Achieving Directional-Stationarity from a Single Random Direction Step

Dan Greenstein **Nadav Hallak**
Faculty of Data and Decision Sciences, Technion
{sdngreen, ndvhllk}@technion.ac.il

Abstract

This paper addresses the challenge of obtaining strong optimality guarantees in constrained nonsmooth nonconvex optimization under mild regularity conditions, namely local Lipschitz continuity and existence of directional derivatives. While standard methods typically ensure weak stationarity notions, achieving directional (d-)stationarity remains nontrivial. We show that a random direction exploration step is sufficient to attain d-stationarity. The proposed approach augments any feasible base method with one additional randomized exploration step per iteration, which samples a direction and step size and accepts the candidate based on a function value comparison. The resulting scheme guarantees that all accumulation points are d-stationary almost surely, independently of the behavior of the underlying method. Moreover, under conditions on the base method, it preserves convergence rates of the base method, as established for the difference-of-convex algorithm (DCA), prox-linear-type schemes, and interpolated normalized gradient descent (INGD). The theoretical results are complemented by numerical experiments illustrating the effect and guarantees of the exploration step.

1 Introduction

1.1 Problem formulation

We consider the constrained nonsmooth nonconvex optimization problem

$$\min_{x \in C} h(x), \tag{1}$$

where $h\colon \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz and directionally differentiable, and $C \subseteq \mathbb{R}^n$ is a nonempty closed convex set. This setting covers a broad class of problems, including smooth-minus-convex objectives $h = f - g$ with $f \in C^1$ and g convex, as well as the difference-of-convex (DC) class where both f and g are convex. Importantly, our approach does not require access to such a decomposition and applies directly to the objective h.

A central challenge in nonsmooth nonconvex optimization is to obtain strong and meaningful optimality guarantees. In DC programming and related settings, several stationarity notions are available, with varying levels of strength (see, e.g., [Pang et al., 2017]). Among these, *directional (d-)stationarity* characterizes points with no feasible descent directions, defined through directional derivatives (Definition 1.3 and Assumption 2).

Directional stationarity is stronger than the criticality notion standard in DC algorithms [de Oliveira, 2020] and than Clarke stationarity [Clarke, 1990]; derivative-free schemes such as MADS [Audet and Dennis Jr., 2006] typically target the latter. See Remark 1.1 for limiting (Mordukhovich) stationarity and for a fuller discussion. It is rarely guaranteed in general constrained nonsmooth nonconvex settings. Existing methods achieving d-stationarity either rely on explicit DC decompositions and

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

access to smooth component functions [Pang et al., 2017, Feng and Yuan, 2026], or are restricted to linear feasibility sets [Beck and Hallak, 2020, 2022], and their guarantees are tightly coupled to the algorithmic dynamics.

In contrast, we propose a minimal exploration step that is independent of the base algorithm and enforces d-stationarity of accumulation points. Hence, any feasible method can be made asymptotically d-stationary without modifying the oracle map itself; the augmentation modifies only the accepted iterate.

Main contributions.

- **Exploration-augmented optimization.** We introduce a random exploration mechanism that augments any feasible optimization method with one additional randomized exploration step per iteration (Algorithms 1 and 2, Section 3).
- **D-stationarity via random exploration.** We prove that, almost surely, all accumulation points are d-stationary, independently of the underlying algorithm (Theorem 4.1, Section 4). The result does not use any structural properties of the underlying method.
- **Rate preservation and oracle compatibility.** We give sufficient conditions under which the augmentation preserves the rate guarantee of the base method (Theorems 5.3 and 5.4, Section 5.2; for their scope, Appendix C.1), and verify them for the difference-of-convex algorithm (DCA, Theorem 5.1), prox-linear-type schemes (Theorem 5.5), and gradient sampling (Theorem 5.6).

1.2 Mathematical preliminaries

We formulate two assumptions on (1) and one supporting lemma that ensure well-defined directional derivatives and the continuity required for the probabilistic argument. Throughout, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$ and S^{n-1} the Euclidean unit sphere; whether the analysis of Section 4 extends to other norms is left open.

Assumption 1 (A1). *h is locally Lipschitz on $\mathbb{R}^n$ and bounded below on C.*

Assumption 2 (A2). *The directional derivative $h'(x;\, d) = \lim_{t\downarrow 0}(h(x+td) - h(x))/t$ exists for every $x \in C$ and $d \in \mathbb{R}^n$.*

Lemma 1.1 (Proof in Appendix A). *Let $x \in C$, and let L be a Lipschitz constant of h on a neighborhood of x. Then $|h'(x;\, d_1) - h'(x;\, d_2)| \le L\,\|d_1 - d_2\|$ for all $d_1, d_2 \in \mathbb{R}^n$.*

These properties hold for a broad class of objectives such as $h = f - g$ with $f \in C^1$ and g convex and finite-valued, as well as in the DC case where both f and g are convex. In both settings, the directional derivative admits the representation $h'(x;\, d) = f'(x;\, d) - g'(x;\, d)$, and continuity in d follows from continuity/convexity respectively [Rockafellar, 1997, Theorems 23.1 and 10.1].

Definition 1.1 (Affine hull subspace). *Given $x^0 \in C$, the linear subspace parallel to $\mathrm{aff}(C)$ is defined by $L := \mathrm{aff}(C - x^0)$; it does not depend on the choice of $x^0 \in C$.*

For $x \in C$ and $v \in \mathbb{R}^n$, denote the maximal feasible step length at x in direction v by

$$\tau(x;\, v) := \sup\bigl\{t \ge 0 : x + tv \in C\bigr\}.$$

Definition 1.2 (Cone of feasible directions). *Let $C \subseteq \mathbb{R}^n$ be convex and $\bar{x} \in C$. The* cone of feasible directions *is $D_C(\bar{x}) := \{w \mid \exists\, \lambda > 0,\ \bar{x} + \lambda w \in C\}$; it is a convex cone.*

Definition 1.3 (Directional (d-)stationarity). *A point $\bar{x} \in C$ is d-stationary for* (1) *if*

$$h'(\bar{x};\, d) \ge 0 \qquad \forall\, d \in D_C(\bar{x}).$$

By Lemma 1.1, the same condition is equivalent to $h'(\bar{x};\, d) \ge 0$ for every $d \in \mathrm{cl}(D_C(\bar{x}))$; see Lemma 4.4.

Remark 1.1 (Comparison with other stationarity notions). *For locally Lipschitz h and closed convex C,* Clarke stationarity *at $\bar{x} \in C$ is $0 \in \partial^C h(\bar{x}) + N_C(\bar{x})$ [Clarke, 1990]. Writing h° for the Clarke directional derivative, one has $h'(\bar{x}; d) \le h^\circ(\bar{x}; d)$ for all d, with equality if h is Clarke regular, hence Clarke stationarity does not in general imply d-stationarity. For convex C, d-stationarity is equivalent to $h'(\bar{x}; d) \ge 0$ for all $d \in T_C(\bar{x})$.* Mordukhovich (limiting) stationarity *($0 \in \partial^L h(\bar{x}) + N_C(\bar{x})$ [Rockafellar and Wets, 2009]) implies Clarke stationarity but remains weaker than d-stationarity in general. For DC objectives $h = f - g$ with f, g convex,* criticality *is again weaker than d-stationarity [de Oliveira, 2020].*

2 Related Work

We summarize the closest strands of the literature in Table 1. To the best of our knowledge, no prior work achieves d-stationarity via random feasible-direction exploration under general convex constraints with oracle-independent guarantees.

DC programming. The standard optimality notion in DC algorithms is criticality, which is strictly weaker than d-stationarity [de Oliveira, 2020]. Pang et al. [2017] initiated a line of work establishing d-stationarity for structured DC problems in which the concave part admits a finite-max representation, by solving one convex subproblem per active component. For this finite-max structure, d-stationarity is equivalent to optimality in one convex program per active component, whereas in our setting the condition of Definition 1.3 ranges over a continuum of directions. This framework has been extended along several directions by Ahn et al. [2017], Tao and Li [2023], Sun and Wu [2024], and Feng and Yuan [2026], all of which rely critically on the finite-max structure of the DC decomposition, explicit access to the DC decomposition and its components, and first-order information, and are inherently algorithm-specific rather than oracle-agnostic. Pang and Tao [2018] extend this approach to multiblock objectives that need not be DC. Their guarantee is blockwise directional stationarity, and their augmented Lagrangian method meets the coupled linear constraints only in the limit.

A complementary line of work [Beck and Hallak, 2020, 2022] establishes d-stationarity for smooth-minus-convex objectives over polyhedral sets using feasible-direction methods, including randomized variants based on fixed finite spanning sets. In contrast, bundle methods [Joki et al., 2018, de Oliveira, 2021] typically guarantee only criticality or Clarke stationarity. Stochastic DC methods, for example the online sampling scheme of Ye et al. [2025], likewise target criticality.

Finally, while the convergence rates of DCA are well understood in the unconstrained setting [Abbaszadehpeivasti et al., 2024], these results concern criticality measures and do not address d-stationarity.

Direct search: GSS, pattern search, and MADS. Generating set search [Kolda et al., 2006] and pattern search [Lewis and Torczon, 2000] poll finitely many directions and yield KKT-type limits in the smooth case and Clarke stationary limits for Lipschitz objectives [Clarke, 1990, Corollary to Proposition 2.4.3]. MADS [Audet and Dennis Jr., 2006] randomizes to asymptotically dense polling (Clarke stationarity); although the direction coverage is conceptually related to our argument, MADS is a standalone mesh-based method with sufficient decrease, whereas we augment an arbitrary oracle and target d-stationarity under general convex constraints. The feasible-direction methods of Beck and Hallak [2020, 2022] reach d-stationarity only under smooth-minus-convex structure over polyhedral sets.

Randomized search outside direct search. Gratton et al. [2015] analyze probabilistic-descent direct search, yielding almost-sure convergence and $O(1/\sqrt{k})$ gradient-norm rates for smooth objectives; Gratton et al. [2019] treat linear constraints analogously. These schemes remain direct-search templates targeting gradient- or Clarke-type stationarity. Classical random search [Solis and Wets, 1981] and sufficient-descent schemes [Wardi, 1989] give value- or stationarity-type guarantees under coverage or internal descent, but are standalone algorithms. Our augmentation instead layers one random feasible-direction probe on an arbitrary oracle and targets d-stationarity of accumulation points under convex constraints.

3 The Random Exploration Procedure

The Random Exploration Procedure (REP) augments a base optimization method with a simple randomized exploration step. At the current iterate, REP samples a direction in the affine hull and a candidate step size, accepting the move only if it yields a feasible sufficient decrease in a regularized objective; it does not rely on gradients or problem structure and uses only function value comparisons. This regularization guarantees sufficient-decrease-type behavior.

Remark 3.1 (Sampling distributions)**.** *The direction distribution $\mathcal{D}_v$ in Algorithm 1 is supported on $S^{n-1} \cap L$ and assigns strictly positive probability to every measurable subset of $S^{n-1} \cap L$ with strictly positive surface measure. The step-size distribution $\mathcal{D}_t$ is supported on $[0, r]$ and assigns*

Table 1: Comparison of related fields.

Field / Work	Target	Mechanism	Nonsmooth DC?
DC methods [Pang et al., 2017]	D-stationarity	Convex subproblem	$g = \max_i \psi_i$, $\psi_i \in C^1$ only
GSS / GPS [Kolda et al., 2006]	Clarke stat.	Finite spanning set	No
MADS [Audet and Dennis Jr., 2006]	Clarke stat.	Dense random poll	No
Prob. direct search [Gratton et al., 2015]	Grad. stat.	Random sphere dir.	No
Random search [Solis and Wets, 1981, Wardi, 1989]	Value / stationarity	Coverage / sufficient descent	No
This work	D-stationarity	Random sphere dir.	Yes

Algorithm 1 Random Exploration Procedure (REP)

Require: $x \in C$; $\gamma > 0$, $r \in (0, \infty)$.

1: Sample $v_{ex} \sim \mathcal{D}_v$ with $\mathrm{supp}(\mathcal{D}_v) = S^{n-1} \cap L$

2: Sample $\hat{t} \sim \mathcal{D}_t$ with $\mathrm{supp}(\mathcal{D}_t) = [0, r]$

3: $t_{ex} \leftarrow \begin{cases} \hat{t} & \text{if } x + \hat{t} v_{ex} \in C \text{ and } h(x + \hat{t} v_{ex}) + \frac{\gamma}{2}(\hat{t})^2 < h(x), \\ 0 & \text{otherwise.} \end{cases}$

4: **Return** $y \leftarrow x + t_{ex} v_{ex}$

positive probability to every interval $[a, b] \subseteq (0, r]$. *Let* $\mathcal{F}_{k-1}$ *denote the* σ*-algebra generated by* $\{x^0, v_{ex}^0, \hat{t}^0, \ldots, x^{k-1}, v_{ex}^{k-1}, \hat{t}^{k-1}, x^k\}$. *We assume that the pairs* $(v_{\mathrm{ex}}^k, \hat{t}^k)$ *are identically distributed, and that* v_{ex}^k *and* $\hat{t}^k$ *are independent of each other and of* $\mathcal{F}_{k-1}$.

Algorithm 2 augments a feasible oracle with a randomized exploration step: the oracle proposes a candidate, while exploration samples a direction and step size to generate a competing feasible point accepted by value comparison. Almost surely, every neighborhood of relative-interior feasible direction at any accumulation point is sampled infinitely often, yielding d-stationarity via Borel–Cantelli.

Algorithm 2 Random exploration procedure implementation

Require: $x^0 \in C$; oracle $\mathcal{O}_C$ over C; $\gamma > 0$, $r \in (0, \infty)$.

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

2: Set $y^{k+1} \leftarrow \mathrm{REP}(x^k)$ and $z^{k+1} \leftarrow \mathcal{O}_C(x^k)$

3: $x^{k+1} \leftarrow \arg\min_{u \in \{y^{k+1},\, z^{k+1}\}} h(u)$

4: **end for**

Remark 3.2 (Implementation cost)**.** *Algorithm 2 requires one feasibility check and at most one function evaluation, which can be skipped when the candidate is infeasible, per exploration step.*

4 Convergence Analysis

This section establishes that the exploration step enforces asymptotic d-stationarity, independently of the oracle; some proofs are deferred to Appendix A.

Theorem 4.1 (Asymptotic d-stationarity)**.** *Let Assumption 1 and Assumption 2 hold, and let* $\{x^k\}$ *be generated by Algorithm 2 with any oracle* $\mathcal{O}_C$ *producing feasible iterates. Then:*

(i) $\sum_{k=0}^{\infty} (t_{ex}^k)^2 < \infty$*, and in particular* $t_{ex}^k \to 0$.

(ii) Almost surely, every accumulation point of $\{x^k\}$ *is d-stationary.*

Remark 4.1 (Existence of accumulation points)**.** *A sufficient condition for the existence of an accumulation point is boundedness of the level set, $\{x \in C : h(x) \le h(x^0)\}$. Indeed, greedy selection ensures $h(x^{k+1}) \le h(x^k)$ for all k, hence $h(x^k) \le h(x^0)$ and $\{x^k\}$ remains in this set. If it is bounded, then $\{x^k\}$ is bounded in $\mathbb{R}^n$ and admits an accumulation point by Bolzano–Weierstrass.*

The proof of Theorem 4.1 views the exploration step as a probe that is tried at every iteration and moves only when it finds sufficient decrease. Such moves are self-limiting: each accepted step lowers h by at least $\frac{\gamma}{2}$ times its squared length, and h is bounded below on C, so the accepted steps must become negligible, which is part (i). A feasible descent direction at an accumulation point x^* would contradict this. Near x^*, a decrease of order t dominates the quadratic acceptance penalty $\frac{\gamma}{2}t^2$, yielding a window of steps, of lengths bounded away from zero, that remain feasible and pass the acceptance test from every iterate sufficiently close to x^* (Figure 1, Appendix A). The iterates come arbitrarily close to x^* infinitely often and the sampling distributions have full support, so the exploration step almost surely falls into this window infinitely often and keeps making steps that do not shrink. Lemma 4.3 supplies the feasibility radius of the window, Lemma 4.2 the repeated sampling, and the decrease estimates in the proof the acceptance.

Throughout this section, let $\mathcal{Q} \subseteq S^{n-1} \cap L$ be a fixed countable dense set.

Lemma 4.2 (Joint density of exploration direction, proximity, and step size)**.** *There exists an event Ω_{BC} with $\mathbb{P}(\Omega_{BC}) = 1$ such that, on Ω_{BC}, for every $0 < a < b \le r$, every $\rho > 0$, every accumulation point x^* of $\{x^k\}$, and every $v \in \mathcal{Q}$, there exists a subsequence $J \subseteq \mathbb{N}$ such that, as $j \to \infty$ along J, it holds that: $x^j \to x^*$, $v_{ex}^j \to v$, $\left\|v_{ex}^j - v\right\| < \rho$, $\hat{t}^j \in [a,\, b]$.*

Lemma 4.3 (Uniform feasibility radius)**.** *Let $x^* \in C$ and $v \in \mathrm{ri}(D_C(x^*)) \cap L$. There exist $\rho > 0$, $\bar{\varepsilon} > 0$, and $\delta > 0$ such that: (i) $B(v,\, \rho) \cap L \subseteq \mathrm{ri}(D_C(x^*))$; (ii) $\tau(x;\, w) \ge \bar{\varepsilon}$ for every $x \in C$ with $\|x - x^*\| < \delta$ and every $w \in B(v,\, \rho) \cap L$.*

Lemma 4.4 (Continuity and extension)**.** *The directional derivative map $d \mapsto h'(x^*;\, d)$ is continuous and positively homogeneous of degree one. Moreover, if $h'(x^*;\, v) \ge 0$ for all $v \in \mathrm{ri}(D_C(x^*))$, then $h'(x^*;\, d) \ge 0$ for all $d \in \mathrm{cl}(D_C(x^*))$.*

Proof of Theorem 4.1. (i) Step-size decay. Since $t = 0$ is always available in the two-point comparison and t_{ex}^k minimizes over $\{0, \hat{t}^k\}$: $h\big(x^k + t_{ex}^k\, v_{ex}^k\big) + \frac{\gamma}{2}(t_{ex}^k)^2 \le h(x^k)$.

By greedy selection, $h(x^{k+1}) \le h(y^{k+1})$, so $h(x^k) - h(x^{k+1}) \ge \frac{\gamma}{2}(t_{ex}^k)^2$. Summing gives $\sum_k (t_{ex}^k)^2 \le \frac{2}{\gamma}(h(x^0) - \inf_C h) < \infty$, hence $t_{ex}^k \to 0$.

(ii) D-stationarity. Let Ω_{BC} be the event from Lemma 4.2, which guarantees simultaneous subsequences for all directions in $\mathcal{Q}$. Fix $\omega \in \Omega_{BC}$ and work pathwise. Let x^* be an accumulation point of $\{x^k(\omega)\}$, and enumerate $\mathcal{Q} \cap \mathrm{ri}(D_C(x^*))$ as $\{v_i\}_{i=0}^{\infty}$. Since $\mathrm{ri}(D_C(x^*)) \cap S^{n-1}$ is relatively open in $S^{n-1} \cap L$ and $\mathcal{Q}$ is dense in $S^{n-1} \cap L$, the set $\{v_i\}$ is dense in $\mathrm{ri}(D_C(x^*)) \cap S^{n-1}$.

Fix $i \in \mathbb{N}$ and suppose for contradiction $h'(x^*;\, v_i) = -\varepsilon < 0$ for some $\varepsilon > 0$.

By Lemma 4.3 applied with $v = v_i$, there exist $\rho_i > 0$, $\bar{\varepsilon}_i > 0$, and $\delta_i > 0$ such that $B(v_i,\, \rho_i) \cap L \subseteq \mathrm{ri}(D_C(x^*))$, and $\tau(x;\, w) \ge \bar{\varepsilon}_i$ for every $x \in C$ with $\|x - x^*\| < \delta_i$ and every $w \in B(v_i,\, \rho_i) \cap L$.

By local Lipschitz continuity of h at x^*, there exist $R_h > 0$ and $L_h > 0$ such that h is L_h-Lipschitz on $B(x^*, R_h)$.

Next choose the comparison scale. By definition of the directional derivative with $h'(x^*;\, v_i) = -\varepsilon$, there exists $\tau_0 > 0$ such that $h(x^* + t v_i) - h(x^*) \le -\frac{\varepsilon}{2}\, t$ for every $t \in (0,\, \tau_0]$. Set

$$t^* := \min\left\{\bar{\varepsilon}_i,\ \tau_0,\ \frac{\varepsilon}{4\gamma},\ r,\ \frac{R_h}{4}\right\} > 0.$$

Now lock direction and step-size sampling simultaneously. Since $\omega \in \Omega_{BC}$ and the parameters are now fixed, Lemma 4.2 gives an infinite subsequence $\mathcal{J}_i \subseteq \mathbb{N}$ along which $x^j \to x^*, v_{ex}^j \to v_i, \left\|v_{ex}^j - v_i\right\| < \rho_i, \hat{t}^j \in [t^*/2,\, t^*]$.

For $j \in \mathcal{J}_i$ sufficiently large, $\left\|x^j - x^*\right\| < \delta_i$; the feasibility bound above then gives $\tau(x^j;\, v_{ex}^j) \ge \bar{\varepsilon}_i \ge t^*$, so $x^j + t v_{ex}^j \in C$ for every $t \in [0,\, t^*]$.

The key point is that the same strict decrease estimate holds uniformly for every $t \in [t^*/2,\, t^*]$. Since $x^j \to x^*$ along $\mathcal{J}_i$, for all sufficiently large $j \in \mathcal{J}_i$ we have $\|x^j - x^*\| < R_h/4$. For such j and any $t \in [0, t^*]$, $\|x^* + tv_i - x^*\| = t \le t^* \le R_h/4 < R_h$, and, using $\|v_{ex}^j\| = 1$,

$$\|x^j + tv_{ex}^j - x^*\| \le \|x^j - x^*\| + t\,\|v_{ex}^j\| < R_h/4 + t^* \le R_h/2 < R_h.$$

Hence both points $x^j + tv_{ex}^j$ and $x^* + tv_i$ lie in $B(x^*, R_h)$, so the L_h-Lipschitz bound is valid. Therefore, for any $t \in [t^*/2,\, t^*]$ and all sufficiently large $j \in \mathcal{J}_i$, decompose

$$\begin{aligned} h(x^j + tv_{ex}^j) - h(x^j) + \tfrac{\gamma}{2}t^2 &= \big[h(x^j + tv_{ex}^j) - h(x^* + tv_i)\big] + \big[h(x^* + tv_i) - h(x^*)\big] \\ &\quad + \big[h(x^*) - h(x^j)\big] + \tfrac{\gamma}{2}t^2 \\ &\le L_h\big(\|x^j - x^*\| + t\,\|v_{ex}^j - v_i\|\big) - \tfrac{\varepsilon}{2}t + L_h\,\|x^j - x^*\| + \tfrac{\gamma}{2}t^2. \end{aligned}$$

Since $t \in [t^*/2,\, t^*]$, $\frac{\varepsilon}{2}t \ge \frac{\varepsilon t^*}{4}$; and $t^* \le \varepsilon/(4\gamma)$ gives $\frac{\gamma}{2}t^2 \le \frac{\gamma}{2}(t^*)^2 \le \frac{\varepsilon t^*}{8}$. Since $x^j \to x^*$ and $v_{ex}^j \to v_i$ along $\mathcal{J}_i$, and since $[t^*/2,\, t^*]$ is compact and $x^j \to x^*$, $v_{ex}^j \to v_i$, we may pass to a tail subsequence (still denoted $\mathcal{J}_i$) on which, uniformly in $t \in [t^*/2,\, t^*]$,

$$L_h\big(\|x^j - x^*\| + t\,\|v_{ex}^j - v_i\|\big) + L_h\,\|x^j - x^*\| \;\le\; 2L_h\,\|x^j - x^*\| + L_h\,t^*\,\|v_{ex}^j - v_i\| \;\le\; \tfrac{\varepsilon t^*}{16}.$$

Therefore, for every $t \in [t^*/2,\, t^*]$ we have that

$$h(x^j + tv_{ex}^j) - h(x^j) + \tfrac{\gamma}{2}t^2 \;\le\; \tfrac{\varepsilon t^*}{16} - \tfrac{\varepsilon t^*}{4} + \tfrac{\varepsilon t^*}{8} \;=\; -\tfrac{\varepsilon t^*}{16} \;<\; 0.$$

Applied at $t = \hat{t}^j \in [t^*/2,\, t^*]$, together with the feasibility established above, this strict negativity forces the two-point comparison to select $t_{ex}^j = \hat{t}^j \ge t^*/2$ for every $j \in \mathcal{J}_i$.

This yields the contradiction: for infinitely many $j \in \mathcal{J}_i$, $t_{ex}^j \ge t^*/2$, so $\sum_k (t_{ex}^k)^2 \ge \sum_{j \in \mathcal{J}_i} (t^*/2)^2 = +\infty$, contradicting part (i). Hence $h'(x^*;\, v_i) \ge 0$.

Since $\{v_i\}$ is dense in $\mathrm{ri}(D_C(x^*)) \cap S^{n-1}$, Lemma 4.4 (continuity and positive homogeneity of $h'(x^*;\, \cdot)$, together with the relative-interior-to-closure extension) yields $h'(x^*;\, d) \ge 0$ for every $d \in \mathrm{cl}(D_C(x^*))$. Hence x^* is d-stationary. As x^* was an arbitrary accumulation point on this fixed $\omega \in \Omega_{BC}$, every accumulation point is d-stationary on Ω_{BC}, proving the almost-sure claim. □

The preceding argument requires only that the oracle outputs feasible points $z^{k+1} \in C$; beyond this requirement, $\mathcal{O}_C$ may be arbitrary, while d-stationarity is enforced solely by the exploration step. In particular, $\mathcal{O}_C$ may be randomized or adaptive, provided each exploration sample $(v_{ex}^k, \hat{t}^k)$ is independent of the σ-algebra $\mathcal{F}_{k-1}$, as assumed in Remark 3.1.

5 Oracle Instantiations

For DCA, prox-linear methods, and gradient sampling (INGD), the oracle supplies a stationarity measure and a rate, while Theorem 4.1 upgrades accumulation points to d-stationarity almost surely. Under the additional assumptions that $C = \mathbb{R}^n$, h is globally Lipschitz, and the exploration step draws its direction and step size independently and uniformly, exploration alone also yields an expected iteration bound, exponential in n, for Goldstein stationarity (Appendix F). Throughout this section, $h^* := \inf_{x \in C} h(x)$, which is finite by Assumption A1, and $\Delta_0 := h(x^0) - h^*$, where x^0 is the initial point of Algorithm 2.

5.1 DCA for constrained DC programming

Consider the constrained difference-of-convex problem

$$\min_{x \in C} h(x) := f_1(x) - f_2(x), \tag{2}$$

where $C \subseteq \mathbb{R}^n$ is closed and convex, $f_1, f_2 \colon \mathbb{R}^n \to \mathbb{R}$ are finite-valued, f_1 is μ_1-strongly convex and f_2 is μ_2-strongly convex with $\mu_1 + \mu_2 > 0$, and h is bounded below on C. We assume that the DCA subproblem below has a solution at every point of C, as it does when $\mu_1 > 0$.

Given $x^k \in C$, the DCA oracle produces

$$z^{k+1} \in \arg\min_{x \in C} \; f_1(x) - \langle g_2^k, x \rangle, \quad g_2^k \in \partial f_2(x^k).$$

The augmented method applies Algorithm 2 with this oracle. The key observation is that the rate analysis of DCA relies only on the descent inequality $h(x^{k+1}) \le h(z^{k+1})$, which is preserved by the greedy selection step regardless of whether the exploration step is accepted; the proof is deferred to Appendix B.

Theorem 5.1 (Rate preservation under DCA)**.** *Under the assumptions above, after N iterations of Algorithm 2 with the DCA oracle,* $\min_{0 \le k \le N-1} \left\|x^k - z^{k+1}\right\| \le \sqrt{\frac{2\,(h(x^0) - h^*)}{(\mu_1 + \mu_2)\,N}}$.

Thus, the augmentation preserves the $O(1/\sqrt{N})$ rate on the natural oracle residual (Remark B.1), while Theorem 4.1 ensures that every accumulation point is almost surely d-stationary, which is strictly stronger than the criticality guarantee of DCA.

5.2 General rate transfer

The proof of Theorem 5.1 uses the DCA oracle only through Lemma B.1 (a lower bound on $h(x^k) - h(z^{k+1})$), and the exploration step only through greedy selection, which gives, whether or not the exploration step is accepted,

$$h(x^{k+1}) \le h(z^{k+1}) \qquad \text{for every } k. \tag{G}$$

As in Algorithm 2, we write $z^{k+1} \leftarrow \mathcal{O}_C(x^k)$: the oracle returns a single point. When its defining rule admits several outputs, $\mathcal{O}_C(x)$ denotes the one it returns; when it is randomized, $\mathcal{O}_C(x)$ is a random variable and z^{k+1} is its realization at iteration k. We consider three properties of $\mathcal{O}_C$.

- (O1) Restartable: $\mathcal{O}_C$ is defined on C, and $\mathcal{O}_C(x)$ depends only on x, not on the iteration index or on state carried across calls. For a randomized oracle, $\mathcal{O}_C(x) = \Psi(x, \xi)$ for a fixed measurable map Ψ, with ξ drawn at each call from the same distribution, independently of the past.
- (O2) Feasible: $\mathcal{O}_C(x) \in C$ for every $x \in C$, as Algorithm 2 requires of every oracle.
- (O3) Non-ascending: $h(\mathcal{O}_C(x)) \le h(x)$ for every $x \in C$.

For a randomized oracle, (O2), (O3) and ($\star$) below are required for every realization of ξ.

Proposition 5.2 (Budget)**.** *Consider a run of Algorithm 2 along which $h(z^{k+1}) \le h(x^k)$ for every k, as holds under (O3). Then, along this run, for every $K \ge 1$ and every $\eta > 0$,*

$$\sum_{k=0}^{K-1} \big(h(x^k) - h(z^{k+1})\big) \le \Delta_0,$$

$$\min_{k<K} \big(h(x^k) - h(z^{k+1})\big) \le \frac{\Delta_0}{K}, \qquad \#\big\{k < K : h(x^k) - h(z^{k+1}) \ge \eta\big\} \le \frac{\Delta_0}{\eta}.$$

The proof (Appendix C) telescopes (G). The exploration randomness does not enter it, so the bounds hold surely, and they hold verbatim for the base method without exploration, for which $z^{k+1} = x^{k+1}$.

Let $\Lambda(x) := h(x) - h(\mathcal{O}_C(x))$ denote the oracle decrement at x, a function on C for a deterministic oracle and a random variable for a randomized one; its value at x^k is the summand in Proposition 5.2. A rate is stated in an optimality measure $\sigma\colon C \times C \to \mathbb{R}_{\ge 0}$ of the base method. Under (O1) and (O3), we say that $\mathcal{O}_C$ admits a monotone calibration for σ if, for some nondecreasing $\varphi\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$,

$$\sigma\big(x, \mathcal{O}_C(x)\big) \le \varphi\big(\Lambda(x)\big) \qquad \text{for every } x \in C. \tag{$\star$}$$

Theorem 5.3 (Rate transfer)**.** *Let $\sigma\colon C \times C \to \mathbb{R}_{\ge 0}$, and let $\varphi\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be nondecreasing. Consider a run of Algorithm 2 along which, for every $k \ge 0$,*

$$h(z^{k+1}) \le h(x^k) \quad \textit{and} \quad \sigma(x^k, z^{k+1}) \le \varphi\big(h(x^k) - h(z^{k+1})\big). \tag{T}$$

Then, along this run, for every $K \ge 1$ and every $\eta > 0$,

$$\min_{k<K} \sigma(x^k, z^{k+1}) \le \varphi\Big(\frac{\Delta_0}{K}\Big), \qquad \#\big\{k < K : \sigma(x^k, z^{k+1}) > \varphi(\eta)\big\} \le \frac{\Delta_0}{\eta}.$$

If $\mathcal{O}_C$ satisfies (O1) and (O3) and admits a monotone calibration for σ, then (T) holds along every run, since $x^k \in C$ and z^{k+1} is the output of $\mathcal{O}_C$ at x^k. This is the checkable sufficient condition for (T): a descent lemma proved at an arbitrary point of C yields ($\star$), and therefore applies at iterates produced by exploration, whereas for an oracle that carries state across calls it holds only along the base method's own trajectory. For the DCA oracle of Section 5.1, with choices of $g_2 \in \partial f_2(x)$ and of the minimizer that depend only on x, (O1)–(O3) hold, and Lemma B.1, whose proof applies at every $x \in C$, is ($\star$) with $\sigma(x, z) = \|x - z\|$ and $\varphi(d) = \sqrt{2d/(\mu_1 + \mu_2)}$; Theorem 5.3 then gives Theorem 5.1 with the same constant. If φ is continuous at 0 with $\varphi(0) = 0$, as here, the first bound of Theorem 5.3 is a convergence rate whose order φ determines; if φ is a step function, the second bounds the number of iterations at which σ exceeds a threshold.

Oracles that certify stationarity rather than measure it admit no calibration with $\varphi(0) = 0$ for a measure σ with $\sigma(x, z) > 0$ whenever x is not stationary, since a call can return $z = x$, so that $\Lambda(x) = 0$, at such a point. For them, Proposition 5.2 yields a hitting time instead. Suppose that the call at iteration k also returns a label $\ell^k \in \{\mathrm{CERT}, \mathrm{PROG}, \mathrm{FAIL}\}$. Let $\mathcal{G}_{k-1}$ be the σ-algebra generated by x^0 and by the exploration samples, oracle outputs and labels of iterations $0, \dots, k-1$ (so $\mathcal{G}_{-1}$ is generated by x^0), so that x^k is $\mathcal{G}_{k-1}$-measurable. Consider the following properties.

(P1) $\ell^k = \mathrm{CERT}$ implies that x^k satisfies the target stationarity criterion.

(P2) $\ell^k = \mathrm{PROG}$ implies $h(x^k) - h(z^{k+1}) \geq \bar{\eta}$, for a constant $\bar{\eta} > 0$.

(P3) $\mathbb{P}(\ell^k = \mathrm{FAIL} \mid \mathcal{G}_{k-1}) \leq q$ almost surely, for every k.

Theorem 5.4 (Certificate within a finite horizon)**.** *Let $\mathcal{O}_C$ satisfy (O3) and (P1)–(P3) with $q < 1/e$. There is a constant κ_q, depending only on q, such that for every $\zeta \in (0, 1)$, with probability at least $1 - \zeta$, some call $\mathcal{O}_C(x^k)$ with $k < \kappa_q \left(1 + \Delta_0/\bar{\eta} + \log(1/\zeta)\right)$ returns* CERT*. An explicit horizon is given in Appendix C.*

The proof is in Appendix C. Under (O1), with the label computed from the same draw, a failure bound at every fixed point of C gives (P3), since the draw at iteration k is independent of $\mathcal{G}_{k-1}$ and x^k is $\mathcal{G}_{k-1}$-measurable.

The conditions of Theorems 5.3 and 5.4 are sufficient for rate transfer, and we do not claim that they are necessary. Since (G) bounds $h(x^{k+1})$ but does not give $x^{k+1} = z^{k+1}$, the analysis of a base method yields (T) only if its descent inequality holds at the points that exploration produces. Condition (O1) fails for oracles that carry state across calls, such as accelerated, bundle and quasi-Newton methods, whose analysis does not in general supply (T) along the augmented run. Deriving (T) or (P3) from the base analysis thus relies on (O1), whereas Theorem 4.1 allows adaptive oracles. Condition (O2) fails for penalty and augmented Lagrangian methods, whose iterates are feasible only in the limit; they lie outside Algorithm 2. Condition ($\star$) can fail under (O1)–(O3), for instance for the distance to a solution set. Condition (O3) fails for methods that may increase h, such as nonmonotone line searches. These cases, and monotone accelerated variants such as MFISTA [Beck and Teboulle, 2009a], which we leave open, are discussed in Appendix C.1.

5.3 Prox-linear Oracle

We next consider composite objectives of the form

$$\min_{x \in \mathbb{R}^n} h(x) := c(F(x)), \tag{3}$$

where h is bounded below, F is continuously differentiable with Lipschitz Jacobian and c is globally Lipschitz and directionally differentiable. We assume in addition that c is *Hadamard directionally differentiable* on $\mathbb{R}^m$: for every $v, w \in \mathbb{R}^m$ the limit $c'(v;\, w) := \lim_{(t,\, w') \to (0^+,\, w)} \big(c(v + tw') - c(v)\big)/t$ exists, and for each fixed v the map $w \mapsto c'(v;\, w)$ is continuous on $\mathbb{R}^m$. Since c is Lipschitz and F is continuously differentiable, $h = c \circ F$ is locally Lipschitz, and the chain rule for Hadamard directional derivatives shows that h is Hadamard directionally differentiable at every $x \in \mathbb{R}^n$, with

$$h'(x;\, d) = c'\big(F(x);\, \nabla F(x)\, d\big). \tag{4}$$

Hence h satisfies Assumptions A1 and A2.

Given x^k, define the local model $m^k(y) := c\big(F(x^k) + \nabla F(x^k)(y - x^k)\big) + \frac{\rho}{2} \left\|y - x^k\right\|^2$, and, because c need not be convex, let the oracle return z^{k+1} from a first-order subproblem variant (Ap-

pendix D) while enforcing

$$m^k(z^{k+1}) \le m^k(x^k) = h(x^k). \tag{5}$$

When ρ is sufficiently large, this implies $h(z^{k+1}) \le h(x^k)$ and $h(x^k) - h(z^{k+1}) \ge \alpha \left\|z^{k+1} - x^k\right\|^2$ for some $\alpha > 0$ (Lemma D.1). Along a run, this is (T) with $\sigma(x, z) = \|x - z\|$ and $\varphi(d) = \sqrt{d/\alpha}$, so Theorem 5.3 gives the following result (proof in Appendix D.1).

Theorem 5.5 (Rate preservation for prox-linear oracles)**.** *Suppose the oracle satisfies* (5) *and that there exists $\alpha > 0$ such that $h(x^k) - h(z^{k+1}) \ge \alpha \left\|z^{k+1} - x^k\right\|^2$ for every k. Then, after N iterations of Algorithm 2, $\min_{0 \le k \le N-1} \left\|x^k - z^{k+1}\right\| \le \sqrt{\frac{h(x^0) - h^*}{\alpha N}}$.*

The prox-linear framework includes a broad class of nonsmooth nonconvex problems; different inner solvers may yield different notions of stationarity (e.g., Clarke or Mordukhovich stationarity). Regardless of these distinctions, Theorem 4.1 upgrades all accumulation points of the augmented method to d-stationarity almost surely.

5.4 Gradient Sampling Oracle and Goldstein Stationarity

Let $C = \mathbb{R}^n$, and let $h \in \mathcal{F}(\Delta, L)$, the class of L-Lipschitz, Hadamard directionally differentiable functions with $h(x^0) - \inf h \le \Delta$ studied by Zhang et al. [2020], who access h through a generalized-gradient oracle O that returns $h(y)$ and an element of $\partial^C h(y)$ (Appendix E). Such h satisfies Assumptions A1 and A2. Rates for this class concern a relaxation of stationarity.

Definition 5.1 (Goldstein stationarity [Goldstein, 1977])**.** *Let $\delta, \varepsilon > 0$. The Goldstein δ-subdifferential of h at x is $\partial_\delta h(x) := \operatorname{conv} \bigcup_{y \in B[x,\delta]} \partial^C h(y)$, and x is (δ, ε)-Goldstein stationary if* $\operatorname{dist}(0, \partial_\delta h(x)) \le \varepsilon$.

The single-pass INGD oracle $\mathcal{O}_C(x)$ runs once, at the anchor x and with fresh randomness, the inner loop of interpolated normalized gradient descent (INGD) [Zhang et al., 2020, Algorithm 1], with $J := \lceil 48L^2/\varepsilon^2 \rceil$ steps. The loop keeps an element m_j of $\partial_\delta h(x)$. It returns x with the label CERT when $\|m_j\| \le \varepsilon$, returns a trial point with the label PROG when that point passes a sufficient-decrease test, and returns x with the label FAIL when J steps pass without either (Appendix E). Since a FAIL call returns $z^{k+1} = x^k$ at a point that need not be stationary, this oracle admits no calibration (Section 5.2), and its rate follows from Theorem 5.4.

Theorem 5.6 (INGD instance; proof in Appendix E)**.** *Let $\delta > 0$, $0 < \varepsilon \le L$ and $\zeta \in (0, 1)$, and run Algorithm 2 with the single-pass INGD oracle. This oracle satisfies (O1)–(O3) and (P1)–(P3) with $\bar{\eta} = \delta\varepsilon/4$ and $q = 1/4$. With probability at least $1 - \zeta$, some call $\mathcal{O}_C(x^k)$ returns* CERT*, so that x^k is (δ, ε)-Goldstein stationary, and the calls up to and including it make at most $800\, L^2\varepsilon^{-2}\big(1 + \Delta_0\delta^{-1}\varepsilon^{-1} + \log(1/\zeta)\big)$ calls to O.*

Up to the constant, the term $\Delta_0 L^2 \varepsilon^{-3} \delta^{-1}$ of this bound is the leading term of the bound of Zhang et al. [2020, Theorem 8] for INGD itself, in which $\log(1/\zeta)$ enters through the factor $\log(4\Delta/(\zeta\delta\varepsilon))$ rather than additively. Exploration makes no calls to O and adds at most one function evaluation per iteration (Remark 3.2), against up to $J + 1$ calls to O per oracle call. Theorem 5.6 certifies an iterate at a finite time; by Theorem 4.1, every accumulation point of $\{x^k\}$ is d-stationary almost surely.

6 Numerical Experiments

Our experiments evaluate whether one exploration step per iteration improves the terminal solution relative to the base oracle across representative benchmarks, using long runs to approach the asymptotic regime implied by the theory. The full details are deferred to Appendix G.

Protocol in brief. For each instance, we compare the base oracle with its augmented version over three seeds, reporting $\Delta = h_{\text{base}} - \operatorname{median}_s h_{\text{aug},s}$, where h_{base} is the objective value at the final iterate of the base oracle, $h_{\text{aug},s}$ is that of the augmented method with seed s, and median_s is the median over the three seeds. Positive values of Δ favor augmentation. Table 2 summarizes win/tie/loss counts, gap statistics, and d-stationarity failures.

Table 2: Headline outcomes across the three focused benchmarks. *w/t/l* denotes win/tie/loss under $\Delta = h_{\text{base}} - \text{median}_s\, h_{\text{aug},s}$. *# non-d-stat (base / aug)* counts runs that fail the benchmark d-stationarity diagnostic. ReLU aggregates two settings (200 paired comparisons total).

Experiment	n_{inst}	w/t/l	mean Δ	med. win Δ	McNemar p	# non-d-stat (base / aug)
Trimmed lasso, λ=1	100	99/1/0	$+0.29$	$+0.27$	$1.6 \cdot 10^{-30}$	$99/4$
LTS, σ_{clean}=4	100	9/91/0	$+3.96$	$+52.2$	$2.0 \cdot 10^{-3}$	$0/0$
ReLU regression	200	44/156/0	$+0.96$	$+1.26$	$5.7 \cdot 10^{-14}$	$0/0$

Trimmed lasso. In trimmed lasso, the difference between criticality and d-stationarity is evident in our diagnostics. In the focused setting ($m = 50$, $n = 100$, $k = 5$, $\lambda = 1$), plain DCA fails the d-stationarity diagnostic on $99/100$ instances versus $4/100$ with augmentation, with objective outcomes $99/1/0$ wins/ties/losses under Δ. Directions are drawn from the gauss-axis mixture specified in Appendix G, which also records matched sphere runs under the same budgets. Uniform sphere sampling accepts virtually no exploration steps and behaves similarly to DCA. The high base failure rate alongside the strong augmentation wins is consistent with a standard DC picture [de Oliveira, 2020]: where the nonsmooth term has nonsingleton subgradient, criticality ($\partial f(x) \cap \partial g(x) \neq \emptyset$) can hold without d-stationarity ($\partial g(x) \subseteq \partial f(x)$).

Least trimmed squares. Focused LTS ($m = 100$, $n = 50$, $q = 10$): zero d-stationarity failures for base and augmented runs, $9/91/0$ wins/ties/losses, and a large win-conditional median Δ ($+52.2$). Augmentation is mostly inert but never hurts and occasionally finds a substantially better basin.

ReLU regression. Two focused ReLU settings (200 comparisons) with a prox-linear inner oracle and an exact DC subdifferential-inclusion d-stationarity check: zero d-stationarity failures, $44/156/0$ wins/ties/losses, mean Δ= + 0.96, McNemar p=$5.7 \cdot 10^{-14}$, showing the same augmentation carries beyond DCA.

7 Conclusions and Limitations

Conclusions. We showed that one additional random feasible-direction step per iteration suffices to enforce almost-sure convergence of accumulation points to d-stationarity, independently of the underlying oracle. The key mechanism is a structural separation: the oracle determines convergence rates and practical performance, while exploration enforces asymptotic d-stationarity at negligible cost. Under the conditions of Theorems 5.3 and 5.4, rate guarantees of the base method are preserved, with the exploration step entering only through the inequality $h(x^{k+1}) \leq h(z^{k+1})$, while d-stationarity follows from the exploration step alone. Rate transfer covers DCA and prox-linear methods through Theorem 5.3, and gradient sampling through Theorem 5.4. These results suggest that lightweight directional exploration can serve as a modular tool for strengthening optimality guarantees. In particular, integrating structured feasible-direction mechanisms with model-based updates is a promising direction for advancing nonsmooth nonconvex optimization.

Limitations. The d-stationarity guarantee is asymptotic. Theorem 4.1 ensures almost-sure d-stationarity of accumulation points, but provides no finite-iteration bound for approaching or certifying it. The rates in Section 5 concern the stationarity measure supplied by the base oracle, not d-stationarity; thus finite-time proximity to a d-stationary point is not guaranteed.

Finite-time performance depends on the sampling distribution $\mathcal{D}_v$. Although the theory requires only full support on $S^{n-1} \cap L$, this may be ineffective in practice. In trimmed lasso, the uniform sphere sampler accepts virtually no exploration steps, whereas an axis-biased Gaussian mixture yields consistent gains by exploiting ℓ_1-type sparsity. When such structure is unavailable and the base oracle is treated as a black box, designing an effective sampler is nontrivial.

Acknowledgments and Disclosure of Funding

No external funding was received for this work. The authors declare no competing financial interests.

References

H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani. On the rate of convergence of the difference-of-convex algorithm (DCA). *Journal of Optimization Theory and Applications*, 202(1):475–496, 2024.

M. Ahn, J.-S. Pang, and J. Xin. Difference-of-convex learning: directional stationarity, optimality, and sparsity. *SIAM Journal on Optimization*, 27(3):1637–1665, 2017.

C. Audet and J. E. Dennis Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.

A. Beck and N. Hallak. On the convergence to stationary points of deterministic and randomized feasible descent directions methods. *SIAM Journal on Optimization*, 30(1):56–79, 2020.

A. Beck and N. Hallak. The regularized feasible directions method for nonconvex optimization. *Operations Research Letters*, 50(5):517–523, 2022.

A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009a.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009b.

F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

D. Davis, D. Drusvyatskiy, Y. T. Lee, S. Padmanabhan, and G. Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

W. de Oliveira. The ABC of DC programming. *Set-Valued and Variational Analysis*, 28(4):679–706, 2020.

W. de Oliveira. A bundle method for nonsmooth DC programming with application to chance-constrained problems. *Computational Optimization and Applications*, 78(2):451–490, 2021.

D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

Z. Feng and Y. Yuan. A perturbed DCA for computing d-stationary points of nonsmooth DC programs, 2026. Preprint, arXiv:2601.02084.

A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13 (1):14–22, 1977.

S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic descent. *SIAM Journal on Optimization*, 25(3):1515–1541, 2015.

S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic feasible descent for bound and linearly constrained problems. *Computational Optimization and Applications*, 72(3):525–559, 2019.

K. Joki, A. M. Bagirov, N. Karmitsa, M. M. Mäkelä, and S. Taheri. Double bundle method for finding Clarke stationary points in nonsmooth DC programming. *SIAM Journal on Optimization*, 28(2):1892–1919, 2018.

T. G. Kolda, R. M. Lewis, and V. Torczon. Stationarity results for generating set search for linearly constrained optimization. *SIAM Journal on Optimization*, 17(4):943–968, 2006.

R. M. Lewis and V. Torczon. Pattern search methods for linearly constrained minimization. *SIAM Journal on Optimization*, 10(3):917–941, 2000.

J.-S. Pang and M. Tao. Decomposition methods for computing directional stationary solutions of a class of nonsmooth nonconvex optimization problems. *SIAM Journal on Optimization*, 28(2): 1640–1669, 2018.

J. S. Pang, M. Razaviyayn, and A. Alvarado. Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118, 2017.

R. T. Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1997.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer, New York, 2009.

A. N. Shiryaev. *Probability-2*, volume 95 of *Graduate Texts in Mathematics*. Springer, 3 edition, 2019.

F. J. Solis and R. J.-B. Wets. Minimization by random search techniques. *Mathematics of Operations Research*, 6(1):19–30, 1981.

Z. Sun and L. Wu. Hybrid algorithms for finding a d-stationary point of a class of structured nonsmooth DC minimization. *SIAM Journal on Optimization*, 34(1):485–506, 2024.

M. Tao and J.-N. Li. Error bound and isocost imply linear convergence of DCA-based algorithms to d-stationarity. *Journal of Optimization Theory and Applications*, 197(1):205–232, 2023.

Y. Wardi. Random search algorithms with sufficient descent for minimization of functions. *Mathematics of Operations Research*, 14(2):343–354, 1989.

Y. Ye, Y. Cui, and J. Wang. An online adaptive sampling algorithm for stochastic difference-of-convex optimization with time-varying distributions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 71959–71978. PMLR, 2025.

J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11173–11182. PMLR, 2020.

A Proofs for Sections 1 and 4

Proof of Lemma 1.1. By hypothesis there is $R > 0$ such that h is L-Lipschitz on the open ball $B(x, R) := \{y \in \mathbb{R}^n : \|y - x\| < R\}$. Fix $d_1, d_2 \in \mathbb{R}^n$ and $0 < t < R/(1 + \max\{\|d_1\|, \|d_2\|\})$. Then $x + td_1, x + td_2 \in B(x, R)$, and therefore

$$\left| \frac{h(x + td_1) - h(x)}{t} - \frac{h(x + td_2) - h(x)}{t} \right| = \frac{|h(x + td_1) - h(x + td_2)|}{t} \leq L \|d_1 - d_2\| .$$

By Assumption A2, since $x \in C$, the two difference quotients converge to $h'(x; d_1)$ and $h'(x; d_2)$ as $t \downarrow 0$, and the bound passes to the limit. Continuity of $d \mapsto h'(x; d)$ follows. □

Figure 1 illustrates part (ii) of the proof of Theorem 4.1.

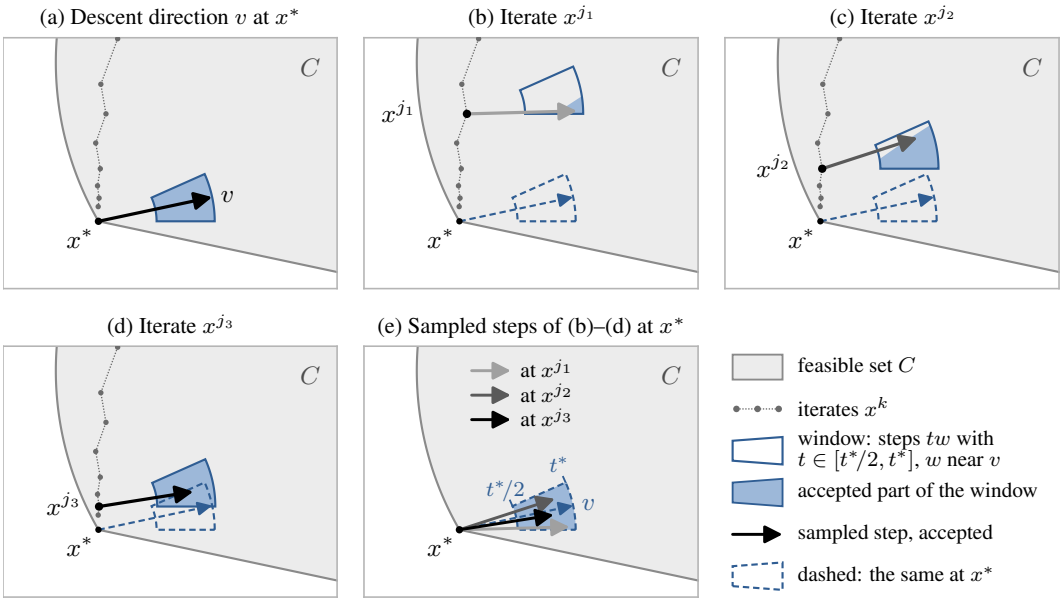


Figure 1: The contradiction argument in part (ii) of the proof of Theorem 4.1, in a two-dimensional example with $L = \mathbb{R}^2$ and a nonsmooth h. (a) Suppose, for contradiction, that an accumulation point x^* of $\{x^k\}$ admits a feasible descent direction. As in the proof, by Lemma 4.4 and continuity of $h'(x^*; \cdot)$, v can be taken among the directions v_i there, which lie in the relative interior of $D_C(x^*)$, with $h'(x^*; v) < 0$; the arrow shows t^*v. The window at a point x consists of the points $x + tw$ with $t \in [t^*/2, t^*]$ and w a unit direction sufficiently close to v. Its shaded part consists of the points $y \in C$ that pass the acceptance test $h(y) + \frac{\gamma}{2}\|y - x\|^2 < h(x)$ of Algorithm 1 at x. At x^*, the whole window passes. (b)–(d) Iterates $x^{j_1}, x^{j_2}, x^{j_3}$ along the subsequence of Lemma 4.2, with the sampled step $\hat{t}^j v^j_{\text{ex}}$ at each (arrow). The dashed outline is the window at x^*. As $x^j \to x^*$, the window at x^j approaches the window at x^*, and for x^j sufficiently close to x^*, the whole window at x^j lies in C (Lemma 4.3) and passes the acceptance test (the estimates in the proof of Theorem 4.1). (e) The three sampled steps, redrawn from x^* at the same scale, inside the window at x^* (dashed). Along the subsequence, $v^j_{\text{ex}} \to v$ and $\hat{t}^j \in [t^*/2, t^*]$, so for all large j the sampled step falls inside the window. Once the whole window at x^j passes the test, each such step is accepted, with $t^j_{\text{ex}} = \hat{t}^j \geq t^*/2$. This happens for infinitely many j, so the accepted step lengths do not tend to zero: $\sum_k (t^k_{\text{ex}})^2 = \infty$, which contradicts part (i). Hence x^* admits no feasible descent direction.

Proof of Lemma 4.2. We assume throughout that $\dim L \geq 1$; otherwise C is a singleton and the result is trivial.

Recall the σ-algebra $\mathcal{F}_{k-1}$ of Remark 3.1: x^k is $\mathcal{F}_{k-1}$-measurable, while v^k_{ex} and $\hat{t}^k$ are independent of $\mathcal{F}_{k-1}$ and of each other.

Fix a point $y \in C$, a radius $\delta > 0$, a spherical cap $V := B[w, \varepsilon'] \cap S^{n-1} \cap L$, for some $w \in S^{n-1} \cap L$ and $\varepsilon' > 0$, and an interval $[a', b'] \subseteq (0, r]$ with $a' < b'$. Set

$$p_v := \mathbb{P}(v_{ex}^k \in V) > 0, \qquad p_t := \mathbb{P}(\hat{t}^k \in [a', b']) > 0,$$

where positivity follows from the full-support assumptions.

Define

$$A_k := \{v_{ex}^k \in V\} \cap \{\hat{t}^k \in [a', b']\} \cap \{\|x^k - y\| < \delta\}.$$

Independence yields $\mathbb{P}(A_k \mid \mathcal{F}_{k-1}) = \mathbf{1}_{\{\|x^k - y\| < \delta\}} \cdot p_v \cdot p_t$.

By the conditional Borel–Cantelli–Lévy lemma [Shiryaev, 2019, Ch. VII, §5, Cor. 2],

$$\Big\{\sum_k \mathbb{P}(A_k \mid \mathcal{F}_{k-1}) = \infty\Big\} = \{A_k \text{ occurs infinitely often}\} \quad \text{a.s.}$$

In particular, whenever x^k visits $B(y, \delta)$ infinitely often, A_k occurs infinitely often almost surely.

Let $\{y_n\}$ be a countable dense subset of C, let $\mathcal{Q} = \{q_i\}_{i \in \mathbb{N}}$, and let $(\delta_m, \varepsilon_l', a_s', b_s')$ range over positive rationals with $a_s' < b_s' \le r$. For each tuple $u = (n, i, m, l, s)$, define the event

$$E_u := \Big\{\|x^k - y_n\| < \delta_m \text{ i.o.} \implies \big(v_{ex}^k \in B[q_i, \varepsilon_l'] \cap S^{n-1} \cap L,\ \hat{t}^k \in [a_s', b_s']\big) \text{ i.o.}\Big\}.$$

By the conditional Borel–Cantelli–Lévy lemma, $\mathbb{P}(E_u) = 1$ for every u: on paths where $\|x^k - y_n\| < \delta_m$ occurs infinitely often, the enriched event occurs infinitely often almost surely; otherwise the implication is vacuous. Set $\Omega_{BC} := \bigcap_u E_u$. Since the index set $\{u\}$ is countable and each E_u has probability 1, we have $\mathbb{P}(\Omega_{BC}) = 1$. On Ω_{BC}, the following grid property holds simultaneously:

For every tuple $(y_n, \delta_m, q_i, \varepsilon_l', a_s', b_s')$, if $\|x^k - y_n\| < \delta_m$ for infinitely many k, then there are infinitely many such k with additionally

$$v_{ex}^k \in B[q_i, \varepsilon_l'] \cap S^{n-1} \cap L, \qquad \hat{t}^k \in [a_s', b_s'].$$

Fix $\omega \in \Omega_{BC}$. Let x^* be any accumulation point of $\{x^k(\omega)\}$ and fix $v \in \mathcal{Q}$. For this fixed ω, write $x^k := x^k(\omega)$ and similarly for sampled variables. Fix arbitrary $\eta \in (0, \rho)$. By density, choose rationals $y_n, \delta_m, \varepsilon_l', a_s', b_s'$ such that $\|y_n - x^*\| < \eta/4, \delta_m \in (\eta/2, 3\eta/4), \varepsilon_l' < \eta/4$, and $a < a_s' < b_s' < b$. Choose i such that $q_i = v$. Since x^* is an accumulation point, $\|x^k - x^*\| < \eta/4$ for infinitely many k. For such indices,

$$\|x^k - y_n\| \le \|x^k - x^*\| + \|x^* - y_n\| < \eta/2 < \delta_m.$$

Hence, by the grid property, there are infinitely many indices k such that: $\|x^k - y_n\| < \delta_m, v_{ex}^k \in B[q_i, \varepsilon_l']$, for $\hat{t}^k \in [a_s', b_s']$. For these indices,

$$\|x^k - x^*\| \le \|x^k - y_n\| + \|y_n - x^*\| < \delta_m + \eta/4 < \eta,$$

and $\|v_{ex}^k - v\| < \varepsilon_l' < \eta < \rho$, while $\hat{t}^k \in [a_s', b_s'] \subseteq (a, b)$.

Let $\eta_q \downarrow 0$ with $\eta_1 < \rho$, and for each q let K_q be the infinite set of indices satisfying the above properties with $\eta = \eta_q$. Choose inductively $j_q \in K_q$ with $j_q > j_{q-1}$. Then, since $\|x^{j_q} - x^*\| < \eta_q$ and $\eta_q \to 0$, we have $x^{j_q} \to x^*$. Similarly, $\|v_{ex}^{j_q} - v\| < \eta_q$ implies $v_{ex}^{j_q} \to v$, while $\|v_{ex}^{j_q} - v\| < \rho$ and $\hat{t}^{j_q} \in [a, b]$ for all q. This completes the proof. □

Proof of Lemma 4.3. Let $k := \dim(L)$. Since $v \in \mathrm{ri}(D_C(x^*))$, there exists $r_0 > 0$ such that

$$B(v, r_0) \cap L \subseteq \mathrm{ri}(D_C(x^*)).$$

Pick $k+1$ directions $d_0, \ldots, d_k \in B(v, r_0) \cap L$ whose convex hull is a k-dimensional simplex in L containing v in its relative interior. Then

$$\mathrm{conv}\{d_0, \ldots, d_k\} \subseteq \mathrm{ri}(D_C(x^*)).$$

Since each $d_j \in D_C(x^*)$, choose $\varepsilon_j > 0$ with $x^* + \varepsilon_j d_j \in C$, and set $\bar{\varepsilon} := \min_j \varepsilon_j > 0$. By convexity of C, $p_j := x^* + \bar{\varepsilon}\, d_j \in C$ for each j, and for any convex combination $d = \sum_j \lambda_j d_j$,

$$x^* + \bar{\varepsilon}\, d = \sum_j \lambda_j\, p_j \in C.$$

Since v lies in the relative interior of the simplex, there exists $\rho_0 > 0$ with

$$B(v, \rho_0) \cap L \subseteq \operatorname{conv}\{d_0, \dots, d_k\},$$

and hence $x^* + \bar{\varepsilon}\, w \in C$ for every $w \in B(v, \rho_0) \cap L$. The map $w \mapsto x^* + \bar{\varepsilon}\, w$ sends $B(v, \rho_0) \cap L$ onto

$$B(x^* + \bar{\varepsilon}\, v,\ \bar{\varepsilon}\rho_0) \cap \operatorname{aff}(C),$$

and therefore

$$B(x^* + \bar{\varepsilon}\, v,\ \bar{\varepsilon}\rho_0) \cap \operatorname{aff}(C) \subseteq C.$$

Set $\rho := \rho_0/2$ and $\delta := \bar{\varepsilon}\rho_0/2$. Then (i) holds because

$$B(v,\ \rho) \cap L \subseteq B(v,\ \rho_0) \cap L \subseteq \operatorname{conv}\{d_0, \dots, d_k\} \subseteq \operatorname{ri}(D_C(x^*)).$$

For (ii), fix $x \in C$ with $\|x - x^*\| < \delta$ and $w \in B(v,\ \rho) \cap L$. Then

$$\|(x + \bar{\varepsilon}\, w) - (x^* + \bar{\varepsilon}\, v)\| \le \|x - x^*\| + \bar{\varepsilon}\, \|w - v\| < \delta + \bar{\varepsilon}\rho = \bar{\varepsilon}\rho_0,$$

and since $x \in C \subseteq \operatorname{aff}(C)$ and $w \in L$, we have $x + \bar{\varepsilon}\, w \in \operatorname{aff}(C)$; hence

$$x + \bar{\varepsilon}\, w \in B(x^* + \bar{\varepsilon}\, v,\ \bar{\varepsilon}\rho_0) \cap \operatorname{aff}(C) \subseteq C.$$

Convexity then gives

$$x + tw = (1 - t/\bar{\varepsilon})\, x + (t/\bar{\varepsilon})\, (x + \bar{\varepsilon}\, w) \in C \qquad \text{for every } t \in [0,\ \bar{\varepsilon}],$$

so $\tau(x;\ w) \ge \bar{\varepsilon}$. □

Proof of Lemma 4.4. Let $K := D_C(x^*)$, a convex cone. By Lemma 1.1, the map $d \mapsto h'(x^*;\ d)$ is continuous. We next show positive homogeneity. For any $\alpha > 0$,

$$h'(x^*;\ \alpha d) = \lim_{t \downarrow 0} \frac{h(x^* + t\alpha d) - h(x^*)}{t} = \alpha \lim_{t \downarrow 0} \frac{h(x^* + td) - h(x^*)}{t} = \alpha h'(x^*;\ d).$$

For the extension statement, recall that for any convex set K,

$$\operatorname{cl}(\operatorname{ri}(K)) = \operatorname{cl}(K),$$

where the relative interior is taken with respect to $\operatorname{aff}(K) = L$. Let $d \in \operatorname{cl}(K)$. Then there exists a sequence $v_m \in \operatorname{ri}(K)$ with $v_m \to d$. By continuity,

$$h'(x^*;\ d) = \lim_{m \to \infty} h'(x^*;\ v_m) \ge 0,$$

since $h'(x^*;\ v_m) \ge 0$ for all m. □

B Proofs for Section 5.1: DCA

Lemma B.1 (Descent at the oracle output)**.** *For each k, $h(x^k) - h(z^{k+1}) \ge \frac{\mu_1+\mu_2}{2} \left\|x^k - z^{k+1}\right\|^2$.*

Proof. By μ_1-strong convexity of f_1 at the subgradient $s^{k+1} \in \partial f_1(z^{k+1})$:

$$f_1(x^k) \ge f_1(z^{k+1}) + \langle s^{k+1},\ x^k - z^{k+1} \rangle + \tfrac{\mu_1}{2} \left\|x^k - z^{k+1}\right\|^2.$$

First-order optimality of the DCA subproblem at z^{k+1} gives $s^{k+1} \in \partial f_1(z^{k+1})$, $n^{k+1} \in N_C(z^{k+1})$, and

$$s^{k+1} + n^{k+1} = g_2^k. \tag{6}$$

Substituting $s^{k+1} = g_2^k - n^{k+1}$ from (6) and using $\langle n^{k+1},\ x^k - z^{k+1} \rangle \le 0$ (since $n^{k+1} \in N_C(z^{k+1})$ and $x^k \in C$):

$$f_1(x^k) \ge f_1(z^{k+1}) + \langle g_2^k,\ x^k - z^{k+1} \rangle + \tfrac{\mu_1}{2} \left\|x^k - z^{k+1}\right\|^2. \tag{7}$$

By μ_2-strong convexity of f_2 with subgradient $g_2^k \in \partial f_2(x^k)$:

$$f_2(x^k) \le f_2(z^{k+1}) + \langle g_2^k,\ x^k - z^{k+1} \rangle - \tfrac{\mu_2}{2} \left\|x^k - z^{k+1}\right\|^2. \tag{8}$$

Subtracting (8) from (7), the $\langle g_2^k,\ \cdot \rangle$ terms cancel: $h(x^k) - h(z^{k+1}) \ge \frac{\mu_1+\mu_2}{2} \left\|x^k - z^{k+1}\right\|^2$. □

Proof of Theorem 5.1. By greedy selection, $h(x^{k+1}) \le h(z^{k+1})$. Combining with Lemma B.1:

$$h(x^k) - h(x^{k+1}) \ge \frac{\mu_1 + \mu_2}{2} \left\|x^k - z^{k+1}\right\|^2.$$

Summing over $k = 0, \dots, N-1$ and telescoping:

$$\frac{\mu_1 + \mu_2}{2} \sum_{k=0}^{N-1} \left\|x^k - z^{k+1}\right\|^2 \le h(x^0) - h(x^N) \le h(x^0) - h^*.$$

Taking the minimum over k (at most the average) and the square root completes the proof. □

Remark B.1 (Oracle residual $\left\|x^k - z^{k+1}\right\|$)**.** *For the DCA oracle, $\left\|x^k - z^{k+1}\right\| = 0$ if and only if x^k is critical (a fixed point of the oracle subproblem). Thus $\left\|x^k - z^{k+1}\right\|$ is a quantitative residual for distance to criticality, and Theorem 5.1 gives the preserved ergodic $O(1/\sqrt{N})$ rate in exactly this residual.*

This rate statement concerns a finite-iteration optimality measure, whereas the asymptotic guarantee of the augmented method (Theorem 4.1) is almost-sure d-stationarity of accumulation points. There is no conflict: the base algorithm converges to criticality, while augmentation upgrades the limit notion to d-stationarity, which is strictly stronger in general (Remark 1.1).

C Proofs for Section 5.2: rate transfer

Proof of Proposition 5.2. By (G), $h(x^k) - h(z^{k+1}) \le h(x^k) - h(x^{k+1})$ for every k. Summing over $k = 0, \dots, K-1$ telescopes the right-hand side to $h(x^0) - h(x^K) \le \Delta_0$, since $x^K \in C$ and $h \ge h^*$ on C. The second bound holds because a minimum is at most an average. For the third, the summands are nonnegative along the run, so $\eta \cdot \#\{k < K : h(x^k) - h(z^{k+1}) \ge \eta\} \le \sum_{k=0}^{K-1} \big(h(x^k) - h(z^{k+1})\big) \le \Delta_0$. □

Proof of Theorem 5.3. Write $\delta_k := h(x^k) - h(z^{k+1})$, which is nonnegative by (T). Let $k^* \in \arg\min_{k<K} \delta_k$. By Proposition 5.2, $\delta_{k^*} \le \Delta_0/K$, and by (T) and monotonicity of φ,

$$\min_{k<K} \sigma(x^k, z^{k+1}) \le \sigma(x^{k^*}, z^{k^*+1}) \le \varphi(\delta_{k^*}) \le \varphi(\Delta_0/K).$$

For the second claim, let $k < K$ satisfy $\sigma(x^k, z^{k+1}) > \varphi(\eta)$. If $\delta_k \le \eta$, then (T) and monotonicity of φ give $\sigma(x^k, z^{k+1}) \le \varphi(\delta_k) \le \varphi(\eta)$, a contradiction. Hence $\delta_k > \eta$, and Proposition 5.2 bounds the number of such k by Δ_0/η. □

Proof of Theorem 5.4. Since (P3) also holds with any larger $q < 1/e$, we may assume $q > 0$. Fix $c > eq/(1-eq)$ depending only on q (for instance $c := 2eq/(1-eq)$, for which $\rho = (1+eq)/2$ below), and set

$$\rho := e(1+1/c)q, \qquad T := \lceil \Delta_0/\bar{\eta} \rceil, \qquad m := \max\left\{\lceil cT \rceil, \lceil \log(1/\zeta)/\log(1/\rho) \rceil\right\}, \qquad N := T+m.$$

We show that, with probability at least $1-\zeta$, some call $\mathcal{O}_C(x^k)$ with $k < N$ returns CERT. Since $0 < eq < 1$, the condition $c > eq/(1-eq)$ is equivalent to $e(1+1/c)q < 1$, so $\rho \in (0,1)$, and $m \ge 1$ because $\zeta < 1$. By (P2), every $k < N$ with $\ell^k = \mathrm{PROG}$ satisfies $h(x^k) - h(z^{k+1}) \ge \bar{\eta}$. By (O3) and the third bound of Proposition 5.2 with $\eta = \bar{\eta}$, there are at most $\Delta_0/\bar{\eta} \le T$ such k, on every realization. Let E be the event that no call $\mathcal{O}_C(x^k)$ with $k < N$ returns CERT. On E, each of these N calls returns PROG or FAIL, so at least $N - T = m$ of them return FAIL. Hence $E \subseteq \bigcup_S E_S$, where S ranges over the subsets of $\{0, \dots, N-1\}$ with m elements and $E_S := \{\ell^k = \mathrm{FAIL} \text{ for every } k \in S\}$.

Fix $S = \{k_1 < \cdots < k_m\}$. The event $\{\ell^{k_i} = \mathrm{FAIL},\ i < m\}$ belongs to $\mathcal{G}_{k_m-1}$, since every such k_i satisfies $k_i \le k_m - 1$. By the tower property and (P3),

$$\mathbb{P}(E_S) = \mathbb{E}\Big[\mathbf{1}_{\{\ell^{k_i} = \mathrm{FAIL},\ i<m\}}\, \mathbb{P}\big(\ell^{k_m} = \mathrm{FAIL} \mid \mathcal{G}_{k_m-1}\big)\Big] \le q\, \mathbb{P}\big(\ell^{k_i} = \mathrm{FAIL},\ i < m\big),$$

and iterating down through $k_{m-1}, \dots, k_1$ gives $\mathbb{P}(E_S) \le q^m$. A union bound over the $\binom{N}{m}$ sets S and the bound $\binom{N}{m} \le (eN/m)^m$ give $\mathbb{P}(E) \le (eNq/m)^m$. Since $m \ge cT$, we have $N = T + m \le (1 + 1/c)m$, and therefore

$$\mathbb{P}(E) \le \big(e(1+1/c)q\big)^m = \rho^m \le \zeta,$$

the last inequality because $m \ge \log(1/\zeta)/\log(1/\rho)$. Finally, $\lceil a \rceil \le a + 1$ and $\max\{a, b\} \le a + b$ for $a, b \ge 0$ give

$$N \le (1+c)T + \frac{\log(1/\zeta)}{\log(1/\rho)} + 2 \le (1+c)\frac{\Delta_0}{\bar\eta} + \frac{\log(1/\zeta)}{\log(1/\rho)} + c + 3 \le \kappa_q\Big(1 + \frac{\Delta_0}{\bar\eta} + \log(1/\zeta)\Big),$$

with $\kappa_q := \max\{c + 3,\ 1/\log(1/\rho)\}$, which depends only on q. □

C.1 Scope of rate transfer

Theorems 5.3 and 5.4 give sufficient conditions for rate transfer. Both use the exploration step only through (G), which bounds $h(x^{k+1})$ by $h(z^{k+1})$ but does not give $x^{k+1} = z^{k+1}$. The rate analysis of a base method therefore yields (T) along a run of Algorithm 2 only if its descent inequality holds at every point passed to the oracle, including points produced by exploration, and the same holds for (P3). In the cases below, Theorems 5.3 and 5.4 do not establish that the rate of the base method is preserved. We do not claim that it is lost.

Oracles that carry state. Condition (O1) fails when the output of the oracle at x depends on state carried across calls. Accelerated methods are the main example: FISTA [Beck and Teboulle, 2009b] takes a proximal gradient step from a point extrapolated from its two previous iterates, and its $O(1/k^2)$ analysis uses that each iterate is this step. Bundle methods build their model from cutting planes collected at earlier iterates, and quasi-Newton methods update a curvature matrix from differences of successive iterates and gradients. For such an oracle, $\mathcal{O}_C(x)$ is not determined by x, so Λ is not a function on C and $(\star)$ is not well posed. Theorem 5.3 still holds along any run that satisfies (T). The analysis of the base method, however, proves its descent inequality along its own trajectory and with its own state, and does not in general supply (T) at an iterate produced by exploration.

With a descent inequality proved at every point of C, condition (O1) yields (T) at such iterates. It is not necessary for (T). A prox-linear oracle fails (O1) when its inner solver is warm-started at the previous output and the point it returns depends on that output. If every returned point still satisfies (5), then for $\rho > L_c L_F$ Lemma D.1 gives $h(x^k) - h(z^{k+1}) \ge \alpha \left\|x^k - z^{k+1}\right\|^2$ with $\alpha = \frac{1}{2}(\rho - L_c L_F) > 0$. That decrease is the quadratic half of the hypothesis of Theorem 5.5, and it is (T) for $\sigma(x, z) = \|x - z\|$ and $\varphi(d) = \sqrt{d/\alpha}$, along every run.

For Theorem 5.4, (O1) is what turns a failure bound at every fixed point of C into (P3). The single-pass INGD oracle of Section 5.4 restarts its inner loop at each call, and Lemma E.2 bounds its failure probability at every fixed anchor. Lemma E.2 does not cover a variant that carries the vector m_j across calls.

Theorem 4.1 imposes no such condition. It allows randomized and adaptive oracles, provided each exploration sample $(v_{ex}^k, \hat t^k)$ is independent of the σ-algebra $\mathcal{F}_{k-1}$, as assumed in Remark 3.1.

Monotone accelerated methods are a boundary case. MFISTA [Beck and Teboulle, 2009a] takes as its next iterate the better, by function value, of the proximal gradient step and the previous iterate, and it retains the $O(1/k^2)$ rate of FISTA for convex problems. Its output depends on the extrapolation point carried from earlier iterates, and so it fails (O1). Whether its analysis admits the exploration point as a further candidate is open.

Iterates feasible only in the limit. Condition (O2) fails for methods whose iterates are feasible only in the limit, such as penalty, augmented Lagrangian and ADMM-type methods. Algorithm 2 requires $z^{k+1} \in C$, so such a method lies outside the scheme, and Theorem 4.1, which also assumes feasible oracle outputs, does not apply to it. Proposition 5.2 uses feasibility as well, through $h(x^K) \ge h^*$. Moreover, the analysis of such a method controls a merit function, such as a penalty or augmented Lagrangian function, rather than h, so (O3) need not hold. Whether an exploration step can be combined with such methods, for example through their merit function, is open.

Measures not calibrated to the decrement. Condition ($\star$) constrains the pair $(\mathcal{O}_C, \sigma)$. It is required at every $x \in C$ and, for a randomized oracle, for every realization of ξ. The residual $\|x - z\|$ of Sections 5.1 and 5.3 admits a calibration because Lemmas B.1 and D.1 bound Λ below by a multiple of its square. A measure such as the distance to a solution set can admit no calibration even when $\mathcal{O}_C$ satisfies (O1)–(O3), as the following example shows.

Remark C.1. *Let $n = 1$, $C = \mathbb{R}$ and $h(x) := \min\{|x|, 1\}$. Then h is 1-Lipschitz, directionally differentiable and bounded below, so Assumptions A1 and A2 hold, and $\arg\min h = \{0\}$. Let $\mathcal{O}_C(x) := 0$ if $|x| \le 1$, and $\mathcal{O}_C(x) := x$ if $|x| > 1$, where h is constant. Then $\mathcal{O}_C$ is a map on C, so (O1) holds, and (O2) holds since $C = \mathbb{R}$. Condition (O3) holds because $h(\mathcal{O}_C(x)) \le h(x)$ for every x: if $|x| \le 1$, then $\mathcal{O}_C(x) = 0$ and $h(0) = 0 \le h(x)$, and if $|x| > 1$, then $\mathcal{O}_C(x) = x$ and $h(\mathcal{O}_C(x)) = h(x)$. Let $\sigma(x, z) := \operatorname{dist}(x, \arg\min h) = |x|$. For every x with $|x| > 1$, $\Lambda(x) = 0$ and $\sigma(x, \mathcal{O}_C(x)) = |x|$. Hence ($\star$) would require $|x| \le \varphi(0)$ for every such x, which no $\varphi\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ satisfies. Thus ($\star$) is not implied by (O1)–(O3). The base method with this oracle has no rate in σ either when $|x^0| > 1$, so the example shows that ($\star$) restricts the pair $(\mathcal{O}_C, \sigma)$, not that augmentation loses a rate.*

Nonmonotone methods. Condition (O3) fails for base methods whose iterations may increase h, such as nonmonotone line-search methods. The non-ascent part of (T) is used by the third bound of Proposition 5.2 and by Theorem 5.3, whose φ is defined on $\mathbb{R}_{\ge 0}$, and Theorem 5.4 uses (O3) through the same bound. Greedy selection does not restore non-ascent of the oracle output: it gives $h(x^{k+1}) \le h(y^{k+1}) \le h(x^k)$, not $h(z^{k+1}) \le h(x^k)$.

D Proofs for Section 5.3: prox-linear-type oracles

This appendix supports Section 5.3. Throughout, z^{k+1} denotes the candidate produced by a prox-linear-type oracle at the current iterate x^k of Algorithm 2 (unconstrained $C = \mathbb{R}^n$ in (3)). As in Section 5.3, $c\colon \mathbb{R}^m \to \mathbb{R}$ is globally Lipschitz, with constant L_c, and $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable, with ∇F Lipschitz with constant L_F in the operator norm. We write $\ell^k(y) := F(x^k) + \nabla F(x^k)(y - x^k)$ and $m^k(y) := c(\ell^k(y)) + \frac{\rho}{2}\left\|y - x^k\right\|^2$. The descent calculation is classical for Lipschitz composite models [Drusvyatskiy and Paquette, 2019].

D.1 Lemma D.1 and Theorem 5.5

Lemma D.1 (Quadratic descent from model decrease)**.** *Under the Lipschitz assumptions on c and F and with $\rho > L_c L_F$, if* (5) *holds at iteration k, then*

$$h(x^k) - h(z^{k+1}) \;\ge\; \tfrac{1}{2}(\rho - L_c L_F)\left\|z^{k+1} - x^k\right\|^2. \tag{9}$$

Proof. From $m^k(z^{k+1}) \le m^k(x^k) = h(x^k)$,

$$h(x^k) - c(\ell^k(z^{k+1})) \;\ge\; \tfrac{\rho}{2}\left\|z^{k+1} - x^k\right\|^2. \tag{10}$$

By the Lipschitz Jacobian estimate for F [Drusvyatskiy and Paquette, 2019], $\left\|F(z^{k+1}) - \ell^k(z^{k+1})\right\| \le \frac{L_F}{2}\left\|z^{k+1} - x^k\right\|^2$, and Lipschitzness of c gives $c(\ell^k(z^{k+1})) \ge c(F(z^{k+1})) - L_c\left\|F(z^{k+1}) - \ell^k(z^{k+1})\right\| \ge h(z^{k+1}) - \frac{L_c L_F}{2}\left\|z^{k+1} - x^k\right\|^2$. Substituting into (10) yields (9). □

Proof of Theorem 5.5. By hypothesis, $h(z^{k+1}) \le h(x^k)$ and $\left\|x^k - z^{k+1}\right\| \le \sqrt{(h(x^k) - h(z^{k+1}))/\alpha}$ for every k, which is (T) with $\sigma(x, z) = \|x - z\|$ and $\varphi(d) = \sqrt{d/\alpha}$. Theorem 5.3 with $K = N$ gives the bound. When $\rho > L_c L_F$, Lemma D.1 shows that (5) implies the quadratic decrease in the hypothesis, with

$$h(x^k) - h(z^{k+1}) \ge \alpha\left\|x^k - z^{k+1}\right\|^2, \qquad \alpha = \tfrac{1}{2}(\rho - L_c L_F). \tag{QD}$$

□

Assumption 3 (PL-BDD)**.** *The sublevel set $\{x \in \mathbb{R}^n : h(x) \le h(x^0)\}$ is bounded.*

With $C = \mathbb{R}^n$, this is the level set of Remark 4.1, so $\{x^k\}$ is bounded and has an accumulation point.

Reading guide for the next theorems. The next three theorems identify what vanishing prox-linear residual $\left\|z^{k+1}-x^k\right\|$ implies at accumulation points under three inner-termination templates: DC, Clarke, and Mordukhovich. This is the same residual that appears in Lemma D.1 and Theorem 5.5, so the $O(1/\sqrt{N})$ bound is a rate statement in a concrete optimality metric. In the augmented algorithm, accumulation points are upgraded further to almost-sure d-stationarity (Theorem 4.1), which is strictly stronger in general (Remark 1.1; see also Remark B.1).

D.2 DC outer function

Assumption 4 (PL-DC1)**.** *The outer function $c\colon \mathbb{R}^m \to \mathbb{R}$ is DC: $c(y) = f(y) - g(y)$ with $f, g\colon \mathbb{R}^m \to \mathbb{R}$ convex and finite-valued, and c is globally Lipschitz with constant $L_c > 0$.*

Assumption 5 (PL-DC2)**.** *The inner map $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and ∇F is L_F-Lipschitz in operator norm, as in Lemma D.1.*

Assumption 6 (PL-DC3)**.** *The parameter satisfies $\rho > L_c L_F$.*

Inner termination (DC criticality of the model). Fix k and decompose $m^k = \varphi^k - \psi^k$ with $\varphi^k(y) := f(\ell^k(y)) + \frac{\rho}{2}\left\|y - x^k\right\|^2$ and $\psi^k(y) := g(\ell^k(y))$ (both convex). The oracle output z^{k+1} is *DC critical* for m^k if

$$\nabla F(x^k)^\top \xi^{k+1} + \rho(z^{k+1} - x^k) = \nabla F(x^k)^\top \eta^{k+1} \tag{11}$$

for some $\xi^{k+1} \in \partial f(\ell^k(z^{k+1}))$ and $\eta^{k+1} \in \partial g(\ell^k(z^{k+1}))$, and satisfies (5).

Theorem D.2 (DC criticality at accumulation points)**.** *Under Assumptions 4–6 and 3, suppose each oracle output z^{k+1} satisfies* (11) *and* (5)*. Then every accumulation point x^* of $\{x^k\}$ satisfies*

$$\nabla F(x^*)^\top \partial f(F(x^*)) \;\cap\; \nabla F(x^*)^\top \partial g(F(x^*)) \;\neq\; \emptyset, \tag{12}$$

i.e., DC criticality of $h = f \circ F - g \circ F$ at x^*.

Proof. Each z^{k+1} satisfies (5). Assumptions 4 and 5 give the Lipschitz hypotheses of Lemma D.1, and $\rho > L_c L_F$ by Assumption 6, so the lemma applies at every k. With greedy selection, $h(x^{k+1}) \le h(z^{k+1})$, and therefore $h(x^k) - h(x^{k+1}) \ge \frac{1}{2}(\rho - L_c L_F)\left\|z^{k+1} - x^k\right\|^2$ for every k. Since h is bounded below (Section 5.3), summing over k gives $\sum_k \left\|z^{k+1} - x^k\right\|^2 < \infty$, hence $\left\|z^{k+1} - x^k\right\| \to 0$. Also $h(x^{k+1}) \le h(x^k)$, so $h(x^k) \le h(x^0)$ for all k and $\{x^k\}$ lies in a compact set by Assumption 3; thus $\{x^k\}$ has accumulation points.

Let $x^{k_j} \to x^*$. Since $\left\|z^{k_j+1} - x^{k_j}\right\| \to 0$, also $z^{k_j+1} \to x^*$. Along k_j, $\nabla F(x^{k_j}) \to \nabla F(x^*)$. Set $y^j := \ell^{k_j}(z^{k_j+1}) = F(x^{k_j}) + \nabla F(x^{k_j})(z^{k_j+1} - x^{k_j}) \to F(x^*)$. For $\xi^{k_j+1} \in \partial f(y^j)$ and $\eta^{k_j+1} \in \partial g(y^j)$, the sequences $\{\xi^{k_j+1}\}$ and $\{\eta^{k_j+1}\}$ are bounded: for all large j, y^j lies in a fixed compact set, and convex finite f, g are Lipschitz on compacts, hence their subdifferentials are uniformly bounded there. Passing to a subsequence, $\xi^{k_j+1} \to \xi^*$ and $\eta^{k_j+1} \to \eta^*$, and outer semicontinuity of convex subdifferentials [Rockafellar, 1997, Theorem 24.4] yields $\xi^* \in \partial f(F(x^*))$ and $\eta^* \in \partial g(F(x^*))$. The proximal term $\rho(z^{k_j+1} - x^{k_j}) \to 0$. Passing to the limit in (11) gives $\nabla F(x^*)^\top \xi^* = \nabla F(x^*)^\top \eta^*$, which is (12). □

D.3 Clarke outer function

Assumption 7 (PL-C1)**.** *The function $c\colon \mathbb{R}^m \to \mathbb{R}$ is globally Lipschitz with constant $L_c > 0$ (hence locally Lipschitz and therefore Clarke subdifferentiable).*

Assumption 8 (PL-C2)**.** *The map F satisfies the same smoothness as in Lemma D.1, and $\nabla F(x)$ is surjective (full row rank m) for every $x \in \mathbb{R}^n$.*

Assumption 9 (PL-C3)**.** *$\rho > L_c L_F$, and $h = c \circ F$ is bounded below on $\mathbb{R}^n$ (equivalently, $h^* > -\infty$ in the notation of Section 5).*

Inner termination (Clarke stationarity of the model). The oracle output z^{k+1} satisfies

$$0 \in \partial^C\Big(c \circ \ell^k + \tfrac{\rho}{2}\left\|\cdot - x^k\right\|^2\Big)(z^{k+1}), \tag{13}$$

together with (5). Under surjectivity of $\nabla F(x^k)$, the Clarke chain rule [Clarke, 1990, Theorem 2.3.10] and sum rule [Clarke, 1990, Proposition 2.3.3] expand (13) to

$$0 \in \nabla F(x^k)^\top \partial^C c(\ell^k(z^{k+1})) + \rho(z^{k+1} - x^k). \tag{14}$$

Theorem D.3 (Clarke stationarity at accumulation points)**.** *Under Assumptions 7–9 and 3, suppose each z^{k+1} satisfies* (13) *and* (5)*. Then every accumulation point x^* of $\{x^k\}$ satisfies $0 \in \partial^C h(x^*)$, i.e., Clarke stationarity of $h = c \circ F$.*

Proof. Greedy monotonicity and Assumption 3 imply that $\{x^k\}$ is bounded, hence it has accumulation points. Let $x^{k_j} \to x^*$. The argument in the proof of Theorem D.2 applies under Assumptions 7–9 in place of Assumptions 4–6, so $\left\|z^{k_j+1} - x^{k_j}\right\| \to 0$, hence $z^{k_j+1} \to x^*$. Choose $v^{k_j+1} \in \partial^C c(\ell^{k_j}(z^{k_j+1}))$ with $\nabla F(x^{k_j})^\top v^{k_j+1} + \rho(z^{k_j+1} - x^{k_j}) = 0$. The Clarke subdifferential is locally bounded [Clarke, 1990, Proposition 2.1.2(a)]; pass to a subsequence with $v^{k_j+1} \to v^*$, and outer semicontinuity of $\partial^C c$ yields $v^* \in \partial^C c(F(x^*))$. Taking limits, $\nabla F(x^*)^\top v^* = 0$, i.e., $0 \in \nabla F(x^*)^\top \partial^C c(F(x^*))$. Under surjectivity, the Clarke chain rule for $c \circ F$ [Clarke, 1990, Theorem 2.3.10] gives $\partial^C h(x^*) = \nabla F(x^*)^\top \partial^C c(F(x^*))$ (equality), hence $0 \in \partial^C h(x^*)$. □

D.4 Mordukhovich (limiting) outer function

Inner termination (Mordukhovich stationarity of the model). Replace (13) by

$$0 \in \partial^L\Big(c \circ \ell^k + \tfrac{\rho}{2}\left\|\cdot - x^k\right\|^2\Big)(z^{k+1}), \tag{15}$$

with (5), where ∂^L denotes the limiting (Mordukhovich) subdifferential. Under surjectivity of $\nabla F(x^k)$, the limiting chain rule [Rockafellar and Wets, 2009, Exercise 10.7] and the smooth-sum rule for ∂^L [Rockafellar and Wets, 2009, Exercise 10.10] yield the expansion

$$0 \in \nabla F(x^k)^\top \partial^L c(\ell^k(z^{k+1})) + \rho(z^{k+1} - x^k). \tag{16}$$

Theorem D.4 (Mordukhovich stationarity at accumulation points)**.** *Under Assumptions 7–9 and 3, suppose each z^{k+1} satisfies* (15) *and* (5)*. Then every accumulation point x^* of $\{x^k\}$ satisfies $0 \in \partial^L h(x^*)$, i.e., Mordukhovich stationarity of $h = c \circ F$.*

Proof. Greedy monotonicity and Assumption 3 imply that $\{x^k\}$ is bounded, hence it has accumulation points. Let $x^{k_j} \to x^*$; the argument in the proof of Theorem D.2 applies under Assumptions 7–9 in place of Assumptions 4–6, so $\left\|z^{k_j+1} - x^{k_j}\right\| \to 0$, hence $z^{k_j+1} \to x^*$. Pick $v^{k_j+1} \in \partial^L c(\ell^{k_j}(z^{k_j+1}))$ with $\nabla F(x^{k_j})^\top v^{k_j+1} + \rho(z^{k_j+1} - x^{k_j}) = 0$. Since $\partial^L c(y) \subseteq \partial^C c(y)$ for locally Lipschitz c [Rockafellar and Wets, 2009, Theorems 8.49 and 9.13] and $\partial^C c$ is locally bounded, the v^{k_j+1} are bounded; pass to a convergent subsequence $v^{k_j+1} \to v^*$. Outer semicontinuity of $\partial^L c$ gives $v^* \in \partial^L c(F(x^*))$, and taking limits yields $0 \in \nabla F(x^*)^\top \partial^L c(F(x^*))$. Under surjectivity, $\partial^L h(x^*) = \nabla F(x^*)^\top \partial^L c(F(x^*))$ [Rockafellar and Wets, 2009, Exercise 10.7], hence $0 \in \partial^L h(x^*)$. □

Remark D.1 (Comparison)**.** *The descent step is common (Lemma D.1); only the subdifferential calculus in the limit and the final chain rule differ. Since $\partial^L c(y) \subseteq \partial^C c(y)$ always, Mordukhovich stationarity of h is stronger than Clarke stationarity, and* (15) *is a stronger inner requirement than* (13)*. Assumption 3 is imposed uniformly in Theorems D.2–D.4 so that greedy iterates admit accumulation points; it is not needed for the algebraic form of the limit along a convergent subsequence.*

E Proofs for Section 5.4: INGD

Throughout this section, $C = \mathbb{R}^n$. Hence the subspace of Definition 1.1 is $\mathbb{R}^n$, $D_C(x) = \mathbb{R}^n$ for every x, and $\bar{x}$ is d-stationary if and only if $h'(\bar{x};\, d) \geq 0$ for every $d \in \mathbb{R}^n$. In this section, L denotes the Lipschitz constant of h.

Let $h \in \mathcal{F}(\Delta, L)$ [Zhang et al., 2020, (2)]: h is L-Lipschitz on $\mathbb{R}^n$, Hadamard directionally differentiable at every point, and $h(x^0) - \inf h \leq \Delta$. Then h satisfies Assumptions A1 and A2, and $\Delta_0 \leq \Delta$. Every $g \in \partial^C h(y)$ satisfies $\|g\| \leq L$ [Clarke, 1990, Proposition 2.1.2(a)]. Following

Zhang et al. [2020, Assumption 1(a)], the method accesses h through a generalized-gradient oracle O, which, given $y, d \in \mathbb{R}^n$, returns $h(y)$ and a vector $g \in \partial^C h(y)$ with $\langle g, d\rangle = h'(y; d)$. We assume that O is deterministic and that the map $(y, d) \mapsto g$ is Borel measurable; the conditional expectations in the proof of Lemma E.2, as in that of Zhang et al. [2020, Lemma 13], require it.

The symbols of Zhang et al. [2020, Algorithm 1] denote other objects in this paper and are replaced as follows.

Zhang et al. [2020]	here	reason
x_t	x	one pass at a fixed anchor; t denotes step sizes
$x_{t,k}$	w_j	x^k denotes the iterates of Algorithm 2
$m_{t,k}$, $g_{t,k+1}$, $\beta_{t,k}$	m_j, g_{j+1}, β_j	k is the iteration index of Algorithm 2
$y_{t,k+1}$	u_{j+1}	y^{k+1} is the exploration candidate
K	J	K is the horizon in Proposition 5.2
γ	ζ	γ is the parameter of Algorithm 1
O	O	$\mathcal{O}_C$ is the oracle of Algorithm 2

Fix $\delta > 0$ and $\varepsilon > 0$, set $J := \lceil 48L^2/\varepsilon^2 \rceil$, and let $\xi := (\lambda_2, \dots, \lambda_{J+1})$, where $\lambda_2, \dots, \lambda_{J+1}$ are independent and uniform on $[0, 1]$. Given $x \in \mathbb{R}^n$, one call of the single-pass INGD oracle computes $\Psi(x, \xi) = (\mathcal{O}_C(x), \ell)$ as follows. If $L = 0$, return (x, CERT). Otherwise:

1. Set $m_1 := g$, where $(h(x), g) = \mathsf{O}(x, 0)$.
2. For $j = 1, \dots, J$:
   (a) if $\|m_j\| \le \varepsilon$, return (x, CERT);
   (b) set $w_j := x - \delta\, m_j / \|m_j\|$; if $h(w_j) - h(x) < -\frac{\delta}{4}\|m_j\|$, return (w_j, PROG);
   (c) set $u_{j+1} := x + \lambda_{j+1}(w_j - x)$, let g_{j+1} be the vector returned by $\mathsf{O}(u_{j+1}, -m_j)$, and set $m_{j+1} := \beta_j m_j + (1 - \beta_j) g_{j+1}$, where $\beta_j := (4L^2 - \|m_j\|^2)/(4L^2 + 2\|m_j\|^2)$.
3. Return (x, FAIL).

These are lines 4–17 of Zhang et al. [2020, Algorithm 1] at the anchor x, in the symbols above, with its three outcomes (termination, the descent test, exhaustion of the loop) labelled CERT, PROG and FAIL. There, the normalized step precedes the test of step (a); here the order is reversed, so that w_j is defined, since $\|m_j\| > \varepsilon > 0$ in step (b). The returned point and label are unchanged. If $L = 0$, the call returns before step 1 and makes no call to O. Otherwise a call makes at most $J + 1$ calls to O, one in step 1 and one per execution of step (c), and at most J further function evaluations, at the points w_j.

The map $\Psi\colon \mathbb{R}^n \times [0,1]^J \to \mathbb{R}^n \times \{\mathrm{CERT}, \mathrm{PROG}, \mathrm{FAIL}\}$ is jointly Borel measurable: it is composed of h, which is continuous, of O and of arithmetic operations, through finitely many branches on Borel sets. In Algorithm 2, the call at iteration k computes $(z^{k+1}, \ell^k) = \Psi(x^k, \xi^k)$, where $\xi^0, \xi^1, \dots$ are independent copies of ξ, independent of the exploration samples. Thus (O1) holds. Since x^0 is fixed and $\mathcal{G}_{k-1}$ is generated by functions of $(v_{\mathrm{ex}}^i, \hat{t}^i, \xi^i)_{i<k}$, the sample ξ^k is independent of $\mathcal{G}_{k-1}$. For the same reason, $(v_{\mathrm{ex}}^k, \hat{t}^k)$ is independent of the σ-algebra $\mathcal{F}_{k-1}$ of Remark 3.1, as the end of Section 4 requires.

Lemma E.1 (One call at a fixed anchor; adapted from the proof of Zhang et al. [2020, Theorem 8])**.** *Let $x \in \mathbb{R}^n$. For every realization of ξ, the following hold.*

(i) Every vector m_j computed by the call satisfies $\|m_j\| \le L$ and $m_j \in \partial_\delta h(x)$, and every β_j computed lies in $[1/2, 1]$.

(ii) If $\ell \in \{\mathrm{CERT}, \mathrm{FAIL}\}$, then $\mathcal{O}_C(x) = x$. If $\ell = \mathrm{PROG}$, then $h(x) - h(\mathcal{O}_C(x)) > \delta\varepsilon/4$.

(iii) If $\ell = \mathrm{CERT}$, then x is (δ, ε)-Goldstein stationary.

Proof. (i) If $L = 0$, the call returns before any m_j or β_j is computed. Now assume $L > 0$. We prove part (i) by induction on j. Since $m_1 \in \partial^C h(x)$ and $x \in B[x, \delta]$, we have $m_1 \in \partial_\delta h(x)$ and $\|m_1\| \le L$. Suppose that m_j satisfies (i) and that step (c) is executed at j. The function $s \mapsto (4L^2 - s)/(4L^2 + 2s)$ decreases from 1 at $s = 0$ to $1/2$ at $s = L^2$, so $\beta_j \in [1/2, 1]$. Since $\|w_j - x\| = \delta$, the point u_{j+1} lies on the segment $[x, w_j] \subseteq B[x, \delta]$, so $g_{j+1} \in \partial^C h(u_{j+1}) \subseteq$

$\partial_\delta h(x)$ and $\|g_{j+1}\| \le L$. Hence m_{j+1}, a convex combination of m_j and g_{j+1}, lies in the convex set $\partial_\delta h(x)$ and satisfies $\|m_{j+1}\| \le L$.

(ii) The first claim holds by construction. If the call returns PROG at step j, then step (a) did not return at j, so $\|m_j\| > \varepsilon$, and step (b) gives $h(w_j) - h(x) < -\frac{\delta}{4}\|m_j\| < -\delta\varepsilon/4$.

(iii) If $L = 0$, then h is constant, so $0 \in \partial^C h(x) \subseteq \partial_\delta h(x)$ and $\operatorname{dist}(0, \partial_\delta h(x)) = 0 \le \varepsilon$. If the call returns CERT at step j, then $\|m_j\| \le \varepsilon$, and $m_j \in \partial_\delta h(x)$ by (i), so $\operatorname{dist}(0, \partial_\delta h(x)) \le \varepsilon$. □

Lemma E.2 (Per-call failure bound; adapted from Zhang et al. [2020, Lemma 13])**.** *For every $x \in \mathbb{R}^n$, the probability that $\Psi(x, \xi)$ has label* FAIL *is at most* $1/16$*.*

Proof. If $L = 0$, the call returns CERT, so the probability is 0. Assume $L > 0$. Fix x. For the proof only, set $m_{j'} := 0$ for every $j' > j_0$ if the call returns at step j_0, so that $m_1, \dots, m_J$ are defined on every realization. Let $\mathcal{H}_1$ be the trivial σ-algebra and $\mathcal{H}_j := \sigma(\lambda_2, \dots, \lambda_j)$ for $j \ge 2$. Then m_j and w_j are $\mathcal{H}_j$-measurable, and λ_{j+1} is independent of $\mathcal{H}_j$. Let $D_j \in \mathcal{H}_j$ be the event that step (c) is executed at j. On D_j, $\|m_j\| > \varepsilon$ and $h(w_j) - h(x) \ge -\frac{\delta}{4}\|m_j\|$.

On D_j, $w_j - x$ is a positive multiple of $-m_j$, so the defining property of O and positive homogeneity of $h'(u_{j+1};\, \cdot)$ give $\langle g_{j+1},\, w_j - x\rangle = h'(u_{j+1};\, w_j - x)$. Since u_{j+1} is uniform on $[x, w_j]$ given $\mathcal{H}_j$, Zhang et al. [2020, Lemma 3] yields, on D_j,

$$\mathbb{E}\big[\langle g_{j+1},\, w_j - x\rangle \mid \mathcal{H}_j\big] = \int_0^1 h'\big(x + \lambda(w_j - x);\, w_j - x\big)\, d\lambda = h(w_j) - h(x).$$

Since $m_j = -(\|m_j\|/\delta)(w_j - x)$, it follows that, on D_j,

$$\mathbb{E}\big[\langle g_{j+1},\, m_j\rangle \mid \mathcal{H}_j\big] = -\frac{\|m_j\|}{\delta}\big(h(w_j) - h(x)\big) \le \frac{\|m_j\|^2}{4}.$$

Using $\|g_{j+1}\| \le L$, and that $\beta_j \in [0, 1]$ is $\mathcal{H}_j$-measurable, we obtain on D_j

$$\mathbb{E}\big[\,\|m_{j+1}\|^2 \mid \mathcal{H}_j\big] \le \chi(\beta_j),$$

where $a := \|m_j\|^2$ and $\chi(\beta) := \beta^2 a + (1-\beta)^2 L^2 + \frac{1}{2}\beta(1-\beta)a$. Expanding,

$$\chi(\beta) = \beta^2\big(L^2 + \tfrac{a}{2}\big) + \beta\big(\tfrac{a}{2} - 2L^2\big) + L^2.$$

The derivative $2\beta(L^2 + a/2) + a/2 - 2L^2$ vanishes at $\beta = (4L^2 - a)/(4L^2 + 2a)$, the value of step (c), and the second derivative $2L^2 + a$ is positive. The vertex value is

$$\chi(\beta_j) = L^2 - \frac{(4L^2 - a)^2}{16(L^2 + a/2)} = \frac{L^2 a - a^2/16}{L^2 + a/2}.$$

For $a \in [0, L^2]$,

$$\frac{L^2 a - a^2/16}{L^2 + a/2} \le \Big(1 - \frac{a}{3L^2}\Big)a,$$

since for $a > 0$ it rearranges to $a/6 + a/16 \ge a^2/(6L^2)$, and multiplying through by $48L^2/a$ gives $11L^2 \ge 8a$. Thus $\chi(\beta_j) \le \theta(a)$, where $\theta(s) := (1 - s/(3L^2))\, s$; the comparison uses $\|m_j\| \le L$ (Lemma E.1). Off D_j, $m_{j+1} = 0$. Since $\theta \ge 0$ on $[0, L^2]$, it follows that $\mathbb{E}[\|m_{j+1}\|^2 \mid \mathcal{H}_j] \le \theta(\|m_j\|^2)$ on every realization.

Let $v_j := \mathbb{E}[\|m_j\|^2]/L^2 \in [0, 1]$. Taking expectations and applying Jensen's inequality to the concave function θ gives $v_{j+1} \le v_j - v_j^2/3$. Since $s \mapsto s - s^2/3$ is increasing on $[0, 1]$, induction from $v_1 \le 1$ gives $v_j \le 3/(j+2)$: if $v_j \le 3/(j+2)$, then $v_{j+1} \le 3/(j+2) - 3/(j+2)^2 = 3(j+1)/(j+2)^2 \le 3/(j+3)$. In particular, $\mathbb{E}[\|m_J\|^2] \le 3L^2/(J+2) \le \varepsilon^2/16$, since $J \ge 48L^2/\varepsilon^2$. If the label is FAIL, step (a) did not return at $j = J$, so $\|m_J\| > \varepsilon$. By Markov's inequality, the label is FAIL with probability at most $\mathbb{E}[\|m_J\|^2]/\varepsilon^2 \le 1/16$. □

Proof of Theorem 5.6. Set $T := \lceil 4\Delta_0/(\delta\varepsilon)\rceil$ and $N := T + \max\{3T, \lceil 11\log(1/\zeta)\rceil\}$. We show that, with probability at least $1 - \zeta$, some call $\mathcal{O}_C(x^k)$ with $k < N$ returns CERT, and that the calls

$\mathcal{O}_C(x^0), \ldots, \mathcal{O}_C(x^{N-1})$ make at most $N(J+1)$ calls to O. Condition (O2) holds since $C = \mathbb{R}^n$, and (O1) was shown above. By Lemma E.1(ii), (O3) and (P2) hold for every realization, with $\bar{\eta} = \delta\varepsilon/4$. By Lemma E.1(iii), (P1) holds with (δ, ε)-Goldstein stationarity as the target criterion. By Lemma E.2, the call at any fixed x has label FAIL with probability at most $1/16$. Since ξ^k is independent of $\mathcal{G}_{k-1}$, x^k is $\mathcal{G}_{k-1}$-measurable, and Ψ is jointly measurable, Fubini's theorem gives $\mathbb{P}(\ell^k = \text{FAIL} \mid \mathcal{G}_{k-1}) \leq 1/16 \leq 1/4$ almost surely, which is (P3) with $q = 1/4 < 1/e$. Theorem 5.4 therefore applies. Its proof in Appendix C admits any $c > eq/(1-eq) = e/(4-e)$; we take $c = 3$, for which $\rho = e/3$. Since $\lceil \Delta_0/\bar{\eta} \rceil = T$ and $1/\log(3/e) < 11$, the horizon of that proof is at most N, and a longer horizon only enlarges the event that some call returns CERT. This proves the claim on CERT. Each call makes at most $J+1$ calls to O. If $\varepsilon \leq L$, then $J+1 \leq 48L^2/\varepsilon^2 + 2 \leq 50L^2/\varepsilon^2$. Moreover, $\lceil a \rceil \leq a+1$ and $\max\{a,b\} \leq a+b$ for $a, b \geq 0$ give $N \leq 4T + 11\log(1/\zeta) + 1 \leq 16\Delta_0/(\delta\varepsilon) + 11\log(1/\zeta) + 5$. Multiplying the two bounds gives $N(J+1) \leq 800\, L^2\varepsilon^{-2}\big(1 + \Delta_0\delta^{-1}\varepsilon^{-1} + \log(1/\zeta)\big)$. Since the certifying call has index $k < N$, the calls up to and including it are among $\mathcal{O}_C(x^0), \ldots, \mathcal{O}_C(x^{N-1})$, which proves the bound of the theorem. □

Remark E.1. *If $\varepsilon > L$, every point is (δ, ε)-Goldstein stationary, since $\partial^C h(x) \subseteq \partial_\delta h(x)$ contains a vector of norm at most L. If $L = 0$, the call returns* CERT *before step 1. If $L > 0$, then $J \geq 1$ and, by Lemma E.1(i), $\|m_1\| \leq L < \varepsilon$, so every call returns* CERT *at $j = 1$. The restriction $\varepsilon \leq L$ in Theorem 5.6 excludes only this case.*

Remark E.2. *Proposition 5.2 holds verbatim for the base method without exploration (Section 5.2), and the proof of Theorem 5.4 uses Algorithm 2 only through Proposition 5.2, (O3) and (P1)–(P3). Up to its first certificate, the base method with the single-pass INGD oracle performs the steps of INGD [Zhang et al., 2020, Algorithm 1] with its outer loop left unbounded: an exhausted pass is restarted at the same anchor with fresh randomness, and an accepted step moves the anchor to w_j. Hence the bound of Theorem 5.6 also holds for INGD in this form, with $\log(1/\zeta)$ entering additively rather than through the factor $\log(4\Delta/(\zeta\delta\varepsilon))$ of Zhang et al. [2020, Theorem 8].*

F A finite-time bound for the exploration step

This section makes two assumptions beyond Assumptions A1 and A2: $C = \mathbb{R}^n$, and h is Lipschitz on $\mathbb{R}^n$, not only locally Lipschitz. In this section, L denotes the Lipschitz constant of h, and we assume $L > 0$; if $L = 0$, then h is constant and every point is (δ, ε)-Goldstein stationary. Since $C = \mathbb{R}^n$, the subspace of Definition 1.1 is $\mathbb{R}^n$, and every oracle is feasible. Assumption A1 gives boundedness below: $h^* := \inf_{x \in C} h(x)$ is finite, and $\Delta_0 := h(x^0) - h^*$, where x^0 is the initial point of Algorithm 2. Assumption A2 is not used. The Goldstein δ-subdifferential $\partial_\delta h$ and (δ, ε)-Goldstein stationarity are as in Definition 5.1, and $\mathcal{S}_{\delta,\varepsilon}$ denotes the set of (δ, ε)-Goldstein stationary points.

At a point x, the methods of Zhang et al. [2020] and Davis et al. [2022] seek a vector of $\partial_\delta h(x)$ that either certifies that x is (δ, ε)-Goldstein stationary or yields a decrease of order $\delta\varepsilon$ along a step of length δ. The exploration step uses function values only and certifies nothing, but at every point outside $\mathcal{S}_{\delta,\varepsilon}$ it samples such a step with a probability bounded below by a quantity that is exponentially small in n.

Lemma F.1. *For every $x \in \mathbb{R}^n$, the set $\partial_\delta h(x)$ is nonempty, convex and compact, and it is contained in $B[0, L]$. If $x_i \to x$, $w_i \in \partial_\delta h(x_i)$ and $w_i \to w$, then $w \in \partial_\delta h(x)$. Consequently, $\mathcal{S}_{\delta,\varepsilon}$ is closed.*

Proof. For every y, the set $\partial^C h(y)$ is nonempty and every $g \in \partial^C h(y)$ satisfies $\|g\| \leq L$ [Clarke, 1990, Proposition 2.1.2(a)], so $\partial_\delta h(x)$ is a nonempty convex subset of $B[0, L]$. By Carathéodory's theorem, every $w \in \partial_\delta h(x)$ can be written as $w = \sum_{j=0}^{n} \lambda_j g_j$, with $(\lambda_j)_{j=0}^n$ in the unit simplex of $\mathbb{R}^{n+1}$, $g_j \in \partial^C h(y_j)$ and $y_j \in B[x, \delta]$. Let x_i, w_i and w be as in the statement, and write each $w_i = \sum_{j=0}^{n} \lambda_{ij} g_{ij}$, with $(\lambda_{ij})_{j=0}^n$ in the unit simplex, $g_{ij} \in \partial^C h(y_{ij})$ and $y_{ij} \in B[x_i, \delta]$. These sequences are bounded, so along a subsequence $\lambda_{ij} \to \lambda_j$, $y_{ij} \to y_j$ and $g_{ij} \to g_j$ for every j. Then $y_j \in B[x, \delta]$, $g_j \in \partial^C h(y_j)$ [Clarke, 1990, Proposition 2.1.5(b)], and $w = \sum_j \lambda_j g_j \in \partial_\delta h(x)$. Taking $x_i = x$ for every i shows that $\partial_\delta h(x)$ is closed, hence compact. Finally, let $x_i \in \mathcal{S}_{\delta,\varepsilon}$ with $x_i \to x$. By compactness, there are $w_i \in \partial_\delta h(x_i)$ with $\|w_i\| \leq \varepsilon$. Along a subsequence, $w_i \to w$, with $\|w\| \leq \varepsilon$ and $w \in \partial_\delta h(x)$, so $x \in \mathcal{S}_{\delta,\varepsilon}$. □

Lemma F.2. *Let $x \notin \mathcal{S}_{\delta,\varepsilon}$, let $\bar{g}$ be the element of minimal norm of $\partial_\delta h(x)$, and let $\hat{g} := \bar{g}/\|\bar{g}\|$. Then $\|\bar{g}\| > \varepsilon$, and the following hold.*

(i) For every $t \in [0, \delta]$, $h(x - t\hat{g}) \le h(x) - t\|\bar{g}\|$.

(ii) For every $v \in S^{n-1}$ with $\|v + \hat{g}\| \le \varepsilon/(2L)$ and every $t \in (0, \delta]$, $h(x+tv) < h(x) - t\varepsilon/2$.

Proof. By Lemma F.1, $\bar{g}$ exists and is unique, and $\|\bar{g}\| = \operatorname{dist}(0, \partial_\delta h(x)) > \varepsilon$. Since $\bar{g}$ is the projection of 0 onto the closed convex set $\partial_\delta h(x)$, every $w \in \partial_\delta h(x)$ satisfies $\langle w - \bar{g}, \bar{g}\rangle \ge 0$, that is, $\langle w, \hat{g}\rangle \ge \|\bar{g}\|$.

(i) The case $t = 0$ is trivial. Let $t \in (0, \delta]$. By Lebourg's mean value theorem [Clarke, 1990, Theorem 2.3.7], there are a point y on the open segment between x and $x - t\hat{g}$ and a vector $w \in \partial^C h(y)$ with $h(x - t\hat{g}) - h(x) = -t\langle w, \hat{g}\rangle$. Since $\|y - x\| < t \le \delta$, $w \in \partial_\delta h(x)$, and therefore $h(x - t\hat{g}) - h(x) \le -t\|\bar{g}\|$.

(ii) By L-Lipschitz continuity of h, (i) and $\|\bar{g}\| > \varepsilon$,

$$h(x + tv) \le h(x - t\hat{g}) + Lt\|v + \hat{g}\| \le h(x) - t\|\bar{g}\| + \frac{t\varepsilon}{2} < h(x) - \frac{t\varepsilon}{2}. \qquad \square$$

For $t = \delta$, the bound in (i) is Goldstein [1977, Lemma 3.1] applied to the direction $-\hat{g}$, with Goldstein's q equal to $\|\bar{g}\|$.

Lemma F.3. *Let v be uniformly distributed on S^{n-1}, and let $s > 0$. The probability $\pi_n(s) := \mathbb{P}(\|v - u\| \le s)$ is the same for every $u \in S^{n-1}$, and $\pi_n(s) \ge \big(s/(s+2)\big)^n$.*

Proof. The uniform distribution on S^{n-1} is invariant under orthogonal maps, which preserve distances and act transitively on S^{n-1}, so $\pi_n(s)$ does not depend on u. Let $u_1, \dots, u_M \in S^{n-1}$ satisfy $\|u_i - u_j\| > s$ for $i \ne j$. The open balls of radius $s/2$ centred at the u_i are pairwise disjoint and contained in the open ball of radius $1 + s/2$ centred at 0, so comparing volumes gives $M(s/2)^n \le (1 + s/2)^n$, that is, $M \le (1 + 2/s)^n$. Choose such points with M maximal. Then every $v \in S^{n-1}$ satisfies $\|v - u_i\| \le s$ for some i, so $1 \le \sum_{i=1}^M \mathbb{P}(\|v - u_i\| \le s) = M\pi_n(s)$, and $\pi_n(s) \ge (1 + 2/s)^{-n} = \big(s/(s+2)\big)^n$. $\square$

Proposition F.4. *Let $C = \mathbb{R}^n$, let h be L-Lipschitz on $\mathbb{R}^n$ with $L > 0$ and bounded below, and let $\{x^k\}$ be generated by Algorithm 2 with any oracle $\mathcal{O}_C$, randomized or adaptive, where in Algorithm 1 the direction v_{ex}^k is uniform on S^{n-1}, the step size $\hat{t}^k$ is uniform on $[0, r]$, and v_{ex}^k and $\hat{t}^k$ are independent of each other and of the σ-algebra $\mathcal{F}_{k-1}$ of Remark 3.1. Let $\delta := r$ and $\varepsilon > \gamma r$, and let $Z := \#\{k \ge 0 : x^k \notin \mathcal{S}_{\delta,\varepsilon}\}$ be the number of iterates that are not (δ, ε)-Goldstein stationary. Then*

$$\mathbb{E}[Z] \le \frac{8\Delta_0}{\delta\varepsilon\,\pi_n\big(\varepsilon/(2L)\big)} \le \frac{8\Delta_0}{\delta\varepsilon}\Big(1 + \frac{4L}{\varepsilon}\Big)^n.$$

In particular, almost surely all but finitely many iterates are (δ, ε)-Goldstein stationary, and the index $\nu := \inf\{k \ge 0 : x^k \in \mathcal{S}_{\delta,\varepsilon}\}$ of the first such iterate satisfies $\mathbb{E}[\nu] \le \mathbb{E}[Z]$.

These distributions satisfy Remark 3.1, so Theorem 4.1 applies to the same run when Assumption A2 holds.

Proof. By greedy selection, $h(x^{k+1}) \le h(y^{k+1})$, and $h(y^{k+1}) \le h(x^k)$ by the definition of t_{ex}^k in Algorithm 1, for every k. Let A be the set of $(x, v, t) \in \mathbb{R}^n \times S^{n-1} \times [0, r]$ with $h(x + tv) + \frac{\gamma}{2}t^2 < h(x)$ and $h(x + tv) \le h(x) - \delta\varepsilon/4$. It is a Borel set, since h is continuous. Let $A_k := \{(x^k, v_{\mathrm{ex}}^k, \hat{t}^k) \in A\}$. On A_k, the exploration step is accepted, since $C = \mathbb{R}^n$, so $y^{k+1} = x^k + \hat{t}^k v_{\mathrm{ex}}^k$ and $h(x^{k+1}) \le h(y^{k+1}) \le h(x^k) - \delta\varepsilon/4$. Since $h(x^k)$ is nonincreasing and $h \ge h^*$, on every realization,

$$\sum_{k \ge 0} \mathbf{1}_{A_k} \le \frac{4\Delta_0}{\delta\varepsilon}.$$

Let $x \notin \mathcal{S}_{\delta,\varepsilon}$, and let $\hat{g}$ be as in Lemma F.2. If $v \in S^{n-1}$ satisfies $\|v + \hat{g}\| \le \varepsilon/(2L)$ and $t \in [\delta/2, \delta]$, then Lemma F.2(ii) gives $h(x + tv) < h(x) - t\varepsilon/2 \le h(x) - \delta\varepsilon/4$, and, since $\gamma t \le \gamma r < \varepsilon$,

$$h(x + tv) + \frac{\gamma}{2}t^2 < h(x) - \frac{t}{2}(\varepsilon - \gamma t) < h(x),$$

so $(x, v, t) \in A$. Let v be uniform on S^{n-1} and t uniform on $[0, r] = [0, \delta]$, independent. By Lemma F.3 with $u = -\hat{g}$, $a(x) := \mathbb{P}\big((x, v, t) \in A\big) \ge \frac{1}{2}\pi_n\big(\varepsilon/(2L)\big) =: p$.

The pair $(v_{\mathrm{ex}}^k, \hat{t}^k)$ is independent of $\mathcal{F}_{k-1}$ and x^k is $\mathcal{F}_{k-1}$-measurable, so by Fubini's theorem a is measurable and $\mathbb{P}(A_k \mid \mathcal{F}_{k-1}) = a(x^k)$ almost surely. The event $\{x^k \notin \mathcal{S}_{\delta,\varepsilon}\}$ belongs to $\mathcal{F}_{k-1}$, because $\mathcal{S}_{\delta,\varepsilon}$ is closed (Lemma F.1), and $a(x^k) \ge p$ on it. Hence, for every $N \ge 1$,

$$p\,\mathbb{E}\Big[\sum_{k<N} \mathbf{1}\{x^k \notin \mathcal{S}_{\delta,\varepsilon}\}\Big] \le \sum_{k<N} \mathbb{E}\big[\mathbf{1}\{x^k \notin \mathcal{S}_{\delta,\varepsilon}\}\, a(x^k)\big] = \sum_{k<N} \mathbb{E}\big[\mathbf{1}\{x^k \notin \mathcal{S}_{\delta,\varepsilon}\}\, \mathbf{1}_{A_k}\big] \le \frac{4\Delta_0}{\delta\varepsilon}.$$

Letting $N \to \infty$, monotone convergence gives $\mathbb{E}[Z] \le 4\Delta_0/(p\,\delta\varepsilon)$, which is the first bound. The second follows from Lemma F.3 with $s = \varepsilon/(2L)$, for which $(s + 2)/s = 1 + 4L/\varepsilon$. Since $\mathbb{E}[Z] < \infty$, Z is finite almost surely. Finally, $x^0, \dots, x^{\nu-1} \notin \mathcal{S}_{\delta,\varepsilon}$, so $\nu \le Z$. □

Remark F.1. *Proposition F.4 bounds, in expectation, the number of iterates that are not (δ, ε)-Goldstein stationary, and hence the index of the first one that is. Algorithm 2 does not identify the stationary iterates: unlike the INGD oracle of Section 5.4, the exploration step computes no element of $\partial_\delta h$ and returns no certificate. The bound uses no property of the oracle. For $\mathcal{O}_C(x) = x$, each iteration makes at most one function evaluation (Remark 3.2), so the bound on $\mathbb{E}[\nu]$ also bounds the expected number of function evaluations made before the first (δ, ε)-Goldstein stationary iterate, apart from that of $h(x^0)$. The target is (δ, ε)-Goldstein stationarity, not d-stationarity, and the bound requires $C = \mathbb{R}^n$ and a global Lipschitz constant. We do not establish a finite-iteration bound under Assumptions A1 and A2 alone, or for a constrained set C.*

Remark F.2. *The proof of Proposition F.4 uses the uniform distribution of v_{ex}^k only through the bound $\mathbb{P}(\|v_{\mathrm{ex}}^k - u\| \le s) \ge \pi_n(s)$ for every $u \in S^{n-1}$, with $s = \varepsilon/(2L)$. For any other direction distribution $\mathcal{D}_v$ on S^{n-1}, fixed over the iterations, the same proof gives the first bound with $\pi_n(s)$ replaced by $\inf_{u \in S^{n-1}} \mathcal{D}_v(\{v : \|v - u\| \le s\})$ whenever this infimum is positive. If the infimum is zero, that argument gives no finite bound. This infimum is at most the average of $\mathcal{D}_v(\{v : \|v - u\| \le s\})$ over u uniform on S^{n-1}, which equals $\pi_n(s)$ by Fubini's theorem and Lemma F.3. Hence no fixed direction distribution improves the bound that this argument gives. We do not claim that the exponential dependence on n is necessary. Direction distributions that vary with k and adapt to the local geometry of h are left for future work.*

G Additional experimental details and discussion

The purpose of this appendix is to document, in one place, when the exploration step changes the outcome and whether those changes are favorable. Across all experiments, we compare a deterministic base oracle to its augmented version (oracle step plus one randomized exploration step), using the same initialization per instance and three augmentation seeds. We report win/tie/loss against the base solution using the signed gap

$$\Delta \;=\; h_{\mathrm{base}} - \mathrm{median}_s\, h_{\mathrm{aug},s},$$

with tolerance 10^{-12}. Positive Δ means augmentation is better. In addition to mean and median gap (wins), we report McNemar's exact one-sided p-value on discordant pairs (wins vs. losses), which directly tests the directional claim "augmentation helps more often than it hurts." This choice is deliberate: our data contain many ties in regimes where the base oracle is already strong, and tied-heavy settings can make signed-rank tests less aligned with this directional question.

Shared mechanics. All experiments use $C = \mathbb{R}^n$, so no projection is needed. The exploration step uses $\gamma = 1$, $r = 1$, accepts a candidate $x + \hat{t}v$ when

$$h(x + \hat{t}v) + \tfrac{\gamma}{2}\hat{t}^2 < h(x),$$

and then greedily keeps the better of the oracle point and accepted exploration point. The step size is sampled as $\hat{t} \sim \mathrm{Unif}([0, r])$ in all reported experiments. The sphere sampler uses $v \sim$

$\mathrm{Unif}(\mathbb{S}^{n-1})$. The gauss-axis sampler first chooses a coordinate index i uniformly, then samples $g \sim \mathcal{N}(0, I + (\mu^2 - 1)e_i e_i^\top)$, and sets $v = g/\|g\|$; at $\mu = 1$ this reduces to sphere sampling, while larger μ concentrates samples near coordinate axes. Whenever both samplers are run, we state explicitly which sampler is reported in the aggregate table.

G.1 D-stationarity diagnostics

The counts "# non-d-stat" in Table 2 flag terminal iterates that fail a numerical stationarity test run once per random instance (on the final iterate of each run). The three benchmarks use different objectives, so the stationarity checks differ; each is defined in the following paragraphs.

Trimmed lasso (exact DC inclusion test). We use the standard DC decomposition $h(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 - \lambda \operatorname{top}_k(|x|) = f(x) - g(x)$ with f, g convex. For such a model, x is d-stationary if and only if $\partial g(x) \subseteq \partial f(x)$ [de Oliveira, 2020], where $\partial f(x)$ and $\partial g(x)$ denote the convex subdifferentials of f and g at x. This inclusion is equivalent to a single nonnegative scalar being zero: writing $\operatorname{dist}_2(u, S) := \inf_{v \in S} \|u - v\|_2$ for nonempty closed convex $S \subseteq \mathbb{R}^d$ and $u \in \mathbb{R}^d$ (in particular $d = 1$ for each coordinate projection below),

$$\partial g(x) \subseteq \partial f(x) \quad \Longleftrightarrow \quad \sup_{\xi \in \partial g(x)} \operatorname{dist}_2\big(\xi,\, \partial f(x)\big) = 0, \tag{17}$$

where the supremum is taken over all subgradients $\xi \in \partial g(x)$ (the set $\partial g(x)$ is compact for this g). The diagnostic below evaluates the right-hand quantity in (17) in closed form: every $\xi \in \partial g(x)$ is determined by a top-k support pattern for the magnitudes $|x_i|$, and the distance to $\partial f(x)$ decouples across coordinates because $\partial f(x)$ is a Cartesian product of the one-dimensional sets in (18).

Write the residual gradient and the kth largest magnitude threshold as

$$r := A^\top(Ax - b) \in \mathbb{R}^n, \qquad \tau := |x|_{(k)},$$

where $|x|_{(k)}$ denotes the kth order statistic of $(|x_1|, \dots, |x_n|)$ (so $\operatorname{top}_k(|x|) = \sum_{j=1}^k |x|_{(j)}$). The convex subdifferential of $f(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$ is a Cartesian product of one-dimensional sets; equivalently, its ith coordinate projection is

$$(\partial f(x))_i \;=\; \begin{cases} \{\, r_i + \lambda \operatorname{sgn}(x_i) \,\}, & x_i \neq 0, \\ [\, r_i - \lambda,\; r_i + \lambda \,], & x_i = 0. \end{cases} \tag{18}$$

Fix a tie tolerance $\delta_{\mathrm{tie}} := 10^{-10}$ and define the top-k magnitude partition

$$\begin{aligned} S_{\mathrm{high}} &:= \{\, i : |x_i| > \tau + \delta_{\mathrm{tie}} \,\}, \\ S_{\mathrm{tie}} &:= \{\, i : \big||x_i| - \tau\big| \le \delta_{\mathrm{tie}} \,\}, \\ S_{\mathrm{low}} &:= \{\, i : |x_i| < \tau - \delta_{\mathrm{tie}} \,\}. \end{aligned}$$

When $|S_{\mathrm{tie}}| = 0$ (equivalently, the top-k magnitude set is uniquely determined up to the tolerance δ_{tie}, so there is no nontrivial ambiguity at the cutoff τ), $\partial g(x)$ is a singleton $\{\xi\}$. The unique $\xi \in \partial g(x)$ has coordinates $\xi_i = \lambda \operatorname{sgn}(x_i)$ if i lies among the k largest magnitudes and $\xi_i = 0$ otherwise. For this ξ, the coordinate projections of $\partial f(x)$ in (18) yield $\delta_i := \operatorname{dist}_2\big(\xi_i, (\partial f(x))_i\big)$, and

$$\sup_{\eta \in \partial g(x)} \operatorname{dist}_2\big(\eta,\, \partial f(x)\big) \;=\; \operatorname{dist}_2\big(\xi,\, \partial f(x)\big) \;=\; \sqrt{\sum_{i=1}^n \delta_i^2} \;=:\; \operatorname{gap}(x),$$

the second equality being the product geometry of (18). We declare *pass* if $\delta_i \le \varepsilon$ for every i, with

$$\varepsilon := 10^{-6}.$$

When $|S_{\mathrm{tie}}| > 0$ and the top-k support is genuinely ambiguous (equivalently, $|S_{\mathrm{high}}| < k < |S_{\mathrm{high}}| + |S_{\mathrm{tie}}|$), the set $\partial g(x)$ is a polytope generated by choosing which $n_{\mathrm{pick}} := k - |S_{\mathrm{high}}|$ indices inside S_{tie} join the mandatory indices S_{high} in the top-k magnitude set. On S_{high} and S_{low} the corresponding $\xi_i \in \partial g(x)$ are fixed at $\xi_i = \lambda \operatorname{sgn}(x_i)$ and $\xi_i = 0$, respectively. In particular, on S_{low} with $x_i = 0$ one has $\operatorname{dist}_2\big(0, (\partial f(x))_i\big) = \max\{0, |r_i| - \lambda\}$, whereas on S_{low} with $x_i \neq 0$

one has $\operatorname{dist}_2\big(0, (\partial f(x))_i\big) = |r_i + \lambda \operatorname{sgn}(x_i)|$ because $(\partial f(x))_i$ is the singleton $\{r_i + \lambda \operatorname{sgn}(x_i)\}$ (this is generally *not* the same as $|r_i|$). On S_{tie} each admissible vertex uses $p_i \in \{0, 1\}$ to indicate whether coordinate i is included in the top-k support; for each $i \in S_{\text{tie}}$ define the squared coordinate distance if p_i is chosen,

$$D_i(p)^2 := \begin{cases} \Big(\max\big\{\, 0,\ |r_i| - \lambda(1-p) \,\big\}\Big)^2, & x_i = 0,\ p \in \{0,1\}, \\ \Big(\lambda \operatorname{sgn}(x_i)(p-1) - r_i\Big)^2, & x_i \neq 0,\ p \in \{0,1\}, \end{cases}$$

where the lower branch is the squared distance from $\xi_i = \lambda \operatorname{sgn}(x_i)\, p$ to the singleton $(\partial f(x))_i = \{r_i + \lambda \operatorname{sgn}(x_i)\}$. Because $(\partial f(x))_i$ is an interval (or a point) independently across i, for any $\xi \in \partial g(x)$ compatible with tie indicators $(p_i)_{i \in S_{\text{tie}}}$ one has

$$\operatorname{dist}_2\big(\xi,\, \partial f(x)\big)^2 \;=\; \sum_{i=1}^{n} D_i(p_i)^2,$$

with the understanding that p_i is fixed on $S_{\text{high}} \cup S_{\text{low}}$ by the forced pattern ($p_i = 1$ on S_{high}, $p_i = 0$ on S_{low}) and only the tied coordinates vary. Thus maximizing $\operatorname{dist}_2(\xi, \partial f(x))$ over $\xi \in \partial g(x)$ reduces to maximizing this separable squared distance over the tie polytope. The maximizer is attained at a $0/1$ vertex; it is found by the greedy rule $M_i := D_i(1)^2 - D_i(0)^2$, sort $\{M_i\}_{i \in S_{\text{tie}}}$ in decreasing order, set $p_i = 1$ on the first n_{pick} indices and $p_i = 0$ on the remaining tied indices. With (p_i) fixed, let $\xi \in \partial g(x)$ be the corresponding vertex, set δ_i on S_{tie} to $\sqrt{D_i(p_i)^2}$, combine with the already-fixed coordinates on S_{high} and S_{low}, and define

$$\operatorname{gap}(x) := \sqrt{\sum_{i=1}^{n} \delta_i^2} \;=\; \sup_{\eta \in \partial g(x)} \operatorname{dist}_2\big(\eta,\, \partial f(x)\big).$$

The same pass criterion $\delta_i \le \varepsilon$ coordinatewise with $\varepsilon = 10^{-6}$ certifies $\operatorname{gap}(x) = 0$ up to floating-point tolerance, hence (17).

Least trimmed squares (exact smooth-f DC test). Here $h(x) = \frac{1}{2}\|Ax - b\|^2 - \frac{1}{2}\operatorname{top}_q((Ax - b)^2) = f(x) - g(x)$ with f smooth convex and g convex. In this template, d-stationarity is equivalent to g being differentiable at x and satisfying $\nabla f(x) = \nabla g(x)$: if g is not differentiable, $\partial g(x)$ is not a singleton and cannot be contained in the singleton $\partial f(x) = \{\nabla f(x)\}$. We form squared residuals r_i^2, partition indices into above-, at-, and below-threshold groups relative to the top-q cutoff, and set $\nabla f(x) = A^\top r$. If there is no residual tie, the top-q active set is unique, $\nabla g(x)$ is explicit, and the gap is $\|\nabla f(x) - \nabla g(x)\|_2$. If there are ties, we first test one admissible top-q completion; if it already violates gradient match beyond 10^{-6}, the point is declared non-d-stationary immediately. Otherwise we measure nondifferentiability by the maximum, over admissible completions T of the tied block, of $\|\nabla g_T(x) - \nabla g_{T_{\text{ref}}}(x)\|_2$, where $\nabla g_T(x) = A^\top w$ with $w_i = r_i$ on T and $w_i = 0$ off T. Exhaustive enumeration is used when the number of completions is at most 8192; for larger tie sets a deterministic pair of extreme completions plus 256 random completions (estimating the witness). The acceptance tolerance is 10^{-6} on the resulting gap.

ReLU regression (exact DC inclusion check). For ReLU regression we now use an explicit DC decomposition and test d-stationarity through convex-subdifferential inclusion. With

$$h(x) = \tfrac{1}{2}\sum_{i=1}^{m}\big(\max\{0, a_i^\top x\} - b_i\big)^2,$$

define $h = f - g$ by

$$f(x) := \tfrac{1}{2}\sum_{i=1}^{m}\big(\max\{0, a_i^\top x\}\big)^2 + \sum_{i=1}^{m} \alpha_i \max\{0, a_i^\top x\},$$

$$g(x) := \sum_{i=1}^{m} \beta_i \max\{0, a_i^\top x\},$$

where $\beta_i := b_i$ for $b_i > 0$ and 0 otherwise, and $\alpha_i := -b_i$ for $b_i < 0$ and 0 otherwise. Then $\alpha_i \beta_i = 0$ for all i, and d-stationarity is the DC condition $\partial g(x) \subseteq \partial f(x)$.

Write $z := Ax$, and use the numerical kink bandwidth

$$\tau_{\mathrm{kink}}(x) := 10^{-6} \max\{\, 1,\ \|z\|_\infty \,\}.$$

Set $I_+(x) := \{i : z_i > \tau_{\mathrm{kink}}(x)\}$ and $I_0(x) := \{i : |z_i| \le \tau_{\mathrm{kink}}(x)\}$, and split $I_0(x) = I_{0,\alpha}(x) \sqcup I_{0,\beta}(x)$ with

$$I_{0,\alpha}(x) := \{i \in I_0(x) : \alpha_i > 0\}, \qquad I_{0,\beta}(x) := \{i \in I_0(x) : \beta_i > 0\}.$$

Define

$$\begin{aligned} g_0(x) &:= \sum_{i \in I_+(x)} \beta_i a_i, \\ f_0(x) &:= \sum_{i \in I_+(x)} \big(z_i + \alpha_i\big) a_i, \\ D(x) &:= g_0(x) - f_0(x), \end{aligned}$$

and generator matrices

$$G_\beta(x) := \big[\beta_i a_i\big]_{i \in I_{0,\beta}(x)}, \qquad F_\alpha(x) := \big[\alpha_i a_i\big]_{i \in I_{0,\alpha}(x)}.$$

Then $\partial g(x) \subseteq \partial f(x)$ is equivalent to

$$D(x) + G_\beta(x)\lambda \in \big\{F_\alpha(x)\mu : \mu \in [0,1]^{|I_{0,\alpha}(x)|}\big\} \quad \forall \lambda \in [0,1]^{|I_{0,\beta}(x)|}.$$

The implemented diagnostic evaluates the exact vertex form of this condition:

$$\Gamma_{\mathrm{ReLU}}(x) := \max_{\lambda \in \{0,1\}^{|I_{0,\beta}(x)|}} \min_{\mu \in [0,1]^{|I_{0,\alpha}(x)|}} \big\| D(x) + G_\beta(x)\lambda - F_\alpha(x)\mu \big\|_2.$$

Each inner problem is a box-constrained least-squares subproblem, and the outer maximum is over binary vertices. We declare *pass* iff

$$\Gamma_{\mathrm{ReLU}}(x) \le 10^{-6}.$$

When $|I_{0,\beta}(x)|$ exceeds the implementation cap (set to 16), the run is marked non-d-stationary for reporting rather than using an uncontrolled approximation.

G.2 Trimmed Lasso: Resolving the Failure Mode

The trimmed-lasso objective

$$h(x) = \tfrac{1}{2}\|Ax - b\|^2 + \lambda\big(\|x\|_1 - \mathrm{top}_k|x|\big)$$

is the regime where plain DCA is known to converge to critical points that can fail d-stationarity. We focus on $\lambda = 1$ with $m = 50$, $n = 100$, $k = 5$, noise std. 0.1, and $N = 5000$ outer iterations. Both samplers are run; the aggregate table reports gauss-axis ($\mu = 300$).

λ	n	win/tie/loss	mean gap	median gap (wins)	McNemar p	# non-d-stat (base/aug)
1.00	100	99/1/0	+2.94e-01	+2.69e-01	1.58e-30	99/4

The central empirical message is clear: DCA fails the d-stationarity diagnostic on $99/100$ instances, while augmentation with gauss-axis reduces this to $4/100$. At the objective level, augmentation wins on 99 instances, ties on one, and never loses. McNemar's exact one-sided test gives $p \approx 1.6 \times 10^{-30}$.

Sphere runs were still checked for completeness, but we do not plot them here because they add little visual information in this focused regime: the sphere variant accepts no exploration moves and is practically indistinguishable from DCA. Instead, we show only the gauss-axis plots in Figures 2 and 3.

The reason this happens is structural. In high dimension, a random sphere direction has typical ℓ_1 mass of order $\sqrt{n}$ after ℓ_2 normalization. For trimmed lasso, the nonconvex sparsity term is

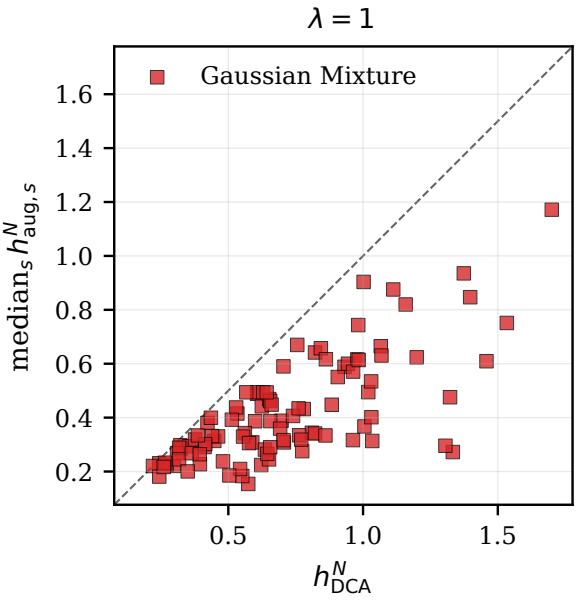


Figure 2: Trimmed lasso scatter at $\lambda = 1$. Horizontal axis: final DCA objective. Vertical axis: median augmented objective over three seeds. Points below the diagonal favor augmentation. (Sphere sampler omitted from display: in this regime it produces no accepted exploration moves and overlays DCA.)

driven by coordinate-level behavior, so this "diffuse" direction tends to spread motion over many coordinates and rarely aligns with the sparse descent geometry near stuck points. The acceptance test then almost never fires in finite budgets. The axis-biased Gaussian mixture is designed to correct exactly this mismatch: each sampled direction is still absolutely continuous on the sphere and keeps full support (so the theorem assumptions are unchanged), but it places much larger probability on directions close to coordinate axes, which are precisely the directions that can unlock descent for this ℓ_1-structured objective. Empirically, this sampler-design change is the difference between inert exploration and systematic improvement.

G.3 Least Trimmed Squares: Non-Disruptive Behavior with Occasional Escapes

For least trimmed squares,

$$h(x) = \tfrac{1}{2}\|Ax - b\|^2 - \tfrac{1}{2}\,\text{top}_q((Ax - b)^2),$$

we focus on $\sigma_{\text{clean}} = 4$ with $m = 100$, $n = 50$, $q = 10$, outlier std. 10, and $N = 5000$. For the focused LTS report, exploration directions are sampled from the sphere distribution $v \sim \text{Unif}(\mathbb{S}^{n-1})$.

σ_{clean}	n	win/tie/loss	mean gap	median gap (wins)	McNemar p	# non-d-stat (base/aug)
4.00	100	9/91/0	+3.96e+00	+5.22e+01	1.95e-03	0/0

This experiment supports a different part of the story. Here the base DCA oracle is already reliable in first-order terms: both base and augmented runs have zero d-stationarity failures in the focused setting. Augmentation is therefore expected to be mostly neutral, and that is exactly what we observe (91 ties). Yet on the discordant subset, all outcomes favor augmentation (9 wins, 0 losses), yielding $p \approx 2.0 \times 10^{-3}$ by McNemar's exact test. In short, augmentation is non-disruptive where the oracle already behaves well, but can still find a better basin when one is nearby.

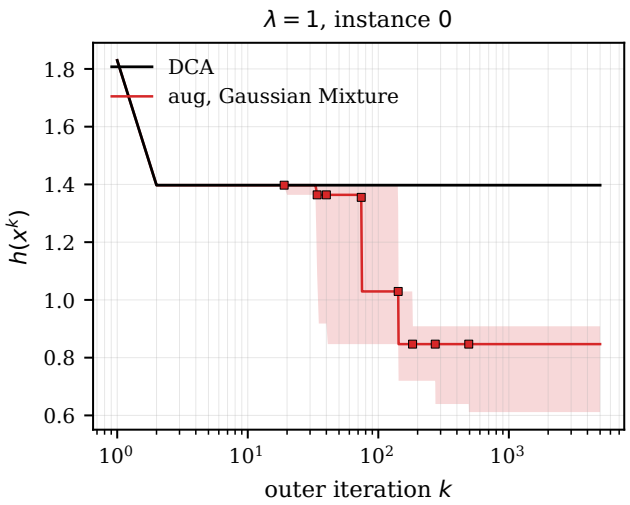


Figure 3: Trimmed lasso trajectory on a representative instance. Black: DCA. Colored band/curve: min/median/max over augmented seeds. The Gaussian-mixture sampler breaks away from the DCA plateau. (Sphere sampler omitted from display: trajectory essentially overlays DCA.)

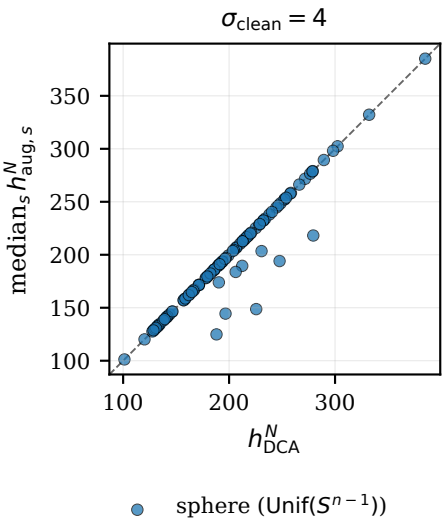


Figure 4: LTS scatter at $\sigma_{\text{clean}} = 4$ (sphere reporting). Most points lie on the diagonal (ties), with a smaller set below (augmentation wins).

G.4 ReLU Regression: Transfer Beyond DC Oracles

The ReLU experiment is intentionally outside the DC setup:

$$h(x) = \tfrac{1}{2} \sum_i \big(\text{relu}(a_i^\top x) - b_i\big)^2,$$

optimized by a prox-linear oracle ($\rho_{\text{prox}} = 0.1$), then augmented by the same exploration mechanism used elsewhere. We aggregate two focused settings, $(q, \rho_b) \in \{(0.2, 2), (0.4, 2)\}$, with 100 instances per setting (200 paired comparisons total), $m = 200$, $n = 50$, and $N_{\text{outer}} = 1000$. For the focused ReLU report, exploration directions are likewise sampled as $v \sim \text{Unif}(\mathbb{S}^{n-1})$.

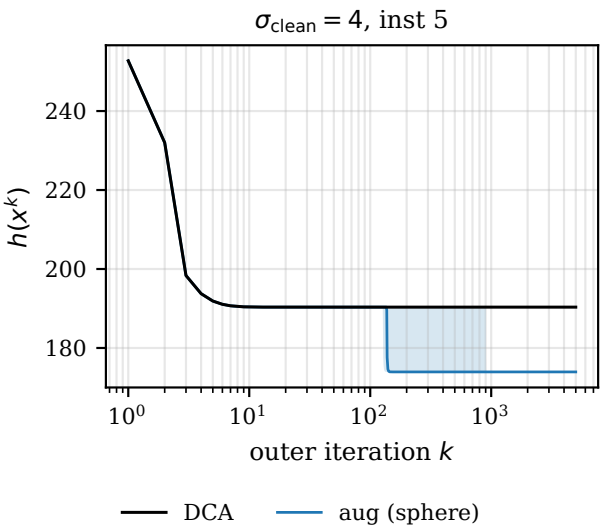


Figure 5: LTS representative trajectory where augmentation improves over the DCA plateau.

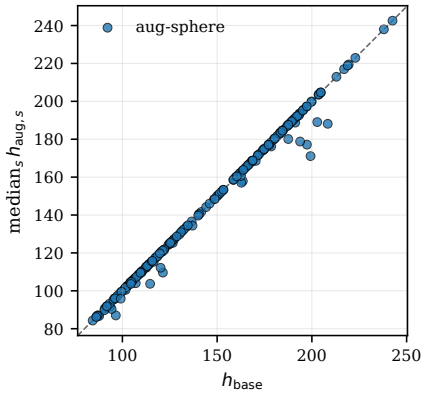


Figure 6: ReLU scatter (sphere reporting). Most pairs tie, with a smaller strict-win set.

setting	n	win/tie/loss	mean gap	median gap (wins)	McNemar p	# non-d-stat (base/aug)
focused sweep	200	44/156/0	+9.58e-01	+1.26e+00	5.68e-14	0/0

As in LTS, augmentation is largely neutral but never harmful in the focused report: 44 wins, 156 ties, 0 losses, with zero d-stationarity failures for both base and augmented methods. McNemar's exact one-sided test gives $p \approx 5.7 \times 10^{-14}$. Since the inner oracle here is prox-linear (not DCA), this result is direct empirical evidence for the paper's oracle-agnostic claim.

G.5 Cross-Experiment Interpretation

Taken together, the three experiments show a consistent pattern with distinct roles. Trimmed lasso demonstrates the failure mode and its correction: augmentation converts a high non-d-stationarity regime into a mostly d-stationary one, with strong objective improvements. LTS and ReLU demonstrate the complementary regime: when the base oracle is already first-order sound, augmentation is mostly neutral, never harmful in the focused runs, and occasionally beneficial.

This is the practical reading of the theoretical separation in the paper. Oracle choice controls local descent structure and rates; exploration controls asymptotic d-stationarity behavior in an oracle-

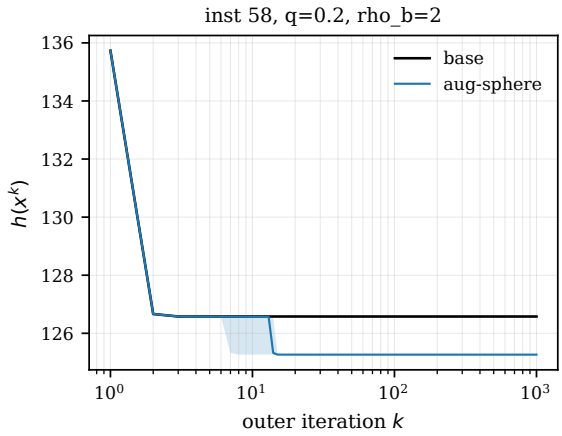


Figure 7: ReLU representative trajectory (base prox-linear vs. augmented seeds).

agnostic way. The experiments were designed so that each benchmark contributes one part of that statement, and the combined suite supports it without requiring any algorithm-specific retuning of the exploration step.

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 and introduction explicitly state the proposed random exploration augmentation, the almost-sure d-stationarity guarantee, and the accompanying rate-preservation claims. The scope is limited to the assumptions and settings formalized in the manuscript.

   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: We have included a "Limitations" section which discusses the limitation of our work.

   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 draft states assumptions A1–A3 and provides full theorem statements and proofs in the main text for the central convergence results. Lemmas and theorems are numbered and cross-referenced.

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: The experimental setting is described in detail in section G.

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: [No]

   Justification: The manuscript's central contribution is theoretical and does not hinge on empirical code/data artifacts. This checklist item is therefore not applicable to the main claims.

   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 full details of the experimental setup are in section G.

   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 paper provides statistical significance results using McNemar's test in section G.

   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 experiments were run on a laptop (CPU setting) and each focused run completed within several minutes; this compute scale is sufficient for reproducing the reported results.

   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 work is methodological optimization research and does not involve human participants or sensitive personal data. The authors commit to complying with NeurIPS ethical standards in preparation, reporting, and release practices.

   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 submission is foundational optimization theory without a direct deployment pathway discussed in the current draft. A broader-impact discussion can be added in a revised version if required by venue policy.

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 manuscript does not release high-risk models or scraped datasets. Hence safeguards for such releases are not applicable.

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: No third-party data/model asset release is central to the paper's theoretical claims in the current submission version. Therefore this item is not applicable.

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 current draft does not introduce or release a new dataset/model asset as a core contribution. This question is therefore not applicable.

    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 work does not involve crowdsourcing or human-subject studies. As such, this item does not apply.

    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: No human-subject data collection or experimentation is conducted in this research. Therefore IRB-related disclosure is not applicable.

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 are not part of the proposed optimization method, theoretical analysis, or experimental methodology. Consequently, declaration is not required under the stated policy.

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.