

Randomization Beats Gradient Descent in the Plane

Xiaoyu Li

University of New South Wales

xiaoyu.li2@unsw.edu.au

Abstract

We study the number of exact gradient queries required to find an ε -stationary point of an L -smooth, bounded-below function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ when function values are unavailable. If $f(x_0) - \inf f \leq \Delta$ and $q = \sqrt{\Delta L}/\varepsilon$, gradient descent uses $O(q^2)$ queries. We give a randomized gradient-only algorithm that returns a queried point X with $\|\nabla f(X)\| \leq \varepsilon$ with probability at least $1 - \rho$ after

$$O\left(q^{4/3} \log^{1/3} \frac{2}{\rho}\right)$$

queries. The argument has one central device: a single stratified line integral estimates every relative function value used on a planar cut. A maximal-prefix inequality avoids a union cost over the marked points, and a robust gradient-flow trap absorbs the remaining comparison error across geometrically shrinking scales. We also determine, up to universal constants, the complete accuracy–confidence law for the underlying Lipschitz integration problem; in particular, its $2/3$ accuracy exponent is unimprovable even for one endpoint. Finally, the exact gradient at the output certifies success. Independent restarts therefore give a Las Vegas algorithm with zero error and expected $O(q^{4/3})$ queries. The optimal planar exponent remains between 1 and $4/3$.

1 Introduction

Why does gradient descent pay a quadratic price for stationarity when the domain is only the plane? The usual argument is exact but dimension-blind. If f has L -Lipschitz gradient and $\|\nabla f(x)\| > \varepsilon$, then one gradient step of length $1/L$ decreases the value by more than $\varepsilon^2/(2L)$. An initial gap of size Δ can finance at most $2\Delta L/\varepsilon^2$ such steps. With

$$q := \frac{\sqrt{\Delta L}}{\varepsilon}, \tag{1.1}$$

the resulting query bound is $O(q^2)$. Nothing in this calculation distinguishes a curve in $\mathbb{R}^2$ from one in a space whose dimension grows with $1/\varepsilon$.

The distinction matters. In sufficiently high dimension the quadratic dependence is minimax-optimal for randomized first-order methods (Carmon et al., 2020). In one dimension, by contrast, randomization gives a rate linear in q , and this is sharp (Chewi et al., 2023). The plane is the first case in which this order disappears: level sets can surround points, gradient trajectories can turn, and a cut can separate a region without being totally ordered. It was therefore natural to ask whether the quadratic lower bound already returns in dimension two.

There is also a reason to expect the opposite. With function values, planar gradient-flow trapping can repeatedly cut a rectangle while preserving a certificate that a stationary point remains inside (Bubeck and Mikulincer, 2020; Hollender and Zampetakis, 2025). The obstacle for a gradient-only

method is not the geometry of the cut but the information it consumes: the trap asks which points on the cut lie below the current anchor, while the oracle never reports a value.

The argument is organized around a single principle:

one randomized line integral should pay for every comparison on a planar cut.

Indeed, gradients determine potential differences along a path. If a path of length H is divided into k equal strata and the tangential derivative is sampled once in each stratum, the integration error has root-mean-square size

$$\frac{LH^2}{k^{3/2}}, \quad (1.2)$$

rather than the deterministic scale LH^2/k . At the first trapping scale, the square has side $H \asymp \Delta/\varepsilon$ and the comparisons only need accuracy of order εH . Thus

$$\frac{LH^2}{\varepsilon H} \asymp \frac{\Delta L}{\varepsilon^2} = q^2, \quad k \asymp (q^2)^{2/3} = q^{4/3}. \quad (1.3)$$

Later rectangles are smaller and can be sampled more cheaply. The complete proof makes this scale calculation uniform over every comparison, preserves strict boundary inequalities under approximation, and sums both query costs and failure probabilities over the adaptive recursion.

Our main result is the following. For every failure probability $\rho \in (0, 1/2)$, a randomized adaptive algorithm using exact gradients and no function values returns a queried ε -stationary point after

$$O\left(q^{4/3} \log^{1/3} \frac{2}{\rho}\right) \quad (1.4)$$

queries. The assumptions are only global L -smoothness and the valid gap bound $f(x_0) - \inf f \leq \Delta$; queries may be adaptive and nonlocal. At constant confidence the rate is $O(q^{4/3})$, which rules out a universal quadratic randomized lower bound in the planar model.

The proof has three parts. First, a maximal martingale inequality controls all marked prefixes of one polygonal path in a single event. This is the point at which one tour replaces many unrelated value estimates. Second, we make the planar trap stable under additive comparison error. A conservative threshold protects the inherited boundary, and a one-dimensional net-to-face estimate protects the new cut. Third, the spatial scale, the reachability loss, and the failure probability are all allocated geometrically. The last allocation is important: a uniform division of ρ over the rounds would create an artificial $\log \log q$ term.

The integration primitive has an independent sharp statement. Let $N_{\text{int}}(L, H, \eta, \rho)$ be the smallest number of point queries needed to integrate every L -Lipschitz function on $[0, H]$ to error η with failure probability at most ρ . In the nontrivial regime,

$$N_{\text{int}}(L, H, \eta, \rho) = \Theta\left(1 + \min\left\{\frac{LH^2}{\eta}, \left(\frac{LH^2}{\eta}\right)^{2/3} \log^{1/3} \frac{1}{\rho}\right\}\right). \quad (1.5)$$

We give a self-contained fixed-confidence lower bound by hiding independent signs under disjoint triangular tents, and obtain the full confidence law from the sharp probabilistic integration theorem of [Kunsch and Rudolf \(2019\)](#). This result identifies the exact limit of improving the estimator inside a single cut. A better planar exponent, if it exists, must reuse information across scales or replace the trapping architecture.

There is one further consequence of exact access. The last query reveals the output gradient itself. A failed trial is therefore observable rather than merely unlikely. Restarting with fresh randomness yields a Las Vegas algorithm that never returns an incorrect point, terminates almost surely, and uses $O(q^{4/3})$ queries in expectation. Its number of trials has a geometric tail.

Relation to randomized integration. The $k^{-3/2}$ error in (1.2) is classical for stratified integration of Lipschitz functions (Yue and Mao, 1999; Carpentier and Munos, 2012). Worst-case optimality of the accuracy exponent goes back to randomized information-based complexity (Baran et al., 2006), and the sharp dependence on confidence is known (Kunsch and Rudolf, 2019). We do not claim these one-dimensional rates as new. The role of the integration lemma here is more specific: one estimator controls every marked prefix of an adaptively selected polygonal path, with no confidence penalty for the number of points subsequently compared by the trap.

Relation to low-dimensional stationarity. The geometric half of the argument descends from gradient-flow trapping. Bubeck and Mikulincer (2020) developed Cut-and-Flow on compact domains with value access, and Hollender and Zampetakis (2025) gave a detailed treatment of the planar unconstrained geometry and related oracle bounds. Our result keeps the topological certificate but reconstructs only the relative values that the certificate actually uses. Results on adaptive rounds use a zeroth-order oracle and a different complexity measure (Zhou et al., 2025). Numerically similar exponents under higher-order smoothness (Hinder, 2019) or comparison access with a Lipschitz Hessian (Wang et al., 2026) concern different oracle models. The comparison held fixed here is exact gradients only, mere Lipschitzness of the gradient, dimension two, and total query count.

Scope. The exponent $4/3$ is an upper bound, not an optimality theorem for planar stationarity. The embedded one-dimensional obstruction leaves the interval $[1, 4/3]$ open. Exact gradients, known valid bounds on L and Δ , unrestricted real computation between queries, and nonlocal sampling are part of the model. The proof neither covers noisy gradients nor extends formally to dimension three, where a cut is a surface rather than a path.

Organization. Section 2 fixes the oracle model and the trapping invariant. Section 3 proves simultaneous prefix integration and its sharp lower counterpart. Section 4 turns approximate cut values into a valid new trap. Section 5 gives the planar algorithm, the complete query accounting, and the Las Vegas conversion. The final section isolates what the scale calculation proves, what it does not prove, and the questions it leaves behind.

2 The oracle model and the trapping invariant

All norms are Euclidean. We fix a target accuracy $\varepsilon > 0$ and make the assumptions explicit because each has a separate use in the proof.

Assumption 2.1 (Planar gradient-query model). The input consists of $x_0 \in \mathbb{R}^2$ and valid numbers $L, \Delta > 0$ such that:

(A1) *Smoothness.* The function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is continuously differentiable and

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad (x, y \in \mathbb{R}^2).$$

(A2) *Initial gap.* $f(x_0) - \inf_{x \in \mathbb{R}^2} f(x) \leq \Delta$.

(A3) *Information.* A query at x returns the exact vector $\nabla f(x)$ and no function value. Queries may be randomized, adaptive, and nonlocal.

(A4) *Accounting.* Every gradient evaluation is counted, computation between queries is free, and the output must be a point whose gradient was queried.

The algorithm also receives the requested failure probability. We use the dimensionless parameter q from (1.1). The statements below are written for $q \geq q_0$, where q_0 is universal; the complementary bounded regime is absorbed by the constants.

For comparison, smoothness gives

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2. \quad (2.1)$$

Thus ordinary gradient descent needs at most $2q^2$ unsuccessful steps. The remainder of the paper improves this exponent without changing [Assumption 2.1](#).

2.1 Unreachable boundaries

The invariant is most transparent when stated in metric form.

Definition 2.2 (Unreachability and traps). For $\lambda > 0$, a point y is λ -unreachable from x if

$$f(y) > f(x) - \lambda\|y - x\|. \quad (2.2)$$

A pair (R, x) is a λ -trap if $R \subset \mathbb{R}^2$ is a compact full-dimensional rectangle, $x \in \text{int } R$, and every point of ∂R is λ -unreachable from x .

The name comes from the following elementary gradient-flow argument. It is the only topological input needed later.

Lemma 2.3 (A trapped rectangle contains a small gradient). *If (R, x) is a λ -trap, then some $u \in R$ satisfies $\|\nabla f(u)\| \leq \lambda$.*

Proof. Suppose instead that $\|\nabla f\| > \lambda$ throughout R , and follow the gradient flow $\dot{\gamma} = -\nabla f(\gamma)$ from x . The globally Lipschitz vector field has a global solution. While the trajectory remains in R ,

$$\frac{d}{dt}f(\gamma(t)) = -\|\nabla f(\gamma(t))\|^2 < -\lambda^2.$$

Compactness forces a finite first exit time τ . At $y = \gamma(\tau) \in \partial R$,

$$\begin{aligned} f(y) - f(x) &= -\int_0^\tau \|\nabla f(\gamma(t))\|^2 dt \\ &< -\lambda \int_0^\tau \|\dot{\gamma}(t)\| dt \leq -\lambda\|y - x\|, \end{aligned}$$

contradicting (2.2). □

The gap assumption creates the first trap without evaluating a function value. This point is easy to miss: the values are used in the proof of the invariant, not by the algorithm.

Lemma 2.4 (Initialization from a gap bound). *Let $\lambda > 0$ and*

$$R(\lambda) = \left\{y \in \mathbb{R}^2 : \|y - x_0\|_\infty \leq 2\Delta/\lambda\right\}. \quad (2.3)$$

Then $(R(\lambda), x_0)$ is a strict λ -trap.

Proof. Every $y \in \partial R(\lambda)$ satisfies $\|y - x_0\| \geq 2\Delta/\lambda$, whereas $f(y) \geq \inf f \geq f(x_0) - \Delta$. Hence

$$f(y) \geq f(x_0) - \Delta > f(x_0) - \lambda\|y - x_0\|.$$

□

2.2 The geometry of one cut

Let R be a rectangle whose longest side has length r . Trisect a longest side; the two cut faces are parallel segments E_1, E_2 . Two facts will be used repeatedly. First, for any anchor $x \in R$, at least one of the two faces lies at distance at least $r/6$ from x . Second, the two faces are separated by $r/3$. A cut retains one of the two overlapping two-thirds subrectangles. Its inherited boundary comes from ∂R ; its only new boundary is one of E_1, E_2 .

With exact values one could compare the anchor to the minima of the two cuts and preserve the trap directly. We shall instead place finite nets on both segments, traverse them by one polygonal path, and estimate every needed relative value from gradients. The next two sections quantify the two possible losses: stochastic integration error at the net points and the deterministic extension from a net to the entire new face.

3 One path estimates every marked value

Although the oracle never returns $f(x)$, it determines differences of f along a path. Let $\gamma : [0, H] \rightarrow \mathbb{R}^2$ be an absolutely continuous unit-speed path. At every s at which the derivative exists,

$$f(\gamma(s)) - f(\gamma(0)) = \int_0^s \langle \nabla f(\gamma(u)), \gamma'(u) \rangle du. \quad (3.1)$$

We use polygonal paths and mark every corner, so the tangent is constant on the interior of each integration interval.

The important point is not merely to estimate the final integral. One trapping round compares the anchor to many net points visited in path order. Estimating those values separately would discard the common information in their prefixes. The following lemma keeps it.

Lemma 3.1 (Simultaneous prefix integration). *Let $\gamma : [0, H] \rightarrow \mathbb{R}^d$ be a unit-speed polygonal path, and let $0 = s_0 < s_1 < \dots < s_m = H$ contain every corner and every point at which a potential difference is requested. For any $\eta > 0$ and $\rho \in (0, 1)$, exact gradient queries produce estimates $\hat{D}_0, \dots, \hat{D}_m$ such that*

$$\mathbb{P} \left[\max_{0 \leq j \leq m} |\hat{D}_j - (f(\gamma(s_j)) - f(\gamma(0)))| \leq \eta \right] \geq 1 - \rho \quad (3.2)$$

using at most

$$m + C \left(\frac{LH^2}{\eta} \right)^{2/3} \log^{1/3} \frac{2}{\rho} \quad (3.3)$$

queries, where C is universal. The conclusion remains valid conditionally when the path and its marks are selected from an earlier transcript and the integration samples are fresh.

Proof. On a straight piece of γ , put

$$h(s) = \langle \nabla f(\gamma(s)), \gamma'(s) \rangle.$$

The path has unit speed and its tangent is constant on this piece, so

$$|h(s) - h(t)| \leq \|\nabla f(\gamma(s)) - \nabla f(\gamma(t))\| \leq L|s - t|. \quad (3.4)$$

Fix an integer $K \geq 1$. Begin with the intervals $[s_{j-1}, s_j]$ and subdivide each into pieces of length at most H/K . If the original interval has length a_j , it contributes at most $\lceil Ka_j/H \rceil$ pieces; hence there are at most $K + m$ refined intervals. Write them in path order as I_i , with lengths ℓ_i .

Independently sample U_i uniformly from I_i , query $\nabla f(\gamma(U_i))$, and estimate the integral on I_i by

$$\hat{I}_i = \ell_i h(U_i).$$

The error

$$\xi_i = \hat{I}_i - \int_{I_i} h(s) ds$$

has mean zero. By (3.4),

$$\begin{aligned} |\xi_i| &\leq \int_{I_i} |h(U_i) - h(s)| ds \\ &\leq L \int_{I_i} |U_i - s| ds \leq \frac{L\ell_i^2}{2} := b_i. \end{aligned} \tag{3.5}$$

Consequently,

$$\sum_i b_i^2 \leq \frac{L^2}{4} \left(\max_i \ell_i \right)^3 \sum_i \ell_i \leq \frac{L^2 H^4}{4K^3} := B^2. \tag{3.6}$$

Let $S_k = \sum_{i \leq k} \xi_i$. Hoeffding's lemma shows that, for every $\theta > 0$,

$$M_k = \exp \left(\theta S_k - \frac{\theta^2}{2} \sum_{i \leq k} b_i^2 \right)$$

is a nonnegative supermartingale. If $S_k > \eta$ for some k , then $M_k \geq \exp(\theta\eta - \theta^2 B^2/2)$. Ville's inequality, followed by $\theta = \eta/B^2$, gives

$$\mathbb{P} \left(\max_k S_k > \eta \right) \leq \exp \left(-\frac{\eta^2}{2B^2} \right).$$

Applying the same argument to $-S_k$ and using (3.6),

$$\mathbb{P} \left(\max_k |S_k| > \eta \right) \leq 2 \exp \left(-\frac{2\eta^2 K^3}{L^2 H^4} \right). \tag{3.7}$$

Every marked point is a refined-interval endpoint. Taking

$$K \geq \left(\frac{LH^2}{\eta} \right)^{2/3} \left(\frac{1}{2} \log \frac{2}{\rho} \right)^{1/3}$$

and rounding upward proves (3.2)–(3.3) by (3.1).

For an adaptive path, condition on the transcript preceding its construction. The path, marks, and partition are then fixed, while the new uniform samples retain exactly the conditional independence and centering used above. Thus (3.7) holds conditionally with the same bound. $\square$

The proof explains both exponents in (3.3). Each of K strata contributes an error of size LH^2/K^2 , and centered independent errors add at square-root scale. The result is $LH^2/K^{3/2}$; solving for K gives the power 2/3, while the exponential tail gives the cube root of the confidence logarithm. A constant component of the gradient is integrated exactly. Only its Lipschitz variation is paid for.

3.1 The accuracy exponent is sharp

The preceding lemma asks for every prefix, but its stochastic exponent is already forced by the much easier task of estimating one integral. The following proof is included because it shows the obstruction in the same smooth-potential language used by the algorithm.

Proposition 3.2 (Fixed-confidence lower bound). *There are universal constants $c, c_0 > 0$ with the following property. Let $0 < \eta \leq c_0 LH^2$. If a randomized adaptive algorithm, given point-query access to an arbitrary L -Lipschitz function $h : [0, H] \rightarrow \mathbb{R}$, uses at most n queries and satisfies*

$$\mathbb{P}\left(\left|\hat{I} - \int_0^H h(s) ds\right| \leq \eta\right) \geq \frac{2}{3} \quad (3.8)$$

for every h , then

$$n \geq c \left(\frac{LH^2}{\eta}\right)^{2/3}. \quad (3.9)$$

Proof. It suffices to treat the regime in which $R = LH^2/\eta$ exceeds a universal constant. Fix a sufficiently small universal $\alpha > 0$, put $M = \lfloor \alpha R^{2/3} \rfloor$, and partition $[0, H]$ into M cells of width $w = H/M$. On cell j , let ψ_j be the triangular tent that vanishes at the endpoints, has height $Lw/2$ at the midpoint, and is zero off that cell. The interiors of the supports are disjoint, every signed sum

$$h_\sigma = \sum_{j=1}^M \sigma_j \psi_j, \quad \sigma_j \in \{-1, +1\}, \quad (3.10)$$

is L -Lipschitz, and one tent contributes signed area

$$\int_0^H \sigma_j \psi_j(s) ds = \sigma_j a, \quad a = \frac{Lw^2}{4} = \frac{LH^2}{4M^2}. \quad (3.11)$$

Give the signs the product Rademacher distribution. Consider a deterministic adaptive algorithm making at most $n < M/2$ queries. An interior query in one cell reveals at most that cell's sign, while a query at a cell boundary reveals none. Conditional on the complete transcript, at least $K = M - n \geq M/2$ signs remain independent and uniform. The unknown part of the integral is therefore aS_K , where S_K is a sum of K independent Rademacher variables.

For every $u \in \mathbb{R}$ and $r \geq 0$, the interval $[u - r, u + r]$ contains at most $r + 2$ points of the parity lattice supporting S_K . The central atom is largest, and the standard central-binomial estimate yields

$$\sup_u \mathbb{P}(|S_K - u| \leq r) \leq (r + 2) \frac{\binom{K}{\lfloor K/2 \rfloor}}{2^K} \leq (r + 2) \sqrt{\frac{2}{K}}. \quad (3.12)$$

This applies to any estimate selected from the transcript with $r = \eta/a = 4M^2/R$. Since $K \geq M/2$ and $M \leq \alpha R^{2/3}$, the right-hand side of (3.12) is at most

$$8\alpha^{3/2} + \frac{4}{\sqrt{M}}.$$

Choose α small enough, and absorb the finitely many smaller values of R into the constants, so that this quantity is strictly below $2/3$. Every deterministic algorithm with fewer than $M/2$ queries then has average success below $2/3$ under the signed-tent distribution.

If a randomized algorithm satisfied (3.8) for every input, its success averaged over both its internal randomness and this input distribution would be at least $2/3$. Fixing a random seed with at least that average contradicts the deterministic bound. Hence $n \geq M/2$, which is (3.9) after changing universal constants.

Finally, every h_σ is the derivative of $g(s) = \int_0^s h_\sigma(u) du$, whose derivative is L -Lipschitz. The hard family therefore belongs to the same one-dimensional smooth-potential oracle encountered on a straight segment in Lemma 3.1. $\square$

3.2 The complete confidence law

Let $N_{\text{int}}(L, H, \eta, \rho)$ denote the smallest worst-case number of point queries needed to integrate every L -Lipschitz $h : [0, H] \rightarrow \mathbb{R}$ to additive error η with failure probability at most ρ .

Corollary 3.3 (Sharp accuracy–confidence tradeoff). *Uniformly for $0 < \eta \leq c_0 LH^2$ and $0 < \rho \leq 1/3$,*

$$N_{\text{int}}(L, H, \eta, \rho) = \Theta \left(1 + \min \left\{ \frac{LH^2}{\eta}, \left(\frac{LH^2}{\eta} \right)^{2/3} \log^{1/3} \frac{1}{\rho} \right\} \right), \quad (3.13)$$

with universal constants.

Proof. Put

$$R = \frac{LH^2}{\eta}, \quad \ell = \log \frac{1}{\rho}.$$

We first remove a minor normalization issue. One query at the origin of the interval determines the arbitrary constant part of h . With

$$g(u) = \frac{h(Hu) - h(0)}{LH}, \quad 0 \leq u \leq 1,$$

the function g is 1-Lipschitz, $g(0) = 0$, and

$$\int_0^H h(s) ds = Hh(0) + LH^2 \int_0^1 g(u) du. \quad (3.14)$$

Thus the normalized accuracy is R^{-1} ; the initial query accounts for the additive 1 in (3.13).

For the normalized one-dimensional Lipschitz class, the optimal probabilistic integration error of Kunsch and Rudolf (2019) states that the smallest radius achievable with n point queries and failure probability ρ is, up to universal constants,

$$e_n(\rho) \asymp n^{-1} \min \left\{ 1, \sqrt{\frac{\ell}{n}} \right\}. \quad (3.15)$$

It remains only to invert this relation. Define

$$n_\star = \min \left\{ R, R^{2/3} \ell^{1/3} \right\}.$$

If $\ell \leq R$, then $n_\star = R^{2/3} \ell^{1/3} \geq \ell$, so the stochastic branch of (3.15) applies and

$$e_{n_\star}(\rho) \asymp \frac{\sqrt{\ell}}{n_\star^{3/2}} = \frac{1}{R}.$$

If $\ell \geq R$, then $n_\star = R \leq \ell$, so the deterministic branch applies and again $e_{n_\star}(\rho) \asymp n_\star^{-1} = R^{-1}$. Since $e_n(\rho)$ is nonincreasing in n , the two-sided constants in (3.15) imply

$$n = \Theta(1 + n_\star) = \Theta\left(1 + \min\left\{R, R^{2/3}\ell^{1/3}\right\}\right),$$

which is (3.13) after substituting R and ℓ .

The two upper-bound mechanisms can also be read directly from the proof. When $\ell \leq R$, the one-endpoint case of Lemma 3.1 uses $O(R^{2/3}\ell^{1/3})$ queries. When $\ell \geq R$, midpoint quadrature on n equal subintervals has deterministic error at most $LH^2/(4n)$ and hence uses $O(R)$ queries. The lower half of (3.15) proves that neither branch can be improved by more than a universal factor; Proposition 3.2 recovers its constant-confidence consequence without invoking the external result. $\square$

At moderate confidence, the randomized branch is smaller and both powers in (3.3) are optimal. Once $\log(1/\rho) \gtrsim LH^2/\eta$, deterministic quadrature is preferable and removes randomness altogether. The additional content of Lemma 3.1 is not a better one-point integrator: it is the simultaneous and conditionally adaptive prefix guarantee required by the planar recursion.

4 Approximate comparisons preserve a planar trap

Consider a λ -trap (R, x) , and let r be the length of a longest side of R . Denote its two trisection segments by E_1, E_2 , and place an endpoint-including δ -net S_i on each E_i . Suppose that for every $z \in S_1 \cup S_2$ we have an estimate $\widehat{d}(z)$ of

$$d(z) = f(z) - f(x)$$

with common error at most η . We call a point a conservative descent candidate when

$$A := \left\{z \in S_1 \cup S_2 : \widehat{d}(z) \leq -\lambda\|z - x\| - 2\eta\right\}. \quad (4.1)$$

If A is empty, retain the two-thirds rectangle whose new face is farther from x . If A is nonempty, choose a minimizer $z_\star$ of $\widehat{d}$ over A , make it the new anchor, and retain the two-thirds rectangle whose new face is the cut opposite $z_\star$.

Two losses must be controlled. The term proportional to η/r is the price of uncertain comparisons. The term proportional to $L\delta^2/r$ is the price of observing only a net. We isolate the second one first.

Lemma 4.1 (From a net to an entire face). *Let E be a segment, let S be an endpoint-including δ -net of E , and let $\ell = \text{dist}(x, E) > 0$. If every $z \in S$ is μ -unreachable from x , then every $y \in E$ is*

$$\left[\mu + \frac{\delta^2}{2\ell} \left(L + \frac{2\mu}{\ell}\right)\right]\text{-unreachable from } x. \quad (4.2)$$

Proof. On E define

$$\phi(y) = f(y) - f(x) + \mu\|y - x\|.$$

Let $y_\star$ minimize ϕ on the segment. If $y_\star$ is an endpoint, then it belongs to S and $\phi(y_\star) > 0$, so the conclusion is immediate. We may therefore assume that $y_\star$ lies in the relative interior of E , where the tangential derivative of ϕ vanishes.

Choose $z \in S$ with $\|z - y_\star\| \leq \delta$ and write $y(t) = y_\star + t(z - y_\star)$. For nonzero vectors u, v whose norms are at least ℓ ,

$$\left\| \frac{u}{\|u\|} - \frac{v}{\|v\|} \right\| \leq \frac{2}{\ell} \|u - v\|. \quad (4.3)$$

Thus the gradient of ϕ along E is $L + 2\mu/\ell$ -Lipschitz. If $\psi(t) = \phi(y(t))$, then $\psi'(0) = 0$ and

$$|\psi'(t) - \psi'(0)| \leq t \|z - y_\star\|^2 \left(L + \frac{2\mu}{\ell} \right).$$

Integrating and using $\phi(z) > 0$ gives

$$\phi(y_\star) > -\frac{\delta^2}{2} \left(L + \frac{2\mu}{\ell} \right).$$

Every $y \in E$ satisfies $\phi(y) \geq \phi(y_\star)$ and $\|y - x\| \geq \ell$. Dividing the additive loss by this last distance is exactly (4.2). $\square$

The minimization of ϕ , rather than f alone, is the useful viewpoint. At an interior minimizer the tangential linear term vanishes, which turns smoothness into a quadratic net error. At an endpoint there is no approximation error because the endpoints are queried.

Lemma 4.2 (Robust planar trisection). *On the event*

$$\max_{z \in S_1 \cup S_2} |\hat{d}(z) - d(z)| \leq \eta, \quad (4.4)$$

the update above returns a rectangle R' and an anchor $x' \in \text{int } R'$ forming a λ' -trap, where

$$\lambda' \leq \lambda + \frac{18\eta}{r} + \frac{3\delta^2}{r} \left[L + \frac{12}{r} \left(\lambda + \frac{18\eta}{r} \right) \right]. \quad (4.5)$$

In particular, if $\lambda + 18\eta/r \leq \varepsilon/2$ and $r \geq \varepsilon/(2\sqrt{2}L)$, then

$$\lambda' - \lambda \leq \frac{18\eta}{r} + \frac{75L\delta^2}{r}. \quad (4.6)$$

Proof. There are two cases.

No conservative descent candidate. Suppose A is empty. Choose the two-thirds rectangle whose new face E has distance at least $r/6$ from x . For every net point z on this face,

$$\hat{d}(z) > -\lambda \|z - x\| - 2\eta,$$

and (4.4) implies

$$d(z) > -\lambda \|z - x\| - 3\eta \geq -\left(\lambda + \frac{18\eta}{r} \right) \|z - x\|. \quad (4.7)$$

The inherited boundary remains λ -unreachable because the anchor has not moved.

A conservative descent candidate exists. Let $z_\star \in \arg \min_{z \in A} \hat{d}(z)$ and set $x' = z_\star$. Membership in A and (4.4) give the genuine descent margin

$$f(z_\star) - f(x) \leq -\lambda \|z_\star - x\| - \eta. \quad (4.8)$$

For a point y on the inherited part of the old boundary,

$$\begin{aligned} f(y) - f(z_\star) &> -\lambda\|y - x\| + \lambda\|z_\star - x\| + \eta \\ &\geq -\lambda\|y - z_\star\| + \eta. \end{aligned} \tag{4.9}$$

Thus the inherited faces remain strictly λ -unreachable.

An endpoint of either cut lies on the old boundary. If such a point belonged to A , the good event would imply $f(z) - f(x) \leq -\lambda\|z - x\| - \eta$, contradicting the old strict trap. Hence $z_\star$ lies in the relative interior of its cut and in the interior of the retained rectangle.

Retain the two-thirds rectangle whose new face is the opposite cut. Every net point z on this face lies at distance at least $r/3$ from $z_\star$. If $z \in A$, approximate minimality gives

$$f(z) - f(z_\star) \geq -2\eta. \tag{4.10}$$

If $z \notin A$, then

$$\begin{aligned} f(z) - f(z_\star) &> -\lambda\|z - x\| - 3\eta + \lambda\|z_\star - x\| + \eta \\ &\geq -\lambda\|z - z_\star\| - 2\eta. \end{aligned} \tag{4.11}$$

Both estimates imply strict $(\lambda + 6\eta/r)$ -unreachability. We shall use the larger common value

$$\mu = \lambda + \frac{18\eta}{r}. \tag{4.12}$$

In both cases of the update, the new-face net is therefore μ -unreachable, and the anchor lies at distance at least $r/6$ from that face.

Apply [Lemma 4.1](#) with $\ell \geq r/6$. Since $1/(2\ell) \leq 3/r$ and $2\mu/\ell \leq 12\mu/r$, the resulting reachability parameter is at most the right-hand side of [\(4.5\)](#). This also dominates the parameter on the inherited boundary and proves the first claim.

For the second claim, assume $\mu \leq \varepsilon/2$ and $r \geq \varepsilon/(2\sqrt{2}L)$. Then $12\mu/r \leq 12\sqrt{2}L$, and hence

$$\frac{3\delta^2}{r} \left(L + \frac{12\mu}{r} \right) \leq \frac{75L\delta^2}{r},$$

where the numerical constant is deliberately loose. Substituting [\(4.12\)](#) yields [\(4.6\)](#). $\square$

The endpoint case deserves an explicit check. A candidate at a cut endpoint would lie on the old boundary, and moving the anchor there would destroy the interior condition in [Definition 2.2](#). The strict margin in [\(4.1\)](#) rules this out. This argument and the quadratic extension in [Lemma 4.1](#) are the two places where a seemingly harmless non-strict comparison would be insufficient.

5 Randomized stationary-point search in the plane

We now assemble the two preceding ingredients. The theorem is stated before the parameter schedule so that the roles of the parameters remain visible: L controls local variation, Δ buys the initial trap, ε is the target gradient scale, and ρ is used only in the randomized line integrals.

Theorem 5.1 (Randomized gradient-only search in the plane). *Under [Assumption 2.1](#), there are universal constants $C, q_0 > 0$ such that, for $q \geq q_0$ and $\rho \in (0, 1/2)$, an adaptive randomized algorithm using only exact gradient queries returns one of its queried points X with*

$$\mathbb{P}(\|\nabla f(X)\| \leq \varepsilon) \geq 1 - \rho \tag{5.1}$$

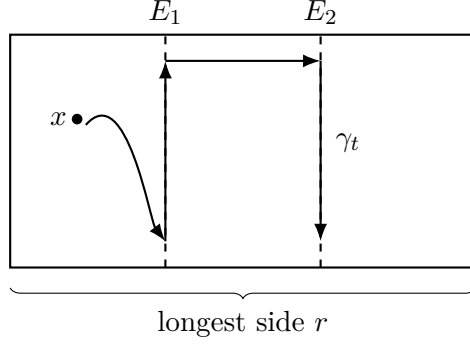


Figure 1: One randomized trapping round. The path begins at the anchor and visits the ordered nets on both trisection faces. It is not a descent path; its sole purpose is to reconstruct, in one event, the relative values used by the robust update.

after at most

$$Cq^{4/3} \log^{1/3} \frac{2}{\rho} \quad (5.2)$$

gradient queries.

Proof idea. Start with the large square supplied by [Lemma 2.4](#). At each scale, traverse the nets on both trisection faces by the path in [Figure 1](#). The integration lemma reconstructs all relative values used by the cut, and the robust trisection lemma replaces the old trap by a smaller one. The integration tolerance and the net spacing spend a geometrically decaying reachability budget. Every two rounds the spatial scale contracts by 2/3, making the query costs summable. The same geometric decay pays for an increasingly stringent confidence request at later rounds, which is why no factor depending on the number of rounds survives.

5.1 The procedure and its schedule

Set $\lambda_0 = \varepsilon/8$ and

$$R_0 = \left\{ y \in \mathbb{R}^2 : \|y - x_0\|_\infty \leq 2\Delta/\lambda_0 \right\}. \quad (5.3)$$

By [Lemma 2.4](#), (R_0, x_0) is a λ_0 -trap without a function-value query. At round t , let r_t be the longest side of R_t and put $i = \lfloor t/2 \rfloor$. The four schedules are

$$\begin{aligned} a_t &= \frac{\varepsilon}{32} \left(\frac{3}{4} \right)^i, & \eta_t &= \frac{a_t r_t}{72}, \\ \delta_t^2 &= \frac{a_t r_t}{300L}, & \rho_t &= \rho 2^{-(t+1)}. \end{aligned} \quad (5.4)$$

The variable a_t is not an oracle parameter; it is the reachability loss reserved for the round.

Algorithm 1 Randomized planar gradient-only search

Input: $x_0, L, \Delta, \varepsilon$, and failure target $\rho \in (0, 1/2)$.

1. Set $\lambda_0 = \varepsilon/8$, form R_0 by (5.3), and set $t = 0$.
2. While the longest side r_t of R_t satisfies $r_t > \varepsilon/(2\sqrt{2}L)$, compute the four parameters in (5.4). Let $E_{t,1}, E_{t,2}$ be the two trisection faces orthogonal to a longest side, and place an endpoint-including δ_t -net $S_{t,j}$ on each $E_{t,j}$.
3. Starting at x_t , follow a polygonal path to one endpoint of $E_{t,1}$, traverse $E_{t,1}$, bridge to the corresponding endpoint of $E_{t,2}$, and traverse $E_{t,2}$. Mark every point in $S_{t,1} \cup S_{t,2}$ and every path corner. Apply Lemma 3.1 with tolerance η_t and failure target ρ_t to obtain estimates $\hat{d}_t(z)$ of $f(z) - f(x_t)$ for all $z \in S_{t,1} \cup S_{t,2}$.
4. Form the conservative candidate set

$$A_t = \left\{ z \in S_{t,1} \cup S_{t,2} : \hat{d}_t(z) \leq -\lambda_t \|z - x_t\| - 2\eta_t \right\}.$$

If $A_t = \emptyset$, let R_{t+1} be the two-thirds subrectangle whose new face is farther from x_t , and set $x_{t+1} = x_t$. Otherwise choose $z_t^* \in \arg \min_{z \in A_t} \hat{d}_t(z)$, let R_{t+1} be the two-thirds subrectangle whose new face is the trisection face opposite z_t^* , and set $x_{t+1} = z_t^*$.

5. Set

$$\lambda_{t+1} = \lambda_t + a_t, \tag{5.5}$$

increment t , and return to Step 2.

6. When the loop ends, set $T = t$, query $\nabla f(x_T)$, and return x_T .
-

Ties in the choice of a longest side, an endpoint, or a minimizing candidate are resolved by any fixed rule. Thus the only randomized operation in Algorithm 1 is the fresh stratified sampling inside Lemma 3.1; the cut itself is the deterministic robust update of Lemma 4.2.

Proof of Theorem 5.1. We verify that Algorithm 1 preserves the trap invariant, then establish termination, success probability, and query complexity.

Initialization. The longest side of R_0 is

$$r_0 = \frac{4\Delta}{\lambda_0} = \frac{32\Delta}{\varepsilon}. \tag{5.6}$$

The trap property follows from Lemma 2.4. Neither $f(x_0)$ nor $\inf f$ is queried or numerically used.

One round uses gradients only. The polygonal tour described above lies inside the convex rectangle and has length at most

$$\left(\sqrt{2} + 1 + \frac{1}{3} + 1 \right) r_t < 4r_t. \tag{5.7}$$

By Lemma 3.1, fresh gradient samples reconstruct $f(z) - f(x_t)$ at every marked net point with common error at most η_t , except on an event of conditional probability at most ρ_t . No absolute value of f is reconstructed.

Invariant induction. Assume that the integration event at round t is good. The geometric series in (5.4) gives

$$\lambda_t \leq \lambda_0 + \sum_{s \geq 0} a_s = \frac{\varepsilon}{8} + 2 \sum_{i \geq 0} \frac{\varepsilon}{32} \left(\frac{3}{4} \right)^i = \frac{3\varepsilon}{8}. \tag{5.8}$$

Moreover,

$$\lambda_t + \frac{18\eta_t}{r_t} \leq \frac{3\varepsilon}{8} + \frac{a_t}{4} \leq \frac{49\varepsilon}{128} < \frac{\varepsilon}{2}. \tag{5.9}$$

The round is active only while $r_t > \varepsilon/(2\sqrt{2}L)$, so (4.6) applies. The schedules were chosen so that its two terms are equal:

$$\lambda' - \lambda_t \leq \frac{18}{r_t} \frac{a_t r_t}{72} + \frac{75L}{r_t} \frac{a_t r_t}{300L} = \frac{a_t}{2} < a_t. \quad (5.10)$$

Unreachability is monotone in its parameter; hence the new pair is a λ_{t+1} -trap. The endpoint argument in Lemma 4.2 also guarantees $x_{t+1} \in \text{int } R_{t+1}$. This closes the induction.

Deterministic termination and output. Each round shortens a currently longest side by one third. If the side lengths before a cut are $a \geq b$, then either $2a/3$ remains longest and the next cut makes it $4a/9$, or b becomes longest and is cut to at most $2a/3$. Thus every two rounds reduce the longest side by at least $2/3$. The deterministic cap

$$\bar{T} = 2 \left\lceil \log_{3/2} \frac{r_0}{\varepsilon/(2\sqrt{2}L)} \right\rceil + 2 = O(1 + \log q) \quad (5.11)$$

therefore suffices.

At termination, the rectangle has diameter at most $\varepsilon/(2L)$. Lemma 2.3 supplies a point $y \in R_T$ with $\|\nabla f(y)\| \leq \lambda_T$. Smoothness and (5.8) yield

$$\|\nabla f(x_T)\| \leq \lambda_T + L\|x_T - y\| \leq \frac{3\varepsilon}{8} + \frac{\varepsilon}{2} = \frac{7\varepsilon}{8}. \quad (5.12)$$

The final gradient query makes x_T an admissible output.

Success probability. Conditioned on the transcript before round t , the new path is fixed and its samples are fresh. Iterated conditioning and a union bound give

$$\mathbb{P}(\text{some active integration event fails}) \leq \sum_{t \geq 0} \rho 2^{-(t+1)} = \rho. \quad (5.13)$$

On the complement of this event the preceding induction proves (5.12), hence (5.1).

Query complexity. An endpoint-including δ_t -net of one cut contains $O(r_t/\delta_t + 1)$ points. During an active round,

$$\frac{Lr_t}{a_t} \geq \frac{L\varepsilon/(2\sqrt{2}L)}{\varepsilon/32} = \frac{16}{\sqrt{2}} > 1,$$

so the additive constant is absorbed and the total number of marked points obeys

$$m_t = O\left(\frac{r_t}{\delta_t}\right) = O\left(\sqrt{\frac{Lr_t}{a_t}}\right). \quad (5.14)$$

Combining (5.7), (5.4), and Lemma 3.1, the stochastic part of round t costs

$$O\left[\left(\frac{Lr_t}{a_t}\right)^{2/3} \log^{1/3} \frac{2}{\rho_t}\right]. \quad (5.15)$$

Let $i = \lfloor t/2 \rfloor$. From (5.6), the two-round contraction, and the definition of a_t ,

$$\frac{Lr_t}{a_t} \leq C \frac{\Delta L}{\varepsilon^2} \left(\frac{2/3}{3/4}\right)^i = Cq^2 \left(\frac{8}{9}\right)^i. \quad (5.16)$$

Put $\beta = (8/9)^{2/3}$ and $A = \log(4/\rho)$. For $t \in \{2i, 2i+1\}$,

$$\log \frac{2}{\rho_t} \leq A + 2(i+1) \log 2.$$

Using $(u + v)^{1/3} \leq u^{1/3} + v^{1/3}$, the sum of (5.15) is at most

$$Cq^{4/3} \sum_{i \geq 0} \beta^i (A + 2(i + 1) \log 2)^{1/3} \leq Cq^{4/3} \left[A^{1/3} \sum_{i \geq 0} \beta^i + (2 \log 2)^{1/3} \sum_{i \geq 0} \beta^i (i + 1)^{1/3} \right]. \quad (5.17)$$

Both series are universal constants. Since $\rho < 1/2$, the final display is $O(q^{4/3} \log^{1/3}(2/\rho))$.

Similarly, (5.14) and (5.16) give

$$\sum_t m_t \leq Cq \sum_{i \geq 0} \left(\frac{8}{9}\right)^{i/2} = O(q),$$

which is absorbed by (5.2) for $q \geq q_0$. The final gradient query is absorbed as well. This proves the theorem. $\square$

5.2 Why the round count disappears

A uniform allocation $\rho_t \asymp \rho/\bar{T}$ would insert $\log^{1/3} \bar{T} = \log^{1/3} \log q$ into the complexity. The geometric allocation in (5.4) requests higher confidence in later rounds, but (5.16) makes those rounds cheaper by a geometric factor. Equation (5.17) couples the two effects before taking a worst case over rounds. This order of summation, rather than a stronger concentration inequality, removes the spurious $\log \log q$ term.

Corollary 5.2 (No universal quadratic lower bound in the plane). *In the model of Assumption 2.1, randomized planar gradient-query complexity is $O(q^{4/3})$ at constant success probability. Consequently, a universal $\Omega(q^2)$ randomized lower bound does not hold in dimension two.*

5.3 Exact certification and zero error

The final gradient query has a second use. It reveals whether a trial has actually succeeded, even on the exceptional integration event. This turns the Monte Carlo theorem into a Las Vegas theorem without changing the expected exponent.

Theorem 5.3 (Zero-error planar search). *Under Assumption 2.1, there are universal constants $C, q_0 > 0$ such that, for $q \geq q_0$, a Las Vegas exact-gradient algorithm terminates almost surely, returns a queried point X satisfying $\|\nabla f(X)\| \leq \varepsilon$ with zero error, and uses at most $Cq^{4/3}$ gradient queries in expectation. If N is the number of trials, then*

$$\mathbb{P}(N > k) \leq 2^{-k} \quad (k = 0, 1, 2, \dots). \quad (5.18)$$

Proof. Run the algorithm of Theorem 5.1 with failure target $1/3$. After increasing the universal constant, one trial uses at most $Q = Cq^{4/3}$ queries deterministically. Inspect the exact final gradient. If its norm is at most ε , return the point; otherwise discard the trial and start again from x_0 with fresh randomness.

Conditional on any transcript consisting of k failed trials, the next trial succeeds with probability at least $2/3$. Iterated conditioning gives

$$\mathbb{P}(N > k) \leq 3^{-k} \leq 2^{-k}.$$

In particular, $\mathbb{P}(N = \infty) = 0$ and

$$\mathbb{E}N = \sum_{k \geq 0} \mathbb{P}(N > k) \leq \frac{3}{2}.$$

The algorithm returns only after observing $\|\nabla f(X)\| \leq \varepsilon$, so its error probability is zero. Its expected query count is at most $Q\mathbb{E}N \leq (3/2)Cq^{4/3}$, which is absorbed into the universal constant. $\square$

The conversion is elementary, but it is specific to an observable success condition. Exact gradients are used once to build the stochastic line integral and once more to certify the final point. The second use is what turns a high-probability speedup into a zero-error one.

6 What the scale calculation reveals

The proof separates four roles that gradient descent merges into one trajectory. Their interaction is summarized in [Table 1](#).

Table 1: The four currencies in the planar argument.

Role	Quantitative statement	Consequence
Gradient information	path error $LH^2/k^{3/2}$	$k \asymp (LH^2/\eta)^{2/3}$
Planar geometry	two cuts and one trapped boundary	one path serves a full round
Scale amortization	$r_{t+2} \leq (2/3)r_t$	round costs form a geometric sum
Exact certification	$\nabla f(x_T)$ is observed	zero-error restart

The first row is a numerical-integration fact, the second is topological, the third is an accounting choice, and the fourth is an oracle property. None by itself gives the theorem. The planar speedup appears because the four fit together at the same scale: a cut is one-dimensional, all of its comparisons are prefixes of one tour, and smaller traps demand proportionally cheaper tours.

What the result changes. The quadratic descent bound is not a universal law of two-dimensional smooth optimization. It is the cost of certifying progress along a single local trajectory. In the present oracle model, exact gradients also contain integrable global information. Randomized quadrature extracts enough of that information to drive a topological localization scheme, and [Theorem 5.1](#) shows that its total cost is strictly subquadratic.

The statement should not be read as a claim that integration alone solves the problem. [Proposition 3.2](#) and [Corollary 3.3](#) show that the accuracy and confidence exponents of one path are already sharp. The new optimization mechanism is the composition: simultaneous prefix control, robust trapping, and geometric amortization. This also identifies where a future improvement must enter.

The remaining exponent gap. A one-dimensional hard instance can be embedded in the plane, giving the known linear obstruction, whereas [Theorem 5.1](#) gives exponent $4/3$. The present frontier is therefore

$$1 \leq \text{planar randomized exponent} \leq \frac{4}{3}. \quad (6.1)$$

An $\Omega(q^{4/3})$ lower bound would have to rule out information reuse across successive cuts; the signed-tent construction only rules out a better integrator on one cut. Conversely, an $O(q)$ algorithm would

have to reuse gradients across scales, tolerate coarser relative values, or replace the rectangular trapping certificate.

Why the argument is planar. In two dimensions a separator is a segment, so a sufficiently fine net can be ordered along a path whose length is comparable to the diameter of the current rectangle. In three dimensions a separator is a surface. A one-dimensional tour that visits a fine surface net is much longer than the diameter, and the balance in (1.3) changes. A higher-dimensional extension therefore needs either a different certificate or a genuinely multidimensional reconstruction primitive; merely repeating the planar proof is not enough.

The role of exact gradients. Exactness is used twice. Within each stratum it gives a centered integration sample with no extra noise term. At the end it gives an exact success test, which is the reason Theorem 5.3 is possible. With stochastic gradients, averaging can reduce the first error but cannot certify the final event with finite cost without an additional separation condition. Noisy integration, robust comparison, and terminal testing would have to be budgeted together.

The proof leaves three concrete questions.

1. **What is the planar exponent?** Is the upper bound $q^{4/3}$ optimal, or can gradient information be shared across scales to reach the embedded one-dimensional rate $O(q)$?
2. **Can the speedup survive noise or locality?** For a variance-bounded stochastic-gradient oracle, is there a complexity bound with the same geometric factor plus a separate statistical term? Does any subquadratic rate remain if every query must stay near the current iterate?
3. **What replaces a path in higher dimension?** Can relative values on a surface be reconstructed at a cost compatible with trapping, or is there an oracle lower bound showing that the planar mechanism ends in dimension two?

The first question is the direct continuation of this paper. The integration lower bound makes it precise: changing constants or concentration inequalities inside a round will not close (6.1). A sharper result must change how information is transported from one scale to the next.

Conclusion. An exact gradient oracle does not reveal function values, but in the plane it reveals enough line-integral information to simulate the comparisons of a gradient-flow trap. Stratification makes those comparisons subquadratic; robust trisection makes them composable; geometric scheduling makes their cost summable. The resulting $O(q^{4/3})$ rate is therefore not an isolated estimate but the quantitative expression of one geometric fact: in the plane, an entire cut can be read through one path.

Computational assistance

Large language models assisted with literature discovery, proof exploration and checking, symbolic and numerical sanity checks, drafting, and editing. They were not treated as sources; all cited claims are tied to the literature. The author remains responsible for the final manuscript.

References

- Ilya Baran, Erik D. Demaine, and Dmitriy A. Katz. Optimally adaptive integration of univariate Lipschitz functions. In *Proceedings of the 7th Latin American Symposium on Theoretical Informatics*, pages 142–153, 2006. doi: 10.1007/11682462_17.
- Sébastien Bubeck and Dan Mikulincer. How to trap a gradient flow. In *Conference on Learning Theory*, 2020.
- Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 2020.
- Alexandra Carpentier and Rémi Munos. Adaptive stratified sampling for monte-carlo integration of differentiable functions. In *Advances in Neural Information Processing Systems*, volume 25, 2012.
- Sinho Chewi, Sébastien Bubeck, and Adil Salim. On the complexity of finding stationary points of smooth functions in one dimension. In *Algorithmic Learning Theory*, 2023.
- Oliver Hinder. Cutting plane methods can be extended into nonconvex optimization. *arXiv preprint arXiv:1805.08370*, 2019.
- Alexandros Hollender and Manolis Zampetakis. The computational complexity of finding stationary points in non-convex optimization. *Mathematical Programming*, 2025. doi: 10.1007/s10107-024-02139-3.
- Robert J. Kunsch and Daniel Rudolf. Optimal confidence for Monte Carlo integration of smooth functions. *Advances in Computational Mathematics*, 45(5–6):3095–3122, 2019. doi: 10.1007/s10444-019-09728-3.
- Helin Wang, Chenyi Zhang, Xiwen Tao, Yexin Zhang, and Tongyang Li. Finding stationary points by comparisons. In *International Conference on Machine Learning*, 2026.
- Rong-Xian Yue and Shi-Song Mao. On the variance of quadrature over scrambled nets and sequences. *Statistics & Probability Letters*, 44(3):267–280, 1999. doi: 10.1016/S0167-7152(99)00018-8.
- Huanjian Zhou, Andi Han, Akiko Takeda, and Masashi Sugiyama. The adaptive complexity of finding a stationary point. In *Conference on Learning Theory*, 2025.