Black-Box Followers, White-Box Leaders: Partial Zeroth-Order Methods for MPECs

Miriam Fischer
Department of Computing
Imperial College London
London, United Kingdom
m.fischer21@imperial.ac.uk

Dario Paccagnan
Department of Computing
Imperial College London
London, United Kingdom
d.paccagnan@imperial.ac.uk

Abstract

We study mathematical programs with equilibrium constraints, in which a leader knows their own cost function, but lacks a model of the followers' response. Instead, the leader can only query this response at specific points. While this setting precludes the use of gradient-based methods, existing zeroth-order approaches treat the composed objective *entirely* as a black box, deploying zeroth-order tools across both the leader and follower. Such approaches are inefficient, as they discard information the leader already possesses about their own cost function. In this work we instead propose to deploy zeroth-order tools *only* where they are truly needed: to handle the unknown, non-smooth followers' response. Specifically, we first propose PZOS, an algorithm that combines exact partial gradients of the leader's cost with zeroth-order Jacobian estimates of the followers' response in a chain-rule-inspired manner, and establish that it achieves a lower variance bound than the black-box baseline. Second, we introduce the partial Goldstein subdifferential, a stationarity notion tailored to this composite structure, and prove convergence of our algorithm to both standard and partial Goldstein stationary points. Finally, we validate our method on two application domains – toll optimization in routing games and defense-attack investment in security games – demonstrating consistent improvements over black-box baselines in convergence speed and objective value, with robust performance even under few queries per iteration.

1 Introduction

Mathematical programs with equilibrium constraints (MPECs) arise naturally in settings where a leader optimizes their own objective while anticipating the equilibrium response of one or more followers. In many practical instances – toll optimization in transportation networks, pricing in competitive markets, resource allocation in multi-agent systems – the leader has a precise understanding of their own cost, but possesses no analytical model of how followers will respond, and can only observe the followers' response by querying it at specific points. For instance, in toll optimization, the toll-setter knows their revenue objective precisely but only observes how drivers respond to the tolls that are in place.

Bilevel optimization problems and MPECs are commonly tackled via gradient-based methods. These methods exploit knowledge of the followers' model to obtain gradient information on the composed objective – for instance by differentiating through the lower-level optimality conditions [1, 12]. While highly effective, such methods are not applicable to the setting we consider, where the leader has no access to the followers' model.

Zeroth-order methods offer a natural alternative [4, 7, 17, 18, 20]. Specifically, they sidestep the need for derivatives by constructing gradient estimates from sole function evaluations, and provide convergence certificates to approximate stationary points. However, existing approaches tailoring

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

these tools to MPECs apply zeroth-order techniques to the entire composite objective – the leader's cost evaluated at the followers' response – treating it as a black box, even though the leader's own cost is smooth and its partial gradients are available. Such approaches discard structural information the leader already possesses, resulting in possibly high variance and slow convergence.

Against this backdrop, in this work we propose an estimator for MPECs that employs zeroth-order tools *only* where they are truly needed, namely to the followers' unknown, non-smooth response. Inspired by the chain rule, our estimator keeps the leader's partial gradients exact, and uses zeroth-order methods solely to estimate the Jacobian of the followers' response. Two technical challenges arise. First, since the followers' response is not available in closed form, partial gradients of the leader's cost and the Jacobian estimate of the followers' response must be evaluated at different points, rendering the resulting estimator *biased*. Second, owing to the non-differentiability of the followers' response, convergence to standard notions of stationarity is not achievable, and one must resort to the more general notion of (δ, ε)-Goldstein stationarity. We resolve both challenges and make the following three contributions.

C1 **Algorithm exploiting upper-level structure.** We propose `PZOS`, an algorithm that combines exact partial gradients of the leader's cost with zeroth-order Jacobian estimates of the followers' response in a chain-rule-inspired manner. Unlike `ZOS` – the black-box method that applies zeroth-order tools to the entire composed objective –, `PZOS` applies zeroth-order tools only to the followers' response, fully exploiting knowledge of the leader's cost. We prove that this yields a lower variance bound than the black-box baseline `ZOS`.

C2 **Convergence guarantees.** We prove convergence of `PZOS` to both (δ, ε)-Goldstein stationary points and (δ, ε)-partial Goldstein stationary points of the composed objective. The latter notion, which we introduce in this work, is tailored to composite objectives where the leader's cost is smooth. Unlike the standard Goldstein subdifferential, which allows variation in all components of the composed objective over a neighborhood, the partial variant fixes the partial gradients of the leader's cost at the current point and allows variation only in the Jacobian of the followers' response. Finally, we show that also `ZOS` converges to this partial notion.

C3 **Empirical validation.** Across 225 routing game instances and 427 security game instances (Section 5), `PZOS` consistently outperforms the baseline `ZOS` in convergence speed, final objective value, and variance. Notably, `PZOS` achieves robust performance from a single two-point evaluation per iteration, whereas `ZOS` requires averaging over multiple two-point evaluations to attain comparable results. Importantly, we demonstrate the broader applicability of our estimator beyond `PZOS`. In particular, we show that incorporating it into variance-reduced methods, such as that of Chen et al. [4], can empirically improve their performance as well.

The remainder of the paper is organized as follows. In Sec. 1.1, we discuss related work. In Sec. 2, we formalize the problem setting. In Sec. 3 we present `PZOS`. In Sec. 4 we introduce the notion of (partial) Goldstein stationarity and state our main convergence result. We then validate our approach on routing and security games in Sec. 5. Proofs and additional details are deferred to the appendix.

1.1 Related Work

Zeroth-order methods and Goldstein stationarity. Zeroth-order optimization addresses settings where only function evaluations are available [21]. Interestingly, in the non-smooth non-convex setting of interest to this work, convergence to Clarke stationary points is not achievable in finite time, and even convergence to near Clarke stationary points requires exponentially many queries [16, 32]. This has motivated the study of the relaxed notion of (δ, ε)-Goldstein stationarity. Lin et al. [18] first establish convergence to (δ, ε)-Goldstein stationary points via zeroth-order smoothing; Chen et al. [4] improve the sample complexity via variance reduction, while Liu et al. [20] extend to constrained settings. Most recently, Kornowski and Shamir [17] provide an algorithm with optimal convergence rate. In this context, our work aims to provide better practical performance while retaining provable convergence guarantees, as opposed to improving on these best-possible convergence rates.

Zeroth-order methods for MPECs and bilevel optimization. Classical approaches to solving MPECs, e.g., gradient-based methods [12], interior-point methods [3, 23], sequential quadratic programming [11, 22], and penalty/relaxation methods [3, 8, 14] require explicit knowledge of the followers' model and are inapplicable in our setting, where the leader can only query the followers' response. Existing zeroth-order methods for MPECs [6, 7, 27] reformulate the MPEC as a single-level

problem and apply zeroth-order methods to the entire composite objective, without distinguishing between the leader's known, smooth cost and the unknown, non-smooth followers' response y^*. Perhaps the closest work to ours is Sow et al. [25], who smooth only the inner solution mapping in bilevel optimization. However, their setting differs in two key aspects. First, their lower-level response is unconstrained, which allows Gaussian smoothing. Second, they assume the lower-level response to be differentiable, which results in convergence guarantees to standard stationary points. Neither carries over to our setting, where the constrained lower-level problem requires feasibility-preserving smoothing, and its non-differentiability necessitates the (δ, ε)-Goldstein stationarity framework.

2 Problem Formulation

We consider mathematical programs with equilibrium constraints, where a leader aims at minimizing a cost function $f(x, y)$, with $f : \mathbb{R}^{\mathrm{d_x}} \times \mathbb{R}^{\mathrm{d_y}} \to \mathbb{R}$, over their decision variable x, while anticipating the response y of one or more followers. We model the followers' response as the solution to a variational inequality parametrized by the leader's decision x, with mapping $G : \mathbb{R}^{\mathrm{d_x}} \times \mathbb{R}^{\mathrm{d_y}} \to \mathbb{R}^{\mathrm{d_y}}$ and closed, convex feasible set $\mathcal{Y} \subseteq \mathbb{R}^{\mathrm{d_y}}$. This formulation captures, e.g., Nash equilibria. The problem of interest thus reads as

$$\min_{x \in \mathbb{R}^{\mathrm{d_x}},\, y \in \mathcal{Y}} f(x, y) \quad \text{subject to} \quad \langle G(x, y), y' - y \rangle \geq 0 \quad \forall y' \in \mathcal{Y}\,. \tag{1}$$

Throughout this manuscript, we assume that the parametric variational inequality admits a unique solution for each given x, which we denote $y^*(x)$. This assumption is satisfied in a broad range of problems of interest, including Cournot competition [7], traffic routing [30], and security games [9, 15]. We write $F(x) \coloneqq f(x, y^*(x))$ for the composite objective. Crucially, since the lower-level problem is constrained by $y \in \mathcal{Y}$, the response $y^*(x)$ (and hence F) is in general non-differentiable.

We work under two standard assumptions. The first is a bounded suboptimality condition, common in zeroth-order optimization [17, 18]. The second collects regularity conditions: Lipschitz continuity and differentiability of f are standard in this literature [7, 18], with the latter enabling exploitation of the leader's partial gradients. For y^*, we assume only Lipschitz continuity.

Assumption 1 (Bounded suboptimality)**.** *Given an initial point $x_0 \in \mathbb{R}^{\mathrm{d_x}}$, there exists a constant $\Delta > 0$ such that $f(x_0, y^*(x_0)) - \inf_x f(x, y^*(x)) \leq \Delta$.*

Assumption 2 (Regularity)**.**

1. *f is L_f-Lipschitz with bounded partial derivatives $\|\nabla_x f\|, \|\nabla_y f\| \leq L_f$.*
2. *f is differentiable with L_g-Lipschitz continuous gradient.*
3. *$y^*(x)$ is L_y-Lipschitz continuous in x.*

Before introducing our approach, we remark that our results also allow for follower responses beyond the variational inequality in (1), such as approximate equilibria or responses of boundedly rational followers, as long as the followers' response y^* is unique for each x, lies in a closed, convex set $\mathcal{Y}$, and is Lipschitz continuous (Assumption 2).

3 Our Partial Gradient Zeroth-Order Estimator and Algorithm

Existing approaches to solve (1) treat the composed objective $F(x) = f(x, y^*(x))$ as a black box and apply zeroth-order methods to it directly [7]. The key idea is to construct the smooth surrogate

$$\widetilde{F}_\mu(x) \coloneqq \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[F(x + \mu u)],$$

where $\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})$ denotes the uniform distribution over the unit ball $\{u \in \mathbb{R}^{\mathrm{d_x}} : \|u\| \leq 1\}$, obtained by averaging F over random perturbations of magnitude at most $\mu > 0$. This surrogate has two crucial properties: it is smooth, enabling gradient descent [10], and its gradient $\nabla \widetilde{F}_\mu$ admits an unbiased estimator via finite differences along randomly sampled directions, namely the two-point estimator [24]

$$\widetilde{g}_t \coloneqq \frac{\mathrm{d_x}}{2\mu}\big[f(x_t + \mu v_t, y^*(x_t + \mu v_t)) - f(x_t - \mu v_t, y^*(x_t - \mu v_t))\big] v_t, \tag{2}$$

where $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ is sampled uniformly from the unit sphere $\{v \in \mathbb{R}^{\mathrm{d_x}} : \|v\| = 1\}$, which requires only evaluations of f and y^*. Gradient descent on $\widetilde{F}_\mu$ with gradient estimator (2) and update

$x_{t+1} = x_t - \alpha\widetilde{g}_t$ yields ZOS (Alg. 2 in Appendix A), which is precisely the algorithm of Lin et al. [18] applied to the composed objective F, and which serves as our baseline throughout.

The black-box estimator $\widetilde{g}_t$, as well as the algorithms [4, 17, 18, 20] utilizing it, discards the available partial gradient information on the leader's cost f. Our work addresses this inefficiency. We propose to replace *only* the followers' response $y^*(x)$ with its smooth surrogate, as this is the only component for which gradients are unavailable, while keeping the leader's cost f intact. Specifically, for a fixed parameter $\mu > 0$, we define the smoothed response

$$y_\mu(x) \coloneqq \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x+\mu u)], \qquad \text{and correspondingly} \qquad F_\mu(x) \coloneqq f(x, y_\mu(x)). \tag{3}$$

Unlike the black-box smooth surrogate, F_μ does not replace the entire objective, but evaluates the function f at the smoothed response y_μ, allowing us to exploit the partial gradients $\nabla_x f$ and $\nabla_y f$.

Building on these ideas, our algorithm PZOS (Alg. 1) proceeds as follows. At each iteration t, we sample a random direction v_t from the unit sphere $\mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ and compute an estimate H_t of the Jacobian $Jy_\mu(x_t)$ via finite differences. Note that $Jy_\mu(x)$ is a matrix, since y^* maps into $\mathbb{R}^{\mathrm{d_y}}$ rather than $\mathbb{R}$, and is the only quantity that requires a zeroth-order estimate:

$$H_t \coloneqq \frac{\mathrm{d_x}}{2\mu}\big(y^*(x_t+\mu v_t) - y^*(x_t-\mu v_t)\big)v_t^\top. \tag{4}$$

Combining this with exact partial gradients of f at $(x_t, y^*(x_t))$ gives the direction

$$g_t \coloneqq \nabla_x f(x_t, y^*(x_t)) + H_t^\top \nabla_y f(x_t, y^*(x_t)), \tag{5}$$

and we update $x_{t+1} = x_t - \alpha g_t$. PZOS thus keeps the outer algorithm of ZOS in place, but replaces its estimator $\widetilde{g}_t$ (2) with g_t (5).

Algorithm 1 Partial Zeroth-Order Smoothing (PZOS)

Input: Initial $x_0 \in \mathbb{R}^{\mathrm{d_x}}$, step size α, smoothing parameter $\mu > 0$, iterations T
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ \\ Random direction uniformly from unit sphere
  Evaluate $y^*(x_t)$, $y^*(x_t+\mu v_t)$, $y^*(x_t-\mu v_t)$
  $H_t = \frac{\mathrm{d_x}}{2\mu}\left[\left(y^*(x_t+\mu v_t) - y^*(x_t-\mu v_t)\right)v_t^\top\right]$ \\ Compute Jacobian estimate
  $g_t = \nabla_x f(x_t, y^*(x_t)) + H_t^\top \nabla_y f(x_t, y^*(x_t))$ \\ Compute gradient estimate
  $x_{t+1} = x_t - \alpha g_t$ \\ Update step (minimization); for maximization, use $+$
end for
Output: x^R where $R \in \{0, 1, \ldots, T-1\}$ is uniformly sampled

We present the first upshot of our approach, before moving to the main convergence result. Specifically, we show that g_t achieves a lower second-moment bound than its black-box counterpart $\widetilde{g}_t$, with the gap growing with the problem dimension $\mathrm{d_x}$.

Lemma 1 (Second-moment bound)**.** *Under Assumptions 1 and 2, let g_t be the estimator from* (5) *and let $\widetilde{g}_t$ be the black-box estimator from* (2)*. Then,*

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \le k_2\,\mathrm{d_x}\,L_f^2 L_y^2 + L_f^2(1+2L_y),$$
$$\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \le k_2\,\mathrm{d_x}\,L_f^2 L_y^2 + k_2\,\mathrm{d_x}\,L_f^2(1+2L_y),$$

where $k_2 \ge 1$ is a universal constant from [24, Lem. 10], and the first bound is no larger than the second for $\mathrm{d_x} \ge 1$.

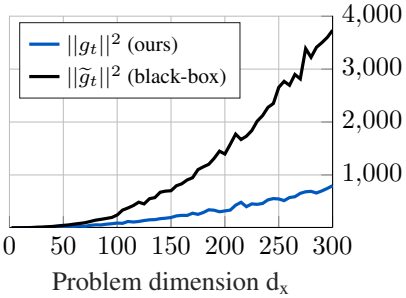


Figure 1: Empirical second moments as a function of problem dimension.

Figure 1 illustrates Lem. 1 empirically: it plots the empirical second moments $\mathbb{E}[\|g_t\|^2]$ and $\mathbb{E}[\|\widetilde{g}_t\|^2]$ as a function of the problem dimension $\mathrm{d_x}$, on security game instances (Sec. 5.2) over 3500 evaluations. Since the bounds in Lem. 1 hold uniformly in x_t, they extend directly to these unconditional averages, and the empirical gap between $\|g_t\|^2$ and $\|\widetilde{g}_t\|^2$ grows with $\mathrm{d_x}$.

4 Main Result

Below, we introduce (δ, ε)-Goldstein stationarity [18], the standard stationarity target for non-smooth, zeroth-order optimization, along with a novel partial counterpart tailored to our composite structure, and prove convergence of PZOS to both.

Clarke gradient and Jacobian. Due to the non-smoothness of F, standard gradients are unavailable and we work instead with the Clarke generalized gradient [5], defined as the convex hull of all limiting gradients at nearby differentiable points.[1] For vector-valued functions such as y^*, the analogous object is the Clarke generalized Jacobian $\partial y^*(x)$, defined likewise with limiting Jacobians. Since f is smooth, Clarke's chain rule applied to $F(x) = f(x, y^*(x))$ yields

$$\partial F(x) = \big\{\nabla_x f(x, y^*(x)) + M^\top \nabla_y f(x, y^*(x)) : M \in \partial y^*(x)\big\}. \tag{6}$$

This expression serves as the foundation for the stationarity notions introduced next.

Goldstein stationarity. A natural target for optimization is a point where the Clarke gradient contains an element of small norm – a (near) Clarke stationary point. However, finding such a point may require exponentially many queries [16]. The (δ, ε)-Goldstein stationarity notion [18] offers a tractable relaxation: rather than requiring a small element in the Clarke gradient at a single point, it requires this of the convex hull of all Clarke gradients within a δ-ball around x, which we denote $\mathbb{B}(x, \delta) = \{x' \in \mathbb{R}^{\mathrm{d_x}} : \|x' - x\| \le \delta\}$. The δ-Goldstein subdifferential is defined as:

$$\partial_\delta^{\mathrm{G}} F(x) = \mathrm{conv}\Big\{\nabla_x f(z, y^*(z)) + M^\top \nabla_y f(z, y^*(z)) : z \in \mathbb{B}(x, \delta),\ M \in \partial y^*(z)\Big\}. \tag{7}$$

In this expression, both the partial gradients of f and the Clarke Jacobian of y^* vary as z ranges over the δ-ball. Since in our setting the partial gradients of f are known and smooth, it is natural to consider a variant that fixes them at the center point and allows variation only in the Clarke Jacobian of y^*. This leads to the novel *δ-partial Goldstein subdifferential*:

$$\partial_\delta^{\mathrm{P}} F(x) = \Big\{\nabla_x f(x, y^*(x)) + M^\top \nabla_y f(x, y^*(x)) : M \in \mathrm{conv}\big\{\cup_{z \in \mathbb{B}(x,\delta)} \partial y^*(z)\big\}\Big\}. \tag{8}$$

We are now ready to define the corresponding stationarity notions.

Definition 1 (Goldstein stationarity)**.** *For $\delta, \varepsilon > 0$:*

- *x is a (δ, ε)-Goldstein stationary point of F if* $\min\{\|g\| : g \in \partial_\delta^G F(x)\} \le \varepsilon$.
- *x is a (δ, ε)-partial Goldstein stationary point of F if* $\min\{\|g\| : g \in \partial_\delta^P F(x)\} \le \varepsilon$.

The two notions are conceptually distinct: $\partial_\delta^{\mathrm{G}} F(x)$ varies all components of the Clarke gradient of F over the δ-ball, whereas $\partial_\delta^{\mathrm{P}} F(x)$ anchors the smooth part and varies only the uncertain part. Neither notion is in general stronger than the other – a point may be (δ, ε)-Goldstein stationary without being (δ, ε)-partial Goldstein stationary, and vice versa. For example, whenever x lies within a δ-neighborhood of a Clarke stationary point, the Goldstein subdifferential at x contains zero, making x trivially (δ, ε)-Goldstein stationary for every $\varepsilon \ge 0$ regardless of the objective gap; Example 1 shows that the partial Goldstein subdifferential does not necessarily suffer from this issue. Which stationary notion provides more meaningful convergence guarantees depends on the problem instance and interaction between x, y^*, and the outer objective f; we provide a discussion of both notions in Appendix B. We proceed with our main result, which establishes convergence of PZOS to both notions.

Theorem 1 (Convergence of PZOS)**.** *Let x^R be chosen uniformly at random from the iterates $\{x_0, \ldots, x_{T-1}\}$ of Alg. 1 under Assumptions 1 and 2, with smoothing parameter $\mu < 1$, step size $\alpha = \Theta(\mu^{1/2} / (T^{1/2}\ \mathrm{d_x}^{3/4}))$. Let $C_p = L_g(1 + L_y)L_y$ and $C_f = (1 + L_y)L_g(1 + 2L_y)$. Then:*

- ***Partial Goldstein stationarity:*** *Choosing $\mu = \min\Big(\delta, \frac{\varepsilon}{\sqrt{2}C_p}\Big)$, x^R is, in expectation, a (δ, ε)-partial Goldstein stationary point of F after $T = \mathcal{O}\Big(\mathrm{d_x}^{3/2} / (\mu\, \varepsilon^4)\Big)$ iterations, i.e.,*

$$\mathbb{E}\left[\min\left\{\|g\| : g \in \partial_\delta^P F(x^R)\right\}\right] \le \varepsilon.$$

[1] Formally, the Clarke generalized gradient of a Lipschitz continuous function $h : \mathbb{R}^n \to \mathbb{R}$ at x is $\partial h(x) = \mathrm{conv}\{\lim_{x_k \to x} \nabla h(x_k) : h \text{ is differentiable at } x_k\}$, and the Clarke generalized Jacobian of a Lipschitz continuous vector-valued function $h : \mathbb{R}^n \to \mathbb{R}^m$ is defined analogously as $\partial h(x) = \mathrm{conv}\{\lim_{x_k \to x} Jh(x_k) : h \text{ is differentiable at } x_k\}$.

- ***Goldstein stationarity:*** *Choosing* $\mu = \min\left(\delta, \frac{\varepsilon}{2C_f}\right)$, x^R *is, in expectation, a* (δ, ε)*-Goldstein stationary point of* F *after* $T = \mathcal{O}\left({\mathrm{d_x}}^{3/2}/(\mu\,\varepsilon^4)\right)$ *iterations, i.e.,*

$$\mathbb{E}\left[\min\left\{\|g\| : g \in \partial_\delta^G F(x^R)\right\}\right] \leq \varepsilon.$$

Before providing a sketch of the proof, we complement this result with a convergence guarantee for ZOS to partial Goldstein stationarity. Convergence of ZOS to Goldstein stationarity after $\mathcal{O}({\mathrm{d_x}}^{3/2}\,\delta^{-1}\varepsilon^{-4})$ iterations immediately follows from Lin et al. [18], choosing $\mu = \delta$. Taken together, both ZOS and PZOS converge to partial Goldstein and Goldstein stationarity by choosing the smoothing parameter μ suitably. Table 1 in Appendix A.2 summarizes the resulting convergence rates.

Proposition 1 (Convergence of ZOS to (δ, ε)-partial Goldstein stationary point)**.** *Under Assumptions 1 and 2, let* x^R *be chosen uniformly at random from the iterates of* ZOS *with step size* $\alpha = \Theta(\mu^{1/2}/(T^{1/2}\,{\mathrm{d_x}}^{3/4}))$ *and smoothing parameter* $\mu = \min(\delta, \varepsilon/(2\widetilde{C}_p)) < 1$*, where* $\widetilde{C}_p = L_g(1+L_y)^2$*. Then, after* $T = \mathcal{O}({\mathrm{d_x}}^{3/2}/(\mu\varepsilon^4))$ *iterations,* $\mathbb{E}[\min\{\|g\| : g \in \partial_\delta^P F(x^R)\}] \leq \varepsilon$*.*

The proof is provided in Appendix C.3.3.

Proof sketch of Theorem 1. The proof of Thm. 1 proceeds in three main steps. We highlight here the main challenges and our approach to resolve them. A detailed proof can be found in Appendix C.

Step 1: Smoothness of y_μ (Appendix C.1). We extend classical smoothing results for scalar functions to the vector-valued setting $\mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y}}$, establishing that $y_\mu(x)$ is continuously differentiable with an explicit Jacobian formula.

Step 2: Smoothness of F_μ and biased estimator (Appendix C.2). We show that F_μ is L_F-smooth, enabling a descent lemma analysis. However, since $y_\mu(x_t)$ is not available in closed form, the partial gradients of f cannot be evaluated at $(x_t, y_\mu(x_t))$, so g_t is a *biased* estimator of $\nabla F_\mu(x_t)$, with bias of order μ. This bias cannot be eliminated: even an unbiased estimator $\hat{y}$ of $y_\mu(x_t)$ would yield a biased estimator of $\nabla F_\mu(x_t)$, since $\mathbb{E}[\nabla_y f(x_t, \hat{y})] \neq \nabla_y f(x_t, y_\mu(x_t))$ in general. This bias also does not vanish as $T \to \infty$, and would prevent convergence to (δ, ε)-Goldstein stationary points if μ were set equal to δ as in the standard black-box approach, since the bias term $\mathcal{O}(\mu) = \mathcal{O}(\delta)$ would then exceed the target tolerance ε whenever $\delta \gg \varepsilon$. We circumvent this by decoupling μ from δ, choosing μ sufficiently small relative to ε so that the bias term $\mathcal{O}(\mu)$ remains below the target tolerance, while still ensuring $\mu \leq \delta$ to transfer convergence guarantees from F_μ to F.

Step 3.A: Convergence to partial Goldstein stationarity (Appendix C.3). Since g_t is a biased estimator, the standard approach of establishing convergence for the surrogate function is unavailable. Instead, we show via a hyperplane separation argument that $Jy_\mu(x) \in \partial_\delta^{\mathrm{G}}\, y^*(x)$ for $\delta \geq \mu$. Since $\partial_\delta^{\mathrm{P}} F(x)$ fixes the partial gradients of f at the center point $(x, y^*(x))$ and varies only the Clarke Jacobian of y^*, this implies $\mathbb{E}[g_t \mid x_t] \in \partial_\delta^{\mathrm{P}} F(x_t)$ directly. Combining this containment with the descent lemma for F_μ then yields convergence to (δ, ε)-partial Goldstein stationarity.

Step 3.B: Convergence to Goldstein stationarity (Appendix C.3). The Goldstein subdifferential $\partial_\delta^{\mathrm{G}} F(x)$ varies both the partial gradients of f and the Clarke Jacobian of y^* over the δ-ball, so $\mathbb{E}[g_t \mid x_t]$ does not lie directly in $\partial_\delta^{\mathrm{G}} F(x_t)$. Instead, we apply Carathéodory's theorem to show that $\nabla F_\mu(x)$ lies within distance $C_f\mu$ of $\partial_\delta^{\mathrm{G}} F(x)$, and carry out the descent lemma analysis on $\nabla F_\mu(x_t)$ directly. Choosing μ sufficiently small then ensures the additional approximation error remains below the target tolerance ε, which yields the desired result.

5 Experiments

We validate our method on two problem classes chosen to reflect our problem setting: toll optimization in routing games (Sec. 5.1) and defense-attack investment in security games (Sec. 5.2). Beyond evaluating PZOS, we investigate the broader applicability of our estimator by incorporating it into other zeroth-order algorithms, demonstrating improvements over their black-box counterparts (Sec. 5.3). Finally, we compare the algorithms in terms of oracle calls to y^*, highlighting the computational advantages of PZOS.

5.1 Routing games

A central authority sets tolls $\tau \in \mathbb{R}^{|E|}$ on the edges of a directed graph $G = (V, E)$ and observes the resulting traffic flow, but has no analytical model of how drivers respond. The network is used by k user groups, each with a fixed origin-destination pair, a travel demand $d_i > 0$, and a sensitivity $s_i > 0$ to tolls, which is unknown to the central authority. Each user group perceives the cost of edge $e \in E$ as $\ell_e(f_e) + \tau_e/s_i$, where the latency ℓ_e depends on the aggregated edge flow f_e on that edge and represents the travel time, and τ_e/s_i represents the perceived toll cost. Each user group routes their demand on minimum perceived-cost paths only, resulting in a Wardrop equilibrium [30]. Under standard conditions, the equilibrium edge flow $f^*(\tau) \in \mathbb{R}^{|E|}$ is unique for given toll vector τ. The central authority sets tolls to maximize the resulting toll revenue:

$$\max_{\tau \in \mathbb{R}^{|E|}} \sum_{e \in E} \tau_e \cdot f_e^*(\tau) - \lambda \|\tau\|^2, \tag{9}$$

where $\lambda\|\tau\|^2$ is a regularization term to discourage excessive tolling. The outer objective is continuously differentiable in (τ, f), whereas the equilibrium edge flow $f^*(\tau)$ is in general non-differentiable in τ [19]; non-differentiability arises when small perturbations of τ change the set of minimum-cost paths used by some user group. This fits our problem setting as the leader's objective is smooth and known, while the followers' response $f^*(\tau)$ is non-smooth and accessible only through queries.

Baseline and Setup. We compare PZOS against ZOS, which treats the entire problem setting as a black box. Both algorithms share the outer algorithm of Lin et al. [18] and differ *solely* in whether they discard (ZOS, estimator $\widetilde{g}_t$ (2)) or utilize (PZOS, estimator g_t (5)) the partial gradients of the leader's objective, so that any difference in performance is attributable to the gradient estimator alone. We argue that this is the most natural comparison for our research question whether utilizing these gradients improves performance, as any difference in performance is then attributable to the gradient estimator alone, and cannot be caused by the employed outer algorithm.

Both algorithms are run using 225 randomly generated network instances with 6-25 vertices, 12-119 edges, and 2-5 user groups, with demands in $\mathcal{U}([3, \dots, 12])$, and toll sensitivities in $\mathcal{U}([0.1, 2])$. Edge latency functions are affine, $\ell_e(f_e) = a_e f_e + b_e$, with $a_e \in \mathcal{U}([1, 5])$ and $b_e \in \mathcal{U}([1, 10])$. We initialize each instance with toll vector $\tau_0 = 0$. We set $\lambda = 1$, $\mu = 0.5$, $\alpha = 0.7/|E|$, and $T = 150$ iterations.[2] For a given instance, both algorithms are run with the same starting point and noise directions $v_t \sim \mathcal{U}(\mathbb{S}^{d_x})$, so that performance differences are not attributable to sampling variation. Since objective values depend on problem dimensions, objectives are normalized to $[0, 1]$ via $(F(x_t) - F_{\text{lowest}})/(F_{\text{highest}} - F_{\text{lowest}})$, where F_{highest} (F_{lowest}) are the highest (lowest) objective values achieved by either algorithm on that run.[3]

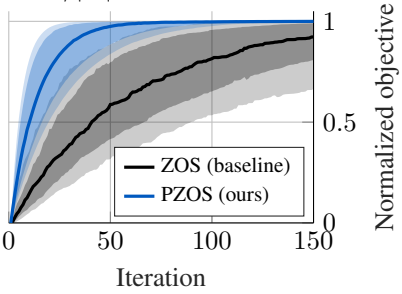


Figure 2: Algorithm performance on heterogeneous routing games. Median normalized objective trajectories (higher is better) with interquartile range (dark shading) and 10th–90th percentile range (light shading).

PZOS outperforms ZOS across all runs. PZOS consistently outperforms ZOS across all 225 runs in convergence speed and median objective value (Fig. 2). It also exhibits lower objective variability across runs, which is consistent with the tighter variance bound shown in Lem. 1. The performance gap is consistent across runs: the interquartile range of PZOS lies entirely above that of ZOS, with the best runs of ZOS barely matching the worst runs of PZOS. Figure 3 further confirms this on a per-instance basis: at every snapshot $t \in \{10, 25, 50\}$, every instance lies above the diagonal, showing that PZOS returns higher (better) objective values on all routing instances. As PZOS requires one oracle call more per iteration than ZOS, we also compare them per oracle call (Fig. 9 in Appendix A.3, left panel). PZOS continues to outperform ZOS per oracle call, confirming the gains originate from exploiting the known structure of the leader's objective, not from an additional oracle call.

[2]The step size was selected to give both algorithms stable and competitive performance and was not tuned specifically to PZOS. If anything, α was chosen conservatively relative to what PZOS alone could achieve, in order to stabilize ZOS. A sensitivity analysis over μ is provided in Figure 12, confirming that PZOS outperforms ZOS across all tested values of μ, with neither algorithm showing strong sensitivity to this parameter.

[3]The experiments are run on a 13-inch MacBook Pro with 2.3 GHz Dual-Core Intel Core i5 processor and 8 GB 2133 MHz RAM. The code is implemented in MATLAB.

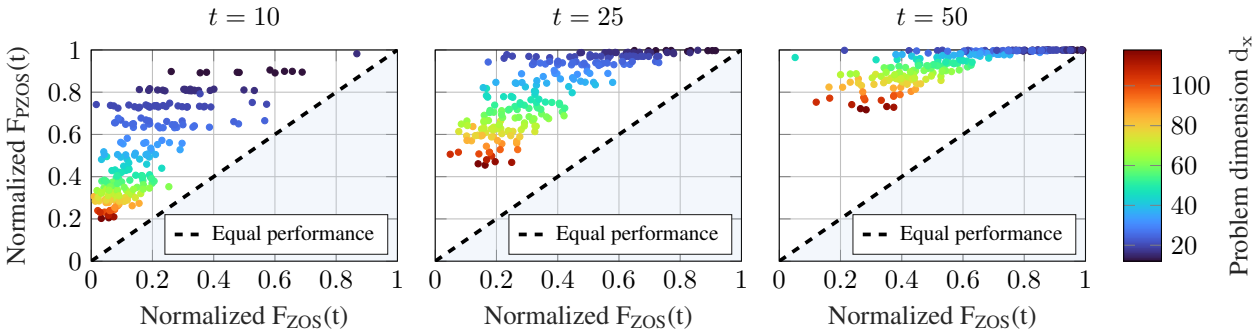


Figure 3: Normalized objective value (higher is better) of the baseline ZOS (x-axis) and our method PZOS (y-axis) per routing game instance, at iteration 10 (left panel), 25 (central panel), and 50 (right panel). Points above the diagonal favor PZOS.

In addition, we compare PZOS to a batched variant ZOS-B of ZOS, where the estimator is evaluated as the average of Q independent estimates $\{\widetilde{g}_{t,q}\}_{q=1}^{Q}$. Even with four oracle calls per iteration compared to PZOS's three, ZOS-B does not close the gap to PZOS (Figs. 8 and 9 in Appendix A.3).

5.2 Security games

We consider a Stackelberg security game in which a defender (leader) allocates investments across n targets to minimize the damage from an anticipated attack of an attacker (follower). The defender moves first, committing to a defense investment $x \in \mathbb{R}^n_{\geq 0}$ across the n targets. The attacker observes x and responds by allocating attack efforts $y \in \{y \in \mathbb{R}^n_{\geq 0} : \sum_i y_i \leq B_a\}$, where $\sum_i y_i \leq B_a$ is an attack budget constraint. An attack on target i is successful with probability given by the contest success function [13, 28] $p_i(x_i, y_i) = \frac{y_i}{x_i + y_i + b_i}$, where $b_i > 0$ is a baseline security level. The defender minimizes expected damage plus investment cost, anticipating the attacker's best response, which maximizes expected gain minus effort cost:

$$\begin{aligned} &\min_{x \geq 0} \ \sum_{i=1}^{n} v_i \, p_i(x_i, y_i^*(x)) + \tfrac{c_i^{\mathrm{D}} x_i^2}{2} && \text{(defender)} \\ &\text{s.t. } \ y^*(x) = \operatorname*{arg\,max}_{y \geq 0,\ \sum_{i=1}^n y_i \leq B_a} \ \sum_{i=1}^{n} w_i \, p_i(x_i, y_i) - \tfrac{c_i^{\mathrm{A}} y_i^2}{2} && \text{(attacker)} \end{aligned} \tag{10}$$

where $w_i, v_i > 0$ are the values of target i to the attacker and the defender, and $c_i^{\mathrm{A}}, c_i^{\mathrm{D}} > 0$ are quadratic effort and investment cost coefficients. In particular, the attacker's valuation and cost parameters are unknown to the defender, so the defender has no analytical model of the attacker's strategy and can only commit to an investment and observes the resulting attack. The attacker's objective is strictly concave, which guarantees a unique response $y^*(x)$ for each x. Further, y^* is continuous but possibly non-differentiable whenever the set of targets receiving positive attack effort changes under a binding budget constraint.[4] As in routing games, this fits our problem setting, as the leader's objective is smooth and known, while the attacker's response $y^*(x)$ is non-smooth.

Baseline and Setup. We use the same baseline and normalization as in Sec. 5.1; since the defender minimizes, lower normalized values are better. We use 427 randomly generated instances with $n \in [2, 300]$ targets, $v_i, w_i \sim \mathcal{U}([1, 5])$, $b_i \sim \mathcal{U}([0.1, 0.5])$, $c_i^{\mathrm{D}}, c_i^{\mathrm{A}} \sim \mathcal{U}([0.1, 0.3])$. The attacker's budget is set to ensure the budget constraint is binding.[5] We set $\mu = 0.1$, $T = 500$, and diminishing step size $\alpha_t = 0.05/\sqrt{t}$ ($n \leq 100$) or $0.03/\sqrt{t}$ ($n > 100$).[6] The initial defense investment x_0 is the defender's optimal investment assuming the attacker exerts effort $y_i = b_i$ on each target, which

[4] Uniqueness and continuity of $y^*(x)$ follow from strict concavity of the attacker's objective and Berge's Maximum Theorem [26, Thm. 9.17].

[5] Specifically, $B_a = \frac{1}{2} \sum_i y_{\mathrm{sym},i}(x_{\mathrm{sym}})$, where $y_{\mathrm{sym},i}(x_{\mathrm{sym}})$ is the attacker's optimal unconstrained effort assuming the defender mirrors the attacker's effort, $x_{\mathrm{sym}} = y_{\mathrm{sym}}$. We scale this by 0.5, which in our experiments is sufficient to ensure the constraint is binding.

[6] The smaller step size for $n > 100$ was chosen to ensure stability of ZOS at higher dimensions, where its larger estimator variance would otherwise lead to instability with the step size used for $n \leq 100$; PZOS remains stable under both step sizes.

requires no knowledge of the attacker's budget or cost structure. Both algorithms enforce $x \geq 0$ via projection after each gradient step; in practice, however, the projection is not activated.

PZOS outperforms ZOS across all runs. The results mirror those of the routing games: PZOS consistently outperforms ZOS across runs in convergence speed, median objective value and variability across instances (Fig. 4), over a wide range of problem dimensions $n \in [2, 300]$. These findings confirm that the advantages of PZOS are not problem-specific, but reflect a systematic benefit of exploiting the known leader's objective.

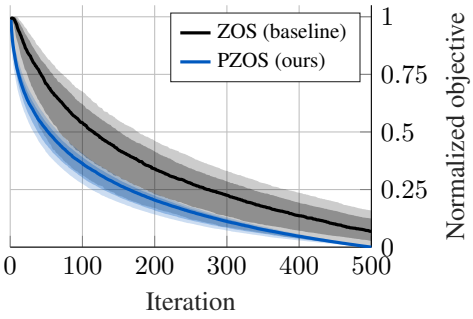


Figure 4: Algorithm performance on security games. Median normalized objective with interquartile range (dark shading) and 10th–90th percentile (light shading).

Performance gap increases with dimension. Figure 5 examines how the two algorithms behave as the problem dimension grows. While PZOS maintains a low and stable objective value across all dimensions, ZOS degrades substantially as n increases, with the gap between the two methods widening monotonically at every snapshot $t \in \{50, 150, 250\}$. This suggests that confining zeroth-order estimation to the Jacobian of y^* instead of the entire composed objective becomes increasingly significant as the problem dimension grows. This is consistent with Lem. 1, under which the second-moment gap between g_t and $\widetilde{g}_t$ grows linearly in the problem dimension.

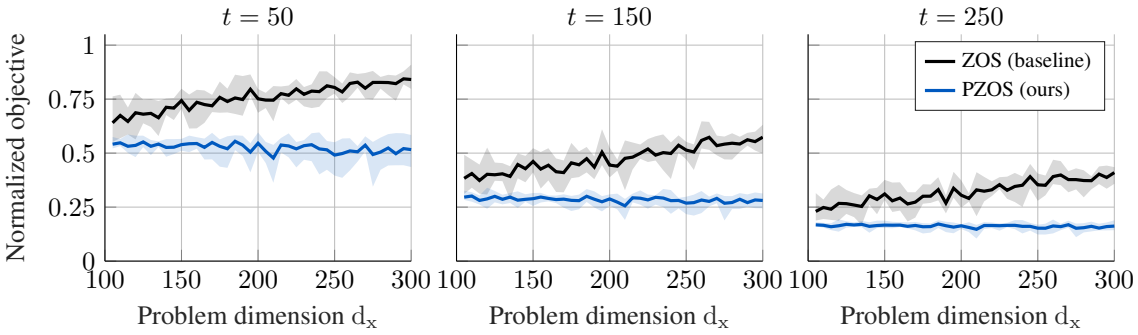


Figure 5: Normalized objective value (lower is better) achieved by the baseline ZOS (black) and our method PZOS (blue) as a function of problem dimension, for 287 security game instances with targets $n \in [100, 300]$, at iterations 50 (left panel), 150 (center), 250 (right). Shading shows min/max range.

5.3 Using our estimator g_t within other outer algorithms

As noted in Sec. 3, ZOS is the algorithm of Lin et al. [18], using $\widetilde{g}_t$ as a black-box gradient estimator. PZOS keeps the outer algorithm of Lin et al. [18] intact, but replaces $\widetilde{g}_t$ with g_t, which utilizes the partial gradients of f. Subsequent methods, including those of Chen et al. [4], Kornowski and Shamir [17], refine the outer algorithm while retaining the black-box estimator $\widetilde{g}_t$. To separate the effect of the estimator from that of the outer algorithm, we additionally consider the refined algorithm of Chen et al. [4], denoted `Chen`, alongside our variant `P-Chen`, which replaces $\widetilde{g}_t$ with g_t. We evaluate both on the two problem classes.

Setup. The method of Chen et al. [4] recomputes the gradient estimate from a large batch of b' samples every m iterations and updates it recursively from mini-batches of b samples in between. For routing games, we use the same 225 instances and parameters as in Sec. 5.1, with $b' = 12$, $m = 4$ and $b = 2$. For security games, we use 240 random instances with $n \in [5, 150]$ targets, generated and parametrized as in Sec. 5.2 (including μ and step sizes), with $b' = 4$, $m = 2$ and $b = 1$, and run all methods for $T = 250$ iterations. Objectives are normalized jointly over the four methods.

Performance by iteration. We first observe that incorporating our estimator into the refined algorithm of Chen et al. [4] also improves its performance, with `P-Chen` outperforming `Chen` (Fig. 10 in Appendix A.3). This mirrors the improvement of PZOS over ZOS and demonstrates that the benefits of g_t extend beyond the outer algorithm of Lin et al. [18]. In addition, we observe that the additional

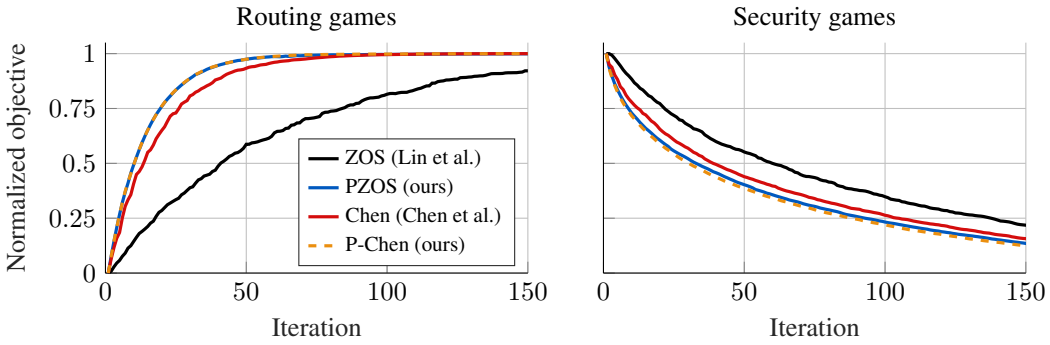


Figure 6: Median normalized objective for ZOS, PZOS, Chen, and P-Chen, on routing (left, higher is better) and security games (right, lower is better) instances of Sec. 5.3. ZOS and Chen use black-box estimator $\widetilde{g}_t$ (2), PZOS and P-Chen use our estimator g_t (5).

variance reduction offered by the refined outer algorithm of Chen et al. [4] yields only marginal gains over the outer algorithm of Lin et al. [18] when using our estimator g_t. While Chen substantially outperforms ZOS with the black-box estimator $\widetilde{g}_t$, P-Chen performs only slightly better than PZOS (Fig. 6). One possible explanation is that the smaller second-moment bound of g_t (Lem. 1) leaves limited room for further improvements through variance reduction in the outer algorithm. Notably, PZOS even outperforms Chen until both have essentially converged on both problem classes, further highlighting the importance of the gradient estimator.

Performance by oracle calls. While P-Chen and PZOS achieve comparable performance per iteration, P-Chen requires substantially more oracle calls. In our experiments, P-Chen uses 13 oracle calls per iteration on routing games and 7 on security games, compared to only 3 for PZOS. Consequently, when comparing the normalized objective value against the number of oracle calls to y^* rather than iterations, PZOS substantially outperforms P-Chen (Fig. 7). This advantage is particularly relevant when evaluating y^* is expensive, e.g., because it requires computing an equilibrium or observing the deployed system. In such cases, PZOS achieves comparable performance at a fraction of the oracle cost.

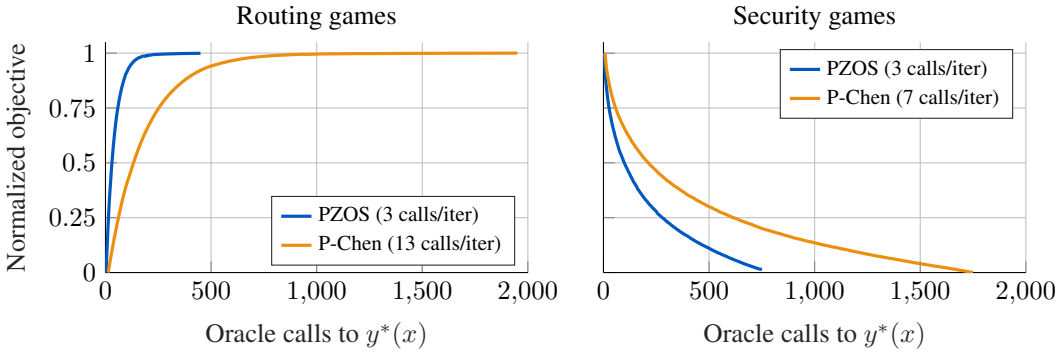


Figure 7: Median normalized objective against the number of oracle calls to y^* for PZOS and P-Chen, on routing (left, higher is better) and security games (right, lower is better) instances of Sec. 5.3.

Limitations and future work. Our algorithm is developed for the setting in which y^* is constrained, but x is unconstrained. In Appendix A, we test both algorithms in the constrained setting by requiring $\tau \geq 0$ in the routing objective (9), enforcing feasibility via projection onto $\mathbb{R}^{|E|}_{\geq 0}$ after each update step. Both ZOS and PZOS remain stable under this modification, with PZOS still achieving visibly better performance than ZOS (Figure 13). Extending the theory to the constrained setting $x \in \mathcal{X}$ would require $y^*(x)$ to be well defined on $\mathcal{X}_\mu = \mathcal{X} + \mathbb{B}(0, \mu)$ rather than only on $\mathcal{X}$. One could then extend the partial Goldstein stationarity notion analogously to the constrained Goldstein stationarity notion of Cui et al. [7], Liu et al. [20].

Acknowledgments and Disclosure of Funding

This work was partially supported by: the EPSRC grant EP/Y001001/1, funded by the International Science Partnerships Fund (ISPF) and UKRI; an Imperial-MIT seed fund; a Google DeepMind Scholarship awarded to Miriam Fischer.

References

[1] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pages 136–145. PMLR, 2017.

[2] Jean-Pierre Aubin and Arrigo Cellina. *Differential inclusions. Set-valued maps and viability theory*. Springer, 1984.

[3] Hande Y Benson, Arun Sen, David F Shanno, and Robert J Vanderbei. Interior-point algorithms, penalty methods and equilibrium problems. *Computational Optimization and Applications*, 34 (2):155–182, 2006.

[4] Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, pages 5219–5233. PMLR, 2023.

[5] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[6] Shisheng Cui and Uday V Shanbhag. On the computation of equilibria in monotone and potential stochastic hierarchical games. *Mathematical Programming*, 198(2):1227–1285, 2023.

[7] Shisheng Cui, Uday V Shanbhag, and Farzad Yousefian. Complexity guarantees for an implicit smoothing-enabled method for stochastic MPECs. *Mathematical Programming*, 198(2):1153–1225, 2023.

[8] Victor DeMiguel, Michael P Friedlander, Francisco J Nogales, and Stefan Scholtes. A two-sided relaxation scheme for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization*, 16(2):587–609, 2005.

[9] Edith Elkind, Abheek Ghosh, and Paul W Goldberg. Continuous-time best-response and related dynamics in tullock contests with convex costs. *arXiv preprint arXiv:2402.08541*, 2024.

[10] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.

[11] Roger Fletcher, Sven Leyffer, Danny Ralph, and Stefan Scholtes. Local convergence of SQP methods for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization*, 17(1):259–286, 2006.

[12] Panagiotis D Grontas, Giuseppe Belgioioso, Carlo Cenedese, Marta Fochesato, John Lygeros, and Florian Dörfler. Big hype: Best intervention in games via distributed hypergradient descent. *IEEE Transactions on Automatic Control*, 69(12):8338–8353, 2024.

[13] Kjell Hausken and Vicki M Bier. Defending against multiple different attackers. *European Journal of Operational Research*, 211(2):370–384, 2011.

[14] Tim Hoheisel, Christian Kanzow, and Alexandra Schwartz. Theoretical and numerical comparison of relaxation methods for mathematical programs with complementarity constraints. *Mathematical Programming*, 137(1):257–288, 2013.

[15] David Iliaev, Sigal Oren, and Ella Segev. A tullock-contest-based approach for cyber security investments. *Annals of Operations Research*, 320(1):61–84, 2023.

[16] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

[17] Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122):1–14, 2024.

[18] Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35:26160–26175, 2022.

[19] P.O. Lindberg and Leonid Engelson. Tolled multi-class traffic equilibria and toll sensitivities. *EURO Journal on Transportation and Logistics*, 4(2):197–222, 2015. ISSN 2192-4376.

[20] 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, 2024.

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

[22] Liqun Qi and Zengxin Wei. On the constant positive linear dependence condition and its application to SQP methods. *SIAM Journal on Optimization*, 10(4):963–981, 2000.

[23] Arvind U Raghunathan and Lorenz T Biegler. An interior point method for mathematical programs with complementarity constraints (mpccs). *SIAM Journal on Optimization*, 15(3): 720–750, 2005.

[24] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

[25] Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for Hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022.

[26] Rangarajan K Sundaram. *A first course in optimization theory*. Cambridge University Press, 1996.

[27] Haochen Tao, Shisheng Cui, Zhuo Li, and Jian Sun. A zeroth-order stochastic implicit method for bilevel-structured actor-critic schemes. *Science China Information Sciences*, 68(5):150204, 2025.

[28] Gordon Tullock. Efficient rent seeking. In James M. Buchanan, Robert D. Tollison, and Gordon Tullock, editors, *Toward a Theory of the Rent-Seeking Society*, pages 97–112. Texas A&M University Press, College Station, TX, 1980.

[29] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2nd edition, 2018.

[30] Hai Yang and Hai-Jun Huang. The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. *Transportation Research Part B: Methodological*, 38(1):1–15, 2004.

[31] Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

[32] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.

Technical appendices and supplementary material

A Supplementary algorithms and experiments

A.1 Algorithms

Algorithm 2 Zeroth-Order Smoothing (ZOS)

Input: Initial $x_0 \in \mathbb{R}^{\mathrm{d_x}}$, step size α, smoothing parameter $\mu > 0$, iterations T
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ \\ Random direction uniformly from unit sphere
  Evaluate $y^*(x_t + \mu v_t)$, and $y^*(x_t - \mu v_t)$
  $\widetilde{g}_t = \frac{\mathrm{d_x}}{2\mu}[f(x_t + \mu v_t, y^*(x_t + \mu v_t)) - f(x_t - \mu v_t, y^*(x_t - \mu v_t))]\, v_t$ \\ Gradient estimate
  $x_{t+1} = x_t - \alpha \widetilde{g}_t$ \\ Update step (minimization); for maximization, use $+$
end for
Output: x^R where $R \in \{0, 1, \ldots, T-1\}$ is uniformly sampled

Algorithm 3 Zeroth-Order Smoothing with batch size Q (ZOS-B)

Input: Initial $x_0 \in \mathbb{R}^{\mathrm{d_x}}$, step size α, smoothing parameter $\mu > 0$, iterations T, batch size Q
for $t = 0, 1, \ldots, T-1$ **do**
  for $q = 1, \ldots, Q$ **do**
    Sample $v_{t,q} \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ \\ Random direction uniformly from unit sphere
    Evaluate $y^*(x_t + \mu v_{t,q})$, and $y^*(x_t - \mu v_{t,q})$
    $\widetilde{g}_{t,q} = \frac{\mathrm{d_x}}{2\mu}[f(x_t + \mu v_{t,q}, y^*(x_t + \mu v_{t,q})) - f(x_t - \mu v_{t,q}, y^*(x_t - \mu v_{t,q}))]\, v_{t,q}$ \\ Gradient est.
  end for
  Compute gradient average: $\widetilde{g}_t = \frac{1}{Q}\sum_{q=1}^{Q} \widetilde{g}_{t,q}$ \\ Average of gradient estimates
  Update: $x_{t+1} = x_t - \alpha \widetilde{g}_t$ \\ Update step (minimization); for maximization, use $+$
end for
Output: x^R where $R \in \{0, 1, \ldots, T-1\}$ is uniformly sampled

Algorithm 4 Partial Zeroth-Order Smoothing with batch size Q (PZOS-B)

Input: Initial $x_0 \in \mathbb{R}^{\mathrm{d_x}}$, step size α, smoothing parameter $\mu > 0$, iterations T, batch size Q
for $t = 0, 1, \ldots, T-1$ **do**
  Evaluate $y^*(x_t)$
  for $q = 1, \ldots, Q$ **do**
    Sample $v_{t,q} \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ \\ Random direction uniformly from unit sphere
    Evaluate $y^*(x_t + \mu v_{t,q})$, and $y^*(x_t - \mu v_{t,q})$
    Compute Jacobian estimate: $H_{t,q} = \frac{\mathrm{d_x}}{2\mu}((y^*(x_t + \mu v_{t,q}) - y^*(x_t - \mu v_{t,q}))\, v_{t,q}^\top)$
  end for
  Compute Jacobian average: $\bar{H}_t = \frac{1}{Q}\sum_{q=1}^{Q} H_{t,q}$
  Compute gradient estimate: $g_t = \nabla_x f(x_t, y^*(x_t)) + \bar{H}_t^\top \nabla_y f(x_t, y^*(x_t))$ \\ Gradient estimate
  Update: $x_{t+1} = x_t - \alpha g_t$ \\ Update step (minimization); for maximization, use $+$
end for
Output: x^R where $R \in \{0, 1, \ldots, T-1\}$ is uniformly sampled

Algorithm 5 Recursive variance-reduced zeroth-order method with estimator $\mathcal{E} = \widetilde{g}$ (Chen) or $\mathcal{E} = g$ (P-Chen)

Input: Initial $x_0 \in \mathbb{R}^{\mathrm{d_x}}$, step size α, smoothing parameter $\mu > 0$, iterations T, batch sizes b' and b, period m, estimator $\mathcal{E}$
for $t = 0, 1, \ldots, T-1$ **do**
  if $t \bmod m = 0$ **then**
    Sample $v_{t,1}, \ldots, v_{t,b'} \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ \\ Batch of size b'
    $\zeta_t = \frac{1}{b'} \sum_{q=1}^{b'} \mathcal{E}(x_t; v_{t,q})$
  else
    Sample $v_{t,1}, \ldots, v_{t,b} \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ \\ Batch of size b
    $\zeta_t = \zeta_{t-1} + \frac{1}{b} \sum_{q=1}^{b} \big[\mathcal{E}(x_t; v_{t,q}) - \mathcal{E}(x_{t-1}; v_{t,q})\big]$ \\ Recursive update
  end if
  $x_{t+1} = x_t - \alpha\, \zeta_t$ \\ Update step (minimization); for maximization, use $+$
end for
Output: x^R where $R \in \{0, 1, \ldots, T-1\}$ is uniformly sampled

`Chen` **($\mathcal{E} = \widetilde{g}$).** This is the black-box method of Chen et al. [4], denoted `Chen`, with the two-point estimator (2)

$$\widetilde{g}(x; v) \coloneqq \frac{\mathrm{d_x}}{2\mu} \big[f(x + \mu v, y^*(x + \mu v)) - f(x - \mu v, y^*(x - \mu v))\big] v.$$

`P-Chen` **($\mathcal{E} = g$).** This is our variant `P-Chen`, which is not part of Chen et al. [4]. `P-Chen` keeps the outer algorithm of Chen et al. [4] in place but replaces the inner estimator $\widetilde{g}$ by our estimator (5)

$$g(x; v) \coloneqq \nabla_x f(x, y^*(x)) + H(x; v)^\top \nabla_y f(x, y^*(x)), \quad H(x; v) \coloneqq \frac{\mathrm{d_x}}{2\mu} \big(y^*(x + \mu v) - y^*(x - \mu v)\big) v^\top.$$

The convergence guarantees of Chen et al. [4] in Table 1 apply to $\mathcal{E} = \widetilde{g}$ (`Chen`) only.

A.2 Comparison of convergence rates

Method	Estimator	Goldstein	Partial Goldstein‡
`ZOS` (Lin et al. [18])	$\widetilde{g}_t$ (black box)	$\mathrm{d_x}^{3/2}\, \delta^{-1} \varepsilon^{-4}$	$\mathrm{d_x}^{3/2}\, \varepsilon^{-4} \max\{\delta^{-1}, \widetilde{C}_p \varepsilon^{-1}\}^\ddagger$
`Chen` (Chen et al. [4])	$\widetilde{g}_t$ (black box)	$\mathrm{d_x}^{3/2}\, \delta^{-1} \varepsilon^{-3}$	–
Kornowski and Shamir [17]	$\widetilde{g}_t$ (black box)	$\mathrm{d_x}\, \delta^{-1} \varepsilon^{-3}$	–
`PZOS` (ours)	g_t (partial)‡	$\mathrm{d_x}^{3/2}\, \varepsilon^{-4} \max\{\delta^{-1}, C_f \varepsilon^{-1}\}^\ddagger$	$\mathrm{d_x}^{3/2}\, \varepsilon^{-4} \max\{\delta^{-1}, C_p \varepsilon^{-1}\}^\ddagger$

Table 1: Iteration complexity (up to $\mathcal{O}(\cdot)$) to reach a (δ, ε)-Goldstein or (δ, ε)-partial Goldstein stationary point, with $\mathrm{d_x}$ the leader's dimension. Estimators: black-box $\widetilde{g}_t$ (2) and partial g_t (5). Rates of the form $\mathrm{d_x}^{3/2}\, \varepsilon^{-4} \max\{\delta^{-1}, C\varepsilon^{-1}\}$ follow from $T = \mathcal{O}(\mathrm{d_x}^{3/2}\, \mu^{-1} \varepsilon^{-4})$ with $\mu = \Theta(\min\{\delta, \varepsilon/C\})$, where C is the constant in the respective entry: $C_f = (1 + L_y) L_g (1 + 2L_y)$, $C_p = L_g (1 + L_y) L_y$, $\widetilde{C}_p = L_g (1 + L_y)^2$. Results new in this paper are indicated by ‡.

A.3 Additional experiments and graphs

Effect of batch size. A standard approach to reduce the variance of `ZOS` is to average $Q \geq 1$ independent estimates $\{\widetilde{g}_{t,q}\}_{q=1}^{Q}$ per iteration when evaluating $\widetilde{g}_t$, at the cost of additional oracle calls to y^* and f (`ZOS-B`). We can define an equivalent variant of `PZOS` (Alg. 4). We compare `PZOS` to `ZOS-B` on the routing instances of Sec. 5.1, where we give `ZOS` a significant advantage by allowing it to use $Q \in \{1, 2, 4\}$ independent samples per iteration, while keeping `PZOS` at $Q = 1$. Even so, `ZOS-B` with $Q = 2$ (using 4 oracle calls per iteration) remains below `PZOS` (using 3 oracle calls) by iteration (Fig. 8). `ZOS-B` with $Q = 4$ closes the gap to `PZOS` after around 90 iterations, but requires 8 oracle calls per iteration compared to `PZOS`'s 3. Plotting the same comparison with oracle calls to y^* on the x-axis instead of iterations (Fig. 9) shows that `PZOS` needs far fewer oracle calls to y^* than `ZOS` to approach convergence across all batch sizes; in particular it remains ahead of `ZOS-B` with $Q = 4$. This is a significant advantage in particular when evaluating $y^*(x)$ is expensive.

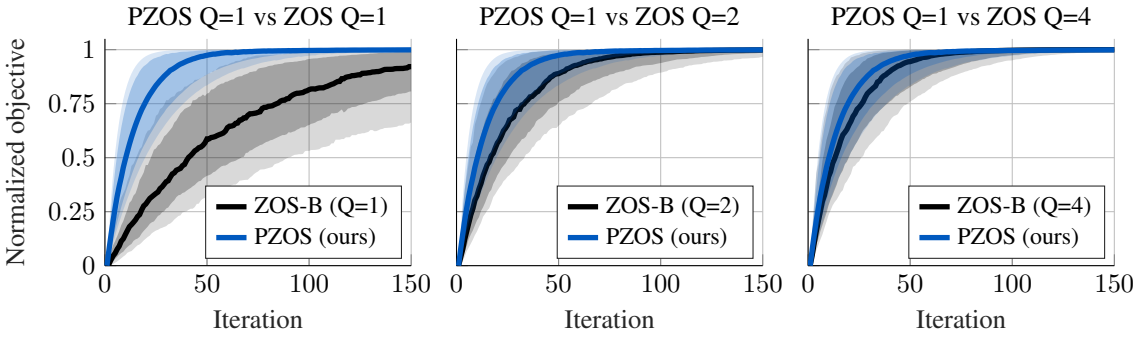


Figure 8: Effect of batch size Q on algorithm performance for heterogeneous routing games: `PZOS` at $Q = 1$ compared against `ZOS` at $Q \in \{1, 2, 4\}$, showing median normalized objective with IQR (dark shading) and 10th–90th percentile bands (light shading), on the same 225 runs as Sec. 5.1.

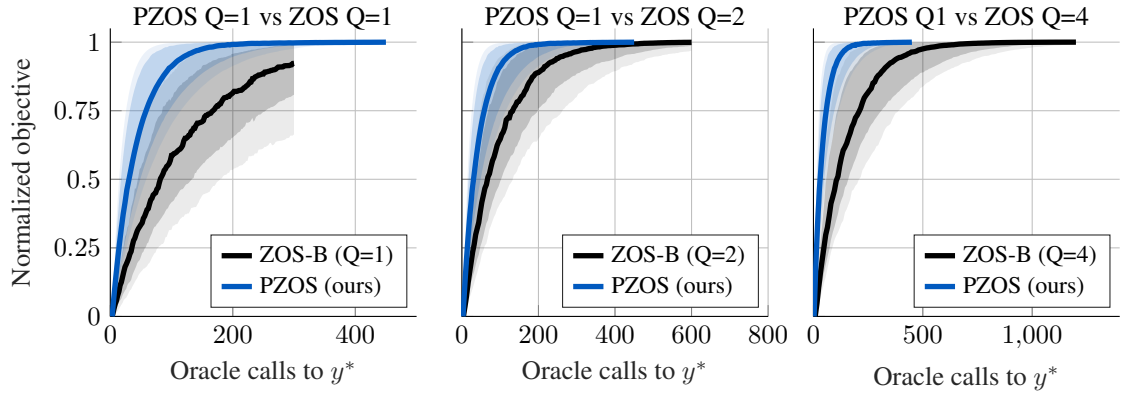


Figure 9: Normalized objective value (higher is better) for the heterogeneous routing game of Sec. 5.1 of `PZOS` (3 oracle calls to y^* per iteration) vs. `ZOS-B` with batch size $Q \in \{1, 2, 4\}$ ($2Q$ calls to y^* per iteration), as a function of oracle calls to y^*. Median with interquartile range (dark shading) and 10–90th percentile bands (light shading).

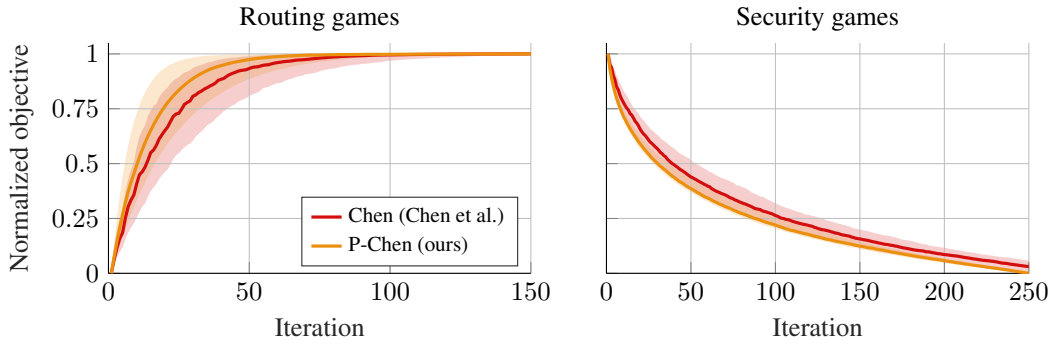


Figure 10: Black-box estimator $\widetilde{g}_t$ (`Chen`) and our estimator g_t (`P-Chen`) for outer algorithm of Chen et al. [4], on routing (left, higher is better) and security games (right, lower is better) instances of Sec. 5.3. Median objective value with interquartile range.

Oracle calls of all algorithms. `ZOS` uses 2 evaluations of y^* at any iteration, `PZOS` uses 3. `ZOS-B` uses $2Q$ evaluations per iteration, `PZOS-B` uses $2Q + 1$. For `Chen`, an iteration with $t \bmod m = 0$ uses $2b'$ calls to y^* and every other iteration uses $4b$ calls, since the same directions v_t are evaluated at both x_t and x_{t-1}. This gives $(2b' + 4b(m-1))/m$ calls per iteration on average. For `P-Chen`, each iteration additionally evaluates $y^*(x_t)$, while $y^*(x_{t-1})$ is reused from the previous iteration, which adds one call per iteration.

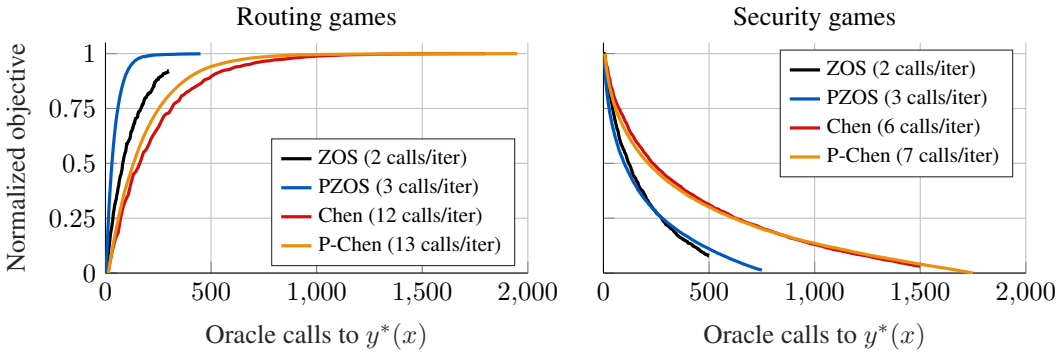


Figure 11: Median normalized objective against the number of oracle calls to y^* for `ZOS`, `PZOS`, `Chen`, and `P-Chen`, on routing (left, higher is better) and security games (right, lower is better) instances of Sec. 5.3.

Figure 11 compares `ZOS`, `PZOS`, `Chen`, and `P-Chen` per oracle call on the instances of Sec. 5.3. On both problem instances, `PZOS` is not only better than `Chen` by iterations (Fig. 6), but also by oracle calls to y^*. Equally, `PZOS` is significantly better than `P-Chen` by oracle calls to y^* on both instances. Further, `PZOS` remains ahead of `ZOS` by oracle calls for the routing game instances, for the security game instances, `PZOS` is ahead of `ZOS` up to about 240 oracle calls, and comparable thereafter.

Sensitivity to μ. Figure 12 shows that `PZOS` outperforms `ZOS` for all tested values $\mu \in \{0.1, 0.5, 0.9, 2\}$ of the smoothing parameter μ on the routing instances of Sec. 5.1.

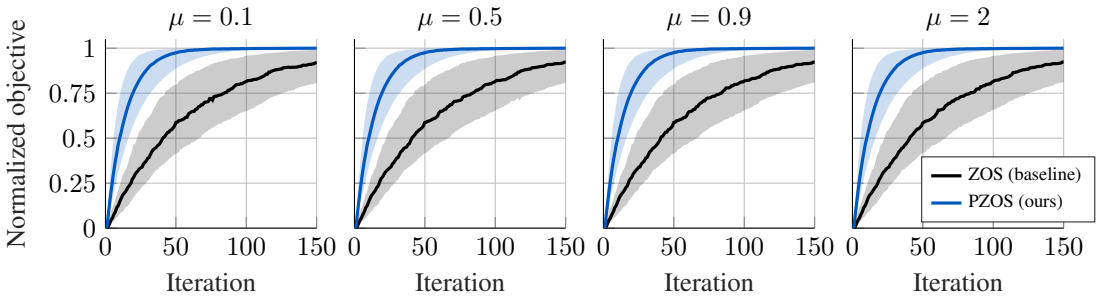


Figure 12: Normalized objective value (higher is better) for smoothing parameter $\mu = \{0.1, 0.5, 0.9, 2\}$ on the same 225 runs as Sec. 5.1. Median with interquartile range.

Constrained leader. Figure 13 repeats the routing experiment of Sec. 5.1 with $\tau \geq 0$ enforced via projection onto $\mathbb{R}^{|E|}_{\geq 0}$ after each update.

Second moments with error bars. Figure 14 reproduces Figure 1 with error bars.

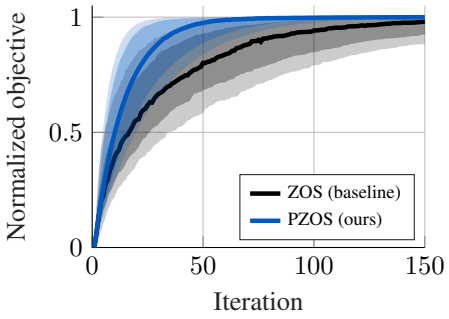


Figure 13: Routing games of Sec. 5.1 with non-negativity constraint $\tau \geq 0$ enforced via projection. Median normalized objective trajectories (higher is better) with interquartile range (dark shading) and 10th–90th percentile range (light shading).

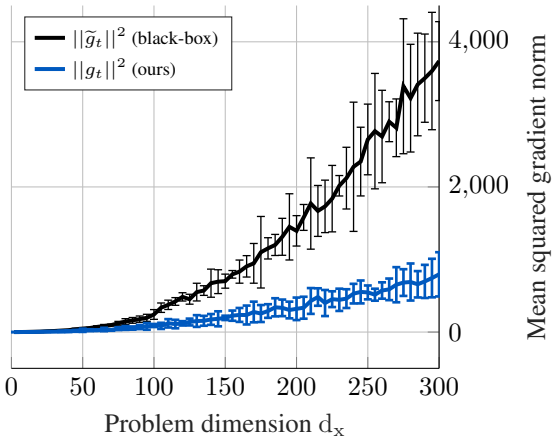


Figure 14: Mean $\pm$ 2 standard deviations of $\|g_t\|^2$ and $\|\widetilde{g}_t\|^2$ on security instances of Sec. 5.2, with 3500 samples per problem dimension.

B Goldstein Stationarity: Detailed Framework and Examples

This section provides technical details underlying the notion of Goldstein and partial Goldstein stationary points in contribution C2. Specifically, we first recall the Clarke generalized Jacobian and chain rule, then introduce the corresponding Goldstein and partial Goldstein subdifferential and stationarity notions. We conclude with two illustrative examples showing that neither notion is in general stronger than the other.

B.1 Clarke Generalized Jacobians

We begin by recalling the Clarke generalized Jacobian for vector-valued functions, which extends the Clarke generalized gradient to mappings $G : \mathbb{R}^n \to \mathbb{R}^m$.

Definition 2 (Clarke generalized Jacobian, [5, Def. 2.6.1])**.** *Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued function with components $G = (g_1(x), g_2(x), \ldots, g_m(x))$, where each $g_i : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz. The Clarke generalized Jacobian $\partial G(x)$ of G at a point $x \in \mathbb{R}^n$ is defined as the convex hull of all matrices that are limits of Jacobians at sequences of points converging to x where G is differentiable:*

$$\partial G(x) = \operatorname{conv}\left\{ \lim_{k\to\infty} JG(x_k) : x_k \to x,\ G \text{ is differentiable at each } x_k \right\}, \tag{11}$$

where $JG(x) \in \mathbb{R}^{m\times n}$ denotes the Jacobian matrix of partial derivatives of G at a point x where those partial derivatives exist.

The Clarke generalized Jacobian is a nonempty, convex, compact subset of $\mathbb{R}^{m\times n}$, and the set-valued mapping $x \mapsto \partial G(x)$ is upper semicontinuous [5, Prop. 2.6.2].

Theorem 2 (Clarke's Generalized Chain Rule, [5, Thm. 2.6.6])**.** *Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz near x and let $f : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz near $G(x)$. Define the function $h : \mathbb{R}^n \to \mathbb{R}$ by $h(x) = f(G(x))$. Then:*

$$\partial h(x) \subseteq \operatorname{conv}\{vM : v \in \partial f(G(x)),\ M \in \partial G(x)\}, \tag{12}$$

where $\partial f(G(x)) \subseteq \mathbb{R}^{1\times m}$ denotes the Clarke generalized gradient of f at $G(x)$, treated as a set of row vectors. If, in addition, f is strictly differentiable at $G(x)$, $\partial f(G(x)) = \{\nabla f(G(x))^\top\}$ and equality holds:

$$\partial h(x) = \{vM : v \in \partial f(G(x)),\ M \in \partial G(x)\}. \tag{13}$$

For a definition of strict differentiability, see [5, Section 2.2]. For our purposes, it suffices to note that continuous differentiability implies strict differentiability.

Proposition 2 ([5, Corollary to Prop. 2.2.1])**.** *If $f : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable at y, then f is strictly differentiable at y.*

Remark 1 (Gradient convention)**.** *Following the convention in [5, Remark 2.6.3, Prop. 2.6.4], the Clarke generalized gradient $\partial f(y)$ of a scalar-valued function $f : \mathbb{R}^m \to \mathbb{R}$ consists of* row vectors *in $\mathbb{R}^{1\times m}$. This ensures dimensional consistency in the chain rule of Thm. 2, where elements of $\partial f(G(x))$ multiply elements of $\partial G(x) \subseteq \mathbb{R}^{m\times n}$ from the left to yield elements in $\mathbb{R}^{1\times n}$. However, throughout this manuscript, we treat gradients as* column vectors*. Accordingly, we reformulate the chain rule* (13) *of Thm. 2 as*

$$\partial h(x) = \left\{ M^\top \nabla f(G(x)) : M \in \partial G(x) \right\}, \tag{14}$$

where elements of $\partial h(x)$ are now column vectors in $\mathbb{R}^n$. We use this column-vector convention for all Clarke generalized gradients hereafter.

B.2 Goldstein and partial Goldstein stationary points

We first introduce the δ-Goldstein subdifferential in the general setting, which enlarges the Clarke generalized gradient by aggregating over a δ-neighborhood. We then specialize to composite functions of the form $h(x) = f(G(x))$, where we introduce both the standard Goldstein subdifferential and our novel partial counterpart.

Definition 3 (δ-Goldstein subdifferential)**.** *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. For $\delta > 0$, the δ-Goldstein subdifferential of h at $x \in \mathbb{R}^n$ is defined as*

$$\partial_\delta h(x) = \operatorname{conv}\left\{ \cup_{z\in\mathbb{B}(x,\delta)} \partial h(z) \right\},$$

where $\mathbb{B}(x,\delta) = \{z \in \mathbb{R}^n : \|z - x\| \le \delta\}$ denotes the closed ball of radius δ centered at x in the ℓ_2-norm. For a locally Lipschitz vector-valued function $G : \mathbb{R}^n \to \mathbb{R}^m$, we define the δ-Goldstein subdifferential as

$$\partial_\delta G(x) = \operatorname{conv}\left\{ \bigcup_{z\in\mathbb{B}(x,\delta)} \partial G(z) \right\} \subseteq \mathbb{R}^{m\times n}. \tag{15}$$

For composite functions of the form $h(x) = f(G(x))$, we distinguish between two notions of Goldstein subdifferentials.

Definition 4 ((Partial) Goldstein subdifferential for composite functions)**.** *Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz near x and let $f : \mathbb{R}^m \to \mathbb{R}$ be continuously differentiable near $G(x)$. Define $h(x) = f(G(x))$. For $\delta > 0$:*

(i) The δ-Goldstein subdifferential *of h at x is*

$$\partial_\delta^G h(x) = \operatorname{conv}\left\{ \bigcup_{z\in\mathbb{B}(x,\delta)} \partial h(z) \right\} = \operatorname{conv}\left\{ \bigcup_{z\in\mathbb{B}(x,\delta)} \left\{ M^\top \nabla f(G(z)) : M \in \partial G(z) \right\} \right\},$$

where the second equality follows from Thm. 2, Prop. 2, and Rem. 1 and Eq. (14)*.*

(ii) The δ-partial Goldstein subdifferential *of h at x is*

$$\partial_\delta^P h(x) = \left\{ M^\top \nabla f(G(x)) : M \in \partial_\delta^G G(x) \right\},$$

where $\partial_\delta^G G(x) = \operatorname{conv}\left\{ \bigcup_{z\in\mathbb{B}(x,\delta)} \partial G(z) \right\}$.

We are now ready to present the notion of (partial) Goldstein stationary points.

Definition 5 ((δ,ε)-Goldstein stationary points for composite functions)**.** *Let $h(x) = f(G(x))$ be as in Def. 4. A point $x \in \mathbb{R}^n$ is called:*

(i) A (δ,ε)-Goldstein stationary point *if* $\min\left\{ \|g\| : g \in \partial_\delta^G h(x) \right\} \le \varepsilon$.

(ii) A (δ,ε)-partial Goldstein stationary point *if* $\min\left\{ \|g\| : g \in \partial_\delta^P h(x) \right\} \le \varepsilon$.

The distinction between Goldstein and partial Goldstein subdifferential lies in where the gradient of f is evaluated: the subdifferential evaluates ∇f at all points $G(z)$ for $z \in \mathbb{B}(x, \delta)$, while the partial subdifferential evaluates ∇f only at the center point $G(x)$. There is no general inclusion relationship between $\partial_\delta^{\mathrm{G}} h(x)$ and $\partial_\delta^{\mathrm{P}} h(x)$, and neither stationarity notion in Def. 5 is a stronger or weaker stationarity condition than the other. We illustrate this with examples in Appendix B.4.

B.3 Application to our problem setting

We now specialize the general framework to our setting. Recall from (1) and Assumption 2 that our objective function is $F(x) = f(x, y^*(x))$, where $f : \mathbb{R}^{\mathrm{d_x}} \times \mathbb{R}^{\mathrm{d_y}} \to \mathbb{R}$ is smooth (i.e., differentiable with Lipschitz-continuous gradient) and $y^* : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y}}$ is Lipschitz continuous but potentially non-smooth.

To apply the chain rule framework, we write $F(x) = f(G(x))$ where $G : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_x} + \mathrm{d_y}}$ is defined by

$$G(x) = \begin{pmatrix} x \\ y^*(x) \end{pmatrix}.$$

The function G is locally Lipschitz since both the identity mapping $x \mapsto x$ and y^* are Lipschitz. At points where y^* is differentiable, the classical Jacobian of G is

$$JG(x) = \begin{pmatrix} I_{\mathrm{d_x}} \\ Jy^*(x) \end{pmatrix} \in \mathbb{R}^{(\mathrm{d_x} + \mathrm{d_y}) \times \mathrm{d_x}},$$

where $I_{\mathrm{d_x}}$ denotes the $\mathrm{d_x} \times \mathrm{d_x}$ identity matrix. The Clarke generalized Jacobian of G is therefore

$$\partial G(x) = \left\{ \begin{pmatrix} I_{\mathrm{d_x}} \\ M \end{pmatrix} : M \in \partial y^*(x) \right\},$$

where $\partial y^*(x) = \operatorname{conv}\{\lim_{k\to\infty} Jy^*(x_k) : x_k \to x,\ y^* \text{ differentiable at each } x_k\}$ according to (11).

Since f is continuously differentiable by Assumption 2 (and hence strictly differentiable by Prop. 2), applying Thm. 2 with the column-vector convention (14) yields:

$$\begin{aligned} \partial F(x) &= \left\{ \begin{pmatrix} I_{\mathrm{d_x}} & M^\top \end{pmatrix} \begin{pmatrix} \nabla_x f(x, y^*(x)) \\ \nabla_y f(x, y^*(x)) \end{pmatrix} : M \in \partial y^*(x) \right\} \\ &= \left\{ \nabla_x f(x, y^*(x)) + M^\top \nabla_y f(x, y^*(x)) : M \in \partial y^*(x) \right\}. \end{aligned} \tag{16}$$

Using this characterization, we can now express the Goldstein subdifferentials for our objective function $F(x) = f(x, y^*(x))$ for problem setting (1). Under Assumption 2, for $\delta > 0$:

(i) The *δ-Goldstein subdifferential* of F at $x \in \mathbb{R}^{\mathrm{d_x}}$ is

$$\partial_\delta^{\mathrm{G}} F(x) = \operatorname{conv} \left\{ \bigcup_{z \in \mathbb{B}(x,\delta)} \left\{ \nabla_x f(z, y^*(z)) + M^\top \nabla_y f(z, y^*(z)) : M \in \partial y^*(z) \right\} \right\}. \tag{17}$$

(ii) The *δ-partial Goldstein subdifferential* of F at $x \in \mathbb{R}^{\mathrm{d_x}}$ is

$$\partial_\delta^{\mathrm{P}} F(x) = \left\{ \nabla_x f(x, y^*(x)) + M^\top \nabla_y f(x, y^*(x)) : M \in \partial_\delta^{\mathrm{G}} y^*(x) \right\}, \tag{18}$$

where $\partial_\delta^{\mathrm{G}} y^*(x) = \operatorname{conv}\left\{ \bigcup_{z \in \mathbb{B}(x,\delta)} \partial y^*(z) \right\}$.

Applying Definition 5 of (δ, ε)-(partial) Goldstein stationary points to (17) and (18) yields the stationarity notions we use throughout our convergence analysis in Appendix C.3. In general, a point may be(δ, ε)-Goldstein stationary but not (δ, ε)-partial Goldstein stationary, and vice versa; the examples in the following section illustrate this.

B.4 Comparison between Goldstein and partial Goldstein stationarity

Example 1 exhibits a point that is (δ, ε)-Goldstein stationary but not (δ, ε)-partial Goldstein stationary.

Example 1 ((δ, ε)-Goldstein stationary, but not (δ, ε)-partial Goldstein stationary)**.** *Define $f(x, y) = (x-1)^2 + y$ and $y^*(x) = |x|$, so that $F(x) = f(x, y^*(x)) = (x-1)^2 + |x|$. The follower response y^* is Lipschitz and non-differentiable at $x = 0$; f is smooth. The composite F is non-differentiable at $x = 0$ with one-sided derivatives $F'(0^+) = -1$ and $F'(0^-) = -3$, giving Clarke gradient*

$$\partial F(0) = \operatorname{conv}\left\{ \lim_{z \to 0^+} F'(z),\ \lim_{z \to 0^-} F'(z) \right\} = \operatorname{conv}\{-1, -3\} = [-3, -1],$$

whereas F is continuously differentiable on $(-\infty, 0)$ and $(0, \infty)$, so $\partial F(x) = \{F'(x)\}$ for $x \neq 0$ by Prop. 2, and

$$F'(x) = \begin{cases} 2x - 1, & x > 0, \\ 2x - 3, & x < 0. \end{cases}$$

The global minimizer of F is at $x = \frac{1}{2}$, where $F(\frac{1}{2}) = \frac{3}{4}$. Below, we analyze the Goldstein and partial Goldstein subdifferential of F at $x = 0$.

Goldstein subdifferential. *For $\delta > 0$, F is differentiable on $\mathbb{B}(0, \delta) \setminus \{0\}$ with $F'(z) = 2z - 1$ for $z > 0$ and $F'(z) = 2z - 3$ for $z < 0$. Taking the convex hull of Clarke gradients over $\mathbb{B}(0, \delta)$, the Goldstein subdifferential at $x = 0$ equals*

$$\partial^G_\delta F(0) = \operatorname{conv}\left\{ \cup_{z \in \mathbb{B}(0,\delta)} \partial F(z) \right\} = [-2\delta - 3,\ 2\delta - 1].$$

If $\delta < \frac{1}{2}$, all elements in $[-2\delta - 3,\ 2\delta - 1]$ are negative and $\min\{|g| : g \in \partial^G_\delta F(0)\} = 1 - 2\delta > 0$. If $\delta \geq \frac{1}{2}$, $\mathbb{B}(0, \delta)$ includes the minimizer $z = \frac{1}{2}$, so $0 \in \partial^G_\delta F(0)$. Thus, $\partial^G_\delta F(0)$ contains zero if $\delta \geq \frac{1}{2}$, and $x = 0$ is a (δ, ε)-Goldstein stationary point for every $\delta \geq \frac{1}{2}$ and $\varepsilon \geq 0$.

Partial Goldstein subdifferential. *$y^*(x) = |x|$ is continuously differentiable on $\mathbb{R} \setminus \{0\}$, with $(y^*)'(z) = 1$ for $z > 0$ and $(y^*)'(z) = -1$ for $z < 0$, thus $\partial y^*(z) = \operatorname{sgn}(z)$ for $z \neq 0$ by Proposition 2 and [5, Prop. 2.2.4]. It follows that $\partial y^*(0) = \operatorname{conv}\{-1, 1\} = [-1, 1]$. Since at any $z \neq 0$, $\partial y^*(z)$ is a singleton contained in $[-1, 1]$, which equals $\partial y^*(0)$, the Goldstein subdifferential of y^* at $x = 0$ is identical to the Clarke gradient/Jacobian of y^* at $x = 0$, i.e., $\partial^G_\delta y^*(0) = [-1, 1] = \partial y^*(0)$, for all $\delta > 0$. Combined with the partial gradients $\nabla_x f = -2$ and $\nabla_y f = 1$ at $(x, y^*(x)) = (0, 0)$, the partial Goldstein subdifferential of F at $x = 0$ equals*

$$\begin{aligned} \partial^P_\delta F(0) &= \left\{ \nabla_x f(0, y^*(0)) + M \cdot \nabla_y f(0, y^*(0)) : M \in \partial^G_\delta y^*(0) \right\} \\ &= \{-2 + M : M \in [-1, 1]\} = [-3,\ -1]. \end{aligned}$$

This interval is independent of δ, and $\min\{|g| : g \in \partial^P_\delta F(0)\} = 1 > 0$ for any $\delta > 0$. Hence, $x = 0$ is not *a (δ, ε)-partial Goldstein stationary point for any $\varepsilon < 1$, regardless of the value of δ.*

In Example 1, the Goldstein subdifferential grows with δ, and, for $\delta \geq \frac{1}{2}$, contains the Clarke gradient at the minimizer $z = \frac{1}{2}$, which equals zero. The partial Goldstein subdifferential, on the other hand, is constant in δ, as the partial gradients $\nabla_x f$ and $\nabla_y f$ are fixed and non-zero at $x = 0$, and the Goldstein subdifferential of y^* at $x = 0$ does not grow with δ. We turn to the reverse situation, where the partial Goldstein subdifferential grows with δ, while the Goldstein subdifferential (for sufficiently small δ) remains constant.

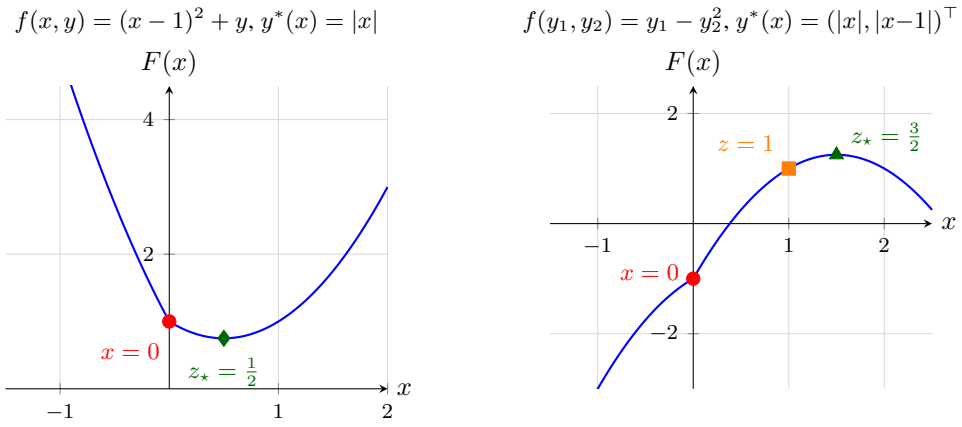
Example 2 ((δ, ε)-partial Goldstein stationary, but not (δ, ε)-Goldstein stationary)**.** *Define $f(y_1, y_2) = y_1 - y_2^2$ and $y^*(x) = \big(|x|,\ |x - 1|\big)^\top$, so that $F(x) = f(x, y^*(x)) = |x| - (x-1)^2$. The follower response y^* is Lipschitz and non-differentiable at $x = 0$ and $x = 1$. However, the non-differentiability of y^* at $x = 1$ does not propagate to F, which is non-differentiable only at $x = 0$.*

Goldstein subdifferential. *On $\mathbb{R} \setminus \{0\}$, F is continuously differentiable, with*

$$F'(x) = \begin{cases} 3 - 2x, & x > 0, \\ 1 - 2x, & x < 0, \end{cases}$$

hence by Proposition 2 and [5, Prop. 2.2.4], $\partial F(z) = \{F'(z)\}$ for $z \neq 0$. Further, at $x = 0$, $\partial F(0) = [1, 3]$. Taking the convex hull of Clarke gradients over $\mathbb{B}(0, \delta)$, we obtain

$$\partial^G_\delta F(0) = \begin{cases} [1, 3], & \delta \in [0, 1], \\ [3 - 2\delta, 1 + 2\delta], & \delta > 1. \end{cases}$$



(a) Example 1: $F(x) = (x-1)^2 + |x|$

(b) Example 2: $F(x) = |x| - (x-1)^2$

Figure 15: Graphs of $F(x) = f(x, y^*(x))$ for Examples 1 and 2. **Left:** $F(x) = (x-1)^2 + |x|$ is non-differentiable at $x = 0$ (red dot). For $\delta \geq \frac{1}{2}$, the Goldstein subdifferential reaches the minimizer $z_\star = \frac{1}{2}$ (green diamond) and contains zero, while the partial Goldstein subdifferential equals $[-3, -1]$ for all $\delta > 0$. **Right:** $F(x) = |x| - (x-1)^2$ is non-differentiable at $x = 0$ (red dot). The non-differentiability of y^* at $z = 1$ (orange square) is invisible in F. For $\delta \in [1, \frac{3}{2})$, the partial Goldstein subdifferential spans $[-1, 3]$. The Goldstein subdifferential only includes Clarke gradients of F in $\mathbb{B}(0, \delta)$, which are strictly positive until the Clarke gradient at the genuine critical point $z_\star = \frac{3}{2}$ (green triangle) of F enters the Goldstein subdifferential for $\delta \geq \frac{3}{2}$.

For $\delta \in (0, 1)$, $\min\{|g| : g \in \partial_\delta^G F(0)\} = 1 > 0$; for $\delta \in [1, \frac{3}{2})$, $\min\{|g|\} = 3 - 2\delta > 0$. For $\delta \geq \frac{3}{2}$, $\mathbb{B}(0, \delta)$ includes $z = \frac{3}{2}$, where $F'(\frac{3}{2}) = 0$, hence $0 \in \partial_\delta^G F(0)$. Therefore, $x = 0$ is not *a (δ, ε)-Goldstein stationary point for $\delta \in (0, 1)$ and $\varepsilon < 1$, nor for $\delta \in [1, \frac{3}{2})$ and $\varepsilon < 3 - 2\delta$.*

Partial Goldstein subdifferential. *The Jacobian of y^* at differentiable points $z \notin \{0, 1\}$ equals*

$$Jy^*(z) = \begin{cases} (-1,-1)^\top, & z < 0, \\ (1,-1)^\top, & 0 < z < 1, \\ (1,\ 1)^\top, & z > 1, \end{cases}$$

with Clarke Jacobians $\partial y^*(0) = \operatorname{conv}\{(-1,-1)^\top, (1,-1)^\top\}$ and $\partial y^*(1) = \operatorname{conv}\{(1,-1)^\top, (1,1)^\top\}$. The Goldstein subdifferential of y^* over $\mathbb{B}(0, \delta)$ depends on whether the point $x = 1$ at which y^* is non-differentiable lies within $\mathbb{B}(0, \delta)$:

If $\delta < 1$, only the non-differentiable point $z = 0$ lies in $\mathbb{B}(0, \delta)$, so

$$\partial_\delta^G\, y^*(0) = \operatorname{conv}\left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} a \\ -1 \end{pmatrix} : a \in [-1, 1] \right\}.$$

Observing that $\nabla_x f = 0$ and $\nabla_y f(0, 1) = (1, -2)^\top$ at $x_0 = 0$, $y^*(0) = (0, 1)^\top$, the partial subdifferential at $x = 0$ equals $\partial_\delta^P F(0) = \{a \cdot 1 + (-1) \cdot (-2) : a \in [-1, 1]\} = [1, 3]$, with $\min\{|g|\} = 1 > 0$.

If $\delta \geq 1$, the point $z = 1$ at which y^* is non-differentiable lies in the ball $\mathbb{B}(0, \delta)$, and the Goldstein subdifferential of y^* becomes

$$\partial_\delta^G\, y^*(0) = \operatorname{conv}\left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} a \\ b \end{pmatrix} : -1 \leq b \leq a \leq 1 \right\}.$$

Combined with partial derivatives $\nabla_x f = 0$ and $\nabla_y f(0, 1) = (1, -2)^\top$ at $(x, y) = (0, (0, 1)^\top)$, the partial subdifferential of F at $x = 0$ equals

$$\partial_\delta^P F(0) = \{a - 2b : -1 \leq b \leq a \leq 1\}.$$

Evaluating $a - 2b$ at the three vertices $(-1, -1)$, $(1, -1)$, and $(1,\ 1)$ of the triangle $\{-1 \leq b \leq a \leq 1\}$ gives values 1, 3, and -1, respectively. The convex hull of these values is $[-1, 3]$, and therefore

$\partial^{\mathrm{P}}_{\delta} F(0) = [-1, 3]$*, which contains zero. Hence* $x = 0$ *is a* (δ, ε)*-partial Goldstein stationary point for every* $\delta \geq 1$ *and* $\varepsilon \geq 0$*.*

Comparing Goldstein and partial Goldstein subdifferential based on the different regimes of δ*, we have the following characterization of Goldstein and partial Goldstein stationarity at* $x = 0$*: For* $\delta \in [1, \frac{3}{2})$*:* $x = 0$ *is* (δ, ε)*-partial Goldstein stationary for all* $\varepsilon \geq 0$*, but not* (δ, ε)*-Goldstein stationary for* $\varepsilon < 3 - 2\delta$*. For* $\delta \geq \frac{3}{2}$*,* $x = 0$ *is both* (δ, ε)*-partial Goldstein stationary and* (δ, ε)*-Goldstein stationary for all* $\varepsilon \geq 0$*.*

Taken together, the two examples show that neither Goldstein nor partial Goldstein subdifferential includes the other, and that neither stationarity notion implies the other. In Example 1, the nearby minimizer $z = \frac{1}{2}$ does not change the partial Goldstein subdifferential at the non-differentiable point $x = 0$, as the Goldstein subdifferential of y^* is constant at $x = 0$. However, once the ball $\mathbb{B}(0, \delta)$ includes this minimizer, its Clarke gradient contributes to the Goldstein subdifferential of F at $x = 0$, causing it to include zero. In Example 2, the non-differentiability of y^* at $z = 1$ does not affect differentiability of F itself, but contributes Clarke Jacobians to $\partial^{\mathrm{G}}_{\delta} y^*(0)$ once $\delta \geq 1$, causing the partial Goldstein subdifferential to grow from $[1, 3]$ to $[-1, 3]$, hence including zero. The Goldstein subdifferential, however, remains strictly positive until $\mathbb{B}(0, \delta)$ includes the true critical point of F at $z = \frac{3}{2}$.

C Proofs of theoretical results

This section details the proofs of Lemma 1, Thm. 1, and Prop. 1.

Appendix C.1 establishes smoothing properties of the equilibrium response y_μ. Besides differentiability of y_μ with $Jy_\mu(x) = \frac{\mathrm{d_x}}{\mu} \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x + \mu v)\, v^\top]$ (Lem. 3), we show that y_μ inherits the Lipschitz constant L_y of y^* (Lem. 4) and remains in $\mathcal{Y}$ (Lem. 2).

Appendix C.2 establishes smoothness of the surrogate F_μ (Lem. 6) and that its distance to F at any given x is bounded by $L_f L_y \mu$ (Prop. 3). Further, we show that H_t is an unbiased estimator of $Jy_\mu(x_t)$ (Lem. 7), so the biasedness $C_p \mu$ of g_t as an estimator of $\nabla F_\mu(x_t)$ is entirely due to evaluating ∇f at $(x_t, y^*(x_t))$ instead of $(x_t, y_\mu(x_t))$ (Lem. 8). We analyze the second moment of the gradient estimator g_t (Lem. 9), and provide the proof of the second-moment comparison to $\widetilde{g}_t$ of Lemma 1 in Lemma 10.

Appendix C.3 establishes convergence of PZOS to (δ, ε)-partial Goldstein and (δ, ε)-Goldstein stationary points (Thms. 4 and 5), and of ZOS to (δ, ε)-partial Goldstein stationary points (Prop. 4). Lemma 11 shows $Jy_\mu(x) \in \partial_\mu y^*(x) \subseteq \partial_\delta y^*(x)$ for $\mu \leq \delta$, utilized in Thm. 4. For Thm. 5, Lem. 12 bounds the distance of ∇F_μ to $\partial^{\mathrm{G}}_{\mu} F$ by $C_f \mu$. Likewise, for Prop. 4, Lem. 13 bounds the distance of $\nabla \widetilde{F}_\mu$ to $\partial^{\mathrm{P}}_{\mu} F$ by $\widetilde{C}_p \mu$.

C.1 Smoothing Properties of the Equilibrium Response

We establish that the smoothed response $y_\mu(x) = \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x + \mu u)]$ is continuously differentiable and derive an explicit expression for its Jacobian matrix $Jy_\mu(x)$. Subsequently, we establish several fundamental properties of y_μ and Jy_μ. We build on the following classical result for smoothing scalar-valued functions, which we later extend to vector-valued settings.

Theorem 3 ([31, Lem. 8], [10, Lemma 2.1])**.** *Let* $h : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}$ *be a continuous function, and let* $\mu > 0$ *be a given scalar. Define* $h_\mu : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}$ *by*

$$h_\mu(x) = \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[h(x + \mu u)].$$

Then h_μ *is differentiable, and for any* $x \in \mathbb{R}^{\mathrm{d_x}}$*, the gradient* $\nabla_x h_\mu(x) \in \mathbb{R}^{\mathrm{d_x}}$ *is given by*

$$\nabla_x h_\mu(x) = \frac{\mathrm{d_x}}{\mu} \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[h(x + \mu v) v].$$

If in addition h *is* L*-Lipschitz continuous, then* $\nabla_x h_\mu$ *is* $\frac{k_1 L \sqrt{\mathrm{d_x}}}{\mu}$*-Lipschitz continuous, where* $k_1 > 0$ *is a constant, and therefore* h_μ *is continuously differentiable.*

We extend uniform smoothing to the vector-valued and constrained setting $\mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y}}$ and define, for a vector-valued and constrained mapping $y^* : \mathbb{R}^{\mathrm{d_x}} \to \mathcal{Y} \subseteq \mathbb{R}^{\mathrm{d_y}}$ and a given scalar $\mu > 0$, the

mapping $y_\mu : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y}}$ by

$$y_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x+\mu u)] = \begin{pmatrix} \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y_1^*(x+\mu u)] \\ \vdots \\ \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y_{\mathrm{d_y}}^*(x+\mu u)] \end{pmatrix}, \tag{19}$$

where u is uniformly distributed over the unit ball $\mathbb{B}^{\mathrm{d_x}} = \{u \in \mathbb{R}^{\mathrm{d_x}} : \|u\| \le 1\}$ and $y_i^*(x)$ denotes the i-th component of $y^*(x) \in \mathcal{Y} \subseteq \mathbb{R}^{\mathrm{d_y}}$.

Since $y_\mu(x)$ is defined as an expectation, it is not immediate that $y_\mu(x)$ remains in $\mathcal{Y}$; averages of elements of an arbitrary set need not lie in the set themselves. Below, we show that if $\mathcal{Y}$ is closed and convex, the smoothed response $y_\mu(x)$ remains in the set $\mathcal{Y}$.

Lemma 2 (Containment of smoothed response in feasible response set)**.** *Under Assumption 2, let $y^* : \mathbb{R}^{\mathrm{d_x}} \to \mathcal{Y}$ be the unique response of the followers for given $x \in \mathbb{R}^{\mathrm{d_x}}$, where y^* is L_y-Lipschitz continuous and $\mathcal{Y} \subseteq \mathbb{R}^{\mathrm{d_y}}$ is closed and convex. Then $y_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x+\mu u)] \in \mathcal{Y}$.*

Proof. Fix any $x \in \mathbb{R}^{\mathrm{d_x}}$ and $\mu > 0$. By assumption, $y^*(z) \in \mathcal{Y}$ for all $z \in \mathbb{R}^{\mathrm{d_x}}$. Since y^* is L_y-Lipschitz by Assumption 2, it is continuous, and therefore the integral

$$y_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x+\mu u)] = \frac{1}{\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}})}\int_{\mathbb{B}^{\mathrm{d_x}}} y^*(x+\mu u)\,du \tag{20}$$

is well-defined and finite, since it is an integral of a continuous function over a compact set.

Suppose for contradiction that $y_\mu(x) \notin \mathcal{Y}$. Since $\mathcal{Y}$ is closed and convex, and $\{y_\mu(x)\}$ (a singleton) is a closed, convex, and bounded (therefore compact and convex) set disjoint from $\mathcal{Y}$, the strict hyperplane separation theorem guarantees the existence of $b \in \mathbb{R}^{\mathrm{d_y}}$ and $\alpha \in \mathbb{R}$ such that:

$$\langle b, y_\mu(x)\rangle > \alpha \quad \text{and} \quad \langle b, y\rangle < \alpha \quad \forall y \in \mathcal{Y}\,. \tag{21}$$

Using (20), we express the left-hand side of the first inequality in (21) as

$$\langle b, y_\mu(x)\rangle = \left\langle b, \frac{1}{\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}})}\int_{\mathbb{B}^{\mathrm{d_x}}} y^*(x+\mu u)\,du \right\rangle = \frac{1}{\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}})}\int_{\mathbb{B}^{\mathrm{d_x}}} \langle b, y^*(x+\mu u)\rangle\,du.$$

Since $y^*(z) \in \mathcal{Y}$ for all $z \in \mathbb{R}^{\mathrm{d_x}}$, thus $y^*(x+\mu u) \in \mathcal{Y}$ for all $u \in \mathbb{B}^{\mathrm{d_x}}$, the second inequality in (21) implies $\langle b, y^*(x+\mu u)\rangle < \alpha$ for all $u \in \mathbb{B}^{\mathrm{d_x}}$. Therefore, we can bound the integral:

$$\frac{1}{\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}})}\int_{\mathbb{B}^{\mathrm{d_x}}} \langle b, y^*(x+\mu u)\rangle\,du \le \frac{1}{\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}})}\int_{\mathbb{B}^{\mathrm{d_x}}} \alpha\,du = \frac{1}{\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}})}\alpha\,\mathrm{Vol}(\mathbb{B}^{\mathrm{d_x}}) = \alpha.$$

This implies $\langle b, y_\mu(x)\rangle \le \alpha$, which contradicts $\langle b, y_\mu(x)\rangle > \alpha$. Therefore, we conclude that $y_\mu(x) \in \mathcal{Y}$. □

Equipped with Lem. 2, we can generalize Thm. 3 to vector-valued and constrained mappings, and provide an explicit expression for the Jacobian of the smoothed mapping.

Lemma 3 (Jacobian of the smoothed response)**.** *Let $y^* : \mathbb{R}^{\mathrm{d_x}} \to \mathcal{Y}$ be the unique response of the followers for given $x \in \mathbb{R}^{\mathrm{d_x}}$ and suppose that y^* is L_y-Lipschitz continuous, where $\mathcal{Y} \subseteq \mathbb{R}^{\mathrm{d_y}}$ is closed and convex. For a given scalar $\mu > 0$, define $y_\mu : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y}}$ by $y_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x+\mu u)]$ (see Eq. (19)). Then y_μ is continuously differentiable over $\mathbb{R}^{\mathrm{d_x}}$. Moreover, for any $x \in \mathbb{R}^{\mathrm{d_x}}$, the Jacobian matrix $Jy_\mu(x) \in \mathbb{R}^{\mathrm{d_y}\times\mathrm{d_x}}$ is given by*

$$Jy_\mu(x) = \frac{\mathrm{d_x}}{\mu}\,\mathbb{E}_{v\sim\mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x+\mu v)\,v^\top\Big], \tag{22}$$

where $\mathbb{S}^{\mathrm{d_x}} = \{v \in \mathbb{R}^{\mathrm{d_x}} : \|v\| = 1\}$ denotes the unit sphere and the expectation in (22) is understood component-wise.

Proof. Let $y_i^*(x)$, for $i \in \{1, \dots, \mathrm{d_y}\}$, denote the i-th component of the vector $y^*(x) \in \mathcal{Y} \subseteq \mathbb{R}^{\mathrm{d_y}}$, and let $[y_\mu]_i$ denote the i-th component of y_μ. Since $y^*(x)$ is L_y-Lipschitz continuous, each component y_i^* is also L_y-Lipschitz continuous, as for all $x_1, x_2 \in \mathbb{R}^{\mathrm{d_x}}$, $|y_i^*(x_1) - y_i^*(x_2)| \leq \|y^*(x_1) - y^*(x_2)\| \leq L_y \|x_1 - x_2\|$. As $[y_\mu(x)]_i = \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y_i^*(x + \mu u)]$, and y_i^* is L_y-Lipschitz, Thm. 3 applied to each component $[y_\mu]_i$ of y_μ yields that $[y_\mu]_i$ is continuously differentiable over $\mathbb{R}^{\mathrm{d_x}}$, with gradient

$$\nabla_x [y_\mu]_i(x) = \frac{\mathrm{d_x}}{\mu} \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y_i^*(x + \mu v)\, v]. \tag{23}$$

As each component $[y_\mu]_i$ of y_μ is continuously differentiable, so is y_μ. The i-th row of its Jacobian $Jy_\mu(x)$ equals $\nabla_x [y_\mu]_i(x)^\top$, so that utilizing the identity $y_i^*(x + \mu v) \cdot v_j = [y^*(x + \mu v)\, v^\top]_{ij}$, we obtain

$$[Jy_\mu(x)]_{ij} = \frac{\mathrm{d_x}}{\mu} \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y_i^*(x + \mu v)\, v_j] = \frac{\mathrm{d_x}}{\mu} \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[\big[y^*(x + \mu v)\, v^\top\big]_{ij}\Big],$$

which is (22). □

The following lemma summarizes key properties of the smoothed equilibrium response $y_\mu(x)$. These results extend known smoothing properties for scalar functions to the vector-valued setting $\mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y}}$.

Lemma 4 (Properties of the smoothed equilibrium response)**.** *Let $y^* : \mathbb{R}^{\mathrm{d_x}} \to \mathcal{Y}$ be the unique followers' response for given $x \in \mathbb{R}^{\mathrm{d_x}}$ and suppose that y^* is L_y-Lipschitz continuous. Let $y_\mu(x)$ be defined as in* (3)*, and let $Jy_\mu(x) = \frac{\mathrm{d_x}}{\mu} \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x + \mu v)\, v^\top\Big]$. Then, for any $x, z \in \mathbb{R}^{\mathrm{d_x}}$:*

(i) $\|y_\mu(x) - y^*(x)\| \leq L_y \mu$.

(ii) $\|y_\mu(x) - y_\mu(z)\| \leq L_y \|x - z\|$.

(iii) $\|Jy_\mu(x) - Jy_\mu(z)\| \leq \frac{k_1 L_y \sqrt{\mathrm{d_x}}}{\mu} \|x - z\|$*, and $k_1 > 0$ is the constant in Thm. 3.*

Proof. **Part (i):** For any $x \in \mathbb{R}^{\mathrm{d_x}}$:

$$\begin{aligned}
\|y_\mu(x) - y^*(x)\| &= \big\|\mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x + \mu u)] - y^*(x)\big\| = \big\|\mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x + \mu u) - y^*(x)]\big\| \\
&\leq \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[\|y^*(x + \mu u) - y^*(x)\|] \leq \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[L_y \|\mu u\|] \\
&= L_y \mu \cdot \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[\|u\|] = L_y \mu \cdot \tfrac{\mathrm{d_x}}{\mathrm{d_x} + 1} \leq L_y \mu.
\end{aligned}$$

where the first inequality follows from Jensen's inequality, the second inequality follows from Lipschitz continuity of y^*, and $\mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[\|u\|] = \mathrm{d_x}/(\mathrm{d_x} + 1)$ follows from [29, Exercise 0.8].

Part (ii): For any $x, z \in \mathbb{R}^{\mathrm{d_x}}$:

$$\begin{aligned}
&\|y_\mu(x) - y_\mu(z)\| = \big\|\mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x + \mu u)] - \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(z + \mu u)]\big\| \\
&= \big\|\mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x + \mu u) - y^*(z + \mu u)]\big\| \leq \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[\|y^*(x + \mu u) - y^*(z + \mu u)\|] \\
&\leq \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[L_y \|x - z\|] = L_y \|x - z\|,
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the second inequality follows from Lipschitz continuity of y^*.

Part (iii): By definition of the spectral norm, for any matrix $A \in \mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}$,

$$\|A\|_2 = \|A^\top\|_2 = \sup_{w \in \mathbb{R}^{\mathrm{d_y}}, \|w\| = 1} \|A^\top w\|.$$

Therefore, it suffices to show that for every unit vector $w \in \mathbb{R}^{\mathrm{d_y}}$, $\big\|\big[Jy_\mu(x) - Jy_\mu(z)\big]^\top w\big\| \leq k_1 \frac{L_y \sqrt{\mathrm{d_x}}}{\mu} \|x - z\|$. Fix such a vector w, and define $h_w : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}$ by $h_w(\cdot) = w^\top y^*(\cdot)$. Since y^* is L_y-Lipschitz and $\|w\| = 1$, we have for all $x, z \in \mathbb{R}^{\mathrm{d_x}}$:

$$|h_w(x) - h_w(z)| = |w^\top (y^*(x) - y^*(z))| \leq \|w\| \, \|y^*(x) - y^*(z)\| \leq L_y \|x - z\|,$$

where the first inequality follows by Cauchy-Schwarz, and the second inequality follows as $\|w\| = 1$ and $y^*(\cdot)$ is L_y-Lipschitz. Hence h_w is L_y-Lipschitz. Then, applying uniform smoothing to h_w, by Thm. 3 the function

$$\begin{aligned} h_{w,\mu}(x) &= \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[h_w(x+\mu u)] = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[w^\top y^*(x+\mu u)] \\ &= w^\top \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x+\mu u)] = w^\top y_\mu(x). \end{aligned}$$

is continuously differentiable, with $k_1 L_y \sqrt{\mathrm{d_x}}/\mu$-Lipschitz-continuous gradient

$$\nabla h_{w,\mu}(x) = \frac{\mathrm{d_x}}{\mu}\, \mathbb{E}_{v\sim\mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[h_w(x+\mu v)\, v]. \tag{24}$$

We now show that $\nabla h_{w,\mu}(x) = J y_\mu(x)^\top w$. Substituting $h_w(\cdot) = w^\top y^*(\cdot) = \sum_{i=1}^{\mathrm{d_y}} w_i\, y_i^*(\cdot)$ into (24) and exchanging the sum with the expectation:

$$\begin{aligned} \nabla h_{w,\mu}(x) &= \frac{\mathrm{d_x}}{\mu}\, \mathbb{E}_{v\sim\mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\left[\sum_{i=1}^{\mathrm{d_y}} w_i\, y_i^*(x+\mu v)\, v\right] = \sum_{i=1}^{\mathrm{d_y}} w_i \frac{\mathrm{d_x}}{\mu}\, \mathbb{E}_{v\sim\mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y_i^*(x+\mu v)\, v] \\ &= \sum_{i=1}^{\mathrm{d_y}} w_i\, \nabla_x [y_\mu]_i(x) \;=\; J y_\mu(x)^\top w, \end{aligned} \tag{25}$$

where $\frac{\mathrm{d_x}}{\mu}\, \mathbb{E}_{v\sim\mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y_i^*(x+\mu v)\, v] = \nabla_x [y_\mu]_i(x)$ by (23) and the last equality uses $[J y_\mu(x)^\top w]_j = \sum_{i=1}^{\mathrm{d_y}} w_i\, [\nabla_x [y_\mu]_i(x)]_j$. Therefore, by (25):

$$\left\| \left[J y_\mu(x) - J y_\mu(z)\right]^\top w \right\| = \|\nabla h_{w,\mu}(x) - \nabla h_{w,\mu}(z)\| \le \frac{k_1 L_y \sqrt{\mathrm{d_x}}}{\mu} \|x - z\|.$$

Since this holds for every unit vector $w \in \mathbb{R}^{\mathrm{d_y}}$,

$$\left\| J y_\mu(x) - J y_\mu(z) \right\|_2 = \sup_{\|w\|=1} \left\| \left[J y_\mu(x) - J y_\mu(z)\right]^\top w \right\| \le k_1\, \frac{L_y \sqrt{d_x}}{\mu}\, \|x - z\|.$$

This completes the proof. □

Lemma 5. *Assume that $y_\mu(x)$ is differentiable and L_y-Lipschitz. Then $\|J y_\mu(x)\|_2 \le L_y$.*

Proof. As y_μ is differentiable by assumption, the directional derivative $D_v y_\mu(x) = \lim_{t\to 0} \frac{y_\mu(x+tv) - y_\mu(x)}{t}$ of y_μ in direction v at x exists and equals $J y_\mu(x) v$. It follows that

$$\begin{aligned} \|J y_\mu(x) v\| &= \left\| \lim_{t\to 0} \frac{y_\mu(x+tv) - y_\mu(x)}{t} \right\| \\ &= \lim_{t\to 0} \left\| \frac{y_\mu(x+tv) - y_\mu(x)}{t} \right\| \le \lim_{t\to 0} \frac{L_y \|tv\|}{|t|} = L_y \|v\|, \end{aligned}$$

where the second equality follows from the continuity of the norm $\|\cdot\|$ and the existence of the limit $\lim_{t\to 0} \frac{y_\mu(x+tv) - y_\mu(x)}{t} = J y_\mu(x) v$, and the inequality follows as $y_\mu(x)$ is L_y-Lipschitz by assumption.

Then, for arbitrary unit vectors v with $\|v\| = 1$, $\|J y_\mu(x) v\| \le L_y$, and by definition of the spectral norm:

$$\|J y_\mu(x)\|_2 = \sup_{v\in\mathbb{R}^{\mathrm{d_x}}, v\neq 0} \frac{\|J y_\mu(x) v\|_2}{\|v\|_2} = \sup_{v\in\mathbb{R}^{\mathrm{d_x}}, \|v\|_2 = 1} \|J y_\mu(x) v\|_2 \le L_y. \qquad \square$$

C.2 Smoothness of surrogate function and estimator properties

We now analyze the surrogate function $F_\mu(x) = f(x, y_\mu(x))$, with $y_\mu(x)$ defined in (3). We first establish that the distance between $F_\mu(x)$ and $F(x)$ is bounded by a constant depending on μ (Prop. 3), then establish that F_μ is differentiable and smooth under Assumption 2 (Lem. 6).

Proposition 3. *Under Assumption 2, let $F(x) = f(x, y^*(x))$ and $F_\mu(x) = f(x, y_\mu(x))$, where y_μ is defined as in* (3)*. Then $|F_\mu(x) - F(x)| \le L_f L_y \mu$.*

Proof. By Assumption 2, $f(x, y)$ is L_f-Lipschitz continuous in y for fixed x and by Lem. 4(i), $\|y_\mu(x) - y^*(x)\| \le L_y \mu$, hence

$$|F_\mu(x) - F(x)| = |(f(x, y_\mu(x)) - f(x, y^*(x))| \le L_f \|y_\mu(x) - y^*(x)\| \le L_f L_y \mu. \qquad \square$$

The following lemma establishes that F_μ is smooth under Assumption 2. Unlike the smoothness of $\widetilde{F}_\mu(x, y^*(x)) = \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[f(x + \mu u, y^*(x + \mu u))]$, which follows directly from the smoothing properties in Thm. 3, the smoothness of F_μ does not follow immediately since F_μ does not smooth $f(x, y^*(x))$ directly. Instead, the proof exploits the smoothness of f in both x and y, together with the smoothness of y_μ.

Lemma 6. *Under Assumption 2, $F_\mu(x) = f(x, y_\mu(x))$ is differentiable and L_F-smooth, with $L_F = L_g(1 + L_y)^2 + \frac{k_1 L_f L_y \sqrt{\mathrm{d_x}}}{\mu}$, and $k_1 > 0$ is the constant in Lem. 4(iii).*

Proof of Lem. 6. As f is differentiable by Assumption 2 and y_μ is differentiable by Lem. 3, the composition $F_\mu(x) = f(x, y_\mu(x))$ is differentiable. By the chain rule:

$$\nabla F_\mu(x) = \nabla_x f(x, y_\mu(x)) + J y_\mu(x)^\top \nabla_y f(x, y_\mu(x)). \tag{26}$$

To show that F_μ is L_F-smooth, we need to show that for any $x_1, x_2 \in \mathbb{R}^{\mathrm{d_x}}$:

$$\|\nabla F_\mu(x_1) - \nabla F_\mu(x_2)\| \le L_F \|x_1 - x_2\|.$$

Using (26), we have

$$\begin{aligned}
&\|\nabla F_\mu(x_1) - \nabla F_\mu(x_2)\| \\
&= \|\nabla_x f(x_1, y_\mu(x_1)) + J y_\mu(x_1)^\top \nabla_y f(x_1, y_\mu(x_1)) - \nabla_x f(x_2, y_\mu(x_2)) - J y_\mu(x_2)^\top \nabla_y f(x_2, y_\mu(x_2))\| \\
&\le \|\nabla_x f(x_1, y_\mu(x_1)) - \nabla_x f(x_2, y_\mu(x_2))\| + \|J y_\mu(x_1)^\top \nabla_y f(x_1, y_\mu(x_1)) - J y_\mu(x_2)^\top \nabla_y f(x_2, y_\mu(x_2))\|.
\end{aligned}$$

For the first term, since ∇f is L_g-Lipschitz by Assumption 2 and y_μ is L_y-Lipschitz by Lem. 4(ii):

$$\begin{aligned}
\|\nabla_x f(x_1, y_\mu(x_1)) - \nabla_x f(x_2, y_\mu(x_2))\| &\le L_g \|(x_1, y_\mu(x_1)) - (x_2, y_\mu(x_2))\| \\
&\le L_g(\|x_1 - x_2\| + \|y_\mu(x_1) - y_\mu(x_2)\|) \\
&\le L_g(1 + L_y)\|x_1 - x_2\|. \qquad (27)
\end{aligned}$$

For the second term, we add and subtract $J y_\mu(x_1)^\top \nabla_y f(x_2, y_\mu(x_2))$:

$$\begin{aligned}
&\|J y_\mu(x_1)^\top \nabla_y f(x_1, y_\mu(x_1)) - J y_\mu(x_2)^\top \nabla_y f(x_2, y_\mu(x_2))\| \\
&= \|J y_\mu(x_1)^\top \nabla_y f(x_1, y_\mu(x_1)) - J y_\mu(x_1)^\top \nabla_y f(x_2, y_\mu(x_2)) \\
&\quad + J y_\mu(x_1)^\top \nabla_y f(x_2, y_\mu(x_2)) - J y_\mu(x_2)^\top \nabla_y f(x_2, y_\mu(x_2))\| \\
&\le \|J y_\mu(x_1)^\top [\nabla_y f(x_1, y_\mu(x_1)) - \nabla_y f(x_2, y_\mu(x_2))]\| + \|[J y_\mu(x_1) - J y_\mu(x_2)]^\top \nabla_y f(x_2, y_\mu(x_2))\|.
\end{aligned}$$

Bounding the first summand:

$$\begin{aligned}
\|J y_\mu(x_1)^\top [\nabla_y f(x_1, y_\mu(x_1)) - \nabla_y f(x_2, y_\mu(x_2))]\| &\le \|J y_\mu(x_1)\|_2 \|\nabla_y f(x_1, y_\mu(x_1)) - \nabla_y f(x_2, y_\mu(x_2))\| \\
&\le L_y \cdot L_g \|(x_1, y_\mu(x_1)) - (x_2, y_\mu(x_2))\| \quad \text{(by Lem. 5)} \\
&\le L_y L_g (1 + L_y)\|x_1 - x_2\| \quad \text{(by (27))}.
\end{aligned}$$

Bounding the second summand:

$$\begin{aligned}
\|[J y_\mu(x_1) - J y_\mu(x_2)]^\top \nabla_y f(x_2, y_\mu(x_2))\| &\le \|J y_\mu(x_1) - J y_\mu(x_2)\|_2 \|\nabla_y f(x_2, y_\mu(x_2))\| \\
&\le \frac{k_1 L_y \sqrt{\mathrm{d_x}}}{\mu} \|x_1 - x_2\| \cdot L_f \quad \text{(by Lem. 4(iii) and Assumption 2)}.
\end{aligned}$$

Then, combining all bounds,

$$\|\nabla F_\mu(x_1) - \nabla F_\mu(x_2)\| \leq L_g(1+L_y)\|x_1 - x_2\| + L_y L_g(1+L_y)\|x_1 - x_2\| + \frac{k_1 L_f L_y \sqrt{\mathrm{d_x}}}{\mu}\|x_1 - x_2\|$$
$$= \left[L_g(1+L_y)^2 + \frac{k_1 L_f L_y \sqrt{\mathrm{d_x}}}{\mu}\right]\|x_1 - x_2\|$$

This establishes that F_μ has Lipschitz-continuous gradient with constant $L_F = L_g(1+L_y)^2 + \frac{k_1 L_f L_y \sqrt{\mathrm{d_x}}}{\mu}$. □

C.2.1 Properties of the Zeroth-Order Estimators

We now turn to the analysis of the estimators H_t and g_t of PZOS.

Lemma 7. *Under Assumption 2, $H_t = \frac{\mathrm{d_x}}{2\mu}\left(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t)\right) v_t^\top$, with $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$, is an unbiased estimator of $Jy_\mu(x_t)$, i.e.,*

$$\mathbb{E}[\, H_t \mid x_t \,] = Jy_\mu(x_t),$$

where $Jy_\mu(x_t) = \frac{\mathrm{d_x}}{\mu}\,\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x_t + \mu v_t)\, v_t^\top\Big]$.

Proof. Given x_t, the randomness in H_t arises solely from $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$, as $y^*(x_t + \mu v_t)$ is uniquely determined for any given $x_t + \mu v_t$. Further observe that $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$ has a symmetric distribution, i.e., v_t and $-v_t$ have the same distribution. Therefore, $\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x_t + \mu(-v_t))\,(-v_t^\top)\Big] = \mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x_t + \mu v_t)\, v_t^\top\Big]$. It follows that

$$\begin{aligned}
\mathbb{E}[H_t \mid x_t] &= \mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\left[\frac{\mathrm{d_x}}{2\mu}(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t))\, v_t^\top\right] \\
&= \frac{\mathrm{d_x}}{2\mu}\,\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t))\, v_t^\top\Big] \\
&= \frac{\mathrm{d_x}}{2\mu}\Big(\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x_t + \mu v_t)\, v_t^\top] - \mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x_t - \mu v_t)\, v_t^\top]\Big) \\
&= \frac{\mathrm{d_x}}{2\mu}\Big(\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x_t + \mu v_t)\, v_t^\top] + \mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x_t + \mu(-v_t))\,(-v_t^\top)]\Big) \\
&= \frac{\mathrm{d_x}}{2\mu}\Big(\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x_t + \mu v_t)\, v_t^\top] + \mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}[y^*(x_t + \mu v_t)\, v_t^\top]\Big) \\
&= \frac{\mathrm{d_x}}{\mu}\,\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x_t + \mu v_t)\, v_t^\top\Big].
\end{aligned}$$

By Lem. 3, $Jy_\mu(x_t) = \frac{\mathrm{d_x}}{\mu}\,\mathbb{E}_{v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})}\Big[y^*(x_t + \mu v_t)\, v_t^\top\Big]$, hence $\mathbb{E}[\, H_t \mid x_t \,] = Jy_\mu(x_t)$ follows. □

Despite the unbiasedness of H_t, the estimator g_t of PZOS is *not* an unbiased estimator of the gradient of F_μ. We recall by (26) that ∇F_μ is given by

$$\nabla F_\mu(x_t) = \nabla_x f(x_t, y_\mu(x_t)) + Jy_\mu(x_t)^\top \nabla_y f(x_t, y_\mu(x_t)).$$

In contrast, the expected value of the estimator g_t of PZOS for given x_t is given by

$$\begin{aligned}
\mathbb{E}[\, g_t \mid x_t \,] &= \mathbb{E}[\, \nabla_x f(x_t, y^*(x_t)) + H_t^\top \nabla_y f(x_t, y^*(x_t)) \mid x_t \,] \\
&= \nabla_x f(x_t, y^*(x_t)) + \mathbb{E}[\, H_t \mid x_t \,]^\top \nabla_y f(x_t, y^*(x_t)) \\
&= \nabla_x f(x_t, y^*(x_t)) + Jy_\mu(x_t)^\top \nabla_y f(x_t, y^*(x_t)), \qquad (28)
\end{aligned}$$

where we use $\mathbb{E}[H_t \mid x_t] = Jy_\mu(x_t)$ by Lem. 7. Therefore, g_t is a *biased* estimator of $\nabla F_\mu(x_t)$. Below, we bound the distance of $\mathbb{E}[g_t | x_t]$ to $\nabla F_\mu(x_t)$ by a constant that depends on μ.

Lemma 8. *Let $\{g_t\}_{t=0}^{T-1}$ and $\{x_t\}_{t=0}^{T-1}$ be generated by Alg. 1. Under Assumption 2,*

$$\|\,\mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t)\| \le L_g(1+L_y)L_y\mu =: C_p\mu.$$

Proof. By (26) and (28):

$$\begin{aligned}&\mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\\ &= \nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_\mu(x_t)) + Jy_\mu(x_t)^\top[\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_\mu(x_t))],\end{aligned}$$

hence

$$\begin{aligned}&\|\,\mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\|\\ &= \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_\mu(x_t)) + Jy_\mu(x_t)^\top[\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_\mu(x_t))]\|\\ &\le \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_\mu(x_t))\| + \|Jy_\mu(x_t)^\top[\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_\mu(x_t))]\|.\end{aligned}$$

For the first term, by L_g-Lipschitz continuity of ∇f and $\|y^*(x_t) - y_\mu(x_t)\| \le L_y\mu$ by Lem. 4(i):

$$\begin{aligned}\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_\mu(x_t))\| &\le L_g\|(x_t, y^*(x_t)) - (x_t, y_\mu(x_t))\|\\ &= L_g\|y^*(x_t) - y_\mu(x_t)\| \le L_g L_y\mu.\end{aligned}$$

For the second term, using $\|Ax\| \le \|A\|_2\|x\|$, $\|Jy_\mu(x_t)\|_2 \le L_y$ by Lem. 5, $\|y^*(x_t) - y_\mu(x_t)\| \le L_y\mu$ by Lem. 4 (i), and L_g-Lipschitz continuity of $\nabla_y f$:

$$\begin{aligned}&\|Jy_\mu(x_t)^\top[\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_\mu(x_t))]\|\\ &\le \|Jy_\mu(x_t)\| \cdot \|\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_\mu(x_t))\|\\ &\le L_y \cdot L_g\|y^*(x_t) - y_\mu(x_t)\| \le L_y \cdot L_g L_y\mu = L_g L_y^2\mu.\end{aligned}$$

Combining both terms,

$$\|\,\mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\| \le L_g L_y\mu + L_g L_y^2\mu = L_g(1+L_y)L_y\mu. \qquad \square$$

We now establish a bound on the second moment of the estimator g_t of PZOS.

Lemma 9. *Let $\{g_t\}_{t=0}^{T-1}$ and $\{x_t\}_{t=0}^{T-1}$ be generated by PZOS. Under Assumption 2,*

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \le k_2\,\mathrm{d_x}\,L_f^2 L_y^2 + L_f^2(1+2L_y) =: \sigma_{\mathrm{d_x}}^2,$$

and k_2 is the universal constant from [24, Lem. 10].

Proof. Recall that the estimator of PZOS is

$$g_t = \nabla_x f(x_t, y^*(x_t)) + H_t^\top \nabla_y f(x_t, y^*(x_t)),$$

where $H_t = \frac{d_x}{2\mu}(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t))\, v_t^\top$ with $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$. Conditioned on x_t, the terms $\nabla_x f(x_t, y^*(x_t))$ and $\nabla_y f(x_t, y^*(x_t))$ are deterministic and only H_t (through v_t) is random. Expanding the squared norm:

$$\begin{aligned}\mathbb{E}[\|g_t\|^2 \mid x_t] &= \mathbb{E}\big[\|\nabla_x f(x_t, y^*(x_t)) + H_t^\top \nabla_y f(x_t, y^*(x_t))\|^2 \;\big|\; x_t\big]\\ &= \|\nabla_x f(x_t, y^*(x_t))\|^2 &(29)\\ &\quad + 2\big\langle \nabla_x f(x_t, y^*(x_t)),\ \mathbb{E}[H_t^\top \nabla_y f(x_t, y^*(x_t)) \mid x_t]\big\rangle &(30)\\ &\quad + \mathbb{E}\big[\|H_t^\top \nabla_y f(x_t, y^*(x_t))\|^2 \;\big|\; x_t\big]. &(31)\end{aligned}$$

We bound each term.

Term (29): By Assumption 2, $\|\nabla_x f(x_t, y^*(x_t))\| \le L_f$, so

$$\|\nabla_x f(x_t, y^*(x_t))\|^2 \le L_f^2. \tag{32}$$

Term (30): By Assumption 2, $\|\nabla_x f(x_t, y^*(x_t))\|, \|\nabla_y f(x_t, y^*(x_t))\| \le L_f$, by Lem. 7, $\|\,\mathbb{E}[H_t \mid x_t]\|_2 = \|Jy_\mu(x_t)\|_2$, and by Lem. 5, $\|Jy_\mu(x_t)\|_2 \le L_y$. Applying the Cauchy-Schwarz inequality $\langle a, b\rangle \le |\langle a, b\rangle| \le \|a\| \cdot \|b\|$ to $a = \nabla_x f$ and $b = H_t^\top \nabla_y f$ and combining, this yields:

$$\begin{aligned}2\big\langle \nabla_x f(x_t, y^*(x_t)),\ \mathbb{E}[H_t^\top \nabla_y f(x_t, y^*(x_t)) \mid x_t]\big\rangle &\le 2\|\nabla_x f(x_t, y^*(x_t))\|\,\|\,\mathbb{E}[H_t^\top \nabla_y f(x_t, y^*(x_t)) \mid x_t]\|\\ &\le 2L_f \cdot \|\,\mathbb{E}[H_t \mid x_t]^\top \nabla_y f(x_t, y^*(x_t))\|\\ &\le 2L_f \cdot \|\,\mathbb{E}[H_t \mid x_t]\|_2 \cdot \|\nabla_y f(x_t, y^*(x_t))\|\\ &\le 2L_f \cdot L_y L_f = 2L_f^2 L_y. &(33)\end{aligned}$$

Term (31): Fix $\nabla_y f(x_t, y^*(x_t))$ and define the scalar function $h_x(z) = \nabla_y f(x_t, y^*(x_t))^\top y^*(z)$. Since y^* is L_y-Lipschitz and $\|\nabla_y f(x_t, y^*(x_t))\| \le L_f$ by Assumption 2:

$$\begin{aligned}|h_x(z) - h_x(z')| &= |\nabla_y f(x_t, y^*(x_t))^\top (y^*(z) - y^*(z'))| \\ &\le \|\nabla_y f(x_t, y^*(x_t))\| \, \|y^*(z) - y^*(z')\| \\ &\le L_f L_y \|z - z'\|,\end{aligned}$$

so h_x is $(L_f L_y)$-Lipschitz. We observe that

$$\begin{aligned}H_t^\top \nabla_y f(x_t, y^*(x_t)) &= \left[\frac{d_x}{2\mu}(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t))\, v_t^\top\right]^\top \nabla_y f(x_t, y^*(x_t)) \\ &= \frac{\mathrm{d_x}}{2\mu}\big[\nabla_y f(x_t, y^*(x_t))^\top \big(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t)\big)\big] v_t \\ &= \frac{\mathrm{d_x}}{2\mu}\big[h_x(x_t + \mu v_t) - h_x(x_t - \mu v_t)\big] v_t,\end{aligned}$$

equals the standard two-point gradient estimator $\frac{\mathrm{d_x}}{2\mu}\big[h_x(x_t + \mu v_t) - h_x(x_t - \mu v_t)\big] v_t$ of Shamir [24] for the scalar Lipschitz function h_x at x_t with smoothing parameter μ and random direction $v_t \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$. Specifically, by [24, Lem. 10] and [18, Appendix D.1], for such a two-point estimator of an L-Lipschitz function $h : \mathbb{R}^{\mathrm{d_x}} \to \mathbb{R}$, we have

$$\mathbb{E}\left[\left\|\frac{\mathrm{d_x}}{2\mu}(h(x + \mu v) - h(x - \mu v))v\right\|^2\right] \le k_2 \, \mathrm{d_x} \, L^2,$$

where $k_2 > 0$ is a constant. Substituting $L = L_f L_y$ and $h = h_x$ gives

$$\mathbb{E}\big[\|H_t^\top \nabla_y f(x_t, y^*(x_t))\|^2 \;\big|\; x_t\big] \le k_2 \, \mathrm{d_x} \, L_f^2 L_y^2. \tag{34}$$

Substituting (32), (33), and (34) into (29)–(31):

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \le k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + L_f^2 + 2L_f^2 L_y = k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + L_f^2(1 + 2L_y). \qquad \square$$

We utilize Lem. 9 for a comparison of second-moment bounds between the estimator g_t of PZOS and the estimator $\widetilde{g}_t$ of ZOS.

Lemma 10 (Restatement of Lemma 1)**.** *Under Assumption 2 and v_t sampled uniformly from the unit sphere $\mathbb{S}^{\mathrm{d_x}}$, let*

$$g_t = \nabla_x f(x_t, y^*(x_t)) + H_t^\top \nabla_y f(x_t, y^*(x_t)), \text{ with } \quad H_t = \frac{\mathrm{d_x}}{2\mu}(y^*(x_t + \mu v_t) - y^*(x_t - \mu v_t))\, v_t^\top$$

be the estimator used in PZOS, and let

$$\widetilde{g}_t = \frac{\mathrm{d_x}}{2\mu}\big[f(x_t + \mu v_t, y^*(x_t + \mu v_t)) - f(x_t - \mu v_t, y^*(x_t - \mu v_t))\big] v_t$$

be the standard two-point zeroth-order gradient estimator of ZOS. Then the conditional second moment bounds satisfy:

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \le k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + L_f^2(1 + 2L_y) =: \sigma_{\mathrm{d_x}}^2, \tag{35}$$

$$\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \le k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + k_2 \, \mathrm{d_x} \, L_f^2(1 + 2L_y) =: \widetilde{\sigma}_{\mathrm{d_x}}^2, \tag{36}$$

where $k_2 \ge 1$ is the universal constant from [24, Lem. 10]. The same bounds hold for the conditional variances of the two estimators:

$$\mathbb{E}[\|g_t - \mathbb{E}[g_t \mid x_t]\|^2 \mid x_t] \le k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + L_f^2(1 + 2L_y), \tag{37}$$

$$\mathbb{E}[\|\widetilde{g}_t - \mathbb{E}[\widetilde{g}_t \mid x_t]\|^2 \mid x_t] \le k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + k_2 \, \mathrm{d_x} \, L_f^2(1 + 2L_y), \tag{38}$$

Since $k_2 \ge 1$ and $\mathrm{d_x} \ge 1$, the bounds for g_t are at least as tight as those for $\widetilde{g}_t$.

Proof. We first recall the relevant second moment bounds, then show that $k_2 \geq 1$. The bounds on the conditional variances follow by noting that the conditional variance is no larger than the second moment.

By Lem. 9, the estimator g_t of PZOS satisfies

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \leq k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + L_f^2(1 + 2L_y). \tag{39}$$

For the estimator $\widetilde{g}_t$ of ZOS, we apply a general result on two-point zeroth-order estimators. Specifically, Shamir [24, Lem. 10] show that, for any L-Lipschitz function $h : \mathbb{R}^d \to \mathbb{R}$, the two-point estimator $\widetilde{g} = \frac{d}{2\mu}\big[h(x + \mu v) - h(x - \mu v)\big]v$, with $v \sim \mathcal{U}(\mathbb{S}^d)$ has second moment no higher than $k_2 \, d \, L^2$ (i.e., $\mathbb{E}[\|\widetilde{g}\|^2 \mid x] \leq k_2 \, d \, L^2$), where $k_2 > 0$ is a universal constant. $\widetilde{g}_t$ is precisely this two-point estimator applied to the composite function $F(x) = f(x, y^*(x))$, which is $L_f(1+L_y)$-Lipschitz as

$$\begin{aligned} |F(x) - F(z)| &= |f(x, y^*(x)) - f(z, y^*(z))| \\ &\leq |f(x, y^*(x)) - f(z, y^*(x))| + |f(z, y^*(x)) - f(z, y^*(z))| \\ &\leq L_f \|x - z\| + L_f L_y \|x - z\| = L_f(1 + L_y)\|x - z\|. \end{aligned} \tag{40}$$

This gives:

$$\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \leq k_2 \, \mathrm{d_x} \, L_f^2 (1 + L_y)^2 = k_2 \, \mathrm{d_x} \, L_f^2 L_y^2 + k_2 \, \mathrm{d_x} \, L_f^2 (1 + 2L_y). \tag{41}$$

It remains to show that $k_2 \geq 1$. To do so, we observe that the bound $\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \leq k_2 \, \mathrm{d_x} \, L^2$ from [24, Lem. 10] must hold for every L-Lipschitz function. We present a specific function for which $\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] = \mathrm{d_x} \, L^2$, which forces $k_2 \geq 1$. For any $L > 0$, consider the linear function $h(x) = L e_1^\top x$, with $e_1 = (1, 0, \ldots, 0)^\top$, which is exactly L-Lipschitz. The two-point estimator $\widetilde{g}$ at any point x evaluates to

$$\widetilde{g} = \frac{\mathrm{d_x}}{2\mu}\big[h(x + \mu v) - h(x - \mu v)\big]v = \frac{\mathrm{d_x}}{2\mu}\big[2L\mu v_1\big]v = \mathrm{d_x} \, L v_1 v,$$

where $v = (v_1, \ldots, v_{\mathrm{d_x}})^\top \sim \mathcal{U}(\mathbb{S}^{\mathrm{d_x}})$. Since $\|v\| = 1$, $\mathbb{E}[\|\widetilde{g}\|^2 \mid x] = \mathbb{E} \, \| \, \mathrm{d_x} \, L v_1 v\|^2 = {\mathrm{d_x}}^2 \, L^2 \, \mathbb{E}[v_1^2]$. By rotational symmetry of the uniform distribution on $\mathbb{S}^{\mathrm{d_x}}$, all components v_i are identically distributed, so $\mathbb{E}[v_1^2] = \mathbb{E}[v_2^2] = \ldots = \mathbb{E}[v_{\mathrm{d_x}}^2]$. Since $\sum_{i=1}^{\mathrm{d_x}} v_i^2 = 1$ on the unit sphere, taking expectations gives $\mathrm{d_x} \, \mathbb{E}[v_1^2] = 1$, hence $\mathbb{E}[v_1^2] = 1/\mathrm{d_x}$. It follows that $\mathbb{E}[\|\widetilde{g}\|^2 \mid x] = {\mathrm{d_x}}^2 \, L^2 \cdot \frac{1}{\mathrm{d_x}} = \mathrm{d_x} \, L^2$. Since $\mathbb{E}[\|\widetilde{g}\|^2 \mid x] \leq k_2 \, \mathrm{d_x} \, L^2$ must hold, this implies $k_2 \geq 1$.

We conclude by noting that the variance of both estimators is upper bounded by their respective second moment, since

$$\mathbb{E}[\|g_t - \mathbb{E}[g_t \mid x_t]\|^2 \mid x_t] = \mathbb{E}[\|g_t\|^2 \mid x_t] - \| \, \mathbb{E}[g_t \mid x_t]\|^2 \leq \mathbb{E}[\|g_t\|^2 \mid x_t],$$

and analogously for $\widetilde{g}_t$. □

C.3 Convergence to Goldstein Stationary Points

We turn to the final step to obtain our convergence results. To do so, we first establish that the Jacobian of the smoothed mapping y_μ lies within the μ-Goldstein subdifferential of y^* (Lem. 11). It then follows that the estimator g_t of PZOS lies in the δ-partial Goldstein subdifferential of F, for $\delta \geq \mu$. For the Goldstein subdifferential, we show that ∇F_μ is close to the Goldstein subdifferential of F, with a distance that depends on the Lipschitz constants of f, ∇f, and y^*, as well as the smoothing parameter μ. We then prove our convergence results.

Lemma 11 (Jacobian of smoothed response y_μ lies in μ-Goldstein subdifferential of y^*)**.** *Let $y^*(x) : \mathbb{R}^{\mathrm{d_x}} \to \mathcal{Y}$ be L_y-Lipschitz continuous and single-valued for given x, and $\mathcal{Y}$ be a closed and convex set. Define $y_\mu(x) = \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[y^*(x + \mu u)]$. Then for all $x \in \mathbb{R}^{\mathrm{d_x}}$:*

$$J y_\mu(x) \in \partial_\mu^G \, y^*(x) \coloneqq \operatorname{conv}\left\{ \bigcup_{z \in \mathbb{B}(x,\mu)} \partial y^*(z) \right\},$$

where ∂y^* denotes the Clarke generalized Jacobian of y^*.

Proof. The proof proceeds in two steps. First, we show that the Jacobian of the smoothed mapping y_μ can be expressed as the expectation of the Jacobian of the original mapping y^* over the uniform distribution on the unit ball $\mathbb{B}^{\mathrm{d_x}}$ at points where y^* is differentiable. Second, we show that this Jacobian lies within the convex hull of Clarke generalized Jacobians $\partial y^*(z)$ of points $z \in \mathbb{B}(x, \mu)$.

Before we begin, we recall that, since y^* is L_y-Lipschitz continuous, by Rademacher's Theorem it is differentiable almost everywhere. Therefore, the set

$$\Omega \coloneqq \{u \in \mathbb{B}^{\mathrm{d_x}} : y^* \text{ is not differentiable at } x + \mu u\}$$

of points at which y^* is non-differentiable has Lebesgue measure zero. Let $p(u)$ denote the probability density function of the uniform distribution on $\mathbb{B}^{\mathrm{d_x}}$. Since Ω has measure zero, $\int_{\mathbb{B}^{\mathrm{d_x}}} p(u)\, du = \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} p(u)\, du = 1$.

Step 1: Expressing $Jy_\mu(x)$ as an expectation of $Jy^*(z)$. We show that, for every $x \in \mathbb{R}^{\mathrm{d_x}}$,

$$Jy_\mu(x) = \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} Jy^*(x + \mu u)\, p(u)\, du.$$

We recall by Lem. 3 that y_μ is continuously differentiable. Let e_j denote the j-th standard basis vector in $\mathbb{R}^{\mathrm{d_x}}$. For each $i \in \{1, \ldots, \mathrm{d_y}\}$ and $j \in \{1, \ldots, \mathrm{d_x}\}$, the partial derivative of the i-th component $y_{\mu,i}$ of y_μ with respect to x_j is defined by the limit of the difference quotient:

$$\frac{\partial y_{\mu,i}}{\partial x_j}(x) = \lim_{t \to 0} \frac{y_{\mu,i}(x + te_j) - y_{\mu,i}(x)}{t}.$$

Substituting the definition of $y_\mu(x) = \int_{\mathbb{B}^{\mathrm{d_x}}} y^*(x + \mu u) p(u)\, du$ from (20), we have:

$$\frac{\partial y_{\mu,i}}{\partial x_j}(x) = \lim_{t \to 0} \int_{\mathbb{B}^{\mathrm{d_x}}} \frac{y_i^*(x + te_j + \mu u) - y_i^*(x + \mu u)}{t} p(u)\, du. \tag{42}$$

We proceed to show that Lebesgue's dominated convergence theorem applies, allowing us to interchange the limit and the integral. To do so, we verify that the difference quotient (i) converges pointwise almost everywhere and (ii) is dominated by an integrable function.

(i) *Pointwise limit almost everywhere:* For all $u \in \mathbb{B}^{\mathrm{d_x}} \setminus \Omega$, the point $x + \mu u$ lies in the set where y^* is differentiable. For such u, the difference quotient converges pointwise to the partial derivative:

$$\lim_{t \to 0} \frac{y_i^*(x + te_j + \mu u) - y_i^*(x + \mu u)}{t} = \frac{\partial y_i^*}{\partial x_j}(x + \mu u).$$

Since Ω has measure zero, this convergence holds for almost every $u \in \mathbb{B}^{\mathrm{d_x}}$.

(ii) *Dominated by integrable function:* By L_y-Lipschitz continuity of y^*, for all $t \neq 0$, the difference quotient is bounded:

$$\left\| \frac{y_i^*(x + te_j + \mu u) - y_i^*(x + \mu u)}{t} \right\| \leq \frac{L_y \|(x + te_j + \mu u) - (x + \mu u)\|}{|t|} = \frac{L_y \|te_j\|}{|t|} = L_y.$$

Since the constant function L_y is integrable over $\mathbb{B}^{\mathrm{d_x}}$ (hence $\mathbb{B}^{\mathrm{d_x}} \setminus \Omega$), the conditions for the dominated convergence theorem are satisfied. Thus, we can interchange the limit and the integral in (42):

$$\begin{aligned}
\frac{\partial y_{\mu,i}}{\partial x_j}(x) &= \lim_{t \to 0} \int_{\mathbb{B}^{\mathrm{d_x}}} \frac{y_i^*(x + te_j + \mu u) - y_i^*(x + \mu u)}{t} p(u)\, du \\
&= \lim_{t \to 0} \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} \frac{y_i^*(x + te_j + \mu u) - y_i^*(x + \mu u)}{t} p(u)\, du \\
&= \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} \frac{\partial y_i^*}{\partial x_j}(x + \mu u)\, p(u)\, du,
\end{aligned}$$

where the second equality holds because Ω has Lebesgue measure zero and the difference quotient is bounded by L_y, so $\int_\Omega (\cdot)\, p(u)\, du = 0$. The final equality applies the dominated convergence theorem

on $\mathbb{B}^{\mathrm{d_x}} \setminus \Omega$: the pointwise limit exists for all $u \in \mathbb{B}^{\mathrm{d_x}} \setminus \Omega$, and the constant L_y is integrable over this set since $\int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} p(u)\, du = 1$.

Collecting all partial derivatives into the Jacobian matrix yields:

$$Jy_\mu(x) = \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} Jy^*(x + \mu u)\, p(u)\, du = \mathbb{E}_{u \sim \mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[Jy^*(x + \mu u)], \tag{43}$$

where the expectation is taken over the points at which y^* is differentiable.

Step 2: $Jy_\mu(x)$ lies in the convex hull of Clarke generalized Jacobians. Define the set $K_\mu \coloneqq \operatorname{conv}\left\{\bigcup_{z \in \mathbb{B}(x,\mu)} \partial y^*(z)\right\} \subseteq \mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}$. We show that $Jy_\mu(x) \in K_\mu$ via a contradiction argument based on the strict hyperplane separation theorem.

First, we establish that K_μ is closed, compact, and convex. The Clarke generalized Jacobian $\partial y^*(z)$ is a nonempty, convex, compact subset of $\mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}$ for each z, and the set-valued mapping $z \mapsto \partial y^*(z)$ is upper semicontinuous [5, Prop. 2.6.2]. Since $\mathbb{B}(x, \mu)$ is compact and the image of a compact set under an upper semicontinuous map with compact values is compact [2, Prop. 3], the set $\bigcup_{z \in \mathbb{B}(x,\mu)} \partial y^*(z)$ is compact. In the finite-dimensional Euclidean space of matrices $\mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}$, the convex hull of a compact set is compact and closed, hence K_μ is closed, compact, and convex.

Suppose for contradiction that $Jy_\mu(x) \notin K_\mu$. We apply the strict hyperplane separation theorem in the finite-dimensional Hilbert space $(\mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}, \langle \cdot, \cdot \rangle_F)$, where the Frobenius inner product $\langle M, A \rangle_F = \operatorname{Tr}(M^\top A)$ corresponds to the Euclidean inner product $\langle \operatorname{vec}(M), \operatorname{vec}(A) \rangle$ under the isomorphism between $\mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}$ and $\mathbb{R}^{\mathrm{d_y} \cdot \mathrm{d_x}}$ induced by the vectorization operator $\operatorname{vec} : \mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}} \to \mathbb{R}^{\mathrm{d_y} \cdot \mathrm{d_x}}$. Since K_μ is closed, compact, and convex, and $\{Jy_\mu(x)\}$ (a singleton) is a closed set disjoint from K_μ by assumption, the strict hyperplane separation theorem guarantees that there exists a matrix $M \in \mathbb{R}^{\mathrm{d_y} \times \mathrm{d_x}}$ and a scalar $\alpha \in \mathbb{R}$ such that:

$$\langle M, Jy_\mu(x) \rangle_F > \alpha \quad \text{and} \quad \langle M, A \rangle_F < \alpha \quad \forall A \in K_\mu. \tag{44}$$

On the set $\mathbb{B}^{\mathrm{d_x}} \setminus \Omega$, the classical Jacobian $Jy^*(x + \mu u)$ exists and, by definition of the Clarke generalized Jacobian, satisfies $Jy^*(x + \mu u) \in \partial y^*(x + \mu u)$. Since $x + \mu u \in \mathbb{B}(x, \mu)$, we have $Jy^*(x + \mu u) \in K_\mu$ for all $u \in \mathbb{B}^{\mathrm{d_x}} \setminus \Omega$. By the second inequality in (44),

$$\langle M, Jy^*(x + \mu u) \rangle_F < \alpha \quad \text{for all } u \in \mathbb{B}^{\mathrm{d_x}} \setminus \Omega.$$

Using (43), we express the left-hand side of the first inequality in (44) as

$$\begin{aligned} \langle M, Jy_\mu(x) \rangle_F = \left\langle M, \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} Jy^*(x + \mu u)\, p(u)\, du \right\rangle_F &= \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} \langle M, Jy^*(x + \mu u) \rangle_F\, p(u)\, du \\ &\leq \int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} \alpha\, p(u)\, du = \alpha, \end{aligned}$$

where the final equality uses $\int_{\mathbb{B}^{\mathrm{d_x}} \setminus \Omega} p(u)\, du = 1$ as Ω has measure zero. This gives $\langle M, Jy_\mu(x) \rangle \leq \alpha$, which contradicts $\langle M, Jy_\mu(x) \rangle > \alpha$ from (44). Therefore, our initial assumption that $Jy_\mu(x) \notin K_\mu$ must be false, and we conclude that $Jy_\mu(x) \in K_\mu = \operatorname{conv}\left\{\bigcup_{z \in \mathbb{B}(x,\mu)} \partial y^*(z)\right\}$. □

C.3.1 Distance to Goldstein Subdifferential

The following lemma bounds the distance of the gradient of the surrogate function F_μ to elements in the Goldstein subdifferential $\partial_\mu^{\mathrm{G}} F(x)$.

Lemma 12 (Distance of ∇F_μ to Goldstein subdifferential)**.** *Let $F_\mu(x) = f(x, y_\mu(x))$ and $y_\mu(x)$ be defined as in Eq.* (3)*. Under Assumption 2, for any $x \in \mathbb{R}^{d_x}$ and $\mu > 0$, there exists $\bar{q} \in \partial_\mu^G F(x)$ such that*

$$\|\nabla F_\mu(x) - \bar{q}\| \leq (1 + L_y) L_g (1 + 2L_y) \mu =: C_p \mu.$$

Proof. By Lem. 11, $Jy_\mu(x) \in \operatorname{conv}\left\{\bigcup_{z\in\mathbb{B}(x,\mu)} \partial y^*(z)\right\}$. By Carathéodory's theorem, there exist $k \leq d_y d_x + 1$ matrices $M_1, \ldots, M_k$, where each $M_i \in \partial y^*(z_i)$ for some $z_i \in \mathbb{B}(x,\mu)$ (where z_i need not be distinct), and weights $\lambda_i \geq 0$ with $\sum_{i=1}^k \lambda_i = 1$ such that

$$Jy_\mu(x) = \sum_{i=1}^k \lambda_i M_i. \tag{45}$$

Define, for each $i \in \{1,\ldots,k\}$, $q_i = \nabla_x f(z_i, y^*(z_i)) + M_i^\top \nabla_y f(z_i, y^*(z_i))$, and let $\bar{q} = \sum_{i=1}^k \lambda_i q_i$, with identical weights as in (45). By recalling the application of Clarke's generalized chain rule to our problem setting (see (16)), we observe that q_i lies in the Clarke generalized gradient of $F(z_i)$, $q_i \in \partial F(z_i)$, for each i. Then, as $z_i \in \mathbb{B}(x,\mu)$ for each i, and $\bar{q}$ is a convex combination of q_i's, by definition of the Goldstein subdifferential in (17):

$$\bar{q} \in \operatorname{conv}\left\{\cup_{z\in\mathbb{B}(x,\mu)} \partial F(z)\right\} = \partial_\mu^{\mathrm{G}} F(x).$$

We now bound $\|\nabla F_\mu(x) - \bar{q}\|$. Using the chain rule (26) for F_μ and (45):

$$\begin{aligned}
\nabla F_\mu(x) - \bar{q} &= \nabla_x f(x, y_\mu(x)) + Jy_\mu(x)^\top \nabla_y f(x, y_\mu(x)) - \sum_{i=1}^k \lambda_i q_i \\
&= \nabla_x f(x, y_\mu(x)) + \left(\sum_{i=1}^k \lambda_i M_i\right)^\top \nabla_y f(x, y_\mu(x)) - \sum_{i=1}^k \lambda_i q_i \\
&= \sum_{i=1}^k \lambda_i \left[\nabla_x f(x, y_\mu(x)) + M_i^\top \nabla_y f(x, y_\mu(x)) - q_i\right].
\end{aligned} \tag{46}$$

For each $i \in \{1,\ldots,k\}$, we bound:

$$\begin{aligned}
&\|\nabla_x f(x, y_\mu(x)) + M_i^\top \nabla_y f(x, y_\mu(x)) - q_i\| \\
&= \|\nabla_x f(x, y_\mu(x)) - \nabla_x f(z_i, y^*(z_i)) + M_i^\top [\nabla_y f(x, y_\mu(x)) - \nabla_y f(z_i, y^*(z_i))]\| \\
&\leq \|\nabla_x f(x, y_\mu(x)) - \nabla_x f(z_i, y^*(z_i))\| + \|M_i\|_2 \|\nabla_y f(x, y_\mu(x)) - \nabla_y f(z_i, y^*(z_i))\|.
\end{aligned}$$

For the first term:

$$\begin{aligned}
&\|\nabla_x f(x, y_\mu(x)) - \nabla_x f(z_i, y^*(z_i))\| \\
&\leq L_g \|(x, y_\mu(x)) - (z_i, y^*(z_i))\| \leq L_g(\|x - z_i\| + \|y_\mu(x) - y^*(z_i)\|) \\
&\leq L_g(\mu + \|y_\mu(x) - y^*(x)\| + \|y^*(x) - y^*(z_i)\|) \leq L_g(\mu + L_y\mu + L_y\mu) = L_g\mu(1 + 2L_y),
\end{aligned}$$

where the first inequality follows by L_g-Lipschitz continuity of $\nabla_x f$ with respect to $w = (x, y)$, the second follows by the triangle inequality, the third by $\|x - z_i\| \leq \mu$ since $z_i \in \mathbb{B}(x,\mu)$ and adding and subtracting $y^*(x)$, and the fourth by $\|y_\mu(x) - y^*(x)\| \leq L_y\mu$ by Lem. 4(i) and $\|y^*(x) - y^*(z_i)\| \leq L_y\|x - z_i\| \leq L_y\mu$ by L_y-Lipschitz continuity of y^*, and $z_i \in \mathbb{B}(x,\mu)$.

Similarly, for the term $\|\nabla_y f(x, y_\mu(x)) - \nabla_y f(z_i, y^*(z_i))\|$, by L_g-Lipschitz continuity of $\nabla_y f$ and equivalent reasoning as above:

$$\|\nabla_y f(x, y_\mu(x)) - \nabla_y f(z_i, y^*(z_i))\| \leq L_g(\|x - z_i\| + \|y_\mu(x) - y^*(z_i)\|) \leq L_g\mu(1 + 2L_y).$$

For the second term, we have $\|M_i\|_2 \leq L_y$. This follows from the fact that $M_i \in \partial y^*(z_i)$, and elements of the Clarke generalized Jacobian are convex combinations of limits of classical Jacobians. Particularly, since $\|Jy^*(x)\|_2 \leq L_y$ at any point x at which $y^*(x)$ is differentiable (as $y^*(x)$ is L_y-Lipschitz), the bound is preserved when taking the limit by continuity of the spectral norm. As $M_i \in \partial y^*(z_i)$ is a convex combination of such limits along sequences converging to z_i, convexity of the spectral norm implies $\|M_i\|_2 \leq L_y$ for all $M_i \in \partial y^*(z_i)$.

Combining all three bounds:

$$\begin{aligned}
&\|\nabla_x f(x, y_\mu(x)) + M_i^\top \nabla_y f(x, y_\mu(x)) - q_i\| \\
&\leq \|\nabla_x f(x, y_\mu(x)) - \nabla_x f(z_i, y^*(z_i))\| + \|M_i\|_2 \|\nabla_y f(x, y_\mu(x)) - \nabla_y f(z_i, y^*(z_i))\| \\
&\leq L_g\mu(1 + 2L_y) + L_y \cdot L_g\mu(1 + 2L_y) = (1 + L_y)L_g(1 + 2L_y)\mu.
\end{aligned}$$

Taking the convex combination with weights $\lambda_i \geq 0$ for all $i \in \{1,\ldots,k\}$ and $\sum_i \lambda_i = 1$ and recalling (46) yields $\|\nabla F_\mu(x) - \bar{q}\| \leq (1 + L_y)L_g(1 + 2L_y)\mu$. □

The following lemma is the analogue of Lem. 12 for the black-box surrogate $\widetilde{F}_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[F(x+\mu u)]$ and the partial Goldstein subdifferential. It is used to establish convergence of ZOS to (δ,ε)-partial Goldstein stationary points (Prop. 4).

Lemma 13 (Distance of $\nabla\widetilde{F}_\mu$ to partial Goldstein subdifferential)**.** *Let $\widetilde{F}_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[F(x+\mu u)]$ with $F(x) = f(x, y^*(x))$. Under Assumption 2, for any $x \in \mathbb{R}^{\mathrm{d_x}}$ and $\mu > 0$, there exists $\bar{v} \in \partial^P_\mu F(x)$ such that*

$$\|\nabla\widetilde{F}_\mu(x) - \bar{v}\| \le L_g(1+L_y)^2\mu =: \widetilde{C}_p\mu.$$

Proof. Since y^* is L_y-Lipschitz, it is differentiable almost everywhere by Rademacher's Theorem, so the set $\Omega \coloneqq \{u \in \mathbb{B}^{\mathrm{d_x}} : y^* \text{ is not differentiable at } x+\mu u\}$ has Lebesgue measure zero. For $u \in \mathbb{B}^{\mathrm{d_x}}\setminus\Omega$, let $z = x+\mu u$. Since f is continuously differentiable and y^* is differentiable at z, F is differentiable at z, and the chain rule yields

$$\nabla F(z) = \nabla_x f(z, y^*(z)) + Jy^*(z)^\top \nabla_y f(z, y^*(z)).$$

Since F is $L_f(1+L_y)$-Lipschitz and differentiable at $x+\mu u$ for all $u \in \mathbb{B}^{\mathrm{d_x}}\setminus\Omega$, Step 1 of the proof of Lem. 11, applied to F in place of y^* (also see [18, Thm. 3.1]) yields

$$\nabla\widetilde{F}_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[\nabla F(x+\mu u)],$$

where the expectation is taken over $u \in \mathbb{B}^{\mathrm{d_x}}\setminus\Omega$.

Define $\bar{p} = \nabla_x f(x, y^*(x)) + Jy_\mu(x)^\top \nabla_y f(x, y^*(x))$. Since $Jy_\mu(x) \in \partial^{\mathrm{G}}_\mu y^*(x)$ by Lem. 11, $\bar{p} \in \partial^{\mathrm{P}}_\mu F(x)$ by definition of the partial Goldstein subdifferential in (18). Further, $Jy_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[Jy^*(x+\mu u)]$ by (43), so that $Jy_\mu(x)^\top\nabla_y f(x, y^*(x)) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[Jy^*(z)^\top\nabla_y f(x, y^*(x))]$. Hence

$$\begin{aligned}
&\|\nabla\widetilde{F}_\mu(x) - \bar{p}\| \\
&\quad= \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}\left[\nabla_x f(z, y^*(z)) - \nabla_x f(x, y^*(x)) + Jy^*(z)^\top\left(\nabla_y f(z, y^*(z)) - \nabla_y f(x, y^*(x))\right)\right]\right\| \\
&\quad\le \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}\left[\|\nabla_x f(z, y^*(z)) - \nabla_x f(x, y^*(x))\| + \|Jy^*(z)\|_2\,\|\nabla_y f(z, y^*(z)) - \nabla_y f(x, y^*(x))\|\right] \\
&\quad\le \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}\left[L_g(1+L_y)\mu + L_yL_g(1+L_y)\mu\right] = L_g(1+L_y)^2\mu,
\end{aligned}$$

where the first inequality follows by Jensen's inequality and the triangle inequality, and the second by L_g-Lipschitz continuity of $\nabla_x f$ and $\nabla_y f$ together with $\|z-x\| \le \mu$ and $\|y^*(z) - y^*(x)\| \le L_y\|z-x\| \le L_y\mu$ as y^* is L_y-Lipschitz by Assumption 2, which together give $\|(z, y^*(z)) - (x, y^*(x))\| \le (1+L_y)\mu$, and $\|Jy^*(z)\|_2 \le L_y$ at points of differentiability as y^* is L_y-Lipschitz by similar reasoning as in Lem. 5 . □

C.3.2 Proof of Theorem 1

We now state and prove our convergence of PZOS to (δ,ε)-partial Goldstein stationarity (Thm. 4) and to (δ,ε)-Goldstein stationarity (Thm. 5), for suitably chosen smoothing parameters μ.

Theorem 4 (Convergence of PZOS to (δ,ε)-partial Goldstein stationary point)**.** *Let x^R be chosen uniformly at random from the iterates $\{x_0, \ldots, x_{T-1}\}$ generated by PZOS (Alg. 1). Under Assumptions 1 and 2, with step size $\alpha = \sqrt{2(\Delta + 2L_fL_y\mu)/(TL_F\sigma_{\mathrm{d_x}}^2)}$, smoothing parameter $\mu = \min(\delta, \varepsilon/(\sqrt{2}C_p)) < 1$, and $C_p = L_g(1+L_y)L_y$, x^R is, in expectation, a (δ,ε)-partial Goldstein stationary point of F after $T = \mathcal{O}\left({\mathrm{d_x}}^{3/2}/(\mu\,\varepsilon^4)\right)$ iterations, i.e.,*

$$\mathbb{E}\left[\min\left\{\|g\| : g \in \partial^P_\delta F(x^R)\right\}\right] \le \varepsilon$$

Proof. By Lem. 6, $F_\mu(x) = f(x, y_\mu(x))$ is L_F-smooth, with $L_F = L_g(1+L_y)^2 + \frac{k_1L_fL_y\sqrt{\mathrm{d_x}}}{\mu}$-Lipschitz continuous gradient. By the descent lemma for F_μ and update step $x_{t+1} = x_t - \alpha g_t$ by PZOS:

$$\begin{aligned}
F_\mu(x_{t+1}) &\le F_\mu(x_t) + \langle\nabla F_\mu(x_t), x_{t+1} - x_t\rangle + \frac{L_F}{2}\|x_{t+1} - x_t\|^2 \\
&= F_\mu(x_t) - \alpha\langle\nabla F_\mu(x_t), g_t\rangle + \frac{\alpha^2L_F}{2}\|g_t\|^2.
\end{aligned}$$

Taking the conditional expectation given x_t:

$$\mathbb{E}[F_\mu(x_{t+1}) \mid x_t] \le F_\mu(x_t) - \alpha\langle \nabla F_\mu(x_t), \mathbb{E}[g_t \mid x_t]\rangle + \frac{\alpha^2 L_F}{2}\,\mathbb{E}[\|g_t\|^2 \mid x_t]. \tag{47}$$

Rewriting the inner product:

$$\begin{aligned}
\langle \nabla F_\mu(x_t), \mathbb{E}[g_t \mid x_t]\rangle &= \langle \mathbb{E}[g_t|x_t] - (\mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t)),\, \mathbb{E}[g_t|x_t]\rangle \\
&= \|\,\mathbb{E}[g_t|x_t]\|^2 - \langle \mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t), \mathbb{E}[g_t|x_t]\rangle \\
&\ge \frac{1}{2}\|\,\mathbb{E}[g_t|x_t]\|^2 - \frac{1}{2}\|\,\mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t)\|^2 \\
&\ge \frac{1}{2}\|\,\mathbb{E}[g_t|x_t]\|^2 - \frac{1}{2}C_p^2\mu^2,
\end{aligned} \tag{48}$$

where the first inequality uses $\langle a, b\rangle \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ for any $a, b \in \mathbb{R}^{\mathrm{d_x}}$, thus $\langle \mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t), \mathbb{E}[g_t|x_t]\rangle \le \frac{1}{2}\|\,\mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t)\|^2 + \frac{1}{2}\|\,\mathbb{E}[g_t|x_t]\|^2$, and the second inequality uses $\|\,\mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\| \le C_p\mu$ by Lem. 8.

Further, we recall by Lem. 9 that $\mathbb{E}[\|g_t\|^2 \mid x_t] \le k_2\,\mathrm{d_x}\,L_f^2 L_y^2 + L_f^2(1 + 2L_y) =: \sigma_{\mathrm{d_x}}^2$, where $k_2 > 0$ is a constant. Substituting into (47) yields:

$$\mathbb{E}[F_\mu(x_{t+1}) \mid x_t] \le F_\mu(x_t) - \frac{\alpha}{2}\|\,\mathbb{E}[g_t \mid x_t]\|^2 + \frac{\alpha}{2}C_p^2\mu^2 + \frac{\alpha^2 L_F}{2}\sigma_{\mathrm{d_x}}^2.$$

Taking expectations, where $\mathbb{E}[\mathbb{E}[F_\mu(x_{t+1}) \mid x_t]] = \mathbb{E}[F_\mu(x_{t+1})]$ by the law of total expectation, and rearranging gives for each $t = 0, 1, \ldots, T-1$:

$$\frac{\alpha}{2}\,\mathbb{E}[\|\,\mathbb{E}[g_t \mid x_t]\|^2] \le \mathbb{E}[F_\mu(x_t)] - \mathbb{E}[F_\mu(x_{t+1})] + \frac{\alpha}{2}C_p^2\mu^2 + \frac{\alpha^2 L_F}{2}\sigma_{\mathrm{d_x}}^2.$$

Then, summing over $t = 0, \ldots, T-1$, telescoping $\sum_{t=0}^{T-1}(\mathbb{E}[F_\mu(x_t)] - \mathbb{E}[F_\mu(x_{t+1})])$ to $F_\mu(x_0) - \mathbb{E}[F_\mu(x_T)]$ (as x_0 is deterministic) and dividing by $\alpha T/2$ yields:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\,\mathbb{E}[g_t \mid x_t]\|^2\big] \le \frac{2\big(F_\mu(x_0) - \mathbb{E}[F_\mu(x_T)]\big)}{\alpha T} + C_p^2\mu^2 + \alpha L_F\sigma_{\mathrm{d_x}}^2. \tag{49}$$

We now bound $F_\mu(x_0) - \mathbb{E}[F_\mu(x_T)]$:

$$\begin{aligned}
F_\mu(x_0) - \mathbb{E}[F_\mu(x_T)] &\le F(x_0) + L_f L_y\mu - \mathbb{E}[F_\mu(x_T)] \\
&\le F(x_0) + L_f L_y\mu - \inf_{x\in\mathbb{R}^{\mathrm{d_x}}} F_\mu(x) \\
&\le F(x_0) + L_f L_y\mu - \inf_{x\in\mathbb{R}^{\mathrm{d_x}}} F(x) + L_f L_y\mu \\
&\le \Delta + 2L_f L_y\mu =: \Delta_\mu,
\end{aligned}$$

where the first inequality follows as $|F_\mu(x) - F(x)| \le L_f L_y\mu$ for all $x \in \mathbb{R}^{\mathrm{d_x}}$ by Prop. 3, the second inequality follows as $\mathbb{E}[F_\mu(x_T)] \ge \inf_x F_\mu(x)$ by definition of the infimum of F_μ, the third inequality follows as $F_\mu(x) \ge F(x) - L_f L_y\mu \ge \inf_{x\in\mathbb{R}^{\mathrm{d_x}}} F(x) - L_f L_y\mu$ for any $x \in \mathbb{R}^{\mathrm{d_x}}$ by Prop. 3, and the last inequality follows as by Assumption 1, $f(x_0, y^*(x_0)) - \inf_{x\in\mathbb{R}^{\mathrm{d_x}}} f(x, y^*(x)) \le \Delta$, so $F(x_0) - \inf_{x\in\mathbb{R}^{\mathrm{d_x}}} F(x) \le \Delta$.

Substituting into (49):

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\,\mathbb{E}[g_t \mid x_t]\|^2\big] \le \frac{2\,\Delta_\mu}{\alpha T} + C_p^2\mu^2 + \alpha L_F\sigma_{\mathrm{d_x}}^2. \tag{50}$$

Choice of α. We choose the step size to balance the first and third terms of the bound in (50). Setting the first and third term equal gives

$$\alpha = \sqrt{\frac{2\,\Delta_\mu}{T L_F \sigma_{\mathrm{d_x}}^2}},$$

and we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\,\mathbb{E}[g_t \mid x_t]\|^2\big] \le 2\sqrt{\frac{2\,\Delta_\mu L_F \sigma_{\mathrm{d_x}}^2}{T}} + C_p^2\mu^2. \tag{51}$$

Containment of $\mathbb{E}[g_t \mid x_t]$ in the partial Goldstein subdifferential. By Lems. 7 and 11, $\mathbb{E}[H_t \mid x_t] = Jy_\mu(x_t) \in \partial_\mu^{\mathrm{G}}\, y^*(x_t) \subseteq \partial_\delta^{\mathrm{G}}\, y^*(x_t)$, where the inclusion holds since $\mathbb{B}(x_t,\mu) \subseteq \mathbb{B}(x_t,\delta)$ for $\mu \le \delta$. We recall by (18) that the partial Goldstein subdifferential is given by

$$\partial_\delta^{\mathrm{P}} F(x) = \left\{\nabla_x f(x, y^*(x)) + M^\top \nabla_y f(x, y^*(x)) : M \in \partial_\delta^{\mathrm{G}}\, y^*(x)\right\}.$$

As $\mu = \min(\delta, \varepsilon/(\sqrt{2}C_p)) \le \delta$, thus $Jy_\mu(x_t) \in \partial_\delta^{\mathrm{G}}\, y^*(x_t)$ by Lem. 11, $\mathbb{E}[g_t \mid x_t] = \nabla_x f(x_t, y^*(x_t)) + Jy_\mu(x_t)^\top \nabla_y f(x_t, y^*(x_t))$ is an element of $\partial_\delta^{\mathrm{P}} F(x_t)$, i.e., $\mathbb{E}[g_t \mid x_t] \in \partial_\delta^{\mathrm{P}} F(x_t)$. As an immediate consequence:

$$\min\{\|g\| : g \in \partial_\delta^{\mathrm{P}} F(x_t)\} \le \|\,\mathbb{E}[g_t \mid x_t]\| \text{ for all } t = 0, 1, \ldots, T-1. \tag{52}$$

Output guarantee. x^R is chosen uniformly at random from iterates $\{x_0, \ldots, x_{T-1}\}$, and $R \in \{0, \ldots, T-1\}$ is its corresponding time index. Let $\bar{g}_R$ denote $\mathbb{E}[g_t \mid x_t]$ evaluated at $t = R$. Then $\mathbb{E}\big[\|\bar{g}_R\|^2\big] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\,\mathbb{E}[g_t \mid x_t]\|^2\big]$. By (52) at $t = R$ and applying Jensen's inequality $\mathbb{E}[\sqrt{X}] \le \sqrt{\mathbb{E}[X]}$ as $\sqrt{\cdot}$ is concave:

$$\mathbb{E}\big[\min\big\{\|g\| : g \in \partial_\delta^{\mathrm{P}} F(x^R)\big\}\big] \le \mathbb{E}\big[\|\bar{g}_R\|\big] \le \sqrt{\mathbb{E}[\|\bar{g}_R\|^2]} = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\,\mathbb{E}[g_t \mid x_t]\|^2\big]}. \tag{53}$$

Substituting (51) into (53):

$$\mathbb{E}\big[\min\big\{\|g\| : g \in \partial_\delta^{\mathrm{P}} F(x^R)\big\}\big] \le \left(2\sqrt{\frac{2\,\Delta_\mu\, L_F\, \sigma_{\mathrm{d_x}}^2}{T}} + C_p^2\mu^2\right)^{1/2}. \tag{54}$$

For x^R to be a (δ,ε)-partial Goldstein stationary point, the right-hand side of (54) must be no higher than ε. Squaring both sides of the inequality, $2\sqrt{\frac{2\,\Delta_\mu\, L_F\, \sigma_{\mathrm{d_x}}^2}{T}} + C_p^2\mu^2 \le \varepsilon^2$, and we allocate $\varepsilon^2/2$ to each term of the left-hand side.

Choice of μ. The constraint $C_p^2\mu^2 \le \varepsilon^2/2$ requires $\mu \le \varepsilon/(\sqrt{2}\,C_p)$. Further, we require $\mu \le \delta$, to guarantee that $Jy_\mu(x_t) \in \partial_\delta^{\mathrm{G}}\, y^*(x_t)$ in the containment above. Hence we set $\mu = \min\Big(\delta,\ \frac{\varepsilon}{\sqrt{2}\,C_p}\Big)$.

Choice of T. It remains to ensure $2\sqrt{2\,\Delta_\mu\, L_F\, \sigma_{\mathrm{d_x}}^2/T} \le \varepsilon^2/2$, i.e., $T \ge \dfrac{32\,\Delta_\mu\, L_F\, \sigma_{\mathrm{d_x}}^2}{\varepsilon^4}$.

Since $\mathrm{d_x} \ge 1$, we have $\sigma_{\mathrm{d_x}}^2 = \mathcal{O}(\mathrm{d_x})$, $L_F = \mathcal{O}(1 + \sqrt{\mathrm{d_x}}/\mu)$, and $\Delta_\mu = \mathcal{O}(1+\mu)$, hence

$$\frac{32\,\Delta_\mu\, L_F\, \sigma_{\mathrm{d_x}}^2}{\varepsilon^4} = \mathcal{O}\left(\frac{(1+\mu)(1+\sqrt{\mathrm{d_x}}/\mu)\ \mathrm{d_x}}{\varepsilon^4}\right) = \mathcal{O}\left(\frac{\mathrm{d_x} + {\mathrm{d_x}}^{3/2}/\mu + \mu\,\mathrm{d_x} + {\mathrm{d_x}}^{3/2}}{\varepsilon^4}\right).$$

Under the assumption $\mu \le 1$, all four terms are absorbed by ${\mathrm{d_x}}^{3/2}/\mu$, i.e., $T = \mathcal{O}\left(\dfrac{{\mathrm{d_x}}^{3/2}}{\mu\,\varepsilon^4}\right)$. □

Theorem 5 (Convergence of PZOS to (δ,ε)-Goldstein stationary point)**.** *Let x^R be chosen uniformly at random from the iterates $\{x_0, \ldots, x_{T-1}\}$ generated by PZOS. Under Assumptions 1 and 2, with step size $\alpha = \sqrt{2(\Delta + 2L_f L_y \mu)/(T L_F \sigma_{\mathrm{d_x}}^2)}$, smoothing parameter $\mu = \min(\delta, \varepsilon/(2C_f)) < 1$, and $C_p = L_g(1+L_y)L_y$ and $C_f = (1+L_y)L_g(1+2L_y)$, x^R is, in expectation, a (δ,ε)-Goldstein stationary point of F after $T = \mathcal{O}\Big({\mathrm{d_x}}^{3/2}/(\mu\,\varepsilon^4)\Big)$ iterations, i.e.,*

$$\mathbb{E}\left[\min\left\{\|g\| : g \in \partial_\delta^G\, F(x^R)\right\}\right] \le \varepsilon.$$

Proof. We follow the proof of Thm. 4 and only state the changes to therein. As in the proof of Thm. 4, we apply the descent lemma to F_μ. However, as $\mathbb{E}[g_t \mid x_t]$ need not lie in $\partial_\delta^{\mathrm{G}} F(x_t)$, we bound $\|\nabla F_\mu(x_t)\|$ instead of $\|\,\mathbb{E}[g_t \mid x_t]\|$. To this end, we start at (47) and replace (48) by

$$\begin{aligned}\langle \nabla F_\mu(x_t), \mathbb{E}[g_t \mid x_t]\rangle &= \langle \nabla F_\mu(x_t), \nabla F_\mu(x_t)\rangle + \langle \nabla F_\mu(x_t)), \mathbb{E}[g_t|x_t] - \nabla F_\mu(x_t)\rangle \\ &= \|\nabla F_\mu(x_t)\|^2 + \langle \nabla F_\mu(x_t), \mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\rangle \\ &\geq \frac{1}{2}\|\nabla F_\mu(x_t)\|^2 - \frac{1}{2}\|\,\mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\|^2 \\ &\geq \frac{1}{2}\|\nabla F_\mu(x_t)\|^2 - \frac{1}{2}C_p^2\mu^2,\end{aligned}$$

where the first inequality uses $\langle a, b\rangle \geq -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$ for any $a, b \in \mathbb{R}^{\mathrm{d_x}}$ and the second inequality uses $\|\,\mathbb{E}[g_t \mid x_t] - \nabla F_\mu(x_t)\| \leq C_p\mu$ by Lem. 8.

The remaining steps up to (51) are unchanged, with $\mathbb{E}[g_t \mid x_t]$ replaced by $\nabla F_\mu(x_t)$, giving step size $\alpha = \sqrt{\frac{2\,\Delta_\mu}{T L_F \sigma_{\mathrm{d_x}}^2}}$ and

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F_\mu(x_t)\|^2\big] \leq 2\sqrt{\frac{2\,\Delta_\mu L_F \sigma_{\mathrm{d_x}}^2}{T}} + C_p^2\mu^2. \tag{55}$$

As we work with the Goldstein subdifferential and $\nabla F_\mu(x_t)$ instead of the partial Goldstein subdifferential and $\mathbb{E}[g_t \mid x_t]$, we use the following distance bound of ∇F_μ to the Goldstein subdifferential instead of the containment of $\mathbb{E}[g_t \mid x_t]$ in the partial Goldstein subdifferential:

Distance of $\nabla F_\mu(x_t)$ to Goldstein subdifferential. By Lem. 12, for each x_t there exists $\bar{q}_t \in \partial_\mu^{\mathrm{G}} F(x_t) \subseteq \partial_\delta^{\mathrm{G}} F(x_t)$ (since $\mu \leq \delta$) with $\|\nabla F_\mu(x_t) - \bar{q}_t\| \leq C_f\mu$. By the triangle inequality:

$$\begin{aligned}\min\big\{\|g\| : g \in \partial_\delta^{\mathrm{G}} F(x_t)\big\} \leq \|\bar{q}_t\| &= \|\nabla F_\mu(x_t) + \bar{q}_t - \nabla F_\mu(x_t)\| \\ &\leq \|\nabla F_\mu(x_t)\| + \|\bar{q}_t - \nabla F_\mu(x_t)\| \leq \|\nabla F_\mu(x_t)\| + C_f\mu.\end{aligned} \tag{56}$$

Output guarantee. x^R is chosen uniformly at random from iterates $\{x_0, \ldots, x_{T-1}\}$, and $R \in \{0, \ldots, T-1\}$ is it corresponding time index. Since (56) holds for every iterate, it holds in particular at $t = R$. Taking expectations gives

$$\mathbb{E}\big[\min\big\{\|g\| : g \in \partial_\delta^{\mathrm{G}} F(x^R)\big\}\big] \leq \mathbb{E}[\|\nabla F_\mu(x^R)\|] + C_f\mu. \tag{57}$$

Further, $\mathbb{E}[\|\nabla F_\mu(x^R)\|^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_\mu(x_t)\|^2]$. By Jensen's inequality as in (53), with $\nabla F_\mu(x^R)$ in place of $\bar{g}_R$, and by (55),

$$\mathbb{E}[\|\nabla F_\mu(x^R)\|] \leq \left(2\sqrt{\frac{2\,\Delta_\mu\, L_F\, \sigma_{\mathrm{d_x}}^2}{T}} + C_p^2\mu^2\right)^{1/2}. \tag{58}$$

Combining (57) and (58):

$$\mathbb{E}\big[\min\big\{\|g\| : g \in \partial_\delta^{\mathrm{G}} F(x^R)\big\}\big] \leq \left(2\sqrt{\frac{2\,\Delta_\mu)\, L_F\, \sigma_{\mathrm{d_x}}^2}{T}} + C_p^2\mu^2\right)^{1/2} + C_f\mu. \tag{59}$$

For x^R to be a (δ, ε)-Goldstein stationary point, the right-hand side of (59) must be no higher than ε. We allocate $\varepsilon/2$ to each summand.

Choice of μ. The constraint $C_f\mu \leq \varepsilon/2$ requires $\mu \leq \varepsilon/(2C_f)$. We also need $\mu \leq \delta$ for the Goldstein subdifferential containment $\partial_\mu^{\mathrm{G}} F(x) \subseteq \partial_\delta^{\mathrm{G}} F(x)$. Therefore we set $\mu = \min\left(\delta,\ \frac{\varepsilon}{2C_f}\right)$.

With this choice, the term $C_p^2\mu^2$ satisfies $C_p^2\mu^2 \leq C_p^2\varepsilon^2/(4C_f^2) < \varepsilon^2/16$, as

$$\frac{C_f}{C_p} = \frac{(1+L_y)L_g(1+2L_y)}{L_g(1+L_y)L_y} = \frac{1+2L_y}{L_y} > 2,$$

so $C_p < C_f/2$ strictly. If we can ensure that $2\sqrt{\frac{2\,\Delta_\mu\,L_F\,\sigma_{\mathrm{d_x}}^2}{T}} \le \frac{\varepsilon^2}{8}$, it follows that the two terms in the square root of (59) contribute at most $\sqrt{\varepsilon^2/16 + \varepsilon^2/8} < \varepsilon/2$.

Choice of T. It remains to ensure

$$2\sqrt{\frac{2\,\Delta_\mu\,L_F\,\sigma_{\mathrm{d_x}}^2}{T}} \le \frac{\varepsilon^2}{8}, \qquad \text{i.e.,} \qquad T \ge \frac{512\,\Delta_\mu\,L_F\,\sigma_{\mathrm{d_x}}^2}{\varepsilon^4}.$$

As in Thm. 4, $\sigma_{\mathrm{d_x}}^2 = \mathcal{O}(\mathrm{d_x})$, $L_F = \mathcal{O}(1 + \sqrt{\mathrm{d_x}}/\mu)$, and $\Delta_\mu = \mathcal{O}(1+\mu)$, hence

$$\frac{512\,\Delta_\mu\,L_F\,\sigma_{\mathrm{d_x}}^2}{\varepsilon^4} = \mathcal{O}\left(\frac{(1+\mu)(1+\sqrt{\mathrm{d_x}}/\mu)\;\mathrm{d_x}}{\varepsilon^4}\right) = \mathcal{O}\left(\frac{\mathrm{d_x} + {\mathrm{d_x}}^{3/2}/\mu + \mu\,\mathrm{d_x} + {\mathrm{d_x}}^{3/2}}{\varepsilon^4}\right).$$

Under the assumption $\mu \le 1$, all four terms are absorbed by ${\mathrm{d_x}}^{3/2}/\mu$, i.e., $T = \mathcal{O}\left(\frac{{\mathrm{d_x}}^{3/2}}{\mu\,\varepsilon^4}\right)$. □

C.3.3 Proof of Proposition 1

We prove that the black-box method ZOS also converges to (δ,ε)-partial Goldstein stationary points, with the same rate as PZOS but larger constant.

Proposition 4 (Restatement of Proposition 1, Convergence of ZOS to (δ,ε)-partial Goldstein stationary point)**.** *Let x^R be chosen uniformly at random from the iterates $\{x_0,\dots,x_{T-1}\}$ generated by ZOS, i.e., $x_{t+1} = x_t - \alpha\widetilde{g}_t$ with $\widetilde{g}_t$ as in* (2)*. Under Assumptions 1 and 2, with step size $\alpha = \sqrt{2(\Delta + L_f(1+L_y)\mu)/(TL_{\widetilde{F}}\widetilde{\sigma}_{\mathrm{d_x}}^2)}$ and smoothing parameter $\mu = \min(\delta,\varepsilon/(2\widetilde{C}_p)) < 1$, where $\widetilde{C}_p = L_g(1+L_y)^2$, x^R is, in expectation, a (δ,ε)-partial Goldstein stationary point of F after $T = \mathcal{O}\left({\mathrm{d_x}}^{3/2}/(\mu\,\varepsilon^4)\right)$ iterations, i.e.,*

$$\mathbb{E}\left[\min\left\{\|g\| : g \in \partial_\delta^P F(x^R)\right\}\right] \le \varepsilon.$$

Proof. The proof follows the structure of the proof of Thm. 5, with the surrogate $\widetilde{F}_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[F(x+\mu u)]$ in place of F_μ, the estimator $\widetilde{g}_t$ in place of g_t, Lem. 13 in place of Lem. 12, $\widetilde{\sigma}_{\mathrm{d_x}}^2$ in place of $\sigma_{\mathrm{d_x}}^2$, and $\widetilde{\Delta}_\mu$ in place of Δ_μ.

Since F is $L_f(1+L_y)$-Lipschitz by (40), by Thm. 3 $\widetilde{F}_\mu$ is $L_{\widetilde{F}}$-smooth with $L_{\widetilde{F}} = k_1\sqrt{\mathrm{d_x}}L_f(1+L_y)/\mu$ for a constant $k_1 > 0$. In addition, $|\widetilde{F}_\mu(x) - F(x)| \le L_f(1+L_y)\mu$ for all $x \in \mathbb{R}^{\mathrm{d_x}}$ by [18, Prop. 2.3]. Further, $\widetilde{g}_t$ is an unbiased estimator of $\nabla\widetilde{F}_\mu(x_t)$ by [24, Lem. 10], i.e., $\mathbb{E}[\widetilde{g}_t \mid x_t] = \nabla\widetilde{F}_\mu(x_t)$, and by Lem. 1,

$$\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \le k_2\,\mathrm{d_x}\,L_f^2L_y^2 + k_2\,\mathrm{d_x}\,L_f^2(1+2L_y) =: \widetilde{\sigma}_{\mathrm{d_x}}^2.$$

Applying the descent lemma to $\widetilde{F}_\mu$ and with update step $x_{t+1} = x_t - \alpha\widetilde{g}_t$ when utilizing ZOS,

$$\widetilde{F}_\mu(x_{t+1}) \le \widetilde{F}_\mu(x_t) - \alpha\langle\nabla\widetilde{F}_\mu(x_t),\widetilde{g}_t\rangle + \frac{\alpha^2L_{\widetilde{F}}}{2}\|\widetilde{g}_t\|^2.$$

Taking the conditional expectation given x_t:

$$\begin{aligned}\mathbb{E}\left[\widetilde{F}_\mu(x_{t+1}) \mid x_t\right] &\le \widetilde{F}_\mu(x_t) - \alpha\langle\nabla\widetilde{F}_\mu(x_t),\mathbb{E}[\widetilde{g}_t \mid x_t]\rangle + \frac{\alpha^2L_{\widetilde{F}}}{2}\,\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \\ &\le \widetilde{F}_\mu(x_t) - \alpha\|\nabla\widetilde{F}_\mu(x_t)\|^2 + \frac{\alpha^2L_{\widetilde{F}}}{2}\widetilde{\sigma}_{\mathrm{d_x}}^2,\end{aligned}$$

where the second inequality uses that $\widetilde{g}_t$ is an unbiased estimator of $\nabla\widetilde{F}_\mu(x_t)$ and that $\mathbb{E}[\|\widetilde{g}_t\|^2 \mid x_t] \le \widetilde{\sigma}_{\mathrm{d_x}}^2$. In contrast to Thms. 4 and 5 no bias term arises in this equation as the inner product equals $\|\nabla\widetilde{F}_\mu(x_t)\|^2$ exactly as $\mathbb{E}[\widetilde{g}_t \mid x_t] = \nabla\widetilde{F}_\mu(x_t)$. Taking expectations, summing over $t =$

$0, 1, \ldots, T-1$, telescoping $\sum_{t=0}^{T-1}(\mathbb{E}[\widetilde{F}_\mu(x_t) - \mathbb{E}[\widetilde{F}_\mu(x_{t+1})])$ to $\widetilde{F}_\mu(x_0) - \mathbb{E}[\widetilde{F}_\mu(x_T)]$ and dividing by αT gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|^2] \le \frac{\widetilde{F}_\mu(x_0) - \mathbb{E}[\widetilde{F}_\mu(x_T)]}{\alpha T} + \frac{\alpha L_{\widetilde{F}}}{2}\widetilde{\sigma}^2_{\mathrm{d_x}}. \tag{60}$$

We now bound the term $\widetilde{F}_\mu(x_0) - \mathbb{E}[\widetilde{F}_\mu(x_T)]$. As noted above, $|\widetilde{F}_\mu(x) - F(x)| \le L_f(1+L_y)\mu$ for any $x \in \mathbb{R}^{\mathrm{d_x}}$, hence $\widetilde{F}_\mu(x_0) \le F(x_0) + L_f(1+L_y)\mu$. Further, since $\widetilde{F}_\mu(x) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{B}^{\mathrm{d_x}})}[F(x+\mu u)]$, $\widetilde{F}_\mu(x) \ge \inf_{x'\in\mathbb{R}^{\mathrm{d_x}}} F(x')$ for all $x \in \mathbb{R}^{\mathrm{d_x}}$. Together with Assumption 1, this gives

$$\widetilde{F}_\mu(x_0) - \mathbb{E}[\widetilde{F}_\mu(x_T)] \le F(x_0) + L_f(1+L_y)\mu - \inf_{x\in\mathbb{R}^{\mathrm{d_x}}} F(x) \le \Delta + L_f(1+L_y)\mu =: \widetilde{\Delta}_\mu.$$

Substituting into (60) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|^2] \le \frac{\widetilde{\Delta}_\mu}{\alpha T} + \frac{\alpha L_{\widetilde{F}}}{2}\widetilde{\sigma}^2_{\mathrm{d_x}}. \tag{61}$$

Choice of α. Setting the two terms in (61) equal gives $\alpha = \sqrt{\dfrac{2\widetilde{\Delta}_\mu}{T L_{\widetilde{F}}\widetilde{\sigma}^2_{\mathrm{d_x}}}}$. With this step size, both terms each equal $\sqrt{\widetilde{\Delta}_\mu L_{\widetilde{F}}\widetilde{\sigma}^2_{\mathrm{d_x}}/(2T)}$, and hence

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla\widetilde{F}_\mu(x_t)\|^2\big] \le \sqrt{\frac{2\widetilde{\Delta}_\mu L_{\widetilde{F}}\widetilde{\sigma}^2_{\mathrm{d_x}}}{T}}. \tag{62}$$

Distance of $\nabla\widetilde{F}_\mu$ to partial Goldstein subdifferential. By Lem. 13, for each x_t there exists $\bar{p}_t \in \partial^{\mathrm{P}}_\mu F(x_t) \subseteq \partial^{\mathrm{P}}_\delta F(x_t)$ (since $\mu \le \delta$) with $\|\nabla\widetilde{F}_\mu(x_t) - \bar{p}_t\| \le \widetilde{C}_p\mu$, and $\widetilde{C}_p = L_g(1+L_y)^2$. By the triangle inequality,

$$\min\big\{\|g\| : g \in \partial^{\mathrm{P}}_\delta F(x_t)\big\} \le \|\bar{p}_t\| \le \|\nabla\widetilde{F}_\mu(x_t)\| + \widetilde{C}_p\mu. \tag{63}$$

Output guarantee. Since x^R is chosen uniformly at random over the iterates $\{x_0, \ldots, x_{T-1}\}$, $\mathbb{E}[\|\nabla\widetilde{F}_\mu(x^R)\|^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|^2]$. From Jensen's inequality for concave functions and the QM-AM inequality, it follows that

$$\begin{aligned}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x^R)\|] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{\|\nabla\widetilde{F}_\mu(x_t)\|^2}\right] \\ &\le \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|^2]} \le \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|^2]},\end{aligned}$$

Together with (62), this gives

$$\mathbb{E}\big[\min\big\{\|g\| : g \in \partial^{\mathrm{P}}_\delta F(x^R)\big\}\big] \le \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\widetilde{F}_\mu(x_t)\|^2]} + \widetilde{C}_p\mu \le \left(\frac{2\widetilde{\Delta}_\mu L_{\widetilde{F}}\widetilde{\sigma}^2_{\mathrm{d_x}}}{T}\right)^{1/4} + \widetilde{C}_p\mu. \tag{64}$$

For x^R to be a (δ,ε)-partial Goldstein stationary point, the right-hand side of (64) must be no higher than ε. We allocate $\varepsilon/2$ to each term.

Choice of μ. The constraint $\widetilde{C}_p\mu \le \varepsilon/2$ requires $\mu \le \varepsilon/(2\widetilde{C}_p)$ and the containment $\partial^{\mathrm{P}}_\mu F(x_t) \subseteq \partial^{\mathrm{P}}_\delta F(x_t)$ requires $\mu \le \delta$. Hence we set $\mu = \min\Big(\delta,\ \varepsilon/(2\widetilde{C}_p)\Big)$.

Choice of T. It remains to ensure

$$\left(\frac{2\widetilde{\Delta}_\mu L_{\widetilde{F}}\widetilde{\sigma}^2_{\mathrm{d_x}}}{T}\right)^{1/4} \le \varepsilon/2 \qquad \text{i.e.,} \qquad T \ge \frac{32\,\widetilde{\Delta}_\mu L_{\widetilde{F}}\widetilde{\sigma}^2_{\mathrm{d_x}}}{\varepsilon^4}.$$

For $\mu < 1$, $\widetilde{\sigma}_{\mathrm{d_x}}^2 = k_2\,\mathrm{d_x}\,L_f^2 L_y^2 + k_2\,\mathrm{d_x}\,L_f^2(1+2L_y) = \mathcal{O}(\mathrm{d_x})$, $L_{\widetilde{F}} = k_1\sqrt{\mathrm{d_x}}L_f(1+L_y)/\mu = \mathcal{O}(\sqrt{\mathrm{d_x}}/\mu)$, and $\widetilde{\Delta}_\mu = \Delta + L_f(1+L_y)\mu = \mathcal{O}(1)$, hence

$$\frac{32\,\widetilde{\Delta}_\mu L_{\widetilde{F}}\widetilde{\sigma}_{\mathrm{d_x}}^2}{\varepsilon^4} = \mathcal{O}\left(\frac{(1+\mu)(\sqrt{\mathrm{d_x}}/\mu)\;\mathrm{d_x}}{\varepsilon^4}\right)$$

and $T = \mathcal{O}\left(\frac{{\mathrm{d_x}}^{3/2}}{\mu\,\varepsilon^4}\right)$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract states the conceptual novelty of the paper, the theoretical convergence results, and empirical validation of improved performance using partial information.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the paragraph 'Limitations'. Assumptions are clearly stated, and justification for these assumptions being standard is given. The contributions are compared to the state of the art, and how the proposed algorithm scales with problem dimension is explicitly discussed in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The convergence results are thoroughly proven in the appendix. A proof sketch is provided in the main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducibility is provided via fixed random seeds, and code to reproduce the experimental results is provided as supplementary material. The used algorithms are clearly described in pseudocode in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code to reproduce the experimental results is provided, and fixed random seeds are used and provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are described in the main paper, and the used hyperparameters are stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The plots show median, interquartile range, and 10th and 90th percentiles across all runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper describes the computer resources used for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the Code of ethics due to its theoretical nature. Experiments do not require human subjects, all data is generated in code.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is foundational research. We are not aware of negative applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

Justification: The paper does not use existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce new assets, although an implementation of the proposed algorithm is provided.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not involved in the core methodological development. Claude Code was used for code assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.