Inexact-Proximal Accelerated Gradient Method for Stochastic Nonconvex Constrained Optimization Problems

Morteza Boroun, Afrooz Jalilzadeh*

Abstract

Stochastic nonconvex optimization problems with nonlinear constraints have a broad range of applications in intelligent transportation, cyber-security, and smart grids. In this paper, first, we propose an inexact-proximal accelerated gradient method to solve a nonconvex stochastic composite optimization problem where the objective is the sum of smooth and nonsmooth functions, the constraint functions are assumed to be deterministic and the solution to the proximal map of the nonsmooth part is calculated inexactly at each iteration. We demonstrate an asymptotic sublinear rate of convergence for stochastic settings using increasing sample-size considering the error in the proximal operator diminishes at an appropriate rate. Then we customize the proposed method for solving stochastic nonconvex optimization problems with nonlinear constraints and demonstrate a convergence rate guarantee. Numerical results show the effectiveness of the proposed algorithm.

1 INTRODUCTION

There is a rapid growth in the global urban population and the concept of smart cities is proposed to manage the impact of this surge in urbanization. Intelligent transportation, cyber-security, and smart grids are playing vital roles in smart city projects which are highly influenced by big data analytic and effective use of machine learning techniques [19]. As data gets more complex and applications of machine learning algorithms for decision-making broaden and diversify, recent research has been shifted to constrained optimization problems with nonconvex objectives [14] to improve efficiency and scalability in smart city projects.

Consider the following constrained optimization problem with a stochastic and nonconvex objective:

$$\begin{aligned} \min_{x\in X} \quad & f(x) \triangleq \mathbb{E}[F(x,\zeta(\omega))] \\ \text{s.t.} \quad & \phi_i(x) \leq 0, \quad i=1,\ldots,m, \end{aligned} \tag{1}$$

where $\zeta : \Omega \to \mathbb{R}^o$, $F : \mathbb{R}^n \times \mathbb{R}^o \to \mathbb{R}$, and $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the associated probability space. We consider function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is smooth and possibly nonconvex, $\phi_i(x) : \mathbb{R}^n \to \mathbb{R}$ are deterministic, convex, and smooth for all i, and set X is convex and compact. To solve this problem, first we propose an algorithm for solving the following composite optimization problem

$$\min_{x\in\mathbb{R}^n} g(x) \triangleq f(x) + h(x), \tag{2}$$

*Department of Systems and Industrial Engineering, The University of Arizona, 1127 E James Rogers Way, Tucson, AZ. Email: morteza@email.arizona.edu, afrooz@arizona.edu.

where $h(x) : \mathbb{R}^n \to \mathbb{R}$ is a convex function and possibly nonsmooth. Using the indicator function $\mathbb{I}_\Theta(\cdot)$, where $\mathbb{I}_\Theta(x) = 0$ if $x \in \Theta$ and $\mathbb{I}_\Theta(x) = +\infty$ if $x \notin \Theta$, one can write problem (1) in the form of problem (2) by choosing $h(x) = \mathbb{I}_\Theta(x)$ and $\Theta = \{x \mid x \in X,\ \phi_i(x) \leq 0,\ \forall i = 1, \ldots, m\}$. Moreover, we show that how to customize the proposed method to solve problem (1). Indeed, proximal-gradient methods are an appealing approach for solving (2) due to their computational efficiency and fast theoretical convergence guarantee. In deterministic and convex regime, subgradient methods have been shown to have a convergence rate of $\mathcal{O}(1/\sqrt{T})$, however, proximal-gradient methods can achieve a faster rate of $\mathcal{O}(1/T)$, where T is the total number of iterations. Each iteration of a proximal-gradient method requires solving the following:

$$\operatorname{prox}_{\gamma,h}(y) = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}}\{h(u) + \frac{1}{2\gamma}\|u - y\|^2\}. \tag{3}$$

In many scenarios, computing the exact solution of the proximal operator may be expensive or may not have an analytic solution. In this work, we propose a gradient-based scheme to solve the nonconvex optimization problem (2) by computing the proximal operator inexactly at each iteration.

Next, we introduce important notations that we use throughout the paper and then briefly summarize the related research.

1.1 Notations

We denote the optimal objective value (or solution) of (2) by g^* (or x^*) and the set of the optimal solutions by X^*, which is assumed to be nonempty. For any $a \in \mathbb{R}$, we define $[a]_+ = \max\{0, a\}$. $\mathbb{E}[\bullet]$ denotes the expectation with respect to the probability measure $\mathbb{P}$ and $\mathcal{B}(s) = \{x \in \mathbb{R}^n \mid \|x\| \leq s\}$. $\Pi_\Theta(\cdot)$ denotes the projection onto convex set Θ and $\mathbf{relint}(X)$ denotes the relative interior of the set X. Throughout the paper, $\tilde{\mathcal{O}}$ is used to suppress all the logarithmic terms.

1.2 Related Works

There has been a lot of studies on first-order methods for convex optimization with convex constraints, see [18, 20] for deterministic constraints and [1, 10] for stochastic constraints. Nonconvex optimization problems without constraints or with easy-to-compute projection on the constraint set have been studied by [3, 21, 9]. When the function f in problem (2) is convex and h is a nonsmooth function, [17] showed that even with errors in the computation of the gradient and the proximal operator, the inexact proximal-gradient method achieves the same convergence rates as the exact counterpart, if the magnitude of the errors is controlled in an appropriate rate. In nonconvex setting, assuming the proximal operator has an exact solution, [4] obtained a convergence rate of $\mathcal{O}(1/T)$, using accelerated gradient scheme for deterministic problems and in stochastic regime using increasing sample-size they obtained the same convergence rate. Inspired by these two works, we present accelerated inexact proximal-gradient framework that can solve problems (1) and (2). In deterministic regime, [8] analyzed the iteration-complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs with iteration complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$. Inexact proximal-point penalty method introduced by [13] and [11] can solve nonlinear constraints with complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ and $\tilde{\mathcal{O}}(\epsilon^{-3})$ for affine equality constraints and nonconvex constraints, respectively. Recently, [12] showed complexity result of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ for deterministic problems with nonconvex objective and convex constraints with nonlinear functions to achieve ϵ-KKT point. In stochastic regime, [2] has studied functional constrained optimization problems and obtained a non-asymptotic convergence rate of $\mathcal{O}(\epsilon^{-2})$ for stochastic problems with convex constraints to achieve ϵ^2-KKT point. In this paper, we obtain the same convergence rate under weaker assumptions. In particular, in contrast to [2], our analysis does not require the objective function to be Lipschitz and we prove an asymptotic convergence rate result. Next, we outline the contributions of our paper.

1.3 Contributions

In this paper, we consider a stochastic nonconvex optimization problem with convex nonlinear constraints. We propose an inexact proximal accelerated gradient (IPAG) method where at each iteration the projection onto the nonlinear constraints is solved inexactly. By improving the accuracy of the approximate solution of the proximal subproblem (projection step) at an appropriate rate and ensuring feasibility at each iteration combined with a variance reduction technique, we demonstrate a convergence rate of $\mathcal{O}(1/T)$, where T is the total number of iterations, and the oracle complexity (number of sample gradients) of $\mathcal{O}(1/\epsilon^2)$ to achieve an ϵ-first-order optimality of problem (1). To accomplish this task, first we analyze the proposed method for the composite optimization problem (2) which can be specialized to (1) using an indicator function. Moreover, our proposed method requires weaker assumptions compare to [2].

Next, we state the main definitions and assumptions that we need for the convergence analysis. In Section 2, we introduce the IPAG algorithm to solve the composite optimization problem and then in Section 2.1 we show that IPAG method can be customized to solve a nonconvex stochastic optimization problem with nonlinear constraints (1). Finally, in section 3 we present some empirical experiments to show the benefit of our proposed scheme in comparison with a competitive scheme.

1.4 Assumptions and Definitions

Let ρ be the error in the calculation of the proximal objective function achieved by $\tilde{x}$, i.e.,

$$\frac{1}{2\gamma}\|\tilde{x}-y\|^2 + h(\tilde{x}) \leq \rho + \min_{x\in\mathbb{R}^n}\left\{\frac{1}{2\gamma}\|x-y\|^2 + h(x)\right\}, \tag{4}$$

and we call $\tilde{x}$ a ρ-approximate solution to the proximal problem. Next, we define ρ-subdifferential and then we state a lemma to characterize the elements of the ρ-subdifferential of h at x.

Definition 1 (ρ-subdifferential)**.** *Given a convex function $h(x):\mathbb{R}^n\to\mathbb{R}$ and a positive scalar ρ, the ρ-approximate subdifferential of $h(x)$ at a point $x\in\mathbb{R}^n$, denoted as $\partial_\rho h(x)$, is*

$$\partial_\rho h(x) = \{d\in\mathbb{R}^n : h(y)\geq h(x) + \langle d, y-x\rangle - \rho\}.$$

Therefore, when $d\in\partial_\rho h(x)$, we say that d is a ρ-subgradient of $h(x)$ at point x.

Lemma 1. *If $\tilde{x}$ is a ρ-approximate solution to the proximal problem* (3) *in the sense of* (4)*, then there exists v such that $\|v\|\leq\sqrt{2\gamma\rho}$ and*

$$\tfrac{1}{\gamma}\left(y-\tilde{x}-v\right)\in\partial_\rho h(\tilde{x}).$$

Proof of Lemma 1 can be found in [17]. Throughout the paper, we exploit the following basic lemma.

Lemma 2. *Given a symmetric positive definite matrix Q, we have the following for any ν_1, ν_2, ν_3:*

$$(\nu_2 - \nu_1)^T Q(\nu_3 - \nu_1) = \frac{1}{2}(\|\nu_2 - \nu_1\|_Q^2 + \|\nu_3 - \nu_1\|_Q^2 - \|\nu_2 - \nu_3\|_Q^2), \text{ where } \|\nu\|_Q \triangleq \sqrt{\nu^T Q \nu}.$$

In our analysis we use the following lemma [4].

Lemma 3. *Given a positive sequence α_k, define $\Gamma_k = 1$ for $k = 1$ and $\Gamma_k = (1 - \alpha_k)\Gamma_{k-1}$ for $k > 1$. Suppose a sequence $\{\chi_k\}_k$ satisfies $\chi_k \leq (1 - \alpha_k)\chi_{k-1} + \lambda_k$, where $\lambda_k > 0$. Then for any $k \geq 1$, we have that $\chi_k \leq \Gamma_k \sum_{j=1}^k \gamma_j / \Gamma_j$.*

The following assumptions are made throughout the paper.

Assumption 1. *The following statements hold:*

(i) A slater point of problem (1) is available, i.e., there exists $x^\circ \in \mathbb{R}^n$ such that $\phi_i(x^\circ) < 0$ for all $i = 1, \ldots, m$ and $x^\circ \in \mathbf{relint}(X)$.

(ii) Function f is smooth and weakly-convex with Lipschitz continuous gradient, i.e. there exists $L, \ell \geq 0$ such that $-\frac{\ell}{2}\|y - x\|^2 \leq f(x) - f(y) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2$.

(iii) There exists $C > 0$ such that $\|prox_{\gamma,h}(y)\| \leq C$ for any $\gamma > 0$ and $y \in \mathbb{R}^n$.

(iv) $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$ holds a.s., where $\xi_k \triangleq \nabla f(z_k) - \nabla F(z_k, \omega_k)$. Also, there exists $\tau > 0$ such that $\mathbb{E}[\|\bar{\xi}_k\|^2 \mid \mathcal{F}_k] \leq \frac{\tau^2}{N_k}$ holds a.s. for all k and $\mathcal{F}_k \triangleq \sigma\left(\{z_0, \bar{\xi}_0, z_1, \bar{\xi}_1 \ldots, z_{k-1}, \bar{\xi}_{k-1}\}\right)$, where $\bar{\xi}_k \triangleq \frac{\sum_{j=1}^{N_k} \nabla f(z_k) - \nabla F(z_k, \omega_{j,k})}{N_k}$.

Note that Assumption 1 is a common assumption in nonconvex and stochastic optimization problems and it holds for many real-world problems such as problem of non-negative principal component analysis and classification problem with nonconvex loss functions [16].

2 CONVERGENCE ANALYSIS

In this section, we propose an inexact-proximal accelerated gradient scheme for solving problem (2) assuming that an inexact solution to the proximal subproblem exists through an inner algorithm $\mathcal{M}$. Later in section 2.1, we show that how the inexact solution can be calculated at each iteration for problem (1). Since problem (2) is nonconvex, we demonstrate the rate result in terms of $\|z - \text{prox}_{\lambda h}(z - \lambda \nabla f(z))\|$ which is a standard termination criterion for solving constrained or composite nonconvex problems [15, 5, 4]. For problem (1), the first-order optimality condition is equivalent to find z^* such that $z^* = \Pi_\Theta(z^* - \lambda \nabla f(z^*))$ for some $\lambda > 0$. Hence, we show the convergence result in terms of ϵ-first-order optimality condition for a vector z, i.e., $\|z - \Pi_\Theta(z - \lambda \nabla f(z))\|^2 \leq \epsilon$.

Assumption 2. *For a given $c \in \mathbb{R}^n$ and $\gamma > 0$, consider the problem $\tilde{u} \triangleq prox_{\gamma h}(c)$. An algorithm $\mathcal{M}$ with an initial point u_0, output u and convergence rate of $\mathcal{O}(1/t^2)$ within t steps exists, such that $\|u - \tilde{u}\|^2 \leq (a_1\|u_0 - \tilde{u}\|^2 + a_2)/t^2$ for some $a_1, a_2 > 0$.*

Algorithm 1 Inexact-proximal Accelerated Gradient Algorithm (IPAG)

input: $x_0, y_0 \in \mathbb{R}^n$, positive sequences $\{\alpha_k, \gamma_k, \lambda_k\}_k$ and Algorithm $\mathcal{M}$ satisfying Assumption 2;
for $k = 1 \ldots T$ **do**
(1) $z_k = (1-\alpha_k)y_{k-1} + \alpha_k x_{k-1}$;
(2) $x_k \approx \text{prox}_{\gamma_k h}\left(x_{k-1} - \gamma_k(\nabla f(z_k) + \bar{\xi}_k)\right)$ (solved inexactly by algorithm $\mathcal{M}$ with q_k iterations);
(3) $y_k \approx \text{prox}_{\lambda_k h}\left(z_k - \lambda_k(\nabla f(z_k) + \bar{\xi}_k)\right)$ (solved inexactly by algorithm $\mathcal{M}$ with p_k iterations);
end for
Output: z_N where N is randomly selected from $\{T/2, \ldots, T\}$ with $\text{Prob}\{N = k\} = \frac{1}{\sum_{k=\lfloor T/2 \rfloor}^{T} \frac{1-L\lambda_k}{16\lambda_k\Gamma_k}}\left(\frac{1-L\lambda_k}{16\lambda_k\Gamma_k}\right)$.

Suppose the solutions of proximal operators $\tilde{x}_k \triangleq \text{prox}_{\gamma_k h}\left(x_{k-1} - \gamma_k(\nabla f(z_k) + \bar{\xi}_k)\right)$ and $\tilde{y}_k \triangleq \text{prox}_{\lambda_k h}(z_k - \lambda_k(\nabla f(z_k) + \bar{\xi}_k))$ are not available exactly, instead an e_k-subdifferential solution x_k and ρ_k-subdifferential solution y_k are available, respectively. In particular, given $\bar{\xi}_k$ for the proximal subproblem in step (2) and (3) of Algorithm 1 at iteration k, Assumption 2 immediately implies that after q_k and p_k steps of Algorithm $\mathcal{M}$ with initial point x_{k-1} and y_{k-1}, we have $e_k = \gamma_k(c_1\|x_{k-1} - \tilde{x}_k\|^2 + c_2)/q_k^2$ and $\rho_k = \lambda_k(b_1\|y_{k-1} - \tilde{y}_k\|^2 + b_2)/p_k^2$, for some $c_1, c_2, b_1, b_2 > 0$ where γ_k, λ_k represents strong convexity of the subproblems, respectively. Later, in Section 2.1, we show the existence of Algorithm $\mathcal{M}$ such that it satisfies Assumption 2.

Remark 1. *Note that from Assumption 1(iii) and 2, we can show the following for all $k > 0$:*

$$\|x_k - \tilde{x}_k\|^2 \leq \tfrac{1}{q_k^2}\left[2c_1(\|x_{k-1} - \tilde{x}_{k-1}\|^2 + \|\tilde{x}_{k-1} - \tilde{x}_k\|^2) + c_2\right] \leq \|x_{k-1} - \tilde{x}_{k-1}\|^2 + \tfrac{8C^2+c_2}{q_k^2}$$
$$\implies \|x_k - \tilde{x}_k\|^2 \leq \|x_0 - \tilde{x}_0\|^2 + \sum_{j=1}^{k} \tfrac{8C^2+c_2}{q_j^2} \implies \|x_k\| \leq C + \sqrt{\|x_0 - \tilde{x}_0\|^2 + \tilde{C}} \triangleq B_1, \quad (5)$$

where $\tilde{C} \triangleq \sum_{j=1}^{k} \frac{8C^2+c_2}{q_j^2}$ and we used the fact that $\|\tilde{x}_k\| \leq C$. Similarly for step (3) of Algorithm 1, there exist $B_2, B_3 > 0$ such that the followings hold for all $k > 0$,

$$\|y_k\| \leq B_2, \qquad \|z_k\| \leq B_3. \quad (6)$$

Next, we state our main lemma that provides a bridge towards driving rate statements.

Lemma 4. *Consider Algorithm 1 and suppose Assumption 1 and 2 hold and choose stepsizes α_k, γ_k and λ_k such that $\alpha_k\gamma_k \leq \lambda_k$. Let $\hat{y}_k \approx prox_{\lambda_k h}(z_k - \lambda_k\nabla f(z_k))$ in the sense of (4) and $\hat{y}_k^r \triangleq prox_{\lambda_k h}(z_k - \lambda_k\nabla f(z_k))$ for any $k \geq 1$, then the following holds for all $T > 0$.*

$$\mathbb{E}[\|\hat{y}_N - z_N\|^2 + \|\hat{y}_N^r - z_N\|^2]$$
$$\leq \left(\sum_{k=\lfloor T/2 \rfloor}^{T} \tfrac{1-L\lambda_k}{16\lambda_k\Gamma_k}\right)^{-1} \left[\tfrac{\alpha_1}{2\gamma_1\Gamma_1}\|x_0 - x^*\|^2 + \tfrac{\ell}{2}\sum_{k=1}^{T} \tfrac{\alpha_k}{\Gamma_k}\left[2B_3^2 + C^2 + \alpha_k(1-\alpha_k)(2B_2^2 + B_1^2)\right]\right.$$
$$\left. + \sum_{k=1}^{T}\left(\tfrac{\lambda_k\tau^2}{\Gamma_k N_k(1-L\lambda_k)} + \tfrac{2e_k}{\Gamma_k} + \tfrac{B_1^2+C^2}{\gamma_k\Gamma_k} + \tfrac{\rho_k(1+k)}{\Gamma_k} + \tfrac{B_1^2+B_2^2}{k\lambda_k\Gamma_k} + \tfrac{5\lambda_k\tau^2(1-L\lambda_k)}{8\Gamma_k N_k} + \tfrac{\rho_k(1-L\lambda_k)}{\lambda_k^2\Gamma_k}\right)\right]. \quad (7)$$

Proof. First of all from the fact that $\nabla f(x)$ is Lipschitz, for any $k \geq 1$, the following holds:

$$f(y_k) \leq f(z_k) + \langle \nabla f(z_k), y_k - z_k \rangle + \tfrac{L}{2}\|y_k - z_k\|^2. \tag{8}$$

Using Assumption 1(ii), for any $\alpha_k \in (0,1)$ one can obtain the following:

$$\begin{aligned}
&f(z_k) - [(1-\alpha_k)f(y_{k-1}) + \alpha_k f(x)] \\
&= \alpha_k[f(z_k) - f(x)] + (1-\alpha_k)[f(z_k) - f(y_{k-1})] \\
&\leq \alpha_k[\langle \nabla f(z_k), z_k - x\rangle + \tfrac{\ell}{2}\|z_k - x\|^2] + (1-\alpha_k)[\langle \nabla f(z_k), z_k - y_{k-1}\rangle + \tfrac{\ell}{2}\|x_u - y_{k-1}\|^2] \\
&= \langle \nabla f(z_k), z_k - \alpha_k x - (1-\alpha_k)y_{k-1}\rangle + \tfrac{\ell\alpha_k}{2}\|z_k - x\|^2 + \tfrac{\ell(1-\alpha_k)}{2}\|z_k - y_{k-1}\|^2 \\
&\leq \langle \nabla f(z_k), z_k - \alpha_k x - (1-\alpha_k)y_{k-1}\rangle + \tfrac{\ell\alpha_k}{2}\|z_k - x\|^2 + \tfrac{\ell\alpha_k^2(1-\alpha_k)}{2}\|y_{k-1} - x_{k-1}\|^2,
\end{aligned} \tag{9}$$

where in the last inequality we used the fact that $z_k - y_{k-1} = \alpha_k(x_{k-1} - y_{k-1})$. From Lemma 1, if e_k be the error in the proximal map of update x_k in Algorithm 1 there exists v_k such that $\|v_k\| \leq \sqrt{2\gamma_k e_k}$ and $\frac{1}{\gamma_k}\left(x_{k-1} - x_k - \gamma_k(\nabla f(z_k) + \bar{\xi}_k) - v_k\right) \in \partial_{e_k} h(x_k)$. Therefore, from Definition 1, the following holds:

$$\begin{aligned}
&h(x) \geq h(x_k) + \langle \tfrac{1}{\gamma_k}(x_{k-1} - x_k) - \nabla f(z_k) - \bar{\xi}_k - \tfrac{1}{\gamma_k}v_k, x - x_k\rangle - e_k \\
&\quad \Longrightarrow \langle \nabla f(z_k) + \bar{\xi}_k, x_k - x\rangle + h(x_k) \leq h(x) - \tfrac{1}{\gamma_k}\langle v_k, x_k - x\rangle + e_k + \tfrac{1}{\gamma_k}\langle x_{k-1} - x_k, x_k - x\rangle.
\end{aligned}$$

From Lemma 2, we have that $\frac{1}{\gamma_k}\langle x_{k-1} - x_k, x_k - x\rangle = \frac{1}{2\gamma_k}[\|x_{k-1} - x\|^2 - \|x_k - x_{k-1}\|^2 - \|x_k - x\|^2]$, therefore,

$$\begin{aligned}
&\langle \nabla f(z_k) + \bar{\xi}_k, x_k - x\rangle + h(x_k) \\
&\quad \leq h(x) - \tfrac{1}{\gamma_k}\langle v_k, x_k - x\rangle + e_k + \tfrac{1}{2\gamma_k}[\|x_{k-1} - x\|^2 - \|x_k - x_{k-1}\|^2 - \|x_k - x\|^2].
\end{aligned} \tag{10}$$

Similarly if ρ_k be the error of computing the proximal map of update y_k in Algorithm 1, then there exists w_k such that $\|w_k\| \leq \sqrt{2\lambda_k\rho_k}$ and one can obtain the following:

$$\begin{aligned}
&\langle \nabla f(z_k) + \bar{\xi}_k, y_k - x\rangle + h(y_k) \\
&\quad \leq h(x) - \tfrac{1}{\lambda_k}\langle w_k, y_k - x\rangle + \rho_k + \tfrac{1}{2\lambda_k}[\|z_k - x\|^2 - \|y_k - z_k\|^2 - \|y_k - x\|^2].
\end{aligned} \tag{11}$$

Letting $x = \alpha_k x_k + (1-\alpha_k)y_{k-1}$ in (11) for any $\alpha_k \geq 0$, the following holds:

$$\begin{aligned}
&\langle \nabla f(z_k) + \bar{\xi}_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + h(y_k) \\
&\quad \leq h(\alpha_k x_k + (1-\alpha_k)y_{k-1}) - \tfrac{1}{\lambda_k}\langle w_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + \rho_k \\
&\quad\quad + \tfrac{1}{2\lambda_k}[\|z_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\|^2 - \|y_k - z_k\|^2].
\end{aligned}$$

From convexity of h and step (1) of algorithm 1 we obtain:

$$\begin{aligned}
&\langle \nabla f(z_k) + \bar{\xi}_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + h(y_k) \\
&\quad \leq \alpha_k h(x_k) + (1-\alpha_k)h(y_{k-1}) - \tfrac{1}{\lambda_k}\langle w_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + \rho_k \\
&\quad\quad + \tfrac{1}{2\lambda_k}[\alpha_k^2\|x_k - x_{k-1}\|^2 - \|y_k - z_k\|^2.
\end{aligned} \tag{12}$$

Multiplying (10) by α_k and then sum it up with (12) gives us the following

$$\langle \nabla f(z_k) + \bar{\xi}_k, y_k - \alpha_k x - (1-\alpha_k)y_{k-1}\rangle + h(y_k)$$

$$
\begin{aligned}
&\leq (1-\alpha_k)h(y_{k-1}) + \alpha_k h(x) - \tfrac{\alpha_k}{2\gamma_k}[\|x_{k-1}-x\|^2 - \|x_k - x\|^2] - \tfrac{1}{\gamma_k}\langle v_k, x_k - x\rangle \\
&\quad + e_k + \underbrace{\tfrac{\alpha_k(\gamma_k\alpha_k-\lambda_k)}{2\gamma_k\lambda_k}}_{\text{term (a)}}\|x_k - x_{k-1}\|^2 - \tfrac{1}{2\lambda_k}\|y_k - z_k\|^2 - \tfrac{1}{\lambda_k}\langle w_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + \rho_k.
\end{aligned}
\tag{13}
$$

By choosing γ_k such that $\alpha_k\gamma_k \leq \lambda_k$, one can easily confirm that term (a)≤ 0. Now combining (8), (9) and (13) and using the facts that $g(x) = f(x) + h(x)$ and $z_k = y_{k-1} + \alpha_k(x_{k-1} - y_{k-1})$, we get the following:

$$
\begin{aligned}
g(y_k) &\leq (1-\alpha_k)g(y_{k-1}) + \alpha_k g(x) - \tfrac{1}{2}(\tfrac{1}{\lambda_k} - L)\|y_k - z_k\|^2 + \overbrace{\langle \bar{\xi}_k, \alpha_k(x - x_{k-1}) + z_k - y_k\rangle}^{\text{term (b)}} \\
&\quad + \tfrac{\alpha_k}{2\gamma_k}[\|x_{k-1}-x\|^2 - \|x_k - x\|^2] + \tfrac{\ell\alpha_k}{2}\|x_{md} - x\|^2 + \tfrac{\ell\alpha_k^2(1-\alpha_k)}{2}\|y_{k-1} - x_{k-1}\|^2 \\
&\quad - \tfrac{1}{\gamma_k}\langle v_k, x_k - x\rangle + e_k - \tfrac{1}{\lambda_k}\langle w_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + \rho_k.
\end{aligned}
\tag{14}
$$

Moreover one can bound term (b) as follows using the Young's inequality.

$$
\begin{aligned}
\langle \bar{\xi}_k, \alpha_k(x - x_{k-1}) + z_k - y_k\rangle &= \langle \bar{\xi}_k, \alpha_k(x - x_{k-1})\rangle + \langle \bar{\xi}_k, z_k - y_k\rangle \\
&\leq \langle \bar{\xi}_k, \alpha_k(x - x_{k-1})\rangle + \tfrac{\lambda_k}{1-L\lambda_k}\|z_k - y_k\|^2 + \tfrac{1-L\lambda_k}{4\lambda_k}\|\bar{\xi}_k\|^2.
\end{aligned}
\tag{15}
$$

Using (15) in (14), we get the following.

$$
\begin{aligned}
g(y_k) &\leq (1-\alpha_k)g(y_{k-1}) + \alpha_k g(x) - \tfrac{1}{4}(\tfrac{1}{\lambda_k} - L)\|y_k - z_k\|^2 + \langle \bar{\xi}_k, \alpha_k(x - x_{k-1})\rangle + \tfrac{\lambda_k}{1-L\lambda_k}\|\bar{\xi}_k\|^2 \\
&\quad + \tfrac{\alpha_k}{2\gamma_k}[\|x_{k-1}-x\|^2 - \|x_k - x\|^2] + \tfrac{\ell\alpha_k}{2}\|x_{md} - x\|^2 + \tfrac{\ell\alpha_k^2(1-\alpha_k)}{2}\|y_{k-1} - x_{k-1}\|^2 \\
&\quad - \tfrac{1}{\gamma_k}\langle v_k, x_k - x\rangle + e_k - \tfrac{1}{\lambda_k}\langle w_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + \rho_k.
\end{aligned}
$$

Subtract $g(x)$ from both sides, using lemma 3, assuming $\frac{\alpha_k}{\lambda_k\Gamma_k}$ is a non-decreasing sequence and summing over k from $k = 1$ to T, the following can be obtained.

$$
\begin{aligned}
&\tfrac{g(x_T)-g(x)}{\Gamma_T} + \sum_{k=1}^{T} \tfrac{1-L\lambda_k}{4\lambda_k\Gamma_k}\|y_k - z_k\|^2 \\
&\leq \tfrac{\alpha_1}{2\gamma_1\Gamma_1}\|x_0 - x\|^2 - \tfrac{\alpha_{T+1}}{2\gamma_{T+1}\Gamma_{T+1}}\|x_T - x\|^2 + \tfrac{\ell}{2}\sum_{k=1}^{T}\tfrac{\alpha_k}{\Gamma_k}\left[\|z_k - x\|^2 + \alpha_k(1-\alpha_k)\|y_{k-1} - x_{k-1}\|^2\right] \\
&\quad + \sum_{k=1}^{T}\tfrac{\alpha_k}{\Gamma_k}\langle \bar{\xi}_k, x - x_{k-1}\rangle + \sum_{k=1}^{T}\tfrac{\lambda_k}{\Gamma_k(1-L\lambda_k)}\|\bar{\xi}_k\|^2 \\
&\quad - \sum_{k=1}^{T}\left[\tfrac{1}{\gamma_k\Gamma_k}\langle v_k, x_k - x\rangle + \tfrac{e_k}{\Gamma_k} - \tfrac{1}{\lambda_k\Gamma_k}\langle w_k, y_k - \alpha_k x_k - (1-\alpha_k)y_{k-1}\rangle + \tfrac{\rho_k}{\Gamma_k}\right].
\end{aligned}
$$

Letting $x = x^*$ and using Assumption 1(iii), inequalities (5) and (6) and the fact that $\|v_k\| \leq \sqrt{2\gamma_k e_k}$ and $\|w_k\| \leq \sqrt{2\lambda_k\rho_k}$, we can simplify the above inequality as follows:

$$
\tfrac{g(x_T)-g(x^*)}{\Gamma_T} + \sum_{k=1}^{T}\tfrac{1-L\lambda_k}{4\lambda_k\Gamma_k}\|y_k - z_k\|^2 \leq \tfrac{\alpha_1}{2\gamma_1\Gamma_1}\|x_0 - x^*\|^2 + \tfrac{\ell}{2}\sum_{k=1}^{T}\tfrac{\alpha_k}{\Gamma_k}\left[2B_3^2 + C^2 + \alpha_k(1-\alpha_k)(2B_2^2 + B_1^2)\right]
$$

$$+\sum_{k=1}^{T}\tfrac{\alpha_k}{\Gamma_k}\langle\bar{\xi}_k, x^*-x_{k-1}\rangle+\sum_{k=1}^{T}\tfrac{\lambda_k}{\Gamma_k(1-L\lambda_k)}\|\bar{\xi}_k\|^2$$
$$+\sum_{k=1}^{T}\left(\tfrac{2e_k}{\Gamma_k}+\tfrac{B_1^2+C^2}{\gamma_k\Gamma_k}+\tfrac{\rho_k(1+k)}{\Gamma_k}+\tfrac{B_1^2+B_2^2}{k\lambda_k\Gamma_k}\right).$$

Using the fact that $g(x_T)-g(x^*)\geq 0$, taking conditional expectation from both sides and applying Assumption 1(iv) on the conditional first and second moments, we get the following.

$$\sum_{k=1}^{T}\tfrac{1-L\lambda_k}{4\lambda_k\Gamma_k}\mathbb{E}[\|y_k-z_k\|^2\mid\mathcal{F}_k]\leq\tfrac{\alpha_1}{2\gamma_1\Gamma_1}\|x_0-x^*\|^2+\tfrac{\ell}{2}\sum_{k=1}^{T}\tfrac{\alpha_k}{\Gamma_k}\left[2B_3^2+C^2+\alpha_k(1-\alpha_k)(2B_2^2+B_1^2)\right]$$
$$+\sum_{k=1}^{T}\tfrac{\lambda_k\tau^2}{\Gamma_k N_k(1-L\lambda_k)}+\sum_{k=1}^{T}\left(\tfrac{2e_k}{\Gamma_k}+\tfrac{B_1^2+C^2}{\gamma_k\Gamma_k}+\tfrac{\rho_k(1+k)}{\Gamma_k}+\tfrac{B_1^2+B_2^2}{k\lambda_k\Gamma_k}\right).\tag{16}$$

To bound the left-hand side we use the following inequality by defining $y_k^r\triangleq\operatorname{prox}_{\lambda_k h}\left(z_k-\lambda_k(\nabla f(z_k)+\bar{\xi}_k)\right)$ and $\hat{y}_k^r\triangleq\operatorname{prox}_{\lambda_k h}\left(z_k-\lambda_k\nabla f(z_k)\right)$.

$$\begin{aligned}\|y_k-z_k\|^2&=\frac{1}{2}\|y_k-z_k\|^2+\frac{1}{2}\|y_k-z_k\|^2\\&\geq\frac{1}{4}\|\hat{y}_k-z_k\|^2-\frac{1}{2}\|\hat{y}_k-y_k\|^2+\frac{1}{4}\|\hat{y}_k^r-z_k\|^2-\frac{1}{2}\|\hat{y}_k^r-y_k\|^2\\&\geq\frac{1}{4}\|\hat{y}_k-z_k\|^2+\frac{1}{4}\|\hat{y}_k^r-z_k\|^2-\frac{3}{2}\|\hat{y}_k-\hat{y}_k^r\|^2-\frac{5}{2}\|\hat{y}_k^r-y_k^r\|^2-\frac{5}{2}\|y_k^r-y_k\|^2,\end{aligned}$$

where we used the fact that for any $a,b\in\mathbb{R}$, we have that $(a-b)^2\geq\frac{1}{2}a^2-b^2$ and for any $a_i\in\mathbb{R}$, $(\sum_{i=1}^m a_i)^2\leq m\sum_{i=1}^m a_i^2$. From Assumption 1(iv), we know that $\|\hat{y}_k^r-y_k^r\|^2\leq\lambda_k^2\tau^2/N_k$, also we know that $\|\hat{y}_k-\hat{y}_k^r\|^2\leq\rho_k/\lambda_k$ and similarly $\|y_k-y_k^r\|^2\leq\rho_k/\lambda_k$. Therefore, one can conclude that $\|y_k-z_k\|^2\geq\frac{1}{4}\|\hat{y}_k-z_k\|^2+\frac{1}{4}\|\hat{y}_k^r-z_k\|^2-\frac{5}{2}\lambda_k^2\tau^2/N_k-4\rho_k/\lambda_k$. Hence, by taking another expectation from (16) and then using this bound, the following can be obtained.

$$\begin{aligned}&\sum_{k=1}^{T}\tfrac{1-L\lambda_k}{16\lambda_k\Gamma_k}\mathbb{E}[\|\hat{y}_k-z_k\|^2+\|\hat{y}_k^r-z_k\|^2]\\&\quad\leq\tfrac{\alpha_1}{2\gamma_1\Gamma_1}\|x_0-x^*\|^2+\tfrac{\ell}{2}\sum_{k=1}^{T}\tfrac{\alpha_k}{\Gamma_k}\left[2B_3^2+C^2+\alpha_k(1-\alpha_k)(2B_2^2+B_1^2)\right]\\&\quad+\sum_{k=1}^{T}\left(\tfrac{\lambda_k\tau^2}{\Gamma_k N_k(1-L\lambda_k)}+\tfrac{2e_k}{\Gamma_k}+\tfrac{B_1^2+C^2}{\gamma_k\Gamma_k}+\tfrac{\rho_k(1+k)}{\Gamma_k}+\tfrac{B_1^2+B_2^2}{k\lambda_k\Gamma_k}+\tfrac{5\lambda_k\tau^2(1-L\lambda_k)}{8\Gamma_k N_k}+\tfrac{\rho_k(1-L\lambda_k)}{\lambda_k^2\Gamma_k}\right).\end{aligned}$$

Using the fact that $\sum_{k=\lfloor T/2\rfloor}^{T}A_t\leq\sum_{k=1}^{T}A_t$ where $A_t=\frac{1-L\lambda_k}{16\lambda_k\Gamma_k}\mathbb{E}[\|\hat{y}_k-z_k\|^2+\|\hat{y}_k^r-z_k\|^2]$, dividing both side by $\sum_{k=\lfloor T/2\rfloor}^{T}\frac{1-L\lambda_k}{16\lambda_k\Gamma_k}$ and using definition of N in Algorithm 1, the desired result can be obtained. □

We are now ready to prove our main rate results.

Theorem 1. *Let $\{y_k, x_k, z_k\}$ generated by Algorithm 1 such that at each iteration $k \geq 1$, e_k-approximate solution of step (2) and ρ_k-approximate solution of step (3) are available through an inner algorithm $\mathcal{M}$. Suppose Assumption 1 and 2 hold and we select the parameters in Algorithm 1 as $\alpha_k = \frac{2}{k+1}$, $\gamma_k = \frac{k}{4L}$, $\lambda_k = \frac{1}{2L}$, $\Gamma_k = \frac{2}{k(k+1)}$ and $N_k = k+1$. Then for $B = B_1^2 + B_2^2 + B_3^2 + C^2$, the following holds for all $T > 0$.*

$$\mathbb{E}[\|\hat{y}_N - z_N\|^2 + \|\hat{y}_N^r - z_N\|^2] \leq \frac{128}{LT^3}\left[2BT(T+1)\left(\frac{\ell}{4} + \frac{13\tau^2}{64LB} + 4L\right) + \sum_{k=1}^{T}\left(\frac{2e_k}{\Gamma_k} + \frac{\rho_k(1+k)}{\Gamma_k} + \frac{4L^2\rho_k}{\Gamma_k}\right)\right], \tag{17}$$

where $\hat{y}_k \approx \text{prox}_{\lambda_k h}(z_k - \lambda_k \nabla f(z_k))$ in the sense of (4), and $\hat{y}_k^r = \text{prox}_{\lambda_k h}(z_k - \lambda_k \nabla f(z_k))$ for any $k \geq 1$.

Proof. Using the definition of λ_k and Γ_k, we get the following.

$$\sum_{k=\lfloor T/2\rfloor}^{T} \frac{1 - L\lambda_k}{16\lambda_k\Gamma_k} = \sum_{k=\lfloor T/2\rfloor}^{T} \frac{Lk(k+1)}{32} = \frac{L}{32}\left[\frac{7T^3}{24} + T^2 + \frac{5T}{6}\right] \geq \frac{LT^3}{128}. \tag{18}$$

Next, using the definition of parameters specified in the statement of the theorem we have that

$$\sum_{k=1}^{T}\frac{\alpha_k}{\Gamma_k} = \sum_{k=1}^{T} k = \frac{T(T+1)}{2}, \qquad \sum_{k=1}^{T}\frac{\tau^2}{\Gamma_k N_k} = \sum_{k=1}^{T}\frac{\tau^2 k}{2} = \frac{\tau^2 T(1+T)}{4},$$
$$\sum_{k=1}^{T}\frac{1}{\gamma_k\Gamma_k} = \sum_{k=1}^{T} 2L(k+1) = 2LT(T+3), \qquad \sum_{k=1}^{T}\frac{1}{k\lambda_k\Gamma_k} = \sum_{k=1}^{T} L(k+1) = LT(T+3). \tag{19}$$

Using (18) and (19) in (7) and the fact that $\alpha_k(1-\alpha_k) \leq 1$, $\frac{T+3}{T+1} \leq 2$ and defining $B = B_1^2 + B_2^2 + B_3^2 + C^2$ we get the desired result. □

Corollary 1. *Let $\{y_k, x_k, z_k\}$ be generated by Algorithm 1 such that at each iteration $k \geq 1$, e_k-approximate solution of step (2) and ρ_k-approximate solution of step (3) are calculated by an inner algorithm $\mathcal{M}$ where $e_k = \gamma_k(c_1\|x_{k-1} - \tilde{x}_k\|^2 + c_2)/q_k^2$ and $\rho_k = \lambda_k(b_1\|y_{k-1} - \tilde{y}_k\|^2 + b_2)/p_k^2$. Suppose Assumptions 1 and 2 hold and $p_k = k+1$ and $q_k = k$. If we choose the stepsize parameters as in Theorem 1, then the following holds for all $T \geq 1$.*

$$\begin{aligned}
&\mathbb{E}[\|\hat{y}_N - z_N\|^2 + \|\hat{y}_N^r - z_N\|^2] \leq \frac{D_1}{T} + \frac{D_2}{T^2}, \\
&D_1 \triangleq \frac{128}{L}\left[4B\left(\frac{\ell}{4} + \frac{13\tau^2}{64LB} + 4L\right) + \left(\frac{c_1(2B_1^2 + C^2) + c_2}{L}\right) + \left(\frac{b_1(2B_2^2 + C^2) + b_2}{4L}\right)\right], \\
&D_2 \triangleq 128\left(b_1(2B_2^2 + C^2) + b_2\right),
\end{aligned} \tag{20}$$

where $\hat{y}_k \approx \text{prox}_{\lambda_k h}(z_k - \lambda_k \nabla f(z_k))$ in the sense of (4) and $\hat{y}_k^r = \text{prox}_{\lambda_k h}(z_k - \lambda_k \nabla f(z_k))$ for any $k \geq 1$. The oracle complexity (number of gradient samples) to achieve $\mathbb{E}[\|\hat{y}_N - z_N\|^2 + \|\hat{y}_N^r - z_N\|^2] \leq \epsilon$ is $\mathcal{O}(1/\epsilon^2)$.

Proof. Using the definition of the stepsizes, p_k, e_k, and ρ_k one can obtain the following:

$$\sum_{k=1}^{T}\frac{2e_k}{\Gamma_k} \leq \frac{c_1(2B_1^2 + C^2) + c_2}{4L}\sum_{k=1}^{T}(k+1) = \left(\frac{c_1(2B_1^2 + C^2) + c_2}{4L}\right)T(T+3).$$

$$\sum_{k=1}^{T} \frac{\rho_k(1+k)}{\Gamma_k} \leq \frac{b_1(2B_2^2+C^2)+b_2}{4L} \sum_{k=1}^{T} k = \left(\frac{b_1(2B_2^2+C^2)+b_2}{8L}\right) T(T+1).$$

$$\sum_{k=1}^{T} \frac{\rho_k}{\Gamma_k} = \left(\frac{b_1(2B_2^2+C^2)+b_2}{4L}\right) \sum_{k=1}^{T} \frac{k(k+1)}{(k+1)^2} \leq \left(\frac{b_1(2B_2^2+C^2)+b_2}{4L}\right) \sum_{k=1}^{T} 1 = \left(\frac{b_1(2B_2^2+C^2)+b_2}{4L}\right) T.$$

Using the above inequalities in (17), we get the desired convergence result. Additionally, the total number of sample gradients of the objective is $\sum_{k=1}^{T} N_k = \sum_{k=1}^{T}(k+1) = T(T+3)$ and the total number of gradients of the constraint is $\sum_{k=1}^{T} p_k + q_k = \sum_{k=1}^{T} 2k+1 = T(T+2)$. From (20), we have that $\mathbb{E}[\|\tilde{y}_N - z_N\|^2] \leq \mathcal{O}(1/T) = \epsilon$, hence, $\sum_{k=1}^{T} N_k = \mathcal{O}(1/\epsilon^2)$ and similarly $\sum_{k=1}^{T} p_k + q_k = \mathcal{O}(1/\epsilon^2)$. □

In the next corollary, we justify our choice of measure. We show that if $\mathbb{E}[\|\hat{y}_N^r - z_N\|^2] \leq \epsilon$, then the first order optimality condition for problem (2) holds within a ball with radius $\sqrt{\epsilon}$.

Corollary 2. *Under the premises of Corollary 1, after running Algorithm 1 for $T \geq D/\epsilon$ iterations, where $D \triangleq D_1 + D_2$, the following holds.*

$$0 \in \mathbb{E}[\nabla f(\hat{y}_N^r)] + \mathbb{E}[\partial h(\hat{y}_N^r)] + \mathcal{B}\left(3L\sqrt{\epsilon}\right).$$

Proof. Suppose $\hat{y}_N^r$ is a solution of $\text{prox}_{\lambda_N h}(z_N - \lambda_N \nabla f(z_N))$. Then $0 \in \partial h(\hat{y}_N^r) + \nabla f(z_N) + (\hat{y}_N^r - z_N)/\lambda$. Adding and subtracting $\nabla f(\hat{y}_N^r)$ form the right-hand side of the above inequality, gives the following:

$$0 \in \partial h(\hat{y}_N^r) + \nabla f(z_N) + 1/\lambda(\hat{y}_N^r - z_N) \pm \nabla f(\hat{y}_N^r). \tag{21}$$

Moreover, using the fact that $T \geq D/\epsilon$ and $\mathbb{E}[\|\hat{y}_N^r - z_N\|^2] \leq \frac{D}{T} = \epsilon$ one can show the following result.

$$\mathbb{E}\left[\|\nabla f(z_N) - \nabla f(\hat{y}_N^r) + 1/\lambda(\hat{y}_N^r - z_N)\|\right] \leq \mathbb{E}\left[L\|\hat{y}_N^r - z_N\| + 1/\lambda\|\hat{y}_N^r - z_N\|\right] \leq 3L\sqrt{\epsilon},$$

where we use the fact that $\lambda = 1/(2L)$. Using the above inequality and taking expectation from (21) the desired result can be obtained. □

In the next section, we show how Algorithm 1 can be customized to solve problem (1).

2.1 Constrained Optimization

Recall that problem (1) can be written in a composite form using an indicator function, i.e. problem (1) is equivalent to $\min_x g(x) = f(x) + h(x)$, where $h(x) = \mathbb{I}_\Theta(x)$ and $\Theta = \{x \mid x \in X,\ \phi_i(x) \leq 0,\ \forall i = 1, \ldots, m\}$. In step (2) and (3) of Algorithm 1, one needs to compute the proximal operators inexactly which are of the following form:

$$\min_{u \in X} \quad \frac{1}{2\gamma}\|u - y\|^2 \quad \text{s.t.} \quad \phi_i(u) \leq 0, \quad i = 1, \ldots, m, \tag{22}$$

for some $y \in \mathbb{R}^n$. Problem (22) has a strongly convex objective function with convex constraints, and there has been variety of methods developed to solve such problems. One of the efficient methods for solving large-scale convex constrained optimization problem with strongly convex objective that satisfies Assumption 2 is first-order primal-dual scheme that guarantees a convergence rate of $\mathcal{O}(1/\sqrt{\epsilon})$ in terms of suboptimality and infeasibility, e.g., [7, 6]. Next, we discuss some details of implementing such schemes as an inner algorithm for solving the subproblems in step (2) and (3) of Algorithm 1.

Based on Corollary 1, to obtain a convergence rate of $\mathcal{O}(1/T)$, one needs to find an e_k- and ϵ_k-approximated solution in the sense of (4). Note that since the nonsmooth part of the objective function, $h(x)$, in the proximal subproblem is an indicator function, (4) implies that the approximate solution of the subproblem has to be feasible, otherwise the indicator function on the left-hand side of (4) goes to infinity. However, the first-order primal-dual methods mentioned above find an approximate solution which might be infeasible. To remedy this issue, let x° be a slater feasible point of (22) (i.e., $\phi_i(x^\circ) < 0$ for all $i = 1, \dots, m$) and let $\hat{x}$ be the output of the inner algorithm $\mathcal{M}$ such that it is ϵ-suboptimal and ϵ-infeasible, then $\tilde{x} = \kappa x^\circ + (1-\kappa)\hat{x}$ is a feasible point of (22) for $\kappa \triangleq \max_i \frac{[\phi_i(\hat{x})]_+}{[\phi_i(\hat{x})]_+ - \phi_i(x^\circ)}$ which is $\mathcal{O}(\epsilon)$-suboptimal, see the next lemma for the proof.

Algorithm 2 IPAG for constrained optimization

input: $x^\circ, x_0, y_0 \in \mathbb{R}^n$ and positive sequences $\{\alpha_k, \gamma_k, \lambda_k\}_k$, and Algorithm $\mathcal{M}$ satisfying Assumption 2;

for $k = 1 \dots T$ **do**

(1) $z_k = (1-\alpha_k)y_{k-1} + \alpha_k x_{k-1}$;

(2) $x \approx \Pi_\Theta\left(x_{k-1} - \gamma_k(\nabla f(z_k) + \bar{\xi}_k)\right)$ (solved inexactly by algorithm $\mathcal{M}$ with q_k iterations);

(3) $y \approx \Pi_\Theta\left(z_k - \lambda_k(\nabla f(z_k) + \bar{\xi}_k)\right)$ (solved inexactly by algorithm $\mathcal{M}$ with p_k iterations);

(4) $\kappa = \max_i \frac{[\phi_i(x)]_+}{[\phi_i(x)]_+ - \phi_i(x^\circ)}$ and $\tilde{\kappa} = \max_i \frac{[\phi_i(y)]_+}{[\phi_i(y)]_+ - \phi_i(x^\circ)}$;

(5) $x_k = \kappa x^\circ + (1-\kappa)x$;

(6) $y_k = \tilde{\kappa} x^\circ + (1-\tilde{\kappa})y$;

end for

Output: z_N where N is randomly selected from $\{T/2, \dots, T\}$ with $\mathrm{Prob}\{N = k\} = \frac{1}{\sum_{k=\lfloor T/2 \rfloor}^{T} \frac{1-L\lambda_k}{16\lambda_k\Gamma_k}}\left(\frac{1-L\lambda_k}{16\lambda_k\Gamma_k}\right)$.

Lemma 5. *Let x° be a strictly feasible point of (22) and $\hat{x}$ be the output of an inner algorithm $\mathcal{M}$ such that it is ϵ-suboptimal and ϵ-infeasible solution of (22). Then $\tilde{x} = \kappa x^\circ + (1-\kappa)\hat{x}$ is a feasible point of (22) and an $\mathcal{O}(\epsilon)$-approximate solution in the sense of (4) where $\kappa = \max_i \frac{[\phi_i(\hat{x})]_+}{[\phi_i(\hat{x})]_+ - \phi_i(x^\circ)}$.*

Proof. Let x^* be the optimal solution of (22). Since $\hat{x}$ is ϵ-suboptimal and ϵ-infeasible solution, $\hat{x} \in X$ and the following holds:

$$\left|\tfrac{1}{2\gamma}\|\hat{x} - y\|^2 - \tfrac{1}{2\gamma}\|x^* - y\|\right| \leq \epsilon, \quad \text{and} \quad [\phi_i(\hat{x})]_+ \leq \epsilon, \ \forall i \in \{1, \dots, m\}.$$

Since X is a convex set and $x^\circ, \hat{x} \in X$, then clearly $\kappa x^\circ + (1-\kappa)\hat{x} \in X$ for any $\kappa \in [0,1]$. Moreover, $\phi_i(x^\circ) < 0$ for all i, hence $\kappa = \max_i \frac{[\phi_i(\hat{x})]_+}{[\phi_i(\hat{x})]_+ - \phi_i(x^\circ)} \in [0,1]$ and $\kappa \leq \frac{\epsilon}{\min_i\{-\phi_i(x^\circ)\}}$. From convexity of $\phi_i(\cdot)$, one can show the following for all $i = 1, \dots, m$.

$$\phi_i(\tilde{x}) \leq \kappa\phi_i(x^\circ) + (1-\kappa)\phi_i(\hat{x}) \leq 0,$$

where we used the definition of κ. Hence, $\tilde{x}$ is a feasible point of (22). Next, we verify $\tilde{x}$ satisfies (4).

$$\tfrac{1}{2\gamma}\|\tilde{x} - y\|^2 + \mathbb{I}_\Theta(\tilde{x}) - \tfrac{1}{2\gamma}\|x^* - y\|^2 - \mathbb{I}_\Theta(x^*)$$

$$
\begin{aligned}
&= \tfrac{1}{2\gamma}\|\tilde{x} - y \pm x^\circ\|^2 - \tfrac{1}{2\gamma}\|x^* - y\|^2 \\
&\le \tfrac{\kappa^2}{2\gamma}\|x^\circ - y\|^2 + \tfrac{(1-\kappa)^2}{2\gamma}\|\hat{x} - y\|^2 + \tfrac{\kappa(1-\kappa)}{\gamma}\|x^\circ - y\|^2\|\hat{x} - y\|^2 - \tfrac{1}{2\gamma}\|x^* - y\|^2 \\
&= \tfrac{\kappa^2}{2\gamma}\|x^\circ - y\|^2 + \tfrac{\kappa(1-\kappa)}{\gamma}\|x^\circ - y\|^2\|\hat{x} - y\|^2 + (1-\kappa^2)\left[\tfrac{1}{2\gamma}\|\hat{x} - y\| - \tfrac{1}{2\gamma}\|x^* - y\|\right] \\
&\quad - \tfrac{1-(1-\kappa^2)}{2\gamma}\|x^* - y\|^2 \\
&\le \tfrac{\kappa^2}{2\gamma}\|x^\circ - y\|^2 + \tfrac{\kappa(1-\kappa)}{\gamma}\|x^\circ - y\|^2\|\hat{x} - y\|^2 + \epsilon \le \mathcal{O}(\epsilon),
\end{aligned}
$$

where we used the fact that $\hat{x}, x^*$ are feasible, $\hat{x}$ is ϵ-suboptimal and $\kappa \le \frac{\epsilon}{\min_i\{-\phi_i(x^\circ)\}}$. □

In the following corollary, we show that the output of Algorithm 2 is feasible to problem (1) and satisfies ϵ-first-order optimality condition.

Corollary 3. *Consider problem* (1). *Suppose Assumption 1 and 2 hold and let* $\{y_k, x_k, z_k\}$ *be generated by Algorithm 2 such that the stepsizes and parameters are chosen as in Corollary 1. Then the iterates are feasible and* $\mathbb{E}\left[\|z_N - \Pi_\Theta\left(z_N - \lambda_N \nabla f(z_N)\right)\|^2\right] \le \mathcal{O}(\epsilon)$ *holds with an oracle complexity* $\mathcal{O}(1/\epsilon^2)$.

Proof. From Lemma 5 we know that the iterates are feasible and from Corollary 1, we conclude that $\mathbb{E}[\|\hat{y}_N^r - z_N\|^2] \le \epsilon$ with an oracle complexity $\mathcal{O}(1/\epsilon^2)$. Considering problem (1), definition of $\hat{y}_N^r$ is equivalent to $\hat{y}_N^r = \Pi_\Theta\left(z_N - \lambda_N \nabla f(z_N)\right)$ which implies the desired result. □

3 NUMERICAL EXPERIMENTS

The goal of this section is to present some computational results to compare the performance of the IPAG method with another competitive scheme. For Algorithm 2, we consider accelerated primal-dual algorithm with backtracking (APDB) method introduced by [6] as the inner algorithm $\mathcal{M}$. In particular, APDB is a primal-dual scheme with a convergence guarantee of $\mathcal{O}(1/T^2)$ in terms of suboptimality and infeasibility when implemented for solving (22) which satisfies the requirements of Corollary 3, i.e., produces approximate solutions for the proximal subproblems.

Example. The IPAG method is benchmarked against the inexact constrained proximal point algorithm (ICPP) introduced by [2]. Consider the following stochastic quadratic programming problem:

$$
\begin{aligned}
\min_{-10 \le x \le 10} \quad & f(x) \triangleq -\tfrac{\epsilon}{2}\|DBx\|^2 + \tfrac{\tau}{2}\mathbb{E}[\|Ax - b(\xi)\|^2] \\
\text{s.t.} \quad & \tfrac{1}{2}x^T Q_i x + d_i^T x - c_i \le 0, \quad \forall i = 1 \ldots m,
\end{aligned}
$$

where $A \in \mathbb{R}^{p\times n}$, $p = n/2$, $B \in \mathbb{R}^{n\times n}, D \in \mathbb{R}^{n\times n}$ is a diagonal matrix, $b(\xi) = b + \omega \in \mathbb{R}^{p\times 1}$, where the elements of ω have an i.i.d. standard normal distribution. The entries of matrices A, B, and vector b are generated by sampling from the uniform distribution $U[0,1]$ and the diagonal entries of matrix D are generated by sampling from the discrete uniform distribution $U\{1{,}1000\}$. Moreover, $(\delta, \tau) \in \mathbb{R}^2_{++}$, $Q_i \in \mathbb{R}^{n\times n}$, $d_i \in \mathbb{R}^{n\times 1}$ and $c_i \in \mathbb{R}$ for all $i \in \{1, \ldots, m\}$. We chose scalers δ and τ such that $\lambda_{min}(\nabla^2 f) < 0$, i.e., minimum eigenvalue of the Hessian is negative. Note that Assumption 1(i) holds for $x^\circ = \mathbf{0}$, where $\mathbf{0}$ is the vector of zeros.

In Table 1 (left), we compared the objective value, CPU time, and infeasibility (Infeas.) of our proposed method with ICPP [2] and in Table 1 (right) we compared the methods for different

		IPAG			ICPP		
n	m	$f(x_T)$	Infeas.	CPU(s)	$f(x_T)$	Infeas.	CPU(s)
100	25	-6.78e+5	0	12.10	-4.85e+4	3.56e-1	32.99
100	50	-8.53e+5	0	31.76	-2.42e+4	3.23e-1	65.79
100	75	-4.18e+5	0	52.43	-2.16e+4	3.75e-1	110.53
200	25	-3.22e+6	0	65.56	-1.81e+5	2.56e-1	132.18
200	50	-1.85e+6	0	90.49	-8.45e+4	4.54e-1	208.84
200	75	-1.33e+6	0	138.75	-7.78e+4	3.93e-1	287.20

			IPAG	ICPP
n	m	std.	$f(x_T)$	$f(x_T)$
100	25	1	-6.7866e+5	-4.8563e+4
100	25	5	-6.5288e+5	-4.8596e+4
100	25	10	-6.2336e+5	-4.8528e+4
200	50	1	-1.8552e+6	-8.4550e+4
200	50	5	-1.8452e+6	-8.5264e+4
200	50	10	-1.8383e+6	-8.6096e+4

Table 1: Comparing IPAG and ICPP.

choices of standard deviation (std.) of ω. To have a fair comparison, we fixed the oracle complexity (i.e. the number of computed gradients is equal for both methods). As it can be seen in the table, for different choices of m and n, IPAG scheme outperforms ICPP. For instance, when we have 25 constraints and $n = 100$, the objective value for our scheme reaches $f(x_T) = -6.78e + 5$ which is significantly smaller than $-4.85e+4$ for ICPP method. Note that our scheme, in contrast to ICPP, obtains a feasible solution at each iteration. Similar behavior can be observed for different choices of the standard deviation in Table 1 (right). standard deviation.

References

[1] Basu, K., Nandy, P.: Optimal convergence for stochastic optimization with multiple expectation constraints. arXiv preprint arXiv:1906.03401 (2019)

[2] Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. arXiv preprint arXiv:1908.02734 (2019)

[3] Ghadimi, S., Lan, G.: Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization **23**(4), 2341–2368 (2013)

[4] Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming **156**(1-2), 59–99 (2016)

[5] Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL **32611** (2014)

[6] Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm with line search for general convex-concave saddle point problems. SIAM Journal on Optimization **31**(2), 1299–1329 (2021)

[7] He, N., Juditsky, A., Nemirovski, A.: Mirror prox algorithm for multi-term composite minimization and semi-separable problems. Computational Optimization and Applications **61**(2), 275–319 (2015)

[8] Kong, W., Melo, J.G., Monteiro, R.D.: Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. SIAM Journal on Optimization **29**(4), 2566–2593 (2019)

[9] Lan, G., Yang, Y.: Accelerated stochastic algorithms for nonconvex finite-sum and multiblock optimization. SIAM Journal on Optimization **29**(4), 2753–2784 (2019)

[10] Lan, G., Zhou, Z.: Algorithms for stochastic optimization with expectation constraints. arXiv preprint arXiv:1604.03887 (2016)

[11] Li, Z., Chen, P.Y., Liu, S., Lu, S., Xu, Y.: Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In: International Conference on Artificial Intelligence and Statistics, pp. 2170–2178. PMLR (2021)

[12] Li, Z., Xu, Y.: Augmented lagrangian based first-order methods for convex and nonconvex programs: nonergodic convergence and iteration complexity. arXiv preprint arXiv:2003.08880 (2020)

[13] Lin, Q., Ma, R., Xu, Y.: Inexact proximal-point penalty methods for constrained non-convex optimization. arXiv preprint arXiv:1908.11518 (2019)

[14] Ma, K., Bai, Y., Yang, J., Yu, Y., Yang, Q.: Demand-side energy management based on nonconvex optimization in smart grid. Energies **10**(10), 1538 (2017)

[15] Nemirovski, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)

[16] Pham, N.H., Nguyen, L.M., Phan, D.T., Tran-Dinh, Q.: Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. Journal of Machine Learning Research **21**(110), 1–48 (2020)

[17] Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 1458–1466 (2011)

[18] Tran-Dinh, Q., Cevher, V.: A primal-dual algorithmic framework for constrained convex minimization. arXiv preprint arXiv:1406.5403 (2014)

[19] Ullah, Z., Al-Turjman, F., Mostarda, L., Gagliardi, R.: Applications of artificial intelligence and machine learning in smart cities. Computer Communications **154**, 313–323 (2020)

[20] Xu, Y.: Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. Mathematical Programming pp. 1–46 (2019)

[21] Zhang, S., He, N.: On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. arXiv preprint arXiv:1806.04781 (2018)