Lower Bounds for Feasibility and Stationarity in First-Order Nonconvex Constrained Optimization

Qiankun Shi* Xiao Wang†

June 28, 2026

Abstract

We study oracle-complexity lower bounds for first-order methods applied to smooth equality-constrained nonconvex optimization, separating two components of approximate KKT accuracy: feasibility and stationarity. Under iteration-wise Jacobian regularity, we prove a lower bound of order $\Omega(\frac{L_c\Delta_c}{\sigma^2} + \log\log(\frac{\sigma^2}{L_c\epsilon}))$ to achieve ϵ-feasibility. For stationarity, we construct a nonlinear equality-constrained hard instance whose multiplier-minimized stationarity residual reduces exactly to the gradient of a scalar function, yielding the lower bound $\Omega((L_f + \frac{GL_c}{\sigma})(\Delta_f + \frac{G\Delta_c}{\sigma})\epsilon^{-2})$ to reach ϵ-stationarity. A reduction using disjoint supports transfers zero-chain lower bounds to any deterministic first-order methods and, for the stationarity construction, preserves the ℓ_∞-norm objective gradient bound. We also verify tightness of the lower bound regarding feasibility via a damped Newton method and regarding the stationarity, up to the stated dimension dependence, via a prox-linear method. Finally, we establish the union lower bound for smooth equality-constrained nonconvex problems and show its tightness through a two-stage method.

1 Introduction

We consider lower-bound complexity theory for first-order methods applied to smooth equality-constrained optimization problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad c(x) = 0, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are Lipschitz continuously differentiable and possibly nonconvex and nonlinear.

For unconstrained nonconvex optimization, lower-bound complexity theory is already well developed. A standard conclusion is that first-order methods require

$$\Omega(L\Delta\epsilon^{-2})$$

oracle calls to find a point x such that $\|\nabla f(x)\| \le \epsilon$, where L is a gradient Lipschitz constant and Δ is the initial objective gap. The underlying mechanism is typically captured through zero-chain hard instances, which restrict how first-order methods reveal coordinates [3, 4].

The constrained setting of (1) introduces a further layer of difficulty. In particular, two natural and distinct accuracy requirements arise. The first is *feasibility*, measured in this paper by producing

*Sun Yat-sen University, Guangzhou, China. (shiqk@mail2.sysu.edu.cn)
†Sun Yat-sen University, Guangzhou, China. (wangx936@mail.sysu.edu.cn)

a point with small constraint violation. The second is *stationarity*: reducing a constrained first-order stationarity residual, minimized over multipliers. An approximate KKT guarantee requires both feasibility and stationarity, but the two requirements capture different aspects of the problem and can therefore be studied separately at the lower-bound level. To formulate lower bounds for these two requirements, we also specify the information available to the method. We use a local oracle model: the algorithm accesses the problem only through an oracle that returns the objective value, gradient, constraint value, and constraint Jacobian at each queried point. It is not given projection or exact feasibility-restoration oracles for the original nonlinear constraints. Formal definitions of the problem class, algorithm class, oracle model, and complexity measures are deferred to Section 2.

Under the local first-order oracle model, we establish separate lower bounds for ϵ-feasibility and ϵ-stationarity (see Definitions 3 and 4), with different parameter dependences:

$$\begin{aligned} &\epsilon\text{-feasibility:} \quad \Omega\left(\frac{L_c\Delta_c}{\sigma^2} + \log\log\frac{\sigma^2}{L_c\epsilon}\right), \\ &\epsilon\text{-stationarity:} \quad \Omega\left(\left(L_f + \frac{GL_c}{\sigma}\right)\left(\Delta_f + \frac{G\Delta_c}{\sigma}\right)\epsilon^{-2}\right). \end{aligned}$$

Here L_f and L_c are the Lipschitz constants of ∇f and ∇c, respectively; G bounds the ℓ_∞-norm of the objective gradient; and σ lower bounds the smallest singular value of ∇c. The quantities Δ_f and Δ_c bound the initial objective gap and the initial constraint violation.

The feasibility lower bound is proved through two constraint-only constructions: one for reducing a large initial constraint violation and one for the high-accuracy phase near feasibility. The stationarity lower bound uses a different construction: a nonlinear equality-constrained hard instance whose minimized stationarity residual reduces to the gradient of a scalar reduced function via an exact elimination argument, thereby using the zero-chain lower bound in the constrained setting. To extend this lower bound to any deterministic first-order method, we use a reduction based on disjoint supports, which preserves the ℓ_∞-norm gradient bound. We also design two first-order algorithms: a damped Newton method that attains the feasibility lower bound up to constants, and an exact penalty method that attains the stationarity lower bound up to the dimension dependence stated in Section 4.4. The same separated analysis also gives a lower bound for finding approximate KKT points. Since an approximate KKT point must satisfy both feasibility and stationarity, its oracle complexity is bounded below by the maximum of the feasibility and stationarity lower bounds, with the corresponding tolerances. This consequence is recorded in Section 5.

We now briefly position these results relative to existing literature. Existing unconstrained lower bounds provide the conceptual foundation for our stationarity analysis, but they do not directly capture the role of nonlinear equality constraints, Jacobian regularity, or the complexity of approaching the feasible set. Existing first-order analyses for constrained nonconvex optimization mainly provide upper bounds for approximate stationary or KKT points. Without matching lower bounds, it remains unclear whether these rates are optimal, and the separate costs of feasibility and stationarity are not visible.

1.1 Related work

Lower bounds for unconstrained optimization. Oracle-complexity lower bounds for first-order methods originate in the information-based framework of Nemirovski and Yudin [12] and the optimality theory developed by Nesterov for smooth and nonsmooth convex minimization [13, 14]. In the nonconvex setting, Carmon et al. introduced zero-chain hard instances and established tight first-order lower bounds for finding first-order stationary points, including the classical $\Omega(L\Delta\epsilon^{-2})$

scaling under Lipschitz-continuous gradients, with further extensions to scenarios with higher-order smoothness and using stochastic oracles [3, 4, 1]. These works provide the main conceptual basis for our lower bound for stationarity. However, they are intrinsically unconstrained and therefore do not capture the role of nonlinear constraints, Jacobian regularity, or the complexity of approaching the feasible set.

Upper bounds for constrained nonconvex optimization. For constrained nonconvex problems, most first-order results take the form of upper bounds for approximate KKT or Fritz–John points obtained through penalty, augmented Lagrangian, or inexact proximal-point frameworks. Representative examples include improved inexact augmented Lagrangian and proximal-penalty schemes, which achieve complexities such as $\widetilde{O}(\epsilon^{-5/2})$, $\widetilde{O}(\epsilon^{-3})$, or $O(\epsilon^{-4})$ under different structural and regularity assumptions [9, 10, 8, 7, 16]. When equality constraints define a smooth manifold and geometric operations such as tangent projection and retraction are available, Riemannian first-order and trust-region methods recover the familiar $O(\epsilon^{-2})$ and $O(\epsilon^{-3})$ iteration guarantees for first- and second-order criticality, respectively [2]. Classical nonlinear-programming complexity analyses for barrier and trust-region methods also provide evaluation-complexity guarantees for approximate constrained critical points [5, 6]. In contrast, we study lower bounds that treat feasibility and stationarity separately under a strictly local oracle. This oracle excludes projection or exact feasibility-restoration oracles for the original nonlinear constraints.

Lower bounds for constrained nonconvex optimization. Compared with the unconstrained theory, lower bounds for constrained problems remain much less developed. For convex-concave bilinear saddle-point problems, and hence for affine-constrained smooth convex optimization as a special case, Ouyang and Xu showed that affine constraints fundamentally change the first-order complexity: the generic $O(1/t)$ rate cannot in general be accelerated to the unconstrained $O(1/t^2)$ rate, and even strong convexity does not restore unconstrained-style linear convergence [15]. More recently, Liu et al. established first-order lower bounds for composite nonconvex problems with affine equality constraints, including lower bounds with explicit dependence on the condition number of the affine constraint matrix [11]. Our results differ from this line in two respects. First, we study genuinely nonlinear equality constraints rather than affine ones. Second, we obtain a lower bound for approximate KKT points by analyzing the feasibility and stationarity requirements separately, rather than treating them only as a combined condition. To the best of our knowledge, this form of separated lower-bound characterization for smooth nonlinear constraints has not been established in the existing first-order literature.

1.2 Contributions

The contributions of this paper are summarized as follows.

(i) **Separated lower bounds.** We establish separate oracle-complexity lower bounds for feasibility and stationarity. The feasibility bound contains a global term depending on the initial infeasibility and a local high-accuracy term of order $\log\log(\frac{\sigma^2}{L_c\epsilon})$. The stationarity bound has ϵ^{-2}-dependence and is proved by reducing the minimized stationarity residual to the gradient of a scalar reduced function, thereby using the zero-chain lower bound in the constrained setting. To extend the lower bounds to any deterministic first-order method, we use a reduction based on disjoint supports that preserves the ℓ_∞-norm gradient bound.

(ii) **Tightness via two algorithms.** We design two first-order methods: a damped Newton

method attaining the feasibility lower bound up to constants and a prox-linear method attaining the stationarity lower bound up to the stated dimension dependence.

(iii) **KKT lower bound.** We obtain a lower bound for finding approximate KKT points by combining the separated feasibility and stationarity lower bounds.

2 Preliminaries

This section sets out the notation and oracle model used in the lower-bound arguments. We first define the problem class, the local oracle, deterministic first-order methods, and the auxiliary notion of zero-respecting sequences and methods. We then state the Jacobian regularity condition along the queries, followed by the two accuracy conditions and the associated oracle-complexity measures. Finally, we collect the chain-based ingredients and the reduction using disjoint supports used in the feasibility and stationarity lower-bound arguments.

2.1 Problem class

We consider equality-constrained problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{subject to} \qquad c(x)=0, \tag{2}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ and $c:\mathbb{R}^n\to\mathbb{R}^m$ are differentiable. Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm, and matrix norms are the induced operator norms. We write $\nabla c(x)$ for the Jacobian of c at x, and use $\sigma_{\min}(J)$ for the smallest singular value of a matrix J. The initial point is denoted by x_0.

Definition 1. For given finite parameters $G, L_f, L_c > 0$ and $\Delta_f, \Delta_c \geq 0$, let

$$\mathcal{P}_\infty(G, L_f, L_c, \Delta_f, \Delta_c)$$

denote the class of instances $P=(f,c,x_0)$ satisfying:

1. **Gradient bound:** $\|\nabla f(x)\|_\infty \leq G, \quad \forall x\in\mathbb{R}^n$.
2. **Smoothness:** $\|\nabla f(x)-\nabla f(y)\| \leq L_f\|x-y\|, \quad \|\nabla c(x)-\nabla c(y)\| \leq L_c\|x-y\|, \quad \forall x,y\in\mathbb{R}^n$.
3. **Initial objective gap and feasibility:** The feasible set $\mathcal{X}=\{x\in\mathbb{R}^n : c(x)=0\}$ is nonempty, its feasible infimum $f^\star=\inf_{x\in\mathcal{X}} f(x)$ is finite, and $f(x_0)-f^\star\leq\Delta_f, \|c(x_0)\|\leq\Delta_c$.

Classical unconstrained lower bounds for approximate stationarity are parameterized by smoothness and an initial objective gap, without a separate global gradient bound. In the constrained setting, we also track the objective-gradient bound, because the multiplier-minimized residual combines objective-gradient and Jacobian information. The zero-chain functions used in proving unconstrained stationary-point lower bounds [3, 4] have a uniformly bounded ℓ_∞-norm gradient, whereas their Euclidean gradient norm may grow with the dimension. We therefore use the ℓ_∞-norm bound $\|\nabla f\|_\infty \leq G$. However, this ℓ_∞-norm choice is not orthogonally invariant: the usual adversarial rotation argument can enlarge ℓ_∞ norms of gradients. This motivates the reduction using disjoint supports developed in Proposition 2, which extends the zero-chain lower bounds to any deterministic first-order methods while preserving the ℓ_∞-norm gradient bound.

2.2 Local oracle and deterministic first-order methods

We use a local oracle for differentiable equality-constrained instances $P \in \mathcal{P}_\infty$. A query at $x \in \mathbb{R}^n$ returns

$$\mathcal{O}_P(x) = \big(f(x),\, \nabla f(x),\, c(x),\, \nabla c(x)\big). \tag{3}$$

The oracle is local and first-order, and returns only the quantities in (3).

Given an instance $P = (f, c, x_0)$ and its oracle $\mathcal{O}_P$, the method A starts from x_0 and generates query points $\{x_t\}_{t\geq 0}$. We say that A is a deterministic first-order method if, for every $t \geq 0$, there exists a mapping A_t such that

$$x_{t+1} = \mathsf{A}_t\big(x_0, \mathcal{O}_P(x_0), x_1, \mathcal{O}_P(x_1), \ldots, x_t, \mathcal{O}_P(x_t)\big). \tag{4}$$

At step t, the next query is determined only by the problem class parameters $(G, L_f, L_c, \Delta_f, \Delta_c)$, the accuracy ϵ, the iterate sequence $\{x_i\}_{i=0}^t$, and the oracle history $\{\mathcal{O}_P(x_i)\}_{i=0}^t$. No other information about f or c is available to the method. In particular, apart from $\mathcal{O}_P$, the method is not given projection or exact feasibility-restoration oracles for the original nonlinear constraints. We denote by $\mathcal{A}$ the resulting class of deterministic first-order methods.

Zero-respecting sequences and methods. To state the zero-respecting condition, we use the following auxiliary notion for sequences. Informally, each new iterate may be nonzero only on coordinates already present in the initial point or exposed by past first-order oracle information. Formally[1], given an instance $P = (f, c, x_0)$, a sequence $\{x_t\}_{t\geq 0}$ is called *zero-respecting* if, for every $t \geq 0$,

$$\operatorname{supp}(x_{t+1}) \subseteq \operatorname{supp}(x_0) \cup \bigcup_{s\leq t} \Big(\operatorname{supp}(\nabla f(x_s)) \cup \operatorname{supp}\big(\operatorname{Range}(\nabla c(x_s)^\top)\big)\Big). \tag{5}$$

We say that a method is zero-respecting on P only when its generated sequence satisfies condition (5). The passage to any deterministic methods in $\mathcal{A}$ is handled by the reduction in Proposition 2.

2.3 Iteration-wise regularity condition

The lower-bound statements below use Jacobian regularity along the algorithm's queries. This condition is not part of the base class $\mathcal{P}_\infty$, because it is imposed relative to a fixed algorithm and to a given number of oracle calls. Instead, it is stated as follows.

Definition 2 (Iteration-wise σ-regularity)**.** Fix a deterministic first-order algorithm A, an integer $T \geq 1$, and an instance $P = (f, c, x_0) \in \mathcal{P}_\infty(G, L_f, L_c, \Delta_f, \Delta_c)$. Run A on P for T oracle calls, producing query points $x_0, \ldots, x_T$. We say that P satisfies *iteration-wise σ-regularity* if

$$\sigma_{\min}(\nabla c(x_t)) \geq \sigma, \qquad t = 0, 1, \ldots, T.$$

Thus the condition requires the constraint Jacobian to have full row rank at each queried point, with its smallest singular value uniformly bounded below by σ. In the lower-bound statements below, this condition is imposed only along the query points generated within the given number of oracle calls.

[1]Here, we use notations: for $v \in \mathbb{R}^n$, let $\operatorname{supp}(v) = \{i : v_i \neq 0\}$; for a subspace $V \subseteq \mathbb{R}^n$, let $\operatorname{supp}(V) = \bigcup_{v\in V} \operatorname{supp}(v)$.

2.4 Accuracy conditions and oracle-complexity measures

With the oracle model, algorithm class, and iteration-wise regularity fixed, we now define the two accuracy conditions and the corresponding oracle-complexity measures. These conditions separate feasibility from the multiplier-minimized first-order stationarity residual. For each condition, the corresponding oracle-complexity measure counts the local oracle calls required to find a point satisfying the given condition.

Definition 3 (ϵ-feasibility)**.** Given $\epsilon > 0$, a point $x \in \mathbb{R}^n$ is called *ϵ-feasible* if $\|c(x)\| \le \epsilon$.

Definition 4 (ϵ-stationarity)**.** For $x \in \mathbb{R}^n$, define the multiplier-minimized stationarity residual

$$\mathcal{R}_{\mathrm{stat}}(x) = \min_{\lambda \in \mathbb{R}^m} \|\nabla f(x) + \nabla c(x)^\top \lambda\|. \tag{6}$$

Given $\epsilon > 0$, a point $x \in \mathbb{R}^n$ is called *ϵ-stationary* if $\mathcal{R}_{\mathrm{stat}}(x) \le \epsilon$.

To turn these conditions into oracle-complexity measures, we evaluate them at the query points generated by the algorithm. For $\mathsf{A} \in \mathcal{A}$ and $P = (f, c, x_0)$, write x_t for the t-th query point generated by A on P. We define

$$T_\epsilon^{\mathrm{feas}}(\mathsf{A}, P) = \inf\{t \ge 0 : \|c(x_t)\| \le \epsilon\}, \tag{7}$$

$$T_\epsilon^{\mathrm{stat}}(\mathsf{A}, P) = \inf\{t \ge 0 : \mathcal{R}_{\mathrm{stat}}(x_t) \le \epsilon\}, \tag{8}$$

with the convention that $\inf \emptyset = +\infty$, and refer to these as the feasibility and stationarity complexities of A on P. With this setup, the worst-case deterministic oracle complexities for $\mathcal{A}$ over $\mathcal{P}_\infty$ are

$$\mathcal{T}_\epsilon^{\mathrm{feas}}(\mathcal{A}, \mathcal{P}_\infty) = \inf_{\mathsf{A} \in \mathcal{A}} \sup_{P \in \mathcal{P}_\infty} T_\epsilon^{\mathrm{feas}}(\mathsf{A}, P), \tag{9}$$

$$\mathcal{T}_\epsilon^{\mathrm{stat}}(\mathcal{A}, \mathcal{P}_\infty) = \inf_{\mathsf{A} \in \mathcal{A}} \sup_{P \in \mathcal{P}_\infty} T_\epsilon^{\mathrm{stat}}(\mathsf{A}, P). \tag{10}$$

For KKT accuracy, feasibility and multiplier-minimized stationarity must hold at the same query point. Since these are two distinct residual requirements, we use two accuracy parameters, one for each residual, and count the first query index at which both requirements are met. For $\epsilon_c, \epsilon_s > 0$, define

$$T^{\mathrm{KKT}}(\epsilon_c, \epsilon_s; \mathsf{A}, P) = \inf\{t \ge 0 : \|c(x_t)\| \le \epsilon_c \text{ and } \mathcal{R}_{\mathrm{stat}}(x_t) \le \epsilon_s\}. \tag{11}$$

At the class level, set

$$\mathcal{T}^{\mathrm{KKT}}(\epsilon_c, \epsilon_s; \mathcal{A}, \mathcal{P}_\infty) = \inf_{\mathsf{A} \in \mathcal{A}} \sup_{P \in \mathcal{P}_\infty} T^{\mathrm{KKT}}(\epsilon_c, \epsilon_s; \mathsf{A}, P). \tag{12}$$

The single-accuracy notation is the diagonal case:

$$T_\epsilon^{\mathrm{KKT}}(\mathsf{A}, P) = T^{\mathrm{KKT}}(\epsilon, \epsilon; \mathsf{A}, P), \qquad \mathcal{T}_\epsilon^{\mathrm{KKT}}(\mathcal{A}, \mathcal{P}_\infty) = \mathcal{T}^{\mathrm{KKT}}(\epsilon, \epsilon; \mathcal{A}, \mathcal{P}_\infty).$$

2.5 Chain constructions and reduction using disjoint supports

We record the auxiliary zero-chain facts and the deterministic reduction used in the lower-bound proofs. The zero-chain φ is used directly in the feasibility construction of Section 3; a scaled zero-chain is used in the stationarity construction of Section 4. The reduction using disjoint supports then transfers lower bounds for zero-respecting sequences to any deterministic first-order methods while preserving the ℓ_∞-norm objective-gradient bound.

Proposition 1. *There exist universal constants $\ell, G_0 > 0$ such that, for any $d \geq 1$, there exists a twice continuously differentiable function $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfying:*

(i) $\varphi(0) = 0$, $\left\|\nabla^2 \varphi(z)\right\| \leq \ell$, *and* $\|\nabla \varphi(z)\|_\infty \leq G_0$ *for all* $z \in \mathbb{R}^d$.

(ii) For all $z \in \mathbb{R}^d$ *such that* $|z_d| \leq \frac{1}{2}$, *the gradient norm satisfies* $\|\nabla \varphi(z)\| \geq 1$.

(iii) If for some non-negative $k < d$, $z_i = 0$ *for all* $i > k$, *then* $\nabla_i \varphi(z) = 0$ *for all* $i > k + 1$.

(iv) For any non-negative $k \leq d$, *every* z *with* $z_i = 0$ *for all* $i > k$ *satisfies* $\varphi(z) \geq -12k$, *and there exists such a point* z *that* $\varphi(z) \leq -2k$.

Moreover, with $\Delta = \varphi(0) - \inf_z \varphi(z)$, *there exists a constant* $c_0 > 0$ *such that for any* $\epsilon > 0$ *and any dimension* $d \geq \lceil c_0 \frac{\ell \Delta}{\epsilon^2} \rceil$, *one can find a scaled zero-chain* $\varphi : \mathbb{R}^d \to \mathbb{R}$ *from this family, such that any zero-respecting method applied to* φ *and initialized at* 0 *requires at least*

$$T \geq c_0 \frac{\ell \Delta}{\epsilon^2}$$

oracle calls before querying a point z^* *satisfying* $\|\nabla \varphi(z^*)\| \leq \epsilon$.

The function φ in Proposition 1 is obtained by shifting the unscaled zero-chain $\bar{f}_d$ from [3]. Here $\bar{f}_d$ is the chain of dimension d, and we set $\varphi(z) = \bar{f}_d(z) - \bar{f}_d(0)$. This constant shift gives $\varphi(0) = 0$ and leaves the gradient, Hessian, and zero-chain support condition unchanged. The ℓ_∞-norm gradient bound in item (i) follows from the uniform bound on each one-dimensional component of this zero-chain. Item (ii) is the large gradient property applied when the last chain coordinate remains inactive, and item (iii) is the first-order zero-chain support condition written in coordinate form. Item (iv) records the function-value bounds on points supported on the first k chain coordinates; the case $k = d$ gives the global gap bound used when the chain is scaled in Section 4. The final lower-bound consequence is the standard scaled zero-chain lower bound for zero-respecting first-order methods [3]; the dimension is chosen as part of this consequence.

We next state the reduction using disjoint supports that transfers the zero-chain lower bounds from zero-respecting sequences to any deterministic first-order methods. The standard adversarial rotation argument preserves Euclidean norms, but it need not preserve the ℓ_∞-norm objective-gradient bound in $\mathcal{P}_\infty$, since one coordinate of Uv may contain contributions from many entries of v. The reduction below avoids this by choosing the columns of U in disjoint coordinate blocks.

Proposition 2. *Let* $\varphi : \mathbb{R}^d \to \mathbb{R}$ *be the zero-chain from Proposition 1. Fix* $\mathsf{A} \in \mathcal{A}$, $T < d$, *and set* $n = d(T + 2)$. *Then there exists a matrix* $U \in \mathbb{R}^{n \times d}$ *with* $U^\top U = I$, *which may depend on* φ, A, *and* T, *such that, for the lifted function* $\varphi_U : \mathbb{R}^n \to \mathbb{R}$, $\varphi_U(x) = \varphi(U^\top x)$, *the query points* x_t *generated by* A *satisfy*

$$\operatorname{supp}(U^\top x_t) \subseteq \{1, \ldots, t\}, \qquad 0 \leq t \leq T. \tag{13}$$

It also holds that for any $x \in \mathbb{R}^n$,

$$\|\nabla \varphi_U(x)\|_\infty \leq \|\nabla \varphi(U^\top x)\|_\infty, \qquad \inf_{x \in \mathbb{R}^n} \varphi_U(x) = \inf_{z \in \mathbb{R}^d} \varphi(z).$$

Moreover, when φ *is chosen as in the final consequence of Proposition 1, with* $\Delta = \varphi(0) - \inf_z \varphi(z) = \varphi_U(0) - \inf_x \varphi_U(x)$, *any deterministic first-order method applied to* φ_U *requires at least*

$$T \geq c_0 \frac{\ell \Delta}{\epsilon^2}$$

oracle calls before querying a point x^* *satisfying* $\|\nabla \varphi_U(x^*)\| \leq \epsilon$.

Proof. Split $\mathbb{R}^n$ into disjoint coordinate blocks $B_1, \dots, B_d$, each of size $T+2$. We construct unit vectors q_j with $\operatorname{supp}(q_j) \subseteq B_j$, and then set $U = [q_1 \cdots q_d]$. At query t, after the algorithm has produced x_t and before the oracle response at x_t is fixed, choose q_{t+1} inside the block B_{t+1} and orthogonal to the block restrictions of $x_0, \dots, x_t$. This is possible because at most $t+1 \le T+1$ linear constraints are imposed in a space of dimension $T+2$. After the oracle response at x_T has been fixed, choose all remaining vectors q_j in their own blocks and orthogonal to the block restrictions of $x_0, \dots, x_T$.

The orthogonality conditions give $q_j^\top x_t = 0$ for every $j > t$ and every $t \le T$; hence $z_t = U^\top x_t$ satisfies (13). By Proposition 1(iii), the first-order response at x_t depends only on $q_1, \dots, q_{t+1}$. Later choices of q_j, $j > t+1$, do not change the previous projected points or first-order responses. Since A is deterministic, the completed matrix U generates the same sequence on φ_U.

The vectors q_j have disjoint supports and unit Euclidean norm, so $U^\top U = I$. For any $v \in \mathbb{R}^d$, at most one term in $Uv = \sum_j v_j q_j$ contributes to the i-th coordinate, for each $i = 1, \dots, n$. Hence

$$|(Uv)_i| \le \max_j |v_j|,$$

which proves $\|Uv\|_\infty \le \|v\|_\infty$. Finally, for any $x \in \mathbb{R}^n$, the chain rule gives $\nabla\varphi_U(x) = U\nabla\varphi(U^\top x)$, and therefore $\|\nabla\varphi_U(x)\| = \|\nabla\varphi(U^\top x)\|$ because $U^\top U = I$. The same disjoint-support bound gives

$$\|\nabla\varphi_U(x)\|_\infty \le \|\nabla\varphi(U^\top x)\|_\infty.$$

For any $x \in \mathbb{R}^n$, $\varphi_U(x) = \varphi(U^\top x)$, while for any $z \in \mathbb{R}^d$, $\varphi_U(Uz) = \varphi(z)$ since $U^\top U = I$. Hence the two infima are equal.

The final claim follows by applying the same construction to the scaled zero-chain φ chosen in Proposition 1. Scaling preserves the zero-chain support condition, so the projected sequence $z_t = U^\top x_t$ is zero-respecting for φ. If a deterministic method queried x_t with $\|\nabla\varphi_U(x_t)\| \le \epsilon$ before $\frac{c_0 \ell \Delta}{\epsilon^2}$ oracle calls, then $\nabla\varphi_U(x_t) = U\nabla\varphi(z_t)$ and $U^\top U = I$ would give $\|\nabla\varphi(z_t)\| \le \epsilon$, contradicting the lower-bound consequence in Proposition 1. □

3 Lower bound for feasibility

In this section we isolate feasibility by setting $f \equiv 0$. Thus only the constraint values and Jacobians carry relevant oracle information. The feasibility lower bound separates into two parts. The first is a global phase, where reducing a large initial constraint violation costs $\Omega(L_c \Delta_c \sigma^{-2})$ oracle calls. The second is a local high-accuracy part, where a scalar resisting construction yields the additional lower bound $\Omega(\log\log(\frac{\sigma^2}{L_c \epsilon}))$ for reaching ϵ-feasibility.

The main difficulty in the global phase is compatibility with iteration-wise Jacobian regularity. A standard zero-chain has the desired support condition, but after sufficiently many coordinates are activated it may enter regions where the chain gradient vanishes; the induced constraint Jacobian would then fail the required lower bound. We therefore need a construction that keeps the Jacobian bounded below along the actual queried iteration while retaining the value lower bound from the zero-chain.

3.1 Feasibility lower bound I: Global reduction

To maintain iteration-wise regularity in the global construction, we choose the dimension d larger than the number of oracle calls, i.e. $d > T$, under consideration. The zero-chain is scaled so that its Hessian bound matches L_c and its large gradient property gives the Jacobian lower bound σ.

By Proposition 2, any deterministic first-order method has projected queries supported only on the first k zero-chain coordinates at the k-th query. Thus the last chain coordinate remains inactive throughout the first T queries, and the induced constraint Jacobian stays uniformly bounded below. The same support restriction gives the value lower bound from the zero-chain: at the k-th projected query, the constraint value is bounded below by $\Delta_c - \mathcal{O}(\frac{k\sigma^2}{L_c})$. Consequently, reaching an ϵ-feasible point from initial violation Δ_c requires $\Omega(L_c\Delta_c\sigma^{-2})$ oracle calls.

Lemma 1. *Fix $\sigma, L_c > 0$, $\Delta_c > 0$, and $\epsilon \in (0, \frac{\Delta_c}{2})$. For every deterministic first-order method $\mathsf{A} \in \mathcal{A}$ and every integer $T \geq 0$, there exists a scalar-constraint instance $P \in \mathcal{P}_\infty$, with constraint function c, such that, if $\{x_k\}_{k\geq 0}$ is the sequence generated by A on P, then P is σ-regular along these queries, i.e., $\sigma_{\min}(\nabla c(x_k)) \geq \sigma$ for all $0 \leq k \leq T$, and every $k \leq T$ satisfying $|c(x_k)| \leq \epsilon$ obeys*

$$k \geq \frac{L_c(\Delta_c - \epsilon)}{12\sigma^2\ell},$$

where ℓ is introduced in Proposition 1. In particular, since $\epsilon < \frac{\Delta_c}{2}$, the global feasibility reduction requires $\Omega(L_c\Delta_c\sigma^{-2})$ oracle calls.

Proof. We first construct a scaled zero-chain constraint and prove the required results in the zero-respecting chain setting, and then apply the reduction using disjoint supports reduction to any deterministic first-order method A. Set

$$R = \frac{\sigma\ell}{L_c}, \qquad \rho = \frac{\sigma^2\ell}{L_c}, \qquad d = T + \left\lceil \frac{\Delta_c}{2\rho} \right\rceil + 1.$$

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be the zero-chain from Proposition 1, and take $f \equiv 0$, initial point 0, and scalar constraint $c : \mathbb{R}^d \to \mathbb{R}$ given by

$$c(y) = \Delta_c + \rho\varphi\left(\frac{y}{R}\right).$$

We now verify the smoothness and feasibility properties of this construction. By construction, the initial constraint violation is exactly $c(0) = \Delta_c$. From Proposition 1(i), using the chain rule and the Hessian bound $\|\nabla^2\varphi(z)\| \leq \ell$,

$$\|\nabla^2 c(y)\| \leq \frac{\rho}{R^2}\ell = \frac{\frac{\sigma^2\ell}{L_c}}{(\frac{\sigma\ell}{L_c})^2}\ell = L_c, \qquad \forall y \in \mathbb{R}^d.$$

By activating the first $d-1$ coordinates, Proposition 1(iv) gives a point $z \in \mathbb{R}^d$, supported on $\{1, \ldots, d-1\}$, such that $\varphi(z) \leq -2(d-1)$. Hence $y = Rz$ satisfies

$$c(y) \leq \Delta_c - 2\rho(d-1) \leq 0.$$

Since $c(0) = \Delta_c > 0$, continuity along the segment $[0, y]$ gives a point $y^\star$, still supported on $\{1, \ldots, d-1\}$, with $c(y^\star) = 0$. Hence this construction has a nonempty feasible set.

Next, consider any sequence $\{y^k\}_{k=0}^T$ generated by a zero-respecting method on this chain construction, starting from 0. By Proposition 1(iii), induction gives $\mathrm{supp}(y^k) \subseteq \{1, \ldots, k\}$ for all $0 \leq k \leq T$. Since $k \leq T < d$, the last coordinate of y^k is zero. With $z^k = \frac{y^k}{R}$, we have $|z_d^k| = 0 \leq \frac{1}{2}$. By Proposition 1(ii), $\|\nabla\varphi(z^k)\| \geq 1$, and therefore

$$\|\nabla c(y^k)\| = \frac{\rho}{R}\left\|\nabla\varphi(z^k)\right\| \geq \frac{\frac{\sigma^2\ell}{L_c}}{\frac{\sigma\ell}{L_c}} = \sigma.$$

This proves the required σ-regularity condition along every such sequence.

The same support inclusion and Proposition 1(iv) yield $\varphi(z^k) \geq -12k$. Substituting this into the constraint definition gives

$$c(y^k) \geq \Delta_c - 12k\rho.$$

If $|c(y^k)| \leq \epsilon$, then $c(y^k) \leq \epsilon$, and hence

$$k \geq \frac{\Delta_c - \epsilon}{12\rho} = \frac{L_c(\Delta_c - \epsilon)}{12\sigma^2\ell}.$$

It remains to transfer this lower bound to any deterministic first-order method A. The reduction in Proposition 2 applies to c, since scaling and adding a constant preserve the first-order support condition. Applying this reduction to A with $n = d(T+2)$, we obtain a matrix $U \in \mathbb{R}^{n\times d}$ with disjoint supports and $U^\top U = I$. For the lifted instance with initial point $x_0 = 0$,

$$f_U(x) \equiv 0, \qquad c_U(x) = c(U^\top x),$$

the sequence $y^k = U^\top x_k$ satisfies

$$\operatorname{supp}(y^k) \subseteq \{1, \ldots, k\}, \qquad 0 \leq k \leq T.$$

The lift preserves membership in $\mathcal{P}_\infty$: the objective is still zero, $c_U(0) = \Delta_c$, $Uy^\star$ is feasible, and

$$\begin{aligned}\|\nabla c_U(x) - \nabla c_U(x')\| &= \|U(\nabla c(U^\top x) - \nabla c(U^\top x'))\| \\ &\leq \|\nabla c(U^\top x) - \nabla c(U^\top x')\| \\ &\leq L_c\|U^\top(x - x')\| \leq L_c\|x - x'\|.\end{aligned}$$

Moreover, for every $k \leq T$,

$$|c_U(x_k)| = |c(y^k)|, \qquad \|\nabla c_U(x_k)\| = \|U\nabla c(y^k)\| = \|\nabla c(y^k)\|.$$

Since the constraint is scalar, $\sigma_{\min}(\nabla c_U(x_k)) = \|\nabla c_U(x_k)\|$, so the lifted instance is σ-regular along $x_0, \ldots, x_T$. The identity $|c_U(x_k)| = |c(y^k)|$ transfers the lower bound to the iterates x_k. Taking $(f_U, c_U, 0)$ as the constructed instance proves the claim.

□

3.2 Feasibility lower bound II: Local high-accuracy

The lemma in above subsection gives the global lower bound for reducing a large initial constraint violation. We now consider the local phase $|c(x)| \lesssim \frac{\sigma^2}{L_c}$, where a scalar construction forces only quadratic residual reduction per query.

Lemma 2. *Fix $\sigma, L_c > 0$, set $\Delta_c = \frac{\sigma^2}{L_c}$, and let $\epsilon \in (0, \Delta_c)$. For every deterministic first-order method and every integer $T \geq 0$, there exists a one-dimensional instance (f, c, x_0), with $x_0 = 0$ and $f \equiv 0$, such that $c(0) = \Delta_c$, $c \in C^{1,1}(\mathbb{R})$, $|c'(x) - c'(y)| \leq L_c|x - y|$ for all $x, y \in \mathbb{R}$, and $c'(x) \leq -\sigma$ for all $x \in \mathbb{R}$. Then any queried iterate x satisfying $|c(x)| \leq \epsilon$ requires*

$$\Omega\left(\log\log\frac{\sigma^2}{L_c\epsilon}\right)$$

first-order oracle calls.

Proof. It suffices to prove the claim for pure feasibility instances, so $f \equiv 0$ throughout. For any deterministic first-order method A and any given number T of oracle calls, we construct a resisting oracle and realize its answers by a single scalar constraint. Since the construction fixes the constraint on successive intervals, we first record the cubic interpolation estimate used to connect these fixed pieces. For $u < v$, set $m = v - u$. Given endpoint values C_u, C_v and endpoint derivatives -2σ, define

$$H(s) = 3s^2 - 2s^3, \qquad \eta = C_v - C_u + 2\sigma m, \qquad p(x) = C_u - 2\sigma(x-u) + \eta H\left(\frac{x-u}{m}\right).$$

Then $p(u) = C_u$, $p(v) = C_v$, and $p'(u) = p'(v) = -2\sigma$. Since $0 \le H' \le \frac{3}{2}$ and $|H''| \le 6$ on $[0,1]$,

$$|p''(x)| \le \frac{6|\eta|}{m^2}, \qquad p'(x) \le -2\sigma + \frac{3|\eta|}{2m}.$$

Thus, whenever $|\eta| \le \frac{L_c m^2}{48}$ and $m \le \frac{\sigma}{L_c}$, the interpolant satisfies $|p''| \le \frac{L_c}{8}$ and $p' \le -\sigma$.

At each stage the oracle maintains an active bracket $[a, b]$ and a residual level $r > 0$:

$$c(a) = r, \qquad c(b) = -r, \qquad c'(a) = c'(b) = -2\sigma, \qquad b - a = \frac{r}{\sigma}. \tag{14}$$

Along with (14), we maintain a fixed-region condition: outside the current open bracket, the function has already been fixed by the exterior rays and by pieces created in previous updates, and every point in that region has residual at least r. Initially $r_0 = \Delta_c = \frac{\sigma^2}{L_c}$ and $[a_0, b_0] = [0, \frac{\sigma}{L_c}]$. The oracle returns $c(0) = r_0$, $c'(0) = -2\sigma$, and fixes the exterior affine rays

$$c(x) = r_0 - 2\sigma x \quad (x \le 0), \qquad c(x) = -r_0 - 2\sigma(x - b_0) \quad (x \ge b_0).$$

We now describe the update from a bracket $[a, b]$ satisfying (14) and the fixed-region condition above. Queries in $\mathbb{R} \setminus (a, b)$, including repeated and endpoint queries, are answered by the fixed function and have residual at least r. For an effective query $x \in (a, b)$, let

$$d = x - a, \qquad e = b - x, \qquad \delta = \min\{d, e\},$$

so that $d + e = \frac{r}{\sigma}$. Define

$$A = r - 2\sigma d = -r + 2\sigma e = \sigma(e - d), \qquad \theta = \frac{L_c \delta^2}{384}.$$

The oracle returns $c'(x) = -2\sigma$. For the function value, write $y = c(x)$ and set

$$y = \begin{cases} A, & |A| \ge \frac{\theta}{2}, \\ A - \theta, & |A| < \frac{\theta}{2}. \end{cases}$$

This value satisfies $y \ne 0$ and $|y| < r$. If $y = A$, then $|y| = \sigma|e - d| < r$. Otherwise,

$$|y| < \frac{3}{2}\theta = \frac{L_c \delta^2}{256} \le \frac{L_c r^2}{1024\sigma^2} \le \frac{r}{1024},$$

where $\delta \le \frac{r}{2\sigma}$ and $r \le r_0 = \frac{\sigma^2}{L_c}$.

Let $r_+ = |y|$. If $y > 0$, set $[a_+, b_+] = [x, x + \frac{y}{\sigma}]$. This case can occur only when $y = A$; hence $A > 0$, $e > d$, and

$$x < x + \frac{y}{\sigma} = x + \frac{A}{\sigma} = b - d < b.$$

If $y < 0$, set $[a_+, b_+] = [x + \frac{y}{\sigma}, x]$. For $y = A$, the left endpoint is $a + e > a$. For $y = A - \theta$, the left endpoint is $a + e - \frac{\theta}{\sigma}$. Since $\delta \le \frac{r}{2\sigma} \le \frac{\sigma}{2L_c}$, this point is also greater than a, because

$$\frac{\theta}{\sigma} = \frac{L_c\delta^2}{384\sigma} \le \frac{\delta}{384} \le \frac{e}{384} < e.$$

Thus $a < x + \frac{y}{\sigma} < x$, since $y < 0$. Hence $[a_+, b_+] \subset (a, b)$ in all cases, and the endpoint data are chosen so that (14) holds with r_+.

We next define c on the two intervals in $[a, b] \setminus [a_+, b_+]$. If $y = A$, both intervals are affine with slope -2σ. The only nonlinear case is $y = A - \theta$, where $y < 0$, $a_+ = a + e - \frac{\theta}{\sigma}$, and $b_+ = x$. The two intervals then have lengths $e - \frac{\theta}{\sigma}$ and e. On the left interval, $C_u = r$, $C_v = r_+ = \theta - A$, and $m = e - \frac{\theta}{\sigma}$, so

$$\eta = (\theta - A) - r + 2\sigma\left(e - \frac{\theta}{\sigma}\right) = -\theta.$$

On the right interval, $C_u = y = A - \theta$, $C_v = -r$, and $m = e$, so

$$\eta = -r - (A - \theta) + 2\sigma e = \theta.$$

Thus $|\eta| = \theta$ on both nonlinear pieces. Since $\frac{\theta}{\sigma} \le \frac{\delta}{384} \le \frac{e}{384}$, the length m of each nonlinear piece satisfies

$$\frac{\delta}{2} \le m \le b - a \le \frac{\sigma}{L_c}.$$

Together with $|\eta| = \theta$, this gives $|\eta| \le \frac{L_c m^2}{48}$, so the interpolation estimate yields $|c''| \le \frac{L_c}{8}$ and $c' \le -\sigma$ there. Moreover, $c(a_+) = r_+$ and $c(b_+) = -r_+$. Together with (14), $r > r_+$, and monotonicity of the new pieces, this ensures $|c| \ge r_+$ on the fixed region after the update.

We next bound the decrease of the active level in one effective query, where the new level is $r_+ = |y|$. If $\delta \le \frac{r}{4\sigma}$, then $\frac{\theta}{2} \le \frac{r}{12288}$, so the oracle uses $y = A$, and hence

$$r_+ = |A| = \sigma|e - d| = \sigma((d + e) - 2\delta) = r - 2\sigma\delta \ge \frac{r}{2} \ge \frac{L_c}{12288\sigma^2}r^2,$$

using $r \le r_0 = \frac{\sigma^2}{L_c}$. If $\delta > \frac{r}{4\sigma}$, then by construction either $|y| = |A| \ge \frac{\theta}{2}$ or $|y| = |A - \theta| > \frac{\theta}{2}$, and therefore

$$r_+ \ge \frac{\theta}{2} = \frac{L_c\delta^2}{768} > \frac{L_c r^2}{12288\sigma^2}.$$

Combining the two cases, every effective query satisfies

$$r_+ \ge \frac{L_c}{12288\sigma^2}r^2. \tag{15}$$

Run the resisting oracle for the first T oracle calls of A. This fixes finitely many affine or Hermite pieces outside a nested active bracket. After the last call, complete the remaining bracket by

$$c(x) = r - 2\sigma(x - a), \qquad x \in [a, b].$$

Together with the exterior affine rays, these pieces define a global C^1 function: adjacent pieces have matching endpoint values and derivative -2σ. On Hermite pieces, c' is $\frac{L_c}{8}$-Lipschitz, and on affine pieces it is constant; hence c' is globally L_c-Lipschitz. Also $c' \le -\sigma$, $c(0) = \Delta_c$, and the final bracket contains a zero of c. Thus the realized instance is feasible and satisfies the required regularity conditions. Since the realized function agrees with all oracle answers, determinism forces A to make the same first T oracle calls.

Let r_k be the active level after k effective calls. From (15) and $r_0 = \frac{\sigma^2}{L_c}$,

$$r_k \geq \frac{\sigma^2}{L_c} 12288^{1-2^k}.$$

If an ϵ-feasible point is queried within the first $N \leq T$ oracle calls, then some $k \leq N$ satisfies $r_k \leq \epsilon$: this follows from the fixed-region condition for non-effective calls and from the definition $r_+ = |y|$ for effective calls. Therefore

$$12288^{1-2^k} \leq \frac{L_c \epsilon}{\sigma^2},$$

and hence

$$k \geq \log_2 \left(1 + \frac{\log(\frac{\sigma^2}{L_c \epsilon})}{\log 12288} \right).$$

Since $N \geq k$, the claimed $\Omega(\log\log(\frac{\sigma^2}{L_c \epsilon}))$ lower bound follows. □

3.3 Feasibility lower bound and tightness

Combining the global reduction above with the local high-accuracy construction gives the full feasibility lower bound. We then show that this scaling is tight up to absolute constants.

Theorem 1. *Fix $\sigma, L_c > 0$, $\Delta_c \geq \frac{\sigma^2}{L_c}$, and $\epsilon \in (0, \min\{\frac{\Delta_c}{2}, \frac{\sigma^2}{L_c}\})$. There exists a constant $c > 0$ such that, for any deterministic first-order method $\mathsf{A} \in \mathcal{A}$ and any T satisfying*

$$T \leq c \left(\frac{L_c \Delta_c}{\sigma^2} + \log\log \frac{\sigma^2}{L_c \epsilon} \right),$$

there exists a scalar constraint instance $P = (f, c, x_0) \in \mathcal{P}_\infty$, with $f \equiv 0$ and $x_0 = 0$, such that P is σ-regular along the first T queries of A, and no point among these queries satisfies $|c(x)| \leq \epsilon$.

Proof. For any fixed method A, Lemma 1 gives a construction with global-phase length $\Omega(\frac{L_c \Delta_c}{\sigma^2})$. The local-phase construction in Lemma 2 gives length $\Omega(\log\log(\frac{\sigma^2}{L_c \epsilon}))$. Both constructions satisfy the regularity condition along the queried points. Choosing the harder of the two instances gives the maximum of the two lower bounds, and $\max\{a, b\} \geq \frac{a+b}{2}$ gives the claimed scaling. □

To show that the lower bound in Theorem 1 is tight, we analyze the complexity of a residual-based damped Newton method for solving the nonlinear system $c(x) = 0$.

At each iteration k, the method computes the minimum-norm Newton direction using the Moore-Penrose pseudoinverse of the constraint Jacobian:

$$p_k = -\nabla c(x_k)^\dagger c(x_k). \tag{16}$$

The iterate is then updated via

$$x_{k+1} = x_k + \alpha_k p_k, \tag{17}$$

where the damping parameter is chosen adaptively based on the current residual norm:

$$\alpha_k = \min\left\{1,\ \frac{\sigma^2}{L_c \|c(x_k)\|}\right\}. \tag{18}$$

Note that $p_k \in \mathrm{Range}(\nabla c(x_k)^\top)$. This ensures that the sequence $\{x_k\}$ is generated using only linear combinations of the historical constraint gradients, strictly adhering to the first-order algorithm class defined in (4).

We now show that this method attains the lower bounds up to absolute constants.

Theorem 2. *Given an initial point x_0 with $\|c(x_0)\| = \Delta_c$ and an accuracy parameter $\epsilon \in (0, \frac{\sigma^2}{2L_c})$. Suppose the constraint function c has L_c-Lipschitz continuous Jacobians, and $\sigma_{\min}(\nabla c(x_k)) \geq \sigma > 0$ for any k. Then the damped Newton method (16)–(18) produces an ϵ-feasible point in at most*

$$\left\lceil \frac{2L_c\Delta_c}{\sigma^2} \right\rceil + \left\lceil \log_2 \log_2 \left(\frac{2\sigma^2}{L_c\epsilon} \right) \right\rceil = O\left(\frac{L_c\Delta_c}{\sigma^2} + \log\log \frac{\sigma^2}{L_c\epsilon} \right)$$

iterations.

Proof. Let $r_k = \|c(x_k)\|$. By the properties of the pseudoinverse, $\nabla c(x_k)p_k = -c(x_k)$. The regularity assumption implies that $\left\|\nabla c(x_k)^\dagger\right\| = \frac{1}{\sigma_{\min}(\nabla c(x_k))} \leq \frac{1}{\sigma}$. Therefore, the norm of the step is bounded by $\|p_k\| \leq \left\|\nabla c(x_k)^\dagger\right\| \|c(x_k)\| \leq \frac{r_k}{\sigma}$. Then by the L_c-Lipschitz continuity of ∇c we obtain

$$r_{k+1} = \|c(x_{k+1})\| \leq \|c(x_k) + \alpha_k \nabla c(x_k)p_k\| + \frac{L_c}{2}\alpha_k^2\|p_k\|^2 \leq (1-\alpha_k)r_k + \frac{L_c}{2\sigma^2}\alpha_k^2 r_k^2, \tag{19}$$

The iteration complexity naturally splits into two phases governed by the damping rule (18):

Phase I: Global damping ($r_k > \frac{\sigma^2}{L_c}$). Here, $\alpha_k = \frac{\sigma^2}{L_c r_k} < 1$. Substituting α_k into (19) gives:

$$r_{k+1} \leq \left(1 - \frac{\sigma^2}{L_c r_k}\right) r_k + \frac{L_c}{2\sigma^2}\left(\frac{\sigma^2}{L_c r_k}\right)^2 r_k^2 = r_k - \frac{\sigma^2}{2L_c}.$$

The residual strictly decreases by a constant amount $\frac{\sigma^2}{2L_c}$ at each iteration. Starting from $r_0 = \Delta_c$, the method must enter the local region $r_k \leq \frac{\sigma^2}{L_c}$ in at most $\left\lceil \frac{2L_c\Delta_c}{\sigma^2} \right\rceil$ iterations.

Phase II: Local quadratic convergence ($r_k \leq \frac{\sigma^2}{L_c}$). Once the residual is sufficiently small, the step size switches to full Newton steps ($\alpha_k = 1$). The recursion (19) simplifies to:

$$r_{k+1} \leq \frac{L_c}{2\sigma^2} r_k^2.$$

Defining the scaled residual $s_k = \frac{L_c}{2\sigma^2} r_k$, we obtain $s_{k+1} \leq s_k^2$. If k' is the entry index of this phase, then $s_{k'} \leq \frac{1}{2}$. By induction, after ℓ local iterations,

$$s_{k'+\ell} \leq \left(\frac{1}{2}\right)^{2^\ell}.$$

To achieve $r_k \leq \epsilon$, it is enough that $s_k \leq \frac{L_c\epsilon}{2\sigma^2}$. Thus this phase terminates in at most

$$\left\lceil \log_2 \log_2 \left(\frac{2\sigma^2}{L_c\epsilon} \right) \right\rceil$$

iterations.

Summing the iteration bounds of the two phases completes the proof. □

Remark 1. The global phase bound $O(L_c\Delta_c\sigma^{-2})$ exactly matches the global lower bound established in Lemma 1, while the local phase bound $O(\log\log(\frac{\sigma^2}{L_c\epsilon}))$ matches the local lower bound derived in Lemma 2. This confirms that our feasibility lower-bound scaling is tight up to absolute constants.

4 Lower bound for stationarity

We construct a nonlinear equality-constrained instance that uses the unconstrained zero-chain lower bound in the constrained setting. To separate stationarity from feasibility, we work with the multiplier-minimized stationarity residual, which contains no primal feasibility term. The proof first builds the constrained instance, then evaluates this residual by an exact multiplier elimination, and finally extends the resulting lower bound to any deterministic methods via the reduction using disjoint supports.

4.1 Constrained hard instance

We define an auxiliary constrained instance on variables $(u, \tau) \in \mathbb{R}^d \times \mathbb{R}$ whose feasible reduction is a scaled zero-chain. The scalar coordinate τ is introduced so that the constraint Jocobian has a uniform lower bound σ for its singular values. When the constraint is eliminated, this component converts the linear objective term $G\tau$ into the additional scale $\frac{GL_c}{\sigma}$ in the reduced zero-chain. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be the function from Proposition 1. For scaling parameters $a, r > 0$, set

$$\psi(u) = \frac{a}{L_f}\varphi(\frac{u}{r}).$$

Consider the constrained problem

$$\begin{aligned} \min_{u\in\mathbb{R}^d,\, \tau\in\mathbb{R}} \quad & F(u,\tau) = L_f\psi(u) + G\tau, \\ \text{subject to} \quad & c(u,\tau) = \sigma\tau - L_c\psi(u) = 0. \end{aligned} \tag{20}$$

Initialize at $u_0 = 0$ and $\tau_0 = -\frac{\Delta_c}{\sigma}$. On the feasible set, the constraint eliminates the scalar variable:

$$\tau = \frac{L_c}{\sigma}\psi(u).$$

Thus the reduced feasible objective is

$$\Lambda(u) = F\left(u, \frac{L_c}{\sigma}\psi(u)\right) = \alpha\psi(u), \quad \text{where} \quad \alpha = L_f + \sigma^{-1}GL_c. \tag{21}$$

We now set the scaling parameters. Let $\Delta = -\inf_z \varphi(z)$. Since $L_f\psi(u) = a\varphi(\frac{u}{r})$ and $\varphi(0) = 0$, we have $L_f\psi(0) = 0$ and $\inf_u L_f\psi(u) = -a\Delta$. Hence $F(u_0, \tau_0) = -\frac{G\Delta_c}{\sigma}$, and the identity $F(u_0, \tau_0) - \inf_u \Lambda(u) = \Delta_f$ is equivalent to $-\frac{G\Delta_c}{\sigma} + \frac{\alpha}{L_f}a\Delta = \Delta_f$. Thus, except for the trivial case $\Delta_f = \Delta_c = 0$, we set

$$a = \frac{L_f}{\alpha\Delta}\left(\Delta_f + \sigma^{-1}G\Delta_c\right), \qquad r = \max\left\{\sqrt{\frac{a\ell}{L_f}}, \frac{aG_0}{G}\right\},$$

where ℓ and G_0 are the constants in Proposition 1.

We next verify that the instance belongs to $\mathcal{P}_\infty(G, L_f, L_c, \Delta_f, \Delta_c)$. Its derivatives are

$$\nabla F(u,\tau) = \begin{bmatrix} L_f\nabla\psi(u) \\ G \end{bmatrix} = \begin{bmatrix} \frac{a}{r}\nabla\varphi(\frac{u}{r}) \\ G \end{bmatrix}, \qquad \nabla c(u,\tau) = \begin{bmatrix} -L_c\nabla\psi(u)^\top & \sigma \end{bmatrix}.$$

The Hessian bound from Proposition 1 gives

$$\|L_f\nabla^2\psi(u)\| \le \frac{a\ell}{r^2} \le L_f, \qquad \|L_c\nabla^2\psi(u)\| \le \frac{L_c}{L_f}\frac{a\ell}{r^2} \le L_c.$$

Thus ∇F is globally L_f-Lipschitz continuous, and ∇c is globally L_c-Lipschitz continuous. Moreover,

$$\|\nabla F(u,\tau)\|_\infty \le \max\left\{\frac{aG_0}{r}, G\right\} \le G, \qquad \forall (u,\tau).$$

The feasible set is nonempty, its feasible infimum is finite, and the initial constraint violation and objective gap are exactly

$$|c(u_0,\tau_0)| = \Delta_c, \qquad F(u_0,\tau_0) - \inf_{u\in\mathbb{R}^d} \Lambda(u) = \Delta_f.$$

In addition, the construction satisfies

$$\sigma_{\min}(\nabla c(u,\tau)) = \sqrt{L_c^2 \|\nabla\psi(u)\|^2 + \sigma^2} \ge \sigma \quad \text{for all } (u,\tau).$$

4.2 Exact residual reduction

By (6), the multiplier-minimized stationarity residual minimizes the norm of the Lagrangian gradient over the multiplier $\lambda \in \mathbb{R}$. For our single-constraint instance, it is given by

$$\begin{aligned}
\mathcal{R}_{\text{stat}}(u,\tau)^2 &= \min_{\lambda\in\mathbb{R}} \left\| \nabla F(u,\tau) + \nabla c(u,\tau)^\top \lambda \right\|^2 \\
&= \min_{\lambda\in\mathbb{R}} \left(\|L_f \nabla\psi(u) - \lambda L_c \nabla\psi(u)\|^2 + |G + \sigma\lambda|^2 \right) \\
&= \min_{\lambda\in\mathbb{R}} \left((L_f - \lambda L_c)^2 \|\nabla\psi(u)\|^2 + |G+\sigma\lambda|^2 \right)
\end{aligned} \tag{22}$$

Note that the function to minimize in (22) is quadratic and strictly convex. It is easy to obtain the unique minimizer $\lambda^* = \frac{L_c L_f q - \sigma G}{L_c^2 q + \sigma^2}$, where $q = \|\nabla\psi(u)\|^2$. Since $\alpha = L_f + \sigma^{-1} G L_c$ as in (21), it gives

$$L_f - L_c\lambda^* = \frac{\sigma^2\alpha}{L_c^2 q + \sigma^2}, \qquad G + \sigma\lambda^* = \frac{\sigma L_c \alpha q}{L_c^2 q + \sigma^2}.$$

The optimal value of the minimizatin problem in (22) is $\frac{\sigma^2\alpha^2 q}{L_c^2 q + \sigma^2}$. We thus obtain

$$\mathcal{R}_{\text{stat}}(u,\tau) = \frac{\sigma\alpha}{\sqrt{L_c^2 \|\nabla\psi(u)\|^2 + \sigma^2}} \|\nabla\psi(u)\|. \tag{23}$$

After the residual identity, it remains to lower bound the gradient information in the u-block. Indeed, the u-components of $\nabla F(u,\tau)$ and $\nabla c(u,\tau)^\top$ are

$$L_f \nabla\psi(u) = \frac{a}{r}\nabla\varphi(\frac{u}{r}), \qquad -L_c\nabla\psi(u) = -\frac{L_c}{L_f}\frac{a}{r}\nabla\varphi(\frac{u}{r}),$$

respectively. Hence the first-order information in the u-block has the same zero-chain support structure as $\nabla\varphi$, while τ does not reveal additional coordinates of $\nabla\varphi$.

4.3 Stationarity lower bound

Theorem 3. *Fix positive G, L_f, L_c, σ and nonnegative Δ_f, Δ_c. Suppose the accuracy parameter satisfies*

$$0 < \epsilon \le \min\left\{ \frac{G}{2\sqrt{2}\,G_0}, \frac{L_f\sigma + GL_c}{\sqrt{2}L_c} \right\},$$

where G_0 is the constant in Proposition 1. Then, for every deterministic first-order method $\mathsf{A} \in \mathcal{A}$, there exists, in sufficiently large dimension, a scalar-constraint instance $P = (f, c, x_0) \in \mathcal{P}_\infty(G, L_f, L_c, \Delta_f, \Delta_c)$. This instance satisfies $\sigma_{\min}(\nabla c(x)) \geq \sigma$ for all x, and A requires at least

$$\Omega\left(\frac{(L_f + \sigma^{-1}GL_c)(\Delta_f + \sigma^{-1}G\Delta_c)}{\epsilon^2}\right) \tag{24}$$

oracle calls before querying an ϵ-stationary point.

Proof. If $\Delta_f + \sigma^{-1}G\Delta_c = 0$, the claimed lower bound is trivial. Hence assume $\Delta_f + \sigma^{-1}G\Delta_c > 0$. Given any deterministic first-order method $\mathsf{A} \in \mathcal{A}$, we prove the stationarity lower bound by contradiction using the construction (20). Suppose that A queries an ϵ-stationary point within T oracle calls satisfying

$$T < \frac{c_0}{4}\frac{\alpha(\Delta_f + \sigma^{-1}G\Delta_c)}{\epsilon^2}, \tag{25}$$

where c_0 is the constant from Proposition 1 and $\alpha = L_f + \sigma^{-1}GL_c$. Choose the dimension d such that

$$d \geq \left\lceil \frac{c_0}{2}\frac{\alpha(\Delta_f + \sigma^{-1}G\Delta_c)}{\epsilon^2} \right\rceil > T, \qquad 2d \geq \frac{L_f^2 G_0^2}{\alpha \ell G^2}\left(\Delta_f + \sigma^{-1}G\Delta_c\right),$$

where $\alpha = L_f + \sigma^{-1}GL_c$ and ℓ is the constant in Proposition 1. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be the chain in Proposition 1, and set $\Delta = -\inf_z \varphi(z)$. Recall that the scaling parameters a and r defined in Section 4.1 satisfy

$$a = \frac{L_f}{\alpha\Delta}\left(\Delta_f + \sigma^{-1}G\Delta_c\right), \qquad r = \max\left\{\sqrt{\frac{a\ell}{L_f}}, \frac{aG_0}{G}\right\}.$$

Since $\Delta \geq 2d$ by Proposition 1(iv), the choice of d gives $(\frac{aG_0}{G})^2 \leq \frac{a\ell}{L_f}$, and thus $r = \sqrt{\frac{a\ell}{L_f}}$.

Let (u, τ) be an ϵ-stationary query among the first T oracle calls. Then by (23), ϵ-stationarity implies

$$\frac{\sigma^2\alpha^2\left\|\nabla\psi(u)\right\|^2}{L_c^2\left\|\nabla\psi(u)\right\|^2 + \sigma^2} \leq \epsilon^2.$$

Since $\epsilon L_c \leq \frac{\alpha\sigma}{\sqrt{2}}$, rearranging yields

$$\|L_f\nabla\psi(u)\| \leq \frac{\sqrt{2}L_f\epsilon}{\alpha}.$$

Using $L_f\nabla\psi(u) = \frac{a}{r}\nabla\varphi(\frac{u}{r})$, this query also satisfies

$$\left\|\nabla\varphi(\frac{u}{r})\right\| \leq \epsilon_\varphi, \quad \text{where} \quad \epsilon_\varphi = \frac{r}{a}\frac{\sqrt{2}L_f\epsilon}{\alpha}.$$

By Propositions 1 and 2, this requires at least

$$T \geq c_0\frac{\ell\Delta}{\epsilon_\varphi^2} = c_0\ell\Delta\left(\frac{a}{r}\frac{\alpha}{\sqrt{2}L_f\epsilon}\right)^2 = \frac{c_0}{2}\frac{\alpha(\Delta_f + \sigma^{-1}G\Delta_c)}{\epsilon^2} > \frac{c_0}{4}\frac{\alpha(\Delta_f + \sigma^{-1}G\Delta_c)}{\epsilon^2},$$

oracle calls, contradicting (25). Therefore, there is an admissible instance on which A has no ϵ-stationary query within any T satisfying (25). Equivalently, A requires $\Omega(\frac{\alpha(\Delta_f+\sigma^{-1}G\Delta_c)}{\epsilon^2})$ oracle calls. Substituting $\alpha = L_f + \sigma^{-1}GL_c$ gives (24). □

4.4 Tightness of stationarity bounds

We next show that the stationarity lower-bound scaling is tight, up to the dimension dependence appearing below, by analyzing a prox-linear method. For $\rho > 0$, define the penalty function

$$\Phi_\rho(x) = f(x) + \rho \|c(x)\|.$$

Given the current iterate x_k and a model parameter $\beta > 0$, update

$$x_{k+1} = x_k + p_k, \quad \text{where} \quad p_k = \arg\min_{p \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), p \rangle + \rho \|c(x_k) + \nabla c(x_k) p\| + \frac{\beta}{2} \|p\|^2 \right\}. \tag{26}$$

Let $\mathbb{B} = \{u \in \mathbb{R}^m : \|u\| \le 1\}$. According to the optimality condition there exists some $u_k \in \mathbb{B}$ such that

$$0 = \nabla f(x_k) + \rho \nabla c(x_k)^\top u_k + \beta p_k, \quad \text{thus} \quad p_k = -\beta^{-1}\big(\nabla f(x_k) + \rho \nabla c(x_k)^\top u_k\big). \tag{27}$$

Therefore, each step uses only the current first-order oracle output and remains within the algorithm class (4). Throughout the remainder of this subsection, ∂ denotes the Clarke subdifferential.

Lemma 3. *Let*

$$\Omega_\rho = \{x \in \mathbb{R}^n : \Phi_\rho(x) \le \Phi_\rho(x_0)\}, \qquad f^\star = \inf\{f(x) : c(x) = 0\}.$$

Assume that $\|\nabla f(x)\|_\infty \le G$ *and* $\sigma_{\min}(\nabla c(x)) \ge \sigma > 0$ *hold on a neighborhood of* Ω_ρ. *If* $\rho \ge \frac{\sqrt{n}\, G}{\sigma}$, *then*

$$\Phi_\rho(x) \ge f^\star, \qquad \forall x \in \Omega_\rho.$$

Consequently,

$$\Phi_\rho(x_0) - \inf_{x \in \Omega_\rho} \Phi_\rho(x) \le \Delta_f + \rho \Delta_c.$$

Proof. Fix any $x \in \Omega_\rho$. If $c(x) = 0$, then $\Phi_\rho(x) = f(x) \ge f^\star$ obviously. Assume now that $c(x) \ne 0$, and consider the normal flow

$$\dot\gamma(\tau) = -\nabla c(\gamma(\tau))^\dagger c(\gamma(\tau)), \quad \text{where} \quad \gamma(0) = x.$$

As long as $\gamma(\tau) \in \Omega_\rho$, the chain rule and the identity $\nabla c(\gamma(\tau)) \nabla c(\gamma(\tau))^\dagger = I$ give

$$\frac{d}{d\tau} c(\gamma(\tau)) = \nabla c(\gamma(\tau)) \dot\gamma(\tau) = -c(\gamma(\tau)),$$

and therefore

$$c(\gamma(\tau)) = e^{-\tau} c(x), \qquad \|\dot\gamma(\tau)\| \le \left\| \nabla c(\gamma(\tau))^\dagger \right\| \|c(\gamma(\tau))\| \le \sigma^{-1} e^{-\tau} \|c(x)\|.$$

It follows that

$$\frac{d}{d\tau} \Phi_\rho(\gamma(\tau)) = \langle \nabla f(\gamma(\tau)), \dot\gamma(\tau) \rangle + \rho \frac{d}{d\tau} \|c(\gamma(\tau))\| \le \left(\frac{\sqrt{n}\, G}{\sigma} - \rho \right) e^{-\tau} \|c(x)\| \le 0.$$

Thus, on any interval on which $\gamma(\tau) \in \Omega_\rho$, we have $\frac{d}{d\tau} \Phi_\rho(\gamma(\tau)) \le 0$. Since $\gamma(0) = x \in \Omega_\rho$, it follows by a continuation argument that $\gamma(\tau) \in \Omega_\rho$ for all $\tau \ge 0$. The bound on $\|\dot\gamma(\tau)\|$ implies that $\gamma(\tau)$ is Cauchy and converges to some $\bar x$ with

$$\|x - \bar x\| \le \int_0^\infty \|\dot\gamma(\tau)\| \, d\tau \le \sigma^{-1} \|c(x)\|.$$

Since $c(\gamma(\tau)) = e^{-\tau}c(x)$, we also have $c(\bar{x}) = 0$. Finally,

$$f(x) \geq f(\bar{x}) - \int_0^\infty \|\nabla f(\gamma(\tau))\| \, \|\dot{\gamma}(\tau)\| \, d\tau \geq f^\star - \frac{\sqrt{n}\, G}{\sigma} \|c(x)\|.$$

Therefore,

$$\Phi_\rho(x) = f(x) + \rho \|c(x)\| \geq f^\star + \left(\rho - \frac{\sqrt{n}\, G}{\sigma}\right) \|c(x)\| \geq f^\star.$$

Applying this at $x = x_0$ gives

$$\Phi_\rho(x_0) - \inf_{x \in \Omega_\rho} \Phi_\rho(x) \leq \Phi_\rho(x_0) - f^\star = \Delta_f + \rho \Delta_c.$$

The proof is completed. □

Theorem 4. *Suppose f has L_f-Lipschitz gradient and ∇c is L_c-Lipschitz. Given initial gaps Δ_f, Δ_c, an accuracy parameter $0 < \epsilon \leq \sqrt{n}\, G$, and constants $G, \sigma > 0$, choose $\rho = \frac{\sqrt{n}\, G + 2\epsilon}{\sigma}$, $L_\rho = L_f + \rho L_c$, and $\beta = 2L_\rho$, and assume $\epsilon \leq \frac{18 L_\rho^2}{L_c}$. Let $\Omega_\rho = \{x \in \mathbb{R}^n : \Phi_\rho(x) \leq \Phi_\rho(x_0)\}$. If $\|\nabla f(x)\|_\infty \leq G$ and $\sigma_{\min}(\nabla c(x)) \geq \sigma$ hold on a neighborhood of Ω_ρ, then the prox-linear method (26) produces a point x_{k+1} satisfying*

$$\|c(x_{k+1})\| \leq \epsilon, \qquad \mathcal{R}_{\text{stat}}(x_{k+1}) \leq \epsilon,$$

in at most

$$O\left(\frac{\left(L_f + \sigma^{-1}\sqrt{n}\, G L_c\right)\left(\Delta_f + \sigma^{-1}\sqrt{n}\, G \Delta_c\right)}{\epsilon^2}\right)$$

iterations. Thus Theorem 3 is tight up to a dimension factor.

Proof. For every $p \in \mathbb{R}^n$, the smoothness assumptions imply

$$f(x_k + p) \leq f(x_k) + \langle \nabla f(x_k), p \rangle + \frac{L_f}{2} \|p\|^2,$$

and

$$\|c(x_k + p) - c(x_k) - \nabla c(x_k) p\| \leq \frac{L_c}{2} \|p\|^2.$$

Since $u \mapsto \rho \|u\|$ is ρ-Lipschitz continuous, we obtain

$$\Phi_\rho(x_k + p) \leq f(x_k) + \langle \nabla f(x_k), p \rangle + \rho \|c(x_k) + \nabla c(x_k) p\| + \frac{L_\rho}{2} \|p\|^2.$$

Setting $p = p_k$ and comparing with $p = 0$ in (26) give

$$\Phi_\rho(x_{k+1}) \leq \Phi_\rho(x_k) - \frac{\beta - L_\rho}{2} \|p_k\|^2 = \Phi_\rho(x_k) - \frac{L_\rho}{2} \|p_k\|^2,$$

where the last inequality uses $\beta = 2L_\rho$. In particular, we have $x_k \in \Omega_\rho$ for all k. Summing over $k = 0, \ldots, N-1$ and invoking Lemma 3 yield

$$\frac{L_\rho}{2} \sum_{k=0}^{N-1} \|p_k\|^2 \leq \Phi_\rho(x_0) - f^\star = \Delta_f + \rho \Delta_c.$$

Hence there exists some $k < N$ such that

$$\|p_k\|^2 \leq \frac{2(\Delta_f + \rho\Delta_c)}{NL_\rho}. \tag{28}$$

Now from (27), there exist a vector $\lambda_k = \rho u_k$ such that

$$\|\nabla f(x_k) + \nabla c(x_k)^\top \lambda_k\|^2 = \beta^2 \|p_k\|^2$$

Therefore, if

$$N \geq 8\,\frac{L_\rho(\Delta_f + \rho\Delta_c)}{\epsilon^2},$$

then (28) implies

$$\|\nabla f(x_k) + \nabla c(x_k)^\top \lambda_k\| \leq \epsilon. \tag{29}$$

We now convert (29) into feasibility and stationarity for the original constrained problem. If $\|c(x_k) + \nabla c(x_k)p_k\| \neq 0$, then $\|u_k\| = 1$ and $\|\lambda_k\| = \rho$ by (27). Using $\|\nabla f(x_k)\| \leq \sqrt{n}\,G$ and $\sigma_{\min}(\nabla c(x_k)) \geq \sigma$, we obtain

$$\|\nabla f(x_k) + \nabla c(x_k)^\top \lambda_k\| \geq \rho\sigma - \sqrt{n}\,G = 2\epsilon,$$

which contradicts (29). Hence it holds that $\|c(x_k) + \nabla c(x_k)p_k\| = 0$. Now we obtain

$$\|c(x_{k+1})\| = \|c(x_{k+1}) - c(x_k) - \nabla c(x_k)p_k\| \leq \frac{L_c}{2}\|p_k\|^2 \leq \epsilon,$$

where the last inequality uses $N \geq \frac{L_c(\Delta_f + \rho\Delta_c)}{L_\rho \epsilon}$. Meanwhile, due to the Lipschitz continuity of ∇f and ∇c, it holds that

$$\mathcal{R}_{\text{stat}}(x_{k+1}) \leq \|\nabla f(x_{k+1}) + \nabla c(x_{k+1})^\top \lambda_k\| \leq (L_f + \rho L_c + \beta)\|p_k\| = 3L_\rho\|p_k\| \leq \epsilon$$

thanks to $N \geq 18(L_f + \rho L_c)(\Delta_f + \rho\Delta_c)\epsilon^{-2}$. The proof is completed. □

5 Lower bound for KKT accuracy

Having established lower bounds for feasibility and stationarity separately, we now record their implication for approximate KKT points. Such a point must satisfy both the feasibility condition and the multiplier-minimized stationarity condition at the same queried point, as in (11).

Corollary 1. *Fix $G, L_f, L_c, \sigma > 0$, $\Delta_f \geq 0$, and $\Delta_c \geq \frac{\sigma^2}{L_c}$. For this statement, write $\mathcal{P}_\infty$ for $\mathcal{P}_\infty(G, L_f, L_c, \Delta_f, \Delta_c)$. Suppose $0 < \epsilon_c < \min\{\frac{\Delta_c}{2}, \frac{\sigma^2}{L_c}\}$ and $0 < \epsilon_s \leq \min\{\frac{G}{2\sqrt{2}\,G_0}, \frac{L_f\sigma + GL_c}{\sqrt{2}L_c}\}$, where G_0 is the constant in Proposition 1. Then*

$$\mathcal{T}^{\text{KKT}}(\epsilon_c, \epsilon_s; \mathcal{A}, \mathcal{P}_\infty) \geq \Omega\left(\max\left\{\frac{L_c\Delta_c}{\sigma^2} + \log\log\frac{\sigma^2}{L_c\epsilon_c}, \frac{(L_f + \sigma^{-1}GL_c)(\Delta_f + \sigma^{-1}G\Delta_c)}{\epsilon_s^2}\right\}\right).$$

Proof. For every deterministic method A and instance P, a query that satisfies the simultaneous KKT condition also satisfies each separated condition. Hence

$$T^{\text{KKT}}(\epsilon_c, \epsilon_s; \mathsf{A}, P) \geq T^{\text{feas}}_{\epsilon_c}(\mathsf{A}, P), \qquad T^{\text{KKT}}(\epsilon_c, \epsilon_s; \mathsf{A}, P) \geq T^{\text{stat}}_{\epsilon_s}(\mathsf{A}, P).$$

Taking the supremum over $P \in \mathcal{P}_\infty$ and then the infimum over $\mathsf{A} \in \mathcal{A}$ gives

$$\mathcal{T}^{\text{KKT}}(\epsilon_c, \epsilon_s; \mathcal{A}, \mathcal{P}_\infty) \geq \max\left\{\mathcal{T}^{\text{feas}}_{\epsilon_c}(\mathcal{A}, \mathcal{P}_\infty), \mathcal{T}^{\text{stat}}_{\epsilon_s}(\mathcal{A}, \mathcal{P}_\infty)\right\}.$$

The claimed bound follows from Theorems 1 and 3. □

Algorithm 1 A two-stage method for approximate KKT points

Require: Problem $P = (f, c, x_0)$, accuracies $\epsilon_s, \epsilon_c > 0$, parameters $G, L_f, L_c, \sigma, \Delta_f, \Delta_c$

Set $\rho = \dfrac{\sqrt{n}\,G + \epsilon_s}{\sigma}$, $\beta = 2L_\rho$ and $N_1 = \lceil \frac{72 L_\rho(\Delta_f + \rho\Delta_c)}{\epsilon_s^2} \rceil$ with $L_\rho = L_f + \rho L_c$.

Stage I: prox-linear exact-penalty phase

for $k = 0, 1, \ldots, N_1 - 1$ **do**

  Set $x_{k+1} = x_k + p_k$ with

$$p_k \in \arg\min_{p \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), p \rangle + \rho \, \|c(x_k) + \nabla c(x_k) p\| + \frac{\beta}{2} \|p\|^2 \right\}.$$

end for

Set $z = x_{\bar{k}} + p_{\bar{k}}$ with $\bar{k} \in \arg\min_{0 \le k < N_1} \|p_k\|$.

Stage II: damped Newton feasibility correction

Set $z_0 = z$ and $j = 0$.

while $\|c(z_j)\| > \epsilon_c$ **do**

  Compute the minimum-norm Newton direction $d_j = -\nabla c(z_j)^\dagger c(z_j)$.

  Choose the damping parameter $\alpha_j = \min\{1, \frac{\sigma^2}{L_c \|c(z_j)\|}\}$.

  Set $z_{j+1} = z_j + \alpha_j d_j$.

  Set $j = j + 1$.

end while

return $y = z_j$.

The lower bound in Corollary 1 is a valid consequence for the simultaneous approximate KKT condition, but it does not rule out simultaneous feasibility and stationarity on one run. The first query satisfying feasibility and the first query satisfying stationarity need not coincide. A lower bound of that form would require a single hard-instance construction that rules out, along the same run, every queried point satisfying both conditions.

The corresponding upper-bound statement is also separated in nature. The prox-linear method can first reduce the stationarity residual while producing a point with very small constraint violation. A subsequent damped Newton correction then reduces the constraint violation to the desired level, and the small total Newton displacement keeps the stationarity residual below the prescribed tolerance. The resulting two-stage procedure is summarized in Algorithm 1.

Theorem 5. *Fix $P = (f, c, x_0) \in \mathcal{P}_\infty$ with parameters $G, L_f, L_c, \Delta_f, \Delta_c$, and suppose $\sigma_{\min}(\nabla c(x)) \ge \sigma > 0$ for any $x \in \mathbb{R}^n$. Let $0 < \epsilon_s \le \sqrt{n}\,G$ and $0 < \epsilon_c \le \frac{\sigma^2}{2L_c}$. Then Algorithm 1 returns a point y satisfying*

$$\|c(y)\| \le \epsilon_c, \qquad \mathcal{R}_{\mathrm{stat}}(y) \le \epsilon_s$$

in at most

$$O\left(\max\left\{ \frac{L_c \Delta_c}{\sigma^2} + \log\log \frac{\sigma^2}{L_c \epsilon_c}, \frac{(L_f + \sigma^{-1}\sqrt{n}\,G L_c)(\Delta_f + \sigma^{-1}\sqrt{n}\,G\Delta_c)}{\epsilon_s^2} \right\} \right)$$

first-order oracle calls.

Proof. Run the prox-linear method (26) with $\rho = \frac{\sqrt{n}\,G + \epsilon_s}{\sigma}$, $L_\rho = L_f + \rho L_c$ and $\beta = 2L_\rho$. The proof of Theorem 4 gives, after

$$N_1 = O\left(\frac{L_\rho(\Delta_f + \rho\Delta_c)}{\epsilon_s^2} \right)$$

iterations, an iterate $z = x_k + p_k$ such that

$$\mathcal{R}_{\text{stat}}(z) \le \frac{\epsilon_s}{2}, \qquad \|p_k\| \le \frac{\epsilon_s}{6L_\rho}, \qquad c(x_k) + \nabla c(x_k)p_k = 0.$$

The Lipschitz continuity of ∇c then gives

$$\|c(z)\| \le \frac{L_c}{2}\|p_k\|^2 \le \frac{L_c\epsilon_s^2}{72L_\rho^2} =: \delta.$$

Starting from z, run the damped Newton method (16)–(18) on the equation $c(x) = 0$. Since

$$L_\rho \ge \frac{L_c\epsilon_s}{\sigma}, \qquad \delta \le \frac{\sigma^2}{72L_c},$$

where the second inequality follows by substituting the first one into $\delta = \frac{L_c\epsilon_s^2}{72L_\rho^2}$, the Newton correction starts in the local region. If $z_0 = z$ and $r_j = \|c(z_j)\|$, the local argument in Theorem 2 gives

$$r_{j+1} \le \frac{L_c}{2\sigma^2}r_j^2 \le \frac{1}{2}r_j, \qquad \|z_{j+1} - z_j\| \le \frac{r_j}{\sigma}.$$

Hence the total displacement of the Newton correction is bounded by

$$\|y - z\| \le \sum_{j\ge 0}\frac{r_j}{\sigma} \le \frac{2\delta}{\sigma},$$

and after

$$N_2 = O\left(\log\log\frac{\sigma^2}{L_c\epsilon_c}\right)$$

Newton iterations we obtain $\|c(y)\| \le \epsilon_c$. It remains to check stationarity at y. Choose λ_z such that

$$\|\nabla f(z) + \nabla c(z)^\top\lambda_z\| \le \frac{\epsilon_s}{2}.$$

Using $\|\nabla f(z)\| \le \sqrt{n}\,G$ and $\sigma_{\min}(\nabla c(z)) \ge \sigma$, we have

$$\|\lambda_z\| \le \frac{\sqrt{n}\,G + \frac{\epsilon_s}{2}}{\sigma}.$$

Therefore, it follows from $L_\rho \ge \frac{L_c\epsilon_s}{\sigma}$ that

$$\begin{aligned}\mathcal{R}_{\text{stat}}(y) &\le \|\nabla f(y) + \nabla c(y)^\top\lambda_z\| \\ &\le \mathcal{R}_{\text{stat}}(z) + \big(L_f + L_c\|\lambda_z\|\big)\|y - z\| \\ &\le \frac{\epsilon_s}{2} + L_\rho\frac{2\delta}{\sigma} = \frac{\epsilon_s}{2} + \frac{L_c\epsilon_s^2}{36\sigma L_\rho} \le \frac{\epsilon_s}{2} + \frac{\epsilon_s}{36} < \epsilon_s.\end{aligned}$$

Therefore, the total number of oracle calls is

$$N_1 + N_2 = O\left(\frac{L_\rho(\Delta_f + \rho\Delta_c)}{\epsilon_s^2} + \log\log\frac{\sigma^2}{L_c\epsilon_c}\right).$$

Since $\epsilon_s \le \sqrt{n}\,G$, the first term is bounded by

$$O\left(\frac{(L_f + \sigma^{-1}\sqrt{n}\,GL_c)(\Delta_f + \sigma^{-1}\sqrt{n}\,G\Delta_c)}{\epsilon_s^2}\right).$$

Moreover, under $\epsilon_s \le \sqrt{n}\,G$, the above term is no smaller than $\frac{L_c\Delta_c}{\sigma^2}$. The stated bound follows. □

Under the assumptions of Theorem 5, the KKT lower bound above is matched up to the same dimension dependence that appears in the stationarity upper bound. This theorem does not prove a lower bound that rules out simultaneous feasibility and stationarity on one run; it shows that the two steps of the algorithm can be combined without changing the displayed rates.

6 Conclusion

We established iteration-complexity lower bounds for smooth equality-constrained nonconvex optimization for two distinct conditions: feasibility and stationarity. Under iteration-wise Jacobian regularity, we derived a lower bound for reaching an ϵ-feasible point. For stationarity, we constructed a nonlinear hard instance satisfying a global Jacobian lower bound and yielding a lower bound for reaching an ϵ-stationary point under an objective-gradient ℓ_∞-norm bound. The reduction using disjoint supports transfers the zero-chain lower bounds to any deterministic first-order methods while preserving the ℓ_∞-norm gradient bound. The upper bounds show that the feasibility lower bound is attained up to constants and that the stationarity lower bound is attained up to the dimension dependence stated in Section 4.4. We also recorded the lower-bound consequence for simultaneous approximate KKT points and showed that a two-stage method consisting of the prox-linear method followed by damped Newton correction matches the consequence obtained from the separate feasibility and stationarity bounds up to the same stationarity dimension dependence. A natural direction for future work is to understand whether an analogous separation persists under weaker regularity assumptions, for more general constraint classes, or in a sharp characterization of simultaneous KKT complexity.

References

[1] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2):165–214, 2023.

[2] Nicolas Boumal, P.-A. Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[3] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1–2):71–120, 2020.

[4] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1–2):315–355, 2021.

[5] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 144(1–2):93–106, 2014.

[6] Oliver Hinder and Yinyu Ye. Worst-case iteration bounds for log barrier methods on problems with nonconvex constraints. *Mathematics of Operations Research*, 49(4):2402–2424, 2024.

[7] Zhichao Jia and Benjamin Grimmer. First-order methods for nonsmooth nonconvex functional constrained optimization with or without Slater points. *SIAM Journal on Optimization*, 35(2):1300–1329, 2025.

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

[9] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2170–2178. PMLR, 2021.

[10] Qihang Lin, Runchao Ma, and Yangyang Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational Optimization and Applications*, 82(1):175–224, 2022.

[11] Wei Liu, Qihang Lin, and Yangyang Xu. Lower complexity bounds of first-order methods for affinely constrained composite nonconvex problems. *Mathematics of Operations Research*, 51:1659–1682, 2026.

[12] Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Chichester, UK, 1983.

[13] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2 edition, 2018.

[14] Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.

[15] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1–2):1–35, 2021.

[16] Mehmet Fatih Sahin, Armin Eftekhari, Ahmet Alacaoglu, Fabian Latorre, and Volkan Cevher. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. In *Advances in Neural Information Processing Systems*, volume 32, pages 13943–13955, 2019.