Abstract

We propose a class of optimal-rate primal-dual algorithms for minimization of the sum of three convex functions with a linear operator. We first establish the optimal convergence rates for solving the saddle-point reformulation of the problem only using first-order information under deterministic and stochastic settings, respectively. We then proceed to show that the proposed algorithm class achieves these rates. The studied algorithm class does not require matrix inversion and is simple to implement. To our knowledge, this is the first work to establish and attain the optimal rates for this class of problems with minimal assumptions. Numerical experiments show that our method outperforms state-of-the-art methods.

1 Introduction

We consider an optimization problem involving sum of three convex functions in the following form:

$$\min_{x \in \mathbb{R}^p} f(x) + g(x) + h(Kx).$$

(P)

The functions $f : \mathbb{R}^p \to \mathbb{R}$, $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, and $h : \mathbb{R}^l \to \mathbb{R} \cup \{+\infty\}$ are assumed to be convex, closed, and proper; $K \in \mathbb{R}^{l \times p}$ is a linear operator. We assume $f$ is differentiable with $L_f$-Lipschitz gradient $\nabla f$. The functions $g$ and $h$ are not necessarily smooth. We assume that these possibly nonsmooth functions are “proximable,” i.e., the proximity operators $\text{prox}_g$ and $\text{prox}_h$ where

$$\text{prox}_g(u) = \arg \min_{x \in \mathbb{R}^p} \left\{ \phi(x) + (1/2)\|x - u\|_2^2 \right\}$$

are easy to evaluate; $\|x\|_2$ denotes the standard Euclidean norm of vector $x$.

Many important problems in statistics and machine learning can be formulated as problem (P). The following illustrates a few examples.

**Sparse generalized lasso** The generalized lasso (Tibshirani and Taylor, 2011) with an additional sparsity-inducing penalty is formulated as

$$\min_{x \in \mathbb{R}^p} \sum_{i=1}^n l_i(a_i^T x, b_i) + \lambda_1 \|x\|_1 + \lambda_2 \|Dx\|_1,$$

(1)

where the set $\{(a_i, b_i) : a_i \in \mathbb{R}^p, b_i \in \mathbb{R}, i = 1, \ldots, n\}$ constitutes a training sample, $l_i : \mathbb{R}^2 \to \mathbb{R}$ is the loss function that may depend on the sample index, $D \in \mathbb{R}^{l \times p}$ is the structure-inducing matrix, and $\| \cdot \|_1$ is the $\ell_1$ norm. In linear regression, $l_i(\cdot; b_i) = (\cdot - b_i)^2$, and the first term in (1) (corresponding to $f$ in (P)) has Lipschitz-continuous gradients with modulus $L_f = \|A^T\|_2$, with $A = [a_1, \ldots, a_n]^T$; $\|M\|_2$ is the standard operator norm of matrix $M$ (maximum singular value). In logistic regression, $l_i(\cdot; b_i) = -b_i(\cdot) - \log(1 + e^{-(\cdot)})$ and $L_f = \frac{1}{2} \|A^T\|_2$.

**Elastic net** The elastic net (Zou and Hastie, 2005) regularized regression use a linear combination of $\ell_1$ and $\ell_2$ penalties in order to promote both sparsity of solution and the grouping effect that highly correlated variables are selected or unselected together. The relevant optimization problem is

$$\min_{x \in \mathbb{R}^p} \frac{\lambda_2}{2} \|x\|_2^2 + \lambda_1 \|x\|_1 + l(Ax, b),$$

(2)

where the data matrix $A$ is the same as in the sparse generalized lasso, and $b = (b_1, \ldots, b_n)^T$. The loss function $l$ may also be nonsmooth, e.g., $l(Ax, b) = \|Ax - b\|_2$ (Belloni et al., 2011).

**PET image reconstruction** In positron emission tomography (PET), photon emissions from a radioactive tracer inside the brain are counted and the location-dependent emission rates are estimated. In this task,
the Radon transform [Jain 1989] is often discretized as matrix $A$. This results in a regularized nonnegative least squares problem, which can be written as
\[
\min_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - b\|_2^2 + \delta_+(x) + \lambda \|Dx\|_1,
\] (3)
where $x$ is the unknown emission map (image), $b$ is the vector of counts, and $\delta_+$ is the indicator function of the nonnegative orthant defined by $\delta_+(x) = 0$ if $x_1, \ldots, x_p \geq 0$ and $\delta_+(x) = +\infty$ otherwise. The $D$ is a discrete gradient operator encoding penalty on total variation.

1.1 Primal-dual formulation

Our major interest in this paper is efficient methods of solving (P) using only first-order information (i.e., $\nabla f$, $\text{prox}_g$, and $\text{prox}_h$) and matrix-vector multiplications (i.e., $Ku$ and $K^Tv$). For this purpose we propose Algorithm [Bauschke and Combettes 2011] named OS3X, and show that it achieves an optimal rate of convergence. We allow a stochastic setting in which the evaluation of the gradient $\nabla f$ is noisy. Such methods have gained a tremendous attention recently due to the advent of high-dimensional, “big data”.

If $h \equiv 0$, then the famous proximal gradient algorithm can be employed (Beck and Teboulle 2009). In deterministic setting, this method has an optimal convergence rate of $O(L_f/N^2)$, where $N$ is the number of iterations. However, in general, the presence of $h$ with $K \neq I$ invalidates the assumption of easy proximability. In this case it is often advantageous to reformulate problem (P) as a saddle point problem
\[
\min_{x \in X} \max_{y \in Y} \mathcal{L}(x, y) := f(x) + g(x) + \langle Kx, y \rangle - h^*(y),
\] (PD)
where $X = \text{dom} g$ and $Y = \text{dom} h^*$, denote the effective domains of the functions $g$ and $h^*(v) = \sup_{u \in \mathbb{R}^l} \langle u, v \rangle - h(u)$, the convex conjugate of $h$, with $\text{dom} \phi = \{u : \phi(u) < \infty\}$; $\langle u, v \rangle$ denotes the standard inner product $u^Tv$. Under a mild regularity condition, e.g., that $0$ is included in the relative interior of $K \text{ dom } g - \text{ dom } h$, a solution $(x^*, y^*)$ to (PD) exists. Furthermore, $x^*$ is a (primal) solution to (P), and $y^*$ is a solution to the associated dual
\[
\max_{y \in Y^*} \left( -(f + g)^*(-K^Ty) - h^*(y) \right). \tag{D}
\] (Bauschke and Combettes 2011, Theorem 19.1 and Proposition 19.18). In the sequel, we assume that (PD) has a solution and seek an algorithm that efficiently finds it.

1.2 Optimal rate of convergence

In both deterministic and stochastic settings, we derive the optimal rates of convergence for solving the saddle point problem (PD), as follows.

**Deterministic setting** In the special case of $g \equiv 0$, Chen et al. (2014) showed that (PD) can be solved at the optimal rate
\[
O \left( \frac{L_f}{N^2} + \frac{L_K}{N} \right),
\] (4)
where $L_K$ is an upper bound of $\|K\|_2$. It turns out, this rate is also optimal for the general case $g \neq 0$, in the following sense.

1. The optimal rate of solving $\min_{x \in X} (f(x) + g(x))$ by using any first-order method is $O(L_f/N^2)$ [Nesterov 2004], e.g., by using FISTA (Beck and Teboulle 2009).

2. For sufficiently large $p$, there exist $b \in Y \subset \mathbb{R}^l$ and $K \in \mathbb{R}^{l \times p}$ such that $h^*(y) = \langle b, y \rangle$ and the rate of convergence for solving $\min_{x \in X} \max_{y \in Y} (\langle Kx, y \rangle - h^*(y)) = \min_{x \in X} \max_{y \in Y} (\|Kx - b\|_2)$ is $O(L_K/N)$ (Nemirovsky 1992, Nemirovski 2004).

**Stochastic setting** The case that even the first-order information on the objective of (P) cannot be obtained exactly can be modeled by a stochastic oracle, which provides unbiased estimators of the first-order information. To be precise, at the $k$-th iteration we assume that the oracle returns a stochastic gradient $\hat{\nabla} f(x^k)$ independently from the previous iteration so that $\mathbb{E}[\hat{\nabla} f(x^k)] = \nabla f(x^k)$. If the variance of these estimators are uniformly bounded, i.e., $\mathbb{E}[\|\hat{\nabla} f(x^k) - \nabla f(x^k)\|_2^2] \leq \chi^2$, Chen et al. (2014) also showed that when $g \equiv 0$, (PD) can be solved in expectation at the optimal rate
\[
O \left( \frac{L_f}{N^2} + \frac{L_K}{N} + \frac{\chi}{\sqrt{N}} \right). \tag{5}
\]
Like the deterministic setting, this rate is also optimal for $g \neq 0$: for the first two terms the discussion of the deterministic setting above carries over. For the last term the argument by Chen et al. (2014) still applies.

1.3 Contributions

The major contributions of this paper are 1) establishing optimal rates (4) and (5), as already made above, 2) showing that Algorithm [Bauschke and Combettes 2011] achieves these optimal rates in their respective settings, for suitable choices of parameters, and 3) demonstrating its superior practical performance to other state-of-the-art algorithms.
for solving $\text{(PD)}$ in both deterministic and stochastic settings. Closeness to a solution to $\text{(PD)}$ is measured by the duality gap between $\text{(P)}$ and $\text{(D)}$. To our knowledge, this is the first work to establish and attain optimal-rate convergence under the general template $\text{(PD)}$ with minimal assumptions, e.g., the absence of strong convexity.

2 Related works

There is a vast literature on first-order methods for solving $\text{(PD)}$ under the deterministic setting. While the Alternating Directions Method of Multipliers (ADMM, Boyd et al., 2010) may be used to tackle $\text{(PD)}$, this method usually involves matrix inversion subproblems, which becomes quickly intractable as the dimension $p$ increases. In the direction of avoiding this difficulty, for $g \equiv 0$, the Primal Dual Hybrid Gradient method (PDHG, Zhu and Chan, 2008; Esser et al., 2010; Combettes and Pock, 2011) has been widely studied. Condat (2013) and Vu (2013) extend PDHG for the general case of $g \not\equiv 0$. These methods fall into the forward-backward operator splitting scheme (Bauschke and Combettes, 2011) and achieve the usual $O(1/N)$-rate. Another forward-backward splitting method for $g \equiv 0$, by Loris and Verhulst (2011), is subsumed by the Primal-Dual Fixed-Point algorithm (PDFP, Chen et al., 2016) for the general case. Other operator splitting approaches for $g \not\equiv 0$ include the Davis-Yin three-operator splitting (Davis and Yin, 2017) for $K = I$), Asymmetric Forward-Backward-Adjoint splitting (AFBA, Latafat and Patrinos, 2017) and Primal-Dual 3-Operator splitting (PD3O, Yan, 2018). The latter two include the above forward-backward splitting methods for $g \equiv 0$ as special cases, and allow general $K$. Acceleration by using variable step sizes and inertia has been studied (Combettes and Vu, 2014; Lorenz and Pock, 2015; Boţ et al., 2015; Goldstein et al., 2015; Combettes and Pock, 2016). Despite the reduction of the constant, they all remain in the $O(1/N)$ regime or require strong convexity.

On the other hand, interests in stochastic first-order methods for $\text{(PD)}$ in general settings appear to be rather recent. When $h \equiv 0$, stochastic versions of the proximal gradient method were considered (Hu et al., 2009; Lin et al., 2014; Nitanda, 2014; Rosasco et al., 2014; Atchadé et al., 2017). For the two-function problem ($K \not\equiv I$ but $g \equiv 0$), mirror-prox algorithms have been considered (Nemirovski et al., 2009; Juditsky et al., 2011; Lan, 2012). Ouyang and Gray (2014) developed a near-optimal algorithm under a strong convexity assumption on $f$ and smoothing of $g$. Zhong and Kwok (2014) achieved a similar rate to (5) under strong convexity. Without additional assumptions on $f$ or $g$ but assuming $K = I$, Yurtsever et al. (2016) introduced a stochastic variant of the Davis-Yin three-operator splitting. For general $K$, the Stochastic Primal-Dual algorithm for Three-composite Convex Minimization method (SPDTCM, Zhao and Cevher, 2018) is proposed. This method can be seen as a stochastic version of Combettes and Pock (2016), and has the rate of $O(Lf/N + LK/N + \sqrt{\tau}/\sqrt{N})$. Note that this rate is not optimal.

3 Algorithm OS3X

Our algorithm OS3X is presented in a separate panel as Algorithm 1. If the gradient evaluation in step (6e) is noisy, then iteration (6f) generates a stochastic sequence. Otherwise, it is deterministic. Algorithm 1 includes many other algorithms as special cases. If $\rho_k \equiv 1$, $\theta_k \equiv 0$, $\sigma_k \equiv \gamma$, $\tau_k \equiv \tau$ (constant step sizes) and $B = K$, then it reduces to the dual version of the extended PDHG by Condat (2013) and Vu (2013); if $g \equiv 0$, then it reproduces the class of optimal two-function algorithms by Ko et al. (2019+), which extends the optimal algorithm by Chen et al. (2014) ($B = 0$). As claimed, this algorithm involves only the evaluation of $\nabla f_j$, $\text{prox}_{g_j}$, and $\text{prox}_{h_k}$ and matrix-vector multiplications; Moreau’s identity $x = \text{prox}_{h_k}(x) + \text{prox}_{g_j}(x)$ converts the evaluation of $\text{prox}_{h_k}$ to that of $\text{prox}_{g_j}$.

In the next section, we show that for the following choices of the matrix $B$, relaxation parameter sequences $\{\rho_k\}$, $\{\theta_k\}$, and step size sequences $\{\sigma_k\}$, $\{\tau_k\}$, iteration (6f) converges at the rate (4) for the deterministic

---

**Algorithm 1 Optimal Sum-of-3-function Acceleration (OS3X)**

**Input:** Initial point $(x^1, y^1)$; positive sequences $\rho_k$, $\theta_k$, $\tau_k$, and $\sigma_k$; matrix $B \in \mathbb{R}^{d \times p}$; number of iterations $N$.

**Initialization:** Put $x^0 = x^1$, $y^0 = y^1$.

**Main loop:**

for $k = 1, 2, \ldots, N$ do

\[ \hat{u}^k = K\hat{x}^k + \theta_k K(x^k - \hat{x}^{k-1}) \]  
\[ x^k_{\text{md}} = (1 - \rho_k)\hat{x}^k + \rho_k \hat{u}^k \]  
\[ y^{k+1} = \text{pro}_{x_k h_k}(y^k + \sigma_k \hat{u}^k) \]  
\[ \tilde{v}^{k+1} = K^T(y^{k+1} + B^T(y^{k+1} - \hat{y}^k)) - \theta_k B^T(y^k - \hat{y}^{k-1}) \]  
\[ \hat{z}^{k+1} = \text{pro}_{x_k g_k}(\hat{x}^k - \theta_k (\nabla f(x_{\text{md}}) + \tilde{v}^{k+1})) \]  
\[ x^{k+1} = (1 - \rho_k)\hat{x}^k + \rho_k x_{\text{md}}^{k+1} \]  
\[ y^{k+1} = (1 - \rho_k)y^k + \rho_k \hat{y}^{k+1} \]

end for

**Output:** $(x^N, y^N)$
setting, and at the rate \(5\) for the stochastic setting, which are respectively optimal.

**Choice of the matrix parameter.** While the choice of matrices \(B\) can be made quite flexible, the choices of \(B = 0\) and \(B = -K\) make the steps \(6a\) the simplest. In general \(B = \alpha K\) for some scalar \(\alpha\) yields a simple update rule.

**Choice of the relaxation parameters.** For all cases, we choose the relaxation parameter \(\rho_k\) and the extrapolation parameter \(k\) as
\[
\rho_k = \frac{2}{k + 1}, \quad k = \frac{k - 1}{k}.
\]

**Choice of the step sizes.**
1) **Bounded domains, deterministic:** in the deterministic setting, if the diameter of the domains \(\mathcal{X}\) and \(\mathcal{Y}\) can be estimated so that
\[
\sup_{x, x' \in \mathcal{X}} \|x - x'\|^2_2 \leq 2\Omega_X^2, \quad \sup_{y, y' \in \mathcal{Y}} \|y - y'\|^2_2 \leq 2\Omega_Y^2,
\]
we consider increasing primal step sizes \(\{\tau_k\}\) and a constant dual step size \(\sigma_k\):
\[
\tau_k = \frac{k\Omega_X}{2P_1L_f\Omega_X + kP_2L_K\Omega_Y}, \quad \sigma_k = \frac{\Omega_Y}{\Omega_XL_K}
\]
for some properly chosen positive constants \(P_1\) and \(P_2\). The primal step size \(\tau_k\) increases over iterates, while the dual step size \(\sigma_k\) is kept a constant. The choice for \(P_1\) and \(P_2\) is discussed in §4.

2) **Unbounded domains, deterministic:** for the cases where bounds for primal or dual variables are unknown, we assume that the horizon \(N\) is known in advance and consider increasing step sizes:
\[
\tau_k = \frac{k}{2P_1L_f + P_2N\Omega_K}, \quad \sigma_k = \frac{k}{N\Omega_K}.
\]

3) **Bounded domains, stochastic:** now for the stochastic setting, where the domain bounds \(8\) are known, our choice of the step sizes is
\[
\tau_k = \frac{\Omega_Xk}{2P_1L_f\Omega_X + P_2L_K\Omega_Yk + P_3\chi k^{3/2}},
\]
\[
\sigma_k = \frac{\Omega_Y}{L_K\Omega_X},
\]
for some positive constants \(P_1\), \(P_2\), and \(P_3\). The choices for \(P_1\), \(P_2\), and \(P_3\) that achieve the optimal rates are discussed in §4.

4) **Unbounded domains, stochastic:** on the other hand, if bounds \(8\) are unknown, the choices are:
\[
\tau_k = \frac{k}{2P_1L_f + P_2L_K(N - 1) + P_3\sqrt{N - 1}\chi},
\]
\[
\sigma_k = \frac{(N - 1)L_K + P_3\sqrt{N - 1}\chi}{N\Omega_K}.
\]

As in the deterministic counterpart \(10\), we assume that the horizon \(N\) is known in advance.

**Remark 1.** Using four algorithm parameters \((\rho_k, \theta_k, \tau_k, \sigma_k)\) is common in accelerated algorithms, e.g., Chen et al. (2014), Zhao and Cehreli (2018), and Ko et al. (2019+). Dependence of step sizes \(\sigma_k\) and \(\tau_k\) on \(\Omega_X\) and \(\Omega_Y\) also appears in Chen et al. (2014). In fact, we can always overestimate either bound so that \(\Omega_X = \Omega_Y\). In this case, step sizes in \(9\) become independent of \(\Omega_X\) and \(\Omega_Y\). For \(11\), we can choose \(P_3 = \Omega_X\) or \(\Omega_Y\) and again make the step sizes independent of these bounds; see Corollary 3. Estimation of the Lipschitz constant \(L_f\) and the bound \(L_K\) can be carried out by backtracking and the power method.

## 4 Convergence analysis

In this section, we show that Algorithm 7 achieves the theoretically optimal rate of convergence for each of the four settings discussed in §3. We define the pre-gap function \(\tilde{G}(\tilde{x}, z) := L(\tilde{x}, y) - L(x, y)\), and the duality gap function \(G^*(\tilde{x}) := \sup_{z \in \mathcal{Z}} G(z, \tilde{z})\), where \(z = (x, y), \tilde{z} = (\tilde{x}, \tilde{y})\), and \(\mathcal{Z} = \mathcal{X} \times \mathcal{Y}\). Nonnegativity of \(G^*(\tilde{x})\) guarantees that \(\tilde{x}\) is a solution to \(\tilde{P}\) under the assumption that \(\tilde{P}\) has a solution. Convergence is thus measured by how fast \(G^*(\tilde{x})\) approaches to zero. When \(\mathcal{Z}\) is unbounded, however, the gap \(G^*(\tilde{x})\) may tend to positive infinity. In this case, we consider the perturbed gap function instead:
\[
\tilde{G}(\tilde{x}, v) := \sup_{z \in \mathcal{Z}} G(z, \tilde{z}) - \langle v, \tilde{z} - z \rangle.
\]

It is known that there always exists a perturbation vector \(v\) that makes function \(13\) finite (Monteiro and Svaiter, 2011). We find a sequence of vanishing perturbation vectors \(\{v^k\}\) that make \(\tilde{G}(\tilde{x}^k, v^k)\) small.

### 4.1 Deterministic setting

#### 4.1.1 Bounded domains

We first consider the case in which the bound for \(\mathcal{X}\) and \(\mathcal{Y}\) are known. Under this assumption, we have the following bound for the duality gap:

**Theorem 1.** Let \(\{x^k, y^k\}\) be the sequence generated by Algorithm 7. Assume for some \(\Omega_X, \Omega_Y > 0\),
\[
\sup_{x, x' \in \mathcal{X}} \|x - x'\|^2_2 \leq 2\Omega_X^2, \quad \sup_{y, y' \in \mathcal{Y}} \|y - y'\|^2_2 \leq 2\Omega_Y^2,
\]
and the parameter sequences \(\{\rho_k\}, \{\theta_k\}, \{\tau_k\}\), and \(\{\sigma_k\}\) satisfy \(\rho_1 = 1\) and
\[
\rho_{k+1} = \rho_k \theta_{k+1},
\]
\[\begin{align}
(14) & \quad \rho_{k+1} = \rho_k \theta_{k+1}, \\
(15a) & \quad \rho_{k+1} = \rho_k \theta_{k+1},
\end{align}\]
Then for all \( q \in (0, 1), r \in (0, 1) \). Further suppose that
\[
0 < \theta_k \leq \min(\tau_k - 1, \sigma_k - 1, \sigma_k),
\]
max(\( \tau_k - 1, \sigma_k - 1, \sigma_k \)) \leq 1.
(16)
Then for all \( k \geq 1 \),
\[
G^*(z^{k+1}) \leq \frac{\rho_k}{\tau_k} \Omega_X^2 + \frac{\rho_k}{\sigma_k} \Omega_Y^2.
\]
(17)
For the discussed choice of the algorithm parameters, we obtain the claimed optimal convergence rate.

**Corollary 1.** Assume that \( ||B||_2 \leq bL_K \) for some \( b > 0 \). Choose the parameter sequences \( \{\rho_k\}, \{\theta_k\} \) as in (7), and \( \{\tau_k\}, \{\sigma_k\} \) as in (9). Finally, if
\[
P_1 = \frac{1}{1-q}, \quad P_2 = \max \left\{ \frac{1}{1-q}, \frac{b^2}{1-\tau^2}, 1 \right\}
\]
holds, then
\[
\epsilon_{n+1} \leq \left( \frac{4P_1 L_f}{N} + \frac{2P_2 b}{N} \right) \left[ \frac{2q-1}{1-1/2r-1} R^2 \right. \]
(24)
and
\[
\epsilon_{n+1} \leq \left( \frac{4P_1 L_f}{N} + \frac{2P_2 b}{N} \right) \left[ \frac{2q-1}{1-1/2r-1} R^2 \right. \]
(25)

**Remark 2.** Setting \( B = 0 \) satisfies the assumption in Corollary 2. In this case Algorithm 2 resembles Chen et al. (2014) for sum of two functions. It is interesting that other selections such as \( B = K \) or \( B = -K \) also achieve the optimal rate.

### 4.1.2 Unbounded domains

Now we consider the case where the bounds for \( \mathcal{X} \) or \( \mathcal{Y} \) are not known in advance.

**Theorem 2.** Suppose that \( \{z^k\} = \{(x^k, y^k)\} \) are generated by Algorithm 1. If the parameter sequences \( \{\rho_k\}, \{\theta_k\}, \{\tau_k\}, \) and \( \{\sigma_k\} \) satisfy (15) and
\[
\theta_k = \tau_k - 1/\tau_k = \sigma_k - 1/\sigma_k \leq 1
\]
(20)
for some \( 0 < q < 1 \) and \( 0 < r < 1/2 \). Then there exists a vector \( v^{k+1} \) such that for any \( k \geq 1 \),
\[
\hat{G}(z^{k+1}, v^{k+1}) \leq \frac{\rho_k}{\tau_k} \left( 2 + \frac{q}{1-q} + \frac{2r+1}{1-2r} \right) R^2,
\]
(21)
and
\[
\|v^{k+1}\|_2 \leq \left( \frac{\rho_k}{\tau_k} \|\hat{x} - \hat{x}\|_2 + \frac{\rho_k}{\tau_k} \|\hat{y} - \hat{y}\|_2 \right) + \left[ \frac{\rho_k}{\tau_k} \left( \mu + \frac{\tau_k}{\mu} \nu \right) + 2\rho_k(\mu L_K + \nu \|B\|_2) \right] R,
\]
(22)
where \( (\hat{x}, \hat{y}) \) is a pair of solutions to (PD), and
\[
R = \sqrt{\|\hat{x} - \hat{x}\|_2^2 + \tau_1/\sigma_1 \|\hat{y} - \hat{y}\|_2^2}, \quad \mu = 1/\sqrt{1-\frac{q}{2}}, \nu = \sqrt{2\sigma_1/\tau_1(1-2r)}.
\]
For the choice of parameters given by (7) and (10), we obtain the optimal rate.

**Corollary 2.** Assume that \( ||B||_2 \leq bL_K \) for some \( b > 0 \). Choose the parameter sequences \( \{\rho_k\}, \{\theta_k\} \) as in (7), and \( \{\tau_k\}, \{\sigma_k\} \) as in (10). Finally, if
\[
P_1 = \frac{1}{1-q}, \quad P_2 = \max \left\{ \frac{1}{1-q}, \frac{b^2}{1-\tau^2}, 1 \right\}
\]
holds, then
\[
\epsilon_{n+1} \leq \left( \frac{4P_1 L_f}{N} + \frac{2P_2 b}{N} \right) \left[ \frac{2q-1}{1-1/2r-1} R^2 \right. \]
(24)
and
\[
\epsilon_{n+1} \leq \left( \frac{4P_1 L_f}{N} + \frac{2P_2 b}{N} \right) \left[ \frac{2q-1}{1-1/2r-1} R^2 \right. \]
(25)

### 4.2 Stochastic setting

Recall the assumptions stated above equation (5):
\[
E[\nabla f(x^k)] = \nabla f(x^k), \quad E[\|\nabla f(x^k) - \nabla f(x^k)\|_2^2] \leq \chi^2.
\]
(27)

#### 4.2.1 Bounded domains

When the bounds for \( \mathcal{X} \) and \( \mathcal{Y} \) are known in advance, the following holds.

**Theorem 3.** Let \( \{z^k\} = \{(x^k, y^k)\} \) be the sequence generated by Algorithm 2, where \( \nabla f \) satisfies condition (27). Suppose that the boundedness condition (8) holds for some \( \Omega_X, \Omega_Y > 0 \). Also assume that for all \( k \geq 1 \), the parameter sequences \( \{\rho_k\}, \{\theta_k\}, \{\tau_k\}, \) and \( \{\sigma_k\} \) satisfy \( \rho_1 = 1, \{15a\}, \{16\} \), and
\[
(s-q)/\tau_k - L_f \rho_k - L_k^2 \sigma_k/r \geq 0,
\]
\[
(t-r)/\sigma_k - \tau_k \|B\|_2^2/q \geq 0
\]
(28)
for some \( q, r, s, t \in (0, 1) \). Then the following holds.
The claimed optimal rate is obtained as follows.

**Corollary 3.** Let \( \{z^k\} = \{(x^k, y^k)\} \) be the sequence generated by (6), where \( \hat{\nabla} f \) satisfies condition (27). Assume the condition (8) holds. In the stochastic variant of Algorithm 7, suppose \( \|B\|_2 \leq bL_k \) for some \( b > 0 \), and the parameters are set as in (7) and (11). Let \( P_1, P_2, \) and \( P_3 \) be constants such that

\[
P_1 = \frac{1}{s - q}, \quad P_2 \geq \max \left\{ \frac{1}{r(s-q)}, \frac{\bar{V}^2}{\nu(t-r)} \right\},
\]

\[
P_3 > 0,
\]

where \( q, r, s, t \in (0,1), q < s, r < t \). Then we have

\[
E[\mathcal{G}^*(z^{k+1}, y^{k+1})] \leq \frac{8P_1L_1\Omega_k^2}{4(k+1)} + \frac{4\Omega_k\Omega_1L_k(P_2+1)}{\sqrt{2}p_k^{1/2}k^{1/2}} + \left( 4P_3 + \frac{\sqrt{2}(s-2)}{3P_1(s-1)} \right) \frac{\sqrt{2}p_k^{1/2}}{\sqrt{k}}
\]

for any \( k \geq 1 \).

**Remark 4.** Zhao and Cevher (2018: Remark 3), who achieve the rate \( O(L_f/N + L_k/N + \chi/\sqrt{N}) \), suggest that the rate for the smooth part \( f \) may be improved to \( O(L_f/N^2) \). We have shown that this is indeed possible and the resulting rate is optimal.

### 4.2.2 Unbounded domains

Now we consider the case where bounds for \( \mathcal{X} \) and \( \mathcal{Y} \) are unavailable.

**Theorem 4.** Let \( \{z^k\} = \{(x^k, y^k)\} \) be the sequence generated by Algorithm 7, where \( \hat{\nabla} f \) satisfies the condition (27). Suppose that for all \( k \geq 1 \), the parameter sequences \( \{\rho_k\} \), \( \{\theta_k\} \), and \( \{\sigma_k\} \) satisfy \( \rho_1 = 1 \), \( \theta_1 = \frac{1}{L_1} \), \( \sigma_1 = \frac{1}{L_2} \), and \( \tau_1 = \frac{1}{L_3} \) for some \( q, s, t \in (0,1) \) and \( r \in (0,1/2) \). Then there exists a perturbation vector \( v^{k+1} \) such that

\[
E[\tilde{\mathcal{G}}(z^{k+1}, v^{k+1})] \leq \frac{\rho_k}{\rho_k} \left[ \left( 6 + \frac{4q}{1-q} + \frac{4(r+1/2)}{1/2-r} \right) R^2 + \left( \frac{5}{2} + \frac{2q}{1-q} + \frac{2(r+1/2)}{1/2-r} \right) \frac{S^2}{\sigma_k} \right]
\]

for each \( k \geq 1 \). Furthermore,

\[
E[\|v^{k+1}\|_2^2] \leq \frac{2\rho_k\|x-x_k\|_2^2}{\tau_k} + \frac{2\rho_k\|y-y_k\|_2^2}{\sigma_k}
\]

The desired optimal rate can be obtained as follows.

**Corollary 4.** Assume that the condition (27) holds. In the stochastic variant of Algorithm 7, suppose the horizon (number of iterations) \( N \geq 1 \) is given, \( \|B\|_2 \leq bL_k \) for some \( b > 0 \), and the parameters are set as in (7) and (11). Let \( P_1, P_2, \) and \( P_3 \) be constants such that

\[
P_1 = \frac{1}{s - q}, \quad P_2 \geq \max \left\{ \frac{1}{r(s-q)}, \frac{\bar{V}^2}{\nu(t-r)} \right\}, \quad P_3 = \sqrt{\frac{2-r}{1-s}} \tilde{R}
\]

for some \( \tilde{R} > 0 \), where \( q, s, t \in (0,1), r \in (0,1/2), q < s, \) and \( r < t \). Then we have

\[
E[\tilde{\mathcal{G}}(z^N, v^N)] \leq \left( \frac{4P_1L_1}{N(N-1)} + \frac{2P_3L_k}{N} + \frac{\sqrt{2-r}}{1-s} \frac{2\chi/\sqrt{N}}{1\sqrt{N-1}} \right) \times \left[ \left( 6 + \frac{4q}{1-q} + \frac{4(r+1/2)}{1/2-r} \right) R^2 + \frac{1}{3} \left( \frac{5}{2} + \frac{2q}{1-q} + \frac{2(r+1/2)}{1/2-r} \right) \tilde{R}^2 \right],
\]

\[
E[\|v^N\|_2^2] \leq \left( \frac{4P_1L_1}{N(N-1)} + \frac{2P_3L_k}{N} + \frac{\sqrt{2-r}}{1-s} \frac{2\chi/\sqrt{N}}{1\sqrt{N-1}} \right) \times \left[ 2\tilde{R} \left( 1 + \frac{\sqrt{\chi}}{\sigma_k} \right) + (\sqrt{2\tilde{R}} + \tilde{R}/\sqrt{3}) \left( 1 + \mu + \left( \frac{\sqrt{\chi}}{\sigma_k} + \nu \right) \right) \right] + \frac{4L_k}{N} \left( \sqrt{2\tilde{R}} + \tilde{R}/\sqrt{3} \right) (\mu + bv).
\]

### 4.3 Outline of the proofs

The following proposition is a key in proving the above results.

**Proposition 1.** Assume that \( \rho_k \leq 1 \) for any \( k \). If \( z^k = (x^k, y^k) \) is generated by (6), then for any \( z = (x, y) \in \mathcal{Z} \),

\[
\rho_k^{-1} \mathcal{G}(z^{k+1}, z) - \rho_k^{-1} \mathcal{G}(z^k, z) \leq \langle \nabla f(x^k_{\text{med}}), \hat{x}^{k+1} - x \rangle + \frac{\rho_kL_f}{2} \|\hat{x}^{k+1} - \hat{x}^k\|_2^2 + g(\hat{x}^{k+1}) - g(x) + h^*(y^{k+1}) - h^*(y) + \langle K \hat{x}^{k+1}, y \rangle - \langle K x, y \rangle
\]

for each \( k \geq 1 \).
Lemmas B.1 and C.1 in Supplementary Material are derived from Proposition 1. Theorems 1–4 follow from these lemmas. Detailed proofs are provided in Supplementary Material.

5 Numerical experiments

5.1 Setup

We compare the practical performance of OS3X (Algorithm 1) with the benchmark methods. For the deterministic setting, we consider Condat-Vu (CV), PDFP, AFBA, PD3O, and SPDTCM without noisy gradients. For the stochastic setting, we compare OS3X with SPDTCM with noise. We tested with two instances of PD, namely graph-guided fused lasso and overlapping group elastic net. We averaged 10 separate runs for all stochastic experiments. For each experiment, primal gap versus the number of epochs is shown. An epoch was defined as (cumulative number of data points used in the estimation of $\nabla f$)/(number of data points in the dataset). The primal gap is the difference between the objective value at the epoch and the optimal objective value, approximated by the objective value after 100000 epochs of deterministic OS3X. We tested three instances of OS3X: $B = 0$, $B = -0.5K$, and $B = -K$. The algorithms were implemented in Matlab R2017a on a machine with two Intel E5-2650 v4 processors and 256 GB RAM.

Parameter selection In the deterministic setting, we chose $q = 0.3$ and $r = 0.7$ from Corollary 2 and Corollary 3, and set $P_1 = 0.9$ for OS3X. For stochastic setting, $(q, r, s, t)$ from Corollary 3 were chosen as (0.3, 0.3, 0.7, 0.7). The variance $\chi$ was set 1000. For CV, PDFP, AFBA, and PD3O, we chose $\tau = 1.9/L_f$ and $\sigma = 1/(4\tau)$. Finally, for SPDTCM, we used the constant parameter recipe as provided by Zhao and Cevher (2018).

Stochastic gradient At iteration $k$, the stochastic gradient $\nabla f(x^k)$ was obtained from a random subsample of $A$. For a random permutation $\pi$, we define a subsample $\tilde{A} := A_{\pi(1); \pi(n_s); \cdot}$ (in Matlab notation), where $n_s = \lfloor 0.2n \rfloor$. Thus for the quadratic loss, we have $\nabla f(x^k) = (n/n_s)\tilde{A}^T(\tilde{A}x - b)$.

5.2 Graph-guided sparse fused lasso

The graph-guided fused lasso is formulated as

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda_1\|x\|_1 + \lambda_2\|Dx\|_1,$$

where $D$ is the difference matrix on a given undirected graph. The data were generated following the transcription factor (TF) model of Zhu (2017). The graph had $J$ fully connected subgraphs of size $T$, where each subgraph had one node designated as TF and the rest were regulatory targets. TF variables were sampled independently from $\mathcal{N}(0, 1)$. Target genes were sampled so that each target gene and the corresponding TF has a bivariate normal with zero mean, unit variance, and correlation of 0.7. Target genes were conditionally independent given the TF. For $j$-th subgraph, we chose

$$x_i = \begin{cases} (-1)^{j+1} \frac{2^{i+1}}{2^j} & \text{if } j = 1, \ldots, J_a, \\ 0 & \text{otherwise} \end{cases},$$

where $i = (j - 1)r + 1, \ldots, jr$, and $J_a$ is the number of active subgraphs. Response $b_i$ was sampled so that $b_i = Ax_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 100^2)$. In addition, we added random edges between active nodes and inactive nodes. For each active node, we added edges connecting this node and $J - 1$ distinct inactive nodes. We used $T = 10$, $J_a = 20$, and $J = 1000$ so that $p = 10000$. The data matrix $A$ was sampled i.i.d. from $\mathcal{N}(0, 1)$. Penalty parameters were set $\lambda_1 = 1 = \lambda_2$. Domain boundaries were estimated as $\Omega_X = 200$, $\Omega_Y = 450$. All the iterates remained within these boundaries. For stochastic unbounded parameter settings, we chose $\Bar{R} = 100$. The results are shown in Figure 1(a-c). The convergence speed gap between OS3X and the other methods is clear (note the log-log scale). Using the parameters with known bounds is faster than the parameters that do not involve bound assumption, but we still achieve faster convergence compared to other methods without the bound assumption. There was no noticeable difference between the choices of $B$.

5.3 Overlapping group elastic net

The overlapping group elastic net problem with a quadratic loss with an additional ridge penalty is:

$$\min_x \frac{1}{2}\|b - Ax\|_2^2 + \frac{\lambda_1}{2}\|x\|_2^2 + \lambda_2\sum_{j=1}^{100}\sqrt{|G_j|}\|x_{G_j}\|_2.$$

The test dataset was generated based on Chen et al. (2012). We defined 100 groups of 100 variables of adjacent indices, with 10 overlaps of adjacent groups. i.e., $G_j = \{90(j - 1) + 1, \ldots, 90j + 10\}$, thus $p = 9010$. We set $x_j = (-1)^j \exp(-(j - 1)/100)$ for $j = 1, \ldots, p$. We sampled each element of $A$ i.i.d. from $\mathcal{N}(0, 1)$, and added a Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to $Ax$ to generate $b = Ax + \epsilon$. The sample size was $n = 5000$. We chose $\lambda_1 = 0.1$, $\lambda_2 = 0.3$, and set $\Omega_X = 20$, $\Omega_Y = 45$. For stochastic case with unbounded parameter setting, we chose $\Bar{R} = 50$. The results are shown in Figure 1(d-f). All the instances of OS3X converged faster than SPDTCM. Stochastic variants of OS3X start slowly, but they surpass SPDTCM eventually.
6 Discussion

It is interesting that the middle-step aggregation strategy for accelerating PDHG-type algorithms for a three-function sum achieves the optimal rate. (This strategy is the key idea of Chen et al. (2014).) Our results thus provide a partial answer to the popularity of the base algorithm by Condat (2013) and Vũ (2013).

There remain several avenues of future research. First, in this work we maintain a minimal assumption on the convexity of the functions since the interest is in the worst-case rates. How the bounds of our algorithm class can be improved with additional assumptions, e.g., the strong convexity of $g$ (Ghadimi and Lan, 2012), would be of interest. Second, in the unbounded settings we assume the horizon $N$ is known in advance. Using step sizes that depend on $N$ at least dates back to Nesterov (2005): achieving optimal rates without this information is a challenging task (Zhao and Cevher, 2018) report a factor of $\log N$ slowdown in the asymptotic rate). However, in many scenarios (e.g., early stopping) the knowledge of $N$ is unavailable, horizon-independent convergence analysis is warranted. Third, techniques for estimating the problem parameters $L_f$ and $L_K$ and combining them with algorithm parameter selection will have an important practical impact.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2014R1A4A1007895).

References


