Three Operator Splitting with a Nonconvex Loss Function

Alp Yurtsever [1] **Varun Mangalick** [1] **Suvrit Sra** [1]

Abstract

We consider the problem of minimizing the sum of three functions, one of which is nonconvex but differentiable, and the other two are convex but possibly nondifferentiable. We investigate the Three Operator Splitting method (TOS) of Davis & Yin (2017) with an aim to extend its theoretical guarantees for this nonconvex problem template. In particular, we prove convergence of TOS with nonasymptotic bounds on its nonstationarity and infeasibility errors. In contrast with the existing work on nonconvex TOS, our guarantees do not require additional smoothness assumptions on the terms comprising the objective; hence they cover instances of particular interest where the nondifferentiable terms are indicator functions. We also extend our results to a stochastic setting where we have access only to an unbiased estimator of the gradient. Finally, we illustrate the effectiveness of the proposed method through numerical experiments on quadratic assignment problems.

1. Introduction

We study nonconvex optimization problems of the form:

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and potentially nonconvex, whereas g and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are proper lower-semicontinuous convex functions (potentially nonsmooth). Further, we assume that the domain of g, that is, $\mathrm{dom}(g) = \{x \in \mathbb{R}^n : g(x) < +\infty\}$, is bounded.

Template (1) enjoys a rich number of applications in optimization, machine learning, and statistics. Nonconvex losses arise naturally in several maximum likelihood estimation (McLachlan & Krishnan, 1996) and M-estimation problems (Ollila & Tyler, 2014; Maronna et al., 2019), in problems with a matrix factorization structure (Zass & Shashua, 2007), in certain transport and assignment problems (Koopmans & Beckmann, 1957; Peyré et al., 2019), among countless others. The nonsmooth terms in (1) can be used as regularizers, *e.g.*, to promote joint behavior such as sparsity and low-rank (Richard et al., 2012). Moreover, we can also split a complex regularizer into simpler terms for computational advantages, *e.g.*, in group lasso with overlaps (Jacob et al., 2009), structured sparsity (El Halabi & Cevher, 2015), or total variation (Barbero & Sra, 2018).

We obtain an important special case by choosing the nonsmooth terms g and h in (1) as indicator functions of closed and convex sets $\mathcal{G}$ and $\mathcal{H} \subseteq \mathbb{R}^n$. In this case, (1) turns into

$$\min_{x\in\mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{G} \cap \mathcal{H}. \tag{2}$$

We are particularly interested in the setting where $\mathcal{G}$ and $\mathcal{H}$ are *simple* in the sense that we can project onto these sets efficiently, but not so easily onto their intersection. Some examples include learning with correlation matrices (Higham & Strabić, 2016), power assignment in wireless networks (De Berg et al., 2010), graph transduction (Shivanna et al., 2015), graph matching (Zaslavskiy et al., 2008), and quadratic assignment (Koopmans & Beckmann, 1957; Loiola et al., 2007).

An effective way to solve (1) for convex f with Lipschitz gradients is the Three Operator Splitting (TOS) method (Davis & Yin, 2017), whose convergence has been well-studied (see §1.1). But for nonconvex f, convergence properties of TOS are less understood (again, see §1.1). This gap motivates us to develop nonasymptotic convergence guarantees for TOS. Beyond theoretical progress, we highlight the potential empirical value of TOS by evaluating it on a challenging nonconvex problem, the quadratic assignment problem (QAP).

Contributions. We summarize our contributions towards the convergence analysis of nonconvex TOS below.

- ▷ We first discuss how to quantify convergence of TOS to first-order stationary points for both templates (1) and (2). Specifically, we propose to measure approximate stationarity based on a variational inequality. Thereafter, we prove that the associated non-stationarity error is smaller than ϵ (in expectation over a random iteration counter) after $T = \mathcal{O}(1/\epsilon^3)$ iterations (and gradient evaluations).

[1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence to: Alp Yurtsever <alpy@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

- ▷ We extend our analysis to stochastic optimization where we have access only to an unbiased estimate of the gradient ∇f. In this case, we prove that the error is smaller than ϵ (in expectation) after $T = \mathcal{O}(1/\epsilon^3)$ iterations. The corresponding algorithm requires drawing $\mathcal{O}(1/\epsilon^5)$ *i.i.d.* stochastic gradients.

Finally, we evaluate TOS on the quadratic assignment problem using the well-known QAPLIB benchmark library (Burkard et al., 1997). Remarkably, TOS performs significantly better than the theory suggests: we find that it converges locally linearly. Understanding this behavior could be a potentially valuable question for future study.

1.1. Related Works

Davis & Yin (2017) introduce TOS for solving the monotone inclusion of three operators, one of which is co-coercive. TOS gives us a simple algorithm for (1) when f is smooth and convex, since the gradient of a smooth convex function is co-coercive. At each iteration, TOS evaluates the gradient of f and the proximal operators of g and h once, separately. TOS extends various previous operator splitting schemes such as the forward-backward splitting, Douglas-Rachford splitting, Forward-Douglas-Rachford splitting (Briceño-Arias, 2015), and the Generalized Forward-Backward splitting (Raguet et al., 2013).

The original algorithm of Davis & Yin (2017) requires knowledge of the smoothness constant of f; Pedregosa & Gidel (2018) introduce a variant of TOS with backtracking line-search that bypasses this restriction. Zong et al. (2018) analyze convergence of TOS with inexact oracles where both the gradient and proximity oracles can be noisy.

Existing work on TOS applied to nonconvex problems limits itself to the setting where at least two terms in (1) have Lipschitz continuous gradients. Under this assumption, Liu & Yin (2019) identify an envelope function for TOS, which permits one to interpret TOS as gradient descent for this envelope under a variable metric. Their envelope generalizes the well-known Moreau envelope as well as the envelopes for Douglas-Rachford and Forward-Backward splitting introduced in (Patrinos et al., 2014) and (Themelis et al., 2018).

Bian & Zhang (2020) present convergence theory for TOS under the same smoothness assumptions. They show that the sequence generated by TOS with a carefully chosen step-size converges to a stationary point of (1). They also prove asymptotic convergence rates under the assumption that the Kurdyka-Łojasiewicz property holds (see Definition 2.3 in (Bian & Zhang, 2020)).

Our focus is significantly different from these prior works on nonconvex TOS. In contrast to the settings of (Liu & Yin, 2019) and (Bian & Zhang, 2020), we do *not* impose any assumption on the smoothness of g and h. However, we do assume that the nonsmooth terms g and h are convex and the problem domain is bounded.

In particular, our setting includes nonconvex minimization over the intersection of two simple convex sets, which covers important applications such as the quadratic assignment problem and graph matching. Note that these problems are challenging for TOS even in the convex setting, because the intermediate estimates of TOS can be infeasible and the known guarantees on the convergence rate of TOS fail, see the discussion in Section 3.2 in (Pedregosa & Gidel, 2018).

Finally, Yurtsever et al. (2016), Cevher et al. (2018), Zhao & Cevher (2018), and Pedregosa et al. (2019) propose and analyze stochastic variants of TOS and related methods in the convex setting. We are unaware of any prior work on nonconvex stochastic TOS.

Notation. Before moving onto the theoretical development, let us summarize here key notation used throughout the paper. We use $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean inner product associated with the norm $\|\cdot\|$. The distance between a point $x \in \mathbb{R}^n$ and a set $\mathcal{G} \subseteq \mathbb{R}^n$ is defined as $\mathrm{dist}(x, \mathcal{G}) := \inf_{y \in \mathcal{G}} \|x - y\|$; the projection of x onto $\mathcal{G}$ is given by $\mathrm{proj}_{\mathcal{G}}(x) := \arg\min_{y \in \mathcal{G}} \|x - y\|$. We denote the indicator function of $\mathcal{G}$ by $\iota_{\mathcal{G}} : \mathbb{R}^n \to \{0, +\infty\}$, that takes 0 for any $x \in \mathcal{G}$ and $+\infty$ otherwise. The proximal operator (or prox-operator) of a function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined by $\mathrm{prox}_g(x) := \arg\min_{y \in \mathbb{R}^n} \{g(y) + \frac{1}{2}\|x - y\|^2\}$. Recall that the prox-operator for the indicator function is the projection, *i.e.*, $\mathrm{prox}_{\iota_{\mathcal{G}}}(x) = \mathrm{proj}_{\mathcal{G}}(x)$.

2. Basic Setup: Approximate Stationarity

We begin our analysis by setting up the notion of approximate stationarity that we will use to judge convergence. For unconstrained minimization of smooth functions, gradient norm is a widely used standard measure. But the gradient norm is unsuitable in our case because of the presence of constraints and nonsmooth terms in the cost.

Related work on operator splitting for nonconvex optimization typically considers the norm of a proximal gradient, or uses some other auxiliary differentiable function that converges to zero as we get closer to a first-order stationary point. See, for instance, the *envelope functions* introduced by Patrinos et al. (2014), Themelis et al. (2018) and Liu & Yin (2019), or the *energy function* defined by Bian & Zhang (2020). However, these functions can characterize stationary points of (1) only under additional smoothness assumptions on g and h. They fail to capture important applications where both g and h are nonsmooth.

In contrast, we consider a simple measure based on the variational inequality characterization of first-order stationarity.

Definition 1 (Stationary point)**.** *$\bar{z} \in \mathrm{dom}(\phi)$ is a first-order stationary point of (1) if, for all $x \in \mathrm{dom}(\phi)$,*

$$\langle \nabla f(\bar{z}), \bar{z} - x \rangle + g(\bar{z}) - g(x) + h(\bar{z}) - h(x) \leq 0. \quad (3)$$

See Lemma 4 in the supplementary material for the technical details on condition (3).

We consider a perturbation of the bound in (3) to define an approximately stationary point.

Definition 2 (ϵ-stationary point)**.** *We say $\bar{z} \in \mathrm{dom}(\phi)$ is an ϵ-stationary point of (1) if, for all $x \in \mathrm{dom}(\phi)$,*

$$\langle \nabla f(\bar{z}), \bar{z} - x \rangle + g(\bar{z}) - g(x) + h(\bar{z}) - h(x) \leq \epsilon. \quad (4)$$

This is a natural extension of the notion of suboptimal solutions in terms of function values used in convex optimization. Similar measures for stationarity appear in the literature for various problems; see *e.g.*, (He & Yuan, 2015; Nouiehed et al., 2019; Malitsky, 2019; Song et al., 2020).

TOS is particularly advantageous for (1) when the proximal operators of g and h are easy to evaluate separately but the proximal operator of their sum is difficult. For (2), this corresponds to optimization over $\mathcal{G} \cap \mathcal{H}$ by using only projections onto the individual sets and not onto their intersection. In this setting, we can achieve a feasible solution only in an asymptotic sense. Finding a *feasible* ϵ-stationary solution is an unrealistic goal. Thus, for (2), we consider a relaxation of Definition 2 that permits approximately feasible solutions.

Definition 3 (ω-feasible ϵ-stationary point)**.** *We say $\bar{z} \in \mathcal{G}$ is an ω-feasible ϵ-stationary point of (2) if*

$$\mathrm{dist}(\bar{z}, \mathcal{H}) \leq \omega, \quad \textit{and} \quad (5)$$

$$\langle \nabla f(\bar{z}), \bar{z} - x \rangle \leq \epsilon, \quad \forall x \in \mathcal{G} \cap \mathcal{H}. \quad (6)$$

Remark 1. *For simplicity, we measure infeasibility of $\bar{z}$ via $\mathrm{dist}(\bar{z}, \mathcal{H})$. This is suitable because the estimates of TOS remain in $\mathcal{G}$ by definition. We can also consider a slightly stronger notion of approximate feasibility given by $\mathrm{dist}(\bar{z}, \mathcal{G} \cap \mathcal{H})$. However, this requires additional regularity conditions on $\mathcal{G}$ and $\mathcal{H}$ to avoid pathological examples. See, for instance, Lemma 1 in (Hoffmann, 1992) or Definition 2 in (Kundu et al., 2018).*

The directional derivative condition (6) is often used in the analysis of conditional gradient methods, and it is known as the *Frank-Wolfe gap* in this literature. See (Jaggi, 2013; Lacoste-Julien, 2016; Reddi et al., 2016b; Yurtsever et al., 2019) for some examples.

Approximately feasible solutions are widely considered in the analysis of primal-dual methods (but usually in the convex setting), see (Yurtsever et al., 2018; Kundu et al., 2018) and the references therein. Remark that TOS can also be viewed as a primal-dual method (Pedregosa & Gidel, 2018).

Problem (2) is challenging for TOS because of the infeasibility of the intermediate estimates, even when f is convex. Davis & Yin (2017) avoid this issue by evaluating the terms h and $(f+g)$ at two different points, $x \in \mathcal{H}$ and $z \in \mathcal{G}$.

However, $\big(f(z) + g(z)\big) + h(x)$ can be equal to the optimal objective value even when neither x nor z is close to a solution. We can address this issue by introducing a condition on the distance between x and z. The following definition of an α-close and β-stationary pair of points is crucial for our analysis.

Definition 4 (α-close β-stationary pair)**.** *We say that $(\bar{x}, \bar{z}) \in \mathrm{dom}(h) \times \mathrm{dom}(g)$ are α-close and β-stationary points of (1) if, for all $x \in \mathrm{dom}(\phi)$,*

$$\|\bar{z} - \bar{x}\| \leq \alpha, \qquad \textit{and} \quad (7)$$

$$\langle \nabla f(\bar{z}), \bar{x} - x \rangle + g(\bar{z}) - g(x) + h(\bar{x}) - h(x) \leq \beta. \quad (8)$$

α-close β-stationary points $(\bar{x}, \bar{z})$ yield approximate solutions to (1) and (2) under appropriate assumptions.

Observation 1. ***(i).*** *Let h be Lipschitz continuous on $\mathbb{R}^n$ with constant L_h. Assume that $\|\nabla f(z)\|$ is bounded by G_f for all $x \in \mathrm{dom}(g)$. Suppose that the points $(\bar{x}, \bar{z})$ are α-close and β-stationary. Then, $\bar{z}$ is an ϵ-stationary point with $\epsilon = \alpha(G_f + L_h) + \beta$ as per Definition 2.*

(ii). *Let g and h be indicators of closed convex sets $\mathcal{G}$ and $\mathcal{H}$ respectively. Assume that $\|\nabla f(z)\|$ is bounded by G_f for all $x \in \mathcal{G}$. Suppose that the points $(\bar{x}, \bar{z}) \in \mathcal{H} \times \mathcal{G}$ are α-close and β-stationary. Then, $\bar{z}$ is an α-feasible ϵ-stationary point with $\epsilon = \alpha G_f + \beta$ as per Definition 3.*

Proof. ***(i).*** Since h is Lipschitz, we have

$$h(\bar{x}) - h(x) \geq h(\bar{z}) - h(x) - L_h \|\bar{z} - \bar{x}\|. \quad (9)$$

And since $\|\nabla f(z)\|$ is bounded, we have

$$\langle \nabla f(\bar{z}), \bar{x} - x \rangle \geq \langle \nabla f(\bar{z}), \bar{z} - x \rangle - G_f \|\bar{z} - \bar{x}\|. \quad (10)$$

We get (4) with $\epsilon = \alpha(G_f + L_h) + \beta$ by using (9) and (10) in (8) and bounding $\|\bar{z} - \bar{x}\|$ by (7).

(ii). We get (5) with $\omega = \alpha$ since

$$\mathrm{dist}(\bar{z}, \mathcal{H}) = \inf_{x \in \mathcal{H}} \|\bar{z} - x\| \leq \|\bar{z} - \bar{x}\|. \quad (11)$$

$h(\bar{x}) = g(\bar{z}) = h(x) = g(x) = 0$ since $\bar{x} \in \mathcal{H}$, $\bar{z} \in \mathcal{G}$, and $x \in \mathcal{G} \cap \mathcal{H}$. Then, (6) follows from (8) by using (10). $\square$

We are now ready to present and analyze the algorithm.

3. TOS with a Nonconvex Loss Function

This section establishes convergence guarantees of TOS for solving Problems (1) and (2). The method is detailed in Algorithm 1.

Algorithm 1 Three Operator Splitting (TOS)

Input: Initial point $y_1 \in \mathbb{R}^n$, step-size sequence $\{\gamma_t\}_{t=1}^T$
for $t = 1, 2, \ldots, T$ **do**
$z_t = \mathrm{prox}_{\gamma_t g}(y_t)$
$x_t = \mathrm{prox}_{\gamma_t h}(2z_t - y_t - \gamma_t \nabla f(z_t))$
$y_{t+1} = y_t - z_t + x_t$
end for
Return: Draw τ uniformly at random from $\{1, 2, \ldots, T\}$ and output z_τ.

Theorem 1. *Consider Problem* (1) *under the following assumptions:*
(i) *The domain of g has finite diameter D_g,*

$$\|x - y\| \le D_g, \quad \forall x, y \in \mathrm{dom}(g).$$

(ii) *g is L_g-Lipschitz continuous on its domain,*

$$g(x) - g(y) \le L_g \|x - y\|, \quad \forall x, y \in \mathrm{dom}(g).$$

(iii) *The gradient of f is bounded by G_f on the domain of g,*

$$\|\nabla f(x)\| \le G_f, \quad \forall x \in \mathrm{dom}(g).$$

(iv) *h is L_h-Lipschitz continuous on $\mathbb{R}^n$,*

$$h(x) - h(y) \le L_h \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Choose $y_1 \in \mathrm{dom}(g)$. Then, z_τ returned by TOS (Algorithm 1) after T iterations with the fixed step-size $\gamma_t = \gamma = \frac{D_g}{2(G_f + L_g + L_h)T^{2/3}}$ satisfies

$$\mathbb{E}_\tau[\langle \nabla f(z_\tau), z_\tau - x\rangle + g(z_\tau) - g(x) + h(z_\tau) - h(x)] \le \frac{4D_g(G_f + L_g + L_h)}{T^{1/3}}, \qquad \forall x \in \mathrm{dom}(g). \tag{12}$$

Proof sketch. We start by writing the optimality conditions for the proximal steps on x_t and z_t. Through algebraic modifications, we show that, for all $x \in \mathrm{dom}(\phi)$,

$$\langle \nabla f(z_t), x_t - x\rangle + g(z_t) - g(x) + h(x_t) - h(x) \le \frac{1}{2\gamma}\Big(\|y_t - x\|^2 - \|y_{t+1} - x\|^2 - \|x_t - z_t\|^2\Big). \tag{13}$$

We take the average of this inequality over $t = 1, 2, \ldots T$. The inverted terms with y_t and y_{t+1} cancel out since the step-size is fixed. As a result, we know that (x_τ, z_τ) satisfy (8) with $\beta = D_g^2/(2\gamma T)$ in expectation.

We also need to show that (x_τ, z_τ) satisfy the proximity condition (7). To this end, we extend (13) as

$$-(G_f + L_h)\|x_t - z_t\| - (G_f + L_g + L_h)D_g \le \frac{1}{2\gamma}\Big(\|y_t - x\|^2 - \|y_{t+1} - x\|^2 - \|x_t - z_t\|^2\Big), \tag{14}$$

by using the boundedness of the domain *(i)*, boundedness of the gradient norm *(iii)*, and Lipschitz continuity of g and h *(ii, iv)*. Again, we take the average over t and eliminate the inverted terms. This leads to a second order inequality of $\mathbb{E}_\tau[\|z_\tau - x_\tau\|]$. By solving this inequality, we get an upper bound on $\mathbb{E}_\tau[\|z_\tau - x_\tau\|]$ in terms of the problem constants G_f, L_g, L_h, D_g, total number of iterations T, and step-size γ. By choosing γ carefully, we ensure that (x_τ, z_τ) are close and approximately stationary as per Definition 4. We complete the proof by using Observation 1 *(i)*. □

Our proof is a nontrivial extension of the convergence guarantees of TOS to the nonconvex problems. The prior analysis for the convex setting is based on a fixed point characterization of TOS and on Fejér monotonicity of $\|y_t - y_\star\|$, where $y_\star$ denotes the fixed point of TOS, see Proposition 2.1 in (Davis & Yin, 2017). Unfortunately. this desirable feature is lost when we drop the convexity of f. Our approach of proving proximity between x_τ and z_τ via second-order inequality (14) is nonstandard.

Remark 2. *We highlight several points about Theorem 1:*

1. *When D_g, G_f, L_g, or L_h is not known, one can use $\gamma_t = \frac{\gamma_0}{T^{2/3}}$ for any $\gamma_0 > 0$. The convergence rate in* (12) *still holds but with different constants. We chose the specific step-size in Theorem 1 in order to simplify the bounds.*
2. *Assumption* (iii) *holds automatically if f is smooth since $\mathrm{dom}(g)$ is bounded.*
3. *We can relax assumption* (iv) *as follows: h is L_h-Lipschitz continuous on $\mathrm{dom}(h)$, and $\mathrm{dom}(h) \supseteq \mathrm{dom}(g)$.*
4. *We can slightly tighten the constants in* (12)*. We defer the details to the supplementary material.*
5. *Our guarantees hold in expectation for the estimation at a randomly drawn iteration. This is a common technique in the nonconvex analysis. For example, see (Reddi et al., 2016b;a; Yurtsever et al., 2019) and the references therein.*

Corollary 1. *Consider Problem* (1) *under the following assumptions:*
(i) *g is the indicator function of a convex closed bounded set $\mathcal{G} \subseteq \mathbb{R}^n$ with a finite diameter $D_\mathcal{G} := \sup_{x,y\in\mathcal{G}} \|x - y\|$.*
(ii) *∇f is bounded on $\mathcal{G}$,* i.e., *$\|\nabla f(x)\| \le G_f$, $\forall x \in \mathcal{G}$.*
(iii) *h is L_h-Lipschitz continuous on $\mathbb{R}^n$.*

Choose $y_1 \in \mathcal{G}$. Then, z_τ returned by TOS after T iterations with the fixed step-size $\gamma_t = \frac{D_\mathcal{G}}{2(G_f + L_h)T^{2/3}}$ satisfies

$$\mathbb{E}_\tau[\langle \nabla f(z_\tau), z_\tau - x\rangle + h(z_\tau) - h(x)] \le \frac{4D_\mathcal{G}(G_f + L_h)}{T^{1/3}}, \quad \forall x \in \mathcal{G}. \tag{15}$$

Proof. Corollary 1 follows from Theorem 1 with $\mathrm{dom}(g) = \mathrm{dom}(\phi) = \mathcal{G}$. Assumptions *(i)* and *(ii)* in Theorem 1 hold with $D_g = D_\mathcal{G}$ and $L_g = 0$. We have $g(z_\tau) = g(x) = 0$ because z_τ and x belong to $\mathcal{G}$. □

Remark 3. *The ϵ-approximate solution (in expectation) that we consider in (12) and (15) reminds the Frank-Wolfe gap (in expectation) used in (Reddi et al., 2016b; Yurtsever et al., 2019). When h is missing and g is the indicator function, the Frank-Wolfe gap quantifies the error by $\mathbb{E}_\tau\big[\max_{x\in\mathcal{G}}\langle\nabla f(z_\tau), z_\tau - x\rangle\big]$. (15) holds for all $x \in \mathcal{G}$ so we can take the maximum over x and get the bound on $\max_{x\in\mathcal{G}}\mathbb{E}_\tau[\langle\nabla f(z_\tau), z_\tau - x\rangle + h(z_\tau) - h(x)]$. Note that $\max_{x\in\mathcal{G}}\mathbb{E}_\tau[\,\cdot\,] \le \mathbb{E}_\tau[\max_{x\in\mathcal{G}}(\,\cdot\,)]$. We leave the question whether similar guarantees hold for $\mathbb{E}_\tau[\max_{x\in\mathcal{G}}(\,\cdot\,)]$ open.*

Theorem 1 does not apply to Problem (2) because indicator functions fail Lipschitz continuity assumption *(iv)* in Theorem 1. The next theorem establishes convergence guarantees of TOS for Problem (2).

Theorem 2. *Consider Problem (2) under the following assumptions:*

(i) *$\mathcal{G} \subseteq \mathbb{R}^n$ is a bounded closed convex set with a finite diameter $D_\mathcal{G} := \sup_{x,y\in\mathcal{G}}\|x-y\|$.*

(ii) *∇f is bounded on $\mathcal{G}$,* i.e., *$\|\nabla f(x)\| \le G_f,\ \forall x \in \mathcal{G}$.*

(iii) *$\mathcal{H} \subseteq \mathbb{R}^n$ is a closed convex set.*

Then, z_τ returned by TOS (Algorithm 1) after T iterations with the fixed step-size $\gamma_t = \frac{D_\mathcal{G}}{2G_f T^{2/3}}$ satisfies

$$\mathbb{E}_\tau[\mathrm{dist}(z_\tau, \mathcal{H})] \le \frac{3D_\mathcal{G}}{T^{1/3}}, \tag{16}$$

$$\mathbb{E}_\tau[\langle\nabla f(z_\tau), z_\tau - x\rangle] \le \frac{4G_f D_\mathcal{G}}{T^{1/3}}, \quad \forall x \in \mathcal{G}\cap\mathcal{H}.$$

Proof sketch. The analysis is similar to the proof of Theorem 1. We use Observation 1 *(ii)* once we show that (x_τ, z_τ) are close and approximately stationary. □

3.1. Extensions for More Than Three Functions

Consider the extension of Problem (1) with an arbitrary number of nonsmooth terms (equivalently, an extension of Problem (2) with an arbitrary number of constraints):

$$\min_{x\in\mathbb{R}^n} \quad f(x) + \sum_{i=1}^m g_i(x). \tag{17}$$

One can solve this problem with TOS via a product-space formulation (see Section 6.1 in (Briceño-Arias, 2015)). We introduce slack variables $x^{(0)}, x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ and reformulate Problem (17) as

$$\begin{aligned}\min_{x^{(i)}\in\mathbb{R}^n} \quad & f(x^{(0)}) + \sum_{i=1}^m g_i(x^{(i)}) \\ \text{subj. to} \quad & x^{(0)} = x^{(1)} = \ldots = x^{(m)}.\end{aligned} \tag{18}$$

Problem (18) is an instance of Problem (1) in $\mathbb{R}^{(m+1)n}$. We can use TOS for solving this problem. Algorithm 2 in the supplementary material describes the algorithm steps.

4. Stochastic Nonconvex TOS

In this section, the differentiable term is the expectation of a function of a random variable, *i.e.*, $f(x) = \mathbb{E}_\xi \tilde{f}(x,\xi)$, where ξ is a random variable with distribution $\mathcal{P}$:

$$\min_{x\in\mathbb{R}^n} \quad \phi(x) := \mathbb{E}_\xi \tilde{f}(x,\xi) + g(x) + h(x). \tag{19}$$

This template covers a large number of applications in machine learning and statistics, including the finite-sum formulations that arise in M-estimation and empirical risk minimization problems.

In this setting, we replace $\nabla f(z_t)$ in Algorithm 1 with the following estimator:

$$u_t := \frac{1}{|Q_t|}\sum_{\xi\in Q_t}\nabla\tilde{f}(z_t,\xi), \tag{20}$$

where Q_t is a set of $|Q_t|$ *i.i.d.* samples from distribution $\mathcal{P}$.

Theorem 3. *Consider Problem (19). Instate the assumptions of Theorem 1. Further, assume that the following conditions hold:*

(v) *$\nabla\tilde{f}(x,\xi)$ is an unbiased estimator of $\nabla f(x)$,*

$$\mathbb{E}_\xi[\nabla\tilde{f}(x,\xi)] = \nabla f(x), \quad \forall x \in \mathbb{R}^n.$$

(vi) *$\nabla\tilde{f}(x,\xi)$ has bounded variance: For some $\sigma < +\infty$,*

$$\mathbb{E}_\xi[\|\nabla\tilde{f}(x,\xi) - \nabla f(x)\|^2] \le \sigma^2, \quad \forall x \in \mathbb{R}^n.$$

Consider TOS (Algorithm 1) with the stochastic gradient estimator (20) instead of $\nabla f(z_t)$. Choose the algorithm parameters

$$\gamma_t = \frac{D_g}{2(G_f + L_g + L_h)T^{2/3}} \quad \text{and}$$

$$|Q_t| = \left\lceil \frac{T^{2/3}}{2(G_f + L_g + L_h)^2} \right\rceil.$$

Then, z_τ returned by the algorithm after T iterations satisfies, $\forall x \in \mathrm{dom}(\phi)$,

$$\begin{aligned}&\mathbb{E}_\tau\mathbb{E}[\langle\nabla f(z_t), z_t - x\rangle + g(z_t) - g(x) + h(z_t) - h(x)] \\ &\le D_g(G_f + L_g + L_h)\left(\frac{\sqrt{4+2\sigma^2}}{T^{1/2}} + \frac{4+\sqrt{2}+\sigma^2}{T^{1/3}}\right).\end{aligned}$$

Similar to Corollary 1, we can specify guarantees for the case where g is an indicator function and h is L_h-Lipschitz continuous. We skip the details.

Next, analogous to Problem (2), we consider the nonconvex expectation minimization problem over the intersection of convex sets:

$$\min_{x\in\mathbb{R}^n} \quad f(x) := \mathbb{E}_\xi\tilde{f}(x,\xi) \quad \text{subj. to} \quad x \in \mathcal{G}\cap\mathcal{H}. \tag{21}$$

The next theorem presents convergence guarantees of TOS for this problem.

Theorem 4. *Consider Problem (21). Instate the assumptions of Theorems 2 and 3. Consider TOS (Algorithm 1) with the stochastic gradient estimator (20) instead of $\nabla f(z_t)$. Choose the algorithm parameters*

$$\gamma_t = \frac{D_{\mathcal{G}}}{2G_f T^{2/3}} \quad \text{and} \quad |Q_t| = \left\lceil \frac{T^{2/3}}{2G_f^2} \right\rceil.$$

Then, z_τ returned by the algorithm after T iterations satisfies, $\forall x \in \mathcal{G} \cap \mathcal{H}$,

$$\mathbb{E}_\tau \mathbb{E}[\mathrm{dist}(z_\tau, \mathcal{H})] \leq D_{\mathcal{G}} \left(\frac{\sqrt{4+2\sigma^2}}{T^{1/2}} + \frac{2+\sqrt{2}}{T^{1/3}} \right),$$

$$\mathbb{E}_\tau \mathbb{E}[\langle \nabla f(z_\tau), z_\tau - x \rangle] \leq G_f D_{\mathcal{G}} \left(\frac{\sqrt{4+2\sigma^2}}{T^{1/2}} + \frac{4+\sqrt{2}+\sigma^2}{T^{1/3}} \right).$$

Corollary 2. *Under the assumptions listed in Theorem 3 (resp. Theorem 4), TOS returns an ϵ-stationary point in expectation as per Definition 2 (resp. ϵ-feasible ϵ-stationary point as per Definition 3) after $T \leq \mathcal{O}(1/\epsilon^3)$ iterations. In total, this algorithm requires drawing $\mathcal{O}(1/\epsilon^5)$* i.i.d. *samples from distribution $\mathcal{P}$.*

Proof. $\epsilon \leq \mathcal{O}(1/T^{1/3})$ implies $T \leq \mathcal{O}(1/\epsilon^3)$ iteration complexity. At each iteration, we use $|Q_t| = \Omega(T^{2/3})$ stochastic gradients, so the total stochastic gradients complexity is $\sum_{t=0}^{T} |Q_t| = (T+1)|Q_t| = \Omega(T^{5/3}) \leq \mathcal{O}(1/\epsilon^5)$. ☐

Reducing the stochastic gradient complexity of TOS (Algorithm 1) via variance reduction techniques (see, for example, (Roux et al., 2012; Johnson & Zhang, 2013; Defazio et al., 2014; Nguyen et al., 2017; Fang et al., 2018)) can be a valuable extension. We leave this for a future study.

5. Numerical Experiments

This section demonstrates the empirical performance of TOS on the quadratic assignment problem (QAP).

QAP is a challenging formulation in the NP-hard problem class (Sahni & Gonzalez, 1976). We focus on the *relax-and-round* strategy proposed in (Vogelstein et al., 2015). This strategy requires solving a nonconvex optimization problem over the Birkhoff polytope (*i.e.*, the set of doubly stochastic matrices). First, we will summarize the main steps of this *relax-and-round* strategy and explain how we can use TOS in this procedure. Then, we will compare the performance of TOS against the Frank-Wolfe method (FW) (Frank & Wolfe, 1956; Jaggi, 2013; Lacoste-Julien, 2016) used in (Vogelstein et al., 2015).

5.1. Problem Description

Given the cost matrices A and $B \in \mathbb{R}^{n\times n}$, the goal in QAP is to align these matrices by finding a permutation matrix that minimizes a quadratic objective:

$$\begin{aligned} &\min_{X \in \mathbb{R}^{n\times n}} && \mathrm{trace}(AXB^\top X^\top) \\ &\text{subj. to} && X \in \{0,1\}^{n\times n}, \;\; X1_n = X^\top 1_n = 1_n, \end{aligned} \tag{22}$$

where 1_n denotes the n-dimensional vector of ones.

The challenge comes from the combinatorial nature of the feasible region. (22) is NP-Hard, so Vogelstein et al. (2015) focus on its continuous relaxation:

$$\begin{aligned} &\min_{X \in \mathbb{R}^{n\times n}} && \mathrm{trace}(AXB^\top X^\top) \\ &\text{subj. to} && X \in [0,1]^{n\times n}, \;\; X1_n = X^\top 1_n = 1_n. \end{aligned} \tag{23}$$

(23) is a quadratic optimization problem over the Birkhoff polytope. Remark that the quadratic objective is nonconvex in general.

The relax-and-round strategy of (Vogelstein et al., 2015) involves two main steps:
1. Finding a local optimal solution of (23).
2. Rounding the solution to the closest permutation matrix.

Solving (23). Projecting an arbitrary matrix onto the Birkhoff polytope is computationally challenging and the standard algorithms in the constrained nonconvex optimization literature are inefficient for (23).

Vogelstein et al. (2015) employ the FW algorithm to overcome this challenge. FW does not require projections. Instead, at each iteration, it requires solving a linear assignment problem (LAP). The arithmetic cost of LAP by using the Hungarian method or the Jonker-Volgenant algorithm (Kuhn, 1955; Munkres, 1957; Jonker & Volgenant, 1987) is $\mathcal{O}(n^3)$.

In this paper, we suggest TOS for solving (23) instead of FW. To apply TOS, we can split the Birkhoff polytope in two different ways.

One, we can consider the intersection of row-stochastic matrices and column-stochastic matrices:

$$\begin{aligned} \mathcal{G} &= \{X \in [0,1]^{n\times n} : X1_n = 1_n\} \\ \mathcal{H} &= \{X \in [0,1]^{n\times n} : X^\top 1_n = 1_n\}. \end{aligned} \tag{Split 1}$$

In this case, the projector onto $\mathcal{G}$ (resp., $\mathcal{H}$) requires projecting each row (resp., column) onto the unit simplex separately. The arithmetic cost of projecting each row (resp., column) is $\mathcal{O}(n)$ (Condat, 2016), and we can project multiple rows (resp., columns) in parallel.

Two, we can consider the following scheme studied in (Zass & Shashua, 2006; Lu et al., 2016; Pedregosa & Gidel, 2018):

$$\begin{aligned} \mathcal{G} &= [0,1]^{n\times n} \\ \mathcal{H} &= \{X \in \mathbb{R}^{n\times n} : X1_n = X^\top 1_n = 1_n\}. \end{aligned} \tag{Split 2}$$

In this case, the projection onto $\mathcal{G}$ truncates the entries and the projection onto $\mathcal{H}$ has a closed-form solution:

$$\mathrm{proj}_{\mathcal{H}}(X) = X + \left(\frac{1}{n}I + \frac{1_n^\top X 1_n}{n^2} I - \frac{1}{n}X\right) 1_n 1_n^\top - \frac{1}{n} 1_n 1_n^\top X,$$

where I denotes the identity matrix. We present a derivation of this projection operator in the supplementary material.

Rounding. The solution of (23) does not immediately yield a feasible point for QAP (22). We need a rounding step.

Suppose X_τ is a solution to (23). A natural strategy is choosing the closest permutation matrix to X_τ. We can find this permutation matrix by solving

$$\begin{aligned} \max_{X \in \mathbb{R}^{n\times n}} \quad & \langle X_\tau, X\rangle \\ \text{subj. to} \quad & X \in [0,1]^{n\times n}, \ \ X 1_n = X^\top 1_n = 1_n. \end{aligned} \tag{24}$$

We present the derivation of this folklore formulation in the supplementary material. (24) is an instance of LAP. Hence, it can be solved in $\mathcal{O}(n^3)$ arithmetic operations via the Hungarian method or the Jonker-Volgenant algorithm.

5.2. Numerical Results

Implementation details. For FW, we use the exact line-search (greedy) step-size as in (Vogelstein et al., 2015). For solving LAP, we employ an efficient implementation of the Hungarian method (Ciao, 2011).

For TOS (Algorithm 1), we output the last iterate instead of the random variable x_τ. We use $\gamma_t = 1/L_f$ step-size (L_f denotes the smoothness constant of f) instead of the more conservative step-size that our theory suggests (which depends on T). $1/L_f$ is the standard rule in convex optimization, and in our experience, it works well for nonconvex problems too.

We start both methods from the same initial point y_1. We construct y_1 by projecting a random matrix with *i.i.d.* standard Gaussian entries onto the Birkhoff polytope via 1000 iterations of the alternating projections method.

Quality of solution. Given a prospective solution $X_t \in \mathcal{G}$, we compute the following errors:

$$\text{infeasibility err.} = \frac{\mathrm{dist}(X_t, \mathcal{H})}{\sqrt{n}} \tag{25}$$

$$\text{nonstationarity err.} = \frac{|\max_{X\in\mathcal{G}\cap\mathcal{H}} \langle \nabla f(X_t), X_t - X\rangle|}{\max\{f(X_t), 1\}}$$

Infeasibility error is always 0 for FW. We evaluate these errors only at iterations $t = 1, 2, 4, 8, \ldots$ to avoid extra computation.

We evaluate the quality of the rounded solution $\tilde{X}_t$ by using the following formula:

$$\text{assignment err.} = \frac{f(\tilde{X}_t) - f(\tilde{X}_{\text{best}})}{\max\{f(\tilde{X}_{\text{best}}), 1\}}, \tag{26}$$

where $\tilde{X}_{\text{best}}$ is the best solution known for (22). $\tilde{X}_{\text{best}}$ is unknown in normal practice, but it is available for the QAPLIB benchmark problems.

Observations. Figure 1 compares the empirical performance of TOS and FW for solving (23) with `chr12a` and `esc128` datasets from QAPLIB. In particular, *TOS exhibits locally linear convergence*, whereas FW converges with sublinear rates. We observed qualitatively similar behavior also with the other datasets in QAPLIB.

Computing the gradient dominates the runtime of TOS. Instead, for FW, the bottleneck is solving the LAP subproblems. As a result, TOS is especially advantageous against FW when A and B are sparse.

Next, we examine the quality of the rounded solutions we obtain after solving (23) with TOS and FW. We initialize both methods from the same point and we stop them at the same level of accuracy, when infeasibility and nonstationarity errors drop below 10^{-5} (recall that the infeasibility error is always 0 for FW). We round the final estimates to the closest permutation matrix and evaluate the assignment error (26). Figure 2 presents the results of this experiment for the 134 datasets in QAPLIB.

Remarkably, TOS gets a *better* solution on 83 problems; TOS and FW perform the same on 16; and FW outperforms TOS on 35 instances. The largest margin appears on the `chr15b` dataset where TOS scores 0.744 lower assignment error than FW. On the other extreme, the assignment error of the FW solution is 0.253 lower than TOS on the `chr15c` dataset. On average (over datasets), TOS outperforms FW in assignment error by a margin of 0.046.

Computational environment. Experiments are performed in MATLAB R2018a on a MacBook Pro Late 2013 with 2.6 GHz Quad-Core Intel Core i7 CPU and 16 GB 1600 MHz DDR3 memory. The source code is available online[1].

Other solvers for QAP. The literature covers numerous approaches for tackling QAP, including *(i)* exact solution methods with branch-and-bound, dynamic programming, and cutting plane methods, *(ii)* heuristics and metaheuristics based on local and tabu search, simulated annealing, genetic algorithms, and neural networks, and *(iii)* lower bound approximation methods via spectral bounds, mixed-integer linear programming, and semidefinite programming relaxations. An extensive comparison with these methods is beyond the scope of our paper. We refer to the comprehensive survey of Loiola et al. (2007) for more details.

[1] https://github.com/alpyurtsever/NonconvexTOS

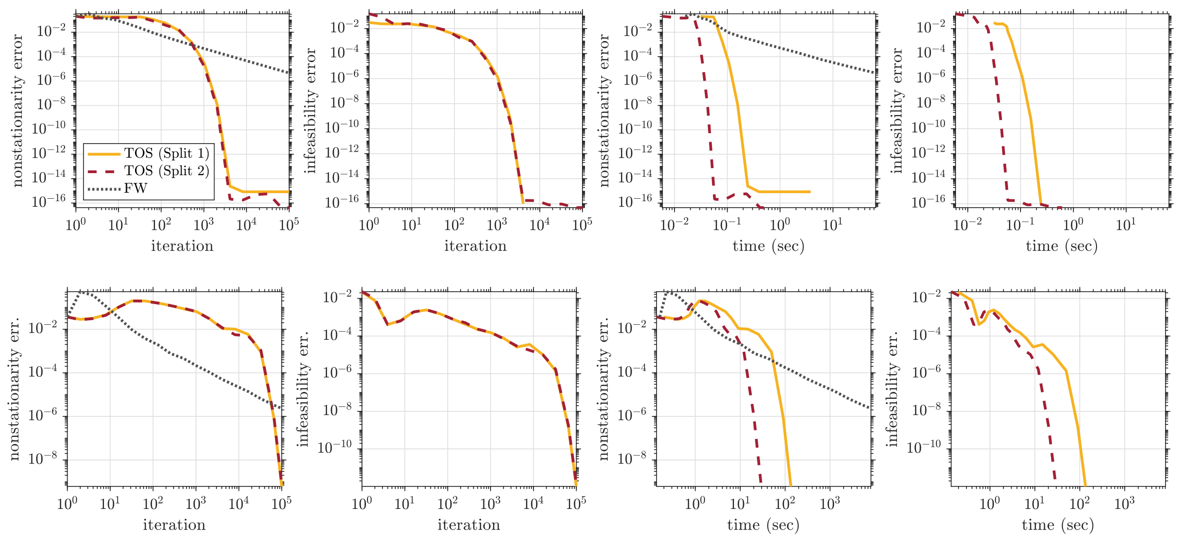

Figure 1. Empirical convergence of TOS for two different formulations ((Split 1) and (Split 2)) compared against FW for solving the relaxed QAP formulation (23). The [top] row corresponds to the results for the chr12a dataset and the [bottom] row for the esc128 dataset (from QAPLIB). In both cases, TOS exhibits locally linear convergence whereas FW converges sublinearly.

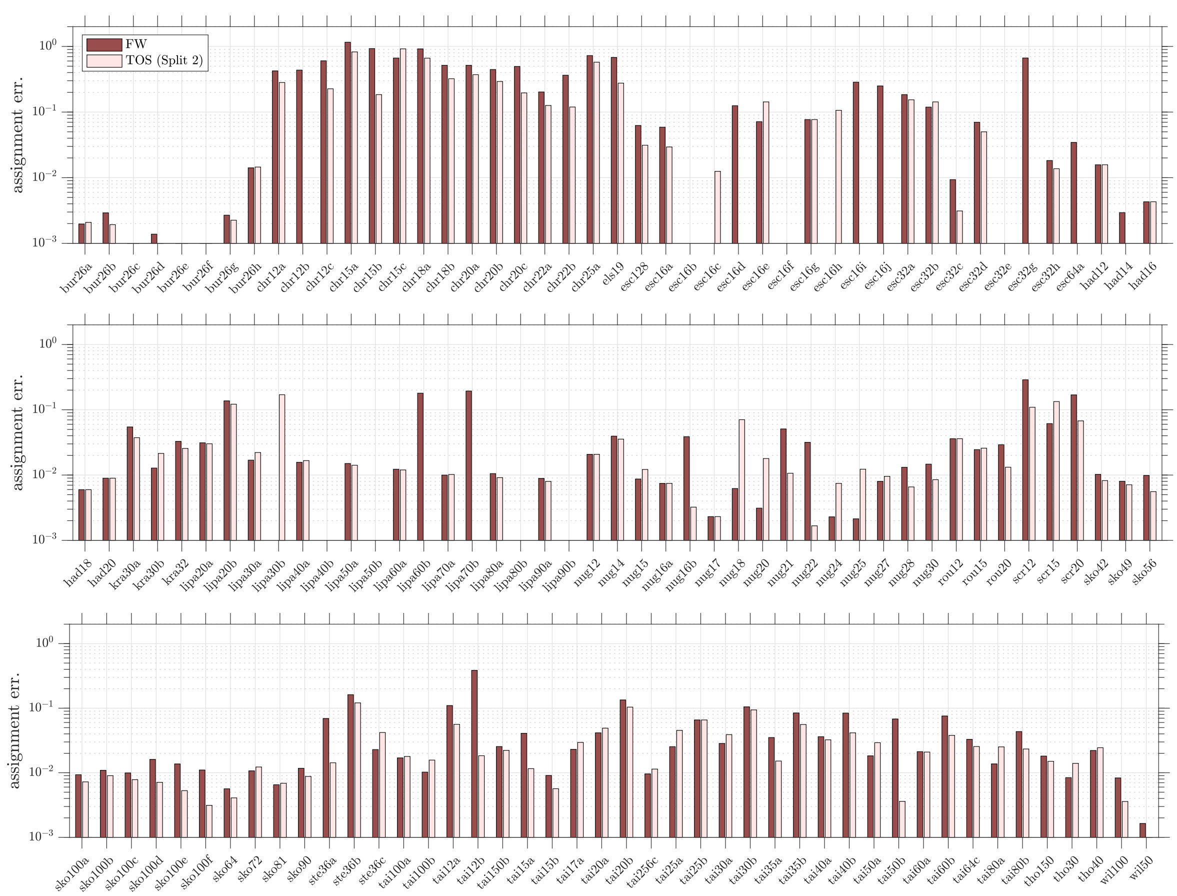

Figure 2. Assignment cost (see (26)) achieved by FW and TOS with the relax-and-round strategy for solving QAP. Smaller values are better, zero means a perfect estimation. Out of 134 QAP instances in the QAPLIB library, TOS outperforms FW on 83 problems; FW is better on 35; and the two methods get the same results on 16 instances.

6. Conclusions

We establish the convergence guarantees of TOS for minimizing the sum of three functions, one differentiable but potentially nonconvex and two convex but potentially nonsmooth. In contrast with the existing results, our analysis permits both nonsmooth terms to be indicator functions. Moreover, we extend our analysis for stochastic problems where we have access only to an unbiased estimator of the gradient of the differentiable term.

We present numerical experiments on QAPs. The empirical performance of the proposed method is promising. In our experience, the method converges to a stationary point with locally linear rates.

We conclude our paper with a short list of open questions and follow-up directions:

(i) We assume that $\mathrm{dom}(g)$ is bounded. This assumption is needed in our analysis since Definition 2 requires (4) to hold for all x in $\mathrm{dom}(g)$. We can potentially drop this assumption by adopting a relaxed notion of stationarity where the inequality holds only on a feasible neighborhood of $\bar{z}$. Such measures are used in recent works for different problem models, *e.g.*, see Definition 2.3 in (Nouiehed et al., 2019) and Definition 1 in (Song et al., 2020).

(ii) We did not explicitly use the smoothness of the differentiable term in our analysis. One can potentially derive tighter guarantees by using the smoothness or under additional assumptions such as the Kurdyka-Łojasiewicz property.

(iii) For the stochastic setting, we can improve the stochastic gradient complexity by using variance reduction techniques.

(iv) Developing an efficient implementation (that benefits from parallel computation) with an aim to investigate the full potential of TOS for solving QAP and other nonconvex problems such as the constrained and regularized neural networks is an important piece of future work.

Acknowledgements

AY received support from the Early Postdoc.Mobility Fellowship P2ELP2_187955 from Swiss National Science Foundation and partial postdoctoral support from NSF-CAREER grant IIS-1846088. SS acknowledges support from an NSF BIGDATA grant (1741341) and an NSF CAREER grant (1846088).

References

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

Bian, F. and Zhang, X. A three-operator splitting algorithm for nonconvex sparsity regularization. *arXiv preprint arXiv:2006.08951*, 2020.

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

Burkard, R. E., Karisch, S. E., and Rendl, F. QAPLIB–a quadratic assignment problem library. *Journal of Global optimization*, 10(4):391–403, 1997.

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.

Ciao, Y. Hungarian algorithm for linear assignment problems (v2.3). (Retrieved December 6, 2019), 2011. URL https://www.mathworks.com/matlabcentral/fileexchange/20652-hungarian-algorithm-for-linear-assignment-problems-v2-3.

Condat, L. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1):575–585, 2016.

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

De Berg, M., Van Nijnatten, F., Sitters, R., Woeginger, G. J., and Wolff, A. The traveling salesman problem under squared Euclidean distances. *arXiv preprint arXiv:1001.0236*, 2010.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

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

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2): 95–110, 1956.

He, B. and Yuan, X. On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. *Numerische Mathematik*, 130(3):567–577, 2015.

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.

Jacob, L., Obozinski, G., and Vert, J.-P. Group lasso with overlap and graph lasso. In *Proceedings of the 26th International Coference on Machine Learning*, pp. 433–440, 2009.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435. PMLR, 2013.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

Jonker, R. and Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

Koopmans, T. C. and Beckmann, M. Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pp. 53–76, 1957.

Kuhn, H. W. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

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.

Lacoste-Julien, S. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Liu, Y. and Yin, W. An envelope for Davis–Yin splitting and strict saddle-point avoidance. *Journal of Optimization Theory and Applications*, 181(2):567–587, 2019.

Loiola, E. M., de Abreu, N. M. M., Boaventura-Netto, P. O., Hahn, P., and Querido, T. A survey for the quadratic assignment problem. *European journal of operational research*, 176(2):657–690, 2007.

Lu, Y., Huang, K., and Liu, C.-L. A fast projected fixed-point algorithm for large graph matching. *Pattern Recognition*, 60:971–982, 2016.

Malitsky, Y. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, pp. 1–28, 2019.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. *Robust statistics: Theory and methods*. John Wiley & Sons, 2019.

McLachlan, G. J. and Krishnan, T. *The EM Algorithm and Extentions*. Wiley, 1996.

Munkres, J. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pp. 2613–2621. PMLR, 2017.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Ollila, E. and Tyler, D. E. Regularized m-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070, 2014.

Patrinos, P., Stella, L., and Bemporad, A. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In *53rd IEEE Conference on Decision and Control*, pp. 4234–4239. IEEE, 2014.

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.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pp. 314–323. PMLR, 2016a.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. Stochastic Frank-Wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1244–1251. IEEE, 2016b.

Richard, E., Savalle, P.-A., and Vayatis, N. Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Coference on Machine Learning*, pp. 51–58, 2012.

Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems*, 25, 2012.

Sahni, S. and Gonzalez, T. P-complete approximation problems. *Journal of the ACM (JACM)*, 23(3):555–565, 1976.

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.

Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33, 2020.

Themelis, A., Stella, L., and Patrinos, P. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

Vogelstein, J. T., Conroy, J. M., Lyzinski, V., Podrazik, L. J., Kratzer, S. G., Harley, E. T., Fishkind, D. E., Vogelstein, R. J., and Priebe, C. E. Fast approximate quadratic programming for graph matching. *PLOS one*, 10(4): e0121002, 2015.

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., Sra, S., and Cevher, V. Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pp. 7282–7291. PMLR, 2019.

Zaslavskiy, M., Bach, F., and Vert, J.-P. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (12):2227–2242, 2008.

Zass, R. and Shashua, A. Doubly stochastic normalization for spectral clustering. *Advances in neural information processing systems*, 19, 2006.

Zass, R. and Shashua, A. Nonnegative sparse PCA. In *Advances in neural information processing systems*, pp. 1561–1568, 2007.

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.

Zong, C., Tang, Y., and Cho, Y. J. Convergence analysis of an inexact three-operator splitting algorithm. *Symmetry*, 10(11):563, 2018.