LOGARITHM-FREE CORRUPTION BOUNDS FOR TSALLIS-INF

Anonymous authors
Paper under double-blind review

ABSTRACT

Tsallis-INF attains logarithmic regret in stochastic bandits and minimax regret in adversarial bandits. For stochastic losses corrupted by an adaptive adversary with expected total budget C, existing self-bounding analyses give a corruption term of the form $\sqrt{CH\log(\cdot)}$, where $H=\sum_{i\neq i^\star}\Delta_i^{-1}$ is the sum of inverse gaps, whereas the corruption lower bound of Ito (2021a) has no logarithm. We remove the upper-bound logarithm for the unchanged algorithm. For both the importance-weighted and reduced-variance versions, without knowledge of T, C, or the gaps, the pseudo-regret on the observed losses is $O(\min\{\sqrt{KT},\, G_T+\sqrt{CH}\})$, where $G_T=H\log_+((K-1)T/H^2)$ is the known corruption-free rate and may be replaced by the per-arm sum $\sum_{i\neq i^\star}\log_+(\Delta_i^2T)/\Delta_i$. The bound holds under a per-round condition on conditional gaps that contains adaptive corruption of i.i.d. losses and lets the identity of the optimal arm change. Under expected or pathwise budgets, worst-case lower bounds for prescribed gap multisets match the corruption term within $O(\sqrt{\log K})$ in a stated regime, and up to constants for equal gaps. For fixed K and equal fixed gaps in $(0,1/4)$, $C=\sqrt{T}$ gives the minimax rate $\Theta(T^{1/4})$, improving the prior upper bound by $\sqrt{\log T}$. The logarithmic factor is unnecessary for Tsallis-INF in the corrupted-stochastic and conditional-gap models; the corresponding question for the full self-bounding class remains open. The proof directly bounds the integrated sampling trajectory with a potential, avoiding a conversion from the terminal self-bounding inequality. As an application, against an arbitrary opponent in a finite zero-sum game with a strict pure Nash equilibrium under bandit feedback, the Nash regret of Tsallis-INF improves from $O(H\log T/\gamma)$ to $O(G_T+H/\gamma)$, where γ is the minimum column gap. We give gap-dependent lower bounds for $K\times 2$ games and obtain the minimax rate $\Theta(\min\{\sqrt{T},\, \log_+(T\Delta^2)/\Delta+1/(\Delta\gamma)\})$ for $K=2$ in a small-gap regime.

1 INTRODUCTION

Stochastic and adversarial bandits describe two different sources of uncertainty. Fixed mean gaps permit logarithmic stochastic regret (Lai & Robbins, 1985; Auer et al., 2002a), whereas arbitrary loss sequences have minimax regret of order $\min\{T,\sqrt{KT}\}$ (Auer et al., 2002b; Audibert & Bubeck, 2010). Best-of-both-worlds algorithms (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017) seek both guarantees without knowing the environment in advance. Tsallis-INF (Zimmert & Seldin, 2019; 2021), an implicitly normalized forecaster (Audibert & Bubeck, 2010) with $1/2$-Tsallis-entropy regularization (Abernethy et al., 2015), attains both guarantees with a single update.

Adversarial corruption provides a concrete intermediate model: an adversary modifies an otherwise stochastic loss sequence, subject to a total budget C (Lykouris et al., 2018). The attainable corruption cost depends on the regret objective. Elimination algorithms bound *clean* pseudo-regret with an additive term linear in C (Gupta et al., 2019), most recently $O(C)$ without the factor K of BARBAR (Hu & Chen, 2025), and a linear term is unavoidable for that objective (Gupta et al., 2019, Section 5.1). Regret on the *corrupted* losses, which differs from clean regret by at most $2C$, can instead scale as $\sqrt{C}$: for prediction with expert advice, Amir et al. (2020) obtain $O(\log N/\Delta+C)$ on the uncorrupted losses, whereas Ito (2021a) shows that anytime Hedge attains $O(\log N/\Delta+\sqrt{C\log N/\Delta})$ on the

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.

corrupted losses, with a matching lower bound. We use expected pseudo-regret on the corrupted losses, the objective of the self-bounding line we compare against (Zimmert & Seldin, 2021; Masoudian & Seldin, 2021).

For bandits, the self-bounding analysis of Zimmert & Seldin (2021) places stochastic bandits, stochastically constrained adversaries, and corrupted stochastic bandits in a common framework, with regret $O(H \log T + \sqrt{CH \log T})$ for $H = \sum_{i \neq i^\star} \Delta_i^{-1}$; Masoudian & Seldin (2021) sharpen both terms but retain a logarithm inside the square-root corruption cost. Ito (2021a) records this upper bound against $\Omega(\min\{H + \sqrt{CH}, \sqrt{KT}\})$ for equal gaps in the self-bounding regime and conjectures that the gap closes from the lower-bound side: that the lower bound can be improved for consistent policies in the sense of Lai & Robbins (1985), and that the upper bound of Zimmert & Seldin (2021) is tight up to a constant factor. Under full information the corruption term carries no horizon logarithm; bandits were the outstanding case. For the stochastic term a logarithm must indeed remain. Whether the corruption term needs one is the question we answer, for Tsallis-INF itself.

It does not. Let $i^\star$ be the unique clean optimal arm, with gaps $\Delta_i \in (0, 1]$ sorted as $\Delta_{(1)} \leq \cdots \leq \Delta_{(K-1)}$. Define

$$H = \sum_{i \neq i^\star} \frac{1}{\Delta_i}, \qquad B_T = \sum_{i \neq i^\star} \frac{\log_+(\Delta_i^2 T)}{\Delta_i},$$
$$G_T = H \log_+\left(\frac{(K-1)T}{H^2}\right), \qquad \log_+ x = \max\{1, \log x\}, \tag{1}$$
$$H_* = \max_{1 \leq m \leq K-1} \frac{m}{\Delta_{(m)}}, \qquad \mathsf{h}_{K-1} = \sum_{m=1}^{K-1} \frac{1}{m}.$$

All logarithms are natural. For both the importance-weighted (IW) and reduced-variance (RV) versions of Tsallis-INF, we prove

$$\mathrm{Reg}_T \leq O\Big(\min\Big\{\sqrt{KT},\, \min\{B_T, G_T\} + \sqrt{CH}\Big\}\Big), \tag{2}$$

where C bounds the expected total ℓ_∞ corruption. The algorithm is unchanged and needs no knowledge of T, C, or the gaps. The guarantee permits adaptive corruptions and a changing optimal arm, and more generally holds under a budget on conditional gap shortfalls (Section 2). On the corrupted-stochastic model, and on the conditional-gap model containing it, the corruption term of the Zimmert–Seldin bound is therefore not tight for Tsallis-INF, contrary to the direction conjectured above; whether the same holds on the full self-bounding class remains open.

Contributions. Theorem 1 removes the logarithm from the corruption term for the unchanged Tsallis-INF algorithm. A strengthened self-bounding baseline for the same algorithm, derived in Appendix I by combining the general conversion of Masoudian & Seldin (2021) with the $O(K)$-remainder stability bound (Lemma 7), is $O(\min\{B_T, G_T\} + \sqrt{CHL_C})$ with $L_C = \log_+((K-1)T/(CH))$; the stochastic terms are theirs, and we remove $\sqrt{L_C}$.

The guarantee holds under a per-round budget on conditional gap shortfalls (Section 2), which contains adaptive corruption of i.i.d. losses and sits inside the self-bounding class of the prior analyses. The improvement is established on this conditional-gap model. The proof controls the sampling trajectory with a potential function (Section 5), avoiding the terminal self-bounding conversion that produces the logarithm.

To establish necessity, for every bandit algorithm we construct instances with prescribed gap multisets, unknown arm assignment, and regret $\Omega(\sqrt{CH_*})$ under expected or pathwise budgets in (4) (Propositions 3 and 13). Since $H_* \leq H \leq \mathsf{h}_{K-1} H_*$, the corruption terms match within $O(\sqrt{\log K})$; uniform gaps give $H_* = H$ and agreement up to constants.

As an application, against an arbitrary opponent in a finite zero-sum game with a strict pure equilibrium, Nash regret under bandit feedback improves from $O(H \log_+ T/\gamma)$ (Ito et al., 2026) to $O(G_T + H/\gamma)$, where γ is the smallest column gap (Corollary 4). For $K \times 2$ games in the small-gap regime of Theorem 5, a lower bound of $\Omega(H_*/\gamma)$ holds for sufficiently long horizons. The H/γ term matches it within $O(\log K)$, and up to constants for uniform gaps. An additional stochastic lower bound gives the full minimax rate for 2×2 games in this regime.

Table 1: Observed-loss pseudo-regret. Upper rows bound Tsallis-INF by $O(\text{stochastic term} + \text{corruption term})$, with a shared $O(\sqrt{KT})$ cap. Corruption terms vanish at $C = 0$; the self-bounding constant is $2C$. Checks mark applicability to the classes in (3); dashes mean no claim. The third row follows from Appendix I. Adapting Ito (2021a), the lower row applies to every learner on worst-case instances in $\mathsf{Cor}(C)$, hence all three classes, with prescribed gap multisets in (4). It compares corruption terms; uniform gaps give $H_* = H$.

Result	Stochastic term	Corruption term	Cor	CG	SB
Zimmert & Seldin (2021), RV, their Cor. 8	$H \log_+ T$	$\sqrt{CH \log_+ T}$	✓	✓	✓
Masoudian & Seldin (2021), RV, their Thm. 3 as published	$G_T + K \log_+ T$	$\sqrt{CHL_C}$	✓	✓	✓
Their Thm. 4 with Lemma 7 as input, IW and RV (our Prop. 14)	$\min\{B_T, G_T\}$	$\sqrt{CHL_C}$	✓	✓	✓
This paper, IW and RV (Thm. 1)	$\boldsymbol{\min\{B_T, G_T\}}$	$\boldsymbol{\sqrt{CH}}$	✓	✓	—
Lower bound, any learner (Props. 3 and 13)	—	$\Omega(\sqrt{CH_*})$	✓	✓	✓

Three model classes. We compare guarantees on three nested classes with the same reference arm, gaps, and budget. $\mathsf{Cor}(C)$: i.i.d. losses that an adaptive adversary may alter, with expected total ℓ_∞ change at most C. $\mathsf{CG}(C)$: the conditional gaps $\delta_{t,i} = \mathbb{E}_{t-1}[\ell_t(i) - \ell_t(i^\star)]$ may fall short of Δ_i, but the largest per-round shortfall sums to at most $2C$ in expectation; this is our model. At $C = 0$ it is the fixed conditional-gap condition of Wei & Luo (2018), stated per arm, which contains the stochastically constrained regime of Zimmert & Seldin (2021); the budget charges violations of that condition. $\mathsf{SB}(2C)$: only the inequality $\mathrm{Reg}_T \geq \mathbb{E}\sum_{t,i\neq i^\star} \Delta_i p_{t,i} - 2C$ is required, the self-bounding class of Zimmert & Seldin (2021); Masoudian & Seldin (2021). For fixed reference gaps and horizon,

$$\mathsf{Cor}(C) \subseteq \mathsf{CG}(C) \subseteq \mathsf{SB}(2C) \tag{3}$$

(Section 2.3). Table 1 records on which classes each bound is proved.

When it matters. The lower bounds require $0 < \Delta_i < 1/4$. Choose m_* attaining H_* and set $D = \Delta_{(m_*)}$. Their budget and horizon range is

$$\frac{2(m_* + 1)}{D} \leq C \leq DT. \tag{4}$$

Within this range, if also $C \geq H \log_+^2((K-1)T/H^2)$, then $G_T \leq \sqrt{CH}$. The full upper and worst-case lower bounds are therefore within $O(\sqrt{\log K})$; for uniform gaps the minimax regret is $\Theta(\sqrt{CH})$. For fixed $K \geq 2$, equal fixed gaps $0 < \Delta < 1/4$, and $C = \sqrt{T}$, these conditions hold for all sufficiently large T. The prior $O(T^{1/4}\sqrt{\log T})$ guarantee becomes the minimax-optimal $O(T^{1/4})$. Learning retains its necessary logarithm; the corruption term carries none.

Implication for adversarial game learning. Ito et al. (2026) show that Tsallis-INF achieves logarithmic Nash regret against an arbitrary opponent in a finite zero-sum game with a strict pure-strategy Nash equilibrium, under *uninformed* bandit feedback: the learner observes neither the matrix nor the opponent's action. Their bound is $O(H \log_+ T/\gamma)$, with H the sum of inverse row gaps and γ the smallest column gap. Their proof charges the opponent's deviations from the equilibrium column as corruption of the equilibrium row's bandit problem; substituting our bound into the same reduction gives $O(\min\{B_T, G_T\} + H/\gamma)$, up to reward normalization (Corollary 4), so the improvement is exactly the removed logarithm. Complementing Theorem 2 of Ito et al. (2026) with a finite-horizon stochastic lower bound, obtained by the classical two-instance argument (Bubeck et al., 2013; Lattimore & Szepesvári, 2020), gives matching minimax bounds for 2×2 games in the stated small-gap regime (Theorem 5).

Why a new proof is needed. The standard FTRL stability bound controls regret by $K + S_T$, where $S_T = \mathbb{E}\sum_{t,i\neq i^\star} \sqrt{p_{t,i}/t}$. The self-bounding argument uses the budget only through the constraint (9), added λ times to the stability bound: the same global λ multiplies the slack $2\lambda C$ and

divides the per-round maximum $O(H/(\lambda t))$, whose harmonic sum is $(H/\lambda)\log_+(\cdot)$, and optimizing λ leaves that logarithm inside $\sqrt{CHL_C}$ (Appendix I). Our proof instead bounds S_T directly using the per-round gap drift. A potential with a Lagrange-multiplier correction has sensitivity at most ρ to each shortfall c_t, even when a corrupted arm temporarily looks best. We count the trajectory with gradually increasing weights whose total cost is $O(H/\rho)$. A sharper multiplier estimate bounds its residual cost by $O(\rho S_T)$, which is absorbed into the left side for sufficiently small ρ. Balancing H/ρ against ρC gives $\sqrt{CH}$; the logarithm in the stochastic term is paid separately by potential and remainder terms and is never divided by ρ. For nonnegative conditional gaps, a short counting proof gives $O(\min\{B_T, G_T\} + \sqrt{CH})$ (Appendix F). Negative gaps leave a residual bounded by $2C$ in that proof; the trajectory argument avoids this charge.

Direct analyses of the $1/2$-Tsallis FTRL family have also been used to study sampling simple regret and best-arm identification in stochastic bandits (Zhan et al., 2026b;a). Our target is the cumulative stability sum S_T under corruption. Appendix N gives a broader comparison, organized by regret objective, self-bounding analysis, trajectory analysis, and game-learning setting.

2 Setting and Algorithm

2.1 Conditional Gaps and Regret

We state the model at the generality the proof uses; Section 2.2 verifies it for adaptive corruption of i.i.d. losses. Let $K \geq 2$ and $T \geq 1$. At round t, the learner selects a probability vector p_t in the simplex $\mathcal{P}_K = \{p \in [0,1]^K : \sum_i p_i = 1\}$, samples an arm I_t, and observes only its loss $\ell_t(I_t)$. Let $\mathcal{F}_{t-1}$ contain all past randomness and observations, so that p_t is $\mathcal{F}_{t-1}$-measurable. The environment generates $\ell_t \in [0,1]^K$ before the current action. It may depend on $\mathcal{F}_{t-1}$ and on p_t, but the learner uses fresh randomization satisfying $\Pr(I_t = i \mid \mathcal{F}_{t-1}, \ell_t) = p_{t,i}$. Write $\mathbb{E}_{t-1}$ for conditional expectation given $\mathcal{F}_{t-1}$.

Fix a reference arm $i^\star$ and numbers $0 < \Delta_i \leq 1$ for $i \neq i^\star$. Set $\Delta_{i^\star} = 0$. Define the conditional gaps and their common shortfall by

$$\delta_{t,i} = \mathbb{E}_{t-1}[\ell_t(i) - \ell_t(i^\star)], \qquad c_t = \max_{i \neq i^\star}(\Delta_i - \delta_{t,i})^+, \qquad \mathbb{E}\sum_{t=1}^{T} c_t \leq 2C. \tag{5}$$

Here $x^+ = \max\{x, 0\}$. The gaps $\delta_{t,i}$ may be negative: the reference arm need not remain optimal during corruption. The fixed vector Δ and the budget C are used only in the analysis. In words, c_t is the largest shortfall of a conditional gap below its reference value, and (5) bounds the expected total shortfall by $2C$. At $C = 0$ it forces $\delta_{t,i} \geq \Delta_i$ in every round: the fixed conditional-gap condition of Wei & Luo (2018) with per-arm gaps, which contains the stochastically constrained regime of Zimmert & Seldin (2021).

We distinguish the reference-arm regret from the usual pseudo-regret:

$$R_T = \mathbb{E}\sum_{t=1}^{T}[\ell_t(I_t) - \ell_t(i^\star)] = \mathbb{E}\sum_{t=1}^{T}\sum_{i \neq i^\star} p_{t,i}\delta_{t,i}, \tag{6}$$

$$\mathrm{Reg}_T = \mathbb{E}\sum_{t=1}^{T}\ell_t(I_t) - \min_{j \in [K]}\mathbb{E}\sum_{t=1}^{T}\ell_t(j). \tag{7}$$

Thus $R_T \leq \mathrm{Reg}_T$; the main upper bound holds for Reg_T, whose minimum is outside the expectation.

2.2 Stochastic Losses with Corruption

Let $\ell_t^0 \in [0,1]^K$ be independent and identically distributed clean loss vectors, independent of the past, with means μ_i, a unique best arm $i^\star$, and gaps $\Delta_i = \mu_i - \mu_{i^\star} > 0$ for $i \neq i^\star$. After seeing the clean vector, the past, and p_t, but not the drawn arm I_t, a nonanticipating adversary may replace it by $\ell_t \in [0,1]^K$, as in Lykouris et al. (2018); Zimmert & Seldin (2021). Suppose

$$\mathbb{E}\sum_{t=1}^{T}\|\ell_t - \ell_t^0\|_\infty \leq C. \tag{8}$$

A pathwise budget $\sum_t \|\ell_t - \ell_t^0\|_\infty \le C$ is a special case. If $d_t = \mathbb{E}_{t-1}\|\ell_t - \ell_t^0\|_\infty$, then $\delta_{t,i} \ge \Delta_i - 2d_t$ for every $i \ne i^\star$. Consequently $c_t \le 2d_t$, and (5) follows.

The regret in (7) uses the *observed, corrupted* losses. The clean pseudo-regret $R_T^0 = \mathbb{E}\sum_t \Delta_{I_t}$ is a different quantity: $|R_T^0 - R_T| \le 2C$. Our result therefore implies $R_T^0 \le O(\min\{B_T, G_T\} + \sqrt{CH}) + 2C$.

2.3 How the models relate

Fix T, $i^\star$, and Δ. The classes $\mathsf{Cor}(C)$ and $\mathsf{CG}(C)$ impose (8) and (5) for every learner; following Zimmert & Seldin (2021); Masoudian & Seldin (2021), $\mathsf{SB}(2C)$ imposes only the self-bounding constraint

$$\mathrm{Reg}_T \ge \mathbb{E}\sum_{t=1}^{T}\sum_{i\ne i^\star} \Delta_i p_{t,i} - 2C. \tag{9}$$

Theorem 1 and the prior analyses need the respective condition only for the learner analyzed. The corruption reduction above gives $\mathsf{Cor}(C) \Rightarrow \mathsf{CG}(C)$, and (5) gives $\mathrm{Reg}_T \ge R_T \ge \mathbb{E}\sum_{t,i\ne i^\star}\Delta_i p_{t,i} - \mathbb{E}\sum_t c_t \ge \mathbb{E}\sum_{t,i\ne i^\star}\Delta_i p_{t,i} - 2C$, which yields (3). Neither converse holds in general, even at $C = 0$ (Appendix M): the self-bounding constraint lets an early excess offset a later shortfall, whereas conditional gaps charge every shortfall before the current action, and this per-round structure is what the proof uses. No i.i.d. clean representation is needed once (5) is verified.

2.4 Tsallis-INF

Initialize $\widehat{L}_0 = 0$. For $a_t = \sqrt{t}/\kappa$, use the follow-the-regularized-leader update

$$p_t = \arg\min_{p\in\mathcal{P}_K}\left\{\langle p, \widehat{L}_{t-1}\rangle - 2a_t\sum_{i=1}^{K}\sqrt{p_i}\right\}, \qquad \widehat{L}_t = \widehat{L}_{t-1} + \widehat{\ell}_t. \tag{10}$$

Every coordinate of the minimizer is positive. We analyze importance weighting (IW), with $\kappa = 1$ and $b_{t,i} = 0$, and reduced variance (RV), with $\kappa = 2$ and $b_{t,i} = \frac{1}{2}\mathbf{1}\{p_{t,i} \ge 16/t\}$. In both cases, after sampling $I_t \sim p_t$ and observing $\ell_t(I_t)$, form the unbiased estimate

$$\widehat{\ell}_{t,i} = b_{t,i} + \frac{\mathbf{1}\{I_t = i\}}{p_{t,i}}(\ell_t(i) - b_{t,i}). \tag{11}$$

These choices correspond to learning rates $\eta_t = 2/\sqrt{t}$ for IW and $\eta_t = 4/\sqrt{t}$ for RV in the normalization of Zimmert & Seldin (2021). The RV threshold is $\eta_t^2 = 16/t$. The algorithm uses neither T, C, nor Δ.

3 Main result and lower bound

Theorem 1 (Logarithm-free corruption dependence)**.** *For either version in (10)–(11), assume (5). There are universal constants M_1, M_2 such that, for every K, T, C and positive reference gap vector,*

$$S_T := \mathbb{E}\sum_{t=1}^{T}\sum_{i\ne i^\star}\sqrt{p_{t,i}/t} \le M_1\big(\min\{B_T, G_T\} + \sqrt{CH}\big), \tag{12}$$

$$\mathrm{Reg}_T \le M_2 \min\left\{\sqrt{KT},\ \min\{B_T, G_T\} + \sqrt{CH}\right\}. \tag{13}$$

Corollary 2 (Adaptive corruption of stochastic losses)**.** *In the model of Section 2.2, the upper bounds in Theorem 1 hold under either the expected budget (8) or the corresponding pathwise budget. The identity of the optimal arm may change.*

Proof. The inequality $c_t \le 2d_t$ verifies (5). □

The trajectory statement (12) is stronger than a bound on the final regret alone: it controls the cumulative quantity in the stability analysis of Tsallis-INF and yields regret against every fixed

comparator, including the best arm in expectation. When $C = 0$, the bound is $O(\min\{B_T, G_T\})$. Both stochastic terms follow from the same potential argument, started at order $K + H^2/(K-1)$, by bounding one scalar remainder per arm or in aggregate. No additive $K \log T$ term is required. For $K-1$ equal gaps Δ, it gives

$$\mathrm{Reg}_T \le O\left(\min\left\{\sqrt{KT}, \frac{K-1}{\Delta}\log_+\left(\frac{\Delta^2 T}{K-1}\right) + \sqrt{\frac{C(K-1)}{\Delta}}\right\}\right).$$

The constants are not optimized. The analysis start enforces $a_t \ge 1024\sqrt{K}$ for the local estimates. For $\delta_{t,i} \ge 0$, Appendix F avoids this potential initialization and gives an explicit bound in terms of the FTRL stability constants. Theorem 1 concerns growth orders in (C, T).

3.1 Lower bound for prescribed gaps

Ito (2021a) proves the lower bound $\Omega(\min\{H + \sqrt{CH}, \sqrt{KT}\})$ for equal gaps in the self-bounding regime. We extend the construction to prescribed nonuniform clean gaps under an actual corruption budget. The complexity H_* in (1) satisfies

$$H_* \le H \le \mathsf{h}_{K-1} H_* \le (1 + \log(K-1)) H_*. \tag{14}$$

Indeed, $m/\Delta_{(m)} \le H$ and $1/\Delta_{(m)} \le H_*/m$. For uniform gaps Δ, $H_* = H = (K-1)/\Delta$.

Proposition 3 (Necessity of the corruption term)**.** *Let $K \ge 2$, $T \ge 1$ be integers and fix a gap multiset $0 < \Delta_{(1)} \le \cdots \le \Delta_{(K-1)} < 1/4$. Choose m_* attaining H_*, write $D = \Delta_{(m_*)}$, and suppose $2(m_*+1)/D \le C \le DT$. For every bandit algorithm, there exist an i.i.d. clean loss distribution with independent Bernoulli coordinates and exactly this gap multiset, and a nonanticipating corruption strategy satisfying (8) for which*

$$\mathrm{Reg}_T = R_T \ge \frac{1}{64}\sqrt{CH_*}.$$

The reference arm remains uniquely optimal in conditional expectation on every round. The learner may know T, C, and the gap multiset, but not its assignment to the arms.

Appendix G uses a shared corruption coin to modify the $m_* + 1$ candidate arms at expected cost at most $2D$ per round. Stopping the construction gives the same bound with constant $1/128$ under a pathwise budget (Proposition 13). By (14), the corruption terms in the upper and lower bounds differ by at most $O(\sqrt{\log K})$ in this regime. If also $C \ge H[\log_+((K-1)T/H^2)]^2$, then $G_T \le \sqrt{CH}$, so the full regret bounds differ by at most the same factor; for uniform gaps they match up to universal constants.

4 Refining upper and lower bounds for adversarial games

Consider an unknown finite zero-sum game with learner actions $[K]$, opponent actions $\mathcal{Y}$ with $|\mathcal{Y}| \ge 2$, and mean rewards $u(i, y) \in [0, 1]$. Feedback is *uninformed*: the learner observes only a reward $r_t \in [0, 1]$ with conditional mean $u(I_t, Y_t)$ given both actions and the past, and never the opponent's action Y_t. The opponent is adaptive but chooses its action without observing the learner's current random draw; the two draws are conditionally independent given the past. Suppose $(i^\star, y^\star)$ is a strict pure-strategy Nash equilibrium, and write

$$v = u(i^\star, y^\star), \quad \Delta_i = v - u(i, y^\star) > 0 \ (i \ne i^\star), \quad \Gamma_y = u(i^\star, y) - v, \quad \gamma = \min_{y \ne y^\star} \Gamma_y > 0.$$

The pure equilibrium makes v both the Nash value and the pure maximin value. We study *Nash regret*, the regret against this value (Maiti et al., 2025): $\mathrm{NR}_T = \mathbb{E}\sum_{t=1}^T [v - u(I_t, Y_t)]$. It coincides here with the pure-strategy maximin regret (PSMR) of Ito et al. (2026). Here H, H_*, and G_T from (1) use the row gaps Δ_i. Ito et al. (2026, Theorem 1) prove the bound $O(H \log_+ T/\gamma)$ in this model.

Corollary 4 (Refinement of the strict-equilibrium upper bound)**.** *Under the preceding assumptions, either version of Tsallis-INF applied to losses $1 - r_t$ satisfies*

$$\mathrm{NR}_T \le O\left(\min\left\{\sqrt{KT},\ G_T + \frac{H}{\gamma}\right\}\right). \tag{15}$$

The learner needs neither the game matrix nor the opponent's actions. The same conclusion holds with G_T replaced by $\min\{B_T, G_T\}$.

The proof follows the reduction in the proof of Theorem 1 of Ito et al. (2026). Deviations from $y^\star$ can reduce the row gaps, but they also improve the reward of the equilibrium row above v. If $N_T = \mathbb{E}\sum_t \mathbf{1}\{Y_t \neq y^\star\}$, the conditional shortfall budget is at most N_T, while this improvement is at least γN_T. Combining the first fact with a corruption term of order $\sqrt{N_T H \log T}$ yields their $O(H \log_+ T/\gamma)$; substituting Theorem 1 instead gives $\mathrm{NR}_T \le O(G_T + \sqrt{N_T H}) - \gamma N_T$, and maximizing over $N_T \ge 0$ gives (15). The improvement is thus a direct consequence of the logarithm-free corruption term; the conditional-gap model also absorbs the martingale reward noise without constructing an i.i.d. coupling. Appendix J supplies the complete reduction, including reward normalization, and replaces $1/\gamma$ by $\chi = \max_{y:\Gamma_y>0} d_y/\Gamma_y \le 1/\gamma$, the largest ratio of the gap reduction $d_y = \max_{i\neq i^\star}(\Delta_i - u(i^\star, y) + u(i, y))^+$ to the column gap Γ_y over columns with $\Gamma_y > 0$; the same argument covers some games with several minimizing columns.

Lower bounds for prescribed row gaps. Theorem 2 of Ito et al. (2026) gives the lower bound $\Omega(\min\{\sqrt{T}, 1/(\Delta\gamma)\})$ for 2×2 games in a small-gap regime. We extend the construction to any prescribed multiset of row gaps, placing the smallest m gaps and the optimal row among $m+1$ candidate rows. A separate argument in the style of Bubeck et al. (2013, Theorem 6), with Bernoulli rewards and a fixed opponent, supplies the logarithm of the stochastic term and matches the full upper bound when $K = 2$.

Theorem 5 (Gap-dependent lower bounds and a matching two-row rate)**.** *Let $K \ge 2$, $T \ge 1$ be integers, $0 < \gamma \le 1/26$, and $0 < \Delta_{(1)} \le \cdots \le \Delta_{(K-1)} \le 1/26$. Over uninformed $K \times 2$ games with rewards in $[0, 1]$, a strict pure equilibrium, this multiset of row gaps, and column gap γ, the minimax Nash regret is at least*

$$\Omega\left(\max_{1\le m\le K-1} \min\left\{T, \sqrt{(m+1)T}, \frac{m+1}{\Delta_{(m)}\gamma}\right\}\right). \tag{16}$$

For uniform gaps Δ, this is $\Omega(\min\{T, \sqrt{KT}, K/(\Delta\gamma)\})$. For $K = 2$ with row gap Δ, the minimax order is

$$\Theta\left(\min\left\{\sqrt{T}, \frac{\log_+(T\Delta^2)}{\Delta} + \frac{1}{\Delta\gamma}\right\}\right). \tag{17}$$

The lower bounds hold even if T, γ, and the gap multiset are known, with its assignment to rows unknown. Either Tsallis-INF version attains the two-row upper bound without these inputs.

The supremum defining this minimax statement ranges over game matrices, admissible reward laws, and nonanticipating opponents; the matrix, its Nash value, and the opponent's actions are unknown to the learner, and the two lower-bound constructions may use different games in the class. Appendix L proves the statement. If $m_* \in \arg\max_m m/\Delta_{(m)}$ and $T \ge (m_*+1)/(\Delta^2_{(m_*)}\gamma^2)$, then (16) gives $\Omega(H_*/\gamma)$. By (14), the H/γ upper-bound term and this lower bound differ by at most $\mathsf{h}_{K-1} \le 1 + \log(K-1)$ in their gap dependence; uniform gaps give $H_* = H$. For nonuniform gaps, the simpler finite-horizon lower bound $\Omega(\min\{T, \sqrt{KT}, H/\gamma\})$ can fail (Appendix J.4). Full matching for $K > 2$ is not established here. For general games with a positive value gap ζ between the mixed and pure maximin values, Ito et al. (2026) also give an $O(K/\zeta)$ bound on PSMR, which then differs from Nash regret; Appendix J records the precise relation.

5 How the trajectory removes the corruption logarithm

The stability bound $\mathrm{Reg}_T \le O(K + S_T)$ is the final step. We first bound the trajectory itself: for every analysis parameter $0 < \rho \le \rho_0$, where ρ_0 is a universal constant,

$$S_T \le O\left(\min\{B_T, G_T\} + \frac{H}{\rho} + \rho C\right). \tag{18}$$

A potential with sensitivity at most ρ pays for a weighted part of the trajectory. The cost of these weights is $O(H/\rho)$, while corruption costs $O(\rho C)$. Optimizing ρ gives $O(H + \sqrt{CH})$, and $B_T, G_T \ge H$. All weights and ρ are used only in the analysis.

5.1 Counting future sampling with bounded sensitivity

For $L \in \mathbb{R}^K$, define $u_i = L_i + z(L, a_t)$, where the Lagrange multiplier $z(L, a_t)$ of the simplex constraint is determined by $\sum_j a_t^2/u_j^2 = 1$ and $u_j > 0$. The KKT conditions of (10) give $p_i = a_t^2/u_i^2$ and $u_i \geq a_t$. Thus the current sampling cost is $1/u_i = \kappa\sqrt{p_i/t}$. If u grew at rate Δ_i, its integrated cost over the next t units of time would be

$$f_i(u, t) = \int_0^t \frac{\mathrm{d}s}{u + \Delta_i s} = \frac{1}{\Delta_i} \log\left(1 + \frac{\Delta_i t}{u}\right). \tag{19}$$

Its gap drift satisfies $-\Delta_i \partial_u f_i = 1/u - 1/(u + \Delta_i t)$, recovering the current cost up to a controlled remainder.

Counting every nonreference arm with full weight can make the potential too sensitive to corruption. We use increasing weights $\theta(t) = (\theta_i(t))_{i \neq i^\star} \in [0, 1]^{K-1}$ with $\|\theta(t)/\Delta\|_2 \leq \rho a_t$. For $m = K - 1$ equal gaps Δ, this simply gives $\theta_i(t) = \min\{1, \rho a_t \Delta/\sqrt{m}\}$: early sampling is counted with a smaller weight, which reaches one as the scale a_t grows. For general gaps, choose a common $v(t) \geq 0$ by

$$\sigma(v) = \left(\sum_{i \neq i^\star} \frac{e^{-2v/\Delta_i}}{\Delta_i^2}\right)^{1/2}, \qquad \sigma(v(t)) = \min\{\rho a_t, \sigma(0)\}, \tag{20}$$
$$\theta_i(t) = e^{-v(t)/\Delta_i}.$$

These weights satisfy the norm constraint and increase to one. The exponential curve makes their total change easy to control, as shown below. Define

$$V_t(L) = \sum_{i \neq i^\star} \frac{\theta_i(t)}{\Delta_i} \log\left(1 + \frac{\Delta_i t}{u_i}\right) + \rho(u_{i^\star} - a_t), \qquad u_i = L_i + z(L, a_t). \tag{21}$$

We evaluate it at $L = \widehat{L}_{t-1}$ before round t. The last term corrects for the Lagrange multiplier and is nonnegative.

Indeed, let $h_i = -\partial_u f_i(u_i, t)$, $q_i = p_i^{3/2}/\sum_j p_j^{3/2}$, and $D = \rho - \sum_{i \neq i^\star} \theta_i(t) h_i$. Cauchy–Schwarz gives $\sum_{i \neq i^\star} \theta_i(t) h_i \leq \|\theta(t)/\Delta\|_2/a_t \leq \rho$. Since $\partial_i z = -q_i$, the sensitivities are $w_i = -\partial_i V_t = \theta_i(t) h_i + D q_i \geq 0$ for $i \neq i^\star$, and $\sum_{i \neq i^\star} w_i = \rho - D q_{i^\star} \leq \rho$. Translation invariance and unbiasedness therefore imply $\nabla V_t \cdot \mathbb{E}_{t-1} \widehat{\ell}_t \leq -\sum_{i \neq i^\star} \Delta_i w_i + \rho c_t$. Even a reversal of the optimal arm costs at most ρc_t.

5.2 One quantity pays for the weights

Let $\theta(0) = 0$ and define the total weighted change

$$\mathcal{L} = \sum_{n \geq 1} \sqrt{n}\, \|\theta(n) - \theta(n-1)\|_2.$$

This controls both costs of weighting. Since $1 - \theta(t) = \sum_{n > t} [\theta(n) - \theta(n-1)]$, the omitted trajectory is at most $2\mathcal{L}$ by exchanging sums and using $\sum_{t < n} t^{-1/2} \leq 2\sqrt{n}$. The cost of increasing the coefficients in V_t is at most $\kappa\mathcal{L}$, because $f_i(u, n) \leq n/u$. Along the exponential curve the derivative norm is $\sigma(v)$, so Lemma 8 gives

$$\mathcal{L} \leq \frac{\kappa}{\rho} \int_0^\infty \sigma(v)^2\, \mathrm{d}v + m = \frac{\kappa H}{2\rho} + m = O(H/\rho). \tag{22}$$

The additional m accounts for integer times. Thus one elementary integral pays for both the omitted sampling and the coefficient changes.

5.3 Noise, initialization, and summation

The Lagrange multiplier creates an update remainder, but this remainder shrinks as the reference arm's probability approaches one. Lemma 9 bounds its expected loss-update remainder by

$O(\sum_{i\neq i^\star}\sqrt{p_i}/a_t)$. The scale update has the same bound after subtracting $a_{t+1}-a_t$; the potential supplies this subtraction through $-\rho(a_{t+1}-a_t)$. The remaining multiplier error is multiplied by $D\le\rho$, so its total cost is $O(\rho S_T)$. For sufficiently small ρ_0 this moves to the left side of the trajectory bound. The second derivative and the explicit time dependence leave $O(H_{a_t}/t)$, where $H_a=\sum_{i\neq i^\star}\min\{\Delta_i^{-1},a\}$.

Initialization needs a further algorithmic property beyond $\mathrm{Reg}_T\le O(K+S_T)$: Lemma 7 controls the reference arm's deficit relative to the *random minimum* $\min_j\widehat{L}_n(j)$, $\mathbb{E}[\widehat{L}_n(i^\star)-\min_j\widehat{L}_n(j)]\le O(\sqrt{Kn})+\mathbb{E}\sum_{t\le n}c_t$. It bounds the initial correction; past and subsequent corruption charges together are at most $2\rho C$.

Start at $\tau=\Theta(K+H^2/(K-1))$, large enough for the local geometry. Early sampling costs $O(H)$, and $\log(1+x)\le x$ bounds the initial logarithmic potential by $\kappa\sqrt{(K-1)\tau}=O(H)$. The common remainder $\sum_{t=\tau}^{T}H_{a_t}/t$ is $O(B_T)$ by splitting each arm's sum at Δ_i^{-2}, and $O(G_T)$ by $H_{a_t}\le H$ and $\tau\ge H^2/(K-1)$. Together with the $O(H/\rho)$ weight charges and absorption of $O(\rho S_T)$, this proves (18); FTRL gives regret. Appendix A supplies the full proof.

6 Discussion

Under a conditional gap-shortfall budget, the corruption cost of Tsallis-INF is $O(\sqrt{CH})$, with no logarithm, while the cost of learning keeps its necessary logarithm. In games this separates the two costs and removes the logarithm from the H/γ term. Our lower bound permits prescribed nonuniform row gaps, matches the $1/(\Delta\gamma)$ dependence for uniform gaps, and gives full matching bounds for 2×2 games in the stated small-gap regime.

Limitations. The proof uses per-round conditional gap information, which is stronger than the terminal self-bounding constraint. The logarithm-free bound therefore applies to the corrupted-stochastic and conditional-gap models; its extension to the full self-bounding class remains open. The analysis requires a distinguished reference arm with strictly positive reference gaps and does not cover multiple clean optima. The guarantee concerns expected pseudo-regret on the observed losses. It gives no high-probability counterpart; the implied clean-regret upper bound includes an additional $2C$.

The lower bounds are worst-case statements for prescribed gap multisets with unknown arm assignments, subject to the stated gap, budget, and horizon conditions. For nonuniform gaps, the lower complexity H_* can be smaller than H: the remaining factors are at most $O(\sqrt{\log K})$ for the bandit corruption term and for sufficiently long horizons, $O(\log K)$ for the game's H/γ term. Full minimax matching in games is proved for the stated 2×2 small-gap class. The Nash-regret corollary assumes a strict pure equilibrium; it does not supply a corresponding logarithmic guarantee for general mixed-equilibrium games.

The analysis targets regret orders, and its constants are not optimized. In particular, $a_t\ge 1024\sqrt{K}$ is a threshold for the local proof rather than a condition on running the algorithm; such conservative constants limit numerical comparisons at practical horizons. We analyze exact FTRL minimizers; probability clipping and numerical approximation require additional error control. No empirical performance evaluation is provided.

Open problems. Further questions are the transfer of the trajectory argument to other regularizers and to linear bandits, MDPs, or feedback graphs, tight lower bounds for arbitrary gap vectors and the stochastic term in games with more than two learner actions, and the extension to the bilinear-game algorithm of Ito et al. (2026), whose bound depends on the action-set dimension.

AI use statement

Generative AI tools were used to assist with theoretical ideation, formulating mathematical claims, developing and writing proofs, and reviewing mathematical arguments. They were also used for literature search and comparison, manuscript drafting and revision, and preparation of the LaTeX

submission. The authors are responsible for reviewing AI-assisted content and for the correctness of all claims, proofs, and citations.

ETHICS STATEMENT

This paper presents a theoretical analysis of bandit algorithms and learning in zero-sum games. The study involves no human participants, personal data, or real-world deployments. Any deployment in applications affecting people would require a separate assessment of the relevant privacy, fairness, and safety considerations.

REPRODUCIBILITY STATEMENT

The two algorithm versions and the conditional-gap model are given in Section 2. Appendix A proves the main theorem with a unified trajectory argument yielding both stochastic complexities; Appendices B–E prove the local estimates and FTRL guarantees. Appendix G constructs the expected-budget lower bound and Appendix H its pathwise refinement. Appendices J–L prove the game upper and lower bounds, and Appendix I derives the per-arm prior bound used in the comparison. Appendix M separates the model classes, and Appendix F treats the case $\delta_{t,i} \geq 0$ with an explicit bound in terms of the FTRL stability constants. All claims are theoretical; no empirical results are used as evidence in this paper.

REFERENCES

Jacob D. Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, and Roi Livni. Prediction with corrupted expert advice. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 116–120. PMLR, 2016.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 42.1–42.23. PMLR, 2012.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 122–134. PMLR, 2013.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5503–5570. PMLR, 2023.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1562–1578. PMLR, 2019.

Zicheng Hu and Cheng Chen. A near-optimal, scalable and parallelizable framework for stochastic bandits robust to adversarial corruptions and beyond. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2502.07514.

Shinji Ito. On optimal robustness to adversarial corruption in online decision problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7409–7420, 2021a.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2552–2583. PMLR, 2021b.

Shinji Ito, Haipeng Luo, Taira Tsuchiya, and Yue Wu. Instance-dependent regret bounds for learning two-player zero-sum games with bandit feedback. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2858–2892. PMLR, 2025.

Shinji Ito, Haipeng Luo, Arnab Maiti, Taira Tsuchiya, and Yue Wu. Adversarial learning in games with bandit feedback: Logarithmic pure-strategy maximin regret. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3666–3692. PMLR, 2026.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30918–30978, 2023.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Wei-Cheng Lee and Francesco Orabona. A best-of-both-worlds proof for Tsallis-INF without fenchel conjugates. arXiv preprint arXiv:2511.11211, 2025.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122. ACM, 2018.

Arnab Maiti, Kevin Jamieson, and Lillian J. Ratliff. On the limitations and possibilities of Nash regret minimization in zero-sum matrix games under noisy feedback. In *Algorithmic Game Theory (SAGT 2025)*, Lecture Notes in Computer Science, pp. 3–20. Springer, 2025.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the Tsallis-INF algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3330–3350. PMLR, 2021.

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 279–289. PMLR, 2021.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1287–1295. PMLR, 2014.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Taira Tsuchiya, Shinji Ito, and Haipeng Luo. Corrupted learning dynamics in games. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5506–5552. PMLR, 2025.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1263–1291. PMLR, 2018.

Jingxin Zhan, Yuze Han, and Zhihua Zhang. Does 1/2-Tsallis-INF also work well for best-arm identification? arXiv preprint arXiv:2608.15365, 2026a. Version 1, August 15, 2026.

Jingxin Zhan, Yuze Han, and Zhihua Zhang. Last-iterate analyses of FTRL with the 1/2-Tsallis entropy in stochastic bandits. arXiv preprint arXiv:2510.22819, 2026b. Version 3, revised August 15, 2026; first posted in 2025.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 467–475. PMLR, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

A FULL TRAJECTORY ANALYSIS AND PROOF OF THE MAIN THEOREM

The theorem concerns the exact minimizer in (10); an implementation with numerical clipping requires a separate approximation analysis. We first state the algorithmic inputs, then construct deterministic weights that control the potential's sensitivity. One path-length bound pays for both costs of weighting. A sharper multiplier estimate makes the remaining update error a small multiple of S_T itself. A single initialization and summation argument gives both B_T and G_T by bounding the same scalar remainder in two ways. Throughout, M denotes a universal positive constant whose value may increase between inequalities, and $m = K - 1$.

A.1 ALGORITHMIC INPUTS: RESPONSE, UPDATE CONTROL, AND INITIALIZATION

The closed form of (10) turns the sampling cost into reciprocal denominators. Local update control bounds the discrete remainder, while FTRL stability provides initialization and the regret bound.

For $L \in \mathbb{R}^K$ and $a > 0$, let $z = z(L, a)$ be the unique solution of

$$u_i = L_i + z > 0, \qquad \sum_{i=1}^{K} \frac{a^2}{u_i^2} = 1. \tag{23}$$

The KKT conditions of (10) give $p_i = a^2/u_i^2$, so $u_i \geq a$. Define

$$Z = \sum_{i=1}^{K} p_i^{3/2}, \qquad q_i = \frac{p_i^{3/2}}{Z}, \qquad U_2 = \sum_{i=1}^{K} p_i^2, \qquad p_{\max} = \max_i p_i.$$

Implicit differentiation yields

$$\partial_i z = -q_i, \qquad \partial_a z = \frac{1}{Z}, \qquad D^2 z[x, x] = 3 \sum_{i=1}^{K} \frac{q_i}{u_i} (x_i - q \cdot x)^2. \tag{24}$$

Thus z is convex in L. We will use

$$K^{-1/2} \leq Z \leq 1, \qquad U_2 \leq Z, \qquad p_{\max} \frac{U_2}{Z^2} \leq 1. \tag{25}$$

The last inequality follows from $U_2 \leq \sqrt{p_{\max}} Z$ and $Z \geq p_{\max}^{3/2}$.

Lemma 6 (Local update bounds)**.** *Suppose $a = \sqrt{t}/\kappa \geq 1024\sqrt{K}$, $1 \leq \kappa \leq 2$, and $x = b + e_j v_j$, where $0 \leq b_i \leq 1/2$ and $v_j = (\ell_j - b_j)/p_j$ with $0 \leq \ell_j \leq 1$. Assume a negative v_j is possible only if $p_j \geq 16/t$. For every point $L + sx$, $0 \leq s \leq 1$, at fixed a,*

$$u_i^s \geq u_i/2, \qquad q_i^s \leq 20 q_i, \qquad Z^s \geq Z/4. \tag{26}$$

Conditional on L and the loss vector, when j is sampled from p,

$$\mathbb{E} x_i^2 \leq 3/p_i, \qquad \mathbb{E}\big[(z(L + x, a) - z(L, a))^2\big] \leq M(1 + U_2/Z^2), \tag{27}$$

$$\mathbb{E}[z(L + x, a) - z(L, a) + q \cdot x] \leq M/(aZ). \tag{28}$$

The loose numerical cutoff makes these comparisons uniform in K. Only $O(K)$ initial rounds fall outside it, at an $O(K)$ total cost. In particular, the cutoff has no algorithmic role.

Lemma 7 (FTRL stability and the minimum of the loss estimates)**.** *For either estimator, define*

$$\widehat{\mathcal{R}}_n = \mathbb{E}\left[\sum_{t=1}^{n} \langle p_t, \widehat{\ell}_t \rangle - \min_j \widehat{L}_n(j)\right].$$

Against every nonanticipating environment, universal constants c_0, c_1, c_2 satisfy

$$\mathrm{Reg}_n \leq \widehat{\mathcal{R}}_n \leq c_2 \sqrt{Kn}, \qquad \mathrm{Reg}_T \leq \widehat{\mathcal{R}}_T \leq c_1 S_T + c_0 K. \tag{29}$$

Under (5), for $n \leq T$,

$$\mathbb{E}[\widehat{L}_n(i^\star) - \min_j \widehat{L}_n(j)] \leq \mathbb{E} \sum_{t=1}^{n} c_t + c_2 \sqrt{Kn} \leq 2C + c_2 \sqrt{Kn}. \tag{30}$$

These are two distinct inputs: the trajectory bound in (29) will convert S_T into regret at the end, whereas (30) will initialize the potential at an intermediate, possibly corrupted state. An ordinary pseudo-regret bound does not by itself control the random minimum $\min_j \widehat{L}_n(j)$. Keeping the prefix sum in (30) also separates past corruption from subsequent drift charges. The proofs of Lemmas 6 and 7 appear in Appendices B and C.

A.2 One path length pays for the weighting

Fix $0 < \rho \le 1$, with a smaller universal upper bound chosen below. For $v \ge 0$, define

$$\vartheta_i(v) = e^{-v/\Delta_i} \quad (i \ne i^\star), \qquad \sigma(v) = \left(\sum_{i \ne i^\star} \frac{e^{-2v/\Delta_i}}{\Delta_i^2} \right)^{1/2}, \qquad R = \sigma(0). \tag{31}$$

For real $t > 0$, choose the unique $v(t) \ge 0$ such that

$$\sigma(v(t)) = \min\{\rho a_t, R\}, \qquad \theta_i(t) = \vartheta_i(v(t)) = e^{-v(t)/\Delta_i}. \tag{32}$$

Set $v(0) = +\infty$ and $\theta(0) = 0$. Each weight increases continuously to one, and all equal one for $t \ge t_* := \kappa^2 R^2/\rho^2$. By construction,

$$\left\| \frac{\theta(t)}{\Delta} \right\|_2 = \min\{\rho a_t, R\} \le \rho a_t, \qquad \|\vartheta'(v)\|_2 = \sigma(v). \tag{33}$$

The vectors in this subsection contain only the m nonreference coordinates, and division by Δ is coordinatewise. For equal gaps $\Delta_i = \Delta$, the weights reduce to $\theta_i(t) = \min\{1, \rho a_t \Delta/\sqrt{m}\}$. The norm constraint limits sensitivity to corruption; smaller early weights will leave part of the trajectory unpaid. The exponential curve is useful because its speed equals the same sensitivity norm. These weights are used only in the analysis.

Lemma 8 (A common budget for both weighting costs)**.** *Let $d_n = \theta(n) - \theta(n-1)$ and $\mathcal{L} = \sum_{n \ge 1} \sqrt{n}\, \|d_n\|_2$. Then*

$$\mathcal{L} \le \frac{\kappa H}{2\rho} + m. \tag{34}$$

For arbitrary positive probability vectors at integer times,

$$\sum_{t \ge 1} \sum_{i \ne i^\star} (1 - \theta_i(t)) \sqrt{p_{t,i}/t} \le 2\mathcal{L} \le \frac{\kappa H}{\rho} + 2m. \tag{35}$$

For $u_{n,i} = a_n/\sqrt{p_{n,i}}$, define the charge for increasing the weights at time n by

$$\mathcal{J}_n = \sum_{i \ne i^\star} \frac{d_{n,i}}{\Delta_i} \log\left(1 + \frac{\Delta_i n}{u_{n,i}}\right). \tag{36}$$

Then, pathwise,

$$\sum_{n \ge 1} \mathcal{J}_n \le \kappa \mathcal{L} \le \frac{\kappa^2 H}{2\rho} + \kappa m. \tag{37}$$

Both cost bounds also hold when restricted to any subset of times.

Proof. First, the two costs are bounded by the same path length for any coordinatewise nondecreasing weights that eventually equal one. The identity $1 - \theta(t) = \sum_{n > t} d_n$ and Cauchy–Schwarz give

$$\begin{aligned} \sum_{t \ge 1} \sum_{i \ne i^\star} (1 - \theta_i(t)) \sqrt{p_{t,i}/t} &\le \sum_{t \ge 1} \frac{\|1 - \theta(t)\|_2}{\sqrt{t}} \\ &\le \sum_{n \ge 2} \|d_n\|_2 \sum_{t < n} \frac{1}{\sqrt{t}} \le 2\mathcal{L}. \end{aligned}$$

Also, $\log(1 + y) \le y$ gives $\mathcal{J}_n \le \kappa\sqrt{n} \sum_{i \ne i^\star} d_{n,i} \sqrt{p_{n,i}} \le \kappa\sqrt{n}\, \|d_n\|_2$.

It remains to bound $\mathcal{L}$ for the exponential curve. Write $v_n = v(n)$. On every nonempty interval (v_n, v_{n-1}), $\kappa^2\sigma(v)^2/\rho^2 > n-1$, hence $\sqrt{n} \le \kappa\sigma(v)/\rho + 1$. This includes $n = 1$, where $v_0 = \infty$, and the last interval crossing t_*. The vector triangle inequality and (33) imply

$$\mathcal{L} \le \sum_n \sqrt{n} \int_{v_n}^{v_{n-1}} \sigma(v)\,\mathrm{d}v \le \frac{\kappa}{\rho}\int_0^\infty \sigma(v)^2\,\mathrm{d}v + \int_0^\infty \sigma(v)\,\mathrm{d}v.$$

The first integral is

$$\int_0^\infty \sigma(v)^2\,\mathrm{d}v = \sum_{i\neq i^\star} \frac{1}{\Delta_i^2}\int_0^\infty e^{-2v/\Delta_i}\,\mathrm{d}v = \frac{H}{2}. \tag{38}$$

The second is at most $\sum_{i\neq i^\star}\int_0^\infty \Delta_i^{-1}e^{-v/\Delta_i}\,\mathrm{d}v = m$. This proves all three bounds. Restricting the times preserves them because all summands are nonnegative. □

A.3 The potential and its sensitivity

For $u_i = L_i + z(L, a_t)$, set

$$f_i(u,t) = \frac{1}{\Delta_i}\log(1+\Delta_i t/u), \qquad V_t(L) = \sum_{i\neq i^\star}\theta_i(t)f_i(u_i,t) + \rho(u_{i^\star} - a_t). \tag{39}$$

This potential is nonnegative and translation invariant. Its scalar terms integrate future reciprocal denominators, $f_i(u,t) = \int_0^t (u+\Delta_i s)^{-1}\,\mathrm{d}s$, and satisfy

$$-\Delta_i\partial_u f_i(u,t) = \frac{1}{u} - \frac{1}{u+\Delta_i t}. \tag{40}$$

Thus gap drift pays for the current sampling cost, up to a smaller positive tail. The shared Lagrange multiplier couples the arms; the correction $\rho(u_{i^\star} - a_t)$ restores the monotonicity needed under corruption. Write

$$h_i = -\partial_u f_i(u_i,t) = \frac{1}{\Delta_i}\left(\frac{1}{u_i} - \frac{1}{u_i+\Delta_i t}\right), \qquad D = \rho - \sum_{i\neq i^\star}\theta_i(t)h_i.$$

Cauchy–Schwarz and (33) give

$$\sum_{i\neq i^\star}\theta_i(t)h_i \le \frac{1}{a_t}\sum_{i\neq i^\star}\frac{\theta_i(t)}{\Delta_i}\sqrt{p_i} \le \rho.$$

Consequently, implicit differentiation yields

$$0 \le D \le \rho, \qquad -\partial_i V_t = \theta_i(t)h_i + Dq_i \ge 0 \quad (i \neq i^\star), \qquad \sum_{i\neq i^\star}(-\partial_i V_t) = \rho - Dq_{i^\star} \le \rho. \tag{41}$$

The reference derivative follows from translation invariance. Unbiasedness and the conditional gap condition (5) therefore imply, at $L = \widehat{L}_{t-1}$,

$$\begin{aligned}\nabla V_t \cdot \mathbb{E}_{t-1}\widehat{\ell}_t &\le -\sum_{i\neq i^\star}\theta_i(t)\Delta_i h_i + \rho c_t \\ &= -\sum_{i\neq i^\star}\frac{\theta_i(t)}{u_i} + \sum_{i\neq i^\star}\frac{\theta_i(t)}{u_i+\Delta_i t} + \rho c_t.\end{aligned} \tag{42}$$

We used $D \ge 0$ to discard the additional favorable gap drift. No ordering of the estimated scores is needed.

A.4 Multiplier noise and one-step control

The following refinement of (28) makes the correction coefficient constant. When the reference arm dominates, the multiplier is nearly linear in its score, so its curvature vanishes with the nonreference probabilities.

Lemma 9 (Multiplier noise tracks the trajectory). *Under the conditions of Lemma 6, put* $x = \widehat{\ell}_t$, $a = a_t$, $a' = a_{t+1}$, $\Delta a = a' - a$, *and*

$$R_z = z(L + x, a) - z(L, a) + q \cdot x, \qquad \varepsilon = z(L + x, a') - z(L + x, a), \qquad s(p) = \sum_{i \neq i^\star} \sqrt{p_i}.$$

Then

$$\mathbb{E}_{t-1} R_z \leq \frac{M s(p)}{a}, \qquad 0 \leq \varepsilon - \Delta a \leq \frac{M s(p)}{a}. \tag{43}$$

The second inequality holds pathwise.

Appendix D derives this estimate from the same local geometry. Its reference-sensitive form is what allows the multiplier noise to be paid by a small multiple of S_T. Define the truncated complexity $H_a = \sum_{i \neq i^\star} \min\{\Delta_i^{-1}, a\}$.

Lemma 10 (Weighted one-step potential bound). *If* $a_t \geq 1024\sqrt{K}$, *then, with* $L = \widehat{L}_{t-1}$,

$$\kappa \sum_{i \neq i^\star} \theta_i(t) \sqrt{p_{t,i}/t} \leq V_t(L) - \mathbb{E}_{t-1} V_{t+1}(L + \widehat{\ell}_t) + \mathbb{E}_{t-1} \mathcal{J}_{t+1} + \rho c_t + \frac{M H_{a_t}}{t} + M \rho \kappa \sum_{i \neq i^\star} \sqrt{p_{t,i}/t}. \tag{44}$$

The denominators in $\mathcal{J}_{t+1}$ *are evaluated at* $(L + \widehat{\ell}_t, a_{t+1})$.

The proof in Appendix E expands the loss and time updates together at the old state. The second derivative and the explicit time dependence cost $O(H_{a_t}/t)$. For the multiplier, retain the negative scale correction:

$$D(R_z + \varepsilon) - \rho(a' - a) = D\{R_z + \varepsilon - (a' - a)\} - (\rho - D)(a' - a).$$

Since $0 \leq D \leq \rho$, Lemma 9 bounds its conditional expectation by $M \rho s(p)/a$. This is the last term in (44), and a sufficiently small universal upper bound on ρ lets us absorb its sum into S_T.

A.5 Initialization, summation, and completion of the main theorem

Let $\tau \leq T$ be any deterministic start with $a_\tau \geq 1024\sqrt{K}$. Summing (44) and discarding the nonnegative terminal potential, then adding the early trajectory and (35), gives

$$S_T \leq 2\sqrt{K\tau} + \frac{\kappa H}{\rho} + 2m + \frac{1}{\kappa} \left\{ \mathbb{E} V_\tau(\widehat{L}_{\tau-1}) + \mathbb{E} \sum_{n=\tau+1}^{T+1} \mathcal{J}_n + \rho \mathbb{E} \sum_{t=\tau}^{T} c_t + M \sum_{t=\tau}^{T} \frac{H_{a_t}}{t} \right\} + M \rho S_T. \tag{45}$$

Fix a universal $\rho_0 \in (0, 1]$ so that the coefficient of the last term is at most $1/2$ for $0 < \rho \leq \rho_0$. All remaining bounds hold uniformly over this range, after absorbing that term into the left side. The state at $T + 1$ is the FTRL response to $\widehat{L}_T$; no extra loss or action is required.

For initialization, the multiplier equation implies

$$u_{i^\star} - a_\tau \leq \widehat{L}_{\tau-1}(i^\star) - \min_j \widehat{L}_{\tau-1}(j) + a_\tau \sqrt{K}.$$

Together with (30), this gives

$$\rho \mathbb{E}[u_{i^\star} - a_\tau] \leq \rho \mathbb{E} \sum_{t < \tau} c_t + O(\sqrt{K\tau}). \tag{46}$$

The prefix charge here and the subsequent corruption charges in (45) sum to at most $2\rho C$. All terms are integrable: the estimators satisfy $\mathbb{E}_{t-1}|\widehat{\ell}_{t,j}| \leq 3/2$, so finite-time scores and their minimum are integrable. The logarithmic part of V_t is bounded by $\sum_{i \neq i^\star} t/u_i \leq \kappa\sqrt{mt}$, and the displayed deficit bound controls its correction.

Completion of Theorem 1. Fix $0 < \rho \le \rho_0$ and choose the common starting time

$$\tau = \left\lceil 4 \cdot 1024^2 K + \frac{H^2}{m} \right\rceil, \qquad m = K - 1. \tag{47}$$

Since $H \ge m \ge K/2$, both $\sqrt{K\tau}$ and $\sqrt{m\tau}$ are $O(H)$. If $T < \tau$, the elementary bound $S_T \le 2\sqrt{KT} = O(H)$ already implies the claimed trajectory bound, because $B_T, G_T \ge H$. Suppose henceforth that $T \ge \tau$. Then $a_\tau \ge 1024\sqrt{K}$, so (45) applies. This starting time is used only in the analysis.

At the initial state, $\theta_i(\tau) \le 1$ and $\log(1+x) \le x$ give the pathwise bound

$$\sum_{i \ne i^\star} \theta_i(\tau) f_i(u_i, \tau) \le \sum_{i \ne i^\star} \frac{\tau}{u_i} = \kappa\sqrt{\tau} \sum_{i \ne i^\star} \sqrt{p_{\tau,i}} \le \kappa\sqrt{m\tau} = O(H). \tag{48}$$

Together with (46), this yields $\mathbb{E}V_\tau(\widehat{L}_{\tau-1}) \le O(H) + \rho\mathbb{E}\sum_{t<\tau} c_t$. Equation (37) bounds all subsequent weight increases, including the final charge at $T+1$, by $O(H/\rho)$ pathwise. Substituting into (45), combining the prefix and later corruption charges into $2\rho C$, and absorbing $M\rho S_T$ gives the common inequality

$$S_T \le M\left(\frac{H}{\rho} + \rho C + \sum_{t=\tau}^{T} \frac{H_{a_t}}{t}\right), \qquad 0 < \rho \le \rho_0, \tag{49}$$

where M is independent of ρ.

Only the scalar remainder needs two estimates. For the per-arm bound, extend the range to $1 \le t \le T$ and split each arm's sum at Δ_i^{-2}:

$$\sum_{t=1}^{T} \frac{\min\{\Delta_i^{-1}, a_t\}}{t} \le \sum_{t=1}^{T} \frac{\min\{\Delta_i^{-1}, \sqrt{t}\}}{t} = O\left(\frac{\log_+(\Delta_i^2 T)}{\Delta_i}\right). \tag{50}$$

Thus $\sum_{t=\tau}^{T} H_{a_t}/t = O(B_T)$. For the aggregate bound, use $H_{a_t} \le H$ and $\tau \ge H^2/m$:

$$\sum_{t=\tau}^{T} \frac{H_{a_t}}{t} \le H\big(1 + \log(T/\tau)\big) \le 2H\log_+(mT/H^2) = 2G_T. \tag{51}$$

Applying both estimates to (49) gives

$$S_T \le M\left(\min\{B_T, G_T\} + H/\rho + \rho C\right), \qquad 0 < \rho \le \rho_0. \tag{52}$$

This also holds when $T < \tau$ by the early-interval bound above. Choose $\rho = \rho_0$ when $C \le H/\rho_0^2$, and $\rho = \sqrt{H/C}$ otherwise. The cost $H/\rho + \rho C$ is $O(H)$ in the first case and $O(\sqrt{CH})$ in the second. Since $B_T, G_T \ge H$, we obtain

$$S_T = O\left(\min\{B_T, G_T\} + \sqrt{CH}\right).$$

Finally, $K \le 2H$ and the two FTRL guarantees (29) give

$$\mathrm{Reg}_T = O\left(\min\left\{\sqrt{KT},\ \min\{B_T, G_T\} + \sqrt{CH}\right\}\right),$$

which completes the proof of Theorem 1. □

B Proof of Lemma 6

Proof. All expectations in this proof are conditional on the current state and the loss vector; only the sampled coordinate is random. For a desired segment point, first add θb and then its sparse residual. Translation equivariance and monotonicity of z imply that adding the bounded baseline changes every u_i by at most $1/2$. Call this intermediate state u^b. Since $a \ge 1024\sqrt{K}$,

$$.99u_i \le u_i^b \le 1.01u_i, \qquad q_i^b \le 2q_i.$$

If the sparse residual is positive, the decrease of z is, by its convexity, at most $q_j^b|v_j| \le 2q_j/p_j \le 2\sqrt{K}$. The selected denominator increases and all other denominators decrease. If $q_j^b \le 1/2$, the unselected coordinates retain at least half of $\sum_i (u_i^b)^{-3}$. If $q_j^b > 1/2$, then p_j^b is maximal and at least $1/K$; using $p_j^b \le p_j/.99^2$ gives $|v_j|/u_j^b \le \sqrt{K}/(.99^2 a) < .01$. The cubic denominator sum again retains at least half of its initial value. These bounds prove (26) in this case.

If the residual is negative, its magnitude h satisfies

$$\frac{h}{u_j} \le \frac{1}{2a\sqrt{p_j}} \le \frac{\kappa}{8} \le \frac{1}{4}.$$

Let $v \ge 0$ be the resulting increase in z during the sparse change. As long as $v \le a/2$, all denominators lie between $.74u_i^b$ and $1.51u_i^b$. Hence $q_i^\theta \le 9q_i^b \le 18q_i$. Integrating $\mathrm{d}v/\mathrm{d}h = q_j^\theta$ gives $v \le 20q_j/(2p_j) \le 10\sqrt{K} < a/2$. Continuity prevents the assumed bound from first failing. The same denominator bounds give $u_i^\theta \ge u_i/2$ and $\sum_i (u_i^\theta)^{-3} \ge \frac{1}{4}\sum_i u_i^{-3}$. This proves (26).

Along any such segment,

$$|q^\theta \cdot x| \le \tfrac{1}{2} + 20q_j/p_j, \qquad |z(L+x,a) - z(L,a)| \le \tfrac{1}{2} + 20q_j/p_j.$$

Also $\mathbb{E}x_i^2 \le 1/2 + 2/p_i \le 3/p_i$. Since $\sum_j q_j^2/p_j = U_2/Z^2$, these prove (27). For (28), use the integral Taylor formula in (24), together with $q_i^\theta/u_i^\theta \le 40q_i/u_i = 40p_i^2/(aZ)$. The relevant expected quadratic expression is bounded by a constant times

$$\frac{1}{aZ}\left\{\sum_i p_i^2 \mathbb{E}x_i^2 + U_2(1 + U_2/Z^2)\right\} \le \frac{M}{aZ},$$

because $U_2 \le Z \le 1$. Constants can be enlarged between displays. □

C Proof of Lemma 7

Proof. Define $F_t(L) = \min_p\{\langle p, L\rangle - 2a_t \sum_i \sqrt{p_i}\}$. Its gradient is p, and

$$-D^2 F_t[x,x] = \frac{2}{a_t} Z \operatorname{Var}_q(x) \le \frac{2}{a_t}\sum_i p_i^{3/2} x_i^2. \tag{53}$$

At round t, abbreviate $L = \widehat{L}_{t-1}$, $x = \widehat{\ell}_t$, and $p = p_t$. The stability $r_t = F_t(L) + p \cdot x - F_t(L+x)$ is nonnegative. For $t \ge 16$, the probability of the selected coordinate along an estimator segment is at most $4p_j$. Indeed, the baseline changes its denominator by at most $1/2$; a positive sparse residual increases that denominator, whereas a negative residual decreases it by at most $u_j/4$. Since $a_t \ge 2$, its resulting denominator is at least $u_j/2$. Unselected coordinates of x are bounded by $1/2$, so (53) and averaging over j give

$$\mathbb{E}_{t-1} r_t = O((1 + \sum_i \sqrt{p_i})/a_t) = O(\sqrt{K/t}). \tag{54}$$

For $t < 16$, both versions have zero baseline. Monotonicity of F_t in every coordinate gives $r_t \le p_j x_j = \ell_j \le 1$.

For the stronger late-time bound, apply Lemma 6. It implies $Z^\theta \le 8Z$ and $\operatorname{Var}_{q^\theta}(x) \le 20 \operatorname{Var}_q(x)$: for the latter, evaluate the variational definition of variance at the mean under q. The bounded baseline can be centered at $b_{i^\star}$, giving

$$Z \operatorname{Var}_q(b) \le \tfrac{1}{4}\sum_{i \ne i^\star} p_i^{3/2}.$$

For the sparse part, its averaged quadratic form is at most $\sum_j \sqrt{p_j}(1 - q_j)$. This is $O(\sum_{i \ne i^\star} \sqrt{p_i})$. To check the reference-arm term, if $p_{i^\star} \ge 1/2$, use $1 - q_{i^\star} \le \sum_{i \ne i^\star} p_i^{3/2}/p_{i^\star}^{3/2}$; otherwise the sum of the suboptimal square roots is at least $1/\sqrt{2}$. Thus at $t \ge 4 \cdot 1024^2 K$,

$$\mathbb{E}_{t-1} r_t = O\left(\sum_{i \ne i^\star} \sqrt{p_i/t}\right). \tag{55}$$

Write $L_t = \widehat{L}_t$ and $x_t = \widehat{\ell}_t$. For clarity about the comparator, telescope the exact identity $\langle p_t, x_t\rangle = F_t(L_t) - F_t(L_{t-1}) + r_t$. Evaluating consecutive regularizers at the current minimizer, and evaluating F_n at an endpoint of the simplex, gives pathwise

$$\sum_{t\le n}\langle p_t, x_t\rangle - \min_i L_n(i) \le 2a_1(\sqrt{K}-1) + 2\sum_{t=2}^{n}(a_t - a_{t-1})\left(\sum_i \sqrt{p_{t,i}} - 1\right) + \sum_{t\le n} r_t. \quad (56)$$

Use $\sum_i \sqrt{p_i} - 1 \le \sqrt{K}$ and (54) to get the first part of (29). Use instead $\sum_i \sqrt{p_i} - 1 \le \sum_{i\neq i^\star}\sqrt{p_i}$ and (55); the expected stability of the initial $O(K)$ rounds is $O(K)$ by (54). This proves the bound $\widehat{\mathcal{R}}_T \le c_1 S_T + c_0 K$ in (29) directly against the random minimum $\min_j \widehat{L}_n(j)$. For the comparison to regret, unbiasedness and the elementary inequality $\mathbb{E}\min_i \widehat{L}_T(i) \le \min_i \mathbb{E}\widehat{L}_T(i)$ give

$$\mathrm{Reg}_T = \mathbb{E}\sum_{t=1}^{T}\langle p_t, \widehat{\ell}_t\rangle - \min_i \mathbb{E}\widehat{L}_T(i) \le \widehat{\mathcal{R}}_T. \quad (57)$$

In particular, $R_T \le \mathrm{Reg}_T \le \widehat{\mathcal{R}}_T$. Finally, $\widehat{\mathcal{R}}_n = R_n + \mathbb{E}[\widehat{L}_n(i^\star) - \min_i \widehat{L}_n(i)]$. The conditional gap assumption gives $R_n \ge -\mathbb{E}\sum_{t\le n} c_t$, so the coarse bound yields

$$\mathbb{E}[\widehat{L}_n(i^\star) - \min_i \widehat{L}_n(i)] \le c_2\sqrt{Kn} + \mathbb{E}\sum_{t\le n} c_t \le c_2\sqrt{Kn} + 2C.$$

This proves (30). Keeping the prefix sum explicit lets the initialization and subsequent drift use disjoint parts of the corruption budget. The random estimated comparator remains inside its expectation throughout. □

D Proof of Lemma 9

Proof. First consider the loss update, conditional on the current state and loss vector. If $p_{i^\star} < 1/2$, Cauchy–Schwarz gives

$$\frac{1}{Z} \le \sum_{j=1}^{K}\sqrt{p_j} \le 2\sum_{i\neq i^\star}\sqrt{p_i} = 2s(p).$$

The first assertion then follows from (28).

Suppose instead that $p_{i^\star} \ge 1/2$. At a segment point $L + \theta x$, normalization gives $u_j^\theta \ge a$, and (26) gives $q_j^\theta \le 20 q_j$. Hence

$$D^2 z(L+\theta x, a)[x,x] \le \frac{3}{a}\mathrm{Var}_{q^\theta}(x) \le \frac{60}{a}\mathrm{Var}_q(x). \quad (58)$$

For the second inequality, evaluate the variational definition of variance at the mean under q. Write $x = b + e_j v_j$ as in Lemma 6. Since $|v_j| \le 1/p_j$ and j is sampled from p,

$$\mathbb{E}\,\mathrm{Var}_q(x) \le 2\,\mathrm{Var}_q(b) + 2\sum_{j=1}^{K}\frac{q_j(1-q_j)}{p_j}.$$

Centering the bounded baseline at $b_{i^\star}$ shows that $\mathrm{Var}_q(b) \le (1-q_{i^\star})/4$. For the sparse part,

$$\sum_{i\neq i^\star}\frac{q_i(1-q_i)}{p_i} \le \frac{s(p)}{Z}, \qquad \frac{q_{i^\star}(1-q_{i^\star})}{p_{i^\star}} \le 2(1-q_{i^\star}), \qquad 1 - q_{i^\star} \le \frac{s(p)}{Z}.$$

Since $Z \ge p_{i^\star}^{3/2} \ge 2^{-3/2}$, the expected variance is at most $Ms(p)$. The integral Taylor formula in (58) now bounds $\mathbb{E}R_z$ by $Ms(p)/a$. Averaging over the loss vector proves the conditional expectation bound in (43).

For the scale change, at every state Cauchy–Schwarz gives

$$0 \le \frac{1}{Z} - 1 \le \sum_{j=1}^{K}\sqrt{p_j} - 1 \le \sum_{i\neq i^\star}\sqrt{p_i}.$$

After the loss update each denominator is at least half its old value, by (26). Denominators increase as the scale changes from a to a', because $\partial_a z = 1/Z > 0$. As $a'/a \le \sqrt{2}$, every intermediate probability is therefore at most $8p_j$. It follows pathwise that

$$\varepsilon - \Delta a = \int_a^{a'} \left(\frac{1}{Z(L+x,\alpha)} - 1 \right) \mathrm{d}\alpha \le \sqrt{8}\, \Delta a\, s(p) \le \frac{M s(p)}{a},$$

where $\Delta a = 1/(\kappa^2(a+a')) \le 1/(2a)$. This proves the second assertion of (43). □

E Proof of Lemma 10

Proof. Put $a = a_t$, $a' = a_{t+1}$, $x = \widehat{\ell}_t$, and use

$$z_0 = z(L,a), \quad z_1 = z(L+x,a), \quad z' = z(L+x,a'), \quad \varepsilon = z' - z_1, \quad u_i' = L_i + x_i + z'.$$

Lemma 6 and monotonicity in a imply $u_i' \ge u_i/2$. Normalization also gives $u_i' \ge a' \ge a$. Moreover, $\partial_a z = 1/Z \le \sqrt{K}$ implies

$$0 \le \varepsilon \le (a'-a)\sqrt{K} < 1/2, \qquad \mathbb{E}_{t-1}(u_i' - u_i)^2 \le M(1/p_i + 1 + U_2/Z^2). \tag{59}$$

The moment bound follows by writing $u_i' - u_i = x_i + (z_1 - z_0) + \varepsilon$ and applying (27); all moments on the right are evaluated at the old state.

For a scalar function f, write $\mathcal{B}_f(v,u) = f(v) - f(u) - f'(u)(v-u)$. First expand the new potential with its weights held at $\theta_i(t)$, then add their increases at the new state. The exact composition identity is

$$\begin{aligned} V_{t+1}(L+x) - V_t(L) = \nabla V_t(L)\cdot x &+ \sum_{i\neq i^\star} \theta_i(t) \mathcal{B}_{f_i(\cdot,t)}(u_i', u_i) \\ &+ D\{z_1 - z_0 + q\cdot x + \varepsilon\} - \rho(a'-a) \\ &+ \sum_{i\neq i^\star} \theta_i(t)[f_i(u_i', t+1) - f_i(u_i', t)] + \mathcal{J}_{t+1}. \end{aligned} \tag{60}$$

The same old-state $D = \rho - \sum_{i\neq i^\star} \theta_i(t) h_i$ multiplies the entire multiplier remainder.

Let $e_i = \min\{\Delta_i^{-1}, a\}$, so that $H_a = \sum_{i\neq i^\star} e_i$. The second derivative satisfies

$$f_i''(u,t) = \frac{1}{\Delta_i}\left(\frac{1}{u^2} - \frac{1}{(u+\Delta_i t)^2} \right) \le \frac{\min\{\Delta_i^{-1}, 2t/u\}}{u^2}. \tag{61}$$

Every point between u_i and u_i' is at least a and at least $u_i/2$. Since $t = \kappa^2 a^2$ and $1 \le \kappa \le 2$, the curvature on this interval is at most $M e_i/u_i^2$. Using (59), $0 \le \theta_i(t) \le 1$, and $u_i^{-2} = p_i/a^2$ therefore gives

$$\begin{aligned} \mathbb{E}_{t-1} \sum_{i\neq i^\star} \theta_i(t) \mathcal{B}_{f_i(\cdot,t)}(u_i', u_i) &\le \frac{M}{a^2}\left\{ H_a + \left(\sum_{i\neq i^\star} e_i p_i \right)(1 + U_2/Z^2) \right\} \\ &\le \frac{M H_a}{a^2}. \end{aligned} \tag{62}$$

The last step uses $\sum_{i\neq i^\star} e_i p_i \le H_a p_{\max}$ and $p_{\max} U_2/Z^2 \le 1$ from (25). The explicit time change in (60) is at most

$$\sum_{i\neq i^\star} \frac{1}{u_i' + \Delta_i t} \le \sum_{i\neq i^\star} \min\{1/a, 1/(\Delta_i t)\} = O(H_a/t).$$

The positive tail sum in (42) obeys the same bound.

For the multiplier term, retain the negative scale correction:

$$D\{z_1 - z_0 + q\cdot x + \varepsilon\} - \rho\Delta a = D\{R_z + \varepsilon - \Delta a\} - (\rho - D)\Delta a.$$

Since $0 \le D \le \rho$ by (41), Lemma 9 bounds its conditional expectation by $M\rho s(p)/a$. Substitute (42) into the conditional expectation of (60). Since

$$\frac{1}{u_i} = \kappa\sqrt{p_{t,i}/t}, \qquad \frac{s(p)}{a} = \kappa \sum_{i\neq i^\star} \sqrt{p_{t,i}/t}, \qquad a^{-2} = \kappa^2/t,$$

the preceding bounds give (44). □

F ORDER-PRESERVING CORRUPTION: A SHORT PROOF

We first consider the case in which the reference arm remains optimal in conditional expectation:

$$\delta_{t,i} \geq 0 \qquad (t \leq T,\ i \neq i^\star). \tag{63}$$

This restriction permits a short self-bounding argument. Its main ingredient is a counting inequality that handles arbitrary heterogeneous gaps without paying for the number of gap scales. The argument also identifies the residual term it leaves when corruption reverses the arm ordering.

Lemma 11 (A gap-counting inequality)**.** *For positive numbers $(\Delta_i)_{i\neq i^\star}$, define $N(x) = |\{i \neq i^\star : \Delta_i \leq x\}|$ and $H = \sum_{i\neq i^\star} \Delta_i^{-1}$. For every finite sequence $z_1, \dots, z_T \geq 0$,*

$$\sum_{t=1}^{T} \sqrt{\frac{N(z_t)}{t}} \leq 2\sqrt{H \sum_{t=1}^{T} z_t}. \tag{64}$$

Proof. Define

$$F(x) = \left(\sum_{i\neq i^\star} \min\{x/\Delta_i, 1\}^2\right)^{1/2}, \qquad m(x) = |\{t : z_t \geq x\}|.$$

The function F is absolutely continuous and nondecreasing, with $F(0) = 0$ and $F(x) \geq \sqrt{N(x)}$. Writing $h_i(x) = \min\{x/\Delta_i, 1\}$, Cauchy–Schwarz gives, almost everywhere,

$$F'(x)^2 = \frac{\left(\sum_{i\neq i^\star} h_i(x) h_i'(x)\right)^2}{\sum_{i\neq i^\star} h_i(x)^2} \leq \sum_{i\neq i^\star} h_i'(x)^2 = \sum_{i\neq i^\star} \frac{\mathbf{1}\{x < \Delta_i\}}{\Delta_i^2}.$$

Thus $\int_0^\infty F'(x)^2\,\mathrm{d}x \leq H$. Since $\sum_{t\in A} t^{-1/2} \leq 2\sqrt{|A|}$ for every finite set of positive integers A, integration followed by Cauchy–Schwarz gives

$$\begin{aligned}\sum_{t=1}^{T} \sqrt{\frac{N(z_t)}{t}} \leq \sum_{t=1}^{T} \frac{F(z_t)}{\sqrt{t}} &= \int_0^\infty F'(x) \sum_{t:z_t\geq x} \frac{1}{\sqrt{t}}\,\mathrm{d}x \\ &\leq 2\int_0^\infty F'(x)\sqrt{m(x)}\,\mathrm{d}x \leq 2\sqrt{H\sum_{t=1}^{T} z_t},\end{aligned}$$

where $\int_0^\infty m(x)\,\mathrm{d}x = \sum_t z_t$. □

Theorem 12 (Nonnegative conditional gaps)**.** *Suppose (63) holds, the corruption budget satisfies $\mathbb{E}\sum_{t=1}^{T} c_t \leq 2C$, where $c_t = \max_{i\neq i^\star}(\Delta_i - \delta_{t,i})^+$, and Lemma 7 gives*

$$\widehat{\mathcal{R}}_T \leq c_1 \mathbb{E} \sum_{t=1}^{T} \sum_{i\neq i^\star} \sqrt{p_{t,i}/t} + c_0 K$$

for absolute constants $c_0, c_1 > 0$. Then

$$\mathrm{Reg}_T \leq \sum_{i\neq i^\star} \frac{2c_1^2}{\Delta_i} \left(3 + \log_+ \frac{\Delta_i^2 T}{c_1^2}\right) + 8c_1\sqrt{CH} + 2c_0 K. \tag{65}$$

In particular, for $0 < \Delta_i \leq 1$ and $K \geq 2$,

$$\mathrm{Reg}_T = O\Big(\min\{B_T, G_T\} + \sqrt{CH}\Big).$$

Proof. For any conditional gaps, Lemma 7 and the comparator definitions give

$$R_T \leq \mathrm{Reg}_T \leq \widehat{\mathcal{R}}_T, \qquad \mathrm{Reg}_T \leq 2\widehat{\mathcal{R}}_T - R_T.$$

Together with the regret identity $R_T = \mathbb{E}\sum_t \sum_{i\neq i^\star} \delta_{t,i} p_{t,i}$ and the stability bound, this implies

$$\mathrm{Reg}_T \le 2c_0 K + \mathbb{E}\sum_{t=1}^{T}\sum_{i\neq i^\star}\left(2c_1\sqrt{p_{t,i}/t} - \delta_{t,i}p_{t,i}\right). \tag{66}$$

For each round, partition the nonreference arms into $A_t = \{i \neq i^\star : \delta_{t,i} \le \Delta_i/2\}$ and its complement. Every arm in A_t satisfies $\Delta_i \le 2c_t$. By (63), its negative linear term in (66) can be discarded. Cauchy–Schwarz and the simplex constraint give

$$\sum_{i\in A_t}\left(2c_1\sqrt{p_{t,i}/t} - \delta_{t,i}p_{t,i}\right) \le \frac{2c_1}{\sqrt{t}}\sqrt{|A_t|\sum_{i\in A_t} p_{t,i}} \le 2c_1\sqrt{\frac{N(2c_t)}{t}}. \tag{67}$$

Lemma 11 applies pathwise, even when the c_t are random and adaptive. Jensen's inequality then gives

$$2c_1\mathbb{E}\sum_{t=1}^{T}\sqrt{\frac{N(2c_t)}{t}} \le 4c_1\sqrt{2H\mathbb{E}\sum_{t=1}^{T} c_t} \le 8c_1\sqrt{CH}.$$

For $i \notin A_t$, maximizing the quadratic in $\sqrt{p_{t,i}}$ and using $p_{t,i} \le 1$ separately yields

$$2c_1\sqrt{p_{t,i}/t} - \delta_{t,i}p_{t,i} \le \min\left\{\frac{2c_1}{\sqrt{t}}, \frac{2c_1^2}{t\Delta_i}\right\}.$$

The right-hand side is nonnegative, so it may be summed over all rounds for each arm. Splitting the sum at $t = c_1^2/\Delta_i^2$ and using the elementary square-root and harmonic-sum bounds gives

$$\sum_{t=1}^{T}\min\left\{\frac{2c_1}{\sqrt{t}}, \frac{2c_1^2}{t\Delta_i}\right\} \le \frac{2c_1^2}{\Delta_i}\left(3 + \log_+\frac{\Delta_i^2 T}{c_1^2}\right).$$

Substituting these estimates into (66) proves (65) and the $O(B_T + \sqrt{CH} + K)$ bound.

For an aggregate estimate, write $m = K - 1$. The same good-arm contribution also satisfies

$$\sum_{\substack{i\notin A_t\\ i\neq i^\star}}\left(2c_1\sqrt{p_{t,i}/t} - \delta_{t,i}p_{t,i}\right) \le \min\left\{2c_1\sqrt{m/t}, \frac{2c_1^2 H}{t}\right\}.$$

The first bound uses the simplex constraint, and the second sums the coordinatewise quadratic bounds. Splitting at $c_1^2H^2/m$ as above gives

$$\sum_{t=1}^{T}\min\left\{2c_1\sqrt{m/t}, \frac{2c_1^2 H}{t}\right\} \le 2c_1^2 H\left(3 + \log_+\frac{mT}{c_1^2H^2}\right) = O(G_T).$$

Both splitting estimates remain valid when their cutoff is below one: then only the harmonic bound is needed. Since $c_1 > 0$ is a fixed universal constant, rescaling the arguments of $\log_+$ changes the bounds by at most a universal factor. Combining the two estimates and using $B_T, G_T \ge H \ge K - 1$ absorbs the additive $O(K)$ term. □

The sign condition is the only place where this argument restricts the corruption model. If some $\delta_{t,i} < 0$, the left side of (67) has an additional positive contribution. The same proof, with the comparator inequalities above unchanged, gives, for $0 < \Delta_i \le 1$ and $K \ge 2$,

$$\mathrm{Reg}_T \le O\Big(\min\{B_T, G_T\} + \sqrt{CH}\Big) + \mathcal{N}_T,$$
$$\mathcal{N}_T = \mathbb{E}\sum_{t=1}^{T}\sum_{i\neq i^\star}(-\delta_{t,i})^+ p_{t,i}.$$

The budget alone gives $\mathcal{N}_T \le \mathbb{E}\sum_t c_t \le 2C$, which is too large to recover the desired square-root corruption dependence when C exceeds H. Thus this short argument leaves a potentially linear residual. The main trajectory argument handles reversals without paying this residual term.

G AN EXPECTED-BUDGET LOWER-BOUND CONSTRUCTION

The following proof extends the bandit construction of Ito (2021a) to a prescribed gap multiset using a shared corruption coin. The clean losses remain independent across both arms and rounds.

Proof of Proposition 3. We prove the stronger bound $\mathrm{Reg}_T = R_T \geq \frac{1}{64}\sqrt{C(m+1)/\Delta_{(m)}}$ for every $m \in \{1, \dots, K-1\}$ satisfying $2(m+1)/\Delta_{(m)} \leq C \leq \Delta_{(m)}T$. Taking $m = m_*$ implies the proposition since $(m_*+1)/\Delta_{(m_*)} \geq H_*$. Write $J = m+1$, $D = \Delta_{(m)}$, and set

$$\begin{aligned} n &= \left\lfloor \frac{C}{4D} \right\rfloor, \qquad \varepsilon = \sqrt{\frac{J}{32n}}, \\ \frac{C}{8D} \leq n \leq T, &\qquad n \geq \frac{J}{4D^2}, \quad 0 < \varepsilon \leq D/\sqrt{8} < D < 1/4. \end{aligned} \tag{68}$$

The inequalities follow from $2J/D \leq C \leq DT$ and $\lfloor x \rfloor \geq x/2$ for $x \geq 1$. Fix the learner, including any dependence on the supplied parameters.

Let S contain J candidate optimal arms. For each $a \in S$, assign the multiset $\{\Delta_{(1)}, \dots, \Delta_{(m)}\}$ to $S \setminus \{a\}$. Assign the remaining gaps to arms outside S, identically for every candidate. Write $\Delta_i^{(a)}$ for these assignments and define the clean means by

$$p_a^{(a)} = 1/2 - D, \qquad p_i^{(a)} = 1/2 - D + \Delta_i^{(a)} \quad (i \neq a).$$

All means lie in $(1/4, 3/4)$; the unique optimal arm is a, with exactly the prescribed clean gap multiset. The corrupted prefix means will be

$$q_i^{(a)} = \begin{cases} 1/2 - \varepsilon, & i = a, \\ 1/2, & i \in S \setminus \{a\}, \\ p_i^{(a)}, & i \notin S. \end{cases}$$

Every outside mean is at least $1/2$, and is independent of a. Thus outside arms give no information about the candidate and are at least as costly as the suboptimal arms in S.

Generate the clean vectors independently over time, with independent Bernoulli coordinates of means $p_i^{(a)}$. For $t \leq n$, draw independent shared triggers, independent also of the clean vectors, and define coordinatewise flip probabilities by

$$B_t \sim \mathrm{Bern}(2D), \qquad \rho_i^{(a)} = \frac{q_i^{(a)} - p_i^{(a)}}{2D(1 - p_i^{(a)})} \quad (i \in S). \tag{69}$$

If $B_t = 1$, independently flip each zero coordinate in S to one with probability $\rho_i^{(a)}$; otherwise make no changes. Arms outside S are unchanged, as are all rounds after n. These decisions precede the learner's action. For $i \in S$, $p_i^{(a)} \leq 1/2$ and $0 \leq q_i^{(a)} - p_i^{(a)} \leq D$, so $0 \leq \rho_i^{(a)} \leq 1$. The corrupted mean is exactly $p_i^{(a)} + 2D(1 - p_i^{(a)})\rho_i^{(a)} = q_i^{(a)}$. The common trigger bounds the cost of simultaneous flips:

$$\|\ell_t - \ell_t^0\|_\infty \leq B_t \quad (t \leq n), \qquad \mathbb{E}\sum_{t=1}^{T} \|\ell_t - \ell_t^0\|_\infty \leq 2Dn \leq C/2. \tag{70}$$

Every prefix conditional gap is at least ε; afterwards it is the prescribed positive clean gap.

It remains to show that some candidate is hard to locate. Let P_a be the learner's prefix history law in candidate a, and let Q be the auxiliary law with mean $1/2$ on all arms in S and the unchanged means outside S. The learner observes one Bernoulli loss per round; independence between coordinates is unnecessary for this comparison. Writing $N_a = \sum_{t=1}^{n} \mathbf{1}\{I_t = a\}$, the chain rule gives

$$\mathrm{KL}(Q \| P_a) \leq 4\varepsilon^2 \mathbb{E}_Q N_a.$$

Indeed, only arm a differs, and $\mathrm{kl}(1/2, 1/2 - \varepsilon) = -\frac{1}{2}\log(1 - 4\varepsilon^2) \leq 4\varepsilon^2$ for $\varepsilon < 1/4$. Pinsker's inequality, followed by Jensen's inequality and $\sum_{a \in S} \mathbb{E}_Q N_a \leq n$, gives

$$\frac{1}{J}\sum_{a \in S} \mathbb{E}_{P_a} N_a \leq \frac{n}{J} + \frac{n\varepsilon}{J}\sum_{a \in S}\sqrt{2\mathbb{E}_Q N_a} \leq \frac{n}{J} + n\varepsilon\sqrt{\frac{2n}{J}} \leq \frac{3n}{4}.$$

Choose a candidate $a = i^\star$ for which

$$\mathbb{E}_{P_a}(n - N_a) \geq n/4. \tag{71}$$

Since the prefix gaps are at least ε and all later gaps are positive,

$$R_T \geq \frac{n\varepsilon}{4} = \frac{\sqrt{Jn}}{16\sqrt{2}} \geq \frac{1}{64}\sqrt{\frac{CJ}{D}}.$$

The reference arm is uniquely optimal in conditional expectation on every round, so $\mathrm{Reg}_T = R_T$. □

H A PATHWISE-BUDGET REFINEMENT OF THE LOWER BOUND

The preceding coupling can be stopped at a deterministic corruption cap. This preserves its prescribed gap multiset and gives a pathwise lower bound in the same parameter regime.

Proposition 13 (Necessity under a pathwise corruption budget)**.** *Let $K \geq 2$ and $T \geq 1$ be integers, and prescribe sorted gaps $0 < \Delta_{(1)} \leq \cdots \leq \Delta_{(K-1)} < 1/4$. Choose m_* attaining H_*, write $D = \Delta_{(m_*)}$, and assume $2(m_* + 1)/D \leq C \leq DT$. For every bandit algorithm, there exist an i.i.d. clean loss distribution with independent coordinates and exactly this gap multiset, and a nonanticipating corruption strategy such that*

$$\sum_{t=1}^{T} \|\ell_t - \ell_t^0\|_\infty \leq C \quad \textit{almost surely}, \qquad \mathrm{Reg}_T = R_T \geq \frac{1}{128}\sqrt{CH_*}.$$

The reference arm is uniquely optimal in conditional expectation on every round. The conclusion also holds when the learner knows T, C, and the gap multiset, but not its assignment to the arms.

Proof. As in the proof of Proposition 3, we prove the stronger bound $\mathrm{Reg}_T = R_T \geq \frac{1}{128}\sqrt{C(m+1)/\Delta_{(m)}}$ for every m satisfying $2(m+1)/\Delta_{(m)} \leq C \leq \Delta_{(m)}T$; $m = m_*$ gives the claim. Write $J = m + 1$, $D = \Delta_{(m)}$ and use the same clean vectors and proposed corruptions, with

$$n = \left\lfloor \frac{C}{4D} \right\rfloor, \qquad \varepsilon = \sqrt{\frac{J}{32n}}, \qquad M = \lfloor C \rfloor.$$

The assumptions imply

$$n \geq \frac{C}{8D} \geq \frac{J}{4D^2}, \qquad n \leq T/4, \qquad 0 < \varepsilon \leq D/\sqrt{8} < 1/4.$$

Fix the learner, including its dependence on any supplied parameters. In the *unstopped* process, (71) selects a reference arm $i^\star$ whose prefix nonreference count N_{un} satisfies $\mathbb{E}N_{\mathrm{un}} \geq n/4$. These pulls include arms outside the candidate set.

A proposed corruption uses a shared trigger $B_t \sim \mathrm{Bernoulli}(2D)$ and has ℓ_∞ cost at most B_t, by (70). In the *stopped* process, permit the proposed corruption only on the first M trigger successes, counting a success even if no coordinate changes. Make no changes after round n. Its pathwise cost is at most $M \leq C$. The decision uses the preceding trigger count, the current clean vector, and the current trigger and flip coins, before the learner's action is sampled.

Let $Z = \sum_{t=1}^{n} B_t$ include all prefix triggers, even those suppressed by the cap. Then $Z \sim \mathrm{Bin}(n, 2D)$ and

$$\mu := \mathbb{E}Z = 2Dn \leq C/2.$$

Markov's inequality applied to 2^Z, together with $\mathbb{E}[2^Z] = (1 + 2D)^n \leq e^\mu$, yields

$$\Pr(Z \geq M) \leq e^{\mu - M\log 2} \leq 2e^{-(\log 2 - 1/2)C} < \frac{1}{8}.$$

For the last inequality, $C \geq 2J/D > 16$, $D < 1/4$, and $12\log 2 > 8$.

Call a prefix round *active* if fewer than M trigger successes have occurred before it. Conditional on the past, an active round has gap ε on the other candidate arms and gap $\Delta_i - D + \varepsilon \geq \varepsilon$ outside the

candidate set. After the cap is reached, each gap is its positive clean value Δ_i, which may be smaller than ε. Thus we count only active nonreference pulls, writing their number as N_{active}.

Couple the stopped and unstopped learners using the same clean vectors, triggers, coordinate flip coins, and learner randomization. On $\{Z < M\}$ the cap remains open and their entire prefix histories coincide. Hence

$$N_{\text{active}} \geq N_{\text{un}} - n\mathbf{1}\{Z \geq M\}, \qquad \mathbb{E}N_{\text{active}} \geq n/4 - n/8 = n/8.$$

All other rounds have nonnegative expected regret, so

$$R_T \geq \varepsilon \mathbb{E}N_{\text{active}} \geq \frac{\sqrt{Jn}}{8\sqrt{32}} \geq \frac{1}{128}\sqrt{\frac{CJ}{D}}.$$

Every conditional gap remains positive, giving $\text{Reg}_T = R_T$. □

I Comparison with existing self-bounding bounds

Write $m = K - 1$, so that $G_T = H \log_+(mT/H^2)$, and define

$$L_C = \log_+\left(\frac{mT}{CH}\right) \quad (C > 0), \qquad D_T = K \log_+ T.$$

Corruption terms are defined to be zero at $C = 0$.

The published RV theorem and the general analysis. These are distinct results in Masoudian & Seldin (2021). Their Theorem 3 analyzes RV estimates with $\eta_t = 4/\sqrt{t}$. Its equations (5)–(7) retain an additive remainder of order D_T; their abstract and Table 1 emphasize the leading term G_T. Their Theorem 4 is a general reduction from

$$\text{Reg}_T \leq b \sum_{t=1}^{T} \sum_{i \neq i^\star} \sqrt{\mathbb{E}[p_{t,i}]/t} + D$$

and a self-bounding constraint to refined regret bounds. The parameters b, D are supplied by the algorithm's stability bound; the general analysis does not require $D = O(K \log T)$. For their RV specialization, equation (12) supplies $b = 5/4$ and $D = 14K \log T + 3\sqrt{K}/4 + 15$. The next proposition instead combines their general analysis with Lemma 7. This makes a stronger benchmark available for both versions, still under the broader self-bounding constraint. In particular, neither the G_T dependence nor removal of an additive $K \log T$ term is our claimed corruption-rate improvement.

Proposition 14 (Benchmark bound from the general self-bounding analysis)**.** *For either the IW or RV version, suppose*

$$\text{Reg}_T \geq \sum_{t=1}^{T} \sum_{i \neq i^\star} \Delta_i \mathbb{E}[p_{t,i}] - 2C, \qquad 0 < \Delta_i \leq 1. \tag{72}$$

Then Lemma 7 and the existing self-bounding analysis (Masoudian & Seldin, 2021) imply

$$\text{Reg}_T \leq O\left(\min\left\{\sqrt{KT}, \min\{B_T, G_T\} + \sqrt{CHL_C}\right\}\right). \tag{73}$$

For IW, this conclusion also follows using only the stability lemmas of Zimmert & Seldin (2021) as the input to that analysis.

Proof. Put $x_{t,i} = \mathbb{E}[p_{t,i}]$ and $Q_T = \sum_{t,i \neq i^\star} \sqrt{x_{t,i}/t}$. Lemma 7 and Jensen's inequality give $\text{Reg}_T \leq bQ_T + D$ for a universal $b \geq 1$ and $D = O(K)$. Writing $A = 2C$, equation (14) of Masoudian & Seldin (2021) gives, for $0 < \lambda \leq 1$,

$$\text{Reg}_T \leq \sum_{t,i \neq i^\star} \left(b(1+\lambda)\sqrt{x_{t,i}/t} - \lambda \Delta_i x_{t,i}\right) + \lambda A + (1+\lambda)D. \tag{74}$$

For completeness, the following bounds implement their analysis with truncated logarithms, covering short horizons and all budgets. Cauchy–Schwarz and coordinatewise quadratic maximization bound each round's sum by

$$\min\left\{2b\sqrt{m/t}, \frac{b^2H}{\lambda t}\right\}.$$

Splitting at $t = b^2H^2/(m\lambda^2)$ therefore bounds its time sum by

$$O\left(\frac{b^2H}{\lambda}\log_+\left(\frac{mT\lambda^2}{b^2H^2}\right)\right).$$

If $A = 0$, take $\lambda = 1$. If $A > 0$, put $L = \log_+(mT/(AH))$ and take $\lambda = \min\{1, b\sqrt{HL/A}\}$. When $\lambda < 1$, the logarithm is at most $L + \log L \le 2L$, yielding $O(b\sqrt{AHL} + D)$. When $\lambda = 1$, the same bound is $O(b^2G_T + A + D)$, and $A \le O(b^2G_T)$: this is immediate if $A \le H$; otherwise $L \le G_T/H$ and $A \le b^2HL$. Since b is universal, this proves

$$\mathrm{Reg}_T \le O(G_T + \sqrt{CHL_C} + D). \tag{75}$$

The coordinatewise calculation in the following proof, applied to (74) with $\lambda = 1$, gives $\mathrm{Reg}_T \le O(B_T + C + D)$. The same case split $C/H \le L_C$ or $C/H > L_C$ used there combines this bound with (75), giving $O(B_T + \sqrt{CHL_C} + D)$. Take the smaller of this and (75). Both B_T and G_T are at least $H \ge K - 1$, so $D = O(K)$ is absorbed. The adversarial bound in Lemma 7 gives the cap.

To verify the purely prior-results route for IW, use part 4 of Lemma 11 in Zimmert & Seldin (2021) with $\alpha = 1/2$, $\xi_i = 1$, and distinguished arm $i^\star$. For $\eta_t = 2/\sqrt{t}$ its condition $\eta_t \le 1/4$ holds at $t \ge 64$, and it bounds instantaneous stability by

$$\sum_{i\neq i^\star}\frac{\sqrt{x_{t,i}} + 3x_{t,i}}{\sqrt{t}} \le 4\sum_{i\neq i^\star}\sqrt{x_{t,i}/t}.$$

Their part 1 bounds each earlier stability term by one. The usual FTRL penalty telescope, using $\sum_i \sqrt{p_i} - 1 \le \sum_{i\neq i^\star}\sqrt{p_i}$, adds at most $2Q_T + 2\sqrt{K}$. Thus $\mathrm{Reg}_T \le 6Q_T + 2\sqrt{K} + 63$ supplies the general analysis with $b = 6$ and $D = O(\sqrt{K} + 1)$ using existing lemmas alone. □

A per-arm consequence of the published RV inequalities. For clarity, the following derivation retains the original RV remainder. It establishes that B_T itself is already available from their published analysis, without invoking Lemma 7.

Proposition 15 (Per-arm consequence of prior analysis)**.** *Consider the RV version of Tsallis-INF, and suppose its pseudo-regret satisfies the self-bounding constraint*

$$\mathrm{Reg}_T \ge \sum_{t=1}^{T}\sum_{i\neq i^\star}\Delta_i\mathbb{E}[p_{t,i}] - 2C, \qquad 0 < \Delta_i \le 1.$$

Then the existing RV analysis (Masoudian & Seldin, 2021) implies

$$\mathrm{Reg}_T \le O\left(B_T + \sqrt{CHL_C}\right), \tag{76}$$

where the square-root contribution is defined to be zero at $C = 0$.

Proof. Write $x_{t,i} = \mathbb{E}[p_{t,i}]$. Equation (14) of Masoudian & Seldin (2021), with its parameter $\lambda = 1$, gives

$$\mathrm{Reg}_T \le \sum_{t=1}^{T}\sum_{i\neq i^\star}\left(2b\sqrt{x_{t,i}/t} - \Delta_i x_{t,i}\right) + 2C + 2D, \qquad b = \tfrac{5}{4}, \tag{77}$$

where $D = 14K\log T + \frac{3}{4}\sqrt{K} + 15 = O(D_T)$. For every $x \in [0,1]$ and $d > 0$,

$$2b\sqrt{x/t} - dx \le \min\left\{\frac{2b}{\sqrt{t}}, \frac{b^2}{dt}\right\}.$$

The first inequality uses $x \le 1$ and drops the negative term; the second maximizes a quadratic in $\sqrt{x}$. Summing the first expression up to $t = d^{-2}$ and the second afterwards yields $O(d^{-1} \log_+(d^2 T))$, also when the splitting point exceeds T. Consequently

$$\mathrm{Reg}_T \le O(B_T + C + D_T). \tag{78}$$

This already proves the claim at $C = 0$, up to absorption of D_T.

For $C > 0$, repeating the preceding general analysis with $b = 5/4$ and $D = O(D_T)$ gives, for every horizon and budget,

$$\mathrm{Reg}_T \le O(G_T + \sqrt{CHL_C} + D_T). \tag{79}$$

The factor two between the self-bounding constant and C only changes universal constants. Put $c = C/H$ and $x = (K-1)T/H^2$, so that $G_T/H = \log_+ x$ and $L_C = \log_+(x/c)$. If $c \le L_C$, then $C \le \sqrt{CHL_C}$, and (78) applies. If $c > L_C \ge 1$, then

$$G_T/H \le L_C + \log c \le \sqrt{cL_C} + \sqrt{c} \le 2\sqrt{cL_C},$$

where $\log c \le \sqrt{c}$ for $c \ge 1$. Thus $G_T \le 2\sqrt{CHL_C}$, and (79) applies. In either case, $\mathrm{Reg}_T \le O(B_T + \sqrt{CHL_C} + D_T)$.

Finally, for $d \in (0, 1]$, the inequalities $d \log T = d \log(d^2 T) + 2d \log(1/d)$ and $d \log(1/d) \le 1/e$ imply

$$d \log_+ T \le (1 + 2/e) \log_+(d^2 T).$$

Summing over the $K - 1$ positive gaps and using $K \le 2(K-1)$ shows $D_T = K \log_+ T \le 2(1 + 2/e) B_T$. Absorbing this remainder proves (76). □

J Adversarial games: reduction and a game-dependent refinement

All rewards and utilities in this appendix are normalized to $[0, 1]$. The model, equilibrium, and gaps are those of Section 4 unless stated otherwise. Let $q_{t,y} = \Pr(Y_t = y \mid \mathcal{F}_{t-1})$, where the filtration may include the opponent's past randomization and actions, even though these are not observed by the learner. The current learner draw is independent of the current opponent draw conditional on this history. The reward is conditionally unbiased given both actions and the past. The reduction below follows the proof of Theorem 1 in Ito et al. (2026), which charges the opponent's deviations from $y^\star$ as corruption of the equilibrium row's bandit problem and completes the square; the differences are that Theorem 1 replaces the corruption term of order $\sqrt{N_T H \log T}$ used there, and that the conditional-gap model absorbs the martingale reward noise without an i.i.d. coupling.

J.1 Embedding the feedback model

For each possible learner action, the environment specifies a conditional reward law supported on $[0, 1]$. One may sample a vector with these conditional marginals after choosing Y_t and before sampling the learner's action. This coupling produces the same distribution of the observed reward as the original feedback model; no independence between coordinates is needed. The corresponding loss vector is $\ell_t(i) = 1 - r_t(i)$, and its conditional gaps satisfy

$$\delta_{t,i} = \sum_{y \in \mathcal{Y}} q_{t,y} \big(u(i^\star, y) - u(i, y)\big). \tag{80}$$

Thus the bounded martingale reward noise allowed by Ito et al. (2026) fits the conditional-gap model; an i.i.d. coupling of clean reward vectors is not required.

Define the gap reduction of column y by

$$d_y = \max_{i \ne i^\star} \big(\Delta_i - u(i^\star, y) + u(i, y)\big)^+. \tag{81}$$

Since $\Delta_i = v - u(i, y^\star)$, the expression inside the positive part is $u(i, y) - u(i, y^\star) - \Gamma_y$. Consequently $d_{y^\star} = 0$ and $0 \le d_y \le 1$. Convexity of the maximum and positive part yields

$$c_t = \max_{i \ne i^\star} (\Delta_i - \delta_{t,i})^+ \le \sum_y q_{t,y} d_y \le 1 - q_{t,y^\star}. \tag{82}$$

For $N_T = \mathbb{E}\sum_t \mathbf{1}\{Y_t \neq y^\star\}$, condition (5) therefore holds with $C = N_T/2$.

Let R_T be the reference-arm regret on these normalized losses and set $Q_T = \mathbb{E}\sum_t \Gamma_{Y_t}$. Conditional unbiasedness gives the exact identity

$$\mathrm{NR}_T = R_T - Q_T, \qquad Q_T \geq \gamma N_T. \tag{83}$$

Using $R_T \leq \mathrm{Reg}_T$ and Theorem 1, we obtain

$$\begin{aligned}\mathrm{NR}_T &\leq M_2 G_T + M_2\sqrt{N_T H/2} - \gamma N_T \\ &\leq M_2 G_T + \frac{M_2^2 H}{8\gamma}.\end{aligned} \tag{84}$$

The last inequality follows by completing the square in $\sqrt{N_T}$. Also $Q_T \geq 0$ implies $\mathrm{NR}_T \leq R_T \leq \mathrm{Reg}_T \leq M_2\sqrt{KT}$. This proves Corollary 4. Neither N_T nor Q_T is an algorithm input. Using the full statement of Theorem 1 also replaces G_T by $\min\{B_T, G_T\}$ throughout this argument and in the proposition below.

J.2 Replacing the smallest column gap

The preceding proof charges every nonequilibrium column the same amount, even if it preserves the row gaps. The following statement keeps its actual gap reduction.

Proposition 16 (Bound in terms of the gap ratio χ)**.** *Let $i^\star$ be a pure maximin row of a finite game, with $v = \min_y u(i^\star, y)$ and $\Gamma_y = u(i^\star, y) - v \geq 0$. Choose numbers $0 < \Delta_i \leq 1$ for $i \neq i^\star$, define H, G_T using these numbers, and define d_y by (81). Suppose $d_y = 0$ whenever $\Gamma_y = 0$. Set*

$$\chi = \max_{y:\Gamma_y>0} \frac{d_y}{\Gamma_y},$$

with $\chi = 0$ if there is no column with $\Gamma_y > 0$. Then either version of Tsallis-INF on losses $1 - r_t$ satisfies

$$\mathrm{NR}_T \leq \min\left\{M_2\sqrt{KT},\ M_2 G_T + \frac{M_2^2 \chi H}{8}\right\}. \tag{85}$$

Proof. For any minimizing column y_0 of $i^\star$, we have $\Gamma_{y_0} = d_{y_0} = 0$, so $u(i, y_0) \leq v - \Delta_i < v = u(i^\star, y_0)$ for every $i \neq i^\star$. Since $u(i^\star, y) \geq v$ for every y, the pair $(i^\star, y_0)$ is a pure Nash equilibrium and v is the Nash value. Equation (80) and the first inequality of (82) remain valid for these reference gaps. The assumption implies $d_y \leq \chi\Gamma_y$ for every column, so

$$\mathbb{E}\sum_{t=1}^T c_t \leq \mathbb{E}\sum_{t=1}^T d_{Y_t} \leq \chi Q_T.$$

Apply Theorem 1 with $C = \chi Q_T/2$ and use (83) to get

$$\mathrm{NR}_T \leq M_2 G_T + M_2\sqrt{\chi Q_T H/2} - Q_T \leq M_2 G_T + \frac{M_2^2 \chi H}{8}.$$

The adversarial alternative follows from $Q_T \geq 0$ as before. □

For a strict pure equilibrium, $d_y \leq 1$ implies $\chi \leq 1/\gamma$, recovering Corollary 4. The coefficient can be much smaller. For example, if every column is obtained from $y^\star$ by adding the same nonnegative constant to every row, then $d_y = 0$ for all y and $\chi = 0$, regardless of how small the positive column gaps are.

Proposition 16 also covers games with several minimizing columns. Its condition is equivalent to requiring

$$u(i^\star, y) - u(i, y) \geq \Delta_i > 0 \quad \text{for every } i \neq i^\star \text{ and every } y \text{ with } u(i^\star, y) = v.$$

For instance, the game matrix

$$\begin{pmatrix} 1/2 & 1/2 & 3/4 \\ 1/4 & 1/4 & 1 \end{pmatrix}$$

has two minimizing columns for the first row and no strict pure equilibrium. Nevertheless, choosing $\Delta_2 = 1/4$ gives gap reductions $(0, 0, 1/2)$ and column gaps $(0, 0, 1/4)$, hence $\chi = 2$ and $\mathrm{NR}_T = O(G_T + H)$. This is a sufficient condition for this subclass of non-strict games, not a guarantee for all games with a non-strict equilibrium.

J.3 Normalization, algorithms, and the remaining logarithm

The rewards in Ito et al. (2026) lie in $[-1, 1]$. Writing their rewards and utilities as $\widetilde{r}, \widetilde{u}$, use $r = (1 + \widetilde{r})/2$, $u = (1 + \widetilde{u})/2$, and $\ell = (1 - \widetilde{r})/2$ here. Row gaps, column gaps, gap reductions, and Nash regret are each divided by two; χ is unchanged. Thus all order comparisons above hold in their units up to universal constants.

The learning-rate conventions can also be checked exactly. Their Algorithm 1 uses $\varphi_{1/2}(p) = 2\sum_i(\sqrt{p_i} - p_i)$, $\eta_t^{\mathrm{game}} = 1/(2\sqrt{t})$, and reward estimates $g_{t,i} = 1 - 2\widehat{\ell}_{t,i}$, where $\widehat{\ell}_{t,i}$ is the IW estimator for ℓ. After dividing their maximization objective by two and discarding terms constant over the simplex, it is

$$-\langle p, \widehat{L}_{t-1}\rangle + \frac{1}{\eta_t^{\mathrm{game}}}\sum_i \sqrt{p_i} = -\langle p, \widehat{L}_{t-1}\rangle + 2\sqrt{t}\sum_i \sqrt{p_i}.$$

This is exactly our IW version. Hence the IW consequence sharpens the analysis of their specified algorithm, despite the different numerical learning-rate notation. The RV consequence is for the separate RV version defined in this paper; it is not a claim about their algorithm for bilinear games, whose bound depends on the action-set dimension.

For $K = 2$, our bound reads

$$O\left(\min\left\{\sqrt{T}, \frac{\log_+(\Delta^2 T)}{\Delta} + \frac{1}{\Delta\gamma}\right\}\right).$$

Theorem 2 and Appendix C of Ito et al. (2026) give the $1/(\Delta\gamma)$ lower bound for two rows. Appendix K extends their construction to K rows with any prescribed multiset of small positive row gaps and all integer horizons $T \geq 1$. Appendix L supplements it with a finite-horizon stochastic lower bound and matches the full expression, over the class of two-by-two games with the specified small gaps.

The surviving logarithm in the stochastic term is substantive: Lemma 18 proves the lower bound $\log(1 + 8T\Delta^2)/(32\Delta)$ for every learner and every horizon, using two games with the same gaps and a fixed equilibrium-column opponent. Thus closing the upper–lower comparison gap does not mean removing the necessary logarithm from the stochastic term.

J.4 Nonuniform gaps and finite horizons

The uniform-gap lower bound cannot be extended to every gap profile by simply replacing $(K-1)/\Delta$ with H while keeping the same horizon caps. To see this, we can also apply Theorem 1 with larger reference gaps. For $h \in (0, 1]$, define

$$\bar{\Delta}_i = \max\{\Delta_i, h\}, \qquad H_h = \sum_{i\neq i^\star}\frac{1}{\bar{\Delta}_i}, \qquad B_{T,h} = \sum_{i\neq i^\star}\frac{\log_+(T\bar{\Delta}_i^2)}{\bar{\Delta}_i}.$$

The new conditional shortfalls satisfy

$$c_{t,h} = \max_{i\neq i^\star}(\bar{\Delta}_i - \delta_{t,i})^+ \leq c_t + h.$$

Thus (82) supplies a budget $C_h = (N_T + hT)/2$. Applying Theorem 1, using (83), and completing the square gives

$$\begin{aligned}\mathrm{NR}_T &\leq M_2 B_{T,h} + M_2\sqrt{(N_T + hT)H_h/2} - \gamma N_T \\ &\leq M_2 B_{T,h} + M_2\sqrt{hTH_h/2} + \frac{M_2^2 H_h}{8\gamma}. \end{aligned} \tag{86}$$

The algorithm is unchanged: h and the larger reference gaps are used only in the analysis. Raising tiny gaps reduces the complexity H_h at the cost of an additional shortfall budget.

For a concrete example, let $K \geq 6$, $T = K^4$, and $\gamma = 1/26$. Prescribe one row gap K^{-10} and $K - 2$ row gaps equal to $1/26$. With $h = K^{-2}$,

$$H_h = K^2 + 26(K-2) = O(K^2), \qquad B_{T,h} = K^2 + 26(K-2)\log_+(K^4/26^2) = O(K^2).$$

Also $\sqrt{hTH_h} = O(K^2)$, so (86) gives $\mathrm{NR}_T = O(K^2)$ uniformly over games with these gaps and all admissible opponents. In contrast, the original $H = K^{10} + 26(K-2)$ gives

$$\min\{T, \sqrt{KT}, H/\gamma\} = K^{5/2}.$$

Hence a universal lower bound of this latter order cannot hold for every prescribed gap multiset. In fact, taking $m = 1$ in (16) gives $\Omega(K^2)$ for this profile, so its minimax rate is $\Theta(K^2)$. This finite-horizon example does not exclude an H/γ lower bound at horizons sufficiently large relative to the full gap profile.

J.5 The value-gap branch of the earlier upper bound

Let $v^{\mathrm{pure}} = \max_i \min_y u(i, y)$ and let v^{Nash} be the mixed-strategy value. Set $\zeta = v^{\mathrm{Nash}} - v^{\mathrm{pure}}$. For general games, Nash regret uses the benchmark v^{Nash}, whereas pure-strategy maximin regret uses v^{pure}:

$$\mathrm{PSMR}_T = \mathbb{E}\sum_{t=1}^{T}[v^{\mathrm{pure}} - u(I_t, Y_t)] = \mathrm{NR}_T - \zeta T.$$

The argument of Ito et al. (2026, Theorem 1) also gives, for any game with $\zeta > 0$,

$$\mathrm{PSMR}_T \le M_2\sqrt{KT} - \zeta T \le \frac{M_2^2 K}{4\zeta}.$$

Indeed, an equilibrium mixed row guarantees v^{Nash} against every opponent, so $\mathrm{NR}_T \le \mathrm{Reg}_T$. The $O(K/\zeta)$ bound concerns PSMR, not Nash regret; it is inherited from the adversarial guarantee and is not a new rate improvement.

The sufficient condition is $\zeta > 0$. Section 3.1 of Ito et al. (2026) states that absence of a pure equilibrium implies $\zeta > 0$; this implication does not hold for every finite game, as the following example shows. The game

$$U = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

has $v^{\mathrm{pure}} = v^{\mathrm{Nash}} = 1/2$: the first row guarantees $1/2$, and the uniform column mixture gives every row reward $1/2$. Neither column forms a pure saddle point, since its maximizing row receives zero in the other column. Thus there is no pure equilibrium although $\zeta = 0$. Our strict-equilibrium refinement does not use the implication contradicted by this example.

K A finite-horizon game lower bound for prescribed row gaps

We extend the two-game construction of Ito et al. (2026, Theorem 2 and Appendix C) to an arbitrary prescribed multiset of row gaps. The construction hides the equilibrium row among a selected group of nearly optimal rows. Rows outside this group have the remaining prescribed gaps and reveal no information about the hidden row. All gaps and regret in the statement use rewards normalized to $[0, 1]$.

Lemma 17 (Finite-horizon lower bound for prescribed gaps)**.** *Fix integers $K \ge 2$, $T \ge 1$, sorted row gaps $0 < \Delta_{(1)} \le \cdots \le \Delta_{(K-1)} \le 1/26$, and $0 < \gamma \le 1/26$. For every uninformed bandit learner, there exist a $K \times 2$ game with utilities in $[0, 1]$, a strict pure-strategy Nash equilibrium whose row-gap multiset is exactly $\{\Delta_{(1)}, \ldots, \Delta_{(K-1)}\}$ and whose column gap equals γ, and an oblivious opponent for which*

$$\mathrm{NR}_T \ge \frac{1}{12800} \max_{1 \le m \le K-1} \min\left\{T, \sqrt{(m+1)T}, \frac{m+1}{\Delta_{(m)}\gamma}\right\}. \tag{87}$$

The rewards can be chosen Bernoulli conditional on the action pair. The learner may know T, γ, and the row-gap multiset, but not its assignment to the rows.

Proof. Fix $m \in \{1, \ldots, K-1\}$ and write $J = m + 1$. We prove the corresponding entry of the maximum; choosing a maximizing m then proves the statement. Construct rewards in $[-1, 1]$ first,

writing $d_j = 2\Delta_{(j)}$, $D = d_m$, and $g = 2\gamma$, so $0 < d_j, g \le 1/13$. Let S consist of J candidate rows. For each $a \in S$, assign the gap multiset $\{d_1, \dots, d_m\}$ to $S \setminus \{a\}$, writing $d_i^{(a)}$ for the assigned gap. Assign the remaining gaps to the rows outside S, writing d_i^{out} for these fixed assignments. Define game a by

$$(\widetilde{U}_{i,1}^{(a)}, \widetilde{U}_{i,2}^{(a)}) = \begin{cases} (0, g), & i = a, \\ (-d_i^{(a)}, z_i^{(a)}), & i \in S \setminus \{a\}, \\ (-d_i^{\mathrm{out}}, -1), & i \notin S, \end{cases} \qquad z_i^{(a)} = \frac{(1-\varepsilon)d_i^{(a)} - 4s}{\varepsilon}.$$

Below we choose $\varepsilon \in (0, 1]$ and $s > 0$ so that $z_i^{(a)} \in [-1, 1]$. Every game then has Nash value zero and strict pure equilibrium $(a, 1)$, with the prescribed row-gap multiset and column gap g.

For an integer $1 \le n \le T$, the opponent independently chooses column 2 with probability ε during the first n rounds and always chooses column 1 afterwards. Conditional on the action pair, the reward takes values in $\{-1, +1\}$ with the prescribed game entry as its mean, independently of the past. Write $\mathsf{B}_{\pm}(x)$ for this law with mean x. Since the opponent's action is unobserved, the conditional prefix reward means are

$$\alpha = \varepsilon g \quad \text{on row } a, \qquad \beta = -4s \quad \text{on } S \setminus \{a\}, \qquad \beta_i^{\mathrm{out}} = -(1-\varepsilon)d_i^{\mathrm{out}} - \varepsilon \quad \text{on } i \notin S.$$

Our choices satisfy $\alpha \le s$ and $\varepsilon \ge 4s$, so $\beta_i^{\mathrm{out}} \le \beta$. Thus a pull outside S costs at least as much Nash regret as a nonequilibrium pull inside S, and its law is identical in every candidate game.

We first bound how often the learner finds the equilibrium row. Suppose $\alpha, \beta \in [-1/2, 1/2]$ and $n\delta^2 \le J/8$, where $\delta = \alpha - \beta > 0$. Let P_a denote the learner's prefix history law in game a, and let Q be the auxiliary law with mean β on every row in S and means β_i^{out} outside S. The auxiliary law is used only for comparison and need not belong to the game class. Write $N_a = \sum_{t=1}^{n} \mathbf{1}\{I_t = a\}$. Only row a differs between Q and P_a, so the chain rule gives

$$\mathrm{KL}(Q \| P_a) \le \delta^2 \mathbb{E}_Q N_a.$$

Here we use $\mathrm{KL}(\mathsf{B}_{\pm}(x) \| \mathsf{B}_{\pm}(y)) \le (x-y)^2$ for $x, y \in [-1/2, 1/2]$: the second derivative with respect to x is $1/(1-x^2) \le 4/3$, and the value and first derivative vanish at $x = y$. Since $0 \le N_a \le n$ and $\sum_{a \in S} \mathbb{E}_Q N_a \le n$, Pinsker's inequality and then Jensen's inequality yield

$$\frac{1}{J} \sum_{a \in S} \mathbb{E}_{P_a} N_a \le \frac{n}{J} + \frac{n\delta}{J} \sum_{a \in S} \sqrt{\frac{\mathbb{E}_Q N_a}{2}} \le \frac{n}{J} + n\delta \sqrt{\frac{n}{2J}} \le \frac{3n}{4}.$$

Thus some game has at least $n/4$ expected prefix pulls of nonequilibrium rows. Because all such rows have prefix mean at most $\beta = -4s$, its unnormalized total Nash regret satisfies

$$\widetilde{\mathrm{NR}}_T \ge -n\alpha + \delta\big(n - \mathbb{E}_{P_a} N_a\big) \ge \frac{n}{4}(-3\alpha - \beta) \ge \frac{ns}{4}. \tag{88}$$

The regret after the prefix is nonnegative because the opponent then plays column 1. This comparison applies to any fixed learner policy, including one supplied with T, g and the common row-gap multiset.

Set $\tau = J/(200D^2g^2) \ge 285$. If $T \ge \tau$, choose

$$n = \lfloor \tau \rfloor, \qquad s = Dg, \qquad \varepsilon = D(1 - 4g).$$

Then $n \ge \tau/2$, $\varepsilon \in (0, 1]$, and $\varepsilon \ge 4s$ because $g \le 1/13 < 1/8$. The second-column entry $z_i^{(a)}$ is increasing in $d_i^{(a)} \in (0, D]$, so

$$-\frac{4}{9} \le -\frac{4g}{1-4g} \le z_i^{(a)} \le 1 - D \le 1.$$

All game entries therefore lie in $[-1, 1]$. Also $0 < \alpha = s(1-4g) \le s$, $\beta = -4s$, and $0 < \delta \le 5s$, so $\alpha, \beta \in [-1/2, 1/2]$ and $n\delta^2 \le 25\tau D^2 g^2 = J/8$. Equation (88) gives

$$\widetilde{\mathrm{NR}}_T \ge \frac{nDg}{4} \ge \frac{J}{1600Dg}. \tag{89}$$

If $T < \tau$, choose

$$n = T, \qquad s = \min\left\{\frac{1}{16}, \sqrt{\frac{J}{200T}}\right\}, \qquad \varepsilon = \min\{1, s/g\}.$$

The branch condition and $Dg \le 1/169 < 1/16$ imply $s > Dg$, so $\varepsilon \ge D$. Also $\varepsilon \ge 4s$ because $g \le 1/13$ and $4s \le 1/4$. Consequently

$$-1 \le -\frac{4s}{\varepsilon} \le z_i^{(a)} \le \frac{d_i^{(a)}}{\varepsilon} \le 1.$$

All game entries again lie in $[-1, 1]$. Moreover, $0 < \alpha = \varepsilon g \le s \le 1/16$, $\beta = -4s \ge -1/4$, and $T\delta^2 \le 25Ts^2 \le J/8$. Thus (88) implies

$$\widetilde{\mathrm{NR}}_T \ge \frac{Ts}{4} = \min\left\{\frac{T}{64}, \frac{\sqrt{JT}}{4\sqrt{200}}\right\} \ge \frac{1}{64}\min\{T, \sqrt{JT}\}. \tag{90}$$

Finally map every utility and reward x to $(1 + x)/2$. This gives utilities in $[0, 1]$ and Bernoulli rewards, divides all gaps and Nash regret by two, and preserves the strict pure equilibrium. A normalized learner is simulated above by transforming each reward before passing it to the learner. Since $Dg = 4\Delta_{(m)}\gamma$, the long-horizon bound becomes $J/(12800\Delta_{(m)}\gamma)$ and the short-horizon bound becomes $\min\{T, \sqrt{JT}\}/128$. These imply the desired bound for each m, and hence (87). All candidate games preserve the entire prescribed gap multiset, and their opponent is oblivious. □

For equal gaps, the maximum is attained at $m = K - 1$, recovering the uniform-gap lower bound in Theorem 5. For general gaps, choose $m_* \in \arg\max_m m/\Delta_{(m)}$, using H_* from (1). If $T \ge (m_* + 1)/(\Delta_{(m_*)}^2\gamma^2)$, both horizon caps for m_* in (87) are at least $(m_* + 1)/(\Delta_{(m_*)}\gamma) \ge H_*/\gamma$. Conversely, every term in the maximum is at most $2H_*/\gamma$, because $m+1 \le 2m$. Thus the right-hand side is $\Theta(H_*/\gamma)$ in this regime. The comparison in (14) shows that its gap dependence differs from H/γ by at most $\mathsf{h}_{K-1} \le 1 + \log(K - 1)$, with $H_* = H$ for uniform gaps. If $\Delta_{(K-1)}/\Delta_{(1)}$ is bounded by a universal constant, then $H_* = \Theta(H)$, so the lower bound matches H/γ up to constants in this long-horizon regime.

L A FINITE-HORIZON LOWER BOUND MATCHING THE GAME REFINEMENT

We supplement the product lower bound for prescribed gap multisets in Appendix K with a finite-horizon stochastic lower bound for uniform gaps. Together they give the matching rate for two-by-two games with prescribed gaps, without a consistency assumption on the learner.

For $K \ge 2$, let $\mathfrak{G}_K(\Delta, \gamma)$ be the class of $K \times 2$ games with rewards in $[0, 1]$, a strict pure equilibrium, every nonoptimal row gap exactly Δ, and column gap exactly γ. The feedback contains only the realized reward. For a learner $\mathcal{A}$, write $\mathcal{R}_T(\mathcal{A}; K, \Delta, \gamma)$ for the supremum of its Nash regret over this game class, all adaptive nonanticipating opponents, and all bounded conditionally unbiased reward laws. All lower bounds below continue to hold if the learner is told Δ, γ, and T; the matrix and its Nash value are not supplied.

The argument extends the two-point construction of Bubeck et al. (2013, Theorem 6) (see also Garivier et al. (2019, Theorem 8)), restated for Bernoulli rewards in $[0, 1]$ and a fixed opponent. Its logarithmic term has no K factor: exact uniform gaps couple the alternative row means.

Lemma 18 (A finite-horizon stochastic lower bound)**.** *For all integers $K \ge 2$ and $T \ge 1$, every $0 < \Delta, \gamma \le 1/4$, and every uninformed learner $\mathcal{A}$,*

$$\mathcal{R}_T(\mathcal{A}; K, \Delta, \gamma) \ge \frac{\log(1 + 8T\Delta^2)}{32\Delta}. \tag{91}$$

Proof. Consider the two games defined, for $i \in [K]$, by

$$A_{i1} = 1/2 - \Delta\mathbf{1}\{i \ne 1\}, \quad A_{i2} = A_{i1} + \gamma,$$
$$B_{i1} = 1/2 + \Delta\mathbf{1}\{i = 2\}, \quad B_{i2} = B_{i1} + \gamma.$$

Every entry lies in $[0,1]$. The strict equilibria are respectively $(1,1)$ and $(2,1)$, and every nonoptimal row has gap Δ in its own game. Both column gaps are γ. Let the opponent always play the first column, and give each observed reward the Bernoulli distribution with its matrix entry as mean, independently conditional on the chosen actions.

Let P_A, P_B be the laws of the full learner interaction, and let $N_i = \sum_{t=1}^{T} \mathbf{1}\{I_t = i\}$ and $N_{\neq i} = T - N_i$. The two regrets are

$$R_A = \Delta \mathbb{E}_A N_{\neq 1}, \qquad R_B = \Delta \mathbb{E}_B N_{\neq 2}.$$

The first-row reward distribution is unchanged. The chain rule for relative entropy (Lattimore & Szepesvári, 2020, Lemma 15.1) therefore gives

$$\mathrm{KL}(P_A \| P_B) = \sum_{i \neq 1} d_i \mathbb{E}_A N_i \leq 16\Delta^2 \mathbb{E}_A N_{\neq 1}, \qquad d_i = \mathrm{kl}(A_{i1}, B_{i1}). \tag{92}$$

Here kl is the relative entropy between Bernoulli laws. For the second row,

$$d_2 = 2\Delta \log \frac{1+2\Delta}{1-2\Delta} \leq \frac{8\Delta^2}{1-4\Delta^2} \leq \frac{32}{3}\Delta^2 \leq 16\Delta^2,$$

using $\log((1+x)/(1-x)) \leq 2x/(1-x^2)$ for $0 \leq x < 1$. For $i \geq 3$, the bound $\mathrm{kl}(p,q) \leq (p-q)^2/(q(1-q))$ gives $d_i \leq 4\Delta^2$.

Apply the Bretagnolle–Huber inequality (Bretagnolle & Huber, 1979), in the form of Lattimore & Szepesvári (2020, Theorem 14.2), to the event $\{I_t \neq 1\}$ under the full interaction laws. For every t,

$$P_A(I_t \neq 1) + P_B(I_t = 1) \geq \tfrac{1}{2} \exp\{-16\Delta^2 \mathbb{E}_A N_{\neq 1}\}.$$

Since choosing row 1 is suboptimal in game B, summing over all rounds and setting $r = \max\{R_A, R_B\}$ yields

$$2r \geq R_A + R_B \geq \frac{\Delta T}{2} \exp\{-16\Delta R_A\} \geq \frac{\Delta T}{2} \exp\{-16\Delta r\}.$$

For $z = 16\Delta r \geq 0$, this means $ze^z \geq 4T\Delta^2$. Since $e^{2z} - 1 = 2e^z \sinh z \geq 2ze^z$, we conclude that

$$z \geq \tfrac{1}{2} \log(1 + 8T\Delta^2), \qquad r \geq \frac{\log(1+8T\Delta^2)}{32\Delta}.$$

Both instances and the fixed opponent belong to the specified class, which proves the claim. □

For $K = 2$, the matching part of Theorem 5 asserts that, for every integer $T \geq 1$ and every $0 < \Delta, \gamma \leq 1/26$,

$$\inf_{\mathcal{A}} \mathcal{R}_T(\mathcal{A}; 2, \Delta, \gamma) = \Theta\left(\min\left\{ \sqrt{T}, \frac{\log_+(T\Delta^2)}{\Delta} + \frac{1}{\Delta\gamma} \right\} \right). \tag{93}$$

The infimum is over all uninformed learners. Either version of Tsallis-INF attains the upper bound.

Proof of Theorem 5. For general $K \geq 2$, Lemma 17 proves (16), with constant $c_{\mathrm{p}} = 1/12800$. We now prove the matching claim for $K = 2$. Its upper bound is Corollary 4, because $H = 1/\Delta$. For its lower bound, note that $T \geq 1$ implies

$$\min\left\{ T, \sqrt{2T}, \frac{2}{\Delta\gamma} \right\} \geq \min\left\{ \sqrt{T}, \frac{1}{\Delta\gamma} \right\}.$$

Thus the same lemma gives

$$\mathcal{R}_T(\mathcal{A}; 2, \Delta, \gamma) \geq c_{\mathrm{p}} \min\{\sqrt{T}, 1/(\Delta\gamma)\}. \tag{94}$$

Fix a learner and abbreviate

$$s = \sqrt{T}, \quad x = T\Delta^2, \quad a = \frac{\log_+ x}{\Delta}, \quad b = \frac{1}{\Delta\gamma}, \quad g = \frac{\log(1+x)}{\Delta}.$$

Lemma 18 and (94) imply

$$\mathcal{R}_T(\mathcal{A}; 2, \Delta, \gamma) \geq c_* \max\{g, \min\{s, b\}\}, \qquad c_* = \min\{1/32, c_{\mathrm{p}}\}. \tag{95}$$

If $x < 1$, then $b/s = 1/(\gamma\sqrt{x}) > 1$, so the second entry of this maximum equals s, which is at least $\min\{s, a + b\}$. If $x \geq 1$, then

$$\log_+ x = \max\{1, \log x\} \leq 2\log(1 + x),$$

because $2\log(1 + x) \geq 2\log 2 > 1$ and $x \leq (1 + x)^2$. Consequently $a \leq 2g$ and

$$\min\{s, a + b\} \leq 2g + \min\{s, b\} \leq 3\max\{g, \min\{s, b\}\}.$$

In both cases (95) bounds the target expression from below up to a universal constant. Taking the infimum over learners proves the result. □

The two lower bounds may be witnessed by different games with the same prescribed gaps. This is sufficient for the minimax statement (93); it is not a lower bound for each fixed payoff matrix. The product lower bound holds for arbitrary prescribed gap multisets; full matching for $K > 2$ or for the finer coefficient χ is not claimed. For $K = 2$, the matching logarithmic stochastic term is part of the lower bound, while the unnecessary logarithmic multiplier on the $1/(\Delta\gamma)$ term has been removed from the upper bound.

M THE CONVERSES OF THE MODEL INCLUSIONS CAN FAIL

The implications in (3) do not identify the three classes. The following two-arm examples use $T = 2$, $i^\star = 1$, $\Delta_2 = \Delta \in (0, 1/2]$, and $C = 0$. All losses lie in $[0, 1]$ and the environments are nonanticipating. These examples separate the classes in general; we do not claim strict inclusion for every choice of the parameters.

Conditional gaps need not come from uncorrupted i.i.d. losses. Take the deterministic sequence

$$\ell_1 = (0, \Delta), \qquad \ell_2 = (1/2, 1/2 + \Delta).$$

Both conditional gaps equal Δ, for every learner, so this environment is in $\mathsf{CG}(0)$. If it admitted a representation in $\mathsf{Cor}(0)$, nonnegativity of the corruption cost would force $\ell_t^0 = \ell_t$ almost surely at each round. The two clean loss vectors would then have different deterministic distributions, contradicting the i.i.d. requirement. Thus it is not in $\mathsf{Cor}(0)$.

Self-bounding need not give a conditional-gap budget. Set

$$\ell_1 = (0, 2\Delta), \qquad \ell_2 = \begin{cases} (0, 0), & I_1 = 2, \\ (0, \Delta), & I_1 = 1. \end{cases}$$

The second loss vector depends only on the previous action. Arm 1 has zero loss on both rounds, and all losses are nonnegative, so it is a best fixed arm for every learner. For any learner,

$$\mathrm{Reg}_2 = 2\Delta \Pr(I_1 = 2) + \Delta \Pr(I_1 = 1, I_2 = 2),$$
$$\mathrm{Reg}_2 - \Delta\mathbb{E}[\mathbf{1}\{I_1 = 2\} + \mathbf{1}\{I_2 = 2\}] = \Delta \Pr(I_1 = 2, I_2 = 1) \geq 0.$$

Thus the self-bounding constraint holds uniformly over learners, placing the environment in $\mathsf{SB}(0)$. It also holds at the first-round prefix. In contrast, $c_1 = 0$ and $c_2 = \Delta\mathbf{1}\{I_1 = 2\}$, so

$$\mathbb{E}(c_1 + c_2) = \Delta \Pr(I_1 = 2).$$

This is positive for any learner that initially explores arm 2 (in particular, it equals $\Delta/2$ for either Tsallis-INF version). Hence the environment is not in $\mathsf{CG}(0)$. The initial excess gap pays for a later shortfall in the self-bounding inequality, whereas conditional gaps charge the shortfall even when such compensation is available.

N Related work

We organize the comparison around four questions: how corruption is measured, where the logarithm enters existing regret bounds, what a direct trajectory analysis controls, and how game-learning guarantees depend on feedback and opponent assumptions. Table 1 summarizes the closest upper and lower bounds; Appendix I provides the detailed comparison with existing self-bounding bounds.

N.1 Corruption models and regret objectives

Lykouris et al. (2018) introduced a model in which an adversary corrupts an otherwise stochastic bandit sequence. Comparisons within this literature require specifying the regret objective as well as the corruption budget. The clean pseudo-regret $R_T^0 = \mathbb{E}\sum_t \Delta_{I_t}$ measures the cost of the learner's choices under the original means, even when its observations have been changed. For this objective, BARBAR (Gupta et al., 2019) incurs an additive $O(KC)$ corruption cost, and BARBAT (Hu & Chen, 2025) removes the factor K. The linear dependence on C is unavoidable for clean regret (Gupta et al., 2019, Section 5.1).

Our objective instead uses the observed losses, as in the self-bounding analyses of Zimmert & Seldin (2021); Masoudian & Seldin (2021). Corruption can both mislead the learner and reduce the loss difference between the selected and reference arms. Consequently its effect on this regret need not be linear in its budget. The distinction is already visible with full information: Amir et al. (2020) bound clean regret for corrupted expert advice, whereas Ito (2021a) proves matching square-root corruption dependence for regret on the corrupted losses. In our bandit model, $|R_T^0 - R_T| \le 2C$, so the new theorem gives a clean-regret bound with an additional $2C$ term. It does not contradict clean-regret lower bounds. The realized-regret and positive-part guarantees studied by Lykouris et al. (2018) also differ from our signed expected pseudo-regret.

The environment assumptions matter separately from this choice of objective. Our corruption model allows the adversary to see the clean loss vector and the learner's distribution before choosing the observed vector, but the current arm is sampled afterward. The proof uses only the resulting budget on conditional gap shortfalls. At zero budget, this is a per-arm version of the fixed conditional-gap condition in Wei & Luo (2018). Section 2.3 and Appendix M distinguish this per-round condition from the terminal self-bounding inequality: the latter can offset a late gap shortfall with an earlier excess. Our logarithm-free bound uses the former structure, and its extension to the full self-bounding class remains open.

N.2 Best-of-both-worlds learning and self-bounding analysis

Best-of-both-worlds bandit algorithms obtain stochastic and adversarial guarantees without being told which environment they face (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017). Tsallis-INF (Zimmert & Seldin, 2019; 2021) achieves both with a single FTRL update. Its analysis first bounds regret by a quantity depending on the sampling probabilities, then uses the self-bounding condition to control that quantity. This also yields a corrupted-stochastic guarantee, but with a logarithm inside the square-root corruption term. Masoudian & Seldin (2021) sharpen the stochastic and corruption-dependent logarithms for the same algorithm. Appendix I derives the baseline used here from their general conversion and our stated stability bound. Theorem 1 changes the analysis of the trajectory while keeping the IW and RV algorithm specifications unchanged.

For multiple optimal arms, Ito (2021b, Theorem 2) establishes logarithmic stochastic regret for Tsallis-INF with IW estimates. Jin et al. (2023) subsequently treat general regularizers, using a new learning-rate schedule and an additional log-barrier term. Both multi-armed bandit guarantees under corruption retain logarithmic dependence inside the square-root term and include an additional term for conditional expected loss differences among zero-gap arms. Our positive reference gaps exclude that difficulty, so our result does not subsume these multiple-optimum guarantees.

As an alternative presentation of the basic Tsallis-INF analysis, Lee & Orabona (2025) give a proof of its stochastic and adversarial guarantees without Fenchel conjugates. That simplification concerns the underlying best-of-both-worlds proof; the additional issue here is separating the corruption cost from the stochastic logarithm.

Beyond multi-armed bandits, Dann et al. (2023) give reductions that turn suitable adversarial guarantees into best-of-both-worlds bounds, with applications to linear and contextual bandits, graph feedback, and MDPs. Jin & Luo (2020) use self-bounding arguments for episodic MDPs with known transitions. Their square-root corruption bounds retain the logarithmic dependence of the stochastic bound. Partial monitoring (Tsuchiya et al., 2023) further changes the feedback structure: locally and globally observable problems have different corruption dependence, so they do not share a universal square-root interpolation formula. These works motivate asking whether the logarithm can be removed beyond ordinary bandits. Such an extension requires an analogue of the trajectory control proved here; a terminal regret bound alone does not supply it.

N.3 Direct analyses of the Tsallis-INF trajectory

Zhan et al. (2026b) study the stochastic trajectory of FTRL with $1/2$-Tsallis entropy through Lyapunov functions of the estimated-gap process. Under their learning-rate assumptions, they obtain $O(t^{-1})$ sampling simple regret, which measures the expected gap of the arm sampled at time t, and allow multiple optimal arms. Zhan et al. (2026a) instead study best-arm identification, bounding the probability that the arm minimizing the cumulative importance-weighted loss estimates is not the unique optimal arm. These results control the algorithm at a specified time, rather than only its cumulative regret, and give a direct view of its sampling behavior in a stochastic environment.

Our analysis shares the focus on the trajectory but uses a different target and environmental assumption. We bound $S_T = \mathbb{E}\sum_{t,i\neq i^\star}\sqrt{p_{t,i}/t}$, the cumulative quantity in the FTRL stability bound, under a budget on conditional gap shortfalls. Conditional gaps may become negative and the reference arm may temporarily cease to be optimal. The potential must therefore control both the sampling probabilities and their response to those shortfalls, with per-arm reference-gap dependence. The cited stochastic trajectory results do not provide such a corruption guarantee; conversely, our bound on S_T does not assert their last-iterate or identification guarantees.

N.4 Learning in zero-sum games

For game learning, the relevant distinctions are the opponent's behavior, the feedback, and the regret comparator. With an arbitrary opponent, Nash regret measures shortfall from the value of the game. O'Donoghue et al. (2021); Maiti et al. (2025) study this objective with noisy bandit payoff observations and access to the opponent's realized action. Observing that action identifies the matrix entry being sampled. In the uninformed setting considered by Ito et al. (2026) and here, the learner observes its payoff but neither the payoff matrix nor the opponent's action.

The direct comparison is with Ito et al. (2026), who study pure-strategy maximin regret under both informed and uninformed bandit feedback. When a pure Nash equilibrium exists, the pure maximin value equals the Nash value, so this is Nash regret. Their logarithmic guarantee for the uninformed setting with a strict pure equilibrium charges deviations from the equilibrium column as corruption of a stochastic bandit problem. Section 4 uses the same reduction with our improved corruption bound. The new bandit estimate removes the logarithm from the H/γ term; the reduction and the underlying algorithm are inherited from that work. Our game lower bounds complement this upper bound and give a full matching rate for 2×2 games in the stated small-gap regime (Theorem 5).

A different line of work exploits regularity in the learning dynamics of both players. Ito et al. (2025) prove instance-dependent external-regret bounds when both players use Tsallis-INF under bandit feedback, with logarithmic regret in games with a unique pure equilibrium. This accelerated guarantee concerns self-play, whereas our Nash-regret bound holds against an arbitrary opponent. Tsuchiya et al. (2025) interpolate between prescribed learning dynamics and adversarial deviations under full-information feedback. They budget deviations from the prescribed strategies and corruption of the observed utility vectors. Our game reduction instead measures the opponent's deviations from an equilibrium column and requires no prescribed algorithm for that opponent. Its effective bandit corruption can arise even in a fixed game with uncorrupted payoffs.