Optimal Minimization of the Sum of Three Convex Functions with a Linear Operator

Seyoon Ko
Seoul National University

Joong-Ho Won
Seoul National University

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:

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

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 all 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 ∇f. The functions g and h are not necessarily smooth. We assume that these possibly nonsmooth functions are "proximable," i.e., the proximity operators $\mathbf{prox}_g$ and $\mathbf{prox}_h$ where

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

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

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

where the set $\{(a_i, b_i) : a_i \in \mathbb{R}^p, b_i \in \mathbb{R}, i = 1, \dots, n\}$ 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 ℓ_1 norm. In linear regression, $l_i(\cdot; b_i) = \frac{1}{2}(\cdot - b_i)^2$, and the first term in (1) (corresponding to f in (P)) has Lipschitz-continuous gradients with modulus $L_f = \|A^T A\|_2$, with $A = [a_1, \dots, a_n]^T$; $\|M\|_2$ 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 $L_f = \frac{1}{4}\|A^T A\|_2$.

Elastic net The elastic net (Zou and Hastie, 2005) regularized regression use a linear combination of ℓ_1 and ℓ_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

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

where the data matrix A is the same as in the sparse generalized lasso, and $b = (b_1, \dots, b_n)^T$. 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,

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

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

$$\min_{x\in\mathbb{R}^p} \frac{1}{2}\|Ax-b\|_2^2 + \delta_+(x) + \lambda\|Dx\|_1, \tag{3}$$

where x is the unknown emission map (image), b is the vector of counts, and δ_+ 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., ∇f, $\mathbf{prox}_g$, and $\mathbf{prox}_h$) and matrix-vector multiplications (i.e., Ku and K^Tv). For this purpose we propose Algorithm 1, named OS3X, and show that it achieves an optimal rate of convergence. We allow a stochastic setting in which the evaluation of the gradient ∇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

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{L}(x,y) := f(x) + g(x) + \langle Kx, y\rangle - h^*(y), \tag{PD}$$

where $\mathcal{X} = \mathbf{dom}\, g$ and $\mathcal{Y} = \mathbf{dom}\, h^*$, 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 $\mathbf{dom}\,\phi = \{u : \phi(u) < \infty\}$; $\langle u, v\rangle$ denotes the standard inner product u^Tv. Under a mild regularity condition, e.g., that 0 is included in the relative interior of $K\,\mathbf{dom}\, g - \mathbf{dom}\, h$, a solution $(x^\star, y^\star)$ to (PD) exists. Furthermore, $x^\star$ is a (primal) solution to (P), and $y^\star$ is a solution to the associated dual

$$\max_{y\in\mathcal{Y}*} \big(-(f+g)^*(-K^Ty) - h^*(y)\big) \tag{D}$$

(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

$$O\left(\frac{L_f}{N^2} + \frac{L_K}{N}\right), \tag{4}$$

where L_K is an upper bound of $\|K\|_2$. It turns out, this rate is also optimal for the general case $g \not\equiv 0$, in the following sense.

1. The optimal rate of solving $\min_{x\in\mathcal{X}}(f(x)+g(x))$ 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 $b \in \mathcal{Y} \subset \mathbb{R}^l$ and $K \in \mathbb{R}^{l\times p}$ such that $h^*(y) = \langle b, y\rangle$ and the rate of convergence for solving $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}}(\langle Kx, y\rangle - h^*(y)) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \langle Kx - b, y\rangle$ is $\Omega(L_K/N)$ (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 $\widehat{\nabla f}(x^k)$ independently from the previous iteration so that $\mathbb{E}[\widehat{\nabla f}(x^k)] = \nabla f(x^k)$. If the variance of these estimators are uniformly bounded, i.e., $\mathbb{E}[\|\widehat{\nabla f}(x^k) - \nabla f(x^k)\|_2^2] \leq \chi^2$, 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 $g \not\equiv 0$: 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 1 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 (PD) in both deterministic and stochastic settings. Closeness to a solution to (PD) is measured by the duality gap between (P) and (D). To our knowledge, this is the first work to establish and attain optimal-rate convergence under the general template (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 (PD) under the deterministic setting. While the Alternating Directions Method of Multipliers (ADMM, Boyd et al., 2010) may be used to tackle (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; Chambolle and Pock, 2011) has been widely studied. Condat (2013) and Vũ (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 Verhoeven (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 Vũ, 2014; Lorenz and Pock, 2015; Boţ et al., 2015; Goldstein et al., 2015; Chambolle 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 (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 ($K \neq I$ but $g \equiv 0$), mirror-prox algorithms have been considered (Nemirovski et al., 2009; Juditsky et al., 2011; Lan, 2012). Ouyang and Gray (2012) 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 Chambolle and Pock (2016), and has the rate of $O(L_f/N + L_K/N + \chi/\sqrt{N})$. Note that this rate is not optimal.

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

Input: Initial point (x^1, y^1); positive sequences ρ_k, θ_k, τ_k, and σ_k; matrix $B \in \mathbb{R}^{l\times p}$; number of iterations N.
Initialization: Put $\tilde{x}^0 = \tilde{x}^1 = x^1$, $\tilde{y}^0 = \tilde{y}^1 = y^1$.
Main loop:
for $k = 1, 2, \ldots, N$ **do**

$$\bar{u}^k = K\tilde{x}^k + \theta_k K(\tilde{x}^k - \tilde{x}^{k-1}) \tag{6a}$$

$$x^k_{md} = (1-\rho_k)x^k + \rho_k\tilde{x}^k \tag{6b}$$

$$\tilde{y}^{k+1} = \mathbf{prox}_{\sigma_k h^*}(\tilde{y}^k + \sigma_k\bar{u}^k) \tag{6c}$$

$$\tilde{v}^{k+1} = K^T\tilde{y}^{k+1} + B^T(\tilde{y}^{k+1} - \tilde{y}^k) - \theta_k B^T(\tilde{y}^k - \tilde{y}^{k-1}) \tag{6d}$$

$$\tilde{x}^{k+1} = \mathbf{prox}_{\tau_k g}\left(\tilde{x}^k - \tau_k(\widehat{\nabla f}(x^k_{md}) + \tilde{v}^{k+1})\right) \tag{6e}$$

$$x^{k+1} = (1-\rho_k)x^k + \rho_k\tilde{x}^{k+1} \tag{6f}$$

$$y^{k+1} = (1-\rho_k)y^k + \rho_k\tilde{y}^{k+1} \tag{6g}$$

end for
Output: (x^N, y^N)

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 (6) generates a stochastic sequence. Otherwise, it is deterministic. Algorithm 1 includes many other algorithms as special cases. If $\rho_k \equiv 1$, $\theta_k \equiv 0$, $\sigma_k \equiv \sigma$, $\tau_k \equiv \tau$ (constant step sizes) and $B = K$, then it reduces to the dual version of the extended PDHG by Condat (2013) and Vũ (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 ∇f, $\mathbf{prox}_g$, and $\mathbf{prox}_h$ and matrix-vector multiplications; Moreau's identity $x = \mathbf{prox}_h(x) + \mathbf{prox}_{h^*}(x)$ converts the evaluation of $\mathbf{prox}_{h^*}$ to that of $\mathbf{prox}_h$.

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 (6) converges at the rate (4) for the deterministic

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 step (6d) the simplest. In general $B = \alpha K$ for some scalar α yields a simple update rule.

Choice of the relaxation parameters. For all cases, we choose the relaxation parameter ρ_k and the extrapolation parameter θ_k as

$$\rho_k = \frac{2}{k+1}, \quad \theta_k = \frac{k-1}{k}. \tag{7}$$

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

$$\sup_{x,x' \in \mathcal{X}} \|x - x'\|_2^2 \le 2\Omega_X^2, \ \sup_{y,y' \in \mathcal{Y}} \|y - y'\|_2^2 \le 2\Omega_Y^2, \tag{8}$$

we consider increasing primal step sizes $\{\tau_k\}$ and a constant dual step size σ_k:

$$\tau_k = \frac{k\Omega_X}{2P_1 L_f \Omega_X + kP_2 L_K \Omega_Y}, \quad \sigma_k = \frac{\Omega_Y}{\Omega_X L_K} \tag{9}$$

for some properly chosen positive constants P_1 and P_2. The primal step size τ_k increases over iterates, while the dual step size σ_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:

$$\tau_k = \frac{k}{2P_1 L_f + P_2 N L_K}, \quad \sigma_k = \frac{k}{N L_K}. \tag{10}$$

3) Bounded domains, stochastic: now for the *stochastic* setting, where the domain bounds (8) are known, our choice of the step sizes is

$$\tau_k = \frac{\Omega_X k}{2P_1 L_f \Omega_X + P_2 L_K \Omega_Y k + P_3 \chi k^{3/2}}, \quad \sigma_k = \frac{\Omega_Y}{L_K \Omega_X}, \tag{11}$$

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:

$$\tau_k = \frac{k}{2P_1 L_f + P_2 L_K (N-1) + P_3 N\sqrt{N-1}\chi}, \quad \sigma_k = \frac{k}{(N-1)L_K + P_3 N\sqrt{N-1}\chi}. \tag{12}$$

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 Cevher (2018), and Ko et al. (2019+). Dependence of step sizes σ_k and τ_k on Ω_X and Ω_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 Ω_X and Ω_Y. For* (11)*, we can choose $P_3 = \Omega_X$ or Ω_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 anlaysis

In this section, we show that Algorithm 1 achieves the theoretically optimal rate of convergence for each of the four settings discussed in §3. We define the pre-gap function $\mathcal{G}(\tilde{z}, z) := \mathcal{L}(\tilde{x}, y) - \mathcal{L}(x, \tilde{y})$, and the duality gap function $\tilde{\mathcal{G}}^\star(\tilde{z}) := \sup_{z \in \mathcal{Z}} \mathcal{G}(\tilde{z}, z)$, where $z = (x, y)$, $\tilde{z} = (\tilde{x}, \tilde{y})$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Nonnegativity of $\mathcal{G}^\star(\tilde{z})$ guarantees that $\tilde{x}$ is a solution to (P) under the assumption that (PD) has a solution. Convergence is thus measured by how fast $\mathcal{G}^\star(\tilde{z})$ approaches to zero. When $\mathcal{Z}$ is unbounded, however, the gap $\mathcal{G}^\star(\tilde{z})$ may tend to positive infinity. In this case, we consider the perturbed gap function instead:

$$\tilde{\mathcal{G}}(\tilde{z}, v) := \sup_{z \in Z} \mathcal{G}(\tilde{z}, z) - \langle v, \tilde{z} - z \rangle. \tag{13}$$

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 $\tilde{\mathcal{G}}(\tilde{z}^k, v^k)$ 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 $\{z^k\} = \{(x^k, y^k)\}$ be the sequence generated by Algorithm 1. Assume for some $\Omega_X, \Omega_Y > 0$,*

$$\sup_{x,x' \in \mathcal{X}} \|x - x'\|_2^2 \le 2\Omega_X^2, \quad \sup_{y,y' \in \mathcal{Y}} \|y - y'\|_2^2 \le 2\Omega_Y^2, \tag{14}$$

and the parameter sequences $\{\rho_k\}$, $\{\theta_k\}$, $\{\tau_k\}$, and $\{\sigma_k\}$ satisfy $\rho_1 = 1$ and

$$\rho_{k+1}^{-1} - 1 = \rho_k^{-1} \theta_{k+1}, \tag{15a}$$

$$(1-q)/\tau_k - L_f\rho_k - (1/r)L_K^2\sigma_k \geq 0, \tag{15b}$$
$$(1-r)/\sigma_k - (\tau_k/q)\|B\|_2^2 \geq 0 \tag{15c}$$

for some $q \in (0,1)$, $r \in (0,1)$. *Further suppose that*

$$\begin{aligned} &0 < \theta_k \leq \min(\tau_{k-1}/\tau_k, \sigma_{k-1}/\sigma_k), \\ &\max(\tau_{k-1}/\tau_k, \sigma_{k-1}/\sigma_k) \leq 1. \end{aligned} \tag{16}$$

Then for all $k \geq 1$,

$$\mathcal{G}^\star(z^{k+1}) \leq \tfrac{\rho_k}{\tau_k}\Omega_X^2 + \tfrac{\rho_k}{\sigma_k}\Omega_Y^2. \tag{17}$$

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*

$$P_1 = \tfrac{1}{1-q},\ P_2 = \max\left\{\tfrac{1}{(1-q)r}, \tfrac{b^2}{q(1-r)}\right\} \tag{18}$$

holds, then

$$\mathcal{G}^\star(z^k) \leq \tfrac{4P_1\Omega_X^2}{k(k-1)}L_f + \tfrac{2\Omega_X\Omega_Y(P_2+1)}{k}L_K, \quad \forall k \geq 2. \tag{19}$$

Remark 2. *Setting* $B = 0$ *satisfies the assumption in Corollary 1. In this case Algorithm 1 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* (6). *If the parameter sequences* $\{\rho_k\}$, $\{\theta_k\}$, $\{\tau_k\}$, *and* $\{\sigma_k\}$ *satisfy* (15) *and*

$$\theta_k = \tau_{k-1}/\tau_k = \sigma_{k-1}/\sigma_k \leq 1 \tag{20}$$

for some $0 < q < 1$ *and* $0 < r < 1/2$. *Then there exists a vector* v^{k+1} *such that for any* $k \geq 1$,

$$\tilde{\mathcal{G}}(z^{k+1}, v^{k+1}) \leq \tfrac{\rho_k}{\tau_k}\left(2 + \tfrac{q}{1-q} + \tfrac{2r+1}{1-2r}\right)R^2, \tag{21}$$

and

$$\begin{aligned} \|v^{k+1}\|_2 \leq{}& \left(\tfrac{\rho_k}{\tau_k}\|\hat{x} - \tilde{x}^1\|_2 + \tfrac{\rho_k}{\sigma_k}\|\hat{y} - \tilde{y}^1\|_2\right) \\ &+ \left[\tfrac{\rho_k}{\tau_k}\left(\mu + \tfrac{\tau_1}{\sigma_1}\nu\right) + 2\rho_k(\mu L_K + \nu\|B\|_2)\right]R, \end{aligned} \tag{22}$$

where $(\hat{x}, \hat{y})$ *is a pair of solutions to* (PD), *and*

$$\begin{aligned} &R = \sqrt{\|\hat{x} - \tilde{x}^1\|_2^2 + \tau_1/\sigma_1\|\hat{y} - \tilde{y}^1\|_2^2}, \\ &\mu = 1/\sqrt{1-q}, \nu = \sqrt{2\sigma_1/\tau_1(1-2r)}. \end{aligned} \tag{23}$$

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*

$$P_1 = \tfrac{1}{1-q},\ P_2 = \max\left\{\tfrac{1}{(1-q)r}, \tfrac{b^2}{q(1-r)}, 1\right\} \tag{24}$$

holds, then

$$\epsilon_{N+1} \leq \left(\tfrac{4P_1L_f}{N^2} + \tfrac{2P_2L_K}{N}\right)\left[\tfrac{2-q}{1-q} + \tfrac{r+1/2}{1/2-r}\right]R^2 \tag{25}$$

and

$$\|v^{N+1}\|_2 \leq \tfrac{4P_1DL_f}{N^2} + \tfrac{L_K}{N}\left[2P_2D + 4R(\mu + b\nu)\right], \tag{26}$$

where $D = \|\hat{x} - \tilde{x}^1\|_2 + \|\hat{y} - \tilde{y}^1\|_2 + R\left(\mu + \tau_1\nu/\sigma_1\right)$.

Remark 3. *The* (x^N, y^N) *in Corollary 2 can be considered an approximation to a solution to the problem* (PD) *in a sense that for any pair of positive scalars* (ρ, ε), *there is an* N *such that* $\|v^N\| \leq \rho$ *and* $\epsilon_N \leq \varepsilon$. *This analysis has to do with a* (ρ, ε)*-saddle point of of problem* (PD) *(Monteiro and Svaiter, 2011, Definition 3.10), and* ε*-subgradient of* h^* *for arbitrarily small* ρ *and* ε. *In a nutshell, when the perturbed gap* $\tilde{\mathcal{G}}(\tilde{z}, v)$ *is small and* $\|v\|_2$ *is also small, then* $\tilde{x}$ *has a small optimal gap for the primal* (P) *and* $\tilde{y}$ *has a small optimal gap for dual* (D); *see also Proposition 4 in Ko et al. (2019+) and the discussion thereafter. A very similar idea of a "nearly optimal" solution is used in Goldstein et al. (2015, Eq. (19)).*

4.2 Stochastic setting

Recall the assumptions stated above equation (5):

$$\mathbb{E}[\widehat{\nabla f}(x^k)] = \nabla f(x^k),\ \mathbb{E}[\|\widehat{\nabla f}(x^k) - \nabla f(x^k)\|_2^2] \leq \chi^2. \tag{27}$$

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 1, where* $\widehat{\nabla f}$ *satisfies condition* (27). *Suppose that the boundness condition* (8) *holds for some* Ω_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*

$$\begin{aligned} (s-q)/\tau_k - L_f\rho_k - L_K^2\sigma_k/r \geq 0, \\ (t-r)/\sigma_k - \tau_k\|B\|_2^2/q \geq 0 \end{aligned} \tag{28}$$

for some q, r, s, $t \in (0,1)$. *Then the following holds.*

$$\mathbb{E}[\mathcal{G}^\star(z^{k+1})] \le 2\rho_k\left(\frac{\Omega_X^2}{\tau_k}+\frac{\Omega_Y^2}{\sigma_k}\right) + \frac{\rho_k}{2\gamma_k}\sum_{i=1}^{k}\frac{(2-s)\tau_i\gamma_i}{1-s}\chi^2, \quad (29)$$

where

$$\gamma_k = \begin{cases} 1 & \text{if } k=1 \\ \gamma_{k-1}/\theta_k & \text{if } k\ge 2 \end{cases}. \quad (30)$$

The claimed optimal rate is obtained as follows.

Corollary 3. *Let $\{z^k\}=\{(x^k,y^k)\}$ be the sequence generated by (6), where $\widehat{\nabla f}$ satisfies condition (27). Assume the condition (8) holds. In the stochastic variant of Algorithm 1, suppose $\|B\|_2 \le bL_K$ 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*

$$P_1=\tfrac{1}{s-q},\ \ P_2 \ge \max\left\{\tfrac{1}{r(s-q)},\tfrac{b^2}{q(t-r)}\right\},\ \ P_3>0, \quad (31)$$

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

$$\mathbb{E}[\mathcal{G}^\star(x^{k+1},y^{k+1})] \le \tfrac{8P_1L_f\Omega_X^2}{k(k+1)} + \tfrac{4\Omega_X\Omega_Y L_K(P_2+1)}{k+1} + \left(4P_3+\tfrac{2\sqrt{2}(2-s)}{3P_3(1-s)}\right)\tfrac{\chi\Omega_X}{\sqrt{k}} \quad (32)$$

for any $k\ge 1$.

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 1, where $\widehat{\nabla f}$ satisfies the condition (27). Suppose that for all $k\ge 1$, the parameter sequences $\{\rho_k\}$, $\{\theta_k\}$, $\{\tau_k\}$, and $\{\sigma_k\}$ satisfy $\rho_1=1$, (15a), (20), and (28) 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*

$$\mathbb{E}[\tilde{\mathcal{G}}(z^{k+1},v^{k+1})] \le \tfrac{\rho_k}{\tau_k}\left[\left(6+\tfrac{4q}{1-q}+\tfrac{4(r+1/2)}{1/2-r}\right)R^2 + \left(\tfrac{5}{2}+\tfrac{2q}{1-q}+\tfrac{2(r+1/2)}{1/2-r}\right)S^2\right] \quad (33)$$

for each $k\ge 1$. Furthermore,

$$\mathbb{E}[\|v^{k+1}\|_2] \le \tfrac{2\rho_k\|\hat{x}-x^1\|_2}{\tau_k}+\tfrac{2\rho_k\|\hat{y}-y^1\|_2}{\sigma_k} + \sqrt{2R^2+S^2}\left[\tfrac{\rho_k(1+\mu)}{\tau_k}+\left(\nu+\tfrac{\sigma_1}{\tau_1}\right)\tfrac{\rho_k}{\sigma_k} + 2\rho_k(L_K\mu+\|B\|_2\nu)\right], \quad (34)$$

where $(\hat{x},\hat{y})$ is a pair of solutions for (PD), R, μ, and ν are as defined in (23), and

$$S=\sqrt{\textstyle\sum_{i=1}^{k}(2-s)\tau_i^2\chi^2/(1-s)}. \quad (35)$$

The desired optimal rate can be obtained as follows.

Corollary 4. *Assume that the condition (27) holds. In the stochastic variant of Algorithm 1, suppose the horizon (number of iterations) $N\ge 1$ is given, $\|B\|_2\le bL_K$ for some $b>0$, and the parameters are set as in (7) and (12). Let P_1, P_2, and P_3 be constants such that*

$$P_1=\tfrac{1}{s-q},\ \ P_2\ge\max\left\{\tfrac{1}{r(s-q)},\tfrac{b^2}{q(t-r)},1\right\},\ \ P_3=\sqrt{\tfrac{2-s}{1-s}}/\tilde{R} \quad (36)$$

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

$$\mathbb{E}[\tilde{\mathcal{G}}(z^N,v^N)] \le \left(\tfrac{4P_1L_f}{N(N-1)}+\tfrac{2P_2L_K}{N}+\sqrt{\tfrac{2-s}{1-s}}\tfrac{2\chi/\tilde{R}}{\sqrt{N-1}}\right) \times\left[\left(6+\tfrac{4q}{1-q}+\tfrac{4(r+1/2)}{1/2-r}\right)R^2 + \tfrac{1}{3}\left(\tfrac{5}{2}+\tfrac{2q}{1-q}+\tfrac{4(r+1/2)}{1/2-r}\tilde{R}^2\right)\right],$$

$$\mathbb{E}[\|v^N\|_2] \le \left(\tfrac{4P_1L_f}{N(N-1)}+\tfrac{2P_2L_K}{N}+\sqrt{\tfrac{2-s}{1-s}}\tfrac{2\chi/\tilde{R}}{\sqrt{N-1}}\right) \times\left[2R\left(1+\sqrt{\tfrac{\sigma_1}{\tau_1}}\right) + (\sqrt{2}R+\tilde{R}/\sqrt{3})\left(1+\mu+\left(\sqrt{\tfrac{\sigma_1}{\tau_1}}+\nu\right)\right)\right] + \tfrac{4L_K}{N}(\sqrt{2}R+\tilde{R}/\sqrt{3})(\mu+b\nu).$$

4.3 Outline of the proofs

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

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

$$\begin{aligned}&\rho_k^{-1}\mathcal{G}(z^{k+1},z)-(\rho_k^{-1}-1)\mathcal{G}(z^k,z)\\ &\le\langle\nabla f(x_{md}^k),\tilde{x}^{k+1}-x\rangle+\frac{\rho_kL_f}{2}\|\tilde{x}^{k+1}-\tilde{x}^k\|_2^2\\ &+g(\tilde{x}^{k+1})-g(x)+h^*(\tilde{y}^{k+1})-h^*(y)\\ &+\langle K\tilde{x}^{k+1},y\rangle-\langle Kx,\tilde{y}^{k+1}\rangle.\end{aligned}$$

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-Vũ (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 $\widehat{\nabla f}$)/(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 1 and Corollary 2, and set $P_1 = 0.9$ for OS3X. For stochastic setting, (q, r, s, t) from Corollary 3 and Corollary 4 were chosen as $(0.3, 0.3, 0.7, 0.7)$. The variance χ 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 $\widehat{\nabla f}(x^k)$ was obtained from a random subsample of A. For a random permutation π, we define a subsample $\tilde{A} := A_{\pi(1):\pi(n_s),:}$ (in Matlab notation), where $n_s = \lfloor 0.2n \rfloor$. Thus for the quadratic loss, we have $\widehat{\nabla f}(x^k) = (n/n_s)\tilde{A}^T(\tilde{A}x - b)$.

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

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

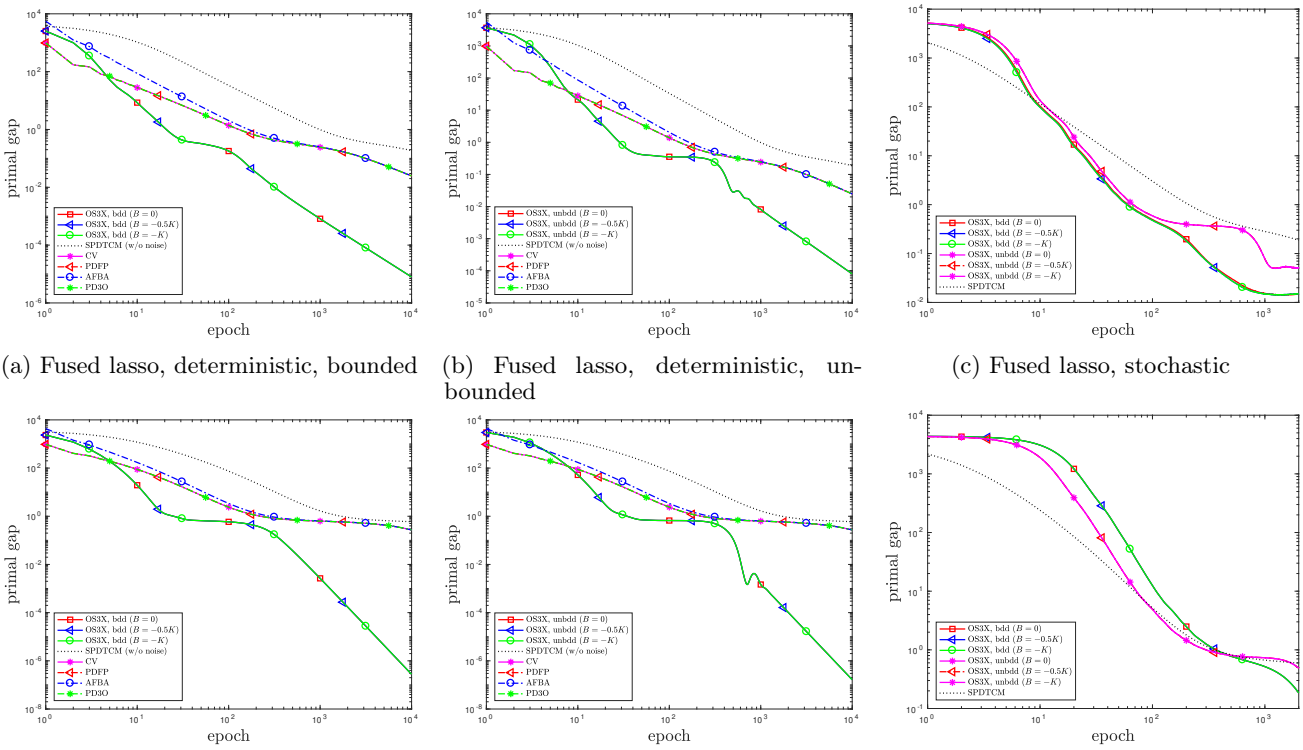
where $i = (j-1)r + 1, \ldots, jr$, and J_a is the number of active subgraphs. Response b_i was sampled so that $b_i = Ax + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 100^2)$. In addition, we added random edges between active nodes and inactive nodes. For each active 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 $\tilde{R} = 100$. 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 $\tilde{R} = 50$. 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.



(a) Fused lasso, deterministic, bounded

(b) Fused lasso, deterministic, unbounded

(c) Fused lasso, stochastic

(d) Group elastic net, deterministic, bounded

(e) Group elastic net, deterministic, unbounded

(f) Group elastic net, stochastic

Figure 1: Convergence of deterministic and stochastic OS3X under various parameter settings and other methods for a sparse graph-guided fused lasso model (a-c) and a overlapping group elastic net (d-f). (a), (d), deterministic OS3X with bounded parameter settings with SPDTCM with deterministic updates, CV, PDFP, AFBA, and PD3O. (b), (e), deterministic OS3X with unbounded parameter settings with SPDTCM with deterministic updates, CV, PDFP, AFBA, and PD3O. (c), (f), stochastic OS3X with bounded and unbounded parameter settings with SPDTCM.

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

Atchadé, Y. F., Fort, G. and Moulines, E. (2017). On perturbed proximal gradient algorithms, *Journal of Machine Learning Research* **18**(1): 310–342.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer Science & Business Media.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear in-

verse problems, *SIAM Journal on Imaging Sciences* **2**(1): 183–202.

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming, *Biometrika* **98**(4): 791–806.

Boţ, R. I., Csetnek, E. R., Heinrich, A. and Hendrich, C. (2015). On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems, *Mathematical Programming* **150**(2): 251–279.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends in Machine Learning* **3**(1): 1–122.

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

Chambolle, A. and Pock, T. (2016). On the ergodic convergence rates of a first-order primal-dual algorithm, *Mathematical Programming* **159**(1-2): 253–287.

Chen, P., Huang, J. and Zhang, X. (2016). A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions, *Fixed Point Theory and Applications* **2016**(54).

Chen, X., Lin, Q., Kim, S., Carbonell, J. G. and Xing, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression, *The Annals of Applied Statistics* **6**(2): 719–752.

Chen, Y., Lan, G. and Ouyang, Y. (2014). Optimal primal-dual methods for a class of saddle point problems, *SIAM Journal on Optimization* **24**(4): 1779–1814.

Combettes, P. L. and Vũ, B. C. (2014). Variable metric forward-backward splitting with applications to monotone inclusions in duality, *Optimization* **63**(9): 1289–1318.

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

Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications, *Set-valued and variational analysis* **25**(4): 829–858.

Esser, E., Zhang, X. and Chan, T. F. (2010). A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science, *SIAM Journal on Imaging Sciences* **3**(4): 1015–1046.

Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework, *SIAM Journal on Optimization* **22**(4): 1469–1492.

Goldstein, T., Li, M. and Yuan, X. (2015). Adaptive primal-dual splitting methods for statistical learning and image processing, *Advances in Neural Information Processing Systems 28*, pp. 2089–2097.

Hu, C., Pan, W. and Kwok, J. T. (2009). Accelerated gradient methods for stochastic optimization and online learning, *Advances in Neural Information Processing Systems*, pp. 781–789.

Jain, A. K. (1989). *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice Hall.

Juditsky, A., Nemirovski, A. and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm, *Stochastic Systems* **1**(1): 17–58.

Ko, S., Yu, D. and Won, J.-H. (2019+). Easily parallelizable and distributable class of algorithms for structured sparsity, with optimal acceleration, *Journal of Computational and Graphical Statistics, accepted for publication, preprint available at arXiv:1702.06234* .

Lan, G. (2012). An optimal method for stochastic composite optimization, *Mathematical Programming* **133**(1-2): 365–397.

Latafat, P. and Patrinos, P. (2017). Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators, *Computational Optimization and Applications* **68**(1): 57–93.

Lin, Q., Chen, X. and Peña, J. (2014). A smoothing stochastic gradient method for composite optimization, *Optimization Methods and Software* **29**(6): 1281–1301.

Lorenz, D. A. and Pock, T. (2015). An inertial forward-backward algorithm for monotone inclusions, *Journal of Mathematical Imaging and Vision* **51**(2): 311–325.

Loris, I. and Verhoeven, C. (2011). On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty, *Inverse Problems* **27**(12): 125007.

Monteiro, R. D. and Svaiter, B. F. (2011). Complexity of variants of tseng's modified fb splitting and korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems, *SIAM Journal on Optimization* **21**(4): 1688–1720.

Nemirovski, A. (2004). Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, *SIAM Journal on Optimization* **15**(1): 229–251.

Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming, *SIAM Journal on optimization* **19**(4): 1574–1609.

Nemirovsky, A. (1992). Information-based complexity of linear operator equations, *Journal of Complexity* **8**(2): 153–175.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*, Vol. 87, Springer Science & Business Media.

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

Nitanda, A. (2014). Stochastic proximal gradient descent with acceleration techniques, *Advances in Neural Information Processing Systems*, pp. 1574–1582.

Ouyang, H. and Gray, A. (2012). Stochastic smoothing for nonsmooth minimizations: accelerating sgd by exploiting structure, *International Coference on International Conference on Machine Learning*, pp. 1523–1530.

Rosasco, L., Villa, S. and Vũ, B. C. (2014). Convergence of stochastic proximal gradient algorithm, *arXiv preprint arXiv:1403.5074* .

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso, *The Annals of Statistics* **39**(3): 1335–1371.

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

Yan, M. (2018). A new primal–dual algorithm for minimizing the sum of three functions with a linear operator, *Journal of Scientific Computing* **76**(3): 1698–1717.

Yurtsever, A., Vũ, B. C. and Cevher, V. (2016). Stochastic three-composite convex minimization, *Advances in Neural Information Processing Systems*, pp. 4329–4337.

Zhao, R. and Cevher, V. (2018). Stochastic three-composite convex minimization with a linear operator, *International Conference on Artificial Intelligence and Statistics*, pp. 765–774.

Zhong, W. and Kwok, J. (2014). Accelerated stochastic gradient method for composite regularization, *International Converence on Artificial Intelligence and Statistics*, pp. 1086–1094.

Zhu, M. and Chan, T. (2008). An efficient primal-dual hybrid gradient algorithm for total variation image restoration, *UCLA CAM Report* pp. 08–34.

Zhu, Y. (2017). An augmented ADMM algorithm with application to the generalized lasso problem, *Journal of Computational and Graphical Statistics* **26.1**: 195–204.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320.