BEST-OF-BOTH-WORLDS CAUSAL BANDITS WITH MEDIATOR FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a bandit setting described by a causal graph $A \to Z \to \ell \leftarrow U$, where A is an action, Z is an observable mediating state separating A from the loss ℓ, and U is an unobserved variable. The setting has been proposed and studied independently in the causality literature under the name of *causal bandits with post-action context*, and in the classical bandit literature under the name of *bandits with mediator feedback*. Similar to both bodies of literature, we assume that the mapping $A \to Z$ is fixed. We propose a best-of-both-worlds algorithm, *Causal Tsallis-INF*, which provides a worst-case regret guarantee in case of adversarial U and instance-dependent regret guarantees for stochastic U, as well as other forms of U for which the regret satisfies the self-bounding constraint of Zimmert & Seldin (2021). Our setting matches the one studied by Eldowa et al. (2024) in the classical bandit literature and generalizes the setting studied in the causal bandit literature, where U is assumed to be stochastic. We advance both strands of literature. Our stochastic regret guarantee improves the logarithmic dependence on the horizon in the EXP4 analysis of Eldowa et al. (2024) and adapts to both action and attainable mediator complexity. Our adversarial guarantee also extends beyond the stochastic assumptions underlying the causal bandit methods. We further provide an empirical evaluation demonstrating that for stochastic U our algorithm performs comparably to state-of-the-art Causal Thompson Sampling and Causal UCB2 algorithms, and that, in contrast to our algorithm, the latter are not robust to non-stochastic U, at the same time, our algorithm consistently outperforms the EXP4 algorithm in all the experiments.

1 INTRODUCTION

We study a bandit setting described by a causal graph $A \to Z \to \ell \leftarrow U$, where A is an action, Z is an observable mediating state separating A from the loss ℓ, and U is an unobserved variable also influencing the loss ℓ. For example, a choice of an education program (A) leads to a set of skills (Z), which in turn impact employment opportunities and wage (ℓ). However, employment and wage may be subject to additional factors (U) affecting the job market, which are independent of A and Z and unknown at the time of taking the action. As another example, a medical treatment (A), e.g., a vaccine, may lead to a known response (Z), e.g., production of antibodies, which in turn will lead to clinical outcome (ℓ). However, clinical outcome may depend on additional factors (U), e.g., virus mutations, which are unknown at the time of taking the action. Importantly for us, different actions A and A' may have an overlap in the set of states Z they produce, leading to information sharing, which can be exploited for faster learning.

The setting has been studied independently in the causality literature under the name *causal bandits* (*with post-action context*) in pure exploration (Lattimore et al., 2016; Sen et al., 2017; Maiti et al., 2022; Shahverdikondori et al., 2026) and cumulative-regret minimization settings (Lu et al., 2020; Nair et al., 2021; Bilodeau et al., 2022; Liu et al., 2024), and in the classical bandit literature under the name of *bandits with mediator feedback* (Papini et al., 2019; Metelli et al., 2021; Eldowa et al., 2024).

For cumulative-regret minimization, Lu et al. (2020) proposed Causal UCB and Causal Thompson Sampling algorithms. They derived a $\widetilde{O}(\sqrt{|\mathcal{Z}|T})$ expected pseudo regret guarantee for their Causal UCB algorithm, where $|\mathcal{Z}|$ is the cardinality of the mediator space $\mathcal{Z}$ and $\widetilde{O}$ notation hides

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.

logarithmic factors. For comparison, the expected regret of the standard UCB algorithm learning directly from A to ℓ (ignoring Z) would have regret growing with the number of actions (Auer et al., 2002). Thus, the result of Lu et al. (2020) shows that if the cardinality of the action space $|\mathcal{A}|$ satisfies $|\mathcal{Z}| \ll |\mathcal{A}|$, then mediator feedback can be helpful for reducing the regret. Nair et al. (2021) showed that expected cumulative regret can even remain bounded by a constant when the known action-to-mediator distributions share the same nonzero support.

Eldowa et al. (2024) relate the problem of bandits with mediator feedback to the problem of bandits with expert advice (Auer et al., 1995). In their setting, the set of actions $\mathcal{A}$ corresponds to a set of policies (experts), where each policy is a distribution on the set of arms $\mathcal{Z}$. At every round the learner observes advice of $|\mathcal{A}|$ experts in the form of distributions over arms $\mathcal{Z}$. The learner then picks an expert and samples an arm Z according to the proposed distribution, and plays that arm to obtain a loss ℓ. It corresponds to the classical adversarial bandits with expert advice, with a constraint that the learner follows the advice of the sampled expert, rather than a mixture of expert advice. Under this constraint, Eldowa et al. (2024) provide a refined analysis of the celebrated EXP4 algorithm of Auer et al. (1995).

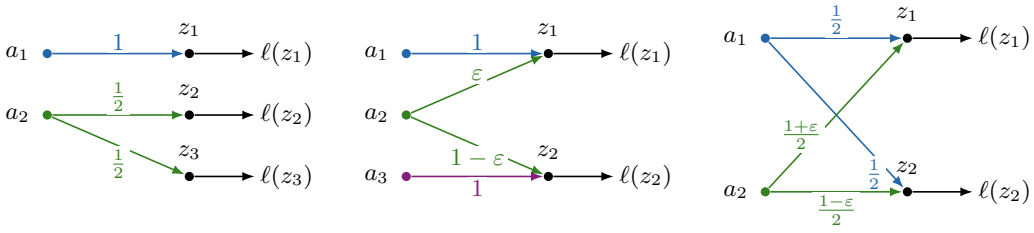
We assume that the mapping $A \to Z$ is known to the learner. In bandits with mediator feedback the assumption that the mapping is known is part of the problem formulation, as described above. In causal bandit literature the mechanism $A \to Z$ is assumed to be stable, and it is common to assume that it is known (Lattimore et al., 2016; Sen et al., 2017; Lu et al., 2020; Nair et al., 2021; Maiti et al., 2022; Bilodeau et al., 2022; Liu et al., 2024; Shahverdikondori et al., 2026), whereas in practice it can be learned from historical data.

Eldowa et al. (2024) assume that the mapping $Z \to \ell$ is adversarial, which in our setting is equivalent to allowing U to be adversarial. They also derive best-of-both-worlds regret guarantees, which provide instance-dependent bounds when $Z \to \ell$ happens to be stochastic, with no need of prior knowledge of the nature of the link. In contrast, causal bandits assume that U is stochastic, and we show that existing causal bandit algorithms fail if this assumption is violated.

As observed in prior work, access to mediator feedback Z may be used to reduce regret growth rate, but in some cases it is still more efficient to learn a direct mapping from A to ℓ. For example, in the adversarial setting the regret bound of EXP4 is $\sqrt{|\mathcal{Z}|T\log|\mathcal{A}|}$. However, if $|\mathcal{A}| \ll |\mathcal{Z}|$, playing an adversarial bandit algorithm directly from A to ℓ (e.g., EXP3 (Auer et al., 1995)) leads to tighter $\sqrt{|\mathcal{A}|T\log|\mathcal{A}|}$ regret guarantee. Similarly, in the stochastic case, if suboptimality gaps in the action space $\mathcal{A}$ are larger than suboptimality gaps in the mediator space $\mathcal{Z}$, it may be more efficient to learn a direct mapping from A to ℓ. A challenge is to design algorithms that would automatically adapt to complexity of a learning instance. In the adversarial setting, McMahan & Streeter (2009) have designed a modification of the EXP4 algorithm achieving $O(\sqrt{\mathcal{C}_0 T\log|\mathcal{A}|})$ regret, where $\mathcal{C}_0$ is a notion of complexity of the mapping $A \to Z$ satisfying $\mathcal{C}_0 \leq \min\{|\mathcal{A}|, |\mathcal{Z}|\}$. Eldowa et al. (2024) have improved the analysis of the original EXP4 algorithm and achieved $O(\sqrt{\mathcal{C}T\log|\mathcal{A}|})$ regret guarantee, where $\mathcal{C}$ is a capacity of the channel $A \to Z$ satisfying $\mathcal{C} \leq \min\{|\mathcal{A}|, |\mathcal{Z}|\}$. Additionally, they have shown that EXP4 achieves a refined (although suboptimal) regret guarantee in terms of suboptimality gaps in the action space $\mathcal{A}$.

In Figure 1 we provide three graphical examples illustrating stochastic cases, where mediator feedback is helpful for reducing the regret, and cases where learning directly from A to ℓ is more efficient. Note that individual graphical elements illustrated in Figure 1 can co-occur in a larger graph, making the complexity of learning dependent on both the known structure of the mapping $A \to Z$ and the unknown outcome mapping $Z \to \ell$. Therefore, designing an algorithm that would adapt to the right form of instance-dependent complexity is a highly non-trivial task. Prior work was only able to exploit small cardinality of $\mathcal{Z}$ and large gaps in the action space, whereas we propose an algorithm that is capable of adapting to the gaps in both the action and the mediator spaces, as well as to the cardinality of both.

Our algorithm, which we name *Causal Tsallis-INF*, is based on Follow-the-Regularized-Leader (FTRL) framework. We design hybrid regularization in the action space and the mediator space. Hybrid regularization has been used in best-of-both-worlds literature in the past (e.g., Zimmert et al., 2019; Zimmert & Seldin, 2020), however, designing an appropriate hybrid regularizer is a highly delicate and non-trivial task, because various parts of the regularizer have to work well together. The EXP4 algorithm analyzed by Eldowa et al. (2024) is based on entropic regularization, which is



(a) Complexity determined by $\Delta_{\mathcal{A}}$. (b) Complexity determined by $\Delta_{\mathcal{Z}}$. (c) Full observability.

Figure 1: The figure provides examples of three major cases illustrating key factors influencing the complexity of learning in the case of stochastic losses. Edge labels are transition probabilities, with $0 < \varepsilon < 1$, and $\ell(z)$ denotes the loss given mediating state z. (a) In this example the expected loss of a_2 is $(\ell(z_2) + \ell(z_3))/2$. The gap between the expected losses of z_2 and z_3 can be arbitrarily small, but it is irrelevant for learning, because the learner cannot target z_2 and z_3 individually. In this case, the regret scales with the inverse of the suboptimality gap in the action space, $\Delta_{\mathcal{A}}$. Large $\Delta_{\mathcal{A}}$ leads to small regret even if $\Delta_{\mathcal{Z}}$ is small. (b) In this case, a_2 can never be the optimal action, however, depending on ε, the gap $\Delta_{\mathcal{A}}$ in the action space can be arbitrarily small. However, the regret growth scales with the inverse of the gap $\Delta_{\mathcal{Z}}$ in the mediator space $\mathcal{Z}$, regardless of the value of $\Delta_{\mathcal{A}}$. (c) In this case, playing either of the two actions reveals information about each of the two states, therefore, there is no need for exploration (no need to play suboptimal actions). Thus, the game is reminiscent of a full information game, and the regret is bounded by an instance-dependent constant (not growing with time).

known to be suboptimal for bandit problems (Zimmert & Seldin, 2021), which is the reason why their stochastic regret bounds are suboptimal by a multiplicative $\log T$ factor. We have used Tsallis entropy regularization with power $\frac{1}{2}$, which is the one that was used by Zimmert & Seldin (2021) to achieve the optimal best-of-both-worlds rates in standard multiarmed bandits. In order to regularize in both the action and the mediator spaces, we had to combine regularization of the distribution over actions with regularization of the induced distribution over mediating states. All prior literature known to us has built hybrid regularizers by summing up individual elements, however, this approach has not worked for our problem. Instead, we had to regularize by the maximum of the two (negative-entropy) regularizers.

We provide additional related literature review in Appendix A.1.

Table 1 provides a comparison of our results to the closest prior work. Below we highlight the main contributions of the paper.

- **Connecting causal bandits and classical bandits.** We show that the causal bandit setting (Lattimore et al., 2016; Lu et al., 2020) and the classical bandits with expert advice setting proposed by Auer et al. (1995) and reinterpreted by Eldowa et al. (2024) are closely related, and that algorithms designed for the latter setting can be applied in the former.
- **Proposing Causal Tsallis-INF - a best of multiple worlds algorithm achieving instance-dependent adaptivity in terms of gaps in the action and in the mediator space, guaranteeing adversarial robustness, and providing refined regret guarantees in intermediate regimes.** We propose *Causal Tsallis-INF*, an FTRL algorithm based on hybrid regularization achieving logarithmic regret scaled by the minimum of inverse suboptimality gaps in the action and mediator space. Additionally, in cases where the optimal action maps to all mediating states with non-negligible probability, the regret of the algorithm is bounded by an instance-dependent constant (not growing with time). Furthermore, the algorithm provides worst-case regret that scales with the minimum of the cardinalities of action and mediator spaces for adversarial losses. The algorithm also provides refined regret guarantees in intermediate regimes, where the regret satisfies a *self-bounding constraint*, which includes adversarially constrained stochastic causal bandits and corrupted causal bandits.
- **Improving on Eldowa et al. (2024).** We improve on the results of Eldowa et al. (2024) in several ways. Most notably: (1) our regret analysis of the stochastic case adapts to the best of the action gap and mediator gap, whereas their analysis only adapts to the action

Table 1: **High-level comparison of pseudo-regret guarantees in our work to prior literature.** $S_{\mathcal{A}} = \sum_{a:\Delta_{\mathcal{A}}(a)>0} \frac{1}{\Delta_{\mathcal{A}}(a)}$, where $\Delta_{\mathcal{A}}(a)$ is a suboptimality gap of action a. $\Delta_{\min}$ is the minimum positive suboptimality gap in the action space, and a^* denotes the optimal action. All instance-dependent bounds in the table assume uniqueness of a^*, although we note that there exist techniques for eliminating this assumption (Ito, 2021). $\boldsymbol{M}$ is a known matrix encoding the stochastic mapping from $\mathcal{A}$ to $\mathcal{Z}$, with $M_{z,a} = \mathbb{P}(Z = z | \operatorname{do}(A = a))$. $S^{\mathrm{a}}_{\mathcal{Z}}$ is a complexity term depending on the structure of $\boldsymbol{M}$ and suboptimality gaps in the mediator space, defined in equation (9). $\mathcal{C}(\boldsymbol{M})$ denotes the chi-squared capacity of the channel defined by $\boldsymbol{M}$ and satisfies $0 \le \mathcal{C}(\boldsymbol{M}) \le \operatorname{rank}(\boldsymbol{M}) - 1 \le \min\{|\mathcal{A}|, |\mathcal{Z}|\} - 1$. The algorithm of Zimmert & Seldin (2021) in the first line of the table does not use mediator feedback Z.

Paper	**Stochastic**		**Adversarial**
	Instance-dependent	Gap-independent	Worst-case
Zimmert & Seldin (2021)	$O(S_{\mathcal{A}} \log T)$	$O\Big(\sqrt{\lvert\mathcal{A}\rvert T}\Big)$	$O\Big(\sqrt{\lvert\mathcal{A}\rvert T}\Big)$
Lu et al. (2020)	—	$\widetilde{O}\Big(\sqrt{\lvert\mathcal{Z}\rvert T}\Big)$	—
Nair et al. (2021)	$O(1)$, if $\forall_{a,z} M_{z,a} > 0$	—	—
Liu et al. (2024)	$O\Big(\frac{\operatorname{rank}(\boldsymbol{M}) \log(\lvert\mathcal{A}\rvert T)}{\Delta_{\min}}\Big)$	$\widetilde{O}\Big(\sqrt{\operatorname{rank}(\boldsymbol{M})\, T}\Big)$	—
Eldowa et al. (2024)	$O\Big(\frac{\mathcal{C}(\boldsymbol{M}) \log T \log(\lvert\mathcal{A}\rvert T)}{\Delta_{\min}}\Big)$	$\widetilde{O}\Big(\sqrt{\mathcal{C}(\boldsymbol{M})\, T}\Big)$	$\widetilde{O}\Big(\sqrt{\mathcal{C}(\boldsymbol{M})\, T}\Big)$
This Paper	$O(1 + \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\} \log T)$ $O(1)$, if $M_{z,a^*} > 0\ \forall z$	$\widetilde{O}\Big(\sqrt{\min\{\lvert\mathcal{A}\rvert, \lvert\mathcal{Z}\rvert\}\, T}\Big)$	$\widetilde{O}\Big(\sqrt{\min\{\lvert\mathcal{A}\rvert, \lvert\mathcal{Z}\rvert\}\, T}\Big)$

gap; (2) our regret in the stochastic setting scales with $\log T$ rather than $(\log T)^2$, because we are using Tsallis entropy regularization instead of entropic regularization behind EXP4, which is suboptimal; (3) we demonstrate experimentally that our Causal Tsallis-INF has dominantly better empirical performance than the EXP4 algorithm studied by Eldowa et al. (2024) in both the stochastic and adversarial cases.

- **Improving on existing causal bandit algorithms.** We improve on Causal UCB, Causal Thompson Sampling (Lu et al., 2020), Causal UCB2 (Nair et al., 2021), and Phase Elimination-Causal UCB (Liu et al., 2024) in their robustness to adversarial losses. We demonstrate experimentally that in certain non-stationary regimes the aforementioned algorithms exhibit substantially worse regret, whereas the regret of Causal Tsallis-INF stays under control.

2 Problem setup

We describe the problem setup in the classical bandit formalism, and then relate it to the causal setup. We let $\mathcal{A}$ be a finite set of actions and $\mathcal{Z}$ a finite set of mediating states. We let $\boldsymbol{M} \in [0,1]^{|\mathcal{Z}| \times |\mathcal{A}|}$ be a known matrix describing the mechanism $A \to Z$, where the columns of $\boldsymbol{M}$ are probability distributions on the mediating states, $\boldsymbol{M}_{\cdot,a} \in \Delta^{|\mathcal{Z}|-1}$, with $M_{z,a}$ being the probability of observing mediating state z when the learner plays action a.

Before the game starts, the environment produces a sequence of losses $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots$ with $\boldsymbol{\ell}_t \in [0,1]^{|\mathcal{Z}|}$, where $\ell_t(z)$ is the loss of mediating state z in round t. The game then proceeds according to the following protocol: (i) The learner picks an action A_t (potentially randomized); (ii) The environment samples $Z_t \sim \boldsymbol{M}_{\cdot,A_t}$ and reveals $(Z_t, \ell_t(Z_t))$ to the learner, and the learner suffers $\ell_t(Z_t)$.

We start with two basic forms of the game. In an *adversarial game* the environment generates $\boldsymbol{\ell}_t$ arbitrarily, and in a *stochastic game* the environment samples each $\boldsymbol{\ell}_t$ independently of the past and of the learner's actions, with a fixed mean $\boldsymbol{\mu}$, so that $\mathbb{E}[\boldsymbol{\ell}_t \mid \mathcal{H}_{t-1}] = \boldsymbol{\mu}$ for all t.

We let $\mathcal{H}_t = \{A_s, Z_s, \ell_s(Z_s)\}_{s=1}^t$ denote the history of the game up to the end of round t and let $\bar{\ell}_t(a) = \langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{\ell}_t \rangle$ denote the expected loss of round t over the mediator, given the loss vector $\boldsymbol{\ell}_t$ and assuming the learner would have picked action a in round t.

The performance of the learner is measured by pseudo-regret, defined as

$$\mathrm{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_t(A_t)\right] - \min_a \mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_t(a)\right],$$

where the expectation is with respect to randomization by the learner and the environment.

Relation to the causal setting. In the causal setting, the formulation above can be described by a causal graph $A \to Z \to \ell \leftarrow U$, where the matrix $\boldsymbol{M}$ represents the mechanism $A \to Z$ with $M_{z,a} := \mathbb{P}(Z = z \mid \mathrm{do}(a))$, and we can see the loss at time t as a function of Z_t and an arbitrary unobserved parameter U_t, so that $\ell_t(Z_t) = \ell(Z_t, U_t)$. Then, stochastic generation of U_t would correspond to the usual stochastic causal bandits, whereas adversarial generation of U_t before the beginning of the game would generalize it to the adversarial setting.

Suboptimality gaps (in the stochastic setting). When each $\boldsymbol{\ell}_t$ is generated from stationary distributions with mean $\boldsymbol{\mu}$, we let $\mu_{\min} = \min_{z\in\mathcal{Z}} \mu(z)$, $z^\star \in \arg\min_{z\in\mathcal{Z}} \mu(z)$, $\bar{\ell}(a) = \langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{\mu}\rangle$, and $a^\star \in \arg\min_{a\in\mathcal{A}} \bar{\ell}(a)$. We define *suboptimality gaps in the mediator space* (or simply *mediator gaps*)

$$\Delta_{\mathcal{Z}}(z) := \mu(z) - \mu_{\min}, \qquad z \in \mathcal{Z} \tag{1}$$

and *suboptimality gaps in the action space* (or simply *action gaps*)

$$\Delta_{\mathcal{A}}(a) := \bar{\ell}(a) - \bar{\ell}(a^\star) = \left\langle \boldsymbol{M}_{\cdot,a} - \boldsymbol{M}_{\cdot,a^\star}, \boldsymbol{\Delta}_{\mathcal{Z}} \right\rangle, \qquad a \in \mathcal{A}\,. \tag{2}$$

In the stochastic setting the pseudo-regret can then be rewritten in terms of suboptimality gaps. Let $\boldsymbol{p}_t$ be the distribution used by the algorithm to sample A_t and let $\boldsymbol{q}_t = \boldsymbol{M}\boldsymbol{p}_t$, then

$$\mathrm{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\Delta_{\mathcal{A}}(a)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{z\in\mathcal{Z}} \big(q_t(z) - M_{z,a^\star}\big)\Delta_{\mathcal{Z}}(z)\right]. \tag{3}$$

3 Causal Tsallis-INF

We start by recalling the Tsallis-INF algorithm of Zimmert & Seldin (2021), which is a best-of-both-worlds algorithm for standard multiarmed bandits, and a departure point for our work.

Tsallis-INF. The Tsallis-INF algorithm (Zimmert & Seldin, 2021) follows the follow-the-regularized-leader (FTRL) framework. At round t it samples A_t according to a distribution $\boldsymbol{p}_t^{\mathrm{TS}}$ that minimizes a parametrized trade-off between weighted cumulative loss estimates $\widehat{\widehat{\boldsymbol{L}}}_{t-1}^{\mathrm{TS}}$ and the negative Tsallis entropy with power $\frac{1}{2}$ of $\boldsymbol{p}_t^{\mathrm{TS}}$:

$$\boldsymbol{p}_t^{\mathrm{TS}} = \operatorname*{arg\,min}_{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}} \left\{ \left\langle \widehat{\widehat{\boldsymbol{L}}}_{t-1}^{\mathrm{TS}}, \boldsymbol{p} \right\rangle - \frac{4}{\eta_t}\sum_{a\in\mathcal{A}} \sqrt{p_a} \right\},$$

where η_t is the learning rate. Zimmert & Seldin (2021) consider both standard Importance-Weighted (IW) loss estimates and Reduced-Variance (RV) loss estimates. In the version with RV loss estimates (the one that we follow) the learning rate is $\eta_t = \frac{4}{\sqrt{t}}$ and the estimates are defined as

$$\widehat{\widehat{\ell}}_t^{\mathrm{TS}}(a) := B_t^{\mathcal{A}}(a) + \mathbb{1}(A_t = a)\frac{\ell_t(Z_t) - B_t^{\mathcal{A}}(a)}{p_t^{\mathrm{TS}}(a)}, \qquad B_t^{\mathcal{A}}(a) := \frac{1}{2}\mathbb{1}(p_t^{\mathrm{TS}}(a) \geq \eta_t^2)\,. \tag{4}$$

The cumulative estimates are updated by $\widehat{\widehat{\boldsymbol{L}}}_t^{\mathrm{TS}} = \widehat{\widehat{\boldsymbol{L}}}_{t-1}^{\mathrm{TS}} + \widehat{\widehat{\boldsymbol{\ell}}}_t^{\mathrm{TS}}$, starting from $\widehat{\widehat{\boldsymbol{L}}}_0^{\mathrm{TS}} = \boldsymbol{0}$. The baseline $B_t^{\mathcal{A}}(a)$ centers the sampled loss at one half when the sampling probability is sufficiently large, thus reducing the variance; otherwise, the estimator reduces to ordinary importance weighting. The RV loss estimates are unbiased, but can take negative values.

Causal Tsallis-INF We propose *Causal Tsallis-INF* algorithm. The algorithm combines three major elements: (i) using mediator feedback to share information across actions; (ii) Tsallis entropy regularization of the distribution on actions; and (iii) Tsallis entropy regularization of the induced distribution on mediating states. Next we explain these elements in more detail.

Sharing information through mediator feedback We use reduced-variance estimators at the *mediator level*. Given an action distribution $\boldsymbol{p}_t$, let $\boldsymbol{q}_t = \boldsymbol{M}\boldsymbol{p}_t$ be the induced mediator distribution. After observing $(Z_t, \ell_t(Z_t))$, define

$$\widehat{\ell}_t(z) := B_t(z) + \mathbb{1}(Z_t = z)\frac{\ell_t(z) - B_t(z)}{q_t(z)}, \qquad B_t(z) := \frac{1}{2}\mathbb{1}(q_t(z) \geq \eta_t^2), \tag{5}$$

where η_t is the learning rate. We will show that regularization in the action space ensures that $p_t(a) > 0$ for all actions. Without loss of generality there are no unreachable mediating states (no $M_{z,a} = 0$ for all a) and the estimator is well-defined and unbiased.

We lift the loss estimates from the mediator space to the action space by taking $\widehat{\widehat{\boldsymbol{\ell}}}_t := \boldsymbol{M}^\top \widehat{\boldsymbol{\ell}}_t$. The cumulative loss estimates are defined by $\widehat{\widehat{\boldsymbol{L}}}_t := \widehat{\widehat{\boldsymbol{L}}}_{t-1} + \widehat{\widehat{\boldsymbol{\ell}}}_t$ with $\widehat{\widehat{\boldsymbol{L}}}_0 = \mathbf{0}$, where $\mathbf{0}$ is the zero vector.

Action-space regularization To adapt to problem complexity in the action space, we use negative Tsallis entropy regularization in the action space. The negative Tsallis entropy

$$\psi_{\mathcal{A}}(\boldsymbol{p}) := 1 - \sum_{a \in \mathcal{A}} \sqrt{p_a}$$

is convex, nonpositive, and vanishes at every pure action: $\psi_{\mathcal{A}}(\boldsymbol{e}_a) = 0$, where $\boldsymbol{e}_a$ is a unit vector in direction a. We use $\psi_{\mathcal{A}}(\boldsymbol{p}_t)$ as one part of our hybrid regularizer.

Mediator-space regularization. To adapt to problem complexity in the mediator space, we use negative Tsallis entropy regularization in the mediator space. In contrast to $\psi_{\mathcal{A}}(\boldsymbol{p})$, the negative Tsallis entropy in the mediator space is defined by

$$\psi_{\mathcal{Z}}(\boldsymbol{q}) := 1 - \sum_{z \in \mathcal{Z}} \sqrt{q_z},$$

does not necessarily vanish at pure actions: we may have $\psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{e}_a) < 0$ due to randomness in $\boldsymbol{M}_{\cdot,a}$. Therefore, we apply calibration before combining it with regularization in the action space.

Calibrated mediator-space regularization. We define the calibrated mediator regularizer, the negative Jensen gap of the convex function $\psi_{\mathcal{Z}}$,

$$\psi_{\mathcal{J}}(\boldsymbol{p}) := \psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{p}) - \sum_{a \in \mathcal{A}} p_a \psi_{\mathcal{Z}}(\boldsymbol{M}_{\cdot,a}) = -\sum_{z \in \mathcal{Z}} \sqrt{(\boldsymbol{M}\boldsymbol{p})_z} + \sum_{a \in \mathcal{A}} p_a \sum_{z \in \mathcal{Z}} \sqrt{M_{z,a}},$$

which adjusts for the stochasticity of the columns of $\boldsymbol{M}$. Convexity of $\psi_{\mathcal{Z}}$ gives $\psi_{\mathcal{J}}(\boldsymbol{p}) \leq 0$. Moreover, $\psi_{\mathcal{J}}(\boldsymbol{e}_a) = 0$ for every $a \in \mathcal{A}$. Thus, $\psi_{\mathcal{A}}$ and $\psi_{\mathcal{J}}$ are both nonpositive and vanish at pure actions.

Our hybrid regularizer $\psi_\wedge(\boldsymbol{p}) := \max\{\psi_{\mathcal{A}}(\boldsymbol{p}), \psi_{\mathcal{J}}(\boldsymbol{p})\}$ takes the maximum of $\psi_{\mathcal{A}}(\boldsymbol{p})$ and $\psi_{\mathcal{J}}(\boldsymbol{p})$, therefore, calibration is important, because if $\psi_{\mathcal{Z}}(\boldsymbol{M}_{\cdot,a}) < 0$, then, for distributions $\boldsymbol{p}$ sufficiently close to $\boldsymbol{e}_a$, the uncalibrated quantity $\psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{p})$ is smaller than $\psi_{\mathcal{A}}(\boldsymbol{p})$, so the maximum always selects $\psi_{\mathcal{A}}(\boldsymbol{p})$, preventing correct adaptation to complexity in the mediator space.

Curvature and stability. Our proof of the main theorem controls how strongly noisy loss estimates can change the action probabilities. The action-space regularizer $\psi_{\mathcal{A}}$ provides curvature for this purpose, but when $\psi_{\mathcal{J}}(\boldsymbol{p}) > \psi_{\mathcal{A}}(\boldsymbol{p})$ it is dormant in our hybrid regularizer $\psi_\wedge(\boldsymbol{p}) = \max\{\psi_{\mathcal{A}}(\boldsymbol{p}), \psi_{\mathcal{J}}(\boldsymbol{p})\}$. The curvature of the Jensen term $\psi_{\mathcal{J}}(\boldsymbol{p})$ depends on the mediator probabilities. An action can then have very small probability while all mediators it produces still have substantial probability, so the Jensen term does not provide the same curvature as the action regularizer. We therefore add an auxiliary action regularizer to $\psi_\wedge(\boldsymbol{p})$, thus enforcing curvature in the action space throughout our stability analysis:

$$\psi_\alpha(\boldsymbol{p}) := \frac{1 - \sum_{a \in \mathcal{A}} p_a^\alpha}{\alpha(1-\alpha)}, \qquad \alpha \in (1/2, 1)\,.$$

Special cases offer alternative ways to control regret. Under perfect controllability ($\boldsymbol{M}_{\cdot,a^\star} = \boldsymbol{e}_{z^\star}$), every suboptimal action produces some mediators that the optimal action never produces. Their probabilities become small as the learner concentrates on the optimal action, making the mediator entropy itself strongly curved in the relevant directions. Under common full support, every mediator is observed with probability bounded away from zero regardless of the action distribution, allowing a different argument based on concentration of cumulative loss estimates.

Causal Tsallis-INF. The final hybrid regularizer is a parametrized combination of $\psi_\wedge(\boldsymbol{p})$ and $\psi_\alpha(\boldsymbol{p})$. Our *Causal Tsallis-INF* samples A_t according to

$$\boldsymbol{p}_t = \operatorname*{arg\,min}_{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}} \left\{ \langle \widehat{\widetilde{\boldsymbol{L}}}_{t-1}, \boldsymbol{p}\rangle + \sqrt{t}\big[c\,\psi_\wedge(\boldsymbol{p}) + \rho\lambda_\alpha\psi_\alpha(\boldsymbol{p})\big] \right\}. \tag{6}$$

Here $c > 1/4$, $\rho > 0$, and $\alpha \in (1/2, 1)$ are fixed regularization parameters, and λ_α uses the dimension-dependent calibration in equation (45). A complete pseudo-code of the algorithm is given in Appendix A.2.

4 Main results

This section presents worst-case and instance-dependent regret bounds for Causal Tsallis-INF. Our instance-dependent analysis is based on the self-bounding technique developed by Zimmert & Seldin (2021), so we start by revisiting their self-bounding condition on the regret.

Self-bounding Zimmert & Seldin (2021) have introduced an *adversarial regime with a self-bounding constraint* for standard multiarmed bandits, which gracefully interpolates between adversarial and stochastic bandits, and includes stochastic bandits as a special case. They showed that whenever an adversarial problem satisfies this constraint, their Tsallis-INF algorithm achieves refined regret guarantees. We generalize the self-bounding constraint to our causal bandit setting, and show that Causal Tsallis-INF achieves refined regret guarantees whenever it is satisfied.

We say that a causal adversarial regime satisfies a $(\boldsymbol{\Delta}_\mathcal{A}, C, T)$ *self-bounding constraint* with $\boldsymbol{\Delta}_\mathcal{A} \in [0,1]^{|\mathcal{A}|}$ and $C \geq 0$, if the pseudo-regret of any algorithm satisfies

$$\mathrm{Reg}_T \geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\Delta_\mathcal{A}(a)\right] - C\,. \tag{7}$$

The condition should be satisfied at time T, but there is no requirement for it to be satisfied for $t < T$. Stochastic causal bandits are a special case of a causal adversarial regime satisfying the self-bounding constraint with $C = 0$ and $\boldsymbol{\Delta}_\mathcal{A}$ being the vector of suboptimality gaps, as can be immediately seen by comparing equation (7) to equation (3). Other special cases generalizing the corresponding settings in multiarmed bandits include *causal stochastically constrained adversarial bandits* and *causal stochastic bandits with adversarial corruptions*. In the former, the conditional mean mediator losses $\mathbb{E}[\ell_t(z) \mid \mathcal{H}_{t-1}]$ may vary with time, while their pairwise differences remain fixed almost surely. Consequently, the conditional action gaps remain fixed, giving a fixed $\boldsymbol{\Delta}_\mathcal{A}$ and $C = 0$. In the latter, a stochastic causal bandit instance undergoes corruption that does not perturb the regret by more than C.

We compare our self bounding condition to Zimmert & Seldin (2021) and Eldowa et al. (2024) in more details in Appendix A.5.

Complexity measures We define the following notions of complexity for games satisfying the self-bounding constraint, including stochastic games as a special case. We use

$$S_\mathcal{A} := \sum_{a:\Delta_\mathcal{A}(a)>0} \frac{1}{\Delta_\mathcal{A}(a)} \tag{8}$$

to denote the complexity term arising from suboptimality gaps in the action space. To discuss the complexity in the mediator space, we let $a^\star \in \{a : \Delta_\mathcal{A}(a) = 0\}$ be an optimal action, which is assumed to be unique, and let $\boldsymbol{q}^\star := \boldsymbol{M}_{\cdot,a^\star}$ and $\mathcal{U} := \mathcal{Z} \setminus \mathrm{supp}(\boldsymbol{q}^\star)$, where $\mathrm{supp}(\boldsymbol{q}^\star)$ is the set of non-zero entries of $\boldsymbol{M}_{\cdot,a^\star}$, and $\Gamma(\boldsymbol{p}) := \langle \boldsymbol{p}, \boldsymbol{\Delta}_\mathcal{A}\rangle$, which in the stochastic case equals $\langle \boldsymbol{M}\boldsymbol{p} - \boldsymbol{q}^\star, \boldsymbol{\Delta}_\mathcal{Z}\rangle$. The complexity in the mediator space is driven by the *attainable mediator complexity*

$$S^{\mathrm{a}}_\mathcal{Z} := \sup_{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}:\Gamma(\boldsymbol{p})>0} \frac{\left(\sum_{z\in\mathcal{U}}\sqrt{(\boldsymbol{M}\boldsymbol{p})_z}\right)^2}{\Gamma(\boldsymbol{p})}, \tag{9}$$

with $S^{\mathrm{a}}_\mathcal{Z} := 0$ when $\mathcal{U} = \varnothing$. If $\mathcal{U} = \varnothing$, it means that every mediator is observed with positive probability when playing $a^\star$. In such cases the algorithm does not need to play suboptimal actions

to gather information, and the resulting regret bound in the theorem below does not grow with time. Interestingly, it is *not* the sum of inverse gaps, $S_{\mathcal{Z}} = \sum_{z \neq z^\star} \frac{1}{\Delta_{\mathcal{Z}}(z)}$, that defines the complexity in the mediator space, and the relation between $S_{\mathcal{Z}}$ and $S^{\mathrm{a}}_{\mathcal{Z}}$ can go in both ways, as discussed in Appendix A.3. Importantly, the instance-dependent regret of Causal Tsallis-INF depends on $\min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}$, so it adapts to the easier of the spaces, as shown in the following theorem.

Theorem 4.1 (Main result)**.** *Causal Tsallis-INF with $\mathbf{p}_t$ defined in equation (6), reduced-variance estimator defined in equation (5), and learning rate $\eta_t = 4/\sqrt{t}$ satisfies the following regret guarantee for* adversarial losses

$$\mathrm{Reg}_T = O\Big(\sqrt{\min\{|\mathcal{A}|, |\mathcal{Z}|\}\, T}\Big)\,. \tag{10}$$

Additionally, if a causal bandit regime satisfies a $(\mathbf{\Delta}_{\mathcal{A}}, C, T)$ self-bounding constraint defined in equation (7) and $a^\star \in \{a : \Delta_{\mathcal{A}}(a) = 0\}$ is unique, then there exists a finite constant $K \geq 0$, depending only on $\boldsymbol{M}, \mathbf{\Delta}_{\mathcal{A}}, \alpha, c$, and $\rho\lambda_\alpha$, such that for

$$B_T = K + \kappa_c \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}(1 + \log T)$$

the regret of Causal Tsallis-INF satisfies

$$\mathrm{Reg}_T \leq B_T + 2\sqrt{C B_T}, \tag{11}$$

where $\kappa_c = 2\,(c/2 + \nu_c)^2$ and $\nu_c = \max\{2/c,\, 1/(2(c - 1/4))\}$; the smallest value is $\kappa_2 = 8$. Stochastic causal bandits satisfy $(\mathbf{\Delta}_{\mathcal{A}}, 0, T)$ self-bounding constraint and the guarantee in equation (11) with $C = 0$. Moreover, for stochastic losses *and unique $a^\star$, if $M_{z,a^\star} > 0$ for every $z \in \mathcal{Z}$, then*

$$\mathrm{Reg}_T \leq K.$$

A proof of Theorem 4.1 is given in Appendix C . The assumption on uniqueness of $a^\star$ in the self-bounding regret analysis is common in best-of-both-worlds literature, and there are tools to eliminate it (Ito, 2021). However, in our setting the identity of $a^\star$ has a non-trivial impact on the complexity $S^{\mathrm{a}}_{\mathcal{Z}}$ in the mediator space in equation (9). We leave treatment of multiple optimal actions to future work.

The complexity term $S_{\mathcal{A}}$ in the action space accounts for the cost of direct exploration of suboptimal actions. The complexity term $S^{\mathrm{a}}_{\mathcal{Z}}$ accounts for the cost of exploring mediators outside of the support of the optimal action $a^\star$. More precisely, equation (9) compares Tsallis exploration mass to its contribution to the regret. Our Causal Tsallis-INF algorithm pays the minimum of the two costs, demonstrating its agility in taking the easier way to identifying $a^\star$.

Finally, we note that the self-bounding constraint is *not* a modeling assumption, it is used only in the analysis, and not used by the algorithm. If it happens to be satisfied, Causal Tsallis-INF enjoys refined regret guarantee, whereas the adversarial regret guarantee in equation (10) holds always.

5 Experiments

We propose a synthetic experiment comparing Causal Tsallis-INF (CTsallis-INF) to a number of baselines. We use Bernoulli losses whose means depend on the observed mediator. The known transition matrix is $M_{z,a} = 0.7\mathbb{1}(z = a \bmod |\mathcal{Z}|) + 0.3/|\mathcal{Z}|$, where actions and mediators are indexed from zero, and $(|\mathcal{A}|, |\mathcal{Z}|) = (80, 30)$. Thus, each action has a dominant mediator, with additional feedback shared across all mediators. Every suboptimal action has a gap 0.07, whereas every suboptimal mediator has a gap 0.1. We experiment with two different environments. A stationary environment, where the means are fixed, and a switching environment, where the means change, while $\boldsymbol{M}$, the optimal actions, and all the gaps remain fixed, thus generating a causal stochastically constrained adversarial regime. In *deterministic switching* environment the means alternate between $(0, 0.1, \ldots, 0.1)$ and $(0.9, 1, \ldots, 1)$in segments, where segment $k = 0, 1, \ldots$ lasts for $\lceil 50(1.6)^k \rceil$ rounds.

We compare CTsallis-INF to the Tsallis-INF algorithm of Zimmert & Seldin (2021), which works with a direct map from actions to losses, the EXP4 algorithm with mediator feedback (EXP4MF) studied by Eldowa et al. (2024), the PE algorithm of Liu et al. (2024), the CUCB and the CTS algorithms of Lu et al. (2020), and the CUCB2 algorithm of Nair et al. (2021). Additionally, we compare

to CTsallis-Action, which only applies regularization in the action space, CTsallis-Mediator, which only applies regularization in the mediator space, and CTsallis-INF (no ψ_α), which omits the ψ_α part of the regularizer in CTsallis. These three ablations of CTsallis allow to discern the impact of the different parts of the hybrid regularizer. In the experiments CTsallis is competitive with the best performing algorithms under stationarity and remains stable under nonstationarity, whereas CTS and CUCB2 deteriorate sharply; CUCB, PE, and EXP4MF exhibit high regret in both regimes. Classical Tsallis-INF that does not use information about mediator can achieve better performance than some algorithms that have been designed to use such information in stationary case and to all beyond the CTsallis family in the nonstationary setting.

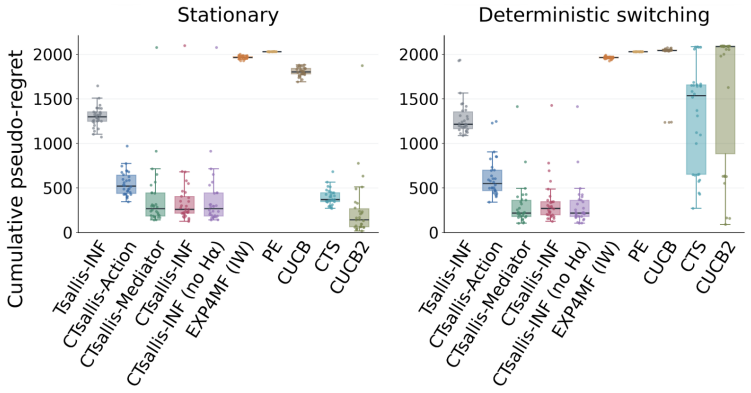


Figure 2: Cumulative pseudo-regret at $T = 30{,}000$ over 30 runs.

In Appendix A.4 we provide additional experiments. We extend main experiment to other configurations of size of $|\mathcal{A}|$, and $|\mathcal{Z}|$, but the results do not change qualitatively. We also corroborate results from the main experiment on wider family of causal graphs $\boldsymbol{M}$ that satisfy our assumptions. Furthermore, we confirm our results about stochastic case from synthetic data simulations on real world data coming from early education STAR program (Mosteller, 1995). Moreover, we show that CTsallis-INF is not sensitive to natural form of misspecification of $\boldsymbol{M}$, whereas CUCB2 is. We also confirm empirically the adaptivity to the more favorable complexity in stochastic case presented in the main result theorem. [1].

6 Conclusions

Our paper connects the literature on bandits with mediator feedback with causal bandits literature, and shows that the former can be used to relax the stochasticity assumption behind the latter. We have proposed Causal Tsallis-INF, a causal bandit algorithm that benefits from mediator feedback, automatically adapts to the problem complexity in the action and mediator space, and exhibits robustness to adversarial losses and refined regret guarantees in intermediate regimes between stochastic and adversarial. In stochastic regimes it performs comparably to state-of-the-art CTS algorithm, but in contrast to stochastic causal bandit algorithms it retains good regret guarantees when stochasticity is violated. It also has tighter regret guarantees in stochastic case and dominates the EXP4 algorithm in empirical performance in both regimes.

Future work Natural extensions include learning with unknown or drifting mechanisms $\boldsymbol{M}$, extending instance-dependent guarantees to multiple optimal actions, and studying more general causal graphs.

[1]The code can be found under `https://anonymous.4open.science/r/Mediator-FC43/README.md`

AI DISCLOSURE

In this work, we used generative AI tools to conceptualize parts of the proofs of the theorems provided [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs] including suggestion on parts of the proofs write-up, as well as implement parts of code for conducting experiments [implement methods]. AI was also used for finding related work pointers. However, AI was not used for making literature review. We have not used generative AI tools for other tasks with required disclosure [propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results]. In particular, the main ideas and observations of this article are confidently the work of human authors. Finally, AI was not used either to formulate questions for surveys or interviews, nor to create or modify scientific figures or images, nor suggest experimental parameters, nor transcribe recordings of research material, nor summarize or analyse existing literature, nor discover research topics or identify gaps, nor format references, nor suggest a structure for a research paper, nor propose a title or keywords for a research paper.

We have reviewed all AI-assisted work carefully. AI was definitely not idea generator nor autonomous co-author. The final outcome of this article is driven by the work of human authors. We checked any artifacts of LLM-generated ideas for potential plagiarism through a manual literature survey, LLM-generated parts of the code were verified and tested for correctness by 2 authors, LLM-generated parts of the proof were checked by 3 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Reda Alami. Non-stationary causal bandits. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2015.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Annual IEEE Symposium on Foundations of Computer Science*, 1995.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

Blair Bilodeau, Linbo Wang, and Dan Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems (NIPS)*, 2022.

Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information capacity regret bounds for bandits with mediator feedback. *Journal of Machine Learning Research*, 2024.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Yeahoon Kwon, Yesong Choe, Soungmin Park, Neil Dhir, and Sanghack Lee. Non-stationary structural causal bandits. *Advances in Neural Information Processing Systems (NIPS)*, 2025.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems (NIPS)*, 2016.

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

Ziyi Liu, Idan Attias, and Daniel M Roy. Causal bandits: The pareto optimal frontier of adaptivity, a reduction to linear bandits, and limitations around unknown marginals. *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2020.

Aurghya Maiti, Vineet Nair, and Gaurav Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2022.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. In *Advances in Neural Information Processing Systems (NIPS)*, 2024.

H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

Alberto Maria Metelli, Matteo Papini, Pierluca D'Oro, and Marcello Restelli. Policy optimization as online learning with mediator feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Frederick Mosteller. The Tennessee study of class size in the early school grades. *The future of children*, 1995.

Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Francesco Orabona. Online learning: A modern introduction using convex optimization. *arXiv preprint arXiv:1912.13213*, 2026.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Mohammad Shahverdikondori, Amir Mohammad Abouei, Alireza Rezaeimoghadam, and Negar Kiyavash. Pure exploration beyond reward feedback: The role of post-action context. *Advances in Neural Information Processing Systems (NIPS)*, 2026.

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

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

A Appendix

A.1 Additional related literature

Our work builds heavily on best-of-both-worlds analysis techniques based on self-bounding, pioneered by Zimmert & Seldin (2021), as well as the idea of using hybrid regularizers (Zimmert et al., 2019; Zimmert & Seldin, 2020; Masoudian et al., 2024).

Mediator feedback has some similarities to graph feedback (Alon et al., 2015), because playing an action reveals some information about other actions in accordance with a predetermined problem structure, but the two settings are not exactly the same. In graph feedback playing an action reveals the loss of neighboring actions, whereas in mediator feedback it provides some information, but not necessarily complete information about the loss of other actions. We leave closer study of potential connections between the two settings to future work. By associating every edge in an undirected feedback graph with a mediator variable and assuming that the mediator variable is observed it is possible to obtain an approximate bidirectional mapping between the two settings, but it somewhat distorts the observation probabilities.

Both Eldowa et al. (2024) and Liu et al. (2024) notice that bandits with mediator feedback resemble linear bandits due to linear relation between the loss and the distribution on mediating states given an action. Liu et al. (2024) use linear-bandit reduction to establish their regret guarantees. However, the mediator feedback differs from standard linear-bandit: the learner observes both the sampled mediator and its loss. Indeed, Eldowa et al. (2024) provide a lower bound demonstrating existence of problem instances, where access to mediator feedback yields qualitatively lower regret guarantees compared to what can be achieved with plain scalar feedback of linear bandits. A more favorable lower bound in mediator feedback presence than linear bandits is also established in the best arm identification task (Shahverdikondori et al., 2026) (their algorithm exploits attainable mediator distributions and illustrates the benefit of retaining mediator observations).

A related but different adaptation question is whether the post-action context is the mediator (satisfies the separator condition versus there is a direct influence of action on the loss $A \to \ell$ that goes around Z) (Bilodeau et al., 2022; Liu et al., 2024). These works seek guarantees that adapt to the presence or absence of this structure. We assume separation throughout and study adaptation to the loss regime and to the more favorable exploration complexity in the action or attainable mediator space.

Nonstationary causal bandits have also been studied with two different angles. (Alami, 2022) demonstrate adaptation to piecewise-stationary environments experimentally, but leave theoretical regret bounds to future work. Kwon et al. (2025) address which variables to intervene on when interventions affect both current and future rewards. Their guarantees concern the intervention search space, while learning performance is evaluated empirically without regret bounds.

A.2 Pseudocode of the variants and the final proposed Causal Tsallis-INF and ablations

In the main text appear *CTsallis-INF* and its ablations *CTsallis-Action* and *CTsallis-Mediator*. We give a common pseudocode for all three variants in Algorithm 1.

For CTsallis-Action there exists a closed form solution for the choice of probability distribution over actions. That is not the luxury we possess once we add regularization in the context space. We describe the details of solving the optimization problem of selecting the probability over actions when context regularizer term is present in Appendix B.

A.3 Comparing the two mediator complexities.

We have two mediator complexities $S_{\mathcal{Z}}$ and $S^{\mathrm{a}}_{\mathcal{Z}}$. The two complexities use different reference points. $S_{\mathcal{Z}}$ measures mediator gaps relative to the best *mediator*, whereas $S^{\mathrm{a}}_{\mathcal{Z}}$ uses the excess loss relative to the best *action*. Moreover, $S_{\mathcal{Z}}$ counts every suboptimal mediator, whereas $S^{\mathrm{a}}_{\mathcal{Z}}$ counts only mediators outside $\mathrm{supp}(\boldsymbol{M}_{\cdot,a^\star})$. Mediators in this support continue to be observed when the learner exploits the optimal action, without incurring action regret.

Algorithm 1 Causal Tsallis-INF: Action, Mediator, and Jensen variants

Require: Known mechanism $\boldsymbol{M}$; variant $v \in \{\text{Action}, \text{Mediator}, \text{INF}\}$; coefficient $c > 1/4$
Require: For variant INF: $\alpha \in (1/2, 1)$, $\rho > 0$, and $\lambda_\alpha > 0$
if $v = \text{Action}$ **then**
  $R(\boldsymbol{p}) \leftarrow c\psi_{\mathcal{A}}(\boldsymbol{p})$
else if $v = \text{Mediator}$ **then**
  $R(\boldsymbol{p}) \leftarrow c\psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{p})$
else
  $R(\boldsymbol{p}) \leftarrow c\max\{\psi_{\mathcal{A}}(\boldsymbol{p}), \psi_{\mathcal{J}}(\boldsymbol{p})\} + \rho\lambda_\alpha\psi_\alpha(\boldsymbol{p})$
end if
Initialize $\widehat{\widetilde{\boldsymbol{L}}}_0 \leftarrow \boldsymbol{0}$
for $t = 1, 2, \ldots$ **do**
  Compute $\boldsymbol{p}_t \in \underset{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}}{\arg\min} \left\{\left\langle \widehat{\widetilde{\boldsymbol{L}}}_{t-1}, \boldsymbol{p}\right\rangle + \sqrt{t}\, R(\boldsymbol{p})\right\}$
  $\boldsymbol{q}_t \leftarrow \boldsymbol{M}\boldsymbol{p}_t$
  $B_t(z) \leftarrow \frac{1}{2}\mathbb{1}(q_t(z) \geq 16/t)$ for every $z \in \mathcal{Z}$
  Sample $A_t \sim \boldsymbol{p}_t$, play A_t, and observe $(Z_t, \ell_t(Z_t))$
  $\widehat{\boldsymbol{\ell}}_t \leftarrow \boldsymbol{B}_t$
  $\widehat{\ell}_t(Z_t) \leftarrow B_t(Z_t) + \dfrac{\ell_t(Z_t) - B_t(Z_t)}{q_t(Z_t)}$
  $\widehat{\widetilde{\boldsymbol{L}}}_t \leftarrow \widehat{\widetilde{\boldsymbol{L}}}_{t-1} + \boldsymbol{M}^\top\widehat{\boldsymbol{\ell}}_t$
end for

We say that $z^\star$ is perfectly controllable if some action produces it with probability one. Under our assumption that the optimal action is unique, this implies $\boldsymbol{q}^\star = \boldsymbol{M}_{\cdot,a^\star} = \boldsymbol{e}_{z^\star}$. In this case, Cauchy–Schwarz gives

$$S^{\mathrm{a}}_{\mathcal{Z}} \leq \sum_{z\neq z^\star} \frac{1}{\Delta_{\mathcal{Z}}(z)}\,.$$

Equality holds if every mediator can be produced with probability one by some action: mixing these actions can then generate any distribution over mediators, including a distribution attaining equality in the Cauchy–Schwarz bound.

Nevertheless, there is no general ordering between the two complexities. The complexity $S^{\mathrm{a}}_{\mathcal{Z}}$ can be smaller: some mediators need no additional exploration, and the mechanism restricts which distributions over the remaining mediators can be generated. In particular, full mediator support of the optimal action gives $S^{\mathrm{a}}_{\mathcal{Z}} = 0$. If the best mediator is perfectly controllable, so that $\boldsymbol{M}_{\cdot,a^\star} = \boldsymbol{e}_{z^\star}$, the two reference losses coincide: $\langle \boldsymbol{M}\boldsymbol{p} - \boldsymbol{M}_{\cdot,a^\star}, \boldsymbol{\mu}\rangle = \sum_{z\neq z^\star}(\boldsymbol{M}\boldsymbol{p})_z\Delta_{\mathcal{Z}}(z)$. Cauchy–Schwarz then gives $S^{\mathrm{a}}_{\mathcal{Z}} \leq S_{\mathcal{Z}}$. Equality holds when every mediator can be produced with probability one by some action, for example, when $\boldsymbol{M} = \boldsymbol{I}$. Conversely, $S^{\mathrm{a}}_{\mathcal{Z}}$ can be larger because two actions can have nearly equal average losses while producing very different mediator distributions. A suboptimal action may produce a high-loss mediator that the optimal action never produces, but compensate by also producing the lowest-loss mediator more often. The probability of the unsupported mediator then remains substantial, while the difference between the two actions' average losses is small. In the definition of $S^{\mathrm{a}}_{\mathcal{Z}}$, this gives a substantial numerator divided by a small action gap.

Numerical example. Consider two actions and three mediators with

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}, \qquad \boldsymbol{M} = \begin{pmatrix} 1/2 & 3/4-\varepsilon \\ 1/2 & 0 \\ 0 & 1/4+\varepsilon \end{pmatrix}, \qquad 0 < \varepsilon < 3/4\,.$$

The first action has mean loss $1/4$, whereas the second has mean loss $1/4 + \varepsilon$. Thus $a^\star = 1$ is uniquely optimal, $\Delta_{\mathcal{A}}(2) = \varepsilon$, and $\mathcal{U} = \{3\}$. For every $\boldsymbol{p}$ with $p(2) > 0$,

$$\Gamma(\boldsymbol{p}) = \langle \boldsymbol{M}\boldsymbol{p} - \boldsymbol{M}_{\cdot,a^\star}, \boldsymbol{\mu}\rangle = \varepsilon p(2), \qquad \left(\sum_{z\in\mathcal{U}} \sqrt{(\boldsymbol{M}\boldsymbol{p})_z}\right)^2 = (1/4+\varepsilon)p(2)\,.$$

Consequently,

$$S_{\mathcal{Z}} = 3, \qquad S^{\mathrm{a}}_{\mathcal{Z}} = \frac{1/4 + \varepsilon}{\varepsilon} = 1 + \frac{1}{4\varepsilon}\,.$$

Three choices illustrate all possible orderings:

ε	$S_{\mathcal{Z}}$	$S^{\mathrm{a}}_{\mathcal{Z}}$	Comparison
$1/4$	3	2	$S_{\mathcal{Z}} > S^{\mathrm{a}}_{\mathcal{Z}}$
$1/8$	3	3	$S_{\mathcal{Z}} = S^{\mathrm{a}}_{\mathcal{Z}}$
$1/20$	3	6	$S_{\mathcal{Z}} < S^{\mathrm{a}}_{\mathcal{Z}}$

The mediator losses, and hence $S_{\mathcal{Z}}$, are unchanged across these examples. As $\varepsilon \downarrow 0$, the actions become nearly tied while the second action still assigns approximately $1/4$ of its probability to the unsupported mediator. Thus $S^{\mathrm{a}}_{\mathcal{Z}} \to \infty$ while $S_{\mathcal{Z}} = 3$.

For $\varepsilon = 1/4$, the second action produces mediators 1 and 3 with probabilities $1/2$ each. Relative to the optimal action, it replaces mediator 2 by the higher-loss mediator 3. Its action gap is $1/4$, and its probability of producing the only unsupported mediator is $1/2$, giving $S^{\mathrm{a}}_{\mathcal{Z}} = 2$. Unlike $S_{\mathcal{Z}}$, this complexity does not separately count mediator 2, which is already observed under the optimal action.

For $\varepsilon = 1/20$, the second action instead produces mediators 1 and 3 with probabilities $7/10$ and $3/10$. Its increased probability of producing the best mediator partly compensates for its probability of producing the worst mediator. Its mean loss is therefore only $3/10$, compared with $1/4$ for the optimal action. The small action gap $1/20$, together with unsupported probability $3/10$, gives $S^{\mathrm{a}}_{\mathcal{Z}} = 6$.

A.4 Remaining experiments

Robustness to nonstationarity. We start with a synthetic experiment comparing Causal Tsallis-INF (CTsallis-INF) to a number of baselines. We use Bernoulli losses whose means depend on the observed mediator. The known transition matrix is $M_{z,a} = 0.7\mathbb{1}(z = a \bmod |\mathcal{Z}|) + 0.3/|\mathcal{Z}|$, where actions and mediators are indexed from zero, and $(|\mathcal{A}|, |\mathcal{Z}|) \in \{(80, 30), (30, 80), (60, 60)\}$. Thus, each action has a dominant mediator, with additional feedback shared across all mediators. We experiment with two different environments. A stationary environment, where the means are fixed, and a switching environment, where the means change, while $\boldsymbol{M}$, the optimal actions, and all the gaps remain fixed, thus generating a causal stochastically constrained adversarial regime. In *deterministic switching* environment the means alternate between $(0, 0.1, \ldots, 0.1)$ and $(0.9, 1, \ldots, 1)$in segments, where segment $k = 0, 1, \ldots$ lasts for $\lceil 50(1.6)^k \rceil$ rounds. Using the same schedule we provide experiment showcasing the same behavior when loss changes are random. In the *random piecewise stationary* environment the means in segment k are defined by $\boldsymbol{\mu}_t = \mathrm{clip}_{[0,1]}(\boldsymbol{\mu}^0 + \beta_k \mathbf{1})$, where $\boldsymbol{\mu}^0 = (0.45, 0.55, \ldots, 0.55)$ and $\beta_k \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(-0.45, 0.45)$, while the segments are defined as before.

We present the results in Figure 3. CTsallis-INF is competitive in all stochastic settings to the best performing algorithms under stationarity. In the cases when $|\mathcal{Z}| \geq |\mathcal{A}|$ CTS performs even worse than CTsallis-INF. Moreover, CTsallis-INF remains stable under nonstationarity, whereas CTS and CUCB2 deteriorate sharply. The algorithms CUCB, PE, and EXP4MF have high regret in both regimes. Classical Tsallis-INF that does not use information about mediator can achieve better performance than some algorithms that have been designed to use such information in stationary case and to all beyond the CTsallis family in the nonstationary setting. In this experiment, for this game horizon CTsallis-Action performs worse than CTsallis-Mediator - action gaps are smaller here than the mediator gaps. Moreover, we show that even in $|\mathcal{Z}| > |\mathcal{A}|$ the causal bandit algorithms perform better than their classical counterparts.

Adaptation between action and mediator spaces. We test the property of CTsallis-INF of adaptation to the more favorable complexity. The data generating process looks as follows. For $n \in \{8, 12, 16\}$, let V and E denote the vertices and edges of the complete graph K_n. We compare two orientations: vertex actions with edge mediators, and edge actions with vertex mediators, respectively:

$$M^{\mathrm{V}}_{e,v} = \frac{\mathbb{1}(v \in e)}{n-1}, \qquad M^{\mathrm{E}}_{v,e} = \frac{\mathbb{1}(v \in e)}{2}\,.$$

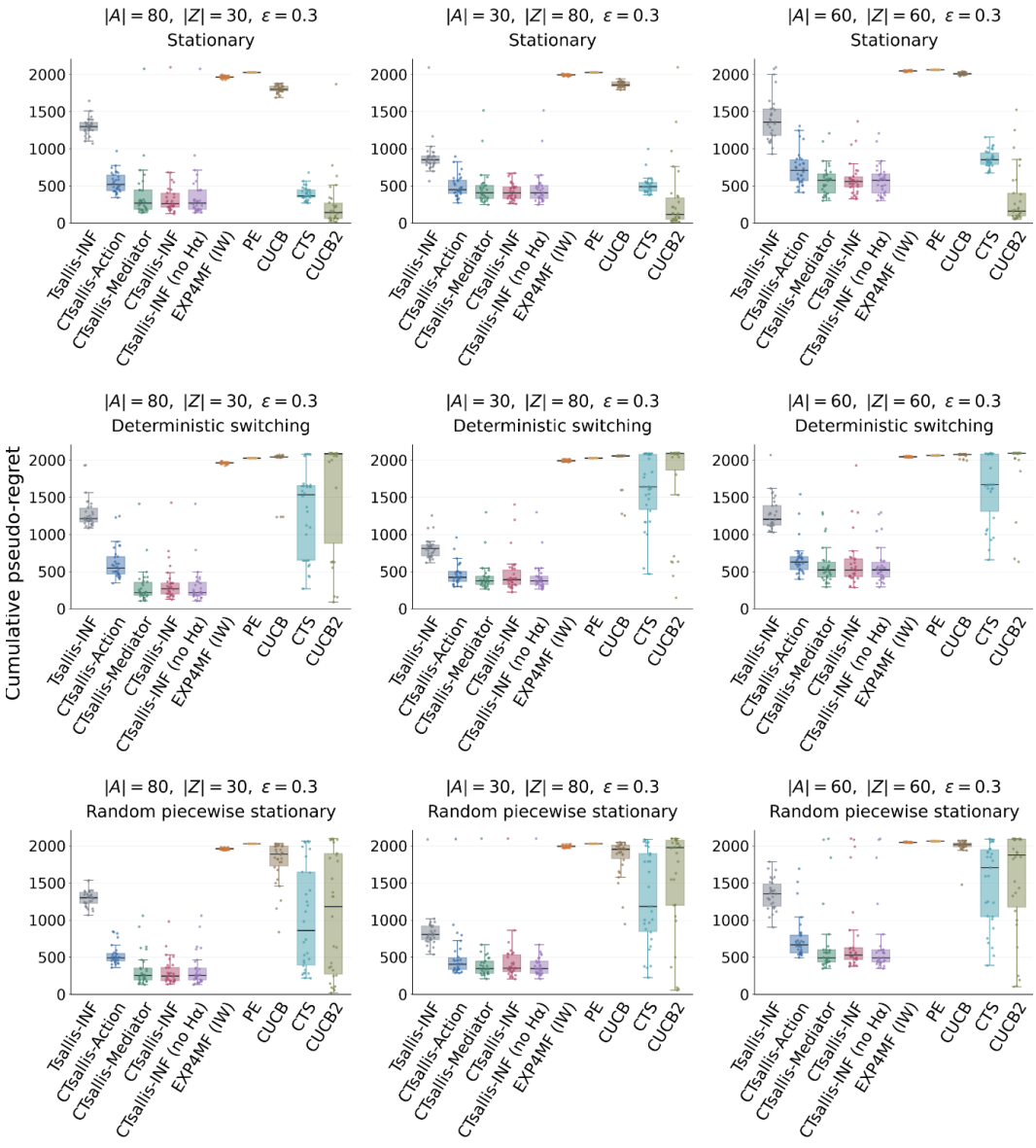


Figure 3: Robustness to nonstationarity. Cumulative pseudo-regret at $T = 30{,}000$ over 30 runs.

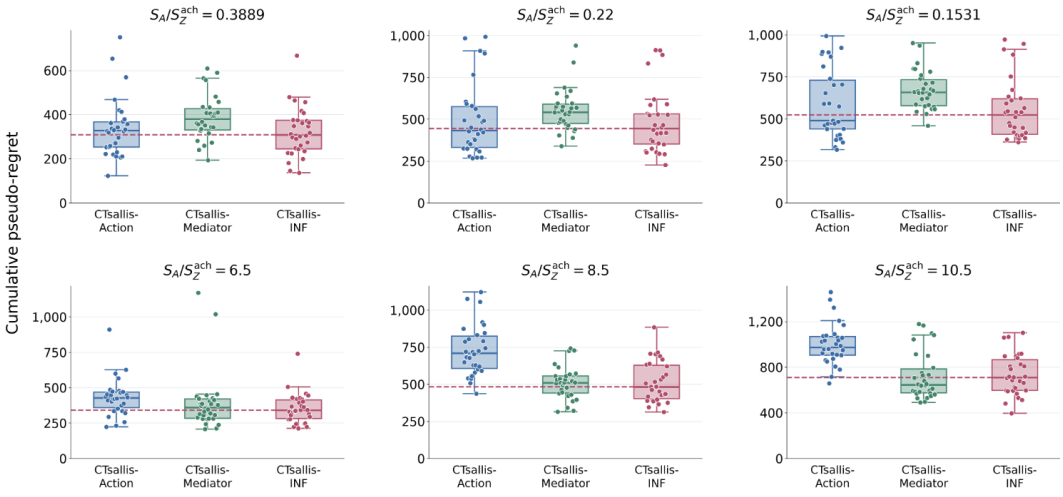


Figure 4: Pseudo-regret at T=50,000 out of 30 runs. Dashed red segments mark the CTsallis-INF sample medians.

An action therefore reveals a uniformly sampled incident edge or endpoint. The cardinalities are $(n, \binom{n}{2})$ and $(\binom{n}{2}, n)$. For each setting, choose an optimal vertex $v^\star$ or edge $e^\star$ uniformly once and keep it fixed across seeds. With $b = 0.35$ and $D = 0.02$, define stationary mediator means

$$\mu^{\mathrm{V}}(e) = b + D\frac{n-1}{n-2}\mathbb{1}(v^\star \notin e), \qquad \mu^{\mathrm{E}}(v) = b + 2D\mathbb{1}(v \notin e^\star)\,.$$

Observations satisfy $Z_t \sim \boldsymbol{M}_{\cdot,A_t}$ and $\ell_t(Z_t) \mid Z_t = z \sim \mathrm{Bernoulli}(\mu(z))$. Both orientations have a unique optimal action and minimum action gap D. Their attainable-complexity ratios are

$$\frac{S_{\mathcal{A}}}{S_{\mathcal{Z}}^{\mathrm{a}}} = \begin{cases} \dfrac{2(n-1)}{(n-2)^2}, & \text{vertex actions,} \\ \dfrac{n+5}{2}, & \text{edge actions.} \end{cases}$$

Thus increasing n strengthens the contrast between the two regimes. We compare cumulative pseudo-regret of CTsallis-Action, CTsallis-Mediator and CTsallis-INF over 30 paired runs at $T = 50{,}000$. All methods use $c = 0.7$ and the RV cutoff $16/t$; CTsallis-INF additionally uses $\rho = 0.015$ and $\alpha = 2/3$.

We portray the results in Figure 4. CTsallis-INF always stays closer to the better of the CTsallis-Action and CTsallis-Mediator showing adaptivity.

Random graphs. To provide evidence that the conclusions from the main text are not dependent on the particular graph we replicate them on wider family of graph structure. In particular we generate random graphs, according to particular structure and depart from full observability regime. The second experiment samples 50 stationary action–mediator bipartite graphs, with both cardinalities in $\{10, 30, 80\}$. Actions and mediators are randomly partitioned into two balanced communities. Edges are sampled independently with probability $d(1 + r)$ within communities and $d(1 - r)$ across communities: $d \in \{0.05, 0.2, 0.4\}$ controls density, while $r \in \{0, 0.5, 1\}$ controls community structure. At $r = 0$ the community labels have no effect; at $r = 1$ there are no cross-community edges. Isolated vertices receive a uniformly sampled allowable edge. Each action's transition probabilities on its support follow $\mathrm{Dirichlet}(\kappa, \ldots, \kappa)$, with $\kappa \in \{0.3, 1, 10\}$ controlling how evenly its probability mass is distributed. Independently of the graph, mediator losses are Bernoulli with means $\mu(z) = \frac{1}{2} + \lambda x_z$, where $x_z \overset{\text{iid}}{\sim} \mathrm{Uniform}[-1, 1]$ and $\lambda \in \{0.03, 0.1, 0.3\}$ controls the loss scale. For each selected instance, the mediator means alternate between $\boldsymbol{\mu} - \min_z \mu(z)\mathbf{1}$ and $\boldsymbol{\mu} + (1 - \max_z \mu(z))\mathbf{1}$, starting with the former and following the geometric segment schedule of the main experiment. These common shifts preserve the transition matrix, all action gaps, and the optimal actions.

We plot the results in Figure 5. Across 70 stationary environments at $T = 30{,}000$, CTsallis-INF has the lowest mean regret normalized by uniform-policy regret across the ablated variants. Overall, in this stationary environment generator CTS is performing the best, while CUCB, EXP4MF and PE incur substantially higher regret. CUCB2 can run only under perfect overlap. Adversarial regime affects CUCB and CTS by making their performance worse than classic Tsallis-INF that does not use information about the mediator.

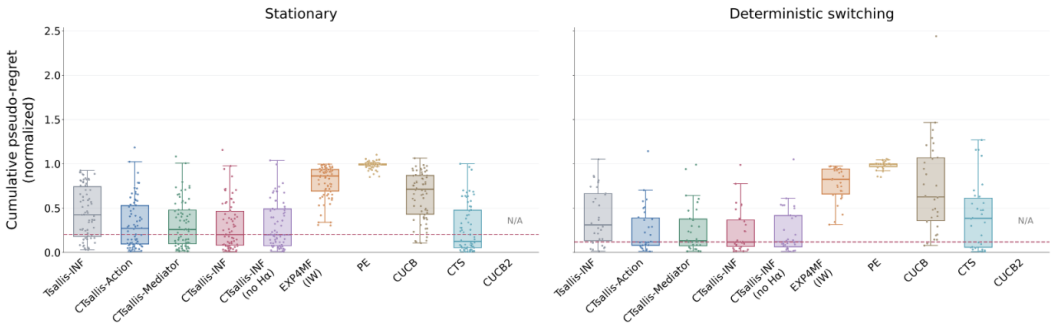


Figure 5: Pseudo-regret at $T = 30{,}000$, normalized by the expected regret of uniform action selection on each instance.

Robustness to misspecified mechanism. This experiment evaluates robustness to estimation error in the mechanism. We compare the true $\boldsymbol{M}$ with estimates initialized from 500 or 2000 samples, obtained by drawing actions uniformly and mediators from their corresponding mechanism distributions. The learner updates the mechanism estimate online as further samples are collected.

We present the results in Figure 6. CTsallis family generally retains the benefits of mediator feedback when the transition mechanism is estimated online. Nevertheless, performance depends on the environment and the amount of initial calibration data. Particularly when mediators outnumber actions the estimator of $\mu(z)$ plays a bigger role. Increasing the initial sample size mitigates this degradation. CTsallis-Mediator seems to be more sensitive to misspecification. CUCB2 is highly sensitive to misspecification.

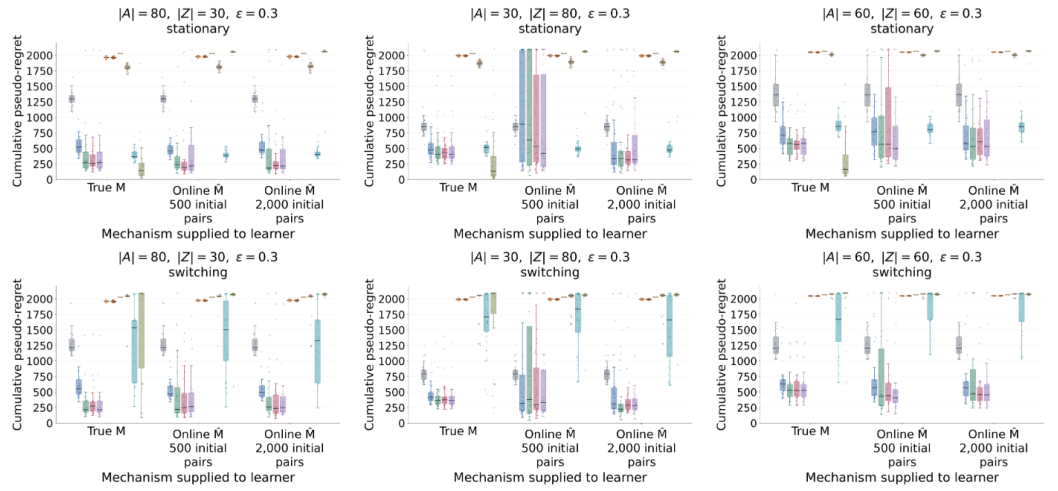


Figure 6: Pseudo-regret at $T = 30{,}000$.

Empirical evaluation on Project STAR. For real world data experiment we use the public `STAR` data from Tennessee's randomized class-size study (1985–1989) (Mosteller, 1995). It consists of 4,298 students with recorded kindergarten class type and observed first-grade reading and mathematics scores. The three actions are the recorded kindergarten class types: regular, regular with an

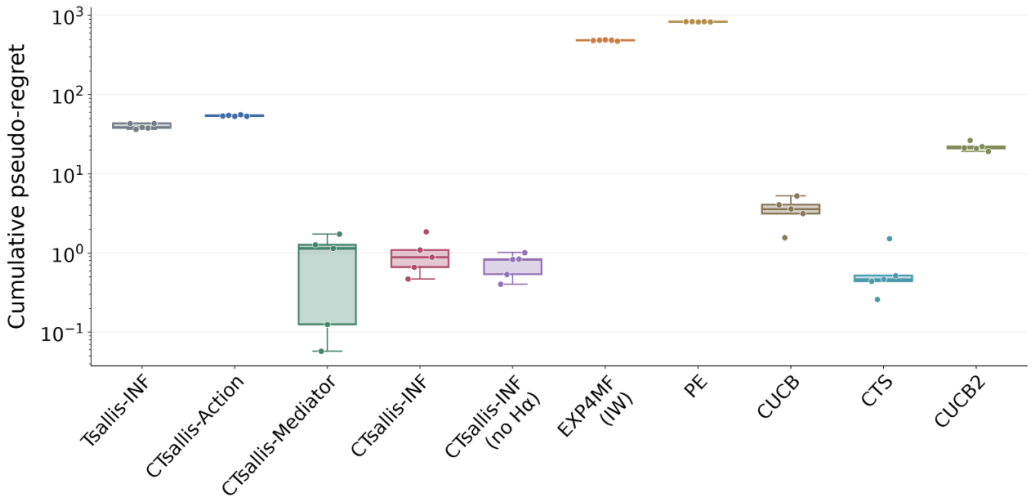


Figure 7: Cumulative pseudo-regret on the stationary empirical STAR benchmark at $T = 20{,}000$, on a logarithmic vertical axis.

aide, and small. The mediator is average kindergarten reading and mathematics achievement (discretized). The outcome is the average of first-grade reading and mathematics scores. We convert it into a loss in $[0, 1]$ by subtracting its midrank empirical cumulative probability from one. The cumulative probability is the fraction of reference outcomes strictly below the observed outcome plus half the fraction equal to it. Higher achievement therefore corresponds to lower loss. This data fits our problem setup. Class size and staffing plausibly affects subsequent kindergarten achievement that in turn plausibly transmit their influence to later achievement. However, recorded results need not capture all treatment pathways. It is possible that $A \to \ell$ exists. The experiment consequently evaluates robustness on a stationary empirical environment with potentially imperfect mediation.

At each round, we sample an entire record uniformly with replacement from the selected action, revealing its mediator and loss together. Learners receive the exact empirical transition matrix $\boldsymbol{M}$. We report empirical pseudo regret at $T = 20{,}000$ over five runs.

We report the results in Figure 7. Mean cumulative pseudo-regret is 1.00 for CTsallis-INF, 0.87 for CTsallis-Mediator, and 0.73 for CTsallis-INF without ψ_α, compared with 40.09 for Tsallis-INF and 54.26 for CTsallis-Action; CTS achieves 0.64. The best action feedback covers every mediator, which explains the strong mediator-based performance, which in turn is inherited by adaptive CTsallis-INF.

Hyperparameter calibration. We have run a couple of preliminary runs of stationary settings as in the experiment *Robustness to nonstationarity* to see what ranges of hyperparameter values of c and ρ from Equation 6 lead to what level of performance. Subsequently, outside of this preliminary run, the parameters were held fixed - not selected separately for each environment or gap regime! Across the reported experiments, we use a single configuration of hyperparameters. The performance of these values remains strong across all the experiments. For CTsallis-INF, we use

$$\sqrt{t}\big[c\,\psi_\wedge(\boldsymbol{p}) + \rho\lambda_\alpha\psi_\alpha(\boldsymbol{p})\big], \quad c = 0.7, \quad \rho = 0.015, \quad \text{and} \quad \alpha = 2/3,$$

with the dimension-dependent calibration of λ_α in equation (45) and RV cutoff $16/t$. CTsallis-Action and CTsallis-Mediator use the same main coefficient and mediator-estimator cutoff, while the no-ψ_α ablation changes only ρ to zero. These comparisons therefore assess fixed configurations across different information structures, rather than configurations optimized separately for each instance.

Note that the main theorem requires any $c > 1/4$ to retain the logarithmic rates in the theorem. The current proof for the regret bounds gives $c = 2$ as the value that minimizes the coefficient κ_c of the $\log T$ term in the regret bound. However, we find that values of $c < 1$ lead to much better

Table 2: Notation correspondence between this paper and Eldowa et al. (2024) for the fixed-mechanism setting: $\theta_a = \boldsymbol{M}_{\cdot,a}$ and $x = z$.

Object	Our notation	Eldowa et al. (2024)
Choices and their number	$a \in \mathcal{A}$, $\lvert\mathcal{A}\rvert$	$\theta \in \Theta$, N
Mediators and their number	$z \in \mathcal{Z}$, $\lvert\mathcal{Z}\rvert$	$x \in \mathcal{X}$, K
Transition probability	$M_{z,a}$	$\theta_a(x)$
Sampling probabilities	$p_t(a)$	$p_t(\theta_a)$
Selected choice	A_t	$\vartheta_t = \theta_{A_t}$
Latent loss vector	$\boldsymbol{\ell}_t \in [0,1]^{\lvert\mathcal{Z}\rvert}$	$\ell_t \in [0,1]^K$
Observed feedback	$(Z_t, \ell_t(Z_t))$	$(X_t, \ell_t(X_t))$
Mediator distribution	$\boldsymbol{M}\boldsymbol{p}_t$	$\psi_t = \sum_{\theta\in\Theta} p_t(\theta)\theta$
Distributions obtainable by mixing choices	$\{\boldsymbol{M}\boldsymbol{p} : \boldsymbol{p} \in \Delta^{\lvert\mathcal{A}\rvert-1}\}$	$\mathrm{co}(\Theta)$
Action loss	$\bar{\ell}_t(a) = \langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{\ell}_t \rangle$	$\ell_t(\theta_a)$
Conditional mean mediator loss	$\mathbb{E}[\ell_t(z) \mid \mathcal{H}_{t-1}]$	$\mathbb{E}_t[\ell_t(x)]$
Conditional mean action loss	$\mathbb{E}[\bar{\ell}_t(a) \mid \mathcal{H}_{t-1}]$	$\mathbb{E}_t[\ell_t(\theta_a)]$
Expected pseudo-regret	Reg_T	R_T
Conditional expectation	$\mathbb{E}[\cdot \mid \mathcal{H}_{t-1}]$	$\mathbb{E}_t[\cdot]$
Simplex over choices	$\Delta^{\lvert\mathcal{A}\rvert-1}$	Δ_N

Our $\boldsymbol{\ell}_t$ and their ℓ_t both denote the latent loss vector; the learner observes only its coordinate at the sampled mediator. The action loss $\bar{\ell}_t(a)$ averages over the mediator distribution while holding $\boldsymbol{\ell}_t$ fixed. It is therefore still random when $\boldsymbol{\ell}_t$ is random.

performance consistently across all experiments we have ever run. The observed mismatch between the theoretically best parameter and the experimentally reasonable choice motivates a sharper finite-time analysis. Nevertheless, this observation does not invalidate the existing guarantees.

A.5 Comparison of self-bounding conditions

We compare our causal self-bounding constraint with the conditions of Zimmert & Seldin (2021) and Eldowa et al. (2024). Our definition uses the same aggregate regret inequality as Zimmert & Seldin (2021). A condition on action gaps conditional on the observed history is sufficient for this inequality. The unconditional gap condition displayed by Eldowa et al. (2024) is different and, under temporal dependence, does not by itself imply the aggregate inequality.

Zimmert & Seldin (2021) and our definition. In our notation, the self-bounding constraint of Zimmert & Seldin (2021, Section 2, equation (1)) takes the form

$$\mathrm{Reg}_T \geq \sum_{t=1}^{T} \sum_{a\in\mathcal{A}} \Delta_{\mathcal{A}}(a)\, \mathbb{P}(A_t = a) - C = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a\in\mathcal{A}} p_t(a)\Delta_{\mathcal{A}}(a)\right] - C. \tag{12}$$

This is precisely the inequality in our definition equation (7). Both definitions require it for every algorithm at the specified horizon T, with the same fixed $\boldsymbol{\Delta}_{\mathcal{A}} \in [0,1]^{\lvert\mathcal{A}\rvert}$ and $C \geq 0$; neither requires the corresponding inequality at every earlier horizon. When proving a bound for a particular algorithm, it suffices that the inequality holds for that algorithm.

The extension concerns the feedback model: our learner observes $(Z_t, \ell_t(Z_t))$, with Z_t sampled from $\boldsymbol{M}_{\cdot,A_t}$. When $\boldsymbol{M} = \boldsymbol{I}_{\lvert\mathcal{A}\rvert}$, this reduces to ordinary bandit feedback and the aggregate constraint is exactly that of Zimmert & Seldin (2021). For general $\boldsymbol{M}$, the aggregate constraint has the same form, while mediator observations can provide information about other actions. Our analysis exploits this information through $S^{\mathrm{a}}_{\mathcal{Z}}$, in addition to the action complexity $S_{\mathcal{A}}$.

The entries of $\boldsymbol{\Delta}_{\mathcal{A}}$ are coefficients in the self-bounding inequality; outside the stochastic setting they need not equal actual expected loss gaps in any particular round. Our theorem, like the gap-dependent guarantee of Zimmert & Seldin (2021, Theorem 1), assumes a unique zero entry $a^\star$.

This additional assumption belongs to the regret guarantee, rather than to the definition of the self-bounding regime. In a corrupted instance, $a^\star$ need not be the best fixed action for the corrupted losses.

A sufficient conditional gap condition. Suppose $\Delta_{\mathcal{A}}(a^\star) = 0$ and, for every $a \neq a^\star$ and $t \leq T$,

$$\mathbb{E}\big[\bar{\ell}_t(a) - \bar{\ell}_t(a^\star) \,\big|\, \mathcal{H}_{t-1}\big] \geq \Delta_{\mathcal{A}}(a) - \xi_t, \qquad \sum_{t=1}^{T} \xi_t \leq C, \tag{13}$$

almost surely, where $\xi_t \geq 0$ is measurable with respect to $\mathcal{H}_{t-1}$. Here $\bar{\ell}_t(a) = \langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{\ell}_t \rangle$ keeps its definition from the problem setup; hence the conditional expectation on the left equals $\langle \boldsymbol{M}_{\cdot,a} - \boldsymbol{M}_{\cdot,a^\star}, \mathbb{E}[\boldsymbol{\ell}_t \mid \mathcal{H}_{t-1}] \rangle$. The learner's sampling probabilities are determined before the current loss is observed, so

$$\begin{aligned}
\mathrm{Reg}_T &\geq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{a \neq a^\star} p_t(a) \mathbb{E}\big[\bar{\ell}_t(a) - \bar{\ell}_t(a^\star) \,\big|\, \mathcal{H}_{t-1}\big] \right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \neq a^\star} p_t(a) \Delta_{\mathcal{A}}(a) - \sum_{t=1}^{T} \xi_t (1 - p_t(a^\star)) \right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} p_t(a) \Delta_{\mathcal{A}}(a) \right] - C.
\end{aligned}$$

The first inequality compares regret with the particular fixed action $a^\star$; it remains valid if some other fixed action has smaller expected cumulative loss. Thus equation (13) implies the aggregate constraint for the learner under consideration. If it holds for every admissible learner with the same $\boldsymbol{\Delta}_{\mathcal{A}}$ and C, it establishes the regime-level definition equation (7). It is a sufficient way to verify that definition, not an additional assumption of our theorem.

Eldowa et al. (2024) unconditional gaps. For the correspondence with Eldowa et al. (2024), take $\theta_a = \boldsymbol{M}_{\cdot,a}$ and identify their outcome x with our mediator z. Their policy loss $\ell_t(\theta_a)$ is our $\bar{\ell}_t(a)$; both average over the mediator while keeping the latent loss vector $\boldsymbol{\ell}_t$ fixed. Their Section 4 assumes, in our notation,

$$\mathbb{E}[\bar{\ell}_t(a) - \bar{\ell}_t(a^\star)] \geq \Delta - \xi_t \quad (a \neq a^\star), \qquad \sum_{t=1}^{T} \xi_t \leq C, \tag{14}$$

where $\Delta \in (0, 1]$ is a common gap lower bound and $\xi_t \geq 0$ are deterministic. Their proof of Theorem 3 uses the aggregate lower bound

$$\mathrm{Reg}_T \geq \Delta \sum_{t=1}^{T} \mathbb{E}[1 - p_t(a^\star)] - C\,. \tag{15}$$

This is the common-gap version of equation (12), obtained by setting $\Delta_{\mathcal{A}}(a) = \Delta$ for $a \neq a^\star$ and $\Delta_{\mathcal{A}}(a^\star) = 0$. More generally, our aggregate condition implies equation (15) with $\Delta = \min_{a \neq a^\star} \Delta_{\mathcal{A}}(a)$.

However, equation (14) alone does not justify equation (15) when losses are dependent across rounds. The learner can select an action more often on histories where its conditional gap is small, so in general

$$\mathbb{E}\big[p_t(a)(\bar{\ell}_t(a) - \bar{\ell}_t(a^\star))\big] \neq \mathbb{E}[p_t(a)]\, \mathbb{E}[\bar{\ell}_t(a) - \bar{\ell}_t(a^\star)]\,.$$

An unconditional gap describes the action's average cost across possible histories, whereas regret weights each history by how often the learner selects the action there. Conditioning the gap assumption on $\mathcal{H}_{t-1}$, or imposing equation (15) directly, supplies the needed justification. In the usual i.i.d. model the current loss vector is independent of the past, so the unconditional and history-conditional expected gaps coincide.

A two-round illustration. Let $\boldsymbol{M} = \boldsymbol{I}_2$ and $T = 2$. Before play, choose either

$$\boldsymbol{\ell}_1 = \boldsymbol{\ell}_2 = (0, 1/2) \qquad \text{or} \qquad \boldsymbol{\ell}_1 = \boldsymbol{\ell}_2 = (0, 1),$$

each with probability $1/2$. Action 1 is uniquely optimal in both cases, and the unconditional gap of action 2 is $3/4$ in each round. Thus equation (14) holds with $a^\star = 1$, $\Delta = 3/4$, and $C = 0$.

Consider a learner that plays action 2 in round 1 and repeats it only if its observed loss is $1/2$; otherwise it switches to action 1. The two equally likely trajectories are

Loss of action 2	(A_1, A_2)	Total loss	Selections of action 2
$1/2$	$(2, 2)$	1	2
1	$(2, 1)$	1	1

Consequently,

$$\mathrm{Reg}_2 = 1 < \frac{9}{8} = \frac{3}{4}\sum_{t=1}^{2}\mathbb{E}[1 - p_t(1)]\,.$$

The unconditional condition holds, but the aggregate condition fails with $\boldsymbol{\Delta}_{\mathcal{A}} = (0, 3/4)$ and $C = 0$. One learner violating it suffices to rule out the regime-level self-bounding constraint at these parameters.

In round 2 the learner selects action 2 with probability $1/2$, and whenever it does so the loss is $1/2$. The actual expected regret contribution is therefore $1/4$, whereas charging the unconditional gap would give $(1/2)(3/4) = 3/8$. The first observation reveals which sequence was chosen; the environment never changes that sequence in response to play. The example identifies a missing implication in the argument, not a violation of the regret upper bound for their algorithm. Indeed, this environment satisfies a conditional gap bound, and therefore the aggregate constraint for every learner, with the smaller vector $\boldsymbol{\Delta}_{\mathcal{A}} = (0, 1/2)$ and $C = 0$.

Regret guarantees. For $|\mathcal{A}| \geq 2$, the bounds in Zimmert & Seldin (2021, Theorem 1), applied directly to the actions while ignoring the mediator identity and retaining the observed loss, imply

$$\mathrm{Reg}_T = O\Big(S_{\mathcal{A}}(1 + \log T) + \sqrt{CS_{\mathcal{A}}(1 + \log T)}\Big)\,.$$

Under the aggregate condition used in their proof, Eldowa et al. (2024, Theorem 3) gives, for $T, |\mathcal{A}| \geq 2$,

$$\mathrm{Reg}_T = O\left(\frac{\mathcal{C}(\boldsymbol{M})}{\Delta}\log T \log(|\mathcal{A}|T) + \sqrt{\frac{C\mathcal{C}(\boldsymbol{M})}{\Delta}\log T \log(|\mathcal{A}|T)} + \log|\mathcal{A}|\right)\,.$$

Our Theorem 4.1 gives

$$\mathrm{Reg}_T \leq B_T + 2\sqrt{CB_T}, \qquad B_T = K + \kappa_c \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}(1 + \log T)\,.$$

Here the complexities are those of equation (8)–equation (9), and K is independent of T and C for fixed $\boldsymbol{M}, \boldsymbol{\Delta}_{\mathcal{A}}, \alpha, c, \rho\lambda_\alpha$. The three bounds have square-root dependence on the self-bounding budget. Relative to the action-based Tsallis-INF bound, our logarithmic coefficient also adapts to the mediator complexity. Relative to the displayed EXP4 bound, our guarantee has $\log T$ rather than $(\log T)^2$ growth when $C = 0$ and the instance is fixed, but the complexity coefficients are not uniformly ordered. If $\boldsymbol{M}_{\cdot,a^\star}$ has full support, then $S^{\mathrm{a}}_{\mathcal{Z}} = 0$ and our bound becomes $K + 2\sqrt{CK}$, which is constant in T when C is fixed.

B Practical optimization for CTsallis-Mediator and CTsallis-INF

The FTRL updates are convex optimization problems over the action simplex. Tsallis-INF and CTsallis-Action have closed form solution to the optimization problem. However, neither CTsallis-Mediator nor CTsallis-INF does. CTsallis-Mediator can be optimized by working-set Newton steps, while CTsallis-INF admits an exact scalar dual formulation with smooth Newton subproblems. The latter formulation handles the maximum of the two regularizer terms without the need of smoothing it. We describe the details below.

Fix a round and divide the FTRL objective by $\sqrt{t}$ which does not change its minimizers. Write $\boldsymbol{c} = \widehat{\widehat{\boldsymbol{L}}}_{t-1}/\sqrt{t}$, let $\beta > 0$ denote the coefficient of the principal regularizer term, and write $\lambda \geq 0$ for the coefficient of ψ_α. All minimizations below are over $\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}$, and $\boldsymbol{q} = \boldsymbol{M}\boldsymbol{p}$. In cases where we know $\boldsymbol{M}$, we remove unreachable mediator rows before optimization.

B.1 CTSALLIS-MEDIATOR: WORKING-SET NEWTON METHOD

The mediator-only objective is

$$f_{\mathrm{M}}(\boldsymbol{p}) = \langle \boldsymbol{c}, \boldsymbol{p} \rangle + \beta \psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{p}) \,.$$

Optimizing directly over $\boldsymbol{p}$ enforces attainability of $\boldsymbol{q} = \boldsymbol{M}\boldsymbol{p}$ and returns the action distribution used for sampling. At any feasible point with $q_z > 0$ for every remaining mediator, its gradient and Hessian are

$$\nabla f_{\mathrm{M}}(\boldsymbol{p}) = \boldsymbol{c} - \frac{\beta}{2} \boldsymbol{M}^\top \boldsymbol{q}^{-1/2}, \qquad \nabla^2 f_{\mathrm{M}}(\boldsymbol{p}) = \frac{\beta}{4} \boldsymbol{M}^\top \operatorname{diag}(\boldsymbol{q}^{-3/2}) \boldsymbol{M},$$

where powers of vectors are coordinatewise. Action probabilities may be zero: differentiability here requires only positive mediator probabilities.

The working set initially consists of actions with positive probability in a feasible starting distribution. On this set, the solver computes a Newton direction whose coordinates sum to zero, thereby preserving the simplex equality constraint. Diagonal scaling and a small positive damping term stabilize the linear solve when the restricted Hessian is singular or poorly conditioned, which does not change the objective being minimized itself.

If the Newton direction is not a feasible descent direction, the solver instead transfers mass from a positive-probability action with a large gradient coordinate to an action with the smallest gradient coordinate. Backtracking chooses a step that keeps $\boldsymbol{p} \geq 0$, preserves $\boldsymbol{q} > 0$, and decreases the original objective. These steps permit actions to enter or leave the support. A previous solution provides a warm start; an initial feasible policy with $\boldsymbol{M}\boldsymbol{p} > 0$ must be supplied to the working-set routine.

Stopping criterion. Let $\boldsymbol{g}_{\mathrm{M}} = \nabla f_{\mathrm{M}}(\boldsymbol{p})$. Convexity gives the global objective-error bound

$$0 \leq f_{\mathrm{M}}(\boldsymbol{p}) - \min_{\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}} f_{\mathrm{M}}(\boldsymbol{u}) \leq \langle \boldsymbol{g}_{\mathrm{M}}, \boldsymbol{p} \rangle - \min_{a \in \mathcal{A}} g_{\mathrm{M}}(a) \,.$$

The solver accepts a policy only when this bound is below its tolerance. The minimum is taken over all actions, including those outside the working set, so stationarity on a restricted face alone is insufficient. This linearization bound is also used in Frank–Wolfe algorithms, but its use as a stopping certificate does not require Frank–Wolfe updates.

B.2 CTSALLIS-INF: A SCALAR WEIGHT SEARCH WITH NEWTON METHOD SUBPROBLEMS

For $\lambda > 0$, consider the normalized hybrid objective

$$f_{\mathrm{J}}(\boldsymbol{p}) = \langle \boldsymbol{c}, \boldsymbol{p} \rangle + \beta \max\{\psi_{\mathcal{A}}(\boldsymbol{p}), \psi_{\mathcal{J}}(\boldsymbol{p})\} + \lambda \psi_{\alpha}(\boldsymbol{p}) \,.$$

Its nonsmoothness can be handled by introducing a scalar v:

$$\begin{aligned} &\min_{\boldsymbol{p}, v} \ \langle \boldsymbol{c}, \boldsymbol{p} \rangle + \beta v + \lambda \psi_{\alpha}(\boldsymbol{p}) \\ &\text{subject to } \boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}, \qquad v \geq \psi_{\mathcal{A}}(\boldsymbol{p}), \quad v \geq \psi_{\mathcal{J}}(\boldsymbol{p}) \,. \end{aligned} \tag{16}$$

Because $\beta > 0$, an optimizer chooses $v = \max\{\psi_{\mathcal{A}}(\boldsymbol{p}), \psi_{\mathcal{J}}(\boldsymbol{p})\}$. Both inequality constraints are convex constraints, since their right-hand sides are convex.

The two inequality multipliers sum to β. Writing them as $\beta\theta$ and $\beta(1-\theta)$ gives

$$\begin{aligned} f_\theta(\boldsymbol{p}) &= \langle \boldsymbol{c}, \boldsymbol{p} \rangle + \beta \big[\theta \psi_{\mathcal{A}}(\boldsymbol{p}) + (1-\theta)\psi_{\mathcal{J}}(\boldsymbol{p})\big] + \lambda \psi_{\alpha}(\boldsymbol{p}), \\ d(\theta) &= \min_{\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}} f_\theta(\boldsymbol{p}), \qquad \min_{\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}} f_{\mathrm{J}}(\boldsymbol{p}) = \max_{\theta \in [0,1]} d(\theta) \,. \end{aligned} \tag{17}$$

The equality follows from strong duality: take a strictly positive action distribution and choose v strictly above both regularizer values to satisfy Slater's condition (Vandenberghe & Boyd, 2004, Section 5.2.3). Thus, the weighted regularizer in f_θ is an auxiliary optimization problem, not a replacement of the hybrid regularizer by a fixed sum.

Inner Newton solve. For each fixed θ, the term $\lambda\psi_\alpha$ makes the minimizer $\boldsymbol{p}_\theta$ unique and strictly positive. Consequently $\boldsymbol{M}\boldsymbol{p}_\theta$ is positive on every reachable mediator, and the inner objective is smooth where the solver operates. Set $h_a = \sum_{z\in\mathcal{Z}} \sqrt{M_{z,a}}$, so that $\psi_{\mathcal{J}}(\boldsymbol{p}) = -\sum_z \sqrt{q_z} + \langle \boldsymbol{h}, \boldsymbol{p}\rangle$. Then

$$\nabla f_\theta(\boldsymbol{p}) = \boldsymbol{c} - \frac{\beta\theta}{2}\boldsymbol{p}^{-1/2} - \beta(1-\theta)\left(\frac{1}{2}\boldsymbol{M}^\top \boldsymbol{q}^{-1/2} - \boldsymbol{h}\right) - \frac{\lambda}{1-\alpha}\boldsymbol{p}^{\alpha-1},$$

$$\nabla^2 f_\theta(\boldsymbol{p}) = \frac{\beta\theta}{4}\operatorname{diag}(\boldsymbol{p}^{-3/2}) + \frac{\beta(1-\theta)}{4}\boldsymbol{M}^\top \operatorname{diag}(\boldsymbol{q}^{-3/2})\boldsymbol{M} + \lambda \operatorname{diag}(\boldsymbol{p}^{\alpha-2}),$$

where the operations on the vectors are element-wise. Writing these as $\boldsymbol{g}_\theta$ and $\boldsymbol{K}_\theta$, respectively, the Newton direction $\boldsymbol{d}$ solves

$$\boldsymbol{K}_\theta \boldsymbol{d} + \boldsymbol{g}_\theta + \nu\mathbf{1} = \mathbf{0}, \qquad \mathbf{1}^\top \boldsymbol{d} = 0.$$

The implementation uses diagonal scaling and Cholesky solves. Each inner solve is warm-started from the preceding policy.

Outer weight search. The dual function is concave and, for exact inner minimizers, satisfies

$$d'(\theta) = \beta\left[-\psi_{\mathcal{J}}(\boldsymbol{p}_\theta) + \psi_{\mathcal{A}}(\boldsymbol{p}_\theta)\right].$$

Its derivative is nonincreasing, so the maximizing weight can be found by a bracketed scalar search. The implementation first tries the previous weight and the endpoints, then bisects the remaining interval. If $\psi_{\mathcal{A}}(\boldsymbol{p}_\theta) > \psi_{\mathcal{J}}(\boldsymbol{p}_\theta)$, it increases the weight on $\psi_{\mathcal{A}}$; if the inequality is reversed, it decreases that weight. An interior maximizing weight equalizes the two regularizer values. At an endpoint, the corresponding one-sided derivative condition suffices.

A certificate for the original nonsmooth objective. Accurate minimization of f_θ alone does not certify the hybrid update. For a feasible positive $\boldsymbol{p}$ and any $\theta \in [0,1]$, define

$$\mathcal{E}(\boldsymbol{p},\theta) = \beta\left[\max\{\psi_{\mathcal{A}}(\boldsymbol{p}), \psi_{\mathcal{J}}(\boldsymbol{p})\} - \theta\psi_{\mathcal{A}}(\boldsymbol{p}) - (1-\theta)\psi_{\mathcal{J}}(\boldsymbol{p})\right] + \langle \nabla f_\theta(\boldsymbol{p}), \boldsymbol{p}\rangle - \min_{a\in\mathcal{A}} [\nabla f_\theta(\boldsymbol{p})]_a .$$

The first term measures the discrepancy between the weighted and hybrid objectives; the second bounds the inner optimization error. Together,

$$0 \le f_{\mathrm{J}}(\boldsymbol{p}) - \min_{\boldsymbol{u}\in\Delta^{|\mathcal{A}|-1}} f_{\mathrm{J}}(\boldsymbol{u}) \le \mathcal{E}(\boldsymbol{p},\theta).$$

Indeed, $f_{\mathrm{J}}(\boldsymbol{u}) \ge f_\theta(\boldsymbol{u})$ for every feasible $\boldsymbol{u}$, and the tangent to the convex function f_θ supplies a lower bound on its minimum. The implementation checks this combined certificate, with a small floating-point allowance, before accepting a policy. It reports failure if its iteration or line-search limits are reached without certification.

Conic alternative and the variant without ψ_α. A conic optimizer can solve the same convex formulations to obtain an initial policy or an alternative when Newton does not return a certified solution. Square-root terms admit exact lifted constraints $x_a \ge 0$, $x_a^2 \le p_a$ and $y_z \ge 0$, $y_z^2 \le (\boldsymbol{M}\boldsymbol{p})_z$. For the hybrid problem, replace the two lower bounds on v by $1 - \sum_a x_a$ and $-\sum_z y_z + \langle \boldsymbol{h}, \boldsymbol{p}\rangle$. For $\lambda > 0$, the additional powers in ψ_α can be represented using power cones. For $\lambda = 0$, the resulting formulation is a second-order cone program and allows boundary solutions with some $p_a = 0$. The interior Newton routine described above assumes $\lambda > 0$; it must not be applied unchanged to this boundary case.

Numerical accuracy. Neither routine imposes a fixed positive floor on action probabilities or smooths the maximum. If the normalized objective-error certificate is at most ε_t, the corresponding error in the original round-t FTRL objective is at most $\sqrt{t}\,\varepsilon_t$. Note that the regret theorems from the main text concern exact minimizers.

Computational cost. For dense $\boldsymbol{M}$, a full gradient costs $O(|\mathcal{Z}||\mathcal{A}|)$ operations. With k actions in the mediator working set, forming and factoring its Newton matrix cost $O(|\mathcal{Z}|k^2 + k^3)$. A full hybrid inner Newton iteration costs $O(|\mathcal{Z}||\mathcal{A}|^2 + |\mathcal{A}|^3)$, in addition to line-search evaluations. Total cost also depends on the number of inner iterations and outer weight evaluations; warm starts can reduce both.

Table 3: Where each term of the regret decomposition equation (28) is bounded.

Term	Adversarial rate (Section C.5)	Gap-dependent rates (Sections C.6–C.8)
Penalty part	at most $D\sqrt{T}$ by equation (25)	Lemma C.15 for $-\psi_\wedge$; Proposition C.18 for $-\psi_\alpha$
Stability of the baseline $\bar{\boldsymbol{B}}_t$	Lemmas C.13 and C.12	Lemma C.16
Stability of the residual $\boldsymbol{x}_t$, Z_t produced by $a^\star$	Lemmas C.13 and C.12	Lemma C.16
Stability of the residual $\boldsymbol{x}_t$, Z_t not produced by $a^\star$	Lemmas C.13 and C.12	Lemma C.17
Combination	Theorem C.14	Proposition C.18, Lemma C.19, and Theorem C.21

C PROOFS

This appendix proves Theorem 4.1. Its precise versions, with explicit constants, are Theorem C.14 for adversarial losses, Theorem C.21 under the self-bounding constraint equation (7), and Corollary C.22 for stochastic losses.

The analysis follows the self-bounding approach of Zimmert & Seldin (2021, Section 7.1). An FTRL decomposition (Lemma C.2) bounds the regret by a penalty, given by the values of the regularizer along the iterates, and by one-round stability terms, which measure how much the loss estimate of a single round can gain against the curvature of the regularizer. Bounding the stability terms is the main obstacle, for two reasons. First, the hybrid regularizer $\psi_\wedge = \max\{\psi_\mathcal{A}, \psi_\mathcal{J}\}$ in equation (6) need not be differentiable when $\psi_\mathcal{A} = \psi_\mathcal{J}$. Second, the reduced-variance estimates equation (5) can be negative. We therefore write each estimate as a baseline, which is known before the round, plus a residual, which is nonzero only in the row of the mechanism matrix that belongs to the sampled mediator. Whichever of $\psi_\mathcal{A}$ and $\psi_\mathcal{J}$ is active, $\psi_\wedge$ is curved like a square root in the coordinate of each single mediator, and this curvature controls the residual. The auxiliary regularizer ψ_α in equation (6) supplies curvature in the action coordinates. Together, these give a uniform bound on the stability terms, and hence the adversarial rate. For the gap-dependent rates, we bound both the penalty and the stability terms by quantities that vanish as the iterates concentrate on a fixed action. Positive gaps control these quantities by $\Gamma(\boldsymbol{p}_t)$, and self-bounding completes the proof.

Table 3 shows which result bounds each term of the regret decomposition equation (28). Sections C.1–C.4 contain the tools shared by all bounds: the setting and the estimator, the decomposition itself, the curvature of the regularizer, and one-dimensional bounds on the stability terms.

C.1 PRELIMINARIES

We recall the setting, the estimator, and the algorithm of the main text.

Setting. The mechanism $\boldsymbol{M}$ has columns $\boldsymbol{M}_{\cdot,a} \in \Delta^{|\mathcal{Z}|-1}$, and, as in the main text, $\mathcal{Z}$ contains only reachable mediators. Let $\mathcal{F}_t$ be the σ-field generated by the learner's internal randomness used to compute $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ and by A_s, Z_s, and $\ell_s(Z_s)$ for $s \le t$, and write $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. In round t, the learner computes an $\mathcal{F}_{t-1}$-measurable distribution $\boldsymbol{p}_t \in \Delta^{|\mathcal{A}|-1}$ and samples an action $A_t \sim \boldsymbol{p}_t$; the environment samples a mediator $Z_t \sim \boldsymbol{M}_{\cdot,A_t}$; and the learner observes Z_t and the loss $\ell_t(Z_t) \in [0,1]$, which is the Z_t-coordinate of a loss vector $\boldsymbol{\ell}_t \in [0,1]^{|\mathcal{Z}|}$. Given $\mathcal{F}_{t-1}$, the mediator Z_t has distribution $\boldsymbol{q}_t = \boldsymbol{M}\boldsymbol{p}_t$. We assume that

$$\boldsymbol{\ell}_t \text{ is conditionally independent of } (A_t, Z_t) \text{ given } \mathcal{F}_{t-1}; \tag{18}$$

the vector $\boldsymbol{\ell}_t$ may depend on $\mathcal{F}_{t-1}$. In the adversarial game, $\boldsymbol{\ell}_t$ is fixed before the game starts; in the stochastic game, $\boldsymbol{\ell}_t$ is independent of $(\mathcal{F}_{t-1}, A_t, Z_t)$ with $\mathbb{E}_{t-1}[\boldsymbol{\ell}_t] = \boldsymbol{\mu}$; and the regime with a self-bounding constraint equation (7) restricts such loss processes further. The loss vector of the actions is $\bar{\boldsymbol{\ell}}_t = \boldsymbol{M}^\top \boldsymbol{\ell}_t$, with coordinates $\bar{\ell}_t(a) = \langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{\ell}_t \rangle$. Since $\boldsymbol{p}_t$ is $\mathcal{F}_{t-1}$-measurable, equation (18)

gives $\mathbb{E}_{t-1}[\bar{\ell}_t(A_t)] = \mathbb{E}_{t-1}[\langle \boldsymbol{M}\boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle]$, so the pseudo-regret Reg_T is

$$\mathrm{Reg}_T = \max_{b\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{M}\boldsymbol{p}_t - \boldsymbol{M}_{\cdot,b}, \boldsymbol{\ell}_t\rangle\right], \tag{19}$$

where the comparator b is fixed before the expectation is taken.

Estimator. With $\eta_t^2 = 16/t$, the baseline $\boldsymbol{B}_t$ with coordinates $B_t(z)$, the mediator estimate $\widehat{\boldsymbol{\ell}}_t$ with coordinates $\widehat{\ell}_t(z)$, the action estimate $\widehat{\bar{\boldsymbol{\ell}}}_t$, and the cumulative estimate $\widehat{\bar{\boldsymbol{L}}}_t$ of equation (5) are

$$\begin{gathered} B_t(z) = \tfrac{1}{2}\mathbb{1}\{q_t(z) \geq 16/t\}, \qquad \widehat{\ell}_t(z) = B_t(z) + \frac{\mathbb{1}\{Z_t = z\}}{q_t(z)}\left(\ell_t(z) - B_t(z)\right), \\ \widehat{\bar{\boldsymbol{\ell}}}_t = \boldsymbol{M}^\top \widehat{\boldsymbol{\ell}}_t, \quad \text{and} \quad \widehat{\bar{\boldsymbol{L}}}_t = \sum_{s=1}^{t} \widehat{\bar{\boldsymbol{\ell}}}_s, \ \text{ with } \ \widehat{\bar{\boldsymbol{L}}}_0 = \boldsymbol{0}. \end{gathered} \tag{20}$$

Regularizers and update. As in Orabona (2026), we write regularizers as convex functions. The regularizers of the main text are, for $\alpha \in (1/2, 1)$,

$$\begin{gathered} \psi_{\mathcal{A}}(\boldsymbol{p}) = 1 - \sum_{a\in\mathcal{A}} \sqrt{p_a}, \qquad \psi_{\mathcal{Z}}(\boldsymbol{q}) = 1 - \sum_{z\in\mathcal{Z}} \sqrt{q_z}, \qquad \psi_\alpha(\boldsymbol{p}) = \frac{1 - \sum_{a\in\mathcal{A}} p_a^\alpha}{\alpha(1-\alpha)}, \\ \psi_{\mathcal{J}}(\boldsymbol{p}) = \psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{p}) - \sum_{a\in\mathcal{A}} p_a \psi_{\mathcal{Z}}(\boldsymbol{M}_{\cdot,a}), \quad \text{and} \quad \psi_\wedge = \max\{\psi_{\mathcal{A}}, \psi_{\mathcal{J}}\}, \end{gathered}$$

with $\Lambda = \rho\lambda_\alpha$, the update equation (6) is FTRL with the regularizers $\psi_t = \sqrt{t}\,\psi$, where $\psi = c\psi_\wedge + \Lambda\psi_\alpha$:

$$\boldsymbol{p}_t = \operatorname*{arg\,min}_{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}} \left\langle \widehat{\bar{\boldsymbol{L}}}_{t-1}, \boldsymbol{p}\right\rangle + \psi_t(\boldsymbol{p}). \tag{21}$$

If $|\mathcal{A}| = 1$ or $|\mathcal{Z}| = 1$, the loss $\bar{\ell}_t(a)$ does not depend on a, and the regret is zero; in these cases we set $\lambda_\alpha = 1$. Henceforth, $|\mathcal{A}| \geq 2$ and $|\mathcal{Z}| \geq 2$. Throughout, $c > 1/4$, $\rho > 0$, and $\lambda_\alpha > 0$ are arbitrary; only the second part of Theorem C.14 uses the calibration equation (45). We write

$$\mathcal{V} = \left\{\boldsymbol{w} \in \mathbb{R}^{|\mathcal{A}|} : \langle \boldsymbol{1}, \boldsymbol{w}\rangle = 0\right\}$$

for the tangent space of $\Delta^{|\mathcal{A}|-1}$.

Properties of the estimator. By equation (18), for every $z \in \mathcal{Z}$, $\mathbb{E}_{t-1}[\mathbb{1}\{Z_t = z\}\ell_t(z)] = q_t(z)\mathbb{E}_{t-1}[\ell_t(z)]$, and Corollary C.5 below shows that $q_t(z) > 0$. Hence, the factor $q_t(z)$ cancels the importance weight $1/q_t(z)$ in equation (20), and

$$\mathbb{E}_{t-1}[\widehat{\boldsymbol{\ell}}_t] = \mathbb{E}_{t-1}[\boldsymbol{\ell}_t], \qquad \mathbb{E}_{t-1}[\widehat{\bar{\boldsymbol{\ell}}}_t] = \mathbb{E}_{t-1}[\bar{\boldsymbol{\ell}}_t], \quad \text{and} \quad \mathbb{E}_{t-1}[|\widehat{\ell}_t(z)|] \leq \tfrac{3}{2}, \tag{22}$$

where the last bound uses $B_t(z) \leq 1/2$ and $|\ell_t(z) - B_t(z)| \leq 1$. The centered observed loss $\varepsilon_t = \ell_t(Z_t) - B_t(Z_t)$ satisfies

$$\varepsilon_t \in [0,1] \text{ if } q_t(Z_t) < 16/t, \quad \text{and} \quad |\varepsilon_t| \leq \tfrac{1}{2} \text{ if } q_t(Z_t) \geq 16/t; \tag{23}$$

in particular, $\varepsilon_t < 0$ only if $q_t(Z_t) \geq 16/t$. We split the estimate into a predictable baseline and a residual supported on the row of the sampled mediator:

$$\widehat{\bar{\boldsymbol{\ell}}}_t = \bar{\boldsymbol{B}}_t + \boldsymbol{x}_t, \qquad \bar{\boldsymbol{B}}_t = \boldsymbol{M}^\top \boldsymbol{B}_t, \quad \text{and} \quad \boldsymbol{x}_t = \frac{\varepsilon_t}{q_t(Z_t)} \boldsymbol{M}^\top_{Z_t,\cdot}. \tag{24}$$

Each coordinate $\bar{B}_t(a) = \langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{B}_t\rangle$ lies in $[0, 1/2]$, while $\boldsymbol{x}_t$ can have either sign.

Range of the regularizer. Each of $\psi_{\mathcal{A}}$, $\psi_{\mathcal{J}}$, $\psi_\wedge$, and ψ_α is convex and nonpositive on $\Delta^{|\mathcal{A}|-1}$ and vanishes at every $\boldsymbol{e}_a$. Since $\psi_{\mathcal{Z}} \leq 0$ on $\Delta^{|\mathcal{Z}|-1}$, we have $\psi_{\mathcal{J}}(\boldsymbol{p}) \geq \psi_{\mathcal{Z}}(\boldsymbol{M}\boldsymbol{p}) \geq 1 - \sqrt{|\mathcal{Z}|}$; also $\psi_{\mathcal{A}} \geq 1 - \sqrt{|\mathcal{A}|}$ and, by Jensen's inequality, $\psi_\alpha \geq -D_\alpha$. Therefore

$$\begin{gathered} -D \leq \psi \leq 0, \qquad \text{where} \\ D_\alpha = \frac{|\mathcal{A}|^{1-\alpha} - 1}{\alpha(1-\alpha)} \quad \text{and} \quad D = c\min\left\{\sqrt{|\mathcal{A}|} - 1, \sqrt{|\mathcal{Z}|} - 1\right\} + \Lambda D_\alpha. \end{gathered} \tag{25}$$

Subgradients, Bregman divergences, and stability. Let f be convex on $\Delta^{|\mathcal{A}|-1}$. A vector $\boldsymbol{\xi} \in \mathbb{R}^{|\mathcal{A}|}$ is a subgradient of f at $\boldsymbol{p}$ if $f(\boldsymbol{u}) \geq f(\boldsymbol{p}) + \langle \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{p} \rangle$ for every $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$, and $\partial f(\boldsymbol{p})$ denotes the set of subgradients. Since $\langle \mathbf{1}, \boldsymbol{u} - \boldsymbol{p} \rangle = 0$, we have $\partial f(\boldsymbol{p}) + \mathbb{R}\mathbf{1} = \partial f(\boldsymbol{p})$. For $\boldsymbol{\xi} \in \partial f(\boldsymbol{p})$, we extend the definition of Bregman divergence to non-differentiable functions as

$$B_f^{\boldsymbol{\xi}}(\boldsymbol{u}; \boldsymbol{p}) = f(\boldsymbol{u}) - f(\boldsymbol{p}) - \langle \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{p} \rangle \geq 0 .$$

When f is differentiable on a neighborhood of $\boldsymbol{p}$ in $\mathbb{R}^{|\mathcal{A}|}$, we write $B_f(\boldsymbol{u}; \boldsymbol{p})$ for the choice $\boldsymbol{\xi} = \nabla f(\boldsymbol{p})$. For $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{A}|}$, the stability function

$$\operatorname{stab}_{f,\boldsymbol{p}}^{\boldsymbol{\xi}}(\boldsymbol{v}) = \sup_{\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}} \left\{ \langle \boldsymbol{v}, \boldsymbol{p} - \boldsymbol{u} \rangle - B_f^{\boldsymbol{\xi}}(\boldsymbol{u}; \boldsymbol{p}) \right\} \tag{26}$$

is the largest gain of the linear loss $\boldsymbol{v}$ against the curvature of f at $\boldsymbol{p}$. It is nonnegative (take $\boldsymbol{u} = \boldsymbol{p}$), and it does not change when $\boldsymbol{v}$ or $\boldsymbol{\xi}$ is shifted by a multiple of $\mathbf{1}$.

Lemma C.1 (Calculus of the stability function)**.** *Let f, f' be convex on $\Delta^{|\mathcal{A}|-1}$, $\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}$, $\boldsymbol{\xi} \in \partial f(\boldsymbol{p})$, $\boldsymbol{\xi}' \in \partial f'(\boldsymbol{p})$, and $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^{|\mathcal{A}|}$.*

(a) If $B_f^{\boldsymbol{\xi}}(\boldsymbol{u}; \boldsymbol{p}) \geq B_{f'}^{\boldsymbol{\xi}'}(\boldsymbol{u}; \boldsymbol{p})$ for every $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$, then $\operatorname{stab}_{f,\boldsymbol{p}}^{\boldsymbol{\xi}}(\boldsymbol{v}) \leq \operatorname{stab}_{f',\boldsymbol{p}}^{\boldsymbol{\xi}'}(\boldsymbol{v})$.

(b) $\operatorname{stab}_{f+f',\boldsymbol{p}}^{\boldsymbol{\xi}+\boldsymbol{\xi}'}(\boldsymbol{v} + \boldsymbol{v}') \leq \operatorname{stab}_{f,\boldsymbol{p}}^{\boldsymbol{\xi}}(\boldsymbol{v}) + \operatorname{stab}_{f',\boldsymbol{p}}^{\boldsymbol{\xi}'}(\boldsymbol{v}')$, and $\operatorname{stab}_{\theta f,\boldsymbol{p}}^{\theta \boldsymbol{\xi}}(\theta \boldsymbol{v}) = \theta \operatorname{stab}_{f,\boldsymbol{p}}^{\boldsymbol{\xi}}(\boldsymbol{v})$ for $\theta \geq 0$.

(c) Suppose that $\boldsymbol{p} \in \operatorname{relint} \Delta^{|\mathcal{A}|-1}$ and that $f(\boldsymbol{u}) - \sum_{a \in \mathcal{A}} \phi_a(u_a)$ is affine in $\boldsymbol{u}$, where each ϕ_a is convex on $[0, \infty)$ and differentiable near p_a. Then

$$\operatorname{stab}_{f,\boldsymbol{p}}^{\nabla f(\boldsymbol{p})}(\boldsymbol{v}) \leq \sum_{a \in \mathcal{A}} \sup_{y \geq 0} \left\{ v_a(p_a - y) - B_{\phi_a}(y; p_a) \right\} .$$

(d) Suppose that $f(\boldsymbol{u}) - \sum_{z \in \mathcal{Z}} \phi_z((\boldsymbol{M}\boldsymbol{u})_z)$ is affine in $\boldsymbol{u}$, where each ϕ_z is convex on $[0, \infty)$ and differentiable near $q_z = (\boldsymbol{M}\boldsymbol{p})_z > 0$. Then, for every $\boldsymbol{w} \in \mathbb{R}^{|\mathcal{Z}|}$,

$$\operatorname{stab}_{f,\boldsymbol{p}}^{\nabla f(\boldsymbol{p})}\left(\boldsymbol{M}^\top \boldsymbol{w}\right) \leq \sum_{z \in \mathcal{Z}} \sup_{y \geq 0} \left\{ w_z(q_z - y) - B_{\phi_z}(y; q_z) \right\} .$$

Proof. Part (a) is immediate. For part (b), $B_{f+f'}^{\boldsymbol{\xi}+\boldsymbol{\xi}'} = B_f^{\boldsymbol{\xi}} + B_{f'}^{\boldsymbol{\xi}'}$, so the supremum of the sum is at most the sum of the suprema; the scaling identity follows from $B_{\theta f}^{\theta \boldsymbol{\xi}} = \theta B_f^{\boldsymbol{\xi}}$. In part (c), the affine part does not affect the divergence, so $B_f(\boldsymbol{u}; \boldsymbol{p}) = \sum_{a \in \mathcal{A}} B_{\phi_a}(u_a; p_a)$ and $\langle \boldsymbol{v}, \boldsymbol{p} - \boldsymbol{u} \rangle = \sum_{a \in \mathcal{A}} v_a(p_a - u_a)$; enlarging the domain from $\Delta^{|\mathcal{A}|-1}$ to $[0, \infty)^{\mathcal{A}}$ decouples the coordinates. In part (d), the affine part does not affect the divergence, so $B_f(\boldsymbol{u}; \boldsymbol{p}) = \sum_{z \in \mathcal{Z}} B_{\phi_z}((\boldsymbol{M}\boldsymbol{u})_z; q_z)$, while $\left\langle \boldsymbol{M}^\top \boldsymbol{w}, \boldsymbol{p} - \boldsymbol{u} \right\rangle = \langle \boldsymbol{w}, \boldsymbol{q} - \boldsymbol{M}\boldsymbol{u} \rangle$; since $\boldsymbol{M}\boldsymbol{u} \in [0, \infty)^{|\mathcal{Z}|}$, the same enlargement applies. □

C.2 The FTRL decomposition

All regret bounds start from the following decomposition. It is the FTRL regret decomposition of Zimmert & Seldin (2021, Section 7.1), written with the conjugate of the regularizer as in Orabona (2026); see also Lee & Orabona (2025, Section 3). It splits the regret into stability terms s_t, which measure how much the loss estimate of round t can gain against the curvature of ψ_t at $\boldsymbol{p}_t$, and a penalty given by the values of ψ along the iterates.

Lemma C.2 (FTRL decomposition)**.** *Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T \in \mathbb{R}^{|\mathcal{A}|}$, $\boldsymbol{L}_0 = \mathbf{0}$, and $\boldsymbol{L}_t = \sum_{s=1}^t \boldsymbol{v}_s$, and let $\boldsymbol{p}_t$ minimize $\langle \boldsymbol{L}_{t-1}, \boldsymbol{p} \rangle + \psi_t(\boldsymbol{p})$ over $\Delta^{|\mathcal{A}|-1}$. Let $\psi_t^\star(\boldsymbol{w}) = \max_{\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}} \langle \boldsymbol{w}, \boldsymbol{p} \rangle - \psi_t(\boldsymbol{p})$ for $\boldsymbol{w} \in \mathbb{R}^{|\mathcal{A}|}$, and*

$$s_t = \langle \boldsymbol{v}_t, \boldsymbol{p}_t \rangle + \psi_t^\star(-\boldsymbol{L}_t) - \psi_t^\star(-\boldsymbol{L}_{t-1}) .$$

Then $-\boldsymbol{L}_{t-1} \in \partial \psi_t(\boldsymbol{p}_t)$ and $s_t = \operatorname{stab}_{\psi_t,\boldsymbol{p}_t}^{-\boldsymbol{L}_{t-1}}(\boldsymbol{v}_t) \geq 0$. Moreover, for every $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$,

$$\sum_{t=1}^T \langle \boldsymbol{v}_t, \boldsymbol{p}_t - \boldsymbol{u} \rangle \leq \sum_{t=1}^T s_t + D + \psi(\boldsymbol{u}) + \sum_{t=2}^T \left(\sqrt{t} - \sqrt{t-1} \right) \left[\psi(\boldsymbol{u}) - \psi(\boldsymbol{p}_t) \right] . \tag{27}$$

Proof. Optimality of $\boldsymbol{p}_t$ means $\psi_t(\boldsymbol{u}) + \langle \boldsymbol{L}_{t-1}, \boldsymbol{u}\rangle \geq \psi_t(\boldsymbol{p}_t) + \langle \boldsymbol{L}_{t-1}, \boldsymbol{p}_t\rangle$ for all $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$, that is, $-\boldsymbol{L}_{t-1} \in \partial\psi_t(\boldsymbol{p}_t)$; it also gives $\psi_t^\star(-\boldsymbol{L}_{t-1}) = -\langle \boldsymbol{L}_{t-1}, \boldsymbol{p}_t\rangle - \psi_t(\boldsymbol{p}_t)$. Substituting this value and the definition of $\psi_t^\star(-\boldsymbol{L}_t)$ into s_t,

$$s_t = \sup_{\boldsymbol{u}\in\Delta^{|\mathcal{A}|-1}} \{\langle \boldsymbol{v}_t, \boldsymbol{p}_t - \boldsymbol{u}\rangle - [\psi_t(\boldsymbol{u}) - \psi_t(\boldsymbol{p}_t) + \langle \boldsymbol{L}_{t-1}, \boldsymbol{u} - \boldsymbol{p}_t\rangle]\} = \mathrm{stab}_{\psi_t,\boldsymbol{p}_t}^{-\boldsymbol{L}_{t-1}}(\boldsymbol{v}_t)\ .$$

For the decomposition, sum the definition of s_t over t and rearrange:

$$\sum_{t=1}^{T} \langle \boldsymbol{v}_t, \boldsymbol{p}_t\rangle = \sum_{t=1}^{T} s_t + \psi_1^\star(\boldsymbol{0}) - \psi_T^\star(-\boldsymbol{L}_T) + \sum_{t=2}^{T} \left[\psi_t^\star(-\boldsymbol{L}_{t-1}) - \psi_{t-1}^\star(-\boldsymbol{L}_{t-1})\right]\ .$$

We bound the three remaining terms. First, $\psi_1^\star(\boldsymbol{0}) = -\min_{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}} \psi(\boldsymbol{p}) \leq D$ by equation (25). Second, $-\psi_T^\star(-\boldsymbol{L}_T) \leq \langle \boldsymbol{L}_T, \boldsymbol{u}\rangle + \sqrt{T}\,\psi(\boldsymbol{u})$. Third, evaluating the maximum that defines $\psi_{t-1}^\star(-\boldsymbol{L}_{t-1})$ at $\boldsymbol{p}_t$ gives

$$\psi_t^\star(-\boldsymbol{L}_{t-1}) - \psi_{t-1}^\star(-\boldsymbol{L}_{t-1}) \leq -\left(\sqrt{t} - \sqrt{t-1}\right)\psi(\boldsymbol{p}_t)\ .$$

Inserting these bounds and writing $\sqrt{T}\,\psi(\boldsymbol{u}) = \psi(\boldsymbol{u}) + \sum_{t=2}^{T}\left(\sqrt{t} - \sqrt{t-1}\right)\psi(\boldsymbol{u})$ proves equation (27). □

We apply Lemma C.2 with $\boldsymbol{v}_t = \widehat{\widehat{\boldsymbol{\ell}}}_t$ and a pure comparator $\boldsymbol{u} = \boldsymbol{e}_b$, for which $\psi(\boldsymbol{e}_b) = 0$. By equation (22) and the predictability of $\boldsymbol{p}_t$, $\mathbb{E}[\langle \widehat{\widehat{\boldsymbol{\ell}}}_t, \boldsymbol{p}_t - \boldsymbol{e}_b\rangle] = \mathbb{E}[\langle \boldsymbol{M}\boldsymbol{p}_t - \boldsymbol{M}_{\cdot,b}, \boldsymbol{\ell}_t\rangle]$, the estimates being integrable. The resulting bound does not depend on b, so maximizing over b gives

$$\mathrm{Reg}_T \leq \mathbb{E}\left[\sum_{t=1}^{T} s_t\right] + D + \sum_{t=2}^{T}\left(\sqrt{t} - \sqrt{t-1}\right)\mathbb{E}\left[-\psi(\boldsymbol{p}_t)\right], \tag{28}$$

where s_t is the stability term with $\boldsymbol{v}_t = \widehat{\widehat{\boldsymbol{\ell}}}_t$. We call the first sum the stability part and the remaining terms the penalty part. The penalty part is at most $D\sqrt{T}$ by equation (25), so the adversarial rate only needs a uniform bound on $\mathbb{E}_{t-1}[s_t]$ (Section C.5); the gap-dependent rates need bounds on both parts that vanish as $\boldsymbol{p}_t$ approaches the vertex $\boldsymbol{e}_{a^\star}$ corresponding to a reference action $a^\star$ (Section C.6).

C.3 Structure of the regularizer

The stability term s_t is a supremum involving the Bregman divergence of $\psi_t = \sqrt{t}\,(c\psi_\wedge + \Lambda\psi_\alpha)$ at $\boldsymbol{p}_t$. Two issues complicate its analysis. First, $\psi_\wedge$ is a maximum and is not differentiable when $\psi_{\mathcal{A}} = \psi_{\mathcal{J}}$. Second, we must repeatedly split the divergence of a sum among its summands and discard a smooth part. This section provides the corresponding tools.

- Lemma C.3 splits a subgradient of $f_1 + f_2$ when f_2 is smooth. It is applied with $f_2 = \Lambda\sqrt{t}\,\psi_\alpha$ (Corollary C.5), with $f_2 = -\sqrt{(\boldsymbol{M}\cdot)_z}$ (Lemma C.7), with $f_2 = \kappa\psi_\alpha$ (Lemma C.10), and with $f_2 = -\sigma\sqrt{(\boldsymbol{M}\cdot)_z}$ (Lemma C.11).
- Lemma C.4 shows that the iterates are in the relative interior of the simplex and that every subgradient of $\psi_\wedge$ is, up to a multiple of $\mathbf{1}$, a convex combination of the gradients of the active functions; Corollary C.5 fixes the subgradient used in the rest of the appendix.
- Lemmas C.6 and C.7 show that, whichever function is active, $\psi_\wedge$ is curved like a square root in each mediator coordinate. This controls the residual $\boldsymbol{x}_t$ of equation (24), which is supported on the row of the sampled mediator.

Lemma C.3 (Removing a smooth summand)**.** *Let $f = f_1 + f_2$, where f_1 and f_2 are convex on $\Delta^{|\mathcal{A}|-1}$ and f_2 is differentiable on a neighborhood in $\mathbb{R}^{|\mathcal{A}|}$ of $\boldsymbol{p} \in \mathrm{relint}\,\Delta^{|\mathcal{A}|-1}$. If $\boldsymbol{\xi} \in \partial f(\boldsymbol{p})$, then $\boldsymbol{\xi} - \nabla f_2(\boldsymbol{p}) \in \partial f_1(\boldsymbol{p})$ and $B_f^{\boldsymbol{\xi}}(\boldsymbol{u};\boldsymbol{p}) = B_{f_1}^{\boldsymbol{\xi}-\nabla f_2(\boldsymbol{p})}(\boldsymbol{u};\boldsymbol{p}) + B_{f_2}(\boldsymbol{u};\boldsymbol{p})$.*

Proof. Fix $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$ and let $\boldsymbol{u}_\tau = \boldsymbol{p} + \tau(\boldsymbol{u} - \boldsymbol{p})$ for $\tau \in (0,1]$. Convexity of f_1 and f_2 gives

$$0 \leq f(\boldsymbol{u}_\tau) - f(\boldsymbol{p}) - \tau\langle \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{p}\rangle \leq \tau[f_1(\boldsymbol{u}) - f_1(\boldsymbol{p})] + f_2(\boldsymbol{u}_\tau) - f_2(\boldsymbol{p}) - \tau\langle \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{p}\rangle\ .$$

Dividing by τ and letting $\tau \to 0$ gives $f_1(\boldsymbol{u}) \geq f_1(\boldsymbol{p}) + \langle \boldsymbol{\xi} - \nabla f_2(\boldsymbol{p}), \boldsymbol{u} - \boldsymbol{p}\rangle$. The identity for the divergences follows by expanding both sides. □

Lemma C.4 (Interior minimizers and active subgradients). *For every $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{A}|}$, $\sigma \geq 0$, and $\kappa > 0$, the function $\langle \boldsymbol{v}, \boldsymbol{p} \rangle + \sigma\psi_\wedge(\boldsymbol{p}) + \kappa\psi_\alpha(\boldsymbol{p})$ has a unique minimizer over $\Delta^{|\mathcal{A}|-1}$, and this minimizer lies in* $\operatorname{relint}\Delta^{|\mathcal{A}|-1}$.

Moreover, let $\boldsymbol{p} \in \operatorname{relint}\Delta^{|\mathcal{A}|-1}$, then

(a) Every $\boldsymbol{\xi} \in \partial\psi_\wedge(\boldsymbol{p})$ has the form

$$\boldsymbol{\xi} = \theta\nabla\psi_{\mathcal{A}}(\boldsymbol{p}) + (1-\theta)\nabla\psi_{\mathcal{J}}(\boldsymbol{p}) + d\mathbf{1}, \qquad \text{with } \theta \in [0,1] \text{ and } d \in \mathbb{R}, \tag{29}$$

where $\theta = 1$ if $\psi_{\mathcal{A}}(\boldsymbol{p}) > \psi_{\mathcal{J}}(\boldsymbol{p})$ and $\theta = 0$ if $\psi_{\mathcal{A}}(\boldsymbol{p}) < \psi_{\mathcal{J}}(\boldsymbol{p})$.

(b) For $\boldsymbol{\xi}$ and θ as in part (a) and every $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$,

$$B^{\boldsymbol{\xi}}_{\psi_\wedge}(\boldsymbol{u};\boldsymbol{p}) \geq \theta B_{\psi_{\mathcal{A}}}(\boldsymbol{u};\boldsymbol{p}) + (1-\theta)B_{\psi_{\mathcal{J}}}(\boldsymbol{u};\boldsymbol{p})\,. \tag{30}$$

Proof. The objective is continuous on a compact set, so it has a minimizer, which is unique because ψ_α is strictly convex. Suppose that a minimizer $\boldsymbol{p}$ has $p_a = 0$, and let $\boldsymbol{p}_\tau = (1-\tau)\boldsymbol{p} + \tau\boldsymbol{e}_a$. Convexity and $\psi_\wedge(\boldsymbol{e}_a) = 0$ give $\psi_\wedge(\boldsymbol{p}_\tau) - \psi_\wedge(\boldsymbol{p}) \leq -\tau\psi_\wedge(\boldsymbol{p})$, and $(1-\tau)^\alpha \geq 1-\tau$ gives

$$\psi_\alpha(\boldsymbol{p}_\tau) - \psi_\alpha(\boldsymbol{p}) \leq -\frac{\tau^\alpha - \tau\sum_{b\in\mathcal{A}} p_b^\alpha}{\alpha(1-\alpha)}\,.$$

Hence, the objective changes by at most a constant multiple of τ minus $\kappa\tau^\alpha/[\alpha(1-\alpha)]$, which is negative for small enough $\tau > 0$, which contradicts minimality.

For part (a), every $z \in \mathcal{Z}$ has $(\boldsymbol{M}\boldsymbol{p})_z > 0$, so $\psi_{\mathcal{A}}$ and $\psi_{\mathcal{J}}$ are differentiable on a neighborhood of $\boldsymbol{p}$ in $(0,\infty)^{\mathcal{A}}$. Let $\mathcal{I}_{\mathrm{act}}$ be the set of active functions, those among $\psi_{\mathcal{A}}, \psi_{\mathcal{J}}$ that equal $\psi_\wedge$ at $\boldsymbol{p}$. For $\boldsymbol{w} \in \mathcal{V}$, the directional derivative of the maximum is $\psi'_\wedge(\boldsymbol{p};\boldsymbol{w}) = \max_{f\in\mathcal{I}_{\mathrm{act}}}\langle\nabla f(\boldsymbol{p}), \boldsymbol{w}\rangle$, and the subgradient inequality at $\boldsymbol{u} = \boldsymbol{p} + \tau\boldsymbol{w}$ gives $\langle\boldsymbol{\xi}, \boldsymbol{w}\rangle \leq \psi'_\wedge(\boldsymbol{p};\boldsymbol{w})$. If the orthogonal projection of $\boldsymbol{\xi}$ onto $\mathcal{V}$ were outside the convex hull of the projections of $\nabla f(\boldsymbol{p})$, $f \in \mathcal{I}_{\mathrm{act}}$, a separating direction $\boldsymbol{w} \in \mathcal{V}$ would give $\langle\boldsymbol{\xi}, \boldsymbol{w}\rangle > \max_{f\in\mathcal{I}_{\mathrm{act}}}\langle\nabla f(\boldsymbol{p}), \boldsymbol{w}\rangle$, a contradiction. This proves equation (29). For part (b), the choice of θ gives $\psi_\wedge(\boldsymbol{p}) = \theta\psi_{\mathcal{A}}(\boldsymbol{p}) + (1-\theta)\psi_{\mathcal{J}}(\boldsymbol{p})$, while $\psi_\wedge(\boldsymbol{u}) \geq \theta\psi_{\mathcal{A}}(\boldsymbol{u}) + (1-\theta)\psi_{\mathcal{J}}(\boldsymbol{u})$; subtracting the affine function $\boldsymbol{u} \mapsto \psi_\wedge(\boldsymbol{p}) + \langle\boldsymbol{\xi}, \boldsymbol{u}-\boldsymbol{p}\rangle$ gives equation (30). □

Corollary C.5 (The decisions). *For every $t \geq 1$, $\boldsymbol{p}_t \in \operatorname{relint}\Delta^{|\mathcal{A}|-1}$, $q_t(z) > 0$ for every $z \in \mathcal{Z}$, and $\boldsymbol{p}_t$ is $\mathcal{F}_{t-1}$-measurable. Moreover, there exist an $\mathcal{F}_{t-1}$-measurable $\theta_t \in [0,1]$ and $\boldsymbol{\xi}^\wedge_t = \theta_t\nabla\psi_{\mathcal{A}}(\boldsymbol{p}_t) + (1-\theta_t)\nabla\psi_{\mathcal{J}}(\boldsymbol{p}_t) \in \partial\psi_\wedge(\boldsymbol{p}_t)$ such that*

$$-\widehat{\widetilde{\boldsymbol{L}}}_{t-1} \in c\sqrt{t}\,\boldsymbol{\xi}^\wedge_t + \Lambda\sqrt{t}\,\nabla\psi_\alpha(\boldsymbol{p}_t) + \mathbb{R}\mathbf{1}, \tag{31}$$

where $\theta_t = 1$ if $\psi_{\mathcal{A}}(\boldsymbol{p}_t) > \psi_{\mathcal{J}}(\boldsymbol{p}_t)$ and $\theta_t = 0$ if $\psi_{\mathcal{A}}(\boldsymbol{p}_t) < \psi_{\mathcal{J}}(\boldsymbol{p}_t)$.

Proof. Equation 31 is the first-order optimality condition of equation (21). Apply Lemma C.4 with $\boldsymbol{v} = \widehat{\widetilde{\boldsymbol{L}}}_{t-1}$, $\sigma = c\sqrt{t}$, and $\kappa = \Lambda\sqrt{t}$, inductively in t starting from $\widehat{\widetilde{\boldsymbol{L}}}_0 = \mathbf{0}$: the minimizer is in the relative interior of the simplex, so $q_t(z) > 0$ and $\widehat{\widetilde{\boldsymbol{L}}}_t$ is finite. The minimizer depends continuously on $\boldsymbol{v}$: for a convergent sequence of coefficients, compactness gives limit points of the minimizers, and uniqueness identifies each of them with the minimizer for the limit coefficient. Hence, $\boldsymbol{p}_t$ is $\mathcal{F}_{t-1}$-measurable. By Lemma C.2, $-\widehat{\widetilde{\boldsymbol{L}}}_{t-1} \in \partial\psi_t(\boldsymbol{p}_t)$, and Lemma C.3 with $f_1 = c\sqrt{t}\,\psi_\wedge$ and $f_2 = \Lambda\sqrt{t}\,\psi_\alpha$ gives $-\widehat{\widetilde{\boldsymbol{L}}}_{t-1} - \Lambda\sqrt{t}\,\nabla\psi_\alpha(\boldsymbol{p}_t) \in c\sqrt{t}\,\partial\psi_\wedge(\boldsymbol{p}_t)$. Then, Lemma C.4(a) gives equation (31).

If $\psi_{\mathcal{A}}(\boldsymbol{p}_t) \neq \psi_{\mathcal{J}}(\boldsymbol{p}_t)$, then θ_t is determined. Otherwise, projecting equation (31) onto $\mathcal{V}$ gives an affine equation in $\theta_t \in [0,1]$; we take its unique solution if there is one, and $\theta_t = 0$ if every $\theta_t \in [0,1]$ solves it. The solution is a ratio of continuous functions of $(\boldsymbol{p}_t, \widehat{\widetilde{\boldsymbol{L}}}_{t-1})$ on the measurable set where it is unique, so in both cases θ_t is $\mathcal{F}_{t-1}$-measurable. □

The next two lemmas show that $\psi_\wedge$ has square-root curvature in every single mediator coordinate, whichever of $\psi_{\mathcal{A}}$ and $\psi_{\mathcal{J}}$ is active. The first collects elementary inequalities for the rows of $\boldsymbol{M}$.

Lemma C.6 (Row moments). *For every $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$ and $z \in \mathcal{Z}$, with $y = (\boldsymbol{M}\boldsymbol{u})_z$,*

$$\sum_{a\in\mathcal{A}} u_a^{3/2} M_{z,a}^2 \le \min\left\{ y^{3/2},\ y\sum_{a\in\mathcal{A}} \sqrt{u_a} M_{z,a} \right\} \quad \text{and} \quad \sum_{a\in\mathcal{A}} u_a^{2-\alpha} M_{z,a}^2 \le y^{2-\alpha}\left(\sum_{a\in\mathcal{A}} M_{z,a}\right)^{\alpha}. \tag{32}$$

Proof. We use $M_{z,a} \le 1$ and $u_a M_{z,a} \le y$. For the first bound, $u_a^{3/2} M_{z,a}^2 \le (u_a M_{z,a})^{3/2} \le u_a M_{z,a}\sqrt{y}$, which sums to $y^{3/2}$, and $u_a^{3/2} M_{z,a}^2 = (u_a M_{z,a})\sqrt{u_a} M_{z,a} \le y\sqrt{u_a} M_{z,a}$. For the second bound, by Hölder's inequality with exponents $1/(1-\alpha)$ and $1/\alpha$,

$$\begin{aligned}\sum_{a\in\mathcal{A}} u_a^{2-\alpha} M_{z,a}^2 &= \sum_{a\in\mathcal{A}} (u_a M_{z,a})^{2-\alpha} M_{z,a}^{\alpha} \le y \sum_{a\in\mathcal{A}} (u_a M_{z,a})^{1-\alpha} M_{z,a}^{\alpha} \\ &\le y^{2-\alpha}\left(\sum_{a\in\mathcal{A}} M_{z,a}\right)^{\alpha}.\end{aligned}$$

□

Lemma C.7 (Curvature in one mediator coordinate). *For every $z \in \mathcal{Z}$, the function $\psi_\wedge + \sqrt{(\boldsymbol{M}\cdot)_z}$ is convex on $\Delta^{|\mathcal{A}|-1}$. Consequently, for every $\boldsymbol{p} \in \operatorname{relint}\Delta^{|\mathcal{A}|-1}$, $\boldsymbol{\xi} \in \partial\psi_\wedge(\boldsymbol{p})$, and $\boldsymbol{u} \in \Delta^{|\mathcal{A}|-1}$,*

$$B_{\psi_\wedge}^{\boldsymbol{\xi}}(\boldsymbol{u};\boldsymbol{p}) \ge \frac{\left(\sqrt{(\boldsymbol{M}\boldsymbol{u})_z} - \sqrt{(\boldsymbol{M}\boldsymbol{p})_z}\right)^2}{2\sqrt{(\boldsymbol{M}\boldsymbol{p})_z}}. \tag{33}$$

Proof. Let $\boldsymbol{q} = \boldsymbol{M}\boldsymbol{p}$. For $\boldsymbol{w} \in \mathbb{R}^{|\mathcal{A}|}$, the Cauchy–Schwarz inequality and Lemma C.6 give $(\sum_{a\in\mathcal{A}} M_{z,a} w_a)^2 \le (\sum_{a\in\mathcal{A}} p_a^{-3/2} w_a^2)(\sum_{a\in\mathcal{A}} p_a^{3/2} M_{z,a}^2) \le q_z^{3/2} \sum_{a\in\mathcal{A}} p_a^{-3/2} w_a^2$. Hence,

$$\boldsymbol{w}^\top \nabla^2 \left[\psi_{\mathcal{A}}(\boldsymbol{p}) + \sqrt{q_z}\right] \boldsymbol{w} = \frac{1}{4}\left(\sum_{a\in\mathcal{A}} p_a^{-3/2} w_a^2 - \frac{\left(\sum_{a\in\mathcal{A}} M_{z,a} w_a\right)^2}{q_z^{3/2}}\right) \ge 0.$$

So, $\psi_{\mathcal{A}} + \sqrt{(\boldsymbol{M}\cdot)_z}$ is convex on $\operatorname{relint}\Delta^{|\mathcal{A}|-1}$. The function $\psi_{\mathcal{J}} + \sqrt{(\boldsymbol{M}\cdot)_z}$ is the sum of $-\sqrt{(\boldsymbol{M}\cdot)_{z'}}$ over $z' \ne z$ and an affine function, hence convex. Their pointwise maximum $\psi_\wedge + \sqrt{(\boldsymbol{M}\cdot)_z}$ is convex on the relative interior and, by continuity, on $\Delta^{|\mathcal{A}|-1}$. Lemma C.3 with $f_1 = \psi_\wedge + \sqrt{(\boldsymbol{M}\cdot)_z}$ and $f_2 = -\sqrt{(\boldsymbol{M}\cdot)_z}$ gives $B_{\psi_\wedge}^{\boldsymbol{\xi}}(\boldsymbol{u};\boldsymbol{p}) \ge B_{f_2}(\boldsymbol{u};\boldsymbol{p})$, and a direct computation of B_{f_2} gives equation (33). □

C.4 One-dimensional and path estimates

This section reduces the stability term s_t of equation (28) to one-dimensional problems. Recall from equation (24) that $\widehat{\boldsymbol{\ell}}_t = \bar{\boldsymbol{B}}_t + \boldsymbol{x}_t$, where the baseline $\bar{\boldsymbol{B}}_t$ is predictable with coordinates in $[0, 1/2]$ and the residual $\boldsymbol{x}_t = (\varepsilon_t/q_t(Z_t))\boldsymbol{M}_{Z_t,\cdot}^\top$ is supported on the row of the sampled mediator. By Lemma C.1(c) and (d), the stability of each part is at most a sum of one-dimensional suprema $\sup_{y\ge 0}\{v(x-y) - \kappa B(y;x)\}$, where B is the Bregman divergence of $-\sqrt{\cdot}$ (curvature of $\psi_{\mathcal{A}}$ and $\psi_{\mathcal{Z}}$) or of $\phi_\alpha(x) = -x^\alpha/[\alpha(1-\alpha)]$ (curvature of ψ_α). Lemma C.8 evaluates these suprema. Moreover, negative scores need care: for the square root, the supremum equation (34) is finite only if $\kappa + 2v\sqrt{x} > 0$. Corollary C.9 shows that the cutoff $16/t$ in the baseline guarantees this, because the centered loss ε_t is negative only if $q_t(Z_t) \ge 16/t$, and then $|\varepsilon_t| \le 1/2$ by equation (23). Lemmas C.10 and C.11 treat the curvature of ψ_α, which must be evaluated away from $\boldsymbol{p}_t$ as well.

Lemma C.8 (Scalar curvature bounds). *Let $\phi_\alpha(x) = -x^\alpha/[\alpha(1-\alpha)]$ with $\alpha \in (0,1)$, and $Q_\alpha = 2^{2-\alpha}$.*

(a) For $x > 0$, $\kappa > 0$, and $\kappa + 2v\sqrt{x} > 0$,

$$\sup_{y\ge 0}\left\{ v(x-y) - \kappa\frac{\left(\sqrt{y} - \sqrt{x}\right)^2}{2\sqrt{x}} \right\} = \frac{2v^2 x^{3/2}}{\kappa + 2v\sqrt{x}}. \tag{34}$$

(b) Let $0 < x \le 1$ and $\kappa > 0$. If $v \ge 0$, then $\sup_{y\ge 0}\{v(x-y)-\kappa B_{\phi_\alpha}(y;x)\} \le x^{2-\alpha}v^2/(2\kappa)$. If $v \ge -\bar{v}$ for some $\bar{v} \ge 0$ with $\kappa \ge 2\bar{v}$, then the same supremum is at most $Q_\alpha x^{2-\alpha}v^2/(2\kappa)$.

Proof. (a) With $y = xr^2$, the objective becomes $vx(1-r^2) - \frac{1}{2}\kappa\sqrt{x}(1-r)^2$, a strictly concave quadratic in $r \ge 0$ maximized at $r = \kappa/(\kappa + 2v\sqrt{x})$. Substituting this value gives equation (34).

(b) Replace v by sv with $s \in [0,1]$ and let $\Phi(s)$ be the supremum. The objective is concave in y, and its stationary point y_s satisfies

$$y_s^{\alpha-1} = x^{\alpha-1} + \frac{(1-\alpha)\, sv}{\kappa}\,.$$

If $v \ge 0$, then $y_s \le x$. If $v \in [-\bar{v}, 0)$ and $\kappa \ge 2\bar{v}$, then, since $x^{\alpha-1} \ge 1$, the right side is at least $x^{\alpha-1}(1-(1-\alpha)/2) \ge 2^{\alpha-1}x^{\alpha-1} > 0$, where the first inequality bounds the convex function $r \mapsto 2^{-r}$ on $[0,1]$ by its chord; hence $y_s \le 2x$. In both cases, y_s is a finite positive stationary point of a concave function and thus the maximizer. The envelope theorem gives $\Phi(0) = \Phi'(0) = 0$ and $\Phi''(s) = v^2 y_s^{2-\alpha}/\kappa$, so $\Phi(1) = \int_0^1 (1-s)\,\Phi''(s)\,ds \le \frac{1}{2}\max_{s\in[0,1]}\, y_s^{2-\alpha}v^2/\kappa$. The two bounds on y_s give the two claims. □

The residual $\boldsymbol{x}_t$ enters through the following instance of part (a), with $x = q_t(z)$ and score $\gamma = \varepsilon_t/q_t(z)$. The cutoff $16/t$ keeps negative scores small relative to the curvature $d\sqrt{t}$, so the denominator in equation (34) stays positive.

Corollary C.9 (Signed square-root bound)**.** *For $d > 1/4$, let $\nu_d = \max\{2/d,\ 1/(2(d-1/4))\}$. Let $t \ge 1$, $x \in (0,1]$, and $\varepsilon \in \mathbb{R}$ with either $\varepsilon \in [0,1]$, or $|\varepsilon| \le 1/2$ and $x \ge 16/t$. Then, with $\gamma = \varepsilon/x$,*

$$\sup_{y\ge 0}\left\{\gamma(x-y) - d\sqrt{t}\,\frac{\left(\sqrt{y}-\sqrt{x}\right)^2}{2\sqrt{x}}\right\} \le \frac{\nu_d}{\sqrt{tx}}\,. \tag{35}$$

Proof. If $\varepsilon \ge 0$, then $d\sqrt{t} + 2\gamma\sqrt{x} \ge d\sqrt{t}$ and $2\gamma^2 x^{3/2} = 2\varepsilon^2/\sqrt{x} \le 2/\sqrt{x}$. If $\varepsilon < 0$, then $2\gamma\sqrt{x} = 2\varepsilon/\sqrt{x} \ge -1/\sqrt{x} \ge -\sqrt{t}/4$, so $d\sqrt{t} + 2\gamma\sqrt{x} \ge (d-1/4)\sqrt{t} > 0$, and $2\gamma^2 x^{3/2} \le 1/(2\sqrt{x})$. In both cases Lemma C.8(a) applies and gives equation (35). □

For the curvature of ψ_α, the divergence at $\boldsymbol{p}_t$ alone does not control the stability function, because a negative residual can move the minimizer to points where ψ_α is flatter. The next lemma adds the loss $\boldsymbol{v}$ gradually, as $s\boldsymbol{v}$ for $s \in [0,1]$, and follows the resulting path of minimizers $\boldsymbol{u}_s$; the stability function is then an integral along this path, controlled by the curvature of ψ_α at $\boldsymbol{u}_s$.

Lemma C.10 (The response of a minimizing path)**.** *Let $\Psi = f_0 + \kappa\psi_\alpha$, where f_0 is continuous and convex on $\Delta^{|\mathcal{A}|-1}$ and $\kappa > 0$. Let $\boldsymbol{p} \in \operatorname{relint}\Delta^{|\mathcal{A}|-1}$, $\boldsymbol{\xi} \in \partial\Psi(\boldsymbol{p})$, and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{A}|}$, and for $s \in [0,1]$ let*

$$\boldsymbol{u}_s = \operatorname*{arg\,min}_{\boldsymbol{u}\in\Delta^{|\mathcal{A}|-1}}\ \langle s\boldsymbol{v} - \boldsymbol{\xi}, \boldsymbol{u}\rangle + \Psi(\boldsymbol{u})\,.$$

If $\boldsymbol{u}_s \in \operatorname{relint}\Delta^{|\mathcal{A}|-1}$ for every $s \in [0,1]$, then $\boldsymbol{u}_0 = \boldsymbol{p}$, the path $s \mapsto \boldsymbol{u}_s$ is Lipschitz, and its derivative $\dot{\boldsymbol{u}}_s$, which exists for almost every s, satisfies

$$\operatorname{stab}^{\boldsymbol{\xi}}_{\Psi,\boldsymbol{p}}(\boldsymbol{v}) = \int_0^1 (1-s)\left[-\langle \boldsymbol{v}, \dot{\boldsymbol{u}}_s\rangle\right] ds \quad \textit{and} \quad 0 \le -\langle \boldsymbol{v}, \dot{\boldsymbol{u}}_s\rangle \le \frac{1}{\kappa}\sum_{a\in\mathcal{A}} u_s(a)^{2-\alpha}\, v_a^2\,. \tag{36}$$

Proof. On $\operatorname{relint}\Delta^{|\mathcal{A}|-1}$, $\nabla^2\psi_\alpha(\boldsymbol{u}) = \operatorname{diag}(u_a^{\alpha-2}) \succeq \mathbf{I}$, and by continuity ψ_α is 1-strongly convex on $\Delta^{|\mathcal{A}|-1}$ with respect to $\|\cdot\|_2$. Adding the quadratic-growth inequalities of the minimizers at s and r gives

$$\kappa\|\boldsymbol{u}_s - \boldsymbol{u}_r\|_2^2 \le -(s-r)\langle \boldsymbol{v}, \boldsymbol{u}_s - \boldsymbol{u}_r\rangle \le |s-r|\,\|\boldsymbol{v}\|_2\|\boldsymbol{u}_s - \boldsymbol{u}_r\|_2,$$

so the path is Lipschitz. Since $\boldsymbol{\xi} \in \partial\Psi(\boldsymbol{p})$, the point $\boldsymbol{p}$ minimizes the objective at $s = 0$, and uniqueness gives $\boldsymbol{u}_0 = \boldsymbol{p}$. Let $F(s)$ be the minimum value at s. Evaluating the objectives at $\boldsymbol{u}_s$ and $\boldsymbol{u}_{s+\tau}$ gives $\tau\langle \boldsymbol{v}, \boldsymbol{u}_{s+\tau}\rangle \le F(s+\tau) - F(s) \le \tau\langle \boldsymbol{v}, \boldsymbol{u}_s\rangle$, so taking the limit for $\tau \to 0$ gives

$F'(s) = \langle \boldsymbol{v}, \boldsymbol{u}_s \rangle$. Expanding the Bregman divergence shows $\mathrm{stab}^{\boldsymbol{\xi}}_{\Psi,\boldsymbol{p}}(\boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{p} \rangle + F(0) - F(1)$, and hence

$$\mathrm{stab}^{\boldsymbol{\xi}}_{\Psi,\boldsymbol{p}}(\boldsymbol{v}) = -\int_0^1 \langle \boldsymbol{v}, \boldsymbol{u}_s - \boldsymbol{u}_0 \rangle \, ds = -\int_0^1 \int_0^s \langle \boldsymbol{v}, \dot{\boldsymbol{u}}_r \rangle \, dr \, ds = \int_0^1 (1-r) \left[- \langle \boldsymbol{v}, \dot{\boldsymbol{u}}_r \rangle \right] dr \,.$$

By Lemma C.3 applied to the objective at s, the vector $\boldsymbol{\xi} - s\boldsymbol{v} - \kappa \nabla \psi_\alpha(\boldsymbol{u}_s)$ is a subgradient of f_0 at $\boldsymbol{u}_s$. Monotonicity of the subgradients of f_0 at $\boldsymbol{u}_s$ and $\boldsymbol{u}_{s+\tau}$ gives

$$-\tau \langle \boldsymbol{v}, \boldsymbol{u}_{s+\tau} - \boldsymbol{u}_s \rangle \geq \kappa \langle \nabla \psi_\alpha(\boldsymbol{u}_{s+\tau}) - \nabla \psi_\alpha(\boldsymbol{u}_s), \boldsymbol{u}_{s+\tau} - \boldsymbol{u}_s \rangle \,.$$

At a point of differentiability, dividing by τ^2 and letting $\tau \to 0$ gives $-\langle \boldsymbol{v}, \dot{\boldsymbol{u}}_s \rangle \geq \kappa \sum_{a \in \mathcal{A}} (u_s(a))^{\alpha-2} (\dot{u}_s(a))^2 \geq 0$. By the weighted Cauchy–Schwarz inequality,

$$\begin{aligned} \langle \boldsymbol{v}, \dot{\boldsymbol{u}}_s \rangle^2 &\leq \left(\sum_{a \in \mathcal{A}} (u_s(a))^{2-\alpha} v_a^2 \right) \left(\sum_{a \in \mathcal{A}} (u_s(a))^{\alpha-2} (\dot{u}_s(a))^2 \right) \\ &\leq \left(\sum_{a \in \mathcal{A}} (u_s(a))^{2-\alpha} v_a^2 \right) \frac{-\langle \boldsymbol{v}, \dot{\boldsymbol{u}}_s \rangle}{\kappa}, \end{aligned}$$

which gives the upper bound in equation (36) after division by $-\langle \boldsymbol{v}, \dot{\boldsymbol{u}}_s \rangle$ when it is positive; the case in which it vanishes is immediate. □

To use this lemma for the residual $\boldsymbol{x}_t = \gamma \boldsymbol{M}^\top_{z,\cdot}$, with $\gamma = \varepsilon_t / q_t(z)$ and $z = Z_t$, we need to know how far the path moves. The bound of Lemma C.10 with $\boldsymbol{v} = \gamma \boldsymbol{M}^\top_{z,\cdot}$ depends on the path only through $y_s = (\boldsymbol{M}\boldsymbol{u}_s)_z$, the probability that $\boldsymbol{u}_s$ assigns to the sampled mediator (Lemma C.6). The next lemma shows that y_s cannot increase much along the path.

Lemma C.11 (Signed residual paths)**.** *Let $\Psi = \sigma\psi_\wedge + \kappa\psi_\alpha$ with $\sigma, \kappa > 0$, let $\boldsymbol{p} \in \mathrm{relint}\, \Delta^{|\mathcal{A}|-1}$, $\boldsymbol{q} = \boldsymbol{M}\boldsymbol{p}$, and $\boldsymbol{\xi} \in \partial\Psi(\boldsymbol{p})$, and fix $z \in \mathcal{Z}$ and $\gamma \in \mathbb{R}$. Let $(\boldsymbol{u}_s)_{s \in [0,1]}$ be the path of Lemma C.10 with $\boldsymbol{v} = \gamma \boldsymbol{M}^\top_{z,\cdot}$, and $y_s = (\boldsymbol{M}\boldsymbol{u}_s)_z$.*

(a) If $\gamma \geq 0$, then $y_s \leq q_z$.

(b) If $\gamma < 0$ and $1 + 2\gamma\sqrt{q_z}/\sigma > 0$, then

$$y_s \leq \frac{q_z}{\left(1 + 2s\gamma\sqrt{q_z}/\sigma\right)^2} \,. \tag{37}$$

(c) If $\sigma = c\sqrt{t}$ and $\gamma = \varepsilon/q_z$, where $\varepsilon \in [-1/2, 0)$ and $q_z \geq 16/t$, then $y_s \leq q_z \left(1 - s/(4c)\right)^{-2}$.

Proof. By Lemma C.4, the whole path lies in $\mathrm{relint}\, \Delta^{|\mathcal{A}|-1}$. Adding the minimizing inequalities at s and 0 gives $s\gamma(y_s - q_z) \leq 0$, which proves part (a) and shows that $y_s \geq q_z$ when $\gamma < 0$. By Lemma C.7, $\Psi = f_z - \sigma\sqrt{(\boldsymbol{M}\cdot)_z}$ with $f_z = \sigma(\psi_\wedge + \sqrt{(\boldsymbol{M}\cdot)_z}) + \kappa\psi_\alpha$ convex. Lemma C.3 shows that $\boldsymbol{\xi} - s\boldsymbol{v} + \frac{\sigma}{2} y_s^{-1/2} \boldsymbol{M}^\top_{z,\cdot}$ and $\boldsymbol{\xi} + \frac{\sigma}{2} q_z^{-1/2} \boldsymbol{M}^\top_{z,\cdot}$ are subgradients of f_z at $\boldsymbol{u}_s$ and $\boldsymbol{p}$, respectively. Monotonicity of ∂f_z gives

$$\left(s\gamma - \frac{\sigma}{2\sqrt{y_s}} + \frac{\sigma}{2\sqrt{q_z}} \right) (y_s - q_z) \leq 0 \,.$$

If $y_s > q_z$, dividing by $y_s - q_z$ gives $1/\sqrt{y_s} \geq 1/\sqrt{q_z} + 2s\gamma/\sigma$, whose right side is positive for $s \in [0,1]$ by hypothesis; inverting gives equation (37), which also holds if $y_s = q_z$. For part (c), $2s\gamma\sqrt{q_z}/\sigma = 2s\varepsilon/(c\sqrt{tq_z}) \geq -s/(4c)$, because $|\varepsilon| \leq 1/2$ and $\sqrt{tq_z} \geq 4$; since $c > 1/4$, part (b) applies. □

C.5 A UNIFORM STABILITY BOUND AND THE ADVERSARIAL RATE

By equation (28), the adversarial rate follows once $\mathbb{E}_{t-1}[s_t] \le G/\sqrt{t}$ for a constant G that does not depend on the losses. We bound s_t through the split $\widehat{\widetilde{\boldsymbol{\ell}}}_t = \bar{\boldsymbol{B}}_t + \boldsymbol{x}_t$, allocating the curvature of ψ_t between the baseline and the residual. There are two natural allocations, which give Lemmas C.13 and C.12:

- *Action allocation*: half of $\Lambda\sqrt{t}\,\psi_\alpha$ controls the baseline, and the rest of ψ_t controls the residual. The resulting bound scales with $|\mathcal{A}|^\alpha/\Lambda$.
- *Mediator allocation*: ψ_α is discarded, and $c\sqrt{t}\,\psi_\wedge$ is split between the baseline and the residual. The resulting bound scales with $\sqrt{|\mathcal{Z}|}$.

The constant G in equation (39) is the smaller of the two bounds, and the calibration equation (45) of λ_α turns it into a dependence on $\min\{|\mathcal{A}|, |\mathcal{Z}|\}$.

Splitting the stability term. Write $\psi_t = \psi_t^{\mathrm{b}} + \psi_t^{\mathrm{x}}$, where

$$\psi_t^{\mathrm{b}} = \frac{\Lambda\sqrt{t}}{2}\psi_\alpha \qquad \text{and } \psi_t^{\mathrm{x}} = c\sqrt{t}\,\psi_\wedge + \frac{\Lambda\sqrt{t}}{2}\psi_\alpha\,.$$

With $\boldsymbol{\xi}_t^\wedge$ from Corollary C.5, split the subgradient in equation (31) accordingly:

$$\boldsymbol{\xi}_t^{\mathrm{b}} = \frac{\Lambda\sqrt{t}}{2}\nabla\psi_\alpha\left(\boldsymbol{p}_t\right) \quad \text{and} \quad \boldsymbol{\xi}_t^{\mathrm{x}} = c\sqrt{t}\,\boldsymbol{\xi}_t^\wedge + \boldsymbol{\xi}_t^{\mathrm{b}}\,.$$

For shortness of notation, we write $\mathrm{stab}_t^{\mathrm{b}} = \mathrm{stab}^{\boldsymbol{\xi}_t^{\mathrm{b}}}_{\psi_t^{\mathrm{b}},\boldsymbol{p}_t}$ and $\mathrm{stab}_t^{\mathrm{x}} = \mathrm{stab}^{\boldsymbol{\xi}_t^{\mathrm{x}}}_{\psi_t^{\mathrm{x}},\boldsymbol{p}_t}$. Since $\boldsymbol{\xi}_t^{\mathrm{b}} + \boldsymbol{\xi}_t^{\mathrm{x}} \in -\widehat{\widetilde{\boldsymbol{L}}}_{t-1} + \mathbb{R}\mathbf{1}$, Lemmas C.2 and C.1(b) applied to equation (24) give

$$s_t \le \mathrm{stab}_t^{\mathrm{b}}(\bar{\boldsymbol{B}}_t) + \mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t), \tag{38}$$

where s_t is the stability term of Lemma C.2 with $\boldsymbol{v}_t = \widehat{\widetilde{\boldsymbol{\ell}}}_t$.

Constants. Recall ν_d from Corollary C.9, and define

$$\begin{aligned} \bar{c} = \frac{c+1/4}{2}, \qquad I_{c,\alpha} = 2\int_0^1 \frac{1-s}{\left(1-s/\left(4c\right)\right)^{4-2\alpha}}\,ds, \qquad \omega_{c,\alpha} = \max\left\{1, I_{c,\alpha}/4\right\}, \\ \text{and} \quad G = \min\left\{\frac{\omega_{c,\alpha}|\mathcal{A}|^\alpha + 1/4}{\Lambda}, \frac{1}{2\left(c-\bar{c}\right)} + \nu_{\bar{c}}\sqrt{|\mathcal{Z}|}\right\}. \end{aligned} \tag{39}$$

Since $c > 1/4$, we have $1/4 < \bar{c} < c$, and the substitution $y = 1 - s/\left(4c\right)$ gives the finite value

$$I_{c,\alpha} = \frac{16c^2}{\left(1-\alpha\right)\left(3-2\alpha\right)}\left[\left(1-\frac{1}{4c}\right)^{-2(1-\alpha)} - 1 - \frac{1-\alpha}{2c}\right].$$

We start with the mediator allocation, whose residual bound is also used in Lemma C.16.

Lemma C.12 (Mediator allocation)**.** *Fix $t \ge 1$, condition on $\mathcal{F}_{t-1}$, and write $\boldsymbol{p} = \boldsymbol{p}_t$ and $\boldsymbol{q} = \boldsymbol{q}_t$.*

(a) For every $d \in (1/4, c]$ and every observation $Z_t = z$, $\mathrm{stab}^{d\sqrt{t}\boldsymbol{\xi}_t^\wedge}_{d\sqrt{t}\psi_\wedge,\boldsymbol{p}}(\boldsymbol{x}_t) \le \nu_d/\sqrt{tq_z}$. In particular,

$$\mathbb{E}_{t-1}\left[\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t)\right] \le \frac{\nu_c}{\sqrt{t}}\sum_{z\in\mathcal{Z}}\sqrt{q_t(z)}\,. \tag{40}$$

(b) $\mathbb{E}_{t-1}[s_t] \le [1/(2(c-\bar{c})) + \nu_{\bar{c}}\sqrt{|\mathcal{Z}|}]/\sqrt{t}$.

Proof. Fix an observation $Z_t = z$ and $\ell_t(z)$. Then, ε_t obeys equation (23), and $\boldsymbol{x}_t = \gamma\boldsymbol{M}_{z,\cdot}^\top$ with $\gamma = \varepsilon_t/q_z$.

Part (a). For every $d \in (1/4, c]$, Lemma C.7 gives $d\sqrt{t}\, B_{\psi_\wedge}^{\boldsymbol{\xi}_t^\wedge}(\boldsymbol{u};\boldsymbol{p}) \geq d\sqrt{t}\,(\sqrt{(\boldsymbol{M}\boldsymbol{u})_z} - \sqrt{q_z})^2/\left(2\sqrt{q_z}\right)$. By Lemma C.1(a) and (d), with $\phi_z(y) = -d\sqrt{t}\sqrt{y}$ and $\phi_{z'} = 0$ for $z' \neq z$, and Corollary C.9,

$$\mathrm{stab}_{d\sqrt{t}\psi_\wedge,\boldsymbol{p}}^{d\sqrt{t}\boldsymbol{\xi}_t^\wedge}(\boldsymbol{x}_t) \leq \sup_{y\geq 0}\left\{\gamma\,(q_z - y) - d\sqrt{t}\,\frac{\left(\sqrt{y}-\sqrt{q_z}\right)^2}{2\sqrt{q_z}}\right\} \leq \frac{\nu_d}{\sqrt{tq_z}}\,. \tag{41}$$

For $d = c$, dropping the nonnegative divergence of $\frac{\Lambda\sqrt{t}}{2}\psi_\alpha$ (Lemma C.1(a)) shows that $\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t) \leq \nu_c/\sqrt{tq_z}$. Since $Z_t = z$ with conditional probability q_z, averaging proves equation (40).

Part (b). By equation (31), $B_{\psi_t}^{-\widehat{\bar{\boldsymbol{L}}}_{t-1}} = (c-\bar{c})\sqrt{t}\, B_{\psi_\wedge}^{\boldsymbol{\xi}_t^\wedge} + \bar{c}\sqrt{t}\, B_{\psi_\wedge}^{\boldsymbol{\xi}_t^\wedge} + \Lambda\sqrt{t}\, B_{\psi_\alpha}$. Dropping the last divergence, Lemma C.1(a) and (b) give

$$s_t \leq \mathrm{stab}_{(c-\bar{c})\sqrt{t}\psi_\wedge,\boldsymbol{p}}^{(c-\bar{c})\sqrt{t}\boldsymbol{\xi}_t^\wedge}\left(\bar{\boldsymbol{B}}_t\right) + \mathrm{stab}_{\bar{c}\sqrt{t}\psi_\wedge,\boldsymbol{p}}^{\bar{c}\sqrt{t}\boldsymbol{\xi}_t^\wedge}(\boldsymbol{x}_t)\,.$$

By part (a) with $d = \bar{c}$, the second term has conditional expectation at most $\nu_{\bar{c}}\sum_{z\in\mathcal{Z}}\sqrt{q_z}/\sqrt{t} \leq \nu_{\bar{c}}\sqrt{|\mathcal{Z}|/t}$. For the first term, let $\sigma_{\mathrm{b}} = (c-\bar{c})\sqrt{t}$. Lemma C.4(b) and Lemma C.1(a) and (b), with $\bar{\boldsymbol{B}}_t = \theta_t\bar{\boldsymbol{B}}_t + (1-\theta_t)\,\bar{\boldsymbol{B}}_t$, bound it by

$$\theta_t\,\mathrm{stab}_{\sigma_{\mathrm{b}}\psi_{\mathcal{A}},\boldsymbol{p}}^{\nabla(\sigma_{\mathrm{b}}\psi_{\mathcal{A}})(\boldsymbol{p})}\left(\bar{\boldsymbol{B}}_t\right) + (1-\theta_t)\,\mathrm{stab}_{\sigma_{\mathrm{b}}\psi_{\mathcal{J}},\boldsymbol{p}}^{\nabla(\sigma_{\mathrm{b}}\psi_{\mathcal{J}})(\boldsymbol{p})}\left(\boldsymbol{M}^\top\boldsymbol{B}_t\right)\,.$$

By Lemma C.1(c) and Lemma C.8(a) with nonnegative scores, the first supremum is at most $\frac{2}{\sigma_{\mathrm{b}}}\sum_{a\in\mathcal{A}} p_a^{3/2}(\bar{B}_t(a))^2 \leq \frac{1}{2\sigma_{\mathrm{b}}}$. By Lemma C.1(d) with $\phi_z(y) = -\sigma_{\mathrm{b}}\sqrt{y}$ and Lemma C.8(a), the second is at most $\frac{2}{\sigma_{\mathrm{b}}}\sum_{z\in\mathcal{Z}} q_z^{3/2}(B_t(z))^2 \leq \frac{1}{2\sigma_{\mathrm{b}}}$. Hence, $\mathbb{E}_{t-1}[s_t] \leq [1/(2(c-\bar{c})) + \nu_{\bar{c}}\sqrt{|\mathcal{Z}|}]/\sqrt{t}$. □

Lemma C.13 (Action allocation)**.** *For every $t \geq 1$, $\mathbb{E}_{t-1}[s_t] \leq (\omega_{c,\alpha}|\mathcal{A}|^\alpha + 1/4)/(\Lambda\sqrt{t})$.*

Proof. Condition on $\mathcal{F}_{t-1}$, write $\boldsymbol{p} = \boldsymbol{p}_t$ and $\boldsymbol{q} = \boldsymbol{q}_t$, and fix an observation $Z_t = z$ and $\ell_t(z)$, so that $\boldsymbol{x}_t = \gamma\boldsymbol{M}_{z,\cdot}^\top$ with $\gamma = \varepsilon_t/q_z$. We bound the two terms of equation (38). The function ψ_t^{b} differs from $\sum_{a\in\mathcal{A}}\frac{\Lambda\sqrt{t}}{2}\phi_\alpha(u_a)$ by a constant. Lemma C.1(c) and Lemma C.8(b) with $\kappa = \Lambda\sqrt{t}/2$ and $v = \bar{B}_t(a) \geq 0$ give

$$\mathrm{stab}_t^{\mathrm{b}}\left(\bar{\boldsymbol{B}}_t\right) \leq \frac{1}{\Lambda\sqrt{t}}\sum_{a\in\mathcal{A}} p_a^{2-\alpha}\left(\bar{B}_t(a)\right)^2 \leq \frac{1}{4\Lambda\sqrt{t}}\,. \tag{42}$$

For the residual, apply Lemma C.10 to ψ_t^{x} with $f_0 = c\sqrt{t}\,\psi_\wedge$, $\kappa = \Lambda\sqrt{t}/2$, $\boldsymbol{\xi} = \boldsymbol{\xi}_t^{\mathrm{x}}$, and $\boldsymbol{v} = \boldsymbol{x}_t$; the path is interior by Lemma C.4. With $y_s = (\boldsymbol{M}\boldsymbol{u}_s)_z$, Lemma C.6 gives $\sum_{a\in\mathcal{A}}(u_s(a))^{2-\alpha}(x_t(a))^2 \leq \gamma^2 y_s^{2-\alpha}(\sum_{a\in\mathcal{A}} M_{z,a})^\alpha$, and therefore

$$\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t) \leq \frac{2\varepsilon_t^2}{\Lambda\sqrt{t}\,q_z^2}\left(\sum_{a\in\mathcal{A}} M_{z,a}\right)^\alpha \int_0^1 (1-s)\,y_s^{2-\alpha}\,ds\,.$$

If $\varepsilon_t \geq 0$, then $\varepsilon_t^2 \leq 1$ and $y_s \leq q_z$ by Lemma C.11(a), so the integral is at most $q_z^{2-\alpha}/2$. If $\varepsilon_t < 0$, then $\varepsilon_t^2 \leq 1/4$ and $y_s \leq q_z(1 - s/(4c))^{-2}$ by Lemma C.11(c), so the integral is at most $q_z^{2-\alpha}I_{c,\alpha}/2$ by the definition of $I_{c,\alpha}$. In both cases

$$\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t) \leq \frac{\omega_{c,\alpha}}{\Lambda\sqrt{t}}q_z^{-\alpha}\left(\sum_{a\in\mathcal{A}} M_{z,a}\right)^\alpha\,.$$

Averaging over Z_t and applying Hölder's inequality with $\sum_{z\in\mathcal{Z}} q_z = 1$ and $\sum_{z\in\mathcal{Z}}\sum_{a\in\mathcal{A}} M_{z,a} = |\mathcal{A}|$,

$$\mathbb{E}_{t-1}\left[\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t)\right] \leq \frac{\omega_{c,\alpha}}{\Lambda\sqrt{t}}\sum_{z\in\mathcal{Z}} q_z^{1-\alpha}\left(\sum_{a\in\mathcal{A}} M_{z,a}\right)^\alpha \leq \frac{\omega_{c,\alpha}|\mathcal{A}|^\alpha}{\Lambda\sqrt{t}}\,.$$

Together with equation (38) and equation (42), this proves the lemma. □

Since G is the smaller of the two bounds, Lemmas C.12(b) and C.13 give, for every $t \geq 1$,

$$\mathbb{E}_{t-1}[s_t] \leq \frac{G}{\sqrt{t}}\,. \tag{43}$$

Inserting equation (43) and the penalty bound $D\sqrt{T}$ into equation (28) gives the adversarial rate.

Theorem C.14 (Adversarial regret)**.** *Let $c > 1/4$, $\rho > 0$, $\lambda_\alpha > 0$, and $\alpha \in (1/2, 1)$, and recall D from equation (25), $\nu_{\bar{c}}$ from Corollary C.9, and $\bar{c}$, $\omega_{c,\alpha}$, and G from equation (39). For every loss sequence $(\boldsymbol{\ell}_t)$ satisfying equation (18) and every integer $T \geq 1$,*

$$\mathrm{Reg}_T \leq (D + 2G)\sqrt{T}\,. \tag{44}$$

If, in addition,

$$\lambda_\alpha = \min\left\{\sqrt{\frac{|\mathcal{A}|^\alpha}{D_\alpha}},\ \frac{\sqrt{|\mathcal{Z}|} - 1}{D_\alpha}\right\}, \tag{45}$$

where $|\mathcal{Z}|$ counts reachable mediators, then

$$\mathrm{Reg}_T \leq C_{\mathrm{adv}}\sqrt{\min\{|\mathcal{A}|, |\mathcal{Z}|\}\,T}, \quad \textit{where}$$
$$C_{\mathrm{adv}} = c + \frac{\rho + \max\left\{\frac{1}{(c-\bar{c})\left(\sqrt{2}-1\right)} + 2\left(2+\sqrt{2}\right)\nu_{\bar{c}},\ \frac{2(\omega_{c,\alpha}+1/4)}{\rho}\right\}}{\sqrt{\alpha\,(1-\alpha)}}\,. \tag{46}$$

Proof. In equation (28), $0 \leq -\psi \leq D$ bounds the penalty part by $D + \sum_{t=2}^{T}(\sqrt{t} - \sqrt{t-1})D = D\sqrt{T}$, and equation (43) with $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T}$ bounds the stability part by $2G\sqrt{T}$. This proves equation (44).

For equation (46), let $v = \sqrt{|\mathcal{A}|^\alpha D_\alpha}$, $D_{\mathcal{Z}} = \sqrt{|\mathcal{Z}|} - 1$, and $w = \min\{v, D_{\mathcal{Z}}\} = \lambda_\alpha D_\alpha$, so that $D = c\min\{\sqrt{|\mathcal{A}|} - 1, D_{\mathcal{Z}}\} + \rho w$. If $\lambda_\alpha = D_{\mathcal{Z}}/D_\alpha$, we use the second branch of G. For $|\mathcal{Z}| \geq 2$ we have $\sqrt{|\mathcal{Z}|} \leq (2+\sqrt{2})D_{\mathcal{Z}}$ and $1 \leq D_{\mathcal{Z}}/(\sqrt{2}-1)$, so

$$2G \leq \left[\frac{1}{(c-\bar{c})\left(\sqrt{2}-1\right)} + 2\left(2+\sqrt{2}\right)\nu_{\bar{c}}\right] w\,.$$

If $\lambda_\alpha = \sqrt{|\mathcal{A}|^\alpha / D_\alpha}$, we use the first branch of G; since $\Lambda = \rho\lambda_\alpha$ and $|\mathcal{A}|^\alpha \geq 1$,

$$2G \leq \frac{2\,(\omega_{c,\alpha}|\mathcal{A}|^\alpha + 1/4)}{\rho}\sqrt{\frac{D_\alpha}{|\mathcal{A}|^\alpha}} \leq \frac{2\,(\omega_{c,\alpha} + 1/4)}{\rho} w\,.$$

Finally, $v^2 = (|\mathcal{A}| - |\mathcal{A}|^\alpha)/[\alpha(1-\alpha)] \leq |\mathcal{A}|/[\alpha(1-\alpha)]$ and $D_{\mathcal{Z}} \leq \sqrt{|\mathcal{Z}|} \leq \sqrt{|\mathcal{Z}|/[\alpha(1-\alpha)]}$ give $w \leq \sqrt{\min\{|\mathcal{A}|, |\mathcal{Z}|\}/[\alpha(1-\alpha)]}$, and $\min\{\sqrt{|\mathcal{A}|}-1, D_{\mathcal{Z}}\} \leq \sqrt{\min\{|\mathcal{A}|, |\mathcal{Z}|\}}$. Substituting into equation (44) proves equation (46). □

C.6 Local bounds near a reference action

The uniform bound equation (43) holds for every $\boldsymbol{p}_t$ and therefore gives at best a $\sqrt{T}$ rate. A logarithmic rate requires bounds on both parts of equation (28) that vanish as $\boldsymbol{p}_t$ concentrates on a fixed action, so that they can be charged to $\Gamma(\boldsymbol{p}_t)$, as in the self-bounding analysis of Zimmert & Seldin (2021). We therefore fix a reference action $a^\star$, which becomes the unique zero of the gap vector in equation (7) in Sections C.7 and C.8, and measure the distance of $\boldsymbol{p}$ from $\boldsymbol{e}_{a^\star}$ by the probability $r(\boldsymbol{p})$ of the other actions and by two square-root widths. The penalty part is bounded by Lemma C.15. The stability part is bounded by Lemma C.16, for the baseline and for mediators that $a^\star$ produces, and by Lemma C.17, for mediators that $a^\star$ never produces. Proposition C.18 collects these bounds.

Fix a deterministic reference action $a^\star \in \mathcal{A}$. Let

$$\begin{gathered} \boldsymbol{q}^\star = \boldsymbol{M}_{\cdot,a^\star}, \qquad \operatorname{supp}(\boldsymbol{q}^\star) = \{z \in \mathcal{Z} : q_z^\star > 0\}, \qquad \mathcal{U} = \mathcal{Z} \setminus \operatorname{supp}(\boldsymbol{q}^\star), \\ r(\boldsymbol{p}) = 1 - p_{a^\star}, \qquad W_{\mathcal{A}}(\boldsymbol{p}) = \sum_{a \in \mathcal{A}\setminus\{a^\star\}} \sqrt{p_a}, \\ W_{\mathcal{U}}(\boldsymbol{p}) = \sum_{z \in \mathcal{U}} \sqrt{(\boldsymbol{M}\boldsymbol{p})_z}, \quad \text{and} \quad W_\wedge(\boldsymbol{p}) = \min\left\{W_{\mathcal{A}}(\boldsymbol{p}), W_{\mathcal{U}}(\boldsymbol{p})\right\}, \end{gathered} \tag{47}$$

where empty sums are defined to be zero. A mediator in $\mathcal{U}$ is never produced by $a^\star$, so it is observed only through the other actions; $r(\boldsymbol{p})$ is the probability of these actions, and $W_{\mathcal{A}}$ and $W_{\mathcal{U}}$ are the Tsallis widths of the exploration in the action space and in $\mathcal{U}$. The penalty part of equation (28) involves $-\psi_\wedge(\boldsymbol{p}_t)$ and $-\psi_\alpha(\boldsymbol{p}_t)$. The next lemma bounds $-\psi_\wedge$ near $\boldsymbol{e}_{a^\star}$ by the widths plus a multiple of r; the bound on $-\psi_\alpha$ is part of the proof of Proposition C.18. The lemma also gives a lower bound on $-\psi_{\mathcal{J}}$, which Lemma C.17 uses when $\psi_{\mathcal{J}}$ is the active part of $\psi_\wedge$. It uses the finite nonnegative constants

$$\beta_+ = \max\left\{0, \max_{a \neq a^\star}\left[\psi_{\mathcal{Z}}(\boldsymbol{M}_{\cdot,a}) - \psi_{\mathcal{Z}}(\boldsymbol{q}^\star) + \sum_{z \in \operatorname{supp}(\boldsymbol{q}^\star)} \frac{M_{z,a} - q_z^\star}{2\sqrt{q_z^\star}}\right]\right\} \quad \text{and}$$

$$\beta_- = \max_{a \neq a^\star} \sum_{z \in \mathcal{U}} \sqrt{M_{z,a}}\,.$$

Lemma C.15 (Calibration at the reference action)**.** *For every $\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}$,*

$$W_{\mathcal{U}}(\boldsymbol{p}) - \beta_- r(\boldsymbol{p}) \le -\psi_{\mathcal{J}}(\boldsymbol{p}) \le W_{\mathcal{U}}(\boldsymbol{p}) + \beta_+ r(\boldsymbol{p}) \quad \textit{and} \quad -\psi_\wedge(\boldsymbol{p}) \le W_\wedge(\boldsymbol{p}) + \beta_+ r(\boldsymbol{p})\,. \tag{48}$$

If $\psi_{\mathcal{J}}(\boldsymbol{p}) \ge \psi_{\mathcal{A}}(\boldsymbol{p})$, then $W_{\mathcal{U}}(\boldsymbol{p}) \le W_\wedge(\boldsymbol{p}) + \beta_- r(\boldsymbol{p})$.

Proof. Let $\boldsymbol{q} = \boldsymbol{M}\boldsymbol{p}$. The tangent inequality $\sqrt{q_z} \le \sqrt{q_z^\star} + (q_z - q_z^\star)/(2\sqrt{q_z^\star})$ for $z \in \operatorname{supp}(\boldsymbol{q}^\star)$ gives

$$-\psi_{\mathcal{Z}}(\boldsymbol{q}) \le -\psi_{\mathcal{Z}}(\boldsymbol{q}^\star) + W_{\mathcal{U}}(\boldsymbol{p}) + \sum_{z \in \operatorname{supp}(\boldsymbol{q}^\star)} \frac{q_z - q_z^\star}{2\sqrt{q_z^\star}}\,.$$

Substitute $\boldsymbol{q} - \boldsymbol{q}^\star = \sum_{a \in \mathcal{A}\setminus\{a^\star\}} p_a(\boldsymbol{M}_{\cdot,a} - \boldsymbol{q}^\star)$ and add $\sum_{a \in \mathcal{A}} p_a \psi_{\mathcal{Z}}(\boldsymbol{M}_{\cdot,a})$ to both sides. Writing $-\psi_{\mathcal{Z}}(\boldsymbol{q}^\star) + p_{a^\star}\psi_{\mathcal{Z}}(\boldsymbol{q}^\star) = -\sum_{a \in \mathcal{A}\setminus\{a^\star\}} p_a \psi_{\mathcal{Z}}(\boldsymbol{q}^\star)$, the right side becomes $W_{\mathcal{U}}(\boldsymbol{p})$ plus $\sum_{a \in \mathcal{A}\setminus\{a^\star\}} p_a$ times the bracket in β_+, which proves the upper bound on $-\psi_{\mathcal{J}}$. For the lower bound, write

$$-\psi_{\mathcal{J}}(\boldsymbol{p}) = \sum_{z \in \mathcal{Z}} \left(\sqrt{q_z} - \sum_{a \in \mathcal{A}} p_a \sqrt{M_{z,a}}\right);$$

each summand is nonnegative by concavity of the square root, and keeping only $z \in \mathcal{U}$, where $M_{z,a^\star} = 0$, gives $-\psi_{\mathcal{J}}(\boldsymbol{p}) \ge W_{\mathcal{U}}(\boldsymbol{p}) - \beta_- r(\boldsymbol{p})$. Next, $-\psi_{\mathcal{A}}(\boldsymbol{p}) = W_{\mathcal{A}}(\boldsymbol{p}) + \sqrt{p_{a^\star}} - 1 \le W_{\mathcal{A}}(\boldsymbol{p})$, so

$$\begin{aligned} -\psi_\wedge(\boldsymbol{p}) &= \min\left\{-\psi_{\mathcal{A}}(\boldsymbol{p}), -\psi_{\mathcal{J}}(\boldsymbol{p})\right\} \le \min\left\{W_{\mathcal{A}}(\boldsymbol{p}), W_{\mathcal{U}}(\boldsymbol{p}) + \beta_+ r(\boldsymbol{p})\right\} \\ &\le W_\wedge(\boldsymbol{p}) + \beta_+ r(\boldsymbol{p})\,. \end{aligned}$$

Finally, if $\psi_{\mathcal{J}}(\boldsymbol{p}) \ge \psi_{\mathcal{A}}(\boldsymbol{p})$, the lower bound gives $W_{\mathcal{U}}(\boldsymbol{p}) - \beta_- r(\boldsymbol{p}) \le -\psi_{\mathcal{J}}(\boldsymbol{p}) \le -\psi_{\mathcal{A}}(\boldsymbol{p}) \le W_{\mathcal{A}}(\boldsymbol{p})$; together with $W_{\mathcal{U}}(\boldsymbol{p}) \le W_{\mathcal{U}}(\boldsymbol{p}) + \beta_- r(\boldsymbol{p})$ this proves the last claim. □

We now bound the stability part near $\boldsymbol{e}_{a^\star}$, separating observations in $\operatorname{supp}(\boldsymbol{q}^\star)$ from those in $\mathcal{U}$. A mediator in $\operatorname{supp}(\boldsymbol{q}^\star)$ is also observed when $a^\star$ is played, so its contribution vanishes linearly in $r(\boldsymbol{p}_t)$ (Lemma C.16); a mediator in $\mathcal{U}$ must be explored through other actions, and its contribution is a width (Lemma C.17). Define

$$\begin{gathered} \bar{v}_\star = \max\left\{2, \max_{z \in \operatorname{supp}(\boldsymbol{q}^\star)} 1/q_z^\star\right\}, \qquad t_0 = \max\left\{1, \left\lceil (4\bar{v}_\star/\Lambda)^2 \right\rceil\right\}, \\ \chi_\star = \max_{a \neq a^\star} \sum_{z \in \operatorname{supp}(\boldsymbol{q}^\star)} \frac{(M_{z,a} - q_z^\star)^2}{q_z^\star}, \quad \text{and} \quad C_\star = \max\left\{\frac{2Q_\alpha \chi_\star}{\Lambda}, 2\nu_c\sqrt{|\mathcal{Z}|}\right\}, \end{gathered} \tag{49}$$

where $Q_\alpha = 2^{2-\alpha}$ is as in Lemma C.8. The set $\operatorname{supp}(\boldsymbol{q}^\star)$ is nonempty, so these constants are finite.

Lemma C.16 (The baseline and supported mediators)**.** *For every $t \geq t_0$,*

$$\mathrm{stab}_t^{\mathrm{b}}\left(\bar{\boldsymbol{B}}_t\right) \leq \frac{Q_\alpha}{4\Lambda\sqrt{t}} r\left(\boldsymbol{p}_t\right) \quad \textit{and} \quad \mathbb{E}_{t-1}[\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t)\,\mathbb{1}\{Z_t \in \mathrm{supp}(\boldsymbol{q}^\star)\}] \leq \frac{C_\star}{\sqrt{t}} r\left(\boldsymbol{p}_t\right). \tag{50}$$

Proof. Write $\boldsymbol{p} = \boldsymbol{p}_t$ and $\boldsymbol{q} = \boldsymbol{q}_t$. For $t \geq t_0$ we have $\Lambda\sqrt{t}/2 \geq 2\bar{v}_\star \geq 4$.

Baseline. Subtracting $\bar{B}_t(a^\star)\mathbf{1}$ does not change $\mathrm{stab}_t^{\mathrm{b}}(\bar{\boldsymbol{B}}_t)$; it makes the coordinate of $a^\star$ zero and every other coordinate at most $1/2$ in absolute value. Lemma C.1(c) and the signed part of Lemma C.8(b) with $\kappa = \Lambda\sqrt{t}/2$ and $\bar{v} = 1/2$ give

$$\mathrm{stab}_t^{\mathrm{b}}\left(\bar{\boldsymbol{B}}_t\right) \leq \frac{Q_\alpha}{\Lambda\sqrt{t}} \sum_{a \in \mathcal{A}\setminus\{a^\star\}} p_a^{2-\alpha}\left(\bar{B}_t(a) - \bar{B}_t(a^\star)\right)^2 \leq \frac{Q_\alpha}{4\Lambda\sqrt{t}} r(\boldsymbol{p}).$$

Residual on $\mathrm{supp}(\boldsymbol{q}^\star)$ *when* $p_{a^\star} \geq 1/2$. Fix $Z_t = z \in \mathrm{supp}(\boldsymbol{q}^\star)$; then $q_z \geq p_{a^\star} q_z^\star \geq q_z^\star/2$. Subtracting the coordinate of $a^\star$ from $\boldsymbol{x}_t$ gives the scores

$$v_a = \varepsilon_t \frac{M_{z,a} - q_z^\star}{q_z} \quad \text{for } a \neq a^\star \quad \text{and} \quad v_{a^\star} = 0.$$

If $\varepsilon_t \geq 0$, then $v_a \geq -\varepsilon_t q_z^\star/q_z \geq -2$; if $\varepsilon_t < 0$, then $|\varepsilon_t| \leq 1/2$ and $|M_{z,a} - q_z^\star| \leq 1$ give $v_a \geq -1/q_z^\star$. Thus, $v_a \geq -\bar{v}_\star$. We drop the divergence of $c\sqrt{t}\,\psi_\wedge$ from ψ_t^{x} (Lemma C.1(a)) and apply Lemma C.1(c) and the signed part of Lemma C.8(b) with $\kappa = \Lambda\sqrt{t}/2 \geq 2\bar{v}_\star$ and $\bar{v} = \bar{v}_\star$. Since $\varepsilon_t^2 \leq 1$, averaging over $Z_t \in \mathrm{supp}(\boldsymbol{q}^\star)$ gives

$$\begin{aligned}\mathbb{E}_{t-1}\left[\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t)\,\mathbb{1}\{Z_t \in \mathrm{supp}(\boldsymbol{q}^\star)\}\right] &\leq \frac{Q_\alpha}{\Lambda\sqrt{t}} \sum_{a \in \mathcal{A}\setminus\{a^\star\}} p_a^{2-\alpha} \sum_{z \in \mathrm{supp}(\boldsymbol{q}^\star)} \frac{\left(M_{z,a} - q_z^\star\right)^2}{q_z} \\ &\leq \frac{2Q_\alpha \chi_\star}{\Lambda\sqrt{t}} r(\boldsymbol{p}),\end{aligned}$$

where we used $p_a^{2-\alpha} \leq p_a$ and $q_z \geq q_z^\star/2$.

Residual on $\mathrm{supp}(\boldsymbol{q}^\star)$ *when* $p_{a^\star} < 1/2$. Then, $r\left(\boldsymbol{p}\right) > 1/2$, and equation (40) together with $\sum_{z \in \mathcal{Z}} \sqrt{q_z} \leq \sqrt{|\mathcal{Z}|}$ bounds the left side of equation (50) by $\nu_c\sqrt{|\mathcal{Z}|/t} \leq 2\nu_c\sqrt{|\mathcal{Z}|}\, r\left(\boldsymbol{p}\right)/\sqrt{t}$. The definition of $C_\star$ covers both cases. □

Lemma C.17 (Mediators outside the reference support)**.** *Recall ν_c from Corollary C.9. For every $t \geq 1$,*

$$\mathbb{E}_{t-1}[\mathrm{stab}_t^{\mathrm{x}}\left(\boldsymbol{x}_t\right)\,\mathbb{1}\{Z_t \in \mathcal{U}\}] \leq \frac{\nu_c}{\sqrt{t}}\left[W_\wedge(\boldsymbol{p}_t) + \beta_- r(\boldsymbol{p}_t)\right]. \tag{51}$$

Proof. Write $\boldsymbol{p} = \boldsymbol{p}_t$, $\boldsymbol{q} = \boldsymbol{q}_t$, and $\theta = \theta_t$. We drop the divergence of $\frac{\Lambda\sqrt{t}}{2}\psi_\alpha$ from ψ_t^{x} and apply Lemma C.4(b) and Lemma C.1(a) and (b), with $\boldsymbol{x}_t = \theta\boldsymbol{x}_t + (1-\theta)\boldsymbol{x}_t$:

$$\mathrm{stab}_t^{\mathrm{x}}(\boldsymbol{x}_t) \leq \theta\,\mathrm{stab}_{c\sqrt{t}\psi_{\mathcal{A}},\boldsymbol{p}}^{\nabla\left(c\sqrt{t}\psi_{\mathcal{A}}\right)(\boldsymbol{p})}(\boldsymbol{x}_t) + (1-\theta)\,\mathrm{stab}_{c\sqrt{t}\psi_{\mathcal{J}},\boldsymbol{p}}^{\nabla\left(c\sqrt{t}\psi_{\mathcal{J}}\right)(\boldsymbol{p})}(\boldsymbol{x}_t).$$

Fix $Z_t = z \in \mathcal{U}$ and write $\varepsilon = \varepsilon_t$.

Action term. By Lemma C.1(c), the first stability function is at most the sum over $a \in \mathcal{A}$ of the suprema in Lemma C.8(a) with $x = p_a$, $v = \varepsilon M_{z,a}/q_z$, and $\kappa = c\sqrt{t}$. If $\varepsilon \geq 0$, the denominator is at least $c\sqrt{t}$ and $\varepsilon^2 \leq 1$. If $\varepsilon < 0$, then $\varepsilon^2 \leq 1/4$, $q_z \geq 16/t$, and $M_{z,a}\sqrt{p_a} \leq \sqrt{M_{z,a}p_a} \leq \sqrt{q_z}$, so the denominator is at least $c\sqrt{t} - \sqrt{t}/4$. In both cases the summand for a is at most $\nu_c p_a^{3/2} M_{z,a}^2/(q_z^2\sqrt{t})$. Averaging over $Z_t \in \mathcal{U}$ and applying Lemma C.6 bounds the action term by

$$\frac{\nu_c}{\sqrt{t}} \sum_{z \in \mathcal{U}} \frac{\sum_{a \in \mathcal{A}} p_a^{3/2} M_{z,a}^2}{q_z} \leq \frac{\nu_c}{\sqrt{t}} \min\left\{\sum_{z \in \mathcal{U}}\sum_{a \in \mathcal{A}} \sqrt{p_a} M_{z,a},\ W_{\mathcal{U}}(\boldsymbol{p})\right\} \leq \frac{\nu_c}{\sqrt{t}} W_\wedge(\boldsymbol{p}),$$

where the last step uses $M_{z,a^\star} = 0$ and $\sum_{z \in \mathcal{U}} M_{z,a} \leq 1$.

Jensen term. Since $\boldsymbol{x}_t = \boldsymbol{M}^\top(\varepsilon q_z^{-1}\boldsymbol{e}_z)$, Lemma C.1(d) with $\phi_{z'}(y) = -c\sqrt{t}\sqrt{y}$ for every $z' \in \mathcal{Z}$ and Corollary C.9 bound the second stability function by $\nu_c/\sqrt{tq_z}$, and averaging over $Z_t \in \mathcal{U}$ gives $\nu_c W_{\mathcal{U}}(\boldsymbol{p})/\sqrt{t}$. If $1-\theta > 0$, then $\psi_{\mathcal{J}}(\boldsymbol{p}) \geq \psi_{\mathcal{A}}(\boldsymbol{p})$ by Corollary C.5, and Lemma C.15 gives $W_{\mathcal{U}}(\boldsymbol{p}) \leq W_\wedge(\boldsymbol{p}) + \beta_- r(\boldsymbol{p})$.

Both terms are therefore at most $\nu_c[W_\wedge(\boldsymbol{p}) + \beta_- r(\boldsymbol{p})]/\sqrt{t}$, and their weights θ and $1-\theta$ sum to 1. This proves equation (51), including the case $\mathcal{U} = \varnothing$. □

Combining the FTRL decomposition with these local bounds gives a regret bound that holds for every loss sequence.

Proposition C.18 (Master inequality)**.** *For the fixed reference action $a^\star$, let*

$$a_c = \frac{c}{2} + \nu_c, \qquad C_1 = \frac{c}{2}\beta_+ + \nu_c\beta_- + C_\star + \frac{Q_\alpha}{4\Lambda}, \qquad C_2 = \frac{\Lambda\left(|\mathcal{A}|-1\right)^{1-\alpha}}{2\alpha\left(1-\alpha\right)}, \\ \text{and} \quad K_0 = \frac{3}{2}D + 2G\sqrt{t_0-1}\,. \tag{52}$$

For every loss sequence $(\boldsymbol{\ell}_t)$ satisfying equation (18) and every integer $T \geq 1$,

$$\mathrm{Reg}_T \leq K_0 + a_c\sum_{t=1}^{T}\frac{\mathbb{E}[W_\wedge(\boldsymbol{p}_t)]}{\sqrt{t}} + C_1\sum_{t=1}^{T}\frac{\mathbb{E}[r(\boldsymbol{p}_t)]}{\sqrt{t}} + C_2\sum_{t=1}^{T}\frac{\mathbb{E}\left[(r(\boldsymbol{p}_t))^\alpha\right]}{\sqrt{t}}\,. \tag{53}$$

Proof. We start from equation (28). For $t \geq 2$,

$$0 \leq \sqrt{t} - \sqrt{t-1} - \frac{1}{2\sqrt{t}} = \frac{1}{2\sqrt{t}\left(\sqrt{t}+\sqrt{t-1}\right)^2} \leq \frac{1}{2}\left(\frac{1}{\sqrt{t-1}} - \frac{1}{\sqrt{t}}\right).$$

Since $0 \leq -\psi \leq D$, the excess over $1/\left(2\sqrt{t}\right)$ contributes at most $D/2$, and adding the nonnegative term for $t=1$ gives

$$\mathrm{Reg}_T \leq \frac{3}{2}D + \mathbb{E}\left[\sum_{t=1}^{T} s_t\right] + \frac{c}{2}\sum_{t=1}^{T}\frac{\mathbb{E}\left[-\psi_\wedge(\boldsymbol{p}_t)\right]}{\sqrt{t}} + \frac{\Lambda}{2}\sum_{t=1}^{T}\frac{\mathbb{E}\left[-\psi_\alpha(\boldsymbol{p}_t)\right]}{\sqrt{t}}\,. \tag{54}$$

Lemma C.15 gives $-\psi_\wedge \leq W_\wedge + \beta_+ r$. Using $p_{a^\star}^\alpha \leq 1$ and then concavity of $x \mapsto x^\alpha$ over the remaining $|\mathcal{A}|-1$ coordinates,

$$-\psi_\alpha\left(\boldsymbol{p}\right) \leq \frac{\sum_{a\in\mathcal{A}\setminus\{a^\star\}} p_a^\alpha}{\alpha\left(1-\alpha\right)} \leq \frac{\left(|\mathcal{A}|-1\right)^{1-\alpha}}{\alpha\left(1-\alpha\right)}\left(r(\boldsymbol{p})\right)^\alpha\,.$$

For $t \geq t_0$, equation (38) and Lemmas C.16 and C.17 give

$$\mathbb{E}_{t-1}[s_t] \leq \frac{\nu_c W_\wedge(\boldsymbol{p}_t)}{\sqrt{t}} + \left(\nu_c\beta_- + C_\star + \frac{Q_\alpha}{4\Lambda}\right)\frac{r(\boldsymbol{p}_t)}{\sqrt{t}}\,.$$

For $t < t_0$, equation (43) and $\sum_{t=1}^{t_0-1} t^{-1/2} \leq 2\sqrt{t_0-1}$ bound the total contribution by $2G\sqrt{t_0-1}$; the local upper bounds are nonnegative, so their sums may be extended to all rounds. Inserting these bounds into equation (54) proves equation (53). □

C.7 Gaps and complexities

Proposition C.18 bounds the regret by sums of $W_\wedge(\boldsymbol{p}_t)$, $r(\boldsymbol{p}_t)$, and $(r(\boldsymbol{p}_t))^\alpha$ weighted by $1/\sqrt{t}$. To make these sums logarithmic in T, we compare each of the three quantities with $\Gamma(\boldsymbol{p}_t)$, whose expected sum over $t \leq T$ is at most $\mathrm{Reg}_T + C$ under the self-bounding constraint equation (7). Lemma C.19 gives the comparisons, and the complexities $S_{\mathcal{A}}$, $S_{\mathcal{Z}}^{\mathrm{a}}$, and $S_\wedge$ are the best constants in them.

Let $\boldsymbol{\Delta}_{\mathcal{A}} \in [0,1]^{|\mathcal{A}|}$ be the gap vector in equation (7), and let $a^\star$ be its unique zero, so that $\Delta_{\mathcal{A}}(a^\star) = 0$ and $\Delta_{\mathcal{A}}(a) > 0$ for every $a \neq a^\star$. We use this $a^\star$ as the reference action of Section C.6. Define the minimum gap and the function Γ of the main text by

$$\Delta_{\min} = \min_{a\neq a^\star}\ \Delta_{\mathcal{A}}(a) \quad \text{and} \quad \Gamma\left(\boldsymbol{p}\right) = \sum_{a\in\mathcal{A}\setminus\{a^\star\}} p_a\Delta_{\mathcal{A}}(a)\,. \tag{55}$$

Since $\Delta_{\mathcal{A}}(a^\star) = 0$, we have $\Gamma(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{\Delta}_{\mathcal{A}} \rangle$. With the widths in equation (47), define the complexities

$$S_{\mathcal{A}} = \sum_{a \in \mathcal{A} \setminus \{a^\star\}} \frac{1}{\Delta_{\mathcal{A}}(a)}, \qquad S^{\mathrm{a}}_{\mathcal{Z}} = \sup_{\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1} : \Gamma(\boldsymbol{p}) > 0} \frac{(W_{\mathcal{U}}(\boldsymbol{p}))^2}{\Gamma(\boldsymbol{p})},$$
$$\text{and} \quad S_{\wedge} = \sup_{\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1} : \Gamma(\boldsymbol{p}) > 0} \frac{(W_{\wedge}(\boldsymbol{p}))^2}{\Gamma(\boldsymbol{p})}; \tag{56}$$

$S_{\mathcal{A}}$ and $S^{\mathrm{a}}_{\mathcal{Z}}$ agree with equation (8) and equation (9), and $S^{\mathrm{a}}_{\mathcal{Z}} = S_{\wedge} = 0$ if $\mathcal{U} = \varnothing$.

Lemma C.19 (Widths and gaps)**.** *For every $\boldsymbol{p} \in \Delta^{|\mathcal{A}|-1}$,*

$$\Gamma(\boldsymbol{p}) \geq \Delta_{\min} r(\boldsymbol{p}), \qquad (W_{\mathcal{A}}(\boldsymbol{p}))^2 \leq S_{\mathcal{A}} \Gamma(\boldsymbol{p}),$$
$$(W_{\mathcal{U}}(\boldsymbol{p}))^2 \leq S^{\mathrm{a}}_{\mathcal{Z}} \Gamma(\boldsymbol{p}), \quad \textit{and} \quad (W_{\wedge}(\boldsymbol{p}))^2 \leq S_{\wedge} \Gamma(\boldsymbol{p}). \tag{57}$$

Moreover, $S_{\wedge} \leq \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}$ and $S^{\mathrm{a}}_{\mathcal{Z}} \leq |\mathcal{U}|/\Delta_{\min}$.

Proof. The first inequality holds because every gap is at least $\Delta_{\min}$. By the Cauchy–Schwarz inequality,

$$(W_{\mathcal{A}}(\boldsymbol{p}))^2 = \left(\sum_{a \in \mathcal{A} \setminus \{a^\star\}} \sqrt{p_a \Delta_{\mathcal{A}}(a)} \frac{1}{\sqrt{\Delta_{\mathcal{A}}(a)}} \right)^2 \leq \Gamma(\boldsymbol{p}) S_{\mathcal{A}}.$$

Since $M_{z,a^\star} = 0$ for $z \in \mathcal{U}$ and the other columns sum to one, $(W_{\mathcal{U}}(\boldsymbol{p}))^2 \leq |\mathcal{U}| \sum_{z \in \mathcal{U}} (\boldsymbol{M}\boldsymbol{p})_z \leq |\mathcal{U}| r(\boldsymbol{p}) \leq |\mathcal{U}| \Gamma(\boldsymbol{p})/\Delta_{\min}$; this gives $S^{\mathrm{a}}_{\mathcal{Z}} \leq |\mathcal{U}|/\Delta_{\min}$. When $\Gamma(\boldsymbol{p}) > 0$, the remaining bounds in equation (57) follow from the definitions of the suprema, and $W_{\wedge}^2 \leq \min\{W_{\mathcal{A}}^2, W_{\mathcal{U}}^2\}$ gives $S_{\wedge} \leq \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}$. When $\Gamma(\boldsymbol{p}) = 0$, positivity of the gaps forces $\boldsymbol{p} = \boldsymbol{e}_{a^\star}$, where all widths vanish. □

C.8 Causal self-bounding

Lemma C.20 bounds the weighted sums of $r(\boldsymbol{p}_t)$ and $(r(\boldsymbol{p}_t))^\alpha$ by the square root of $\sum_{t=1}^T r(\boldsymbol{p}_t)$. Theorem C.21 combines this with Proposition C.18 and the self-bounding constraint, and closes the argument by solving the resulting quadratic inequality in $\sqrt{\mathrm{Reg}_T}$. Stochastic losses are the special case $C = 0$ (Corollary C.22).

Throughout this section, $\boldsymbol{\Delta}_{\mathcal{A}}$, $a^\star$, and the quantities equation (55) and equation (56) are as in Section C.7, $C \geq 0$, and $T \geq 1$ is a horizon at which the $(\boldsymbol{\Delta}_{\mathcal{A}}, C, T)$ self-bounding constraint equation (7) holds. Since $\Delta_{\mathcal{A}}(a^\star) = 0$, the inner sum in equation (7) is $\sum_{a \in \mathcal{A}} p_t(a) \Delta_{\mathcal{A}}(a) = \Gamma(\boldsymbol{p}_t)$, so the constraint reads

$$\sum_{t=1}^T \mathbb{E}\left[\Gamma(\boldsymbol{p}_t)\right] \leq \mathrm{Reg}_T + C. \tag{58}$$

Only this inequality, at this single horizon T, is used below. Appendix A.5 gives a sufficient per-round condition for it.

Lemma C.20 (Weighted probability sums)**.** *For $x_1, \ldots, x_T \in [0,1]$ and $\alpha \in (1/2, 1)$,*

$$\sum_{t=1}^T \frac{x_t}{\sqrt{t}} \leq 2\sqrt{\sum_{t=1}^T x_t} \quad \textit{and} \quad \sum_{t=1}^T \frac{x_t^\alpha}{\sqrt{t}} \leq \frac{2\alpha}{\sqrt{2\alpha - 1}} \sqrt{\sum_{t=1}^T x_t}. \tag{59}$$

Proof. Let $X = \sum_{t=1}^T x_t$; if $X = 0$, both bounds are immediate. Let $f(s) = x_t$ for $s \in (t-1, t]$ and $f(s) = 0$ for $s > T$. Since $s^{-1/2} \geq t^{-1/2}$ on $(t-1, t]$, the sums in equation (59) are at most $\int_0^\infty s^{-1/2} f(s)\, ds$ and $\int_0^\infty s^{-1/2} f(s)^\alpha\, ds$, where f takes values in $[0,1]$ and $\int_0^\infty f(s)\, ds = X$.

For the first integral, let $f_0 = \mathbb{1}_{(0,X)}$. Then, $(s^{-1/2} - X^{-1/2})(f(s) - f_0(s)) \leq 0$ for almost every s; integrating and using $\int_0^\infty (f - f_0) = 0$ gives $\int_0^\infty s^{-1/2} f(s)\, ds \leq \int_0^X s^{-1/2}\, ds = 2\sqrt{X}$.

For the second integral, let $y = (2\alpha - 1)X$, $m = 1/(2(1-\alpha)) > 1$, and $f_\alpha(s) = \min\{1, (y/s)^m\}$. Direct integration gives

$$\int_0^\infty f_\alpha(s)\, ds = y\left(1 + \frac{1}{m-1}\right) = X \quad \text{and}$$
$$\int_0^\infty s^{-1/2} f_\alpha(s)^\alpha\, ds = \sqrt{y}\left(2 + \frac{1}{m-1}\right) = \frac{2\alpha}{\sqrt{2\alpha-1}}\sqrt{X},$$

where both integrals converge because $m > 1$ and $m\alpha - 1/2 = m - 1 > 0$. For each $s > 0$, $f_\alpha(s)$ maximizes the concave function $u \mapsto s^{-1/2}u^\alpha - \alpha u/\sqrt{y}$ over $[0,1]$, as its derivative shows. Evaluating this inequality at $u = f(s)$, integrating, and using $\int_0^\infty f(s)ds = \int_0^\infty f_\alpha(s)ds$ gives $\int_0^\infty s^{-1/2} f(s)^\alpha\, ds \le \frac{2\alpha}{\sqrt{2\alpha-1}}\sqrt{X}$. □

Theorem C.21 (Regret under causal self-bounding)**.** *Under the assumptions of this section, recall a_c, C_1, C_2, and K_0 from equation (52) with reference action $a^\star$, let $\kappa_c = 2a_c^2$, and define the constant K_{sb} and the bound B_T by*

$$K_{\mathrm{sb}} = 2K_0 + \frac{2}{\Delta_{\min}}\left(2C_1 + \frac{2\alpha C_2}{\sqrt{2\alpha-1}}\right)^2 \quad \textit{and} \quad B_T = K_{\mathrm{sb}} + \kappa_c \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}(1 + \log T)\,.$$

Then,

$$\mathrm{Reg}_T \le B_T + 2\sqrt{CB_T}\,. \tag{60}$$

Proof. By Lemma C.19, Jensen's inequality, and the Cauchy–Schwarz inequality over t,

$$\sum_{t=1}^T \frac{\mathbb{E}[W_\wedge(\boldsymbol{p}_t)]}{\sqrt{t}} \le \sqrt{S_\wedge}\sum_{t=1}^T \sqrt{\frac{\mathbb{E}[\Gamma(\boldsymbol{p}_t)]}{t}} \le \sqrt{S_\wedge\,(1+\log T)\sum_{t=1}^T \mathbb{E}\left[\Gamma(\boldsymbol{p}_t)\right]}\,.$$

Also $\sum_{t=1}^T \mathbb{E}[r(\boldsymbol{p}_t)] \le \sum_{t=1}^T \mathbb{E}[\Gamma(\boldsymbol{p}_t)]/\Delta_{\min}$ and $\mathbb{E}[(r(\boldsymbol{p}_t))^\alpha] \le (\mathbb{E}[r(\boldsymbol{p}_t)])^\alpha$. Lemma C.20 with $x_t = \mathbb{E}[r(\boldsymbol{p}_t)]$ and Proposition C.18 give

$$\begin{aligned}\mathrm{Reg}_T &\le K_0 + A_T\sqrt{\sum_{t=1}^T \mathbb{E}\left[\Gamma(\boldsymbol{p}_t)\right]} \le K_0 + A_T\sqrt{\mathrm{Reg}_T + C}, \quad \text{where}\\ A_T &= a_c\sqrt{S_\wedge(1+\log T)} + \frac{2C_1 + 2\alpha C_2/\sqrt{2\alpha-1}}{\sqrt{\Delta_{\min}}}\,.\end{aligned} \tag{61}$$

The second inequality is equation (58), which also shows $\mathrm{Reg}_T + C \ge \sum_{t=1}^T \mathbb{E}[\Gamma(\boldsymbol{p}_t)] \ge 0$. If $\mathrm{Reg}_T < 0$, equation (60) is immediate. Otherwise, $\sqrt{\mathrm{Reg}_T + C} \le \sqrt{\mathrm{Reg}_T} + \sqrt{C}$ and $A_T\sqrt{\mathrm{Reg}_T} \le (A_T^2 + \mathrm{Reg}_T)/2$ give $\mathrm{Reg}_T \le 2K_0 + A_T^2 + 2A_T\sqrt{C}$. Since $A_T^2 \le 2a_c^2 S_\wedge(1 + \log T) + (2/\Delta_{\min})(2C_1 + 2\alpha C_2/\sqrt{2\alpha-1})^2$ and $S_\wedge \le \min\{S_{\mathcal{A}}, S^{\mathrm{a}}_{\mathcal{Z}}\}$, we have $2K_0 + A_T^2 \le B_T$ and $A_T \le \sqrt{B_T}$, which proves equation (60). □

Corollary C.22 (Stochastic losses)**.** *Suppose that, for every learner, equation (18) holds with $\mathbb{E}_{t-1}[\boldsymbol{\ell}_t] = \boldsymbol{\mu}$ almost surely for every $t \ge 1$, for a fixed $\boldsymbol{\mu} \in [0,1]^{|\mathcal{Z}|}$ (this includes the stochastic game of the main text), and that $a^\star$ uniquely minimizes $\langle \boldsymbol{M}_{\cdot,a}, \boldsymbol{\mu}\rangle$ over $a \in \mathcal{A}$. Let $\Delta_{\mathcal{A}}(a) = \langle \boldsymbol{M}_{\cdot,a} - \boldsymbol{q}^\star, \boldsymbol{\mu}\rangle$, which is the action gap equation (2). Then, the $(\boldsymbol{\Delta}_{\mathcal{A}}, 0, T)$ self-bounding constraint equation (7) holds for every learner and every integer $T \ge 1$. Consequently, the iterates equation (21) satisfy $\mathrm{Reg}_T \le B_T$ for every integer $T \ge 1$, with B_T as in Theorem C.21. If, in addition, $M_{z,a^\star} > 0$ for every $z \in \mathcal{Z}$, then $\mathrm{Reg}_T \le K_{\mathrm{sb}}$ for every integer $T \ge 1$.*

Proof. The columns of $\boldsymbol{M}$ are distributions and $\boldsymbol{\mu} \in [0,1]^{|\mathcal{Z}|}$, so $\Delta_{\mathcal{A}}(a) \le 1$; optimality and uniqueness of $a^\star$ give $\Delta_{\mathcal{A}}(a^\star) = 0$ and $\Delta_{\mathcal{A}}(a) > 0$ for $a \ne a^\star$. Since $\boldsymbol{p}_t$ is $\mathcal{F}_{t-1}$-measurable, the expectation in equation (19) equals $\mathbb{E}[\sum_{t=1}^T \langle \boldsymbol{M}\boldsymbol{p}_t - \boldsymbol{M}_{\cdot,b}, \boldsymbol{\mu}\rangle]$, whose maximum over b is attained at $b = a^\star$. As $\langle \boldsymbol{M}\boldsymbol{p} - \boldsymbol{q}^\star, \boldsymbol{\mu}\rangle = \sum_{a\in\mathcal{A}} p_a \Delta_{\mathcal{A}}(a) = \Gamma(\boldsymbol{p})$,

$$\mathrm{Reg}_T = \sum_{t=1}^T \mathbb{E}\left[\Gamma\left(\boldsymbol{p}_t\right)\right]\,.$$

This is equation (58) with $C = 0$, and Theorem C.21 gives $\mathrm{Reg}_T \le B_T$. If $M_{z,a^\star} > 0$ for every $z \in \mathcal{Z}$, then $\mathcal{U} = \varnothing$, so $S^{\mathrm{a}}_{\mathcal{Z}} = 0$ and $B_T = K_{\mathrm{sb}}$. □