FULL LENGTH PAPER

Series A

Level constrained first order methods for function constrained optimization

Digvijay Boob[1] (iD) **· Qi Deng[2] · Guanghui Lan[3]**

Abstract

We present a new feasible proximal gradient method for constrained optimization where both the objective and constraint functions are given by summation a smooth, possibly nonconvex function and a convex simple function. The algorithm converts the original problem into a sequence of convex subproblems. Formulating those subproblems requires the evaluation of at most one gradient-value of the original objective and constraint functions. Either exact or approximate subproblems solutions can be computed efficiently in many cases. An important feature of the algorithm is the constraint level parameter. By carefully increasing this level for each subproblem, we provide a simple solution to overcome the challenge of bounding the Lagrangian multipliers and show that the algorithm follows a strictly feasible solution path till convergence to the stationary point. We develop a simple, proximal gradient descent type analysis, showing that the complexity bound of this new algorithm is comparable to gradient descent for the unconstrained setting which is new in the literature. Exploiting this new design and analysis technique, we extend our algorithms to some more challenging constrained optimization problems where (1) the objective is a stochastic or finite-sum function, and (2) structured nonsmooth functions replace smooth components of both

---

✉ Digvijay Boob
  dboob@smu.edu

✉ Qi Deng
  contact.qdeng@gmail.com

  Guanghui Lan
  george.lan@isye.gatech.edu

[1] Operations Research and Engineering Management, Southern Methodist University, Dallas, TX, USA

[2] Antai College of Economics & Management, Shanghai Jiao Tong University, Shanghai, China

[3] Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

🖄 Springer

objective and constraint functions. Complexity results for these problems also seem to be new in the literature. Finally, our method can also be applied to convex function constrained problems where we show complexities similar to the proximal gradient method.

Keywords Function constrained optimization · Proximal gradient descent · Moving balls method · Stochastic optimization

Mathematics Subject Classification 90C26 · 90C30 · 90C06 · 90C51 · 49M37

1 Introduction

In this paper, we study the following constrained optimization problem

$$
\begin{aligned}
\min_{x \in \mathbb{R}^d} \quad & \psi_0(x) := f_0(x) + \chi_0(x) \\
\text{s.t.} \quad & \psi_i(x) := f_i(x) + \chi_i(x) \le \eta_i, \quad i = 1, \dots, m.
\end{aligned}
\tag{1.1}
$$

where $\psi_i(x)$ is a composite function that sums up functions $f_i(x)$ and $\chi_i(x)$. Here, $f_i, i = 0, 1, \dots, m$, are smooth functions, $\chi_0(x)$ is a proper, convex, lower-semicontinuous (lsc) function and $\chi_i(x), i = 1, \dots, m$, are convex continuous functions over the domain of χ_0 (i.e. dom_{χ_0}). We consider that $\chi_i, i = 0, \dots, m$ are 'simple' functions, namely, a feasible optimization problem of the form below

$$
\min_{x \in \mathbb{R}^d} \left\{ \|x - a^0\|^2 + \chi_0(x) : \|x - a^i\|^2 + \chi_i(x) \le b^i, i = 1, \dots, m. \right\}
\tag{1.2}
$$

can be solved to efficiently obtain either an exact solution or an inexact solution of desired accuracy. Note that if $\chi_i = 0, i = 1, \dots, m$, then (1.2) becomes a proximal operator for function χ_0 on the intersection of balls. If we further assume $\chi_0 = 0$, then (1.2) is a special type of quadratically constrained quadratic programming (QCQP) that can be solved efficiently because all the Hessians are identity matrices. In addition, we consider the case where constraints $f_i, i = 1, \dots, m$, are structured nonsmooth functions which can be approximated by smooth functions (also called smoothable functions). Note that problem (1.1) covers a variety of convex and nonconvex function constrained optimization depending on the assumptions of $f_i, i = 0, \dots, m$.

Nonlinear optimization with function constraints is a classical topic in continuous optimization. While the earlier study focused on the asymptotic performance, recent work has put more emphasis on the complexity analysis of algorithms, mainly driven by the growing interest in large-scale optimization and machine learning. For most of our discussion on the complexity analysis, we generally require convergence to an ϵ-approximate KKT point (c.f. Definition 3). Penalty methods [9, 25, 33], including augmented Lagrangian methods [22–24, 34], is one popular approach for constrained optimization. In [8], Cartis et al. presented an exact penalty method by minimizing a sequence of convex composition functions. When the penalty weight is bounded, this method solves $\mathcal{O}(1/\epsilon)$ trust region subproblems. If the penalty weight is unbounded,

the complexity is of $\mathcal{O}(1/\epsilon^{2.5})$ to reach an ϵ-KKT point. In a subsequent work [9], the authors provided a target following method that achieves the complexity of $\mathcal{O}(1/\epsilon)$, regardless of the growth of the penalty parameter. In [33], Wang et al. extended the penalty method for solving constrained problems where the objective takes the expectation form. Sequential quadratic programming (SQP) is another important approach for constrained optimization. Typically, SQP involves linearization of the constraints, quadratic approximation of the objective, and possibly some trust region constraint for the convergence guarantee [6, 7]. The recent work [12] established a unified convergence analysis of SQP (GSQP) in more general settings where the feasibility and constraint qualification may or may not hold. Different from the standard SQP, the Moving Balls Approximation (MBA) method [1] follows a feasible solution path and transforms the initial problem into a diagonal QCQP. A subsequent work [3] presented a unified analysis of MBA and other variants of SQP methods. Under the assumption of Kurdyka-Łojasiewicz (KL) property, they establish the global convergence rates which depend on the Łojasiewicz exponent.

Despite much progress in prior works, there are some significant remaining issues. Specifically, most of the analysis is carried out for only smooth optimization and requires that the exact optimal solution of the convex subproblem is readily available. Unfortunately, both assumptions can be unrealistic in many large-scale applications. To overcome these issues, [4, 25, 26] presented some new proximal point algorithms that iteratively solve strongly convex proximal subproblems inexactly using first-order methods. A significant computational advantage is that first-order methods only need to compute a relatively easy proximal gradient mapping in each iteration. In particular, [4] proposed to solve the proximal point subproblem by a new first-order primal-dual method called ConEx. Under some strict feasibility assumption, they derived the total complexities of the overall algorithm for which the objective and constraints can be either stochastic or deterministic, and either nonsmooth or smooth. Notably, for nonconvex and smooth constrained problems, inexact proximal point [4] requires $\mathcal{O}(1/\epsilon^{1.5})$ function/gradient evaluations. A similar complexity bound is obtained by the proximal point penalty method [25] when a feasible point is available. Nevertheless, at this point, it may be difficult to directly compare the efficiency of the proximal point with the earlier approach, given that very different oracles are employed in each method. The inexact proximal point method appears to be less efficient in terms of gradient and function value computations since first-order penalty method [9] and a variant of SQP [12] (where the surrogate is formed by first-order approximation) has an $\mathcal{O}(1/\epsilon)$ complexity bound. Nevertheless, it might be more efficient if the corresponding proximal mapping is much easier to solve than the subproblems in penalty or SQP methods.

In this paper, we attempt to alleviate some of the aforementioned issues in solving nonconvex constrained optimization. Our main contribution is the development of a novel **L**evel **C**onstrained **P**roximal **G**radient (LCPG) method for constrained optimization, based on the following key ideas.

First, we convert the original problem (1.1) into a sequence of simple convex subproblems of the form (1.2) for which an exact or an approximate solution can be computed efficiently. In particular, solving the subproblem requires at most one gradient and function value computation for $f_i, i = 0, \ldots, m$. This phenomenon is quite

similar to simple single-loop methods even though LCPG method can be multi-loop since we allow for an inexact solution of (1.2) using some kind of iterative scheme.

Second, starting from a strictly feasible initial point and carefully controlling the feasibility levels of the subproblem constraints, we ensure that LCPG follows a strictly feasible solution path. This also allows us to deal with nonsmooth constraints where χ_i is not necessarily 0 and further extends LCPG to the inexact case where the subproblem admits an approximate solution. Even though subtle, the level-control design is crucial in bounding the Lagrange multipliers under the well-known Mangasarian–Fromovitz constraint qualification (MFCQ) [4, 27]. Subsequently, we also show asymptotic convergence of the LCPG method.

Third, we offer a new insight into the complexity analysis of LCPG as a gradient descent type method, which could be of independent interest. When the objective and constraints are nonconvex composite, we aim to find a first-order ϵ-KKT point (c.f. Definition 3) under the aforementioned MFCQ assumption. We can show that LCPG method converges in $O(1/\epsilon)$ iterations. Furthermore, each subproblem requires at most one function-value and gradient computation. The net outcome is that *gradient complexity* of our method is of $O(1/\epsilon)$. Notice that the number of iterations required by the proximal point method under MFCQ is also $O(1/\epsilon)$ (see [4, Theorem 5]). However, each iteration of this method requires $O(1/\epsilon^{0.5})$ gradient computation, and hence its total gradient complexity can be bounded by $O(1/\epsilon^{1.5})$. This is much worse than LCPG method. We compare with some significant lines of work in Table 1.

Exploiting the intrinsic connection between LCPG and proximal gradient (without function constraints), we extend LCPG to a variety of cases. (1) We can show a similar $O(1/\epsilon)$ gradient complexity for an inexact LCPG method for which the subproblem is solved to a pre-specified accuracy. If we assume $\chi_i = 0$, then the corresponding subproblem (1.2) (i.e. diagonal QCQP) can be efficiently solved by a customized interior point method in logarithmic time. In the more general setting when $\chi_i \neq 0$, we propose to solve (1.2) by the first-order method ConEx, which has very cheap iterations. (2) We also extend LCPG method to stochastic (LCSPG) and variance-reduced (LCSVRG) variants when f_0 is either stochastic or finite-sum function, respectively. LCSPG and LCSVRG require $O(1/\epsilon^2)$ (similar to SGD [15]) and $O(\sqrt{n}/\epsilon)$ (similar to SVRG [18]) stochastic gradients, respectively, where n is the number of components in the finite-sum objective. The complexity of variants of LCPG method for stochastic cases can also be seen in Table 2. (3) We consider the case when function $f_i, i = 0, 1, \ldots, m$ are nondifferentiable but contain a smooth saddle structure (referred to as *structured nonsmooth*). We extend LCPG method for such nonsmooth nonconvex function constrained problem using Nesterov's smoothing scheme [29]. In this case, LCPG method requires $O(1/\epsilon^2)$ gradients.

We show that the GD-type analysis of the LCPG method can be extended to the convex case. In particular, when the objective and constraint functions are convex, we show that LCPG method requires $O(1/\epsilon)$ gradient computations for smooth and composite constrained problems, and this complexity improves to $O(\log(1/\epsilon))$ when the objective is smooth and strongly-convex. Furthermore, we develop the complexity of inexact variants of LCPG method by leveraging the analysis of gradient descent with inexact projection oracles [31]. Inexact LCPG method maintains the gradient complexity of $O(1/\epsilon)$ and $O(\log(1/\epsilon))$ for convex and strongly convex problems, respectively.

Table 1 Comparison of algorithm function/gradient evaluation complexities

Algorithms	Function type	Composite?	Inexact?	CQ	Criteria	Gradient complexity[†]
MBA[1]	**str cvx**	No	No	Feasibility+MFCQ	Opt. gap	$\mathcal{O}(\log(1/\epsilon))$
GSQP [12]	**noncvx**	No	No	Extended MFCQ	KKT	$\mathcal{O}(1/\epsilon)$
IPP +ConEx [4]	**noncvx**	Yes	Yes	Strong feasibility	KKT	$\mathcal{O}(1/\epsilon^{1.5})$
IPPP [25]	**noncvx**	No	Yes	Feasibility + Nonsingularity	KKT	$\mathcal{O}(1/\epsilon^{1.5})$
				Nonsingularity		$\mathcal{O}(1/\epsilon^2)$
IQRC [26]	**noncvx**	Yes*	Yes	Uniform Slater	KKT	$\mathcal{O}(1/\epsilon^{1.5})$
LCPG (this work)	**cvx**	Yes	Yes	Slater+MFCQ	Opt. gap	$\mathcal{O}(1/\epsilon)$
	str cvx				Opt. gap	$\mathcal{O}(\log(1/\epsilon))$
	noncvx				KKT	$\mathcal{O}(1/\epsilon)$

cvx: convex, **str cvx:** strongly convex, and **noncvx:** nonconvex. For convex problems, we consider the complexity to reach a feasible solution with $O(\epsilon)$-optimality gap. For nonconvex problems, we consider the complexity to reach an approximate KKT solution that satisfies $\|\partial \mathcal{L}\|_-^2 \le \epsilon$. Note that different works have quite different error measurements of the complementary slackness. For example, in our translation, [12, 25] requires an $\mathcal{O}(\sqrt{\epsilon})$ error on the complementary slackness and feasibility. Our measure requires 0-feasibility error and $\mathcal{O}(\epsilon)$-complementary slackness error. * IQRC does not explicitly discuss the composite case. Their subproblem oracle can be upgraded to handle proximal cases relatively easily

† Different methods have different costs for solving the subproblem. Some methods require explicit gradient computations for solving the subproblems and hence, are expected to be quite computationally costly. Some methods (including ours) have simple subproblems. See Remark 11 for a detailed discussion. Hence, comparing total computational complexity is not possible. We instead compare gradient complexities to provide a realistic estimate of the computational effort of each of these methods

Throughout our analysis, we require that the Lagrange multipliers for the convex subproblems of type (1.2) are bounded. This problem is addressed in different ways in arguably all works in the literature. In this paper, we show that under the assumption of MFCQ, Lagrange multipliers associated with the sequence of subproblems remain bounded by a quantity specified as B. Even then, the value of B cannot be estimated a priori. Fortunately, this bound is not needed in the implementation of our methods. However, it plays a role in complexity analysis. Hence, our comparison with the existing complexity literature (e.g., proximal point method of [4]) is valid when bound B on the sequence of Lagrange multipliers largely depends on the problem itself and not on the sequence of subproblems. One can easily see that such uniform bounds on Lagrange multipliers hold under the strong feasibility constraint qualification [4], a similar uniform Slater's condition [26] or for nonsmooth nonconvex relaxation in the application of sparsity constrained optimization [5]. The problem of comparing bounds B on Lagrange multipliers requires getting into specific applications, which is not the purpose of this paper. Hence, throughout our comparison with existing literature, we assume that bound B for different methods is of a similar order.

Comparison with MBA method Auslender et al. [1] provided a Moving Balls Approximation (MBA) method for smooth constrained problems, i.e. $\chi_i(x)$, $i = 0, \ldots, m$, are not present. They use Lipschitz continuity of constraint gradients along with MFCQ to

Table 2 Total number of stochastic gradient evaluations to obtain $\mathcal{O}(\varepsilon, \varepsilon)$ randomized KKT points in the finite sum problem

	LCPG	LCSPG	LCSVRG
Complexity	$\mathcal{O}(n\varepsilon^{-1})$	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(n^{1/2}\varepsilon^{-1})$

ensure that the subproblems satisfy Slater's conditions (see [1, Proposition 2.1(iii)]). A similar result is also used in [35] where they provide a line-search version of MBA for functions satisfying certain KL properties. The MBA method was studied for semialgebraic functions in [3] where they used the KL-property of semi-algebraic functions. The work [1] also provides the complexity guarantee for constrained programs with a smooth and strongly convex objective. Our results differ from the past studies in the following several aspects. First, we do not assume any KL property on the class of functions, hence making the method applicable to a wider class of problems. Second, we show complexity analysis for a variety of cases, e.g., stochastic, finite-sum, or structured nonsmooth cases. Note that complexity results are not known for the MBA type method even for the purely smooth problem. Third, we show complexity results for both convex and strongly convex cases which strictly subsumes the results in [1]. Fourth, it should be noted that [1] also considered the efficiency of solving subproblems. They proposed an accelerated gradient method that obtains $\mathcal{O}(1/\sqrt{\epsilon})$ complexity for solving the dual of the QCQP subproblem. However, it is unclear what accuracy is enough for ensuring asymptotic convergence of the whole algorithm.

Comparison with generalized SQP The work [12] developed the first complexity analysis of the generalized SQP (GSQP) method by using a novel ghost penalty approach. Different from our feasible method, they consider a general setting where the feasibility and constraint qualification may or may not hold. They show that SQP-type methods have an $\mathcal{O}(1/\epsilon^2)$ complexity for reaching some ϵ-approximate generalized stationary point. Under an extended MFCQ condition, they established an improved complexity $\mathcal{O}(1/\epsilon)$ for reaching the scaled-KKT point, which matches our complexity result under a similar MFCQ assumption. Notably, both their analysis and ours rely on MFCQ to show that the global upper bound (constant B) on the multipliers of the subproblems exists. However, to obtain the best $\mathcal{O}(1/\epsilon)$ complexity, GSQP explicitly relies on the value of such unknown upper bound to determine the stepsize, which appears to be rather challenging in practical use. In contrast, our algorithm does not involve the constant B in the algorithm implementation; we only require the Lipschitz constants of the gradients, which is standard for gradient descent methods.

Outline This paper is organized as follows: Sect. 2 describes notations and assumptions. It also provides various definitions used throughout the paper. Section 3 presents the LCPG method which uses exact solutions of subproblems. It also establishes asymptotic convergence and convergence rate results. Section 4.1 and Sect. 4.2 provides the LCSPG and LCSVRG method for stochastic and finite-sum problems, respectively. Section 5 shows the extension of LCPG for nonsmooth nonconvex function constraints. Section 6 introduces the inexact LCPG method and provides its complexity analysis when the subproblems are inexactly solved by an interior point method or first-order method. Finally, Sect. 7 extends LCPG method for convex optimization problems and establishes its complexity for both strongly convex and convex problems.

2 Notations and assumptions

Notations. $\mathbb{R}_+^n$ stands for the non-negative orthant in $\mathbb{R}^n$. We use $\|\cdot\|$ to express the Euclidean norm. For a set $\mathcal{X}$, we define $\|\mathcal{X}\|_- = \text{dist}(0, \mathcal{X}) = \inf \{\|x\|, x \in \mathcal{X}\}$. If $\mathcal{X}$ is a convex set then we denote its normal cone at x as $N_{\mathcal{X}}(x)$. Furthermore, we denote the dual cone of the normal cone at x as $N_{\mathcal{X}}^*(x)$. Let e be a vector full of elements one. For simplicity, we denote $[m] = \{1, 2, \ldots, m\}$, $f(x) = [f_1(x), \ldots, f_m(x)]^T$, $\chi(x) = [\chi_1(x), \ldots, \chi_m(x)]^T$, and $\psi(x) = [\psi_1(x), \psi_2(x), \ldots, \psi_m(x)]^T$. For vectors $x, y \in \mathbb{R}^m$, $x \leq y$ is understood as $x_i \leq y_i$ for $i \in [m]$.

Assumption 1 (General) We assume that the optimal value of problem (1.1) is finite: $\psi_0^* > -\infty$. Furthermore, the objective and constraint functions have the following properties.

1: χ_0 is a proper, convex, and lower semi-continuous (lsc) function. Moreover, we assume that for all $i = 1, \ldots, m$, the function $\chi_i(x)$ is convex continuous over dom_{χ_0}.
2: $f_i(x)$ is L_i-Lipschitz smooth on dom_{χ_0}: $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$ for any $x, y \in \text{dom}_{\chi_0}$. For brevity, we denote $L = [L_1, \ldots, L_m]^T$.
3: The feasible set for (1.1), i.e., $\bigcap_{i \in [m]} \{x : \psi_i(x) \leq \eta_i\} \cap \text{dom}_{\chi_0}$ is nonempty and compact.[1]

The Lagrangian function of problem (1.1) is denoted by

$$\mathcal{L}(x, \lambda) = \psi_0(x) + \sum_{i=1}^m \lambda_i [\psi_i(x) - \eta_i]. \tag{2.1}$$

For functions ψ_i, we denote its subdifferential as

$$\partial \psi_i(x) = \{\nabla f_i(x)\} + \partial \chi_i(x), i = 0, \ldots, m,$$

where the sum is in the Minkowski sense. Note that this definition of the subdifferential for nonconvex functions was first proposed in [4]. Moreover, $\partial \psi_i = \{\nabla f_i\}$ when ψ_i is a "purely" smooth nonconvex function and $\partial \psi_i = \partial \chi_i$ when ψ_i is a nonsmooth convex function. Hence, it is a valid definition for the subdifferential of a nonconvex function. Below, we define the KKT condition using the above subdifferential.

Definition 1 (KKT condition) We say $x \in \text{dom}_{\chi_0}$ is a KKT point of problem (1.1) if x is feasible and there exists a vector $\lambda \in \mathbb{R}_+^m$ such that

$$0 \in \partial_x \mathcal{L}(x, \lambda), \quad 0 = \sum_{i=1}^m \lambda_i [\psi_i(x) - \eta_i]. \tag{2.2}$$

The values $\{\lambda_i\}$ are called Lagrange multipliers.

---

[1] This assumption is used to ensure the limit point of a certain sequence in the analysis. There are other conditions that can ensure the existence of a limit point as we will discuss in Remark 1.

It is known that the KKT condition is necessary for optimality under the assumption of certain constraint qualifications (c.f. [2]). Our result will be based on a variant of the Mangasarian–Fromovitz constraint qualification, which is formally given below.

Definition 2 (MFCQ) We say that a point x satisfies the Mangasarian–Fromovitz constraint qualification for (1.1) if there exists a vector $z \in -N^*_{\text{dom}\chi_0}(x)$ such that

$$\max_{v \in \partial \psi_i(x)} \langle v, z \rangle < 0, \quad i \in \mathcal{A}(x), \tag{2.3}$$

where $\mathcal{A}(x) = \{i : 1 \leq i \leq m, \psi_i(x) = \eta_i\}$.

Proposition 1 (*Necessary condition*) *Let x be a local optimal solution of Problem* (1.1). *If it satisfies MFCQ* (2.3), *then there is a vector $\lambda \in \mathbb{R}^m_+$ such that the KKT condition* (1) *holds.*

Next, we introduce some optimality measures before formally presenting any algorithms. It is natural to characterize algorithm performance by measuring the error of satisfying the KKT condition. Towards this goal, we have the following definition.

Definition 3 We say that x is an ϵ type-I (approximate) KKT point if it is feasible (i.e. $\psi(x) \leq \eta$), and there exists a vector $\lambda \in \mathbb{R}^m_+$ satisfying the following conditions:

$$\|\partial_x \mathcal{L}(x, \lambda)\|^2_- \leq \epsilon$$

$$-\sum_{i=1}^{m} \lambda_i [\psi_i(x) - \eta_i] \leq \epsilon.$$

Moreover, x is a randomized ϵ type-I KKT point if both x and λ are feasible randomized primal-dual solutions that satisfy

$$\mathbb{E}[\|\partial_x \mathcal{L}(x, \lambda)\|^2_-] \leq \epsilon$$

$$\mathbb{E}[-\sum_{i=1}^{m} \lambda_i [\psi_i(x) - \eta_i]] \leq \epsilon,$$

where the expectation is taken over the randomness of x and λ.

Besides the above definition, we invoke a second optimality measure which will help analyze the performance of a proximal algorithm (see, for example, [4]). Therein, it is arguably more convenient to approach the proximity of some nearly stationary points.

Definition 4 We say that x is a (ϵ, ν) type-II KKT point if there exists an ϵ type-I KKT point $\hat{x}$ and $\|x - \hat{x}\|^2 \leq \nu$. Similarly, x is a randomized (ϵ, ν) type-II KKT point if $\hat{x}$ is a random vector and $\mathbb{E}[\|x - \hat{x}\|^2] \leq \nu$.

3 A proximal gradient method

We present the level constrained proximal gradient (LCPG) method in Algorithm 1. The main idea of this algorithm is to turn the original nonconvex problem into a sequence of relatively easier subproblems that involve some convex surrogate functions $\psi_i^k(x)$ $(0 \leq i \leq m)$ and variable constraint levels η^k:

$$
\begin{aligned}
\min_{x \in \mathbb{R}^d} \quad & \psi_0^k(x) \\
\text{s.t.} \quad & \psi_i^k(x) \leq \eta_i^k, \quad i \in [m].
\end{aligned}
\tag{3.1}
$$

Above, we take the surrogate function $\psi_i^k(x)$ $(0 \leq i \leq m)$ by partially linearizing $\psi_i(x)$ at x^k and adding the proximal term $\frac{L_i}{2}\|x - x^k\|^2$:

$$
\psi_i^k(x) := f_i(x^k) + \langle \nabla f_i(x^k), x - x^k \rangle + \frac{L_i}{2}\|x - x^k\|^2 + \chi_i(x). \tag{3.2}
$$

It should be noted that our algorithm may not be well-defined if it were to be initialized by an infeasible solution x^0. Furthermore, we require the initial point to be strictly feasible with respect to the nonlinear constraints $\psi(x) \leq \eta$. Therefore, we explicitly state this assumption below and assume it holds throughout the paper.

Assumption 2 (Strict feasibility) There exist a point $x^0 \in \mathrm{dom}_{\chi_0}$ and a vector $\eta^0 \in \mathbb{R}^m$ satisfying

$$
\psi_i(x^0) < \eta^0 < \eta.
$$

With a strictly feasible solution, we assume that the constraint levels $\{\eta^k\}$ are incrementally updated and converge to the constraint levels for the original problem:

$$
\lim_{k \to \infty} \eta_i^k = \eta_i, \quad i \in [m].
$$

Algorithm 1 Level constrained proximal gradient method (LCPG)

1: **Input:** x^0, η^0;
2: **for** k=0,1,2,...,K **do**
3:     For $i = 0, 1, \ldots, m$, set $\psi_i^k(x)$ by (3.2);
4:     Update x^{k+1} by solving (3.1);
5:     Choose $\delta^k > 0$ and update η^{k+1} by

$$
\eta_i^{k+1} = \eta_i^k + \delta_i^k < \eta_i, \qquad\qquad i = 1, 2, \ldots, m. \tag{3.3}
$$

6: **end for**
7: **Output:** $x^{\hat{k}+1}$ where $\hat{k}$ is a random index sampled from $\{0, 1, 2, ..., K\}$.

The following Lemma will be used many times throughout the rest of the paper.

Lemma 1 *[Three-point inequality] Let* $g : \mathbb{R}^d \rightarrow (-\infty, \infty]$ *be a proper lsc convex function and*

$$x^+ = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ g(z) + \tfrac{\gamma}{2} \|z - y\|^2 \right\},$$

then for any $x \in \mathbb{R}^d$, *we have*

$$g(x^+) - g(x) \leq \tfrac{\gamma}{2} \big(\|x - y\|^2 - \|x^+ - y\|^2 - \|x - x^+\|^2 \big). \tag{3.4}$$

Next, we present some important properties of the generated solutions in the following theorem.

Proposition 2 *Suppose that Assumption 2 holds, then Algorithm 1 has the following properties.*

1. *The sequence* $\{x^k\}$ *is well-defined and is feasible for problem* (1.1). $\{x^k\}$ *satisfies the sufficient descent property:*

$$\tfrac{L_0}{2} \|x^{k+1} - x^k\|^2 \leq \psi_0(x^k) - \psi_0(x^{k+1}). \tag{3.5}$$

   Moreover, the sequence of objective values $\{\psi_0(x^k)\}$ *is monotonically decreasing and* $\lim_{k \to \infty} \psi_0(x^k)$ *exists.*
2. *There exists a vector* $\lambda^{k+1} \in \mathbb{R}_+^m$ *such that the KKT condition holds:*

$$\partial \psi_0^k(x^{k+1}) + \sum_{i=1}^m \lambda_i^{k+1} \partial \psi_i^k(x^{k+1}) \ni 0 \tag{3.6}$$
$$\lambda_i^{k+1} \big[\psi_i^k(x^{k+1}) - \eta_i^k \big] = 0, \quad i \in [m].$$

Proof Part 1). First, we show that $\{x^k\}$ is a well-defined sequence, namely, $\mathcal{X}_k \cap \operatorname{dom}_{\chi_0}$ is a nonempty set where $\mathcal{X}_k = \{x : \psi_i^k(x) \leq \eta_i^k\}$. This result clearly holds for $k = 0$ by Assumption 2. We show the general case ($k > 0$) by induction. Suppose that x^k is well-defined, i.e., $\mathcal{X}_{k-1} \cap \operatorname{dom}_{\chi_0}$ is nonempty. Then, by the definition of ψ_i^k, ψ_i^{k-1} and the definition x^k, we have $x^k \in \operatorname{dom}_{\chi_0}$ and

$$\psi_i^k(x^k) = \psi_i(x^k) \leq \psi_i^{k-1}(x^k) \leq \eta_i^{k-1} < \eta_i^k \quad \text{for all } i \in [m]. \tag{3.7}$$

Here, the first inequality follows due to the smoothness of $f_i(x)$ which ensures for all x,

$$f_i(x) \leq f_i(x^{k-1}) + \langle \nabla f_i(x^{k-1}), x - x^{k-1} \rangle + \tfrac{L_i}{2} \|x - x^{k-1}\|^2, \quad \forall i \in [m].$$

This is equivalent to $x^k \in \mathcal{X}_k \cap \operatorname{dom}_{\chi_0}$, implying that $\mathcal{X}_k \cap \operatorname{dom}_{\chi_0}$ is nonempty. We conclude that x^{k+1} is well-defined. Hence, by induction, we conclude that $\{x^k\}$ is a well-defined sequence. Furthermore, in view of $x^k \in \operatorname{dom}_{\chi_0}$, relation (3.7) and the

fact that $\eta_i^k < \eta_i$, we have $x^k \in \mathrm{dom}_{\chi_0} \cap \{x : \psi_i(x) \leq \eta_i, \ i = 1, \ldots, m\}$. Hence, the whole sequence $\{x^k\}$ remains feasible for the original problem.

Now, let us apply Lemma 1 with $g(x) = \langle \nabla f_0(x^k), x \rangle + \chi_0(x) + \mathbf{1}_{\mathcal{X}_k}(x)$, $y = x^k$ and $\gamma = L_0$. Then, for any $x \in \mathrm{dom}_{\chi_0} \cap \mathcal{X}_k$, we have

$$\langle \nabla f_0(x^k), x^{k+1} - x \rangle + \chi_0(x^{k+1}) - \chi_0(x)$$
$$\leq \tfrac{L_0}{2} \big(\|x - x^k\|^2 - \|x - x^{k+1}\|^2 - \|x^k - x^{k+1}\|^2 \big).$$

Placing $x = x^k$ in the above relation, we have

$$\langle \nabla f_0(x^k), x^{k+1} - x^k \rangle + \chi_0(x^{k+1}) - \chi_0(x^k) \leq -L_0 \|x^{k+1} - x^k\|^2. \qquad (3.8)$$

Moreover, since $f_0(\cdot)$ is Lipschitz smooth, we have that

$$f_0(x^{k+1}) \leq f_0(x^k) + \langle f_0(x^k), x^{k+1} - x^k \rangle + \tfrac{L_0}{2} \|x^{k+1} - x^k\|^2.$$

Summing up the above two inequalities and using the definition $\psi_0 = f_0 + \chi_0$, we conclude (3.5). Hence, the sequence $\{\psi_0(x^k)\}$ is monotonically decreasing. The convergence of $\{\psi_0(x^k)\}$ follows from the lower boundedness assumption.

Part 2). Note that (3.7) ensures the strict feasibility of x^k w.r.t. the constraint set $\mathcal{X}_k$ for the kth subproblem. Therefore, Slater's condition for (3.1) and the optimality of x^{k+1} imply that there must exist a vector $\lambda^{k+1} \in \mathbb{R}_+^m$ satisfying KKT condition (3.6). Hence, we complete the proof. $\qquad \square$

In order to show convergence to the KKT solutions, we need the following constraint qualifications.

Assumption 3 [Uniform MFCQ] All the feasible points of problem (1.1) satisfy MFCQ.

We state the main asymptotic convergence property of LCPG in the following theorem.

Theorem 1 *Suppose that Assumption 3 holds, then we have the following conclusions.*

1. *The dual solutions $\{\lambda^{k+1}\}$ are bounded from above. Namely, there exists a constant $B > 0$ such that*
$$\sup_{0 \leq k \leq \infty} \|\lambda^{k+1}\| < B. \qquad (3.9)$$

2. *Every limit point of Algorithm 1 is a KKT point.*

Proof Part 1). First, we can immediately see that $\{x^k\}$ is a bounded sequence and hence the limit point exists. This result is from Assumption 1.3: and the feasibility of x^k for problem (1.1) (c.f. Proposition 2, Part 1). Without loss of generality, we can assume $\lim_{k \to \infty} x^k = \bar{x}$. For the sake of contradiction, suppose that λ^{k+1} is unbounded, then passing to a subsequence if necessary, we can assume $\|\lambda^{k+1}\| \to \infty$

for simplicity. Note that we also have $\lim_{k\to 0}\|x^{k+1} - x^k\|^2 = 0$ due to the sufficient descent property (3.5). From the KKT condition (3.6), we have

$$\psi_0^k(x) + \langle \lambda^{k+1}, \psi^k(x) - \eta^k \rangle \geq \psi_0^k(x^{k+1}) + \langle \lambda^{k+1}, \psi^k(x^{k+1}) - \eta^k \rangle, \quad \forall x \in \text{dom.} \atop \chi_0 \tag{3.10}$$

Let $X := \bigcap_{i\in[m]}\{x : \psi_i(x) \leq \eta_i\} \cap \text{dom}_{\chi_0}$ be the feasible domain for problem (1.1). Due to the fact that $x^k \in X$ (Proposition 2), boundedness of X (Assumption 1.3:) and strong convexity of ψ_0^k, there exists $l_0 \in \mathbb{R}$ such that $X \subset \{x : \psi_0^k(x) < l_0\}$ for all k. Then, using (3.10) for all $x \in \text{dom}_{\chi_0} \cap \{\psi_0^k(x) \leq l_0\}$ and dividing both sides by $\|\lambda^{k+1}\|$, we have

$$\psi_0^k(x)/\|\lambda^{k+1}\| + \langle \lambda^{k+1}/\|\lambda^{k+1}\|, \psi^k(x) \rangle$$
$$\geq \psi_0^k(x^{k+1})/\|\lambda^{k+1}\| + \langle \lambda^{k+1}/\|\lambda^{k+1}\|, \psi^k(x^{k+1}) \rangle. \tag{3.11}$$

Let us take $k \to \infty$ on both sides of (3.11). Note that for all $x \in \text{dom}_{\chi_0} \cap \{\psi_0^k(x) \leq l_0\}$, we have

$$\lim_{k\to\infty} \psi_0^k(x)/\|\lambda^{k+1}\| = 0, \quad \lim_{k\to\infty} \psi_0^k(x^{k+1})/\|\lambda^{k+1}\| = 0, \tag{3.12}$$

$$\lim_{k\to\infty} \psi_i^k(x) = f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x} \rangle + \frac{L_i}{2}\|x - \bar{x}\|^2 + \chi_i(x), \ i \in [m], \tag{3.13}$$

$$\lim_{k\to\infty} \psi_i^k(x^{k+1}) = f_i(\bar{x}) + \chi_i(\bar{x}) = \psi_i(\bar{x}), \ i \in [m], \tag{3.14}$$

where (3.12) is due to boundedness of $\psi_0^k(x)$ on $\text{dom}_{\chi_0} \cap \{\psi_0^k(x) \leq l_0\}$ and boundedness of $\psi_0^k(x^{k+1})$ since $x^{k+1} \in X$ which is a bounded set. Moreover, (3.13) and (3.14) use the continuity of $f_i(x)$ and $\chi_i(x)$, $i \in [m]$. Next, we consider the sequence $\{u^k = \lambda^{k+1}/\|\lambda^{k+1}\|\}$. Since $\|u^k\|$ is a bounded sequence, it has a convergent subsequence. Let $\bar{u}$ be a limit point of $\{u^k\}$ and the subsequence $\{j_k\} \subseteq \{1, 2, ..., \}$ such that $\lim_{k\to\infty} u^{j_k} = \bar{u}$. Since the subsequence of a convergent sequence is also convergent, we pass to the subsequence j_k in (3.11) and apply (3.12), (3.13) and (3.14), yielding

$$\sum_{i=1}^{m} \bar{u}_i \left[\langle \nabla f_i(\bar{x}), x - \bar{x} \rangle + \frac{L_i}{2}\|x - \bar{x}\|^2 + \chi_i(x) \right] \geq \sum_{i=1}^{m} \bar{u}_i \chi_i(\bar{x}), \tag{3.15}$$

for all $x \in \text{dom}_{\chi_0} \cap \{\psi_0^k(x) \leq l_0\}$. Hence, $\bar{x}$ minimizes $\sum_{i=1}^{m} \bar{u}_i \left[\langle \nabla f_i(\bar{x}), x - \bar{x} \rangle + \frac{L_i}{2}\|x - \bar{x}\|^2 + \chi_i(x) \right]$ on $\text{dom}_{\chi_0} \cap \{\psi_0^k(x) \leq l_0\}$. Now noting $\bar{x} \in X \subset \{\psi_0^k(x) < l_0\}$ and using the stationarity condition for optimality of $\bar{x}$, we have

$$0 \in \sum_{i=1}^{m} \bar{u}_i \left[\nabla f_i(\bar{x}) + \partial \chi_i(\bar{x}) \right] + N_{\text{dom}_{\chi_0}}(\bar{x}) = \sum_{i=1}^{m} \bar{u}_i \partial \psi_i(\bar{x}) + N_{\text{dom}_{\chi_0}}(\bar{x}), \tag{3.16}$$

where we dropped the constraint $\psi_0^k(x) \leq l_0$ due to complementary slackness and the fact that $\psi_0^k(\bar{x}) < l_0$.

Let $\mathcal{B} = \{i : \bar{u}_i > 0\}$, then we must have $\lim_{k \to \infty} \lambda_i^{j_k} = \lim_{k \to \infty} \bar{u}_i \|\lambda^{j_k}\| = \infty$ for $i \in \mathcal{B}$. Based on complementary slackness, we have $\psi_i^{j_k}(x^{j_k+1}) = \eta_i^{j_k}$ for any $i \in \mathcal{B}$ for large enough k. Due to (3.14), we have the limit: $\psi_i(\bar{x}) = \eta_i$. Therefore, the ith constraint is active at $\bar{x}$, i.e. $i \in \mathcal{A}(\bar{x})$. In view of (3.16), there exists subgradients $v_i \in \partial\psi_i(\bar{x}), i \in [m]$ and $v_0 \in N_{\mathrm{dom}_{\chi_0}}(\bar{x})$ such that

$$0 = v_0 + \sum_{i \in \mathcal{B}} \bar{u}_i v_i. \tag{3.17}$$

However, equation (3.17) contradicts the MFCQ assumption. This is because MFCQ guarantees the existence of $z \in -N^*_{\mathrm{dom}_{\chi_0}}(\bar{x})$ such that $\langle z, v_i \rangle < 0$ for all $i \in \mathcal{A}(\bar{x})$, which implies

$$0 = \langle z, v_0 + \sum_{i \in \mathcal{B}} \bar{u}_i v_i \rangle \le \sum_{i \in \mathcal{B}} \bar{u}_i \langle z, v_i \rangle \le \sum_{i \in \mathcal{B}} \bar{u}_i \max_{v \in \partial\psi_i(\bar{x})} \langle z, v \rangle < 0,$$

where the first inequality follows since $z \in -N_{\mathrm{dom}_{\chi_0}}(\bar{x})$ and $v_0 \in N_{\mathrm{dom}_{\chi_0}}(\bar{x})$ implying that $\langle z, v_0 \rangle \le 0$, the second inequality follow since $\bar{u}_i \ge 0$ and $v_i \in \partial\psi_i(\bar{x})$, and the last strict inequality follows due to uniform MFCQ (c.f. Assumption 3 and relation (2.3)) and $\bar{u}_i > 0$ for at least one $i \in \mathcal{B}$. This clearly leads to a contradiction. Hence, we conclude that $\{\lambda^{k+1}\}$ is a bounded sequence and conclude the proof.

Part 2). Without loss of generality, we assume that $\bar{x}$ is the only limit point. Since $\{\lambda^{k+1}\}$ is a bounded sequence, there exists a limit point $\bar{\lambda}$. Passing to a subsequence if necessary, we have $\lambda^{k+1} \to \bar{\lambda}$.

From the optimality condition $0 \in \partial_x \mathcal{L}_k(x^{k+1}, \lambda^{k+1})$, for any x, we have

$$\langle \nabla f_0(x^k) + \sum_{i=1}^{m} \lambda_i^{k+1} \nabla f_i(x^k), x^{k+1} - x \rangle + \chi_0(x^{k+1}) - \chi_0(x) + \langle \lambda^{k+1}, \chi(x^{k+1}) - \chi(x) \rangle$$
$$\le \frac{L_0 + \langle \lambda^{k+1}, L \rangle}{2} \left[\|x - x^k\|^2 - \|x^{k+1} - x^k\|^2 - \|x - x^{k+1}\|^2 \right]. \tag{3.18}$$

Let us take $k \to \infty$ on both sides of (3.18). We note that $\lim_{k \to \infty} \|x^k - x^{k+1}\| = 0$ due to Proposition 2, $\lim_{k \to \infty} \chi_i(x^{k+1}) = \chi_i(\bar{x})$ due to the continuity of χ_i ($i \in [m]$), and $\chi_0(\bar{x}) \le \liminf_{k \to \infty} \chi_0(x^k)$ due to the lower semi-continuity of $\chi_0(\cdot)$. It then follows that

$$\langle \nabla f_0(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla f_i(\bar{x}), \bar{x} - x \rangle + \chi_0(\bar{x}) - \chi_0(x) + \langle \bar{\lambda}, \chi(\bar{x}) - \chi(x) \rangle \le 0. \tag{3.19}$$

In other words, $\bar{x}$ is the minimizer of convex optimization problem $\min_x \langle \nabla f_0(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla f_i(\bar{x}), x \rangle + \chi_0(x) + \bar{\lambda}[\chi(x) - \eta]$ over dom_{χ_0}. Hence we have

$$0 \in \nabla f_0(\bar{x}) + \partial\chi_0(\bar{x}) + \langle \bar{\lambda}, \partial\psi(\bar{x}) \rangle. \tag{3.20}$$

In addition, using the complementary slackness, we have

$$
\begin{aligned}
0 &= \lim_{k\to\infty} \sum_{i=1}^{m} \lambda_i^{k+1} \big[\psi_i^k(x^{k+1}) - \eta_i^k \big] \\
&= \lim_{k\to\infty} \sum_{i=1}^{m} \lambda_i^{k+1} \big[f_i(x^k) + \langle \nabla f_i(x^k), x^{k+1} - x^k \rangle + \tfrac{L_i}{2} \| x^{k+1} - x^k \|^2 + \chi_i(x^{k+1}) - \eta_i^k \big] \\
&= \sum_{i=1}^{m} \bar{\lambda}_i \big[f_i(\bar{x}) + \chi_i(\bar{x}) - \eta_i \big] = \langle \bar{\lambda}, \psi(\bar{x}) - \eta \rangle,
\end{aligned}
\tag{3.21}
$$

due to the convergence $\lim_{k\to\infty} \lambda^{k+1} = \bar{\lambda}$, $\lim_{k\to\infty} \eta^k = \eta$, $\lim_{k\to\infty} \chi(x^{k+1}) = \chi(\bar{x})$ and $\lim_{k\to\infty} \| x^{k+1} - x^k \| = 0$. Putting (3.20) and (3.21) together, we conclude that $(\bar{x}, \bar{\lambda})$ satisfies the KKT condition. □

Remark 1 To show the existence of a limit point $\bar{x}$, we use Assumption 1.3: to ensure that the sequence x^k remains in a bounded domain. For the sake of conciseness, we henceforth assume the existence of a limit point $\bar{x}$ and do not delve into the technical assumption used to ensure this condition. Moreover, it should be noted that the boundedness property can be obtained under other assumptions, e.g., assuming the compactness of sublevel set $\{x : \psi_0(x) \le \psi_0(x^0)\}$ and using the sufficient descent condition (3.5), we can immediately show the existence of $\bar{x}$. However, it appears to be more challenging to show convergence via this approach when sufficient descent condition fails (e.g., in the forthcoming stochastic optimization).

3.1 Dependence of *B* on the constraint qualification.

In Theorem 1, we proved existence of a bound B on the dual multiplier. However, the size of that bound still remains unknown. Through Example 1 below, we observe that the limiting behaviour of the sequence λ^k (which largely governs the size of B) is closely tied to the magnitude of the number $c(\bar{x})$, where

$$
c(x) := - \min_{\|z\| \le 1} \max_{v \in \partial \psi_i(x)} \langle v, z \rangle.
$$

Here, the inner max follows from the relation (2.3) and outer min tries to find the best possible z that ensures MFCQ. It is observed that if $c(\bar{x})$ is a large positive number, then MFCQ is strongly satisfied and B is a reasonable bound. In contrast, if $c(\bar{x})$ is close to 0, then B can get quite large.

Example 1 Consider a two dimensional optimization problem with SCAD constraint: $\min_x \psi_0(x)$ subject to $\psi_1(x) \le \eta_1$ where $\psi_0(x) = 7 - x_1$ and $\psi_1(x) = \beta \|x\|_1 -$

$\sum_{i=1}^{2} h_{\beta,\theta}(x_i)$. Note that

$$h_{\beta,\theta}(u) = \begin{cases} 0 & \text{if } |u| \leq \beta; \\ \frac{(|u|-\beta)^2}{2(\theta-1)} & \text{if } \beta \leq |u| \leq \beta\theta; \\ \beta|u| - \frac{(\theta+1)\beta^2}{2} & \text{if } |u| \geq \beta\theta \end{cases} \tag{3.22}$$

This function fits our framework with the smoothness parameter $\frac{1}{\theta-1}$. Lets consider $\beta = 1, \theta = 5$, the level $\eta_1 = 3$ and limit point $\bar{x} = (5, 0)$. Clearly, the constraint is active at $\bar{x}$ and h is $\frac{1}{4}$-smooth function. Then, $\partial\psi(\bar{x}) = \{\begin{bmatrix} 0 \\ t \end{bmatrix} : t \in [-1, 1]\}$ as per the definition. Then, one can see that uniform MFCQ is violated at the limit point $\bar{x}$. Indeed,

$$\max_{v \in \partial\psi(x)} \langle v, z \rangle = \max_{t \in [-1,1]} tz_2 = |z_2| \not< 0,$$

implying $c(\bar{x}) = 0$. Furthermore, no λ can be found satisfying the KKT condition

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} 0 \\ t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

for all $t \in [-1, 1]$. Hence, as we get close to this limit point, bound on $\|\lambda^k\|$ will get arbitrarily large. Easy way to see this fact is to construct a subproblem at the limit point itself. After observing the feasible region for the subproblem at $(5, 0)$, it is clear that it has only one feasible solution $(5, 0)$ which gives rise to degeneracy. See Fig. 1 for more details. Figure 1a shows the well-behaved subproblem at an interior point while Fig. 1b show the degeneracy at the limit.

However, as we change level η_1 to any value either above or below 3, we do not get any violation of MFCQ. It also gives nondegenerate feasible sets at limit point and λ^k remains bounded for all k. See Fig. 2 below for more details. In particular, if $\eta_1 = 2.5 < 3$, then $\bar{x} = (3, 0)$ is the limit point. At this point, we have $\partial\psi(\hat{x}) = \{\begin{bmatrix} 0.5 \\ t \end{bmatrix} : t \in [-1, 1]\}$. Moreover, we have

$$\max_{v \in \partial\psi(\bar{x})} \langle v, z \rangle = 0.5z_1 + \max_{t \in [-1,1]} tz_2 = 0.5z_1 + |z_2|.$$

Choosing the unit vector $z = (z_1, z_2) = (-1, 0)$, we obtain that point $\bar{x}$ satisfies MFCQ with $c(\bar{x}) = 0.5$. Hence, even when the search point reaches the limit point $\hat{x}$ (i.e., $\epsilon \to 0$), the λ^k still exists. (See, in particular, Fig. 2b whose subproblem at $\bar{x}$ has a nonempty interior).

The view of the above example, we see that the limiting behavior of $\|\lambda^k\|$ (and by implication the order of B) is closely related to the "strength" of the constraint qualification MFCQ at the limit point. In order to get an apriori bound B, we use a somewhat stronger yet verifiable constraint qualification called as *strong feasibility* which is a slight modification of the CQ proposed in [4, Assumption 3].
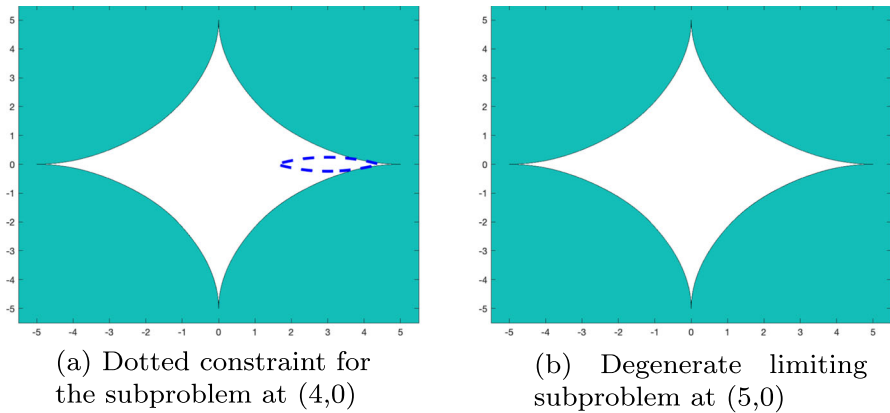
(a) Dotted constraint for the subproblem at (4,0)

(b) Degenerate limiting subproblem at (5,0)

Fig. 1 The nonconvex constraint $\psi_1(x) \leq \eta_1$ where $\eta_1 = 3$. The dotted blue curves are the subproblem constraint for two different points. Since the MFCQ is violated at $(5, 0)$, the subproblem reduces to a single feasible point at the limit point $(5, 0)$

(a) Dotted constraint for the subproblem at (2,0)

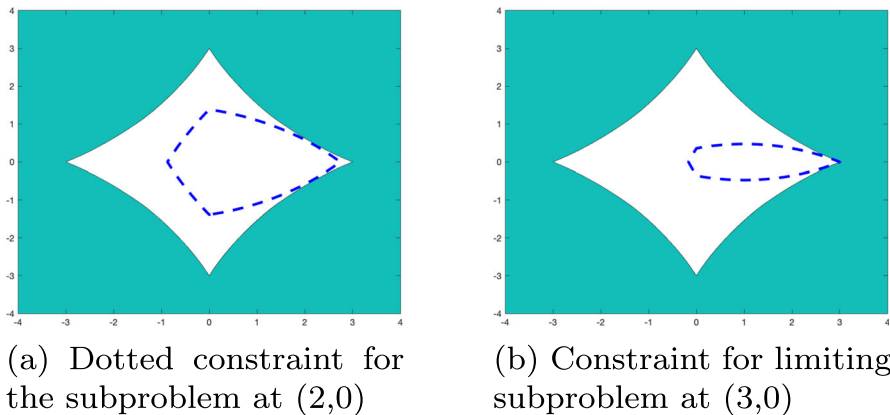(b) Constraint for limiting subproblem at (3,0)

Fig. 2 The nonconvex constraint $\psi_1(x) \leq \eta_1$ where $\eta_1 = 2.5$. The dotted blue curves are subproblem constraint for two different points. Since the MFCQ is satisfied, the limiting subproblem constraint at $(3, 0)$ is still a full dimensional set with nonempty interior

Assumption 4 [Strong feasibility CQ] There exists $\hat{x} \in \mathcal{X} := \bigcap_{i \in [m]} \{x : \psi_i(x) \leq \eta_i\} \cap \mathrm{dom}_{\chi_0}$ such that

$$\psi_i(\hat{x}) \leq \eta_i^0 - 2L_i D_{\mathcal{X}}^2 \tag{3.23}$$

where $D_{\mathcal{X}} := \max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|$ is the diameter of the set $\mathcal{X}$.

In view of Assumption 1.3:, we note that $\mathcal{X}$ is a bounded set. Hence, $D_{\mathcal{X}}$ and (consequently) Assumption 4 are well-defined. See [4] for a connection between Assumption 4 and Assumption 3. Below, we show that strong feasibility CQ leads to a fixed apriori bound on λ^k.

Lemma 2 *Suppose Assumption 4 is satisfied. Then,* $\|\lambda^k\|_1 \leq B := \frac{\psi_0(\hat{x}) - \psi_0^* + L_0 D_{\mathcal{X}}^2}{L_{\min} D_{\mathcal{X}}^2}$

Proof Note that

$$\psi_i^k(\hat{x}) = f_i(x^k) + \langle \nabla f_i(x^k), \hat{x} - x^k \rangle + \frac{L_i}{2} \|\hat{x} - x^k\|^2 + \chi_i(\hat{x})$$
$$\leq f_i(\hat{x}) + \chi_i(\hat{x}) + L_i \|\hat{x} - x^k\|^2$$
$$= \psi_i(\hat{x}) + L_i \|\hat{x} - x^k\|^2 \leq \eta_i^0 - L_i D_{\mathcal{X}}^2 < \eta_i^k - L_i D_{\mathcal{X}}^2, \quad (3.24)$$

where first inequality uses $f_i(\hat{x}) \geq f_i(x^k) + \langle \nabla f_i(x^k), \hat{x} - x^k \rangle - \frac{L_i}{2} \|\hat{x} - x^k\|^2$ (follows due to L_i-smoothness of f_i), and second inequality follows by Assumption 4 along with the fact that $x^k \in \mathcal{X}$ (see Proposition 2).

In view of (3.24), we have strict feasibility of subproblem (3.1) for all k implying that λ^{k+1} exists. Hence, we have $x^{k+1} = \operatorname{argmin}_x \psi_0^k(x) + \langle \lambda^{k+1}, \psi^k(x) \rangle$. Then, for all $x \in \operatorname{dom}_{\chi_0}$, we have

$$\psi_0^k(x^{k+1}) = \psi_0^k(x^{k+1}) + \langle \lambda^{k+1}, \psi^k(x^{k+1}) - \eta^k \rangle$$
$$\leq \psi_0^k(x) + \langle \lambda^{k+1}, \psi^k(x) - \eta^k \rangle$$

where equality follows from the complementary slackness of the KKT condition, and inequality is due to optimality of x^{k+1}. Using $x = \hat{x}$ in the above relation and combining with (3.24), we obtain

$$\psi_0^k(\hat{x}) - \psi_0^k(x^{k+1}) \geq \langle \lambda^{k+1}, \eta^k - \psi^k(\hat{x}) \rangle \geq \langle \lambda^{k+1}, L \rangle D_{\mathcal{X}}^2 \geq \|\lambda^{k+1}\|_1 L_{\min} D_{\mathcal{X}}^2$$
$$(3.25)$$

Finally, note that

$$\psi_0^k(\hat{x}) - \psi_0^k(x^{k+1}) \leq \psi_0(\hat{x}) + L_0 \|\hat{x} - x^k\|^2 - \psi_0(x^{k+1}) \leq \psi_0(\hat{x}) - \psi_0^* + L_0 D_{\mathcal{X}}^2$$

where first inequality follows by (3.24) for $i = 0$ and $\psi_0^k(x) \geq \psi_0(x)$, and second inequality follows by the definition of ψ_0^* and $D_{\mathcal{X}}$. Combining the above relation with (3.25), we get the result. Hence, we conclude the proof. □

The discussion above implies that the value of B is intricately related to the constraint qualification. While uniform MFCQ is unverifiable and does not allow for a priori bounds on B, it is widely used in nonlinear programming to ensure the existence of such a bound [2]. As observed in Fig. 1b and Fig. 2b, the actual value of B depends on the closeness of the MFCQ violation at the limit point. This situation is rare, but the current assumptions do not eliminate that possibility. Problems of this nature are ill-conditioned, and to our knowledge, no algorithm can ensure bounds on the dual in such a situation. The existing literature deals with this issue in two ways: One track assumes existence of B (similar to Theorem 1) and performs the complexity or convergence analysis; A second track assumes a stronger constraint qualification that removes the ill-conditioned problems and shows more explicit bound on the dual (similar to Lemma 2. We perform our analysis for both cases. To conclude, we henceforth assume that the bound B is a constant and do not delve into the discussion on

related constraint qualification. To substantiate that the bound B is small, we perform detailed numerical experiments in Sect. 8.

3.2 Convergence rate analysis of LCPG method

Our main goal now is to develop some non-asymptotic convergence rates for Algorithm 1.

Lemma 3 *In Algorithm 1, for $k = 1, 2, ...,$ we have*

$$\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_- \leq 2\big(L_0 + \langle \lambda^{k+1}, L \rangle\big)\|x^{k+1} - x^k\|. \qquad (3.26)$$

Proof Let $\mathcal{L}_k$ be the Lagrangian function of subproblem (3.1):

$$\mathcal{L}_k(x, \lambda) = \psi_0^k(x) + \sum_{i=1}^m \lambda_i [\psi_i^k(x) - \eta_i^k]. \qquad (3.27)$$

Using (2.1) and (3.27), we have

$$
\begin{aligned}
\partial_x & \mathcal{L}_k(x^{k+1}, \lambda^{k+1}) \\
&= \nabla f_0(x^k) + L_0(x^{k+1} - x^k) \\
&\quad + \partial \chi_0(x^{k+1}) + \sum_{i=1}^m \lambda_i^{k+1} \big[\nabla f_i(x^k) + L_i(x^{k+1} - x^k) + \partial_x \chi_i(x^{k+1}) \big] \\
&= \partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1}) + \nabla f_0(x^k) - \nabla f_0(x^{k+1}) \\
&\quad + \sum_{i=1}^m \lambda_i^{k+1} \big[\nabla f_i(x^k) - \nabla f_i(x^{k+1}) \big] \\
&\quad + \big(L_0 + \langle \lambda^{k+1}, L \rangle\big)\big(x^{k+1} - x^k\big).
\end{aligned} \qquad (3.28)
$$

Using the smoothness of $f_i(x)$, the optimality condition $0 \in \partial_x \mathcal{L}_k(x^{k+1}, \lambda^{k+1})$ and the triangle inequality, we obtain

$$\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_- \leq 2L_0\|x^{k+1} - x^k\| + 2\langle \lambda^{k+1}, L \rangle \|x^{k+1} - x^k\|. \qquad (3.29)$$

Hence we conclude the proof. □

In view of Lemma 3, we derive the complexity of LCPG to attain approximate KKT solutions in the following theorem.

Theorem 2 *Let $\alpha_k > 0$ ($k = 0, 1, .., K$) be a non-decreasing sequence and suppose that Assumption 3 holds, then there exists a constant $B > 0$ such that*

$$\sum_{k=0}^K \alpha_k \|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2 \leq 8(L_0 + B\|L\|)^2 D^2 \alpha_K, \qquad (3.30)$$

$$\sum_{k=0}^{K} \alpha_k \langle \lambda^{k+1}, |\psi(x^{k+1}) - \eta| \rangle \leq 2B\|L\|D^2\alpha_K + B\sum_{k=0}^{K} \alpha_k\|\eta - \eta^k\|, \qquad (3.31)$$

where $D = \sqrt{\frac{\psi_0(x^0)-\psi_0^*}{L_0}}$. Moreover, if we choose the index $\hat{k} \in \{0, 1, ..., K\}$ with probability $\mathbb{P}(\hat{k} = k) = \alpha_k/(\sum_{k=0}^{K} \alpha_k)$, then $x^{\hat{k}+1}$ is a randomized ϵ_K type-I KKT point with

$$\epsilon_K = \frac{1}{\sum_{k=0}^{K} \alpha_k} \max\left\{8(L_0 + B\|L\|)^2 D^2\alpha_K, \ 2B\|L\|D^2\alpha_K + B\sum_{k=0}^{K} \alpha_k\|\eta - \eta^k\|\right\}$$
$$(3.32)$$

Proof From the sufficient descent property (3.5), we have

$$\sum_{k=0}^{K} \alpha_k\|x^k - x^{k+1}\|^2 \leq \frac{2}{L_0}\sum_{k=0}^{K} \alpha_k[\psi_0(x^k) - \psi_0(x^{k+1})]$$
$$= \frac{2}{L_0}\left[\alpha_0\psi_0(x^0) + \sum_{k=1}^{K}(\alpha_k - \alpha_{k-1})\psi_0(x^k) - \alpha_K\psi_0(x^{K+1})\right]$$
$$\leq \frac{2\alpha_K}{L_0}[\psi_0(x^0) - \psi_0(x^{K+1})]$$
$$\leq \frac{2\alpha_K}{L_0}[\psi_0(x^0) - \psi_0(x^*)] = 2\alpha_K D^2, \qquad (3.33)$$

where the second inequality uses the monotonicity of sequence $\psi_0(x^k)$. In view of Theorem 1 and Cauchy-Schwarz inequality, we have $\langle \lambda^{k+1}, L \rangle \leq \|\lambda^{k+1}\|\|L\| \leq B\|L\|$. This relation and (3.26) implies

$$\|\partial_x\mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2 \leq 4(L_0 + B\|L\|)^2\|x^k - x^{k+1}\|^2.$$

Combining the above inequality with (3.33) immediately yields (3.30).

Next, we bound the error of complementary slackness. We have

$$\sum_{i=1}^{m} \lambda_i^{k+1}|\psi_i(x^{k+1}) - \eta_i|$$
$$= \sum_{i=1}^{m} \lambda_i^{k+1}|\psi_i^k(x^{k+1}) - \eta_i^k - (\eta_i - \eta_i^k) + \psi_i(x^{k+1}) - \psi_i^k(x^{k+1})|$$
$$\leq \sum_{i=1}^{m} \left[\lambda_i^{k+1}|\psi_i^k(x^{k+1}) - \eta_i^k| + \lambda_i^{k+1}(\eta_i - \eta_i^k)\right.$$
$$\left. + \lambda_i^{k+1}|f_i(x^{k+1}) - f_i(x^k) - \langle \nabla f_i(x^k), x^{k+1} - x^k \rangle - \frac{L_i}{2}\|x^{k+1} - x^k\|^2|\right]$$
$$\leq \sum_{i=1}^{m} \lambda_i^{k+1}(\eta_i - \eta_i^k) + \lambda_i^{k+1}L_i\|x^{k+1} - x^k\|^2$$

$$\leq B\|\eta - \eta^k\| + B\|L\|\|x^{k+1} - x^k\|^2 \tag{3.34}$$

where the first inequality uses the triangle inequality, the second inequality uses complementary slackness and the Lipschitz smoothness of $f_i(\cdot)$, and the last inequality follows from Cauchy-Schwartz inequality and boundedness of λ^{k+1}. Summing up (3.34) weighted by α_k for $k = 0, ..., K$, we have

$$\sum_{k=0}^{K} \alpha_k \langle \lambda^{k+1}, |\psi(x^{k+1}) - \eta| \rangle \leq \sum_{k=0}^{K} \alpha_k \left[B\|L\|\|x^{k+1} - x^k\|^2 + B\|\eta - \eta^k\| \right].$$

Combining the above result with (3.33) gives (3.31). Finally, the fact that $x^{\hat{k}+1}$ is a randomized ϵ_K type-I KKT point for ϵ_K defined in (3.32) is immediately follows from (3.30), (3.31) and Definition 3. □

The following corollary shows that the output of Algorithm 1 is a randomized $\mathcal{O}(1/K)$ KKT point under more specific parameter selection.

Corollary 1 *In Algorithm 1, suppose that all the assumptions of Theorem 2 hold. Set $\delta^k = \frac{\eta - \eta^0}{(k+1)(k+2)}$ and $\alpha_k = k + 1$. Then $x^{\hat{k}+1}$ is a randomized ϵ Type-I KKT point with*

$$\epsilon = \frac{2}{K+2} \max \left\{ 8(L_0 + B\|L\|)^2 D^2, 2B\|L\|D^2 + B\|\eta - \eta^0\| \right\} \tag{3.35}$$

Proof Notice that $\alpha_K = K + 1$, $\sum_{k=0}^{K} \alpha_k = \frac{(K+1)(K+2)}{2}$. Moreover, for $i \in [m]$ and $k \geq 0$, we have

$$\eta_i^k = \eta_i^0 + \sum_{i=0}^{k-1} \delta^i = \eta_i^0 + (\eta_i - \eta_i^0) \sum_{i=0}^{k-1} \frac{1}{(i+1)(i+2)} = \frac{k}{k+1}\eta_i + \frac{1}{k+1}\eta_i^0,$$

which implies that $\sum_{k=0}^{K} \alpha_k \|\eta - \eta^k\| = (K+1)\|\eta - \eta^0\|$. Plugging these values into (3.32) gives us the desired conclusion. □

Remark 2 Corollary 1 shows that the gradient complexity of LCPG for smooth composite constrained problems is on a par with that of gradient descent for unconstrained optimization problems. To the best of our knowledge, this is the first complexity result for a constrained problem where the constraint functions can be nonsmooth and nonconvex. Note that the convergence rate involves the unknown bound B on the Lagrangian multipliers. The presence of such a constant is not new in nonlinear programming literature [1, 8, 9, 12, 13]. Fortunately, we can safely implement LCPG method since the step-size scheme does not rely on B. On the other hand, the bound B is often a problem-dependent quantity. E.g., in [4] authors show a class of problems for which an a priori bound B can be established, or [5] shows the exact value of B for a class of nonconvex relaxations of sparse optimization problems. In such cases, our comparisons are arguably fair. Hence, throughout the paper, we make comparative statements under the assumption that B largely depends on the problem.

4 Stochastic optimization

The goal of this section is to extend our proposed framework to stochastic constrained optimization where the objective f_0 is an expectation function:

$$f_0(x) := \mathbb{E}_{\xi \in \Xi}[F(x, \xi)]. \tag{4.1}$$

Here, $F(x, \xi)$ is differentiable and ξ denotes a random variable following a certain distribution Ξ. Directly evaluating either the objective f_0 or its gradient can be computationally challenging due to the stochastic nature of the problem. To address this, we introduce the following additional assumptions.

Assumption 5 The information of f_0 is available via a stochastic first-order oracle (*SFO*). Given any input x and a random sample ξ, *SFO* outputs a stochastic gradient $\nabla F(x, \xi)$ such that

$$\mathbb{E}[\nabla F(x, \xi)] = \nabla f_0(x), \text{ and } \mathbb{E}[\|\nabla F(x, \xi) - \nabla f_0(x)\|^2] \leq \sigma^2,$$

for some $\sigma \in (0, \infty)$.

4.1 Level constrained stochastic proximal gradient

In Algorithm 2, we present a stochastic variant of `LCPG` for solving problem 1.1 with f_0 defined by (4.1). As observed in (4.2) and (4.3), the `LCSPG` method uses a mini-batch of random samples to estimate the true gradient in each iteration. It should be noted that the value $f_0(x^k)$ is presented in (4.3) only for the ease of description, it is not required when solving (3.1).

Algorithm 2 Level constrained stochastic proximal gradient (`LCSPG`)

1: **Input:** $x^0, \eta^0 < \eta, b_k, \delta^k$;
2: **for** $k = 0, 1, \ldots, K$ **do**
3:     Sample a mini-batch B_k of size b_k and compute

$$G^k = \frac{1}{b_k} \sum_{i=1}^{m_k} \nabla F(x^k, \xi_{i,k}); \tag{4.2}$$

4:     Set $\psi_0^k(x)$ by

$$\psi_0^k(x) := \langle G^k, x \rangle + \frac{\gamma_k}{2} \|x - x^k\|^2 + \chi_0(x); \tag{4.3}$$

5:     For $i = 1, \ldots, m$, set $\psi_i^k(x)$ by (3.2);
6:     Update x^{k+1} by (3.1) and update η^{k+1} by (3.3);
7: **end for**Output: $x^{\hat{k}+1}$ where $\hat{k}$ is a random index sampled from $\{0, 1, 2, ..., K\}$.

Note that the proximal point method of [4] does not need to account for a stochastic nonconvex problem separately since they solve corresponding stochastic convex subproblems using a ConEx method developed in their work. On the contrary, `LCSPG` directly applies to stochastic nonconvex function constrained problems and convex subproblems are deterministic in nature. Hence, we need to develop asymptotic convergence and convergence rates for the `LCSPG` method separately.

Let $\zeta^k = G^k - \nabla f(x^k)$ denote the error of gradient estimation. We have

$$\mathbb{E}\big[\|\zeta^k\|^2\big] = \tfrac{1}{b_k^2} \sum_{i=1}^{b_k} \mathbb{E}_\xi\big[\|\nabla F(x^k, \xi_{i,k}) - \nabla f(x^k)\|^2\big] \leq \tfrac{\sigma^2}{b_k}.$$

The following proposition summarizes some important properties of the generated solutions of `LCSPG`.

Proposition 3 *In Algorithm 2, for any $\beta_k \in (0, 2\gamma_k - L_0)$, we have*

$$\psi_0(x^{k+1}) \leq \psi_0(x^k) - \tfrac{2\gamma_k - \beta_k - L_0}{2}\|x^{k+1} - x^k\|^2 + \tfrac{\|\zeta^k\|^2}{2\beta_k}. \tag{4.4}$$

Moreover, there exists a vector $\lambda^{k+1} \in \mathbb{R}_+^m$ such that the KKT condition (3.6) (with ψ_0^k defined in (4.3)) holds.

Proof By the KKT condition, x^{k+1} is the minimizer of $\mathcal{L}_k(\cdot, \lambda^{k+1})$. Therefore, we have

$$\mathcal{L}_k(x^{k+1}, \lambda^{k+1}) + \tfrac{\gamma_k + \langle \lambda^{k+1}, L \rangle}{2}\|x^{k+1} - x\|^2 \leq \mathcal{L}_k(x, \lambda^{k+1}), \quad \forall x \in \mathcal{X}. \tag{4.5}$$

Placing $x = x^k$ in (4.5) and using (3.27), we have

$$\begin{aligned}
&\langle G^k, x^{k+1} - x^k \rangle + \chi_0(x^{k+1}) - \chi_0(x^k) \\
&\leq \sum_{i=1}^m \lambda_i^{k+1}\big[\psi_i^k(x^k) - \eta_i^k\big] - \lambda_i^{k+1}\big[\psi_i^k(x^{k+1}) - \eta_i^k\big] - \gamma_k\|x^{k+1} - x^k\|^2 \\
&\leq -\gamma_k\|x^{k+1} - x^k\|^2,
\end{aligned} \tag{4.6}$$

where the second inequality is due to the complementary slackness $\lambda_i^{k+1}\big[\psi_i^k(x^{k+1}) - \eta_i^k\big] = 0$ and strict feasibility $\lambda_i^{k+1}\big[\psi_i^k(x^k) - \eta_i^k\big] = \lambda_i^{k+1}\big[\psi_i(x^k) - \eta_i^k\big] < 0$. Using (4.6) and Lipschitz smoothness of f_0, we have

$$\begin{aligned}
\psi_0(x^{k+1}) &\leq f_0(x^k) + \langle \nabla f_0(x^k), x^{k+1} - x^k \rangle + \tfrac{L_0}{2}\|x^{k+1} - x^k\|^2 + \chi_0(x^{k+1}) \\
&= f_0(x^k) + \chi_0(x^{k+1}) + \langle G^k, x^{k+1} - x^k \rangle + \tfrac{L_0}{2}\|x^{k+1} - x^k\|^2 - \langle \zeta^k, x^{k+1} - x^k \rangle \\
&\leq \psi_0(x^k) - \tfrac{2\gamma_k - L_0}{2}\|x^{k+1} - x^k\|^2 - \langle \zeta^k, x^{k+1} - x^k \rangle \\
&= \psi_0(x^k) - \tfrac{2\gamma_k - \beta_k - L_0}{2}\|x^{k+1} - x^k\|^2 + \|\zeta^k\| \cdot \|x^{k+1} - x^k\| - \tfrac{\beta_k}{2}\|x^{k+1} - x^k\|^2 \\
&\leq \psi_0(x^k) - \tfrac{2\gamma_k - \beta_k - L_0}{2}\|x^{k+1} - x^k\|^2 + \tfrac{\|\zeta^k\|^2}{2\beta_k}. \tag{4.7}
\end{aligned}$$

Above, the last inequality uses the fact $-\frac{a}{2}x^2 + bx \leq \frac{b^2}{2a}$ for any $x, b \in \mathbb{R}, a > 0$. Showing the existence of the KKT condition follows a similar argument of proving part 2, Proposition 2. □

We prove a technical result in the following lemma which plays a crucial role in proving dual boundedness.

Lemma 4 *Let $\{X_k\}_{k\geq 1}$ be a sequence of random vectors such that $\mathbb{E}[X_k] = 0$ for all $k \geq 1$ and $\sum_{k=1}^{\infty} \sigma_k^2 \leq M < \infty$ where $\sigma_k := \sqrt{\mathbb{E}[\|X_k\|^2]}$. Then, $\lim_{k\to\infty} X_k = 0$ almost surely* (a.s.).

Proof We prove this result by contradiction. If the result does not hold then there exists $\epsilon > 0$ and $c > 0$ such that

$$\mathbb{P}\left(\limsup_k \|X_k\| \geq \epsilon\right) \geq c. \tag{4.8}$$

However, due to Chebyshev's inequality, we have $\mathbb{P}(\|X_k\| \geq \epsilon) \leq \frac{\sigma_k^2}{\epsilon^2}$. Since σ_k^2 is summable, there exists k_0 such that $\sum_{k=k_0}^{\infty} \mathbb{P}(\|X_k\| \geq \epsilon) \leq \sum_{k=k_0}^{\infty} \frac{\sigma_k^2}{\epsilon^2} < c$. Therefore, we have

$$\mathbb{P}\left(\limsup_k \|X_k\| \geq \epsilon\right) = \mathbb{P}\left(\limsup_{k\geq k_0} \|X_k\| \geq \epsilon\right) \leq \sum_{k=k_0}^{\infty} \mathbb{P}(\|X_k\| \geq \epsilon) < c.$$

The above relation contradicts (4.8). Hence, we have $\lim_{k\to\infty} X_k = 0$ a.s. □

In the following theorem, we present the main asymptotic property of LCSPG.

Theorem 3 *Suppose that $\sum_{k=0}^{\infty} b_k^{-1} < \infty$, then we have that $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$, a.s. Moreover, suppose that Assumption 3 holds, $\gamma_k < \infty$, β_k is lower bounded and $2\gamma_k - \beta_k - L_0 > 0$, then we have that (1) $\sup_k \|\lambda^k\| < \infty$ a.s., and (2) all the limit points of Algorithm 2 satisfy the KKT condition, a.s.*

Proof First, we fix notations. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space defined over the sampling minibatches $B_0, B_1, \ldots,$. Let $\mathbb{E}_k[\cdot]$ be the expectation conditioned on the sub σ-algebra generating $B_0, B_1, \ldots, B_{k-1}$. Applying it to (4.4) gives

$$\mathbb{E}_k[\psi_0(x^{k+1})] \leq \psi_0(x^k) + \frac{\sigma^2}{2b_k\beta_k}.$$

In view of the super-martingale convergence theorem [30], we have that

$$\lim_{k\to\infty} \psi_0(x^k) \text{ exists and is finite a.s., when } \sum_{k=0}^{\infty} (b_k\beta_k)^{-1} < \infty. \tag{4.9}$$

Let $C_{k+1} = \sum_{s=0}^{k} \frac{2\gamma_s - \beta_s - L_0}{2} \|x^{s+1} - x^s\|^2$ for $k \geq 0$ and $C_0 = 0$. We have

$$\mathbb{E}_k[\psi_0(x^{k+1}) + C_{k+1}] \leq \psi_0(x^k) + C_k + \frac{\sigma^2}{2b_k\beta_k}.$$

Applying the super-martingale convergence theorem [30] again we can show that the limit of $\psi_0(x^k) + C_k$ exists a.s. Together with (4.9) and lower-boundedness of β_k and $2\gamma_k - \beta_k - L_0$, we have that $\lim_{k\to\infty} \|x^{k+1} - x^k\|^2 = 0$, a.s.

Next, we prove the boundedness of $\|\lambda^k\|$. Let us consider the events

$$\mathcal{U} = \left\{\omega \in \Omega : \sup_k \|\lambda^k(\omega)\| = \infty\right\}, \quad \mathcal{A} = \left\{\omega \in \Omega : \sup_k \|G^k(\omega)\| < \infty\right\},$$

$$\mathcal{B} = \left\{\omega \in \Omega : \lim_k \|x^{k+1}(\omega) - x^k(\omega)\| = 0\right\}.$$

We just argued $\mathbb{P}(\mathcal{B}) = 1$. It is easy to see that if both conditions (i) $\mathbb{P}(\mathcal{A}) = 1$ and (ii) $\mathcal{U} \subseteq \mathcal{A}^c \cup \mathcal{B}^c$ hold, then we have $\mathbb{P}(\mathcal{U}) \leq \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{B}^c) = 0$. Hence $\{\lambda^k\}$ is a bounded sequence a.s.

Since $\{b_k^{-1}\}$ is summable, we have $\sum_{k=1}^{\infty} \mathbb{E}[\|\zeta^k\|^2] \leq \frac{\sigma^2}{b_k} < \infty$. Hence, using Lemma 4, we have that $\lim_{k\to\infty} \zeta^k = 0$ a.s. Due to the boundedness of $\nabla f(x^k)$, we have that $G^k = \zeta^k + \nabla f(x^k)$ is bounded, a.s.

We prove condition (ii) by contradiction. Suppose that our claim fails. We take an element $\omega \in \mathcal{U} \cap (\mathcal{A} \cap \mathcal{B})$ and then pass to a subsequence $\{j_k\}$ such that $\lim_{k\to\infty} \|\lambda^{j_k}(\omega)\| = \infty$. In the rest of the proof, we skip ω for brevity. Passing to another subsequence if necessary, let $\bar{x}$ be a limit point of $\{x^{j_k}\}$. By our presumption, $\bar{x}$ satisfies MFCQ. Moreover, the KKT condition implies that

$$\langle G^{j_k}, x^{j_k+1}\rangle + \chi_0(x^{j_k+1}) + \frac{\gamma_{j_k}}{2}\|x^{j_k+1} - x^{j_k}\|^2 + \langle \lambda^{j_k+1}, \psi^{j_k}(x^{j_k+1})\rangle$$
$$\leq \langle G^{j_k}, x\rangle + \chi_0(x) + \frac{\gamma_{j_k}}{2}\|x - x^{j_k}\|^2 + \langle \lambda^{j_k+1}, \psi^{j_k}(x)\rangle, \quad \forall x \in \mathrm{dom}_{\chi_0}. \quad (4.10)$$

Dividing both sides by $\|\lambda^{j_k+1}\|$ gives

$$\left[\langle G^{j_k}, x^{j_k+1}\rangle + \chi_0(x^{j_k+1}) + \frac{\gamma_{j_k}}{2}\|x^{j_k+1} - x^{j_k}\|^2\right]/\|\lambda^{j_k+1}\| + \langle u^k, \psi^{j_k}(x^{j_k+1})\rangle$$
$$\leq \left[\langle G^{j_k}, x\rangle + \chi_0(x) + \frac{\gamma_{j_k}}{2}\|x - x^{j_k}\|^2\right]/\|\lambda^{j_k+1}\| + \langle u^k, \psi^{j_k}(x)\rangle, \quad \forall x \in \mathrm{dom}_{\chi_0}. \quad (4.11)$$

where we denote $u^k = \frac{\lambda^{j_k+1}}{\|\lambda^{j_k+1}\|}$. Since $\{u^k\}$ is bounded, passing to a subsequence if needed, we have $\lim_{k\to\infty} u^k = \bar{u}$. Since $\omega \in \mathcal{A} \cap \mathcal{B}$, $\{G^{j_k}\}$ is bounded and $\{\frac{\gamma_{j_k}}{2}\|x^{j_k+1} - x^{j_k}\|^2\}$ converges to 0. Therefore, taking $k \to \infty$ on both sides of (4.11), we have

$$\langle \bar{u}, \chi(\bar{x})\rangle \leq \langle \bar{u}, \psi(\bar{x}) + \langle\nabla\psi(\bar{x}), x - \bar{x}\rangle + \frac{L}{2}\|x - \bar{x}\|^2 + \chi(x)\rangle, \quad \forall x \in \mathrm{dom}_{\chi_0}. \quad (4.12)$$

Analogous to the proof of Theorem 1, it is easy to show that $\bar{x}$ violates MFCQ, which however, contradicts Assumption 3. As a consequence of this argument, we have $\mathcal{U} \subseteq \mathcal{A}^c \cup \mathcal{B}^c$. Hence, we claim that the event $\sup_k \|\lambda^k\| < \infty$ will happen a.s. and complete our proof of the boundedness condition.

Next, we prove asymptotic convergence to KKT solutions. For any random element ω, let $\bar{x}(\omega)$ be any limit point of $\{x^k\}$. Passing to some subsequence if necessary, we

assume that $\lim_{k\to\infty} x^k = \bar{x}$ and $\lim_{k\to\infty} \lambda^{k+1} = \bar{\lambda}$.

$$\langle G^k + \nabla f(x^k)\lambda^{k+1}, x^{k+1} - x \rangle + \chi_0(x^{k+1}) - \chi_0(x) + \langle \lambda^{k+1}, \chi(x^{k+1}) - \chi(x) \rangle$$
$$\leq \tfrac{\gamma_k + \langle \lambda^{k+1}, L \rangle}{2} \left(\|x - x^k\|^2 - \|x^{k+1} - x^k\|^2 - \|x - x^{k+1}\|^2 \right).$$

Moreover, we have

$$\langle G^k, x^{k+1} - x \rangle = \langle \nabla f_0(x^k), x^{k+1} - x \rangle + \langle \zeta^k, x^{k+1} - x \rangle$$
$$= \langle \nabla f_0(x^k), x^{k+1} - x \rangle + \langle \zeta^k, x^{k+1} - x^k \rangle + \langle \zeta^k, x^k - x \rangle$$
$$\geq \langle \nabla f_0(x^k), x^{k+1} - x \rangle - \|\zeta^k\| \|x^{k+1} - x^k\| + \langle \zeta^k, x^k - x \rangle.$$

Combining the above two results, we have

$$\langle \nabla f_0(x^k) + \nabla f(x^k)\lambda^{k+1}, x^{k+1} - x \rangle + \chi_0(x^{k+1}) - \chi_0(x) + \langle \lambda^{k+1}, \chi(x^{k+1}) - \chi(x) \rangle$$
$$\leq \tfrac{\gamma_k + \langle \lambda^{k+1}, L \rangle}{2} \left(\|x - x^k\|^2 - \|x^{k+1} - x^k\|^2 - \|x - x^{k+1}\|^2 \right)$$
$$+ \|\zeta^k\| \|x^{k+1} - x^k\| + \langle \zeta^k, x - x^k \rangle$$
$$\leq \tfrac{\gamma_k + \langle \lambda^{k+1}, L \rangle}{2} \left(\|x - x^k\|^2 - \|x - x^{k+1}\|^2 \right) + \tfrac{\|\zeta^k\|^2}{2(\gamma_k + \langle \lambda^{k+1}, L \rangle)} + \langle \zeta^k, x - x^k \rangle.$$

Taking $k \to \infty$ in the above relation and noting that almost surely we have $\lim_{k\to\infty} \zeta^k = 0$ and $\lim_{k\to\infty} \|x^k - x^{k+1}\| = 0$, then

$$\langle \nabla f_0(\bar{x}) + \nabla f(\bar{x})\lambda^{k+1}, \bar{x} - x \rangle + \chi_0(\bar{x}) - \chi_0(x) + \langle \lambda^{k+1}, \chi(\bar{x}) - \chi(x) \rangle \leq 0, \quad a.s.$$

Using an argument similar to the one in Theorem 1, we can show that $\bar{x}$ is almost surely a KKT point. □

Our next goal is to develop the iteration complexity of Algorithm 2. To achieve this goal, we need to assume that the dual is uniformly bounded, namely, condition (3.9) holds for all the random events. While this condition is stronger than the almost sure boundedness of λ^{k+1} shown by Theorem 3, it is indeed satisfied in many scenarios, e.g., when strong feasibility (Assumption 4) holds or other scenarios described in [4, 5]. We present the main complexity result in the following theorem.

Theorem 4 *Suppose that condition* (3.9) *holds. Then, the sequence* $\{(x^{k+1}, \lambda^{k+1})\}$ *satisfies that*

$$\sum_{k=0}^{K} \tfrac{\alpha_k(2\gamma_k - \beta_k - L_0)}{4(\gamma_k + L_0 + 2B\|L\|)^2} \|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2$$

$$\leq L_0 D^2 \alpha_K + \sum_{k=0}^{K} \left(\tfrac{\alpha_k(2\gamma_k - \beta_k - L_0)}{2(\gamma_k + L_0 + 2B\|L\|)^2} + \tfrac{\alpha_K}{2\beta_k} \right) \|\zeta^k\|^2 \qquad (4.13)$$

$$\sum_{k=0}^{K} \alpha_k (2\gamma_k - \beta_k - L_0)\langle \lambda^{k+1}, |\psi(x^{k+1}) - \eta| \rangle$$

$$\leq 2BL_0 \|L\| D^2 \alpha_K + B\|L\| \sum_{k=0}^{K} \frac{\alpha_K \|\zeta^k\|^2}{\beta_k}$$

$$+ B\sum_{k=0}^{K} \alpha_k (2\gamma_k - \beta_k - L_0)\|\eta - \eta^k\|. \tag{4.14}$$

Proof First, appealing to (4.3), (2.1) and (3.27), we have

$$\partial_x \mathcal{L}_k(x^{k+1}, \lambda^{k+1})$$
$$= \partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1}) + \nabla f_0(x^k) - \nabla f_0(x^{k+1})$$
$$+ \sum_{i=1}^{m} \lambda_i^{k+1}[\nabla f_i(x^k) - \nabla f_i(x^{k+1})]$$
$$+ (\gamma_k + \langle \lambda^{k+1}, L \rangle)(x^{k+1} - x^k) + \zeta^k.$$

It follows that

$$\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-$$
$$\leq \|\nabla f_0(x^k) - \nabla f_0(x^{k+1})\| + \|\zeta^k\|$$
$$+ \sum_{i=1}^{m} \lambda_i^{k+1} \|\nabla f_i(x^k) - \nabla f_i(x^{k+1})\| + (\gamma_k + \langle \lambda^{k+1}, L \rangle)\|x^{k+1} - x^k\|$$
$$\leq (\gamma_k + L_0 + 2\langle \lambda^{k+1}, L \rangle)\|x^{k+1} - x^k\| + \|\zeta^k\|.$$

In view of the above result and basic inequality $(a + b)^2 \leq 2a^2 + 2b^2$, we have

$$\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2 \leq 2(\gamma_k + L_0 + 2B\|L\|)^2 \|x^{k+1} - x^k\|^2 + 2\|\zeta^k\|^2. \tag{4.15}$$

Let us denote an auxiliary sequence $C_k = \begin{cases} \psi_0(x^0) & k = 0 \\ \psi_0(x^k) - \sum_{s=0}^{k-1} \frac{\|\zeta^s\|^2}{2\beta_s} & k > 0 \end{cases}$. Proposition 3 implies that

$$\frac{2\gamma_k - \beta_k - L_0}{2}\|x^{k+1} - x^k\|^2 \leq C_k - C_{k+1}. \tag{4.16}$$

Putting this relation and (4.15) together, we have

$$\frac{2\gamma_k - \beta_k - L_0}{4(\gamma_k + L_0 + 2B\|L\|)^2}\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2 \leq C_k - C_{k+1} + \frac{2\gamma_k - \beta_k - L_0}{2(\gamma_k + L_0 + 2B\|L\|)^2}\|\zeta^k\|^2. \tag{4.17}$$

Summing up (4.17) over $k = 0, 1, \ldots, K$ weighted by α_k leads to

$$\sum_{k=0}^{K} \frac{\alpha_k(2\gamma_k - \beta_k - L_0)}{4(\gamma_k + L_0 + 2B\|L\|)^2} \|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|^2$$

$$\leq \sum_{k=0}^{K} \alpha_k(C_k - C_{k+1}) + \sum_{k=0}^{K} \frac{\alpha_k(2\gamma_k - \beta_k - L_0)}{2(\gamma_k + L_0 + 2B\|L\|)^2} \|\zeta^k\|^2.$$

Moreover, since $\{C_k\}$ is monotonically decreasing, we have

$$\sum_{k=0}^{K} \alpha_k(C_k - C_{k+1}) \leq \alpha_0 C_0 + \sum_{k=1}^{K}(\alpha_k - \alpha_{k-1})C_k - \alpha_K C_{K+1}$$

$$\leq \alpha_K(C_0 - C_{K+1}) \leq L_0 D^2 \alpha_K + \alpha_K \sum_{k=0}^{K} \frac{\|\zeta^k\|^2}{2\beta_k}.$$

Combining the above two relations leads to our first result (4.13).

For the second part, note that (3.34) remains valid in the stochastic setting. Putting (3.34) and (4.16) together, we obtain

$$(2\gamma_k - \beta_k - L_0)\langle \lambda^{k+1}, |\psi(x^{k+1}) - \eta|\rangle \leq 2B\|L\|(C_k - C_{k+1}) + B(2\gamma_k - \beta_k - L_0)\|\eta - \eta^k\|.$$

Multiplying both ends by α_k and then summing up the resulting terms over $k = 0, \ldots, K$ gives (4.14). $\qquad\square$

We next obtain more specific convergence rate by choosing the parameters properly.

Corollary 2 *In Algorithm LCSPG, set $\gamma_k = L_0$, $\beta_k = L_0/2$, $\alpha_k = k + 1$, $b_k = K + 1$ and $\delta^k = \frac{(\eta - \eta^0)}{(k+1)(k+2)}$. Then $x^{\hat{k}+1}$ is a randomized ϵ type-I KKT point with*

$$\epsilon = \frac{4}{K+2} \max \left\{ 8(L_0 + B\|L\|)^2 \left(D^2 + \frac{17\sigma^2}{16L_0^2}\right), 2B\|L\|D^2 + \frac{2B\|L\|\sigma^2}{L_0^2} + \frac{B\|\eta - \eta^0\|}{2} \right\}.$$

Proof Plugging in the value of $\gamma_k, \alpha_k, \beta_k$ in the relation (4.13) and taking expectation over all the randomness, we have

$$\frac{L_0}{32(L_0 + B\|L\|)^2} \sum_{k=0}^{K}(k+1)\mathbb{E}[\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2]$$

$$\leq L_0 D^2(K+1) + \sum_{k=0}^{K}\left(\frac{L_0(k+1)}{16(L_0 + B\|L\|)^2} + \frac{k+1}{L_0}\right)\mathbb{E}[\|\delta^k\|^2]$$

$$\leq L_0 D^2(K+1) + \frac{17\sigma^2}{16L_0}(K+1).$$

Moreover, due to the random sampling of $\hat{k}$, we have

$$\mathbb{E}_{\hat{k}}\big[\|\partial_x \mathcal{L}(x^{\hat{k}+1}, \lambda^{\hat{k}+1})\|_-^2\big] = \tfrac{2}{(K+1)(K+2)} \sum_{k=0}^{K} (k+1)\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2.$$

Combining the above two results, we have

$$\mathbb{E}\big[\|\partial_x \mathcal{L}(x^{\hat{k}+1}, \lambda^{\hat{k}+1})\|_-^2\big] \leq \tfrac{32(L_0+B\|L\|)^2}{K+2}\Big(D^2 + \tfrac{17\sigma^2}{16L_0^2}\Big).$$

Second, plugging in the values of γ_k, β_k and δ^k in (4.14), we have

$$\tfrac{L_0}{2} \sum_{k=0}^{K} (k+1)\langle \lambda^{k+1}, |\psi(x^{k+1}) - \eta|\rangle \leq 2BL_0\|L\|D^2(K+1)$$
$$+ 2B\|L\|\tfrac{\sigma^2(K+1)}{L_0} + \tfrac{BL_0\|\eta-\eta^0\|(K+1)}{2}. \tag{4.18}$$

It then follows from (4.18) and the definition of $\hat{k}$ that

$$\mathbb{E}\big[\langle \lambda^{\hat{k}+1}, |\psi(x^{\hat{k}+1}) - \eta|\rangle\big] \leq \tfrac{4}{K+2}\Big\{2B\|L\|D^2 + \tfrac{2B\|L\|\sigma^2}{L_0^2} + \tfrac{B\|\eta-\eta^0\|}{2}\Big\}.$$

This completes our proof. $\qquad\qquad\square$

Remark 3 In order to obtain some ϵ-error in satisfying the type I KKT condition, LCSPG requires a number of $\mathcal{O}(\varepsilon^{-2})$ calls to the *SFO*, which matches the complexity bound of stochastic gradient descent for unconstrained nonconvex optimization [15]. Moreover, due to minibatching, LCSPG obtains an even better $\mathcal{O}(\varepsilon^{-1})$ complexity in the number of evaluations of $f_i(x)$ and $\nabla f_i(x)$ ($i \in [m]$).

4.2 Level constrained stochastic variance reduced gradient descent

We consider the finite sum problem:

$$f_0(x) = \tfrac{1}{n} \sum_{i=1}^{n} F(x, \xi_i), \tag{4.19}$$

where each $F(x, \xi_i)$ is Lipschitz smooth with the parameter L_0, $i = 1, 2, \ldots, n$. To further improve the convergence performance in this setting, we present a new variant of the stochastic gradient method by extending the stochastic variance reduced gradient descent to the constrained setting.

We present the level constrained stochastic variance-reduced gradient descent (LCSVRG) in Algorithm 3, which extends the nonconvex variance reduced mirror descent (see [20]) to handle nonlinear constraint. Algorithm 3 can be viewed as a double-loop algorithm in which the outer loop computes the full gradient $\nabla f(x^k)$

Algorithm 3 Level constrained stochastic variance-reduced gradient descent (LCSVRG)

1: **Input:** $x^0, x^{-1}, \eta_0 < \eta, T$;
2: **for** $k = 0, 1, \ldots, K$ **do**
3:     **if** $k\%T == 0$ **then**
4:         $G^k = \nabla f_0(x^k)$;
5:     **else**
6:         Sample a mini-batch B_k of size b uniformly at random and compute

$$G^k = \tfrac{1}{b} \sum_{i \in B_k} \left[\nabla F(x^k, \xi_i) - \nabla F(x^{k-1}, \xi_i) \right] + G^{k-1}; \qquad (4.20)$$

7:     **end if**
8:     Set $\psi_0^k(x)$ by (4.3) and set $\psi_i^k(x)$ by (3.2) for $i \in [m]$;
9:     Update x^{k+1} by (3.1) and update η^{k+1} by (3.3);
10: **end for**

---

once every T iterations and the nested loop performs stochastic proximal gradient updates based on an unbiased estimator of the true gradient. In this view, we let k indicate the tth iteration at the rth epoch, for some values t and r. Then we use k and (r, t) interchangeably throughout the rest of this section. We keep the notation ζ^k (or $\zeta^{(r,j)}$) for expressing $G^k - \nabla f(x^k)$ and note that $\zeta^{(r,0)} = 0$.

Our next goal is to develop some iteration complexity results of LCSVRG. We skip the asymptotic analysis since it is similar to that of LCSPG. The following Lemma (see [20, Lemma 6.10]) presents a key insight of Algorithm 3 that the variance is controlled by the point distances. We provide proof for completeness.

Lemma 5 *In Algorithm 3, G^k is an unbiased estimator of $\nabla f_0(x^k)$. Moreover, Let (r, t) correspond to k. If $t > 0$, then we have*

$$\mathbb{E}\big[\|\zeta^{(r,t)}\|^2\big] \leq \frac{L_0^2}{b} \sum_{i=0}^{t-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big].$$

Proof We prove the first part by induction. When $k = 0$, we have $G^0 = \nabla f_0(x^0)$. Then for $k > 0$, if $k\%T == 0$, we have $G^k = \nabla f(x^k)$ by definition. Otherwise, we have

$$\mathbb{E}_k\big[G^k\big] = \nabla f(x^k) - \nabla f(x^{k-1}) + G^{k-1} = \nabla f_0(x^k)$$

by induction hypothesis $\mathbb{E}_{k-1}\big[G^{k-1}\big] = \nabla f(x^{k-1})$.

Next, we estimate the variance of the stochastic gradient. Appealing to (4.20), we have

$$\mathbb{E}_k\big[\|\zeta^k\|^2\big] = \mathbb{E}\big[\big\| \tfrac{1}{b} \sum_{i \in B_k} \left[\nabla F(x^k, \xi_i) - \nabla F(x^{k-1}, \xi_i) \right] + G^{k-1} - \nabla f(x^k)\big\|^2\big]$$

$$= \mathbb{E}\big[\big\| \tfrac{1}{b} \sum_{i \in B_k} \left[\nabla F(x^k, \xi_i) - \nabla F(x^{k-1}, \xi_i) \right] - \left[\nabla f(x^k) - \nabla f(x^{k-1}) \right] + \zeta^{k-1}\big\|^2\big]$$

$$= \mathbb{E} \Big\| \frac{1}{b} \sum_{i \in B_k} \big[\nabla F(x^k, \xi_i) - \nabla F(x^{k-1}, \xi_i) - \nabla f(x^k) + \nabla f(x^{k-1}) \big] \Big\|^2 + \big\| \zeta^{k-1} \big\|^2$$

$$\leq \tfrac{1}{b^2} \sum_{i \in B_k} \mathbb{E}_\xi \| \nabla F(x^k, \xi) - \nabla F(x^{k-1}, \xi) \|^2 + \big\| \zeta^{k-1} \big\|^2$$

$$\leq \tfrac{L_0^2}{b} \| x^k - x^{k-1} \|^2 + \big\| \zeta^{k-1} \big\|^2,$$

where the third equality uses the independence of B_k and ζ^{k-1}, the first inequality uses the bound $\mathrm{Var}(x) \leq \mathbb{E}\|x\|^2$, and the second inequality uses the Lipschitz smoothness of $F(\cdot, \xi)$. Taking expectation over all the randomness generating $B_{(r,1)}, B_{(r,2)}, \ldots, B_{(r,t)}$, we have

$$\mathbb{E}\big[\|\zeta^k\|^2\big] \leq \tfrac{L_0^2}{b} \sum_{i=1}^{t} \mathbb{E}\big[\|x^{(r,i)} - x^{(r,i-1)}\|^2\big].$$

□

The next Lemma shows that the generated solutions satisfy a property of sufficient descent on expectation.

Lemma 6 *Assume that $\gamma_k = \gamma$ and $\beta_k = \beta$ and $\tilde{L} := \frac{2\gamma - \beta - L_0}{2} - \frac{L_0^2(T-1)}{2\beta b} > 0$. Then we have*

$$\tilde{L} \sum_{j=0}^{t} \mathbb{E}\big[\|x^{(r,j+1)} - x^{(r,j)}\|^2\big] \leq \mathbb{E}\big[\psi_0(x^{(r,0)})\big] - \mathbb{E}\big[\psi_0(x^{(r,t+1)})\big], \quad 0 \leq t < T.$$

(4.21)

Proof In view of (3), at the jth iteration of the rth epoch, we have

$$\psi_0(x^{(r,j+1)}) \leq \psi_0(x^{(r,j)}) - \tfrac{2\gamma - \beta - L_0}{2} \|x^{(r,j+1)} - x^{(r,j)}\|^2 + \tfrac{\|\zeta^{(r,j)}\|^2}{2\beta}.$$

Summing up the above result over $j = 0, 1, 2, ..., t$ $(t < T)$ and using Lemma 5, we have

$$\tfrac{2\gamma - \beta - L_0}{2} \sum_{j=0}^{t} \mathbb{E}\big[\|x^{(r,j+1)} - x^{(r,j)}\|^2\big]$$

$$\leq \mathbb{E}\big[\psi_0(x^{(r,0)})\big] - \mathbb{E}\big[\psi_0(x^{(r,t+1)})\big] + \tfrac{L_0^2}{2\beta b} \sum_{j=0}^{t} \sum_{i=0}^{j-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big]$$

$$\leq \mathbb{E}\big[\psi_0(x^{(r,0)})\big] - \mathbb{E}\big[\psi_0(x^{(r,t+1)})\big] + \tfrac{L_0^2 t}{2\beta b} \sum_{i=0}^{t-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big]$$

$$\leq \mathbb{E}\big[\psi_0(x^{(r,0)})\big] - \mathbb{E}\big[\psi_0(x^{(r,t+1)})\big] + \tfrac{L_0^2(T-1)}{2\beta b} \sum_{i=0}^{t-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big].$$

Here we use $\sum_{j=0}^{-1} \cdot = 0$. □

We present the main convergence property of Algorithm 3 in the next theorem.

Theorem 5 *Suppose that condition* (3.9) *and assumptions of Lemma* 6 *hold,* $b \geq 2T$ *and* $K = r_0 T + j_0$ *for some* $r_0, j_0 \geq 0$. *Let* $\{\alpha_k\}$ *be a non-decreasing sequence and* $\{\alpha_{(r,j)}\}$ *be its equivalent form in* (r, j) *notations. Suppose that* $\alpha_{(r,j)} = \alpha_{(r,0)}$ *for* $j = 1, 2, ..., T - 1$. *Then we have*

$$
\sum_{k=0}^{K} \alpha_k \mathbb{E}\big[\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_{-}^2\big]
$$
$$
\leq 8 \tilde{L}^{-1} L_0 (\gamma + L_0 + B\|L\|)^2 D^2 \alpha_{(r_0,0)}, \tag{4.22}
$$
$$
\sum_{k=0}^{K} \alpha_k \mathbb{E}\big[\langle \lambda^{k+1}, |\psi(x^{k+1}) - \eta|\rangle\big]
$$
$$
\leq B \sum_{k=0}^{K} \alpha_k \|\eta - \eta_k\| + B \tilde{L}^{-1} \|L\| L_0 D^2 \alpha_{(r_0,0)}. \tag{4.23}
$$

Moreover, if we take $T = \lceil \sqrt{n} \rceil, b = 8T, \gamma = L_0, \beta = L_0/2,$ *and* $\alpha_k = T \lfloor k/T \rfloor + 1,$ *and set* $\delta^k = \frac{\eta - \eta^0}{(k+1)(k+2)}$. *Then* $x^{\hat{k}+1}$ *is a randomized Type-I* ϵ-KKT *point with*

$$
\epsilon = \frac{K+1}{(K-T+1)^2} \max \big\{ 128(2L_0 + B\|L\|)^2 D^2, B\|\eta - \eta^0\| + 16B\|L\|D^2 \big\}. \tag{4.24}
$$

Proof First, using Lemma 5 and the assumption that $b \geq 2T$, for any $t \leq T - 1$ we have

$$
\sum_{j=0}^{t} \mathbb{E}\big[\|\zeta^{(r,j)}\|^2\big]
$$
$$
\leq \frac{L_0^2}{b} \sum_{j=0}^{t} \sum_{i=0}^{j-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big]
$$
$$
\leq \frac{L_0^2 t}{b} \sum_{j=0}^{t-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big]
$$
$$
\leq \frac{L_0^2}{2} \sum_{j=0}^{t-1} \mathbb{E}\big[\|x^{(r,i+1)} - x^{(r,i)}\|^2\big]. \tag{4.25}
$$

Note that (4.15) still holds. Therefore, combining (4.15) and (4.25) leads to

$$
\sum_{j=0}^{t} \mathbb{E}\big[\|\mathcal{L}(x^{(r,j+1)}, \lambda^{(r,j+1)})\|_{-}^2\big]
$$
$$
\leq 2(\gamma + L_0 + 2B\|L\|)^2 \sum_{j=0}^{t} \mathbb{E}\big[\|x^{(r,j+1)} - x^{(r,j)}\|^2\big]
$$
$$
+ 2 \sum_{j=0}^{t} \mathbb{E}\big[\|\zeta^{(r,j)}\|^2\big]
$$
$$
\leq 8(\gamma + L_0 + B\|L\|)^2 \sum_{j=0}^{t} \mathbb{E}\big[\|x^{(r,j+1)} - x^{(r,j)}\|^2\big].
$$

It then follows from Lemma 6 that

$$
\sum_{j=0}^{t} \mathbb{E}\big[\|\partial_x \mathcal{L}(x^{(r,j+1)}, \lambda^{(r,j+1)})\|_{-}^2\big]
$$
$$
\leq 8 \tilde{L}^{-1} (\gamma + L_0 + B\|L\|)^2 \mathbb{E}\big[\psi_0(x^{(r,0)}) - \psi_0(x^{(r,t+1)})\big]. \tag{4.26}
$$

Let $K = r_0 T + j_0$. Summing up the above inequality weighted by α_k and exchanging the notation $\alpha_k \leftrightarrow \alpha_{(r,j)}$, then we have

$$
\sum_{k=0}^{K} \alpha_k \mathbb{E}\big[\|\partial_x \mathcal{L}(x^{k+1}, \lambda^{k+1})\|_{-}^2\big]
$$

$$\begin{aligned}
&= \sum_{r=0}^{r_0-1}\sum_{j=0}^{T-1}\alpha_{(r,j)}\mathbb{E}\big[\|\partial_x\mathcal{L}(x^{(r,j+1)},\lambda^{(r,j+1)})\|_-^2\big]\\
&\quad + \sum_{j=0}^{j_0}\alpha_{(r_0,j)}\mathbb{E}\big[\|\partial_x\mathcal{L}(x^{(r_0,j+1)},\lambda^{(r_0,j+1)})\|_-^2\big]\\
&\le 8\tilde{L}^{-1}(\gamma+L_0+B\|L\|)^2\Big\{\sum_{r=0}^{r_0-1}\alpha_{(r,0)}\mathbb{E}[\psi_0(x^{(r,0)})-\psi_0(x^{(r+1,0)})]\\
&\quad + \alpha_{(r_0,0)}\mathbb{E}[\psi_0(x^{(r_0,0)})-\psi_0(x^{(r_0,j_0+1)})]\Big\}\\
&\le 8\tilde{L}^{-1}(\gamma+L_0+B\|L\|)^2\alpha_{(r_0,0)}\mathbb{E}[\psi_0(x^{(0,0)})-\psi_0(x^{(r_0,j_0+1)})]\\
&\le 8\tilde{L}^{-1}L_0(\gamma+L_0+B\|L\|)^2D^2\alpha_{(r_0,0)}. \tag{4.27}
\end{aligned}$$

Above, the first inequality applies (4.26) and uses $x^{(r,T)}=x^{(r+1,0)}$ while the second inequality uses the monotonicity of $\{\psi_0(x^k)\}$ and an argument similar to (3.33).

The second part is similar to the argument of Theorem 4. Particularly, combining (3.34) and (4.21) gives

$$\begin{aligned}
\sum_{j=0}^{t}\mathbb{E}\langle\lambda^{(r,j+1)},|\psi_i(x^{(r,j+1)})-\eta|\rangle &\le B\tilde{L}^{-1}\|L\|\mathbb{E}[\psi_0(x^{(r,0)})-\psi_0(x^{(r,t+1)})]\\
&\quad + B\sum_{j=0}^{t}\|\eta-\eta^{(r,j+1)}\| \tag{4.28}
\end{aligned}$$

Consequently, using the above relation and an argument similar to show (3.33), we deduce

$$\begin{aligned}
\sum_{k=0}^{K}&\alpha_k\mathbb{E}\langle\lambda^{k+1},|\psi(x^{k+1})-\eta|\rangle\\
&= \sum_{r=0}^{r_0-1}\sum_{j=0}^{T-1}\alpha_{(r,j)}\mathbb{E}\langle\lambda^{(r,j+1)},|\psi(x^{(r,j+1)})-\eta|\rangle\\
&\quad + \sum_{j=0}^{j_0}\alpha_{(r_0,j)}\mathbb{E}\langle\lambda^{(r_0,j+1)},|\psi(x^{(r_0,j+1)})-\eta|\rangle\\
&= B\sum_{k=0}^{K}\alpha_k\|\eta-\eta_k\|\\
&\quad + B\tilde{L}^{-1}\|L\|\mathbb{E}\Big\{\sum_{r=0}^{r_0-1}\alpha_{(r,0)}[\psi_0(x^{(r,0)})-\psi_0(x^{(r+1,0)})]\\
&\quad + \alpha_{(r_0,0)}[\psi_0(x^{(r_0,0)})-\psi_0(x^{(r_0,j_0+1)})]\Big\}\\
&\le B\sum_{k=0}^{K}\alpha_k\|\eta-\eta_k\| + B\tilde{L}^{-1}\|L\|L_0D^2\alpha_{(r_0,0)}.
\end{aligned}$$

Therefore, we complete the proof of (4.22) and (4.23).

Using the provided parameter setting, we have $\tilde{L} = \frac{2\gamma-L_0-\beta}{2} - \frac{L_0^2(T-1)}{2\beta b} \ge \frac{L_0}{4} - \frac{L_0}{8} = \frac{L_0}{8}$. Moreover, since $\alpha_k = T\lfloor k/T\rfloor+1$, we have $\alpha_k \le T\cdot k/T+1 \le k+1$. It is easy to check

$$\sum_{k=0}^{K}\alpha_k = T+\sum_{k=T}^{K}\big(\lfloor\tfrac{k}{T}\rfloor T+1\big) \ge T+\sum_{k=T}^{K}\big[[\tfrac{k}{T}-1]T+1\big] \ge \frac{(K-T+1)^2}{2}.$$

$\square$

Remark 4 It is interesting to compare the performance of LCSVRG with the other two level constrained first-order methods in the finite sum setting (4.19). Similar to LCPG, LCSVRG runs for $\mathcal{O}(\varepsilon^{-1})$ iterations to compute Type-I ϵ-KKT point.

Moreover, LCSVRG has an appealing feature that the number of stochastic gradient $\nabla F(x, \xi)$ computed can be significantly reduced for a large value of n. Specifically, Algorithm 3 requires a full gradient $\nabla f_0(x)$ every T iterations, which contributes $N_1 = \mathcal{O}(n\lceil \frac{K}{T} \rceil) = \mathcal{O}(\sqrt{n}K)$ stochastic gradient computations. During the other iterations, Algorithm 3 invokes a batch of size $b = \mathcal{O}(T)$ each time, exhibiting a complexity of $N_2 = \mathcal{O}(bK) = \mathcal{O}(\sqrt{n}K)$. Therefore, the total number of stochastic gradient computations is $N = N_1 + N_2 = \mathcal{O}(\sqrt{n}K)$. This is better than the $O(nK)$ stochastic gradients needed by LCPG. Moreover, it is better than the bound $O(K^2)$ of LCSPG when K is at an order larger than $\Omega(\sqrt{n})$, which corresponds to a higher accuracy regime of $\epsilon \ll \frac{1}{\sqrt{n}}$. The complexities of all the proposed algorithms for getting some ε-KKT solutions are listed in Table 2.

Remark 5 While we mainly discuss the finite-sum objective (4.19), it is possible to extend the variance reduction technique to handle the expectation-based objective (4.1) and improve the $\mathcal{O}(\varepsilon^{-2})$ bound of LCSPG to $\mathcal{O}(\varepsilon^{-3/2})$. To achieve this goal, we impose an additional assumption that $F(x, \xi)$ is L_0-Lipschitz smooth for each ξ in the support set. We choose to omit a detailed discussion on this particular extension, as the technical development for this can be readily derived from the arguments in Sec. 6.5.2. [21] and our previous analysis.

5 Smooth optimization of nonsmooth constrained problems

In this section, we consider the constrained problem (1.1) with nonsmooth objective and nonsmooth constraint functions. We assume that f_i ($i \in \{0, 1, ..., m\}$) exhibits a difference-of-convex (DC) structure $f_i(x) := g_i(x) - h_i(x)$: 1) h_i is an L_{h_i}-Lipschitz-smooth convex function and 2) g_i is a structured nonsmooth convex function:

$$g_i(x) = \max_{y_i \in \mathcal{Y}_i} \langle A_i x, y_i \rangle - p_i(y_i),$$

where $A_i \in \mathbb{R}^{a_i \times n}$ is a linear mapping, $\mathcal{Y}_i \subset \mathbb{R}^{a_i}$ is a convex compact set and $p_i : \mathcal{Y}_i \to \mathbb{R}$ is a convex continuous function. In view of such a nonsmooth structure, we can not simply apply the LCPG method, as the crucial quadratic upper-bound on $f_i(x)$ does not hold in the nonsmooth cases. However, as pointed out by Nesterov [29], the nonsmooth convex function g_i can be closely approximated by a smooth convex function. Let us denote $\widehat{y}_i := \arg\min_{y_i \in \mathcal{Y}_i} \|y_i\|$, $D_{\mathcal{Y}_i} := \max_{y_i \in \mathcal{Y}_i} \|y_i - \widehat{y}_i\|$ and define the approximation function

$$g_i^{\beta_i}(x) := \max_{y_i \in \mathcal{Y}_i} \langle A_i x, y_i \rangle - p(y_i) - \frac{\beta_i}{2} \|y_i - \widehat{y}_i\|^2, \ \ f_i^{\beta_i}(x) := g_i^{\beta_i}(x) - h_i(x),$$
$$\text{where } \beta_i > 0.$$

Given some properly chosen smoothing parameter β_i, we propose to apply LCPG to solve the following smooth approximation problem:

$$\min_x \ \psi_0^{\beta_0}(x) = f_0^{\beta_0}(x) + \chi_0(x)$$

$$\text{s.t.} \quad \psi_i^{\beta_i}(x) = f_i^{\beta_i}(x) + \chi_i(x) \le \eta_i \quad i = 1, \dots, m. \tag{5.1}$$

Prior to the analysis of our algorithm, we need to develop some properties of the smooth function $f_i^{\beta_i}$. We first present a key Lemma which builds some important connection between the quadratic approximation of smooth function and Lipschitz smoothness. The proof is left in Appendix A.

Lemma 7 *Suppose $p(\cdot)$ is continuously differentiable function satisfying*

$$-\tfrac{\mu}{2}\|x - y\|^2 \le p(x) - p(y) - \langle \nabla p(y), x - y \rangle \le \tfrac{L}{2}\|x - y\|^2, \tag{5.2}$$

for all x, y. Then, $p(\cdot)$ satisfies

$$\|\nabla p(x) - \nabla p(y)\| \le \max\{L, \mu\}\|x - y\|. \tag{5.3}$$

In smooth approximation, it is shown in [29] that $g_i^{\beta_i}$ is a Lipschitz smooth function and it approximates the function value of g_i with some $\mathcal{O}(\beta_i)$-error:

$$g_i^{\beta_i}(x) \le g_i(x) \le g_i^{\beta_i}(x) + \tfrac{\beta_i D_{\mathcal{Y}_i}^2}{2}, \quad \forall x, \tag{5.4}$$

$$\|\nabla g_i^{\beta_i}(x) - \nabla g_i^{\beta_i}(z)\| \le L_{g_i}^{\beta_i}\|x - z\|, \quad \forall x, z, \quad L_{g_i}^{\beta_i} := \tfrac{\|A_i\|^2}{\beta_i}. \tag{5.5}$$

Similar properties of $f_i^{\beta_i}$ are developed in the following proposition.

Proposition 4 *We have the following properties about the approximation function $f_i^{\beta_i}$ ($\beta_i > 0$).*

1. *Let $\bar{\beta}_i \in [0, \beta_i]$, then we have*

$$f_i^{\beta_i}(x) \le f_i^{\bar{\beta}_i}(x) \le f_i^{\beta_i}(x) + \tfrac{(\beta_i - \bar{\beta}_i)D_{\mathcal{Y}_i}^2}{2}. \tag{5.6}$$

2. *$f_i^{\beta_i}(x)$ has upper curvature $L_{g_i}^{\beta_i}$ and negative lower curvature $-L_{h_i}$, namely,*

$$f_i^{\beta_i}(x) \le f_i^{\beta_i}(y) + \langle \nabla f_i^{\beta_i}(y), x - y \rangle + \tfrac{L_{g_i}^{\beta_i}}{2}\|x - y\|^2, \tag{5.7}$$

$$f_i^{\beta_i}(x) \ge f_i^{\beta_i}(y) + \langle \nabla f_i^{\beta_i}(y), x - y \rangle - \tfrac{L_{h_i}}{2}\|x - y\|^2. \tag{5.8}$$

3. *$f_i^{\beta_i}$ is Lipschitz smooth with modulus $L_i^{\beta_i} := \max\{L_{g_i}^{\beta_i}, L_{h_i}\}$. Namely, for any x, y, we have*

$$\|\nabla f_i^{\beta}(x) - \nabla f_i^{\beta}(y)\| \le L_i^{\beta_i}\|x - y\|. \tag{5.9}$$

Proof Part 1. If $\bar{\beta} < \beta$, then by definition we have $f_i^{\bar{\beta}_i}(x) \ge f_i^{\beta_i}(x)$. On the other hand, using the boundedness of $\mathcal{Y}_i$, we have

$$f_i^{\beta_i}(x) = \max_{y_i \in \mathcal{Y}_i} \langle A_i x, y_i \rangle - p(y_i) - \tfrac{\bar{\beta}_i}{2}\|y_i - \hat{y}_i\|^2 - \tfrac{\beta_i - \bar{\beta}_i}{2}\|y_i - \hat{y}_i\|^2 - h(x)$$

$$\geq \max_{y_i \in \mathcal{Y}_i} \langle A_i x, y_i \rangle - p(y_i) - \frac{\bar{\beta}_i}{2} \|y_i - \hat{y}_i\|^2 - h(x) - \frac{\beta_i - \bar{\beta}_i}{2} D^2_{\mathcal{Y}_i}$$

$$= f_i^{\bar{\beta}_i}(x) - \frac{\beta_i - \bar{\beta}_i}{2} D^2_{\mathcal{Y}_i}.$$

Combining the above two results gives the desired inequality.

Part 2. Since $g_i^{\beta_i}$ and h^i are both convex and smooth functions, we have

$$g_i^{\beta_i}(x_1) \leq g_i^{\beta_i}(x_2) + \langle \nabla g_i^{\beta_i}(x_2), x_1 - x_2 \rangle + \frac{L^{\beta_i} g_i}{2} \|x_1 - x_2\|^2,$$
$$h_i(x_1) \leq h_i(x_2) - \langle \nabla h_i(x_2), x_1 - x_2 \rangle.$$

Summing up the above two inequalities and noting the definition of $f_i^{\beta_i}, \nabla f_i^{\beta_i}$, we conclude that $f_i^{\beta_i}$ has an upper curvature of $L_{g_i}^{\beta_i}$. Similarly, using convexity of $g_i^{\beta_i}$ and smoothness of h_i, we obtain that $f_i^{\beta_i}$ has a negative lower curvature $-L_{h_i}$.

Part 3. The Lipschitz continuity (5.9) is an immediate consequence of part 2) and Lemma 7. □

Remark 6 When $\bar{\beta}_i = 0$, Relation (5.6) reads $f_i^{\beta_i}(x) \leq f_i(x) \leq f_i^{\beta_i}(x) + \frac{\beta_i D^2_{\mathcal{Y}_i}}{2}$. Together with Assumption 2, it can be seen that x^0 is also strictly feasible for problem (5.1). This justifies that LCPG is well-defined for problem (5.1).

Remark 7 The Lipschitz constant of $\nabla f_i^{\beta_i}$ can be derived in a different way. Since $\nabla g_i^{\beta_i}$ and ∇h_i are $L_{g_i}^{\beta_i}$ and L_{h_i} Lipschitz continuous, respectively, we can show by triangle inequality that $\nabla f_i^{\beta_i}(x)$ is $L_{g_i}^{\beta_i} + L_{h_i}$-Lipschitz continuous. In contrast, by exploiting the asymmetry between lower and upper curvature, Proposition 4 derived a slightly sharper bound on the gradient Lipschitz constant.

Throughout this section, we choose specific β_i to ensure $\beta_i D^2_{\mathcal{Y}_i}$ is constant for all $i \in [m]$. Hence, we can define the additive approximation factor above as

$$\nu := \frac{\beta_i D^2_{\mathcal{Y}_i}}{2}, \ i \in [m]. \tag{5.10}$$

Note that (5.4) provides an approximation error for function values, or the so-called zeroth-order oracle of function g_i. However, convergence results for nonconvex optimization are generally given in terms of first-order stationarity measure, implying that we need approximation for first-order oracle for the function f_i and consequently function g_i. Below we discuss a widely used approximate subdifferential for convex functions and generalize it for nonsmooth nonconvex functions.

Definition 5 [ν-subdifferential] We say that a vector $v \in \mathbb{R}^n$ is a ν-subgradient of the convex function $p(\cdot)$ at x if for any z, we have

$$p(z) \geq p(x) + \langle v, z - x \rangle - \nu.$$

The set of all ν-subgradients of p at x is called the ν-subdifferential, denoted by $\partial^\nu p(x)$. Moreover, we define ν subdifferential of nonconvex function f_i as $\partial^\nu f_i(x) := \partial^\nu g_i(x) + \{-\nabla h_i(x)\}$ where the addition of sets is in Minkowski sense.

Finally, we define a generalization of type-I KKT convergence criterion for structured nonsmooth nonconvex function constrained optimization problem:

Definition 6 We say that a point x is an (ϵ, ν) type-III KKT point of (1.1) if there exists $\lambda \geq 0$ satisfying the following conditions:

$$\|\partial^\nu \psi_0(x) + \sum_{i=1}^m \lambda_i \partial^\nu \psi_i(x)\|_-^2 \leq \epsilon, \tag{5.11}$$

$$\sum_{i=1}^m \lambda_i |\psi_i(x) - \eta_i| \leq \epsilon, \tag{5.12}$$

$$\|[\psi(x) - \eta]_+\|_1 \leq \epsilon. \tag{5.13}$$

Moreover, we say that x is a randomized (ϵ, ν) type-III KKT point of (1.1) if (5.11), (5.12) and (5.13) are satisfied in expectation.

The ϵ-subdifferential and the type-III KKT point are essential for associating smooth approximation with the original nonsmooth problem. We build some important properties in the following proposition.

Proposition 5 *Let β_i and ν satisfy (5.10).*

1. *For any $x \in \mathbb{R}^d$, we have $\nabla f_i^{\beta_i}(x) \in \partial^\nu f_i(x)$, $i = 0, 1, 2, ..., m$.*
2. *Suppose that x is a (randomized) Type-I ϵ-KKT point of problem (5.1) and λ is the associated dual variable with bound $\|\lambda\| \leq B$, then x is a (randomized) Type-III $(\bar{\epsilon}, \nu)$-KKT point of problem (1.1) for $\bar{\epsilon} = \max\{\epsilon + B\nu, m\nu\}$.*

Proof Part 1. It suffices to show $\nabla g_i^\beta(x) \in \partial^\nu g_i(x)$. Due to the convexity of $g_i^{\beta_i}$ and (5.4), we have

$$g_i(z) \geq g_i^{\beta_i}(z) \geq g_i^{\beta_i}(x) + \langle \nabla g_i^{\beta_i}(x), z - x \rangle \geq g_i(x) + \langle \nabla g_i^{\beta_i}(x), z - x \rangle - \frac{\beta_i D_{y_i}^2}{2},$$

where the first inequality follows from the first relation in (5.4), and the third inequality follows from second relation in (5.4). Noting the definition of ν-subgradient, we conclude the proof.

Part 2. It suffices to show the conversion of randomized Type-I KKT points to randomized Type-III KKT points. Suppose that x is a randomized Type-I ϵ-KKT point and we have $\|\lambda\|_1 \leq B$. Using Part 1 it is easy to see $\partial \mathcal{L}(x, \lambda) \subseteq \partial^\nu \psi_0(x) + \sum_{i=1}^m \partial^\nu \psi_i(x)$, therefore, we have

$$\mathbb{E}\|\partial^\nu \psi(x) + \sum_{i=1}^m \partial^\nu \psi_i^\nu(x)\|_-^2 \leq \epsilon.$$

Using Proposition 4, we have

$$\lambda_i(\psi_i^{\beta_i}(x) - \eta_i) \leq \sum_{i=1}^m \lambda_i(\psi_i(x) - \eta_i) \leq \lambda_i(\psi_i^{\beta_i}(x) - \eta_i) + \lambda_i \nu \leq \lambda_i \nu.$$

This implies

$$|\lambda_i(\psi_i(x) - \eta_i)| \leq \max\{|\lambda_i(\psi_i^{\beta_i}(x) - \eta_i)|, \lambda_i \nu\}.$$

Summing up this inequality over $i = 1, 2, 3, ..., m$ and taking expectation with all the randomness, we have the

$$\sum_{i=1}^{m} \mathbb{E}|\lambda_i(\psi_i(x) - \eta_i)| \leq \mathbb{E}\sum_{i=1}^{m}|\lambda_i(\psi_i^{\beta_i}(x) - \eta_i)| + \sum_{i=1}^{m}\lambda_i \nu \leq \epsilon + B\nu.$$

Moreover, we have

$$\sum_{i=1}^{m}[\psi_i(x) - \eta_i]_+ = \sum_{i=1}^{m}[\psi_i^{\beta_i}(x) - \eta_i + \psi_i(x) - \psi_i^{\beta_i}(x)]_+$$
$$\leq \sum_{i=1}^{m}[\psi_i(x) - \psi_i^{\beta_i}(x)] \leq m\nu.$$

$\square$

Now, we are ready to discuss the convergence rate of LCPG for nonsmooth nonconvex function constrained optimization.

Theorem 6 *Assume that β_i, ν satisfy (5.10) and set $\delta^k = \frac{\eta - \eta^0}{(k+1)(k+2)}$ when running LCPG to solve problem (5.1). Denote $c_i = \|A_i\|^2 D_{y_i}^2$ $(0 \leq i \leq m)$, $c = [c_1, c_2, ..., c_m]^T$ and let $\|\lambda^k\|_1 \leq B$. Suppose that $\nu = o(\frac{c_i}{L_{h_i}})$ for $i = 0, 1, 2, ..., m$, then $x^{\hat{k}+1}$ is a randomized Type-III $(\bar{\epsilon}, \nu)$-KKT point with*

$$\bar{\epsilon} = \mathcal{O}\left\{\frac{2}{K+2}\left[\left(\frac{8(c_0 + B\|c\|)^2}{c_0 \nu} + \frac{2B\|c\|}{c_0}\right)(\Delta + \nu) + B\|\eta - \eta^0\|\right] + B\nu + m\nu\right\}.$$

Proof Our analysis resembles the proof of Theorem 2. Using a similar argument in (3.33), we have

$$\sum_{k=0}^{K}\alpha_k\|x^k - x^{k+1}\|^2 \leq \frac{2\alpha_K}{L_0}\left[\psi_0^{\beta_0}(x^0) - \psi_0^{\beta_0}(x^{K+1})\right]$$
$$\leq \frac{2\alpha_K}{L_0^{\beta_0}}\left[\psi_0(x^0) - \psi_0(x^{K+1}) + \nu\right]. \qquad (5.14)$$

Combining this result with Lemma 3 we obtain

$$\sum_{k=0}^{K}\alpha_k\|\partial_x\mathcal{L}(x^{k+1}, \lambda^{k+1})\|_-^2 \leq 8\frac{(L_0^{\beta_0} + B\|L^\beta\|)^2}{L_0^{\beta_0}}\alpha_K(\Delta + \nu),$$

$$\sum_{k=0}^{K}\alpha_k\langle\lambda^{k+1}, |\psi(x^{k+1}) - \eta|\rangle \leq 2B\frac{\|L^\beta\|}{L_0^{\beta_0}}\alpha_K(\Delta + \nu) + B\sum_{k=0}^{K}\alpha_k\|\eta - \eta^k\|,$$

(5.15)

where $\Delta = \psi_0(x^0) - \psi_0(x^*), \alpha_k \geq 0$ and $\mathcal{L}^\beta(x, \lambda) := \psi_0^{\beta_0}(x) + \sum_{i=1}^{m}\lambda_i(\psi_i^{\beta_i}(x) - \eta_i)$. Taking $\delta^k = \frac{\eta - \eta^0}{(k+1)(k+2)}$ and $\alpha_k = k + 1$ in (5.15), we see that $x^{\hat{k}+1}$ is a Type-I ϵ-KKT point for

$$\epsilon = \frac{2}{K+2}\max\left\{\frac{8(L_0^{\beta_0} + B\|L^\beta\|)^2}{L_0^{\beta_0}}(\Delta + \nu), \frac{2B\|L^\beta\|}{L_0^{\beta_0}}(\Delta + \nu) + B\|\eta - \eta^0\|\right\}.$$

Noting that $L_{g_i}^{\beta_i} = \frac{\|A_i\|^2}{\beta_i} = \frac{\|A_i\|^2 D_{\mathcal{Y}_i}^2}{2\nu} = \frac{c_i}{2\nu}$ and $\nu = o(\frac{c_i}{L_{h_i}})$, we have $\frac{(L_0^{\beta_0} + B\|L^\beta\|)^2}{L_0^{\beta_0}} = \mathcal{O}\big(\frac{(c_0 + B\|c\|)^2}{c_0 \nu}\big)$ and $\frac{\|L^\beta\|}{L_0^{\beta_0}} = \frac{\|c\|}{c_0}$. Using the definition of $\hat{k}$ and Proposition 5 we obtain the desired result. $\qquad\square$

6 Inexact `LCPG`

`LCPG` requires the exact optimal solution of subproblem (3.1), which, however, poses a great challenge when the subproblem is difficult to solve. To alleviate such an issue, we consider an inexact variant of `LCPG` method for which the update of x^{k+1} only solves problem (3.1) approximately. This section is organized as follows. First, we present a general convergence property of inexact `LCPG` when the subproblem solutions satisfy certain approximation criterion. Next, we analyze the efficiency of inexact `LCPG` when the subproblems are handled by different external solvers. When the subproblem is a quadratically constrained quadratic program (QCQP), we propose an efficient interior point algorithm by exploiting the diagonal structure. When the subproblem has general proximal components, we propose to solve it by first-order methods. Particularly, we consider solving the subproblem by the constraint extrapolation (`ConEx`) method and develop the total iteration complexity of `ConEx`-based `LCPG`.

6.1 Convergence analysis under an inexactness criterion

Throughout the rest of this section, we will denote the exact primal-dual solution of (3.1) as $(\widetilde{x}^{k+1}, \widetilde{\lambda}^{k+1})$. We use the following criterion for measuring the accuracy of subproblem solutions.

Definition 7 We say that a point x is an ϵ-solution of (3.1) if

$$\psi_0^k(x) - \psi_0^k(\widetilde{x}^{k+1}) \le \epsilon, \quad \|[\psi^k(x)]_+\| \le \epsilon, \quad \mathcal{L}_k(x, \widetilde{\lambda}^{k+1}) \le \mathcal{L}_k(\widetilde{x}^{k+1}, \widetilde{\lambda}^{k+1}) + \epsilon.$$

The following theorem shows asymptotic convergence to stationarity for inexact `LCPG` method under mild assumptions. Since the proof is similar to the previous argument, we present the details in Appendix B for the sake of completeness. Note that the theorem applies to a general nonconvex problem and hence applies to convex problems as well.

Theorem 7 *Suppose that Assumption 3 holds and let x^{k+1} be an ϵ_k-solution of (3.1) satisfying $\epsilon_k < \min_{i \in [m]} \delta_i^k$. Then all the conclusions of Theorem 1 still hold. Then the dual sequence $\{\widetilde{\lambda}^k\}$ is uniformly bounded by a constant $B > 0$. Moreover, every limit point of inexact `LCPG` is a KKT point.*

Under the inexactness condition in Definition 7, we establish the complexity of inexact `LCPG` in the following theorem.

Theorem 8 *Under the assumptions of Theorem 7, we have*

$$\sum_{k=0}^{K}\alpha_k\|\partial_x\mathcal{L}(\tilde{x}^{k+1},\tilde{\lambda}^{k+1})\|_-^2 \leq \frac{8(L_0+B\|L\|)^2}{L_0}\tilde{\Delta}, \tag{6.1}$$

$$\sum_{k=0}^{K}\alpha_k\langle\tilde{\lambda}^{k+1},|\psi(\tilde{x}^{k+1})-\eta|\rangle \leq B\sum_{k=0}^{K}\alpha_k\|\eta-\eta^k\| + \frac{2B\|L\|}{L_0}\tilde{\Delta}, \tag{6.2}$$

$$\sum_{k=0}^{K}\alpha_k\|x^k-\tilde{x}^{k+1}\|^2 \leq \frac{2}{L_0}\tilde{\Delta}, \tag{6.3}$$

where $\tilde{\Delta} = \sum_{k=0}^{K}\alpha_k[\psi_0(x^k)-\psi_0(x^{k+1})+\varepsilon_k]$. *Moreover, if we choose the index* $\hat{k} \in \{0,1,\ldots,K\}$ *with probability* $\mathbb{P}(\hat{k}=k) = \alpha_k/(\sum_{i=0}^{K}\alpha_i)$, *then* $x^{\hat{k}}$ *is a randomized* (ϵ,δ) *type-II KKT point with*

$$\epsilon = 1/(\sum_{i=0}^{K}\alpha_i)\max\{\frac{8(L_0+B\|L\|)^2}{L_0}\tilde{\Delta}, B\sum_{k=0}^{K}\alpha_k\|\eta-\eta^k\| + \frac{2B\|L\|}{L_0}\tilde{\Delta}\},$$
$$\delta = 2\tilde{\Delta}/(L_0\sum_{i=0}^{K}\alpha_i). \tag{6.4}$$

In particular, using $\alpha_k = k+1$, $\epsilon_k = \min_{i\in[m]}\frac{\delta_i^k}{2}$ *and* $\delta_i^k = \frac{\eta_i-\eta_i^k}{(k+1)(k+2)}$, *we have* $x^{\hat{k}}$ *is* (ϵ,δ) *type-II KKT point of* (1.1) *where*

$$\epsilon = \frac{2}{K+2}\max\{4(L_0+B\|L\|)^2[2D^2+\frac{\|\eta-\eta^0\|}{L_0}], B\|\eta-\eta^0\|+B\|L\|[2D^2+\frac{\|\eta-\eta^0\|}{L_0}]\},$$
$$\delta = \frac{2}{K+2}(2D^2+\frac{\|\eta-\eta^0\|}{L_0}). \tag{6.5}$$

Proof Using (3.8) with x^{k+1} replaced by $\tilde{x}^{k+1}$ (the optimal solution of problem (3.1)) and adding $f(x^k) + \frac{L_0}{2}\|x^k-\tilde{x}^{k+1}\|^2$ on both sides, we have

$$\begin{aligned}
\frac{L_0}{2}\|x^k-\tilde{x}^{k+1}\|^2 &\leq \psi_0^k(x^k)-\psi_0^k(\tilde{x}^{k+1})\\
&\leq \psi_0(x_k)-\psi_0^k(x^{k+1})+\epsilon_k\\
&\leq \psi_0(x_k)-\psi_0(x^{k+1})+\epsilon_k, \tag{6.6}
\end{aligned}$$

where the second inequality follows from $\psi_0^k(x^k) = \psi_0(x^k)$ as well as x^{k+1} being an ϵ_k-solution (see Definition 7) of subproblem (3.1), and the third inequality follows from the fact that $\psi_0^k(x) \geq \psi_0(x)$ for all $x \in \text{dom}\,\chi_0$.

Using Lemma 3 (again x^{k+1} is replaced by $\tilde{x}^{k+1}$), noting that ϵ_k satisfies the requirements of Theorem 7 implying that $\|\tilde{\lambda}^k\| \leq B$ and using (6.6), we have

$$\begin{aligned}
\|\partial_x\mathcal{L}(\tilde{x}^{k+1},\tilde{\lambda}^{k+1})\|_-^2 &\leq 4(L_0+B\|L\|)^2\|\tilde{x}^{k+1}-x^k\|^2\\
&\leq \frac{8(L_0+B\|L\|)^2}{L_0}[\psi_0(x^k)-\psi_0(x^{k+1})+\varepsilon_k]. \tag{6.7}
\end{aligned}$$

Similar to the argument of (3.34), we have

$$\sum_{i=1}^{m}\tilde{\lambda}_i^{k+1}|\psi_i(\tilde{x}^{k+1})-\eta_i| \leq B\|\eta-\eta^k\| + B\|L\|\|\tilde{x}^{k+1}-x^k\|^2$$

$$\leq B\|\eta - \eta^k\| + \tfrac{2B\|L\|}{L_0}\big[\psi_0(x^k) - \psi_0(x^{k+1}) + \varepsilon_k\big].$$
(6.8)

Multiplying (6.6), (6.7) and (6.8) by α_k and summing over $k = 0, 1, \ldots, K$ give (6.1), (6.2) and (6.3).

We derive a convergence rate based on the specified parameters. First, from relation (6.6), we note that $\psi_0(x^{k+1}) \leq \psi_0(x^k) + \epsilon_k$. Hence, we have by induction that

$$\psi_0(x^{k+1}) \leq \psi_0(x^0) + \sum_{i=0}^{k}\epsilon_i.$$
(6.9)

By setting $\alpha_k = (k+1)$ and $\epsilon_k = \min_{i \in [m]} \tfrac{\delta_i^k}{2} = \min_{i \in [m]} \tfrac{\eta_i - \eta_i^k}{2(k+1)(k+2)}$ for all $k \geq 0$ (note that ϵ_k satisfies the requirement of Theorem 7), we have

$$
\begin{aligned}
\widetilde{\Delta} &= \sum_{k=0}^{K}\alpha_k[\psi_0(x^k) - \psi_0(x^{k+1})] + \sum_{k=0}^{K}\alpha_k\epsilon_k \\
&= \alpha_0\psi_0(x^0) + \sum_{k=0}^{K-1}(\alpha_{k+1} - \alpha_k)\psi_0(x^{k+1}) - \alpha_K\psi_0(x^{K+1}) + \sum_{k=0}^{K}\alpha_k\epsilon_k \\
&\overset{(i)}{\leq} \alpha_0\psi_0(x^0) + \sum_{k=0}^{K-1}(\alpha_{k+1} - \alpha_k)[\psi_0(x^0) + \sum_{i=0}^{k}\epsilon_i] - \alpha_K\psi_0(x^{K+1}) + \sum_{k=0}^{K}\alpha_k\epsilon_k \\
&\overset{(ii)}{=} \alpha_K[\psi_0(x^0) - \psi_0(x^{K+1})] + \sum_{k=0}^{K-1}\sum_{i=0}^{k}\epsilon_i + \sum_{k=0}^{K}\alpha_k\epsilon_k \\
&= \alpha_K[\psi_0(x^0) - \psi_0(x^{K+1})] + \sum_{i=0}^{K-1}\sum_{k=i}^{K-1}\epsilon_i + \sum_{k=0}^{K}\alpha_k\epsilon_k \\
&\overset{(iii)}{=} \alpha_K[\psi_0(x^0) - \psi_0(x^{K+1})] + \sum_{i=0}^{K}(K - i)\epsilon_i + \sum_{k=0}^{K}\alpha_k\epsilon_k \\
&\overset{(iv)}{\leq} \alpha_K[\psi_0(x^0) - \psi_0(x^{K+1})] + \tfrac{\|\eta - \eta^0\|}{2}\sum_{k=0}^{K}\tfrac{K-k}{(k+1)(k+2)} + \tfrac{1}{k+2} \\
&= \alpha_K[\psi_0(x^0) - \psi_0(x^{K+1})] + \tfrac{\|\eta - \eta^0\|}{2}\sum_{k=0}^{K}\tfrac{K+1}{(k+1)(k+2)} \\
&\leq \alpha_K\big[\psi_0(x^0) - \psi_0(x^{K+1}) + \tfrac{\|\eta - \eta^0\|}{2}\big].
\end{aligned}
$$
(6.10)

Here, (i), (ii) follows from (6.9) and $\alpha_{k+1} - \alpha_k = 1(> 0)$, (iii) follows $(K - i)$ is 0 at $i = K$, (iv) follows by observing $\epsilon_k \leq \tfrac{\|\eta - \eta^0\|}{2(k+1)(k+2)}$ and last inequality follows since $\sum_{k=0}^{\infty}\tfrac{1}{(k+1)(k+2)} = 1$ and $\alpha_K = K + 1$.

Applying same arguments as those in Corollary 1, we have $\sum_{k=0}^{K}\alpha_k\|\eta - \eta^k\| = \alpha_K\|\eta - \eta^0\|$. Using this relation along with $\sum_{k=0}^{K}\alpha_k = \tfrac{(K+1)(K+2)}{2}$ and (6.10) inside (6.4), we have (6.5). Hence, we conclude the proof. $\square$

Remark 8 Compared to the convergence result (3.31) for exact LCPG, we have to control the accumulated error in $\tilde{\Delta}$ for the inexact case (6.4). However, we need an even more stringent condition on the error to ensure asymptotic convergence. Specifically, we assume ε_k to be smaller than the level increments δ_i^k to ensure that each subsequent subproblem is strictly feasible. As long as the subproblems are solved deterministically with sufficient accuracy, we can ensure such feasibility as well as the boundedness of the dual.

Remark 9 Note that the convergence analysis of the inexact method for the stochastic case will go through in a similar fashion. In particular, the subproblems of LCSPG are

still deterministic in nature. Hence, a deterministic error can be easily incorporated into the analysis of the stochastic outer loop. In particular, Proposition 3 will have an additional ϵ_k in the RHS. We can use $\epsilon_k = \min_{i \in [m]} \frac{\delta_i^k}{2}$ to ensure the strict feasibility. Following the analysis in Theorem 4, we will get the additional term $\sum_{k=0}^{K} \alpha_k \epsilon_k$. Note that we have identical policies for α_k in the above analysis and Corollary 2. Furthermore, since δ_k used above and in Corollary 2 are the same, we have identical values of ϵ_k as well. Following the above development, we can easily bound the additional $\sum_{k=0}^{K} \alpha_k \epsilon_k$ term.

6.2 Solving the subproblem with the interior point method

Our goal is to develop an efficient interior point algorithm to solve problem (3.1) when $\chi_i(x) = 0$, $i \in [m]$. Without loss of generality, we express the subproblem as the following QCQP:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & g_0(x) := \tfrac{L_0}{2} \|x - a_0\|^2 \\ \text{s.t.} \quad & g_i(x) := \tfrac{L_0}{2} \|x - a_i\|^2 - b_i \leq 0, \quad i \in [m]. \end{aligned} \tag{6.11}$$

We assume that the initial solution $\hat{x}$ of such problem is strictly feasible, namely, there exists $\delta > 0$ such that

$$g_i(\hat{x}) \leq -\delta, \quad i = 1, 2, \ldots m. \tag{6.12}$$

Let $e_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{d+1}$. With a slight abuse of notation, we can formulate (6.11) as the following problem

$$\begin{aligned} \min \quad & e_1^T u \\ \text{s.t.} \quad & \tilde{g}_0(u) = g_0(x) - \eta \leq 0, \\ & \tilde{g}_i(u) = g_i(x) \leq 0, \ i \in [m], \\ & \tilde{g}_{m+1}(u) = \tfrac{L_{m+1}}{2} \|u - (0, a_{m+1})^T\|^2 - b_{m+1} \leq 0, \\ & u = (\eta, x) \in \mathbb{R} \times \mathbb{R}^d. \end{aligned} \tag{6.13}$$

Here we set artificial variables $L_{m+1} = 1$, $a_{m+1} = 0$ and $b_{m+1} = \frac{1}{2} R^2$ for some sufficiently large R. We explicitly add such a ball constraint to ensure bound on (η, x). Note that the bound R always exists since our domain is compact and the objective is Lipschitz continuous. Our goal is to apply the path-following method to solve (6.13). We denote

$$\phi(u) = -\sum_{i=0}^{m+1} \log -\tilde{g}_i(u) \tag{6.14}$$

Since each $\tilde{g}_i(u)$ is convex quadratic in u, $\phi(u)$ is a self-concordant barrier with $\upsilon = m + 2$. The key idea of the path-following algorithm is to approximately solve a sequence of penalized problems

$$\min_u \phi_\tau(u) := \tau \eta + \phi(u) \tag{6.15}$$

with increased values of τ, and generate a sequence of strictly feasible solution u_τ close to the central path-a trajectory composed of the minimizers $u_\tau^* = \text{argmin}_u \phi_\tau(u)$.

We apply a standard path-following algorithm (See [28, Chapter 4]) for solving (6.13), and outline the overall procedure in Algorithm 4. This algorithm consists of two main steps:

1. Initialization: We seek a solution u^0 near the analytic center (i.e. minimizer of $\phi(u)$). To this end, we solve a sequence of auxiliary problems $\hat{\phi}_\tau(u) = \tau w^T u + \phi(u)$ where $w = -\nabla\phi(\hat{u})$. It can be readily seen that $\hat{u}$ is in the central path of this auxiliary problem with $\tau = 1$. Performing a reverse path-following scheme (decreasing rather than increasing τ), we gradually converge to the analytic center.
2. Path-following: We solve a sequence of penalized problems with an increasing value of τ by a damped version of Newton's method, which ensures the solutions in the proximity of the central path.

Algorithm 4 Path-following Interior Point Method ([28])

$\triangleright$ Newton decrement $n(f, u) := \sqrt{\nabla f(u)^T [\nabla^2 f(u)]^{-1} \nabla f(u)}$
1: **Input:** $\hat{u}, \kappa \in (0, 1), \gamma > 0, \varepsilon$; Set $\tau_0 = 1, u^0 = \hat{u}$;
$\triangleright$ Phase Zero: Approximate analytic center
2: **for** $i = 0, 1, \ldots$ **do**
3:     $\tau_{i+1} = \left(1 + \frac{\gamma}{\sqrt{v}}\right)^{-1} \tau_i$;
4:     Obtain u^{i+1} from calling $\text{Newton}(\phi_{\tau_{i+1}}, u^i, \tau_{i+1}, \kappa/2)$;
5:     **if** $n(\phi, u^{i+1}) \le \frac{3}{4}\kappa$ **then**
6:         Set $u^* = u^{i+1}$ and **Break;**
7:     **end if**
8: **end for**
$\triangleright$ Phase One: Path-following scheme
9: Set $u^0 = u^*, \tau_0 = \max\{\tau : n(\phi_\tau, u^0) \le \kappa\}, s = \left\lceil \frac{\sqrt{v}}{\gamma} \ln \frac{2v}{\tau_0 \varepsilon} \right\rceil - 1$;
10: **for** $i = 0, 1, \ldots, s$ **do**
11:     $\tau_{i+1} = \left(1 + \frac{\gamma}{\sqrt{v}}\right)\tau_i$;
12:     Obtain u^{i+1} **from calling** $\text{Newton}(\phi_{\tau_{i+1}}, u^i, \tau_{i+1}, \kappa)$;
13: **end for**
14: **Output:** u^{s+1}.
$\triangleright$ Damped Newton method for solving the subproblem.
15: **function** $\text{Newton}(f, v^0, \tau, \epsilon)$
16:     **for** $s = 0, 1, 2, \ldots$ **do**
17:         **if** $n(f, v^s) \le \epsilon$ **then Break;**
18:         **end if**
19:         $v^{s+1} = v^s - \frac{1}{1+n(f,v^s)}[\nabla^2 f(v^s)]^{-1}\nabla f(v^s)$;
20:     **end for**
21: **end function**

6.2.1 Solving the Newton equation

First, we calculate the gradient and Hessian map of $\phi_t(\cdot)$:

$$\nabla\phi_\tau(u) = \tau e_1 + \sum_{i=0}^{m+1} \theta_i \nabla\tilde{g}_i(u),$$

$$\nabla^2 \phi_\tau(u) = \sum_{i=0}^{m+1} \theta_i^2 \nabla \tilde{g}_i(u) \tilde{g}_i(u)^{\mathrm{T}} + \sum_{i=0}^{m+1} \theta_i \nabla^2 \tilde{g}_i(u) = NN^{\mathrm{T}} + \Gamma,$$

where $\theta_i = -\tilde{g}_i(u)^{-1}$, and

$$N = \left[\theta_0 \nabla \tilde{g}_0(u), \ldots, \theta_{m+1} \nabla \tilde{g}_{m+1}(u) \right] \in \mathbb{R}^{(d+1) \times (m+2)},$$

$$\Gamma = \begin{bmatrix} \theta_{m+1} L_{m+1} & 0 \\ 0 & \sum_{i=0}^{m+1} \theta_i L_i I_d \end{bmatrix} \in \mathbb{R}^{(d+1) \times (d+1)}.$$

Note that computing the gradient $\nabla \phi_t(u)$ takes $\mathcal{O}(dm)$, hence the computation burden is from forming and solving the Newton systems. This is divided into two cases.

1. $m < d$. Then the Hessian is the sum of a low rank matrix and a diagonal matrix. Based on the Sherman-Morrison-Woodbury formula, we have

$$[\nabla^2 \phi_\tau(u)]^{-1} = \Gamma^{-1} - \Gamma^{-1} N (I + N^{\mathrm{T}} \Gamma^{-1} N)^{-1} N^{\mathrm{T}} \Gamma^{-1}. \qquad (6.16)$$

Computing the product $N^{\mathrm{T}} \Gamma^{-1} N$ takes $\mathcal{O}(m^2 d)$ while performing Cholesky factorization takes $\mathcal{O}(m^3)$. Therefore, the overall complexity of each Newton step is $\mathcal{O}(m^3 + m^2 d) = \mathcal{O}(m^2 d)$.

2. $m \geq d$. In such case, we can directly compute NN^{T} in $\mathcal{O}(md^2)$ and then perform Cholesky factorization $\nabla^2 \phi_\tau(x) = LL^{\mathrm{T}}$ in $\mathcal{O}(d^3)$, followed by two triangle systems. Hence the overall complexity of a Newton step is $\mathcal{O}(d^3 + md^2) = \mathcal{O}(md^2)$.

Due to the above discussion, the cost of computing each Newton system is

$$\mathcal{O}\left(\min\{d, m\} \cdot md \right). \qquad (6.17)$$

6.2.2 Complexity

Before deriving the complexity of solving the subproblems, we require some additional assumptions. We assume that $M = \max_x \left\{ \|\nabla g_i(x)\| \right\}$ and $\max g_0(x) - \min_x g_0(x) \leq V$. Note that these assumptions are easily satisfied if we assume functions in the original problem have bounded level sets.

According to [28, Theorem 4.5.1], the complexity of interior point methods depends not only on the time to follow the central path, but also on the time to arrive near the analytic center from an arbitrary initial point. Let us put it in the context of Algorithm 1. Despite the strict feasibility guarantee, we do not know whether x^k is near the analytic center of each subproblem. It remains to show how to control the complexity of approximating the analytic center.

To measure the strict feasibility of the initial point, we use the Minkowsky function of the domain, which is defined by $\pi_x(y) = \inf\{t > 0 : x + t^{-1}(y - x) \in D\}$ for any given x in the interior of the domain. With the help of the Minkowsky function, we bound the distance between the initial point and the boundary in the following proposition.

Proposition 6 *Let $\hat{u} = (g_0(\hat{\eta}), \hat{x})$ where $\hat{\eta} = g_0(\hat{x}) + \delta$. If $\|u - \hat{u}\| \leq \frac{\delta}{M+1}$, then u is feasible for problem (6.13). Moreover, we have*

$$\pi_{u^*}(\hat{u}) \leq \frac{(M+1)R}{(M+1)R+\delta}, \tag{6.18}$$

where u^* is defined in phase zero of Algorithm 4.

Proof We have

$$|\tilde{g}_0(u) - \tilde{g}_0(\hat{u})| \leq |g_0(x) - g_0(\hat{x})| + |\eta - \hat{\eta}| \leq \frac{M}{M+1}\delta + \frac{\delta}{M+1} = \delta,$$

Analogously, for $i = 0, 1, \ldots, m$, we have

$$|g_i(x) - g_i(\hat{x})| \leq M\|x - \hat{x}\| \leq \frac{M}{M+1}\delta \leq \delta.$$

Using triangle inequality, we have $\tilde{g}_i(u) = g_i(x) \leq g_i(\hat{x}) + \delta = 0$. The last constraint in (6.13) is trivially satisfied for sufficiently large R. Therefore, u is a feasible point of (6.13).

Let $t^+ = \frac{(M+1)\|\hat{u}-u^*\|}{(M+1)\|\hat{u}-u^*\|+\delta}$, then from the above analysis, we know that the point

$$u^+ = u^* + \frac{1}{t^+}(\hat{u} - u^*) = \hat{u} + \frac{\delta(\hat{u}-u^*)}{(M+1)\|\hat{u}-u^*\|}$$

must be a feasible solution. Using the last constraint $\|u\| \leq R$, we immediately obtain the bound (6.18). $\qquad\square$

Using [28, Theorem 4.5.1] and Proposition 6, we can derive the total complexity of solving the diagonal QCQP.

Theorem 9 *Under the assumptions of Proposition 6, the total number of Newton steps to get an ε solution is*

$$N_\varepsilon = \mathcal{O}(1)\sqrt{m + 2}\ln\left(\frac{(m+2)V((M+1)R+\delta)}{\delta\varepsilon} + 1\right).$$

Corollary 3 *In the inexact* LCPG *method, assume that the subproblems are solved by Algorithm 4 and the returned solution satisfies the inexactness requirement in Theorem 8. Then, to get an $\mathcal{O}(\epsilon, \epsilon)$ Type-II KKT point, the overall arithmetic cost of Algorithm 4 is*
$$\mathcal{T} = \mathcal{O}\left(\min\{m, d\} \cdot m^{1.5}d \cdot \frac{1}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)\right).$$

Proof According to Theorem 8, the total number of LCPG is $K = \mathcal{O}(1/\varepsilon)$. In the kth iteration of LCPG, we set the error criteria $\nu = \mathcal{O}(\frac{1}{k^2})$ and $\varepsilon = \mathcal{O}(\frac{1}{k^2})$. Theorem 9 implies that the number of Newton steps is $N_k = \mathcal{O}(\sqrt{m}\ln(k))$. Therefore, the total number of Newton steps in LCPG is $T_K = \sum_{k=0}^{K}N_k = \mathcal{O}\left(\sqrt{m}\frac{1}{\varepsilon}\ln\left(\frac{1}{\varepsilon}\right)\right)$. Combining this result with (6.17) gives us the desired bound. $\qquad\square$

Remark 10 First, at the kth step of LCPG, we need $\log(k)$ iterations of interior point methods, of which the complexity order is equally contributed by the two phases of IPM. Specifically, we first require $\mathcal{O}(\ln(k))$ Newton steps to pull the iterates from near the boundary to the proximity of the central path, and then require $\mathcal{O}(\ln(k))$ to obtain an $\mathcal{O}(1/k^2)$-accurate solution. Second, it is interesting to consider the case when the constraint is far less than the feature dimensionality, namely, $m \ll d$. We observe that the total computation

$$\mathcal{O}\left(dm^{2.5}K \ln K\right)$$

is linear in dimensionality. Third, despite the simplicity, the basic barrier method offers a relatively stronger approximate solution than what is needed in Theorem 8, the feasibility of the solution path allows us to weaken the assumption to $\hat{\varepsilon}_k = 0$. Nevertheless, besides our approach, it is possible to employ long-step and infeasible primal-dual interior point methods which may give a better empirical performance.

6.3 Solving subproblems with the first-order method

In this section, we use a previously proposed ConEx method [4] to solve the subproblem (3.1) when general proximal functions χ_i are present. Then, we analyze the overall complexity of LCPG method with ConEx method as a subproblem solver. First, we formally state the extended version of problem (6.11) as follows:

$$\begin{aligned}
\min_{x \in X} \quad & \phi_0(x) := g_0(x) + \chi_0(x) \\
\text{s.t.} \quad & \phi_i(x) := g_i(x) + \chi_i(x) \le 0, \quad i = 1, \ldots, m.
\end{aligned} \tag{6.19}$$

For the application of ConEx for the subproblem, we need access to a convex compact set X such that $\cap_i \operatorname{dom} \chi_i \subseteq X$. Moreover, X is a "simple" set in the sense that it allows easy computation of the proximal operator of $\chi_0(x) + \sum_{i=1} w_i \chi_i(x)$ for any given weights $w_i, i = 1, \ldots, m$. Such assumptions are not very restrictive as many machine learning and engineering problems explicitly seek the optimal solution from a bounded set. Under these assumptions, we apply ConEx to solve the subproblem (3.1) of LCPG. We now reproduce a simplified optimality guarantee of the ConEx method below without necessarily going into the details of the algorithm.

Theorem 10 *[4] Let x be the output of ConEx after T iterations for problem (6.19). Assume that ϕ_0 is a strongly convex function and $(\widetilde{x}, \widetilde{\lambda})$ is the optimal primal-dual solution. Moreover, Let B be a parameter of the ConEx method which satisfies $B > \|\widetilde{\lambda}\|$. Then, the solution x satisfies*

$$\phi_0(x) - \phi_0(\widetilde{x}) \le O\left(\tfrac{1}{T^2}(B^2 + \|\widetilde{\lambda}\|^2)\right),$$
$$\|[\phi(x)]_+\| \le O\left(\tfrac{1}{T^2}(B^2 + \|\widetilde{\lambda}\|^2)\right).$$

Even though `ConEx` can be applied to a wider variety of convex function constrained problems, it has two vital and intricate issues that need to be addressed in our context:

1. The solution path of `ConEx` can be arbitrarily infeasible in the early iterations, while the successive iterations make the solutions infeasibility smaller. Note that the approximation criterion in Definition 7 requires guarantees on the amount of infeasibility. This implies `ConEx` has to run a significant number of iterations before getting sufficiently close to the feasible set.
2. Since `ConEx` is a primal-dual method, its convergence guarantees depend on the optimal dual solution λ^*. Moreover, a bound on the dual, $B(> \|\lambda^*\|)$, is required to implement the algorithm to achieve an accelerated convergence rate of $O(1/T^2)$ for strongly convex problems.

From Theorem 10, it is clear that `ConEx` requires a bound B. This requirement naturally leads to two cases: (1) bound B can be estimated apriori, e.g., see Lemma 2; and (2) bound B is known to exist but cannot be estimated, e.g., see Theorem 1. Both cases have different convergence rates for the subproblem which leads to different overall computational complexity.

Case 1: B can be estimated apriori. In this case, we do not need to estimate B^k as in (6.20). Using the bound B, we can get accelerated convergence of `ConEx` in accordance with Theorem 10 which leads to better performance of the `LCPG` method. The corollary below formally states the total computational complexity of `LCPG` method for this case.

Corollary 4 *If an explicit value of B is known, the `LCPG` method with `ConEx` as subproblem solver obtains $O(\frac{1}{K}, \frac{1}{K})$ type-II KKT point in $O(K^2)$ computations.*

Proof According to Theorem 10, the required `ConEx` iterations for each subproblem can be bounded by

$$T^k = O(\frac{B}{\sqrt{\epsilon_k}}).$$

Since B is a constant, we have $T^k = O(\epsilon_k^{-1/2}) = O(k)$. Finally, we have total computations $\sum_{k=1}^{K} T^k = O(K^2)$. Hence, we conclude the proof. □

Case 2: B is known to exist but cannot be estimated. For the subproblem (3.1), we can easily find $B^k > \|\widetilde{\lambda}^{k+1}\|$ by using the difference in levels of successive iterations. This bound is weak, especially in the limiting case as it does not take into account

Proposition 7 *For subproblem (3.1), we have*

$$\|\widetilde{\lambda}^{k+1}\| \leq \frac{\psi_0(x^k) - \psi_0^*}{\min_{i \in [m]} \delta_i^k}. \tag{6.20}$$

Proof By Slater's condition, we know that $\widetilde{\lambda}^{k+1}$ exists. Then, due to saddle point property of $(\widetilde{x}^{k+1}, \widetilde{\lambda}^{k+1})$, we have for all $x \in X$

$$\psi_0^k(x) + \widetilde{\lambda}^{k+1}[\psi^k(x) - \eta^{k+1}] \geq \psi_0^k(\widetilde{x}^{k+1}) + \widetilde{\lambda}^{k+1}[\psi^k(\widetilde{x}^{k+1}) - \eta^{k+1}]$$

$$= \psi_0^k(\widetilde{x}^{k+1}),$$

where the equality follows by complementary slackness. Using $x = x^k$ in the above relation and noting that x^k satisfies $\psi(x^k) \leq \psi^{k-1}(x^k) \leq \eta^k$, we have $\eta^{k+1} - \psi^k(x^k) = \eta^{k+1} - \psi(x^k) \geq \eta^{k+1} - \eta^k = \delta^k$ implying that

$$\psi_0^k(x^k) - \psi_0^k(\widetilde{x}^{k+1}) \geq \langle \widetilde{\lambda}^{k+1}, \delta^k \rangle \geq \|\lambda^{k+1}\|_1 \cdot \min_{i \in [m]} \delta_i^k \geq \|\lambda^{k+1}\| \cdot \min_{i \in [m]} \delta_i^k,$$

where second inequality follows from $\widetilde{\lambda}^{k+1} \geq 0$ and $\delta^k > 0$ and last inequality follows due to the fact that $\|\widetilde{\lambda}^{k+1}\|_1 \geq \|\widetilde{\lambda}^{k+1}\|$. We can further upper bound the LHS of the above relation as follows

$$\psi_0^k(x^k) - \psi_0^k(\widetilde{x}^{k+1}) = \psi_0(x^k) - \psi_0^k(\widetilde{x}^{k+1}) \leq \psi_0(x^k) - \psi_0(\widetilde{x}^{k+1}) \leq \psi_0(x^k) - \psi_0^*,$$

where the last inequality follows since $\widetilde{x}^{k+1}$ is feasible for the original problem (1.1). Combining the above two relations, we obtain (6.20). Hence, we conclude the proof.
□

We now state the final computation complexity of LCPG with ConEx which uses the bound in (6.20).

Corollary 5 *If an explicit value of B is not known, the LCPG method with ConEx as subproblem solver obtains $O(\frac{1}{K}, \frac{1}{K})$ type-II KKT point in $O(K^4)$ computations.*

Proof Using Proposition 7, we can set $B^k := \frac{\psi_0(x^k) - \psi_0^*}{\delta_{i_*}^k}$ where $i_* := \mathrm{argmin}_{i \in m} \eta_i - \eta_i^0$. Then, required ConEx iterations T^k can be bounded by

$$T^k = O\left(\frac{B^k}{\sqrt{\epsilon_k}}\right).$$

Finally, in view of (6.9) and the fact that $\sum_{i=0}^{\infty} \epsilon_i \leq \|\eta - \eta^0\|$ implies that $B^k \leq \frac{1}{\delta_{i_*}^k}[\psi_0(x^0) - \psi_0^* + \|\eta - \eta^0\|]$ for all k. Moreover, for all $k \leq K$, we have $\epsilon_k = \frac{\delta_{i_*}^k}{2}$. Hence, we get $T^k = O(\epsilon_k^{-3/2}) = O(k^3)$. Finally, we have $\sum_{k=1}^K T^k = O(K^4)$ which is the overall computational complexity of LCPG method with ConEx as subproblem solver to obtain $(O(\frac{1}{K}), O(\frac{1}{K}))$ type-II KKT point.
□

Remark 11 [Gradient complexity vs. computational complexity] Note that evaluating the gradient of $\psi_i^k(x)$ is relatively simple since it does not involve any new computation of $\nabla f_i(x)$. In that sense, the entire inner loop requires only one ∇f_i computation; hence the total gradient complexity of ∇f_i equals the total outer loops of inexact LCPG. On the other hand, inner loop computation does contribute to the problem's computational complexity. However, such iterations are expected to be very cheap given the ease of obtaining gradients for the QP subproblem (3.1) with identity hessian matrices.

7 LCPG for convex optimization

In this section, we establish the complexity of LCPG (i.e., Algorithm 1) when the objective f_0 and constraint $f_i, i \in [m]$ are convex. In particular, we consider two convex problems, depending on whether f_0 is convex or strongly convex. To provide a combined analysis of the two cases, we assume the following:

Assumption 6 $f_0(x)$ is μ_0-convex function for some $\mu_0 \geq 0$. Namely,

$$f_0(x) \geq f_0(y) + \langle \nabla f_0(y), x - y \rangle + \tfrac{\mu_0}{2}\|x - y\|^2, \quad \text{for any } x, y \in \mathbb{R}^d.$$

Note that if $\mu_0 = 0$ then f_0 is simply a convex function. Now we provide the convergence rate of LCPG to optimality.

For more generality, we consider an inexact variant of LCPG for which an approximate solution in terms of Definition 7 is returned in each iteration. Let $(\widetilde{x}^{k+1}, \widetilde{\lambda}^{k+1})$ be the saddle-point solution $\mathcal{L}_k(x, \widetilde{\lambda})$, i.e., $\widetilde{x}^{k+1}$ is an exact solution of the subproblem (3.1). First, we extend the three-point inequality in Lemma 1 for an inexact solution.

Lemma 8 *Let z^+ be an ϵ-approximate solution of problem $\min_{x \in \mathbb{R}^d}\{g(x) + \tfrac{\gamma}{2}\|x - z\|^2\}$ where $g(x)$ is a proper, lsc. and convex function. Then,*

$$g(z^+) - g(x) \leq \tfrac{\gamma}{2}\big[\|z - x\|^2 - \|z^+ - x\|^2 - \|z^+ - z\|^2\big] + \epsilon + \sqrt{2\gamma\epsilon}\,\|z^+ - x\|. \quad (7.1)$$

Proof First, let x^+ be the optimal solution of $\min_{x \in \mathbb{R}^d}\{g(x) + \tfrac{\gamma}{2}\|x - z\|^2\}$. In view of Lemma 1, for any x, we have

$$g(x^+) + \tfrac{\gamma}{2}\|x^+ - z\|^2 + \tfrac{\gamma}{2}\|x - x^+\|^2 \leq g(x) + \tfrac{\gamma}{2}\|x - z\|^2. \quad (7.2)$$

Placing $x = z^+$ above, we have

$$g(x^+) + \tfrac{\gamma}{2}\|x^+ - z\|^2 + \tfrac{\gamma}{2}\|z^+ - x^+\|^2 \leq g(z^+) + \tfrac{\gamma}{2}\|z^+ - z\|^2. \quad (7.3)$$

On the other hand, by the definition of ϵ-solution, we have

$$g(z^+) + \tfrac{\gamma}{2}\|z^+ - z\|^2 \leq g(x^+) + \tfrac{\gamma}{2}\|x^+ - z\|^2 + \epsilon. \quad (7.4)$$

Combining the above two inequalities gives

$$\tfrac{\gamma}{2}\|z^+ - x^+\|^2 \leq \epsilon. \quad (7.5)$$

Summing up (7.2) and (7.4) again and then rearranging the terms, we get

$$
\begin{aligned}
g(z^+) - g(x) &\leq \tfrac{\gamma}{2}\|x - z\|^2 - \tfrac{\gamma}{2}\|z^+ - z\|^2 - \tfrac{\gamma}{2}\|x - x^+\|^2 + \epsilon \\
&\leq \tfrac{\gamma}{2}\|x - z\|^2 - \tfrac{\gamma}{2}\|z^+ - z\|^2 - \tfrac{\gamma}{2}\|x - z^+\|^2 + \gamma\|x - z^+\|\|z^+ - x^+\| + \epsilon,
\end{aligned}
$$

where the last inequality uses the fact that $-\frac{1}{2}\|a+b\|^2 \le -\frac{1}{2}\|a\|^2 - \langle a, b \rangle \le -\frac{1}{2}\|a\|^2 + \|a\|\|b\|$ with $a = x - z^+$ and $b = z^+ - x^+$. Finally, combining the above two results gives the desired inequality (7.1). $\qquad\square$

Using the above lemma, we provide the main convergence property of LCPG for convex optimization.

Lemma 9 *Let x be feasible solution. Then, we have*

$$
\psi_0(x^{k+1}) - \psi_0(x) \le \langle \widetilde{\lambda}^{k+1}, \psi(x) - \eta^k \rangle + \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0}{2}\|x^k - x\|^2
$$
$$
- \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{2}\|x^{k+1} - x\|^2 + 2\epsilon_k + \sqrt{2(L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle)\epsilon_k}\|x^{k+1} - x\|.
$$
(7.6)

Proof Note that

$$
\begin{aligned}
\psi_0(x^{k+1}) &\overset{(i)}{\le} \psi_0^k(x^{k+1}) \overset{(ii)}{\le} \psi_0^k(\widetilde{x}^{k+1}) + \epsilon_k \\
&\overset{(iii)}{=} \psi_0^k(\widetilde{x}^{k+1}) + \langle \widetilde{\lambda}^{k+1}, \psi^k(\widetilde{x}^{k+1}) - \eta^k \rangle + \epsilon_k \\
&\le \psi_0^k(x^{k+1}) + \langle \widetilde{\lambda}^{k+1}, \psi^k(x^{k+1}) - \eta^k \rangle + \epsilon_k \\
&\overset{(iv)}{\le} \psi_0^k(x) + \langle \widetilde{\lambda}^{k+1}, \psi^k(x) - \eta^k \rangle - \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{2}\|x^{k+1} - x\|^2 \\
&\quad + 2\epsilon_k + \sqrt{2(L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle)\epsilon_k}\|x^{k+1} - x\|,
\end{aligned}
$$
(7.7)

where (i) follows from the definition of ψ_0^k, (ii) follows since x^{k+1} is an ϵ_k solution of (3.1), (iii) follows by complementary slackness for the optimal primal-dual solution for (3.1) and (iv) follows from Lemma 8. In particular, we use $g(x) + \frac{\gamma}{2}\|x - z\|^2 = \psi_0^k(x) + \langle \widetilde{\lambda}^{k+1}, \psi^k(x) - \eta^k \rangle$ with $z = x^k, z^+ = x^{k+1}, \epsilon = \epsilon_k$ and $\gamma = L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle$. Note that x^{k+1} is an ϵ_k-approximate solution for $\min_{x \in \mathbb{R}^d} \psi_0^k(x) + \langle \widetilde{\lambda}^{k+1}, \psi^k(x) - \eta^k \rangle$ due to Definition 7.

Finally, note that

$$
\psi_0^k(x) + \frac{\mu_0}{2}\|x - x^k\|^2 \le \psi_0(x) + \frac{L_0}{2}\|x - x^k\|^2,
$$
$$
\psi_i^k(x) \le \psi_i(x) + \frac{L_i}{2}\|x - x^k\|^2.
$$

Using the above two relations in (7.7), we obtain (7.6). Hence, we conclude the proof. $\qquad\square$

Let x^* be an optimal solution of (1.1) and $\widetilde{D} := \max\{\|x - y\| : x, y \in \text{dom } \chi_0, \psi_i(x) \le \eta_i, \psi_i(y) \le \eta_i, \text{ for all } i \in [m]\}$. Now, we show convergence rate guarantees.

Theorem 11 *Consider general convex optimization problems with $\mu_0 = 0$. Suppose Assumption 3 is satisfied and set $\delta^k = \frac{(\eta - \eta^0)}{(k+1)(k+2)}$. Then we have*

$$
\psi_0(\bar{x}_K) - \psi_0(x^*) \le \frac{L_0 + B\|L\|}{(K+1)}\left[\widetilde{D}^2 + \frac{(4B+2)\|\eta - \eta^0\|}{L_0} + \widetilde{D}\sqrt{\frac{\|\eta - \eta^0\|}{L_0}} + \frac{\|\eta - \eta^0\|}{L_0}\frac{\log K}{K}\right]
$$
(7.8)

Proof From Lemma 9 with $\mu_0 = 0$ for convex part and $\psi(x^*) \leq \eta$, we have

$$
\begin{aligned}
&\psi_0(x^{k+1}) - \psi_0(x^*) \\
&= \langle \widetilde{\lambda}^{k+1}, \eta - \eta^k \rangle + \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{2} \|x^k - x^*\|^2 - \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{2} \|x^{k+1} - x^*\|^2 \\
&\quad + 2\epsilon_k + \sqrt{2(L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle)\epsilon_k} \|x^{k+1} - x\|.
\end{aligned}
$$

Dividing both sides by $\frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{2}$, we have

$$
\begin{aligned}
\frac{2[\psi_0(x^{k+1}) - \psi_0(x^*)]}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle} &\leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \\
&\quad + \frac{2\langle \widetilde{\lambda}^{k+1}, \eta - \eta^k \rangle}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle} + \frac{4\epsilon_k}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle} + \sqrt{\frac{8\epsilon_k}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}} \|x^{k+1} - x^*\|
\end{aligned}
$$

Note that the sequence $\{\widetilde{\lambda}^{k+1}\}$ is uniformly bounded above such that $\|\widetilde{\lambda}^{k+1}\| \leq B$ for all $k \geq 0$. Using this fact and the above relation, we have

$$
\frac{2[\psi_0(x^{k+1}) - \psi_0(x^*)]}{L_0 + B\|L\|} \leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 + \frac{2B\|\eta - \eta^k\|}{L_0} + \frac{4\epsilon_k}{L_0} + \sqrt{\frac{8\epsilon_k}{L_0}} \|x^{k+1} - x^*\|.
\tag{7.9}
$$

Using $\delta_k = \frac{\eta - \eta^0}{(k+1)(k+2)}$ and $\epsilon_k = \frac{\|\eta - \eta^0\|}{2(k+1)(k+2)}$, we have x^k is strictly feasible solutions for (3.1) for all k. Hence, under Assumption 3, we can follow the steps of Theorem 1 to show uniform bound B on sequence $\{\|\widetilde{\lambda}^k\|\}$. Using these values in (7.9), we have

$$
\begin{aligned}
\frac{2[\psi_0(x^{k+1}) - \psi_0(x^*)]}{L_0 + B\|L\|} &\leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 + \frac{2B\|\eta - \eta^0\|}{L_0(k+1)} \\
&\quad + \frac{2\|\eta - \eta^0\|}{L_0(k+1)(k+2)} + \sqrt{\frac{\|\eta - \eta^0\|}{L_0}} \frac{1}{k+1} \|x^{k+1} - x^*\|.
\end{aligned}
\tag{7.10}
$$

Due to the optimality of the exact solution $\widetilde{x}^{k+1}$, we have $\psi_0^k(\widetilde{x}^{k+1}) \leq \psi_0^k(x^k) = \psi_0(x^k)$. We also have $\psi_0(x^{k+1}) \leq \psi_0^k(x^{k+1}) \leq \psi_0^k(\widetilde{x}^{k+1}) + \epsilon_k$. Combining these two relations, we get:

$$
\psi_0(x^{k+1}) \leq \psi_0(x^k) + \epsilon_k.
$$

Effectively, inexact `LCPG` method is almost (up to an additive error of ϵ_k) a descent method. Using this relation recursively, we have

$$
\begin{aligned}
\psi_0(x^K) &\leq \psi_0(x^{k+1}) + \sum_{i=k+1}^{K-1} \epsilon_i. \\
&\leq \psi_0(x^{k+1}) + \frac{\|\eta - \eta^0\|}{2} \sum_{i=k+1}^{K-1} \frac{1}{(i+1)(i+2)} \\
&= \psi_0(x^{k+1}) + \frac{\|\eta - \eta^0\|}{2} \frac{(K-k-1)}{(k+2)(K+1)}.
\end{aligned}
$$

Using the above relation in (7.10), we have

$$
\frac{2[\psi_0(x^K) - \psi_0(x^*)]}{L_0 + B\|L\|}
$$

$$
\leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 + \frac{2B\|\eta - \eta^0\|}{L_0(k+1)}
$$

$$
+ \frac{2\|\eta - \eta^0\|}{L_0(k+1)(k+2)} + \sqrt{\frac{\|\eta - \eta^0\|}{L_0}} \frac{1}{k+1} \|x^{k+1} - x^*\|
$$

$$
+ \frac{\|\eta - \eta^0\|}{L_0 + B\|L\|} \frac{K - k - 1}{(k+2)(K+1)}.
$$

Multiplying the above relation by $k + 1$ and summing from $k = 0$ to $K - 1$, we have

$$
\frac{K(K+1)[\psi_0(x_K) - \psi_0(x^*)]}{L_0 + B\|L\|} \leq \sum_{k=0}^{K-1} \|x^k - x^*\|^2 + \frac{2B\|\eta - \eta^0\|K}{L_0} + \frac{\|\eta - \eta^0\| \log(K+2)}{L_0}
$$

$$
+ \sqrt{\frac{\|\eta - \eta^0\|}{L_0}} \sum_{k=0}^{K-1} \|x^{k+1} - x^*\| + \frac{\|\eta - \eta^0\|K}{L_0 + B\|L\|}
$$

$$
\leq K\widetilde{D}^2 + \frac{2(B+1)\|\eta - \eta^0\|K}{L_0} + \sqrt{\frac{\|\eta - \eta^0\|}{L_0}} K\widetilde{D}.
$$

After rearranging, this relation implies (7.8). Hence, we conclude the proof. □

Theorem 12 *Consider strongly convex problems ($\mu_0 > 0$) and suppose that Assumption 3 is satisfied. Set $\delta^k = \rho^k(1 - \rho)(\eta - \eta^0)$ where $\rho = \frac{L_0 - \mu_0}{2(L_0 - a\mu_0)}$, $2\epsilon_k \leq a(1 - \rho)\rho^k\|\eta - \eta^0\|$ and $a \in (0, 1)$. Then we have*

$$
\psi_0(x^K) - \psi_0(x^*) \leq \exp\left(-\frac{(1-a)\mu_0 K}{L_0 + B\|L\| - a\mu_0}\right)(L_0 + B\|L\| - \mu_0)
$$

$$
\left\{\left[\frac{(4B+1)}{2(L_0 - \mu_0)} + \frac{L_0 + B\|L\| + 2a\mu_0}{\mu_0(L_0 - \mu_0)}(1 - \rho)\right]\|\eta - \eta^0\| + \frac{1}{2}\|x^0 - x^*\|^2\right\}.
$$

$$(7.11)$$

Moreover, if $\epsilon_k = 0$, we have

$$
\psi_0(x^K) - \psi_0(x^*) \leq \exp\left(-\frac{\mu_0 K}{L_0 + B\|L\|}\right)(L_0 + B\|L\| - \mu_0)
$$

$$
\left\{\left[\frac{(4B+1)}{2(L_0 - \mu_0)}\right]\|\eta - \eta^0\| + \frac{1}{2}\|x^0 - x^*\|^2\right\}.
$$

$$(7.12)$$

Proof Proceeding similar to the convex case, using Lemma 9, we obtain

$$
\psi_0(x^{k+1}) - \psi_0(x^*)
$$

$$
= \langle \widetilde{\lambda}^{k+1}, \eta - \eta^k \rangle + \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0}{2} \|x^k - x^*\|^2
$$

$$
- \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{2} \|x^{k+1} - x^*\|^2
$$

$$
+ 2\epsilon_k + \sqrt{2(L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle)\epsilon_k} \|x^{k+1} - x\|.
$$

For $0 < a < 1$, we have

$$\sqrt{2(L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle)\epsilon_k} \|x^{k+1} - x\| \leq \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{a\mu_0}\epsilon_k + \frac{a\mu_0}{2}\|x^{k+1} - x^*\|^2.$$

Combining the above two results, we have

$$\begin{aligned}
\psi_0(x^{k+1}) &- \psi_0(x^*) \\
&\leq \langle \widetilde{\lambda}^{k+1}, \eta - \eta^k \rangle + \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0}{2}\|x^k - x^*\|^2 \\
&\quad - \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - a\mu_0}{2}\|x^{k+1} - x^*\|^2 \\
&\quad + \left(2 + \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle}{a\mu_0}\right)\epsilon_k.
\end{aligned}$$

Let us denote

$$\Gamma_k = \begin{cases} 1 & \text{if } k = 0; \\ \frac{L_0 + \langle \widetilde{\lambda}^k, L \rangle - a\mu_0}{L_0 + \langle \widetilde{\lambda}^k, L \rangle - \mu_0}\Gamma_{k-1} & \text{if } k \geq 1. \end{cases}$$

Multiplying both sides of the above inequality by $\frac{\Gamma_k}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0}$ and noting that $\eta - \eta^k = \rho^k(\eta - \eta^0)$ (follows by the choice of δ^k), we obtain

$$\begin{aligned}
\frac{\Gamma_k}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0}&\left[\psi_0(x^{k+1}) - \psi_0(x^*)\right] \\
&\leq \frac{\Gamma_k}{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0}\rho^k \langle \widetilde{\lambda}^{k+1}, \eta - \eta^0 \rangle \\
&\quad + \frac{\Gamma_k}{2}\|x^* - x^k\|^2 - \frac{\Gamma_{k+1}}{2}\|x^* - x^{k+1}\|^2 \\
&\quad + \frac{L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle + 2a\mu_0}{a\mu_0(L_0 + \langle \widetilde{\lambda}^{k+1}, L \rangle - \mu_0)}\Gamma_k\epsilon_k. \quad (7.13)
\end{aligned}$$

Since $\|\widetilde{\lambda}^k\| \leq B$, we have $\left(\frac{L_0 + B\|L\| - a\mu_0}{L_0 + B\|L\| - \mu_0}\right)^k \leq \Gamma_k \leq \left(\frac{L_0 - a\mu_0}{L_0 - \mu_0}\right)^k$. Moreover, we have $2\epsilon_k \leq a(1 - \rho)\rho^k\|\eta - \eta^0\|$ and $\rho = \frac{L_0 - \mu_0}{2(L_0 - a\mu_0)}$. Using these relations in (7.13), we have

$$\begin{aligned}
\frac{\Gamma_k}{L_0 + B\|L\| - \mu_0}&\left[\psi_0(x^{k+1}) - \psi_0(x^*)\right] \\
&\leq \frac{B\|\eta - \eta^0\|}{L_0 - \mu_0}\Gamma_k\rho^k + \frac{L_0 + B\|L\| + 2a\mu_0}{a\mu_0(L_0 - \mu_0)}\Gamma_k\epsilon_k + \frac{\Gamma_k}{2}\|x^k - x^*\|^2 - \frac{\Gamma_{k+1}}{2}\|x^{k+1} - x^*\|^2 \\
&\leq \frac{B\|\eta - \eta^0\|}{L_0 - \mu_0}\frac{1}{2^k} + \frac{L_0 + B\|L\| + 2a\mu_0}{\mu_0(L_0 - \mu_0)}\frac{1 - \rho}{2^{k+1}}\|\eta - \eta^0\| + \frac{\Gamma_k}{2}\|x^k - x^*\|^2 - \frac{\Gamma_{k+1}}{2}\|x^{k+1} - x^*\|^2.
\end{aligned}$$
$$(7.14)$$

Similar to the convex part, we also have

$$\begin{aligned}
\psi_0(x^K) &\leq \psi_0(x^{k+1}) \\
&\quad + \sum_{i=k+1}^{K-1}\epsilon_i \leq \psi_0(x^{k+1})
\end{aligned}$$

$$+ \frac{\|\eta-\eta^0\|(1-\rho)}{2}\sum_{i=k+1}^{K-1}\rho^i \leq \psi_0(x^{k+1})$$
$$+ \frac{\|\eta-\eta^0\|\rho^{k+1}}{2}.$$

Using the above relation into (7.14), we have

$$\frac{\Gamma_k}{L_0+B\|L\|-\mu_0}\big[\psi_0(x^K) - \psi_0(x^*)\big] \leq \frac{(4B+1)\|\eta-\eta^0\|}{L_0-\mu_0}\frac{1}{2^{k+2}} + \frac{L_0+B\|L\|+2a\mu_0}{\mu_0(L_0-\mu_0)}\frac{1-\rho}{2^{k+1}}\|\eta - \eta^0\|$$
$$+ \frac{\Gamma_k}{2}\|x^k - x^*\|^2 - \frac{\Gamma_{k+1}}{2}\|x^{k+1} - x^*\|^2.$$

Summing the above relation from $k = 0$ to $K - 1$, we have

$$\frac{\Gamma_{K-1}}{L_0+B\|L\|-\mu_0}\big[\psi_0(x^K) - \psi_0(x^*)\big]$$
$$\leq \sum_{k=0}^{K-1} \frac{\Gamma_k}{L_0+B\|L\|-\mu_0}\big[\psi_0(x^K) - \psi_0(x^*)\big]$$
$$\leq \Big[\frac{(4B+1)}{2(L_0-\mu_0)}$$
$$+ \frac{L_0+B\|L\|+2a\mu_0}{\mu_0(L_0-\mu_0)}(1 - \rho)\Big]\|\eta - \eta^0\| + \frac{1}{2}\|x^0 - x^*\|^2.$$

Note that

$$\Gamma_{K-1}^{-1} \leq \big(1 - \frac{(1-a)\mu_0}{L_0+B\|L\|-a\mu_0}\big)^K \leq \exp\big(-\frac{(1-a)\mu_0 K}{L_0+B\|L\|-a\mu_0}\big).$$

Combining the above two relations we obtain the desired result (7.11). □

8 Numerical study

In this section, we conduct some preliminary studies to examine our theoretical results and the performance of the LCPG method. The experiments are run on CentOS with Intel Xeon (2.60 GHz) and 128 GB memory.

8.1 A simulated study on the QCQP

In the first experiment, we compare LCPG with some established open-source solvers such as CVXPY [11] and DCCP [32]. We consider the penalized Quadratically Constrained Quadratic Program (QCQP) described as follows,

$$\begin{aligned}
\min_{x\in\mathbb{R}^n} \quad & \tfrac{1}{2}x^T Q_0 x + b_0^T x + \alpha\|x\|_1 \\
\text{s.t.} \quad & \tfrac{1}{2}x^T Q_i x + b_i^T x + c_i \leq 0, \quad i = 1, 2, \ldots, m - 1 \\
\text{s.t.} \quad & \|x\| \leq r
\end{aligned} \tag{8.1}$$

where each Q_i ($0 \leq i \leq m$) is an $n \times n$ matrix, $b_0, b_1, \ldots, b_m$ are n-dimensional real vectors, α is a positive weight on the ℓ_1 norm penalty, which helps to promote sparse solution. In the first setting, we consider a convex constrained problem where each Q_i

Table 3 Comparison of algorithms on convex quadratic problems. Running time is measured in seconds

n	LCPG				CVXPY		
	Objective	Time	DNorm	Max DNorm	Objective	Time	DNorm
500	−1.785e+02	9.46e+00	1.773e−01	1.773e−01	-1.785e+02	3.29e+00	1.774e−01
1000	−9.698e+01	1.97e+01	1.478e−01	1.478e−01	-9.698e+01	1.61e+01	1.479e−01
2000	−5.418e+01	6.30e+01	1.117e−01	1.117e−01	-5.419e+01	7.82e+01	1.120e−01
3000	−4.010e+01	1.44e+02	9.525e−02	9.525e−02	-4.011e+01	2.36e+02	9.602e−02
4000	−3.293e+01	2.45e+02	8.255e−02	8.255e−02	-3.294e+01	4.96e+02	8.384e−02

is a positive semidefinite matrix. We set $Q_i = VDV^{\mathrm{T}}$ where V is an $n \times n$ random sparse matrix with density 0.01, and its nonzero entries are uniformly distributed in $[0, 1]$. D is a diagonal matrix whose diagonal elements are uniformly distributed in between $[0, 100]$. We set $b_i = 10e + v$, where e is a vector of ones and $v \in \mathcal{N}(0, I_{n \times n})$ is sampled from standard Gaussian distribution. We set $c_i = -10$ to make $x = 0$ a strictly feasible initial solution. Furthermore, we add a ball constraint to ensure that the domain is a compact set. We set $r = \sqrt{20}$ and $\alpha = 1$. We fix $m = 10$ and explore different dimensions n from the set {500, 1000, 2000, 3000, 4000}.

We solve Problem (8.1) by both CVXPY and LCPG. Both use the initial solution $x = 0$. For CVXPY, we use MOSEK as the internal solver due to its superior performance in quadratic optimization. In LCPG, for simplicity, we also solve the diagonal quadratic subproblem by MOSEK through CVXPY. Note that calling the external API repetitively for each LCPG subproblem only causes more overheads to run LCPG. Nonetheless, as we shall see, the standard IPM solvers can still fully leverage the diagonal structure and exhibit fast convergence.

In Table 3, we present the experiment results of the compared algorithms. The final objective, the norm of the dual solution (DNorm), and for LCPG, the maximum dual norm in the solution path (Max DNorm) are reported. All values represent the average of 5 independent runs. From the results, we observe that while LCPG does not outperform CVXPY for the small-size problem ($n = 500$), LCPG becomes increasingly favorable as the problem dimension increases. This justifies the empirical advantage of our proposed approach as we do not need to construct a full Hessian matrix. Moreover, interestingly, we observe that the dual solution norm $\{\|\lambda^k\|\}$ is increasing, reaching the maximum at the last iteration. This accounts for the equal values of DNorm and Max DNorm. Meanwhile, in all the cases, the dual remains bounded and the reported dual norm closely aligns with the solution returned by CVXPY. This result confirms our intuition that the dual bound is intricately tied to the nature of problems.

In the second setting of this experiment, we examine the performance of LCPG on nonconvex constrained optimization. Specifically, we express Q_i as the difference of two matrices: $Q_i = P_i - S_i$, where P_i is generated in the same manner as Q_i in the first setting, and $S_i = 10I_{n \times n}$. Given the construction of the quadratic components, it is natural to view the function $\frac{1}{2}x^T Q_i x + b_i^T x + c_i$ as a difference of two convex quadratic functions: $\frac{1}{2}x^T P_i x + b_i^T x + c_i - \frac{1}{2}x^T S_i x$. Leveraging this decomposition, we apply the DC programming, and more specifically, the DCCP framework to solve

Table 4 Comparison of algorithms on nonconvex quadratic problems

n	LCPG				DCCP			
	Objective	Time	DNorm	Max DNorm	Objective	Time	DNorm	Max DNorm
500	−3.056e+01	1.525e+01	1.194e−01	1.194e−01	−3.056e+01	1.400e+01	1.199e−01	1.199e−01
1000	−2.112e+01	3.350e+01	7.602e−02	7.608e−02	−2.112e+01	5.665e+01	7.672e−02	7.672e−02
2000	−1.609e+01	1.052e+02	4.502e−02	4.588e−02	−1.609e+01	2.678e+02	4.571e−02	4.571e−02
3000	−1.425e+01	1.918e+02	3.216e−02	3.523e−02	−1.426e+01	6.505e+02	3.301e−02	3.301e−02
4000	−1.329e+01	3.837e+02	2.485e−02	2.738e−02	−1.330e+01	1.201e+03	2.574e−02	2.574e−02
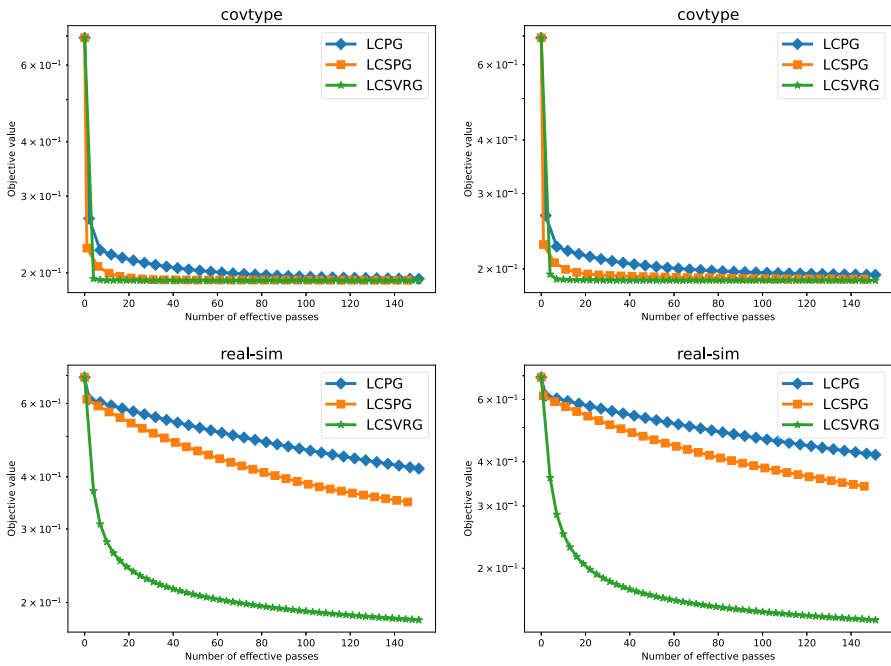
Fig. 3 Comparison of `LCPG`, `LCSPG` and `LCSVRG`. The first row reports the results on covtype (left: $\sigma = 0.4$; right: $\sigma = 0.6$). The second row reports the results on real-sim (left: $\sigma = 0.1$; right: $\sigma = 0.2$)

(8.1). Each convex subproblem of `DCCP` is solved by `MOSEK` through the CVXPY interface. In Table 4, we describe the performance of `LCPG` and the `DCCP` algorithm. It can be observed that `LCPG` compares favorably against the `DCCP` solver. Furthermore, the boundedness of the dual for both algorithms is also observed, which is consistent with our intuition.

8.2 Study of gradient complexities

In the next experiment, our primary goal is to examine the main theoretical results, namely, the gradient complexities of `LCPG`, its stochastic variants `LCSPG`

and `LCSVRG`. We apply all these algorithms to a sparsity-induced finite-sum problem, wherein a nonconvex constraint is incorporated into the supervised learning framework to actively enforce a sparse solution. The optimization problem is as follows

$$
\begin{aligned}
\min_{x \in \mathbb{R}^d} \quad & \psi_0(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \\
\text{s.t.} \quad & \psi_1(x) = \beta \|x\|_1 - g(x) \leq \eta_1,
\end{aligned}
\tag{8.2}
$$

where $f_i(x)$ is a smooth loss function associated with the ith sample, $\psi_1(x)$ is the difference between ℓ_1-penalty and a convex smooth function $g(x)$. Employing a difference-of-convex constraint is seen as a tighter relaxation of the cardinality constraint $\|x\|_0 \leq \kappa$ than the ℓ_1 relaxation. The appealing properties of difference-of-convex penalties have been demonstrated in various studies [5, 14, 16, 17, 36, 37].

In view of the concave structure of $-g(x)$, there is a strong asymmetry between the lower and upper curvature of $-g(x)$, namely, the following

$$
-\frac{L_g}{2}\|y - x\|^2 - \nabla g(y)^T(x - y) - g(y) \leq -g(x) \leq -g(y) - \nabla g(y)^T(x - y) \tag{8.3}
$$

holds for certain $L_g > 0$. Note that this is much stronger than the L_g smoothness condition which adds an extra $\frac{L_g}{2}\|y - x\|^2$ on the right-hand side of (8.3). Due to this feature, one can impose a tighter piece-wise linear surrogate function constraint

$$
\beta \|x\|_1 - g(x^k) - \nabla g(x^k)(x - x^k) \leq \eta_1^k
$$

in the `LCPG` subproblem. It should be noted that our analysis is readily adaptable to accommodate this scenario since it is the smoothness, as opposed to concavity/convexity, that plays a central role in our convergence analysis and that remains valid. An empirical advantage of this approach is that we now have a tractable subproblem solvable in nearly linear time. See more discussion in [5].

Our experiment considers the task of binary classification with logistic loss, denoted by $f_i(x) = \log(1 + \exp(-b_i(a_i^T x)))$, where $a_i \in \mathbb{R}^d$, $b_i \in \{1, -1\}$, $1 \leq i \leq n$. We use the SCAD penalty $g(x) = \sum_{j=1}^{d} h_{\beta,\theta}(x_j)$ where $h_{\beta,\theta}(\cdot)$ is defined in (3.22). We use the real-sim dataset from the LibSVM repository [10] and the covtype data from the UCI repository [19]. For the latter, we formulate a binary classification task by distinguishing class "3" from the other classes. We set $\beta = 2$, $\theta = 5$, and $\eta_1 = \sigma d$, with $\sigma \in \{0.4, 0.6\}$ for covtype and $\sigma \in \{0.1, 0.2\}$ for real-sim dataset. For each algorithm, we use its theoretically suggested batch size and stepsize. for a fair comparison, we count n evaluations of the stochastic gradient as an effective pass over the dataset and plot the objective value over the number of effective passes. Figure 3 plots the convergence result of the compared algorithms. It can be readily seen that `LCSPG` performs better than `LCPG` in terms of the number of gradient samples, and `LCSVRG` achieves the best performance among the three algorithms. The empirical findings further confirm our theoretical complexity analysis.

9 Conclusion

In this work, we presented a new `LCPG` method for nonconvex function constrained optimization which can achieve gradient complexity of the same order as that of unconstrained nonconvex problems. The key ingredient in our algorithm design is the use of constraint levels to ensure the subproblem feasibility, which allows us to overcome a well-known difficulty in bounding the Lagrange multipliers in the presence of nonsmooth constraints. Moreover, a merit of our convergence analysis is its striking similarity with that of gradient descent methods for unconstrained problems. Therefore, we can easily extend our method to minimizing stochastic, finite-sum, and structured nonsmooth functions with nonconvex function constraints; many of the complexity results were not known before. Another important feature of our work is that the method can deal with complex scenarios where the subproblems are not exactly solvable. To the best of our knowledge, existing work on sequential convex optimization (`SQP`, `MBA`) only assumes the subproblems to be exactly solved. We provided a detailed complexity analysis of `LCPG` when the subproblems are inexactly solved by customized interior point method and first-order method. Finally, we clearly distinguished the notion of gradient complexity from that of computational complexity. In terms of gradient complexity, all of our proposed methods are state-of-the-art and easy to implement. Whether the computational complexity can be further improved for composite cases remains an open problem that we leave as a future direction.

A Proof of Lemma 7

Let $\{\delta_n\}_{n \geq 1}$ be a function of $\mathcal{C}^{\infty}$-smooth, real-valued mollifier functions over $\mathbb{R}^d$ where, for every $n \geq 1$, we have: (i) $\delta_n \geq 0$, (ii) $\int \delta_n(\tau) d\tau = 1$, and (iii) and $\delta_n(\tau) = 0$ for τ satisfying $\|\tau\| \geq \frac{1}{n}$. Moreover, we define $p_n = \delta_n * p$. It now follows that

$$
\begin{aligned}
&p_n(x) - p_n(y) - \langle \nabla p_n(y), x - y \rangle \\
=&\delta_n * [p(x) - p(y)] - \langle \delta_n * \nabla p(y), x - y \rangle \\
=&\int_{\tau} \delta_n(\tau)[p(x - \tau) - p(y - \tau)]d\tau - \langle \int_{\tau} \delta_n(\tau)\nabla p(y - \tau)d\tau, x - y \rangle \\
=&\int_{\tau} \delta_n(\tau)[p(x - \tau) - p(y - \tau) - \langle \nabla p(y - \tau), x - y \rangle]d\tau
\end{aligned}
$$

Using (5.2) along with the above relation, noting that $\delta_n \geq 0$ and the fact $\int \delta_n(\tau)d\tau = 1$, we have

$$-\tfrac{\mu}{2}\|x - y\|^2 \leq p_n(x) - p_n(y) - \langle \nabla p_n(y), x - y \rangle \leq \tfrac{L}{2}\|x - y\|^2, \qquad \text{(A.1)}$$

for all x, y. Note that p_n is C^∞-smooth. Hence, using Taylor's theorem, there exist $\xi \in [x, y]$ such that $p_n(x) - p_n(y) - \langle \nabla p_n(y), x - y \rangle = \tfrac{1}{2}\langle x - y, \nabla^2 p_n(\xi)(x - y) \rangle + o(\|x - y\|^2)$. Using this relation along with (A.1) and denoting $v := y - x$, we have

$$-\tfrac{\mu}{2}\|v\|^2 \leq \tfrac{1}{2}\langle v, \nabla^2 p_n(\xi)v \rangle + o(\|v\|^2) \leq \tfrac{L}{2}\|v\|^2 \qquad \text{(A.2)}$$

Now, diving both sides of the above relation by $\|v\|^2$, taking $y \to x$ which implies $\xi \to x$ and $\nabla^2 p_n(\xi) \to \nabla^2 p_n(x)$, we have

$$-\mu \leq \langle \tfrac{v}{\|v\|}, \nabla^2 p_n(x) \tfrac{v}{\|v\|} \rangle \leq L,$$

for all v and x. The above relation is equivalent to the fact that $\|\nabla^2 p_n(x)\| \leq \max\{L, \mu\}$ for all x. From here, for any x, y, we have

$$\begin{aligned}
\|\nabla p_n(x) - \nabla p_n(y)\| &= \|\int_{t=0}^{1} \nabla^2 p(y + t(x - y))(x - y)dt\| \\
&\leq \int_{t=0}^{1} \|\nabla^2 p(y + t(x - y))(x - y)\|dt \\
&\leq \int_{t=0}^{1} \|\nabla^2 p(y + t(x - y))\| \|x - y\|dt \\
&\leq \max\{L, \mu\}\|x - y\|.
\end{aligned}$$

Now, taking $n \to \infty$ and noting that $\nabla p_n(x) \to \nabla p(x)$ for all x, we have (5.3). Hence, we conclude the proof.

B Proof of Theorem 7

First of all, using the definition of ϵ_k and the fact that $\psi_i^k(x^{k+1}) \leq \eta_i^k + \epsilon_k$ for all $i \in [m]$, we have

$$\psi_i^k(x^k) = \psi_i(x^k) \leq \psi_i^{k-1}(x^k) \leq \eta_i^{k-1} + \epsilon_{k-1} < \eta_i^{k-1} + \delta_i^{k-1} = \eta_i^k, \quad i = 1, 2, \ldots, m.$$

Hence x^k is strictly feasible solution of (3.1). Due to Slater condition, there exists a pair of optimal primal and dual solutions, which we denote by $\tilde{x}^{k+1}$ and λ^{k+1}. We first prove the following lemma. Our argument will be based on the following key result.

Lemma 10

$$\lim_{k\to\infty} \|\tilde{x}^{k+1} - x^k\| = 0, \quad \lim_{k\to\infty} \|x^{k+1} - x^k\| = 0 \qquad \text{(B.1)}$$

Proof Using optimality of $\widetilde{x}^{k+1}$, strong convexity of ψ_0^k, feasibility of x_k for the subproblem (3.1) and $\psi_0^k(x^{k+1}) \leq \psi_0^k(\widetilde{x}^{k+1}) + \epsilon_k$, we have,

$$
\begin{aligned}
\psi_0(x^{k+1}) &\leq \psi_0^k(x^{k+1}) \\
&\leq \psi_0^k(\widetilde{x}^{k+1}) + \epsilon_k \\
&\leq \psi_0^k(x^k) - \frac{L_0}{2}\|x^k - \widetilde{x}^{k+1}\|^2 + \epsilon_k \\
&= \psi_0(x^k) - \frac{L_0}{2}\|x^k - \widetilde{x}^{k+1}\|^2 + \epsilon_k.
\end{aligned}
$$

Since ϵ_k is summable, we have that $\|x^k - \widetilde{x}^{k+1}\|^2$ is summable. This implies $\|x^k - \widetilde{x}^{k+1}\| \to 0$. Since $\epsilon_k \to 0$ and $\widetilde{x}^{k+1}$ is a unique optimal solution of (3.1), we have $\|x^{k+1} - \widetilde{x}^{k+1}\| \to 0$. Then, using Cauchy-Schwarz inequality, we have $\|x^{k+1} - x^k\| \leq \|x^{k+1} - \widetilde{x}^{k+1}\| + \|\widetilde{x}^{k+1} - x^k\|$ and hence, $\|x^{k+1} - x^k\| \to 0$. Hence, we conclude the proof. $\qquad\square$

Now, we show boundedness of $\tilde{\lambda}^{k+1}$. Assume, for the sake of contradiction, that $\tilde{\lambda}^{k+1}$ is unbounded. Let $\bar{x}$ be a limit point of $\{x^k\}$. Passing to a subsequence if necessary, we have $x^k \to \bar{x}$. Using Lemma 10, we have $\widetilde{x}^{k+1} \to \bar{x}$ and $x^{k+1} \to \bar{x}$. Then, we have

$$
\begin{aligned}
\psi_0^k(x^{k+1}) + \langle \tilde{\lambda}^{k+1}, \psi^k(x^{k+1})\rangle &\leq \psi_0^k(\widetilde{x}^{k+1}) + \langle \tilde{\lambda}^{k+1}, \psi^k(\widetilde{x}^{k+1})\rangle + \epsilon_k \\
&\leq \psi_0^k(x) + \langle \tilde{\lambda}^{k+1}, \psi^k(x)\rangle + \epsilon_k
\end{aligned}
$$

Note that the above relation is comparable to (3.10) up to an error term of ϵ_k. Following the arguments in the proof of Theorem 1 (Part 1, (3.10) onwards) and noting that ϵ_k is summable, we conclude that $\{\tilde{\lambda}^{k+1}\}$ is bounded.

Now, we prove limit point of $\{x^k\}$ is a KKT point. Since $\mathcal{L}_k(x^{k+1}, \tilde{\lambda}^{k+1}) \leq \mathcal{L}_k(\widetilde{x}^{k+1}, \tilde{\lambda}^{k+1}) + \epsilon_k$, we rewrite (3.18) as

$$
\begin{aligned}
&\left\langle \nabla f_0(x^k) + \textstyle\sum_{i=1}^m \tilde{\lambda}_i^{k+1}\nabla f_i(x^k), x^{k+1} - x\right\rangle + \chi_0(x^{k+1}) - \chi_0(x) \\
&\quad + \langle \tilde{\lambda}^{k+1}, \chi(x^{k+1}) - \chi(x)\rangle \\
&\leq \left\langle \nabla f_0(x^k) + \textstyle\sum_{i=1}^m \tilde{\lambda}_i^{k+1}\nabla f_i(x^k), \widetilde{x}^{k+1} - x\right\rangle \\
&\quad + \chi_0(\widetilde{x}^{k+1}) - \chi_0(x) + \langle \tilde{\lambda}^{k+1}, \chi(\widetilde{x}^{k+1}) - \chi(x)\rangle + \epsilon_k \\
&\leq \frac{L_0 + \langle \tilde{\lambda}^{k+1}, L\rangle}{2}\left[\|x - x^k\|^2 - \|\widetilde{x}^{k+1} - x^k\|^2 - \|x - \widetilde{x}^{k+1}\|^2\right] + \epsilon_k. \qquad\text{(B.2)}
\end{aligned}
$$

Let $\bar{x}$ be a limit point of sequence $\{x^k\}$. Since $\tilde{\lambda}^{k+1}$ is bounded, we assume limit point $\bar{\lambda}$. Without loss of generality, we have $x^k \to \bar{x}$ and $\tilde{\lambda}^{k+1} \to \bar{\lambda}$. Then, in view of Lemma 10, we have $\lim_{k\to\infty} x^{k+1} \to \bar{x}$ and $\lim_{k\to\infty} \widetilde{x}^{k+1} \to \bar{x}$. Taking limit $k \to \infty$ in (B.2), we have

$$
\left\langle \nabla f_0(\bar{x}) + \textstyle\sum_{i=1}^m \bar{\lambda}_i \nabla f_i(\bar{x}), \bar{x} - x\right\rangle + \chi_0(\bar{x}) - \chi_0(x) + \langle \bar{\lambda}, \chi(\bar{x}) - \chi(x)\rangle \leq 0.
$$

Note that the above equation matches with (3.19) exactly. From here, we follow the proof of Theorem 1 (Part 2, (3.19) onwards) to conclude first-order stationarity of

$(\bar{x}, \bar{\lambda})$. A similar argument can show complimentary slackness. Hence, we have that $(\bar{x}, \bar{\lambda})$ is KKT-solution and we conclude the proof.

References

1. Auslender, A., Shefi, R., Teboulle, M.: A moving balls approximation method for a class of smooth constrained minimization problems. SIAM J. Optim. **20**(6), 3232–3259 (2010)
2. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Nashua (1999)
3. Bolte, J., Pauwels, E.: Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. Math. Oper. Res. **41**(2), 442–465 (2016)
4. Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. Mathematical Programming, pp. 1–65 (2022)
5. Boob, D., Deng, Q., Lan, G., Wang, Y.: A feasible level proximal point method for nonconvex sparse constrained optimization. In Advances in Neural Information Processing Systems, volume 33, pp. 16773–16784. Curran Associates, Inc., (2020)
6. Burke, J.: A sequential quadratic programming method for potentially infeasible mathematical programs. J. Math. Anal. Appl. **139**(2), 319–351 (1989)
7. Burke, J.V., Han, S.-P.: A robust sequential quadratic programming method. Math. Program. **43**(1), 277–303 (1989)
8. Cartis, C., Gould, N.I., Toint, P.L.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. **4**, 1721–1739 (2011)
9. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of finding first-order critical points in constrained nonlinear optimization. Math. Program. **144**(1), 93–106 (2014)
10. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1-27:27 (2011)
11. Diamond, S., Boyd, S.: Cvxpy: a python-embedded modeling language for convex optimization. J. Mach. Learn. Res. **17**(1), 2909–2913 (2016)
12. Facchinei, F., Kungurtsev, V., Lampariello, L., Scutari, G.: Ghost penalties in nonconvex constrained optimization: diminishing stepsizes and iteration complexity. Math. Oper. Res. **46**(2), 595–627 (2021)
13. Facchinei, F., Lampariello, L., Scutari, G.: Feasible methods for nonconvex nonsmooth problems with applications in green communications. Math. Program. **164**(1), 55–90 (2017)
14. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
15. Ghadimi, S., Lan, G.: Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**(4), 2341–2368 (2013)
16. Gong, P., Zhang, C., Lu, Z., Huang, J., Ye, J.: A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: International Conference on Machine Learning, pp. 37–45. PMLR (2013)
17. Gotoh, J.-Y., Takeda, A., Tono, K.: Dc formulations and algorithms for sparse optimization problems. Math. Program. **169**, 141–176 (2018)
18. Reddi, S.J., Sra, S., Poczos, B., Smola, A. J.: Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In: Advances in Neural Information Processing Systems, volume 29, (2016)
19. Kelly, M., Longjohn, R., Nottingham, K.: The UCI machine learning repository (2019)
20. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer, New York (2020)
21. Lan, G.: First-Order and Stochastic Optimization Methods for Machine Learning. Springer-Nature, New York (2020)
22. Lan, G., Monteiro, R.D.C.: Iteration-complexity of first-order penalty methods for convex programming. Math. Program. **138**, 115–139 (2013)
23. Lan, G., Monteiro, R.D.C.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. Math. Program. **155**(1–2), 511–547 (2016)
24. Li, Z., Xu, Y.: Augmented Lagrangian-based first-order methods for convex-constrained programs with weakly convex objective. INFORMS J. Optim. **3**(4), 373–397 (2021)
25. Lin, Q., Ma, R., Xu, Y.: Inexact proximal-point penalty methods for constrained non-convex optimization. arXiv:1908.11518, (2019)

26. Ma, R., Lin, Q., Yang, T.: Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In: International Conference on Machine Learning, pp. 6554–6564. PMLR (2020)

27. Mangasarian, O., Fromovitz, S.: The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. J. Math. Anal. Appl. **17**, 37–47 (1967)

28. Nemirovski, A.: Interior point polynomial time methods in convex programming. Lecture notes (2004)

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

30. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. Optimizing Methods in Statistics, pp. 111–135 (1971)

31. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization (2011). arXiv:1109.2415

32. Shen, X., Diamond, S., Gu, Y., Boyd, S.: Disciplined convex-concave programming. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 1009–1014. IEEE (2016)

33. Wang, X., Ma, S., Yuan, Y.: Penalty methods with stochastic approximation for stochastic nonlinear programming. Math. Comput. **86**(306), 1793–1820 (2017)

34. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. Math. Program. (2019)

35. Yu, P., Pong, T.K., Lu, Z.: Convergence rate analysis of a sequential convex programming method with line search for a class of constrained difference-of-convex optimization problems. SIAM J. Optim. **31**(3), 2024–2054 (2021)

36. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)

37. Zhang, C.-H., Zhang, T.: A general theory of concave regularization for high-dimensional sparse estimation problems. Stat. Sci. **27**(4), 576–593 (2012)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.