bar{y}^k$ in SPDHG, we have for the bilinear term in (8.5) that

$$\begin{aligned}\langle A(x - x^k), \hat{y}^{k+1} - \bar{y}^k\rangle &= \langle A(x - x^k), \hat{y}^{k+1} - y^k - P^{-1}(y^k - y^{k-1})\rangle \\ &= \langle A(x - x^k), \hat{y}^{k+1} - y^k\rangle - \langle A(x - x^{k-1}), P^{-1}(y^k - y^{k-1})\rangle \\ &\quad - \langle A(x^{k-1} - x^k), P^{-1}(y^k - y^{k-1})\rangle \\ &= \langle A(x - x^k), P^{-1}(y^{k+1} - y^k)\rangle - \langle A(x - x^{k-1}), P^{-1}(y^k - y^{k-1})\rangle \\ &\quad - \langle A(x^{k-1} - x^k), P^{-1}(y^k - y^{k-1})\rangle \\ &\quad + \langle A(x - x^k), \hat{y}^{k+1} - y^k - P^{-1}(y^{k+1} - y^k)\rangle.\end{aligned} \tag{8.7}$$

On Δ_2, we add and subtract $\|y^k - y\|^2_{\sigma^{-1}P^{-1}} - \|y^{k+1} - y\|^2_{\sigma^{-1}P^{-1}}$ to get

$$-\Delta_2 = -\frac{1}{2}\|y^{k+1} - y\|^2_{\sigma^{-1}P^{-1}} + \frac{1}{2}\|y^k - y\|^2_{\sigma^{-1}P^{-1}} - \frac{1}{2}\|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}} + \epsilon^k, \tag{8.8}$$

where

$$\begin{aligned}\epsilon^k &= \frac{1}{2}\Big[\|y^k - y\|^2_{\sigma^{-1}} - \|\hat{y}^{k+1} - y\|^2_{\sigma^{-1}} - \left(\|y^k - y\|^2_{\sigma^{-1}P^{-1}} - \|y^{k+1} - y\|^2_{\sigma^{-1}P^{-1}}\right)\Big] \\ &= \frac{1}{2}\Big[\|y^k\|^2_{\sigma^{-1}} - \|\hat{y}^{k+1}\|^2_{\sigma^{-1}} - \left(\|y^k\|^2_{\sigma^{-1}P^{-1}} - \|y^{k+1}\|^2_{\sigma^{-1}P^{-1}}\right) \\ &\quad - 2\langle y, y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1})\rangle_{\sigma^{-1}}\Big].\end{aligned} \tag{8.9}$$

We use eqs. (8.6) to (8.8) in (8.5), add and subtract $\frac{1}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}}$ and use the definition $v^{k+1} = y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1})$ from Theorem 4.7 to obtain

$$\begin{aligned}
\mathcal{H}(x^k, y^{k+1}; x, y) \leq & -\frac{1}{2}\|x^k - x\|^2_{\tau^{-1}} + \frac{1}{2}\|x^{k-1} - x\|^2_{\tau^{-1}} \\
& - \langle A(x - x^k), P^{-1}(y^{k+1} - y^k)\rangle + \langle A(x - x^{k-1}), P^{-1}(y^k - y^{k-1})\rangle \\
& - \frac{1}{2}\|x^k - x^{k-1}\|^2_{\tau^{-1}} - \frac{1}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}} \\
& - \langle A(x^k - x^{k-1}), P^{-1}(y^k - y^{k-1})\rangle - \frac{1}{2}\|y^{k+1} - y\|^2_{\sigma^{-1}P^{-1}} \\
& + \frac{1}{2}\|y^k - y\|^2_{\sigma^{-1}P^{-1}} - \frac{1}{2}\|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}} + \frac{1}{2}\|y^k - y^{k-1}\|^2_{\sigma^{-1}P^{-1}} \\
& + \frac{1}{2}\left[\|y^k\|^2_{\sigma^{-1}} - \|\hat{y}^{k+1}\|^2_{\sigma^{-1}} - \left(\|y^k\|^2_{\sigma^{-1}P^{-1}} - \|y^{k+1}\|^2_{\sigma^{-1}P^{-1}}\right)\right] \\
& + f^*(y^k) - f^*(\hat{y}^{k+1}) - (f^*_{P^{-1}}(y^k) - f^*_{P^{-1}}(y^{k+1})) - \langle y, v^{k+1}\rangle_{\sigma^{-1}} \\
& - \langle Ax^k, y^k - \hat{y}^{k+1} - P^{-1}(y^k - y^{k+1})\rangle + D^{P^{-1}-I}_{f^*}(y_k; z) - D^{P^{-1}-I}_{f^*}(y_{k+1}; z). \qquad (8.10)
\end{aligned}$$

The first result follows by the definitions of V_k and V from (4.1), and definition of $\mathcal{E}^k$ from (4.27).

On $\mathcal{E}^k$, we use $\mathbb{E}_k\left[P^{-1}(y^k - y^{k+1})\right] = y^k - \hat{y}^{k+1}$, $\mathbb{E}_k\left[f^*_{P^{-1}}(y^k) - f^*_{P^{-1}}(y^{k+1})\right] = f^*(y^k) - f^*(\hat{y}^{k+1})$ and $\mathbb{E}_k\left[\|y^{k+1} - y^k\|^2_{\sigma^{-1}P^{-1}}\right] = \|\hat{y}^{k+1} - y_k\|^2_{\sigma^{-1}}$

$$\begin{aligned}
\mathbb{E}_k\left[\mathcal{E}^k\right] = & -\frac{1}{2}\|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}} + \frac{1}{2}\mathbb{E}_k\left[\|y^{k+1} - y^k\|^2_{\sigma^{-1}P^{-1}}\right] \\
& + \frac{1}{2}\left(\|y^k\|^2_{\sigma^{-1}} - \|\hat{y}^{k+1}\|^2_{\sigma^{-1}}\right) - \frac{1}{2}\mathbb{E}_k\left[\|y^k\|^2_{\sigma^{-1}P^{-1}} - \|y^{k+1}\|^2_{\sigma^{-1}P^{-1}}\right] \\
& + f^*(y^k) - f^*(\hat{y}^{k+1}) - \mathbb{E}_k\left[f^*_{P^{-1}}(y^k) - f^*_{P^{-1}}(y^{k+1})\right] \\
& - \langle Ax^k, y^k - \hat{y}^{k+1} - \mathbb{E}_k\left[P^{-1}\left(y^k - y^{k+1}\right)\right]\rangle = 0.
\end{aligned}$$

□

8.3 Proof of Theorem 4.1

Proof. At step k of SPDHG in Algorithm 1, we select an index $i_k \in \{1, \ldots, n\}$ randomly with probability p_{i_k} and perform the following step on the dual variable

$$y^{k+1}_{i_k} = \hat{y}^{k+1}_{i_k}, \text{ and } y^{k+1}_i = y^k_i, \forall i \neq i_k. \qquad (8.11)$$

For any $Y \in \mathcal{Y}$ that is measurable with respect to $\mathcal{F}_k$, (8.11) immediately gives

$$\mathbb{E}_k[y^{k+1}] = P\hat{y}^{k+1} + (I - P)\, y^k, \qquad (8.12)$$

$$\mathbb{E}_k\left[\|y^{k+1} - Y\|^2_{\sigma^{-1}}\right] = \|\hat{y}^{k+1} - Y\|^2_{\sigma^{-1}P} + \|y^k - Y\|^2_{\sigma^{-1}(I-P)}. \qquad (8.13)$$

A simple manipulation of (8.12) and plugging in $Y = y$ and $Y = y_k$ in (8.13) gives

$$\hat{y}^{k+1} = P^{-1}\mathbb{E}_k[y^{k+1}] - (P^{-1} - I)y^k \qquad (8.14)$$

$$\|\hat{y}^{k+1} - y\|^2_{\sigma^{-1}} = \mathbb{E}_k\left[\|y^{k+1} - y\|^2_{\sigma^{-1}P^{-1}}\right] - \|y^k - y\|^2_{\sigma^{-1}(P^{-1}-I)} \qquad (8.15)$$

$$\|\hat{y}^{k+1} - y^k\|^2_{\sigma^{-1}} = \mathbb{E}_k\left[\|y^{k+1} - y^k\|^2_{\sigma^{-1}P^{-1}}\right]. \qquad (8.16)$$

The first result follows by taking expectation of the result of Theorem 4.8, after using tower property and the above estimations. On deriving the conclusion, we also use $D_g(x_k; z) + D_{f^*}(\hat{y}_{k+1}; z) = \mathcal{H}(x_k, \hat{y}_{k+1}; x, y)$ and (8.6).

It is straightforward to prove (4.4) and (4.5). Since $y_j^k = y_j^{k-1}, \forall j \neq i_{k-1}$,

$$\begin{aligned}
|\langle Ax, P^{-1}(y^k - y^{k-1})\rangle| &= |\langle A_{i_{k-1}} x, p_{i_{k-1}}^{-1}(y_{i_{k-1}}^k - y_{i_{k-1}}^{k-1})\rangle| \\
&\leq \|A_{i_{k-1}} x\| p_{i_{k-1}}^{-1} \|y_{i_{k-1}}^k - y_{i_{k-1}}^{k-1}\| \\
&= \left(\tau^{1/2} \sigma_{i_{k-1}}^{1/2} p_{i_{k-1}}^{-1/2} \|A_{i_{k-1}}\|\right) \tau^{-1/2} \|x\| p_{i_{k-1}}^{-1/2} \sigma_{i_{k-1}}^{-1/2} \|y_{i_{k-1}}^k - y_{i_{k-1}}^{k-1}\| \\
&\leq \gamma \left(\tau^{-1/2} \|x\| p_{i_{k-1}}^{-1/2} \sigma_{i_{k-1}}^{-1/2} \|y_{i_{k-1}}^k - y_{i_{k-1}}^{k-1}\|\right) \\
&\leq \frac{\gamma}{2} \left(\|x\|_{\tau^{-1}}^2 + \|y_{i_{k-1}}^k - y_{i_{k-1}}^{k-1}\|_{p_{i_{k-1}}^{-1} \sigma_{i_{k-1}}^{-1}}^2\right) \\
&= \frac{\gamma}{2} \left(\|x\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{\sigma^{-1} P^{-1}}^2\right), \qquad (8.17)
\end{aligned}$$

where the last step is due to $y_j^k = y_j^{k-1}, \forall j \neq i_{k-1}$. Plugging in (8.17) into the definitions of $V(z^k - z^{k-1})$ and $V_k(z)$ is sufficient to prove (4.4) and (4.5). □

	Linear convergence	Rates with only convexity	Step sizes for linear convergence*
[6]	f_i^* : μ_i-s.c. g : μ_g-s.c.	Ergodic $\mathcal{O}\left(\frac{1}{k}\right)$for Bregman distance to solution	$\|A_i\|, \mu_i, \mu_g$
[52]	f_i^* : μ_i-s.c. g : μ_g-s.c.	Nonergodic $\mathcal{O}\left(\frac{1}{k}\right)$with bounded dual domain and fixed horizon	$\|A_i\|, \mu_i, \mu_g$
[19]	f_i^* : μ_i-s.c. g : μ_g-s.c.	Randomly selected iterate $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$n^2\tau\sigma_i\|A_i\|^2 < 1$
[29]	F is MS (see (2.3))	×	$n^2\tau\sigma_i\|A_i\|^2 < 1$
This paper	F is MS (see (2.3))	Ergodic $\mathcal{O}\left(\frac{1}{k}\right)$ for primal-dual gap, objective values and feasibility	$n\tau\sigma_i\|A_i\|^2 < 1$

Table 1: Comparison of primal dual coordinate descent methods. s.c. denotes strongly convex, MS denotes metrically subregular. Please see Section 5 for a thorough comparison. Please see Section 2 for comparison of MS and s.c. assumptions. *Step sizes are for optimization with a potentially dense A matrix and uniform sampling: $p_i = 1/n$.

	a.s. convergence	Linear convergence	Ergodic rates
[6]	$D_h(z^k; z^\star) \to 0$, for any $z^\star$ where D_h is Bregman distance generated by $h(z) = f^*(y) + g(x)$	Assumption: f_i^*, g s.c. step sizes depending on μ_i, μ_g	$D_h(z_{av}^k; z^\star) = \mathcal{O}(1/k)$
This paper	$z^k \to z^\star$, for some $z^\star$.	Assumption: F in (2.3) is MS Step sizes: $n\tau\sigma_i\|A_i\|^2 < 1$*	• Restricted primal-dual gap $\mathbb{E}\left[G_{\mathcal{B}}(x_{av}^k, y_{av}^k)\right] = \mathcal{O}(1/k)$ • f is Lipschitz† $\mathbb{E}\left[\|P(x_{av}^k) - P(x^\star)\|\right] = \mathcal{O}(1/k)$ • $f(\cdot) = \delta_b(\cdot)$ $\mathbb{E}\left[\|g(x_{av}^k) - g(x^\star)\|\right] = \mathcal{O}(1/k)$ $\mathbb{E}\left[\|Ax_{av}^k - b\|\right] = \mathcal{O}(1/k)$

Table 2: Comparison of our results and previous results on SPDHG. *Step size condition is for uniform sampling: $p_i = 1/n$. †In this case $P(x) := f(Ax) + g(x)$.

References

[1] Ahmet Alacaoglu, Quoc Tran Dinh, Olivier Fercoq, and Volkan Cevher. Smooth primal-dual coordinate descent algorithms for nonsmooth convex optimization. In *Advances in Neural Information Processing Systems*, pages 5852–5861, 2017.

[2] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18(1):8194–8244, 2017.

[3] Sanjeev Arora, Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and LSTMs. In *International Conference on Learning Representations*, 2018.

[4] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

[5] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Math. Program.*, 129(2):163, 2011.

[6] Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM J. Optim.*, 28(4):2783–2808, 2018.

[7] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[8] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Math. Program.*, 159(1-2):253–287, 2016.

[9] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[11] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.*, 25(2):1221–1248, 2015.

[12] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping ii: mean-square and linear convergence. *Math. Program.*, 174(1-2):433–451, 2019.

[13] Cong Dang and Guanghui Lan. Randomized methods for saddle point computation. *arXiv:1409.8625*, 2014.

[14] Asen L Dontchev and R Tyrrell Rockafellar. Implicit functions and solution mappings. *Springer Monographs in Mathematics. Springer*, 208, 2009.

[15] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018.

[16] Matthias J Ehrhardt, Pawel Markiewicz, Antonin Chambolle, Peter Richtárik, Jonathan Schott, and Carola-Bibiane Schönlieb. Faster pet reconstruction with a stochastic primal-dual hybrid gradient method. In *Wavelets and Sparsity XVII*, volume 10394, page 103941O. SPIE, 2017.

[17] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imag. Sci.*, 3(4):1015–1046, 2010.

[18] Olivier Fercoq, Ahmet Alacaoglu, Ion Necoara, and Volkan Cevher. Almost surely constrained convex optimization. In *International Conference on Machine Learning*, pages 1910–1919, 2019.

[19] Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM J. Optim.*, 29(1):100–134, 2019.

[20] Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM J. Optim.*, 25(4):1997–2023, 2015.

[21] Tom Goldstein and Christoph Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Trans. Inf. Theory*, 64(4):2675–2689, 2018.

[22] Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM J. Imag. Sci.*, 5(1):119–149, 2012.

[23] Alan J Hoffman. On approximate solutions of systems of linear inequalities. *J. Res. Nat. Bur. Stand.*, 49(4), 1952.

[24] Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE conference on decision and control*, pages 3671–3676. IEEE, 2013.

[25] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[26] Daniil Kazantsev, Edoardo Pasca, Mark Basham, Martin Turner, Matthias J Ehrhardt, Kris Thielemans, Benjamin A Thomas, Evgueni Ovtchinnikov, Philip J Withers, and Alun W Ashton. Versatile regularisation toolkit for iterative image reconstruction with proximal splitting algorithms. In *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, volume 11072, page 110722D. SPIE, 2019.

[27] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Math. Program.*, 171(1):167–215, 2018.

[28] Guanghui Lan and Yi Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM J. Optim.*, 28(4):2753–2782, 2018.

[29] Puya Latafat, Nikolaos M Freris, and Panagiotis Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *arXiv:1706.02882v4*, 2019.

[30] Puya Latafat and Panagiotis Patrinos. Primal-dual proximal algorithms for structured convex optimization: A unifying framework. In *Large-Scale & Distrib. Optim.*, pages 97–120. Springer, 2018.

[31] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Math. Program.*, 159(1-2):403–434, 2016.

[32] Pan Liu and Xin Yang Lu. Real order (an)-isotropic total variation in image processing-part ii: Learning of optimal structures. *arXiv:1903.08513*, 2019.

[33] Yongchao Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Partial error bound conditions and the linear convergence rate of the alternating direction method of multipliers. *SIAM J. Numer. Anal.*, 56(4):2095–2123, 2018.

[34] D Russell Luke and Yura Malitsky. Block-coordinate primal-dual method for nonsmooth minimization over linear constraints. In *Large-Scale & Distrib. Optim.*, pages 121–147. Springer, 2018.

[35] Ion Necoara, Peter Richtárik, and Andrei Patrascu. Randomized projection methods for convex feasibility: Conditioning and convergence rates. *SIAM J. Optim.*, 29(4):2814–2852, 2019.

[36] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[37] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.

[38] Evangelos Papoutsellis, Evelina Ametova, Claire Delplancke, Gemma Fardell, Jakob S Jørgensen, Edoardo Pasca, Martin Turner, Ryan Warr, William RB Lionheart, and Philip J Withers. Core imaging library–part ii: Multichannel reconstruction for dynamic and spectral tomography. *arXiv:2102.06126*, 2021.

[39] Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *J. Mach. Learn. Res.*, 18:198–1, 2017.

[40] Jean-Christophe Pesquet and Audrey Repetti. A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis*, 16(12):2453–2490, 2015.

[41] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.*, 144(1-2):1–38, 2014.

[42] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[43] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optim. Method. in Stat.*, pages 233–257. Elsevier, 1971.

[44] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.*, 14(Feb):567–599, 2013.

[45] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International Conference on Machine Learning*, pages 64–72, 2014.

[46] Quoc Tran-Dinh, Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. An adaptive primal-dual framework for nonsmooth convex minimization. *Math. Program. Comp.*, Oct 2019.

[47] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM J. Optim.*, 28(1):96–134, 2018.

[48] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, 24(4):2057–2075, 2014.

[49] Yangyang Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM J. Optim.*, 30(2):1664–1692, 2020.

[50] Wei Hong Yang and Deren Han. Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. *SIAM J. Numer. Anal.*, 54(2):625–640, 2016.

[51] Ian En-Hsu Yen, Kai Zhong, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. Sparse linear programming via primal and dual augmented coordinate descent. In *Advances in Neural Information Processing Systems*, pages 2368–2376, 2015.

[52] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *J. Mach. Learn. Res.*, 18(1):2939–2980, 2017.

[53] Kaiwen Zhou, Fanhua Shang, and James Cheng. A simple stochastic variance reduced algorithm with fast convergence rates. In *International Conference on Machine Learning*, pages 5975–5984, 2018.