MULTINOMIAL POLICIES FOR LARGE-SCALE BANDIT CONVEX OPTIMIZATION ON THE SIMPLEX: ALGORITHM DESIGN AND PERFORMANCE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimization over the simplex arises in a wide range of large-scale resource and workload allocation problems. We study an unknown convex objective through noisy one-point feedback and develop a multinomial-policy method that samples feasible allocations. A geometry-adapted stochastic mirror descent update learns the policy mean, while first-order comparisons connect its gradients to regret for the original objective even when the policy value is nonconvex. We prove sublinear regret with explicit dimension dependence for general convex objectives, sharper bounds under smoothness and, with additional structure and optimizer sparsity. Numerical experiments demonstrate the efficiency of the proposed algorithm, compared with existing approaches.

1 INTRODUCTION

In optimization over the simplex, the decision variables are nonnegative and constrained to sum to one. Resource and workload allocation can take this form when a fixed total is divided among alternatives. Examples include transmit power across wireless channels (Boyd & Vandenberghe, 2004; Wei et al., 2021), computation across edge servers (Ale et al., 2022), and traffic across paths (Wang et al., 2025). Normalization gives a simplex decision, with benefit maximization recast as loss minimization when needed (Patriksson, 2008).

We study stochastic bandit convex optimization on the simplex with a fixed unknown objective and one noisy value per allocation. Multinomial policies generate feasible allocations, and a geometry-adapted mirror descent update learns their means. We connect policy optimization to regret for the original objective, even when the policy value is nonconvex, and establish sublinear guarantees for general, smooth, and sparse cases.

1.1 MOTIVATION

Classical one-point BCO perturbs a point-valued decision and uses the observed value to estimate a smoothed gradient (Flaxman et al., 2005; Saha & Tewari, 2011). Such perturbations are often defined in the ambient space rather than by simplex geometry. Dirichlet-based methods instead provide feasible, sometimes single-evaluation gradient estimators (Lam & Zhang, 2025; Zrnic & Mazumdar, 2022). These motivate regret analysis for feasible randomization under noisy feedback.

Gradient-based stochastic search, including natural evolution strategies and information-geometric optimization, updates a sampling distribution from observed values (Wierstra et al., 2014; Ollivier et al., 2017; Zhou & Hu, 2014; Zhou & Bhatnagar, 2018). We likewise optimize distribution parameters, using the samples as feasible online allocations.

Policy-gradient methods for multi-armed bandits also optimize randomized policies from sampled rewards, with global-convergence guarantees established recently (Mei et al., 2023; 2024; Robertson et al., 2025). A linear objective has a vertex optimum and reduces to identifying a best arm; nonlinear allocation values depend jointly on all shares and cannot be represented by single-arm rewards. Our method extends this perspective to nonlinear convex optimization over simplex-valued allocations.

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.

1.2 Related work

Bandit convex optimization. Bandit convex optimization (BCO) studies online convex decision-making when the learner observes only the function value at the deployed action (Lattimore, 2026). Related regret bounds are summarized in Table 1.

Table 1: Representative regret guarantees for one-point zeroth-order online convex optimization.

Algorithm	Objective properties	Noise	Regret bound
BCO (Flaxman et al., 2005)	Convex, ℓ_2-Lipschitz on compact convex sets	No	$O(KT^{3/4})$
Smooth BCO (Saha & Tewari, 2011)	Convex, ℓ_2-smooth on compact convex sets	No	$\widetilde{O}(KT^{2/3})$
Dirichlet PGD (Zrnic & Mazumdar, 2022)	Convex, ℓ_2-smooth on the simplex	No	$\widetilde{O}(KT^{3/4})$
Negative-entropy MD (Flaxman et al., 2005; Hazan et al., 2016)	Convex, ℓ_1-Lipschitz on the simplex	Conditionally zero-mean & bounded	$\widetilde{O}(KT^{3/4})$
Center-point ellipsoid (Agarwal et al., 2013)	Convex, ℓ_2-Lipschitz on compact convex sets	i.i.d., zero-mean, sub-Gaussian	$\widetilde{O}(K^{16}T^{1/2})$
Surrogate-loss ellipsoid (Lattimore & György, 2021)	Convex on compact convex sets	Conditionally zero-mean sub-Gaussian	$\widetilde{O}(K^{9/2}T^{1/2})$
Online Newton (Fokkema et al., 2024)	Convex on compact convex sets	Conditionally zero-mean sub-Gaussian	$O(K^{3/2}T^{1/2}$ $\mathrm{polylog}(K,T,1/\delta))$ with probability $1-\delta$
Ours	Convex, ℓ_1-Lipschitz on the simplex	Conditionally zero-mean & bounded	$\boldsymbol{\widetilde{O}(K^{1/2}T^{3/4})}$
Ours	Convex, ℓ_1-Lipschitz and ℓ_2-smooth on the simplex	Conditionally zero-mean & bounded	$\boldsymbol{\widetilde{O}(K^{1/3}T^{2/3})}$

Optimization on the simplex. With gradients, Chok & Vasil (2025) obtain feasible sphere-parameterized updates and KL convergence. For nonsmooth stochastic objectives, Gasnikov et al. (2016) analyze two-point mirror descent using shared randomness and simplex-neighborhood queries. For noisy feasible queries, Lam & Zhang (2025) combine Dirichlet-mixture estimators, including single-evaluation variants, with Frank–Wolfe and entropic mirror descent, controlling gradient-estimate bias and variance to prove convergence. With exact smooth-objective values, Zrnic & Mazumdar (2022) use feasible Dirichlet perturbations for projected gradient descent and exponential weights, obtaining $O(T^{-1/4})$ average optimization error. Our unbiased MP score instead targets the policy-value gradient, with approximation control and finite-time cumulative regret. Other methods use recursive pattern search with up to $2K$ evaluations (Das, 2021) or Bayesian models for expensive distributional objectives (Candelieri et al., 2025; Pavesi et al., 2026); their focus differs from one-action cumulative regret.

Gradient-based stochastic search. Adaptive stochastic search treats nonsmooth simulation objectives (Zhou & Hu, 2014; Zhou & Bhatnagar, 2018); natural evolution strategies use natural gradients and information-geometric optimization emphasizes invariance (Wierstra et al., 2014; Ollivier et al., 2017). These largely offline, population-based methods motivate our online one-action regret analysis. Categorical policies also underpin architecture search (Akimoto et al., 2019). MSE3 uses sampled-arm counts and set rewards in sum-max submodular bandits (Pasteris et al., 2024), relying on concavity of expected set rewards; our allocation losses depend on shares, and our general guarantee permits a nonconvex policy value. Dirichlet policies also serve portfolios, battery allocation, offloading, robotic control, and traffic assignment (André & Coqueret, 2020; Tian et al., 2022; Ale et al., 2022; Ma et al., 2024; Wang et al., 2025). We analyze stationary convex bandit regret with one noisy feasible query per round and boundary comparators.

1.3 Our contributions

Our contributions connect policy gradients to regret for the original convex objective, even when the policy value is nonconvex.

First-order bridges from policy gradients to the original objective. For smooth convex objectives, a replace-one comparison bounds the gap to any comparator by a policy-gradient term plus $O(1/n)$ error without MP convexity (Lemma E.4). For general convex Lipschitz objectives, a scaled first-order inequality and positive-floor implicit update control boundary error without extra observations or subgradients (Lemma F.3). A counterexample rules out uniformly vanishing error for the corresponding raw-gradient inequality, but is not a regret lower bound (Proposition F.1).

One-point regret guarantees with sublinear dimension dependence. Under bounded, conditionally zero-mean noise, Theorem 3.1 gives expected pseudo-regret $\widetilde{O}(K^{1/2}T^{3/4})$ for convex ℓ_1-Lipschitz objectives and $\widetilde{O}(K^{1/3}T^{2/3})$ under ℓ_2-smoothness. Both allow boundary optima without MP convexity or sparsity information. Algorithm 1 uses an unbiased multinomial score, a stored noisy reference, matched-Fisher geometry, and a positive probability floor, with one new observation per subsequent round and no score clipping. Table 1 compares oracle assumptions and rates.

Sharper guarantees from optimizer sparsity. Given a known support-size bound s, but not the active coordinates, Theorem 3.2 gives $\widetilde{O}((Ks)^{1/4}T^{3/4})$ regret for Lipschitz MP convex objectives and $\widetilde{O}(\sqrt{KT})$ at a vertex, without assuming sparse gradients or a known face. Comparator-dependent Bernstein errors vanish at vertices; MP convexity removes the nonsmooth correction. The score–Hessian representation and subset-sum plus quadratic (SSQ) class identify applicable objectives (Propositions C.2 and 2.1).

2 Problem Formulation and Multinomial Policies

This section defines multinomial policies and MP convexity, including a nonseparable sufficient class. Supporting approximation and score–Hessian results appear in Appendices C.1 and C.2.

2.1 Problem statement

For $K \geq 2$, let $\Delta_{K-1} := \{x \in \mathbb{R}^K_{\geq 0} : \|x\|_1 = 1\}$. We study $(\mathrm{P}_1) : \min_{x\in\Delta_{K-1}} V(x)$ for an unknown convex objective $V : \Delta_{K-1} \to \mathbb{R}$ that is L-Lipschitz in ℓ_1, with known $L > 0$. At round t, the learner chooses $X_t \in \Delta_{K-1}$ and observes only $Y_t := V(X_t) + \epsilon_t$. Given the information $\mathcal{F}_{t-1}$ available before the round, $\mathbb{E}[\epsilon_t \mid X_t, \mathcal{F}_{t-1}] = 0$ and $|\epsilon_t| \leq \sigma$ almost surely, with known $\sigma > 0$. Write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$, $V^* := \min_{x\in\Delta_{K-1}} V(x)$, and $\mathcal{X}^* := \arg\min_{x\in\Delta_{K-1}} V(x)$.

We optimize a family of simplex-valued policies $\{\mathbb{P}_\theta : \theta \in \Omega\}$ with value $J(\theta) := \mathbb{E}_{X\sim\mathbb{P}_\theta}[V(X)]$. A general family may have $\inf_\theta J(\theta) > V^*$, and J need not inherit convexity or Lipschitz continuity from V. The multinomial family below preserves the Lipschitz constant and has infimum V^* across resolutions; MP convexity identifies when its value is convex in the policy mean. Section 3 gives one-point algorithms with finite-time guarantees.

For any $x^* \in \mathcal{X}^*$, define the pseudo-regret $R_T := \sum_{t=1}^T (V(X_t) - V^*)$. We seek $\mathbb{E}[R_T] = o(T)$ for fixed K and problem constants, while tracking its dependence on K for large-scale problems.

2.2 Multinomial Policies and Value Function

To construct a finite-resolution distribution over the simplex, let $n \geq 1$ be an integer and $m \in \Delta_{K-1}$. A count vector $N \in \mathcal{C}_{K,n} := \{z \in \mathbb{Z}^K_{\geq 0} : \sum_{i=1}^K z_i = n\}$ follows the *multinomial distribution*, written $N \sim \mathrm{Mult}(n, m)$, if $\mathbb{P}(N = z) = \frac{n!}{\prod_{i=1}^K z_i!} \prod_{i=1}^K m_i^{z_i}$ for $z \in \mathcal{C}_{K,n}$, with the convention $0^0 = 1$. Equivalently, N records the counts of n independent categorical draws with probability vector m. This definition includes boundary means: if $m_i = 0$, then $N_i = 0$ almost surely, and if $m = e_i$, then $N = ne_i$ almost surely.

Definition 2.1 (Multinomial policy)**.** Fix an integer resolution $n \geq 1$ and a policy mean $m \in \Delta_{K-1}$. Draw $N \sim \mathrm{Mult}(n, m)$ and set $X := N/n$. The distribution of the simplex-valued action X is called the *multinomial (MP) policy* with parameters (n, m). Its mean is $\mathbb{E}[X] = m$, and its covariance is

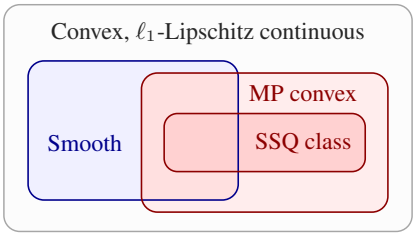


Figure 1: Objective classes in Section 2.2. Smooth and MP convex objectives form overlapping subsets of the convex, ℓ_1-Lipschitz continuous class. SSQ lies inside MP convex and crosses the smooth boundary. The outer, MP convex, and smooth sets correspond to cases (i), (ii), and (iii) of Theorem 3.1, respectively.

$\mathrm{Cov}(X) = \frac{1}{n}\big(\mathrm{Diag}(m) - mm^\top\big)$. For an objective function V, the associated MP value function is

$$F_{V,n}(m) := \mathbb{E}_{N\sim\mathrm{Mult}(n,m)}\left[V\left(\frac{N}{n}\right)\right]. \tag{1}$$

Thus $J(n,m) = F_{V,n}(m)$ for this policy family.

The finite sum (1) is the classical multivariate Bernstein polynomial of V on the simplex (Chang & Davis, 1984; Sauer, 1991). For fixed n, it is a polynomial in m and is continuous on the closed simplex.

The next lemma collects regularity properties used in the regret analysis. Its Lipschitz estimate follows from categorical coupling; resolution monotonicity and the failure to preserve ordinary convexity in several variables are classical Bernstein properties (Sauer, 1991).

Lemma 2.1 (Lipschitz continuity and resolution monotonicity)**.** *Let $V : \Delta_{K-1} \to \mathbb{R}$. If V is L-Lipschitz with respect to $\|\cdot\|_1$, then, for every integer $n \geq 1$ and all $m, m' \in \Delta_{K-1}$,*

$$|F_{V,n}(m) - F_{V,n}(m')| \leq L\|m - m'\|_1. \tag{2}$$

If V is convex, then, for every $m \in \Delta_{K-1}$ and integer $n \geq 1$,

$$V(m) \leq F_{V,n+1}(m) \leq F_{V,n}(m). \tag{3}$$

Even when V is convex, the policy value $F_{V,n}$ need not be convex in the mean m.

The proof is given in Appendix B.1. Example A.1 illustrates the failure of convexity preservation. For continuous convex V, Proposition C.3 gives a self-contained probabilistic proof of the classical Bernstein approximation limit $F_{V,n}(m) \downarrow V(m)$ as $n \to \infty$. Consequently, $\inf_{n\geq 1, m\in\Delta_{K-1}} F_{V,n}(m) = V^*$, which establishes equality of the policy and objective infima.

To obtain regret guarantees, we distinguish smoothness, MP convexity, and optimizer sparsity. Theorem 3.1 first treats all convex Lipschitz objectives, then sets the correction identically to zero under MP convexity and improves the rate under smoothness. Theorem 3.2 uses a known support-size bound for vertex and MP convex refinements; comparator-dependent MP convex refinements are given in Appendix E.4.

Figure 1 summarizes the relationship among the objective classes in this subsection.

We first specify smoothness for objectives satisfying the standing convexity and Lipschitz assumptions. Let $\mathcal{T} := \{h \in \mathbb{R}^K : \mathbf{1}^\top h = 0\}$ denote the tangent space of the simplex.

Definition 2.2 (Smoothness on the closed simplex)**.** A continuous function $V : \Delta_{K-1} \to \mathbb{R}$ is *β-smooth on the closed simplex relative to its affine hull* if it is differentiable on $\mathrm{relint}\,\Delta_{K-1}$ and its relative gradient $\nabla_{\mathrm{aff}} V : \mathrm{relint}\,\Delta_{K-1} \to \mathcal{T}$ extends to a map on Δ_{K-1} satisfying $\|\nabla_{\mathrm{aff}} V(x) - \nabla_{\mathrm{aff}} V(y)\|_2 \leq \beta\|x - y\|_2$ for all $x, y \in \Delta_{K-1}$. At a boundary point m, we use this extension to define $DV(m)[h] := \langle \nabla_{\mathrm{aff}} V(m), h\rangle$ for $h \in \mathcal{T}$. An objective is called *smooth* if it is β-smooth in this sense for some finite $\beta \geq 0$, and *nonsmooth* if no such finite β exists. Thus smooth and nonsmooth describe disjoint objective classes.

For a β-smooth objective, this definition implies the usual descent inequality $V(y) \leq V(x) + DV(x)[y-x] + \frac{\beta}{2}\|y-x\|_2^2$ for all $x, y \in \Delta_{K-1}$.

Proposition C.1 in Appendix C.1 gives the corresponding approximation bounds for MP function values and first derivatives, including at boundary points. These bounds support the smooth regret analysis without imposing MP convexity.

We next use ordinary convexity of the Bernstein transform at every resolution as a structural condition for the regret analysis. We call this property MP convexity. Convexity criteria for Bernstein polynomials on triangles and simplexes have been studied in approximation theory (Chang & Davis, 1984; Sauer, 1991; He, 1995); the definition below specifies the property needed by our policy analysis.

Definition 2.3 (MP convexity)**.** The objective function V is *multinomial policy convex* (*MP convex*) if, for every integer $n \geq 1$, the map $m \mapsto F_{V,n}(m)$ defining the MP value function is convex on Δ_{K-1}.

Proposition C.2 in Appendix C.2 expresses the standard tangent Hessian criterion as an exact likelihood-score identity. To obtain a condition that can be checked from the objective's form, the following SSQ class combines multinomial second moments and subset aggregation with classical univariate Bernstein convexity preservation (Bustamante, 2017). For $v \in \mathbb{R}^K$ and $S \subseteq \{1, \ldots, K\}$, write $v(S) := \sum_{i \in S} v_i$.

Proposition 2.1 (SSQ objectives are MP convex)**.** *Consider an objective function of the form $V(x) = c_0 + \ell^\top x + \frac{1}{2}x^\top Q x + \sum_{g=1}^{G} \phi_g(x(S_g))$, where $c_0 \in \mathbb{R}$, $\ell \in \mathbb{R}^K$, $\varnothing \neq S_g \subsetneq \{1, \ldots, K\}$, $Q = Q^\top$ satisfies $h^\top Q h \geq 0$ for every $h \in \mathcal{T}$, and each $\phi_g : [0,1] \to \mathbb{R}$ is convex. The sets S_g may overlap. We call objectives of this form* subset-sum plus quadratic *(SSQ). Every SSQ objective function is convex on the simplex and is MP convex.*

The proof of Proposition 2.1 is given in Appendix C.3. Thus SSQ objective functions provide an explicit class whose MP value functions are convex for every n, without requiring the objectives to be differentiable. The class includes separable convex objectives by taking singleton sets $S_g = \{g\}$ and $Q = 0$, while overlapping subset penalties and tangent-convex quadratics give genuinely nonseparable examples. For SSQ objective functions, the standing Lipschitz requirement holds, for example, when the ϕ_g are Lipschitz on $[0,1]$. The smooth SSQ subclass includes objectives for which each ϕ_g' is Lipschitz; if its Lipschitz constant is M_g, then one may take $\beta = \|Q\|_{\mathrm{op}} + \sum_g M_g |S_g|$.

The policy mean m directly parameterizes the simplex and preserves the objective's Lipschitz constant by Lemma 2.1. The integer n controls both action granularity and exploration variance: $X \in \Delta_{K-1} \cap n^{-1}\mathbb{Z}^K$ and $\mathrm{Cov}(X) = n^{-1}(\mathrm{Diag}(m) - mm^\top)$. Increasing n refines the action grid and concentrates the policy around m, while the finite-count representation yields the score estimator used in Section 3.

3 Multinomial Policy Mirror Descent

Under the objective and feedback assumptions of Section 2.1, we combine multinomial sampling with a one-point score estimator and mirror descent. Algorithm 1 covers smooth and nonsmooth convex objectives, including refinements when a support-size bound is known. The corresponding theorems state the smoothness, MP convexity, and support conditions for each guarantee. Throughout, *general* means that MP convexity is not required, and *nonsmooth* has the "not smooth" meaning specified in Definition 2.2.

3.1 A multinomial mirror-descent algorithm

Fix $T \geq 2$, let $u := \mathbf{1}/K$, and put $B := 2(L + \sigma)$. The first round plays $X_1 = u$ and stores $Y_{\mathrm{ref}} := Y_1$. For every subsequent observation, the Lipschitz and noise bounds give $|Y_{t+1} - Y_{\mathrm{ref}}| \leq B$. The reference round incurs regret $r_1 := V(u) - V^* \leq 2L(1 - 1/K)$.

Sampling and gradient estimation. For a fixed integer $n \geq 1$, initialize $m_1 = u$. In optimization round $t = 1, \ldots, T-1$, sample $N_t \mid \mathcal{F}_t \sim \mathrm{Mult}(n, m_t)$ and deploy $X_{t+1} := N_t/n$. After observing

Y_{t+1}, form

$$W_t := Y_{t+1} - Y_{\mathrm{ref}} \in [-B, B], \qquad g_{t,k} := W_t \left(\frac{N_{t,k}}{m_{t,k}} - n \right). \tag{4}$$

Conditionally on $\mathcal{F}_t$, this estimator is unbiased for the tangent derivatives of $F_{V,n}$, satisfies $\langle m_t, g_t \rangle = 0$, and obeys $|g_{t,k}| \le Bn/m_{t,k}$ and $\mathbb{E}[\sum_k m_{t,k} g_{t,k}^2 \mid \mathcal{F}_t] \le B^2 n(K-1)$; see Appendix E.2. The score is never clipped. Only one new noisy objective value is used per optimization round.

Mirror update. For $0 < \delta < 1/K$, let $\Delta_\delta := \{z \in \Delta_{K-1} : z_k \ge \delta \ \forall k\}$. For $m \in \Delta_\delta$ and $p \in \Delta_{K-1}$, use negative entropy $R_{\mathrm{F}}(m) := \sum_k m_k \log m_k$ and its KL divergence $D_{R_{\mathrm{F}}}(p, m) = \mathrm{KL}(p\|m) := \sum_{k:p_k>0} p_k \log(p_k/m_k)$, where $0 \log 0 := 0$. For every closed-simplex comparator p,

$$D_{R_{\mathrm{F}}}(p, u) = \log K + \sum_k p_k \log p_k \le \log K. \tag{5}$$

For general convex objectives, use the correction

$$\Phi = \Phi_n, \qquad \Phi_n(m) := -\frac{4L}{\sqrt{n}} \sum_k \sqrt{m_k}. \tag{6}$$

Choose $c \ge 0$ (default 1) independently of K, T and set $\rho_T := (\sqrt{2/\pi} - cT^{-1/8})_+$, where $(x)_+ := \max\{x, 0\}$. This weight is fixed throughout the run and may equal zero. For all other guarantees, including the vertex case, set $\Phi \equiv 0$; the value of ρ_T is then immaterial. All cases of Theorems 3.1 and 3.2 use

$$m_{t+1} = \operatorname*{arg\,min}_{z \in \Delta_\delta} \{\eta \langle g_t, z \rangle + \mathrm{KL}(z\|m_t) + \eta \rho_T \Phi(z)\}. \tag{7}$$

This is an exact constrained minimization; the floor applies to the policy mean, while a sampled action may lie on a simplex face. For $\Phi \equiv 0$, the unique minimizer has the form $m_{t+1,k} = \max\{\delta, \zeta m_{t,k} e^{-\eta g_{t,k}}\}$, where $\zeta > 0$ normalizes the coordinates. The corrected update is likewise uniquely defined on Δ_δ. In either case, coordinate clamping after an unconstrained update is not the prescribed operation. Negative entropy is the normalized Fisher potential of the multinomial policy; the geometry and natural-gradient interpretation are given in Appendices D and E.6.

Parameter choice. Fix tunable multipliers $n_0, \eta_0, \delta_0 > 0$ before the run, each independent of K, T and defaulting to one. They remain fixed within the run; δ_0 multiplies the actual probability floor and is not itself that floor. For the cases of Theorem 3.1, choose

$$n_{\mathrm{base}} := \begin{cases} \lceil (T/\log K)^{1/2} \rceil, & \text{cases (i), (ii),} \\ \lceil (T/(K \log K))^{1/3} \rceil, & \text{case (iii),} \end{cases} \qquad n := \lceil n_0 n_{\mathrm{base}} \rceil, \tag{8}$$

and set

$$\eta := \min \left\{ \frac{\eta_0}{2B} \sqrt{\frac{\log K}{n(K-1)T}}, \quad \frac{1}{2\delta_0 BnK} \right\}, \qquad \delta := \delta_0 \eta B n. \tag{9}$$

The step cap ensures $0 < K\delta \le 1/2$, so $u \in \Delta_\delta$. The outer ceiling makes n a positive integer also when $0 < n_0 < 1$; for integer n_0 it leaves the product unchanged. The resulting resolution can equal 1, so the default rule does not force nonvertex allocations. A requirement $n \ge 2$ can instead be imposed separately and analyzed with the arbitrary-resolution bounds below. All parameters, including ρ_T, and the function Φ remain fixed during the run. Constants in the tuned regret bounds may depend on the fixed multipliers, but not on K, T.

3.2 General convex regret

The next theorem first gives a guarantee for all convex Lipschitz objectives, then states the improvements under MP convexity and smoothness. All three cases require no support information and use the score interface, comparison inequalities, and Fisher stability bounds in Appendices E and F.

Theorem 3.1 (General convex regret)**.** *Suppose V is convex and L-Lipschitz continuous in ℓ_1 on the simplex, the feedback conditions in Section 2.1 hold, and $K, T \ge 2$. Algorithm 1 with parameters* (8)–(9) *satisfies:*

Algorithm 1 Floored multinomial mirror descent with a fixed baseline

1: **Input:** $T \geq 2$, bounds L, σ, fixed positive multipliers n_0, η_0, δ_0 (default 1), and a case of Theorem 3.1 or 3.2; for the latter, a known support-size upper bound $s \in \{1, \dots, K\}$; for case (i) of Theorem 3.1, $c \geq 0$ (default 1).
2: Set $u \leftarrow \mathbf{1}/K$.
3: For Theorem 3.1, choose n, η, δ by (8)–(9); for Theorem 3.2, use (10).
4: In case (i) of Theorem 3.1, set $\rho_T \leftarrow (\sqrt{2/\pi} - cT^{-1/8})_+$ and $\Phi \leftarrow \Phi_n$; otherwise set $\rho_T \leftarrow 0$ and $\Phi \leftarrow 0$.
5: Play $X_1 \leftarrow u$, observe Y_1, store $Y_{\mathrm{ref}} \leftarrow Y_1$, and initialize $m_1 \leftarrow u$.
6: **for** $t = 1, \dots, T-1$ **do**
7: Sample $N_t \sim \mathrm{Mult}(n, m_t)$, play $X_{t+1} \leftarrow N_t/n$, and observe Y_{t+1}.
8: Form the raw score g_t by (4); update m_{t+1} exactly by (7).
9: **end for**

(i) Convex and Lipschitz continuous. *Use $\Phi = \Phi_n$ with weight ρ_T. Then $\mathbb{E}[R_T] = \widetilde{O}(K^{1/2}T^{3/4})$.*

(ii) MP convex and Lipschitz continuous. *If V is MP convex, use $\Phi \equiv 0$. Then $\mathbb{E}[R_T] = \widetilde{O}(K^{1/2}T^{3/4})$.*

(iii) Convex, Lipschitz continuous, and smooth. *If V is β-smooth, use $\Phi \equiv 0$. Then $\mathbb{E}[R_T] = \widetilde{O}(K^{1/3}T^{2/3})$.*

The hidden constants depend only on L, σ, the fixed positive (n_0, η_0, δ_0), the fixed c in (i), and β in (iii), and are independent of K, T. Every fixed $\delta_0 > 0$, including $\delta_0 < 1$, is covered. The logarithmic factors are $(\log K)^{1/4}$ in (i) and (ii), and $(\log K)^{1/3}$ in (iii); the fixed multipliers introduce no additional $\log(TK)$ factor.

The proofs, including explicit logarithmic factors and bounds for arbitrary resolutions, are in Appendices E.4 and F. Case (i) applies whether or not V is smooth or MP convex. Case (ii) sets $\Phi \equiv 0$ under MP convexity without requiring smoothness, while case (iii) improves the rate under smoothness. Comparator-dependent MP convex refinements are given in Appendix E.4. The positive floor permits the signed centered score without clipping. Its analysis uses the feasible comparator $(1 - K\delta)x^* + \delta\mathbf{1}$ and an entropy-stability constant depending on δ_0.

3.3 Regret with a sparse optimizer

Suppose a known integer $s \in \{1, \dots, K\}$ bounds the support size of some optimizer x^*; its support coordinates are unknown. The sparse cases use $\Phi = 0$ and the update (7).

$$n := \begin{cases} 1, & s = 1, \\ \lceil n_0 n_{\mathrm{base}} \rceil, & s \geq 2, \end{cases}$$

$$n_{\mathrm{base}} := \left\lceil \frac{L}{B} \sqrt{\frac{(s-1)T}{(K-1)\log K}} \right\rceil \quad (s \geq 2), \tag{10}$$

$$\eta := \min\left\{ \frac{\eta_0}{2B} \sqrt{\frac{\log K}{n(K-1)T}}, \frac{1}{2\delta_0 BnK} \right\}, \qquad \delta := \delta_0 \eta B n.$$

Here $n_0, \eta_0, \delta_0 > 0$ are tunable real constants selected before the run, with defaults one and independent of K, T, s. The vertex rule $n = 1$ is not multiplied by n_0. In both branches $0 < K\delta \leq 1/2$.

Theorem 3.2 (Regret with a sparse optimizer)**.** *Suppose V is convex and L-Lipschitz continuous in ℓ_1 on the simplex, the feedback conditions in Section 2.1 hold, and $K, T \geq 2$. Let a known integer $s \in \{1, \dots, K\}$ satisfy $|\mathrm{supp}(x^*)| \leq s$ for some $x^* \in \mathcal{X}^*$. Run Algorithm 1 with the sparse parameter rule* (10) *in the indicated case.*

(i) Support bound *$s = 1$. Use $n = 1$ and $\Phi = 0$. Then $\mathbb{E}[R_T] = \widetilde{O}(\sqrt{KT})$.*

(ii) Support bound *$2 \leq s \leq K$. If V is MP convex, use $\Phi = 0$; no smoothness assumption is required. Then $\mathbb{E}[R_T] = \widetilde{O}((Ks)^{1/4}T^{3/4})$.*

The hidden constants depend only on L, σ and the fixed positive (n_0, η_0, δ_0), independently of K, T, s. In both bounds, $\widetilde{O}$ suppresses logarithmic factors in K and T.

4 Numerical Experiments

Figures 2 and 3 compare the proposed algorithm. The curves labeled *Proposed* use its smooth and MP-convex settings, respectively. The four benchmarks are negative-entropy mirror descent (Entropy MD), which combines a one-point tangent-sphere estimate with a KL update (Flaxman et al., 2005; Hazan et al., 2016); Euclidean projected gradient descent (Euclidean PGD), which uses the same estimate with Euclidean projection (Flaxman et al., 2005); simplex zeroth-order PGD (Simplex ZO-PGD), which uses Dirichlet perturbations and projection (Zrnic & Mazumdar, 2022); and projection-free BCO (PF-BCO), which uses a linear minimization oracle and Frank–Wolfe updates (Chen et al., 2019).

For $x \in \Delta_{K-1}$, the two test objectives are

$$h_j(x) = a_j + \sum_{k=1}^{K} \left(\frac{q_{j,k}}{2} x_k^2 + b_{j,k} x_k + e_{j,k} \log(x_k + c_{j,k}) \right), \tag{11}$$

$$f_j(x) = a_j + \sum_{k=1}^{K} \left(\frac{q_{j,k}}{2} x_k^2 + b_{j,k} x_k + e_{j,k} |x_k - c_{j,k}| \right). \tag{12}$$

Appendix H describes their generation. Feedback noise is $\mathrm{Unif}[-0.1, 0.1]$. The curves average 40, 80, and 160 test functions at $K = 10, 100, 1000$, respectively, with five seeds per function. Plotted metrics use true objective values at queried actions.

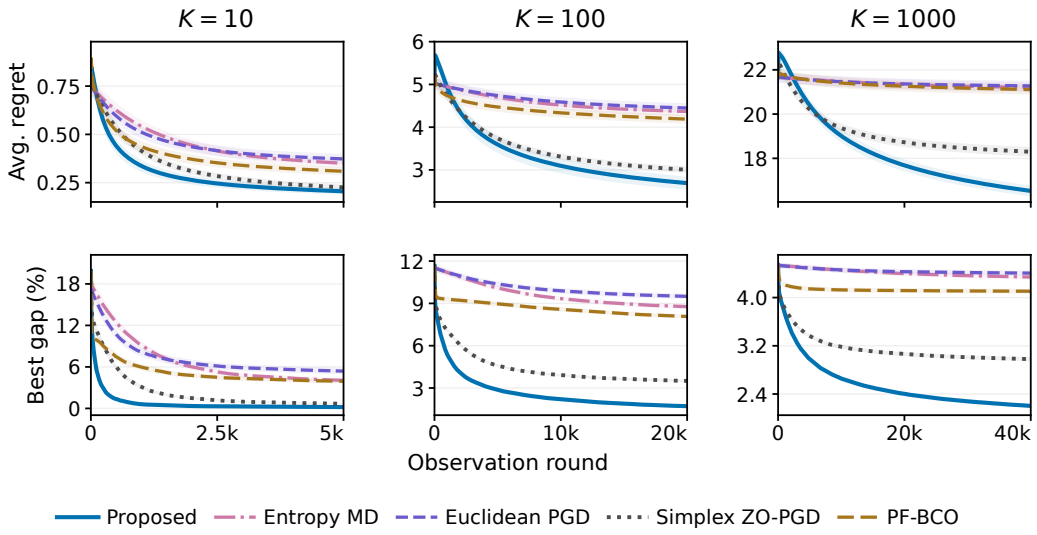


Figure 2: Smooth log-regularized quadratics. Columns use $(K, T) = (10, 5000), (100, 20000), (1000, 40000)$. The upper row shows average cumulative pseudo-regret R_t/t; the lower row shows the best-so-far gap as a percentage of the true optimum. Bands show one standard error across test-function means.

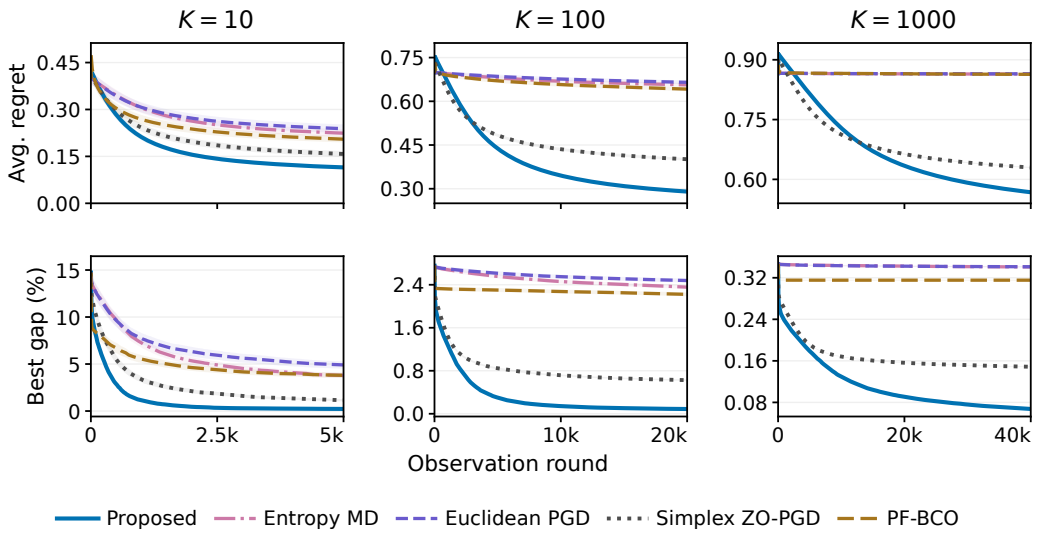


Figure 3: Nonsmooth weighted-ℓ_1 quadratics at the same (K, T) settings as Figure 2. The upper row shows R_t/t; the lower row shows the best-so-far percentage gap. Bands show one standard error across test-function means.

5 Conclusion

We developed a multinomial-policy method for stationary stochastic bandit convex optimization on the simplex. It combines a fixed-reference count score without clipping, a Fisher-geometry mirror update, and a positive probability floor. Under the stated Lipschitz and bounded-noise assumptions, the expected pseudo-regret is $\widetilde{O}(K^{1/2}T^{3/4})$ for convex objectives and $\widetilde{O}(K^{1/3}T^{2/3})$ under smoothness. Both bounds cover boundary optima. A known upper bound on optimizer support yields a $\widetilde{O}(\sqrt{KT})$ vertex guarantee and a support-sensitive MP-convex bound for larger sparse supports. Our score–Hessian characterization and SSQ sufficient class identify when the uncorrected update applies.

Future work will study adversarially varying objectives and extend multinomial-policy methods to feasible sets with additional constraints.

AI Use Statement. Generative AI tools were used to improve the readability and language of the manuscript. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content.

References

Alekh Agarwal, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013. doi: 10.1137/110850827.

Youhei Akimoto, Shinichi Shirakawa, Nozomu Yoshinari, Kento Uchida, Shota Saito, and Kouhei Nishida. Adaptive stochastic natural gradient method for one-shot neural architecture search. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 171–180. PMLR, 2019.

Laha Ale, Scott A King, Ning Zhang, Abdul Rahman Sattar, and Janahan Skandaraniyam. D3pg: Dirichlet ddpg for task partitioning and offloading with constrained hybrid action space in mobile-edge computing. *IEEE Internet of Things Journal*, 9(19):19260–19272, 2022.

Eric André and Guillaume Coqueret. Dirichlet policies for reinforced factor portfolios. *arXiv preprint arXiv:2011.05381*, 2020.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jorge Bustamante. *Bernstein Operators and Their Properties*. Birkhäuser, Cham, 2017. doi: 10.1007/978-3-319-55402-0.

Antonio Candelieri, Andrea Ponti, and Francesco Archetti. Bayesian optimization over the probability simplex. *Annals of Mathematics and Artificial Intelligence*, 93(1):77–91, 2025. doi: 10.1007/s10472-023-09883-w.

G. Z. Chang and P. J. Davis. The convexity of Bernstein polynomials over triangles. *Journal of Approximation Theory*, 40(1):11–28, 1984. doi: 10.1016/0021-9045(84)90132-1.

Lin Chen, Mingrui Zhang, and Amin Karbasi. Projection-free bandit convex optimization. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2047–2056. PMLR, 2019.

James Chok and Geoffrey M. Vasil. Optimization over a probability simplex. *Journal of Machine Learning Research*, 26(73):1–35, 2025.

Priyam Das. Recursive modified pattern search on high-dimensional simplex: A blackbox optimization technique. *Sankhya B*, 83(Suppl 2):440–483, 2021. doi: 10.1007/s13571-020-00236-9.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394. SIAM, 2005.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J. Mayo. Online newton method for bandit convex optimisation extended abstract. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1713–1714. PMLR, 2024.

Alexander V. Gasnikov, Anastasia A. Lagunovskaya, Ilnura N. Usmanova, and Fedor A. Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034, 2016. doi: 10.1134/S0005117916110114.

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

T. X. He. Shape criteria of Bernstein–Bézier polynomials over simplexes. *Computers & Mathematics with Applications*, 30(3–6):317–333, 1995. doi: 10.1016/0898-1221(95)00110-7.

Henry Lam and Junhui Zhang. Distributionally constrained black-box stochastic gradient estimation and optimization. *Operations Research*, 73(5):2680–2694, 2025. doi: 10.1287/opre.2021.0307.

Tor Lattimore. *Bandit Convex Optimisation*. Cambridge University Press, 2026. ISBN 9781009607599. doi: 10.1017/9781009607551.

Tor Lattimore and András György. Improved regret for zeroth-order stochastic convex bandits. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2938–2964. PMLR, 2021.

Jianming Ma, Zhanxiang Cao, and Yue Gao. Constrained dirichlet distribution policy: Guarantee zero constraint violation reinforcement learning for continuous robotic control. *IEEE Robotics and Automation Letters*, 9(12):11690–11697, 2024. doi: 10.1109/LRA.2024.3490392.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24325–24360. PMLR, 2023.

Jincheng Mei, Bo Dai, Alekh Agarwal, Sharan Vaswani, Anant Raj, Csaba Szepesvári, and Dale Schuurmans. Small steps no more: Global convergence of stochastic gradient bandits for arbitrary learning rates. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2371.

Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017.

Stephen U. Pasteris, Alberto Rumi, Fabio Vitale, and Nicolò Cesa-Bianchi. Sum-max submodular bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2323–2331. PMLR, 2024.

Michael Patriksson. A survey on the continuous nonlinear resource allocation problem. *European Journal of Operational Research*, 185(1):1–46, 2008.

Federico Pavesi, Antonio Candelieri, and Noémie Jaquier. Information theoretic bayesian optimization over the probability simplex. *arXiv preprint arXiv:2603.09793*, 2026. doi: 10.48550/arXiv.2603.09793.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015. doi: 10.1109/TIT.2015.2388583.

Samuel Robertson, Thang D. Chu, Bo Dai, Dale Schuurmans, Csaba Szepesvári, and Jincheng Mei. REINFORCE converges to optimal policies with any learning rate. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 636–642. PMLR, 2011.

Tomas Sauer. Multivariate Bernstein polynomials and convexity. *Computer Aided Geometric Design*, 8(6):465–478, 1991. doi: 10.1016/0167-8396(91)90031-6.

Yuan Tian, Minghao Han, Chetan Kulkarni, and Olga Fink. A prescriptive dirichlet power allocation policy with deep reinforcement learning. *Reliability Engineering & System Safety*, 224:108529, 2022.

Leizhen Wang, Peibo Duan, Cheng Lyu, Zewen Wang, Zhiqiang He, Nan Zheng, and Zhenliang Ma. Scalable and reliable multi-agent reinforcement learning for traffic assignment. *Communications in Transportation Research*, 5:100225, 2025. doi: 10.1016/j.commtr.2025.100225.

Zhiqiang Wei, Xianghao Yu, Derrick Wing Kwan Ng, and Robert Schober. Resource allocation for simultaneous wireless information and power transfer systems: A tutorial overview. *Proceedings of the IEEE*, 110(1):127–149, 2021.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(27):949–980, 2014.

Enlu Zhou and Shalabh Bhatnagar. Gradient-based adaptive stochastic search for simulation optimization over continuous space. *INFORMS Journal on Computing*, 30(1):154–167, 2018. doi: 10.1287/ijoc.2017.0771.

Enlu Zhou and Jiaqiao Hu. Gradient-based adaptive stochastic search for non-differentiable optimization. *IEEE Transactions on Automatic Control*, 59(7):1818–1832, 2014.

Tijana Zrnic and Eric Mazumdar. A note on zeroth-order optimization on the simplex. *arXiv preprint arXiv:2208.01185*, 2022. doi: 10.48550/arXiv.2208.01185.

A Convexity Need Not Be Preserved by a Multinomial Policy

The failure of ordinary convexity preservation by multivariate Bernstein operators is classical (Sauer, 1991). The following example illustrates it in the notation of our policy model.

Example A.1 (A convex objective with a nonconvex policy value)**.** On Δ_2, let $V(x) = |x_1 - x_2|$. This objective is convex and 1-Lipschitz with respect to $\|\cdot\|_1$, but it is not MP convex.

Proof. Enumerating the six count allocations for $n = 2$ gives

$$F_{V,2}(m) = m_1^2 + m_2^2 + m_1 m_3 + m_2 m_3 = m_1 + m_2 - 2m_1 m_2.$$

On the line $m = (r/2, r/2, 1 - r)$, $0 \le r \le 1$, this becomes $F_{V,2}(m) = r - r^2/2$, which is strictly concave. For example, its value at $r = 1/2$ is $3/8$, exceeding the average $1/4$ of its values at $r = 0$ and $r = 1$. Thus convexity of V does not imply MP convexity. □

B Proof of MP Value-Function Regularity

B.1 Proof of Lemma 2.1

Lipschitz continuity. Fix $m, m' \in \Delta_{K-1}$. For each $j \in \{1, \dots, n\}$, let (Z_j, Z_j') be a maximal coupling of categorical random vectors taking values in $\{e_1, \dots, e_K\}$, with respective probability vectors m and m'. Choose these pairs independently across j. Then

$$\mathbb{P}(Z_j \neq Z_j') = \frac{1}{2}\|m - m'\|_1. \tag{13}$$

Writing $X := n^{-1}\sum_{j=1}^n Z_j$ and $X' := n^{-1}\sum_{j=1}^n Z_j'$, the marginals are the multinomial policies with respective means m and m'. Moreover,

$$\begin{aligned}\mathbb{E}\|X - X'\|_1 &\le \frac{1}{n}\sum_{j=1}^n \mathbb{E}\|Z_j - Z_j'\|_1 \\ &= \frac{2}{n}\sum_{j=1}^n \mathbb{P}(Z_j \neq Z_j') = \|m - m'\|_1. \end{aligned} \tag{14}$$

Consequently,

$$|F_{V,n}(m) - F_{V,n}(m')| \le \mathbb{E}|V(X) - V(X')| \le L\|m - m'\|_1,$$

which proves (2), including boundary means.

Monotonicity in the resolution. We give a leave-one-out proof of the classical monotonicity of successive Bernstein approximations for convex functions (Sauer, 1991). Fix $m \in \Delta_{K-1}$ and draw independent categorical vectors $Z_1, \dots, Z_{n+1}$ with probability vector m. Let I be uniform on $\{1, \dots, n+1\}$ and independent of these vectors, and set

$$X_{n+1} := \frac{1}{n+1}\sum_{j=1}^{n+1} Z_j, \qquad Y_n := \frac{1}{n}\sum_{j \neq I} Z_j.$$

Since $\mathbb{E}[Y_n \mid Z_1, \dots, Z_{n+1}] = X_{n+1}$, conditional Jensen's inequality gives

$$V(X_{n+1}) \le \mathbb{E}[V(Y_n) \mid Z_1, \dots, Z_{n+1}].$$

The two marginals are multinomial policies with resolutions $n + 1$ and n, respectively. Taking expectations proves $F_{V,n+1}(m) \le F_{V,n}(m)$, and Jensen's inequality using $\mathbb{E}[X_{n+1}] = m$ gives the remaining lower bound in (3). This argument uses only convexity of V.

Q.E.D.

C Technical Results for Multinomial Policies

C.1 Smooth MP approximation and first-order consistency

Proposition C.1 (Smooth MP approximation and first-order consistency)**.** *Suppose, in addition to the standing convexity and Lipschitz assumptions, that V is β-smooth on the closed simplex relative to its affine hull in the sense of Definition 2.2. For an integer $n \geq 1$, define*

$$q_n := \frac{1}{n}. \tag{15}$$

Then, for every $m \in \Delta_{K-1}$,

$$0 \leq F_{V,n}(m) - V(m) \leq \frac{\beta}{2} q_n \big(1 - \|m\|_2^2\big), \tag{16}$$

and, for every $h \in \mathcal{T}$,

$$|DF_{V,n}(m)[h] - DV(m)[h]| \leq \beta \left(\sqrt{q_n}\, \|h\|_2 + q_n \|h\|_1\right). \tag{17}$$

Consequently, as $n \to \infty$, the MP value function converges uniformly to V on the closed simplex. Its first derivatives also converge uniformly to those of V over all unit Euclidean tangent directions, including at boundary points.

Proof. The value bound (16) follows from Lemma E.1, and the first-order bound (17) follows from Lemma D.2. Since $q_n \to 0$ as $n \to \infty$, the value bound gives uniform convergence on the closed simplex. For fixed K, the inequality $\|h\|_1 \leq \sqrt{K}\|h\|_2$ also makes the first-order bound uniform over all unit Euclidean tangent directions. Both cited lemmas include boundary means. □

C.2 Score–Hessian characterization of MP convexity

Convexity criteria for Bernstein polynomials are classical (Chang & Davis, 1984; Sauer, 1991; He, 1995). The representation below follows by twice differentiating the multinomial likelihood and applying the usual tangent-Hessian criterion for convexity. It expresses that criterion using the same scores as the policy-gradient estimator; it does not require differentiability of V.

For $N \sim \mathrm{Mult}(n, m)$ with $m \in \operatorname{relint} \Delta_{K-1}$, define

$$d_i(N; m) := \frac{N_i}{m_i}, \tag{18a}$$

$$r_i(N; m) := \frac{N_i}{m_i^2}. \tag{18b}$$

Write $d = (d_i)_{i=1}^K$, $r = (r_i)_{i=1}^K$, and

$$\Gamma_n^V(m) := \mathbb{E}\Big[\big(V(N/n) - F_{V,n}(m)\big)\,\big(dd^\top - \mathrm{Diag}(r)\big)\Big]. \tag{19}$$

The centered score is $d - n\mathbf{1}$; it has the same inner product with every tangent direction as d.

Proposition C.2 (Score–Hessian characterization)**.** *For every integer $n \geq 1$, every $m \in \operatorname{relint} \Delta_{K-1}$, and every $h \in \mathcal{T}$,*

$$\begin{aligned} D^2 F_{V,n}(m)[h,h] &= h^\top \Gamma_n^V(m) h \\ &= \mathbb{E}\Big[\big(V(N/n) - F_{V,n}(m)\big)\left((h^\top d)^2 - \sum_{i=1}^K h_i^2 r_i\right)\Big]. \end{aligned} \tag{20}$$

Consequently, V is MP convex if and only if, for every integer $n \geq 1$,

$$\Gamma_n^V(m) \succeq_{\mathcal{T}} 0 \qquad \text{for every } m \in \operatorname{relint} \Delta_{K-1}. \tag{21}$$

The observed-information correction $-\mathrm{Diag}(r)$ is part of the exact Hessian and cannot in general be omitted.

Proof. Fix $n \geq 1$, $m \in \operatorname{relint} \Delta_{K-1}$, and $h \in \mathcal{T}$. For a count vector N, the multinomial mass is

$$p_m(N) = \frac{n!}{\prod_i N_i!} \prod_{i=1}^{K} m_i^{N_i}.$$

Differentiation along the affine path $m + th$ gives

$$D_h \log p_m(N) = h^\top d(N; m), \tag{22a}$$

$$D_h^2 \log p_m(N) = -\sum_{i=1}^{K} h_i^2 r_i(N; m). \tag{22b}$$

The count space is finite, so termwise differentiation of (1) gives

$$D^2 F_{V,n}(m)[h, h] = \mathbb{E}\left[V(N/n) \left((h^\top d)^2 - \sum_{i=1}^{K} h_i^2 r_i \right) \right]. \tag{23}$$

Twice differentiating $\sum_N p_m(N) = 1$ along the same tangent path yields

$$\mathbb{E}\left[(h^\top d)^2 - \sum_{i=1}^{K} h_i^2 r_i \right] = 0. \tag{24}$$

Subtracting $F_{V,n}(m)$ inside the expectation proves (20). The same calculation with two tangent directions gives the bilinear form represented by (19). The polynomial $F_{V,n}$ is convex on the simplex if and only if its Hessian is positive semidefinite on $\mathcal{T}$ throughout the relative interior: the forward implication follows by restricting to interior line segments, and the reverse implication extends to boundary segments by continuity. This proves (21). □

C.3 Proof of Proposition 2.1

The SSQ sufficient condition combines standard multinomial moment and aggregation identities with univariate Bernstein convexity preservation (Bustamante, 2017). We give the elementary calculation to make this construction explicit, including overlapping subsets.

Fix an integer $n \geq 1$, let $N \sim \operatorname{Mult}(n, m)$, and set $X = N/n$. Affine terms remain affine after expectation. The second moment is

$$\mathbb{E}[XX^\top] = \left(1 - \frac{1}{n}\right) mm^\top + \frac{1}{n} \operatorname{Diag}(m). \tag{25}$$

Consequently,

$$\mathbb{E}\left[\frac{1}{2} X^\top Q X \right] = \frac{1 - 1/n}{2} m^\top Q m + \frac{1}{2n} \operatorname{diag}(Q)^\top m. \tag{26}$$

The second term is affine, while the tangent Hessian of the first term is $(1 - 1/n)Q \succeq_{\mathcal{T}} 0$.

For a subset-sum term, fix a convex $\phi : [0, 1] \to \mathbb{R}$ and a nonempty proper subset $S \subset \{1, \ldots, K\}$. Put $q = m(S)$. Aggregation of independent categorical draws gives $Y := N(S) \sim \operatorname{Bin}(n, q)$. Hence

$$\mathbb{E}[\phi(X(S))] = (\mathcal{B}_n \phi)(q) := \sum_{j=0}^{n} \phi(j/n) \binom{n}{j} q^j (1 - q)^{n-j}.$$

For $n \geq 2$, differentiating this polynomial twice yields

$$(\mathcal{B}_n \phi)''(q) = n(n-1) \sum_{j=0}^{n-2} [\phi((j+2)/n) - 2\phi((j+1)/n) + \phi(j/n)] \times \binom{n-2}{j} q^j (1-q)^{n-2-j} \geq 0,$$

because every discrete second difference of ϕ is nonnegative. For $n = 1$ the polynomial is affine. Thus $\mathcal{B}_n \phi$ is convex on $[0, 1]$, and its composition with the affine map $m \mapsto m(S)$ is convex on the simplex. Finite sums preserve convexity, so every SSQ objective is MP convex. The tangent-PSD condition on Q and convexity of each composition $x \mapsto \phi_g(x(S_g))$ also show directly that the original SSQ objective is convex. □

C.4 Limit as the action resolution increases

Proposition C.3 (Limit of the multinomial transform)**.** *Let $V : \Delta_{K-1} \to \mathbb{R}$ be convex and continuous. For every $m \in \Delta_{K-1}$,*

$$F_{V,n}(m) \downarrow V(m) \qquad \text{as } n \to \infty, \tag{27}$$

and the convergence is uniform on the closed simplex. In particular,

$$F_{V,n}(m) \geq V(m), \qquad \inf_{n\geq 1,\, m\in\Delta_{K-1}} F_{V,n}(m) = \min_{m\in\Delta_{K-1}} V(m). \tag{28}$$

Proof. Fix $m \in \Delta_{K-1}$. Let $(Z_j)_{j\geq 1}$ be independent categorical vectors with probability vector m, and put $X_n = n^{-1}\sum_{j=1}^{n} Z_j$. The strong law gives $X_n \to m$ almost surely. Since V is continuous on the compact simplex, it is bounded; dominated convergence therefore yields $F_{V,n}(m) = \mathbb{E}[V(X_n)] \to V(m)$. Lemma 2.1 gives monotonicity and the lower bound. Each $F_{V,n}$ is continuous, so Dini's theorem gives uniform convergence to the continuous function V on the compact simplex. Finally, evaluating at an optimizer of V and letting $n \to \infty$ proves the infimum identity. □

D Multinomial Fisher Information Geometry and First-Order Consistency

This appendix identifies the exact Fisher information of the multinomial policy used by the algorithms in Section 3. Its normalized potential is negative entropy, whose divergence remains finite for boundary comparators and interior base points. We also prove a boundary-uniform first-order approximation result that does not require MP convexity.

D.1 Exact multinomial Fisher geometry

Fix an integer $n \geq 1$. For $N \sim \mathrm{Mult}(n, m)$ and $m \in \operatorname{relint}\Delta_{K-1}$, define

$$S_k(N; m) := \frac{N_k}{m_k} - n. \tag{29}$$

This representative is centered both in expectation and, pathwise, under m: $\mathbb{E}S = 0$ and $\sum_k m_k S_k = 0$. For a tangent direction $h \in \mathcal{T} := \{h : \mathbf{1}^\top h = 0\}$, the scalar likelihood score is $s_h(N; m) = \langle S(N; m), h\rangle$.

Definition D.1 (Normalized multinomial Fisher geometry)**.** For the fixed resolution n, define

$$\phi_n(x) := nx\log x, \qquad \Psi_n(m) := n\sum_k m_k \log m_k, \qquad R_{\mathrm{F}}(m) := \frac{\Psi_n(m)}{n} = \sum_k m_k \log m_k, \tag{30}$$

where $0\log 0 := 0$. For $R := R_{\mathrm{F}}$, an interior base point m, and any $p \in \Delta_{K-1}$, the Bregman divergence is

$$D_R(p, m) := R(p) - R(m) - \langle \nabla R(m), p - m\rangle.$$

The normalized potential R is independent of n; Ψ_n retains the scale of the exact count Fisher information.

Lemma D.1 (Multinomial Fisher geometry)**.** *The scalar information coefficient and the exact tangent information are*

$$\omega_n(x) := \frac{n}{x}, \tag{31}$$

$$\mathcal{I}_n(m)[h, h] := \mathbb{E}[s_h(N; m)^2] = n\sum_{k=1}^{K} \frac{h_k^2}{m_k}. \tag{32}$$

In particular the lower and upper information scales coincide:

$$\underline{\iota}_n = \bar{\iota}_n = n, \qquad \kappa_n = 1, \tag{33}$$

$$\mathcal{I}_n(m)[h, h] = \underline{\iota}_n \sum_k \frac{h_k^2}{m_k} = \bar{\iota}_n \sum_k \frac{h_k^2}{m_k}. \tag{34}$$

The restriction of $\nabla^2\Psi_n(m)$ to $\mathcal{T}$ equals $\mathcal{I}_n(m)$. The potentials in Definition D.1 are continuous on the closed simplex, and $\phi_n'(x) \to -\infty$ as $x \downarrow 0$. For every $p \in \Delta_{K-1}$ and interior m,

$$D_{\Psi_n}(p, m) = n\,\mathrm{KL}(p\|m), \qquad D_{R_{\mathrm{F}}}(p, m) = \mathrm{KL}(p\|m). \tag{35}$$

Consequently, with $u = \mathbf{1}/K$ and $\mathsf{H}(p) := -\sum_k p_k \log p_k$,

$$D_{R_{\mathrm{F}}}(p, u) = \log K - \mathsf{H}(p) \le \log K. \tag{36}$$

Proof. The multinomial mass is $n! \prod_k m_k^{N_k} / \prod_k N_k!$. Differentiating its logarithm in a tangent direction gives $\sum_k N_k h_k / m_k = s_h$. The covariance formula

$$\mathrm{Cov}(N/n) = \frac{1}{n}\big(\mathrm{Diag}(m) - mm^\top\big) \tag{37}$$

and $\sum_k h_k = 0$ imply

$$\mathbb{E}[s_h^2] = n\left(\sum_k \frac{h_k^2}{m_k} - (\mathbf{1}^\top h)^2\right) = n\sum_k \frac{h_k^2}{m_k}.$$

Equivalently, the expected negative coordinate Hessian is

$$\mathbb{E}\frac{N_k}{m_k^2} = \frac{n}{m_k}, \tag{38}$$

which directly gives the scalar coefficient. The Hessian of Ψ_n is $n\,\mathrm{Diag}(1/m_k)$. Expanding its Bregman divergence and using $\sum_k p_k = \sum_k m_k = 1$ proves the KL identities for interior p. Continuity extends them to boundary p. Finally, $\mathrm{KL}(p\|u) = \log K - \mathsf{H}(p)$ proves the budget. □

D.2 Boundary-uniform first-order multinomial consistency

Lemma D.2 (Boundary-uniform C^1 consistency of the multinomial transform)**.** *Suppose V is convex and β-smooth in the sense of Definition 2.2. No MP convexity is required. For $n \ge 1$, put*

$$q_n := \frac{1}{n}, \tag{39}$$

$$\delta_n^{\mathrm{MP}}(V) := \sup_{m\in\Delta_{K-1}} \sup_{\substack{h\in\mathcal{T}\\ \|h\|_2\le 1}} |DF_{V,n}(m)[h] - DV(m)[h]|. \tag{40}$$

Then every $m \in \Delta_{K-1}$ and $h \in \mathcal{T}$ satisfy

$$|DF_{V,n}(m)[h] - DV(m)[h]| \le \beta\left(\frac{\|h\|_2}{\sqrt{n}} + \frac{\|h\|_1}{n}\right), \tag{41}$$

and hence

$$\delta_n^{\mathrm{MP}}(V) \le \beta\left(n^{-1/2} + \frac{\sqrt{K}}{n}\right). \tag{42}$$

The suprema include boundary means and all affine-hull tangent directions.

Proof. The relative smoothness assumptions give

$$\|\nabla_{\mathrm{aff}} V(x) - \nabla_{\mathrm{aff}} V(y)\|_2 \le \beta\|x - y\|_2, \tag{43}$$

$$0 \le V(y) - V(x) - DV(x)[y - x] \le \frac{\beta}{2}\|y - x\|_2^2. \tag{44}$$

Represent the multinomial sample as $X = n^{-1}\sum_{j=1}^n e_{I_j}$ for independent $I_j \sim \mathrm{Categorical}(m)$. Differentiating this finite product distribution gives

$$DF_{V,n}(m)[h] = \sum_{j=1}^n \sum_{i=1}^K h_i \mathbb{E}V(X_{-j} + e_i/n), \qquad X_{-j} := \frac{1}{n}\sum_{\ell\ne j} e_{I_\ell}. \tag{45}$$

Put $z_j = X_{-j} + m/n$. Taylor expansion at z_j gives

$$V(X_{-j} + e_i/n) = V(z_j) + \frac{1}{n} DV(z_j)[e_i - m] + r_{j,i}, \qquad |r_{j,i}| \le \frac{\beta}{2n^2} \|e_i - m\|_2^2. \tag{46}$$

The constant terms cancel since $\sum_i h_i = 0$. Therefore

$$\begin{aligned} DF_{V,n}(m)[h] - DV(m)[h] = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E} \langle \nabla_{\mathrm{aff}} V(z_j) - \nabla_{\mathrm{aff}} V(m), h \rangle \\ + \sum_{j=1}^{n} \sum_{i=1}^{K} h_i \mathbb{E} r_{j,i}. \end{aligned} \tag{47}$$

Independence yields

$$\mathbb{E} \|z_j - m\|_2^2 = \frac{n-1}{n^2} (1 - \|m\|_2^2) \le \frac{1}{n}. \tag{48}$$

Thus the first line is bounded in absolute value by $\beta \|h\|_2 / \sqrt{n}$. Since $\|e_i - m\|_2^2 \le 2$, the second line is bounded by $\beta \|h\|_1 / n$. This proves the mixed-norm bound in the relative interior; $\|h\|_1 \le \sqrt{K} \|h\|_2$ gives the stated defect bound. The transform $F_{V,n}$ is a finite polynomial in m, so its affine first derivatives are continuous at the boundary. The relative gradient of V is continuous by smoothness. Passing to a boundary limit proves both bounds for every affine-hull tangent direction, including directions that add mass to a zero coordinate. □

The proof uses only a Lipschitz relative gradient and an absolute quadratic Taylor remainder; convexity is needed in subsequent objective-gap comparisons. No lower bound on the mean coordinates is required, and $n \to \infty$ suffices for uniform first-order consistency for fixed K.

E Proofs for Multinomial Fisher Mirror Descent

This appendix proves the score interface for Algorithm 1, as well as the objective bridges, stability inequality, and regret guarantees for the matched-Fisher update (7) with $\Phi = 0$, used in cases (ii) and (iii) of Theorem 3.1 and cases (i) and (ii) of Theorem 3.2. The general convex Lipschitz analysis for case (i) of Theorem 3.1, with applied correction $\rho_T \Phi_n$, is given in Appendix F. Throughout, R denotes the normalized exact multinomial Fisher potential $R_{\mathrm{F}} = \Psi_n / n$ in Definition D.1, matched to the fixed sampling resolution n. Its boundary properties and KL bounds follow from Lemma D.1.

Notation. Write $H := T - 1$, $d := K - 1$, $B := 2L + 2\sigma$, and $q_n := 1/n$. For a comparator $p \in \Delta_{K-1}$, put

$$C_1(p) := \sum_{k=1}^{K} \sqrt{p_k (1 - p_k)} \le \sqrt{K - 1}, \qquad C_2(p) := 1 - \|p\|_2^2 \le 1 - 1/K.$$

Both quantities vanish at vertices. For later use, the ℓ_1-diameter of the simplex and Lipschitz continuity give the pathwise range bound

$$0 \le R_T \le \sum_{t=1}^{T} L \|X_t - x^*\|_1 \le 2LT. \tag{49}$$

The proofs use query-round indexing: for $j = 2, \dots, T$, define

$$\rho^{(j)} := m_{j-1}, \qquad N^{(j)} := N_{j-1}, \qquad \widehat{g}_j := g_{j-1}, \qquad W_j^{\mathrm{qry}} := Y_j - Y_{\mathrm{ref}}.$$

The terminal mean is $\rho^{(T+1)} := m_T$. Thus $X_j = N^{(j)}/n$, and the main-text estimator becomes

$$S_{j,k} := \frac{N_k^{(j)}}{\rho_k^{(j)}} - n, \qquad \widetilde{S}_j := S_j - \langle \rho^{(j)}, S_j \rangle \mathbf{1} = S_j, \qquad \widehat{g}_j = W_j^{\mathrm{qry}} S_j.$$

Within this appendix, W_t abbreviates W_t^{qry} and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$; both are indexed by the query round t.

E.1 SMOOTH VALUE APPROXIMATION

Lemma E.1 (Smooth multinomial approximation)**.** *Suppose V is convex and β-smooth on the closed simplex relative to its affine hull. For an integer $n \geq 1$ and $m \in \Delta_{K-1}$,*

$$0 \leq F_{V,n}(m) - V(m) \leq \frac{\beta}{2} q_n (1 - \|m\|_2^2), \qquad q_n := \frac{1}{n}. \tag{50}$$

Proof. For $X = N/n$ with $N \sim \mathrm{Mult}(n, m)$, the standard multinomial covariance identity is

$$\mathrm{Cov}(X) = q_n \big(\mathrm{Diag}(m) - m m^\top\big). \tag{51}$$

Jensen's inequality gives $F_{V,n}(m) \geq V(m)$. The smoothness inequality at m, followed by $\mathbb{E}[X] = m$, gives

$$\begin{aligned} F_{V,n}(m) - V(m) &\leq \mathbb{E}\left[DV(m)[X - m] + \frac{\beta}{2}\|X - m\|_2^2 \right] \\ &= \frac{\beta}{2} \operatorname{tr} \mathrm{Cov}(X) = \frac{\beta}{2} q_n (1 - \|m\|_2^2). \end{aligned}$$

□

E.2 PROOF OF THE SCORE INTERFACE

Lemma E.2 (Multinomial score interface)**.** *Conditionally on $\mathcal{F}_{t-1}$, for every $h \in \mathcal{T} := \{h : \mathbf{1}^\top h = 0\}$,*

$$\mathbb{E}_t \langle \widehat{g}_t, h \rangle = DF_{V,n}(\rho^{(t)})[h], \tag{52}$$

$$\langle \rho^{(t)}, \widehat{g}_t \rangle = 0, \qquad |\widehat{g}_{t,k}| \leq \frac{Bn}{\rho_k^{(t)}}, \tag{53}$$

$$\mathbb{E}_t \sum_{k=1}^K \rho_k^{(t)} \widehat{g}_{t,k}^2 \leq B^2 n d. \tag{54}$$

Proof of Lemma E.2. Fix $t \geq 2$ and condition on $\mathcal{F}_{t-1}$. Write $\rho = \rho^{(t)}$ and let $p_\rho(N)$ denote the mass of $N \sim \mathrm{Mult}(n, \rho)$. Differentiating the finite probability mass function along $h \in \mathcal{T}$ gives

$$D \log p_\rho(N)[h] = \langle S_t, h \rangle. \tag{55}$$

Consequently, the likelihood-score identity and conditional centering of the current noise imply

$$\mathbb{E}_t[Y_t \langle S_t, h \rangle] = DF_{V,n}(\rho)[h].$$

The score has zero conditional mean. Since Y_{ref} is predictable, multiplying the score by the centered feedback gives (52).

For the pathwise properties, the multinomial score is already centered:

$$\widetilde{S}_{t,k} = S_{t,k} = \frac{N_k^{(t)}}{\rho_k} - n, \qquad \sum_k \rho_k S_{t,k} = 0. \tag{56}$$

The feedback bound gives $|W_t| \leq B$. Since $|N_k^{(t)}/\rho_k - n| \leq n/\rho_k$, the centered score can have either sign, satisfies $|\widehat{g}_{t,k}| \leq Bn/\rho_k$, and obeys $\langle \rho, \widehat{g}_t \rangle = 0$.

It remains to prove the weighted second-moment bound. Let $v_k := e_k - \rho \in \mathcal{T}$. Then $\widetilde{S}_{t,k} = \langle S_t, v_k \rangle$. The exact Fisher-information identity in (34) gives

$$\mathbb{E}_t[\widetilde{S}_{t,k}^2] = n \sum_{j=1}^K \frac{(v_k)_j^2}{\rho_j} = n \frac{1 - \rho_k}{\rho_k}.$$

Multiplying by ρ_k, summing over k, and using $W_t^2 \leq B^2$ yields

$$\mathbb{E}_t \sum_{k=1}^K \rho_k \widehat{g}_{t,k}^2 \leq B^2 n \sum_{k=1}^K (1 - \rho_k) = B^2 n (K - 1).$$

□

E.3 PROOFS OF THE OBJECTIVE BRIDGES

Lemma E.3 (Lipschitz bridge under MP convexity)**.** *Suppose V is MP convex. For every interior m and every $p \in \Delta_{K-1}$,*

$$F_{V,n}(m) - V(p) \le DF_{V,n}(m)[m-p] + LC_1(p)\sqrt{q_n}. \tag{57}$$

Proof of Lemma E.3. MP convexity makes $F_{V,n}$ convex, so

$$F_{V,n}(m) - F_{V,n}(p) \le DF_{V,n}(m)[m-p].$$

For $X \sim \mathrm{Mult}(n,p)/n$, Lipschitz continuity, (51), and Jensen's inequality give

$$\begin{aligned} F_{V,n}(p) - V(p) &\le L\mathbb{E}\|X-p\|_1 \\ &\le L\sum_{k=1}^{K}\sqrt{\mathrm{Var}(X_k)} = LC_1(p)\sqrt{q_n}. \end{aligned}$$

Adding the two displays proves (57). Finally,

$$C_1(p)^2 \le K\sum_{k=1}^{K} p_k(1-p_k) = K(1-\|p\|_2^2) \le K-1,$$

which proves the stated uniform bound. □

Lemma E.4 (Smooth replace-one bridge)**.** *Suppose V is convex and β-smooth in the sense of Definition 2.2. No MP convexity condition is required. For every interior m and every $p \in \Delta_{K-1}$,*

$$\begin{aligned} F_{V,n}(m) - V(p) &\le DF_{V,n}(m)[m-p] + \beta q_n(1-m^\top p) \\ &\le DF_{V,n}(m)[m-p] + \beta q_n. \end{aligned} \tag{58}$$

Proof of Lemma E.4. Draw independent labels $I_j \sim \mathrm{Categorical}(m)$, $j = 1,\dots,n$, and write the multinomial action as

$$X = \frac{1}{n}\sum_{j=1}^{n} e_{I_j}. \tag{59}$$

Every observation has deterministic weight $1/n$, so

$$\sum_{j=1}^{n}\left(\frac{1}{n}\right)^2 = \frac{1}{n} = q_n. \tag{60}$$

Independently draw replacement labels $J_j \sim \mathrm{Categorical}(p)$ and set

$$X^{(j)} := X - e_{I_j}/n + e_{J_j}/n.$$

Differentiation of the finite product distribution of the labels gives

$$DF_{V,n}(m)[m-p] = \mathbb{E}\sum_{j=1}^{n}\big[V(X) - V(X^{(j)})\big]. \tag{61}$$

For completeness, the contribution of the jth label to the derivative is multiplied by $(m_k - p_k)/m_k$ on the event $I_j = k$; the subtracted term is exactly the expectation obtained by replacing I_j with an independent draw from p.

Smoothness applied from X to $X^{(j)}$ gives

$$\begin{aligned} V(X) - V(X^{(j)}) \ge \frac{1}{n}DV(X)[e_{I_j} - e_{J_j}] \\ - \frac{\beta}{2}\frac{1}{n^2}\|e_{I_j} - e_{J_j}\|_2^2. \end{aligned}$$

After summing and averaging over the independent replacement labels, the first-order term becomes

$$\mathbb{E}\, DV(X)[X-p] \geq \mathbb{E}[V(X)-V(p)] = F_{V,n}(m) - V(p),$$

where convexity of V gives the inequality. Moreover,

$$\mathbb{E}\|e_{I_j} - e_{J_j}\|_2^2 = 2(1 - m^\top p).$$

Combining the last three displays with (60) yields

$$DF_{V,n}(m)[m-p] \geq F_{V,n}(m) - V(p) - \beta q_n(1 - m^\top p),$$

which is equivalent to (58). □

Lemma E.5 (Smooth bridge under MP convexity)**.** *Suppose V is MP convex and β-smooth in the sense of Definition 2.2. For every interior m and every $p \in \Delta_{K-1}$,*

$$\begin{aligned} F_{V,n}(m) - V(p) \leq\ & DF_{V,n}(m)[m-p] \\ & + \min\left\{ LC_1(p)\sqrt{q_n},\ \frac{\beta}{2} q_n C_2(p) \right\}. \end{aligned} \tag{62}$$

Proof of Lemma E.5. Convexity of $F_{V,n}$ under MP convexity gives

$$F_{V,n}(m) - V(p) \leq DF_{V,n}(m)[m-p] + F_{V,n}(p) - V(p).$$

The approximation term is at most $LC_1(p)\sqrt{q_n}$ by the proof of Lemma E.3, and at most $\frac{\beta}{2} q_n C_2(p)$ by Lemma E.1. Taking their minimum proves (62). Both terms depend only on the comparator and vanish when it is a vertex. □

E.4 Stability and the master theorem

Write $H := T-1$, $d := K-1$, $D := \log K$, and $B := 2(L+\sigma)$. Fix positive multipliers n_0, η_0, δ_0 independently of K, T. For any integer $n \geq 1$ and any budget $D_0 \geq D$, the floored entropy parameters are

$$\eta(D_0) := \min\left\{ \frac{\eta_0}{2B}\sqrt{\frac{D_0}{ndT}},\ \frac{1}{2\delta_0 BnK} \right\}, \qquad \delta(D_0) := \delta_0 \eta(D_0) Bn. \tag{63}$$

The main-text choice (9) is $\eta_{\mathrm{F}}^{\mathrm{arb}} := \eta(D)$ for arbitrary fixed n; its associated absolute floor is $\delta(D)$.

$$\eta_{\mathrm{F}}^{\mathrm{arb}} = \eta(\log K). \tag{64}$$

In particular $0 < K\delta \leq 1/2$ and $u \in \Delta_\delta$, where $\Delta_\delta := \{z \in \Delta_{K-1} : z_k \geq \delta\ \forall k\}$. The input δ_0 is a multiplier, not the absolute floor.

Lemma E.6 (Extended entropy stability for the centered score)**.** *For $a > 0$ put*

$$c(a) := \frac{e^a - 1 - a}{a^2} = \int_0^1 (1-v)e^{av}\, dv. \tag{65}$$

For $m \in \Delta_\delta$ and a centered realized score g from (4)*, with $\delta = \delta_0 \eta Bn$, define $\Gamma_\eta(g,m) := \eta\langle m, g\rangle + \log\sum_k m_k e^{-\eta g_k}$. Then $|\eta g_k| \leq 1/\delta_0$ and*

$$0 \leq \Gamma_\eta(g,m) \leq c(1/\delta_0)\eta^2 \sum_k m_k g_k^2. \tag{66}$$

For the exact constrained Fisher step with $\Phi = 0$, every $p \in \Delta_\delta$ also satisfies

$$\langle g, m-p\rangle \leq \frac{\mathrm{KL}(p\|m) - \mathrm{KL}(p\|m^+)}{\eta} + \frac{\Gamma_\eta(g,m)}{\eta}. \tag{67}$$

Proof. The score interface gives $|g_k| \leq Bn/m_k \leq Bn/\delta$. For every real $x \geq -a$, Taylor's formula yields $e^{-x} \leq 1 - x + c(a)x^2$. Averaging this inequality with weights m_k, using $\log y \leq y - 1$, proves the upper bound on Γ; Jensen's inequality proves its nonnegativity. This argument covers $0 < \delta_0 < 1$, for which $|\eta g_k|$ may exceed one.

The variational inequality for the minimizer on the compact convex set Δ_δ gives the KL three-point bound. Its stability remainder is at most

$$\sup_{z\in\Delta_{K-1}} \{\eta\langle g, m-z\rangle - \mathrm{KL}(z\|m)\} = \Gamma_\eta(g,m),$$

which proves (67), including when some coordinates of m^+ equal the floor. □

Lemma E.7 (Floored centered master regret bound)**.** *Fix $n \geq 1$, $\eta > 0$, $\delta = \delta_0\eta Bn \leq 1/(2K)$, and use* (7) *with $\Phi = 0$ on Δ_δ. Let $p \in \Delta_\delta$ and $q \in \Delta_{K-1}$ be deterministic. Suppose, at every $m \in \Delta_\delta$, the bridge $F_{V,n}(m) - V(q) \leq DF_{V,n}(m)[m-p] + e_n(m,p,q)$ holds. Then*

$$\begin{aligned}\mathbb{E}[R_T] \leq r_1 + H[V(q) - V^*] + \frac{\mathrm{KL}(p\|u)}{\eta} + c(1/\delta_0)\eta B^2 ndH \\ + \mathbb{E}\sum_{t=2}^{T} e_n(\rho^{(t)}, p, q).\end{aligned} \tag{68}$$

In particular, when $q = x^*$, the bound is

$$\mathbb{E}[R_T] \leq r_1 + \frac{\mathrm{KL}(p\|u)}{\eta} + c(1/\delta_0)\eta B^2 ndH + \mathbb{E}\sum_{t=2}^{T} e_n(\rho^{(t)}, p, x^*). \tag{69}$$

Proof. Condition on the history before each sample. The score identity (52) replaces the policy derivative by the expected tangent score. Apply (67) and (66); the conditional second moment is at most B^2nd by (54). Summing telescopes the KL terms, and their terminal value is nonnegative. The first query contributes r_1. Compactness of Δ_δ and bounded feedback justify all expectations. □

Lemma E.8 (A feasible comparator near any optimizer)**.** *For $p \in \Delta_{K-1}$ define*

$$p^\delta := (1 - K\delta)p + \delta\mathbf{1} = (1 - K\delta)p + K\delta u. \tag{70}$$

Then $p^\delta \in \Delta_\delta$, $\mathrm{KL}(p^\delta\|u) \leq D$, and $\|p^\delta - p\|_1 \leq 2K\delta$. Moreover, for any two policy means $a, b \in \Delta_{K-1}$ and every integer $n \geq 1$, $|F_{V,n}(a) - F_{V,n}(b)| \leq L\|a - b\|_1$.

Proof. The first assertions follow from the definition, simplex diameter, and (5). For the last assertion, maximally couple each of the n categorical labels drawn from a and b. Each pair differs with probability $\|a - b\|_1/2$ and has ℓ_1 distance two when it differs. The expected distance between the empirical averages is at most $\|a - b\|_1$; apply Lipschitz continuity of V. □

Put

$$C(\eta_0, \delta_0) := \frac{2}{\eta_0} + \frac{\eta_0}{2}\big[c(1/\delta_0) + 2\delta_0\big] + 2\sqrt{\delta_0}. \tag{71}$$

This quantity is finite for every fixed positive pair. It is not uniform as $\delta_0 \downarrow 0$.

Lemma E.9 (Balanced arbitrary-resolution bound)**.** *Suppose that, for an optimizer p^*, the bridge at its feasible version p^δ has the uniform error $e_n + 2LK\delta$, where $e_n \geq 0$ is independent of the iterate. Use $\eta = \eta(D)$ and $\delta = \delta(D)$ from* (63)*. Then*

$$\mathbb{E}[R_T] \leq r_1 + \frac{D}{\eta} + [c(1/\delta_0) + 2\delta_0]\eta B^2 ndT + Te_n, \tag{72}$$

$$\leq \min\{2LT,\ r_1 + C(\eta_0, \delta_0)B\sqrt{DndT} + Te_n\}. \tag{73}$$

Proof. Apply Lemma E.7 with $q = p^*$. The floor charge obeys $2LK\delta T = 2LK\delta_0\eta BnT \leq 2\delta_0\eta B^2 ndT$, using $B \geq 2L$ and $K \leq 2d$. For $S = \sqrt{DndT}$, the parameter rule gives

$$D/\eta \leq (2B/\eta_0)S + 2\delta_0 BnKD, \qquad [c(1/\delta_0) + 2\delta_0]\eta B^2 ndT \leq \frac{\eta_0}{2}[c(1/\delta_0) + 2\delta_0]BS.$$

Finally combine with the pathwise bound $R_T \leq 2LT \leq BT$ and

$$\min\{BT, 2\delta_0 BnKD\} \leq B\sqrt{2\delta_0 nKDT} \leq 2\sqrt{\delta_0}BS. \tag{74}$$

This proves (73). □

Lemma E.10 (Regret bounds for floored multinomial Fisher mirror descent)**.** *For arbitrary fixed integer $n \geq 1$, use the centered score, $\Phi = 0$, and the parameters in* (63) *with $D_0 = D$. With $C = C(\eta_0, \delta_0)$, the following hold.*

(i) If V is β-smooth, without MP convexity, then

$$\mathbb{E}[R_T] \leq \min\{2LT, r_1 + CB\sqrt{DndT} + \beta T/n\}. \tag{75}$$

(ii) If V is also MP convex, then

$$\mathbb{E}[R_T] \leq \min\{2LT, r_1 + CB\sqrt{DndT} + T\min\{LC_1(p^*)/\sqrt{n}, \beta C_2(p^*)/(2n)\}\}. \tag{76}$$

(iii) If V is MP convex without smoothness, then

$$\mathbb{E}[R_T] \leq \min\{2LT, r_1 + CB\sqrt{DndT} + LC_1(p^*)T/\sqrt{n}\}. \tag{77}$$

Here p^* is any deterministic optimizer, possibly on the boundary.

Proof. For (i), apply Lemma E.4 at p^δ and add $V(p^\delta) - V(p^*) \leq 2LK\delta$. For (ii) and (iii), MP convexity gives $F_{V,n}(m) - F_{V,n}(p^\delta) \leq DF_{V,n}(m)[m - p^\delta]$. Lemma E.8 transfers the policy value from p^δ to p^* at cost $2LK\delta$. The sampling error $F_{V,n}(p^*) - V(p^*)$ is bounded by the terms in Lemmas E.5 and E.3. Now apply (73). □

Proof of Theorem 3.1. Write $D := \log K$, $d := K - 1$, $C := C(\eta_0, \delta_0)$, and $c_+ := 2n_0 + 1$. For every $b \geq 1$,

$$n_0 b \leq \lceil n_0 \lceil b \rceil \rceil \leq c_+ b. \tag{78}$$

The multipliers are fixed positive constants, independent of K, T. Case (i) is established in Appendix F.3. Case (ii) follows from Proposition E.1(b) below. For case (iii), if $T \geq KD$, put $b = (T/(KD))^{1/3} \geq 1$. The chosen resolution obeys $n_0 b \leq n \leq c_+ b$, so (75) gives

$$\mathbb{E}[R_T] \leq \left[(1 + C\sqrt{c_+})B + \beta/n_0\right](KD)^{1/3}T^{2/3}.$$

Here $r_1 \leq B$ is absorbed because the target factor is at least one. For $T < KD$, the pathwise bound yields $\mathbb{E}[R_T] \leq BT \leq B(KD)^{1/3}T^{2/3}$. Thus case (iii) has the asserted $(\log K)^{1/3}$ factor, including short horizons. Case (i), proved in Appendix F, has the asserted $(\log K)^{1/4}$ factor. □

Proposition E.1 (Regret refinements under MP convexity)**.** *Suppose V is MP convex and the standing objective and feedback conditions hold. Run Algorithm 1 with its centered score, exact floored update, and fixed positive n_0, η_0, δ_0. If V is also β-smooth, the case (iii) resolution gives*

$$\mathbb{E}[R_T] = O\big(K^{1/3}(\log K)^{1/3}T^{2/3}\big). \tag{79}$$

Without smoothness, the case (ii) resolution gives

$$\mathbb{E}[R_T] = O\big(\sqrt{K}(\log K)^{1/4}T^{3/4}\big). \tag{80}$$

The constants may depend on fixed $L, \sigma, n_0, \eta_0, \delta_0$ and, for the smooth result, β, but not on K, T.

Proof. Write $D := \log K$, $d := K - 1$, $C := C(\eta_0, \delta_0)$, and $c_+ := 2n_0 + 1$. The arbitrary-resolution comparator refinement is exactly (76):

$$\mathbb{E}[R_T] \leq r_1 + CB\sqrt{DndT} + T\min\{LC_1(p^*)/\sqrt{n}, \beta C_2(p^*)/(2n)\}. \tag{81}$$

For the smooth rate, use $C_2(p^*) \leq 1$ and repeat case (iii) of Theorem 3.1 with $\beta/2$ in place of β. For the nonsmooth MP rate, if $T \geq D$, set $b = \sqrt{T/D} \geq 1$. Then $n_0 b \leq n \leq c_+ b$ and (77) gives

$$\mathbb{E}[R_T] \leq \left[\left(1 + C\sqrt{c_+}\right)B + \frac{L}{\sqrt{n_0}}\right]D^{1/4}\sqrt{K}\,T^{3/4}.$$

We used $C_1(p^*) \leq \sqrt{K-1}$ and $r_1 \leq B$. For $T < D$, the pathwise bound gives $\mathbb{E}[R_T] \leq BT \leq BD^{1/4}\sqrt{K}\,T^{3/4}$. Both rates hold at every horizon without a horizon-dependent resolution multiplier. □

E.5 Implementation of the matched Fisher update

Algorithm 1 uses the normalized information geometry of its multinomial sampling policy in every case:

$$R_{\mathrm{F}}(m) = \frac{\Psi_n(m)}{n} = \sum_{k=1}^{K} r(m_k), \qquad r(x) := x \log x. \tag{82}$$

Its scalar curvature is $r''(x) = 1/x$. Lemma D.1 gives, for interior m and every boundary or interior comparator p,

$$D_{R_{\mathrm{F}}}(p, m) = \mathrm{KL}(p\|m), \qquad \sup_{p \in \Delta_{K-1}} D_{R_{\mathrm{F}}}(p, u) = \log K. \tag{83}$$

Thus the budget $D_0 = \log K$ and the normalized potential are independent of the resolution n and remain fixed throughout a run.

Fisher mirror update with $\Phi = 0$. For $m \in \Delta_\delta$ and finite realized score g, the Karush–Kuhn–Tucker conditions are

$$\log m_k^+ = \log m_k - \eta g_k - \lambda + \tau_k, \quad \tau_k \geq 0, \quad \tau_k(m_k^+ - \delta) = 0, \quad \sum_k m_k^+ = 1. \tag{84}$$

Consequently the exact constrained update has the form

$$m_k^+ = \max\{\delta, \zeta\, m_k \exp(-\eta g_k)\}, \qquad \sum_k m_k^+ = 1,$$

where the scalar $\zeta > 0$ is chosen by normalization. Strict convexity gives uniqueness. This is a constrained minimization, not clamping an unconstrained update after normalization.

Implicit correction for general convex objectives. In case (i) of Theorem 3.1, Algorithm 1 uses the applied correction $\rho_T \Phi_n$, with $\rho_T \geq 0$ fixed within the run, and the same Fisher potential; hence (7) has coordinate conditions

$$r'(m_k^+) - \frac{2\eta\rho_T L\sqrt{q_n}}{\sqrt{m_k^+}} = r'(m_k) - \eta g_k - \lambda + \tau_k, \quad \tau_k \geq 0, \quad \tau_k(m_k^+ - \delta) = 0, \quad \sum_{k=1}^{K} m_k^+ = 1. \tag{85}$$

The derivative of the left-hand side with respect to $m_k^+ \in (0, 1)$ is

$$r''(m_k^+) + \eta\rho_T L\sqrt{q_n}\,(m_k^+)^{-3/2} > 0.$$

The scalar map $f(x) := r'(x) - 2\eta\rho_T L\sqrt{q_n}/\sqrt{x}$ is strictly increasing on $x > 0$, tending to $-\infty$ as $x \downarrow 0$ and $+\infty$ as $x \to \infty$. When $\rho_T > 0$, its inverse can be written using the principal branch W_0 of the Lambert function, defined by $W_0(a)e^{W_0(a)} = a$ for $a \geq 0$. Set

$$\kappa := \eta\rho_T L\sqrt{q_n} = \frac{\eta\rho_T L}{\sqrt{n}} > 0.$$

For coordinates above the floor, $\tau_k = 0$. Since $r'(x) = \log x + 1$, rearranging (85) on those coordinates and setting $y_k = \sqrt{m_k^+}$, and then setting $w_k = \kappa/y_k$ gives

$$w_k e^{w_k} = \frac{\kappa}{\sqrt{m_k}} \exp\left(\frac{\eta g_k + \lambda}{2}\right).$$

Consequently, when $\rho_T > 0$, each coordinate has the closed form

$$m_k^+(\lambda) = \max\left\{\delta, \left[\frac{\kappa}{W_0\left(\frac{\kappa}{\sqrt{m_k}} \exp\left(\frac{\eta g_k + \lambda}{2}\right)\right)}\right]^2\right\}. \tag{86}$$

When $\rho_T = 0$, the correction vanishes and the corresponding formula is $m_k^+(\lambda) = \max\{\delta, m_k \exp(-\eta g_k - \lambda)\}$. The common multiplier is the unique solution of the scalar normalization equation

$$\sum_{k=1}^{K} m_k^+(\lambda) = 1. \tag{87}$$

Indeed, every $m_k^+(\lambda)$ is continuous and nonincreasing in λ, and at the normalized solution at least one coordinate exceeds δ, because $K\delta \le 1/2$. Thus one monotone scalar root solve suffices. The resulting minimizer is unique and belongs to Δ_δ; some coordinates may equal the floor.

E.6 Natural-gradient interpretation in dual coordinates

We make the mirror/natural-gradient correspondence (Raskutti & Mukherjee, 2015) explicit for the multinomial Fisher potential. Use intrinsic coordinates $\theta = (m_1, \dots, m_{K-1})$ on $\Theta := \{\theta > 0 : \mathbf{1}^\top\theta < 1\}$, with $m(\theta) = (\theta, 1 - \mathbf{1}^\top\theta)$ and Jacobian $\mathsf{E} := \big[I_{K-1}; -\mathbf{1}^\top\big]$. Set $\widehat{R}(\theta) := R(m(\theta))$ and $\mathcal{J}(\theta) := F_{V,n}(m(\theta))$. The exact Fisher matrix in these coordinates and the mirror Hessian are

$$\mathsf{G}(\theta) = \mathsf{E}^\top \operatorname{diag}\big(n/m_k(\theta)\big)\mathsf{E}, \qquad \mathsf{H}(\theta) := \nabla^2\widehat{R}(\theta) = \frac{\mathsf{G}(\theta)}{n} \succ 0. \tag{88}$$

Define the dual coordinate $\xi := \nabla\widehat{R}(\theta)$. The inverse $\theta = \theta(\xi)$ is well defined at every positive mean, including points on the floor boundary; its Jacobian in the ambient open simplex is $\mathrm{d}\theta/\mathrm{d}\xi = \mathsf{H}(\theta)^{-1}$.

Fix a round, suppress its index, and write $\widehat{g}_\theta := \mathsf{E}^\top\widehat{g}_t$ at the current $m = \rho^{(t)}$. For $\Phi = 0$, let $\tau \ge 0$ be the vector of active-floor KKT multipliers in (84). Multiplying those conditions by $\mathsf{E}^\top$ gives

$$\xi^+ = \xi - \eta\widehat{g}_\theta + \mathsf{E}^\top\tau. \tag{89}$$

By Lemma E.2, $\mathbb{E}_t[\widehat{g}_\theta] = \nabla_\theta\mathcal{J}(\theta)$. For $\widetilde{J}(\xi) := \mathcal{J}(\theta(\xi))$, the transformed estimator $\widehat{g}_\xi := \mathsf{H}(\theta)^{-1}\widehat{g}_\theta$ therefore satisfies

$$\mathbb{E}_t[\widehat{g}_\xi] = \mathsf{H}(\theta)^{-1}\nabla_\theta\mathcal{J}(\theta) = \nabla_\xi\widetilde{J}(\xi).$$

The normalized Fisher metric transforms to $\mathsf{H}_\xi = (\mathrm{d}\theta/\mathrm{d}\xi)^\top\mathsf{H}(\theta)(\mathrm{d}\theta/\mathrm{d}\xi) = \mathsf{H}(\theta)^{-1}$. Consequently,

$$\mathsf{H}_\xi^{-1}\widehat{g}_\xi = \widehat{g}_\theta, \qquad \xi^+ = \xi - \eta\mathsf{H}_\xi^{-1}\widehat{g}_\xi + \mathsf{E}^\top\tau. \tag{90}$$

When the floor is inactive, $\tau = 0$ and this is exactly a stochastic natural-gradient step on the policy value in dual coordinates. The multinomial Fisher matrix in this chart is $\mathsf{G}_\xi = n\mathsf{H}_\xi$, so the unconstrained part uses step size ηn with the unnormalized metric. These algebraic identities do not require convexity of $F_{V,n}$.

In the intrinsic policy-mean coordinates, Taylor expansion of the inverse mirror map instead gives

$$\theta^+ = \theta - \eta\mathsf{H}(\theta)^{-1}\widehat{g}_\theta + O(\eta^2) = \theta - \eta n\mathsf{G}(\theta)^{-1}\widehat{g}_\theta + O(\eta^2). \tag{91}$$

The expansion is for a fixed interior point with an inactive floor, fixed realized estimator, and fixed n as $\eta \to 0$; its remainder is not asserted to be uniform near an active floor. The constrained update also includes the KKT term in (89).

Finally, in case (i) of Theorem 3.1, put $\widehat{\Phi}_T(\theta) := \rho_T\Phi_n(m(\theta))$. With the active-floor multiplier τ, the optimality conditions give

$$\xi^+ = \xi - \eta\widehat{g}_\theta - \eta\nabla_\theta\widehat{\Phi}_T(\theta^+) + \mathsf{E}^\top\tau. \tag{92}$$

When $\rho_T > 0$, the additional term is evaluated implicitly at the new iterate, yielding a composite Fisher mirror step. At $\rho_T = 0$, (92) reduces to the uncorrected natural-gradient step in (90).

F Proofs for General Convex Lipschitz Objectives

This appendix proves case (i) of Theorem 3.1 for Algorithm 1 with the matched Fisher potential $R = R_{\mathrm{F}}$ and the applied correction $\rho_T\Phi_n$. Throughout, write $F := F_{V,n}$, $q := q_n$, $B := 2L + 2\sigma$, $a_* := \sqrt{2/\pi}$, and Φ_n as in (6). The nonnegative coefficient $\rho_T = (a_* - cT^{-1/8})_+$ is selected before

the run and held fixed throughout it, including when $\rho_T = 0$. The parameters n, η and the matched Fisher potential $R = \Psi_n/n$ are fixed within each finite-horizon run; Definition D.1 in Appendix D specifies R. Only convexity and the standing ℓ_1-Lipschitz and feedback assumptions are imposed on V; neither smoothness nor MP convexity is used. As in Appendix E, we use query-round indexing: $\rho^{(j)} = m_{j-1}$ and $\widehat{g}_j = g_{j-1}$ for $j = 2, \dots, T$ in terms of the algorithm's optimization rounds. We also use that appendix's query-round conventions for W_t and $\mathbb{E}_t$, and its definitions of C_1 and q_n.

F.1 A FIRST-ORDER INEQUALITY FOR CONVEX LIPSCHITZ OBJECTIVES

The multinomial value function need not be convex. The base correction Φ_n is finite on the closed simplex, with interior gradient $[\nabla\Phi_n(m)]_k = -2L\sqrt{q}/\sqrt{m_k}$. The comparison below isolates the gradient term absorbed by the implicit update, without additional feedback. The algorithm imposes the positive floor in (9).

Lemma F.1 (Binomial absolute deviation)**.** *If $Z \sim \mathrm{Bin}(n, p)$ and $\mu = np$, then*

$$\mathbb{E}|Z - \mu| \le a_*\sqrt{\mu} + \frac{1}{4}. \tag{93}$$

Proof. The claim is immediate for $p = 0, 1$. For $0 < p < 1$, put $k := \lfloor \mu \rfloor \le n - 1$. Summing the positive part of $Z - \mu$ and using $j\binom{n}{j} = n\binom{n-1}{j-1}$ gives

$$\mathbb{E}|Z - \mu| = 2\mu(1 - p)\Pr\{\mathrm{Bin}(n - 1, p) = k\}.$$

For fixed k, the logarithm of the ratio of $(1 - p)\Pr\{\mathrm{Bin}(n - 1, p) = k\}$ to $e^{-\mu}\mu^k/k!$ equals

$$H(p) = \sum_{j=1}^{k} \log(1 - j/n) + (n - k)\log(1 - p) + np.$$

For $p \ge k/n$, one has $H'(p) = (k - np)/(1 - p) \le 0$. When $k \ge 1$, a right Riemann sum for the decreasing function $\log(1 - x/n)$ yields

$$\sum_{j=1}^{k} \log(1 - j/n) \le \int_0^k \log(1 - x/n)\,\mathrm{d}x = -k - (n - k)\log(1 - k/n),$$

so $H(k/n) \le 0$. For $k = 0$, $\log(1 - p) \le -p$ gives $H(p) \le 0$ directly. Therefore, writing

$$A_k := \frac{2e^{-(k+1/2)}(k + 1/2)^{k+1/2}}{k!},$$

we obtain $\mathbb{E}|Z - \mu|/\sqrt{\mu} \le A_k$ by maximizing $2e^{-\mu}\mu^{k+1/2}/k!$ over $k \le \mu < k + 1$. Concavity of log shows

$$\log\frac{A_{k+1}}{A_k} = \int_{k+1/2}^{k+3/2} \log x\,\mathrm{d}x - \log(k + 1) \le 0.$$

Hence $A_k \le A_0 = \sqrt{2/e}$. For $k \ge 1$, the Stirling bound $k! \ge \sqrt{2\pi k}(k/e)^k$ gives

$$\log\frac{A_k}{a_*} \le (k + 1/2)\log(1 + 1/(2k)) - 1/2 \le \frac{1}{8k}.$$

Using $e^{1/(8k)} \le (1 - 1/(8k))^{-1} \le 1 + 1/(7k)$, we get

$$(A_k - a_*)\sqrt{\mu} \le \frac{a_*\sqrt{k + 1}}{7k} \le \frac{\sqrt{2}}{7} < \frac{1}{4}.$$

For $k = 0$, $\mu < 1$ and $(A_0 - a_*)\sqrt{\mu} \le \sqrt{2/e} - a_* < 1/4$. These two cases prove (93). □

Lemma F.2 (Bounded subgradients at sampled boundary points)**.** *There exists a finite convex L-Lipschitz function $\widetilde{V} : \mathbb{R}^K \to \mathbb{R}$ agreeing with V on the simplex. At every $x \in \mathbb{R}^K$, a subgradient $v(x) \in \partial\widetilde{V}(x)$ exists and satisfies $\|v(x)\|_\infty \le L$.*

Proof. Define

$$\widetilde{V}(x) := \inf_{y \in \Delta_{K-1}} \{V(y) + L\|x - y\|_1\}.$$

The infimum is finite and attained by compactness. Combining minimizers at two points proves convexity, and the triangle inequality proves the global Lipschitz bound. For x in the simplex, Lipschitz continuity of V gives $V(y) + L\|x - y\|_1 \geq V(x)$, with equality at $y = x$. A finite convex function on $\mathbb{R}^K$ has a subgradient everywhere. Applying its subgradient inequality and the Lipschitz bound at $x \pm te_k$, $t > 0$, gives $|v_k(x)| \leq L$. We choose one such subgradient at each point of the finite multinomial action grid and use the same choice for every direction below; thus measurability presents no issue. □

Lemma F.3 (Raw and scaled-correction first-order inequalities)**.** *For every $m \in \operatorname{relint} \Delta_{K-1}$ and $p \in \Delta_{K-1}$,*

$$\begin{aligned} F(m) - V(p) \leq DF(m)[m - p] + L\sqrt{q}\, C_1(m) \\ + 2L\sqrt{q} \sum_{k=1}^{K} p_k \sqrt{\frac{1 - m_k}{m_k}}. \end{aligned} \tag{94}$$

If additionally $m \in \Delta_\delta$ for $\delta > 0$, then, for every $\rho \geq 0$, put $s_\rho := (a_* - \rho)_+$ and

$$e_\rho(n, \delta) := (1 + 2\rho)L\sqrt{K/n} + \frac{2Ls_\rho}{\sqrt{n\delta}} + \frac{L}{2n\delta}. \tag{95}$$

The scaled comparison is

$$F(m) - V(p) \leq DF(m)[m - p] + \rho\langle \nabla\Phi_n(m), m - p\rangle + e_\rho(n, \delta). \tag{96}$$

Neither statement requires convexity of F or of $F + \rho\Phi_n$.

Proof. Use the independent-label representation in (59): $I_1, \ldots, I_n$ are independent with law $\mathrm{Categorical}(m)$ and $X = n^{-1}\sum_{\ell=1}^{n} e_{I_\ell}$. The finite product distribution makes F a polynomial in m, even when V is nonsmooth. For $i \neq j$, differentiation of the finite product distribution of the labels gives

$$DF(m)[e_i - e_j] = \mathbb{E}\sum_{\ell=1}^{n} \frac{\mathbf{1}\{I_\ell = j\}}{m_j}\big[V(X + (e_i - e_j)/n) - V(X)\big]. \tag{97}$$

Indeed, conditional on the other labels, the ℓth contribution to the derivative replaces that label by i and subtracts its replacement by j; conditioning the displayed summand on $I_\ell = j$ gives exactly this difference. The shifted value is evaluated only on $\{I_\ell = j\}$, where its argument belongs to the simplex; the entire summand is zero otherwise.

Select $v(X)$ from Lemma F.2. Convexity at X bounds (97) below by

$$\mathbb{E}\sum_{\ell=1}^{n} \frac{\mathbf{1}\{I_\ell = j\}}{nm_j}\langle v(X), e_i - e_j\rangle = \mathbb{E}\left[\frac{X_j}{m_j}\langle v(X), e_i - e_j\rangle\right].$$

By (51),

$$\mathbb{E}\left|\frac{X_j}{m_j} - 1\right| \leq \sqrt{q\frac{1 - m_j}{m_j}}.$$

Since $|\langle v(X), e_i - e_j\rangle| \leq 2L$, it follows that

$$DF(m)[e_i - e_j] \geq \mathbb{E}\langle v(X), e_i - e_j\rangle - 2L\sqrt{q}\sqrt{\frac{1 - m_j}{m_j}}.$$

For $i = j$ this inequality also holds. Multiply by $m_i p_j$, sum over i, j, and use $\sum_{i,j} m_i p_j (e_i - e_j) = m - p$ to obtain

$$DF(m)[m - p] \geq \mathbb{E}\langle v(X), m - p\rangle - 2L\sqrt{q}\sum_j p_j \sqrt{\frac{1 - m_j}{m_j}}. \tag{98}$$

On the other hand, convexity, the bounded subgradient, and the covariance identity imply

$$\begin{aligned} F(m) - V(p) &\le \mathbb{E}\langle v(X), X - p\rangle \\ &\le \mathbb{E}\langle v(X), m - p\rangle + L\mathbb{E}\|X - m\|_1 \\ &\le \mathbb{E}\langle v(X), m - p\rangle + L\sqrt{q}\, C_1(m). \end{aligned}$$

Combining this with (98) proves (94), without assuming independence of $v(X)$ and X.

For the scaled comparison, retain the absolute deviation before the Cauchy–Schwarz step leading to (98). Since $N_j \sim \mathrm{Bin}(n, m_j)$, Lemma F.1 gives

$$\mathbb{E}\left|\frac{X_j}{m_j} - 1\right| \le \frac{a_*}{\sqrt{nm_j}} + \frac{1}{4nm_j}.$$

Together with $C_1(m) \le \sqrt{K}$, this sharpens the raw comparison to

$$F(m) - V(p) \le DF(m)[m - p] + L\sqrt{K/n} + \frac{2La_*}{\sqrt{n}} \sum_j \frac{p_j}{\sqrt{m_j}} + \frac{L}{2n} \sum_j \frac{p_j}{m_j}.$$

The exact identity

$$\langle \nabla\Phi_n(m), m - p\rangle = \frac{2L}{\sqrt{n}} \left(\sum_j \frac{p_j}{\sqrt{m_j}} - \sum_j \sqrt{m_j} \right)$$

then yields (96): use $\sum_j p_j/\sqrt{m_j} \le \delta^{-1/2}$, $\sum_j p_j/m_j \le \delta^{-1}$, and $\sum_j \sqrt{m_j} \le \sqrt{K}$. Subgradients are used only in this proof, not by Algorithm 1. □

F.2 Well-defined floored updates and implicit mirror stability

Lemma F.4 (Well-defined updates and finite-horizon integrability)**.** *Let $0 < \delta \le 1/(2K)$ and $m \in \Delta_\delta$. For every finite g and $\eta > 0$, the corrected objective $z \mapsto \eta\langle g, z\rangle + \mathrm{KL}(z\|m) + \eta\rho_T\Phi_n(z)$ has a unique minimizer on Δ_δ. All quantities in the finite-horizon regret argument are integrable.*

Proof. The floored simplex is compact and convex, and the objective is continuous and strictly convex there. Thus it has a unique minimizer, which may have coordinates equal to δ. The fixed positive floor, finite count space, and bounded residual $|Y_t - Y_{\mathrm{ref}}| \le B$ give bounded scores, potential derivatives, and divergences at every round of a fixed run. This justifies the conditional expectations. □

Lemma F.5 (Composite stability with a signed score)**.** *Use query-round indices, a fixed n, η, δ with $\delta = \delta_0\eta Bn \le 1/(2K)$, and $m_2 = u$. Suppose $m_t \in \Delta_\delta$ and g_t is the centered score. Define m_{t+1} by* (7) *with $\Phi = \Phi_n$ and correction multiplier ρ_T. Then every $p \in \Delta_\delta$ satisfies*

$$\sum_{t=2}^{T} \langle g_t + \rho_T\nabla\Phi_n(m_t), m_t - p\rangle \le \frac{\mathrm{KL}(p\|u)}{\eta} + c_{\mathrm{st}}\eta \sum_{t=2}^{T}\sum_k m_{t,k}g_{t,k}^2 + \rho_T\frac{2L}{\sqrt{n}}(\sqrt{K} - 1). \quad (99)$$

Here $c_{\mathrm{st}} := c(1/\delta_0)$ is the stability constant from (65)*; it is distinct from the horizon-weight input c.*

Proof. Set $\widetilde{R} = R_{\mathrm{F}} + \eta\rho_T\Phi_n$ and $h_t = g_t + \rho_T\nabla\Phi_n(m_t)$. The exact constrained update minimizes $\eta\langle h_t, z\rangle + D_{\widetilde{R}}(z, m_t)$ on Δ_δ. Its variational inequality gives the Bregman three-point estimate, even when the minimizer lies on the floor:

$$\begin{aligned} \eta\langle h_t, m_t - p\rangle \le{}& D_{\widetilde{R}}(p, m_t) - D_{\widetilde{R}}(p, m_{t+1}) \\ &+ \eta\langle h_t, m_t - m_{t+1}\rangle - D_{\widetilde{R}}(m_{t+1}, m_t). \end{aligned} \quad (100)$$

For any feasible z, the last line decomposes as

$$\begin{aligned} \eta\langle h_t, m_t - z\rangle - D_{\widetilde{R}}(z, m_t) = {}& \eta\langle g_t, m_t - z\rangle - \mathrm{KL}(z\|m_t) \\ &+ \eta\rho_T[\Phi_n(m_t) - \Phi_n(z)]. \end{aligned} \quad (101)$$

The supremum of the first two terms over the full simplex is $\Gamma_\eta(g_t, m_t)$, which is bounded by (66) using $|\eta g_{t,k}| \le 1/\delta_0$. Sum (100) and telescope. At the terminal mean $a = m_{T+1}$ the correction is ρ_T times

$$D_{\Phi_n}(p,u) - D_{\Phi_n}(p,a) + \Phi_n(u) - \Phi_n(a) = \langle \nabla\Phi_n(a), p-a\rangle \le \frac{2L}{\sqrt{n}}(\sqrt{K}-1),$$

because $\nabla\Phi_n(u)$ is constant across coordinates, $\sum_k \sqrt{a_k} \le \sqrt{K}$, and $\sum_k p_k/\sqrt{a_k} \ge 1$. Drop the nonpositive terminal KL term to prove the claim. □

F.3 Proof of the general convex regret bound

Proof of Theorem 3.1, case (i). Write $H := T-1$, $d := K-1$, $D := \log K$, $B := 2(L+\sigma)$, and $C := C(\eta_0, \delta_0)$ from (71). Let $p = x^*$ be a deterministic optimizer and use its feasible version p^δ from (70). The scaled bridge (96), applied at p^δ, and $V(p^\delta) - V(p) \le 2LK\delta$ give

$$F_{V,n}(m) - V^* \le DF_{V,n}(m)[m - p^\delta] + \rho_T\langle \nabla\Phi_n(m), m - p^\delta\rangle + e_{\rho_T}(n,\delta) + 2LK\delta.$$

The centered score identity (52), Lemma F.5, and the conditional moment bound (54) yield, for every fixed integer $n \ge 1$ and the prescribed η, δ,

$$\begin{aligned}\mathbb{E}[R_T] \le r_1 + \frac{D}{\eta} + c_{\mathrm{st}}\eta B^2 ndH + He_{\rho_T}(n,\delta) + 2LK\delta H \\ + \rho_T C_\Phi,\end{aligned} \tag{102}$$

where $C_\Phi := 2L(\sqrt{K}-1)/\sqrt{n}$. Since $\delta = \delta_0\eta Bn$, $B \ge 2L$, $K \le 2d$, and $H \le T$, this implies

$$\mathbb{E}[R_T] \le r_1 + \frac{D}{\eta} + [c_{\mathrm{st}} + 2\delta_0]\eta B^2 ndT + Te_{\rho_T}(n,\delta) + \rho_T C_\Phi.$$

The parameter and range-bound calculation in Lemma E.9 applies to the first three terms without change; the remaining terms are nonnegative. Thus, for any fixed n and η from (9),

$$\mathbb{E}[R_T] \le \min\left\{2LT,\ r_1 + CB\sqrt{DndT} + Te_{\rho_T}(n,\delta) + \rho_T C_\Phi\right\}. \tag{103}$$

The coefficient satisfies $0 \le \rho_T \le a_*$ and $s_T := (a_* - \rho_T)_+ = \min\{a_*, cT^{-1/8}\}$. Consequently (103) gives the uniform finite-horizon bound

$$\begin{aligned}\mathbb{E}[R_T] \le r_1 + CB\sqrt{DndT} + (1+2a_*)LT\sqrt{K/n} \\ + \frac{2Ls_T T}{\sqrt{n\delta}} + \frac{LT}{2n\delta} + a_* C_\Phi.\end{aligned} \tag{104}$$

The two floor-dependent terms must be handled even when $\rho_T = 0$. Set $h := \delta_0\eta_0 > 0$. The parameter rule gives $\delta = \min\{(h/2)\sqrt{Dn/(dT)}, 1/(2K)\}$, and hence

$$\frac{1}{\sqrt{n\delta}} \le \sqrt{\frac{2}{h}}\left(\frac{d}{D}\right)^{1/4} T^{1/4}n^{-3/4} + \sqrt{\frac{2K}{n}}, \tag{105}$$

$$\frac{LT}{2n\delta} \le \frac{L}{h}\sqrt{\frac{d}{D}}\,T^{3/2}n^{-3/2} + \frac{LKT}{n}. \tag{106}$$

The pathwise range bound $R_T \le BT$ absorbs the last term: for any $A \ge 0$,

$$\min\{BT, A + LKT/n\} \le A + \min\{BT, LKT/n\} \le A + \sqrt{BL}\,T\sqrt{K/n}.$$

For the case (i) resolution in (8), suppose first that $T \ge D$. By (78),

$$n_0\sqrt{T/D} \le n \le (2n_0+1)\sqrt{T/D}.$$

The first contribution from (105) is bounded by

$$2Ls_T\sqrt{\frac{2}{h}}\left(\frac{d}{D}\right)^{1/4} T^{5/4}n^{-3/4} \le \frac{2\sqrt{2/h}}{n_0^{3/4}} cLd^{1/4}D^{1/8}T^{3/4},$$

using $s_T \le cT^{-1/8}$. Its second contribution is at most $2\sqrt{2}\,a_* LT\sqrt{K/n}$, since $s_T \le a_*$. The other terms satisfy

$$B\sqrt{DndT} \le \sqrt{2n_0+1}\,B\sqrt{d}\,D^{1/4}T^{3/4},$$
$$LT\sqrt{K/n} \le n_0^{-1/2}L\sqrt{K}\,D^{1/4}T^{3/4},$$
$$\frac{L}{h}\sqrt{\frac{d}{D}}\,T^{3/2}n^{-3/2} \le \frac{L}{hn_0^{3/2}}\sqrt{d}\,D^{1/4}T^{3/4}.$$

Because $B \ge 2L$ and $K \le 2d$, the last line and the cap charge $\sqrt{BL}\,T\sqrt{K/n}$ have the order of the first line. Also $r_1 \le B$ and $a_* C_\Phi \le 2a_* LT\sqrt{K/n}$. For an explicit finite constant independent of K, T, c, define

$$C_{\mathrm{N}}(n_0,\eta_0,\delta_0) := 2 + C(\eta_0,\delta_0)\sqrt{2n_0+1} + \frac{2+4a_*+2\sqrt{2}\,a_*}{\sqrt{n_0}} + \frac{1}{2hn_0^{3/2}} + \frac{2\sqrt{2/h}}{n_0^{3/4}}. \quad (107)$$

The preceding bounds prove the rate below for $T \ge D$. If $T < D$, the pathwise range bound instead gives $\mathbb{E}[R_T] \le BT \le BD^{1/4}T^{3/4}$, which is bounded by the same right-hand side. Therefore, for all $K, T \ge 2$,

$$\mathbb{E}[R_T] \le C_{\mathrm{N}}(n_0,\eta_0,\delta_0)\left[(B\sqrt{K-1}+L\sqrt{K})(\log K)^{1/4} + cL(K-1)^{1/4}(\log K)^{1/8}\right]T^{3/4}. \quad (108)$$

This proves case (i) of Theorem 3.1 with constants independent of K, T for every fixed positive triple and fixed $c \ge 0$. The constant C_{N} is uniform in c; its displayed c term records the extra cost of $\rho_T < a_*$. No division by ρ_T occurs, so the proof includes the zero-correction case at finite horizons. The floor multiplier must remain positive; the bound is not uniform as $\delta_0 \downarrow 0$. □

F.4 Why the uncorrected first-order bound is insufficient

Proposition F.1 (A boundary obstruction for the raw multinomial gradient)**.** *There is a fixed convex, 1-Lipschitz objective on Δ_2, with unique minimizer p, such that every integer $n \ge 1$ satisfies*

$$\sup_{m \in \operatorname{relint}\Delta_2} \{F_{V,n}(m) - V(p) - DF_{V,n}(m)[m-p]\} \ge 1. \quad (109)$$

Thus a uniformly vanishing additive error for the raw gradient cannot hold over the whole convex Lipschitz class, even as $q_n \to 0$.

Proof. Fix $0 < \chi < 1$ and define

$$V(x) = |x_1 - x_2| - \chi(x_1 + x_2) = \max\{(1,-1,\chi)^\top x, (-1,1,\chi)^\top x\} - \chi.$$

The second expression proves convexity and the ℓ_1-Lipschitz constant 1. Its unique minimizer is $p = (1/2, 1/2, 0)$, with $V(p) = -\chi$. At e_3 the multinomial policy is deterministic, so $F(e_3) = 0$. On either edge joining e_3 to e_i, $i = 1, 2$, the objective is affine and the multinomial mean identity gives $F((1-s)e_3 + se_i) = (1-\chi)s$. Hence $DF(e_3)[e_i - e_3] = 1 - \chi$, and

$$F(e_3) - V(p) - DF(e_3)[e_3 - p] = \chi - (\chi - 1) = 1.$$

The polynomial F and its derivatives are continuous at e_3. Taking $m = (\varepsilon, \varepsilon, 1-2\varepsilon)$ and sending $\varepsilon \downarrow 0$ proves (109).

Both feasible coordinate derivatives at e_3 are strictly positive, so it is a strict local minimum of F on the simplex. However, the Multinomial covariance identity gives

$$F(p) = \mathbb{E}|X_1 - X_2| - \chi \le \sqrt{q_n} - \chi < 0 \qquad \text{if } q_n < \chi^2.$$

Under that condition the boundary local minimum is spurious and F is nonconvex. □

Proposition F.1 excludes a uniform first-order bound; it does not give a regret lower bound. The scaled comparison and positive floor control the remaining boundary-dependent error, including when $\rho_T = 0$. The sharper comparator-dependent errors under MP convexity remain specific to the refinements in Appendix E.4.

G REGRET WITH A SPARSE OPTIMIZER

This appendix proves cases (i) and (ii) of Theorem 3.2. In both sparse cases, Algorithm 1 uses the fixed first-observation reference, the centered score without clipping, and the exact entropy update over Δ_δ. The floor applies to the policy mean; the sampled action may still lie on the boundary. The support bound selects parameters, but the algorithm does not use the support coordinates.

We first record the comparison needed for a floored policy. Write $F_n(m) = \mathbb{E}[V(N/n)]$ for $N \sim \mathrm{Mult}(n, m)$.

Lemma G.1 (Sparse comparison after flooring). *For every $a, b \in \Delta_{K-1}$ and integer $n \geq 1$,*

$$|F_n(a) - F_n(b)| \leq L\|a - b\|_1. \tag{110}$$

If p is an optimizer with at most s nonzero coordinates and $p^\delta := (1 - K\delta)p + \delta\mathbf{1}$, then

$$F_n(p^\delta) - V(p) \leq \frac{L\sqrt{s-1}}{\sqrt{n}} + 2LK\delta. \tag{111}$$

For $s = 1, n = 1$, the sampling term vanishes exactly.

Proof. Maximally couple a categorical draw from a with one from b. The mismatch probability is $\|a - b\|_1/2$. Repeat the coupling independently n times. The empirical frequencies then satisfy $\mathbb{E}\|X_a - X_b\|_1 \leq \|a - b\|_1$, proving (110) by Lipschitz continuity.

Let $S = \mathrm{supp}(p)$ and $j = |S| \leq s$. The coordinate variances of a multinomial average and Cauchy–Schwarz give

$$\begin{aligned} F_n(p) - V(p) &\leq \frac{L}{\sqrt{n}} \sum_{k \in S} \sqrt{p_k(1 - p_k)}, \\ \left(\sum_{k \in S} \sqrt{p_k(1 - p_k)} \right)^2 &\leq j \sum_{k \in S} p_k(1 - p_k) = j(1 - \|p\|_2^2) \leq j - 1 \leq s - 1. \end{aligned} \tag{112}$$

Since $\|p^\delta - p\|_1 \leq 2K\delta$, applying (110) after bounding the sampling error at the truly sparse p proves (111). For a vertex and $n = 1$, $X_p = p$ almost surely. □

Proof of Theorem 3.2, cases (i) and (ii). Let p be a deterministic optimizer with $|\,\mathrm{supp}(p)| \leq s$, and put $d = K-1$, $D = \log K$, $B = 2(L+\sigma)$, $H = T-1$. The first-round regret is $r_1 = V(u) - V(p) \leq 2L \leq B$; pathwise, $R_T \leq 2LT$. The multiplier inputs $n_0, \eta_0, \delta_0 > 0$ are fixed independently of K, T, s. The rule (10) ensures $0 < K\delta \leq 1/2$, so $p^\delta \in \Delta_\delta$ and $\mathrm{KL}(p^\delta \| u) \leq D$.

For $s = 1$, $n = 1$ and

$$F_1(m) = \sum_{k=1}^{K} m_k V(e_k), \qquad F_1(p) = V(p). \tag{113}$$

Thus F_1 is affine, regardless of whether V is MP convex. For $s \geq 2$, MP convexity makes F_n convex. In either case,

$$F_n(m) - V(p^\delta) \leq DF_n(m)[m - p^\delta] + F_n(p^\delta) - V(p^\delta). \tag{114}$$

Apply the centered, floored master inequality (69) with comparator p^δ. The terms involving $V(p^\delta)$ combine into $F_n(p^\delta) - V(p)$; Lemma G.1 then yields

$$\mathbb{E}[R_T] \leq r_1 + \frac{D}{\eta} + c(1/\delta_0)\eta B^2 ndH + \frac{L\sqrt{s-1}\,H}{\sqrt{n}} + 2LK\delta H, \tag{115}$$

where $c(a) = (e^a - 1 - a)/a^2$ is the extended entropy-stability constant. The floor charge satisfies $2LK\delta H \leq \delta_0 \eta B^2 nKH \leq 2\delta_0 \eta B^2 ndT$. Consequently, with $G := c(1/\delta_0) + 2\delta_0$,

$$\mathbb{E}[R_T] \leq r_1 + \frac{D}{\eta} + G\eta B^2 ndT + \frac{L\sqrt{s-1}\,T}{\sqrt{n}}. \tag{116}$$

The score has not been clipped, so there is no clipping-bias term.

The cap in the step-size rule can be absorbed with the pathwise range bound. Set $C(\eta_0, \delta_0) := 2/\eta_0 + (\eta_0/2)G + 2\sqrt{\delta_0}$. Indeed, the rule (10) gives

$$\frac{D}{\eta} \le \frac{2B}{\eta_0}\sqrt{DndT} + 2\delta_0 BnKD, \qquad G\eta B^2 ndT \le \frac{\eta_0 G}{2} B\sqrt{DndT}.$$

Since $K \le 2d$ and $2L \le B$,

$$\min\{2LT, 2\delta_0 BnKD\} \le B\sqrt{2\delta_0 nKDT} \le 2\sqrt{\delta_0}\, B\sqrt{DndT}.$$

Combining the two available regret bounds proves

$$\mathbb{E}[R_T] \le \min\left\{2LT, r_1 + C(\eta_0, \delta_0)B\sqrt{DndT} + \frac{L\sqrt{s-1}\,T}{\sqrt{n}}\right\}. \tag{117}$$

(i) A vertex optimizer. The rule fixes $n = 1$, and the comparator sampling term is zero. As $r_1 \le B$ and $\sqrt{DdT} > 1$ for $K, T \ge 2$,

$$\mathbb{E}[R_T] \le \min\{2LT, [C(\eta_0, \delta_0) + 1]B\sqrt{(K-1)T\log K}\}. \tag{118}$$

This proves case (i) of Theorem 3.2.

(ii) A sparse optimizer under MP convexity. Set $v = (L/B)\sqrt{(s-1)T/(Dd)}$, $U = \sqrt{DdT}$, and $Z = [Dd(s-1)]^{1/4}T^{3/4}$. The integer rule is $n = \lceil n_0\lceil v\rceil\rceil$. If $v \le 1$, then $n = \lceil n_0\rceil \le n_0 + 1$, and the second bound in (117) is at most $r_1 + [C(\eta_0, \delta_0) + 1]BU\sqrt{n_0 + 1}$. If $v > 1$, elementary rounding gives $n_0 v \le n \le (2n_0 + 1)v$, and the same bound is at most

$$r_1 + \left[C(\eta_0, \delta_0)\sqrt{2n_0 + 1} + n_0^{-1/2}\right]\sqrt{BL}\, Z.$$

Adding these two nonnegative regime bounds gives a uniform bound. When $T \ge Dd$, $U \le Z$, $Z \ge 1$, $r_1 \le B \le 2BZ$, and $\sqrt{BL} \le B/\sqrt{2}$. When $T < Dd$, the pathwise bound gives $\mathbb{E}[R_T] \le 2LT \le BZ$ because $s - 1 \ge 1$. Thus, for the dimension- and horizon-independent constant

$$C_{\mathrm{sp}}(n_0, \eta_0, \delta_0) := 2 + [C(\eta_0, \delta_0) + 1]\sqrt{n_0 + 1} + \frac{C(\eta_0, \delta_0)\sqrt{2n_0 + 1} + n_0^{-1/2}}{\sqrt{2}}, \tag{119}$$

we obtain

$$\mathbb{E}[R_T] \le \min\left\{2LT, C_{\mathrm{sp}}(n_0, \eta_0, \delta_0)B[(K-1)(s-1)\log K]^{1/4}T^{3/4}\right\}. \tag{120}$$

For each fixed positive multiplier triple, this gives case (ii) of Theorem 3.2. No lower bound $n_0 \ge 1$ is needed. The guarantee assumes a known support-size upper bound and does not establish adaptation to unknown s. □

H GENERATION OF THE REGULARIZED QUADRATIC TEST FUNCTIONS

The smooth and nonsmooth objectives in (11) and (12) share the same construction of their diagonal quadratic part. At each $K \in \{10, 100, 1000\}$, we start with 100 base quadratics. Their coefficients are independently sampled as $q_{j,k} \sim \mathrm{Unif}[0.5, 1.5]$ and $b_{j,k} \sim \mathrm{Unif}[-1, 0]$. The additive constant a_j makes the *base quadratic* attain minimum value one on Δ_{K-1}. For $K = 100$ and 1000, we append another 100 and 300 quadratics, respectively, using the same coefficient distributions. These appended base quadratics are generated separately for the two regularized families. Thus the numbers of functions at the three dimensions are 100, 200, and 400.

Smooth logarithmic family. For each function and coordinate in (11), we draw $c_{j,k} \sim \mathrm{Unif}[0, 1]$ and $e_{j,k} \sim \mathrm{Unif}[-1, 0]$ independently. All realized offsets $c_{j,k}$ are strictly positive. The draws are fixed before any algorithm run. Because $q_{j,k} - e_{j,k}/(x_k + c_{j,k})^2 \geq q_{j,k} > 0$ on the closed simplex, each resulting objective is smooth and strictly convex.

Nonsmooth weighted-ℓ_1 family. For (12), the fixed centers $c_{j,k}$ and weights $e_{j,k}$ are independently drawn from $\mathrm{Unif}[0, 1]$. The centers are not normalized across coordinates. The positive quadratic coefficients ensure a unique simplex minimizer; the absolute-value terms create coordinatewise kinks.

In both families, the base constant a_j is retained after adding the regularizer: the new minimum is *not* shifted back to one. We compute each new optimum offline by solving the separable simplex optimality conditions for one Lagrange multiplier. The logarithmic coordinates use the monotone derivative equation and multiplier bisection; the weighted-ℓ_1 coordinates use soft thresholding around their centers, nonnegative truncation, and multiplier bisection. The stored solutions are checked for simplex feasibility and the corresponding KKT conditions. Optima are used only to evaluate the plotted metrics, never as feedback to the learning algorithms.

For $K = 10, 100, 1000$, the first 60, 120, and 240 functions are used for parameter selection, and the remaining 40, 80, and 160 functions are held out for evaluation, with five test runs per function.