Three Operator Splitting with Subgradients, Stochastic Gradients, and Adaptive Learning Rates

Alp Yurtsever⋆
Umeå University
alp.yurtsever@umu.se

Alex Gu⋆ & **Suvrit Sra**
Massachusetts Institute of Technology
{gua,suvrit}@mit.edu

Abstract

Three Operator Splitting (TOS) (Davis & Yin, 2017) can minimize the sum of multiple convex functions effectively when an efficient gradient oracle or proximal operator is available for each term. This requirement often fails in machine learning applications: *(i)* instead of full gradients only stochastic gradients may be available; and *(ii)* instead of proximal operators, using subgradients to handle complex penalty functions may be more efficient and realistic. Motivated by these concerns, we analyze three potentially valuable extensions of TOS. The first two permit using subgradients and stochastic gradients, and are shown to ensure a $\mathcal{O}(1/\sqrt{t})$ convergence rate. The third extension AdapTos endows TOS with adaptive step-sizes. For the important setting of optimizing a convex loss over the intersection of convex sets AdapTos attains universal convergence rates, i.e., the rate adapts to the *unknown* smoothness degree of the objective function. We compare our proposed methods with competing methods on various applications.

1 Introduction

We study convex optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \quad \phi(x) := f(x) + g(x) + h(x), \tag{1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ and $g, h : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ are proper, lower semicontinuous and convex functions. Importantly, this template captures constrained problems via indicator functions. To avoid pathological examples, we assume the relative interiors of $\mathrm{dom}(f)$, $\mathrm{dom}(g)$ and $\mathrm{dom}(h)$ have a nonempty intersection.

Problem (1) is motivated by a number of applications in machine learning, statistics, and signal processing, where the three functions comprising the objective ϕ model data fitting, structural priors, or decision constraints. Examples include overlapping group lasso (Yuan et al., 2011), isotonic regression (Tibshirani et al., 2011), dispersive sparsity (El Halabi & Cevher, 2015), graph transduction (Shivanna et al., 2015), learning with correlation matrices (Higham & Strabić, 2016), and multidimensional total variation denoising (Barbero & Sra, 2018).

An important technique for addressing composite problems of the form (1) is *operator splitting* (Bauschke et al., 2011). However, a naive use of the basic proximal-(sub)gradient method

Alp Yurtsever and Alex Gu contributed equally to this paper. The paper is based on the work done while Alp Yurtsever was at Massachusetts Institute of Technology.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

may be unsuitable, since it requires access to the prox-operator of $g + h$, computing which may be vastly more expensive than individual prox-operators of g and h. An elegant, recent method, Three Operator Splitting (TOS, Davis & Yin (2017), see Algorithm 2) offers a practical choice for solving Problem (1) when f is L-smooth, and prox-operators of both g and h are available. Importantly, at each iteration, TOS evaluates the gradient of f and the proximal operators of g and h only once. Moreover, problems comprising more than three functions can be reformulated as an instance of Problem (1) in a product-space and minimized with TOS as long as each term has an efficient gradient oracle or proximal operator (see Section 2).

Unfortunately, TOS is not readily applicable to many optimization problems that arise in machine learning. Most important among those are problems where only access to stochastic gradients is feasible, e.g., when performing large-scale empirical risk minimization and online learning. Moreover, prox-operators for some complex penalty functions are computationally expensive and it may be more efficient to instead use subgradients. For example, proximal operator for the maximum eigenvalue function that appears in dual-form semidefinite programs (e.g., see Section 6.1 in (Ding et al., 2019)) may require computing a full eigendecomposition. In contrast, we can form a subgradient by computing only the top eigenvector via power method or Lanczos algorithm.

Contributions. With the above motivation, this paper contributes three key extensions of TOS. We tackle nonsmoothness in Section 3 and stochasticity in Section 4. These two extensions enable us to use subgradients and stochastic gradients of f in TOS (see Section 2 for a comparison with related work), and satisfy a $\mathcal{O}(1/\sqrt{T})$ convergence rate in function value after T iterations. The third main contribution is AdapTos in Section 5. This extension provides an adaptive step-size rule in the spirit of AdaGrad (Duchi et al., 2011; Levy, 2017) for an important subclass of Problem (1). Notably, for optimizing a convex loss over the intersection of two convex sets, AdapTos ensures universal convergence rates. That is, AdapTos implicitly adapts to the *unknown* smoothness of the problem, and ensures a $\tilde{\mathcal{O}}(1/\sqrt{T})$ rate when the problem is nonsmooth or stochastic, with the rate improving to $\tilde{\mathcal{O}}(1/T)$ if the problem is smooth and a solution lies in the relative interior of the feasible set.

In Section 6, we discuss empirical performance of our methods by comparing them against present established methods on various benchmark problems from COPT Library (Pedregosa et al., 2020) including the overlapping group lasso, total variation deblurring, and sparse and low-rank matrix recovery. We also test our methods on nonconvex optimization to train a neural network model.

Notation. We denote a solution of Problem (1) by $x_\star$ and $\phi_\star = \phi(x_\star)$. The distance between a point $x \in \mathbb{R}^n$ and a closed and convex set $\mathcal{G} \subseteq \mathbb{R}^n$ is $\mathrm{dist}(x, \mathcal{G}) := \min_{y\in\mathcal{G}} \|x - y\|$; the projection of x onto $\mathcal{G}$ is given by $\mathrm{proj}_{\mathcal{G}}(x) := \arg\min_{y\in\mathcal{G}} \|x - y\|$. The proximal operator (or prox-operator) of a function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$\mathrm{prox}_g(x) := \arg\min_{y\in\mathbb{R}^n} \left[g(y) + \frac{1}{2}\|x - y\|^2 \right]$$

The indicator function of $\mathcal{G}$ gives 0 for any $x \in \mathcal{G}$ and $+\infty$ otherwise. Recall that the prox-operator for the indicator function is the projection onto the corresponding set.

2 Background and related work

TOS, proposed recently by Davis & Yin (2017), can be seen as a generic extension of various operator splitting schemes, including the forward-backward splitting, Douglas-Rachford splitting, forward-Douglas-Rachford splitting (Briceño-Arias, 2015), and the generalized forward-backward splitting (Raguet et al., 2013). It covers these aforementioned approaches as special instances when the terms f, g and h in Problem (1) are chosen appropriately. Convergence of TOS is well studied when f has Lipschitz continuous gradients. It ensures $\mathcal{O}(1/t)$ convergence rate in this setting, see (Davis & Yin, 2017) and (Pedregosa, 2016) for the technical details.

Other related methods that can be used for Problem (1) with smooth f are the primal-dual hybrid gradient (PDHG) method (Condat, 2013; Vũ, 2013) and the primal-dual three operator splitting method (Yan, 2018). These methods can handle a slightly more general template where g or h is composed with a linear map, however, they require f to be smooth. The convergence rate of PDHG is analyzed in (Chambolle & Pock, 2016).

Nonsmooth setting. We are unaware of any prior result that permits using subgradients in TOS (or in related methods that can use prox-operator of g and h separately) for Problem (1). The closest approach is the proximal subgradient method which applies if h is missing, and it is covered by our nonsmooth TOS as a special case.

Stochastic setting. There are multiple attempts to devise a stochastic TOS in the literature. Yurtsever et al. (2016) studied Problem (1) under the assumption that f is smooth and strongly convex, and an unbiased gradient estimator with bounded variance is available. Their stochastic TOS has a guaranteed $\mathcal{O}(1/t)$ convergence rate. In (Cevher et al., 2018), they relaxed the strong convexity assumption but assume that the gradient estimator's variance is summable. They show asymptotic convergence with no guarantees on the rate. Later, Pedregosa et al. (2019) proposed a stochastic variance-reduced TOS and analyzed its non-asymptotic convergence guarantees. Their method gets $\mathcal{O}(1/t)$ convergence rate for arbitrary smooth convex f. The rate becomes linear if f is smooth and strongly convex and g (or h) is smooth.

Very recently, Yurtsever et al. (2021) studied TOS on problems where f can be nonconvex but the problem domain is bounded. They have shown that the method finds a first-order stationary point with $\mathcal{O}(1/\sqrt[3]{t})$ convergence rate under a diminishing variance assumption. They use an increasing mini-batch size to ensure that this condition holds.

None of these prior works cover the broad template we consider: f is smooth or Lipschitz continuous and the stochastic first-order oracle has bounded variance. To our knowledge, our paper gives the first analysis for stochastic TOS without strong convexity assumption or variance reduction.

Other related methods include stochastic PDHG in (Zhao & Cevher, 2018) and a stochastic primal-dual method in (Zhao et al., 2019). The latter can be viewed as an extension of stochastic ADMM (Ouyang et al., 2013; Azadi & Sra, 2014) from the sum of two terms to three terms in the objective.

Adaptive step-sizes. The standard writings of TOS and PDHG require the knowledge of the smoothness constant of f for the step-size. Backtracking line-search strategies are proposed to find a suitable step-size when the smoothness constant is unknown, for PDHG in (Malitsky & Pock, 2018) and for TOS in (Pedregosa & Gidel, 2018). A heuristic line-search strategy was also introduced in the original preprint (Davis & Yin, 2015), but it is reported as nonconvergent in (Pedregosa & Gidel, 2018).

The motivation and goals of these line-search strategies significantly differ from our adaptive learning rate. Importantly, these methods work only when f is smooth; they require extra function evaluations, and are thus not suitable for stochastic optimization. And their goal is to estimate the *smoothness constant*. In contrast, our method does not require function evaluations. It can be used for smooth, nonsmooth, or stochastic settings. Most importantly, it can adapt to the unknown *smoothness degree*.

At the heart of our method lie adaptive online learning algorithms (Duchi et al., 2011; Rakhlin & Sridharan, 2013) together with online to offline conversion techniques (Levy, 2017; Cutkosky, 2019). Similar methods appear in the literature for other problem templates with no constraint or a single constraint in (Levy, 2017; Levy et al., 2018; Kavis et al., 2019; Cutkosky, 2019). Our method extends these results to optimization over the intersection of convex sets. When f is nonsmooth or stochastic, the proposed method ensures a convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{t})$, whereas the rate of the same algorithm improves to $\tilde{\mathcal{O}}(1/t)$ if f is smooth and there is a solution in the relative interior of the feasible set.

TOS for more than three functions. TOS can be used for solving problems with more than three convex functions via a product-space reformulation technique (Briceño-Arias, 2015). Consider

$$\min_{x \in \mathbb{R}^d} \quad \sum_{i=1}^{q} \phi_i(x), \tag{2}$$

where each proper and lower semicontinuous function $\phi_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is either smooth or has an efficient prox-operator. Without loss of generality, suppose $\phi_1, \ldots, \phi_p$ are smooth and $\phi_{p+1}, \ldots, \phi_q$ are prox-friendly. Then, we can reformulate (2) in the product-space $\mathbb{R}^{d \times (p+1)}$ as

$$\min_{x_0, x_1, \ldots, x_p \in \mathbb{R}^d} \quad \sum_{i=1}^{p} \phi_i(x_i) + \sum_{i=p+1}^{q} \phi_i(x_0) \quad \text{subject to} \quad x_0 = x_1 = \ldots = x_p. \tag{3}$$

This is an instance of Problem (1) with $n = d \times (p+1)$ and $x = (x_0, x_1, \ldots, x_p)$. We can choose $g(x)$ as the indicator for the equality constraint, $f(x) = \sum_{i=p+1}^{q} \phi_i(x_0)$, and $h(x) = \sum_{i=1}^{p} \phi_i(x_i)$. Then, gradient of f is the sum of gradients of $\phi_p + 1, \ldots, \phi_q$; prox_g is a mapping that averages $x_0, x_1, \ldots, x_p$; and prox_h is the concatenation of the individual prox-operators of $\phi_1, \ldots, \phi_p$.

noend 1 Three Operator Splitting (TOS)

input: initial $y_0 \in \mathbb{R}^n$, functions f, g, h, step-sizes $\gamma_0, \cdots, \gamma_T$
for $t = 0, 1, 2, \ldots, T$ **do**
  $z_t = \text{prox}_{\gamma_t g}(y_t)$
  $x_t = \text{prox}_{\gamma_t h}(2z_t - y_t - \gamma_t u_t)$
  $y_{t+1} = y_t - z_t + x_t$
return: $\bar{x}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} x_\tau$ and $\bar{z}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} z_\tau$

noend 2 Three Operator Splitting (TOS)

input: initial $y_0 \in \mathbb{R}^n$, functions f, g, h, step-sizes $\gamma_0, \cdots, \gamma_T$
for $t = 0, 1, 2, \ldots, T$ **do**
  $z_t = y_t$
  $x_t = 2z_t - y_t - \gamma_t \nabla f(z_t)$
  $y_{t+1} = y_t - z_t + x_t$
return: $\bar{x}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} x_\tau$ and $\bar{z}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} z_\tau$

To our knowledge, TOS has been studied for Problem (1) only when f is smooth. This permits nonsmooth terms in (2) to be processed only via their proximal operator. In this work, by enhancing TOS with a subgradient step for nonsmooth f, we provide the flexibility to choose how to process each nonsmooth ϕ_i, either by the proximal operator or subgradient.

3 TOS for Nonsmooth Setting

Algorithm 2 presents the generalized TOS for Problem (1). It recovers the standard method in (Davis & Yin, 2017) if f is smooth and we choose $u_t = \nabla f(z_t)$. For convenience, we define the mapping

$$\texttt{TOS}_\gamma(y, u) := y - \text{prox}_{\gamma g}(y) + \text{prox}_{\gamma h}\big(2 \cdot \text{prox}_{\gamma g}(y) - y - \gamma u\big) \tag{4}$$

which represents one iteration of Algorithm 2.

The first step of the analysis is the fixed-point characterization of TOS. The following lemma is a straightforward extension of Lemma 2.2 in (Davis & Yin, 2017) to permit subgradients. The proof is similar to (Davis & Yin, 2017), we present it in the supplementary material for completeness.

$$u_t = \text{oracle of} \partial f(z_t), \gamma_t = \mathcal{O}(1/L)$$

$$u_t = \text{oracle of } \partial f(z_t), \gamma_t = \mathcal{O}(1/\sqrt{T})$$

Lemma 1 (Fixed points of TOS). *Let $\gamma > 0$. Then, there exists a subgradient $u \in \partial f(\text{prox}_{\gamma g}(y))$ that satisfies $\texttt{TOS}_\gamma(y, u) = y$ if and only if $\text{prox}_{\gamma g}(y)$ is a solution to Problem (1).*

Ergodic sequence. Convergence of operator splitting methods are often given in terms of ergodic (averaged) sequences. This strategy requires maintaining the running averages of z_t and x_t:

$$\bar{x}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} x_\tau \qquad \text{and} \qquad \bar{z}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} z_\tau. \tag{5}$$

Clearly, we do not need to store the history of x_t and z_t to maintain these sequences. In practice, the last iterate often converges faster than the ergodic sequence. We can evaluate the objective function at both points and return the one with the smaller value.

We are ready to present convergence guarantees of TOS for the nonsmooth setting.

Assumption 1. *There exists $G_f > 0$ such that*

$$\|u\| \leq G_f, \qquad \forall u \in \partial f(x), \ \forall x \in \text{dom}(g). \tag{6}$$

This is a standard assumption in nonsmooth optimization, equivalent to assuming f is G_f-Lipschitz continuous on $\mathrm{dom}(g)$.

Theorem 1. *Consider Problem (1) and let Assumption 1 hold. Employ TOS (Algorithm 2) and choose the update direction and step-size as*

$$\gamma_t = \mathcal{O}(1/L_f)$$
$$\gamma_t = \mathcal{O}(1/\sqrt{T})$$

$u_t \in \partial f(z_t)$ *and* $\gamma_t = \frac{\gamma_0}{\sqrt{T+1}}$ *for some* $\gamma_0 > 0$, *for* $t = 0, 1, \dots, T$.(7)*Then, the following guarantees hold:*

$$\mathbb{E}[f(\bar{z}_T) + g(\bar{z}_T) + h(\bar{x}_T) - \phi_\star] \leq \mathcal{O}(1/\sqrt{T}) \tag{8}$$
$$\mathbb{E}[\|\bar{x}_T - \bar{z}_T\|] \leq \mathcal{O}(1/T) \tag{9}$$

If D and G_f are known, we can optimize the constants in (8) by choosing $\gamma_0 = D/G_f$. This gives $f(\bar{z}_T) + g(\bar{z}_T) + h(\bar{x}_T) - \phi_\star \leq \mathcal{O}(DG_f/\sqrt{T})$ and $\|\bar{x}_T - \bar{z}_T\| \leq \mathcal{O}(D/T)$.

Proof sketch. To prove the bound in (8), we start by writing optimality conditions for the proximal steps on z_t and x_t. Through algebraic modifications and by using convexity of f, g and h, we obtain

$$f(z_t) + g(z_t) + h(x_t) - \phi_\star \leq \frac{1}{2\gamma}\|y_t - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - x_\star\|^2 + \frac{\gamma}{2}\|u_t\|^2, \tag{10}$$

and $\|u_t\|$ is bounded by Assumption 1. We average this inequality over $t = 0, 1, \dots, T$ and use Jensen's inequality to get (8).

The major technical novelty is in the proof for (**??**). TOS operator in (4) with $u_t = \nabla f(z_t)$ for a L_f-smooth function f is α-averaged[1] with coefficient $\alpha = 1/(2 - \gamma L_f)$ (see Proposition 2.1 in (Davis & Yin, 2017)), and the analysis in prior work is based on this property. In particular, α-averagedness implies Fejér monotonicity, i.e., that $\|y_t - y_\star\|$ is non-increasing where $y_\star$ denotes a fixed point of TOS. This yields the convergence of $\|\bar{x}_T - \bar{z}_T\|$ since

$$\|\bar{x}_T - \bar{z}_T\| = \frac{\|y_{T+1} - y_0\|}{T+1} \leq \frac{\|y_{T+1} - y_\star\| + \|y_\star - y_0\|}{T+1} \leq \frac{2\|y_\star - y_0\|}{T+1}. \tag{11}$$

However, when f is nonsmooth and u_t is replaced with a subgradient, TOS operator in (4) is no longer α-averaged, and the bound in (11) along with the standard analysis fail. Our key observation which unlocks the new results is that $\|y_t - y_\star\|$ remains bounded even-though we loose α-averagedness and Fejér monotonicity in this setting. In particular, we show that (see Theorem S.6 in the supplements)

$$\|y_{T+1} - y_\star\| \leq \|y_0 - y_\star\| + 2\gamma_0 G_f, \tag{12}$$

which gives (**??**), similarly to (11). □

Theorem 1 does not immediately yield convergence to a solution of Problem (1) because $f + g$ and h are evaluated at different points in (8). Next corollary solves this issue.

Corollary 1. *We are interested in two particular cases of Theorem 1:*

(i). *Suppose h is G_h-Lipschitz continuous. Then,*

$$\phi(\bar{z}_T) - \phi_\star \leq \mathcal{O}(1/\sqrt{T}) \tag{13}$$

(ii). *Suppose h is the indicator function of a convex set $\mathcal{H} \subseteq \mathbb{R}^n$. Then,*

$$f(\bar{z}_T) + g(\bar{z}_T) - \phi_\star \leq \mathcal{O}(1/\sqrt{T}) \tag{14}$$
$$\mathrm{dist}(\bar{z}_T, \mathcal{H}) \leq \mathcal{O}(1/T) \tag{15}$$

[1] An operator $\mathtt{T} : \mathbb{R}^n \to \mathbb{R}^n$ is α-averaged if $\|\mathtt{T}x - \mathtt{T}y\|^2 \leq \|x - y^2\| - \frac{1-\alpha}{\alpha}\|(x - \mathtt{T}x) - (y - \mathtt{T}y)\|^2$ for some $\alpha \in (0, 1)$ for all $x, y \in \mathbb{R}^n$.

Proof. (i). Since h is G_h-Lipschitz, we know $\phi(\bar{z}_T) \le f(\bar{z}_T) + g(\bar{z}_T) + h(\bar{x}_T) + G_h\|\bar{x}_T - \bar{z}_T\|$.
(ii). $h(\bar{x}_T) = 0$ since $\bar{x}_T \in \mathcal{H}$. Moreover, $\mathrm{dist}(\bar{z}_T, \mathcal{H}) := \inf_{x\in\mathcal{H}} \|\bar{z}_T - x\| \le \|\bar{z}_T - \bar{x}_T\|$. □

Remark 1. *We fix horizon T for the ease of analysis and presentation. Similar guarantees (with extra logarithmic factors) hold for all iterations if we use $\gamma_t = \gamma_0/\sqrt{t+1}$.*

Theorem 1 covers the case in which g is the indicator of a convex set $\mathcal{G} \subseteq \mathbb{R}^n$. By definition, $\bar{z}_T \in \mathcal{G}$ and $x_\star \in \mathcal{G}$, hence $g(\bar{z}_T) = g(x_\star) = 0$. If both g and h are indicator functions, TOS gives an approximately feasible solution in $\mathcal{G}$ and close to $\mathcal{H}$. We could also consider a stronger notion of approximate feasibility, measured by $\mathrm{dist}(\bar{z}_T, \mathcal{G} \cap \mathcal{H})$. However, this requires additional regularity assumptions on $\mathcal{G}$ and $\mathcal{H}$ to avoid pathological examples, see Lemma 1 in (Hoffmann, 1992) and Definition 2 in (Kundu et al., 2018).

Problem (1) captures unconstrained minimization problems when $g = h = 0$. Therefore, the convergence rate in (1) is optimal in the sense it matches the information theoretical lower bounds for the first-order black-box methods, see Section 3.2.1 in (Nesterov, 2003). Remark that the subgradient method can achieve a faster $\mathcal{O}(1/t)$ convergence when f is strongly convex. We leave the convergence rate analysis of TOS for strongly convex f as an open problem.

4 TOS for Stochastic Setting

In this section, we focus on the three-composite stochastic optimization template:

$$\min_{x\in\mathbb{R}^n} \quad \phi(x) := f(x) + g(x) + h(x) \quad \text{where} \quad f(x) := \mathbb{E}_\xi \tilde{f}(x,\xi) \tag{16}$$

and ξ is a random variable. We suppose the following standard assumption holds.

Assumption 2 (Bounded variance)**.** *There exists a stochastic first-order oracle that, upon query at $x \in \mathbb{R}^n$, returns an unbiased estimate $u \in \mathbb{R}^n$ of the (sub)gradient such that*

$$\mathbb{E}_\xi[u|x] = \nabla f(x) \quad \text{and} \quad \mathbb{E}_\xi[\|u - \nabla f(x)\|^2] \le \sigma^2 \ \text{for some } \sigma > 0. \tag{17}$$

The following theorem characterizes the convergence rate of Algorithm 2 for Problem (16).

Theorem 2. *Consider Problem (16) and employ TOS (Algorithm 2) with a fixed step-size $\gamma_t = \gamma = \gamma_0/\sqrt{T+1}$ for some $\gamma_0 > 0$. Let Assumption 1 hold and suppose that we are receiving the update direction u_t from a stochastic first-order oracle that satisfies Assumption 2. Then,*

$$\mathbb{E}[f(\bar{z}_T) + g(\bar{z}_T) + h(\bar{x}_T)] - \phi_\star \le \frac{1}{\sqrt{T+1}}\left(\frac{1}{2\gamma_0}D^2 + \gamma_0(\sigma^2 + G_f^2)\right) \tag{18}$$

$$\text{and} \quad \mathbb{E}[\|\bar{x}_T - \bar{z}_T\|] \le \frac{2}{T+1}\left(D + \gamma_0\left(G_f + \frac{\sigma}{2}\right)\right), \quad \text{where} \quad D = \|y_0 - x_\star\|. \tag{19}$$

If we can estimate D, G_f and σ, then we can optimize the bounds by choosing $\gamma_0 \approx D/\max\{G_f, \sigma\}$. This gives

$$\mathbb{E}[f(\bar{z}_T) + g(\bar{z}_T) + h(\bar{x}_T)] - \phi_\star \le \mathcal{O}(D\max\{G_f,\sigma\}/\sqrt{T})$$
$$\mathbb{E}[\|\bar{x}_T - \bar{z}_T\|] \le \mathcal{O}(D/T)$$

Analogous to Corollary 1, we can derive convergence guarantees when h is Lipschitz continuous or an indicator function from (18) and (19). As in the nonsmooth setting, the rates shown in this section are optimal because Problem (16) covers $g(x) = h(x) = 0$ as a special case.

TOS with stochastic gradients have been already studied under various conditions. In particular, the method is known to get $\mathcal{O}(1/t)$ convergence rate when f is smooth and strongly convex (Yurtsever et al., 2016). We present a detailed discussion on the related works and the comparison in Section 2.

Remark 2. *Similar guarantees hold if we replace Assumption 1 with the smoothness of f.*

5 TOS with Adaptive Learning Rates

In this section, we focus on an important subclass of Problem (1) where g and h are indicator functions of some closed convex sets,

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{G} \cap \mathcal{H}. \tag{20}$$

TOS is effective for Problem (20) particularly when the projection onto $\mathcal{G}$ and $\mathcal{H}$ are easy, but the projection onto their intersection is challenging. Particular examples include the transportation polytopes, doubly nonnegative matrices, and isotonic regression, among many others.

$$u_t = \partial f(z_t), \gamma_t = \frac{\alpha}{\sqrt{\beta + \sum_{\tau=0}^{t-1} \|u_\tau\|^2}}$$

We propose ADAPTOS, a new variant of TOS with adaptive step-size in the spirit of adaptive online learning algorithms and online to batch conversion techniques, see (Duchi et al., 2011; Rakhlin & Sridharan, 2013; Levy, 2017; Levy et al., 2018; Cutkosky, 2019; Kavis et al., 2019) and the references therein. ADAPTOS employs the following step-size rule:

$$\gamma_t = \frac{\gamma_0}{\sqrt{\sum_{\tau=0}^{t} \|u_\tau\|^2}} \quad \text{for some } \gamma_0 > 0. \tag{21}$$

We can employ (21) in Algorithm 2 only for Problem (20) because $z_t = \mathrm{prox}_{\gamma_t g}(y_t) = \mathrm{proj}_{\mathcal{G}}(y_t)$ is the same for any $\gamma_t > 0$. In Section 6, we empirically test ADAPTOS on generic instances of Problem (1). If g is not an indicator function, we use the step-size from the previous iteration (*i.e.*, γ_{t-1}) in the z_t update.

For ADAPTOS, we will also use a second ergodic sequence, with weighted averages as

$$\tilde{x}_t = \frac{1}{\sum_{\tau=0}^{t} \gamma_\tau} \sum_{\tau=0}^{t} \gamma_\tau x_\tau \qquad \text{and} \qquad \tilde{z}_t = \frac{1}{\sum_{\tau=0}^{t} \gamma_\tau} \sum_{\tau=0}^{t} \gamma_\tau z_\tau \tag{22}$$

for proving convergence in function value. This sequence was also considered for TOS with line-search in (Pedregosa & Gidel, 2018).

Theorem 3. *Consider Problem (20) and let Assumption 1 hold. Employ TOS (Algorithm 2) with the update direction $u_t \in \partial f(z_t)$ and adaptive step-size (21). Then, the following bounds hold:*

$$f(\tilde{z}_t) - f_\star \leq \mathcal{O}(1/\sqrt{t}) \tag{23}$$

$$\mathrm{dist}(\tilde{z}_t, \mathcal{H}) \leq \mathcal{O}(1/\sqrt{t}) \tag{24}$$

$$f(\tilde{z}_t) - f_\star \leq \frac{G_f}{2\sqrt{t+1}} \left(\frac{D^2}{\gamma_0} + \gamma_0 \log(G_f^2(t+1)) \right) \quad \textit{and} \tag{25}$$

$$\mathrm{dist}(\bar{z}_t, \mathcal{H}) \leq \frac{2D}{t+1} + \frac{\gamma_0 \sqrt{\log(G_f^2(t+1))}}{\sqrt{t+1}}, \quad \textit{where} \quad D = \|y_0 - y_\star\|. \tag{26}$$

We recommend choosing $\gamma_0 = D$ if D is known. This gives $f(\tilde{z}_t) - f_\star \leq \tilde{\mathcal{O}}(DG_f/\sqrt{t})$ and $\mathrm{dist}(\bar{z}_t, \mathcal{H}) \leq \tilde{\mathcal{O}}(D/\sqrt{t})$.

The next theorem establishes a faster rate for the same algorithm when f is smooth and a solution lies in the interior of the feasible set.

Theorem 4. *Consider Problem (20) and let Assumption 1 hold. Suppose f is differentiable with L_f-Lipschitz continuous gradients on $\mathcal{G}$ and assume that there exists a solution in the interior of the feasible set. Employ TOS (Algorithm 2) with update direction $u_t = \nabla f(z_t)$ and adaptive step-size (21). Then,*

$$f(\tilde{z}_t) - f_\star \leq \mathcal{O}(1/t)$$
$$\mathrm{dist}(\tilde{z}_t, \mathcal{H}) \leq \mathcal{O}(1/t)$$

If $\gamma_0 = D$ is chosen, we get $f(\bar{z}_t) - f_\star \leq \tilde{\mathcal{O}}(L_f D^2/t)$ and $\mathrm{dist}(\bar{z}_t, \mathcal{H}) \leq \tilde{\mathcal{O}}(D/t)$.

The limitation on the location of the solution is inherit from the prior work (Levy, 2017; Levy et al., 2018), and it creates a challenge for extending these results for the regularized problems. We believe this to be a limitation of the analysis, and the method can achieve the fast rates when f is smooth regardless of where the solution lies. Numerical experiments support this claim.

Finally, the next theorem shows that ADAPTOS can successfully handle stochastic (sub)gradients.

Theorem 5. *Consider Problem (20) and let Assumption 1 hold. Employ TOS (Algorithm 2) with adaptive step-size (21). Suppose we are receiving the update direction u_t from a stochastic first-order oracle that satisfies Assumption 2. Then, the following bounds hold:*

$$\mathbb{E}[f(\tilde{z}_t) - f_\star] \leq \mathcal{O}(1/\sqrt{t})$$
$$\mathbb{E}[\mathrm{dist}(\bar{z}_t, \mathcal{H})] \leq \mathcal{O}(1/\sqrt{t})$$

Following the definition in (Nesterov, 2015), we say that an algorithm is universal if it does not require to know whether the objective is smooth or not yet it implicitly adapts to the smoothness of the objective. ADAPTOS attains universal convergence rates for Problem (20). It converges to a solution with $\tilde{\mathcal{O}}(1/\sqrt{t})$ rate (in function value) when f is nonsmooth or stochastic. The rate becomes $\tilde{\mathcal{O}}(1/t)$ if f is smooth and the solution is in the interior of the feasible set.

6 Numerical Experiments

This section demonstrates the empirical performance of the proposed method on a number of smooth and nonsmooth convex optimization templates. We also present an experiment on neural networks. Our experiments are performed in Python 3.7 with Intel Core i9-9820X CPU @ 3.30GHz. The source code is available in the supplemental material.

6.1 Experiments on Convex Optimization with Smooth f

In this subsection, we compare ADAPTOS with TOS, PDHG and their line-search variants TOS-LS and PDHG-LS. Our experiments are based on the benchmarks described in (Pedregosa & Gidel, 2018) and the source code available in COPT Library (Pedregosa et al., 2020) under the new BSD License. We implement ADAPTOS and investigate its performance on three different problems:

▷ Logistic regression with *overlapping group lasso* penalty:

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{N}\sum_{i=1}^{N} \log(1 + \exp(-b_i\langle a_i, x\rangle)) + \lambda \sum_{G \in \mathcal{G}} \sqrt{|G|}\|x_G\| + \lambda \sum_{H \in \mathcal{H}} \sqrt{|H|}\|x_H\|, \tag{27}$$

where $\{(a_1, b_1), \ldots, (a_N, b_N)\}$ is a given set of training examples, $\mathcal{G}$ and $\mathcal{H}$ are the sets of distinct groups and $|\cdot|$ denotes the cardinality. The model we use (from COPT) considers groups of size 10 with 2 overlapping coefficients. In this experiment, we use the benchmarks on synthetic data (dimensions $n = 1002$, $N = 100$) and RCV1 dataset (Lewis et al., 2004) ($n = 677399$, $N = 20242$).

▷ Image recovery with *total variation* penalty:

$$\min_{X \in \mathbb{R}^{m \times n}} \quad \|Y - \mathcal{A}(X)\|_F^2 + \lambda \sum_{i=1}^{m}\sum_{j=1}^{n-1} |X_{i,j+1} - X_{i,j}| + \lambda \sum_{j=1}^{n}\sum_{i=1}^{m-1} |X_{i+1,j} - X_{i,j}|, \tag{28}$$

where Y is a given blurred image and $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ is a linear operator (blur kernel). The benchmark in COPT solves this problem for an image of size 153×115 with a provided blur kernel.

▷ Sparse and low-rank matrix recovery via ℓ_1 *and nuclear-norm* regularizations:

$$\min_{X \in \mathbb{R}^{n \times n}} \quad \frac{1}{N}\sum_{i=1}^{N} \texttt{huber}(b_i - \langle A_i, X\rangle) + \lambda\|X\|_* + \lambda\|X\|_1. \tag{29}$$

We use huber loss. $\{(A_1, b_1), \ldots, (A_N, b_N)\}$ is a given set of measurements and $\|X\|_1$ is the vector ℓ_1-norm of X. The benchmark in COPT considers a symmetric ground truth matrix $X^\natural \in \mathbb{R}^{20 \times 20}$

and noisy synthetic measurements ($N = 100$) where A_i has Gaussian iid entries. $b_i = \langle A_x, X\rangle + \omega_i$, where ω_i is generated from a zero-mean unit variance Gaussian distribution.

At each problem, we consider two different values for the regularization parameter λ. We use all methods with their default parameters in the benchmark. For ADAPTOS, we tune γ_0 by trying the powers of 10, see the supplement for the values found. Figure 1 shows the results of this experiment. In most cases, the performance of ADAPTOS lies in between TOS-LS and PDHG-LS. Remark that TOS-LS is using the extra knowledge of the Lipschitz constant of h.

6.2 Experiments on Convex Optimization with Nonsmooth f

We examine the empirical performance of ADAPTOS for nonsmooth problems on an image impainting and denoising task from (Zeng & So, 2018; Yurtsever et al., 2018). We are given an occluded image (i.e., missing some pixels) of size 517×493, contaminated with salt and pepper noise of 10% density. We use the following template where data fitting is measured in terms of vector ℓ_p-norm:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} \quad \|\mathcal{A}(X) - Y\|_p \quad \text{subject to} \quad \|X\|_* \leq \lambda, \quad 0 \leq X \leq 1, \tag{30}$$

where Y is the observed noisy image with missing pixels. This is essentially a matrix completion problem, $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is a linear map that samples the observed pixels in Y. In particular, we consider (30) with $p = 1$ and $p = 2$. The ℓ_2-loss is common in practice for matrix completion (often in the least-squares form) but it is not very robust against the outliers induced by salt and pepper noise. ℓ_1-loss is known to be more reliable for this task.

The subgradients in both cases have a fixed norm at all points (note that the subgradients are binary valued for ℓ_1-loss, and unit-norm for ℓ_2-loss), hence the analytical ($\gamma_t = \gamma_0/\sqrt{t+1}$) and adaptive (21) step-sizes are the same up to a constant factor. We present ADAPTOS only.

Figure 2 shows the results. The empirical rates for $p = 1$ roughly match our guarantees in Theorem 1. We observe a faster locally linear convergence rate when ℓ_2-loss is used. Interestingly, the ergodic sequence converges faster than the last iterate on $p = 1$ but significantly slower in $p = 2$. The runtime of the two settings are approximately the same, with 81.4 (for ℓ_1) versus 80.1 (for ℓ_2) msec per iteration on average. Despite the slower rates, we found ℓ_1-loss more practical on this problem. A low-accuracy solution obtained by 1000 iterations (approximately 80 sec) on ℓ_1-loss yields a high quality recovery with PSNR 26.19 dB, whereas the PSNR saturates at 21.15 dB for the ℓ_2-formulation. See the supplements for the visuals.

6.3 An Experiment on Neural Networks

In this section, we train a regularized deep neural network to test our methods on nonconvex optimization. We consider a regularized neural network problem formulation in (Scardapane et al., 2017). This problem involves a fully connected neural network with the standard cross-entropy loss function, a ReLu activation for the hidden layers, and the softmax activation for the output layer. Two regularizers are added to this loss function: The first one is the standard ℓ_1 regularizer, and the second is the group sparse regularizer where the outgoing connections of each neuron is considered as a group. The goal is to force all outgoing connections from the same neurons to be simultaneously zero, so that we can safely remove the neurons from the network. This is shown as an effective way to obtain compact networks (Scardapane et al., 2017), which is crucial for the deployment of the learned parameters on resource-constrained devices such as smartphones (Blalock et al., 2020).

We reuse the open source implementation (built with Lasagne framework based on Theano) published in (Scardapane et al., 2017) under BSD-2 License. We strictly follow their experimental setup and instructions, with the MNIST database (LeCun, 1998) containing 70k grayscale images (28×28) of handwritten digits (split 75/25 into train and test partitions). We train a fully connected neural network with 784 input features, three hidden layers ($400/300/100$) and 10-dimensional output layer. Interested readers can find more details on the implementation in the supplementary material or in (Scardapane et al., 2017).

Scardapane et al. (2017) use SGD and Adam with the subgradient of the overall objective. In contrast, our methods can leverage the prox-operators of the regularizers. Figure 3 compares the performance in two terms: the sparsity of the parameters and the accuracy. On the left side, we see the spectrum of weight and neuron magnitudes. The advantage of using prox-operators is outstanding: More

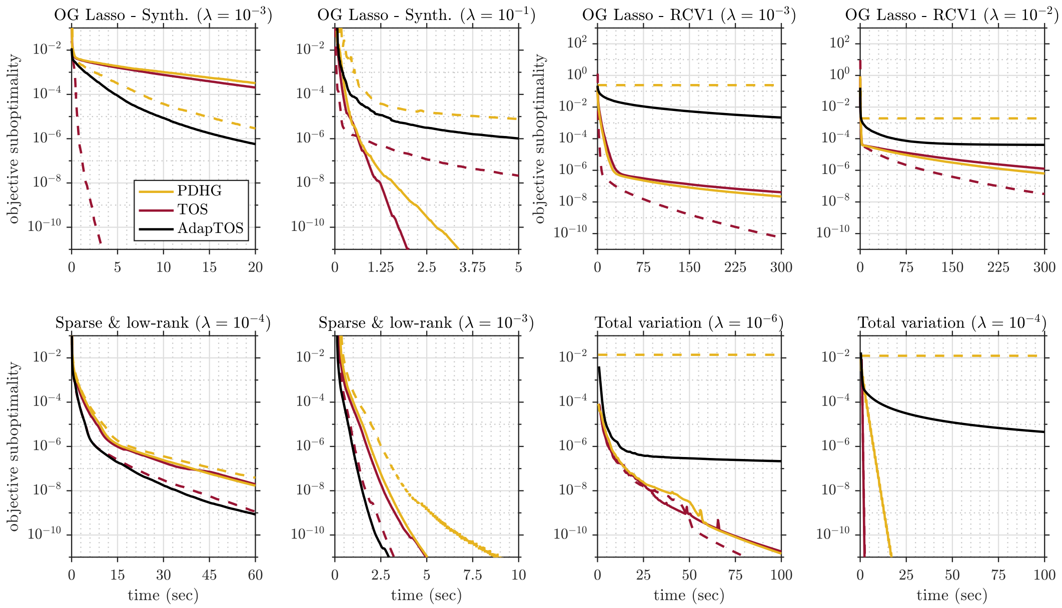

Figure 1: Empirical comparison of 5 algorithms for Problem (1) with smooth f. Dashed lines represent the line-search variants of TOS and PDHG. We observed the performance of AdapTos lies between TOS-LS and PDHG-LS in all cases. TOS and PDHG require the knowledge of the smoothness constant, and TOS-LS uses the Lipschitz constant for one of the nonsmooth terms.

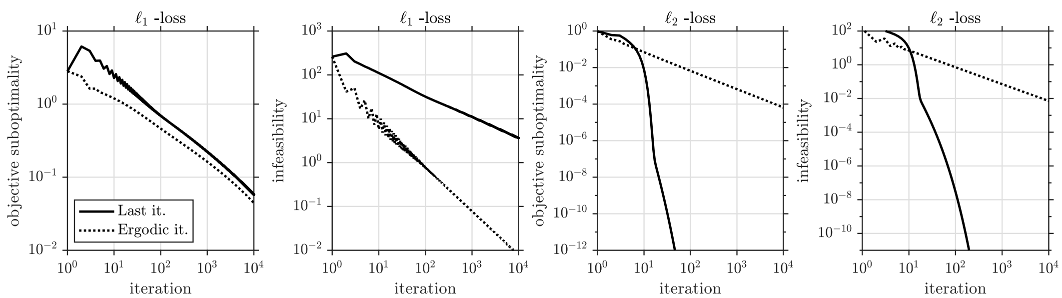

Figure 2: Performance of AdapTos on image impainting and denoising problems with ℓ_1 and ℓ_2-loss functions (mind the difference in the axes limits). The convergence rates on the ℓ_1 norm match the guaranteed rates, $\mathcal{O}(1/\sqrt{t})$ in objective suboptimality and $\mathcal{O}(1/t)$ in infeasibility. The method achieves a locally linear convergence rate on the ℓ_2-loss.

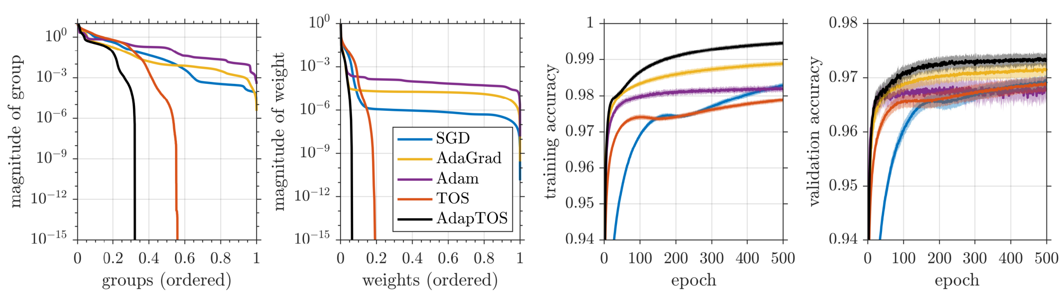

Figure 3: Comparison of methods on training neural networks with group lasso regularization. The outgoing connections of each neuron form a group. The first plot shows the magnitude of weights after 500 epochs. The second plot shows the absolute sum of outgoing weights from each neuron. x-axes are normalized by the total number of weights and neurons in these plots. More than 68% of the neurons are inactive on the network trained by AdapTos. The third and the fourth plots show the training and validation loss. This experiment is performed with 20 random seeds. The solid lines show the average performance and the shaded area represents ± standard deviation from the mean.

than 93% of the weights are zero and 68% of neurons are inactive when trained with AdapTos. In contrast, subgradient based methods can achieve only approximately sparse solutions.

The third and the fourth subplots present the training and test accuracy. Remarkably, AdapTos performs better than the state-of-the-art (both in train and test). Unfortunately, we could not achieve the same performance gain in preliminary experiments with more complex models like ResNet (He et al., 2016), where SGD with momentum shines. Interested readers can find the code for these preliminary experiments in the supplements. We leave the technical analysis and a comprehensive examination of AdapTos for nonconvex problems to a future work.

7 Conclusions

We studied an extension of TOS that permits subgradients and stochastic gradients instead of the gradient step, and we established the convergence guarantees of this method. Moreover, we proposed an adaptive step-size rule (AdapTos) for the minimization of a convex function over the intersection of two convex sets. AdapTos guarantees a nearly optimal $\tilde{\mathcal{O}}(1/\sqrt{t})$ rate on the baseline setting, and it enjoys the faster $\tilde{\mathcal{O}}(1/t)$ rate when the problem is smooth and the solution is in the relative interior of the domain. We present numerical experiments on various benchmark problems. The empirical performance of the method is promising.

We conclude with a short list of open questions and follow-up directions: (i) In parallel to the subgradient method, we believe TOS can achieve $\mathcal{O}(1/t)$ rate guarantees in the nonsmooth setting if f is strongly convex. The analysis remains open. (ii) The faster rate for AdapTos on smooth f requires an extra assumption on the location of the solution. We believe this assumption can be removed, and leave this as an open problem. (iii) We analyzed AdapTos only for a specific subclass of Problem (1) in which g and h are indicator functions. Extending this result for the whole class is a valuable question for future study.

Acknowledgements

Alp Yurtsever received support from the Swiss National Science Foundation Early Postdoc.Mobility Fellowship P2ELP2_187955, from the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and partial postdoctoral support from the NSF-CAREER grant IIS-1846088. Suvrit Sra acknowledges support from an NSF BIGDATA grant (1741341) and an NSF CAREER grant (1846088).

References

Azadi, S. and Sra, S. Towards an optimal stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pp. 620–628. PMLR, 2014.

Barbero, A. and Sra, S. Modular proximal optimization for multidimensional total-variation regularization. *The Journal of Machine Learning Research*, 19(1):2232–2313, 2018.

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Blalock, D., Ortiz, J. J. G., Frankle, J., and Guttag, J. What is the state of neural network pruning? *arXiv preprint arXiv:2003.03033*, 2020.

Briceño-Arias, L. M. Forward-douglas–rachford splitting and forward-partial inverse method for solving monotone inclusions. *Optimization*, 64(5):1239–1261, 2015.

Cevher, V., Vũ, B. C., and Yurtsever, A. Stochastic forward douglas-rachford splitting method for monotone inclusions. In *Large-Scale and Distributed Optimization*, pp. 149–179. Springer, 2018.

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

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

Cutkosky, A. Anytime online-to-batch, optimism and acceleration. In *International Conference on Machine Learning*, pp. 1446–1454. PMLR, 2019.

Davis, D. and Yin, W. A three-operator splitting scheme and its optimization applications. *arXiv preprint arXiv:1504.01032*, 2015.

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

Ding, L., Yurtsever, A., Cevher, V., Tropp, J. A., and Udell, M. An optimal-storage approach to semidefinite programming using approximate complementarity. *arXiv preprint arXiv:1902.03373*, 2019.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

El Halabi, M. and Cevher, V. A totally unimodular view of structured sparsity. In *Artificial Intelligence and Statistics*, pp. 223–231. PMLR, 2015.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Higham, N. J. and Strabić, N. Anderson acceleration of the alternating projections method for computing the nearest correlation matrix. *Numerical Algorithms*, 72(4):1021–1042, 2016.

Hoffmann, A. The distance to the intersection of two convex sets expressed by the distances to each of them. *Mathematische Nachrichten*, 157(1):81–98, 1992.

Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Kundu, A., Bach, F., and Bhattacharya, C. Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 958–967. PMLR, 2018.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Levy, K. Y. Online to offline conversions, universality and adaptive minibatch sizes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

Lewis, D. D., Yang, Y., Russell-Rose, T., and Li, F. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Malitsky, Y. and Pock, T. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432, 2018.

McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. *23rd Annual Conference on Learning Theory*, 2010.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

Ouyang, H., He, N., Tran, L., and Gray, A. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pp. 80–88. PMLR, 2013.

Pedregosa, F. On the convergence rate of the three operator splitting scheme. *arXiv preprint arXiv:1610.07830*, 2016.

Pedregosa, F. and Gidel, G. Adaptive three operator splitting. In *International Conference on Machine Learning*, pp. 4085–4094, 2018.

Pedregosa, F., Fatras, K., and Casotto, M. Proximal splitting meets variance reduction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1–10. PMLR, 2019.

Pedregosa, F., Negiar, G., and Dresdner, G. copt: composite optimization in python. 2020. doi: 10.5281/zenodo.1283339. URL http://openo.pt/copt/.

Raguet, H., Fadili, J., and Peyré, G. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

Rakhlin, A. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pp. 3066–3074, 2013.

Scardapane, S., Comminiello, D., Hussain, A., and Uncini, A. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

Shivanna, R., Chatterjee, B., Sankaran, R., Bhattacharyya, C., and Bach, F. Spectral norm regularization of orthonormal representations for graph transduction. In *Neural Information Processing Systems*, 2015.

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. Nearly-isotonic regression. *Technometrics*, 53(1): 54–61, 2011.

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

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

Yuan, L., Liu, J., and Ye, J. Efficient methods for overlapping group lasso. *Advances in neural information processing systems*, 24:352–360, 2011.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Yurtsever, A., Vũ, B. C., and Cevher, V. Stochastic three-composite convex minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4329–4337, 2016.

Yurtsever, A., Fercoq, O., Locatello, F., and Cevher, V. A conditional gradient framework for composite convex minimization with applications to semidefinite programming. In *International Conference on Machine Learning*, pp. 5727–5736. PMLR, 2018.

Yurtsever, A., Mangalick, V., and Sra, S. Three operator splitting with a nonconvex loss function. In *International Conference on Machine Learning*, pp. 12267–12277. PMLR, 2021.

Zeng, W.-J. and So, H. C. Outlier–robust matrix completion via ℓ_p-minimization. *IEEE Trans. on Sig. Process*, 66(5):1125–1140, 2018.

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

Zhao, R., Haskell, W. B., and Tan, V. Y. An optimal algorithm for stochastic three-composite optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 428–437. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We explicitly state limitations of our methods and its analysis in multiple places in the paper.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We present a new convex optimization technique and its technical analysis, we do not expect any direct societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We listed our assumptions in the beginning of each Theorem.
   (b) Did you include complete proofs of all theoretical results? [Yes] In the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Some details are deferred to the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report error bars for our experiments on neural networks. Or other experiments are on convex and deterministic problems which do not involve any randomness.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the beginning of Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We implemented the proposed methods for the openopt/copt Python library.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We used standard benchmark datasets with citation or synthetic data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The datasets we use do not contain any personal information to the best of our knowledge.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Three Operator Splitting with Subgradients, Stochastic Gradients, and Adaptive Learning Rates

Supplementary Material

A Preliminaries

We will use the following standard results in our analysis.

Lemma S.2 (Prox-theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper closed and convex function. Then, for any $x, u \in \mathbb{R}^n$, the followings are equivalent:*

(i) $u = \mathrm{prox}_f(x)$.

(ii) $x - u \in \partial f(u)$.

(iii) $\langle x - u, y - u\rangle \leq f(y) - f(u)$ *for any* $y \in \mathbb{R}^n$.

Corollary 2 (Firm non-expansivity of the prox-operator)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper closed and convex function. Then, for any $x, u \in \mathbb{R}^n$, the followings hold:*

$$\text{(non-expansivity)} \qquad \|\mathrm{prox}_f(x) - \mathrm{prox}_f(y)\| \leq \|x - y\|.$$

$$\text{(firm non-expansivity)} \qquad \|\mathrm{prox}_f(x) - \mathrm{prox}_f(y)\|^2 \leq \langle x - y, \mathrm{prox}_f(x) - \mathrm{prox}_f(y)\rangle.$$

Lemma S.3 (Lemma 7 in (McMahan & Streeter, 2010))**.** *For any non-negative numbers $a_1, \ldots, a_t$,*

$$\sum_{i=1}^{t} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{t} a_i}.$$

Lemma S.4 (Lemma A.3 in (Levy et al., 2018))**.** *For any non-negative numbers $a_1, \ldots, a_t$,*

$$\sum_{i=1}^{t} \frac{a_i}{1 + \sum_{j=1}^{i} a_j} \leq 1 + \log\left(1 + \sum_{i=1}^{t} a_i\right).$$

Lemma S.5 (Lemma A.2 in (Levy, 2017))**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L_f-smooth function and let $x_\star \in \arg\min_{x \in \mathbb{R}^n} f(x)$. Then,*

$$\|\nabla f(x)\|^2 \leq 2L_f\big(f(x) - f(x_\star)\big), \qquad \forall x \in \mathbb{R}^n.$$

B Fixed Point Characterization

This appendix presents the proof for Lemma 1. This is a straightforward extension of Lemma 2.2 in (Davis & Yin, 2017) to permit subgradients. We will use this lemma in the next section to prove the boundedness of y_t in Algorithm 2.

B.1 Proof of Lemma 1

Define $z = \mathrm{prox}_{\gamma g}(y)$ and $x = \mathrm{prox}_{\gamma h}(2z - y - \gamma u)$. Then, $\mathtt{TOS}_\gamma(y, u) := y - z + x$.

Suppose there exists $u \in \partial f(z)$ such that $\mathtt{TOS}(y, u) = y$. Then, we must have $z = x$. Moreover, by Lemma S.2, we have

$$z = \mathrm{prox}_{\gamma g}(y) \iff y - z \in \gamma \partial g(z), \tag{S.31}$$

$$\text{and} \quad z = x = \mathrm{prox}_{\gamma h}(2z - y - \gamma u) \iff z - y - \gamma u \in \gamma \partial h(x). \tag{S.32}$$

By summing up (S.31) and (S.32), we observe

$$0 \in \gamma(u + \partial g(z) + \partial h(x)) \implies 0 \in \partial f(z) + \partial g(z) + \partial h(z) = \partial \phi(z), \tag{S.33}$$

which proves that z is an optimal solution of Problem (1) since ϕ is convex.

To prove the reverse direction, suppose z is an optimal solution, i.e., there exists $u \in \partial f(z), v \in \partial g(z), w \in \partial h(z)$ such that $u + v + w = 0$. By Lemma S.2, we have

$$z = \text{prox}_{\gamma g}(y) \iff y - z \in \gamma \partial g(z), \tag{S.34}$$

$$\text{and} \quad x = \text{prox}_{\gamma h}(2z - y - \gamma u) \iff 2z - x - y - \gamma u \in \gamma \partial h(x). \tag{S.35}$$

Now, let $y = z + \gamma v$. Then,

$$2z - x - y - \gamma u = z - x - \gamma(u + v) = z - x + \gamma w. \tag{S.36}$$

Therefore, we have $z - x + \gamma w \in \partial h(x)$. Again, due to Lemma S.2, this means $x = \text{prox}_{\gamma h}(z + \gamma w)$. We also know $w \in \partial h(z) \iff z + \gamma w - z \in \partial \gamma h(z) \iff z = \text{prox}_{\gamma h}(z + \gamma w)$. However, since h is convex, its prox-operator is unique, hence, $x = z$ and $\texttt{TOS}_\gamma(y, u) = y$.

C Boundedness Guarantees

Theorem S.6. *Consider Problem (16) and employ TOS (Algorithm 2) for T iterations with a fixed step-size $\gamma_t = \gamma = \gamma_0/\sqrt{T+1}$ for all $0 \leq t \leq T$ and some $\gamma_0 > 0$. Choose the update directions $u_t \in \partial f(z_t)$. Then, we have*

$$\|y_{T+1} - y_\star\| \leq \|y_0 - y_\star\| + 2\gamma_0 G_f, \tag{S.37}$$

where $y_\star$ is a fixed point of TOS.

Proof. By Lemma 1, there exists $u_\star \in \partial f(x_\star)$ such that

$$x_\star = \text{prox}_{\gamma g}(y_\star) = \text{prox}_{\gamma h}(2x_\star - y_\star - \gamma u_\star) = z_\star. \tag{S.38}$$

We decompose $\|y_{t+1} - y_\star\|^2$ as

$$\begin{aligned} \|y_{t+1} - y_\star\|^2 &= \|y_t - z_t + x_t - y_\star + x_\star - x_\star\|^2 \\ &= \|y_t - z_t - y_\star + x_\star\|^2 + \|x_t - x_\star\|^2 + 2\langle x_t - x_\star, y_t - z_t - y_\star + x_\star\rangle. \end{aligned} \tag{S.39}$$

Since $z_t = \text{prox}_{\gamma g}(y_t)$ and $x_\star = \text{prox}_{\gamma g}(y_\star)$, by the firm non-expansivity of the prox-operator, we have

$$\begin{aligned} \|y_t - z_t - y_\star + x_\star\|^2 &= \langle y_t - z_t - y_\star + x_\star, y_t - y_\star\rangle - \langle y_t - z_t - y_\star + x_\star, z_t - x_\star\rangle \\ &\leq \langle y_t - z_t - y_\star + x_\star, y_t - y_\star\rangle. \end{aligned} \tag{S.40}$$

Similarly, since $x_t = \text{prox}_{\gamma h}(2z_t - y_t - \gamma u_t)$ and $x_\star = \text{prox}_{\gamma h}(2x_\star - y_\star - \gamma u_\star)$, by the firm non-expansivity of the prox-operator, we have

$$\|x_t - x_\star\|^2 \leq \langle x_t - x_\star, (2z_t - y_t - \gamma u_t) - (2x_\star - y_\star - \gamma u_\star)\rangle. \tag{S.41}$$

By combining (S.39), (S.40) and (S.41), we get

$$\begin{aligned} \|y_{t+1} - y_\star\|^2 &\leq \langle y_t - z_t + x_t - y_\star, y_t - y_\star\rangle - \gamma\langle x_t - x_\star, u_t - u_\star\rangle \\ &= \langle y_{t+1} - y_\star, y_t - y_\star\rangle - \gamma\langle x_t - x_\star, u_t - u_\star\rangle \\ &= \frac{1}{2}\|y_{t+1} - y_\star\|^2 + \frac{1}{2}\|y_t - y_\star\|^2 - \frac{1}{2}\|y_{t+1} - y_t\|^2 - \gamma\langle x_t - x_\star, u_t - u_\star\rangle. \end{aligned} \tag{S.42}$$

Since $u_t \in \partial f(z_t)$ and $u_\star \in \partial f(x_\star)$, we have

$$\begin{aligned} -\langle x_t - x_\star, u_t - u_\star\rangle &= -\langle z_t - x_\star, u_t - u_\star\rangle - \langle x_t - z_t, u_t - u_\star\rangle \\ &\leq -\langle x_t - z_t, u_t - u_\star\rangle \\ &\leq \frac{1}{2\gamma}\|x_t - z_t\|^2 + \frac{\gamma}{2}\|u_t - u_\star\|^2. \end{aligned} \tag{S.43}$$

where we used Young's inequality at the last line. We use this inequality in (S.42) to obtain

$$\|y_{t+1} - y_\star\|^2 \leq \|y_t - y_\star\|^2 + \gamma^2\|u_t - u_\star\|^2. \tag{S.44}$$

If we sum this inequality from $t = 0$ to T, we get

$$\|y_{T+1} - y_\star\|^2 \leq \|y_0 - y_\star\|^2 + \gamma^2 \sum_{\tau=0}^{T} \|u_\tau - u_\star\|^2. \tag{S.45}$$

Finally, due to Assumption 1, we have $\|u_\tau - u_\star\| \leq 2G_f$, hence

$$\|y_{T+1} - y_\star\|^2 \leq \|y_0 - y_\star\|^2 + 4G_f^2\gamma^2(T+1) = \|y_0 - y_\star\|^2 + 4G_f^2\gamma_0^2. \tag{S.46}$$

We complete the proof by taking the square-root of both sides,

$$\|y_{T+1} - y_\star\| \leq \sqrt{\|y_0 - y_\star\|^2 + 4G_f^2\gamma_0^2} \leq \|y_0 - y_\star\| + 2G_f\gamma_0. \tag{S.47}$$

□

Theorem S.7. *Consider Problem (16) and employ TOS (Algorithm 2) with a fixed step-size $\gamma_t = \gamma = \gamma_0/\sqrt{T+1}$ for some $\gamma_0 > 0$. Let Assumption 1 hold and suppose that we are receiving the update direction u_t from a stochastic first-order oracle that satisfies Assumption 2. Then,*

$$\mathbb{E}[\|y_{T+1} - y_\star\|] \leq \|y_0 - y_\star\| + \gamma_0(2G_f + \sigma), \tag{S.48}$$

where $y_\star$ is a fixed point of TOS.

Proof. The proof is similar to the proof of Theorem S.6. We follow the same steps until (S.42). Then,

$$\begin{aligned} -\langle x_t - x_\star, u_t - u_\star\rangle &= -\langle x_t - z_t, u_t - u_\star\rangle - \langle z_t - x_\star, u_t - \nabla f(z_t)\rangle - \langle z_t - x_\star, \nabla f(z_t) - u_\star\rangle \\ &\leq -\langle x_t - z_t, u_t - u_\star\rangle - \langle z_t - x_\star, u_t - \nabla f(z_t)\rangle. \end{aligned} \tag{S.49}$$

We take the expectation of both sides and get

$$\begin{aligned} -\mathbb{E}[\langle x_t - x_\star, u_t - u_\star\rangle] &\leq -\mathbb{E}[\langle x_t - z_t, u_t - u_\star\rangle] \\ &\leq \frac{1}{2\gamma}\mathbb{E}[\|x_t - z_t\|^2] + \frac{\gamma}{2}\mathbb{E}[\|u_\star - u_t\|^2] \\ &= \frac{1}{2\gamma}\mathbb{E}[\|x_t - z_t\|^2] + \frac{\gamma}{2}\mathbb{E}[\|u_\star - \nabla f(z_t)\|^2] + \frac{\gamma}{2}\mathbb{E}[\|\nabla f(z_t) - u_t\|^2] \\ &\leq \frac{1}{2\gamma}\mathbb{E}[\|x_t - z_t\|^2] + 2\gamma G_f^2 + \frac{\gamma}{2}\sigma^2, \end{aligned} \tag{S.50}$$

where the last line holds due to Assumptions 1 and 2.

Now, we take the expectation of (S.42) and substitute (S.50) into it. This gives

$$\mathbb{E}[\|y_{t+1} - y_\star\|^2] \leq \mathbb{E}[\|y_t - y_\star\|^2] + \gamma^2(4G_f^2 + \sigma^2). \tag{S.51}$$

Finally, we sum this inequality over $t = 0$ to T,

$$\mathbb{E}[\|y_{T+1} - y_\star\|^2] \leq \|y_0 - y_\star\|^2 + \gamma_0^2(4G_f^2 + \sigma^2). \tag{S.52}$$

Remark that $\mathbb{E}[\|y_{T+1} - y_\star\|]^2 \leq \mathbb{E}[\|y_{T+1} - y_\star\|^2]$. We finish the proof by taking the square-root of both sides. □

Next, we assume that f is L_f-smooth instead of Lipschitz continuity.

Theorem S.8. *Consider Problem (16) and let f be L_f-smooth. Employ TOS (Algorithm 2) with a fixed step-size $\gamma_t = \gamma = \gamma_0/\sqrt{T+1}$ for some $\gamma_0 \in [0, \frac{1}{L_f}]$ and suppose that we are receiving the update direction u_t from a stochastic first-order oracle that satisfies Assumption 2. Then,*

$$\mathbb{E}[\|y_{T+1} - y_\star\|] \leq \|y_0 - y_\star\| + \gamma_0(2G_f + \sigma), \tag{S.53}$$

where $y_\star$ is a fixed point of TOS.

Proof. The proof is similar to the proof of Theorem S.7. We take the expectation of (S.49) and get

$$\begin{aligned}
-\mathbb{E}[\langle x_t - x_\star, u_t - u_\star\rangle] &= -\mathbb{E}[\langle x_t - z_t, u_t - u_\star\rangle] - \mathbb{E}[\langle z_t - x_\star, u_t - \nabla f(z_t)\rangle] - \mathbb{E}[\langle z_t - x_\star, \nabla f(z_t) - u_\star\rangle] \\
&= -\mathbb{E}[\langle x_t - z_t, u_t - \nabla f(z_t)\rangle] - \mathbb{E}[\langle x_t - z_t, \nabla f(z_t) - u_\star\rangle] - \mathbb{E}[\langle z_t - x_\star, \nabla f(z_t) - u_\star\rangle] \\
&\leq \frac{\alpha+\beta}{2}\mathbb{E}[\|x_t - z_t\|^2] + \frac{\sigma^2}{2\alpha} + \left(\frac{1}{2\beta} - \frac{1}{L_f}\right)\|\nabla f(z_t) - u_\star\|^2, \qquad \text{(S.54)}
\end{aligned}$$

for any $\alpha > 0$ and $\beta > 0$, where we used the smoothness of f and Assumption 2 in the last line. If we choose $\beta = L_f$ and $\alpha = L_f(\sqrt{T+1}-1)$, then we get

$$\begin{aligned}
-\mathbb{E}[\langle x_t - x_\star, u_t - u_\star\rangle] &\leq \frac{L_f\sqrt{T+1}}{2}\mathbb{E}[\|x_t - z_t\|^2] + \frac{\sigma^2}{2L_f(\sqrt{T+1}-1)} \\
&\leq \frac{L_f\sqrt{T+1}}{2}\mathbb{E}[\|x_t - z_t\|^2] + \frac{\sigma^2}{L_f\sqrt{T+1}}. \qquad \text{(S.55)}
\end{aligned}$$

We take the expectation of (S.42) and substitute (S.55) into it to get

$$\begin{aligned}
\mathbb{E}[\|y_{t+1} - y_\star\|^2] &\leq \mathbb{E}[\|y_t - y_\star\|^2] - \mathbb{E}[\|y_{t+1} - y_t\|^2] + \gamma_0 L_f \mathbb{E}[\|x_t - z_t\|^2] + \frac{2\gamma_0\sigma^2}{L_f(T+1)} \\
&\leq \mathbb{E}[\|y_t - y_\star\|^2] + \frac{2\sigma^2}{L_f^2(T+1)}, \qquad \text{(S.56)}
\end{aligned}$$

where the second line holds by assumption $\gamma_0 < 1/L_f$.

Finally, we sum this inequality over $t = 0$ to T,

$$\mathbb{E}[\|y_{T+1} - y_\star\|^2] \leq \|y_0 - y_\star\|^2 + \frac{2\sigma^2}{L_f^2}. \qquad \text{(S.57)}$$

Remark that $\mathbb{E}[\|y_{T+1} - y_\star\|]^2 \leq \mathbb{E}[\|y_{T+1} - y_\star\|^2]$. We finish the proof by taking the square-root of both sides. ☐

D Convergence Guarantees

This section presents the technical analysis of our main results.

D.1 Proof of Theorem 1

We divide this proof into two parts. In the first part, we will show that the iterates of TOS satisfy,

$$\langle u_t, x_t - x_\star\rangle + g(z_t) - g(x_\star) + h(x_t) - h(x_\star) \leq \frac{1}{2\gamma}\|y_t - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - y_t\|^2, \quad \text{(S.58)}$$

for all $x \in \mathbb{R}^n$.

Since $x_t = \mathrm{prox}_{\gamma h}(2z_t - y_t - \gamma_t u_t)$, by Lemma S.2, we have

$$\langle 2z_t - y_t - \gamma u_t - x_t, x_\star - x_t\rangle \leq \gamma h(x_\star) - \gamma h(x_t). \qquad \text{(S.59)}$$

We rearrange this inequality as follows:

$$\begin{aligned}
\langle u_t, x_t - x_\star\rangle + h(x_t) - h(x_\star) &\leq \frac{1}{\gamma}\langle 2z_t - y_t - x_t, x_t - x_\star\rangle \\
&= \frac{1}{\gamma}\langle z_t - y_t, z_t - x_\star\rangle + \frac{1}{\gamma}\langle z_t - y_t, x_t - z_t\rangle + \frac{1}{\gamma}\langle z_t - x_t, x_t - x_\star\rangle \\
&= \frac{1}{\gamma}\langle z_t - y_t, z_t - x_\star\rangle + \frac{1}{\gamma}\langle y_t + x_t - z_t - x_\star, x_t - z_t\rangle \\
&= \frac{1}{\gamma}\langle z_t - y_t, z_t - x_\star\rangle + \frac{1}{\gamma}\langle y_{t+1} - x_\star, y_{t+1} - y_t\rangle. \qquad \text{(S.60)}
\end{aligned}$$

Then, we use Lemma S.2 once again (for γg) and get

$$\begin{aligned}\langle u_t, x_t - x_\star\rangle + g(z_t) - g(x_\star) + h(x_t) - h(x_\star) &\leq \frac{1}{\gamma}\langle y_{t+1} - x_\star, y_{t+1} - y_t\rangle \\ &\leq \frac{1}{2\gamma}\|y_t - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - y_t\|^2. \quad \text{(S.61)}\end{aligned}$$

This completes the first part of the proof.

In the second part, we will characterize the convergence rate of $f(\bar{z}_t) + g(\bar{z}_t) + h(\bar{x}_t) - \phi_\star$ to 0 by using (S.58). Since f is convex, we have

$$\langle u_t, x_t - x_\star\rangle = \langle u_t, z_t - x_\star\rangle - \langle u_t, z_t - x_t\rangle \geq f(z_t) - f(x_\star) - \frac{1}{2\gamma}\|y_{t+1} - y_t\|^2 - \frac{\gamma}{2}\|u_t\|^2. \quad \text{(S.62)}$$

By combining (S.62) and (S.58), we obtain

$$f(z_t) + g(z_t) + h(x_t) - \phi_\star \leq \frac{1}{2\gamma}\|y_t - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - x_\star\|^2 + \frac{\gamma}{2}\|u_t\|^2. \quad \text{(S.63)}$$

We sum this inequality over $t = 0$ to T:

$$\sum_{\tau=0}^{T}\Big(f(z_\tau) + g(z_\tau) + h(x_\tau) - \phi_\star\Big) \leq \frac{1}{2\gamma}\|y_0 - x_\star\|^2 + \frac{\gamma}{2}\sum_{\tau=0}^{T}\|u_\tau\|^2 \leq \frac{1}{2\gamma}\|y_0 - x_\star\|^2 + \frac{\gamma_0}{2}G_f^2\sqrt{T+1}, \quad \text{(S.64)}$$

where the second inequality holds due to Assumption 1. Finally, we divide both sides by $(T+1)$ and use Jensen's inequality:

$$f(\bar{z}_t) + g(\bar{z}_t) + h(\bar{x}_t) - \phi_\star \leq \frac{1}{2\sqrt{T+1}}\left(\frac{1}{\gamma_0}\|y_0 - x_\star\|^2 + \gamma_0 G_f^2\right). \quad \text{(S.65)}$$

D.2 Proof of Theorem 2

The proof is similar to the proof of Theorem 1. We will only discuss the different steps. Part 1 of the proof is the same, *i.e.,* (S.58) still holds.

We need to consider the randomness of the gradient estimator in the second part. To this end, we modify (S.62) as:

$$\begin{aligned}\mathbb{E}[\langle u_t, x_t - x_\star\rangle] &= \mathbb{E}[\langle \nabla f(z_t), z_t - x_\star\rangle] + \mathbb{E}[\langle u_t - \nabla f(z_t), z_t - x_\star\rangle] - \mathbb{E}[\langle u_t, z_t - x_t\rangle] \\ &\geq \mathbb{E}[f(z_t) - f(x_\star)] - \mathbb{E}[\langle u_t, z_t - x_t\rangle] \\ &= \mathbb{E}[f(z_t) - f(x_\star)] + \mathbb{E}[\langle \nabla f(z_t) - u_t, z_t - x_t\rangle] - \mathbb{E}[\langle \nabla f(z_t), z_t - x_t\rangle] \\ &\geq \mathbb{E}[f(z_t) - f(x_\star)] - \frac{1}{2\gamma}\mathbb{E}[\|z_t - x_t\|^2] - \gamma\mathbb{E}[\|\nabla f(z_t) - u_t\|^2] - \gamma\mathbb{E}[\|\nabla f(z_t)\|^2] \\ &\geq \mathbb{E}[f(z_t) - f(x_\star)] - \frac{1}{2\gamma}\mathbb{E}[\|y_{t+1} - y_t\|^2] - \gamma(\sigma^2 + G_f^2). \quad \text{(S.66)}\end{aligned}$$

Now, we take the expectation of (S.58) and substitute (S.66) into it to obtain

$$\mathbb{E}[f(z_t) + g(z_t) + h(x_t)] - \phi_\star \leq \frac{1}{2\gamma}\mathbb{E}[\|y_t - x_\star\|^2] - \frac{1}{2\gamma}\mathbb{E}[\|y_{t+1} - x_\star\|^2] + \gamma(\sigma^2 + G_f^2). \quad \text{(S.67)}$$

We sum this inequality from $t = 0$ to T and divide both sides by $T+1$. Then, we use Jensen's inequality and get

$$\mathbb{E}[f(\bar{z}_T) + g(\bar{z}_T) + h(\bar{x}_T)] - \phi_\star \leq \frac{1}{\sqrt{T+1}}\left(\frac{1}{2\gamma_0}\|y_0 - x_\star\|^2 + \gamma_0(\sigma^2 + G_f^2)\right). \quad \text{(S.68)}$$

D.3 TOS for the Smooth and Stochastic Setting (Remark 2)

Theorem S.9. *Consider Problem (16) and let f be L_f-smooth. Employ TOS (Algorithm 2) with a fixed step-size $\gamma_t = \gamma = \gamma_0/\sqrt{T+1}$ for some $\gamma_0 \in [0, \frac{1}{2L_f}]$ and suppose that we are receiving the update direction u_t from a stochastic first-order oracle that satisfies Assumption 2. Then,*

$$\mathbb{E}[f(\bar{x}_T) + g(\bar{z}_T) + h(\bar{x}_T)] - \phi_\star \leq \frac{1}{\sqrt{T+1}}\left(\frac{D^2}{2\gamma_0} + \gamma_0\sigma^2\right), \quad \textit{where} \quad D = \|y_0 - x_\star\|. \quad \text{(S.69)}$$

Proof. The proof is similar to the proof of Theorem 1. (S.58) still holds. We modify (S.62) as follows (similar to (S.66)):

$$\begin{aligned}\mathbb{E}[\langle u_t, x_t - x_\star\rangle] &\geq \mathbb{E}[f(z_t) - f(x_\star)] + \mathbb{E}[\langle \nabla f(z_t) - u_t, z_t - x_t\rangle] - \mathbb{E}[\langle \nabla f(z_t), z_t - x_t\rangle] \\ &\geq \mathbb{E}[f(x_t) - f(x_\star)] - \frac{1}{4\gamma}\mathbb{E}[\|z_t - x_t\|^2] - \gamma\mathbb{E}[\|\nabla f(z_t) - u_t\|^2] - \frac{L_f}{2}\|x_t - z_t\|^2 \\ &\geq \mathbb{E}[f(x_t) - f(x_\star)] - \frac{1+2\gamma L_f}{4\gamma}\mathbb{E}[\|y_{t+1} - y_t\|^2] - \gamma\sigma^2. \end{aligned} \tag{S.70}$$

We take the expectation of (S.58) and replace (S.70) into it

$$\begin{aligned}\mathbb{E}[f(x_t) + g(z_t) + h(x_t)] - \phi_\star &\leq \frac{1}{2\gamma}\|y_t - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - x_\star\|^2 + \frac{2\gamma L_f - 1}{4\gamma}\mathbb{E}[\|y_{t+1} - y_t\|^2] + \gamma\sigma^2 \\ &\leq \frac{1}{2\gamma}\|y_t - x_\star\|^2 - \frac{1}{2\gamma}\|y_{t+1} - x_\star\|^2 + \gamma\sigma^2, \end{aligned} \tag{S.71}$$

where the second line holds since we choose $\gamma_0 \in [0, \frac{1}{2L_f}]$.

We sum (S.71) from $t = 0$ to T and divide both sides by $T+1$. We complete the proof by using Jensen's inequality. □

E Convergence Guarantees for AdapTos

In this section, we focus on Problem (20), a special subclass of (1) where g and h represent constraints. In this setting, TOS performs the following steps iteratively for $t = 0, 1, \ldots$:

$$z_t = \mathrm{proj}_{\mathcal{G}}(y_t) \tag{S.72}$$
$$x_t = \mathrm{proj}_{\mathcal{H}}(2z_t - y_t - \gamma_t u_t) \tag{S.73}$$
$$y_{t+1} = y_t - z_t + x_t, \tag{S.74}$$

where γ_t at line (S.73) is chosen according to the adaptive step-size rule (21).

E.1 Proof of Theorem 3

First, we will bound the growth rate of $\|y_{t+1} - y_\star\|$. Similar to the proof of Theorem S.6, we decompose $\|y_{t+1} - y_\star\|^2$ as

$$\begin{aligned}\|y_{t+1} - y_\star\|^2 &= \|y_t - z_t + x_t - y_\star + x_\star - x_\star\|^2 \\ &= \|y_t - z_t - y_\star + x_\star\|^2 + \|x_t - x_\star\|^2 + 2\langle x_t - x_\star, y_t - z_t - y_\star + x_\star\rangle. \end{aligned} \tag{S.75}$$

Since $z_t = \mathrm{proj}_{\mathcal{G}}(y_t)$ and $x_\star = \mathrm{proj}_{\mathcal{G}}(y_\star)$, by the firm non-expansivity, we have

$$\begin{aligned}\|y_t - z_t - y_\star + x_\star\|^2 &= \langle y_t - z_t - y_\star + x_\star, y_t - y_\star\rangle - \langle y_t - z_t - y_\star + x_\star, z_t - x_\star\rangle \\ &\leq \langle y_t - z_t - y_\star + x_\star, y_t - y_\star\rangle. \end{aligned} \tag{S.76}$$

Similarly, since $x_t = \mathrm{proj}_{\mathcal{H}}(2z_t - y_t - \gamma_t u_t)$ and $x_\star = \mathrm{proj}_{\mathcal{H}}(2x_\star - y_\star - \gamma u_\star)$, by the firm non-expansivity,

$$\|x_t - x_\star\|^2 \leq \langle x_t - x_\star, (2z_t - y_t - \gamma_t u_t) - (2x_\star - y_\star - \gamma u_\star)\rangle. \tag{S.77}$$

By combining (S.75), (S.76) and (S.77), we get

$$\begin{aligned}\|y_{t+1} - y_\star\|^2 &\leq \langle y_t - z_t + x_t - y_\star + x_\star - x_\star, y_t - y_\star\rangle - \langle x_t - x_\star, \gamma_t u_t - \gamma u_\star\rangle \\ &= \langle y_{t+1} - y_\star, y_t - y_\star\rangle - \langle x_t - x_\star, \gamma_t u_t - \gamma u_\star\rangle \\ &= \frac{1}{2}\|y_{t+1} - y_\star\|^2 + \frac{1}{2}\|y_t - y_\star\|^2 - \frac{1}{2}\|y_{t+1} - y_t\|^2 - \langle x_t - x_\star, \gamma_t u_t - \gamma u_\star\rangle, \end{aligned} \tag{S.78}$$

where the second line follows the definition of y_{t+1}. Then, we take $\gamma \to 0$ and rearrange (S.78) as follows:

$$\begin{aligned}\|y_{t+1} - y_\star\|^2 &\leq \|y_t - y_\star\|^2 - \|y_{t+1} - y_t\|^2 + 2\gamma_t\langle u_t, x_\star - x_t\rangle \\ &= \|y_t - y_\star\|^2 - \|y_{t+1} - y_t\|^2 + 2\gamma_t\langle u_t, x_\star - z_t\rangle + 2\gamma_t\langle u_t, z_t - x_t\rangle \\ &\leq \|y_t - y_\star\|^2 + 2\gamma_t\langle u_t, x_\star - z_t\rangle + \gamma_t^2\|u_t\|^2 \\ &\leq \|y_t - y_\star\|^2 + 2\gamma_t\|u_t\|\|z_t - x_\star\| + \gamma_t^2\|u_t\|^2 \\ &\leq \|y_t - y_\star\|^2 + 2\gamma_t\|u_t\|\|y_t - y_\star\| + \gamma_t^2\|u_t\|^2 = (\|y_t - y_\star\| + \gamma_t\|u_t\|)^2, \end{aligned} \tag{S.79}$$

where we used the non-expansivity of the projection operator in the last line, *i.e.,* $\|z_t - x_\star\| = \|\mathrm{proj}_{\mathcal{G}}(y_t) - \mathrm{proj}_{\mathcal{G}}(y_\star)\| \le \|y_t - y_\star\|$. Next, we take the square root of both sides to get

$$\begin{aligned}\|y_{t+1} - y_\star\| \le \|y_t - y_\star\| + \gamma_t\|u_t\| \le \|y_0 - y_\star\| + \sum_{\tau=0}^{t}\gamma_\tau\|u_\tau\| \\ = \|y_0 - y_\star\| + \gamma_0\sum_{\tau=0}^{t}\sqrt{\frac{\|u_\tau\|^2}{\sum_{j=0}^{\tau}\|u_j\|^2}} \\ \le \|y_0 - y_\star\| + \gamma_0\sqrt{(t+1)\sum_{\tau=0}^{t}\frac{\|u_\tau\|^2}{\sum_{j=0}^{\tau}\|u_j\|^2}}. \end{aligned} \tag{S.80}$$

By using Lemma S.3 and Assumption 1, we get

$$\|y_{t+1} - y_\star\| \le \|y_0 - y_\star\| + \gamma_0\sqrt{(t+1)\log\left(\sum_{\tau=0}^{t}\|u_\tau\|^2\right)} \le \|y_0 - y_\star\| + \gamma_0\sqrt{(t+1)\log\left(G_f^2(t+1)\right)}. \tag{S.81}$$

Now, we can derive a bound on the infeasibility as follows:

$$\begin{aligned}\mathrm{dist}(\bar{z}_t, \mathcal{H}) \le \|\bar{x}_t - \bar{z}_t\| = \frac{1}{t+1}\|\sum_{\tau=0}^{t}(x_\tau - z_\tau)\| = \frac{1}{t+1}\|y_{t+1} - y_0\| \le \frac{1}{t+1}\left(\|y_{t+1} - y_\star\| + \|y_0 - y_\star\|\right) \\ \le \frac{1}{t+1}\left(2\|y_0 - y_\star\| + \gamma_0\sqrt{(t+1)\log\left(G_f^2(t+1)\right)}\right). \end{aligned} \tag{S.82}$$

Next, we prove convergence in the objective value. Define $s_t = \sum_{\tau=0}^{t}\gamma_t$ and $\tilde{z}_t = \frac{1}{s_t}\sum_{\tau=0}^{t}\gamma_t z_t$. Since f is convex, by Jensen's inequality,

$$f(\tilde{z}_t) - f_\star \le \frac{1}{s_t}\sum_{\tau=0}^{t}\gamma_\tau\left(f(z_\tau) - f_\star\right) \le \frac{1}{s_t}\sum_{\tau=0}^{t}\gamma_\tau\langle u_t, z_\tau - x_\star\rangle. \tag{S.83}$$

From (S.79), we have

$$\gamma_t\langle u_t, z_t - x_\star\rangle \le \frac{1}{2}\|y_t - y_\star\|^2 - \frac{1}{2}\|y_{t+1} - y_\star\|^2 + \frac{1}{2}\gamma_t^2\|u_t\|^2. \tag{S.84}$$

If we substitute (S.84) into (S.83), we obtain

$$\begin{aligned}f(\tilde{z}_t) - f_\star \le \frac{1}{2s_t}\left(\|y_0 - y_\star\|^2 + \sum_{\tau=0}^{t}\gamma_\tau^2\|u_\tau\|^2\right) = \frac{1}{2s_t}\left(\|y_0 - y_\star\|^2 + \gamma_0^2\sum_{\tau=0}^{t}\frac{\|u_\tau\|^2}{\sum_{j=0}^{\tau}\|u_j\|^2}\right) \\ \le \frac{1}{2s_t}\left(\|y_0 - y_\star\|^2 + \gamma_0^2\log(\sum_{\tau=0}^{t}\|u_\tau\|^2)\right) \le \frac{1}{2s_t}\left(\|y_0 - y_\star\|^2 + \gamma_0^2\log(G_f^2(t+1))\right), \end{aligned} \tag{S.85}$$

where the second line uses Lemma S.4 and Assumption 1. Finally, we note that

$$s_t = \sum_{\tau=0}^{t}\gamma_\tau = \sum_{\tau=0}^{t}\frac{\gamma_0}{\sqrt{\sum_{j=0}^{\tau}\|u_j\|^2}} \ge \sum_{\tau=0}^{t}\frac{\gamma_0}{\sqrt{\sum_{j=0}^{\tau}G_f^2}} = \sum_{\tau=0}^{t}\frac{\gamma_0}{\sqrt{G_f^2(\tau+1)}} \ge \sum_{\tau=0}^{t}\frac{\gamma_0}{G_f\sqrt{t+1}} = \frac{\gamma_0\sqrt{t+1}}{G_f}. \tag{S.86}$$

We complete the proof by using (S.86) in (S.85):

$$f(\tilde{z}_t) - f_\star \le \frac{G_f}{2\gamma_0\sqrt{t+1}}\left(\|y_0 - y_\star\|^2 + \gamma_0^2\log(G_f^2(t+1))\right). \tag{S.87}$$

E.2 Proof of Theorem 4

As in the proof of Theorem 3, our first goal is to bound $\|y_{t+1} - y_\star\|$. To this end, we start by

$$\begin{aligned}\|y_{t+1} - y_\star\|^2 &\le \|y_t - y_\star\|^2 - \|y_{t+1} - y_t\|^2 + 2\gamma_t\langle u_t, x_\star - x_t\rangle \\ &= \|y_t - y_\star\|^2 - \|y_{t+1} - y_t\|^2 + 2\gamma_t\langle u_t, x_\star - z_t\rangle + 2\gamma_t\langle u_t, z_t - x_t\rangle \\ &\le \|y_t - y_\star\|^2 + 2\gamma_t\langle u_t, x_\star - z_t\rangle + \gamma_t^2\|u_t\|^2. \end{aligned} \tag{S.88}$$

f is convex and $u_\star = 0$ by assumption. Hence, we have

$$\langle u_t, x_\star - z_t\rangle = \langle u_t - u_\star, x_\star - z_t\rangle \le 0. \tag{S.89}$$

Together with (S.88), this leads to

$$\|y_{t+1} - y_\star\|^2 \le \|y_t - y_\star\|^2 + \gamma_t^2\|u_t\|^2 \le \|y_0 - y_\star\|^2 + \sum_{\tau=0}^{t}\gamma_\tau^2\|u_\tau\|^2 = \|y_0 - y_\star\|^2 + \gamma_0^2\sum_{\tau=0}^{t}\frac{\|u_\tau\|^2}{\sum_{j=0}^{\tau}\|u_j\|^2}. \tag{S.90}$$

By using Lemma S.4 and Assumption 1, we get

$$\begin{aligned}\|y_{t+1} - y_\star\|^2 \le \|y_t - y_\star\|^2 + \gamma_t^2\|u_t\|^2 &\le \|y_0 - y_\star\|^2 + \gamma_0^2\log(\sum_{\tau=0}^{t}\|u_\tau\|^2) \\ &\le \|y_0 - y_\star\|^2 + \gamma_0^2\log(G_f^2(t+1)). \end{aligned} \tag{S.91}$$

We take the square-root of both sides to obtain

$$\|y_{t+1} - y_\star\| \le \|y_0 - y_\star\| + \gamma_0\sqrt{\log(G_f^2(t+1))}. \tag{S.92}$$

This proves that $\|y_t - y_\star\|$ is bounded by a logarithmic growth. We can use this bound to prove convergence to a feasible point, since

$$\mathrm{dist}(\bar{z}_t, \mathcal{H}) \le \|\bar{x}_t - \bar{z}_t\| \le \frac{1}{t+1}\left(\|y_{t+1} - y_\star\| + \|y_0 - y_\star\|\right) \le \frac{1}{t+1}\left(2\|y_0 - y_\star\| + \gamma_0\sqrt{\log(G_f^2(t+1))}\right). \tag{S.93}$$

Next, we analyze the objective suboptimality. From (S.88), we have

$$\langle u_t, z_t - x_\star\rangle \le \frac{1}{2\gamma_t}\|y_t - y_\star\|^2 - \frac{1}{2\gamma_t}\|y_{t+1} - y_\star\|^2 + \frac{\gamma_t}{2}\|u_t\|^2. \tag{S.94}$$

The, since f is convex, we have

$$\begin{aligned}\frac{1}{t+1}\sum_{\tau=0}^{t}(f(z_\tau) - f_\star) &\le \frac{1}{t+1}\sum_{\tau=0}^{t}\langle u_t, z_t - x_\star\rangle \\ &\le \frac{1}{t+1}\left(\frac{1}{2\gamma_0}\|y_0 - y_\star\|^2 + \frac{1}{2}\sum_{\tau=1}^{t}\left(\frac{1}{\gamma_\tau} - \frac{1}{\gamma_{\tau-1}}\right)\|y_\tau - y_\star\|^2 + \sum_{\tau=0}^{t}\frac{\gamma_\tau}{2}\|u_\tau\|^2\right). \end{aligned} \tag{S.95}$$

We focus on the middle term and use (S.91) as

$$\begin{aligned}\sum_{\tau=1}^{t}\left(\frac{1}{\gamma_\tau} - \frac{1}{\gamma_{\tau-1}}\right)\|y_\tau - y_\star\|^2 &\le \left(\|y_0 - y_\star\|^2 + 2\gamma_0^2\log(G_f\sqrt{t})\right)\sum_{\tau=1}^{t}\left(\frac{1}{\gamma_\tau} - \frac{1}{\gamma_{\tau-1}}\right) \\ &= \left(\|y_0 - y_\star\|^2 + 2\gamma_0^2\log(G_f\sqrt{t})\right)\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_0}\right). \end{aligned} \tag{S.96}$$

If we substitute (S.96) back into (S.95), we get

$$\begin{aligned}\frac{1}{t+1}\sum_{\tau=0}^{t}(f(z_\tau) - f_\star) &\le \frac{1}{t+1}\left(\left(\frac{1}{2}\|y_0 - y_\star\|^2 + \gamma_0^2\log(G_f\sqrt{t})\right)\frac{1}{\gamma_t} + \sum_{\tau=0}^{t}\frac{\gamma_\tau}{2}\|u_\tau\|^2\right) \\ &= \frac{1}{t+1}\left(\left(\frac{1}{2\gamma_0}\|y_0 - y_\star\|^2 + \gamma_0 x\log(G_f\sqrt{t})\right)\sqrt{\sum_{\tau=0}^{t}\|u_\tau\|^2} + \frac{\gamma_0}{2}\sum_{\tau=0}^{t}\frac{\|u_\tau\|^2}{\sqrt{\sum_{j=0}^{\tau}\|u_j\|^2}}\right). \end{aligned} \tag{S.97}$$

Then, we use Lemma S.3 and Lemma S.5 to get

$$\frac{1}{t+1}\sum_{\tau=0}^{t}(f(z_\tau)-f_\star) \le \frac{1}{t+1}\left(\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0\log(G_f\sqrt{t})\right)\sqrt{\sum_{\tau=0}^{t}\|u_\tau\|^2}+\gamma_0\sqrt{\sum_{\tau=0}^{t}\|u_\tau\|^2}\right)$$
$$= \frac{1}{t+1}\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0(1+\log(G_f\sqrt{t}))\right)\sqrt{\sum_{\tau=0}^{t}\|u_\tau\|^2}$$
$$\le \frac{1}{t+1}\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0(1+\log(G_f\sqrt{t}))\right)\sqrt{2L_f\sum_{\tau=0}^{t}(f(z_\tau)-f_\star)}. \tag{S.98}$$

We can simplify (S.98) as

$$\sqrt{\frac{1}{t+1}\sum_{\tau=0}^{t}(f(z_\tau)-f_\star)} \le \frac{\sqrt{2L_f}}{\sqrt{t+1}}\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0(1+\log(G_f\sqrt{t}))\right). \tag{S.99}$$

Finally, we take the square of both sides and use Jensen's inequality to achieve the desired bound.

Remark S.3. *Note that we did not use the smoothness assumption for f until (S.98). Therefore, for Lipschitz continuous nonsmooth f, if the solution is in the interior of the feasible set, we get*

$$f(\bar{z}_t)-f_\star \le \frac{1}{t+1}\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0(1+\log(G_f\sqrt{t}))\right)\sqrt{\sum_{\tau=0}^{t}\|u_\tau\|^2}$$
$$\le \frac{1}{t+1}\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0(1+\log(G_f\sqrt{t}))\right)\sqrt{\sum_{\tau=0}^{t}G_f^2}$$
$$= \frac{G_f}{\sqrt{t+1}}\left(\frac{1}{2\gamma_0}\|y_0-y_\star\|^2+\gamma_0(1+\log(G_f\sqrt{t}))\right). \tag{S.100}$$

We used the assumption that $u_\star = 0$ only for the boundedness condition (S.91).

E.3 Proof of Theorem 5

The proof is similar to the proof of Theorem 3. We follow the same steps until (S.80). Then, we take the expectation

$$\mathbb{E}[\|y_{t+1}-y_\star\|] \le \|y_0-y_\star\|+\gamma_0\sqrt{(t+1)\log\left(\sum_{\tau=0}^{t}\mathbb{E}[\|u_\tau\|^2]\right)}. \tag{S.101}$$

By Assumptions 1 and 2, we have

$$\mathbb{E}[\|u_t\|^2] = \mathbb{E}[\|\nabla f(z_t)\|^2]+\mathbb{E}[\|\nabla f(z_t)-u_t\|^2] \le G_f^2+\sigma^2. \tag{S.102}$$

We combine (S.101) and (S.102) to get

$$\mathbb{E}[\|y_{t+1}-y_\star\|] \le \|y_0-y_\star\|+\gamma_0\sqrt{(t+1)\log\left((G_f^2+\sigma^2)(t+1)\right)}. \tag{S.103}$$

Similar to (S.82), we get

$$\mathbb{E}[\mathrm{dist}(\bar{z}_t,\mathcal{H})] \le \frac{1}{t+1}\left(2\|y_0-y_\star\|+\gamma_0\sqrt{(t+1)\log\left((G_f^2+\sigma^2)(t+1)\right)}\right). \tag{S.104}$$

Note that, similar to (S.86), we have

$$\mathbb{E}[s_t] \ge \frac{\gamma_0\sqrt{t+1}}{G_f+\sigma}. \tag{S.105}$$

For the convergence in function value, we modify (S.85) as

$$\begin{aligned}\mathbb{E}[f(\tilde{z}_t) - f_\star] &\leq \mathbb{E}\left[\frac{1}{2s_t}\left(\|y_0 - y_\star\|^2 + \gamma_0^2 \log(\sum_{\tau=0}^{t} \|u_\tau\|^2)\right)\right] \\ &\leq \frac{G_f + \sigma}{2\gamma_0\sqrt{t+1}}\left(\|y_0 - y_\star\|^2 + \gamma_0^2 \log\left(\sum_{\tau=0}^{t} \mathbb{E}[\|u_\tau\|^2]\right)\right) \\ &\leq \frac{G_f + \sigma}{2\gamma_0\sqrt{t+1}}\left(\|y_0 - y_\star\|^2 + \gamma_0^2 \log\left((G_f^2 + \sigma^2)(t+1)\right)\right), \end{aligned} \tag{S.106}$$

where we used (S.102) again in the last line.

F Additional Details on the Numerical Experiments

F.1 Details for Section 6.1

Figure S.4 presents the details on how we tuned γ_0 for ADAPTOS on the experiments we considered in Section 6.1.

For Figure 1, we choose:

- ▷ $\gamma_0 = 10$ for overlapping group lasso with $\lambda = 10^{-3}$ and synthetic data,
- ▷ $\gamma_0 = 1$ for overlapping group lasso with $\lambda = 10^{-1}$ and synthetic data,
- ▷ $\gamma_0 = 100$ for overlapping group lasso with $\lambda = 10^{-3}$ and real-sim dataset,
- ▷ $\gamma_0 = 100$ for overlapping group lasso with $\lambda = 10^{-2}$ and real-sim dataset,
- ▷ $\gamma_0 = 1$ for sparse and low-rank regularization with $\lambda = 10^{-3}$,
- ▷ $\gamma_0 = 1$ for sparse and low-rank regularization with $\lambda = 10^{-4}$,
- ▷ $\gamma_0 = 100$ for total variation deblurring with $\lambda = 10^{-6}$,
- ▷ $\gamma_0 = 100$ for total variation deblurring with $\lambda = 10^{-4}$.

F.2 Details for Section 6.2

Figure S.5 shows the recovered approximations with ℓ_1 and ℓ_2-loss functions along with the original image and the noisy observation. ℓ_1-noise is known to be more reliable against outliers, and it empirically generates a better approximation of the original image with 26.21 dB peak signal to noise ratio (PSNR) against 21.15 dB for the ℓ_2-loss.

F.3 Details for Section 6.3

Let $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{w})$ denote a generic deep neural network with H hidden layers, which takes a vector input x and returns a vector output y, and w represents the column-vector concatenation of all adaptable parameters. kth hidden layer operates on input vector $\boldsymbol{\theta}_k$ and returns $\boldsymbol{\theta}_{k+1}$,

$$\boldsymbol{\theta}_{k+1} = \phi_k(\boldsymbol{W}_k\boldsymbol{\theta}_k + \boldsymbol{b}_k), \quad \text{for } 1 \leq k \leq H, \tag{S.107}$$

where $\boldsymbol{\theta}_1 = \boldsymbol{x}$ denotes the input layer by convention, $\{\boldsymbol{\theta}_k, \boldsymbol{b}_k\}$ are the adaptable parameters of the layer and ϕ_k is an activation function to be applied entry-wise. We use ReLu activation (Glorot et al., 2011) for the hidden layers of the network and the softmax activation function for the output layer. We use the same initial weights as in (Scardapane et al., 2017), which is based on the method described in (Glorot & Bengio, 2010).

Given a set of N training examples $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, we train the network by minimizing

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \quad \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{y}_i, f(\boldsymbol{x}_i, \boldsymbol{w})) + \lambda\|\boldsymbol{w}\|_1 + \lambda\sum_{\boldsymbol{\alpha}\in\Omega}\sqrt{|\boldsymbol{\alpha}|}\|\boldsymbol{\alpha}\|_2, \tag{S.108}$$

with the standard cross-entropy loss given by $L(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w})) = -\sum_{j=1}^{\dim(\boldsymbol{y})} y_j \log(f_j(\boldsymbol{x}, \boldsymbol{w}))$. $\lambda > 0$ is the regularization parameter. We set $\lambda = 10^{-4}$, which is shown to provide the best results in terms of classification accuracy and sparsity in (Scardapane et al., 2017).

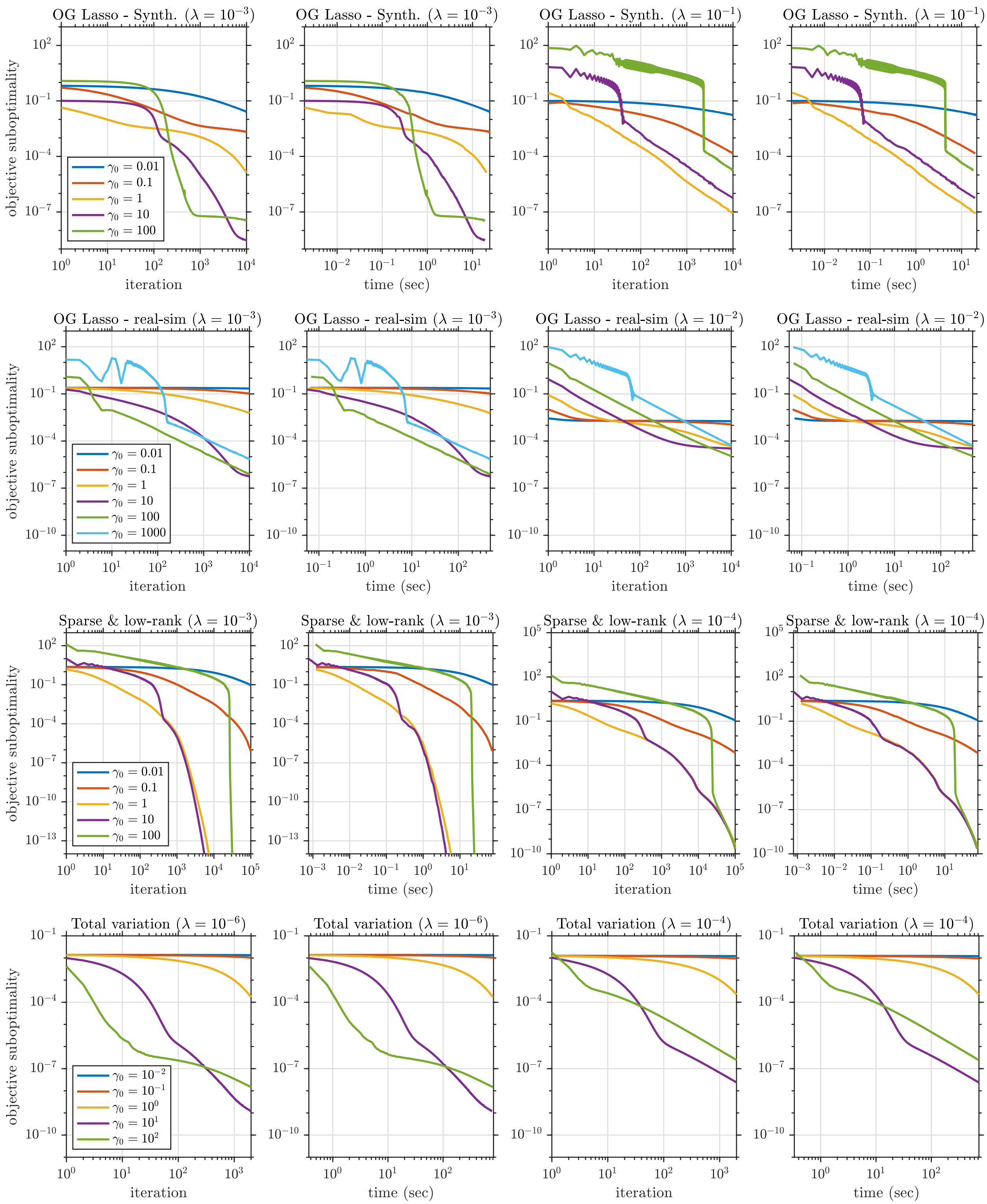

Figure S.4: Empirical performance of ADAPTOS with different choices of γ_0 for the problems with smooth and convex loss function studied in Section 6.1.

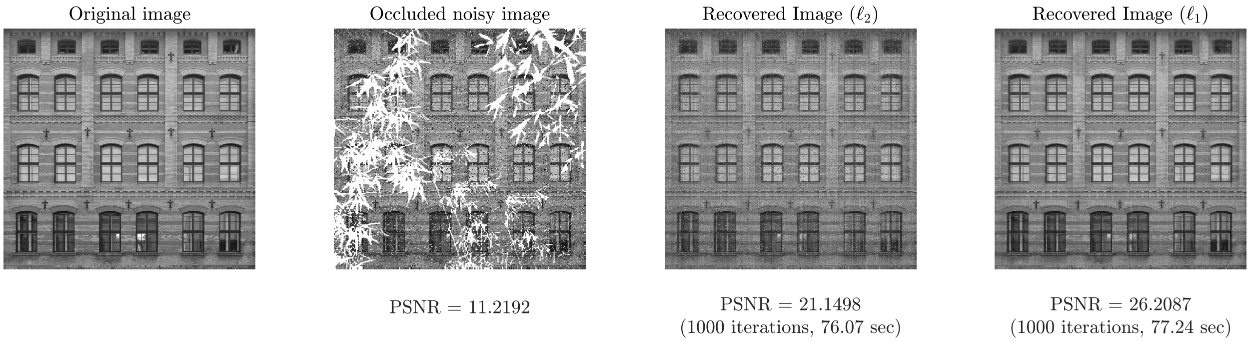

Figure S.5: Comparison or images recovered by minimizing the ℓ_1 and ℓ_2-loss functions described in Section 6.2. ℓ_1-loss empirically gives a better approximation with 5dB higher PSNR.

The first regularizer (ℓ_1 penalty) in (S.108) promotes sparsity on the overall network, while the second regularizer (Group-Lasso penalty, introduced in (Yuan & Lin, 2006)) is used to achieve group-level sparsity. The goal is to force all outgoing connections from the same neurons to be simultaneously zero, so that we can safely remove them and obtain a compact network. To this end, Ω contains the sets of all outgoing connections from each neuron (corresponding to the rows of $\boldsymbol{W}_k$) and single element groups of bias terms (corresponding to the entries of $\boldsymbol{b}_k$).

We compare our methods against SGD, AdaGrad and Adam. We use minibatch size of 400 for all methods. We use the built-in functions in Lasagne for SGD, AdaGrad and Adam. These methods use the subgradient of the overall objective (S.108). All of these methods have one learning rate parameter for tuning. We tune these parameters by trying the powers of 10. We found that $\gamma_0 = 1$ works well for TOS and ADAPTOS. For SGD and AdaGrad, we got the best performance when the learning rate parameter is set to 10^{-2}, and for Adam we got the best results with 10^{-3}.

Remark that subgradient methods are known to destroy sparsity at the intermediate iterations. For instance, the subgradients of ℓ_1 norm are fully dense. In contrast, TOS and ADAPTOS handle the regularizers through their proximal operators. The advantage of using a proximal method instead of subgradients is outstanding. TOS and ADAPTOS result in precisely sparse networks whereas other methods can only get approximately sparse solutions. The comparison becomes especially stark in group sparsity, with no clear discontinuity in the spectrum for other methods.