ONLINE CONVEX OPTIMIZATION WITH DUELING FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Noisy binary comparison between two candidates is a common interface between human and learning systems, especially in modern large language model (LLM) post-training alignment. We study online convex optimization with dueling (pairwise comparison) feedback, where the learner observes only a binary preference between two queried points. We consider adversarial sequences of convex losses and measure regret with the loss at both queried points, under a comparison link with a known nonzero slope at the origin. We propose a simple reduction that converts dueling feedback into approximate gradients, enabling the use of standard first-order methods. We show that regret guarantees transfer under this reduction, yielding $\mathcal{O}(T^{3/4})$ static and adaptive regret, and $\mathcal{O}(T^{3/4}\sqrt{1+P_T/D})$ dynamic regret with unknown comparator path length P_T. For strongly convex losses, the static and adaptive bounds improve to $\widetilde{\mathcal{O}}(T^{2/3})$. For smooth losses, we presents unified dueling ellipsoidal FTRL, and proves $\widetilde{\mathcal{O}}(T^{2/3})$ static regret, which improves to $\widetilde{\mathcal{O}}(\sqrt{T})$ under additional strong convexity.

1 INTRODUCTION

Preference-based learning has become central to modern AI systems, where specifying an explicit reward is difficult, but obtaining binary *dueling* feedback via pairwise comparisons is convenient. In preference-based reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) pipelines, supervision is often reduced to pairwise preferences between candidate outputs, commonly modeled using a Bradley-Terry (logistic) link (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Muldrew et al., 2024; Deng et al., 2025). The same paradigm also appears in robotics (Bıyık et al., 2024; Sadigh et al., 2017), recommendation systems (Hofmann et al., 2013), and clinical treatment (Sui & Burdick, 2017).

The theory of dueling feedback is well-developed for finite action spaces (Yue et al., 2012; Ailon et al., 2014; Bengs et al., 2021; Saha et al., 2021) and contextual settings (Dudík et al., 2015; Saha & Krishnamurthy, 2022; Saha & Schapire, 2025). For continuous domains, Yue & Joachims (2009) bound regret defined on pairwise preference probabilities for a fixed concave utility, Lu et al. (2022) extend and track time-varying utilities through their per-round optimizers, and other works study stationary objectives under different structural and preference assumptions (Kumagai, 2017; Saha et al., 2025; Sharma et al., 2026). Appendix A discuss related works, and Table 2 compares the setting, metric and result of prior works in dueling convex optimization.

Many applications require optimizing time-varying objectives in adversarial environments for dynamic comparators, including adaptive resource allocation (Deb et al., 2024; Xiong et al., 2025), dynamic portfolio optimization (Sutiene et al., 2024), and interactive systems such as recommendation platforms or post-training human-in-the-loop alignment, where user preferences evolve, tasks vary, and preference optimization is increasingly run online(Xiong et al., 2024). The tuned decision is often a continuous parameter, such as reward-mixture weights, or a decoding parameter, and two nearby settings are compared.

Online convex optimization (OCO) provides a principled framework for such adversarial settings, evaluating these decisions against fixed or changing comparators (Zhang et al., 2018a; Garber & Kretzu, 2022). However, these methods rely on gradient or function value, therefore cannot be

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.

directly applied in preference-based settings where only binary comparisons are available. This raises a fundamental question:

Can we transfer adversarial online convex optimization results under limited dueling feedback?

We study regret measured directly on the underlying convex losses and against arbitrary comparator sequences in the adversarial environment. Our focus is a general reduction that makes these OCO guarantees accessible through dueling feedback. The difficulty is fundamentally informational. Consider three feedback models. In the two-point bandit setting, the learner observes noisy function values at two nearby points within the same round, enabling estimation of function differences and hence gradients. In the one-point bandit setting, only a single noisy function value is observed, making gradient estimation possible but significantly noisier. In contrast, under dueling feedback, the learner does not observe function values at all, but only a binary outcome obtained by applying a (possibly noisy) transfer function to the difference of function values at two queried points. Thus, the learner receives only a *1-bit, nonlinear transformation* of the information available in the two-point setting. This severe information constraint limits the ability to estimate gradients and makes it unclear whether standard online convex optimization techniques can be applied in this setting. We propose *Duel-To-First-Order (DTFO)*, a reduction framework that converts dueling feedback into approximate gradient information, enabling the use of standard first-order OCO algorithms. The key challenge is to control estimation errors so that the resulting updates yield regret guarantees on the underlying losses for different comparator benchmarks. This is more challenging than stochastic settings, since the loss functions change arbitrarily over time and estimation errors accumulate into regret. We summarize our contributions to the community and our technical novelties as follows.

Contributions.

- **A unified loss-based interval regret transfer** (Section 3). We use a generalized notion of *expected interval regret* over any fixed time interval $I \subseteq [1, T]$ against an arbitrary sequence of comparators $U = \{u_t\}_{t \in I}$ in Eq. (1). The DTFO wrapper and Theorem 1 convert first-order surrogate regret to expected interval regret, enabling integration of OCO algorithms. One quadratic-surrogate construction covers both convex and strongly convex losses, with the linear surrogate under $\sigma = 0$.
- **Static, dynamic, and adaptive consequences** (Section 4). Instantiate DTFO with first-order methods gives the first four rows of Table 1, including $\mathcal{O}(T^{3/4})$ static and adaptive regret and $\mathcal{O}(T^{3/4}\sqrt{1 + P_T/D})$ dynamic regret with unknown P_T for convex losses, and $\tilde{\mathcal{O}}(T^{2/3}\ell_T^{1/3})$ static and adaptive regret for strongly convex loss.
- **Improved static regret under smoothness** (Section 5). Under additional link bound, we present the unified dueling ellipsoidal FTRL (Algorithm 2) and establish Theorem 6, attaining $\mathcal{O}(T^{2/3}\ell_T^{1/3})$ static regret, and under strong convexity, $\mathcal{O}(\sqrt{T\ell_T})$(Corollaries 1 and 2).

Technical Novelty.

- **Comparison-to-surrogate transfer:** Lemma 1 separates the smoothed underlying-loss gradient from link distortion through the residual $\rho'(\Delta) + \gamma$, giving explicit $O(\delta)$ spherical bias. Retaining loss curvature in a shared quadratic surrogate (Eq. (2)) yields one transfer from pathwise first-order regret to regret at both queried actions, for fixed intervals and comparator paths (Theorem 1).
- **Ellipsoidal bias controlled by exploration geometry:** The comparison identity cancels the inverse exploration matrix before bounding bias, giving $O(\delta^2 \operatorname{tr}(A_t^2))$ error under the stronger link condition (Eq. (11)). A feasible comparison retains curvature without perturbing the comparator (Lemma 2), allowing this error to contract with the exploration ellipsoid in the barrier-FTRL analysis (Theorem 6).

2 PRELIMINARIES

Online Setup and Assumptions. We consider online convex optimization over a convex compact domain $\mathcal{K} \subset \mathbb{R}^d$ with diameter D. An oblivious adversary fixes the losses $f_1, \ldots, f_T$ before the

Loss class	Benchmark	Regret bound	Result
Convex	Static / Adaptive	$\mathcal{O}(T^{3/4})$	Theorem 2
Convex	Dynamic	$\mathcal{O}(T^{3/4}\sqrt{1+P_T/D})$	Theorem 3
Strongly convex	Static	$\mathcal{O}(T^{2/3}\ell_T^{1/3})$	Theorem 4
Strongly convex	Adaptive	$\mathcal{O}(T^{2/3}\ell_T^{2/3})$	Theorem 5
Smooth convex	Static	$\mathcal{O}(T^{2/3}\ell_T^{1/3})$	Corollary 1
Strongly convex, smooth	Static	$\mathcal{O}(\sqrt{T\ell_T})$	Corollary 2

Table 1: Dueling regret guarantees of this work, for fixed problem parameters. Here $\ell_T = 1 + \log T$ and $P_T = \sum_{t=2}^{T} \|w_t^* - w_{t-1}^*\|_2$. The first four rows respectively use DTFO with OGD, Improved Ader, SC-OGD, and binary IFLH, and the last two use ellipsoidal FTRL. Each row uses its stated algorithm and tuning, and the first row gives both benchmarks for one run. The dynamic algorithm does not require P_T. The smooth rows additionally assume $|\rho'''| \le M_3$ and access to a self-concordant barrier. All comparators lie in $\mathcal{K}$.

learner's randomization. Each loss is G-Lipschitz on $\mathcal{K}$ for a fixed $G > 0$ and has a known strong-convexity lower bound $\sigma \ge 0$, where $\sigma = 0$ is general convexity. Let $\mathbb{B}_d$ be the closed unit ball in $\mathbb{R}^d$, and assume $\mathbb{B}_d \subseteq \mathcal{K}$. For a given perturbation radius $\delta > 0$, the δ-shrunken domain $\mathcal{K}_\delta = \{x \in \mathcal{K} \mid x + \delta v \in \mathcal{K},\ \forall v \in \mathbb{B}_d\}$ contains the points in $\mathcal{K}$ with distance at least δ from the boundary, and δ is small enough that $\mathcal{K}_\delta \neq \varnothing$.

Dueling Feedback. The learner does not observe gradient or values directly. Instead, at each round t, it queries two points $w_t^+, w_t^- \in \mathcal{K}$ and receives a noisy binary preference feedback indicating which point has lower loss. Let $\mathcal{F}_{t-1}$ denote the information available just before the current query direction is drawn, including the center w_t and any internal randomization used to choose it. Then,

Definition 1 (Dueling Oracle)**.** *At each step t, when queried with a pair of points $(w^+, w^-) \in \mathcal{K}\times\mathcal{K}$, the dueling oracle returns a noisy binary preference bit $o_t \in \{-1, +1\}$. $o_t = 1$ if w^+ is preferred otherwise $o_t = -1$. The feedback is generated probabilistically according to an underlying **transfer function** $\rho : \mathbb{R} \to [-1, 1]$ such that:*

$$\mathbb{E}[o_t \mid \mathcal{F}_{t-1}, w^+, w^-] = \rho\big(f_t(w^+) - f_t(w^-)\big).$$

The notion of preference indicates that reversing the query order should reverse the preference probabilities. If P is the probability of $o_t = 1$ for (w^+, w^-), it is $1 - P$ for (w^-, w^+), so the conditional means are $2P - 1$ and $1 - 2P$. Thus, the transfer function should be odd (i.e., $\rho(z) = -\rho(-z)$), which implies $\rho(0) = 0$. In addition, we impose a mild regularity condition on the transfer function.

Assumption 1 (Transfer Function)**.** *The function ρ has a strictly negative first derivative $\rho'(0) = -\gamma$, where $\gamma > 0$, and bounded curvature $|\rho''(x)| \le M$ for all $x \in [-2G, 2G]$.*

We also call ρ links. Prior work Saha et al. (2025) allows transfer functions that may be flat at the origin. In contrast, we assume $\rho'(0) = -\gamma < 0$, which ensures that local comparisons provide a usable first-order signal. This condition is satisfied by Bradley-Terry (logistic), Thurstone-Mosteller (Gaussian), and links that are linear on the relevant loss-difference range, and enables the construction of gradient estimators with controlled variance.

Regret. To evaluate the algorithm's performance, we hold the algorithm accountable for the exact points queried to the oracle. For a fixed contiguous interval $I \subseteq [1, T]$ and comparator sequence $U = \{u_t\}_{t\in I}$ fixed independently of the learner's randomization, define the *expected interval regret* for the sequence of losses $\{f_t\}$ by

$$\text{Regret}(U, I) = \mathbb{E}\left[\sum_{t\in I} \frac{f_t(w_t^+) + f_t(w_t^-)}{2}\right] - \sum_{t\in I} f_t(u_t). \tag{1}$$

This general formulation recovers three notions of adversarial regret studied in this work. **Static Regret** evaluates performance over the full horizon $I = [1, T]$ against the single best fixed decision $u_t = w^* \in \mathcal{K}$, $\text{Regret}_S(T) = \max_{w^*\in\mathcal{K}} \text{Regret}(\{w^*\}_{t=1}^T, [1, T])$. **Dynamic Regret** evaluates performance over the full horizon $I = [1, T]$ against an arbitrary, changing sequence of comparators

$W^* = \{w_1^*, \ldots, w_T^*\}$, $\text{Regret}_D(W^*, T) = \text{Regret}(W^*, [1, T])$. **Adaptive Regret** ensures the algorithm performs well over any contiguous sub-interval $I = [t_1, t_2] \subseteq [1, T]$ against the best fixed decision for that specific interval $u_t = w_I^* \in \mathcal{K}$, $\text{Regret}_A(I) = \max_{w_I^* \in \mathcal{K}} \text{Regret}(\{w_I^*\}_{t \in I}, I)$.

Remark 1 (Adaptive and strongly adaptive regret.)**.** *We take the maximum over intervals of the expected regret against each interval's best fixed comparator, with expectation before maximization. Our bounds depend polynomially on the horizon T, whereas strongly adaptive bounds scale sublinearly with interval length up to logarithmic factors in T. This horizon dependence comes from a single global exploration radius tuned to T, whose reciprocal controls surrogate gradient magnitudes. Retuning the radius separately for each interval would not establish a simultaneous guarantee for one run.*

3 DTFO Reduction Framework

Our primary contribution, the Duel-To-First-Order (DTFO) wrapper, acts as a modular interface between any standard first-order online convex optimization algorithm $\mathcal{A}$ and the restrictive noisy 1-bit dueling oracle. It queries two perturbed points on a small sphere around the prediction of $\mathcal{A}$ and turns comparison bit into a surrogate gradient. A single transfer inequality then handles both cases, for fixed, moving, and interval comparators. 3.1 specifies the wrapper, 3.2 establishes its comparison identity and bias bound, and 3.3 transfers surrogate regret to expected interval regret. A known $\sigma \geq 0$ covers both convex and strongly convex losses, and Appendices B-D provide the technical lemmas and full proofs.

3.1 Dueling-to-first-order wrapper

Fix $0 < \delta \leq 1/2$ and run a base algorithm $\mathcal{A}$ on $\mathcal{K}_\delta$. At round t, its prediction w_t determines the query pair $w_t^\pm = w_t \pm \delta U_t$, where U_t is uniform on $\mathbb{S}^{d-1}$. From the comparison bit o_t, DTFO constructs

$$\widehat{g}_t = -\frac{d}{2\gamma\delta} o_t U_t, \qquad q_t(w) = \langle \widehat{g}_t, w \rangle + \frac{\sigma}{2} \|w - w_t\|_2^2. \tag{2}$$

The surrogate loss $q_t(w)$ is linear when $\sigma = 0$ and preserves curvature when $\sigma > 0$. Algorithm 1 gives the wrapper; Appendix E supplies concrete base algorithms.

Algorithm 1 Duel-To-First-Order (DTFO) wrapper $\mathcal{W}(\mathcal{A})$

Require: Base algorithm $\mathcal{A}$, domain $\mathcal{K}$, radius $0 < \delta \leq 1/2$, known $\gamma > 0$ and $\sigma \geq 0$.
1: Initialize $\mathcal{A}$ on $\mathcal{K}_\delta$.
2: **for** $t = 1, \ldots, T$ **do**
3: Receive $w_t \in \mathcal{K}_\delta$ from $\mathcal{A}$.
4: Draw $U_t \sim \text{Unif}(\mathbb{S}^{d-1})$; query $w_t \pm \delta U_t$ and observe $o_t \in \{+1, -1\}$.
5: Form $\widehat{g}_t$ and q_t using Eq. (2).
6: Update $\mathcal{A}$ with q_t, using $\nabla q_t(w) = \widehat{g}_t + \sigma(w - w_t)$ as needed.
7: **end for**

One shared curvature-preserving surrogate. If $\mathcal{A}$ aggregates experts, an expert predicting $w_{t,j}$ uses $\nabla q_t(w_{t,j}) = \widehat{g}_t + \sigma(w_{t,j} - w_t)$, so every expert is updated from the same comparison, which enables first-order meta-learning algorithms with one shared gradient per round. On $\mathcal{K}_\delta$ the surrogate gradients are bounded by $H_\delta = d/(2\gamma\delta) + \sigma D$, and $\nabla q_t(w_t) = \widehat{g}_t$. Additionally, for every realized bit, the surrogate retains the original curvature of the quadratic surrogate loss through $q_t(w_t) - q_t(v_t) = \langle \widehat{g}_t, w_t - v_t \rangle - \frac{\sigma}{2}\|w_t - v_t\|_2^2$, which permits strongly convex base algorithms to exploit the same curvature that appears in the original losses. Moreover, it σ/H_δ^2-exp-concave on $\mathcal{K}_\delta$, because its Hessian is the known matrix σI_d.

3.2 Comparison identity, estimator bounds and bias control

Lemma 1 gives a comparison identity for spherical and ellipsoidal exploration, and then the norm and bias bounds for the spherical estimator. Appendix C proves it in detail using the smoothing lemmas in Appendix B.

Lemma 1 (Comparison identity and spherical estimator bounds)**.** *Under Assumption 1, let f be a convex G-Lipschitz loss on $\mathcal{K}$. Fix x and a symmetric positive-definite matrix A such that $x + \delta A\mathbb{B}_d \subseteq \mathcal{K}$ and $\delta\|A\|_{\mathrm{op}} \le 1$. For $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $V \sim \mathrm{Unif}(\mathbb{B}_d)$, define*

$$\Delta(v) = f(x+\delta Av) - f(x-\delta Av), \qquad m(v) = \frac{\nabla f(x+\delta Av) + \nabla f(x-\delta Av)}{2}, \quad \bar{g} = \mathbb{E}_V m(V).$$

If o is the comparison of $x \pm \delta AU$ and $\widehat{g} = -d\, oA^{-1}U/(2\gamma\delta)$, then

$$\mathbb{E}\widehat{g} = \bar{g} + B, \qquad B = -\frac{1}{\gamma}\mathbb{E}_V[\{\rho'(\Delta(V)) + \gamma\}m(V)]. \tag{3}$$

In particular, for $A = I_d$,

$$\|\widehat{g}\|_2 = G_{\mathrm{surr}} := \frac{d}{2\gamma\delta}, \qquad \|B\|_2 \le \frac{2d}{d+1}\frac{MG^2}{\gamma}\delta. \tag{4}$$

Proof outline. Applying the weak sphere-to-ball identity $d\mathbb{E}_U[h(U)U] = \mathbb{E}_V \nabla h(V)$ from Lemma 4 to $h = \rho \circ \Delta$, with $\nabla\Delta(v) = 2\delta A m(v)$, gives

$$\mathbb{E}\widehat{g} = -\frac{A^{-1}}{2\gamma\delta}\mathbb{E}_V \nabla(\rho\circ\Delta)(V) = -\frac{1}{\gamma}\mathbb{E}_V[\rho'(\Delta(V))m(V)].$$

Thus A^{-1} cancels before any norm bound. For $A = I_d$, use $|\rho'(z)+\gamma| \le M|z|$, $|\Delta(v)| \le 2G\delta\|v\|_2$, $\|m(v)\|_2 \le G$, and $\mathbb{E}\|V\|_2 = d/(d+1)$. The estimator norm follows from $o^2 = 1$. □

Integrate first, expand the link later. The sphere-to-ball identity is classical for one-point estimators, but never applied under the link. We apply it to the composition $\rho \circ \Delta$ of the link with the loss difference, before the link is linearized. The residual in Eq. (3) then involves $\rho'(\Delta) + \gamma$ rather than $\rho(\Delta) + \gamma\Delta$, the matrix A^{-1} in $\widehat{g}$ cancels against the factor A produced by the chain rule, and no derivative of f at a query point is needed. The direct route bounds $|\rho(\Delta) + \gamma\Delta| \le M\Delta^2/2$ inside $\widehat{g}$ and multiplies by the estimator scale $d/(2\gamma\delta)$. It gives only $\|B\|_2 \le dMG^2\delta/\gamma$, a factor $(d+1)/2$ worse than Eq. (4). For ellipsoidal exploration it also retains a factor $\|A^{-1}\|_{\mathrm{op}}$, which grows without bound as the ellipsoid flattens near $\partial\mathcal{K}$. Section 5 relies on this cancellation.

3.3 Unified interval regret transfer

Let $\mathcal{F}_{t-1}$ be the σ-algebra generated by the learner's history before drawing U_t, including past directions and comparisons and any internal randomization used to choose w_t. Thus w_t is $\mathcal{F}_{t-1}$-measurable, while U_t is sampled uniformly independently of $\mathcal{F}_{t-1}$. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and let $\widetilde{f}_t(x) = \mathbb{E}_V f_t(x + \delta V)$ on $\mathcal{K}_\delta$. Define

$$C_{\mathrm{sph}} = G(D+1) + \frac{2d}{d+1}\frac{MG^2D}{\gamma}. \tag{5}$$

Theorem 1 converts any pathwise surrogate-regret bound into an expected interval regret bound, with full proof in Appendix D.

Theorem 1 (DTFO interval regret transfer)**.** *Suppose the losses are G-Lipschitz and σ-strongly convex, where $\sigma \ge 0$, and the feedback satisfies Assumption 1. For any fixed interval $I = [a, b] \subseteq [1, T]$ and comparator sequence $W_I^* = (w_t^*)_{t\in I} \subseteq \mathcal{K}$ independent of the learner's randomization, set $v_t = \Pi_{\mathcal{K}_\delta}(w_t^*)$. Algorithm 1 satisfies*

$$\mathbb{E}\left[\sum_{t\in I}\left(\frac{f_t(w_t^+) + f_t(w_t^-)}{2} - f_t(w_t^*)\right)\right] \le \mathbb{E}\left[\sum_{t\in I}\big(q_t(w_t) - q_t(v_t)\big)\right] + C_{\mathrm{sph}}\delta|I|. \tag{6}$$

Moreover, $\|v_t - w_t^*\|_2 \le D\delta$ and $\sum_{t=a+1}^{b}\|v_t - v_{t-1}\|_2 \le \sum_{t=a+1}^{b}\|w_t^* - w_{t-1}^*\|_2$.

Proof outline. Since $\mathbb{B}_d \subseteq \mathcal{K}$, $(1-\delta)w_t^* \in \mathcal{K}_\delta$. The projection error is therefore at most $\delta\|w_t^*\|_2 \le D\delta$, and non-expansiveness of projection gives the path-length bound. Let $\bar{g}_t = \mathbb{E}_V \nabla f_t(w_t + \delta V)$. Applying strong convexity between $w_t + \delta V$ and $v_t + \delta V$ and averaging yields

$$\begin{aligned}\widetilde{f}_t(w_t) - \widetilde{f}_t(v_t) &\le \langle \bar{g}_t, w_t - v_t\rangle - \frac{\sigma}{2}\|w_t - v_t\|_2^2 \\ &\le \mathbb{E}_t[q_t(w_t) - q_t(v_t)] + D\|B_t\|_2,\end{aligned}$$

where Lemma 1 gives $\mathbb{E}_t \widehat{g}_t = \bar{g}_t + B_t$. The averaged supporting inequality remains valid on $\partial\mathcal{K}_\delta$; see Lemma 3 in Appendix B.

To account for both queried actions, couple a uniform ball point as $V = RU$, where R is independent of the uniform sphere point U and $\mathbb{E}R = d/(d+1)$. Symmetry and Lemma 5 give

$$\mathbb{E}_t \frac{f_t(w_t^+) + f_t(w_t^-)}{2} - \widetilde{f}_t(w_t) \le \frac{G\delta}{d+1}, \qquad \widetilde{f}_t(v_t) - f_t(v_t) \le \frac{dG\delta}{d+1}.$$

Add $f_t(v_t) - f_t(w_t^*) \le GD\delta$ and the bias bound Eq. (4), then sum over I and take total expectations. All loss evaluations are in $\mathcal{K}$. Appendix D gives the full proof. □

If $\mathcal{A}$ guarantees $\sum_{t\in I}[q_t(w_t) - q_t(v_t)] \le R_{\mathcal{A}}(v_{a:b}, I)$ for every realized surrogate sequence, its expected query regret is at most $\mathbb{E}R_{\mathcal{A}}(v_{a:b}, I) + C_{\mathrm{sph}}\delta|I|$. Choosing the comparator and the base bound gives static, dynamic, and adaptive guarantees.

Reductions from bandit feedback to first-order methods usually rely on an unbiased estimate of the smoothed gradient and a single fixed comparator. However, our estimate is biased due to the link function, the surrogate gradient depends on the current bit, and the smoothed loss is centered at w_t while the losses are charged at $w_t^{\pm}$. Theorem 1 handles all three at once. It requires the bound on surrogate sequences, which accommodates the dependence of q_t on the bit. It charges the link bias ($D\|B_t\|_2$), the smoothing of both queried points ($G\delta$, through the coupling $V = RU$), and the comparator projection ($GD\delta$) to one additive term $C_{\mathrm{sph}}\delta|I|$. The statement holds for any fixed interval and comparator path, and projection does not change the path length bound. Static, dynamic, and adaptive guarantees therefore follow from the base algorithm's analysis applied to q_t.

4 Consequences of DTFO Wrapper: Algorithms and Regrets

Using the Duel-To-FO (DTFO) wrapper as a meta-algorithm, we instantiate Theorem 1 with standard first-order algorithms. Section 4.1 handles convex losses using Online Gradient Descent (OGD) and Improved Ader (IA), whereas Section 4.2 uses Strongly-Convex(SC)-OGD and Improved Following the Leading History (IFLH) for strongly convex losses. Appendix E gives the implementation and proof details. Let $0 < \delta \le 1/2$, $\ell_T = 1 + \log T$, and use the constants are as in Section 3. Radius δ is chosen once per run for the whole horizon, and comparators belong to $\mathcal{K}$.

4.1 Convex losses: static, interval, and dynamic regret

With $\sigma = 0$, DTFO applied to projected OGD gives Online Relative Gradient Descent (ORGD). The usual projection argument (Hazan, 2016) applies pathwise to the shared linear surrogates. Theorem 2 gives the static and adaptive regret of ORGD, with the implementation and proof in Appendix E.1.

Theorem 2 (Static and adaptive regret of ORGD)**.** *Under the assumptions of Theorem 1, use $\sigma = 0$ and run Algorithm 3 with $\eta_t = D/(G_{\mathrm{surr}}\sqrt{t})$. For every fixed interval $I = [s, e] \subseteq [1, T]$,*

$$\mathrm{Regret}_A(I) \le \frac{DG_{\mathrm{surr}}}{2}\left(\sqrt{e} + \sum_{t=s}^{e} t^{-1/2}\right) + C_{\mathrm{sph}}\delta|I| \le \frac{3}{2}DG_{\mathrm{surr}}\sqrt{T} + C_{\mathrm{sph}}\delta|I|. \tag{7}$$

A single run with $\delta = cT^{-1/4}$, $0 < c \le 1/2$, therefore has $\mathrm{Regret}_S(T) = \mathcal{O}(T^{3/4})$ and $\max_I \mathrm{Regret}_A(I) = \mathcal{O}(T^{3/4})$.

For moving comparators, Dynamic ORGD uses IA framework (Zhang et al., 2018a) , which calls OGD experts that share the comparison at the mixture point but with a geometric grid of step sizes. Theorem 3 gives the resulting path-length bound, and Appendix E.2 specifies the grid and aggregation rule and proves the guarantee without knowing P_T.

Theorem 3 (Dynamic regret of Dynamic ORGD)**.** *Under the assumptions of Theorem 1, set $\sigma = 0$ and run Algorithm 4 as described. For every comparator path $W^* = (w_t^*)_{t=1}^T$ fixed independently of the learner's randomization, let $P_T = \sum_{t=2}^T \|w_t^* - w_{t-1}^*\|_2$. Then*

$$\mathrm{Regret}_D(W^*, T) \le \mathcal{O}\Big(DG_{\mathrm{surr}}\sqrt{T(1 + P_T/D)}\Big) + C_{\mathrm{sph}}\delta T. \tag{8}$$

With $\delta = cT^{-1/4}$, $0 < c \le 1/2$, this gives $\mathcal{O}(T^{3/4}\sqrt{1 + P_T/D})$ dynamic regret without knowing P_T, using one comparison and $\mathcal{O}(\log T)$ experts per round.

Similar to IA, Dynamic ORGD shares one duel per round at the mixture point across all $\mathcal{O}(\log T)$ experts and adapts to the unknown P_T. However, the radius must be common to all experts, because changing it changes the queried points and the comparison bit. This is why the radius is fixed at $cT^{-1/4}$ independently of P_T (Appendix E.2).

4.2 Strongly convex losses: static and adaptive regret

For $\sigma > 0$, SC-OGD runs Algorithm 3 with step size $\eta_t = 1/(\sigma t)$ (Hazan, 2016) on the quadratic surrogates in Eq. (2). Theorem 4 gives its static regret, with full proof in Appendix E.3.

Theorem 4 (Static regret for strongly convex losses)**.** *Under the assumptions of Theorem 1 with $\sigma > 0$, running Algorithm 3 with step size $\eta_t = 1/(\sigma t)$ satisfies*

$$\mathrm{Regret}_S(T) \le \frac{G_{\mathrm{surr}}^2}{2\sigma}\mathsf{H}_T + C_{\mathrm{sph}}\delta T, \qquad \mathsf{H}_T := \sum_{t=1}^{T}\frac{1}{t}. \tag{9}$$

Choosing $\delta = c(\ell_T/T)^{1/3}$, $0 < c \le 1/2$, gives $\mathcal{O}(T^{2/3}\ell_T^{1/3})$ static regret.

For intervals instead of whole horizon, binary IFLH (Zhang et al., 2018b) runs SC-OGD experts on the shared q_t. Their gradients are bounded by H_δ rather than G_{surr}, and the static and interval guarantees use different algorithms and radii. Theorem 5 gives the adaptive regret, and Appendix E.4 gives the complete implementation and proof.

Theorem 5 (Adaptive regret for strongly convex losses)**.** *Under the assumptions of Theorem 1 with $\sigma > 0$, Algorithm 5 satisfies, for every fixed interval I,*

$$\mathrm{Regret}_A(I) \le \frac{H_\delta^2}{2\sigma} m_I(1 + 3\log T) + C_{\mathrm{sph}}\delta|I|, \qquad m_I = \lceil \log_2 |I| \rceil + 1. \tag{10}$$

If we choose a single radius $\delta = c(\ell_T^2/T)^{1/3}$ with $0 < c \le 1/2$, the adaptive regret is $\max_I \mathrm{Regret}_A(I) = \mathcal{O}(T^{2/3}\ell_T^{2/3})$. The algorithm uses one comparison query and $\mathcal{O}(\log T)$ active experts per round.

The SC-OGD cancellation and the exponential weighting of IFLH apply here because the curvature and exp-concavity of q_t do not depend on the comparison bit (Section 3.1), and they require no smoothness of f_t. The logarithmic surrogate regret of SC-OGD therefore holds pathwise, and the rate improves from $T^{3/4}$ to $T^{2/3}\ell_T^{1/3}$ with no change to the dueling machinery.

5 Smooth Losses and Ellipsoidal Estimators

Ellipsoidal exploration lets the iterates approach the boundary and exploits smoothness there, while a fixed Euclidean sphere would leave $\mathcal{K}$. Assume additionally that each f_t has an L-Lipschitz gradient on $\mathcal{K}$ and that the odd link ρ is three times continuously differentiable on $[-2G, 2G]$, with $|\rho'''| \le M_3$. We keep the known $\sigma \ge 0$ and assume access to a ν-self-concordant barrier Φ with $\nabla^2\Phi \succ 0$ on $\operatorname{int}\mathcal{K}$ (Definition 2), its Hessian, and the Follow-The-Regularized-Leader (FTRL) subroutine. Algorithm 2 gives the Unified Dueling Ellipsoidal FTRL.

Lemma 7 ensures that the queries are feasibility, that $\|A_t\|_{\mathrm{op}} \le D$, and that $\|A_t\|_{\mathrm{op}}^2 \le 1/(\eta\sigma t)$ when $\sigma > 0$. Lemma 8 gives local-norm identity $\|\widehat{g}_t\|_{H_t^{-1}}^2 = \frac{d^2}{4\gamma^2\delta^2}$, where $\|g\|_{H_t^{-1}}^2 = g^\top H_t^{-1} g$. It also bounds the bias B resulting from the comparison identity (Eq. (3)) in Lemma 1.

Curvature known before the duel. The H_t is the Hessian at w_t of the updated FTRL objective $\Phi + \eta\sum_{s=1}^{t} q_s$, including the current surrogate's curvature. It is available before observing the comparison because the unknown feedback enters q_t only through its linear term, while $\nabla^2 q_t = \sigma I_d$ is known. Sampling with it makes the local-norm identity above hold for the matrix that controls FTRL stability.

Algorithm 2 Unified Dueling Ellipsoidal FTRL

Require: $0 < \delta \le \delta_0 := \min\{1/2, 1/D\}$, $0 < \eta \le \gamma\delta/d$, known $\sigma \ge 0$, γ, and barrier Φ.
1: $w_1 \leftarrow \arg\min_{w \in \operatorname{int} \mathcal{K}} \Phi(w)$.
2: **for** $t = 1, \dots, T$ **do**
3: $H_t \leftarrow \nabla^2\Phi(w_t) + \eta\sigma t I_d$; $A_t \leftarrow H_t^{-1/2}$.
4: Draw $U_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ independently of $\mathcal{F}_{t-1}$; query $w_t^{\pm} = w_t \pm \delta A_t U_t$ and observe o_t.
5: $\widehat{g}_t \leftarrow -d\, o_t A_t^{-1} U_t/(2\gamma\delta)$; $q_t(w) \leftarrow \langle \widehat{g}_t, w\rangle + \frac{\sigma}{2}\|w - w_t\|_2^2$.
6: $w_{t+1} \leftarrow \arg\min_{w \in \operatorname{int} \mathcal{K}} \{\Phi(w) + \eta \sum_{s=1}^{t} q_s(w)\}$.
7: **end for**

Section 5.1 bounds the bias of the ellipsoidal gradient estimator, and gives the per-round comparison-to-surrogate transfer. Section 5.2 uses Lemma 2 to establish static-regret bound before tuning for convex and strongly convex cases. Appendix F supplies the full details and proofs.

5.1 Ellipsoidal gradient estimator, Quadratic bias and regret transfer

Condition on $\mathcal{F}_{t-1}$ and apply Lemma 1 with $A = A_t$. Oddness implies $\rho''(0) = 0$, hence $|\rho'(z) + \gamma| \le M_3 z^2/2$ on $[-2G, 2G]$. Using $|\Delta(v)| \le 2G\delta\|A_t v\|_2$ and $\mathbb{E}\|A_t V\|_2^2 = \operatorname{tr}(A_t^2)/(d+2)$ gives the quadratic bias bound proved for Lemma 8 in Appendix F.3:

$$\|\mathbb{E}_t \widehat{g}_t - \bar{g}_t\|_2 \le \frac{2M_3 G^3 \delta^2}{\gamma(d+2)} \operatorname{tr}(A_t^2), \qquad \bar{g}_t = \mathbb{E}_V \nabla f_t(w_t + \delta A_t V). \tag{11}$$

The estimator contains A_t^{-1}, whose norm can grow as the exploration ellipsoid shrinks. Lemma 1 cancels this matrix before bounding the conditional-mean error. Thus, Eq. (11) scales with the squared semiaxis lengths of the exploration ellipsoid. Its dependence on $\delta^2 \operatorname{tr}(A_t^2)$ matches the smoothness error of the queried losses, allowing both errors to decrease with the same geometry. Lemma 2 is the per-round transfer for this geometry, from the queried loss to the surrogate loss. Its proof in Appendix F.4 retains curvature and allows boundary comparators without perturbing them. The transfer requires smoothness only on $\mathcal{K}$.

Lemma 2 (Per-round Ellipsoidal Regret Transfer)**.** *For every fixed $u \in \mathcal{K}$, Algorithm 2 satisfies*

$$\mathbb{E}_t\left[\frac{f_t(w_t^+) + f_t(w_t^-)}{2} - f_t(u)\right] \le \mathbb{E}_t[q_t(w_t) - q_t(u)] + c_{d,\sigma}\delta^2 \operatorname{tr}(A_t^2), \tag{12}$$

where $c_{d,\sigma} = \frac{L}{2d} + \frac{L-\sigma}{2(d+2)} + \frac{2M_3 G^3 D}{\gamma(d+2)}$. Further, $c_{d,\sigma} \operatorname{tr}(A_t^2) \le C\|A_t\|_{\mathrm{op}}^2$ where $C := L + \frac{2M_3 G^3 D}{\gamma}$.

Proof sketch. Apply smoothness and strong convexity at the same feasible point $y = w_t + \delta A_t V$, where V is uniform on the unit ball. Subtracting the inequalities and using $\mathbb{E}_V[V] = 0$ gives

$$f_t(w_t) - f_t(u) \le \langle \bar{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2 + \frac{L-\sigma}{2(d+2)}\delta^2 \operatorname{tr}(A_t^2).$$

This step evaluates the loss only at y, w_t, and u and requires no feasibility of a perturbed comparator $u + \delta A_t V$. Adding the query-smoothing error and Eq. (11) supplies the claim, with the retained negative quadratic term absorbed exactly into $q_t(w_t) - q_t(u)$. Appendix F.4 gives the full proof. □

The exploration ellipsoid lies in $\mathcal{K}$ around w_t, but it need not do so around a boundary comparator u, so these values can lie outside $\mathcal{K}$. Moving u inward would cost a first-order term that does not scale with $\delta^2 \operatorname{tr}(A_t^2)$. Lemma 2 instead applies smoothness and strong convexity at the same feasible point $y = w_t + \delta A_t V$ and compares directly with $f_t(u)$. The negative quadratic term survives exactly, every error is of order $\delta^2 \operatorname{tr}(A_t^2)$, and smoothness is needed only on $\mathcal{K}$. Together these properties make the comparison compatible with quadratic-surrogate FTRL and its shrinking geometry.

5.2 Improved Static Regret for Smooth Losses

Theorem 6 combines the per-round query-to-surrogate transfer with the path-wise FTRL bound into a master static-regret guarantee. Appendix F.6 gives the complete proof.

Theorem 6 (Ellipsoidal FTRL static regret)**.** *Under the assumptions of this section, for every* $0 < \alpha < 1$*, Algorithm 2 satisfies*

$$\mathrm{Regret}_S(T) \le \frac{\nu \log(1/\alpha)}{\eta} + \frac{\eta d^2 T}{2\gamma^2\delta^2} + C\delta^2 \sum_{t=1}^{T} \mathbb{E}\|A_t\|_{\mathrm{op}}^2 + \alpha GDT. \tag{13}$$

Proof outline. The barrier-FTRL bound applied to the quadratic surrogates (Lemma 9) gives, for every interior comparator v,

$$\sum_{t=1}^{T}[q_t(w_t) - q_t(v)] \le \frac{\Phi(v) - \Phi(w_1)}{\eta} + 2\eta \sum_{t=1}^{T} \|\widehat{g}_t\|_{H_t^{-1}}^2.$$

The step-size cap makes the Newton decrement at most $1/2$, which controls the FTRL stability term (Hazan & Levy, 2014). For $u \in \mathcal{K}$, set $v = (1-\alpha)u + \alpha w_1$. Barrier growth bounds $\Phi(v) - \Phi(w_1)$ by $\nu \log(1/\alpha)$, and Lipschitzness costs at most αGDT in the original losses. Combine Lemma 2 with $\|\widehat{g}_t\|_{H_t^{-1}}^2 = \frac{d^2}{4\gamma^2\delta^2}$ and take expectations. □

In Eq. (13), the third term contains both nonlinear-link bias and smoothness error; the other terms come from barrier FTRL and interior approximation of the comparator. At a fixed radius δ, the estimator's local-norm cost is independent of A_t, while the geometric sum is bounded by D^2T for convex losses and by $\mathsf{H}_T/(\eta\sigma)$ under strong convexity. The comparison-feedback error therefore fits the shrinking geometry of the FTRL analysis, leading to the two tunings below.

Set $\alpha = 1/(T+1)$, $B_T = \nu \log(T+1)$, and $V_0 = d^2/(2\gamma^2)$. Let η be constant within each run. Corollary 1 uses $\|A_t\|_{\mathrm{op}} \le D$, and Appendix F.7 proves the tuning below, including the caps.

Corollary 1 (Smooth convex tuning)**.** *For* $\sigma = 0$ *and* $C_{\mathrm{geom}} := CD^2 > 0$*, choose* $\delta = \min\left\{\delta_0, \left(\frac{B_T V_0}{C_{\mathrm{geom}}^2 T}\right)^{1/6}\right\}$*, and* $\eta = \min\left\{\delta\sqrt{\frac{B_T}{V_0 T}}, \frac{\gamma\delta}{d}\right\}$*. For every* T*, regret is at most* $B_T/\eta + V_0\eta T/\delta^2 + C_{\mathrm{geom}}\delta^2 T + GD$*. Once both caps are inactive,* $\mathrm{Regret}_S(T) \le 3(B_T V_0 C_{\mathrm{geom}})^{1/3} T^{2/3} + GD = \mathcal{O}(T^{2/3}\ell_T^{1/3})$.

If $C_{\mathrm{geom}} = 0$, a fixed feasible radius and the capped learning rate in Corollary 1 give $\mathcal{O}(\sqrt{T\ell_T})$ directly. For $\sigma > 0$, the bound $\|A_t\|_{\mathrm{op}}^2 \le 1/(\eta\sigma t)$ gives harmonic geometric error. Corollary 2 uses this improvement, and Appendix F.8 proves its tuning.

Corollary 2 (Strongly convex smooth tuning)**.** *For* $\sigma > 0$*, fix* $0 < \delta \le \delta_0$ *independent of* T *and set* $V_\delta = V_0/\delta^2$*,* $K_T = B_T + \frac{C\delta^2}{\sigma}\mathsf{H}_T$*, and* $\eta = \min\left\{\sqrt{\frac{K_T}{V_\delta T}}, \frac{\gamma\delta}{d}\right\}$*. For every* T*, regret is at most* $K_T/\eta + V_\delta\eta T + GD$*. Once the learning-rate cap is inactive,* $\mathrm{Regret}_S(T) \le 2\sqrt{V_\delta T K_T} + GD = \mathcal{O}(\sqrt{T\ell_T})$.

Harmonic accumulation of the dueling error. Sampling with the regularized Hessian $\nabla^2\Phi(w_t) + \eta\sigma t I_d$ is the device of Hazan & Levy (2014) for value feedback, where it controls the smoothing error. For comparisons bit, we have a second error, the link bias B in Eq. (3). Without Lemma 1 and Eq. (11) that bias would carry $\|A_t^{-1}\|_{\mathrm{op}}$, which grows at least as $\sqrt{\eta\sigma t}$ because $H_t \succeq \eta\sigma t I_d$. Through Lemma 1 and Eq. (11) the link bias has the same $\delta^2 \operatorname{tr}(A_t^2)$ form as the smoothness error, so both accumulate harmonically, and a radius fixed independently of T yields $\mathcal{O}(\sqrt{T\ell_T})$.

6 Conclusion

We studied adversarial online convex optimization with one binary comparison per round and regret measured on the losses at both queried points. The DTFO reduction uses controlled spherical gradient estimator and curvature-preserving quadratic surrogates, yielding a unified interval transfer for convex and strongly convex losses. Its instantiations give static, path-length-dependent dynamic, and adaptive guarantees. For smooth losses, a trace-controlled ellipsoidal bias bound and a feasible comparison argument lead to improved static regret, with strong convexity making the accumulated geometric error harmonic and hence further improvement. Extending the framework to $K > 2$ ranking feedback and more preference models are promising directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools for [edit a research paper to improve readability, suggest a structure for a research paper]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, provide critical ingredients for proving mathematical claims], and [generate synthetic data sets, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results] are not applicable to this work. We have reviewed all AI-assisted work. We have reviewed any sentence that we have refined using LLMs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For our theoretical results, we have provided detailed proofs in Appendices. We refer to the corresponding Appendices near the claim.

REFERENCES

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the Annual Conference on Learning Theory*, 2008.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the Annual Conference on Learning Theory*, 2010.

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the International Conference on Machine Learning*, 2014.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

Erdem Bıyık, Nicolas Huynh, Mykel J. Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning and optimization. *The International Journal of Robotics Research*, 43(5):665–684, 2024.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the Advances in Neural Information Processing Systems*, 2017.

Rohan Deb, Aadirupa Saha, and Arindam Banerjee. Think before you duel: Understanding complexities of preference learning under constrained resources. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2024.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. In *Proceedings of the Advances in Neural Information Processing Systems*, 2025.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the Annual Conference on Learning Theory*, 2015.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the annual ACM-SIAM symposium on Discrete algorithms*, 2005.

Hidde Fokkema, Dirk Van der Hoeven, Tor Lattimore, and Jack J. Mayo. Online newton method for bandit convex optimisation. In *Proceedings of the Annual Conference on Learning Theory*, 2024.

Dan Garber and Ben Kretzu. New projection-free algorithms for online convex optimization with adaptive regret guarantees. In *Proceedings of the Annual Conference on Learning Theory*, 2022.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Kfir Y. Levy. Bandit convex optimization: Towards tight bounds. In *Proceedings of the Advances in Neural Information Processing Systems*, 2014.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten De Rijke. Reusing historical interaction data for faster online learning to rank for ir. In *Proceedings of the ACM international conference on Web search and data mining*, pp. 183–192, 2013.

Wataru Kumagai. Regret analysis for continuous dueling bandit. In *Proceedings of the Advances in Neural Information Processing Systems*, 2017.

Shiyin Lu, Yuan Miao, Ping Yang, Yao Hu, and Lijun Zhang. Non-stationary dueling bandits for online learning to rank. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, 2022.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *Proceedings of the International Conference on Machine Learning*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022.

Mohammad Pedramfar and Vaneet Aggarwal. From linear to linearizable optimization: A novel framework with applications to stationary and non-stationary dr-submodular optimization. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active preference-based learning of reward functions. In *Proceedings of Robotics: Science and Systems*, 2017.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 2022.

Aadirupa Saha and Robert E. Schapire. Efficient and near-optimal algorithm for contextual dueling bandits with offline regression oracles. In *Proceedings of the Advances in Neural Information Processing Systems*, 2025.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *Proceedings of the International Conference on Machine Learning*, 2021.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Dueling convex optimization with general preferences. In *Proceedings of the International Conference on Machine Learning*, 2025.

Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.

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

Mudit Sharma, Shweta Jain, Vaneet Aggarwal, and Ganesh Ghalme. Lipschitz dueling bandits over continuous action spaces. *arXiv preprint arXiv:2604.00523*, 2026.

Yanan Sui and Joel W. Burdick. Correlational dueling bandits with application to clinical treatment in large decision spaces. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.

Kristina Sutiene, Peter Schwendner, Ciprian Sipos, Luis Lorenzo, Miroslav Mirchev, Petre Lameski, Audrius Kabasinskas, Chemseddine Tidjani, Belma Ozturkkal, and Jurgita Cerneviciene. Enhancing portfolio management using artificial intelligence: literature review. *Frontiers in artificial intelligence*, 7, 2024.

Guojun Xiong, Ujwal Dinesha, Debajoy Mukherjee, Jian Li, and Srinivas Shakkottai. Dopl: Direct online preference learning for restless bandits with preference feedback. In *Proceedings of the International Conference on Learning Representations*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the International Conference on Machine Learning*, 2024.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the International Conference on Machine Learning*, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Proceedings of the Advances in Neural Information Processing Systems*, 2018a.

Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Dynamic regret of strongly adaptive methods. In *Proceedings of the International Conference on Machine Learning*, 2018b.

A Related Works

Preference-Based Learning and Dueling Bandits. Preference-based learning using pairwise comparisons was introduced through dueling bandits (Yue et al., 2012), with reductions to standard bandits in Ailon et al. (2014). Extensive work studies stochastic and adversarial dueling bandits in discrete action spaces (Bengs et al., 2021; Saha et al., 2021), as well as contextual extensions (Dudík et al., 2015; Saha & Krishnamurthy, 2022; Saha & Schapire, 2025). These works study finite or contextual action spaces, and the continuous-action extension of Saha & Schapire (2025) is discussed below.

Continuous Dueling under Convex Objectives. Several works extend dueling feedback to continuous domains. Yue & Joachims (2009) consider a compact convex parameter space with a fixed concave utility function and develop Dueling Bandit Gradient Descent (DBGD), obtaining an $O(T^{3/4})$-type regret guarantee measured through pairwise preference probabilities. Kumagai (2017) studies a fixed strongly convex, smooth objective, with preference-regret guarantees called 'Dueling-Bandit(DB)-regret' and a possible conversion to objective regret under link-slope bounds. Lu et al. (2022) extend the DBGD line to non-stationary Online Learning to Rank (OL2R) for time-varying utilities, deriving a path-length-dependent $O(T^{3/4})$-type dynamic preference-regret bound and a meta-learning method that avoids prior knowledge of the path length. More recent work by Saha et al. (2025) considers smooth convex objectives under general preference links in a stationary pure-exploration setting with sample complexity guarantees, while Sharma et al. (2026) studies stochastic Lipschitz dueling bandits over continuous action spaces. The continuous-action model of Saha & Schapire (2025) adopt a contextual approach and assumes linear objectives with respect to actions under stochastic feedback, using a Best-Response(BR)-Regret notation.

Table 2 summarizes these results. These works provide important precursors for continuous dueling optimization, but differ from the adversarial OCO formulation studied here. In Yue & Joachims (2009) and Lu et al. (2022), cumulative regret is measured through pairwise preference probabilities induced by a link function, whereas we measure regret directly in the underlying convex objective values at the queried points. Moreover, our DTFO framework transfers guarantees from general first-order OCO algorithms and supports fixed, changing, and interval comparators within a unified objective-regret formulation.

Table 2: Selected continuous dueling results. The metrics and comparator classes differ: C_T is the path length of per-round optimizers, while $P_T(U)$ is the path length of an arbitrary comparator sequence. Rates suppress dimension, geometry, and link constants.

Reference	Setting and benchmark	Guarantee
Yue & Joachims (2009)	Fixed strictly concave utility; preference regret against its maximizer	$\mathcal{O}(T^{3/4})$
Lu et al. (2022)	Time-varying utilities in the DBGD model; preference regret against per-round maximizers	$\tilde{\mathcal{O}}(T^{3/4}\sqrt{1+C_T})$ Unknown C_T
Kumagai (2017)	Noisy strongly convex, smooth loss; preference regret with objective-value conversion	$\mathcal{O}(\sqrt{T\log T})$
Saha et al. (2025)	Fixed smooth convex loss; optimization error under general preferences	Sample complexity depending on accuracy and transfer function
This work (convex)	Changing convex losses; objective-value regret against arbitrary comparators	$\mathcal{O}(T^{3/4}\sqrt{1+P_T(U)})$ Unknown $P_T(U)$

Bandit Online Convex Optimization. Classical one-point (Flaxman et al., 2005) and two-point (Agarwal et al., 2010; Duchi et al., 2015; Shamir, 2017) methods construct gradient estimates from function evaluations. Our convex static bound has the same $T^{3/4}$ horizon dependence as the classical one-point value-feedback baseline (Flaxman et al., 2005). Computationally efficient online Newton methods attain $\widetilde{\mathcal{O}}(\sqrt{T})$ adversarial regret with scalar-value feedback (Fokkema et al., 2024), using gradient and Hessian estimates of surrogate losses from possibly noisy observations. These

value-based estimators are not directly available from a binary comparison. Our contribution is a reduction from dueling feedback to first-order surrogate regret, giving static, dynamic, and adaptive objective-regret guarantees through standard OCO methods (Zhang et al., 2018a; Garber & Kretzu, 2022). Related reduction frameworks transfer linear or quadratic regret guarantees to broader objectives and connect first-order and value feedback (Pedramfar & Aggarwal, 2024), whereas this work constructs and controls a surrogate gradient from binary comparisons.

In this paper, we study adversarial online convex optimization with general convex objectives under purely comparison-based feedback, and provide the regret guarantees measured directly on objective function value.

B TECHNICAL LEMMAS

The three lemmas below establish curvature preservation, the weak sphere-to-ball identity, and smoothing-error bounds, respectively. Appendices C and D use these ingredients to prove the estimator and interval-transfer results.

Convexity Preservation. To build the theoretical bridge between the true objective functions and the base algorithm's surrogate losses, we rely on the properties of spherically smoothed functions. Throughout this appendix, $\mathcal{K}$ satisfies the domain assumptions, $0 < \delta \leq 1/2$, and U and V are uniform on $\mathbb{S}^{d-1}$ and $\mathbb{B}_d$, respectively. First, we establish that the smoothing operation preserves the convexity and, when present, strong convexity of the original function. This is a standard property in bandit convex optimization (Flaxman et al., 2005; Hazan, 2016).

Lemma 3 (Convexity of Smoothed Loss)**.** *Let $f : \mathcal{K} \to \mathbb{R}$ be G-Lipschitz and σ-strongly convex, with $\sigma \geq 0$. Define*

$$\widetilde{f}(w) = \mathbb{E}_V f(w + \delta V), \qquad w \in \mathcal{K}_\delta.$$

Then $\widetilde{f}$ is σ-strongly convex on $\mathcal{K}_\delta$. Moreover, for any $x, z \in \mathcal{K}_\delta$,

$$\widetilde{f}(x) - \widetilde{f}(z) \leq \langle \bar{g}(x), x - z \rangle - \frac{\sigma}{2}\|x - z\|_2^2, \qquad \text{where } \bar{g}(x) = \mathbb{E}_V \nabla f(x + \delta V), \tag{14}$$

where gradients of f are taken almost everywhere in $\operatorname{int}\mathcal{K}$*. Setting $\sigma = 0$ gives convexity.*

Proof. For $x, z \in \mathcal{K}_\delta$ and $\lambda \in [0, 1]$, both $x + \delta V$ and $z + \delta V$ are feasible, and

$$\lambda x + (1 - \lambda)z + \delta V = \lambda(x + \delta V) + (1 - \lambda)(z + \delta V).$$

Applying strong convexity inside the expectation gives

$$\begin{aligned}\widetilde{f}(\lambda x + (1 - \lambda)z) &\leq \mathbb{E}_V \left[\lambda f(x + \delta V) + (1 - \lambda) f(z + \delta V) - \frac{\sigma}{2}\lambda(1 - \lambda)\|x - z\|_2^2\right] \\ &= \lambda \widetilde{f}(x) + (1 - \lambda)\widetilde{f}(z) - \frac{\sigma}{2}\lambda(1 - \lambda)\|x - z\|_2^2.\end{aligned}$$

This proves the first claim.

For the supporting inequality, note that $x + \delta v$ lies in $\operatorname{int}\mathcal{K}$ whenever $\|v\|_2 < 1$, even when $x \in \partial\mathcal{K}_\delta$. A convex G-Lipschitz function is differentiable almost everywhere in the interior, with gradient norm at most G. Thus, almost surely,

$$f(z + \delta V) \geq f(x + \delta V) + \langle \nabla f(x + \delta V), z - x \rangle + \frac{\sigma}{2}\|x - z\|_2^2.$$

Taking expectations and rearranging proves Eq. (14). □

Gradient Extraction Identity. Next, we establish the gradient extraction identity. We also state its weak sphere-to-ball form, which will be applied to the composed comparison function in the proof of Lemma 1. This is the integration identity underlying spherical smoothing (Flaxman et al., 2005).

Lemma 4 (Gradient Extraction via Spherical Symmetry). *For any Lipschitz scalar function h on the closed unit ball,*

$$d\mathbb{E}_U[h(U)U] = \mathbb{E}_V \nabla h(V), \tag{15}$$

where the right side uses the weak gradient. In particular, for a convex G-Lipschitz function $f : \mathcal{K} \to \mathbb{R}$ and $w \in \mathcal{K}_\delta$, set $\Delta(U) = f(w + \delta U) - f(w - \delta U)$. Then

$$\frac{d}{2\delta}\mathbb{E}_U[\Delta(U)U] = \bar{g}(w) := \mathbb{E}_V \nabla f(w + \delta V). \tag{16}$$

For $w \in \operatorname{int} \mathcal{K}_\delta$, this vector equals $\nabla \widetilde{f}(w)$. Equation Eq. (16) also holds for boundary centers.

Proof. For a continuously differentiable h, apply the divergence theorem to he_i for each coordinate i:

$$\int_{\mathbb{B}_d} \partial_i h(v)\, dv = \int_{\mathbb{S}^{d-1}} h(u) u_i\, dS(u).$$

Since the surface area of $\mathbb{S}^{d-1}$ is d times the volume of $\mathbb{B}_d$, normalization gives Eq. (15). For Lipschitz h, extend h to a Lipschitz function on $\mathbb{R}^d$ and mollify. The mollifications converge uniformly on the closed ball, and their gradients converge in $L^1(\mathbb{B}_d)$ to its weak gradient. Passing to the limit proves the same identity. This auxiliary extension of h does not require the original loss f to be defined outside $\mathcal{K}$.

For the second claim, expanding Δ gives

$$\frac{d}{2\delta}\mathbb{E}_U[\Delta(U)U] = \frac{d}{2\delta}\left(\mathbb{E}_U[f(w + \delta U)U] - \mathbb{E}_U[f(w - \delta U)U]\right). \tag{17}$$

By spherical symmetry, U and $-U$ have the same distribution, so

$$\mathbb{E}_U[f(w - \delta U)U] = -\mathbb{E}_U[f(w + \delta U)U].$$

Substituting into Eq. (17) and applying Eq. (15) to $h(v) = f(w + \delta v)$ yields

$$\frac{d}{2\delta}\mathbb{E}_U[\Delta(U)U] = \frac{d}{\delta}\mathbb{E}_U[f(w + \delta U)U] = \mathbb{E}_V \nabla f(w + \delta V) = \bar{g}(w).$$

At an interior center, differentiation of the ball average is justified by dominated difference quotients: f is differentiable at $w + \delta V$ almost surely, and its difference quotients in each fixed direction are bounded by G times the direction's norm. Thus $\bar{g}(w) = \nabla \widetilde{f}(w)$ there. At boundary centers, the averaged supporting inequality Eq. (14) provides the property needed for regret transfer. □

Controlled Smoothing Error. Finally, we bound the approximation error introduced by the smoothing operation and the difference between the expected query loss and the ball average. Because the original function is G-Lipschitz, the smoothed function over a δ-radius ball is guaranteed to be point-wise close to the true function, allowing us to bound the bias introduced by our wrapper.

Lemma 5 (Approximation Error of Smoothed Loss). *Let $f : \mathcal{K} \to \mathbb{R}$ be G-Lipschitz and let $\widetilde{f}(w) = \mathbb{E}_V f(w + \delta V)$ on $\mathcal{K}_\delta$. For every $w \in \mathcal{K}_\delta$,*

$$|\widetilde{f}(w) - f(w)| \le \frac{d}{d+1} G\delta \le G\delta. \tag{18}$$

Writing $Q(w) = \mathbb{E}_U[f(w + \delta U) + f(w - \delta U)]/2$, we also have

$$Q(w) - \widetilde{f}(w) \le \frac{G\delta}{d+1}. \tag{19}$$

Proof. By the triangle inequality and Lipschitzness,

$$\begin{aligned} |\widetilde{f}(w) - f(w)| &\le \mathbb{E}_V |f(w + \delta V) - f(w)| \\ &\le G\delta\, \mathbb{E}\|V\|_2 = \frac{d}{d+1} G\delta. \end{aligned}$$

For the second claim, represent a uniform ball point as $V = RU$, where R and U are independent and R has density dr^{d-1} on $[0, 1]$. Hence $\mathbb{E}R = d/(d + 1)$. By spherical symmetry, $Q(w) = \mathbb{E}_U f(w + \delta U)$, and coupling the two expectations gives

$$\begin{aligned} Q(w) - \widetilde{f}(w) &= \mathbb{E}_{R,U}[f(w + \delta U) - f(w + \delta RU)] \\ &\le G\delta\, \mathbb{E}(1 - R) = \frac{G\delta}{d+1}. \end{aligned}$$

All points in these expectations lie in $\mathcal{K}$. □

C Proof of Lemma 1

We apply the weak sphere-to-ball identity from Appendix B to $\rho \circ \Delta$, cancel A^{-1}, and isolate the nonlinear bias. Specializing to $A = I_d$ then gives the spherical estimator norm and linear bias bound.

Proof. Fix x and A as in the lemma. When they are chosen using the past, condition on the information available before sampling the current direction throughout this argument. Define

$$\begin{aligned} \Delta(v) &= f(x + \delta Av) - f(x - \delta Av), \\ m(v) &= \frac{\nabla f(x + \delta Av) + \nabla f(x - \delta Av)}{2}, \qquad \bar{g} = \mathbb{E}_V m(V). \end{aligned}$$

Feasibility ensures that $x \pm \delta Av$ lie in $\operatorname{int}\mathcal{K}$ for $\|v\|_2 < 1$. Since A is invertible, the affine maps preserve null sets, so both gradients exist almost everywhere and $\|m(v)\|_2 \le G$. Lipschitzness also gives

$$|\Delta(v)| \le 2G\delta\|Av\|_2 \le 2G,$$

using $\delta\|A\|_{\mathrm{op}} \le 1$.

By dueling oracle definition and the tower property,

$$\mathbb{E}\widehat{g} = -\frac{d}{2\gamma\delta} A^{-1}\mathbb{E}_U[\rho(\Delta(U))U].$$

The function $h(v) = \rho(\Delta(v))$ is Lipschitz on the closed unit ball. The weak chain rule and symmetry $A^\top = A$ give, almost everywhere,

$$\nabla\Delta(v) = 2\delta Am(v), \qquad \nabla h(v) = 2\delta\rho'(\Delta(v))Am(v).$$

Applying Eq. (15) from Lemma 4, we obtain

$$\begin{aligned} \mathbb{E}\widehat{g} &= -\frac{A^{-1}}{2\gamma\delta}\mathbb{E}_V\nabla h(V) \\ &= -\frac{1}{\gamma}\mathbb{E}_V[\rho'(\Delta(V))m(V)] \\ &= \bar{g} - \frac{1}{\gamma}\mathbb{E}_V[\{\rho'(\Delta(V)) + \gamma\}m(V)]. \end{aligned}$$

This is the claimed identity with $B = -\gamma^{-1}\mathbb{E}_V[\{\rho'(\Delta(V)) + \gamma\}m(V)]$. The inverse matrix cancels before any norm bound, and no derivative of f at a sphere query point is required.

For the spherical specialization $A = I_d$, since $o \in \{-1, 1\}$ and $\|U\|_2 = 1$, the estimator has exact magnitude

$$\|\widehat{g}\|_2 = \left\| -\frac{d}{2\gamma\delta} oU \right\|_2 = \frac{d}{2\gamma\delta}.$$

Finally, Assumption 1 implies $|\rho'(z) + \gamma| \le M|z|$ on $[-2G, 2G]$. Therefore,

$$\begin{aligned} \|B\|_2 &\le \frac{1}{\gamma}\mathbb{E}_V[|\rho'(\Delta(V)) + \gamma|\,\|m(V)\|_2] \\ &\le \frac{MG}{\gamma}\mathbb{E}_V|\Delta(V)| \le \frac{2MG^2\delta}{\gamma}\mathbb{E}\|V\|_2 = \frac{2d}{d+1}\frac{MG^2}{\gamma}\delta. \end{aligned}$$

For $A = I_d$, symmetry of V also gives $\bar{g} = \mathbb{E}_V\nabla f(x + \delta V)$, as used in the transfer theorem. □

D Proof of Theorem 1

We first project the comparator path into $\mathcal{K}_\delta$ without increasing its path length. Curvature preservation and Lemma 1 then transfer surrogate regret; the final step adds smoothing and projection errors to obtain $C_{\mathrm{sph}}\delta|I|$.

Proof. Fix an interval $I = [a, b]$ and a comparator sequence $(w_t^*)_{t\in I}$ independently of the learner's randomization. Set $v_t = \Pi_{\mathcal{K}_\delta}(w_t^*)$ and write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. The loss sequence is fixed, w_t is measurable with respect to $\mathcal{F}_{t-1}$, and the comparators and their projections are fixed. By Lemma 1,

$$\mathbb{E}_t\widehat{g}_t = \bar{g}_t + B_t, \qquad \bar{g}_t = \mathbb{E}_V \nabla f_t(w_t + \delta V), \qquad \|B_t\|_2 \le \frac{2d}{d+1}\frac{MG^2}{\gamma}\delta.$$

Lemma 3 gives the supporting inequality with curvature. Since $w_t, v_t \in \mathcal{K}_\delta \subseteq \mathcal{K}$, we have $\|w_t - v_t\|_2 \le D$, and hence

$$\begin{aligned}\widetilde{f}_t(w_t) - \widetilde{f}_t(v_t) &\le \langle \bar{g}_t, w_t - v_t\rangle - \frac{\sigma}{2}\|w_t - v_t\|_2^2 \\ &= \mathbb{E}_t\Big[\langle \widehat{g}_t, w_t - v_t\rangle - \frac{\sigma}{2}\|w_t - v_t\|_2^2\Big] - \langle B_t, w_t - v_t\rangle \\ &\le \mathbb{E}_t[q_t(w_t) - q_t(v_t)] + \frac{2d}{d+1}\frac{MG^2D}{\gamma}\delta.\end{aligned}$$

Taking total expectations and summing gives

$$\mathbb{E}\sum_{t\in I}[\widetilde{f}_t(w_t) - \widetilde{f}_t(v_t)] \le \mathbb{E}\sum_{t\in I}[q_t(w_t) - q_t(v_t)] + \frac{2d}{d+1}\frac{MG^2D}{\gamma}\delta|I|. \tag{20}$$

To bound the regret on the queried points, define $Q_t(w) = \mathbb{E}_U[f_t(w + \delta U) + f_t(w - \delta U)]/2$. Since w_t is chosen before U_t is drawn,

$$\mathbb{E}_t\left[\frac{f_t(w_t^+) + f_t(w_t^-)}{2}\right] = Q_t(w_t).$$

Adding and subtracting the smoothed losses and the loss at the projected comparator gives

$$\begin{aligned}Q_t(w_t) - f_t(w_t^*) = \big(Q_t(w_t) - \widetilde{f}_t(w_t)\big) + \big(\widetilde{f}_t(w_t) - \widetilde{f}_t(v_t)\big) \\ + \big(\widetilde{f}_t(v_t) - f_t(v_t)\big) + \big(f_t(v_t) - f_t(w_t^*)\big).\end{aligned}$$

By Lemma 5, the first and third terms together are at most $G\delta/(d+1) + dG\delta/(d+1) = G\delta$.

For the fourth term, the domain assumptions imply $(1-\delta)w_t^* \in \mathcal{K}_\delta$: for every $z \in \mathbb{B}_d \subseteq \mathcal{K}$, convexity gives $(1-\delta)w_t^* + \delta z \in \mathcal{K}$. Since v_t is the nearest point in $\mathcal{K}_\delta$,

$$\|v_t - w_t^*\|_2 \le \|(1-\delta)w_t^* - w_t^*\|_2 = \delta\|w_t^*\|_2 \le D\delta.$$

Here $\|w_t^*\|_2 \le D$ follows from $0, w_t^* \in \mathcal{K}$. Thus Lipschitzness bounds the fourth term by $GD\delta$.

Taking expectations, summing over I, and substituting Eq. (20) proves

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in I}\left(\frac{f_t(w_t^+) + f_t(w_t^-)}{2} - f_t(w_t^*)\right)\right] \\ &\qquad \le \mathbb{E}\sum_{t\in I}[q_t(w_t) - q_t(v_t)] + \left(G(D+1) + \frac{2d}{d+1}\frac{MG^2D}{\gamma}\right)\delta|I|.\end{aligned}$$

This is the stated bound with C_{sph}. Finally, Euclidean projection onto the closed convex set $\mathcal{K}_\delta$ is nonexpansive, so

$$\sum_{t=a+1}^{b}\|v_t - v_{t-1}\|_2 \le \sum_{t=a+1}^{b}\|w_t^* - w_{t-1}^*\|_2.$$

This also proves the comparator-path assertion.

If the base algorithm has the pathwise guarantee $\sum_{t\in I}[q_t(w_t) - q_t(v_t)] \le R_{\mathcal{A}}(v_{a:b}, I)$, its expectation can be substituted into the preceding bound. This requires the guarantee on the realized surrogate sequence, even though q_t depends on current feedback. The argument applies for every $\sigma \ge 0$. □

E ALGORITHMS AND PROOFS FOR THE DTFO CONSEQUENCES

Sections E.1 and E.2 give ORGD and Dynamic ORGD, together with their static, interval, and path-length bounds. Section E.3 proves the SC-OGD curvature cancellation. Section E.4 specifies IFLH and combines exp-concavity, expert lifetimes, and interval covering. Each analysis ends by applying Theorem 1.

E.1 ORGD: IMPLEMENTATION AND STATIC/ADAPTIVE REGRET

We state the projected update and then telescope its distance potential on an arbitrary interval to prove Theorem 2.

E.1.1 IMPLEMENTATION

Algorithm 3 implements DTFO with projected OGD; the proof immediately below covers the convex step-size schedule.

Algorithm 3 Online Relative Gradient Descent (ORGD)

Require: Convex compact domain $\mathcal{K} \subset \mathbb{R}^d$, horizon T, perturbation radius $0 < \delta \leq 1/2$.
Require: Known slope magnitude $\gamma = -\rho'(0) > 0$, positive step sizes $\{\eta_t\}_{t=1}^T$.
Define $\mathcal{K}_\delta = \{w \in \mathcal{K} : w + \delta \mathbb{B}_d \subseteq \mathcal{K}\}$.
Initialize $w_1 \in \mathcal{K}_\delta$ arbitrarily.
for $t = 1, \dots, T$ **do**
  Sample $u_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$.
  Set $w_t^+ = w_t + \delta u_t$ and $w_t^- = w_t - \delta u_t$.
  Query the dueling oracle with (w_t^+, w_t^-) and receive $o_t \in \{-1, +1\}$.
  Construct $\hat{g}_t = -\dfrac{d}{2\gamma\delta} o_t u_t$.
  Update $w_{t+1} = \Pi_{\mathcal{K}_\delta}(w_t - \eta_t \hat{g}_t)$.
end for

E.1.2 PROOF OF THEOREM 2

Proof. Fix $I = [s, e]$ and $v \in \mathcal{K}_\delta$. Write $a_t = \|w_t - v\|_2^2$. By the non-expansiveness of Euclidean projection, since $v \in K_\delta$,

$$\begin{aligned} a_{t+1} &= \|w_{t+1} - v\|_2^2 = \|\Pi_{K_\delta}(w_t - \eta_t \hat{g}_t) - v\|_2^2 \leq \|w_t - \eta_t \hat{g}_t - v\|_2^2 \\ &= \|w_t - v\|_2^2 - 2\eta_t \langle \hat{g}_t, w_t - v \rangle + \eta_t^2 \|\hat{g}_t\|_2^2 \\ &= a_t - 2\eta_t \langle \hat{g}_t, w_t - v \rangle + \eta_t^2 \|\hat{g}_t\|_2^2. \end{aligned}$$

Rearranging and dividing by $2\eta_t > 0$ gives

$$\langle \hat{g}_t, w_t - v \rangle \leq \frac{a_t - a_{t+1}}{2\eta_t} + \frac{\eta_t}{2} \|\hat{g}_t\|_2^2.$$

Since η_t is non-increasing and $a_t \leq D^2$,

$$\sum_{t=s}^{e} \frac{a_t - a_{t+1}}{2\eta_t} = \frac{a_s}{2\eta_s} + \sum_{t=s+1}^{e} a_t \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) - \frac{a_{e+1}}{2\eta_e} \leq \frac{D^2}{2\eta_e}.$$

Using $\|\hat{g}_t\|_2 \leq G_{\mathrm{surr}}$, we obtain the pathwise interval bound

$$\sum_{t=s}^{e} \langle \hat{g}_t, w_t - v \rangle \leq \frac{D^2}{2\eta_e} + \frac{G_{\mathrm{surr}}^2}{2} \sum_{t=s}^{e} \eta_t = \frac{D G_{\mathrm{surr}}}{2} \left(\sqrt{e} + \sum_{t=s}^{e} \frac{1}{\sqrt{t}} \right) \leq \frac{3}{2} D G_{\mathrm{surr}} \sqrt{T},$$

where the equality results from substituting $\eta_t = \frac{D}{G_{\mathrm{surr}}\sqrt{t}}$ and, then we used $\sum_{t=1}^{e} t^{-1/2} \leq 2\sqrt{e}$ and $e \leq T$. The algorithm is not restarted at s so the starting point w_s is the iterate of the continuing run.

For the fixed loss sequence, choose $w_I^* \in \arg\min_{w\in\mathcal{K}} \sum_{t\in I} f_t(w)$ and set $v = \Pi_{\mathcal{K}_\delta}(w_I^*)$. Applying Theorem 1 to the preceding surrogate bound and substituting $G_{\mathrm{surr}} = d/(2\gamma\delta)$ gives the claimed expected query-regret inequality. The tuning $\delta = cT^{-1/4}$ yields the adaptive rate, and taking $I = [1, T]$ gives the static rate for the same algorithm and parameters. □

E.2 Dynamic ORGD: implementation and dynamic regret

The geometric step-size grid and aggregation rule below define Dynamic ORGD. The proof of Theorem 3 combines an expert path-length bound with the aggregation penalty before applying DTFO.

E.2.1 Implementation

We describe the implementation of Dynamic ORGD before proving its regret bound. Following Zhang et al. (2018a), set $N = \left\lceil \frac{1}{2}\log_2\left(1 + \frac{4T}{7}\right)\right\rceil + 1$, $\eta_i = \frac{2^{i-1}D}{G_{\mathrm{surr}}}\sqrt{\frac{7}{2T}}$, for $i \in [N]$, and use the meta learning rate and initial weights $\lambda = \sqrt{\frac{2}{TG_{\mathrm{surr}}^2 D^2}}$, $p_{1,i} = \frac{N+1}{N\,i(i+1)}$, which sum to one. Centering the shared linear surrogate at w_t changes neither its regret nor the normalized exponential-weight update.

Algorithm 4 Dynamic ORGD (via Improved Ader)

Require: Convex compact domain $\mathcal{K} \subset \mathbb{R}^d$ containing $\mathbb{B}_d$, known diameter bound D, Horizon T, radius $0 < \delta \le 1/2$, known slope magnitude $\gamma = -\rho'(0) > 0$.

Define $\mathcal{K}_\delta = \{w \in \mathcal{K} : w + \delta\mathbb{B}_d \subseteq \mathcal{K}\}$.
Choose $N, \{\eta_i\}_{i=1}^N, \lambda, \{p_{1,i}\}_{i=1}^N$ as stated above.
Initialize $w_{1,i} = 0$ for $i = 1, \dots, N$.
for $t = 1, \dots, T$ **do**
  Form the mixture $w_t = \sum_{i=1}^N p_{t,i} w_{t,i}$.
  Sample $u_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$; set $w_t^{\pm} = w_t \pm \delta u_t$.
  Query (w_t^+, w_t^-), receive $o_t \in \{-1, +1\}$, and set $\hat{g}_t = -\dfrac{d}{2\gamma\delta} o_t u_t$.
  Compute $z_{t,i} = \langle \hat{g}_t, w_{t,i} - w_t\rangle$ for $i = 1, \dots, N$.
  for $i = 1, \dots, N$ **do**
    Update $p_{t+1,i} = \dfrac{p_{t,i}\exp(-\lambda z_{t,i})}{\sum_{j=1}^N p_{t,j}\exp(-\lambda z_{t,j})}$.
    Update $w_{t+1,i} = \Pi_{\mathcal{K}_\delta}(w_{t,i} - \eta_i \hat{g}_t)$.
  end for
end for

Improved Ader adapts to unknown path length by aggregating experts with different step sizes. All experts share the surrogate generated by one duel at a common radius. This construction does not simultaneously provide the feedback that would arise from querying other exploration radii. Changing the radius changes the queried pair and the distribution of its comparison bit. We therefore use the path-length-independent choice $\delta = cT^{-1/4}$ in this instantiation.

E.2.2 Proof of Theorem 3

Proof. Fix W^* and set $v_t = \Pi_{\mathcal{K}_\delta}(w_t^*)$. The non-expansive property of Euclidean projection implies that, for every $t = 1, \dots, T-1$,

$$\|v_{t+1} - v_t\|_2 = \left\|\Pi_{\mathcal{K}_\delta}(w_{t+1}^*) - \Pi_{\mathcal{K}_\delta}(w_t^*)\right\|_2 \le \|w_{t+1}^* - w_t^*\|_2.$$

Therefore, summing both sides,

$$P_T(v_{1:T}) = \sum_{t=1}^{T-1} \|v_{t+1} - v_t\|_2 \le \sum_{t=1}^{T-1} \|w_{t+1}^* - w_t^*\|_2 = P_T(W^*) = P.$$

The domain $\mathcal{K}_\delta$ is compact and convex, contains zero, and has diameter at most D. Thus the experts and their mixture remain feasible, and both queries belong to $\mathcal{K}$. For every realization of the

feedback, the surrogate losses are linear with $\|\nabla q_t(w)\|_2 = G_{\mathrm{surr}}$. The following expert and meta bounds establish the surrogate guarantee pathwise, as in the proof of Theorem 4 of Zhang et al. (2018a).

For each expert, non-expansiveness of Euclidean projection gives, for every t,

$$\|w_{t+1,i} - v_t\|_2^2 \le \|w_{t,i} - \eta_i \hat{g}_t - v_t\|_2^2 = \|w_{t,i} - v_t\|_2^2 - 2\eta_i \langle \hat{g}_t, w_{t,i} - v_t\rangle + \eta_i^2 \|\hat{g}_t\|_2^2.$$

Hence,

$$\langle \hat{g}_t, w_{t,i} - v_t\rangle \le \frac{\|w_{t,i} - v_t\|_2^2 - \|w_{t+1,i} - v_t\|_2^2}{2\eta_i} + \frac{\eta_i}{2}\|\hat{g}_t\|_2^2.$$

Summing over $t = 1, \dots, T$, the potential terms satisfy

$$\sum_{t=1}^{T} \left(\|w_{t,i} - v_t\|_2^2 - \|w_{t+1,i} - v_t\|_2^2\right)$$

$$= \|w_{1,i} - v_1\|_2^2 - \|w_{T+1,i} - v_T\|_2^2 + \sum_{t=2}^{T} \left(\|w_{t,i} - v_t\|_2^2 - \|w_{t,i} - v_{t-1}\|_2^2\right).$$

For each $t \ge 2$,

$$\|w_{t,i} - v_t\|_2^2 - \|w_{t,i} - v_{t-1}\|_2^2 = \langle (w_{t,i} - v_t) + (w_{t,i} - v_{t-1}), v_{t-1} - v_t\rangle \le 2D\|v_t - v_{t-1}\|_2,$$

since all three points lie in $\mathcal{K}_\delta$, whose diameter is at most D. Moreover, $\|w_{1,i} - v_1\|_2^2 \le D^2$, while the terminal potential is non-positive. Therefore, using $P_T(v_{1:T}) \le P$ and $\|\hat{g}_t\|_2 \le G_{\mathrm{surr}}$,

$$\sum_{t=1}^{T} \langle \hat{g}_t, w_{t,i} - v_t\rangle \le \frac{D^2 + 2DP}{2\eta_i} + \frac{\eta_i G_{\mathrm{surr}}^2 T}{2} \le \frac{7D^2 + 4DP}{4\eta_i} + \frac{\eta_i G_{\mathrm{surr}}^2 T}{2}. \tag{21}$$

For the meta algorithm, recall that $z_{t,i} = \langle \hat{g}_t, w_{t,i} - w_t\rangle$ and $|z_{t,i}| \le \|\hat{g}_t\|_2 \|w_{t,i} - w_t\|_2 \le DG_{\mathrm{surr}}$ because both $w_{t,i}$ and w_t lie in $\mathcal{K}_\delta$ with diameter D and $\|\hat{g}_t\|_2 \le G_{\mathrm{surr}}$. Moreover, because $w_t = \sum_i p_{t,i} w_{t,i}$ and $\sum_i p_{t,i} = 1$, $\sum_i p_{t,i} z_{t,i} = \langle \hat{g}_t, \sum_i p_{t,i}(w_{t,i} - w_t)\rangle = 0$. Thus, regarding $z_{t,i}$ as a random variable taking value $z_{t,i}$ with probability $p_{t,i}$, it has mean zero and lies in $[-DG_{\mathrm{surr}}, DG_{\mathrm{surr}}]$. Hence, by Hoeffding's lemma,

$$\log\left(\sum_i p_{t,i} e^{-\lambda z_{t,i}}\right) \le \frac{\lambda^2 D^2 G_{\mathrm{surr}}^2}{2}.$$

Define the normalization factor $Z_t := \sum_j p_{t,j} \exp(-\lambda z_{t,j})$. Given the exponential-weights update $p_{t+1,i} = \frac{p_{t,i}\exp(-\lambda z_{t,i})}{Z_t}$, taking the log on both sides, rearranging and summing over t yields

$$-\lambda \sum_{t=1}^{T} z_{t,i} = \log p_{T+1,i} - \log p_{1,i} + \sum_{t=1}^{T} \log Z_t.$$

Since $p_{T+1,i} \le 1$, we have $\log p_{T+1,i} \le 0$, and therefore

$$-\lambda \sum_{t=1}^{T} z_{t,i} \le \ln \frac{1}{p_{1,i}} + \sum_{t=1}^{T} \log Z_t \le \ln \frac{1}{p_{1,i}} + \frac{\lambda^2 D^2 G_{\mathrm{surr}}^2 T}{2}.$$

where the second inequality results from replacing Z_t and applying Hoeffding's lemma. Dividing by λ and using $-z_{t,i} = \langle \hat{g}_t, w_t - w_{t,i}\rangle$, we obtain

$$\sum_{t=1}^{T} \langle \hat{g}_t, w_t - w_{t,i}\rangle \le \frac{\ln(1/p_{1,i})}{\lambda} + \frac{\lambda D^2 G_{\mathrm{surr}}^2 T}{2} \le \frac{DG_{\mathrm{surr}}\sqrt{2T}}{2}\big[1 + 2\ln(i+1)\big], \tag{22}$$

where $\ln(1/p_{1,i}) \le 2\ln(i+1)$ and substituting $\lambda = \sqrt{\frac{2}{TG_{\mathrm{surr}}^2 D^2}}$. Now define the index $k = \left\lfloor \frac{1}{2}\log_2\left(1 + \frac{4P}{7D}\right)\right\rfloor + 1$, and the number of experts $N = \left\lceil \frac{1}{2}\log_2\left(1 + \frac{4T}{7}\right)\right\rceil + 1$. Since $P \le D(T-1)$, the index k lies in $\{1, \dots, N\}$, and the grid satisfies

$$\eta_k \le \eta^* < 2\eta_k, \qquad \eta^* = \sqrt{\frac{7D^2 + 4DP}{2TG_{\mathrm{surr}}^2}},$$

because $k-1 \le \frac{1}{2}\log_2\left(1+\frac{4P}{7D}\right) < k$ from the definition of k. Substituting $i=k$ in Eq. (21) gives an expert regret of at most $\frac{3G_{\mathrm{surr}}}{4}\sqrt{2T(7D^2+4DP)}$. Adding Eq. (22) bounds $\sum_{t=1}^{T}\langle \hat{g}_t, w_t - v_t\rangle$ by the first two terms in the bound below.

These potential inequalities hold for each realized surrogate sequence, even though it depends on the learner's predictions and observations. In particular, the base argument does not require $\hat{g}_t$ to be an unbiased gradient of the original loss. Taking expectations and applying Theorem 1 adds $C_{\mathrm{sph}}\delta T$, which accounts for estimator bias, smoothing, and comparator projection. Thus the exact query-regret bound is

$$\begin{aligned}\mathrm{Regret}_D(W^*, T) \le{}& \frac{3G_{\mathrm{surr}}}{4}\sqrt{2T(7D^2+4DP)} \\ &+ \frac{DG_{\mathrm{surr}}\sqrt{2T}}{2}\left[1+2\ln(k+1)\right] + C_{\mathrm{sph}}\delta T. \end{aligned} \tag{23}$$

To obtain the order stated in Theorem 3, set $x = P/D \ge 0$. Using $\ln(1+y) \le y$,

$$k+1 \le 2 + \frac{\ln(1+4x/7)}{2\ln 2} \le 2 + \frac{2x}{7\ln 2} \le 2 + x.$$

Since $\ln y \le \sqrt{y}$ for $y \ge 1$, $1+2\ln(k+1) \le (1+2\sqrt{2})\sqrt{1+x}$. Also $7D^2+4DP \le 7D^2(1+x)$, so both first-order terms in Eq. (23) are $\mathcal{O}(DG_{\mathrm{surr}}\sqrt{T(1+P/D)})$. Substituting $G_{\mathrm{surr}} = d/(2\gamma\delta)$ and the single radius $\delta = cT^{-1/4}$, $0 < c \le 1/2$, completes the proof. □

E.3 SC-OGD: PROOF OF THEOREM 4

Algorithm 3 with $\eta_t = 1/(\sigma t)$ implements SC-OGD. The quadratic surrogate cancels the increasing inverse step size, leaving a harmonic regret bound to which Theorem 1 applies.

Proof. We reproduce the strongly convex OGD argument underlying (Hazan et al., 2007, Theorem 1). Fix $v \in \mathcal{K}_\delta$, write $a_t = \|w_t - v\|_2^2$, and set $\eta_t = 1/(\sigma t)$. The projection inequality gives

$$\begin{aligned} q_t(w_t) - q_t(v) &= \langle \hat{g}_t, w_t - v\rangle - \frac{\sigma}{2}a_t \\ &\le \frac{a_t - a_{t+1}}{2\eta_t} - \frac{\sigma}{2}a_t + \frac{\eta_t}{2}\|\hat{g}_t\|_2^2 \\ &= \frac{\sigma}{2}\big((t-1)a_t - t a_{t+1}\big) + \frac{\|\hat{g}_t\|_2^2}{2\sigma t}. \end{aligned}$$

The distance terms telescope to $-\sigma T a_{T+1}/2 \le 0$. Using $\|\hat{g}_t\|_2 \le d/(2\gamma\delta)$ yields the pathwise bound

$$\sum_{t=1}^{T}\big(q_t(w_t) - q_t(v)\big) \le \frac{G_{\mathrm{surr}}^2}{2\sigma}\mathsf{H}_T \le \frac{G_{\mathrm{surr}}^2}{2\sigma}(1+\log T), \qquad \mathsf{H}_T = \sum_{t=1}^{T}\frac{1}{t}.$$

The surrogates also satisfy the global gradient bound $\|\nabla q_t(w)\|_2 \le d/(2\gamma\delta) + \sigma D$. The sharper bound above uses only gradients at the played points and remains valid when the surrogates depend on the learner's observations.

Choose a fixed $w^* \in \arg\min_{w\in\mathcal{K}}\sum_{t=1}^{T} f_t(w)$. Apply Theorem 1 with $I = [1,T]$ and $v = \Pi_{\mathcal{K}_\delta}(w^*)$ to obtain

$$\mathrm{Regret}_S(T) \le \frac{G_{\mathrm{surr}}^2}{2\sigma}\mathsf{H}_T + C_{\mathrm{sph}}\delta T.$$

With $\ell_T = \log(eT)$, choose $\delta = c(\ell_T/T)^{1/3}$, $0 < c \le 1/2$. Since $\ell_T \le T$ and $\mathsf{H}_T \le \ell_T$, this is a feasible radius and both terms are $\mathcal{O}(T^{2/3}\ell_T^{1/3})$ for fixed problem parameters. □

E.4 IFLH WITH SC-OGD EXPERTS: IMPLEMENTATION AND ADAPTIVE REGRET

This subsection specifies the implementation underlying Theorem 5. It uses the binary specialization of IFLH (Zhang et al., 2018b, Algorithm 1 and Lemma 1) on the shared quadratic surrogates q_t, not on the original losses. We first establish exp-concavity and specify expert lifetimes, then give the implementation and its pathwise interval bound. The final step transfers this bound to query regret and tunes the radius.

E.4.1 SURROGATE PROPERTIES AND EXPERT LIFETIMES

Recall $H_\delta = d/(2\gamma\delta) + \sigma D$, and define $\alpha_\delta = \sigma/H_\delta^2$. For every $w \in \mathcal{K}_\delta$,

$$\|\nabla q_t(w)\|_2 \le \|\hat{g}_t\|_2 + \sigma\|w - w_t\|_2 \le H_\delta.$$

Moreover,

$$\alpha_\delta \nabla q_t(w)\nabla q_t(w)^\top \preceq \alpha_\delta H_\delta^2 I_d = \sigma I_d = \nabla^2 q_t(w).$$

It follows by differentiating $\exp(-\alpha_\delta q_t(w))$ that this function is concave. Hence q_t is α_δ-exp-concave, as in (Zhang et al., 2018b, Lemma 2). This property concerns the explicitly known quadratic and it does not assume smoothness of the original f_t.

For an integer $j \ge 1$, define

$$\nu_2(j) = \max\{r \ge 0 : 2^r \text{ divides } j\}, \qquad e_j = j + 2^{\nu_2(j)}. \tag{24}$$

Expert j starts at round j and is active on $[j, e_j - 1]$, truncated at T. This is the base-2 ending-time rule of IFLH. At a given time, there is at most one active expert at each lifetime scale, so

$$|\mathcal{S}_t| \le \lfloor \log_2 t \rfloor + 1.$$

Each expert starts at zero, which belongs to $\mathcal{K}_\delta$ under the unit-ball assumption and $\delta < 1$.

E.4.2 COMPLETE IMPLEMENTATION

The complete implementation is presented in Algorithm 5. The meta algorithm uses learning rate α_δ and combines experts started at different times. An expert starting at j uses local step size $1/[\sigma(t-j+1)]$ and the update $w_{t+1,j} = \Pi_{\mathcal{K}_\delta}\left(w_{t,j} - \frac{\hat{g}_t + \sigma(w_{t,j} - w_t)}{\sigma(t-j+1)}\right)$. Write $p_{t,j}$ for the weights used to form the round-t mixture. The posterior weights $\bar{p}_{t+1,j}$ are normalized after observing round t's surrogate, before any experts are removed at round $t+1$. If no older expert survives, the new expert receives weight one. Otherwise it receives weight $1/t$, with the remaining mass assigned to survivors. This convention includes $t = 1$ and prevents normalization of an empty set.

Algorithm 5 Spherical DTFO with IFLH and SC-OGD Experts

Require: Convex compact $\mathcal{K} \subset \mathbb{R}^d$ containing $\mathbb{B}_d$, diameter bound D, horizon T.
Require: Radius $0 < \delta \le 1/2$, known $\sigma > 0$ and $\gamma = -\rho'(0) > 0$.
Set $\mathcal{K}_\delta = \{w \in \mathcal{K} : w + \delta\mathbb{B}_d \subseteq \mathcal{K}\}$.
Set $H_\delta = d/(2\gamma\delta) + \sigma D$, $\alpha_\delta = \sigma/H_\delta^2$, and $\mathcal{S}_0 = \varnothing$.
for $t = 1, \dots, T$ **do**
  Keep survivors $\mathcal{A}_t = \{j \in \mathcal{S}_{t-1} : t < e_j\}$; discard other experts.
  Start expert t with $w_{t,t} = 0$ and $e_t = t + 2^{\nu_2(t)}$.
  if $\mathcal{A}_t = \varnothing$ **then**
    Set $p_{t,t} = 1$.
  else
    Set $Z_t^{\mathrm{surv}} = \sum_{j \in \mathcal{A}_t} \bar{p}_{t,j}$.
    Set $p_{t,j} = (1 - 1/t)\bar{p}_{t,j}/Z_t^{\mathrm{surv}}$ for $j \in \mathcal{A}_t$, and $p_{t,t} = 1/t$.
  end if
  Set $\mathcal{S}_t = \mathcal{A}_t \cup \{t\}$ and $w_t = \sum_{j \in \mathcal{S}_t} p_{t,j} w_{t,j}$.
  Sample $u_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and set $w_t^{\pm} = w_t \pm \delta u_t$.
  Query (w_t^+, w_t^-), receive $o_t \in \{-1, +1\}$, and set $\hat{g}_t = -d\, o_t u_t/(2\gamma\delta)$.
  For every $j \in \mathcal{S}_t$, compute $z_{t,j} = \langle \hat{g}_t, w_{t,j} - w_t\rangle + \frac{\sigma}{2}\|w_{t,j} - w_t\|_2^2$.
  Set $Z_t = \sum_{j \in \mathcal{S}_t} p_{t,j} \exp(-\alpha_\delta z_{t,j})$.
  for $j \in \mathcal{S}_t$ **do**
    Set $\bar{p}_{t+1,j} = p_{t,j} \exp(-\alpha_\delta z_{t,j})/Z_t$.
    Set $g_{t,j} = \hat{g}_t + \sigma(w_{t,j} - w_t)$.
    Set $w_{t+1,j} = \Pi_{\mathcal{K}_\delta}(w_{t,j} - g_{t,j}/[\sigma(t-j+1)])$.
  end for
end for

The scores satisfy $z_{t,j} = q_t(w_{t,j}) - q_t(w_t)$, so the centered update is identical to exponential weighting by $q_t(w_{t,j})$. All scores and expert gradients are computed from one duel. Mixtures remain in $\mathcal{K}_\delta$, ensuring feasibility of both queries. The number of expert projections is at most $\lfloor \log_2 t \rfloor + 1$ at round t; no extra first-order or dueling-oracle evaluations are needed.

E.4.3 Pathwise interval analysis

Lemma 6 combines expert regret, exponential-weight aggregation, and an interval cover; its proof below holds for every realized surrogate sequence.

Lemma 6 (Interval Regret on the Quadratic Surrogates)**.** *Let $m_I = \lceil \log_2 |I| \rceil + 1$. For every realized feedback sequence generated by Algorithm 5, every interval $I = [a, b]$, and every $v \in \mathcal{K}_\delta$,*

$$\sum_{t=a}^{b} \big(q_t(w_t) - q_t(v)\big) \le \frac{H_\delta^2}{2\sigma} m_I (1 + 3 \log T). \tag{25}$$

Proof. We give the expert, aggregation, and interval-covering arguments explicitly. They specialize the proof of (Zhang et al., 2018b, Theorem 2) to the quadratics and the normalized weights above.

Fix an expert j, a time $b' < e_j$, and $v \in \mathcal{K}_\delta$. Let $\tau = t - j + 1$, $a_t = \|w_{t,j} - v\|_2^2$, and $g_{t,j} = \nabla q_t(w_{t,j})$. Strong convexity and the projection inequality give

$$q_t(w_{t,j}) - q_t(v) \le \frac{\sigma}{2}\big((\tau - 1)a_t - \tau a_{t+1}\big) + \frac{\|g_{t,j}\|_2^2}{2\sigma\tau}.$$

Summing from j to b', the initial distance coefficient is zero and the final distance coefficient is nonpositive. Thus

$$\sum_{t=j}^{b'} \big(q_t(w_{t,j}) - q_t(v)\big) \le \frac{H_\delta^2}{2\sigma}\big(1 + \log(b' - j + 1)\big). \tag{26}$$

The bound uses H_δ, not merely the smaller norm of the estimate at the mixture center.

By exp-concavity and $w_t = \sum_j p_{t,j} w_{t,j}$,

$$Z_t = \sum_{j \in \mathcal{S}_t} p_{t,j} \exp\big(-\alpha_\delta (q_t(w_{t,j}) - q_t(w_t))\big) \le 1.$$

Consequently, for each active j,

$$\alpha_\delta \big(q_t(w_t) - q_t(w_{t,j})\big) \le \log \bar{p}_{t+1,j} - \log p_{t,j}.$$

For $j < t < e_j$, survivor normalization gives $p_{t,j} \ge (1 - 1/t)\bar{p}_{t,j}$, since $Z_t^{\mathrm{surv}} \le 1$. Also $p_{j,j} \ge 1/j$. For every $b' \in [j, e_j - 1] \cap [1, T]$, summation therefore yields

$$\begin{aligned} \alpha_\delta \sum_{t=j}^{b'} \big(q_t(w_t) - q_t(w_{t,j})\big) &\le \log \bar{p}_{b'+1,j} - \log p_{j,j} + \sum_{t=j+1}^{b'} \log \frac{t}{t-1} \\ &\le \log j + \log \frac{b'}{j} = \log b' \le \log T. \end{aligned} \tag{27}$$

The posterior $\bar{p}_{b'+1,j}$ is taken before the next pruning step, so this argument also covers the last round of an expert's lifetime.

Starting at $j_1 = a$, form the segments

$$[j_r, \min\{e_{j_r} - 1, b\}], \qquad j_{r+1} = e_{j_r},$$

until b is covered. These form a disjoint partition of I, each starting at the birth of an expert that remains active throughout that segment. Successive untruncated lengths $2^{\nu_2(j_r)}$ at least double. If there are m segments, the first $m - 1$ are complete, so $|I| \ge (2^{m-1} - 1) + 1 = 2^{m-1}$. In particular $m \le m_I$, as in the binary instance of (Zhang et al., 2018b, Lemma 1).

Add Eq. (26) and Eq. (27) on each segment. Since $\alpha_\delta^{-1} = H_\delta^2/\sigma$, this gives

$$\sum_{t=a}^{b} \big(q_t(w_t) - q_t(v)\big) \le \frac{H_\delta^2}{2\sigma} m (1 + 3 \log T) \le \frac{H_\delta^2}{2\sigma} m_I (1 + 3 \log T).$$

Every step is a deterministic inequality for the realized quadratic losses; no unbiasedness or independence of the surrogate sequence is assumed. □

E.4.4 Proof of Theorem 5 and radius tuning

Proof. The surrogates are σ-strongly convex and their gradients on $\mathcal{K}_\delta$ are bounded by H_δ. Lemma 6 specializes the IFLH analysis of (Zhang et al., 2018b, Theorem 2) to these surrogates: for every realization and every $v \in \mathcal{K}_\delta$,

$$\sum_{t\in I}\big(q_t(w_t) - q_t(v)\big) \le \frac{H_\delta^2}{2\sigma} m_I(1 + 3\log T).$$

Choose a fixed minimizer $w_I^* \in \arg\min_{w\in\mathcal{K}} \sum_{t\in I} f_t(w)$, set $v = \Pi_{\mathcal{K}_\delta}(w_I^*)$, and apply Theorem 1 to obtain the stated query-regret bound with penalty $C_{\mathrm{sph}}\delta|I|$.

For the uniform query-regret rate, let $\ell_T = \log(eT)$ and $\kappa = d/(2\gamma) + \sigma D$, and choose the single radius

$$\delta = c\min\{1, (\ell_T^2/T)^{1/3}\}, \qquad 0 < c \le 1/2.$$

Since $H_\delta \le \kappa/\delta$ and $m_I(1 + 3\log T) = \mathcal{O}(\ell_T^2)$ uniformly over intervals, Theorem 1 gives

$$\max_I \mathrm{Regret}_A(I) \le \mathcal{O}(\ell_T^2/\delta^2) + C_{\mathrm{sph}}\delta T.$$

When $\ell_T^2 \le T$, substituting the chosen radius yields $\mathcal{O}(T^{2/3}\ell_T^{2/3})$. When $\ell_T^2 > T$, the elementary bound $\mathrm{Regret}_A(I) \le GD|I| \le GDT \le GDT^{2/3}\ell_T^{2/3}$ gives the same order. Thus one radius, fixed before the run and shared by all experts, gives the stated rate for every horizon. □

F Proofs for smooth losses and ellipsoidal FTRL

Section F.1 introduces the barrier geometry and conditional notation. Sections F.2–F.4 establish feasibility, quadratic estimator bias, and the one-round transfer. Section F.5 proves the pathwise FTRL bound; Section F.6 combines these ingredients and handles boundary comparators. Sections F.7 and F.8 verify the two tunings.

F.1 Setup and barrier geometry

This appendix proves the static query-regret guarantees of Section 5. Throughout, the losses are fixed before the learner's randomization, G-Lipschitz, L-smooth, and σ-strongly convex on $\mathcal{K}$, with $\sigma \ge 0$. The link satisfies Assumption 1, is odd and three times continuously differentiable on $[-2G, 2G]$, and obeys $|\rho'''| \le M_3$. Algorithm 2 uses $0 < \delta \le \delta_0 := \min\{1/2, 1/D\}$ and $0 < \eta \le \gamma\delta/d$; η is fixed throughout each run. For any fixed feasible comparator $u \in \mathcal{K}$, with the query points for round t being $w_t^+, w_t^- \in \mathcal{K}$, write

$$R_T(u) = \mathbb{E}\sum_{t=1}^{T}\left[\frac{f_t(w_t^+) + f_t(w_t^-)}{2} - f_t(u)\right],$$

and regret R_T as

$$R_T = \max_{u\in\mathcal{K}} R_T(u) = \mathrm{Regret}_S(T).$$

Self-concordant barrier. The use of self-concordant barrier functions as the FTRL regularizer for zeroth-order online learning was pioneered by Abernethy et al. (2008) and extended in Saha & Tewari (2011) to smooth objectives, achieving $\widetilde{\mathcal{O}}(T^{2/3})$ regret. Further leveraging the inverse Hessian geometry, Hazan & Levy (2014) demonstrated how to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret bounds for strongly convex and smooth loss functions for bandit convex optimization. Definition 2 records the derivative inequalities used below for query feasibility, FTRL stability, and comparator approximation.

Definition 2 (Self-concordant barrier)**.** *A convex function $\Phi : \mathrm{int}(\mathcal{K}) \to \mathbb{R}$ is a ν-self-concordant barrier for a bounded convex domain $\mathcal{K} \subset \mathbb{R}^d$ with non-empty interior if it is three times continuously differentiable, blows up at the boundary $\partial\mathcal{K}$ (i.e., $\Phi(x) \to \infty$ as $x \to \partial\mathcal{K}$), and satisfies $\forall x \in \mathrm{int}\,\mathcal{K}$ and $\forall h \in \mathbb{R}^d$:*

1. ***Self-Concordance:*** $|D^3\Phi(x)[h,h,h]| \le 2\left(h^\top \nabla^2\Phi(x)h\right)^{3/2}$,

2. ***Barrier Property:*** $|D\Phi(x)[h]| \le \sqrt{\nu}\left(h^\top \nabla^2\Phi(x)h\right)^{1/2}$,

where D is the differentiation operator and h specifies the direction, thus $D\Phi(x)[h] = \frac{d}{ds}\Phi(x+sh)|_{s=0} = \langle \nabla\Phi(x), h\rangle$, and $D^3\Phi(x)[h,h,h] = \frac{d^3}{ds^3}\Phi(x+sh)|_{s=0}$.

As in Section 5, we assume $\nabla^2\Phi(x) \succ 0$ throughout $\operatorname{int}\mathcal{K}$. Since each realized surrogate is continuous on the compact domain and Φ diverges at its boundary, every FTRL objective attains its unique minimum in the interior. In zeroth-order online learning, FTRL with a Euclidean regularizer may place iterates on the boundary of the domain $\mathcal{K}$, where symmetric exploration with a fixed Euclidean radius becomes infeasible. A self-concordant barrier over $\operatorname{int}\mathcal{K}$ resolves this by keeping the FTRL iterates in the interior. Specifically, for $x \in \operatorname{int}\mathcal{K}$, write the local norm defined by the Hessian $\nabla^2\Phi(x)$ as $\|h\|_{\nabla^2\Phi(x)} := \sqrt{h^\top \nabla^2\Phi(x)h}$, then the corresponding local Dikin ellipsoid

$$\mathcal{E}_x(r) = \{y \in \mathbb{R}^d : \|y - x\|_{\nabla^2\Phi(x)} \le r\}$$

is contained in $\operatorname{int}(\mathcal{K})$ whenever $0 < r < 1$. This Dikin ellipsoid inclusion property allows the exploration geometry to adapt to the domain while maintaining query feasibility. It is also stated in (Hazan & Levy, 2014, Section 2.3, Eq. (3)).

Regularized Hessian, Sampling Matrix, and Spectral Norm. To unify the analysis of smooth convex and smooth strongly-convex objectives, we use the regularized Hessian $H_t = \nabla^2\Phi(w_t) + \eta\sigma t I$ and construct the sampling matrix $A_t = H_t^{-1/2}$ as its symmetric positive definite inverse square root, where $\sigma = 0$ covers the convex case. We write $\|A_t\|_{\mathrm{op}} := \max_{\|v\|_2=1}\|A_t v\|_2$ for its Euclidean operator norm, also called the spectral norm. This quantity measures the maximum stretching of a unit vector by A_t, so $\delta\|A_t\|_{\mathrm{op}}$ bounds the Euclidean displacement of either query resulting from w_t.

To prove Theorem 6 for Algorithm 2, we first establish the feasibility of the perturbed query points constructed by the sampling matrix A_t, and the conditional expectation of the ellipsoidal gradient estimator relative to the ellipsoidal smoothed gradient. For a single-round calculation, condition on the pre-query history $\mathcal{F}_{t-1}$ defined in Section 3, and write $x = w_t$, $A = A_t$, $H = H_t$, and $f = f_t$. These quantities are $\mathcal{F}_{t-1}$-measurable; we abbreviate $\mathbb{E}_t$ by $\mathbb{E}$ in the single-round calculations below. Let U be uniform on the unit sphere independently of this history, and let V be an independent auxiliary point uniform on the solid unit ball. Expectations involving $\widehat{g}$ include both U and the oracle bit. Denote the sampled loss difference $\Delta(v) = f(x+\delta Av) - f(x-\delta Av)$, and denote the ellipsoidal smoothed gradient

$$\bar{g} := \mathbb{E}_V m(V), \quad \text{where } m(v) = \frac{\nabla f(x+\delta Av) + \nabla f(x-\delta Av)}{2}.$$

By symmetry, $\bar{g} = \mathbb{E}_V[\nabla f(x+\delta AV)] = \mathbb{E}_V[\nabla f(x-\delta AV)]$, and it is the gradient at x of the locally defined ball smoothing with A held fixed.

F.2 Query feasibility and spectral bounds

The following lemma establishes query feasibility and the spectral bounds needed for the regret analysis.

Lemma 7 (Feasibility and Technical Bounds)**.** *All queries are feasible and $\|A_t\|_{\mathrm{op}} \le D$. If $\sigma > 0$, $\|A_t\|_{\mathrm{op}}^2 \le 1/(\eta\sigma t)$. Every sampled loss difference lies in $[-2G, 2G]$.*

Proof. Because $H_t \succeq \nabla^2\Phi(w_t) \succ 0$, we have $A_t^\top \nabla^2\Phi(w_t)A_t \preceq A_t^\top H_t A_t = I$. For $w_t \in \operatorname{int}\mathcal{K}$,

$$\begin{aligned}\|(w_t + \delta A_t U_t) - w_t\|_{\nabla^2\Phi(w_t)}^2 &= (\delta A_t U_t)^\top \nabla^2\Phi(w_t)(\delta A_t U_t)\\ &= \delta^2 U_t^\top A_t \nabla^2\Phi(w_t) A_t U_t \le \delta^2\|U_t\|^2 < 1.\end{aligned}$$

Thus, by the Dikin ellipsoid inclusion property stated above, all perturbed query points $w_t^+, w_t^- \in \operatorname{int}\mathcal{K}$, as the same inequality holds for w_t^-. The same calculation holds for every $v \in \mathbb{B}_d$, so $w_t + \delta A_t\mathbb{B}_d \subseteq \operatorname{int}\mathcal{K}$.

For $x \in \operatorname{int} \mathcal{K}$, set $B_x = (\nabla^2 \Phi(x))^{-1/2}$. For a unit vector v and $0 < r < 1$, $\|rB_x v\|_{\nabla^2\Phi(x)} = r < 1$, the Dikin ellipsoid inclusion property therefore gives $x + rB_x v \in \mathcal{K}$. Since $x \in \mathcal{K}$, the diameter assumption implies

$$r\|B_x v\|_2 = \|(x + rB_x v) - x\|_2 \le D$$

Letting r approach 1 from below and maximizing over unit vectors gives $\|B_x\|_{\mathrm{op}} \le D$. Because $H_t \succeq \nabla^2\Phi(w_t)$, we have $A_t^2 = H_t^{-1} \preceq (\nabla^2\Phi(w_t))^{-1} = B_{w_t}^2$. Thus, $\|A_t\|_{\mathrm{op}} \le D$. When $\sigma > 0$, we have $H_t \succeq \eta\sigma t I$, which yields $H_t^{-1} \preceq \frac{1}{\eta\sigma t} I$, and thus $\|A_t\|_{\mathrm{op}}^2 \le \frac{1}{\eta\sigma t}$.

Finally, since f_t are G-Lipschitz continuous, we have $|\Delta_t(v)| \le 2\delta G\|A_t v\| \le 2G$ by $\delta D \le 1$. □

F.3 Estimator identity, quadratic bias, and local norm

Lemma 8 uses the comparison identity from Lemma 1 and the third-derivative link bound to obtain quadratic bias and an exact local-norm identity. The proof below reuses Appendix C.

Lemma 8 (Ellipsoidal Gradient Estimator and Bias Control)**.** *The conditional expectation of the ellipsoid gradient estimator satisfies*

$$\mathbb{E}[\widehat{g}] = -\mathbb{E}_V\left[\frac{\rho'(\Delta(V))}{\gamma} m(V)\right] = \bar{g} + B$$

with the bias term $-\frac{1}{\gamma}\mathbb{E}_V\left[(\rho'(\Delta(V)) + \gamma)\, m(V)\right]$ *bounded by*

$$\|B\|_2 \le \frac{2M_3 G^3 \delta^2}{\gamma(d+2)} \operatorname{tr}(A^2) \le \frac{2M_3 G^3}{\gamma} \delta^2 \|A\|_{\mathrm{op}}^2.$$

Moreover, $\|\widehat{g}\|_{H^{-1}}^2 = d^2/(4\gamma^2\delta^2)$.

Proof. Lemma 1 applies by Lemma 7. Its proof in Appendix C uses the sphere-to-ball identity

$$d\mathbb{E}_U[h(U)U] = \mathbb{E}_V[\nabla_v h(V)], \tag{28}$$

as established in Eq. (15), with $h(v) = \rho(\Delta(v))$. Applying that comparison identity with the present x, A, and f gives

$$\mathbb{E}[\widehat{g}] = -\frac{1}{\gamma}\mathbb{E}_V[\rho'(\Delta(V))m(V)] = \bar{g} + B,$$
$$B = -\frac{1}{\gamma}\mathbb{E}_V[(\rho'(\Delta(V)) + \gamma)m(V)].$$

We now bound the differentiated-link remainder using the additional third-derivative assumption.

Since ρ is odd and three-times continuously differentiable, $\rho''(0) = 0$. Integrating the third-derivative bound twice,

$$\begin{aligned}\rho'(z) - \rho'(0) &= \int_0^z \rho''(s)\, ds = \int_0^z \left(\rho''(0) + \int_0^s \rho'''(r)\, dr\right) ds \\ &= \int_0^z \int_0^s \rho'''(r)\, dr\, ds = z^2 \int_0^1 \int_0^a \rho'''(bz)\, db\, da \\ &= z^2 \int_0^1 (1-b)\rho'''(bz)\, db.\end{aligned}$$

For $|z| \le 2G$ and $b \in [0,1]$, we have $|bz| \le 2G$. Thus, the bound $|\rho'''(bz)| \le M_3$ gives

$$\begin{aligned}|\rho'(z) - \rho'(0)| &= z^2 \left|\int_0^1 (1-b)\rho'''(bz)\, db\right| \le z^2 \int_0^1 (1-b)|\rho'''(bz)|\, db \\ &\le M_3 z^2 \int_0^1 (1-b)\, db = \frac{M_3}{2} z^2.\end{aligned}$$

With $\rho'(0) = -\gamma$, we have $|\rho'(z) + \gamma| \le \frac{M_3}{2} z^2$, for $|z| \le 2G$.

Since f is differentiable and G-Lipschitz, its gradients at the feasible interior points $x \pm \delta Av$ have norm at most G. Consequently,

$$\|m(v)\|_2 = \left\| \frac{\nabla f(x+\delta Av) + \nabla f(x-\delta Av)}{2} \right\|_2 \le \frac{1}{2}\left(\|\nabla f(x+\delta Av)\|_2 + \|\nabla f(x-\delta Av)\|_2\right) \le G.$$

Moreover, by G-Lipschitzness,

$$|\Delta(v)| = |f(x+\delta Av) - f(x-\delta Av)| \le G\|(x+\delta Av)-(x-\delta Av)\|_2 = 2\delta G\|Av\|_2.$$

Since $|\Delta(V)| \le 2G$, the previously established bound $|\rho'(z)+\gamma| \le M_3 z^2/2$ applies. Hence,

$$\begin{aligned}\|B\|_2 &= \frac{1}{\gamma}\left\|\mathbb{E}_V\left[(\rho'(\Delta(V))+\gamma)m(V)\right]\right\|_2 \le \frac{1}{\gamma}\mathbb{E}_V\left[|\rho'(\Delta(V))+\gamma|\,\|m(V)\|_2\right] \\ &\le \frac{M_3}{2\gamma}\mathbb{E}_V\left[\Delta(V)^2\|m(V)\|_2\right] \le \frac{GM_3}{2\gamma}\mathbb{E}_V[\Delta(V)^2] \\ &\le \frac{GM_3}{2\gamma}\mathbb{E}_V\left[4\delta^2G^2\|AV\|_2^2\right] = \frac{2M_3G^3\delta^2}{\gamma}\mathbb{E}_V\left[\|AV\|_2^2\right].\end{aligned}$$

For $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, reflecting coordinate i preserves the distribution of U while reversing the sign of U_iU_j, thus spherical symmetry gives $\mathbb{E}[U_iU_j] = 0,\ \forall(i \ne j)$. Moreover, coordinate symmetry implies that all $\mathbb{E}[U_i^2]$ are equal. Since $\|U\|_2^2 = 1$, $d\,\mathbb{E}[U_1^2] = \sum_{i=1}^d \mathbb{E}[U_i^2] = \mathbb{E}[\|U\|_2^2] = 1$. Therefore, $\mathbb{E}[UU^\top] = \frac{I}{d}$.

For $V \sim \mathrm{Unif}(\mathbb{B}_d)$, write $V = RU$, where $R \in [0,1]$ and U is uniform on the unit sphere, independently of R. The radial distribution satisfies

$$\mathbb{P}(R \le r) = \frac{\mathrm{vol}(r\mathbb{B}_d)}{\mathrm{vol}(\mathbb{B}_d)} = r^d, \qquad 0 \le r \le 1.$$

Hence its density is $p_R(r) = d\,r^{d-1}$, and

$$\mathbb{E}[R^2] = \int_0^1 r^2 p_R(r)\,dr = d\int_0^1 r^{d+1}\,dr = \frac{d}{d+2}.$$

Using independence,

$$\mathbb{E}[VV^\top] = \mathbb{E}[R^2UU^\top] = \mathbb{E}[R^2]\,\mathbb{E}[UU^\top] = \frac{d}{d+2}\frac{I}{d} = \frac{I}{d+2}.$$

Holding A fixed and using $A = A^\top$, we obtain

$$\begin{aligned}\mathbb{E}_V[\|AV\|_2^2] &= \mathbb{E}_V[V^\top A^\top AV] = \mathbb{E}_V[\mathrm{tr}(A^\top AVV^\top)] \\ &= \mathrm{tr}\big(A^\top A\,\mathbb{E}_V[VV^\top]\big) = \frac{\mathrm{tr}(A^\top A)}{d+2} = \frac{\mathrm{tr}(A^2)}{d+2}.\end{aligned}$$

Substituting into the preceding bias estimate yields

$$\begin{aligned}\|B\|_2 &\le \frac{2M_3G^3\delta^2}{\gamma}\mathbb{E}_V[\|AV\|_2^2] = \frac{2M_3G^3\delta^2}{\gamma(d+2)}\,\mathrm{tr}(A^2) \\ &\le \frac{2M_3G^3\delta^2}{\gamma}\frac{d}{d+2}\|A\|_{\mathrm{op}}^2 \le \frac{2M_3G^3\delta^2}{\gamma}\|A\|_{\mathrm{op}}^2,\end{aligned}$$

where we used $\mathrm{tr}(A^2) \le d\,\|A\|_{\mathrm{op}}^2$.

Finally, since $A = H^{-1/2}$, $o^2 = 1$, and $\|U\|_2^2 = 1$, with $\widehat{g} = -\frac{d}{2\gamma\delta}oA^{-1}U$,

$$\begin{aligned}\|\widehat{g}\|_{H^{-1}}^2 &= \widehat{g}^\top H^{-1}\widehat{g} = \|A\widehat{g}\|_2^2 = \|\frac{d}{2\gamma\delta}oU\|_2^2 \\ &= \frac{d^2}{4\gamma^2\delta^2}o^2\|U\|_2^2 = \frac{d^2}{4\gamma^2\delta^2}.\end{aligned}$$

□

F.4 Proof of Lemma 2

We compare the center loss with an arbitrary feasible comparator, then add query smoothing and estimator bias. Identifying the quadratic-surrogate regret proves the transfer inequality; a trace bound gives its operator-norm form.

Proof of Lemma 2. Fix a round t and a fixed comparator $u \in \mathcal{K}$, including boundary points. Condition on $\mathcal{F}_{t-1}$ and write $x = w_t$, $A = A_t$, $f = f_t$, and $\widehat{g} = \widehat{g}_t$. The quantities x, A, f, and u are fixed under this conditioning. All expectations below are conditional on this history; those involving $\widehat{g}$ include the query direction and oracle bit.

We first compare the center loss with the comparator loss. Let $h = \delta AV$ and $y = x + h$, where $V \sim \mathrm{Unif}(\mathbb{B}_d)$. The expectations in this step are over the auxiliary ball point V. By feasibility in Lemma 7, $y \in \mathrm{int}(\mathcal{K})$.

Applying smoothness and strong convexity at the same point y gives respectively

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|_2^2,$$
$$f(u) \ge f(y) + \langle \nabla f(y), u - y\rangle + \frac{\sigma}{2}\|u - y\|_2^2.$$

Subtracting the second inequality from the first, we obtain

$$\begin{aligned} f(x) - f(u) &\le \langle \nabla f(y), (x - y) - (u - y)\rangle + \frac{L}{2}\|x - y\|_2^2 - \frac{\sigma}{2}\|u - y\|_2^2 \\ &= \langle \nabla f(x + h), x - u\rangle + \frac{L}{2}\|h\|_2^2 - \frac{\sigma}{2}\|u - x - h\|_2^2. \end{aligned}$$

Since $\mathbb{E}[V] = 0$ and A is fixed,

$$\mathbb{E}[h] = \delta A\,\mathbb{E}[V] = 0.$$

Expanding the squared distance therefore gives

$$\begin{aligned} \mathbb{E}[\|u - x - h\|_2^2] &= \mathbb{E}\left[\|u - x\|_2^2 - 2\langle u - x, h\rangle + \|h\|_2^2\right] \\ &= \|u - x\|_2^2 - 2\langle u - x, \mathbb{E}[h]\rangle + \mathbb{E}[\|h\|_2^2] \\ &= \|u - x\|_2^2 + \mathbb{E}[\|h\|_2^2]. \end{aligned}$$

Moreover, symmetry of V implies

$$\begin{aligned} \mathbb{E}[\nabla f(x + h)] &= \frac{1}{2}\mathbb{E}\left[\nabla f(x + \delta AV) + \nabla f(x - \delta AV)\right] \\ &= \bar{g}. \end{aligned}$$

Taking expectations in the comparison inequality yields

$$\begin{aligned} f(x) - f(u) &\le \langle \mathbb{E}[\nabla f(x + h)], x - u\rangle + \frac{L}{2}\mathbb{E}[\|h\|_2^2] - \frac{\sigma}{2}\mathbb{E}[\|u - x - h\|_2^2] \\ &= \langle \bar{g}, x - u\rangle - \frac{\sigma}{2}\|x - u\|_2^2 + \frac{L - \sigma}{2}\mathbb{E}[\|h\|_2^2]. \end{aligned}$$

Using $h = \delta AV$ and $\mathbb{E}[VV^\top] = I/(d + 2)$, we have $\mathbb{E}[\|h\|_2^2] = \delta^2\mathbb{E}[\|AV\|_2^2] = \frac{\delta^2\,\mathrm{tr}(A^2)}{d+2}$. Thus,

$$f(x) - f(u) \le \langle \bar{g}, x - u\rangle - \frac{\sigma}{2}\|x - u\|_2^2 + \frac{L - \sigma}{2}\frac{\delta^2\,\mathrm{tr}(A^2)}{d + 2}. \tag{29}$$

This intermediate comparison requires no evaluation of f at $u + h$.

Next, define the expected average query loss by $Q(x) = \mathbb{E}_U\left[\frac{f(x+\delta AU)+f(x-\delta AU)}{2}\right]$, where $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. By L-smoothness at x,

$$f(x + \delta AU) \le f(x) + \delta\langle \nabla f(x), AU\rangle + \frac{L\delta^2}{2}\|AU\|_2^2,$$
$$f(x - \delta AU) \le f(x) - \delta\langle \nabla f(x), AU\rangle + \frac{L\delta^2}{2}\|AU\|_2^2.$$

Averaging these two inequalities cancels the linear terms:

$$\frac{f(x+\delta AU)+f(x-\delta AU)}{2} \le f(x)+\frac{L\delta^2}{2}\|AU\|_2^2.$$

Taking expectations and using $\mathbb{E}[UU^\top]=I/d$ and $A=A^\top$, we obtain

$$Q(x)-f(x) \le \frac{L\delta^2}{2}\mathbb{E}_U[\|AU\|_2^2] = \frac{L\delta^2}{2}\operatorname{tr}\big(A^\top A\,\mathbb{E}_U[UU^\top]\big) = \frac{L\delta^2}{2d}\operatorname{tr}(A^2).$$

Adding this inequality to Eq. (29) yields

$$\begin{aligned} Q(x)-f(u) &= \big(Q(x)-f(x)\big)+\big(f(x)-f(u)\big) \\ &\le \langle \bar{g}, x-u\rangle - \frac{\sigma}{2}\|x-u\|_2^2 + \left(\frac{L}{2d}+\frac{L-\sigma}{2(d+2)}\right)\delta^2\operatorname{tr}(A^2). \end{aligned}$$

By the definition of the bias vector, $B=\mathbb{E}[\widehat{g}]-\bar{g}$, and hence

$$\begin{aligned} \langle \bar{g}, x-u\rangle &= \langle \mathbb{E}[\widehat{g}]-B, x-u\rangle \\ &= \mathbb{E}[\langle \widehat{g}, x-u\rangle] - \langle B, x-u\rangle. \end{aligned}$$

Here the expectation involving $\widehat{g}$ is over both the sampled direction and the oracle bit.

Since $x,u\in\mathcal{K}$ and $\operatorname{diam}(\mathcal{K})\le D$, Cauchy–Schwarz and Lemma 8 give

$$\begin{aligned} -\langle B, x-u\rangle &\le |\langle B, x-u\rangle| \\ &\le \|B\|_2\|x-u\|_2 \\ &\le D\|B\|_2 \\ &\le \frac{2M_3G^3D\delta^2}{\gamma(d+2)}\operatorname{tr}(A^2). \end{aligned}$$

Substituting these bounds into the preceding comparison, we conclude that

$$\begin{aligned} Q(x)-f(u) &\le \mathbb{E}[\langle \widehat{g}, x-u\rangle] - \frac{\sigma}{2}\|x-u\|_2^2 \\ &\quad + \left(\frac{L}{2d}+\frac{L-\sigma}{2(d+2)}+\frac{2M_3G^3D}{\gamma(d+2)}\right)\delta^2\operatorname{tr}(A^2) \\ &= \mathbb{E}[\langle \widehat{g}, x-u\rangle] - \frac{\sigma}{2}\|x-u\|_2^2 + c_{d,\sigma}\delta^2\operatorname{tr}(A^2), \end{aligned}$$

where

$$c_{d,\sigma} := \frac{L}{2d}+\frac{L-\sigma}{2(d+2)}+\frac{2M_3G^3D}{\gamma(d+2)}.$$

Recalling the conditioning and $x=w_t$, we have

$$\begin{aligned} Q(x) &= \mathbb{E}_t\left[\frac{f_t(w_t^+)+f_t(w_t^-)}{2}\right], \\ q_t(w_t)-q_t(u) &= \langle \widehat{g}_t, w_t-u\rangle - \frac{\sigma}{2}\|w_t-u\|_2^2. \end{aligned}$$

Consequently, the preceding bound gives

$$\mathbb{E}_t\left[\frac{f_t(w_t^+)+f_t(w_t^-)}{2}-f_t(u)\right] \le \mathbb{E}_t[q_t(w_t)-q_t(u)] + c_{d,\sigma}\delta^2\operatorname{tr}(A_t^2), \tag{30}$$

which is Eq. (12) in Lemma 2.

It remains to prove the lemma's operator-norm bound. Since A is symmetric positive definite, let $\lambda_1,\dots,\lambda_d>0$ denote its eigenvalues. Then

$$\operatorname{tr}(A^2)=\sum_{i=1}^d \lambda_i^2, \qquad \|A\|_{\mathrm{op}} = \max_{1\le i\le d}\lambda_i.$$

Therefore, $\operatorname{tr}(A^2) = \sum_{i=1}^d \lambda_i^2 \le \sum_{i=1}^d \|A\|_{\mathrm{op}}^2 = d\,\|A\|_{\mathrm{op}}^2$. Moreover, by the definition of $c_{d,\sigma}$,

$$\begin{aligned} d\,c_{d,\sigma} &= \frac{L}{2} + \frac{d}{d+2}\frac{L-\sigma}{2} + \frac{d}{d+2}\frac{2M_3G^3D}{\gamma} \le \frac{L}{2} + \frac{L-\sigma}{2} + \frac{2M_3G^3D}{\gamma} \\ &= L - \frac{\sigma}{2} + \frac{2M_3G^3D}{\gamma} \le L + \frac{2M_3G^3D}{\gamma} =: C, \end{aligned}$$

where we used $d/(d+2) \le 1$ and $0 \le \sigma \le L$.

Combining these inequalities, and noting that $c_{d,\sigma} \ge 0$, gives

$$c_{d,\sigma}\delta^2 \operatorname{tr}(A^2) \le d\,c_{d,\sigma}\delta^2\|A\|_{\mathrm{op}}^2 \le C\delta^2\|A\|_{\mathrm{op}}^2.$$

Dividing by $\delta^2 > 0$ and recalling $A = A_t$ proves the remaining assertion of Lemma 2. □

F.5 Pathwise FTRL regret and curvature

Lemma 9 controls the realized quadratic-surrogate regret. Its proof identifies H_t as the updated objective Hessian, bounds the FTRL movement by self-concordance, and telescopes the surrogate losses.

Lemma 9 (Pathwise surrogate regret)**.** *For every realized estimator sequence from Algorithm 2 and every $u \in \operatorname{int}\mathcal{K}$, given $\frac{\eta d}{2\gamma\delta} \le \frac{1}{2}$,*

$$\sum_{t=1}^T \Big[\langle \widehat{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|^2\Big] \le \frac{\Phi(u) - \Phi(w_1)}{\eta} + 2\eta\sum_{t=1}^T \|\widehat{g}_t\|_{H_t^{-1}}^2. \tag{31}$$

Proof. This is the application of (Hazan & Levy, 2014, Lemma 11) to the quadratic surrogate losses q_t, after subtracting $\Phi(w_1)$ from the regularizer. We give a direct proof to expose the indexing, curvature and step-size requirements, and to cover $\sigma = 0$ explicitly.

Recall that the surrogate loss at round s is $q_s(w) = \langle \widehat{g}_s, w\rangle + \frac{\sigma}{2}\|w - w_s\|_2^2$. When differentiating with respect to w, the previously generated vectors w_s and $\widehat{g}_s$ are held fixed, thus

$$\nabla q_s(w) = \widehat{g}_s + \sigma(w - w_s), \qquad \nabla^2 q_s(w) = \sigma I.$$

Define $F_0(w) := \Phi(w)$, and $F_t(w) := \Phi(w) + \eta\sum_{s=1}^t q_s(w)$ for $t \ge 1$. By the FTRL update, $w_t = \arg\min_{w\in\operatorname{int}(\mathcal{K})} F_{t-1}(w)$, which also holds for $t = 1$, because the initialization is $w_1 = \arg\min_{w\in\operatorname{int}(\mathcal{K})} F_0(w)$. Since w_t is an interior minimizer of the differentiable function F_{t-1}, its first-order optimality condition is $\nabla F_{t-1}(w_t) = 0$. Using $F_t = F_{t-1} + \eta q_t$, we therefore obtain

$$\nabla F_t(w_t) = \nabla F_{t-1}(w_t) + \eta\nabla q_t(w_t) = 0 + \eta\,(\widehat{g}_t + \sigma(w_t - w_t)) = \eta\widehat{g}_t.$$

For the Hessian, every surrogate contributes σI, including the current surrogate q_t. Hence,

$$\nabla^2 F_t(w_t) = \nabla^2\Phi(w_t) + \eta\sum_{s=1}^t \nabla^2 q_s(w_t) = \nabla^2\Phi(w_t) + \eta\sum_{s=1}^t \sigma I = \nabla^2\Phi(w_t) + \eta\sigma t I.$$

Thus,

$$\nabla F_t(w_t) = \eta\widehat{g}_t, \qquad \nabla^2 F_t(w_t) = \nabla^2\Phi(w_t) + \eta\sigma t I = H_t. \tag{32}$$

Although w_t minimizes F_{t-1}, the matrix H_t is the Hessian of the updated objective F_t, evaluated at w_t. Equivalently, $H_t = \nabla^2 F_{t-1}(w_t) + \eta\sigma I$. The additional curvature $\eta\sigma I$ comes from the quadratic part of the current surrogate, whose gradient vanishes at its center but whose Hessian does not. This explains the index t, rather than $t-1$, in the sampling matrix $A_t = H_t^{-1/2}$. Moreover, H_t is available before observing the current feedback, so the unknown vector $\widehat{g}_t$ appears only in a linear term and therefore does not affect the Hessian.

Write $F_t(w) = \Phi(w) + P_t(w)$ where $P_t(w) := \eta\sum_{s=1}^t q_s(w)$. Since each q_s is a quadratic function with Hessian σI, we have $\nabla^2 P_t(w) = \eta\sigma t I \succeq 0$, and $D^3 P_t(w)[v,v,v] = 0$. Thus for every $z \in \operatorname{int}(\mathcal{K})$ and $v \in \mathbb{R}^d$,

$$|D^3 F_t(z)[v,v,v]| = |D^3\Phi(z)[v,v,v]| \le 2\big(v^\top\nabla^2\Phi(z)v\big)^{3/2} \le 2\big(v^\top\nabla^2 F_t(z)v\big)^{3/2},$$

which concludes that F_t satisfies the standard self-concordance inequality. Moreover, $\nabla^2 F_t(z) \succ 0$ because $\nabla^2 \Phi(z) \succ 0$. Let $h := w_{t+1} - w_t$, $r := \|h\|_{H_t}$, and $\lambda := \eta \|\widehat{g}_t\|_{H_t^{-1}} \leq \frac{1}{2}$. Recall that $H_t = \nabla^2 F_t(w_t)$, and $\nabla F_t(w_t) = \eta \widehat{g}_t$. If $h = 0$, the displacement and stability bounds $r \leq \frac{\lambda}{1-\lambda} \leq 2\lambda$ hold immediately.

If $h \neq 0$, then $r > 0$. Both iterates w_t, w_{t+1} lie in $\operatorname{int}(\mathcal{K})$, and this set is convex. Therefore, the entire segment $w_t + sh$, $s \in [0,1]$, lies in the interior. Define its directional curvature by $a(s) := h^\top \nabla^2 F_t(w_t + sh) h$. Positive definiteness gives $a(s) > 0$, and $a(0) = h^\top H_t h = r^2$. Differentiating along the segment gives $a'(s) = D^3 F_t(w_t + sh)[h, h, h]$. By self-concordance,

$$|a'(s)| \leq 2a(s)^{3/2}.$$

To integrate this inequality, define $b(s) := a(s)^{-1/2}$. Then $b'(s) = -\frac{1}{2} a(s)^{-3/2} a'(s)$, and $|b'(s)| \leq \frac{|a'(s)|}{2a(s)^{3/2}} \leq 1$. Hence, for $s \in [0,1]$, $a(s)^{-1/2} = b(s) = b(0) + \int_0^s b'(\tau)\, d\tau \leq b(0) + s = \frac{1}{r} + s$. Since both sides are positive, taking reciprocal squares yields

$$a(s) \geq \left(\frac{1}{r} + s\right)^{-2} = \frac{r^2}{(1+sr)^2}.$$

Next, let $\psi(s) := F_t(w_t + sh)$. Its first two derivatives are $\psi'(s) = \langle \nabla F_t(w_t + sh), h \rangle$ and $\psi''(s) = a(s)$. Since w_{t+1} is an interior minimizer of F_t, first-order condition dictates that $\nabla F_t(w_{t+1}) = 0$, and therefore $\psi'(1) = 0$. The fundamental theorem of calculus gives:

$$\begin{aligned} -\langle \nabla F_t(w_t), h \rangle = \psi'(1) - \psi'(0) &= \int_0^1 a(s)\, ds \geq \int_0^1 \frac{r^2}{(1+sr)^2}\, ds \\ &= \left[-\frac{r}{1+sr}\right]_{s=0}^{s=1} = r - \frac{r}{1+r} = \frac{r^2}{1+r}. \end{aligned}$$

On the other hand, Cauchy–Schwarz in the H_t and H_t^{-1} norms gives

$$-\langle \nabla F_t(w_t), h \rangle \leq |\langle \nabla F_t(w_t), h \rangle| \leq \|\nabla F_t(w_t)\|_{H_t^{-1}} \|h\|_{H_t} = \eta \|\widehat{g}_t\|_{H_t^{-1}} r = \lambda r.$$

Combining the lower and upper bounds, we obtain $\frac{r^2}{1+r} \leq \lambda r$. Since $r > 0$, division by r and rearrangement give $(1-\lambda) r \leq \lambda$. Using $\lambda \leq 1/2$, it follows that

$$r \leq \frac{\lambda}{1-\lambda} \leq 2\lambda,$$

which also holds when $h = 0$. By the definition of q_t and $w_{t+1} = w_t + h$,

$$\begin{aligned} q_t(w_t) - q_t(w_{t+1}) &= \langle \widehat{g}_t, w_t \rangle - \left(\langle \widehat{g}_t, w_t + h \rangle + \frac{\sigma}{2} \|h\|_2^2 \right) \\ &= -\langle \widehat{g}_t, h \rangle - \frac{\sigma}{2} \|h\|_2^2 \leq |\langle \widehat{g}_t, h \rangle| \leq \|\widehat{g}_t\|_{H_t^{-1}} \|h\|_{H_t} \\ &= \|\widehat{g}_t\|_{H_t^{-1}} r \leq 2\lambda \|\widehat{g}_t\|_{H_t^{-1}} = 2\eta \|\widehat{g}_t\|_{H_t^{-1}}^2. \end{aligned} \tag{33}$$

Define $J_t(w) := \frac{\Phi(w)}{\eta} + \sum_{s=1}^t q_s(w)$ and $S_t := \min_{w \in \operatorname{int} \mathcal{K}} J_t(w)$, for $t = 0, \ldots, T$, where the sum is empty when $t = 0$. Since $J_t = F_t / \eta$ and $\eta > 0$, the update rule implies that w_{t+1} minimizes J_t. Therefore,

$$S_t = J_t(w_{t+1}) = J_{t-1}(w_{t+1}) + q_t(w_{t+1}) \geq S_{t-1} + q_t(w_{t+1}),$$

where the inequality follows from $S_{t-1} = \min_{w \in \operatorname{int} \mathcal{K}} J_{t-1}(w)$. Rearranging and summing over t gives

$$\sum_{t=1}^T q_t(w_{t+1}) \leq \sum_{t=1}^T (S_t - S_{t-1}) = S_T - S_0.$$

Evaluating J_T at the comparator u yields

$$S_T \leq J_T(u) = \frac{\Phi(u)}{\eta} + \sum_{t=1}^T q_t(u).$$

Moreover, since w_1 minimizes Φ,

$$S_0 = \min_{w \in \operatorname{int} \mathcal{K}} \frac{\Phi(w)}{\eta} = \frac{\Phi(w_1)}{\eta}.$$

Combining these relations, we obtain

$$\sum_{t=1}^{T} \big[q_t(w_{t+1}) - q_t(u)\big] \le \frac{\Phi(u) - \Phi(w_1)}{\eta}. \tag{34}$$

Add and subtract $q_t(w_{t+1})$ and apply the stability bound Eq. (33):

$$\begin{aligned}\sum_{t=1}^{T} \big[q_t(w_t) - q_t(u)\big] &= \sum_{t=1}^{T} \big[q_t(w_t) - q_t(w_{t+1})\big] + \sum_{t=1}^{T} \big[q_t(w_{t+1}) - q_t(u)\big] \\ &\le 2\eta \sum_{t=1}^{T} \|\widehat{g}_t\|^2_{H_t^{-1}} + \frac{\Phi(u) - \Phi(w_1)}{\eta}.\end{aligned}$$

Using $q_t(w) = \langle \widehat{g}_t, w\rangle + \frac{\sigma}{2}\|w - w_t\|_2^2$, we have

$$\begin{aligned}q_t(w_t) - q_t(u) &= \langle \widehat{g}_t, w_t\rangle - \Big(\langle \widehat{g}_t, u\rangle + \frac{\sigma}{2}\|u - w_t\|_2^2\Big) \\ &= \langle \widehat{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2.\end{aligned}$$

Substituting $q_t(w_t) - q_t(u)$, we have

$$\sum_{t=1}^{T} \Big[\langle \widehat{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2\Big] \le \frac{\Phi(u) - \Phi(w_1)}{\eta} + 2\eta \sum_{t=1}^{T} \|\widehat{g}_t\|^2_{H_t^{-1}}.$$

Every step holds for each realization satisfying the stated step-size condition. Thus the bound is pathwise and requires no unbiasedness assumption on the estimators. □

F.6 Proof of Theorem 6

We combine Lemmas 2 and 9 for an interior comparator, then use barrier growth to approximate any boundary comparator. The approximation cost is charged in the original losses.

Proof. **Interior comparator bound.** Fix a deterministic comparator $u \in \operatorname{int}(\mathcal{K})$. Use the pre-query history $\mathcal{F}_{t-1}$ from Section 3 and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. Under the oblivious-adversary assumption, the loss functions are fixed. Moreover, w_t, A_t, and H_t are $\mathcal{F}_{t-1}$-measurable. Define the realized average query loss by

$$L_t^{\mathrm{q}} := \frac{f_t(w_t^+) + f_t(w_t^-)}{2}.$$

Applying Lemma 2, proved in Appendix F.4, and expanding $q_t(w_t) - q_t(u)$ gives

$$\mathbb{E}_t[L_t^{\mathrm{q}}] - f_t(u) \le \langle \mathbb{E}_t[\widehat{g}_t], w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2 + c_{d,\sigma}\delta^2 \operatorname{tr}(A_t^2).$$

Since $w_t - u$ is $\mathcal{F}_{t-1}$-measurable, the tower property gives

$$\mathbb{E}\left[\langle \mathbb{E}_t[\widehat{g}_t], w_t - u\rangle\right] = \mathbb{E}\left[\mathbb{E}_t\left[\langle \widehat{g}_t, w_t - u\rangle\right]\right] = \mathbb{E}\left[\langle \widehat{g}_t, w_t - u\rangle\right].$$

Taking expectations in the conditional comparison inequality therefore yields

$$\mathbb{E}[L_t^{\mathrm{q}}] - f_t(u) \le \mathbb{E}\left[\langle \widehat{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2\right] + c_{d,\sigma}\delta^2 \mathbb{E}[\operatorname{tr}(A_t^2)].$$

Summing over $t = 1, \dots, T$, we obtain

$$R_T(u) = \sum_{t=1}^{T} \left(\mathbb{E}[L_t^{\mathrm{q}}] - f_t(u)\right) \le \mathbb{E}\left[\sum_{t=1}^{T} \Big(\langle \widehat{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2\Big)\right] + c_{d,\sigma}\delta^2 \sum_{t=1}^{T} \mathbb{E}[\operatorname{tr}(A_t^2)].$$

For the quadratic surrogate losses $q_t(w) = \langle \widehat{g}_t, w\rangle + \frac{\sigma}{2}\|w - w_t\|_2^2$, we have

$$q_t(w_t) - q_t(u) = \langle \widehat{g}_t, w_t - u\rangle - \frac{\sigma}{2}\|w_t - u\|_2^2.$$

Thus, Lemma 9 applies pathwise and gives

$$R_T(u) \le \frac{\Phi(u) - \Phi(w_1)}{\eta} + 2\eta \sum_{t=1}^{T} \mathbb{E}\left[\|\widehat{g}_t\|_{H_t^{-1}}^2\right] + c_{d,\sigma}\delta^2 \sum_{t=1}^{T} \mathbb{E}[\mathrm{tr}(A_t^2)].$$

Finally, Lemma 8 establishes the deterministic identity

$$\|\widehat{g}_t\|_{H_t^{-1}}^2 = \frac{d^2}{4\gamma^2\delta^2}.$$

Substituting this identity yields

$$R_T(u) \le \frac{\Phi(u) - \Phi(w_1)}{\eta} + \frac{\eta d^2 T}{2\gamma^2\delta^2} + c_{d,\sigma}\delta^2 \sum_{t=1}^{T} \mathbb{E}\,\mathrm{tr}(A_t^2). \tag{35}$$

This is the required interior-comparator bound.

Best-fixed-in-hindsight comparator regret. Fix any $u \in \mathcal{K}$ and $0 < \alpha < 1$, and define $u_\alpha := (1 - \alpha)u + \alpha w_1$. We first verify that $u_\alpha \in \mathrm{int}\,\mathcal{K}$. Since $w_1 \in \mathrm{int}\,\mathcal{K}$, there exists $\varepsilon > 0$ such that the open Euclidean ball $B(w_1, \varepsilon)$ is contained in $\mathcal{K}$. By convexity, $B(u_\alpha, \alpha\varepsilon) = (1-\alpha)u + \alpha B(w_1, \varepsilon) \subseteq \mathcal{K}$, and thus the interior-comparator bound Eq. (35) applies to u_α.

We prove the required barrier-growth bound directly from

$$|D\Phi(x)[v]|^2 \le \nu D^2\Phi(x)[v, v],$$

which is the gradient inequality in the definition of a ν-self-concordant barrier. Set $h = u - w_1$ and define

$$\varphi(s) := \Phi(w_1 + sh), \qquad b(s) := \exp\left(-\frac{\varphi(s)}{\nu}\right), \qquad 0 \le s < 1.$$

The segment points $w_1 + sh$ lie in $\mathrm{int}\,\mathcal{K}$ for $s < 1$. By the chain rule and the barrier's gradient inequality,

$$\varphi'(s) = D\Phi(w_1 + sh)[h], \qquad \varphi''(s) = D^2\Phi(w_1 + sh)[h, h], \qquad (\varphi'(s))^2 \le \nu\varphi''(s).$$

Consequently,

$$b''(s) = b(s)\left(\frac{(\varphi'(s))^2}{\nu^2} - \frac{\varphi''(s)}{\nu}\right) \le 0.$$

Thus b is positive and concave. For $0 < s < r < 1$,

$$b(s) \ge \left(1 - \frac{s}{r}\right) b(0) + \frac{s}{r} b(r) \ge \left(1 - \frac{s}{r}\right) b(0).$$

Letting $r \uparrow 1$ gives $b(s) \ge (1 - s)b(0)$. Taking logarithms and rearranging yields

$$\varphi(s) - \varphi(0) \le \nu \log \frac{1}{1 - s}.$$

The choice $s = 1 - \alpha$ proves $\Phi(u_\alpha) - \Phi(w_1) \le \nu \log(1/\alpha)$.

Next, since $\mathcal{K}$ has diameter D, we have $\|u_\alpha - u\|_2 = \alpha\|w_1 - u\|_2 \le \alpha D$, and the G-Lipschitz continuity of each f_t therefore implies

$$f_t(u_\alpha) - f_t(u) \le G\|u_\alpha - u\|_2 \le \alpha GD.$$

For the fixed oblivious loss sequence, the definition of $R_T(u)$ gives the exact identity

$$R_T(u) = R_T(u_\alpha) + \sum_{t=1}^{T} \big[f_t(u_\alpha) - f_t(u)\big].$$

Applying Eq. (35) at u_α, we obtain

$$\begin{aligned} R_T(u) &\le R_T(u_\alpha) + \alpha GDT \\ &\le \frac{\Phi(u_\alpha) - \Phi(w_1)}{\eta} + \frac{\eta d^2 T}{2\gamma^2\delta^2} + c_{d,\sigma}\delta^2 \sum_{t=1}^{T} \mathbb{E}\,\mathrm{tr}(A_t^2) + \alpha GDT \\ &\le \frac{\nu \log(1/\alpha)}{\eta} + \frac{\eta d^2 T}{2\gamma^2\delta^2} + C\delta^2 \sum_{t=1}^{T} \mathbb{E}\|A_t\|_{\mathrm{op}}^2 + \alpha GDT, \end{aligned}$$

where $C = L + \frac{2M_3 G^3 D}{\gamma}$, and the last inequality uses the pointwise bound $c_{d,\sigma}\,\mathrm{tr}(A_t^2) \le C\|A_t\|_{\mathrm{op}}^2$. Since A_t may depend on the random history, its squared operator norm remains inside the expectation.

The resulting upper bound is independent of u. For an oblivious loss sequence, a minimizer

$$u^\star \in \arg\min_{u\in\mathcal{K}} \sum_{t=1}^{T} f_t(u)$$

exists by compactness and continuity, and can be chosen independently of the algorithm's randomness. Consequently,

$$R_T \le \frac{\nu \log(1/\alpha)}{\eta} + \frac{\eta d^2 T}{2\gamma^2\delta^2} + C\delta^2 \sum_{t=1}^{T} \mathbb{E}\|A_t\|_{\mathrm{op}}^2 + \alpha GDT. \tag{36}$$

The comparator approximation is performed on the original losses through $f_t(u_\alpha) - f_t(u)$ and requires no bound on the Euclidean norms of the linear estimators. □

F.7 Proof of Corollary 1

We substitute $\|A_t\|_{\mathrm{op}} \le D$ into the master bound and balance learning and exploration errors. The capped parameters remain feasible at finite horizons, and the zero-geometric-error case is treated separately.

Proof. Let $\sigma = 0$ and $C_{\mathrm{geom}} = CD^2 > 0$. Write $B_T = \nu \log(T+1)$ and $V_0 = d^2/(2\gamma^2)$. Set $\alpha = 1/(T+1)$ and use $\|A_t\|_{\mathrm{op}}^2 \le D^2$ from Lemma 7 in Eq. (36) to obtain

$$R_T \le B_T/\eta + V_0\eta T/\delta^2 + C_{\mathrm{geom}}\delta^2 T + GD. \tag{37}$$

For fixed δ, balancing the first two terms gives $\eta = \delta\sqrt{B_T/(V_0 T)}$ and their sum is $2\sqrt{B_T V_0 T}/\delta$. Minimizing this plus $C_{\mathrm{geom}}\delta^2 T$ gives the uncapped choices

$$\delta_* = \left(\frac{B_T V_0}{C_{\mathrm{geom}}^2 T}\right)^{1/6}, \qquad \eta_* = \delta_*\sqrt{\frac{B_T}{V_0 T}}. \tag{38}$$

At these choices, the first two terms together equal twice the third, so

$$R_T \le 3(B_T V_0 C_{\mathrm{geom}})^{1/3} T^{2/3} + GD = O\Big(T^{2/3}(\log(T+1))^{1/3}\Big). \tag{39}$$

The conditions needed are $\delta_* \le \delta_0$ and $B_T \le T/2$; the latter gives $\eta_* d/(2\gamma\delta_*) \le 1/2$. Both hold eventually for fixed problem constants.

For every finite T, a valid choice is $\delta = \min\{\delta_0, \delta_*\}$ and $\eta = \min\{\delta\sqrt{B_T/(V_0 T)}, \gamma\delta/d\}$; use Eq. (37) directly when a cap is active. If $C_{\mathrm{geom}} = 0$, take a fixed feasible δ and the capped learning rate $\eta = \min\{\delta\sqrt{B_T/(V_0 T)}, \gamma\delta/d\}$. Once the cap is inactive, balancing just the first two terms gives $O(\sqrt{T\log(T+1)})$. This includes affine losses with a locally linear oracle. □

F.8 Proof of Corollary 2

We use $\|A_t\|_{\mathrm{op}}^2 \le 1/(\eta\sigma t)$ to sum the geometric error harmonically, then optimize the learning rate with a fixed feasible radius.

Proof. Let $\sigma > 0$, choose any $0 < \delta \le \delta_0$ fixed independently of T, and set $\alpha = 1/(T+1)$. For $\mathsf{H}_T = \sum_{t=1}^{T} 1/t$, define

$$V_\delta = \frac{d^2}{2\gamma^2\delta^2}, \qquad K_T = \nu \log(T+1) + \frac{C\delta^2}{\sigma}\mathsf{H}_T.$$

Use $\|A_t\|_{\mathrm{op}}^2 \le 1/(\eta\sigma t)$ from Lemma 7 in Eq. (36) to obtain

$$R_T \le V_\delta \eta T + K_T/\eta + GD. \tag{40}$$

Balancing the first two terms gives $\eta = \sqrt{K_T/(V_\delta T)}$; imposing the cap yields $\eta = \min\{\sqrt{K_T/(V_\delta T)}, \gamma\delta/d\}$. Once the cap is inactive,

$$R_T \le 2\sqrt{V_\delta T K_T} + GD = O\Big(\sqrt{T\log(T+1)}\Big), \tag{41}$$

where $\mathsf{H}_T \le 1 + \log T$. The cap is inactive whenever $K_T \le T/2$, which eventually holds for fixed problem constants. □