ZEROTH-ORDER CONSTRAINED OPTIMIZATION FROM A CONTROL PERSPECTIVE VIA FEEDBACK LINEARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing safe derivative-free optimization algorithms under unknown constraints is a fundamental challenge in modern learning and control. Most existing zeroth-order (ZO) approaches typically assume white-box constraints or focus on convex settings, leaving the general case of nonconvex optimization with black-box constraints largely open. We propose a control-theoretic framework for ZO constrained optimization that enforces feasibility without relying on solving costly convex subproblems. Leveraging feedback linearization, we introduce a family of ZO feedback linearization (ZOFL) algorithms applicable to both equality and inequality constraints. Our method requires only noisy, sample-based gradient estimates yet provably guarantees constraint satisfaction under mild regularity conditions. We establish finite-time bounds on constraint violation and further present a midpoint discretization variant that further improves feasibility without sacrificing optimality. Empirical results demonstrate that ZOFL consistently outperforms standard ZO baselines, achieving competitive objective values while maintaining feasibility.

1 INTRODUCTION

Designing safe learning methods is both important and challenging. Safety requires guarantees of feasibility at every step, which in turn demands reliable information about the system's objectives and constraints. In many real-world settings, such information is only accessible through function evaluations—gradients are either unavailable, unreliable, or prohibitively expensive to compute. This makes derivative-free methods natural candidates: they update decisions from sampled outcomes without requiring gradient access. Yet, enforcing strict safety guarantees in these derivative-free settings remains largely unresolved.

Among derivative-free approaches, zeroth-order methods have attracted significant attention due to their simplicity and scalability to high dimensions [49, 40]. The core idea is to build stochastic *gradient estimators* via finite differences of function evaluations and then plug them into standard gradient-based updates [9]. For instance, two-point estimators perturb the decision along random isotropic directions (Gaussian or uniform on the unit sphere) and combine the function values to approximate gradients [40, 48, 44]. When combined with gradient descent, such estimators yield provable converge rates for unconstrained optimization.

In the constrained setting, most existing ZO algorithms assume black-box objectives but *white-box constraints*. Explicit knowledge of the constraint set enables efficient projections or local linearizations, ensuring feasibility. This has led to a variety of algorithms, including projection–gradient-descent [48, 33, 12, 21], Frank–Wolfe–type methods [45, 36, 8], and Sequential Quadratic Programming (SQP)-style approaches [17, 10]. However, in many safe learning settings—such as safe RL or chance-constrained optimization [2, 16, 51]—the constraint functions themselves are *unknown*. Here, only noisy zeroth-order estimates of the constraint gradients are available, and feasibility becomes much harder to enforce. Most existing work in this regime focuses on convex problems and relies on primal–dual schemes [55, 13, 50, 41, 32, 38, 34, 15, 27]. For nonconvex problems, guarantees are scarce: some works provide only empirical evidence [55], while others require solving expensive convex subproblems at each step [41, 32]. Feasibility in these settings remains fragile, typically degrading as noise in the gradient estimators increases [42].

In contrast, in the first-order (FO) setting, SQP methods are known to be highly effective for nonconvex constrained optimization. They often outperform primal–dual approaches in practice, particularly when the number of constraints is small relative to the dimension of the decision variable [22, 35, 52, 3]. Another complementary line of work interprets constrained optimization through a control-theoretic lens [11, 53]. Here, the optimization dynamics are viewed as a controlled system: the primal variables are states, the Lagrange multipliers are control inputs, and feasibility corresponds to stabilizing the system at an equilibrium. This perspective enables the use of nonlinear control tools, such as *feedback linearization (FL)*, to design

algorithms. Recent work shows that, under suitable conditions, FL-based schemes recover SQP-like updates and achieve strong performance on nonconvex problems [53].

Gap and motivation. Despite their promise in first-order optimization, FL and SQP approaches have not been systematically studied in the zeroth-order regime. Extending them is nontrivial: FL and SQP depends critically on precise first-order information, but in the zeroth-order setting only noisy estimates are available, breaking the mechanisms that ensure feasibility. This raises the fundamental question:

> *How can we design zeroth-order methods that handle nonconvex constrained optimization with only noisy gradient estimators, while still providing provable guarantees of constraint satisfaction?*

Our Contributions. We develop a zeroth-order constrained optimization framework that extends feedback linearization ideas to the derivative-free regime, inspired by control and dynamical systems perspective [11, 53]. First, we show how to construct an FL scheme tailored to dynamics evolving under noisy gradient information, in contrast to prior approaches that rely on convex relaxations or primal–dual surrogates. Second, we demonstrate that full Jacobians are unnecessary: it is enough to approximate a small set of Jacobian–vector products, which can be efficiently estimated via two-point zeroth-order queries. Third, we establish theoretical guarantees (Theorem 1 and Theorem 2) showing that constraint violations contract toward zero with high probability, up to controllable approximation and discretization errors. Finally, we provide empirical evidence that our method consistently achieves stronger feasibility performance than standard baselines, while achieving competitive objective values.

Notations. We use $\nabla f(x)$ to denote the gradient of a scalar function $f : \mathbb{R}^n \to \mathbb{R}$ evaluated at the point $x \in \mathbb{R}^n$ and use $\nabla^2 f(x)$ to denote its corresponding Hessian matrix. We use $J_h(x)$ to denote the Jacobian matrix of a function $h : \mathbb{R}^n \to \mathbb{R}^m$ evaluated at $x \in \mathbb{R}^n$, i.e. $[J_h(x)]_{i,j} = \frac{\partial h_i(x)}{\partial x_j}$, $i \in [m], j \in [n]$. Unless specified otherwise, we use $\|\cdot\|$ to denote the L_2 norm of matrices and vectors. We also denote $[x]_+ := \max(x, 0)$ where $\max$ is taken entrywise for a vector x.

2 Preliminaries

We begin by introducing the constrained optimization setup and reviewing prior work from a control perspective, which motivates our approach. We then highlight the challenges unique to the ZO setting.

Problem setup. We consider constrained optimization problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t. } h(x) = 0, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective and $h : \mathbb{R}^n \to \mathbb{R}^m$ encodes equality constraints. Here we assume that f, h are differentiable, and additional assumptions will be introduced where needed to support the analysis. The first-order Karush-Kuhn-Tucker (KKT) conditions are

$$-\nabla f(x) - J_h(x)^\top \lambda = 0, \quad h(x) = 0. \tag{2}$$

Here, $J_h(x)$ denotes the Jacobian of h and $\lambda \in \mathbb{R}^m$ are the Lagrange multipliers.

While we begin by focusing on equality-constrained problems for clarity of exposition, our analysis also extends to problems with *inequality* constraints, which will be studied in Section 4.

2.1 First-order Constrained Optimization: A Control Perspective

Recent works [11, 53] interpret constrained optimization from a control perspective, offering new insights and enabling novel algorithmic designs in the first-order optimization regime. The key idea is to reinterpret Eq. (2) as the equilibrium of a dynamical system. Specifically, define the updates

$$x_{t+1} - x_t = -\eta_t\big(\nabla f(x_t) + J_h(x_t)^\top \lambda_t\big), \quad y_t = h(x_t), \tag{3}$$

where x_t is the system state, y_t the constraint output, and λ_t the control input.

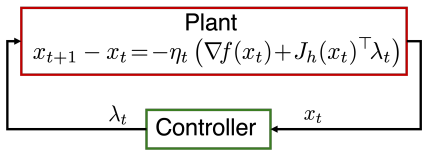


Figure 1: Control Perspective for constrained optimization.

At any equilibrium $(x^\star, \lambda^\star)$ of Eq. (3), we have $\nabla f(x^\star) + J_h(x^\star)^\top \lambda^\star = 0$. If, in addition, $x^\star$ is feasible (i.e., $h(x^\star) = 0$), then $(x^\star, \lambda^\star)$ satisfies the KKT conditions Eq. (2). Hence, the control objective is to design λ_t so that the closed-loop dynamics drive $y_t \to 0$ and stabilize x_t at a feasible equilibrium (see Fig. 1).

To design the controller λ_t to reach a feasible equilibrium, we next introduce the feedback linearization (FL) approach, which is the main focus of this paper.

Feedback linearization (FL). FL is a classical control technique for stabilizing nonlinear systems of the form

$$x_{t+1} - x_t = F(x_t) + G(x_t)\lambda_t, \tag{4}$$

by introducing a new input that cancels the nonlinearities [29, 25]. If $G(x)$ is invertible, one may set $\lambda_t = G(x_t)^{-1}(u_t - F(x_t))$, so that the dynamics reduce to $x_{t+1} - x_t = u_t$, a linear system for which standard stabilizing controllers are available.

Recall the dynamics in Eq. (3). Writing out the constraint evolution gives

$$y_{t+1} - y_t \approx J_h(x_t)(x_{t+1} - x_t) = -\eta_t J_h(x_t)\nabla f(x_t) - \eta_t J_h(x_t)J_h(x_t)^\top \lambda_t, \tag{5}$$

where the terms can be viewed as $F(x_t)$ and $G(x_t)\lambda_t$ in Eq. (4). Choosing

$$\lambda_t = -\left(J_h(x_t)J_h(x_t)^\top\right)^{-1}\left(J_h(x_t)\nabla f(x_t) - Kh(x_t)\right),$$

cancels the nonlinear dependence and yields the linearized dynamics $y_{t+1} - y_t \approx -\eta_t K h(x_t) = -\eta_t K y_t$. By picking K Hurwitz, the constraints converge exponentially to zero.

This design gives the ***first-order feedback linearization (FO-FL)*** method:

FO-FL (Equality Constraints) [11, 53]

$$\begin{aligned} x_{t+1} - x_t &= -\eta_t \left(\nabla f(x_t) + J_h(x_t)^\top \lambda_t\right), \\ \lambda_t &= -\left(J_h(x_t)J_h(x_t)^\top\right)^{-1}\left(J_h(x_t)\nabla f(x_t) - Kh(x_t)\right). \end{aligned} \tag{6}$$

FO-FL has been shown to effectively handle nonlinear dynamics, making it well-suited for nonconvex constrained optimization. Empirical and theoretical studies [11, 46, 53] confirm its strong performance relative to primal–dual methods.

2.2 Zeroth-order Constrained Optimization: Baseline and Challenges

Problem Setup. In many learning and control problems (e.g., safe RL), the gradients of f and h are unavailable; *one can only query their values $f(x)$ and $h(x)$ at selected points x, without access to $\nabla f(x)$ or $J_h(x)$.* Zeroth-order optimization aims to solve Eq. (1) using only such queries.

The absence of first-order information motivates the use of stochastic finite-difference estimators for $\nabla f(x)$ and $J_h(x)$. A standard choice is the two-point estimator:

$$\begin{aligned} \widetilde{\nabla} f(x_t) &= \frac{n}{T_B}\sum_{i=1}^{T_B} \frac{f(x_t + r_1 u_i) - f(x_t - r_1 u_i)}{2r_1}\, u_i, \\ \widetilde{J}_h(x_t) &= \frac{n}{T_B}\sum_{i=1}^{T_B} \frac{h(x_t + r_1 u_i) - h(x_t - r_1 u_i)}{2r_1}\, u_i^\top, \end{aligned} \tag{7}$$

where u_i are drawn i.i.d. from the n-dimensional unit sphere.
These estimators are nearly unbiased in expectation when the radius r_1 is sufficiently small, but can be very noisy, particularly in high dimensions.

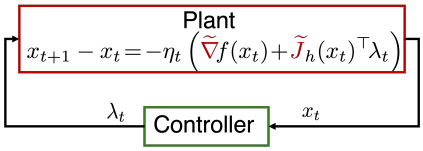


Figure 2: Control Perspective for Zeroth-Order Constrained Optimization

A Zeroth-Order Baseline and Its Limitation. Given the gradient estimator (Eq. (7)), a natural idea is to substitute these estimates directly into the FO-FL updates (Fig. 2). This yields the following ***zeroth-order baseline (ZO-baseline)***:

ZO-baseline for Equality-Constrained Optimization [11]

$$\begin{aligned} x_{t+1} - x_t &= -\eta_t \left(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top \lambda_t\right), \\ \lambda_t &= -\left(\widetilde{J}_h(x_t)\widetilde{J}_h(x_t)^\top\right)^{-1}\left(\widetilde{J}_h(x_t)\widetilde{\nabla} f(x_t) - Kh(x_t)\right). \end{aligned} \tag{8}$$

This approach has been explored in recent work on noisy or biased estimators [42]. However, it suffers from a critical drawback: constraint satisfaction is no longer guaranteed. To see this, note that the constraint dynamics become

$$h(x_{t+1}) - h(x_t) \approx J_h(x_t)(x_{t+1} - x_t)$$
$$= -\eta_t \left(J_h(x_t)\widetilde{\nabla} f(x_t) - \underbrace{\left(J_h(x_t)\widetilde{J}_h(x_t)^\top\right)\left(\widetilde{J}_h(x_t)\widetilde{J}_h(x_t)^\top\right)^{-1}}_{\neq I} \left(\widetilde{J}_h(x_t)\widetilde{\nabla} f(x_t) - Kh(x_t)\right) \right).$$

The mismatch between $J_h(x_t)$ and $\widetilde{J}_h(x_t)$ breaks the exact cancellation property of FO-FL, so the update no longer simplifies to $-Kh(x_t)$. As a result, the iterates are not guaranteed to converge to the feasible set $\{x : h(x) = 0\}$.

This limitation motivates the central question of our work:

> *Can we design zeroth-order methods that enforce constraint satisfaction despite relying on noisy gradient estimates (Eq.* (7)*)?*

In the next section, we show that a refined FL-based design yields a positive answer.

3 Feedback-Linearization-Inspired Zeroth-order Algorithm

From the previous section, we know that simply substituting noisy gradient estimates into the FO-FL scheme does not guarantee feasibility. A more careful design is required. In this section, we will present our algorithm along with the design insight and the theoretical guarantees on the constraint satisfaction.

3.1 Algorithm

Key idea. FL works by introducing a change of input that transforms nonlinear dynamics into a linear system. In the ZO setting, however, the dynamics evolve under a *noisy* gradient descent process (Fig. 2), which prevents the direct use of FO-FL. To recover feasibility, we must rederive the FL scheme for this setting.

Constraint dynamics. Consider the evolution of the constraints:

$$h(x_{t+1}) - h(x_t) \approx J_h(x_t)(x_{t+1} - x_t) = -\eta_t \Big(J_h(x_t)\widetilde{\nabla} f(x_t) - J_h(x_t)\widetilde{J}_h(x_t)^\top \lambda_t \Big). \tag{9}$$

If we choose

$$\lambda_t = -\Big(J_h(x_t)\widetilde{J}_h(x_t)^\top\Big)^{-1}\big(J_h(x_t)\widetilde{\nabla} f(x_t) - Kh(x_t)\big), \tag{10}$$

then Eq. (9) simplifies to

$$h(x_{t+1}) - h(x_t) \approx -\eta_t K h(x_t),$$

which guarantees exponential decay of constraint violations.

Challenge. Eq. (10) requires access to the exact Jacobian $J_h(x_t)$, which is not available in the ZO regime. At first glance, this seems to present a fundamental obstacle.

Insight. A closer examination reveals that the full knowledge of $J_h(x_t)$ is in fact unnecessary: we only need the products $J_h(x_t)\widetilde{\nabla} f(x_t)$ and $J_h(x_t)\widetilde{J}_h(x_t)^\top$. In other words, instead of recovering the entire Jacobian, it suffices to evaluate its inner product along $m+1$ prescribed directions: the estimated gradient $\widetilde{\nabla} f(x_t)$ and the rows of $\widetilde{J}_h(x_t)$. Crucially, these Jacobian–vector products can be efficiently approximated using standard two-point estimators as follows, thereby rendering the scheme implementable in the ZO setting:

$$\begin{aligned} G_f &= \|\widetilde{\nabla} f(x_t)\| \frac{(h(x_t + r_2 v_f) - h(x_t - r_2 v_f))}{2r_2}, \ \text{ where } v_f = \frac{\widetilde{\nabla} f(x_t)}{\|\widetilde{\nabla} f(x_t)\|} \\ [G_h]_{:,i} &= \|\widetilde{\nabla} h_i(x_t)\| \frac{(h(x_t + r_2 v_{h,i}) - h(x_t - r_2 v_{h,i}))}{2r_2}, \ \text{ where } v_{h,i} = \frac{\widetilde{\nabla} h_i(x_t)}{\|\widetilde{\nabla} h_i(x_t)\|}, \end{aligned} \tag{11}$$

Here $\widetilde{\nabla} h_i(x_t)$ is the transpose of the i-th row of $\widetilde{J}_h(x_t)$, i.e. $\widetilde{\nabla} h_i(x_t) = [\widetilde{J}_h(x_t)]_{i,:}^\top$. Then, given G_f, G_h, according to Eq. (10), we can set the Lagrangian multiplier λ to be $\lambda_t = -(G_h)^{-1}(G_f - Kh(x_t))$, which leads to our zeroth-order feedback linearization algorithm (ZOFL). The full procedure is summarized in Algorithm 1.

Algorithm 1 ZOFL (equality constraints)

Input: Initial point x_0, algorithm hyperparameters: $T_G, T_B, r_1, r_2, K, \eta_t$

1: **for** $t = 0, 1, 2, \ldots, T_G$ **do**

2: **Step 1**: Compute gradient estimation $\widetilde{\nabla} f(x_t), \widetilde{J}_h(x_t)$ using Eq. (7).

3: **Step 2**: Given the gradient estimation $\widetilde{\nabla} f(x_t), \widetilde{J}_h(x_t)$, calculate λ_t as follows

- Step 2.1: Compute G_f, G_h that approximate $J_h(x_t)\widetilde{\nabla} f(x_t), J_h(x_t)\widetilde{J}_h(x_t)^\top$ as in Eq. (11).
- Step 2.2: Set $\lambda_t = -{G_h}^{-1}(G_f - Kh(x_t))$

4: **Step 3**: Perform update $x_{t+1} = x_t - \eta_t \left(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top \lambda_t\right)$

5: **end for**

3.2 Theoretical Guarantees on Constraint Satisfaction

Building on the feedback–linearization perspective, the proposed ZOFL algorithm is designed to reduce constraint violations. We now formalize this intuition by showing that, under mild regularity assumptions, the algorithm guarantees constraint satisfaction with high probability.

We begin by stating the assumptions on boundedness, smoothness, and conditioning that will be used throughout the analysis.

Assumption 1 (Bounded iterates)**.** The trajectory $\{x_t\}$ of the algorithm lies inside a compact set $\mathcal{D} \subset \mathbb{R}^n$.

Assumption 2 (Objective regularity)**.** The objective function f is differentiable on $\mathcal{D}$ and satisfies

$$\|\nabla f(x)\| \le L_f, \ \ \forall\, x \in \mathcal{D}.$$

Assumption 3 (Constraint regularity and conditioning)**.** The constraint function h is C^3, i.e., three times continuously differentiable on $\mathcal{D}$, and there exist constants $H, \overline{L}_h, \underline{L}_h, M, R > 0$ such that for all $x \in \mathcal{D}$:

$$\|h(x)\| \le H, \ \ \|J_h(x)\| \le \overline{L}_h, \ \ \sigma_{\min}(J_h(x)) \ge \underline{L}_h, \ \ \|D^2 h(x)\| \le M, \ \ \|D^3 h(x)\|_{\text{diag}} \le R. \ \forall\, x \in \mathcal{D},$$

where $D^2 h(x), D^3 h(x)$ and $\|\cdot\|_{\text{diag}}$ are defined as in Definition 1.

Definition 1 (Second and Third-order directional derivative norm)**.** Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be C^3. Then:

- The second derivative $D^2 f(x)$ is a symmetric bilinear map:

$$D^2 f(x) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m, \quad D^2 f(x)[u, v] := \frac{\partial^2}{\partial s \partial t} f(x + su + tv)\Big|_{s=t=0}.$$

- The third derivative $D^3 f(x)$ is a symmetric trilinear map:

$$D^3 f(x) : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m, \quad D^3 f(x)[u, v, w] := \frac{\partial^3}{\partial s \partial t \partial r} f(x + su + tv + rw)\Big|_{s=t=r=0}.$$

We define the diagonal norms of $D^2 f(x)$ and $D^3 f(x)$ as follows:

$$\|D^2 f(x)\|_{\text{diag}} := \sup_{\|u\|=1} \|D^2 f(x)[u, u]\|, \qquad \|D^3 f(x)\|_{\text{diag}} := \sup_{\|u\|=1} \|D^3 f(x)[u, u, u]\|.$$

With these assumptions in place, we can formally state our main guarantee on constraint satisfaction.

Theorem 1. *Suppose Assumptions 1–3 hold and $K \succ 0$. Run Algorithm 1 with fresh finite-difference directions at each iteration. Fix $\delta \in (0, 1)$ and horizon $T_G \in \mathbb{N}$. If the batch size T_B and probe radii r_1, r_2 satisfy (cf. Appendix Lemma 5)*

$$T_B \ge 32\left(m \log\left(\frac{192\, n\, \overline{L}_h^2}{\underline{L}_h^2}\right) + \log\left(\frac{T_G}{\delta}\right)\right), \qquad r_1 \le \frac{\underline{L}_h}{8\sqrt{2\, \overline{L}_h\, R}}, \qquad r_2 \le \frac{\underline{L}_h}{8\sqrt{2\, n\, \overline{L}_h\, R}},$$

and the stepsizes obey the stability condition $0 < \eta_t\, \lambda_{\min}(K) < 1$ for all t, then with probability at least $1 - \delta$, for all $t = 1, \ldots, T_G$,

$$\|h(x_t)]\| \le \prod_{s=0}^{t-1} (1 - \eta_s \lambda_{\min}(K)) \|h(x_0)\| + C_2 r_2^2 \sum_{s=0}^{t-1} \prod_{\tau=s+1}^{t-1} (1 - \eta_\tau \lambda_{\min}(K))\, \eta_s + C_1 \sum_{s=0}^{t-1} \prod_{\tau=s+1}^{t-1} (1 - \eta_\tau \lambda_{\min}(K))\, \eta_s^2,$$

where

$$C_1 \;=\; M\Big(nL_f + \tfrac{64\,n\,\overline{L}_h\,(nL_f\overline{L}_h + \|K\|\,H)}{\underline{L}_h^2}\Big), \qquad C_2 \;=\; nR\Big(L_f + \tfrac{64\,\overline{L}_h\,(nL_f\overline{L}_h + \|K\|\,H)}{\underline{L}_h^2}\Big).$$

In particular, for a constant step $\eta_t = \eta$,

$$\|h(x_t)\| \;\le\; (1-\eta\lambda_{\min}(K))^t\,\|h(x_0)\| \;+\; \frac{C_2 r_2^2}{\lambda_{\min}(K)} \;+\; \frac{C_1\eta}{\lambda_{\min}(K)}. \tag{12}$$

For a diminishing step $\eta_t = \eta/\sqrt{t}$,

$$\|h(x_t)\| \;\le\; e^{-\eta\lambda_{\min}(K)(\sqrt{t}-1)}\,\|h(x_0)\| \;+\; \frac{2e\,C_2 r_2^2}{\lambda_{\min}(K)} \;+\; \frac{C_1\eta\,e^{2-\eta\sqrt{t}}}{\lambda_{\min}(K)} \;+\; \frac{2e\,C_1\eta}{\lambda_{\min}(K)\sqrt{t+1}}. \tag{13}$$

Remark 1 (Interpretation of Constraint Violation Bound)**.** We now unpack the meaning of the bound in Eq. (12). The first term, $(1-\eta\lambda_{\min}(K))^t\|h(x_0)\|$, decays exponentially in t to zero. This reflects the core effect of the FL design: in the absence of estimation or discretization errors, the constraint dynamics reduce to a simple stable linear system, driving violations to zero at a geometric rate. In this sense, ZOFL inherits the strong feasibility guarantees of first-order FL.

The second term, $\frac{C_2 r_2^2}{\lambda_{\min}(K)} \sim O(r_2^2)$, arises from replacing the exact Jacobian-vector products $J_h(x_t)\nabla f(x_t)$ and $J_h(x_t)J_h(x_t)^\top$ with their ZO approximations G_f, G_h. Because these approximations are based on finite-difference probing with radius r_2, the residual scales quadratically in r_2. This error is fully controllable: if function evaluations of f, h are exact, one can make this term arbitrarily small by shrinking r_2, up to the limits of numerical precision. Thus, this term does not represent a fundamental barrier but rather a trade-off between accuracy and evaluation cost.

The third term, $\frac{C_1\eta}{\lambda_{\min}(K)} \sim O(\eta)$, comes from higher-order terms in the Taylor expansion of the constraint dynamics. Unlike the approximation error, this residual is intrinsic to the Euler discretization used in ZOFL: even with perfect gradient and Jacobian information, a fixed step size η produces a non-vanishing bias. This is the main bottleneck for achieving exact feasibility under constant step sizes.

To mitigate this discretization bias, one can use a diminishing schedule as in Eq. (13) In this case, the residual terms vanish asymptotically, and constraint violations eventually disappear. The trade-off is that the ideal contraction term slows down: instead of exponential decay, the dominant term becomes $e^{-\eta\lambda_{\min}(K)(\sqrt{t}-1)}\|h(x_0)\|$, which decreases subexponentially in t. This mirrors a common theme in stochastic optimization: stronger asymptotic guarantees are possible, but only at the cost of slower transient progress.

In summary, the bound neatly separates three effects: (i) exponential contraction from FL, (ii) a controllable $O(r_2^2)$ error from zeroth-order approximation, and (iii) an $O(\eta)$ residual from discretization. Constant stepsizes yield fast initial reduction but leave a small feasibility gap, while diminishing stepsizes remove the gap but slow down the rate. This trade-off will guide the practical choice of stepsize and probing radius.

We also note that the batch size T_B for the two-point estimator scales only with the number of constraints, $T_B \sim \widetilde{O}(m)$. Consequently, our algorithm is particularly efficient when the number of constraints is smaller than the number of variables, requiring only a small batch size at each iteration.

4 Extension to Inequality-constrained Setting

So far we have focused on equality constraints of the form as in Eq. (1). We now consider the more general problem with inequality constraints:

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t. } h(x) \le 0. \tag{14}$$

The KKT conditions are $\quad -\nabla f(x) - J_h(x)^\top\lambda = 0, \quad h(x) \le 0, \quad \lambda \ge 0, \quad \lambda^\top h(x) = 0. \tag{15}$

First-order FL algorithm. We can again view this as a control problem (Fig. 1), with dynamics

$$x_{t+1} - x_t = -\eta_t\left(\nabla f(x_t) + J_h(x_t)^\top\lambda_t\right), \quad y_t = h(x_t), \quad \lambda_t \ge 0. \tag{16}$$

Compared with the equality case, the difficulty lies in enforcing the non-negativity of multipliers and the complementary slackness condition $\lambda^\top h(x) = 0$. In [53], this is achieved by designing a more intricate FL controller:

FO-FL for Inequality-Constrained Optimization

$$\begin{aligned} x_{t+1} - x_t &= -\eta_t\big(\nabla f(x_t) + J_h(x_t)^\top \lambda_t\big)\,, \\ \lambda_t &= \arg\min_{\lambda \geq 0} \Big\{ \tfrac{1}{2}\lambda^\top J_h(x_t) J_h(x_t)^\top \lambda + \lambda^\top \big(J_h(x_t)\nabla f(x_t) - Kh(x_t)\big)\Big\}. \end{aligned} \tag{17}$$

Unlike the equality-constrained case in Eq. (6), where λ_t admits a closed-form expression, here λ_t is defined implicitly through a quadratic program. This introduces nonsmooth trajectories and complicates the extension to ZO settings.

Naive zeroth-order attempt. In the ZO regime, the dynamics become

$$x_{t+1} - x_t = -\eta_t\big(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top \lambda_t\big).$$

A natural extension of Eq. (17) is to replace gradients with their estimates, which gives:

$$\lambda_t = \arg\min_{\lambda \geq 0} \Big\{ \tfrac{1}{2}\lambda^\top J_h(x_t) \widetilde{J}_h(x_t)^\top \lambda + \lambda^\top \big(J_h(x_t)\widetilde{\nabla} f(x_t) - Kh(x_t)\big)\Big\}. \tag{18}$$

However, this quadratic form is not guaranteed to be symmetric positive definite (since $J_h(x_t)\widetilde{J}_h(x_t)^\top$ need not be symmetric), and the resulting optimization problem may be ill-posed.

Refined derivation. The key is to return to the KKT conditions of Eq. (17). For the exact (first-order) case, λ_t and an auxiliary slack variable s must satisfy

$$J_h(x_t)J_h(x_t)^\top \lambda_t + J_h(x_t)\nabla f(x_t) = Kh(x_t) + s, \quad s^\top \lambda_t = 0, \quad s \geq 0, \quad \lambda_t \geq 0.$$

In the zeroth-order regime, we mirror this structure but replace exact terms with their estimators:

$$J_h(x_t)\widetilde{J}_h(x_t)^\top \lambda_t + J_h(x_t)\widetilde{\nabla} f(x_t) = Kh(x_t) + s, \quad s^\top \lambda_t = 0, \quad s \geq 0, \quad \lambda_t \geq 0. \tag{19}$$

This system defines λ_t without requiring $J_h(x_t)\widetilde{J}_h(x_t)^\top$ to be symmetric positive definite. Our analysis confirms that Eq. (19) provides the correct formulation for ensuring feasibility.

Moreover, as in the equality-constrained case, full Jacobians are not required. It suffices to estimate the products

$$G_f \approx J_h(x_t)\widetilde{\nabla} f(x_t), \qquad G_h \approx J_h(x_t)\widetilde{J}_h(x_t)^\top,$$

which can be obtained from the two-point estimators in Eq. (11). The resulting ZOFL scheme for inequality constraints is summarized below.

Algorithm 2 ZOFL (inequality constraints)

Input: Initial point x_0, algorithm hyperparameters: $T_G, T_B, r_1, r_2, K, \eta$

1: **for** $t = 0, 1, 2, \ldots, T_G$ **do**

2: **Step 1**: Compute gradient estimation $\widetilde{\nabla} f(x_t), \widetilde{J}_h(x_t)$ using Eq. (7).

3: **Step 2**: Given the gradient estimation $\widetilde{\nabla} f(x_t), \widetilde{J}_h(x_t)$, calculate λ_t as follows

- Step 2.1: Compute G_f, G_h that approximate $J_h(x_t)\widetilde{\nabla} f(x_t), J_h(x_t)\widetilde{J}_h(x_t)^\top$ as in Eq. (11).
- Step 2.2: Solve the following equations:

$$G_h\lambda + G_f = Kh(x_t) + s, \quad s^\top \lambda = 0, \quad s \geq 0, \quad \lambda \geq 0$$

Set λ_t to be the solution for λ.

4: **Step 3**: Perform update $x_{t+1} = x_t - \eta\left(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top \lambda_t\right)$

5: **end for**

Our following theoretical analysis further validates Algorithm 2's ability to guarantee constraint satisfaction.

Theoretical guarantees. We now state the main feasibility result. The proof follows the same high-level structure as Theorem 1 but requires sharper bounds on the error terms due to the nonsmooth projection step.

Theorem 2 (Feasibility with Inequality Constraints)**.** *Under Assumption 1, 2 and 3, suppose T_B and r_1, r_2 are chosen as in Theorem 1. Then with probability at least $1-\delta$, the ZOFL algorithm for inequality constraints (Algorithm 2) satisfies*

$$\|[h(x_t)]_+\| \leq \prod_{s=0}^{t-1}(1-\eta_s\lambda_{\min}(K))\|[h(x_0)]_+\| + C_2 r_2^2 \sum_{s=0}^{t-1}\prod_{\tau=s+1}^{t-1}(1-\eta_\tau\lambda_{\min}(K))\,\eta_s + C_1\sum_{s=0}^{t-1}\prod_{\tau=s+1}^{t-1}(1-\eta_\tau\lambda_{\min}(K))\,\eta_s^2,$$

for all $t = 1, \ldots, T_G$, where $[h(x)]_+ = \max\{h(x), 0\}$ denotes the positive part of the constraint. Here the constants are

$$C_1 = M\Big(nL_f + \tfrac{64n\overline{L}_h(nL_f\overline{L}_h+\|K\|H)}{\underline{L}_h^2}\Big), \qquad C_2 = n^2\overline{L}_h^2 R\Big(\tfrac{4096n\overline{L}_h(L_f\overline{L}_h+\|K\|H)}{\underline{L}_h^4} + \tfrac{64L_f}{\underline{L}_h^2}\Big).$$

In particular, for constant step $\eta_t = \eta$,

$$\|[h(x_t)]_+\| \leq (1-\eta\lambda_{\min}(K))^t\|[h(x_0)]_+\| + \tfrac{C_2 r_2^2}{\lambda_{\min}(K)} + \tfrac{C_1\eta}{\lambda_{\min}(K)}.$$

For diminishing step $\eta_t = \eta/\sqrt{t}$, the bound improves asymptotically as in Theorem 1.

The structure of the bound mirrors the equality-constrained case: exponential contraction toward feasibility, plus two residual terms accounting for zeroth-order approximation and discretization. The detailed interpretation in Remark 1 applies here as well, with the caveat that violations are measured via $[h(x_t)]_+$ rather than $h(x_t)$.

5 Exploring midpoint methods for zeroth-order optimization

Algorithm 3 ZOFL-midpoint (equality constraints)

Input: Initial point x_0, algorithm hyperparameters: $T_G, T_B, r_1, r_2, K, \eta$

1: **for** $t = 0, 1, 2, \ldots, T_G$ **do**
2: **Step 1**: Compute gradient estimation $\widetilde{\nabla} f(x_t), \widetilde{J}_h(x_t)$ using Eq. (7).
3: **Step 2**: Given the gradient estimation $\widetilde{\nabla} f(x_t), \widetilde{J}_h(x_t)$, calculate λ_t as follows
- Step 2.1: Compute G_f, G_h that approximate $J_h(x_t)\widetilde{\nabla} f(x_t), J_h(x_t)\widetilde{J}_h(x_t)^\top$ as in Eq. (11).
- Step 2.2: Set $\lambda_t = -{G_h}^{-1}(G_f - Kh(x_t))$

4: **Step 3**: Perform update $x_{\text{mid}} = x_t - \frac{\eta}{2}\left(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top\lambda_t\right)$
5: **Step 4**: Calculate $\widetilde{\nabla} f(x_{\text{mid}}), \widetilde{J}_h(x_{\text{mid}})$ according to Eq. (7) (replace x with x_{mid}) *using the same u_i's as in Step 1*
6: **Step 5**:
- Step 5.1: Recalculate G_f, G_h that approximate $J_h(x_{\text{mid}})\widetilde{\nabla} f(x_{\text{mid}}), J_h(x_{\text{mid}})\widetilde{J}_h(x_{\text{mid}})^\top$ according to Eq. (11) (replace x with x_{mid})
- Step 5.2: Set $\lambda_t = -{G_h}^{-1}(G_f - Kh(x_t))$.

7: **Step 6**: Perform update $x_{t+1} = x_t - \frac{\eta}{2}\left(\widetilde{\nabla} f(x_{\text{mid}}) + \widetilde{J}_h(x_{\text{mid}})^\top\lambda_t\right)$
8: **end for**

In Remark 1, we pointed out that discretization error is a major bottleneck in controlling constraint violation. This error arises from approximating $h(x_{t+1}) - h(x_t)$ using only the first-order term of the Taylor expansion, leading to an $O(\eta)$ residual. A natural question, then, is whether more accurate numerical schemes can reduce this error. Motivated by this, we introduce the midpoint method from numerical analysis (cf. [47]), which achieves a discretization error of $O(\eta^2)$, and develop the midpoint variant of ZOFL (Algorithm 3). Our experiments (Figures 3(a) and 3(b)) demonstrate that this variant achieves improved constraint satisfaction compared to standard ZOFL. However, ZOFL-midpoint requires twice as many function evaluations per iteration, highlighting a trade-off between accuracy and sample efficiency. We further conjecture that the constraint violation bound under the midpoint method scales as $O(\eta^2)$, and leave a rigorous proof of this property as an open question.

6 NUMERICAL VALIDATIONS

We implement the ZOFL and ZOFL-midpoint algorithms (Algorithm 1 and 3) and compare it with the ZO-baseline method (equation 8) along with other baseline algorithms in zeroth-order constrained optimization, namely SZO-ConEx ([41]) and ZOGDA [34].

Equality Constrained. We consider the following nonconvex quadratic programming problem

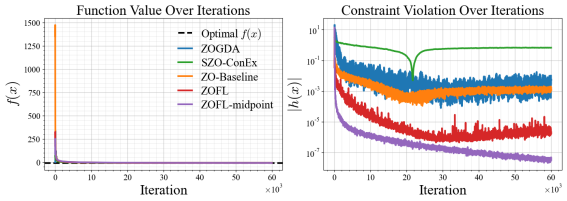
$$\min \tfrac{1}{2}x^\top x + c^\top x \qquad s.t.\ \ \tfrac{1}{2}x^\top x + a^\top x + b = 0,$$

where $x \in \mathbb{R}^{100}$, $b = 20$ and $a, c \in \mathbb{R}^{100}$ are random vectors whose entry are sampled from a standard Gaussian distribution.

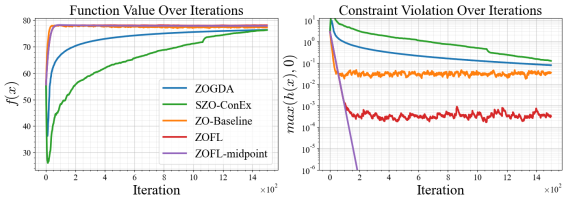
Inequality Constrained. We tested our algorithm on learning an efficient controller for building thermal regulation. We assume that the thermal dynamics to be a linear RC model ([54, 31]) $x_{t+1} = Ax_t + Bu_t + d$, where $x_t = \{x_{1,t}, x_{2,t}, \ldots, x_{n,t}\} \in \mathbb{R}^n$ represents the temperature in each building at time step t, $u_t = \{u_{1,t}, u_{2,t}, \ldots, u_{n,t}\}$ is the thermal power injection and d is the disturbances. We consider the controller $u_{i,t} = k_i x_{i,t} + b_i$ and the optimization problem is given by optimizing the control parameters: $K = \{k_i\}_{i=1}^n, b = \{b_i\}_{i=1}^n$ to minimize the thermal energy subject to the thermal comfort constraint:

$$\begin{aligned} &\min_{K,b} \tfrac{1}{T}\textstyle\sum_{t=0}^{T-1} \tfrac{1}{n}\sum_{i=1}^{n} u_{i,t}^2 \\ s.t.\quad &\tfrac{1}{T}\textstyle\sum_{t=0}^{T-1} \tfrac{1}{n}\sum_{i=1}^{n} \max\left((x_{i,t} - x_{\text{set}}), 0\right)^2 - c \le 0 \\ &x_{t+1} = Ax_t + Bu_t + d, \qquad u_i = k_i x_{i,t} + b_i, \end{aligned}$$

where we set $x_{\text{set}} = 22^{\text{o}}$C and $c = 1.5$.



3(a) Nonconvex quadratic programming with Equality Constraints



3(b) Thermal Control with Thermal Comfort Constraints

Figures 3(a) and 3(b) present the numerical results. Since diminishing step sizes often converge more slowly and are harder to tune in practice, we use a constant step size for the ZO algorithms. The left-hand plots show the cost function values, while the right-hand plots display the constraint violation. From the simulations, we observe that our algorithm, ZOFL, achieves better constraint satisfaction compared to the baseline methods while maintaining a similar cost. Moreover, ZOFL-midpoint further improves constraint satisfaction. These results suggest that, in safety-critical systems where constraint violations can have severe consequences, our algorithms are more favorable as they maintains safer operations.

7 CONCLUSIONS

We introduced a control-theoretic framework for zeroth-order constrained optimization, extending feedback linearization ideas to the derivative-free setting. Building on this perspective, we developed zeroth-order feedback linearization (ZOFL) algorithms that provide rigorous feasibility guarantees for both equality and inequality constraints, and we proposed a midpoint discretization variant that further reduces violation. Our analysis shows that the FL perspective yields exponential contraction of constraint errors, while experiments confirm that ZOFL consistently achieves stronger feasibility with competitive objective values compared to existing baselines.

Despite these contributions, several limitations remain. Our guarantees rely on access to reasonably accurate zeroth-order oracles, and their robustness under biased or highly noisy evaluations is not yet established. Moreover, although we prove finite-time bounds on constraint satisfaction and demonstrate strong empirical behavior, formal convergence to stationary points of the underlying problem remains open. Addressing these challenges, through robust extensions, convergence analysis, and deployment in safety-critical domains, defines a promising direction for future work.

LARGE LANGUAGE MODEL (LLM) USAGE DISCLOSURE

We used a large language model (ChatGPT) to assist with language editing and clarity improvements. All technical content, analysis, and contributions were developed by the authors.

REFERENCES

[1] Luigi Acerbi and Wei Ji Ma. Practical bayesian optimization for model fitting with bayesian adaptive direct search. *Advances in neural information processing systems*, 30, 2017.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization, May 2017. URL http://arxiv.org/abs/1705.10528. arXiv:1705.10528 [cs].

[3] David Applegate, Mateo Díaz, Haihao Lu, and Miles Lubin. Infeasibility detection with primal-dual hybrid gradient for large-scale linear programming. *SIAM Journal on Optimization*, 34(1):459–484, 2024.

[4] Kartik B Ariyur and Miroslav Krstic. *Real-time optimization by extremum-seeking control*. John Wiley & Sons, 2003.

[5] Charles Audet and J. E. Dennis. A Pattern Search Filter Method for Nonlinear Programming without Derivatives. *SIAM Journal on Optimization*, 14(4):980–1010, January 2004. ISSN 1052-6234, 1095-7189. doi: 10.1137/S105262340138983X. URL http://epubs.siam.org/doi/10.1137/S105262340138983X.

[6] Charles Audet and J. E. Dennis. A Progressive Barrier for Derivative-Free Nonlinear Programming. *SIAM Journal on Optimization*, 20(1):445–472, January 2009. ISSN 1052-6234, 1095-7189. doi: 10.1137/070692662. URL http://epubs.siam.org/doi/10.1137/070692662.

[7] F. Augustin and Y. M. Marzouk. NOWPAC: A provably convergent derivative-free nonlinear optimizer with path-augmented constraints, November 2015. URL http://arxiv.org/abs/1403.1931. arXiv:1403.1931 [math].

[8] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order Nonconvex Stochastic Optimization: Handling Constraints, High-Dimensionality and Saddle-Points, January 2019. URL http://arxiv.org/abs/1809.06474. arXiv:1809.06474 [math].

[9] Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

[10] Albert S Berahas, Miaolan Xie, and Baoyu Zhou. A sequential quadratic programming method with high-probability complexity bounds for nonlinear equality-constrained stochastic optimization. *SIAM Journal on Optimization*, 35(1):240–269, 2025.

[11] V. Cerone, S. M. Fosson, S. Pirrera, and D. Regruto. A new framework for constrained optimization via feedback control of lagrange multipliers, 2024. URL https://arxiv.org/abs/2403.12738.

[12] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. ZO-AdaMM: Zeroth-Order Adaptive Momentum Method for Black-Box Optimization, October 2019. URL http://arxiv.org/abs/1910.06513. arXiv:1910.06513 [cs, math, stat].

[13] Xin Chen, Jorge I. Poveda, and Na Li. Model-Free Feedback Constrained Optimization Via Projected Primal-Dual Zeroth-Order Dynamics, June 2022. URL http://arxiv.org/abs/2206.11123. arXiv:2206.11123 [math].

[14] Xin Chen, Yujie Tang, and Na Li. Improve single-point zeroth-order optimization using high-pass and low-pass filters. In *International conference on machine learning*, pages 3603–3620. PMLR, 2022.

[15] Xin Chen, Jorge I. Poveda, and Na Li. Continuous-Time Zeroth-Order Dynamics with Projection Maps: Model-Free Feedback Optimization with Safety Guarantees, March 2023. URL http://arxiv.org/abs/2303.06858. arXiv:2303.06858 [cs, eess, math].

[16] Jingjing Cui, Zhiguo Ding, Yansha Deng, Arumugam Nallanathan, and Lajos Hanzo. Adaptive uav-trajectory optimization under quality of service constraints: A model-free solution. *IEEE Access*, 8: 112253–112265, 2020.

[17] Frank E. Curtis, Xin Jiang, and Qi Wang. Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization, August 2023. URL http://arxiv.org/abs/2308.03687. arXiv:2308.03687 [cs, math].

[18] Kwassi Joseph Dzahini, Michael Kokkolaras, and Sébastien Le Digabel. Constrained stochastic black-box optimization using a progressive barrier and probabilistic estimates. *Mathematical Programming*, 198(1):675–732, March 2023. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-022-01787-7. URL http://arxiv.org/abs/2011.04225. arXiv:2011.04225 [math].

[19] N. Echebest, M. L. Schuverdt, and R. P. Vignau. An inexact restoration derivative-free filter method for nonlinear programming. *Computational and Applied Mathematics*, 36(1):693–718, March 2017. ISSN 0101-8205, 1807-0302. doi: 10.1007/s40314-015-0253-0. URL http://link.springer.com/10.1007/s40314-015-0253-0.

[20] Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *ICML*, volume 2014, pages 937–945, 2014.

[21] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, January 2016. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-014-0846-1. URL http://link.springer.com/10.1007/s10107-014-0846-1.

[22] Nicholas IM Gould, Dominique Orban, and Philippe L Toint. Galahad, a library of thread-safe fortran 90 packages for large-scale nonlinear optimization. *ACM Transactions on Mathematical Software (TOMS)*, 29(4):353–372, 2003.

[23] Robert B. Gramacy, Genetha A. Gray, Sébastien Le Digabel, Herbert K. H. Lee, Pritam Ranjan, Garth Wells, and Stefan M. Wild. Modeling an augmented lagrangian for blackbox constrained optimization. *Technometrics*, 58(1):1–11, 2016. ISSN 0040-1706. URL https://doi.org/10.1080/00401706.2015.1014065. Num Pages: 11 Publisher: American Statistical Association.

[24] Leroy Hazeleger, Dragan Nešić, and Nathan van de Wouw. Sampled-data extremum-seeking framework for constrained optimization of nonlinear dynamical systems. *Automatica*, 142:110415, August 2022. ISSN 0005-1098. doi: 10.1016/j.automatica.2022.110415. URL https://www.sciencedirect.com/science/article/pii/S0005109822002680.

[25] Michael A Henson and Dale E Seborg. Feedback linearizing control. In *Nonlinear process control*, volume 4, pages 149–231. Prentice-Hall Upper Saddle River, NJ, USA, 1997.

[26] José Miguel Hernández-Lobato, Michael A Gelbart, Ryan P Adams, Matthew W Hoffman, and Zoubin Ghahramani. A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 17(160):1–53, 2016.

[27] Chuanhao Hu, Xuan Zhang, and Qiuwei Wu. Gradient-Free Accelerated Event-Triggered Scheme for Constrained Network Optimization in Smart Grids. *IEEE Transactions on Smart Grid*, 15(3): 2843–2855, May 2024. ISSN 1949-3061. doi: 10.1109/TSG.2023.3315207. URL https://ieeexplore.ieee.org/document/10250877/.

[28] Suk-Geun Hwang. Cauchy's interlace theorem for eigenvalues of hermitian matrices. *The American mathematical monthly*, 111(2):157–159, 2004.

[29] Alberto Isidori. *Nonlinear control systems: an introduction*. Springer, 1985.

[30] Bo'az Klartag. Super-gaussian directions of random vectors. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, pages 187–211. Springer, 2017.

[31] Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 67(12):6429–6444, 2022. doi: 10.1109/TAC.2021.3128592.

[32] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Zeroth-Order Optimization for Composite Problems with Functional Constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7453–7461, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i7.20709. URL https://ojs.aaai.org/index.php/AAAI/article/view/20709.

[33] Sijia Liu, Xingguo Li, Pin-Yu Chen, Jarvis Haupt, and Lisa Amini. ZEROTH-ORDER STOCHASTIC PROJECTED GRADIENT DESCENT FOR NONCONVEX OPTIMIZATION. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1179–1183, November 2018. doi: 10.1109/GlobalSIP.2018.8646618. URL https://ieeexplore.ieee.org/document/8646618/.

[34] Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Mingyi Hong, and Una-May O'Reilly. Min-Max Optimization without Gradients: Convergence and Applications to Black-Box Evasion and Poisoning Attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6282–6293. PMLR, November 2020. URL https://proceedings.mlr.press/v119/liu20j.html. ISSN: 2640-3498.

[35] Zhe Liu, Fahim Forouzanfar, and Yu Zhao. Comparison of sqp and al algorithms for deterministic constrained production optimization of hydrocarbon reservoirs. *Journal of Petroleum Science and Engineering*, 171:542–557, 2018.

[36] Zhuanghua Liu, Cheng Chen, Luo Luo, and Bryan Kian Hsiang Low. Zeroth-order methods for constrained nonconvex nonsmooth stochastic optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30842–30872. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24j.html.

[37] Giampaolo Liuzzi, Stefano Lucidi, and Marco Sciandrone. Sequential penalty derivative-free methods for nonlinear constrained optimization. *SIAM Journal on Optimization*, 20(5):2614–2635, 2010. doi: 10.1137/090750639. URL https://doi.org/10.1137/090750639.

[38] Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, Shankar Sastry, and Lillian Ratliff. Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6702–6734. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/maheshwari22a.html.

[39] Juliane Müller and Joshua D. Woodbury. GOSAC: global optimization with surrogate approximation of constraints. *Journal of Global Optimization*, 69(1):117–136, September 2017. ISSN 0925-5001, 1573-2916. doi: 10.1007/s10898-017-0496-y. URL http://link.springer.com/10.1007/s10898-017-0496-y.

[40] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[41] Anthony Nguyen and Krishnakumar Balasubramanian. Stochastic Zeroth-order Functional Constrained Optimization: Oracle Complexity and Applications, October 2022. URL http://arxiv.org/abs/2210.04273. arXiv:2210.04273 [math].

[42] Figen Oztoprak, Richard Byrd, and Jorge Nocedal. Constrained Optimization in the Presence of Noise, October 2021. URL http://arxiv.org/abs/2110.04355. arXiv:2110.04355 [math].

[43] Tony Pourmohamad and Herbert K. H. Lee. The Statistical Filter Approach to Constrained Optimization. *Technometrics*, 62(3):303–312, July 2020. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.2019.1638304. URL https://www.tandfonline.com/doi/full/10.1080/00401706.2019.1638304.

[44] Zhaolin Ren, Yujie Tang, and Na Li. Escaping saddle points in zeroth-order optimization: the power of two-point estimators. In *International Conference on Machine Learning*, pages 28914–28975. PMLR, 2023.

[45] Anit Kumar Sahu and Soummya Kar. Decentralized Zeroth-Order Constrained Stochastic Optimization Algorithms: Frank–Wolfe and Variants With Applications to Black-Box Adversarial Attacks. *Proceedings of the IEEE*, 108(11):1890–1905, November 2020. ISSN 1558-2256. doi: 10.1109/JPROC.2020.3012609. URL https://ieeexplore.ieee.org/document/9170539/.

[46] J. Schropp and I. Singer. A dynamical systems approach to constrained minimization. *Numerical Functional Analysis and Optimization*, 21(3-4):537–551, January 2000. ISSN 0163-0563, 1532-2467. doi: 10.1080/01630560008816971. URL http://www.tandfonline.com/doi/abs/10.1080/01630560008816971.

[47] Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

[48] Yujie Tang, Zhaolin Ren, and Na Li. Zeroth-order feedback optimization for cooperative multi-agent systems. *Automatica*, 148:110741, February 2023. ISSN 0005-1098. doi: 10.1016/j.automatica.2022.110741. URL https://www.sciencedirect.com/science/article/pii/S0005109822006070.

[49] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1356–1365. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/wang18e.html.

[50] Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H. Johansson. Linear Convergence of First- and Zeroth-Order Primal-Dual Algorithms for Distributed Nonconvex Optimization, August 2021. URL http://arxiv.org/abs/1912.12110. arXiv:1912.12110 [math].

[51] Hui Zhang and Pu Li. Chance constrained programming for optimal power flow under uncertainty. *IEEE Transactions on Power Systems*, 26(4):2417–2424, 2011.

[52] Jiawei Zhang and Zhi-Quan Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020.

[53] Runyu Zhang, Arvind Raghunathan, Jeff Shamma, and Na Li. Constrained optimization from a control perspective via feedback linearization, 2025. URL https://arxiv.org/abs/2503.12665.

[54] Xuan Zhang, Wenbo Shi, Xiwang Li, Bin Yan, Ali Malkawi, and Na Li. Decentralized temperature control via hvac systems in energy efficient buildings: An approximate solution procedure. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 936–940. IEEE, 2016.

[55] Yuke Zhou, Ruiyang Jin, Siyang Gao, Jianxiao Wang, and Jie Song. A Zeroth-Order Extra-Gradient Method for Black-Box Constrained Optimization, July 2025. URL http://arxiv.org/abs/2506.20546. arXiv:2506.20546 [math].

A Other Related Works on Derivative Free Constrained Optimization

Beyond zeroth-order (ZO) methods, several other lines of research in derivative-free constrained optimization have been developed.

One classical family of approaches is filter methods [5, 19, 43, 6, 18, 37], which are based on pattern search techniques. These methods iteratively reduce the objective function while attempting to decrease constraint violations, often through a progressive barrier function. While conceptually simple, filter methods generally rely on user-specified surrogate functions to generate candidate points and frequently require solving auxiliary subproblems. As a result, they are not easily generalizable to high-dimensional settings.

Another important class of derivative-free optimization techniques is model-based methods [39, 7, 23], which build local surrogate models of the objective and constraints and optimize them iteratively. Such methods can achieve strong performance in low- to medium-dimensional problems, but their reliance on accurate surrogate models makes them sample-intensive and thus less practical in high-dimensional scenarios.

A different perspective is offered by extremum seeking (ES) [4], which estimates gradients through deterministic perturbations of the system, typically sinusoidal probing signals, as opposed to random perturbations used in two-point estimators. The estimated gradient is then used to drive the system along a descent flow towards an extremum. ES shares a close connection with zeroth-order optimization: it can be interpreted as the continuous-time counterpart of single-point ZO methods [14]. Recent works have begun to extend ES to constrained optimization settings [24, 13], though its relationship to ZO approaches in this regime remains an open direction for future study.

Finally, Bayesian optimization (BO) represents another major branch of derivative-free optimization (see, e.g., [20, 26, 1]). BO adopts a fundamentally different philosophy: it constructs global probabilistic surrogate models (typically Gaussian processes) and leverages acquisition functions to trade off exploration and exploitation. BO is particularly well-suited for low- to medium-dimensional problems where function evaluations are costly, whereas ZO methods are more appropriate in high-dimensional regimes with relatively inexpensive evaluations.

B Proof of Theorem 1

Proof of Theorem 1. From Taylor's expansion and Assumption 3 we know that

$$y_{t+1} - y_t = J_h(x_t)(x_{t+1} - x_t) + \epsilon_t,$$

where

$$\|\epsilon_t\| \le M\|x_{t+1} - x_t\|^2 \underset{\text{Lemma 1}}{\le} \underbrace{M\left(nL_f + \frac{64n\overline{L}_h(nL_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right)}_{:=C_1}\eta_t^2.$$

We define an auxiliary variable $\lambda_t^\star := -\left(J_h(x_t)\widetilde{J}_h(x_t)^\top\right)^{-1}\left(J_h(x_t)\widetilde{\nabla} f(x_t) - Kh(x_t)\right)$, and hence

$$\begin{aligned}
&y_{t+1} - y_t = J_h(x_t)(x_{t+1} - x_t) + \epsilon_t \\
&= -\eta_t J_h(x_t)\left(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top \lambda_t\right) + \epsilon_t \\
&= -\eta_t J_h(x_t)\left(\widetilde{\nabla} f(x_t) + \widetilde{J}_h(x_t)^\top \lambda_t^\star\right) + \eta_t J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t) + \epsilon_t \\
&= -\eta_t\left(J_h(x_t)\widetilde{\nabla} f(x_t) - J_h(x_t)\widetilde{J}(x_t)^\top\left(J_h(x_t)\widetilde{J}(x_t)^\top\right)^{-1}(\widetilde{\nabla} f(x_t) + Kh(x_t))\right) \\
&\quad + \eta_t \underbrace{J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t)}_{:=\Delta_t} + \epsilon_t \\
&= -\eta_t K y_t + \eta_t\Delta_t + \epsilon_t \\
\implies &\quad \|y_{t+1}\| \le (1 - \eta_t\lambda_{\min}(K))\|y_t\| + \eta_t\|\Delta_t\| + C_1\eta_t^2
\end{aligned}$$

Further, from Lemma 2, we have that

$$\|\Delta_t\| \le C_2 r_2^2, \text{ where } C_2 = nR\left(L_f + \frac{64\overline{L}_h(nL_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right)$$

Thus we get that

$$\begin{aligned}
&\|y_{t+1}\| \le (1 - \eta_t\lambda_{\min}(K))\|y_t\| + \eta_t C_2 r_2^2 + C_1\eta_t^2 \\
\implies &\|y_t\| \le \prod_{s=0}^{t-1}(1 - \eta_s\lambda_{\min}(K))\|y_0\| + C_2 r_2^2\sum_{s=0}^{t-1}\prod_{\tau=s+1}^{t-1}(1 - \eta_\tau\lambda_{\min}(K))\eta_s + C_1\sum_{s=0}^{t-1}\prod_{\tau=s+1}^{t-1}(1 - \eta_\tau\lambda_{\min}(K))\eta_s^2
\end{aligned}$$

In particular, if η_t is set to a constant $\eta_t = \eta$, we have

$$\begin{aligned}
\|y_t\| &\le (1 - \eta\lambda_{\min}(K))^t\|y_0\| + C_2 r_2^2\sum_{s=0}^{t-1}(1 - \eta\lambda_{\min}(K))^s\eta + C_1\eta^2\sum_{s=0}^{t-1}(1 - \eta\lambda_{\min}(K))^2 \\
&\le (1 - \eta\lambda_{\min}(K))^t\|y_0\| + \frac{C_2 r_2^2}{\lambda_{\min}(K)} + \frac{C_1\eta}{\lambda_{\min}(K)}
\end{aligned}$$

If $\eta_t = \frac{\eta}{\sqrt{t+1}}$, then we have that

$$\begin{aligned}
\|y_t\| &\leq \prod_{s=0}^{t-1}\left(1-\frac{\eta\lambda_{\min}(K)}{\sqrt{s+1}}\right)\|y_0\| + C_2 r_2^2 \sum_{s=0}^{t-1}\prod_{\tau=s+1}^{t-1}\left(1-\frac{\eta\lambda_{\min}(K)}{\sqrt{\tau+1}}\right)\frac{\eta}{\sqrt{s+1}} \\
&\quad + C_1\sum_{s=0}^{t-1}\prod_{\tau=s+1}^{t-1}\left(1-\frac{\eta\lambda_{\min}(K)}{\sqrt{\tau+1}}\right)\frac{\eta^2}{s} \\
&\underset{\text{Lemma 7}}{\leq} e^{-\eta\lambda_{\min}(K)(\sqrt{t}-1)}\|y_0\| + C_2 r_2^2\eta e^{-\eta\sqrt{t}}\sum_{s=1}^{t} e^{\eta\lambda_{\min}(K)\sqrt{s}}\frac{1}{\sqrt{s}} + C_1\eta^2 e^{-\eta\lambda_{\min}(K)\sqrt{t}}\sum_{s=1}^{t} e^{\eta\lambda_{\min}(K)\sqrt{s}}\frac{1}{s} \\
&\underset{\text{Lemma 8, 9}}{\leq} e^{-\eta\lambda_{\min}(K)(\sqrt{t}-1)}\|y_0\| + \frac{2eC_2r_2^2}{\lambda_{\min}(K)} + \frac{C_1\eta e^{2-\eta\sqrt{t}}}{\lambda_{\min}(K)} + \frac{2eC_1\eta}{\lambda_{\min}(K)\sqrt{t+1}}
\end{aligned}$$

□

B.1 Bounding $\|x_{t+1} - x_t\|$

Lemma 1. *In Algorithm 1 we have that given*

$$T_B \geq 32\left(m\log\left(\frac{192\cdot n\cdot \overline{L}_h^2}{\underline{L}_h^2}\right) + \log\left(\frac{T_G}{\delta}\right)\right) \sim O\left(m\left(\log(n) + \log\left(\frac{\overline{L}_h}{\underline{L}_h}\right)\right) + \log\left(\frac{T_G}{\delta}\right)\right)$$

and

$$r_1 \leq \frac{L_h}{8\sqrt{2\overline{L}_h R}}, \quad r_2 \leq \frac{L_h}{8\sqrt{2n\overline{L}_h R}}$$

then with probability at least $1-\delta$*,*

$$\|x_{t+1} - x_t\| \leq \eta_t\left(nL_f + \frac{64n\overline{L}_h(L_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right)$$

holds for all $t = 1, 2, \ldots, T_G$

Proof. From Assumption 2 and the Cauchy mean value theorem

$$\begin{aligned}
&|f(x_t + r_1u_i) - f(x_t - r_1u_i)| \leq 2r_1\nabla f(x_t + \tilde{r}_1u_i)^\top u_i \leq 2r_1L_f, \\
\implies &\|\widetilde{\nabla} f(x_t)\| \leq nL_f.
\end{aligned}$$

Similarly, from Cauchy mean value inequality we have that

$$\|\widetilde{J}(x_t)\| \leq n\overline{L}_h, \;\; \|G_f\| \leq nL_f\overline{L}_h.$$

Further, from Lemma 5, when

$$r_1 \leq \frac{L_h}{8\sqrt{2\overline{L}_h R}}, \quad r_2 \leq \frac{L_h}{8\sqrt{2n\overline{L}_h R}},$$

we have that

$$\sigma_{\min}(G_h) \geq \frac{\underline{L}_h^2}{64}.$$

Thus

$$\|\lambda_t\| = \|G_h^{-1}(G_f - Kh(x_t))\| \leq \frac{64(L_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}$$

Finally

$$\begin{aligned}
\|x_{t+1} - x_t\| &\leq \eta_t(\|\widetilde{\nabla} f(x)\| + \|\widetilde{J}_h(x_t)\|\|\lambda_t\|) \\
&\leq \eta_t\left(nL_f + \frac{64n\overline{L}_h(nL_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right),
\end{aligned}$$

which completes the proof. □

Lemma 2. *We define an auxiliary variable* $\lambda_t^\star := -\left(J_h(x_t)\widetilde{J}_h(x_t)^\top\right)^{-1}\left(J_h(x)\widetilde{\nabla}f(x) - Kh(x)\right)$. *Under the conditions as stated in Lemma 1, we have that*

$$\|J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t)\| \le nR\left(L_f + \frac{64\overline{L}_h(L_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right)r_2^2$$

Proof. From Taylor's expansion we have that

$$\|G_h - J_h(x_t)\widetilde{J}_h(x_t)^\top\| \le n\overline{L}_h R r_2^2, \quad \|G_f - J_h(x_t)\widetilde{\nabla}f(x_t)\| \le nL_f R r_2^2, \quad \|G_f\| \le nL_f\overline{L}_h.$$

Further from Lemma 5 we have

$$\|G_h^{-1}\| \le \frac{64}{\underline{L}_h^2}.$$

And thus

$$\begin{aligned}
&\|J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t)\| \\
&= \|J_h(x_t)\widetilde{\nabla}f(x_t) - Kh(x_t) + J_h(x_t)\widetilde{J}_h(x_t)^\top G_h^{-1}(G_f - Kh(x_t))\| \\
&= \|J_h(x_t)\widetilde{\nabla}f(x_t) - G_f + (G_h - J_h(x_t)\widetilde{J}_h(x_t)^\top)G_h^{-1}(G_f - Kh(x))\| \\
&\le \|J_h(x_t)\widetilde{\nabla}f(x_t) - G_f\| + \|G_h - J_h(x_t)\widetilde{J}_h(x_t)^\top\|\|G_h^{-1}\|(\|G_f\| + \|K\|H) \\
&\le nL_f R r_2^2 + n\overline{L}_h R r_2^2 \frac{64}{\underline{L}_h^2}(L_f\overline{L}_h + \|K\|H) \\
&= nR\left(L_f + \frac{64\overline{L}_h(nL_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right)r_2^2
\end{aligned}$$

□

C PROOF OF THEOREM 2

Proof of Theorem 2. The proof follows a similar structure as the proof of Theorem 1, with some substantial changes. From Taylor's expansion and Assumption 3 we know that

$$y_{t+1} - y_t = J_h(x_t)(x_{t+1} - x_t) + \epsilon_t,$$

where

$$\|\epsilon_t\| \le M\|x_{t+1} - x_t\|^2 \underset{\text{Lemma 3}}{\le} \underbrace{M\left(nL_f + \frac{64n\overline{L}_h(L_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right)}_{:=C_1}\eta_t^2.$$

We define auxiliary variable $\lambda_t^\star, s^\star$ such that it satisfies the following sets of conditions

$$\left(J_h(x_t)\widetilde{J}_h(x_t)^\top\right)\lambda_t^\star + \left(J_h(x)\widetilde{\nabla}f(x) - Kh(x)\right) = s^\star, \;\; \lambda_t^\star \ge 0, \;\; s^\star \ge 0, \;\; (\lambda_t^\star)^\top s^\star = 0 \tag{20}$$

and hence

$$\begin{aligned}
&y_{t+1} - y_t = J_h(x_t)(x_{t+1} - x_t) + \epsilon_t \\
&= -\eta_t J_h(x_t)\left(\widetilde{\nabla}f(x_t) + \widetilde{J}_h(x_t)^\top\lambda_t\right) + \epsilon_t \\
&= -\eta_t J_h(x_t)\left(\widetilde{\nabla}f(x_t) + \widetilde{J}_h(x_t)^\top\lambda_t^\star\right) + \eta_t\underbrace{J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t)}_{:=\Delta_t} + \epsilon_t \\
&= -\eta_t(Kh(x_t) + s^\star) + \eta_t\Delta_t + \epsilon_t \\
&= -\eta_t K y_t - \eta^t s^\star + \eta_t\Delta_t + \epsilon_t
\end{aligned}$$

Since $s^\star \ge 0$, we have that

$$\|[y_{t+1}]_+\| \le (1 - \eta_t\lambda_{\min}(K))\|[y_t]_+\| + \eta_t\|\Delta_t\| + C_1\eta_t^2$$

Further, from Lemma 4, we have that

$$\|\Delta_t\| \le C_2 r_2^2, \text{ where } C_2 = n^2 \overline{L}_h^2 R \left(\frac{4096 n \overline{L}_h (L_f \overline{L}_h + \|K\| H)}{\underline{L}_h^4} + \frac{64 L_f}{\underline{L}_h^2} \right)$$

Thus the rest of the proof follows exactly the same derivation as the proof of Theorem 1, here we repeat as:

$$\begin{aligned}
&\|[y_{t+1}]_+\| \le (1 - \eta_t \lambda_{\min}(K)) \|[y_t]_+\| + \eta_t C_2 r_2^2 + C_1 \eta_t^2 \\
\implies &\|[y_t]_+\| \le \prod_{s=0}^{t-1} (1 - \eta_s \lambda_{\min}(K)) \|[y_0]_+\| + C_2 r_2^2 \sum_{s=0}^{t-1} \prod_{\tau=s+1}^{t-1} (1 - \eta_\tau \lambda_{\min}(K)) \eta_s + C_1 \sum_{s=0}^{t-1} \prod_{\tau=s+1}^{t-1} (1 - \eta_\tau \lambda_{\min}(K)) \eta_s^2
\end{aligned}$$

In particular, if η_t is set to a constant $\eta_t = \eta$, we have

$$\begin{aligned}
\|[y_t]_+\| &\le (1 - \eta \lambda_{\min}(K))^t \|[y_0]_+\| + C_2 r_2^2 \sum_{s=0}^{t-1} (1 - \eta \lambda_{\min}(K))^s \eta + C_1 \eta^2 \sum_{s=0}^{t-1} (1 - \eta \lambda_{\min}(K))^2 \\
&\le (1 - \eta \lambda_{\min}(K))^t \|[y_0]_+\| + \frac{C_2 r_2^2}{\lambda_{\min}(K)} + \frac{C_1 \eta}{\lambda_{\min}(K)}
\end{aligned}$$

If $\eta_t = \frac{\eta}{\sqrt{t+1}}$, then we have that

$$\begin{aligned}
\|[y_t]_+\| &\le \prod_{s=0}^{t-1} \left(1 - \frac{\eta \lambda_{\min}(K)}{\sqrt{s+1}} \right) \|[y_0]_+\| + C_2 r_2^2 \sum_{s=0}^{t-1} \prod_{\tau=s+1}^{t-1} \left(1 - \frac{\eta \lambda_{\min}(K)}{\sqrt{\tau+1}} \right) \frac{\eta}{\sqrt{s+1}} \\
&\quad + C_1 \sum_{s=0}^{t-1} \prod_{\tau=s+1}^{t-1} \left(1 - \frac{\eta \lambda_{\min}(K)}{\sqrt{\tau+1}} \right) \frac{\eta^2}{s} \\
&\underset{\text{Lemma 7}}{\le} e^{-\eta \lambda_{\min}(K)(\sqrt{t}-1)} \|[y_0]_+\| + C_2 r_2^2 \eta e^{-\eta \sqrt{t}} \sum_{s=1}^{t} e^{\eta \lambda_{\min}(K) \sqrt{s}} \frac{1}{\sqrt{s}} + C_1 \eta^2 e^{-\eta \lambda_{\min}(K) \sqrt{t}} \sum_{s=1}^{t} e^{\eta \lambda_{\min}(K) \sqrt{s}} \frac{1}{s} \\
&\underset{\text{Lemma 8, 9}}{\le} e^{-\eta \lambda_{\min}(K)(\sqrt{t}-1)} \|[y_0]_+\| + \frac{2 e C_2 r_2^2}{\lambda_{\min}(K)} + \frac{C_1 \eta e^{2 - \eta \sqrt{t}}}{\lambda_{\min}(K)} + \frac{2 e C_1 \eta}{\lambda_{\min}(K) \sqrt{t+1}}
\end{aligned}$$

□

C.1 BOUNDING $\|x_{t+1} - x_t\|$

Lemma 3. *In Algorithm 1 we have that given*

$$T_B \ge 32 \left(m \log \left(\frac{192 \cdot n \cdot \overline{L}_h^2}{\underline{L}_h^2} \right) + \log \left(\frac{T_G}{\delta} \right) \right) \sim O \left(m \left(\log(n) + \log \left(\frac{\overline{L}_h}{\underline{L}_h} \right) \right) + \log \left(\frac{T_G}{\delta} \right) \right)$$

and

$$r_1 \le \frac{L_h}{8 \sqrt{2 \overline{L}_h R}}, \quad r_2 \le \frac{L_h}{8 \sqrt{2 n \overline{L}_h R}}$$

then with probability at least $1 - \delta$*,*

$$\|x_{t+1} - x_t\| \le \eta_t \left(n L_f + \frac{64 n \overline{L}_h (L_f \overline{L}_h + \|K\| H)}{\underline{L}_h^2} \right)$$

holds for all $t = 1, 2, \ldots, T_G$

Proof. From Assumption 2 and the Cauchy mean value theorem

$$\begin{aligned}
&|f(x_t + r_1 u_i) - f(x_t - r_1 u_i)| \le 2 r_1 \nabla f(x_t + \tilde{r}_1 u_i)^\top u_i \le 2 r_1 L_f, \\
\implies &\|\widetilde{\nabla} f(x_t)\| \le n L_f.
\end{aligned}$$

Similarly, from Cauchy mean value inequality we have that

$$\|\widetilde{J}(x_t)\| \leq n\overline{L}_h, \;\; \|G_f\| \leq L_f\overline{L}_h.$$

Further, from Lemma 5, when

$$r_1 \leq \frac{L_h}{8\sqrt{2\overline{L}_h R}}, \quad r_2 \leq \frac{L_h}{8\sqrt{2n\overline{L}_h R}},$$

we have that

$$\sigma_{\min}(G_h) \geq \frac{\underline{L}_h^2}{64}.$$

Also, note that λ_t is given by the following sets of equations

$$G_h\lambda_t + G_f = Kh(x_t) + s, \;\; \lambda_t \geq 0, \;\; s \geq 0, \;\; \lambda_t^\top s = 0.$$

Thus let the index set $\mathcal{I}$ be $\mathcal{I} := \{i : s_i = 0\}$, then we have that

$$\begin{aligned}[\lambda_t]_{\mathcal{I}^c} &= 0\\ [G_h]_{\mathcal{I}\mathcal{I}}[\lambda_t]_{\mathcal{I}} + [G_h - Kh(x_t)]_{\mathcal{I}} &= 0.\end{aligned}$$

Thus

$$\|\lambda_t\| = \left\|[G_h]_{\mathcal{I}\mathcal{I}}^{-1}[G_f - Kh(x_t)]_{\mathcal{I}}\right\|$$

From Cauchy's Interlacing Theorem we get that

$$\sigma_{\min}([G_h]_{\mathcal{I}\mathcal{I}}) \geq \frac{\underline{L}_h^2}{64}$$

Thus

$$\|\lambda_t\| \leq \frac{64}{\underline{L}_h^2}\|G_h - Kh(x_t)\| \leq \frac{64n\overline{L}_h(L_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}$$

Finally

$$\begin{aligned}\|x_{t+1} - x_t\| &\leq \eta_t(\|\widetilde{\nabla}f(x)\| + \|\widetilde{J}_h(x_t)\|\|\lambda_t\|)\\ &\leq \eta_t\left(nL_f + \frac{64n\overline{L}_h(L_f\overline{L}_h + \|K\|H)}{\underline{L}_h^2}\right),\end{aligned}$$

which completes the proof. □

Lemma 4. *We define the auxiliary variable $\lambda_t^\star$ as in equation 20. Under the conditions as stated in Lemma 3, we have that*

$$\|J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t)\| \leq n^2\overline{L}_h^2 R\left(\frac{4096nL_f\overline{L}_h^2}{\underline{L}_h^4} + \frac{64L_f}{\underline{L}_h^2}\right)r_2^2$$

Proof. Define

$$\begin{aligned}A &= J_h(x_t)\widetilde{J}_h(x_t)^\top, & b &= J_h(x_t)\widetilde{\nabla}f(x_t) - Kh(x)\\ \Delta A &= G_h - J_h(x_t)\widetilde{J}_h(x_t)^\top, & \Delta b &= G_f - J_h(x_t)\widetilde{\nabla}f(x_t)\end{aligned}$$

By Taylor's expansion,

$$\begin{aligned}\|\Delta A\| &= \|G_h - J_h(x_t)\widetilde{J}_h(x_t)^\top\| \leq n\overline{L}_h R r_2^2,\\ \|\Delta b\| &= \|G_f - J_h(x_t)\widetilde{\nabla}f(x_t)\| \leq nL_f R r_2^2,\\ \|G_f\|,&\|J_h(x_t)\widetilde{\nabla}f(x_t)\| \leq nL_f\overline{L}_h.\end{aligned}$$

For the sake of notational simplicity, in the proof we abbreviate $\lambda_t, \lambda_t^\star$ as $\lambda, \lambda^\star$. We define $A(\alpha), b(\alpha)$ as

$$A(\alpha) := A + \alpha\Delta A, \;\; B(\alpha) = B + \alpha\Delta B$$

and define $\lambda(\alpha)$ to be the solution of

$$A(\alpha)\lambda(\alpha) + b(\alpha) = s(\alpha), \;\; \lambda(\alpha) \geq 0, \;\; s(\alpha) \geq 0, \;\; \lambda(\alpha)^\top s(\alpha) = 0 \tag{21}$$

Then it is clear that $\lambda = \lambda(1), \lambda^\star = \lambda(0)$.

We can find a sequence of $\{\alpha_i\}$ such that $0 = \alpha_0 < \alpha_1 < ... < \alpha_N = 1$ such that with in each interval $\alpha, \alpha' \in (\alpha_i, \alpha_{i+1})$, $\lambda(\alpha)$ and $\lambda(\alpha')$ shares exactly the same support, which we denote as $\mathcal{I}_i$. And from equation 21 we know that for any $\alpha \in [\alpha_1, \alpha_2]$, $\lambda(\alpha)$ can be written as follows:

$$[\lambda(\alpha)]_{\mathcal{I}_i} = [A(\alpha)]^{-1}_{\mathcal{I}_i\mathcal{I}_i} b(\alpha), \;\; [\lambda(\alpha)]_{\mathcal{I}_i^c} = 0$$

And thus we get

$$\begin{aligned}
&\|\lambda(\alpha_{i+1}) - \lambda(\alpha_i)\| \leq \|[A(\alpha_{i+1})]^{-1}_{\mathcal{I}_i\mathcal{I}_i} b(\alpha_{i+1}) - [A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i} b(\alpha_i)\| \\
&\leq \|A(\alpha_{i+1})]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \|A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| A(\alpha_{i+1}) - A(\alpha_i)\| \|b(\alpha_{i+1})\| + \|A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \|b(\alpha_{i+1}) - b(\alpha_i)\| \\
&= (\alpha_{i+1} - \alpha_i) \left(\|A(\alpha_{i+1})]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \|A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \Delta A\| \|b(\alpha_{i+1})\| + \|A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \|\Delta b\| \right) \\
&\leq (\alpha_{i+1} - \alpha_i) \left(\overline{L_h} \|A(\alpha_{i+1})]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \|A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \|b(\alpha_{i+1})\| + L_f \|A(\alpha_i)]^{-1}_{\mathcal{I}_i\mathcal{I}_i}\| \right) nRr_2^2
\end{aligned}$$

Further, from Lemma 5 and Cauchy's interlacing theorem (cf. [28]) we know that for any principal minor of $A(\alpha)$ we have

$$\|[A(\alpha)^{-1}]_{\mathcal{I}\mathcal{I}}\| \leq \frac{64}{\underline{L}_h^2}, \;\; \forall \alpha \in [0,1]$$

and clearly

$$\|b(\alpha)\| \leq nL_f \overline{L}_h + \|K\|H$$

And thus we get

$$\|\lambda(\alpha_{i+1}) - \lambda(\alpha_i)\| \leq (\alpha_{i+1} - \alpha_i) \left(\frac{4096 n \overline{L}_h (L_f \overline{L}_h + \|K\|H)}{\underline{L}_h^4} + \frac{64 L_f}{\underline{L}_h^2} \right) nRr_2^2$$

And thus

$$\|\lambda - \lambda^\star\| = \|\lambda(1) - \lambda(0)\| \leq \left(\frac{4096 n \overline{L}_h (L_f \overline{L}_h + \|K\|H)}{\underline{L}_h^4} + \frac{64 L_f}{\underline{L}_h^2} \right) nRr_2^2.$$

Thus

$$\begin{aligned}
&\|J_h(x_t)\widetilde{J}_h(x_t)^\top(\lambda_t^\star - \lambda_t)\| \leq n\overline{L}_h^2 \|\lambda_t^\star - \lambda_t\| \\
&\leq n^2 \overline{L}_h^2 R \left(\frac{4096 n \overline{L}_h (L_f \overline{L}_h + \|K\|H)}{\underline{L}_h^4} + \frac{64 L_f}{\underline{L}_h^2} \right) r_2^2
\end{aligned}$$

□

D BOUNDING $\lambda_{\min}(J_h(x_t)\widetilde{J}_h(x_t))$

Lemma 5. *In Algorithm 1, we have that given*

$$T_B \geq 32 \left(m \log \left(\frac{192 \cdot n \cdot \overline{L}_h^2}{\underline{L}_h^2} \right) + \log \left(\frac{T_G}{\delta} \right) \right) \sim O\left(m \left(\log(n) + \log \left(\frac{\overline{L}_h}{\underline{L}_h} \right) \right) + \log \left(\frac{T_G}{\delta} \right) \right),$$

then with probability at least $1 - \delta$

$$\begin{aligned}
\lambda_{\min}(J_h(x_t)\widetilde{J}_h(x_t)^\top) &\geq \frac{\underline{L}_h^2}{32} - \overline{L}_h R r_1^2 \\
\sigma_{\min}(G_h) &\geq \frac{\underline{L}_h^2}{32} - \overline{L}_h R (r_1^2 + n r_2^2).
\end{aligned}$$

for all $t = 1, 2, \ldots, T_G$

Proof. From Taylor's expansion and Assumption 3 we have that

$$\widetilde{J}_h(x_t) = \frac{n}{T_B}\sum_{i=1}^{T_b} J_h(x_t)u_i u_i^\top + \epsilon(x_t),$$

where $\|\epsilon(x_t)\| \le Rr_1^2$. Thus

$$J_h(x_t)\widetilde{J}_h(x_t)^\top = \frac{n}{T_B}\sum_{i=1}^{T_b} J_h(x_t)u_i u_i^\top J_h(x_t) + \widetilde{\epsilon}(x_t), \quad \text{where } \|\widetilde{\epsilon}(x_t)\| \le \overline{L}_h R r_1^2$$

Further, from Lemma 6 we have that when

$$T_B \ge 32\left(m\log\left(\frac{192\cdot n\cdot \overline{L}_h^2}{\underline{L}_h^2}\right) + \log\left(\frac{T_G}{\delta}\right)\right) \sim O\left(m\left(\log(n) + \log\left(\frac{\overline{L}_h}{\underline{L}_h}\right)\right) + \log\left(\frac{T_G}{\delta}\right)\right)$$

then with probability at least $1-\delta$

$$\lambda_{\min}\left(\frac{n}{T_B}\sum_{i=1}^{T_B} J_h(x_t)u_i u_i^\top J_h(x_t)\right) \ge \frac{\underline{L}_h^2}{32}, \ \ \forall\, t = 1, 2, \ldots, T_G$$

And thus

$$\lambda_{\min}(J_h(x_t)\widetilde{J}_h(x_t)^\top) \ge \frac{\underline{L}_h^2}{32} - \overline{L}_h R r_1^2$$

Further, from Taylor expansion, we have that

$$\|G_h - J_h(x_t)\widetilde{J}_h(x_t)^\top\| \le n\overline{L}_h R r_2^2$$

and thus

$$\sigma_{\min}(G_h) \ge \frac{\underline{L}_h^2}{32} - \overline{L}_h R(r_1^2 + n r_2^2)$$

□

Proving Lemma 5 will need the following fundamental theorem that uses Small-ball condition to prove anti-concentration:

Theorem 3 (Small-Ball Lower Bound on Minimum Eigenvalue on Empirical Covariance Matrix)**.** *Let $u_1, \ldots, u_N \in \mathbb{R}^n$ be i.i.d. random vectors. Suppose there exist constants $\tau > 0$, $p > 0$, and $K > 0$ such that:*

*1. (**Small-ball condition**) For all $z \in \mathbb{S}^{n-1}$:*

$$\mathbb{P}(|\langle u_i, z\rangle| \ge \tau) \ge p.$$

2. For all $z \in \mathbb{S}^{n-1}$:

$$|\langle u_i, z\rangle|^2 \le K.$$

Then for any $\delta \in (0,1)$, if:

$$N \ge \frac{8}{p^2}\left(n\log\left(\frac{24K}{\tau^2 p}\right) + \log\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1-\delta$,

$$\lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^{N} u_i u_i^\top\right) \ge \frac{\tau^2 p}{4}.$$

Proof. Define:

$$S := \frac{1}{N}\sum_{i=1}^{N} u_i u_i^\top.$$

Then for any $z \in \mathbb{S}^{n-1}$:

$$z^\top S z = \frac{1}{N}\sum_{i=1}^{N}\langle u_i, z\rangle^2.$$

Let $Z_i = \langle u_i, z\rangle^2$. By assumption, $\mathbb{P}(Z_i \geq \tau^2) \geq p$. Define indicator variables $A_i := \mathbb{I}\{Z_i \geq \tau^2\}$. Then $A_i \sim \text{Bernoulli}(\geq p)$, and:

$$z^\top S z \geq \frac{\tau^2}{N}\sum_{i=1}^{N} A_i.$$

By the Chernoff bound, for all $N \geq 1$,

$$\mathbb{P}\left(\sum_{i=1}^{N} A_i < \frac{pN}{2}\right) \leq \exp\left(-\frac{p^2 N}{8}\right).$$

Thus, with probability at least $1 - \exp\left(-\frac{p^2 N}{8}\right)$, for a fixed z,

$$z^\top S z \geq \frac{\tau^2 p}{2}.$$

Now construct an ϵ-net $\mathcal{N}_\epsilon \subset \mathbb{S}^{n-1}$ with $|\mathcal{N}_\epsilon| \leq (3/\epsilon)^n$. Using the union bound:

$$\mathbb{P}\left(\exists z \in \mathcal{N}_\epsilon : z^\top S z < \frac{\tau^2 p}{2}\right) \leq (3/\epsilon)^n \cdot \exp\left(-\frac{p^2 N}{8}\right).$$

To ensure the right hand side is $\leq \delta$, it suffices that:

$$N \geq \frac{8}{p^2}\left(n\log\left(\frac{3}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right).$$

To extend from the net to all z, note:

$$|z^\top S z - \hat{z}^\top S \hat{z}| = |(z+\hat{z})S(z-\hat{z})| \leq 2K\epsilon$$

And thus choosing $\epsilon \sim \frac{\tau^2 p}{8K}$ ensures for all $z \in \mathbb{S}^{n-1}$:

$$z^\top S z \geq \frac{\tau^2 p}{2} - 2K \cdot \epsilon \geq \frac{\tau^2 p}{4}.$$

This concludes the proof. □

The following lemma is an immediate corollary of Theorem 3.

Lemma 6. *Given a fixed matrix $A \in \mathbb{R}^{m\times n}$, and random variables $u_1, u_2, \ldots, u_N \in \mathbb{R}^n$ where u_i is sampled i.i.d. from the unit sphere, we have that when*

$$N \geq 32\left(m\log\left(192 \cdot n \cdot \kappa(AA^\top)\right) + \log\left(\frac{1}{\delta}\right)\right) \sim O\left(m(\log(n) + \log(\kappa(AA^\top)) + \log\left(\frac{1}{\delta}\right)\right).$$

where $\kappa(AA^\top) := \frac{\lambda_{\max}(AA^\top)}{\lambda_{\min}(AA^\top)}$, then with probability at least $1-\delta$,

$$\lambda_{\min}\left(\frac{n}{N}\sum_{i=1}^{N} A u_i u_i^\top A^\top\right) \geq \frac{\lambda_{\min}(AA^\top)}{32}$$

Proof. Define $v_i = \sqrt{n}Au_i$. From the property of random uniform unit sphere vectors [30] we have that for any $z \in \mathbb{R}^n$,

$$\mathbb{P}\left(|u_i^\top z| \geq \frac{1}{2\sqrt{n}}\|z\|\right) \geq 1/2$$

Thus for any $z' \in \mathbb{R}^m$,

$$\mathbb{P}\left(|v_i^\top z'| \geq \frac{1}{2}\sigma_{\min}(A)\|z'\|\right) \geq \mathbb{P}\left(|v_i^\top z'| \geq \frac{1}{2}\|Az'\|\right) = \mathbb{P}\left(|u_i^\top z| \geq \frac{1}{2\sqrt{n}}\|z\|\right) \geq 1/2$$

Thus, the vector v_i's satisfies Condition 1 in Theorem 3 with

$$\tau = \frac{\sigma_{\min}(A)}{2}, \quad p = \frac{1}{2}.$$

Further, given that $u_i \in \mathbb{S}^{n-1}$,

$$|v_i^\top z'| = \sqrt{n}|u_i^\top A^\top z'| \leq \sqrt{n}\|A\|.$$

Thus the vector v_i's satisfies Condition 2 in Theorem 3 with

$$K = n\|A\|^2.$$

Directly applying Theorem 3 will finish the proof. □

E AUXILIARIES

Lemma 7.

$$\prod_{\tau=s}^{t-1}(1 - \frac{\eta}{\sqrt{\tau+1}}) \leq e^{-\eta\left(\sqrt{t}-\sqrt{s+1}\right)}$$

Proof.

$$\prod_{\tau=s}^{t-1}(1 - \frac{\eta}{\sqrt{\tau+1}}) \leq \prod_{\tau=s}^{t-1} e^{-\frac{\eta}{\sqrt{\tau+1}}} \leq e^{-\eta\sum_{\tau=s}^{t-1}\frac{1}{\sqrt{\tau+1}}} \leq e^{-\eta(\sqrt{t}-\sqrt{s+1})}$$

□

Lemma 8.

$$\sum_{s=1}^{t} e^{\eta\sqrt{s}}\frac{1}{\sqrt{s}} \leq \frac{2}{\eta}2e^{\eta\sqrt{t+1}}$$

Proof.

$$\sum_{s=1}^{t} e^{\eta\sqrt{s}}\frac{1}{\sqrt{s}} \leq \int_{s=1}^{t+1} e^{\eta\sqrt{s}}\frac{1}{\sqrt{s}}ds = 2\int_{s=1}^{t+1} e^{\eta\sqrt{s}}d\sqrt{s} \leq \frac{2}{\eta}e^{\eta\sqrt{t+1}}$$

□

Lemma 9.

$$\sum_{s=1}^{t} e^{\eta\sqrt{s}}\frac{1}{s} \leq \frac{e^2}{\eta} + \frac{2}{\eta}\frac{1}{\sqrt{t+1}}e^{\eta\sqrt{t+1}}$$

Proof.

$$\sum_{s=1}^{t} e^{\eta\sqrt{s}}\frac{1}{s} \leq \int_{s=1}^{t+1} e^{\eta\sqrt{s}}\frac{1}{s}ds \leq 2\int_{s=1}^{t+1} e^{\eta\sqrt{s}}\frac{1}{\sqrt{s}}d\sqrt{s}$$

$$= \int_{x=1}^{\sqrt{t+1}} \frac{1}{x}e^{\eta x}dx \leq \int_{x=1}^{\frac{2}{\eta}} e^{\eta x}dx + \int_{x=\frac{2}{\eta}}^{\sqrt{t+1}} \frac{1}{x}e^{\eta x}dx$$

$$\leq \int_{x=1}^{\frac{2}{\eta}} e^{\eta x}dx + \int_{x=\frac{2}{\eta}}^{\sqrt{t+1}} \left(\frac{2}{x} - \frac{2}{\eta x^2}\right)e^{\eta x}dx \leq \frac{e^2}{\eta} + \frac{2}{\eta}\frac{1}{\sqrt{t+1}}e^{\eta\sqrt{t+1}}$$

□