MINIMAX-OPTIMAL UPDATE RULES FOR STOCHASTIC OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the following question in stochastic optimization: given a single unbiased stochastic gradient observation with bounded variance, how should one design an update rule to achieve the strongest possible descent guarantee? We formulate a minimax problem in which the decision variable is a rule that maps a single observation to an update. Under a prescribed bound on the update norm, the goal is to maximize the worst-case expected one-step decrease guaranteed by smoothness. The worst case is taken over all gradient directions and observation distributions consistent with the available information. When the gradient norm is known, we derive the exact optimal descent guarantee and characterize all rules that attain it. When the gradient norm is unknown, we fix a positive calibration threshold and derive the best uniform guarantee for all gradient norms at or above that threshold, along with a fixed rule that attains this guarantee. The same rule yields a finite-time stationarity guarantee without requiring estimation of the current gradient norm at each iteration. These results establish a unified theory of optimal bounded updates under different levels of information about the gradient norm. Numerical experiments support the theoretical results. Code is available at `https://anonymous.4open.science/r/MINIMAX-opt-0808/`.

1 INTRODUCTION

Stochastic optimization updates parameters using noisy gradient observations (Ghadimi & Lan, 2013). A single observation may deviate substantially from the true gradient, and using it directly can produce an excessively large update. Gradient clipping (Koloskova et al., 2023) and normalization (Hübler et al., 2025) are common ways to bound the update norm. These methods motivate a more general question: *given one unbiased observation with bounded variance and a prescribed update norm bound, which update rule provides the strongest possible descent guarantee?*

Existing analyses typically begin with a specified update rule and study how its parameters affect bias and convergence. We instead study the design of the observation-to-update map through a minimax model. The decision variable is a measurable rule T satisfying $\|T(z)\| \le R$ for every observation z. Its payoff is the expected lower bound on one-step objective decrease provided by smoothness. The observation is unbiased and its noise variance is at most σ^2, while its distribution is otherwise unknown. For each rule, we take the worst case over all gradient directions and observation distributions consistent with the available information. Optimizing this guarantee determines both the strongest descent obtainable from a single observation and the rules that attain it.

When the current gradient norm is known, we derive the exact minimax value, construct an attaining clipped linear rule, and characterize all optimal bounded measurable rules. When the norm is unknown, we prescribe a positive calibration threshold m and seek one fixed rule that guarantees progress for every gradient norm at least m. The rule calibrated at the threshold attains the best uniform guarantee over this entire region. Controlling its behavior below the threshold then yields a finite-time stationarity guarantee without estimating the current gradient norm at each iteration. A Gaussian example illustrates how additional distributional information changes the optimal update and attainable descent. Numerical experiments illustrate the exact values, joint calibration, and the response of the fixed rule to different noise shapes.

Overall, the analysis reveals how optimality constrains update design. For a positive variance bound, every optimal bounded measurable rule must be differentiable at the origin with the same local

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

gain. When the update cap is active, tightening it requires reducing the gain before clipping. A rule calibrated only at m also attains the best uniform guarantee over all gradient norms at least m. Together, these findings explain how the available gradient information and the update constraint determine both the attainable descent and the structure of optimal rules.

2 Related Work

Algorithm design through worst-case optimization. Performance estimation problems (PEP) formulate the worst-case performance of a first-order method as an optimization problem (Drori & Teboulle, 2014). Exact interpolation conditions enable sharp bounds over smooth convex function classes (Taylor et al., 2017), and optimizing these bounds leads to improved methods (Kim & Fessler, 2016). The Arcsine stepsize schedule provides a related motivation: a worst-case convergence criterion determines a random stepsize distribution with an equalization property across admissible curvatures (Altschuler & Parrilo, 2024). In stochastic optimization, expected progress also guides learning-rate selection from local gradient statistics (Schaul et al., 2013). Our work applies this design principle to the entire observation-to-update map, optimizing its worst-case smoothness-based one-step descent under an update norm bound.

Robust statistical decisions and distributional uncertainty. Robust statistics studies how distributional uncertainty shapes statistical procedures, from minimax location estimation under contamination (Huber, 1964) to mean estimation under moment assumptions (Catoni, 2012; Devroye et al., 2016). Recent work constructs optimal estimators over moment classes through convex analysis (van Parys & Zwart, 2026). Distributionally robust optimization similarly optimizes decisions against an ambiguity set of distributions, including sets defined by moment information (Delage & Ye, 2010); sharp moment inequalities provide tools for evaluating these distributional extrema (Bertsimas & Popescu, 2005). Here, distributional uncertainty concerns the gradient observation at the current iterate. The decision is a bounded update rule, and its criterion is guaranteed optimization progress, yielding an exact minimax value and a characterization of all optimal rules.

Clipping and bounded stochastic updates. Clipped-SGD theory connects clipping to generalized smoothness (Zhang et al., 2020) and high-probability convergence under heavy-tailed noise (Gorbunov et al., 2020; Cutkosky & Mehta, 2021). Tight analyses quantify how clipping thresholds affect stochastic bias and convergence (Koloskova et al., 2023). Recent work develops gradient quantile clipping (Merad & Gaïffas, 2024), analyzes normalization under bounded central moments (Hübler et al., 2025), and establishes guarantees for general nonlinear transformations under symmetric heavy-tailed noise (Armacki et al., 2025). The parameterizations $\eta \operatorname{clip}_\tau(G)$ and $\operatorname{clip}_{\eta\tau}(\eta G)$ are equivalent. Our contribution is to determine all optimal clipping parameters under a prescribed update bound and establish their optimality over the full class of bounded measurable rules.

3 The Decision Problem for Bounded Stochastic Updates

This section defines the measure of one-step progress and the available information for designing an update rule. All norms are Euclidean. Let $f : \mathbb{R}^d \to \mathbb{R}$ have an L-Lipschitz gradient, where $L > 0$, and let $R > 0$ be the prescribed bound on the update norm.

3.1 One-step descent and the full-information benchmark

Fix the current point x and write $\nabla f(x) = \mu e$, where $\mu > 0$ and $\|e\| = 1$. For an update $x^+ = x - u$, the standard smoothness inequality (Ghadimi & Lan, 2013) gives

$$f(x) - f(x-u) \geq \Phi_e(u) := \mu\langle e, u\rangle - \frac{L}{2}\|u\|^2. \tag{1}$$

The function Φ_e captures both the direction and the magnitude of the update: the first term measures progress along the true gradient, and the second accounts for the curvature cost of the step. For a stochastic update $u = T(G)$, the expectation $\mathbb{E}[\Phi_e(T(G))]$ lower-bounds the expected objective decrease. We use this quantity to evaluate an update rule under a given observation distribution.

Equality in (1) holds for a quadratic objective with Hessian LI_d, so this criterion can equal the actual one-step decrease. We provide the construction details in Appendix A.4.

Suppose that the true gradient is available. Under $\|u\| \le R$, the best update points along e, and its length maximizes $\mu r - Lr^2/2$ over $r \in [0, R]$. Define

$$b_\mu := \min\{R, \mu/L\}, \;\; H_\mu := \max_{\|u\| \le R} \Phi_e(u) = \mu b_\mu - \frac{L}{2} b_\mu^2. \tag{2}$$

The unique optimal update is $b_\mu e$. We use H_μ as the full-information benchmark against which to measure the loss caused by observation noise.

3.2 Available information and the minimax model

Now suppose that only one stochastic gradient observation G is available. The observation is unbiased and its variance is at most σ^2, $\sigma \ge 0$. We write $\mathbb{E}_P$ for expectation under P. For a given direction e, the admissible observation distributions form the class

$$\mathcal{P}_e^{\mu,\sigma} = \left\{ P : \mathbb{E}_P G = \mu e,\ \mathbb{E}_P \|G - \mu e\|^2 \le \sigma^2 \right\}. \tag{3}$$

Under the mean constraint $\mathbb{E}_P G = \mu e$, the variance bound is equivalent to $\mathbb{E}_P \|G\|^2 \le \mu^2 + \sigma^2$. No further distributional assumptions are imposed.

Let $\mathcal{T}_R$ denote all Borel measurable maps $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $\|T(z)\| \le R$ for every z. We initially fix the gradient norm μ and allow the update rule to use (μ, σ, L, R), while the gradient direction and the particular observation distribution remain unknown.

For a given rule, define its worst-case descent guarantee and the optimal value over the rule class by

$$W_\mu(T) = \inf_{\|e\|=1} \inf_{P \in \mathcal{P}_e^{\mu,\sigma}} \mathbb{E}_P[\Phi_e(T(G))], \;\; V_\mu = \sup_{T \in \mathcal{T}_R} W_\mu(T). \tag{4}$$

Here $W_\mu(T)$ is the worst-case descent guarantee of a fixed rule, and V_μ is the largest guarantee attainable within $\mathcal{T}_R$. The order of optimization reflects the available information: the rule is chosen first and must work for every admissible direction and observation distribution. We next derive a universal upper bound, construct a rule that attains it, and characterize all optimal rules.

4 Optimal Updates with a Known Gradient Norm

This section solves the single-step decision problem when the gradient norm μ is known. We derive the exact minimax value, construct a rule that attains it, and characterize all optimal rules. Complete proofs appear in Appendices A and B.

4.1 How much progress can one observation guarantee?

Consider a two-point observation law: $G = ae$ with probability p and $G = 0$ otherwise. Unbiasedness requires $pa = \mu$, and the variance is $\mu^2(1-p)/p$. Thus, within this family, the smallest admissible probability of a nonzero observation is $p_\mu := \mu^2/Q_\mu$, where $Q_\mu := \mu^2 + \sigma^2$. The corresponding nonzero atom is $a_\mu e$, with $a_\mu := Q_\mu/\mu$. Let δ_z denote a unit point mass at z, giving

$$P_e^* = (1 - p_\mu)\delta_0 + p_\mu \delta_{a_\mu e}, \tag{5}$$

The distribution P_e^* has mean μe and uses the full variance budget σ^2, concentrating all directional information in its nonzero observation. For any rule T, choose a unit vector e with $\langle e, T(0) \rangle \le 0$. Then $\Phi_e(T(0)) \le 0$, while $\Phi_e(T(a_\mu e)) \le H_\mu$ because $\|T(a_\mu e)\| \le R$. The expected payoff under P_e^* is therefore at most $p_\mu H_\mu$. Since T is arbitrary, this gives $W_\mu(T) \le p_\mu H_\mu$ for every T, and taking the supremum over the rule class yields $V_\mu \le p_\mu H_\mu$.

To attain this upper bound, a rule must guarantee $p_\mu H_\mu$ under every admissible distribution. The two-point construction suggests two requirements: a zero update at the origin and the deterministic optimal update at the informative atom. We therefore consider $T_\eta(z) = \mathrm{clip}_{b_\mu}(\eta z)$, with the gain $\eta > 0$ chosen to obtain a guarantee over the entire moment class. Here $\mathrm{clip}_c(v) = \min\{1, c/\|v\|\}v$ for

$v \neq 0$, and $\mathrm{clip}_c(0) = 0$. The gain can be chosen using a pointwise lower bound whose expectation depends only on the available moments:

$$\Phi_e(T_\eta(z)) \geq \mu\eta\langle e, z\rangle - \lambda_\eta \|z\|^2, \;\; \lambda_\eta = \frac{\mu^2\eta^2}{4H_\mu}. \tag{6}$$

To verify the bound (6), consider the two clipping regimes. When clipping is inactive, it suffices that $\lambda_\eta \geq L\eta^2/2$, which follows from $H_\mu \leq \mu^2/(2L)$. In the clipped region, set $r = \|z\|$. Using $\langle e, z\rangle \leq r$ reduces the required inequality to the nonnegativity of a square:

$$\lambda_\eta r^2 - \mu\eta r + H_\mu = \frac{(\mu\eta r - 2H_\mu)^2}{4H_\mu} \geq 0.$$

Taking expectations in (6) and applying unbiasedness and the second-moment bound yields

$$W_\mu(T_\eta) \geq \mu^2\eta - \frac{\mu^2 Q_\mu}{4H_\mu}\eta^2. \tag{7}$$

The right-hand side is a concave quadratic in the gain. It is maximized at $k_\mu := 2H_\mu/Q_\mu$, where it equals $p_\mu H_\mu$. At this gain, the lower bound matches the universal upper bound $V_\mu \leq p_\mu H_\mu$, yielding the following theorem.

Theorem 1 (Exact single-step minimax guarantee)**.** *For every $\mu > 0$ and $\sigma \geq 0$, the minimax value and an attaining rule are*

$$V_\mu = \frac{\mu^2}{\mu^2 + \sigma^2} H_\mu, \;\; T_\mu(z) = \mathrm{clip}_{b_\mu}(k_\mu z). \tag{8}$$

The complete proof is given in Appendix A.

The result separates the effects of the geometric constraints and observation uncertainty. Smoothness and the update bound determine H_μ, while the noise determines the fraction p_μ that can be retained. This fraction equals $1/2$ when the variance bound equals the squared gradient norm, and equals 1 in the absence of noise.

Allowing internal randomization does not improve the optimal value. The conditional mean of a randomized update remains in the update ball, and concavity of Φ_e ensures that replacing the random update by this mean cannot reduce its expected payoff. Appendix A.4 proves this extension.

4.2 The structure of all optimal rules

Theorem 1 provides one optimal rule. We now characterize every rule attaining the same guarantee, identifying which responses are required by optimality and where freedom remains. Throughout this subsection, assume $\sigma > 0$ and write $\lambda_\mu = \mu^2 H_\mu / Q_\mu^2$.

We begin with the response at the origin. Every optimal rule must satisfy $T(0) = 0$. Otherwise, under the two-point distribution with direction opposite to $T(0)$, the zero observation would contribute a strictly negative payoff and reduce the total guarantee below $p_\mu H_\mu$.

To constrain the response at an arbitrary observation z, we perturb the two-point law by assigning a small mass to z and adjusting the remaining masses and the nonzero atom to preserve the mean and second moment. For sufficiently small $t > 0$, take

$$P_t = (1 - t - q_t)\delta_0 + t\delta_z + q_t\delta_{w_t}, \;\; q_t = \frac{\|\mu e - tz\|^2}{Q_\mu - t\|z\|^2}, \;\; w_t = \frac{\mu e - tz}{q_t}.$$

This distribution has mean μe and second moment Q_μ. Since $q_t \to p_\mu \in (0, 1)$, all three masses are positive for sufficiently small t.

Since T is optimal, it must guarantee $p_\mu H_\mu$ under P_t. Using $T(0) = 0$ and $\Phi_e(u) \leq H_\mu$ for every $\|u\| \leq R$, we obtain

$$p_\mu H_\mu \leq t\Phi_e(T(z)) + q_t H_\mu.$$

Dividing by t gives $\Phi_e(T(z)) \geq H_\mu \frac{p_\mu - q_t}{t}$. Letting $t \downarrow 0$ and using the derivative of q_t at zero yields $\Phi_e(T(z)) \geq \mu k_\mu \langle e, z\rangle - \lambda_\mu \|z\|^2$. Since this inequality holds for every unit vector e, the identity $\inf_{\|e\|=1}\langle e, v\rangle = -\|v\|$ yields

$$\frac{L}{2}\|T(z)\|^2 + \mu\|T(z) - k_\mu z\| \leq \lambda_\mu \|z\|^2. \tag{9}$$

Conversely, if (9) holds for every z, then for every unit vector e, $\Phi_e(T(z)) \geq \mu k_\mu \langle e, z\rangle - \lambda_\mu \|z\|^2$. Taking expectations under any admissible distribution gives $W_\mu(T) \geq p_\mu H_\mu$, and the reverse inequality follows from Theorem 1. This proves the following characterization.

Theorem 2 (Characterization of all optimal rules)**.** *Suppose $\sigma > 0$. A rule $T \in \mathcal{T}_R$ satisfies $W_\mu(T) = V_\mu$ if and only if* (9) *holds for every $z \in \mathbb{R}^d$.*

The complete proof is given in Appendix B.

The pointwise condition yields two structural consequences. First, it determines the local behavior near the origin. Dropping the nonnegative quadratic term gives $\|T(z) - k_\mu z\| \leq (\lambda_\mu/\mu)\|z\|^2$. Hence

$$T(z) = k_\mu z + O(\|z\|^2) \text{ as } z \to 0, \;\; DT(0) = k_\mu I_d.$$

Thus every optimal rule applies the same first-order scaling to small observations. Although the rule class requires only measurability, optimality itself forces differentiability at the origin with the same derivative.

Second, it fixes the response at the informative atom. Under the two-point distribution P_e^*, optimality requires the payoff at $a_\mu e$ to equal H_μ. Since $b_\mu e$ uniquely maximizes Φ_e over the update ball, the response must satisfy $T(a_\mu e) = b_\mu e$. Appendix B.2 gives the full derivation of both consequences.

When $\sigma = 0$, the observation equals the true gradient. Optimality then requires only $T(\mu e) = b_\mu e$ for every unit direction, and imposes no condition at the origin. This case is treated in Appendix A.3.

4.3 Examples of optimal update design

We illustrate two uses of the minimax model: selecting optimal parameters within a prescribed rule family and designing an optimal rule under a specified noise distribution.

Optimal parameters for Euclidean clipping. Consider $T(z) = \mathrm{clip}_c(\eta z)$ with $\eta > 0$ and $0 < c \leq R$. When $\sigma > 0$, the local condition in Theorem 2 requires $\eta = k_\mu$, while the response at the informative atom requires $c \geq b_\mu$. These conditions are also sufficient.

Corollary 3 (All optimal Euclidean clipping parameters)**.** *Suppose $\sigma > 0$. Euclidean clipping attains the optimum V_μ over the general rule class if and only if $\eta = k_\mu$ and $b_\mu \leq c \leq R$.*

The complete proof is given in Appendix E.1.

For $\sigma > 0$ and $R < \mu/L$, the update bound constrains the best deterministic step. The optimal pair is unique, with $c = R$ and $k_\mu = (2\mu R - LR^2)/(\mu^2 + \sigma^2)$. Tightening the update bound therefore also requires a smaller gain before clipping. The constraint affects both the truncation of large observations and the scaling of small ones, so the two parameters must be chosen jointly. Appendix E further treats active and inactive bounds and the zero-variance case.

A known Gaussian noise model. Continue to assume that the gradient norm is known and its direction is unknown, but now specify $G = \mu e + \tau Z$, where $Z \sim \mathcal{N}(0, I_d)$, $\tau^2 = \sigma^2/d$, and $\sigma > 0$. The total variance remains σ^2, while the admissible distributions are restricted to a Gaussian family with an unknown mean direction.

Appendix F derives the optimal rule and value in every dimension. In one dimension, the rule is

$$T_{\mathrm{G}}(z) = \mathrm{clip}_R\Big(\frac{\mu}{L}\tanh\Big(\frac{\mu z}{\sigma^2}\Big)\Big).$$

For every finite dimension and positive μ, L, R, σ, its optimal value satisfies $V_\mu < V_\mu^{\mathrm{G}} < H_\mu$. The additional Gaussian information thus strictly improves the optimal guarantee, while a loss relative to the full-information benchmark remains. These two examples illustrate the distinct roles of the parameterized rule family and the available distributional information.

5 FIXED UPDATES WITHOUT KNOWING THE CURRENT GRADIENT NORM

The preceding section allows the rule to depend on the current gradient norm μ. During optimization, this norm is generally unknown and changes with the iterate. We therefore seek a single rule that provides a uniform descent guarantee whenever $\mu \geq m$, for a prescribed threshold $m > 0$. The rule may use (m, σ, L, R), but not the current norm. We first solve this uniform design problem and then establish a finite-time stationarity guarantee for the resulting fixed rule. Complete proofs appear in Appendices C and D.

5.1 FROM BOUNDARY CALIBRATION TO AN OPTIMAL UNIFORM GUARANTEE

Fix a positive calibration threshold m, and require a single rule to work for every $\mu \geq m$. The corresponding minimax problem is

$$V_{\geq m} = \sup_{T \in \mathcal{T}_R} \inf_{\mu \geq m} W_\mu(T). \tag{10}$$

The rule may use m, but it cannot depend on the unknown current norm μ. Since the range includes $\mu = m$, the uniform guarantee cannot exceed the boundary optimum: $V_{\geq m} \leq V_m$. The question is whether this upper bound is attained by the rule $T_m(z) = \mathrm{clip}_{b_m}(k_m z)$, which by Theorem 1 is optimal at $\mu = m$.

To answer this question, fix $\mu \geq m$, a unit direction e, and an arbitrary admissible observation $G = \mu e + \xi$, where $\mathbb{E}\xi = 0$ and $\mathbb{E}\|\xi\|^2 \leq \sigma^2$. Keeping the noise distribution fixed, we compare the payoffs as the mean varies from me to μe. For each realization of ξ, define $u_s := T_m(se + \xi)$, $q(s) := \langle e, u_s \rangle$, and $h_\xi(s) := sq(s) - \frac{L}{2}\|u_s\|^2$. Here $q(s)$ is the update component along the gradient direction, and $h_\xi(s)$ is the payoff along the shift.

The directional component $q(s)$ of Euclidean clipping is nondecreasing in s. In the unclipped region, $h'_\xi(s) = q(s) + k_m(s - Lq(s)) \geq q(s)$, because $q(s) \leq b_m$ and $Lb_m \leq m \leq s$. In the clipped region, the update norm is constant, so $h'_\xi(s) = q(s) + sq'(s) \geq q(s)$. Both relations hold almost everywhere. Integrating along the shift and using $q(s) \geq q(m)$ gives

$$h_\xi(\mu) \geq h_\xi(m) + (\mu - m)q(m).$$

Let $F(s) = \mathbb{E}[h_\xi(s)]$. The boundary observation $me + \xi$ satisfies the moment conditions at norm m, so $F(m) \geq V_m$. Since $F(m) = m\mathbb{E}[q(m)] - \frac{L}{2}\mathbb{E}\|u_m\|^2$, we also have $F(m) \leq m\mathbb{E}[q(m)]$. Combining these two estimates,

$$F(\mu) \geq F(m) + (\mu - m)\mathbb{E}[q(m)] \geq \frac{\mu}{m}F(m) \geq \frac{\mu}{m}V_m.$$

The direction and noise distribution were arbitrary, so the comparison holds over the entire admissible class. Appendix C.1 gives the full mean-shift argument, including the clipping transition points.

Theorem 4 (Exact threshold calibration)**.** *For every $m > 0$ and $\sigma \geq 0$, the fixed rule T_m satisfies*

$$W_\mu(T_m) \geq \frac{\mu}{m}V_m \text{ for all } \mu \geq m, \;\; V_{\geq m} = V_m. \tag{11}$$

The complete proof is given in Appendix C.2.

The boundary value therefore determines the best uniform guarantee over the entire region, and the rule calibrated at m attains it. The theorem provides a lower bound on the guarantee that increases linearly with the gradient norm, even though the rule itself does not depend on μ.

When $\sigma > 0$, the zero rule is the only fixed rule whose worst-case smoothness guarantee is nonnegative for every $\mu > 0$; see Appendix C.3. A positive threshold therefore identifies a region where nonzero uniform progress is possible. To analyze a full optimization trajectory, we must also control updates below that threshold.

5.2 FROM ONE-STEP GUARANTEES TO FINITE-TIME STATIONARITY

With $b_m := \min\{R, m/L\}$ and $k_m := (2mb_m - Lb_m^2)/(m^2 + \sigma^2)$ fixed in advance, consider

$$x_{t+1} = x_t - T_m(G_t) = x_t - \mathrm{clip}_{b_m}(k_m G_t). \tag{12}$$

Suppose f is bounded below by $f_{\inf}$ and x_0 is fixed. Let $\mathcal{F}_t$ denote the history before observing G_t, and assume

$$\mathbb{E}[G_t \mid \mathcal{F}_t] = \nabla f(x_t), \;\; \mathbb{E}\big[\|G_t - \nabla f(x_t)\|^2 \mid \mathcal{F}_t\big] \le \sigma^2. \tag{13}$$

These assumptions allow the observation to depend on the history. To analyze the full trajectory, we must control both the descent above the threshold and the possible loss of progress below it.

Consider a single update. Write the true gradient as g, its norm as μ, and the expected smoothness-based payoff as $D := \mathbb{E}[\langle g, T_m(G)\rangle - (L/2)\|T_m(G)\|^2]$. We treat the two cases separately.

If $\mu \ge m$, Theorem 4 and $V_m = k_m m^2/2$ give $D \ge k_m m\mu/2$. If $\mu \le m$, the pointwise certificate for T_m gives

$$\frac{L}{2}\|T_m(z)\|^2 + m\|T_m(z) - k_m z\| \le \lambda_m \|z\|^2,$$

where $p_m := m^2/(m^2+\sigma^2)$ and $\lambda_m := p_m k_m/2$. Since $\|g\| \le m$,

$$\langle g, T_m(z)\rangle - \frac{L}{2}\|T_m(z)\|^2 = k_m\langle g, z\rangle + \langle g, T_m(z) - k_m z\rangle - \frac{L}{2}\|T_m(z)\|^2 \ge k_m\langle g, z\rangle - \lambda_m\|z\|^2.$$

Taking expectations and using $\mathbb{E}G = g$, $\mathbb{E}\|G\|^2 \le \mu^2 + \sigma^2$, and $\lambda_m = p_m k_m/2$ gives

$$D \ge k_m\mu^2 - \frac{p_m k_m}{2}(\mu^2 + \sigma^2).$$

Since $p_m\sigma^2 = (1-p_m)m^2$, this simplifies to

$$D \ge \frac{k_m}{2}\left[(2-p_m)\mu^2 - (1-p_m)m^2\right]. \tag{14}$$

This bound controls the magnitude of a possible negative expected payoff below the threshold. It follows directly from the constructed pointwise bound and therefore also holds when $\sigma = 0$. Appendix D.1 gives the full derivation.

To combine the two regions, set $r := \mu/m \in [0,1]$. The bound above the threshold is linear in μ, so we lower-bound the quadratic expression below the threshold by an affine function:

$$(2-p_m)r^2 - (1-p_m) = 2r^2 - 1 + p_m(1-r^2) \ge 2r^2 - 1 \ge r - \frac{9}{8}.$$

The last inequality follows from $2r^2 - r + 1/8 = 2(r-1/4)^2 \ge 0$. Together with the bound above the threshold, this gives, for every $\mu \ge 0$,

$$D \ge \frac{k_m m}{2}\left(\mu - \frac{9}{8}m\right). \tag{15}$$

Applying this inequality conditionally at each iteration and using smoothness gives

$$\mathbb{E}[f(x_t) - f(x_{t+1}) \mid \mathcal{F}_t] \ge \frac{k_m m}{2}\left(\|\nabla f(x_t)\| - \frac{9}{8}m\right).$$

Taking expectations and summing over time telescopes the objective differences. With $\Delta := f(x_0) - f_{\inf}$,

$$\frac{k_m m}{2}\sum_{t=0}^{N-1}\left(\mathbb{E}\|\nabla f(x_t)\| - \frac{9}{8}m\right) \le \Delta.$$

Let J be independent of the trajectory and uniform on $\{0,\dots,N-1\}$. Then $\mathbb{E}\|\nabla f(x_J)\| = N^{-1}\sum_{t=0}^{N-1}\mathbb{E}\|\nabla f(x_t)\|$, so the average bound becomes a stationarity guarantee.

Theorem 5 (Stationarity with fixed parameters)**.** *Under the preceding assumptions, let $N \ge 1$ and let J be independent of the trajectory and uniform on $\{0,\dots,N-1\}$. Then*

$$\mathbb{E}\|\nabla f(x_J)\| \le \frac{9}{8}m + \frac{2\Delta}{k_m m N}. \tag{16}$$

For any desired accuracy $\varepsilon > 0$, choosing $m = 4\varepsilon/9$ and an integer $N \ge \max\{1, 4\Delta/(k_m m\varepsilon)\}$ ensures $\mathbb{E}\|\nabla f(x_J)\| \le \varepsilon$.

The complete proof is given in Appendix D.3.

The bound separates the calibration accuracy from the iteration budget. For a fixed m, increasing N reduces the second term; a smaller prescribed accuracy is obtained by choosing a smaller threshold before the run. Once calibrated, the rule remains fixed throughout optimization and requires no online estimate of the gradient norm. Appendix D.2 gives a sharper coefficient in place of $9/8$.

6 NUMERICAL EXPERIMENTS

The experiments illustrate the exact minimax value, the coupling between gain and update radius, and the response of the fixed rule to different noise shapes. Numerical moment programs check the analytical values; controlled oracles then isolate the effect of distributional information on single-step descent and optimization trajectories. Protocols and complete comparisons are in Appendix G.

6.1 EXACT VALUES AND JOINT CALIBRATION

For each scalar rule, a finite-support linear program (LP) minimizes the expected smoothness payoff subject to the mean and variance constraints. Its optimum is an upper bound on the unrestricted worst-case payoff. With $L = 1$, $\mu \in \{0.1, 0.2, \ldots, 4\}$, $\sigma \in \{0.5, 2, 5\}$, and $R \in \{0.5, 2\}$, supports containing the least-favorable atom recover V_μ to a maximum absolute discrepancy of 5.23×10^{-11}. Independent grids omitting that atom have a maximum relative discrepancy of 5.54×10^{-6}.

Figure 1 illustrates both calibration results. When $LR/\mu < 1$, the update cap changes the optimal gain: the jointly calibrated rule attains V_μ, while applying the unconstrained SNR gain before clipping gives a smaller worst-case payoff. When $LR/\mu \geq 1$, both choices attain the same value. For the fixed rule calibrated at $m = 1$, the LP values agree with V_m at the boundary and lie above the guarantee $(\mu/m)V_m$ for larger gradient norms.

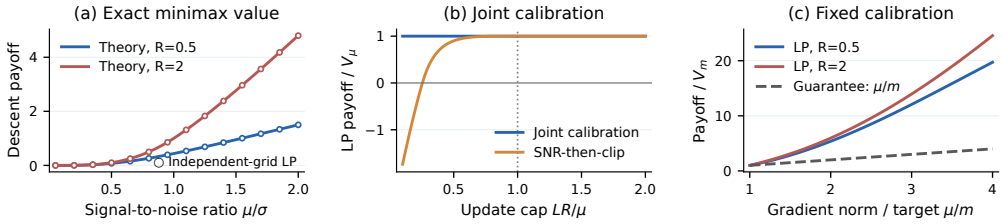

Figure 1: **Exact minimax calibration**, with $L = 1$ and $\sigma = 2$. (a) Theory and independent-grid LP values. (b) Joint calibration and SNR-then-clip at $\mu = 1$; the vertical line marks $LR/\mu = 1$. (c) Fixed calibration at $m = 1$. LP values are support-restricted upper bounds; the dashed curve in (c) is the analytical lower guarantee.

6.2 THE EFFECT OF NOISE SHAPE ON SINGLE-STEP DESCENT

To vary noise shape without changing the available moment information, use $G = g + \xi_q$, where ξ_q is zero with probability $1 - q$ and equals $\pm\sigma v/\sqrt{q}$ with probability $q/2$ each, for a fixed unit vector v. Thus $\mathbb{E}\xi_q = 0$ and $\mathbb{E}\|\xi_q\|^2 = \sigma^2$ for every q. All expectations in Figure 2 are exact three-atom sums of actual objective decrease.

On $f(x) = x^2/2$ at $x = m = 1$, with $\sigma = 2$ and $R = 0.2$, joint calibration uses gain 0.072, compared with 0.2 for SNR-then-clip. At $q = 1$, the two possible observations are 3 and -1. Joint calibration maps them to 0.2 and -0.072, retaining magnitude information and giving decrease $0.052704 \geq V_m = 0.036$. The SNR rule maps them to 0.2 and -0.2, yielding decrease -0.020 through cancellation of the linear terms. This illustrates why the cap enters the optimal gain.

The same noise family is evaluated on a regularized logistic objective at a point with gradient norm $m = 0.05$, using a valid global smoothness bound and variance $\sigma^2 = 0.562515093$. Joint calibration again stays above its guarantee. The SNR rule achieves larger descent on these particular logistic noise laws, consistent with a minimax criterion over the entire moment class rather than pointwise dominance. Appendix G also reports a common-pool moment stress test on this objective.

6.3 FIXED CALIBRATION ALONG CONTROLLED LOGISTIC TRAJECTORIES

The logistic objective uses the Wisconsin Diagnostic Breast Cancer training data. Starting at zero, we run 4000 updates with the same artificial oracle $G_t = \nabla f(w_t) + \xi_{q,t}$, independently sampled over time. The direction v is the normalized initial gradient and remains fixed. We retain $q \in$

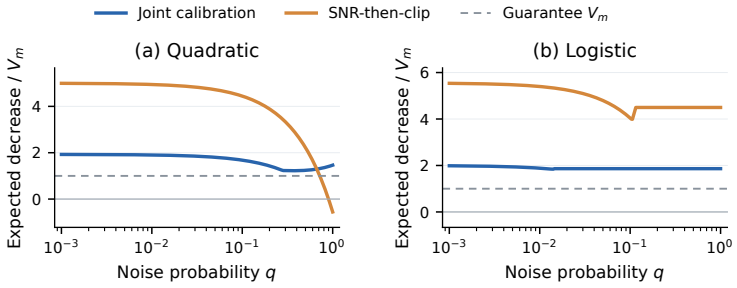

Figure 2: **Fixed moments, varying noise shape.** Actual expected decrease divided by V_m. Both panels use $LR/m = 0.2$, with problem-specific L, m, σ stated in the text. The gain and radius remain fixed as q varies. The dashed line marks the joint rule's guarantee.

$\{0.001, 0.003, 0.01, 0.03, 0.1, 1\}$ and $LR/m \in \{0.2, 2\}$. The variance is identical across all settings. Figure 3 reports the full-horizon average gradient norm and clipping frequency over 64 paired seeds.

The rule clips precisely when $\|G_t\| > b_m/k_m$. Here these thresholds are approximately 6.28 and 11.30. Decreasing q increases each nonzero perturbation's magnitude while decreasing its frequency. In the active-cap setting, clipping occurs at $q \leq 0.01$ and is absent at the larger tested probabilities; under the inactive cap it occurs at $q \leq 0.003$. Thus the experiment exhibits both the clipped and linear regimes of the same fixed rule, without changing its moment calibration. The appendix reports natural-minibatch experiments and comparisons with independently tuned step-decay controls.

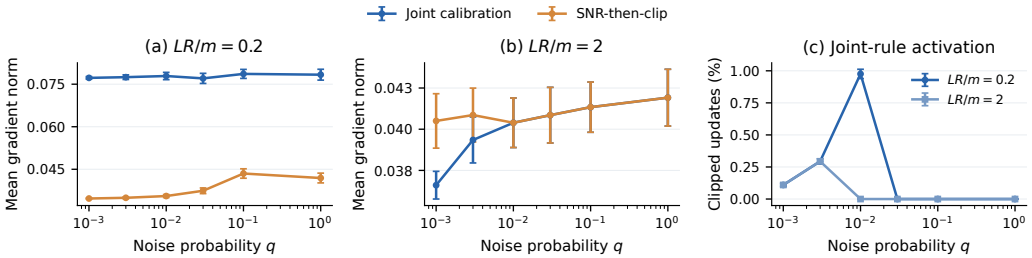

Figure 3: **Controlled logistic trajectories.** (a,b) Full-horizon average gradient norm for the two fixed rules. (c) Fraction of updates clipped by the joint rule. Error bars are pointwise 95% Student-t intervals over 64 paired seeds. The conditional noise variance is fixed; only q changes.

7 Conclusion and Limitations

This work determines the exact minimax descent guarantee from a single unbiased gradient observation under a variance bound and an update norm constraint, and characterizes all attaining rules. Optimality constrains the local response and couples the gain with the update cap. When the gradient norm is unknown, calibration at a prescribed threshold attains the best uniform guarantee above it and yields a finite-time stationarity bound.

The exact optimality results concern one-step descent; designing rules that are optimal over an entire trajectory remains open. The calibration also requires known smoothness and variance bounds. Estimating these quantities online while preserving the guarantees, using multiple observations or past information, and extending the analysis to non-Euclidean update constraints are natural directions for further study.

AI USE STATEMENT

AI was used to assist with coding and review feedback. The authors take responsibility for the final manuscript.

REFERENCES

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by random stepsizes: Hedging, equalization, and the Arcsine stepsize schedule, 2024. URL https://arxiv.org/abs/2412.05790. Revised May 2026; to appear in Foundations of Computational Mathematics.

Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1774–1782, 2025. URL https://proceedings.mlr.press/v258/armacki25a.html.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6: 1345–1382, 2005. URL https://www.jmlr.org/papers/v6/banerjee05a.html.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 2 edition, 1985. doi: 10.1007/978-1-4757-4286-2.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005. doi: 10.1137/S1052623401399903. URL https://www.mit.edu/~dbertsim/papers/MomentProblems/Optimal-inequalities-in-probability-theory-A-convex-optimization-approach-SIAM15.pdf.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454. URL https://www.numdam.org/articles/10.1214/11-AIHP454/.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/26901debb30ea03f0aa833c9de6b81e9-Abstract.html.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010. doi: 10.1287/opre.1090.0741. URL https://pubsonline.informs.org/doi/10.1287/opre.1090.0741.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. doi: 10.1214/16-AOS1440. URL https://arxiv.org/abs/1509.05845.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1–2):451–482, 2014. URL https://arxiv.org/abs/1206.3209.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://arxiv.org/abs/1309.5549.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.10785.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-35/issue-1/Robust-Estimation-of-a-Location-Parameter/10.1214/aoms/1177703732.full.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2413–2421, 2025. URL https://proceedings.mlr.press/v258/hubler25a.html.

Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):81–107, 2016. doi: 10.1007/s10107-015-0949-3. URL https://arxiv.org/abs/1406.5468.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363, 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

Ibrahim Merad and Stéphane Gaïffas. Robust stochastic optimization via gradient quantile clipping. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2309.17316.

National Institute of Standards and Technology. NIST Digital Library of Mathematical Functions. URL https://dlmf.nist.gov/. Sections 10.25 and 10.29; accessed September 20, 2026.

Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. In *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 343–351, 2013. URL https://proceedings.mlr.press/v28/schaul13.html.

Alexander Shapiro. On duality theory of conic linear problems. In Miguel A. Goberna and Marco A. López (eds.), *Semi-Infinite Programming*, volume 57 of *Nonconvex Optimization and Its Applications*. Springer, 2001. doi: 10.1007/978-1-4757-3403-4_7.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 2017. doi: 10.1007/s10107-016-1009-3. URL https://arxiv.org/abs/1502.05666.

Bart P. G. van Parys and Bert Zwart. Optimal estimators for heavy-tailed mean estimation via convex analysis, 2026. URL https://arxiv.org/abs/2606.27899. Version 1.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.11881.

A Proof of the Exact Minimax Guarantee

This appendix proves Theorem 1. The lower bound follows from a pointwise inequality for clipped updates, and the upper bound follows from a two-point observation law. Throughout, $\mu, L, R > 0$ and $\sigma \geq 0$. Recall that $b_\mu = \min\{R, \mu/L\}$, $H_\mu = \mu b_\mu - L b_\mu^2/2$, and $Q_\mu = \mu^2 + \sigma^2$. We use $p_\mu = \mu^2/Q_\mu$, $k_\mu = 2H_\mu/Q_\mu$, $a_\mu = Q_\mu/\mu$, and $\lambda_\mu = \mu^2 H_\mu/Q_\mu^2$. In particular, $H_\mu > 0$ and $Lb_\mu \leq \mu$.

A.1 A pointwise certificate and the optimal gain

The proof uses a quadratic lower certificate whose expectation is controlled by the prescribed moments. This is the weak-duality principle for moment problems (Shapiro, 2001); here the certificate and a matching distribution are constructed explicitly.

Lemma A.1 (Clipping certificate)**.** *For any $\eta > 0$, let $T_\eta(z) = \mathrm{clip}_{b_\mu}(\eta z)$ and $\lambda_\eta = \mu^2\eta^2/(4H_\mu)$. Then, for every unit vector e and every $z \in \mathbb{R}^d$,*

$$\Phi_e(T_\eta(z)) \geq \mu\eta\langle e, z\rangle - \lambda_\eta \|z\|^2.$$

Proof. Write $r = \|z\|$. If $\eta r \leq b_\mu$, then $T_\eta(z) = \eta z$, and the difference between the two sides is

$$\left(\lambda_\eta - \frac{L}{2}\eta^2\right) r^2 = \frac{\eta^2(\mu - Lb_\mu)^2}{4H_\mu} r^2 \geq 0.$$

Here we used $\mu^2 - 2LH_\mu = (\mu - Lb_\mu)^2$.

If $\eta r > b_\mu$, then $T_\eta(z) = b_\mu z/r$. Since $\eta - b_\mu/r > 0$ and $\langle e, z\rangle \leq r$,

$$\begin{aligned}\Phi_e(T_\eta(z)) &= \mu\eta\langle e, z\rangle - \mu\left(\eta - \frac{b_\mu}{r}\right)\langle e, z\rangle - \frac{L}{2}b_\mu^2 \\ &\geq \mu\eta\langle e, z\rangle - \mu\eta r + H_\mu.\end{aligned}$$

The remaining inequality follows from

$$\lambda_\eta r^2 - \mu\eta r + H_\mu = \frac{(\mu\eta r - 2H_\mu)^2}{4H_\mu} \geq 0.$$

The two cases cover every z, including the clipping boundary. □

This proves the certificate (6). For each $P \in \mathcal{P}_e^{\mu,\sigma}$, the moment conditions imply $\mathbb{E}_P\|G\|^2 \leq Q_\mu$. Taking expectations in the certificate gives

$$\mathbb{E}_P[\Phi_e(T_\eta(G))] \geq \mu^2\eta - \frac{\mu^2 Q_\mu}{4H_\mu}\eta^2.$$

The bound is independent of e and P, so it also bounds $W_\mu(T_\eta)$ from below, as stated in (7). The quadratic expression on the right has its unique maximum over $\eta > 0$ at $\eta = k_\mu$, where its value is $p_\mu H_\mu$. Consequently, $T_\mu(z) = \mathrm{clip}_{b_\mu}(k_\mu z)$ satisfies $W_\mu(T_\mu) \geq p_\mu H_\mu$. At this gain, $\lambda_{k_\mu} = \lambda_\mu$.

A.2 A matching least-favorable law

Assume first that $\sigma > 0$, so $0 < p_\mu < 1$. Consider the law $P_e^* = (1 - p_\mu)\delta_0 + p_\mu\delta_{a_\mu e}$. The identities $p_\mu a_\mu = \mu$ and $p_\mu a_\mu^2 = Q_\mu$ imply

$$\mathbb{E}_{P_e^*} G = \mu e, \quad \mathbb{E}_{P_e^*}\|G - \mu e\|^2 = \sigma^2.$$

Thus P_e^* satisfies the moment conditions (3).

Fix any $T \in \mathcal{T}_R$. If $T(0) \neq 0$, choose $e = -T(0)/\|T(0)\|$; otherwise choose any unit vector. This choice ensures $\Phi_e(T(0)) \leq 0$. The deterministic benchmark in (2) also gives $\Phi_e(T(a_\mu e)) \leq H_\mu$. Hence

$$\begin{aligned}W_\mu(T) &\leq \mathbb{E}_{P_e^*}[\Phi_e(T(G))] \\ &= (1 - p_\mu)\Phi_e(T(0)) + p_\mu\Phi_e(T(a_\mu e)) \leq p_\mu H_\mu.\end{aligned}$$

Taking the supremum over T, and combining this upper bound with Appendix A.1, proves $V_\mu = p_\mu H_\mu$ and the optimality of T_μ when $\sigma > 0$.

The same two-point law attains the guarantee for T_μ in every direction. Indeed, $T_\mu(0) = 0$ and

$$k_\mu a_\mu = \frac{2H_\mu}{\mu} = b_\mu \left(2 - \frac{Lb_\mu}{\mu}\right) \geq b_\mu.$$

It follows that $T_\mu(a_\mu e) = b_\mu e$, and therefore $\mathbb{E}_{P_e^*}[\Phi_e(T_\mu(G))] = p_\mu H_\mu$.

A.3 Zero variance

If $\sigma = 0$, the moment conditions force $G = \mu e$ almost surely. Every rule therefore satisfies $W_\mu(T) = \inf_{\|e\|=1} \Phi_e(T(\mu e)) \leq H_\mu$. The deterministic benchmark has the unique maximizer $b_\mu e$: for $\|u\| = r$, Cauchy–Schwarz gives $\Phi_e(u) \leq \mu r - Lr^2/2$, and the latter expression has its unique maximum on $[0, R]$ at $r = b_\mu$.

For the calibrated rule, $k_\mu \mu = 2H_\mu/\mu \geq b_\mu$, so $T_\mu(\mu e) = b_\mu e$. Hence $V_\mu = H_\mu = p_\mu H_\mu$, which completes the proof of Theorem 1 for all $\sigma \geq 0$.

More generally, a rule $T \in \mathcal{T}_R$ is optimal in the zero-variance case if and only if

$$T(\mu e) = b_\mu e \ \text{ for every unit vector } e.$$

Necessity follows because each payoff is bounded above by H_μ, while their infimum must equal H_μ; uniqueness of the deterministic maximizer then determines the update. Sufficiency is immediate. The rule is unrestricted away from this observation sphere, apart from measurability and the norm bound.

A.4 Randomized rules and tightness of the smoothness bound

A randomized rule can be represented by a Borel probability kernel $K(z, \mathrm{d}u)$ supported on $\{u : \|u\| \leq R\}$. Its conditional mean $\overline{T}(z) = \int u\, K(z, \mathrm{d}u)$ is measurable and satisfies $\|\overline{T}(z)\| \leq R$. Since Φ_e is concave,

$$\Phi_e(\overline{T}(z)) \geq \int \Phi_e(u)\, K(z, \mathrm{d}u).$$

After integration over any admissible observation law, the deterministic rule $\overline{T}$ has at least as large an expected payoff as the randomized rule. This comparison holds for every direction and every law, so randomization cannot improve the minimax value.

The smoothness lower bound is also sharp as an objective decrease. Fix a point x, a unit vector e, and a scalar f_0, and define

$$f_e(y) = f_0 + \mu\langle e, y - x\rangle + \frac{L}{2}\|y - x\|^2.$$

This function has gradient μe at x and Hessian LI_d. For every update u,

$$f_e(x) - f_e(x - u) = \mu\langle e, u\rangle - \frac{L}{2}\|u\|^2 = \Phi_e(u).$$

Thus equality holds in (1) for every observation and every rule. In particular, the least-favorable construction above realizes the minimax bound as an actual expected one-step objective decrease for an L-smooth quadratic objective.

B Characterization of All Optimal Rules

Throughout this appendix, $\sigma > 0$. Recall that $Q_\mu = \mu^2 + \sigma^2$, $p_\mu = \mu^2/Q_\mu$, $k_\mu = 2H_\mu/Q_\mu$, $\lambda_\mu = \mu^2 H_\mu/Q_\mu^2$, and $a_\mu = Q_\mu/\mu$. The strict inequality $p_\mu < 1$ allows moment-preserving perturbations of the two-point law and is essential for the necessity argument.

B.1 Proof of Theorem 2

Proof. We first prove sufficiency. Suppose $T \in \mathcal{T}_R$ satisfies the pointwise condition (9). For every unit vector e and every observation z,

$$\begin{aligned}\Phi_e(T(z)) &= \mu k_\mu \langle e, z\rangle + \mu\langle e, T(z) - k_\mu z\rangle - \frac{L}{2}\|T(z)\|^2 \\ &\geq \mu k_\mu \langle e, z\rangle - \mu\|T(z) - k_\mu z\| - \frac{L}{2}\|T(z)\|^2 \\ &\geq \mu k_\mu \langle e, z\rangle - \lambda_\mu \|z\|^2.\end{aligned}$$

Taking expectations under any $P \in \mathcal{P}_e^{\mu,\sigma}$ yields

$$\mathbb{E}_P[\Phi_e(T(G))] \geq \mu^2 k_\mu - \lambda_\mu Q_\mu = \frac{\mu^2 H_\mu}{Q_\mu} = V_\mu.$$

It follows that $W_\mu(T) \geq V_\mu$, with equality by the definition of V_μ.

For necessity, suppose $W_\mu(T) = V_\mu = p_\mu H_\mu$. First, $T(0) = 0$. Otherwise, choose a direction opposite to $T(0)$ and use the two-point law $P_e^* = (1 - p_\mu)\delta_0 + p_\mu \delta_{a_\mu e}$. Its zero observation has strictly negative payoff, its nonzero observation has payoff at most H_μ, and its mass at zero is positive. Its expected payoff would therefore be strictly less than $p_\mu H_\mu$, a contradiction.

Now fix any unit vector e and any $z \in \mathbb{R}^d$. For sufficiently small $t > 0$, define

$$q_t = \frac{\|\mu e - tz\|^2}{Q_\mu - t\|z\|^2}, \;\; w_t = \frac{\mu e - tz}{q_t},$$

and let $P_t = (1 - t - q_t)\delta_0 + t\delta_z + q_t \delta_{w_t}$. As $t \downarrow 0$, the denominator tends to $Q_\mu > 0$, the numerator tends to $\mu^2 > 0$, and $q_t \to p_\mu \in (0, 1)$. Consequently, $q_t > 0$ and $1 - t - q_t > 0$ for all sufficiently small t, so P_t is a probability law. Direct calculation gives

$$\mathbb{E}_{P_t} G = tz + q_t w_t = \mu e, \;\; \mathbb{E}_{P_t}\|G\|^2 = t\|z\|^2 + \frac{\|\mu e - tz\|^2}{q_t} = Q_\mu.$$

Thus $P_t \in \mathcal{P}_e^{\mu,\sigma}$.

Optimality, the identity $T(0) = 0$, and the deterministic upper bound $\Phi_e(u) \leq H_\mu$ for $\|u\| \leq R$ imply

$$p_\mu H_\mu \leq t\Phi_e(T(z)) + q_t \Phi_e(T(w_t)) \leq t\Phi_e(T(z)) + q_t H_\mu.$$

Hence $\Phi_e(T(z)) \geq H_\mu(p_\mu - q_t)/t$. The right derivative of q_t at zero is

$$q_0' = -\frac{2\mu\langle e, z\rangle}{Q_\mu} + \frac{\mu^2\|z\|^2}{Q_\mu^2}.$$

Taking $t \downarrow 0$ therefore gives

$$\Phi_e(T(z)) \geq \mu k_\mu \langle e, z\rangle - \lambda_\mu \|z\|^2.$$

This limit requires no continuity of T, because the moving atom's payoff was bounded by H_μ before taking the limit.

Rearranging the last inequality yields

$$\frac{L}{2}\|T(z)\|^2 + \mu\langle e, k_\mu z - T(z)\rangle \leq \lambda_\mu \|z\|^2.$$

It holds for every unit vector e. Taking the supremum over e gives

$$\frac{L}{2}\|T(z)\|^2 + \mu\|T(z) - k_\mu z\| \leq \lambda_\mu \|z\|^2,$$

which is exactly (9). □

B.2 THE FORCED LOCAL GAIN AND THE INFORMATIVE OBSERVATION

Let T be any optimal rule. Dropping the nonnegative quadratic term from (9) gives the global residual estimate

$$\|T(z) - k_\mu z\| \le \frac{\lambda_\mu}{\mu}\|z\|^2 \text{ for all } z \in \mathbb{R}^d. \tag{B.1}$$

At $z = 0$, this implies $T(0) = 0$. For $z \neq 0$, dividing by $\|z\|$ and letting $z \to 0$ proves Fréchet differentiability at the origin, with $DT(0) = k_\mu I_d$. Equivalently, $T(z) = k_\mu z + O(\|z\|^2)$. This conclusion follows from optimality even though the rule class imposes only measurability.

The update at the informative observation of the two-point law is also determined. For every unit vector e, optimality and $T(0) = 0$ give

$$p_\mu H_\mu \le \mathbb{E}_{P_e^*}[\Phi_e(T(G))] = p_\mu \Phi_e(T(a_\mu e)) \le p_\mu H_\mu.$$

Thus $\Phi_e(T(a_\mu e)) = H_\mu$. The unique maximizer of Φ_e over the radius-R ball is $b_\mu e$, so

$$T(a_\mu e) = b_\mu e \text{ for every unit vector } e. \tag{B.2}$$

Equations (B.1) and (B.2) will determine the optimal parameters within the clipping family in Appendix E.

C FIXED CALIBRATION WITHOUT THE CURRENT GRADIENT NORM

This appendix proves Theorem 4 and the necessity of a positive calibration threshold for a nonzero fixed rule. Throughout, $m > 0$, $b_m = \min\{R, m/L\}$, $H_m = mb_m - Lb_m^2/2$, and $Q_m = m^2 + \sigma^2$. Write $p_m = m^2/Q_m$, $k_m = 2H_m/Q_m$, and $V_m = p_m H_m = k_m m^2/2$, as follows from Theorem 1. The rule calibrated at the threshold is $T_m(z) = \mathrm{clip}_{b_m}(k_m z)$. The first two subsections allow $\sigma = 0$.

C.1 A COMPARISON UNDER SHIFTS OF THE MEAN

Fix a unit vector e and a vector $\xi \in \mathbb{R}^d$. For $s \ge m$, define $u_s = T_m(se + \xi)$, $q(s) = \langle e, u_s\rangle$, and $h_\xi(s) = sq(s) - L\|u_s\|^2/2$. These quantities describe the update and its descent bound when the gradient norm changes while the centered observation remains fixed.

Lemma C.1 (Comparison under a shift of the mean)**.** *For every $\xi \in \mathbb{R}^d$ and $\mu \ge m$,*

$$h_\xi(\mu) \ge h_\xi(m) + (\mu - m)\langle e, u_m\rangle. \tag{C.1}$$

Proof. Euclidean clipping is the projection onto a closed ball and is nonexpansive. Thus $s \mapsto u_s$ is k_m-Lipschitz, and u_s, $q(s)$, and $h_\xi(s)$ are absolutely continuous on every bounded interval. We show that $q'(s) \ge 0$ and $h_\xi'(s) \ge q(s)$ almost everywhere.

If $k_m\|se + \xi\| < b_m$, then $u_s = k_m(se + \xi)$, so $u_s' = k_m e$ and $q'(s) = k_m$. Consequently,

$$h_\xi'(s) = q(s) + k_m\big(s - Lq(s)\big) \ge q(s),$$

because $q(s) \le \|u_s\| \le b_m$ and $Lb_m \le m \le s$.

If $k_m\|se + \xi\| > b_m$, set $r_s = \|se + \xi\|$ and $n_s = (se + \xi)/r_s$. Then $u_s = b_m n_s$ and

$$u_s' = \frac{b_m}{r_s}(I_d - n_s n_s^\top)e, \;\; q'(s) = \frac{b_m}{r_s}\big(1 - \langle e, n_s\rangle^2\big) \ge 0.$$

Since $\langle u_s, u_s'\rangle = 0$, it follows that $h_\xi'(s) = q(s) + sq'(s) \ge q(s)$.

The clipping boundary is determined by $k_m^2\|se + \xi\|^2 = b_m^2$. Its left-hand side is a quadratic polynomial in s with a positive leading coefficient, so the boundary contains at most two values of s. The derivative inequalities therefore hold almost everywhere. Absolute continuity implies that q is nondecreasing, and hence

$$h_\xi(\mu) - h_\xi(m) = \int_m^\mu h_\xi'(s)\,ds \ge \int_m^\mu q(s)\,ds \ge (\mu - m)q(m).$$

This is (C.1). □

C.2 Proof of Theorem 4

Fix $\mu \geq m$, a unit vector e, and any $P \in \mathcal{P}_e^{\mu,\sigma}$. Write $G = \mu e + \xi$, where $\mathbb{E}\xi = 0$ and $\mathbb{E}\|\xi\|^2 \leq \sigma^2$, and retain the preceding definitions for each realization of ξ. Set $F(s) = \mathbb{E}[h_\xi(s)]$. These expectations are finite because $|h_\xi(s)| \leq sb_m + Lb_m^2/2$.

Taking expectations in (C.1) yields

$$F(\mu) \geq F(m) + (\mu - m)\mathbb{E}[q(m)].$$

The shifted observation $me + \xi$ belongs to $\mathcal{P}_e^{m,\sigma}$. Theorem 1 therefore gives $F(m) \geq V_m$. Moreover,

$$F(m) = m\mathbb{E}[q(m)] - \frac{L}{2}\mathbb{E}\|u_m\|^2 \leq m\mathbb{E}[q(m)].$$

Since $\mu - m \geq 0$, combining these inequalities gives

$$F(\mu) \geq \frac{\mu}{m}F(m) \geq \frac{\mu}{m}V_m.$$

Both the direction and the admissible observation law were arbitrary. Thus $W_\mu(T_m) \geq (\mu/m)V_m$ for every $\mu \geq m$.

Finally, the range of norms in (10) contains m, so

$$V_{\geq m} = \sup_{T \in \mathcal{T}_R} \inf_{\mu \geq m} W_\mu(T) \leq \sup_{T \in \mathcal{T}_R} W_m(T) = V_m.$$

The preceding lower bound gives $\inf_{\mu \geq m} W_\mu(T_m) \geq V_m$, proving that $V_{\geq m} = V_m$ and that T_m attains this value. This establishes the theorem.

C.3 Why the calibration threshold must be positive

When the variance bound is positive, the following result excludes a nonzero fixed rule with a nonnegative guarantee at every positive gradient norm.

Proposition C.2 (Necessity of a positive threshold)**.** *Suppose $\sigma > 0$ and $T \in \mathcal{T}_R$ satisfies $W_\mu(T) \geq 0$ for every $\mu > 0$. Then $T \equiv 0$.*

Proof. Suppose instead that $T(z) \neq 0$ for some z. We first construct a zero-mean distribution P_0 such that $q_0 := \mathbb{E}_{P_0}\|G\|^2 < \sigma^2$ and $B_0 := \mathbb{E}_{P_0}\|T(G)\|^2 > 0$.

If $z = 0$, take $P_0 = \delta_0$. If $z \neq 0$, choose $0 < q < \min\{1/2, \sigma^2/(4\|z\|^2)\}$ and set

$$P_0 = q\delta_z + q\delta_{-z} + (1 - 2q)\delta_0.$$

This distribution has mean zero and second moment $q_0 = 2q\|z\|^2 < \sigma^2$. Its positive mass at z ensures that $B_0 > 0$.

Choose a nonzero vector v, taking $v = z$ when $z \neq 0$, and let $P_t = (1 - t)P_0 + t\delta_v$ for $t \in (0, 1)$. Its mean is tv, so define $\mu_t = t\|v\| > 0$ and $e = v/\|v\|$. Its centered second moment is

$$\mathbb{E}_{P_t}\|G - tv\|^2 = (1 - t)q_0 + t(1 - t)\|v\|^2.$$

This expression converges to $q_0 < \sigma^2$ as $t \downarrow 0$, so $P_t \in \mathcal{P}_e^{\mu_t,\sigma}$ for all sufficiently small $t > 0$.

Boundedness of T gives

$$\mathbb{E}_{P_t}\left[\langle tv, T(G)\rangle - \frac{L}{2}\|T(G)\|^2\right] \leq t\|v\|R - \frac{L}{2}(1 - t)B_0.$$

The right-hand side tends to $-LB_0/2 < 0$. For sufficiently small t, the admissible law P_t therefore gives $W_{\mu_t}(T) < 0$, a contradiction. This proof uses only values of T at fixed atoms and requires no continuity assumption. □

The zero rule has guarantee zero for every admissible direction and distribution, so the conclusion is exact. The proposition concerns one fixed rule across all positive gradient norms; choosing a positive threshold permits nonzero uniform progress above that threshold while leaving the behavior below it to be controlled separately.

D FINITE-TIME STATIONARITY OF THE FIXED RULE

This appendix proves Theorem 5. Fix $m > 0$ and retain $T_m(z) = \mathrm{clip}_{b_m}(k_m z)$, where $b_m = \min\{R, m/L\}$, $H_m = mb_m - Lb_m^2/2$, and $k_m = 2H_m/(m^2+\sigma^2)$. Write $p_m = m^2/(m^2+\sigma^2)$ and $\lambda_m = p_m k_m/2$. All results in this appendix hold for $\sigma \geq 0$.

D.1 CONTROL BELOW THE CALIBRATION THRESHOLD

The certificate proved in Appendix A.1, evaluated at gradient norm m and gain k_m, gives, for every unit vector e and every z,

$$m\langle e, T_m(z)\rangle - \frac{L}{2}\|T_m(z)\|^2 \geq mk_m\langle e, z\rangle - \lambda_m\|z\|^2.$$

Rearrange this inequality and maximize its left-hand side over e to obtain

$$\frac{L}{2}\|T_m(z)\|^2 + m\|T_m(z) - k_m z\| \leq \lambda_m\|z\|^2. \tag{D.1}$$

This property follows directly from the certificate for the constructed rule and yields the below-threshold bound (14), including when $\sigma = 0$.

Lemma D.1 (Descent bound below the threshold)**.** *Let $g \in \mathbb{R}^d$ satisfy $\|g\| \leq m$, and suppose that $\mathbb{E}G = g$ and $\mathbb{E}\|G - g\|^2 \leq \sigma^2$. Then*

$$\mathbb{E}\left[\langle g, T_m(G)\rangle - \frac{L}{2}\|T_m(G)\|^2\right] \geq \frac{k_m}{2}\left[(2-p_m)\|g\|^2 - (1-p_m)m^2\right].$$

Proof. Set $\mu = \|g\|$. Cauchy–Schwarz, $\mu \leq m$, and (D.1) imply, for every z,

$$\begin{aligned}\langle g, T_m(z)\rangle - \frac{L}{2}\|T_m(z)\|^2 &= k_m\langle g, z\rangle + \langle g, T_m(z) - k_m z\rangle - \frac{L}{2}\|T_m(z)\|^2 \\ &\geq k_m\langle g, z\rangle - \mu\|T_m(z) - k_m z\| - \frac{L}{2}\|T_m(z)\|^2 \\ &\geq k_m\langle g, z\rangle - \lambda_m\|z\|^2.\end{aligned}$$

Taking expectations and using $\mathbb{E}\|G\|^2 \leq \mu^2 + \sigma^2$ gives

$$\mathbb{E}\left[\langle g, T_m(G)\rangle - \frac{L}{2}\|T_m(G)\|^2\right] \geq k_m\mu^2 - \lambda_m(\mu^2+\sigma^2).$$

Substitute $\lambda_m = p_m k_m/2$ and $p_m\sigma^2 = (1-p_m)m^2$ to obtain the stated bound. The argument also covers $g = 0$. □

D.2 A DESCENT BOUND VALID AT EVERY GRADIENT NORM

Let $D_m(g, G) = \mathbb{E}[\langle g, T_m(G)\rangle - L\|T_m(G)\|^2/2]$. For $\mu := \|g\| \geq m$, equation (11) gives $D_m(g, G) \geq k_m m\mu/2$. Below the threshold, Lemma D.1 provides a quadratic bound. These bounds admit a common affine lower bound that can be summed along a trajectory.

First, put $r = \mu/m \in [0, 1]$ in the below-threshold bound. Since $p_m \in (0, 1]$,

$$(2-p_m)r^2 - (1-p_m) = 2r^2 - 1 + p_m(1-r^2) \geq 2r^2 - 1 \geq r - \frac{9}{8}.$$

The last inequality is equivalent to $2(r - 1/4)^2 \geq 0$. Combining this inequality with the bound above the threshold proves (15), with the constant $9/8$ used in the main text.

Keeping the dependence on p_m yields the following sharper version. Define $C(p) = 1 - p + 1/[4(2-p)]$ for $p \in [0, 1]$.

Lemma D.2 (Uniform affine descent bound)**.** *For any $g \in \mathbb{R}^d$ and any observation satisfying $\mathbb{E}G = g$ and $\mathbb{E}\|G - g\|^2 \leq \sigma^2$,*

$$D_m(g, G) \geq \frac{k_m m}{2}\big(\|g\| - C(p_m)m\big). \tag{D.2}$$

Moreover, $1/4 \leq C(p_m) \leq 9/8$.

Proof. If $\mu := \|g\| \geq m$, the stronger bound $D_m(g, G) \geq k_m m\mu/2$ proves the claim because $C(p_m) \geq 0$. For $0 \leq \mu \leq m$, set $r = \mu/m$. Completing the square gives

$$(2 - p_m)r^2 - (1 - p_m) - r + C(p_m) = (2 - p_m)\left(r - \frac{1}{2(2 - p_m)}\right)^2 \geq 0.$$

Multiplying by $k_m m^2/2$ and applying Lemma D.1 proves (D.2). Finally, $C'(p) = -1 + 1/[4(2 - p)^2] < 0$ on $[0, 1]$, while $C(0) = 9/8$ and $C(1) = 1/4$. □

D.3 Proof of Theorem 5

Consider the iteration (12), with a fixed initial point x_0, an L-smooth objective bounded below by $f_{\inf}$, and the conditional oracle assumptions (13). Let $\mathcal{F}_t$ contain the history before observing G_t, including x_t, and set $\Delta = f(x_0) - f_{\inf}$.

All expectations used below are finite over each fixed horizon N. Indeed, $\|T_m(G_t)\| \leq b_m \leq R$ implies $\|x_t - x_0\| \leq tR \leq NR$ for $t \leq N$. Both f and its continuous gradient are bounded on this deterministic compact ball. The conditional variance bound then also makes $\mathbb{E}\|G_t\|^2$ finite. These observations justify conditional expectations and finite telescoping sums without additional moment assumptions on the iterates.

Corollary D.3 (Stationarity with the noise-dependent constant)**.** *Under the assumptions of Theorem 5, let $N \geq 1$ and let J be independent of the trajectory and uniform on $\{0, \ldots, N - 1\}$. Then*

$$\mathbb{E}\|\nabla f(x_J)\| \leq C(p_m)m + \frac{2\Delta}{k_m m N}. \tag{D.3}$$

Proof. Condition on $\mathcal{F}_t$ and write $g_t = \nabla f(x_t)$. The oracle assumptions give $\mathbb{E}[G_t \mid \mathcal{F}_t] = g_t$ and $\mathbb{E}[\|G_t - g_t\|^2 \mid \mathcal{F}_t] \leq \sigma^2$, so Lemma D.2 applies to the conditional observation law. Smoothness and (D.2) yield

$$\begin{aligned}\mathbb{E}[f(x_t) - f(x_{t+1}) \mid \mathcal{F}_t] &\geq \mathbb{E}\left[\langle g_t, T_m(G_t)\rangle - \frac{L}{2}\|T_m(G_t)\|^2 \,\middle|\, \mathcal{F}_t\right] \\ &\geq \frac{k_m m}{2}\big(\|g_t\| - C(p_m)m\big).\end{aligned}$$

Take expectations, sum from $t = 0$ to $N - 1$, and use the objective lower bound to obtain

$$\frac{k_m m}{2}\sum_{t=0}^{N-1}\big(\mathbb{E}\|\nabla f(x_t)\| - C(p_m)m\big) \leq \mathbb{E}[f(x_0) - f(x_N)] \leq \Delta.$$

After rearrangement,

$$\frac{1}{N}\sum_{t=0}^{N-1}\mathbb{E}\|\nabla f(x_t)\| \leq C(p_m)m + \frac{2\Delta}{k_m m N}.$$

Independence and uniformity of J identify the average on the left with $\mathbb{E}\|\nabla f(x_J)\|$, proving (D.3). □

Since $C(p_m) \leq 9/8$, (D.3) implies the bound (16) in Theorem 5. For a prescribed accuracy $\varepsilon > 0$, take $m = 4\varepsilon/9$. Then $9m/8 = \varepsilon/2$. Choosing an integer

$$N \geq \max\left\{1, \left\lceil \frac{4\Delta}{k_m m \varepsilon} \right\rceil\right\}$$

makes the remaining term $2\Delta/(k_m m N)$ at most $\varepsilon/2$, and therefore $\mathbb{E}\|\nabla f(x_J)\| \leq \varepsilon$. This completes the proof of Theorem 5. The calibration threshold m is determined from the requested accuracy, and the same rule is used throughout the trajectory.

E Optimal Parameters within the Euclidean Clipping Family

This appendix specializes the general minimax problem to the rules $T(z) = \mathrm{clip}_c(\eta z)$, where $\eta > 0$ and $0 < c \leq R$. We first prove Corollary 3, including the zero-variance extension, and then describe the effect of the update bound on the optimal gain and clipping radius.

E.1 Complete characterization of the optimal parameters

Assume first that $\sigma > 0$. We prove that the clipping rule is minimax optimal if and only if $\eta = k_\mu$ and $b_\mu \le c \le R$.

For necessity, the derivative of $\mathrm{clip}_c(\eta z)$ at the origin is ηI_d, because $c > 0$. Equation (B.1) therefore forces $\eta = k_\mu$. At the informative observation, equation (B.2) requires

$$\min\{k_\mu a_\mu, c\} = b_\mu.$$

Since $k_\mu a_\mu = b_\mu(2 - Lb_\mu/\mu) \ge b_\mu$, this condition implies $c \ge b_\mu$. The upper bound $c \le R$ is part of feasibility.

For sufficiency, first suppose $R < \mu/L$. Then $b_\mu = R$, so the admissible interval for an optimal radius consists only of $c = b_\mu$. The rule is exactly T_μ, whose optimality follows from Theorem 1.

Now suppose $R \ge \mu/L$. In this case, $b_\mu = \mu/L$, $H_\mu = \mu^2/(2L)$, and $k_\mu = \mu^2/(LQ_\mu)$. Completing the square gives

$$\Phi_e(u) = H_\mu - \frac{L}{2}\left\|u - \frac{\mu}{L}e\right\|^2.$$

For every $c \ge b_\mu$, the vector $(\mu/L)e$ lies in the radius-c ball. Since clip_c is the Euclidean projection onto that ball, its nonexpansiveness gives

$$\left\|\mathrm{clip}_c(k_\mu z) - \frac{\mu}{L}e\right\| \le \left\|k_\mu z - \frac{\mu}{L}e\right\|.$$

Consequently, $\Phi_e(\mathrm{clip}_c(k_\mu z)) \ge \Phi_e(k_\mu z)$. Taking expectations under any admissible law yields

$$\begin{aligned}\mathbb{E}_P[\Phi_e(\mathrm{clip}_c(k_\mu G))] &\ge \mu^2 k_\mu - \frac{L}{2}k_\mu^2 \mathbb{E}_P\|G\|^2 \\ &\ge \mu^2 k_\mu - \frac{L}{2}k_\mu^2 Q_\mu = \frac{\mu^2 H_\mu}{Q_\mu} = V_\mu.\end{aligned}$$

Thus every $c \in [b_\mu, R]$ is optimal, completing the proof of Corollary 3.

When $\sigma = 0$, Appendix A.3 shows that optimality is equivalent to $T(\mu e) = b_\mu e$ for every direction. For ordinary clipping, this is precisely

$$\min\{\eta\mu, c\} = b_\mu, \;\; \eta > 0, \; 0 < c \le R.$$

This condition is both necessary and sufficient. If $R < \mu/L$, it reduces to $c = R$ and $\eta \ge R/\mu$. If $R \ge \mu/L$, it is equivalent to $\eta \ge 1/L$, $c \ge \mu/L$, and at least one of the equalities $\eta = 1/L$ or $c = \mu/L$. Thus zero variance does not force a unique local gain.

E.2 Joint calibration of the gain and update bound

Return to $\sigma > 0$. If $R < \mu/L$, the update constraint limits the deterministic optimal step, and the unique optimal parameter pair is

$$c = R, \;\; \eta = k_\mu = \frac{2\mu R - LR^2}{\mu^2 + \sigma^2}.$$

For fixed μ, σ, and L, its derivative with respect to R is $2(\mu - LR)/(\mu^2 + \sigma^2) > 0$. Reducing the permitted update length therefore requires reducing the gain applied before clipping. If $R \ge \mu/L$, the gain instead equals $\mu^2/[L(\mu^2 + \sigma^2)]$, and every radius in $[\mu/L, R]$ is optimal. The two expressions agree at $R = \mu/L$.

This distinction can also be seen by first optimizing an unrestricted linear rule. For $T(z) = \eta z$, the exact worst-case payoff under the moment assumptions is $\mu^2\eta - (L/2)\eta^2 Q_\mu$: the second-moment upper bound is attained by an admissible law. Its maximizing gain is $\eta_{\mathrm{lin}} = \mu^2/(LQ_\mu)$. In the active regime,

$$\eta_{\mathrm{lin}} - k_\mu = \frac{(\mu - LR)^2}{L(\mu^2 + \sigma^2)} > 0.$$

Thus selecting the unrestricted linear gain and subsequently clipping at radius R does not produce a minimax-optimal bounded rule when $R < \mu/L$. This is an exact consequence of the necessary local gain condition, rather than only a comparison of sufficient guarantees. For example, at $R = \mu/(2L)$, the optimal bounded-rule gain is $k_\mu = 3\eta_{\mathrm{lin}}/4$.

F Optimal Bounded Updates under Isotropic Gaussian Noise

This appendix specializes the decision problem (4) to a known noise model. Fix $d \geq 1$ and $\mu, L, R, \sigma > 0$. The gradient norm and the noise law are known, while the gradient direction $e \in \mathbb{S}^{d-1}$ is unknown. Specifically, $G = \mu e + \tau Z$, where $Z \sim \mathcal{N}(0, I_d)$ and $\tau^2 = \sigma^2/d$, so that the total noise variance is σ^2. For the same class $\mathcal{T}_R$ of measurable bounded rules, define

$$V_\mu^{\mathrm{G}} = \sup_{T \in \mathcal{T}_R} \inf_{\|e\|=1} \mathbb{E}_e \left[\mu \langle e, T(G) \rangle - \frac{L}{2} \|T(G)\|^2 \right],$$

where $\mathbb{E}_e$ denotes expectation under this Gaussian observation model.

F.1 The optimal rule and its value

Let ω_d be the uniform probability measure on $\mathbb{S}^{d-1}$; for $d = 1$, it assigns probability $1/2$ to each of -1 and 1. For any unit vector v, introduce the normalizing function

$$Z_d(\kappa) = \int_{\mathbb{S}^{d-1}} \exp(\kappa \langle u, v \rangle)\, \omega_d(du), \quad A_d(\kappa) = \frac{d}{d\kappa} \log Z_d(\kappa).$$

Rotational symmetry makes Z_d independent of v. Using the power series and differentiation identities for the modified Bessel function of the first kind (National Institute of Standards and Technology, Sections 10.25 and 10.29), we obtain

$$Z_d(\kappa) = \Gamma(d/2) \left(\frac{2}{\kappa} \right)^{d/2-1} I_{d/2-1}(\kappa), \quad A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} \ (\kappa > 0),$$

with $Z_d(0) = 1$ and $A_d(0) = 0$ by continuity. These identities also hold for $d = 1$, where $Z_1(\kappa) = \cosh \kappa$ and $A_1(\kappa) = \tanh \kappa$. For example, the expression for Z_d follows by expanding the exponential and using

$$\int_{\mathbb{S}^{d-1}} \langle u, v \rangle^{2j}\, \omega_d(du) = \frac{\Gamma(j + 1/2)\Gamma(d/2)}{\sqrt{\pi}\, \Gamma(j + d/2)};$$

odd moments vanish. The resulting series is the defining Bessel series, and differentiation gives the expression for A_d.

For $a \geq 0$, write

$$\psi_R(a) = \begin{cases} a^2/(2L), & 0 \leq a \leq LR, \\ aR - LR^2/2, & a > LR. \end{cases}$$

This is the optimal quadratic descent value for a known gradient of norm a, and in particular $\psi_R(\mu) = H_\mu$.

Theorem F.1 (Optimal update under Gaussian noise)**.** *A continuous minimax-optimal rule is*

$$T_{\mathrm{G}}(z) = \begin{cases} \min\left\{ R, \dfrac{\mu}{L} A_d\left(\dfrac{\mu \|z\|}{\tau^2} \right) \right\} \dfrac{z}{\|z\|}, & z \neq 0, \\ 0, & z = 0. \end{cases} \tag{F.1}$$

Let ρ have the distribution of $\|\mu e + \tau Z\|$, which is independent of the choice of unit direction e. The exact optimal value is

$$V_\mu^{\mathrm{G}} = \mathbb{E}\left[\psi_R\left(\mu A_d\left(\frac{\mu \rho}{\tau^2} \right) \right) \right]. \tag{F.2}$$

A rule in $\mathcal{T}_R$ is minimax-optimal if and only if it agrees with T_{G} Lebesgue-almost everywhere.

Proof. We use the standard decision-theoretic argument that a Bayes rule with constant risk is minimax (Berger, 1985), applied here to the negative descent payoff.

Place the uniform prior ω_d on a random gradient direction U. For a nonzero observation z, write $v = z/\|z\|$ and $\kappa = \mu \|z\|/\tau^2$. The Gaussian likelihood, as a function of u, is proportional to $\exp(\kappa \langle u, v \rangle)$, because $\|z - \mu u\|^2 = \|z\|^2 + \mu^2 - 2\mu \langle u, z \rangle$. The posterior distribution is therefore

$$\mathbb{P}(U \in du \mid G = z) = \frac{\exp(\kappa \langle u, v \rangle)}{Z_d(\kappa)}\, \omega_d(du).$$

For $d \geq 2$, this is the von Mises–Fisher distribution with mean direction v and concentration κ (Banerjee et al., 2005). For $d = 1$, it is an exponential tilt of the two-point uniform law. Symmetry eliminates the posterior mean perpendicular to v, while differentiating Z_d gives its component along v. Thus the posterior mean gradient is

$$M(z) := \mathbb{E}[\mu U \mid G = z] = \mu A_d(\kappa) v, \;\; M(0) = 0.$$

Conditional on $G = z$, the expected descent payoff of an update w is $M(z)^\top w - L\|w\|^2/2$. Completing the square shows that its unique maximizer over $\|w\| \leq R$ is the Euclidean projection of $M(z)/L$ onto the radius-R ball. This is exactly (F.1), and the conditional optimal value is $\psi_R(\|M(z)\|)$.

For every rule, its worst-direction payoff is at most its payoff averaged under ω_d. Optimizing the latter conditionally on each observation therefore bounds V_μ^{G} from above by $\mathbb{E}[\psi_R(\|M(G)\|)]$. On the other hand, $T_{\mathrm{G}}(Qz) = QT_{\mathrm{G}}(z)$ for every orthogonal matrix Q. The isotropy of the noise implies that the expected payoff of this rule is the same for every e. Consequently, its worst-direction payoff equals its prior-average payoff and attains the upper bound. The distribution of $\|G\|$ is the same for every direction and for the prior mixture, giving (F.2).

It remains to characterize equality. By the variational characterization of Euclidean projection (Bauschke & Combettes, 2017), for every z and every feasible w,

$$M(z)^\top T_{\mathrm{G}}(z) - \frac{L}{2}\|T_{\mathrm{G}}(z)\|^2 - \left(M(z)^\top w - \frac{L}{2}\|w\|^2\right) \geq \frac{L}{2}\|w - T_{\mathrm{G}}(z)\|^2.$$

If T is minimax-optimal, its prior-average payoff is at least its worst-direction payoff V_μ^{G} and at most the optimal prior-average payoff, which is the same value. Integrating the displayed inequality with $w = T(G)$ therefore gives $\mathbb{E}\|T(G) - T_{\mathrm{G}}(G)\|^2 = 0$ under the prior mixture. That mixture has a strictly positive density everywhere on $\mathbb{R}^d$, so equality holds Lebesgue-almost everywhere. Conversely, a modification on a Lebesgue-null set changes no expected payoff under any of the Gaussian laws, provided the rule remains in $\mathcal{T}_R$. □

Formula (F.2) is a one-dimensional radial expectation: ρ/τ has the noncentral chi distribution with d degrees of freedom and noncentrality μ/τ. It thus determines the exact value without optimization over rules or directions. In dimension one, the rule has the particularly simple form

$$T_{\mathrm{G}}(z) = \mathrm{clip}_R\Big(\frac{\mu}{L}\tanh\Big(\frac{\mu z}{\sigma^2}\Big)\Big).$$

F.2 Local scaling and saturation

The first two nonzero terms of the normalizing-function expansion are $Z_d(\kappa) = 1 + \kappa^2/(2d) + \kappa^4/[8d(d+2)] + O(\kappa^6)$. Taking its logarithmic derivative gives

$$A_d(\kappa) = \frac{\kappa}{d} - \frac{\kappa^3}{d^2(d+2)} + O(\kappa^5).$$

Because $R > 0$, the norm constraint is inactive near the origin. Substitution in (F.1), together with $d\tau^2 = \sigma^2$, yields

$$T_{\mathrm{G}}(z) = \frac{\mu^2}{L\sigma^2} z + O(\|z\|^3) \text{ as } z \to 0.$$

Hence the local gain of this continuous optimal representative is $\mu^2/(L\sigma^2)$, independently of R. Since the Gaussian model determines optimal rules only almost everywhere, this derivative statement concerns the representative in (F.1).

The function A_d is strictly increasing on $[0, \infty)$, with values in $[0, 1)$ and limit 1. Indeed, its derivative is the variance of $\langle U, v\rangle$ under the tilted sphere distribution, which is positive for every finite κ. The strict upper bound follows from $\langle U, v\rangle \leq 1$ and nondegeneracy of that distribution. To verify the limit, write $S = \langle U, v\rangle$. For any $\delta \in (0, 2)$, the sphere cap has probability $c_\delta := \omega_d(S \geq 1 - \delta/2) > 0$. Under the tilted distribution,

$$\mathbb{P}_\kappa(S \leq 1 - \delta) \leq c_\delta^{-1} \exp(-\kappa\delta/2).$$

Consequently, $0 \leq 1 - A_d(\kappa) \leq \delta + 2\mathbb{P}_\kappa(S \leq 1 - \delta)$, proving $A_d(\kappa) \to 1$. Thus, if $R \geq \mu/L$, the rule never reaches the update cap at a finite observation. If $R < \mu/L$, there is a unique $\kappa_R > 0$ satisfying $A_d(\kappa_R) = LR/\mu$, and saturation occurs exactly when $\|z\| \geq \tau^2\kappa_R/\mu$.

F.3 THE VALUE OF ADDITIONAL DISTRIBUTIONAL INFORMATION

The Gaussian family is a subset of the moment class considered in Theorem 1. Consequently, $V_\mu \le V_\mu^{\mathrm{G}} \le H_\mu$. Both inequalities are strict when the parameters are positive.

Proposition F.2 (Strict comparison with the moment model)**.** *For every finite $d \ge 1$ and $\mu, L, R, \sigma > 0$,*

$$V_\mu < V_\mu^{\mathrm{G}} < H_\mu.$$

Proof. For the upper inequality, $A_d(\kappa) < 1$ at every finite κ, so the posterior mean gradient norm is strictly smaller than μ almost surely. The function ψ_R is strictly increasing on $[0, \infty)$. Equation (F.2) therefore implies $V_\mu^{\mathrm{G}} < \psi_R(\mu) = H_\mu$.

For the lower inequality, consider the moment-optimal rule $T_\mu(z) = \mathrm{clip}_{b_\mu}(k_\mu z)$ from Theorem 1. Every Gaussian law in this appendix satisfies the required moment conditions, so its expected payoff under T_μ is at least V_μ. The same is true of its prior-average payoff. Near the origin, $T_\mu(z) = k_\mu z$, whereas the continuous Gaussian optimizer has the strictly larger local gain

$$k_\mu = \frac{2H_\mu}{\mu^2 + \sigma^2} \le \frac{\mu^2}{L(\mu^2 + \sigma^2)} < \frac{\mu^2}{L\sigma^2}.$$

The two rules consequently differ on a set of positive Lebesgue measure. The strict conditional improvement established in the proof of Theorem F.1 implies that their prior-average payoffs differ by at least $(L/2)\mathbb{E}\|T_\mu(G) - T_{\mathrm{G}}(G)\|^2 > 0$. The Gaussian rule therefore achieves a value strictly above V_μ. □

Additional knowledge of the noise law changes both the attainable descent and the optimal response to an observation. In particular, the coupling between the update cap and the local gain in the moment model need not persist under a specified Gaussian noise law.

G EXPERIMENTAL PROTOCOLS AND COMPLETE COMPARISONS

The supplied package contains all scripts, frozen protocol files, numerical results, LP witnesses, plotting code, and figures in vector PDF and PNG formats. The command `python reproduce.py all` reruns the suite on CPU. Figure regeneration does not rerun optimization. No running-average smoothing is applied. For stochastic runs, curves show means across independent seeds with pointwise 95% Student-t confidence intervals.

G.1 MOMENT PROGRAMS

For each finite support $\{z_i\}$ and scalar response T, the program is

$$\min_{\pi \ge 0} \sum_i \pi_i \left(\mu T(z_i) - \frac{L}{2} T(z_i)^2 \right), \ \sum_i \pi_i = 1, \ \sum_i \pi_i z_i = \mu, \ \sum_i \pi_i z_i^2 \le \mu^2 + \sigma^2.$$

Here $L = 1$. With $s = \max\{1, \mu + \sigma\}$ and $B = \max\{12s, 8, 160(1 + \sigma^2)\}$, the standard support combines 1201 equally spaced points on $[-4s, 4s]$, 401 geometric points in each tail, zero, μ, and specified informative atoms. The independent grid uses 4001 central points and 1201 points in each tail, explicitly removing any point coinciding numerically with $a_\mu = (\mu^2 + \sigma^2)/\mu$. HiGHS primal and dual feasibility tolerances are 10^{-9}; moment constraints are rescaled for conditioning. The suite comprises 744 programs: 480 fixed-norm checks, 62 fixed-threshold checks, and 202 joint-calibration comparisons. The first study covers $\mu = 0.1, 0.2, \ldots, 4$, $\sigma \in \{0.5, 2, 5\}$, and $R \in \{0.5, 2\}$. Fixed-threshold checks use $m = 1$, $\sigma = 2$, both caps, and $\mu = 1, 1.1, \ldots, 4$. Joint-calibration checks use $\mu = m = 1$, $\sigma = 2$, and 101 caps from 0.1 to 2. LP extrema are upper bounds on the unrestricted worst-case payoffs. Feasible negative values establish attainable objective increases independently of grid completeness.

G.2 Controlled noise-shape experiment

The objective is $f(x) = x^2/2$, with $L = m = 1$, $\sigma = 2$, and $R = 0.2$. One-step expectations are evaluated by exact summation over the three-point noise law. The trajectory experiment starts at $x_0 = 5$ and runs 400 updates. The noise parameter changes from $q = 0.01$ to $q = 1$ after 20 updates. We compare five rules: SNR-then-clip throughout, minimax throughout, and switching from SNR-then-clip to minimax after 10, 20, or 40 updates. All 1024 seeds share the same noise across rules. Figure 4 displays the entire horizon.

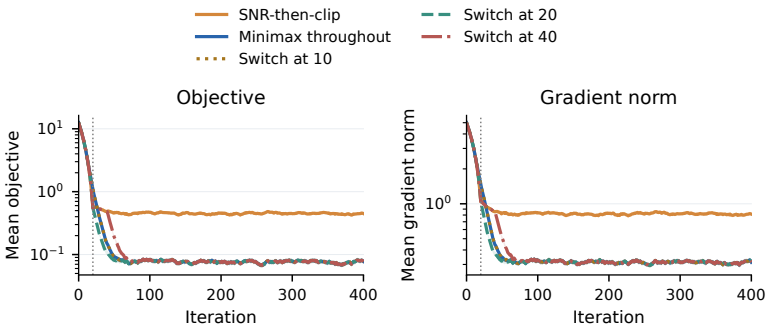

Figure 4: Controlled trajectories over all 400 updates. The vertical dotted line marks the noise change at step 20. Curves and pointwise intervals use 1024 independent paired seeds.

G.3 Logistic objective and valid global constants

The dataset is the Wisconsin Diagnostic Breast Cancer data distributed with scikit-learn (UCI DOI: `https://doi.org/10.24432/C5DW2B`). A stratified 80/20 split with seed 1729 gives 455 training examples and 114 held-out examples. All optimization metrics are evaluated on the training objective. Features are standardized using training statistics. A standardized row z_i is divided by $\max\{1, \|z_i\|\}$, augmented with a constant coordinate, and divided by $\sqrt{2}$, giving $a_i \in \mathbb{R}^{31}$ with $\|a_i\| \le 1$. The objective is

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}\left[\log(1 + e^{a_i^\top w}) - y_i a_i^\top w\right] + \frac{\lambda}{2}\|w\|^2, \quad \lambda = 0.01.$$

Let $h_i(w) = (s(a_i^\top w) - y_i)a_i$, with logistic sigmoid s. A minibatch of $B = 16$ independent indices drawn with probabilities q returns

$$G(w) = \lambda w + \frac{1}{B}\sum_{j=1}^{B}\frac{h_{I_j}(w)}{n q_{I_j}}.$$

Thus $\mathbb{E}[G(w) \mid w] = \nabla f(w)$. Global valid bounds are

$$L = \lambda + \tfrac{1}{4}\lambda_{\max}(A^\top A/n) = 0.138697357, \quad \sigma^2 = \max_q \frac{1}{Bn^2}\sum_i \frac{\|a_i\|^2}{q_i} = 0.562515093.$$

Uniform sampling and two skewed samplers are used. Each skewed sampler puts 90% of its mass uniformly on a subset and 10% uniformly on all training examples. The class subset has label zero; the feature subset consists of the top 10% of standardized feature norms before row normalization. All scenarios preserve the same objective and obey the same variance bound; their actual variances need not be equal. In the changing-sampler scenario, sampling is uniform for $t < 1000$, class-skewed for $1000 \le t < 3000$, and feature-skewed thereafter. Oracle changes are separated from the algorithm switch at step 2000.

G.4 Controlled fixed-variance oracle on the logistic objective

This experiment holds the objective and conditional variance fixed while changing the shape of the stochastic observation. It replaces minibatch sampling with an artificial oracle. With $v =$

$\nabla f(0)/\|\nabla f(0)\|$ fixed throughout a trajectory, define

$$G_t = \nabla f(w_t) + \xi_{q,t}, \;\; \mathbb{P}(\xi_{q,t} = 0) = 1 - q, \; \mathbb{P}\left(\xi_{q,t} = \pm\frac{\sigma}{\sqrt{q}}v\right) = \frac{q}{2}.$$

The draws are independent over time. Consequently,

$$\mathbb{E}[G_t \mid \mathcal{F}_t] = \nabla f(w_t), \;\; \mathbb{E}[\|G_t - \nabla f(w_t)\|^2 \mid \mathcal{F}_t] = \sigma^2.$$

The full gradient defines this controlled oracle; its computational cost is not equated with that of a minibatch oracle.

Fixed protocol. All runs start at zero and use $N = 4000$, $m = 0.05$, $LR/m \in \{0.2, 2\}$, and $q \in \{0.001, 0.003, 0.01, 0.03, 0.1, 1\}$. The rules T_m and SNR-then-clip retain the same parameters for every q. For supplementary trajectory comparisons, a constant gain is selected from $k_m\{1/4, 1/2, 1, 2, 4, 8, 16, 32, 64\}$ using seeds 2000–2005 at the nominal probability $q = 0.01$. Independent seeds 2100–2105 select a post-warmup decay factor from $\{1/2, 1/4, 1/8, 1/16\}$, again only at $q = 0.01$. The hybrid and decay methods share the same warmup for 2000 updates and the same second-stage radius b_m. The selected constant gains are $64k_m$ for both caps, and the selected decay factors are both $1/2$. No parameter is reselected at the other probabilities.

Evaluation uses 64 independent seeds 3000–3063. The same uniforms couple noise across methods and probabilities, preserving the correct marginal distribution in every setting. There are $2 \times 6 \times 5 \times 64 = 3840$ evaluation trajectories. Each method uses $64 \times 4000 = 256{,}000$ artificial-oracle calls per cap and probability. Per cap, constant-gain selection uses 216,000 additional calls and decay selection uses 96,000. The primary metric is the full-horizon mean gradient norm; clipping frequency describes the rule's operating regime. Tables 1–2 report the trajectory metrics and clipping frequencies.

Relation to the calibration. For the joint rule, the observation-space clipping threshold is

$$C_m = \frac{b_m}{k_m} = \frac{m^2 + \sigma^2}{2m - Lb_m},$$

which equals approximately 6.28 and 11.30 for the two caps. The nonzero perturbation magnitude $\sigma/\sqrt{q}$ increases as q decreases, while its occurrence probability decreases. This explains the transition between clipped rare observations and linear responses to more frequent smaller observations. Across the evaluated trajectories, the joint rule clips at the three smallest q values under the active cap and at the two smallest under the inactive cap. At $q = 0.001$, its observed clipping frequency is about 0.110% for both caps, close to the generating probability of 0.1%. At $q = 0.01$, the active-cap frequency is about 0.976%, whereas the inactive-cap rule remains linear.

Exact single-step evaluation. For the logistic panel in Figure 2, a point with $\|\nabla f(w)\| = m$ is located along the line from the deterministic minimizer to zero. The perturbation direction remains the same fixed v as in the trajectory experiment. At each q, both the actual objective decrease and the smoothness payoff are computed by summing over the three observation atoms. A grid of 241 logarithmically spaced probabilities, augmented with the trajectory probabilities, gives 972 rule–cap–probability evaluations. The saved moments are verified independently. Conditional smoothness payoffs are additionally evaluated at trajectory steps 0, 2000, and 3999. Whenever the current norm is at least m and the joint rule is active, the checked payoff satisfies $(\|\nabla f(w_t)\|/m)V_m$. All 1330 applicable checks satisfy this bound.

Trajectory interpretation. The fixed joint rule maintains similar mean gradient norms across the tested probabilities, while its clipping frequency changes. On many specified laws the tuned rules progress faster; at $q = 1$, their larger gains and frequent saturation give larger mean gradient norms. These observations describe how a fixed moment calibration and a nominally selected gain respond to changes in noise shape.

G.5 Natural-minibatch selection, pairing, costs, and diagnostics

All trajectories start at zero. We fix $m = 0.05$, $N = 4000$, and $LR/m \in \{0.2, 2\}$. Constant clipping selects a gain from $k_m\{1/4, 1/2, 1, 2, 4, 8, 16, 32, 64\}$ by minimizing the full-horizon mean gradient

Table 1: Controlled logistic oracle: full-horizon mean gradient norm, mean ± one standard error over 64 paired seeds.

LR/m	q	Joint	SNR	Tuned constant	Hybrid	Tuned decay
0.2	0.001	0.07723 ± 0.00022	0.03469 ± 0.00009	0.00201 ± 0.00001	0.00268 ± 0.00006	0.00206 ± 0.00001
0.2	0.003	0.07748 ± 0.00041	0.03494 ± 0.00017	0.00217 ± 0.00001	0.00334 ± 0.00009	0.00227 ± 0.00001
0.2	0.01	0.07788 ± 0.00064	0.03556 ± 0.00026	0.00260 ± 0.00001	0.00476 ± 0.00014	0.00281 ± 0.00002
0.2	0.03	0.07705 ± 0.00089	0.03742 ± 0.00050	0.00343 ± 0.00002	0.00600 ± 0.00022	0.00379 ± 0.00002
0.2	0.1	0.07863 ± 0.00079	0.04350 ± 0.00083	0.00500 ± 0.00003	0.00681 ± 0.00018	0.00565 ± 0.00005
0.2	1	0.07835 ± 0.00096	0.04191 ± 0.00086	0.24073 ± 0.01541	0.18466 ± 0.01176	0.24068 ± 0.01540
2	0.001	0.03660 ± 0.00042	0.04050 ± 0.00083	0.00096 ± 0.00002	0.00330 ± 0.00021	0.00095 ± 0.00002
2	0.003	0.03935 ± 0.00070	0.04086 ± 0.00082	0.00160 ± 0.00003	0.00590 ± 0.00028	0.00158 ± 0.00003
2	0.01	0.04039 ± 0.00075	0.04039 ± 0.00075	0.00368 ± 0.00005	0.00834 ± 0.00029	0.00349 ± 0.00005
2	0.03	0.04085 ± 0.00085	0.04085 ± 0.00085	0.00849 ± 0.00008	0.01155 ± 0.00038	0.00767 ± 0.00007
2	0.1	0.04134 ± 0.00076	0.04134 ± 0.00076	0.01819 ± 0.00011	0.01705 ± 0.00043	0.01574 ± 0.00011
2	1	0.04191 ± 0.00086	0.04191 ± 0.00086	0.24800 ± 0.00613	0.19147 ± 0.00807	0.23940 ± 0.00777

Table 2: Percentage of clipped updates over the full horizon under the controlled logistic oracle.

LR/m	q	Joint	SNR	Tuned constant	Hybrid	Tuned decay
0.2	0.001	0.110	0.110	0.735	0.735	0.735
0.2	0.003	0.294	0.294	0.920	0.920	0.920
0.2	0.01	0.976	0.976	1.599	1.599	1.599
0.2	0.03	0.000	2.987	3.615	2.110	3.615
0.2	0.1	0.000	9.419	10.499	5.547	10.499
0.2	1	0.000	0.000	100.000	50.000	99.795
2	0.001	0.110	0.110	0.110	0.110	0.110
2	0.003	0.294	0.000	0.294	0.294	0.294
2	0.01	0.000	0.000	0.976	0.493	0.976
2	0.03	0.000	0.000	2.987	1.482	2.987
2	0.1	0.000	0.000	9.861	4.909	9.861
2	1	0.000	0.000	89.096	44.650	90.127

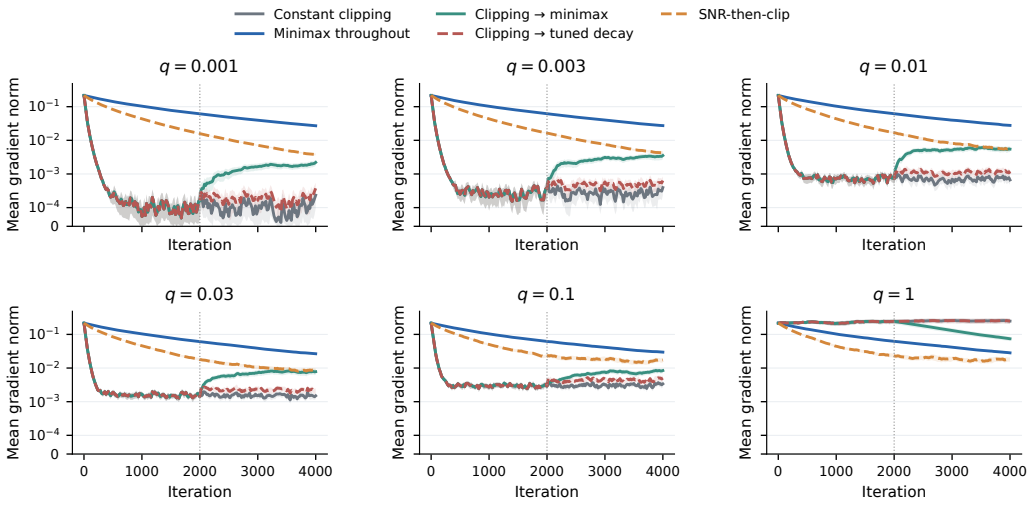

Figure 5: Controlled logistic trajectories with $LR/m = 0.2$. Shading shows pointwise 95% intervals across 64 paired seeds. The vertical line marks the hybrid and decay switch at 2000. The vertical scale is linear below 10^{-4} and logarithmic above; interval endpoints are restricted to the nonnegative domain.

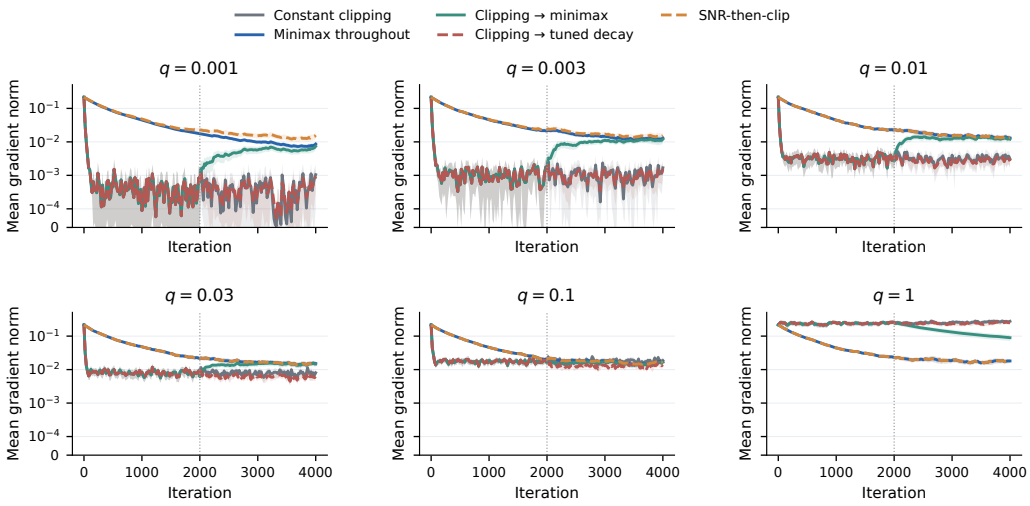

Figure 6: The same controlled-oracle experiment with $LR/m = 2$, using the same vertical scale and interval convention. The joint and SNR rules have the same gain and differ only in their allowed update radius; their trajectories coincide in the tested settings where neither rule clips.

norm under uniform sampling on seeds 0–5. A second search on separate seeds 100–105 selects a decay factor from $\{1/2, 1/4, 1/8, 1/16\}$. Each candidate shares the chosen warmup and switches at $K = 2000$ to radius b_m. The decay search selects a factor $1/8$ under the active cap and $1/16$ under the inactive cap. Selected parameters appear in Table 3.

Evaluation uses seeds 1000–1031, disjoint from both pilot sets. Common sampled indices pair all methods within each scenario and seed. Full gradients are evaluated before every update for scoring; they are never used by update rules or switch decisions. The same full-horizon metric selects pilots and evaluates held-out seeds. Secondary outputs include final gradient norm, final training objective gap, second-half average gradient norm, clipping frequencies, and the maximum step length. A radius-only control switches from R to b_m while retaining the initial gain; an SNR-then-clip fixed rule is also retained.

Per cap, constant-gain selection costs $9 \times 6 \times 4000 = 216{,}000$ minibatch calls. Decay selection adds $4 \times 6 \times 4000 = 96{,}000$ calls. Each evaluated rule costs $32 \times 4000 = 128{,}000$ calls per cap and scenario. The hybrid reuses the warmup selection and needs no second-stage search. Every minibatch call uses 16 example gradients. Oracle-call counts exclude diagnostic full-gradient evaluations.

Table 3: Selected logistic parameters. The same warmup and post-switch radius are shared by both hybrid methods.

LR/m	R	b_m	k_m	η_{warm}	Decay factor
0.2	0.07209943	0.07209943	0.01148456	0.36750585	0.125
2.0	0.72099427	0.36049714	0.03190155	0.51042479	0.0625

The paired hybrid-minus-decay differences are positive in each active-cap scenario, with all four pointwise 95% intervals excluding zero. Under the inactive cap the selected post-decay gain equals k_m exactly, so the two rules and trajectories coincide. The minimax rule, including the hybrid's second phase, never triggers clipping in the evaluated minibatch runs. The calibrated rule therefore operates in its linear regime. Table 6 records the clipping rates and the radius-only control. All matched warmup prefixes agree exactly in the rerun; update-cap violations are below floating-point roundoff.

Table 4: Full-horizon average gradient norms, in units of 10^{-3}. Entries are means $\pm$ one standard error over 32 seeds; lower is better. Radius-only changes only the post-switch radius.

LR/m	Sampler	Constant	T_m	Hybrid	Decay	Radius-only	SNR
0.2	Uniform	7.25 ± 0.04	77.01 ± 0.03	6.35 ± 0.09	6.04 ± 0.04	7.25 ± 0.04	34.63 ± 0.02
0.2	Class-skewed	14.00 ± 0.15	77.02 ± 0.17	11.53 ± 0.28	10.81 ± 0.16	14.00 ± 0.15	34.86 ± 0.12
0.2	Feature-skewed	17.16 ± 0.16	77.20 ± 0.19	15.00 ± 0.38	13.87 ± 0.19	17.16 ± 0.16	35.51 ± 0.17
0.2	Switching	12.20 ± 0.11	77.11 ± 0.07	9.38 ± 0.27	8.88 ± 0.12	12.20 ± 0.11	35.05 ± 0.06
2	Uniform	7.18 ± 0.04	34.63 ± 0.02	5.69 ± 0.05	5.69 ± 0.05	7.18 ± 0.04	34.63 ± 0.02
2	Class-skewed	12.82 ± 0.10	34.86 ± 0.12	9.58 ± 0.15	9.58 ± 0.15	12.82 ± 0.09	34.86 ± 0.12
2	Feature-skewed	17.58 ± 0.13	35.51 ± 0.17	13.10 ± 0.21	13.10 ± 0.21	17.55 ± 0.13	35.51 ± 0.17
2	Switching	12.60 ± 0.08	35.05 ± 0.06	8.51 ± 0.13	8.51 ± 0.13	12.59 ± 0.08	35.05 ± 0.06

Table 5: Paired primary-metric differences in units of 10^{-3}, with pointwise 95% Student-t intervals. Negative values favor the minimax hybrid.

LR/m	Sampler	Hybrid $-$ constant	95% interval	Hybrid $-$ decay	95% interval
0.2	Uniform	-0.906	$[-1.091, -0.722]$	+0.308	$[+0.147, +0.470]$
0.2	Class-skewed	-2.461	$[-2.943, -1.979]$	+0.722	$[+0.367, +1.076]$
0.2	Feature-skewed	-2.151	$[-2.824, -1.478]$	+1.139	$[+0.626, +1.652]$
0.2	Switching	-2.810	$[-3.353, -2.267]$	+0.508	$[+0.130, +0.886]$
2	Uniform	-1.495	$[-1.582, -1.407]$	+0.000	$[+0.000, +0.000]$
2	Class-skewed	-3.237	$[-3.461, -3.013]$	+0.000	$[+0.000, +0.000]$
2	Feature-skewed	-4.475	$[-4.854, -4.095]$	+0.000	$[+0.000, +0.000]$
2	Switching	-4.094	$[-4.332, -3.856]$	+0.000	$[+0.000, +0.000]$

Table 6: Percentage of updates that activate clipping. The minimax-throughout and SNR rules have zero clipping in both halves of all configurations.

LR/m	Sampler	Warmup	Constant post	Radius-only post	Hybrid post	Decay post
0.2	Uniform	0.472	0.003	0.003	0	0
0.2	Class-skewed	11.566	9.786	9.786	0	0
0.2	Feature-skewed	21.150	18.722	18.722	0	0
0.2	Switching	5.273	14.267	14.267	0	0
2	Uniform	0.000	0.000	0.000	0	0
2	Class-skewed	0.000	0.000	0.005	0	0
2	Feature-skewed	0.000	0.000	0.106	0	0
2	Switching	0.000	0.000	0.056	0	0

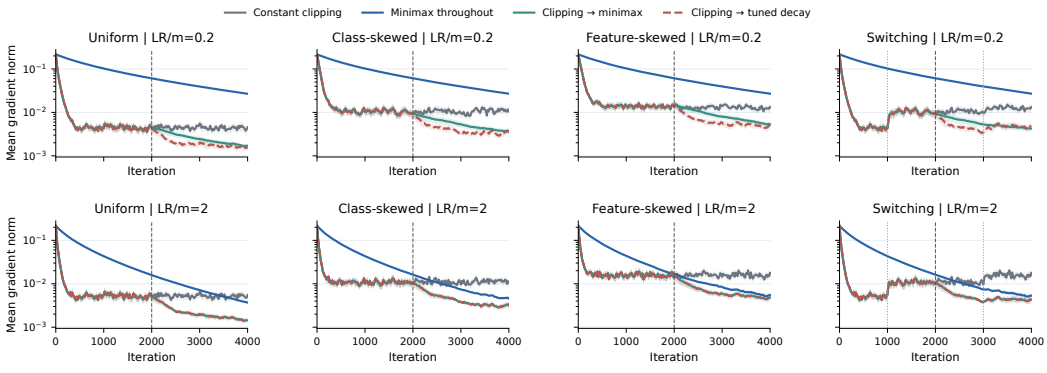

Figure 7: All eight logistic configurations, over the complete 4000-step horizon. The dashed vertical line marks the algorithm switch at 2000; dotted lines in the changing-sampler panels mark oracle changes at 1000 and 3000. Shading gives pointwise 95% intervals over 32 paired seeds. The inactive-cap hybrid and decay curves coincide.

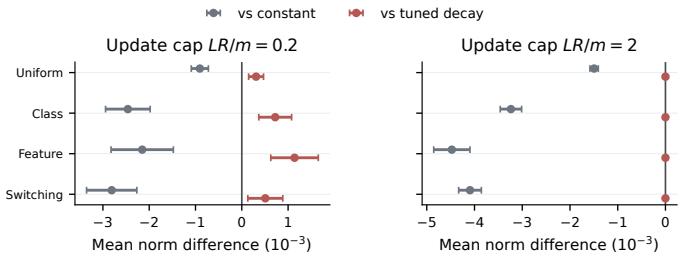

Figure 8: Natural-minibatch trajectory comparisons. Differences are the mean gradient norm of the minimax hybrid minus that of constant clipping or tuned decay. Negative values favor the hybrid. Error bars are pointwise paired 95% intervals.

G.6 Common-pool moment stress tests

A deterministic minimizer is computed to gradient norm below 10^{-8}. On the line joining it to the zero initialization, roots of the gradient norm determine 31 points with $\mu/m = 1, 1.1, \ldots, 4$. At each point, set $e = \nabla f(w)/\mu$. The artificial observation laws have the form $G = ze$, with $\mathbb{E}z = \mu$ and $\mathbb{E}z^2 \le \mu^2 + \sigma^2$. Thus they are legitimate full-dimensional oracles at that point, but need not arise from the minibatch sampler.

For each cap and norm, LPs generate witnesses for the target rule, SNR-then-clip, tuned constant clipping, the tuned post-decay rule, and a known-norm minimax reference. Supports use $z = a_\mu s$, with symmetric logarithmic s grids of 1400 positive points from 10^{-5} to 32, zero, the equality atom, and extra points at half, one, and twice each clipping threshold. Objectives are normalized by V_μ and moment constraints rescaled; HiGHS tolerances are 10^{-9}. The union of all witnesses forms one common pool. Every compared rule is evaluated on every pooled law by exact finite sums of both its smoothness payoff and actual logistic decrease. The underlying atoms, masses, points, and directions are saved for independent checking. At each point and cap, all four rules are cross-evaluated on the same five-law pool: 310 solved moment programs and 1240 cross-evaluations in total. Figure 9 shows all 31 computed points per curve; markers identify every fifth point and straight segments connect neighboring evaluations, without smoothing. Table 7 gives the representative ratios 1, 2, and 4; complete values and witness distributions are supplied in the code package. Each reported minimum is taken over the common five-law pool and is therefore an upper bound on the worst-case actual decrease over the full admissible distribution class.

Table 7: Minimum actual logistic decrease on the common witness pool, divided by V_m. Representative values from the 31-point sweep.

LR/m	μ/m	Joint	SNR	Tuned constant	Tuned decay
0.2	1	+1.0717	-1.5366	-211.2969	-6.4574
0.2	2	+6.1419	+7.8440	-378.5737	+2.4777
0.2	4	+28.2464	+59.0230	-581.1809	+65.6530
2	1	+1.4635	+1.6433	-57.0125	+1.4635
2	2	+6.8976	+7.5231	+8.2062	+6.8976
2	4	+29.7852	+30.8089	+199.4987	+29.7852

G.7 Quadratic trajectories under common noise laws

The objective is $f(x) = x^\top Ax/2$ in dimension 20, with diagonal eigenvalues geometrically spaced from 0.1 to 1. We set $L = \sigma = 1$, $m = 0.2$, $R \in \{0.1, 0.4\}$, and $N = 4000$. Initial points are constructed from Gaussian directions to have gradient norm 0.25 or 4. Independent coordinates of the additive noise are $Z/\sqrt{20}$, $(E-1)/\sqrt{20}$, or $T/\sqrt{60}$, where $Z \sim N(0,1)$, $E \sim \mathrm{Exp}(1)$, and $T \sim t_3$. All three laws have total centered variance one. The Student law has infinite fourth moment.

For every configuration, clipping chooses among gains $\{0.001, 0.003, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 1\}$ and normalization chooses among nine geometric step lengths from 0.0005 to R. Six seeds 102000–

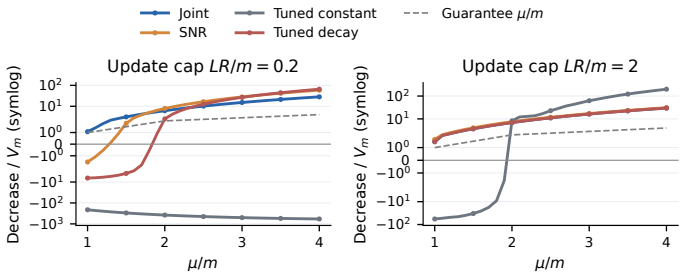

Figure 9: Actual objective-decrease minima on the common witness pool at 31 independently computed gradient norms. The signed logarithmic axis is linear on $[-2, 2]$ and logarithmic outside it; representative values appear in Table 7. The dashed curve is the target rule's theoretical lower guarantee.

102005 select each parameter by the full-horizon average gradient norm; seeds 202000–202031 evaluate it. Common initialization and noise pair methods within each seed. Each tuned baseline costs an additional 216,000 pilot oracle calls per configuration, beyond its 128,000 evaluation calls. The target rule is not tuned. Table 8 reports all 12 configurations; tuned clipping has a smaller mean primary metric in all 12, although some differences are small.

Table 8: Full-horizon mean gradient norm on all quadratic configurations. Entries are means $\pm$ one standard error over 32 seeds.

R	Initial norm	Noise	Minimax	Tuned clipping	Normalization
0.1	0.25	Gaussian	0.07660 ± 0.00030	0.05145 ± 0.00036	0.04885 ± 0.00047
0.1	0.25	Exponential	0.07647 ± 0.00029	0.05117 ± 0.00042	0.06349 ± 0.00069
0.1	0.25	Student	0.07466 ± 0.00030	0.05108 ± 0.00036	0.04448 ± 0.00037
0.1	4	Gaussian	0.14758 ± 0.00251	0.14434 ± 0.00193	0.18119 ± 0.00144
0.1	4	Exponential	0.14757 ± 0.00255	0.14430 ± 0.00195	0.18252 ± 0.00149
0.1	4	Student	0.14592 ± 0.00249	0.14121 ± 0.00194	0.16979 ± 0.00138
0.4	0.25	Gaussian	0.08761 ± 0.00030	0.05145 ± 0.00036	0.04845 ± 0.00053
0.4	0.25	Exponential	0.08752 ± 0.00028	0.05117 ± 0.00042	0.06309 ± 0.00074
0.4	0.25	Student	0.08590 ± 0.00030	0.05116 ± 0.00037	0.04400 ± 0.00041
0.4	4	Gaussian	0.13939 ± 0.00185	0.13895 ± 0.00178	0.17622 ± 0.00190
0.4	4	Exponential	0.13947 ± 0.00187	0.13904 ± 0.00180	0.17896 ± 0.00197
0.4	4	Student	0.13785 ± 0.00185	0.13777 ± 0.00179	0.16585 ± 0.00184

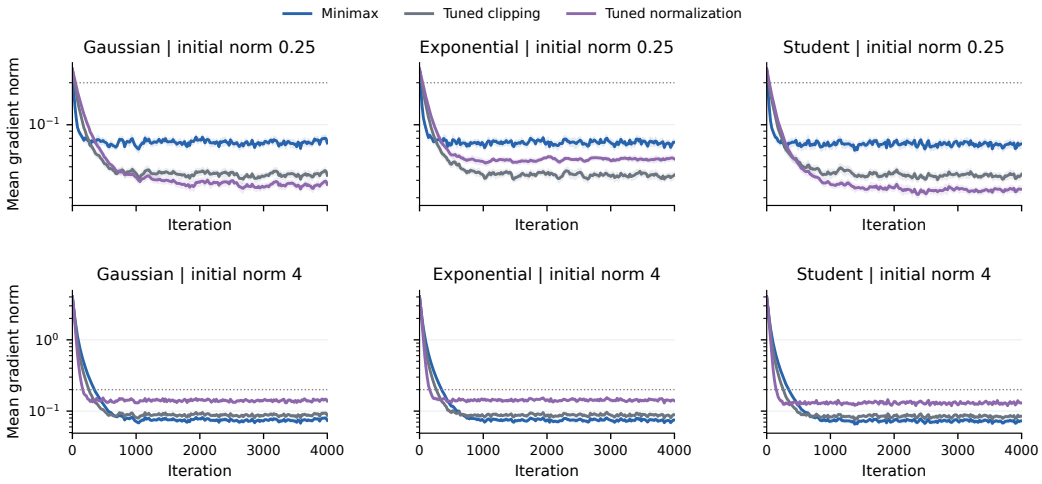

Figure 10: Quadratic trajectories with $R = 0.1$. All 4000 updates and both initial gradient norms are visible. Dotted horizontal lines mark $m = 0.2$; shading shows pointwise 95% intervals over 32 evaluation seeds.

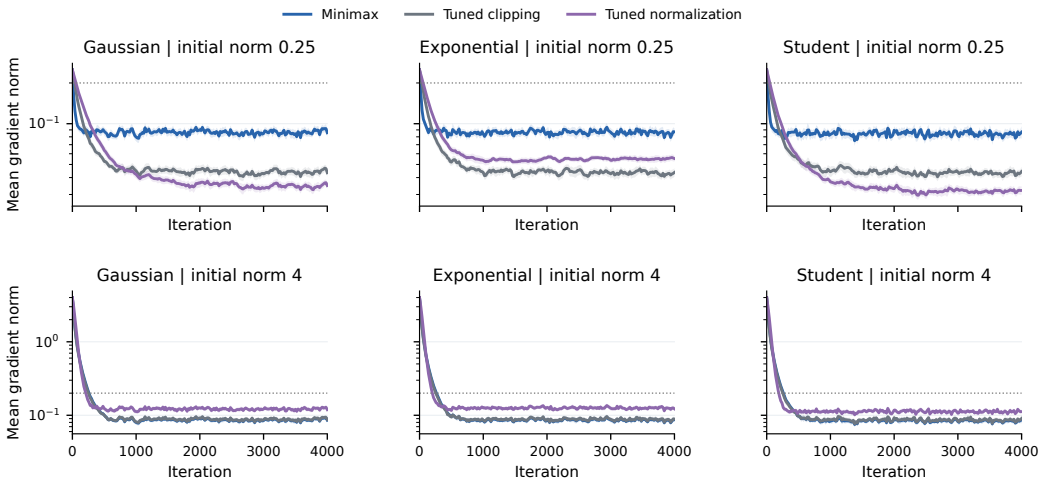

Figure 11: The same quadratic comparison with $R = 0.4$, retaining all noise laws, initializations, and the complete horizon.