PRODuctive bandits: Importance Weighting No More

Julian Zimmert
Google Research
zimmert@google.com

Teodor V. Marinov
Google Research
tvmarinov@google.com

Abstract

Prod is a seminal algorithm in full-information online learning, which has been conjectured to be fundamentally sub-optimal for multi-armed bandits. By leveraging the interpretation of Prod as a first-order OMD approximation, we present the following surprising results: 1. Variants of Prod can obtain optimal regret for adversarial multi-armed bandits. 2. There exists a simple and (arguably) importance-weighting free variant with optimal rate. 3. One can even achieve best-both-worlds guarantees with logarithmic regret in the stochastic regime.

The bandit algorithms in this work use simple arithmetic update rules without the need of solving optimization problems typical in prior work. Finally, the results directly improve the state of the art of incentive-compatible bandits.

1 Introduction

The adversarial multi-armed bandit (MAB) problem is a seminal online learning problem with applications in experimental design, online advertisement and more [Thompson, 1933, Lai and Robbins, 1985, Auer et al., 2002a,b]. MABs are characterized by the limited feedback given to the learner in every round, the so-called bandit feedback, in which the learner only observes the loss of their selected action, unlike in the full information, also known as the experts, setting where the loss of all actions are provided as feedback.

The first nearly optimal algorithm for the adversarial MAB problem is EXP3 [Auer et al., 2002b]. EXP3 is a direct adaptation of the Hedge algorithm [Littlestone and Warmuth, 1994, Cesa-Bianchi et al., 1997, Freund and Schapire, 1997] with importance weighting to handle partial bandit feedback. Hedge and EXP3 are special versions of online mirror descent (OMD), where the Bregman divergence is the KL divergence induced by the Negative entropy potential. The OMD view of online learning [Abernethy et al., 2008] has lead to a wide range of MAB algorithms such as Tsallis-INF [Audibert and Bubeck, 2009] and Logbarrier [Agarwal et al., 2017], which enjoy improved regret guarantees. These guarantees can be attributed to regularizers more suited to the bandit feedback setting, compared to the negative entropy regularizer. A downside of OMD is that usually the mirror descent update is not closed form and requires (approximately) solving optimization problems at every iteration.

Alternative full-information algorithms with simple arithmetic updates are Prod [Cesa-Bianchi et al., 2007, Even-Dar et al., 2008, Gaillard et al., 2014], which enjoys second order regret bounds and Multiplicative Weights Update (MWU) [Arora et al., 2012]. Prod is known to be closely related to Hedge, both of which are different generalizations of the weighted majority algorithm [Littlestone and Warmuth, 1994] to non-binary feedback.

More recently, Freeman et al. [2020] studied incentive-compatible online learning, a setting where experts are not necessarily truthful but make predictions strategically with regard to the agent's algorithm. Motivated by deriving an algorithm where the incentives of experts align with the agent,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2710

they propose WSU which can be seen as an instance of Prod. Freeman et al. [2020] further introduce a bandit adaptation WSU-UX, the first Prod algorithm for bandits. Unfortunately, WSU-UX only enjoys $T^{\frac{2}{3}}$ regret guarantees. This is not merely an issue with their analysis as has been shown via lower bounds [Mortazavi et al., 2024] and leads to the conjecture that this might be a fundamental separation between full-information and bandits.

We make the following contributions for understanding Prod under bandit feedback:

1. We disprove the separation conjecture by providing a simple modification of WSU-UX with nearly optimal $O(\sqrt{KT\log(K)})$ regret guarantees.
2. We present a Prod variant that does not require importance weighting and yet enjoys $O(\sqrt{KT\log(T)})$ regret bounds.
3. We present a Prod variant that achieves best of both worlds regret guarantees, i.e. it enjoys improved $O(\log(T))$ regret bounds when the losses are stochastic, while maintaining worst-case $O(\sqrt{KT})$ regret bounds.

Notation: For $N \in \mathbb{N}$, let $[N] = \{1, \dots, N\}$. We use $\langle \cdot, \cdot \rangle$ to denote the regular Euclidean scalar product and $\Delta(A)$ to denote the probability simplex over a finite set A. O is the standard Landau notation hiding numerical constants, while $\widetilde{O}$ omits polylogaritmic factors as well. The expectation $\mathbb{E}$ is always taken over all randomness of the algorithm, losses and experts, while $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot\,|\mathcal{F}_t]$, where $\mathcal{F}_t$ is the filtration over all randomness up to step t. $\mathbb{I}(E)$ denotes the indicator function function for the event E. For a convex differentiable function F, the Bregman divergence is defined by $D_F(y, x) = F(y) - F(x) - \langle y - x, \nabla F(x)\rangle$.

2 Problem setting and related work

The adversarial bandit problem is formally defined as follows. In every round $t = 1, \dots, T$, an (oblivious) adversary selects a loss $\ell_t \in [0,1]^K$ (it is possible to extend the loss range to $[-1,1]^K$) unknown to the agent. The agent simultaneously selects an expert $A_t \sim \pi_t$, $\pi_t \in \Delta([K])$. The agent incurs and observes the loss ℓ_{t,A_t}, but does not see the losses of other experts. The goal is to minimize the pseudo-regret[1]

$$\mathrm{Reg} = \max_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} - \ell_{t,i}\right] = \max_{u\in\Delta([K])} \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi_t - u, \ell_t\rangle\right],$$

Popular families of algorithms for this problem setting include online mirror descent (OMD) and follow the regularized leader (FTRL). Typically the algorithms use unbiased loss estimates of the loss vector via importance weighting: $\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{\pi_{t,i}}\mathbb{I}(A_t = i)$. We note that other types of importance weighted estimators have been used in literature such as the implicit exploration estimator Kocák et al. [2014], which has improved variance properties. The algorithms are defined by a twice-differentiable convex potential function $F : \mathbb{R}^K \to \mathbb{R}$ and a learning rate schedule η_t, the agent maintains a distribution via

$$\pi_{t+1} = \arg\min_{\pi\in\Delta([K])} \left\langle \pi, \eta_t \hat{\ell}_t \right\rangle - D_F(\pi, \pi_t)\,, \tag{OMD}$$

$$\pi_{t+1} = \arg\min_{\pi\in\Delta([K])} \left\langle \pi, \eta_t \sum_{s=1}^{t} \hat{\ell}_s \right\rangle - F(\pi)\,, \tag{FTRL}$$

OMD optimizes locally given the last loss and, as we will show, is most closely related to Prod. FTRL on the other hand performs a global optimization and is generally considered superior for adaptive bounds with time-dependent learning rates. In some special cases, such as time-independent learning rate with potentials that satisfy $\|\nabla F(x)\| \to \infty$ on the border of the optimization set, both algorithms are equivalent. Common potentials in the bandit literature are given in Table 1 and we refer to their respective Bregman divergences as D_{KL}, D_{TS} and D_{LB} respectively. The negative entropy is the potential which defines Hedge and Exp-3 (and derivatives) [Littlestone and

[1]For the rest of the paper we refer to pseudo-regret as regret for simplicity.

	Negentropy/KL divergence	$1/2$-Tsallis Entropy	Logbarrier
$F(\pi)$	$\sum_{i=1}^K \pi_{t,i} \log(\pi_{t,i})$	$-2\sum_{i=1}^K \sqrt{\pi_{t,i}}$	$-\sum_{i=1}^K \log(\pi_{t,i})$

Table 1: Common potential functions

Warmuth, 1994, Vovk, 1995, Freund and Schapire, 1997, Auer et al., 2002b, Kocák et al., 2014]. The $1/2$-Tsallis Entropy is the key to achieving optimal best-of-both-worlds regret guarantees as was first demonstrated by Zimmert and Seldin [2021]. The Logbarrier potential was used by Agarwal et al. [2017] to first solve the corralling of bandits problem and has found many applications in model-selection problems [Foster et al., 2020], regret bounds which depend on the properties of the loss sequence [Wei and Luo, 2018, Lee et al., 2020b,a] and various other bandit problems.

2.1 Prod family of algorithms

The original version of Prod [Cesa-Bianchi et al., 2007] maintains weights $w_{t,i}$ for each experts which are updated via $w_{t+1,i} = w_{t,i}(1-\eta\ell_{t,i})$ and the agent plays the policy $\pi_{t,i} \propto w_{t,i}$. This framework has been extended to D-Prod [Even-Dar et al., 2008], which shifts the losses in the weight update by the loss of a fixed policy, and ML-Prod [Gaillard et al., 2014] that shifts losses by the mean of the current policy (among other modifications). With a suitable shift in losses, one can ensure that the weights sum up to 1 and hence operate directly on the policy space. In its simplest form, this is

$$\pi_{1,i} = \frac{1}{K}\,, \qquad \pi_{t+1,i} = \pi_{t,i}(1-\eta(\ell_{t,i}-\lambda_t))\,, \qquad \lambda_t = \sum_{j=1}^K \pi_{t,j}\ell_{t,j}\,. \qquad \text{(Vanilla-Prod/WSU)}$$

We refer to this update as Vanilla-Prod to emphasize its connection to the Prod literature, however this algorithm is exactly WSU [Freeman et al., 2020] derived for incentive-compatible online learning. We consider any algorithm a variant of Prod if it performs product updates of the form $\pi_{t+1,i} = \pi_{t,i}(1-\eta L_{t,i}(\ell_t; A_t))$, where $L_{t,i}$ are linear affine functions of the loss. From now on we always assume the initial policy is $\pi_{t,i} = 1/K$, and this holds for all algorithms presented in the paper. The appeal of Prod algorithms lies in their simple arithmetic update rule. A second motivation for using Prod updates is the mentioned incentive-compatibility.

2.2 Incentive-compatible online learning

In the incentive-compatible online learning setting, introduced by Freeman et al. [2020], experts provide recommendations, for example a prediction of whether it will rain on the next day. Each expert has an internal belief and the agent would like to receive each expert's true beliefs in order to learn to follow the best expert. In the simplest setting the experts make predictions about binary outcomes, with the i-th expert having (private) belief $b_{t,i} \in [0,1]$ about the t-th round outcome. The expert's belief is unknown to the agent and the expert only reports a prediction $p_{t,i} \in [0,1]$ about the outcome. Based on the expert predictions, $\{p_{t,i}\}_{i\in[K]}$ the agent makes a prediction based on $\bar{p}_t = \sum_{i=1}^K \pi_{t,i}p_{t,i}$ and incurs a loss $\mathcal{L}(\bar{p}_t, r_t) \in [0,1]$ based on the realized outcome $r_t \in \{0,1\}$. In the weather forecasting example the outcome is the indicator if it rains the next day and the loss is $\mathcal{L}(\bar{p}_t, r_t) = (r_t - \bar{p}_t)^2$. In Freeman et al. [2020] the experts only care about maximizing the probability that they are selected which does not necessarily result in truthful reporting, that is $b_{t,i}$ may differ from $p_{t,i}$. The agent's goal of receiving the true beliefs, $\{b_{t,i}\}_{i\in[K]}$, can be achieved by playing an *incentive compatible strategy* which will always prefer selecting an truthful expert, that is the probability of $\pi_{t+1,i}$ of selecting expert i when the expert reports $b_{t,i}$ may only decrease if the expert reports any other $p_{t,i}$ instead, no matter how the remaining experts act throughout the game. This is made precise in Definition 2.1 of Freeman et al. [2020].

Freeman et al. [2020] show that standard OMD and FTRL algorithms are in fact not incentive compatible even when the loss $\mathcal{L}$ is restricted to be *proper* that is $\mathbb{E}_{r\sim\mathsf{Bern}(b)}[\mathcal{L}(p,r)] \geq \mathbb{E}_{r\sim\mathsf{Bern}(b)}[\mathcal{L}(b,r)]$ for all $p \neq b$. It turns out that any update for π_{t+1} which is linear affine in the proper loss function will lead to incentive compatibility and so the Prod family will ensures that experts report their true believes in this setting, i.e. they are incentive-compatible. The state of the art for incentive-compatible bandits is $T^{\frac{2}{3}}$ regret and any improvement for Prod directly transfers to better rates for this setting as well.

https://doi.org/10.52202/079017-2710

3 Modifying WSU-UX for nearly optimal regret guarantees

We begin by presenting a minimal modification of Algorithm WSU-UX which is sufficient for a regret guarantee of the order $O(\sqrt{KT\log(K)})$.

WSU-UX uses importance-weighted updates and injects a small uniform exploration.

$$\tilde{\pi}_{t,i} = \frac{\gamma}{K} + (1-\gamma)\pi_{t,i}, \qquad A_t \sim \tilde{\pi}_{t,i}, \qquad \hat{\ell}_{t,i} = \mathbb{I}(A_t = i)\frac{\ell_{t,i}}{\tilde{\pi}_{t,i}}$$

$$\pi_{t+1,i} = \pi_{t,i}(1 - \eta(\hat{\ell}_{t,i} - \lambda_t))\,, \qquad \lambda_t = \sum_{j=1}^{K} \pi_{t,j}\hat{\ell}_{t,j}\,, \tag{WSU-UX}$$

where γ is the mixture coefficient. The role of uniform exploration is to ensure that the policy updates are proper i.e. $\pi_{t+1,i} \in (0,1)$. Freeman et al. [2020] uses the following key lemmas in their analysis, which hold for any sequence of losses $\ell_t \in [0,1]^K$. For completeness we restate the results we use below.

Lemma 1 (Lemma 4.1 [Freeman et al., 2020]). *If $\eta K/\gamma \le \frac{1}{2}$, the WSU-UX weights π_t and $\tilde{\pi}_t$ are valid probability distributions for all $t \in [T]$.*

Lemma 2 (Lemma 4.3 [Freeman et al., 2020]). *For WSU-UX, the probability vectors $\{\pi_t\}_{t\in[T]}$ and loss estimators $\hat{\ell}_t$ satisfy the following second order-bound*

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \pi_{t,i}\hat{\ell}_{t,i} - \sum_{t=1}^{T}\hat{\ell}_{t,i^\star} \le \frac{\log(K)}{\eta} + \eta\sum_{t=1}^{T}\hat{\ell}_{t,i^\star}^2 + \eta\sum_{t=1}^{T}\sum_{i=1}^{K}\pi_{t,i}\hat{\ell}_{t,i}^2,$$

where $i^\star$ is the optimal expert/arm.

The bound in Lemma 2 is almost the standard regret bound that appears in the analysis of Hedge, except for the term $\eta\sum_{t=1}^{T}\hat{\ell}_{t,i^\star}^2$. This term is the reason why prior work can not show regret bounds smaller than $T^{\frac{2}{3}}$. Even after taking the expectation over the randomness of the agents actions, this term scales with with $1/\pi_{t,i^\star}$, which is potentially unbounded. Alternatively this term can be written as $\mathbb{E}_t^{j\sim\pi^\star}[\eta\hat{\ell}_{tj}^2]$ (where $\pi^\star$ is the policy picking $i^\star$ with probability 1) and if one could perform a change of measure to $\mathbb{E}_t^{j\sim\tilde{\pi}_t}[\eta\hat{\ell}_{tj}^2]$, this term is immediately controllable.

In fact, change of measure techniques for bandits are now well established [Foster et al., 2020, Luo et al., 2021] by introducing biases to the losses. Assume we construct a bias to the losses $\tilde{\ell}_t = \ell_t + \delta_t$, which satisfies the same regret guarantee, Reg, as the original loss sequence, then running an algorithm over the biased loss sequence $\tilde{\ell}_t$ which selects $A_t \sim \tilde{\pi}_t$ ensures

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,A_t} - \ell_{t,i^\star}\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,i^\star} + \delta_{t,A_t} - \delta_{t,i^\star}\right] \\ &= \mathrm{Reg} + \mathbb{E}\left[\sum_{t=1}^{T}\underbrace{(\mathbb{E}_t^{j\sim\tilde{\pi}_t}[\delta_{t,j}] - \mathbb{E}_t^{j\sim\pi^\star}[\delta_{t,j}])}_{\text{change of measure}}\right].\end{aligned}$$

We introduce now the following modification to the losses

$$\tilde{\ell}_{t,i} = \ell_{t,i}\left(1 - \frac{\eta}{\tilde{\pi}_{t,i}}\right), \qquad \hat{\ell}_{t,i} = \mathbb{I}(A_t = i)\frac{\tilde{\ell}_{t,i}}{\tilde{\pi}_{t,i}}\,. \tag{1}$$

which corresponds to $\delta_{ti} = \frac{\eta\ell_{ti}}{\tilde{\pi}_{t,i}}$. This yields the change of measure term

$$\mathbb{E}_t^{j\sim\tilde{\pi}_t}[\delta_{t,j}] - \mathbb{E}_t^{j\sim\pi^\star}[\delta_{t,j}] = \sum_{i=1}^{K}\eta\ell_{ti} - \mathbb{E}_t^{j\sim\pi^\star}\left[\frac{\eta\ell_{tj}}{\tilde{\pi}_{tj}}\right] \le \eta K - \mathbb{E}_t^{j\sim\pi^\star}[\eta\hat{\ell}_{tj}^2]\,,$$

which is sufficient for controlling the term $\eta\sum_{t=1}^{T}\hat{\ell}_{t,i^\star}^2$ in Lemma 2.

https://doi.org/10.52202/079017-2710

Theorem 1. *Running WSU-UX with the loss estimators in Equation 1 and* $\gamma = \frac{\eta K}{2}, \eta = \Theta(\sqrt{\frac{\log(K)}{KT}})$ *guarantees the following regret bound*

$$\sum_{t=1}^{T} \mathbb{E}[\ell_{t,A_t} - \ell_{t,i^\star}] \leq O(\sqrt{KT\log(K)}).$$

The proof of Theorem 1 is deferred to Appendix B.

3.1 Intuition on biasing the update and the Prod family of algorithms

We provide an intuition in this section about why WSU-UX is not tight and why the bias we choose is able to correct the regret. The modern analysis of OMD (or FTRL) with a divergence function D is follows the template[2]

$$\begin{aligned}\sum_{t=1}^{T} \langle \pi_t^{\text{OMD}} - \pi^\star, \ell_t \rangle &= \sum_{t=1}^{T} \left(\langle \pi_t^{\text{OMD}} - \pi^\star, \ell_t \rangle + \frac{D(\pi^\star, \pi_{t+1}^{\text{OMD}}) - D(\pi^\star, \pi_t^{\text{OMD}})}{\eta} \right) \\ + \frac{D(\pi^\star, \pi_1^{\text{OMD}}) - D(\pi^\star, \pi_{T+1}^{\text{OMD}})}{\eta} &\leq \sum_{t=1}^{T} \left(\underbrace{\langle \pi_t^{\text{OMD}} - \pi_{t+1}^{\text{OMD}}, \ell_t \rangle + \frac{D(\pi_{t+1}^{\text{OMD}}, \pi_t^{\text{OMD}})}{\eta}}_{\text{stability}} \right) + \frac{D(\pi^\star, \pi_1^{\text{OMD}})}{\eta}\,,\end{aligned} \tag{2}$$

where the inequality crucially relies on π_{t+1}^{OMD} being the 1-step OMD update to the previous policy. OMD algorithms like Hedge choose the policy that minimizes the per-step *stability* term in every round, which is what allows for the stability term to be bounded appropriately. If instead of playing π_t^{OMD} an approximate policy $\pi_t \approx \pi_t^{\text{OMD}}$ is played, the template regret analysis can be changed by adding the terms

$$\eta^{-1}(D(\pi^\star, \pi_{t+1}) - D(\pi^\star, \pi_{t+1}^{\text{OMD}}))\,,$$

where π_{t+1}^{OMD} is now the 1-step OMD update from π_t. When D is the KL divergence and the approximate policy π_t is coming from the WSU-UX, this term contributes the undesirable $\eta\hat{\ell}_{t,i^\star}^2$.

We now explain how our loss biasing solves this issue. The Vanilla-Prod/WSU update can be seen as a first order approximation to the Hedge update, that is

$$\pi_{t+1,i}^{\text{OMD}} = \pi_{t,i}\exp(-\eta(\hat{\ell}_{t,i} - \lambda_t)) \underbrace{\approx}_{\text{first-order}} \pi_{t,i}(1 - \eta(\hat{\ell}_{t,i} - \lambda_t)) = \pi_{t,i}\,,$$

where λ_t is a normalization factor. Tuning λ_t such that $\sum_{i=1}^{K} \pi_{t+1,i} = 1$ recovers Vanilla-Prod/WSU. To control the undesirable terms, we have to make the approximation tighter. The loss-biasing introduced in the previous section acts as a correction which brings the Vanilla-Prod/WSU update closer to the second order approximation of the Hedge update. Indeed, we have

$$\begin{aligned}\pi_{t+1,i}^{\text{OMD}} = \pi_{t,i}\exp(-\eta(\hat{\ell}_{t,i} - \lambda_t)) &\underbrace{\approx}_{\text{second-order}} \pi_{t,i}\left(1 - \eta(\hat{\ell}_{t,i} - \lambda_t) + \frac{\eta^2}{2}(\hat{\ell}_{t,i} - \lambda_t)^2\right) \\ &= \pi_{t,i}\left(1 - \eta(\hat{\ell}_{t,i} - \lambda_t)\left(1 - \frac{\eta}{2}(\hat{\ell}_{t,i} - \lambda_t)\right)\right)\,,\end{aligned}$$

and so our loss adjustment in Equation 1, $\tilde{\ell}_{t,i} = \ell_{t,i}(1 - \eta/\tilde{\pi}_{t,i})$, can be seen as a second order correction to the term $\frac{\eta^2}{2}(\hat{\ell}_{t,i} - \lambda_t)^2$. We cannot exactly correct the second order difference with linear update rules, which we address by slightly overcorrecting, i.e. biasing by a larger amount than the second order adjustments implies as necessary. That is, the correction term we use is of the order $\eta/\tilde{\pi}_{t,i}$ instead of $\eta\ell_{t,i}/\tilde{\pi}_{t,i}$. Fortunately, the regret analysis is not sensitive towards this as we have shown in Theorem 1.

[2]This might look very dissimilar from the original Hedge/EXP/MWU analysis, but it is actually equivalent after accounting for the special form of the KL divergence.

https://doi.org/10.52202/079017-2710

4 Importance weighting free adversarial MAB with LB-Prod

While our biased WSU-UX obtains optimal regret, it still has to go through the extra complexity of injecting additional uniform exploration at a rate of γ to ensure proper updates and add bias to the losses. As mentioned in the introduction, prior work proposed other potential functions that have favourable properties for bandit feedback. Using the same linearization argument to derive a Prod version based on the Logbarrier leads to a surprisingly simple algorithm without loss biasing that is arguably importance weighting free. LB-Prod differs from WSU-UX by using the masked loss $\tilde{\ell}_{t,i} = \ell_{t,i}\mathbb{I}(A_t = i)$ instead of the importance weighted loss and a non-symmetric normalization $\lambda_{t,i}$:

$$\pi_{t+1,i} = \pi_{t,i}(1 - \eta(\tilde{\ell}_{t,i} - \lambda_{t,i})), \qquad \lambda_{t,i} = \pi_{t,i}\frac{\pi_{t,A_t}\ell_{t,A_t}}{\sum_{j=1}^{K}\pi_{t,j}^2}\,. \tag{LB-Prod}$$

It is easy to confirm $\tilde{\ell}_{t,i} - \lambda_{t,i} \in [-1, 1]$ via $\pi_{t,i}\pi_{t,A_t} \leq (\pi_{t,i}^2 + \pi_{t,A_t}^2)/2$ yielding proper updates for $\eta < 1$. The following theorem shows that this simple algorithm is rate optimal under the right tuning.

Theorem 2. *For any sequence of losses $\ell_t \in [-1, 1]^K$ and any $\eta < 1$, LB-Prod produces valid distributions $\pi_t \in \Delta([K])$ and its regret is bounded by*

$$\sum_{t=1}^{T}\mathbb{E}[\ell_{t,A_t} - \ell_{t,i^\star}] \leq 2 + \frac{K\log(T)}{\eta} + \frac{2\eta T}{1-\eta}\,.$$

Tuning $\eta = \sqrt{\frac{\log(T)K}{2T}}$ results in a regret bound of $O(\sqrt{KT\log(T)})$ for any $T > \frac{K\log(T)}{2}$. The proof of Theorem 2 is deferred to the end of the section.

4.1 Intuition of LB-Prod

As mentioned before, LB-Prod is the linear approximation of OMD with Bregman divergence induced by the Logbarrier potential, that is D_{LB} induced by the potential fuction $F(x) = -\sum_{i=1}^{K}\log(x_i)$. The one-step Logbarrier OMD update of a policy π_t with importance-weighting is known (see e.g. [Zimmert and Seldin, 2021]) to take the form

$$\pi_{t+1,i}^{\text{OMD}} = \frac{\pi_{t,i}}{1 + \eta\pi_{t,i}(\hat{\ell}_{t,i} - \lambda_t)} \approx \pi_{t,i}(1 - \eta\pi_{t,i}(\hat{\ell}_{t,i} - \lambda_t)) = \pi_{t+1,i}\,,$$

where λ_t is a normalization constant that ensures π_t^{OMD} is a probability distribution. If instead λ_t is tuned so that $\sum_{i=1}^{K}\pi_{t+1,i} = 1$, the LB-Prod update is recovered. This can be be verified by setting $\tilde{\ell}_{t,i} = \pi_{t,i}\hat{\ell}_{t,i}$ and $\lambda_{t,i} = \pi_{t,i}\lambda_t$. The curvature of the Logbarrier regularization is what ensures that the importance weighted loss $\hat{\ell}_{t,i}$ is always multiplied with its probability $\pi_{t,i}$, allowing to run the algorithm on the masked non-weighted loss sequence directly.

Additionally, the second order approximation is

$$\pi_{t+1,i}^{\text{OMD}} = \frac{\pi_{t,i}}{1 + \eta\pi_{t,i}(\hat{\ell}_{t,i} - \lambda_t)} \underbrace{\approx}_{\text{second-order}} \pi_{t,i}(1 - \eta\pi_{t,i}(\hat{\ell}_{t,i} - \lambda_t) + \eta^2\pi_{t,i}^2(\hat{\ell}_{t,i} - \lambda_t)^2)\,.$$

The additional undesirable terms, unlike in the case for WSU-UX, will only contribute ηT regret if not adjusted for as we show next.

4.2 Analysis of LB-Prod

The following technical lemma is proven in the appendix.

Lemma 3. *For any timestep t and arm i, it holds*

$$\mathbb{E}_t[\tilde{\ell}_{t,i} - \lambda_{t,i}] = \pi_{t,i}\left(\ell_{t,i} - c_t\right)\,, \qquad \mathbb{E}_t[(\tilde{\ell}_{t,i} - \lambda_{t,i})^2] \leq 2\pi_{t,i}\,,$$

where $c_t \in [-1, 1]$ is an arm independent constant.

In Section 3.1 we argued that one needs to bound the additional term $\eta^{-1}(D_{LB}(\pi^\star, \pi_{t+1}) - D_{LB}(\pi^\star, \pi_{t+1}^{\text{OMD}}))$ to reduce the analysis to standard OMD. While this term is nicely bounded for LB-Prod, it turns out that it is easier to directly bound the "prototype" of the *stability* term in Equation (2) due to the fact that we have a closed form expression of π_{t+1}.

Lemma 4. *For any time $t \in [T]$ and any $u \in \Delta([K])$, it holds*

$$\langle \pi_t - u, \ell_t \rangle + \mathbb{E}_t \left[\eta^{-1} D_{LB}(u, \pi_{t+1}) \right] - \eta^{-1} D_{LB}(u, \pi_t) \leq \frac{2\eta}{1-\eta} .$$

The proof is an algebraic exercise and deferred to the supplementary material. Finally, we can prove the main regret guarantee.

Proof of Theorem 2. To show that this algorithm outputs proper probability distributions, note that

$$\sum_{i=1}^{K} \pi_{t+1,i} = \left(\sum_{i=1}^{K} \pi_{t,i} \right) - \eta \pi_{t,A_t} \ell_{t,A_t} + \eta \sum_{j=1}^{K} \frac{\pi_{t,A_t} \pi_{t,j}^2}{\sum_{k=1}^{K} \pi_{t,k}^2} \ell_{t,A_t} = \sum_{i=1}^{K} \pi_{t,i} = \cdots = \sum_{i=1}^{K} \pi_{1,i} = 1 .$$

Additionally we have seen that $|\tilde{\ell}_{t,i} - \lambda_{t,i}| \leq 1$, hence for any $\eta < 1$, the probability of any arm is strictly positive. For any comparator $u^\star$, we define $u = u^\star + \frac{1}{T}(\pi_1 - u^\star)$, which satisfies $\sum_{t=1}^{T} \langle u - u^\star, \ell_t \rangle \leq 2$. Using Lemma 4 and Equation (2), we obtain by the telescoping sum of Bregman terms

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \pi_t - u, \ell_t \rangle \right] \leq \frac{4\eta T}{1-\eta} + \eta^{-1} \mathbb{E}\left[D_{LB}(u, \pi_1) - D_{LB}(u, \pi_{T+1}) \right] \leq \frac{4\eta T}{1-\eta} + \frac{K \log(T)}{\eta} .$$

□

4.3 The perturbation analysis

We outline an alternative analysis that reuses established machinery and might be more accessible for some readers. Our analysis begins by viewing the Prod update as an exact OMD update over a perturbed loss sequence. Indeed, there is a sequence of perturbations $\{\epsilon_t\}_{t \in [T]}, \epsilon_t \in \mathbb{R}^K$, such that

$$\pi_{t+1}^{\text{OMD}} = \frac{\pi_{t,i}}{1 + \eta \pi_{t,i} (\hat{\ell}_{t,i} - \epsilon_{t,i} - \lambda_t)} = \pi_{t,i} (1 - \eta \pi_{t,i} (\hat{\ell}_{t,i} - \lambda_t)) = \pi_{t+1,i} .$$

The exact form of $\epsilon_{t,i}$ satisfies the following

$$\epsilon_{t,i} = \frac{\eta \pi_{t,i} (\hat{\ell}_{t,i} - \lambda_t)^2}{1 + \eta \pi_{t,i} (\hat{\ell}_{t,i} - \lambda_t)} , \qquad |\mathbb{E}_t[\epsilon_{t,i}]| = O(\eta) .$$

Since Prod is exactly OMD over the sequence $\hat{\ell}_t - \epsilon_t$, we can decompose the regret as follows

$$\mathbb{E}\left[\sum_{t=1} \left\langle \pi_t - u, \hat{\ell}_t \right\rangle \right] = \mathbb{E}\left[\sum_{t=1} \left\langle \pi_t - u, \hat{\ell}_t - \epsilon_t \right\rangle \right] + \mathbb{E}\left[\sum_{t=1} \langle \pi_t - u, \epsilon_t \rangle \right] = \text{Reg}_{\text{OMD}} + O(\eta T) .$$

The analysis is not entirely straightforward as the loss range for the OMD update becomes $[-1 - O(\eta), 1 + O(\eta)]$ because of the shift introduced by the perturbation of the losses, and this posses some additional difficulties.

5 Best of both worlds algorithms

In applications where the loss is potentially more benign, for example sampled i.i.d. from a a fixed distribution over $[0,1]^K$, it is desirable to obtain faster rates in nice environments while preserving worst-case guarantees. Probably the simplest algorithm with this property is Tsallis-INF [Zimmert and Seldin, 2021], which is FTRL with 1/2-Tsallis entropy and $\eta_t \propto 1/\sqrt{t}$ learning rate.

5.1 TS-Prod

Recall the 1/2-Tsallis regularizer is $F(x) = -\sum_{i=1}^{K} 2\sqrt{x_i}$. Unlike OMD, FTRL is not canonically expressed as a 1-step update of the previous policy. Instead, the 1/2-Tsallis-INF policy is given with a normalization constant Λ_t (see [Zimmert and Seldin, 2021])

$$\pi_{t+1,i}^{\text{FTRL}} = \left(\Lambda_{t+1} + \eta_t \sum_{s=1}^{t} \hat{\ell}_{s,i} \right)^{-2} .$$

https://doi.org/10.52202/079017-2710

Recursively using this expression yields

$$\pi^{\text{FTRL}}_{t+1,i} = \left(\Lambda_{t+1} + \frac{\eta_t}{\eta_{t-1}}\left(\frac{1}{\sqrt{\pi^{\text{FTRL}}_{t,i}}} - \Lambda_t\right) + \eta_t \hat{\ell}_t\right)^{-2} = \pi^{\text{FTRL}}_{t,i}\left(1 + \eta_t\sqrt{\pi^{\text{FTRL}}_{t,i}}\left(\hat{\ell}_t - \frac{\eta_t \xi_t}{\sqrt{\pi^{\text{FTRL}}_{t,i}}} - \lambda_t\right)\right)^{-2},$$

where $\xi_t = \frac{1}{\eta_t^2} - \frac{1}{\eta_t \eta_{t-1}}$ and $\lambda_t = \Lambda_{t+1} - \frac{\eta_t}{\eta_{t-1}}\Lambda_t$. The first order approximation is

$$\pi^{\text{FTRL}}_{t+1,i} \underbrace{\approx}_{\text{1st-order}} \pi^{\text{FTRL}}_{t,i}\left(1 - 2\eta_t\sqrt{\pi^{\text{FTRL}}_{t,i}}\left(\hat{\ell}_{t,i} - \xi_t/\sqrt{\pi^{\text{FTRL}}_{t,i}} - \lambda_t\right)\right).$$

We ensure following the approximation in expectation by directly biasing the losses with $\eta_t \xi_t/\sqrt{\pi_{t,i}}$. Additionally, one needs to perform a second-order correction as discussed in Section 3.1. We omit a formal derivation, but notice that WSU-UX required $\eta/\pi_{t,i}$ correction, while LB-Prod works without correction because the error is of order η. As the intermediate potential between KL and Logbarrier, Tsallis-INF turns out to require a correction of order $\eta/\sqrt{\pi_{t,i}}$, which we tighten by an additional factor of $(1 - \pi_{t,i})$ necessary to ensure stochastic bounds.

With this, we are ready to present

$$\hat{\ell}_{t,i} = \left(\ell_{t,i} - \frac{\eta_t(\xi_t + \gamma(1-\pi_{t,i}))}{\sqrt{\pi_{t,i}}}\right)\frac{\mathbb{I}(A_t = i)}{\pi_{t,i}}, \quad \lambda_t = \sum_{i=1}^{K} \frac{\pi_{t,i}^{\frac{3}{2}}\hat{\ell}_{t,i}}{\sum_{j=1}^{K}\pi_{t,j}^{\frac{3}{2}}}, \quad \xi_t = \frac{1}{\eta_t^2} - \frac{1}{\eta_t\eta_{t-1}},$$

$$\pi_{t+1,i} = \pi_{t,i}\left(1 - 2\eta_t\sqrt{\pi_{t,i}}(\hat{\ell}_{t,i} - \lambda_t)\right). \tag{TS-Prod}$$

Theorem 3. *The regret of TS-Prod with $\eta_t = \frac{1}{\sqrt{K+26t}}$, $\gamma = \frac{13}{2}$ is bounded by $O(\sqrt{KT} + K\log(T))$ in the adversarial setting and by $O\left(\sum_{i\neq i^\star}\frac{\log(T)}{\Delta_i}\right)$ in the stochastic setting.*

5.2 Analysis of TS-Prod

We first show that the loss biasing is sufficient to ensure that the distribution is well defined.

Lemma 5. *If ξ_t is a non-increasing sequence, $\eta_t < \frac{2}{\sqrt{K(\xi_t+\gamma)^2}}$ and $\eta_{t+1}^2 \leq \eta_t^2(1 - 2\gamma\eta_t^2)$ for all t, then the update rule of TS-Prod is proper and satisfies $\pi_{t,i} > (\xi_t + \gamma)^2\eta_t^2$ for any arm and loss sequence at all time steps.*

Next we present the moving parts of the analysis.

$$\mathbb{E}[\sum_{t=1}^{T}\langle \pi_t - u, \ell_t\rangle] = \sum_{t=1}^{T}\mathbb{E}\Bigg[\underbrace{\left\langle \pi_t - u, \ell_t - \hat{\ell}_t\right\rangle}_{\text{change of measure}} + \underbrace{\frac{D_{TS}(u,\pi_t) - D_{TS}(u,\pi_{t+1})}{\eta_t}}_{\text{proto-penalty}} + \underbrace{\left\langle \pi_t - u, \hat{\ell}_t\right\rangle - \frac{D_{TS}(u,\pi_t) - D_{TS}(u,\pi_{t+1})}{\eta_t}}_{\text{proto-stability}}\Bigg]$$

The change of measure is by construction

$$\text{change-of-measure} = \sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{K}\frac{\pi_{t,i} - u_i}{\sqrt{\pi_{t,i}}}\left(\eta_t\xi_t + \eta_t\gamma(1-\pi_{t,i})\right)\right]. \tag{3}$$

We now bound the stability and penalty.

Lemma 6. *For any time t such that $\pi_{t,i} > (\xi_t + \gamma)^2\eta_t^2$, it holds*

$$\left\langle \pi_t - u, \mathbb{E}_t[\hat{\ell}_t]\right\rangle + \mathbb{E}_t\left[\frac{D_{TS}(u,\pi_{t+1})}{\eta_t}\right] - \frac{D_{TS}(u,\pi_t)}{\eta_t} \leq \frac{13}{2}\frac{\eta_t u_i}{\sqrt{\pi_{ti}}}(1 - \pi_{ti}).$$

The tuning of Theorem 3 satisfies the conditions.

Lemma 7. *The proto-penalty is bounded by*

$$\sum_{t=1}^{T} \frac{D_{TS}(u, \pi_t) - D_{TS}(u, \pi_{t+1})}{\eta_t} \leq \sum_{t=1}^{T} \eta_t \xi_t \left(\sum_{i \neq i^\star} 2\sqrt{\pi_{ti}} + \sum_{i=1}^{K} \frac{u_i - \pi_{ti}}{\sqrt{\pi_{ti}}} \right)$$

We are ready to prove the main regret guarantee.

Proof of Theorem 3. By Lemma 5 we have a proper update rule. Using Lemma 6, equation (3), where we tuned $\gamma = \frac{13}{2}$ and Lemma 7 yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \pi_t - u, \ell_t \rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \eta_t \left(\sum_{i=1}^{K} \frac{13}{2} \sqrt{\pi_{t,i}}(1 - \pi_{ti}) + \sum_{i \neq i^\star} \xi_t \sqrt{\pi_{t,i}} \right)\right] .$$

The adversarial regret follows from $\sum_{i=1}^{K} \sqrt{\pi_{t,i}} \leq \sqrt{K}$, $\xi_t \leq 4$, $\sum_{t=1}^{T} \eta_t = O(\sqrt{T})$. For the stochastic regret, the proof follows standard arguments using the self-bounding trick as in Zimmert and Seldin [2021]. For details on the self-bounding trick see the supplementary material. □

5.3 TS-Prod and stabilized OMD

Even though we derived TS-Prod from FTRL, it turns out that one can also interpret the update as an approximation of *stabilized* OMD proposed by Fang et al. [2022]. In Appendix E, we formalize this connection and present a slight variation of TS-Prod. We then analyse this variant via the perturbation technique described in Section 4.3. Our analysis also shows that the stabilized OMD algorithm induced by the $1/2$-Tsallis entropy enjoys best-of-both worlds regret guarantees which to the best of our knowledge is novel.

6 Discussion

We have provided an extensive study of incentive-compatible bandits. We have negatively resolved an open question of whether incentive-compatibility as defined in Freeman et al. [2020] is harder than regular bandits. Using linear approximations, partly with second order corrections, allows to recover results from well studied algorithms in the literature. We even obtain an algorithm with best-of-both-world guarantees. Our algorithms are conceptually simpler than existing bandit algorithms, they update the probability distributions with basic arithmetic operations without the need to solve optimization problems.

Our successes make it likely that one can transfer even more sophisticated methods, such as first-order, second-order, path-norm bounds and online learning with graph feedback to this framework. We leave this investigation to future work.

References

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pages 263–274. Citeseer, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 2009.

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1), 2002b.

 https://doi.org/10.52202/079017-2710

Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

Eyal Even-Dar, Michael Kearns, Yishay Mansour, and Jennifer Wortman. Regret to the best vs. regret to the average. *Machine Learning*, 72(1):21–37, 2008.

Huang Fang, Nicholas JA Harvey, Victor S Portella, and Michael P Friedlander. Online mirror descent and dual averaging: keeping pace in the dynamic case. *The Journal of Machine Learning Research*, 23(1):5271–5308, 2022.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Rupert Freeman, David Pennock, Chara Podimata, and Jennifer Wortman Vaughan. No-regret and incentive-compatible online learning. In *International Conference on Machine Learning*, pages 3270–3279. PMLR, 2020.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

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

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems*, 27, 2014.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1), 1985.

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

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533, 2020a.

Chung-Wei Lee, Haipeng Luo, and Mengxiao Zhang. A closer look at small-loss bounds for bandits with graph feedback. In *Conference on Learning Theory*, pages 2516–2564. PMLR, 2020b.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

Ali Mortazavi, Junhao Lin, and Nishant Mehta. On the price of exact truthfulness in incentive-compatible online learning with bandit feedback: a regret lower bound for wsu-ux. In *International Conference on Artificial Intelligence and Statistics*, pages 4681–4689. PMLR, 2024.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2), 2012.

William R Thompson. On the likelihood that one unknown probability exceeds another in view ooster the evidence of two samples. *Biometrika*, 25(3/4), 1933.

Vladimir G Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60, 1995.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference on Learning Theory*, 2018.

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

Contents

 https://doi.org/10.52202/079017-2710

A Background on FTRL and OMD

We give a brief overview of the template analysis for FTRL and OMD. For an extensive discussion and regret analysis of these two frameworks we refer the interested readers to Chapter 2 of Shalev-Shwartz [2012], Chapter 5 of Hazan et al. [2016] or Chapter 28 of Lattimore and Szepesvári [2020].

A.1 OMD analysis overview

The OMD update

$$\pi_{t+1} = \arg\min_{\pi \in \Delta([K])} \left\langle \pi, \eta_t \hat{\ell}_t \right\rangle - D_F(\pi, \pi_t) \tag{OMD}$$

can be written in two steps as

$$\begin{aligned}\tilde{\pi}_{t+1} &= \arg\min_{\pi \in \mathbb{R}^K} \left\langle \pi, \eta_t \hat{\ell}_t \right\rangle - D_F(\pi, \pi_t) \\ \pi_{t+1} &= \arg\min_{\pi \in \Delta([K])} D_F(\pi, \tilde{\pi}_{t+1}),\end{aligned}$$

where the first step is an unconstrained optimization over the linear loss at time t together with a regularization term given by the Bregman divergence induced by F, and the second step is the Bregman projection onto the probability simplex. Thie first step of the OMD update can be re-written as

$$\nabla F(\tilde{\pi}_{t+1}) = \nabla F(\pi_t) - \eta_t \hat{\ell}_t,$$

Let $F_t = \frac{1}{\eta_t}$. Since π_{t+1} is the minimizer of the OMD update, and $D_F w$ is convex we have that

$$\begin{aligned}\left\langle \pi_{t+1} - \pi, \hat{\ell}_t \right\rangle &\leq \langle u - \pi_{t+1}, \nabla F_t(\pi_{t+1}) - \nabla F_t(\pi_t) \rangle \\ &= D_{F_t}(u, \pi_t) - D_{F_t}(u, \pi_{t+1}) - D_{F_t}(\pi_{t+1}, \tilde{\pi}_t).\end{aligned}$$

Further, it holds that

$$\begin{aligned}\left\langle \pi_t - \pi_{t+1}, \hat{\ell}_t \right\rangle &= D_{F_t}(\pi_{t+1}, \pi_t) + D_{F_t}(\pi_t, \tilde{\pi}_{t+1}) - D_{F_t}(\tilde{\pi}_{t+1}, \pi_{t+1}) \\ &\leq D_{F_t}(\pi_{t+1}, \pi_t) + D_{F_t}(\pi_t, \tilde{\pi}_{t+1}).\end{aligned}$$

Combining the two inequalities together we have that one step of the regret to any $u \in \Delta([K])$ is bounded as

$$\left\langle \hat{\ell}_t, \pi_t - u \right\rangle \leq D_{F_t}(u, \pi_t) - D_{F_t}(u, \pi_{t+1}) + D_{F_t}(\pi_t, \tilde{\pi}_{t+1}).$$

For a fixed step-size $\eta_t = \eta$ the above telescopes to bound the regret as

$$\sum_{t=1}^{T} \left\langle \hat{\ell}_t, \pi_t - u \right\rangle \leq \frac{D_F(u, \pi_1) - D_F(u, \pi_T)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T-1} D_F(\pi_t, \tilde{\pi}_{t+1}).$$

As long as F is twice differentiable, each of the terms can be bounded as $D_F(\pi_t, \tilde{\pi}_{t+1}) \leq O(\eta^2 \|\hat{\ell}_t\|^2_{\nabla^2(F^*)(w_t)})$, where F^* is the Fenchel conjugate of F. Controlling $\|\hat{\ell}_t\|^2_{\nabla^2(F^*)(w_t)}$ in OCO is usually done by assuming some boundedness of the losses. In bandit literature controlling this term is slightly more involved and depends on the choice of F.

When η is not constant, telescoping the above sum does not work and the analysis becomes much more involved. It is possible to construct sequences of losses for which the OMD update does not enjoy sub-linear regret for $\eta_t = \frac{1}{\sqrt{t}}$. Fang et al. [2022] introduce a stabilization term to the OMD update which overcomes this problem and show that this new version does enjoy the standard OMD regret guarantees.

A.2 FTRL analysis overview

The FTRL analysis follows similar ideas, however, the one step regret is bounded as

$$\left\langle \hat{\ell}_t, \pi_t - u \right\rangle \leq (F_t + I_{\Delta([K])})^*(-\hat{L}_{t-1}) - (F_t + I_{\Delta([K])})^*(-\hat{L}_t) + \frac{1}{\eta_t} D_{F^*}(-\hat{L}_t, -\hat{L}_{t-1}),$$

where $\hat{L}_t = \sum_{s=1}^{t-1} \hat{\ell}_t$, and $I_{\Delta[K]}$ is the characteristic function of $\Delta_{[K]}$, i.e., $I_{\Delta[K]}(\pi) = 0$ if $\pi \in \Delta[K]$ and $I_{\Delta[K]}(\pi) = +\infty$ otherwise. The term $D_{F^*}(-\hat{L}_t, -\hat{L}_{t-1})$ can be thought of as the equivalent to $D_F(\pi_t, \tilde{\pi}_{t+1})$ in the OMD analysis. The term $(F_t + I_{\Delta([K])})^*(-\hat{L}_{t-1}) - (F_t + I_{\Delta([K])})^*(-\hat{L}_t)$ needs to be telescoped in an appropriate way. For more details we refer the reader to the penalty term bound of Zimmert and Seldin [2021].

B Missing Proof Section 3

Proof of Theorem 1. WLOG we assume that $T \geq K$. We begin with the bound from Lemma 2. The second and third term in the RHS of the inequality are bounded as is standard in the Exp3 analysis

$$\eta \sum_{t=1}^{T} \sum_{i=1}^{K} \pi_{t,i} \, \mathbb{E}[\hat{\ell}_{t,i}^2 | \mathcal{F}_{t-1}] \leq \frac{\eta T K}{1-\gamma} \leq 2\eta T K, \qquad \eta \sum_{t=1}^{T} \mathbb{E}[\hat{\ell}_{t,i^*}^2] \leq \eta \sum_{t=1}^{T} \mathbb{E}\left[\frac{\tilde{\ell}_{t,i^*}^2}{\tilde{\pi}_{t,i^*}}\right] \leq \eta \sum_{t=1}^{T} \mathbb{E}\left[\frac{\ell_{t,i^*}^2}{\tilde{\pi}_{t,i^*}}\right],$$

where $\mathcal{F}_{t-1}$ is the filtration generated by the random play and randomness of the losses up to time $t-1$. We now consider the expectation of the LHS which evaluates to

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[\pi_{t,i} \hat{\ell}_{t,i}] - \sum_{t=1}^{T} \mathbb{E}[\hat{\ell}_{t,i^*}] &= \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[\pi_{t,i} \tilde{\ell}_{t,i}] - \sum_{t=1}^{T} \mathbb{E}[\tilde{\ell}_{t,i^*}] = \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[\pi_{t,i} \ell_{t,i}] - \sum_{t=1}^{T} \mathbb{E}[\ell_{t,i^*}] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}\left[\frac{\eta \ell_{t,i^*}}{\tilde{\pi}_{t,i}}\right] - \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}\left[\frac{\eta \pi_{t,i} \ell_{t,i}}{\tilde{\pi}_{t,i}}\right] \geq \eta \sum_{t=1}^{T} \mathbb{E}\left[\frac{\ell_{t,i^*}^2}{\tilde{\pi}_{t,i}}\right] - 2\eta T K \\
&\quad + \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[\pi_{t,i} \ell_{t,i}] - \sum_{t=1}^{T} \mathbb{E}[\ell_{t,i^*}].
\end{aligned}$$

Thus combining the bounds on the LHS and RHS we have

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[\pi_{t,i} \ell_{t,i}] - \sum_{t=1}^{T} \mathbb{E}[\ell_{t,i^*}] \leq \frac{\log(K)}{\eta} + 4\eta T K + \eta \sum_{t=1}^{T} \mathbb{E}\left[\frac{\ell_{t,i^*}^2}{\tilde{\pi}_{t,i}}\right] - \eta \sum_{t=1}^{T} \mathbb{E}\left[\frac{\ell_{t,i^*}^2}{\tilde{\pi}_{t,i}}\right].$$

To complete the proof we only note that $\sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[\pi_{t,i} \ell_{t,i}] - \sum_{t=1}^{T} \mathbb{E}[\ell_{t,A_t}] \leq 2T\gamma = \eta K T$. □

C Missing Proofs Section 4

Proof of Lemma 3. The expectation of $\tilde{\ell}_{ti} = \ell_{ti} \mathbb{I}(A_t = i)$ is obviously $\pi_{ti} \ell_{ti}$, hence by the definition of λ_{ti}, we have

$$\mathbb{E}_t[\tilde{\ell}_{ti} - \lambda_{ti}] = \pi_{ti} \left(\ell_{ti} - \frac{\sum_{j=1}^{K} \pi_{tj}^2 \ell_{tj}}{\sum_{j=1}^{K} \pi_{tj}^2} \right).$$

For the second part, we have

$$\mathbb{E}_t[(\tilde{\ell}_{ti} - \lambda_{ti})^2] \leq \mathbb{E}_t[\tilde{\ell}_{ti}^2] + \mathbb{E}_t[\lambda_{ti}^2] = \pi_{ti} \left(\ell_{ti}^2 + \pi_{ti}^2 \frac{\sum_{j=1}^{K} \pi_{tj}^3 \ell_{tj}^2}{\left(\sum_{k=1}^{K} \pi_{tk}^2\right)^2} \right) \leq \pi_{ti} \left(1 + \frac{\pi_{ti} \sum_{j=1}^{K} \pi_{tj}^3}{\left(\sum_{k=1}^{K} \pi_{tk}^2\right)^2} \right).$$

The proof is completed by noting

$$\frac{\pi_{ti} \sum_{j=1}^{K} \pi_{tj}^3}{\left(\sum_{k=1}^{K} \pi_{tk}^2\right)^2} \leq \frac{\pi_{ti} \sum_{j=1}^{K} \pi_{tj}^3}{\left(\sum_{k=1}^{K} \pi_{tk}^3\right)^{\frac{4}{3}}} = \frac{\pi_{ti}}{\left(\sum_{k=1}^{K} \pi_{tk}^3\right)^{\frac{1}{3}}} \leq 1.$$

□

https://doi.org/10.52202/079017-2710

Proof of Lemma 4.

$$
\begin{aligned}
&\langle \pi_t - u, \ell_t \rangle + \mathbb{E}_t \left[\eta^{-1} D_{LB}(u, \pi_{t+1}) \right] - \eta^{-1} D_{LB}(u, \pi_t) \\
&= \langle \pi_t - u, \ell_t \rangle + \mathbb{E}_t \left[\sum_{i=1}^{K} \frac{u_i - \pi_{t+1,i}}{\eta \pi_{t+1,i}} - \frac{u_i - \pi_{t,i}}{\eta \pi_{t,i}} + \frac{1}{\eta} \log \left(\frac{\pi_{t+1,i}}{\pi_{t,i}} \right) \right] \\
&= \langle \pi_t - u, \ell_t \rangle + \mathbb{E}_t \left[\sum_{i=1}^{K} \frac{u_i \left(1 - \frac{\pi_{t+1,i}}{\pi_{t,i}} \right)}{\eta \pi_{t+1,i}} + \frac{1}{\eta} \log \left(1 - \eta(\tilde{\ell}_{t,i} - \lambda_{t,i}) \right) \right] \\
&\leq \langle \pi_t - u, \ell_t \rangle + \mathbb{E}_t \left[\sum_{i=1}^{K} \frac{u_i (\tilde{\ell}_{t,i} - \lambda_{t,i})}{\pi_{t,i}(1 - \eta(\tilde{\ell}_{t,i} - \lambda_{t,i}))} - \tilde{\ell}_{t,i} + \lambda_{t,i} \right] && (\log(1+x) \leq x) \\
&\leq \langle \pi_t - u, \ell_t \rangle + \sum_{i=1}^{K} \left(\frac{u_i}{\pi_{t,i}} - 1 \right) \mathbb{E}_t[\tilde{\ell}_{t,i} - \lambda_{t,i}] + \eta \sum_{i=1}^{K} \frac{u_i}{\pi_{t,i}} \mathbb{E}_t \left[\frac{(\tilde{\ell}_{t,i} - \lambda_{t,i})^2}{1 - \eta} \right] \\
&\leq \langle \pi_t - u, \ell_t \rangle + \sum_{i=1}^{K} (u_i - \pi_{t,i})(\ell_{t,i} - c_t) + \frac{2\eta}{1-\eta} \sum_{i=1}^{K} u_i = \frac{2\eta}{1-\eta} . && \text{(Lemma 3)}
\end{aligned}
$$

□

D Missing Proofs Section 5.2

D.1 Technical Lemmas

Lemma 8.

$$
\min_{x \in [0,1]} f(x) = \min_{x \in [0,1]} \frac{x^3}{1-x} + \sqrt{1-x} \geq \sqrt{\frac{8}{9}} .
$$

Proof. We first show that the optimal point is smaller than $\frac{1}{9}$, by looking at the derivative

$$
f'(x) = \frac{3\sqrt{x} - x^{\frac{3}{2}} - (1-x)^{\frac{3}{2}}}{2(1-x)^2} .
$$

For the enumerator, we have for all $x \geq \frac{1}{9}$:

$$
3\sqrt{x} - x^{\frac{3}{2}} - (1-x)^{\frac{3}{2}} > 3\sqrt{x} - \max\{\sqrt{x}, \sqrt{1-x}\} \qquad \geq \min\{2\sqrt{x}, 3\sqrt{x} - 1\} \geq 0
$$

Hence

$$
\min_{x \in [0,1]} f(x) = \min_{x \in [0,1/9]} f(x) > \min_{x \in [0,1/9]} \sqrt{1-x} = \sqrt{\frac{8}{9}} .
$$

□

Lemma 9. *For any $a, b \geq 0$ such that $a + b \geq 1$, it holds*

$$
\frac{a}{\sqrt{b}} + \frac{b}{\sqrt{a}} \geq \sqrt{2} .
$$

Proof. We can assume w.l.o.g. that $a = 1 - b$, otherwise scale both a and b down and reduce the objective. The resulting problem is symmetric with $a = \frac{1}{2}$ as the unique minimizer resulting in the statement. □

D.2 Minimal probability

Lemma 10. *Assume that $\pi_{ti} > (\xi_t+\gamma)^2\eta_t^2$ holds for all arms, then*

$$\mathbb{E}_t\left[(\hat{\ell}_{ti}-\lambda_t)^2\right] \le \frac{13(1-\pi_{ti})}{8\pi_{t,i}}\,.$$

Proof. Let $\tilde{\ell}_{t,i} = \pi_{t,i}\hat{\ell}_{t,i}$ and $\lambda_{t,i} = \pi_{t,i}\lambda_t$. Note that $\tilde{\ell}_{ti} \in [-1,1]$ by the condition on π_{ti} and $\tilde{\ell}_{ti} = 0$ for $i \neq I_t$ by construction of the loss estimate. Hence

$$\begin{aligned}
\mathbb{E}_t\left[(\tilde{\ell}_{ti}-\lambda_{ti})^2\right] &\le \pi_{ti}\left(1-\frac{\pi_{ti}^{\frac{3}{2}}}{\sum_{j=1}^K \pi_{tj}^{\frac{3}{2}}}\right)^2 \mathbb{E}[\tilde{\ell}_{ti}^2|i=I_t] + \sum_{j\neq i}\pi_{tj}\left(\frac{\pi_{ti}\sqrt{\pi_{tj}}}{\sum_{j=1}^K \pi_{tj}^{\frac{3}{2}}}\right)^2 \mathbb{E}[\tilde{\ell}_{tj}^2|j=I_t] \\
&\le \pi_{ti}\left(\frac{(\sum_{j\neq i}\pi_{tj}^{\frac{3}{2}})^2 + \pi_{ti}\sum_{j\neq i}\pi_{tj}^2}{(\sum_{j=1}^K \pi_{tj}^{\frac{3}{2}})^2}\right) \\
&= \pi_{ti}(1-\pi_{ti})\left(\frac{(1-\pi_{ti})^2(\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^2 + \pi_{ti}(1-\pi_{ti})\sum_{j\neq i}\tilde{\pi}_{tj}^2}{(\pi_{ti}^{\frac{3}{2}} + (1-\pi_{ti})^{\frac{3}{2}}\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^2}\right),
\end{aligned}$$

where $\tilde{\pi}_{tj} = \frac{\pi_{tj}}{1-\pi_{ti}}$. We bound the two terms in the bracket separately, for the first term we have

$$\begin{aligned}
\left(\frac{(1-\pi_{ti})\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}}}{\pi_{ti}^{\frac{3}{2}} + (1-\pi_{ti})^{\frac{3}{2}}\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}}}\right)^2 &\le \left(\frac{(1-\pi_{ti})}{\pi_{ti}^{\frac{3}{2}} + (1-\pi_{ti})^{\frac{3}{2}}}\right)^2 && (\textstyle\sum_{j\neq i}\tilde{\pi}_{tj} = 1) \\
&\le \frac{9}{8} && \text{(Lemma 8)}
\end{aligned}$$

The second term is

$$\begin{aligned}
\frac{\pi_{ti}(1-\pi_{ti})\sum_{j\neq i}\tilde{\pi}_{tj}^2}{(\pi_{ti}^{\frac{3}{2}} + (1-\pi_{ti})^{\frac{3}{2}}\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^2} &\le \frac{\pi_{ti}(1-\pi_{ti})(\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^{\frac{4}{3}}}{(\pi_{ti}^{\frac{3}{2}} + (1-\pi_{ti})^{\frac{3}{2}}\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^2} \\
&= \left(\frac{\pi_{ti}}{\sqrt{1-\pi_{ti}}(\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^{\frac{2}{3}}} + \frac{(1-\pi_{ti})(\sum_{j\neq i}\tilde{\pi}_{tj}^{\frac{3}{2}})^{\frac{1}{3}}}{\sqrt{\pi_{ti}}}\right)^{-2} \\
&\le \frac{1}{2} && \text{(Lemma 9)}
\end{aligned}$$

□

Lemma 11 (Lemma 5)**.** *If ξ_t is a non-increasing sequence, $\eta_t < \frac{2}{\sqrt{K(\xi_t+\gamma)^2}}$ and $\eta_{t+1}^2 \le \eta_t^2(1-2\gamma\eta_t^2)$ for all t, then the update rule of TS-Prod is well defined and satisfies $\pi_{t,i} > (\xi_t+\gamma)^2\eta_t^2$ for any arm and loss sequence at all time steps.*

Proof. The proof follows by induction. At $t=1$ the statement is true by definition. Let the claim hold at time t, then the probability of an arm only decreases when $\hat{\ell}_{t,i} - \lambda_t$ is positive. We look at the cases where $A_t = i$ and $A_t \neq i$ independently.

Case $A_t = i$:

$$\begin{aligned}
\pi_{t+1,i} &> \pi_{t,i}(1-2\eta_t\sqrt{\pi_{t,i}}\hat{\ell}_{t,i}) = \pi_{t,i} - 2\eta_t\left(\sqrt{\pi_{t,i}}\ell_{t,i} - \eta_t(\xi_t+\gamma(1-\pi_{t,i}))\right) \\
&> \pi_{t,i}(1-2\gamma\eta_t^2) - 2\eta_t\sqrt{\pi_{t,i}} + 2\gamma\eta_t^2 \\
&= (1-2\gamma\eta_t^2)\left(\sqrt{\pi_{t,i}} - \frac{\eta_t}{1-2\gamma\eta_t^2}\right)^2 + (2\gamma - \frac{1}{1-2\gamma\eta_t^2})\eta_t^2
\end{aligned}$$

https://doi.org/10.52202/079017-2710

This is a quadratic function in $\pi_{t,i}$ with minimizer $\frac{\eta_t^2}{(1-2\gamma\eta_t^2)^2} < (\xi_t+\gamma)^2\eta_t^2$, hence the value is lower bounded by setting $\pi_{t,i}$ to $(\xi_t+\gamma)^2\eta_t^2$

$$\pi_{t+1,i} > (\xi_t+\gamma)^2\eta_t^2(1-2\gamma\eta_t^2) \geq (\xi_t+\gamma)^2\eta_{t+1}^2 \geq (\xi_{t+1}+\gamma)^2\eta_{t+1}^2$$

Case $A_t \neq i$:

$$\begin{aligned}\pi_{t+1,i} &= \pi_{t,i} - 2\eta_t\pi_{t,i}^{\frac{3}{2}}\left(\frac{\eta_t((\xi_t+\gamma(1-\pi_{t,A_t})) - \ell_{t,A_t}\sqrt{\pi_{t,A_t}}}{\sum_{j=1}^K \pi_{tj}^{\frac{3}{2}}}\right)\\ &> \pi_{t,i} - 2(\xi_t+\gamma)\eta_t^2\pi_{t,i}^{\frac{3}{2}}\sqrt{K}\end{aligned}$$

This is a concave function (in $\pi_{t,i}$) so the minimizer is at either at $\pi_{t,i} = (\xi_t+\gamma)^2\eta_t^2$ or at $\pi_{t,i} = 1$. For the latter, we have have $\pi_{t+1,i} > \frac{1}{2}$, so the minimum is obtained for the first case.

$$\pi_{t+1,i} > (\xi_t+\gamma)^2\eta_t^2 - 2(\xi_t+\gamma)^4\eta_t^5\sqrt{K} > (\xi_t+\gamma)^2\eta_t^2(1-2\gamma\eta_t^2) \geq C_{t+1}^2\eta_{t+1}^2\,.$$

□

Lemma 12 (Lemma 6)**.** *For any time t such that $\pi_{t,i} > (\xi_t+\gamma)^2\eta_t^2$, it holds*

$$\langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \mathbb{E}_t\left[\frac{D_{TS}(u,\pi_{t+1})}{\eta_t}\right] - \frac{D_{TS}(u,\pi_t)}{\eta_t} \leq \sum_{i=1}^K\left(2\eta_t\sqrt{\pi_{t,i}}(1-\pi_{t,i}) - \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\frac{u_i - \pi_{t,i}}{\sqrt{\pi_{t,i}}}\right)\,.$$

Proof of Lemma 6.

$$\begin{aligned}&\langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \mathbb{E}_t\left[\eta_t^{-1}D_{TS}(u,\pi_{t+1})\right] - \eta_t^{-1}D_{TS}(u,\pi_t)\\
&= \langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \mathbb{E}_t\left[\sum_{i=1}^K \frac{u_i - \pi_{t+1,i}}{\eta_t\sqrt{\pi_{t+1,i}}} - \frac{u_i - \pi_{ti}}{\eta_t\sqrt{\pi_{ti}}} + \frac{1}{\eta_t}\left(2\sqrt{\pi_{t+1,i}} - 2\sqrt{\pi_{ti}}\right)\right]\\
&= \langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \sum_{i=1}^K\left(\frac{u_i}{\eta_t\sqrt{\pi_{ti}}}\mathbb{E}_t\left[\sqrt{\frac{\pi_{ti}}{\pi_{t+1,i}}} - 1\right] + \frac{\pi_{ti}}{\eta_t\sqrt{\pi_{ti}}}\mathbb{E}_t\left[\sqrt{\frac{\pi_{t+1,i}}{\pi_{ti}}} - 1\right]\right)\\
&= \langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \sum_{i=1}^K\left(\frac{u_i}{\eta_t\sqrt{\pi_{ti}}}\mathbb{E}_t\left[\sqrt{1 + \frac{2\eta_t\sqrt{\pi_{ti}}(\hat\ell_{ti} - \lambda_t)}{1 - 2\eta_t\sqrt{\pi_{ti}}(\hat\ell_{ti} - \lambda_t)}} - 1\right]\right.\\
&\qquad\left. + \frac{\pi_{ti}}{\eta_t\sqrt{\pi_{ti}}}\mathbb{E}_t\left[\sqrt{1 - 2\eta_t\sqrt{\pi_{ti}}(\hat\ell_{ti} - \lambda_t)} - 1\right]\right)\\
&\leq \langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \sum_{i=1}^K\left(\frac{u_i}{\eta_t\sqrt{\pi_{ti}}}\mathbb{E}_t\left[\eta_t\sqrt{\pi_{ti}}(\hat\ell_{ti} - \lambda_t) + \frac{2\eta_t^2\pi_{ti}(\hat\ell_{ti} - \lambda_t)^2}{1 - 2\eta_t\sqrt{\pi_{ti}}(\hat\ell_{ti} - \lambda_t)}\right]\right.\\
&\qquad\left. + \frac{\pi_{ti}}{\eta_t\sqrt{\pi_{ti}}}\mathbb{E}_t\left[-\eta_t\sqrt{\pi_{ti}}(\hat\ell_{ti} - \lambda_t)\right]\right)\\
&\leq \langle \pi_t - u, \mathbb{E}_t[\ell_t]\rangle + \sum_{i=1}^K\left((u_i - \pi_{ti})\,\mathbb{E}_t\left[\hat\ell_{ti} - \lambda_t\right] + 4\eta_t u_{ti}\sqrt{\pi_{ti}}\,\mathbb{E}_t\left[(\hat\ell_{ti} - \lambda_t)^2\right]\right)\\
&\hspace{20em}\text{(Setting } (\xi_t+\gamma) \geq 4\text{)}\\
&\leq \frac{13}{2}\frac{\eta_t u_i}{\sqrt{\pi_{ti}}}(1 - \pi_{ti})\,. \hspace{14em}\text{(Lemma 10)}\end{aligned}$$

We now show that the requirements of Lemma 5 are satisfied with the tuning of Theorem 3. With $\eta_t = \frac{1}{\sqrt{K+26t}}$, we have $c_t = \left(K + 26t - \sqrt{(K+26t)(K+26t-26)}\right) > 2$, which is monotonically decreasing. $c_t > 2$ and $\gamma = \frac{13}{2}$ ensures that $\eta_t = \frac{1}{\sqrt{K+26t}} < \frac{2}{\sqrt{K(c_t+\gamma)^2}}$. Further we have

$$\frac{\eta_{t+1}^2}{\eta_t^2} = \frac{K+26t}{K+26(t+1)} = 1 - \frac{26}{K+26(t+1)} \leq 1 - \frac{4}{K+26t} = 1 - 4\eta_t^2\,.$$

□

Proof of Lemma 7. Using $\eta_0 = \infty$ and $D_{TS}(u, \Pi_{T+1}) \geq 0$

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{D_{TS}(u,\pi_t) - D_{TS}(u,\pi_{t+1})}{\eta_t} \leq \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) D_{TS}(u,\pi_t) \\
&= \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) (F(u) - F(\pi_t) - \langle u - \pi_t, \nabla F(\pi_t)\rangle) \\
&= \sum_{t=1}^{T} \sum_{i=1}^{K} \eta_t \xi_t \left(2(\sqrt{\pi_{t,i}} - \sqrt{u_{t,i}}) - (u_i - \pi_{ti}) \frac{1}{\sqrt{\pi_{ti}}}\right) \\
&\leq \sum_{t=1}^{T} \eta_t \xi_t \left(\sum_{i \neq i^\star} 2\sqrt{\pi_{t,i}} - \sum_{i=1}^{K} (u_i - \pi_{ti}) \frac{1}{\sqrt{\pi_{ti}}}\right),
\end{aligned}$$

where the last inequality follows from $\sum_{i=1}^{K} \sqrt{u_i} \geq 1 \geq \sqrt{\pi_{t,i^\star}}$. □

D.3 Self-bounding trick

We now quickly describe how to apply the self-bounding trick from Zimmert and Seldin [2021]. Assume that we have a regret bound of the form

$$\sum_{t=1}^{T} \sum_{i \neq i^*} \pi_{t,i} \Delta_i \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{a\pi_{t,i} + b\sqrt{\pi_{t,i}}}{\sqrt{t}},$$

for some positive a and b. The above inequality implies

$$\frac{1}{3} \sum_{t=1}^{T} \sum_{i \neq i^*} \pi_{t,i} \Delta_i \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \pi_{t,i} \left(\frac{a}{\sqrt{t}} - \frac{\Delta_i}{3}\right) + \sqrt{\pi_{t,i}} \left(\frac{b}{\sqrt{t}} - \frac{\Delta_i \sqrt{\pi_{t,i}}}{3}\right).$$

For a fixed i the term $\left(\frac{a}{\sqrt{t}} - \frac{\Delta_i}{3}\right) \leq 0$ if $t \geq \frac{9a^2}{\Delta_i^2}$ and so the maximum regret from

$$\sum_{t=1}^{T} \pi_{t,i} \left(\frac{a}{\sqrt{t}} - \frac{\Delta_i}{3}\right) \leq \sum_{t=1}^{\lfloor \frac{9a^2}{\Delta_i^2} \rfloor} \frac{a}{\sqrt{t}} \leq \frac{6a^2}{\Delta_i}.$$

Further the term $\sqrt{\pi_{t,i}} \left(\frac{b}{\sqrt{t}} - \frac{\Delta_i \sqrt{\pi_{t,i}}}{3}\right) \leq \frac{2b^2}{t\Delta_i}$ for $t \geq \frac{4b^2}{\Delta_i^2}$. This implies

$$\sum_{t=1}^{T} \sqrt{\pi_{t,i}} \left(\frac{b}{\sqrt{t}} - \frac{\Delta_i \sqrt{\pi_{t,i}}}{3}\right) \leq \sum_{t=1}^{\lfloor \frac{4b^2}{\Delta_i^2} \rfloor} \frac{b}{\sqrt{t}} + \sum_{t=1}^{T} \frac{2b^2}{\Delta_i t} \leq \frac{8b^2}{\Delta_i} + \frac{2b^2 \log(T)}{\Delta_i}.$$

Combining the two bounds we have

$$\sum_{t=1}^{T} \sum_{i \neq i^*} \pi_{t,i} \Delta_i \leq O\left(\sum_{i \neq i^*} \frac{b^2 \log(T)}{\Delta_i} + \frac{a^2}{\Delta_i}\right).$$

E TS-Prod and stabilized OMD

The TS-Prod update is a linearization of the FTRL update as explained in Section 5.2. In this section we show a regret bound for the TS-Prod algorithm by linearizing the OMD update as we did for WSU-UX and LB-Prod. This comes with its own set of challenges. First, we believe that a decreasing step-size in the OMD update is important for achieving optimal regret bounds in the stochastic setting. Second, the vanilla OMD update with decreasing step-size might incur linear regret in the adversarial setting. This second issue is resolved by the *stabilized* OMD algorithms proposed by Fang et al. [2022]. TS-Prod turns out to be equivalent to the dual stabilized OMD algorithm of Fang et al. [2022]

https://doi.org/10.52202/079017-2710

and hence will inherit the adversarial regret guarantees. In the rest of the section we sketch the regret analysis for the stochastic setting and how we can reduce the regret analysis in the adversarial setting to that of Fang et al. [2022].

Stabilization as introduced by Fang et al. [2022] is the process of mixing the gradient mapping, $\nabla F(\pi_t)$, of the current iterate with the gradient mapping of the first iterate, $\nabla F(\pi_1)$, in the mirror descent update in the dual space, where $F(x) = -\sum_{i=1}^{K} 2\sqrt{x_i}$. This mixing turns out to be equivalent to the negative biasing of losses in Equation TS-Prod by the ξ_t dependent terms.

The regret analysis begins by defining the perturbations $\{\epsilon_{t,i}\}_{t\in[T],i\in[K]}$ so that

$$\pi_{t+1,i} = \frac{\pi_{t,i}}{(1+\eta_{t+1}\sqrt{\pi_{t,i}}(\hat{L}_{t,i}+\epsilon_{t,i}))^2}, \tag{4}$$

that is $\epsilon_{t,i}$ makes the update in Equation TS-Prod equivalent to the $1/2$-Tsallis mirror descent update. We note that the $\epsilon_{t,i}$ is only defined to assist with the regret analysis and it never needs to be computed for the actual update. The perturbations, $\epsilon_{t,i}$, are well controlled as we show next.

Lemma 13. *For every $t \in [T], i \in [K]$, there exists $\epsilon_{t,i}$ such that $\epsilon_{t,i} \le 4\eta_t\sqrt{\pi_{t,i}}\hat{L}_{t,i}^2$ for all i and $\ell_{t,i}$.*

Lemma 13 allows us to proceed with the analysis for the stochastic and adversarial cases by using the standard regret decomposition into a *penalty* and *stability* terms. In the stochastic case we can bound the two terms in the following way

Lemma 14. *For stochastic losses the penalty term is bounded in expectation by*

$$O\left(\frac{\mathbb{E}\left[\left(\sum_{i\neq i^*}\pi_{t+1,i}\right)^2\right]\sqrt{K}\log(t)}{\sqrt{t}} + \frac{1}{\sqrt{t}} \wedge \frac{\mathbb{E}\left[\left(\sum_{i\neq i^*}\pi_{t+1,i}\right)^2\right]\log(KT)}{\sqrt{t}}\right).$$

Lemma 15. *For stochastic losses the stability term is bounded by*

$$O\left(\frac{1}{\sqrt{t}}\sum_{i=1}^{K}\sqrt{\pi_{t,i}}(1-\pi_{t,i}) + \frac{K\sqrt{\pi_{t,i}}}{t^2} + \frac{K}{t}\right).$$

The stochastic regret bound proof can now be completed by a careful self-bounding argument.

In the adversarial case we reduce the regret bound to that of Fang et al. [2022] in the following way. Let $\Phi = F + I_{\Delta^{K-1}}$ be potential defined by mixing the $1/2$-Tsallis potential together with the indicator function for the probability simplex. The update of Algorithm 2 (Dual Stabilized OMD) can then be written as

$$\begin{aligned}
\hat{w}_{t+1} &= \nabla\Phi(\pi_t) - \eta_t(\tilde{\ell}_t + \epsilon_t),\\
\hat{y}_{t+1} &= \chi_t\hat{w}_{t+1} + (1-\chi_t)\nabla\Phi(\pi_1),\\
\pi_{t+1} &= \nabla\Phi^*(\hat{y}_{t+1}).
\end{aligned}$$

It turns out that this update is equivalent to the OMD update with respect to $\hat{L}_{t,i}$ in Equation 4. This allows us to use the regret bound in Theorem 3 [Fang et al., 2022]. Overall the regret of the perturbed OMD version is bounded as follows.

Theorem 4. *The regret of the algorithm defined by the update in Equation 4 is bounded by*

$$O\left(\sum_{i\neq i^*}\frac{\log(T)}{\Delta_i} + \frac{K\log^2(1/\Delta_{min})}{\Delta_{min}} + K^{3/2}\right)$$

in the stochastic case, where Δ_{min} is the smallest gap between the expected losses. Further the regret in the adversarial setting is bounded by $O(\sqrt{KT})$.

E.1 Proof of Theorem 4

Proof of Lemma 13. We work under the following assumption which is satisfied with the choice of η_t and γ by Lemma 5. Further, we are going to work with the following slight modification of the

losses $\hat{\ell}_t$ in the update of TS-Prod:

$$\hat{\ell}_{t,i} = \left(\ell_{t,i} - \frac{\eta_t(\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\mathbb{I}(A_t=i)}{\pi_{t,i}} - \frac{\eta_t\xi_t}{\sqrt{\pi_{t,i}}}.$$

We quickly check that the variance with this definition of $\hat{\ell}_t$ is bounded by Lemma 10

$$\begin{aligned}
&\mathbb{E}_t\left[\left(\left(\ell_{t,i} - \frac{\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\mathbb{I}(A_t=i)}{\pi_{t,i}} - \frac{\eta_t\xi_t}{\sqrt{\pi_{t,i}}}\right)^2\right] \\
=&\mathbb{E}_t\left[\left(\ell_{t,i} - \frac{\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)^2\frac{1}{\pi_{t,i}} + \frac{(\eta_t\xi_t)^2}{\pi_{t,i}} - 2\left(\ell_{t,i} - \frac{\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\eta_t\xi_t}{\sqrt{\pi_{t,i}}}\right] \\
\leq&\mathbb{E}_t\left[\left(\ell_{t,i} - \frac{\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)^2\frac{1}{\pi_{t,i}} + \frac{(\eta_t\xi_t)^2}{\pi_{t,i}^2} - 2\left(\ell_{t,i} - \frac{\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\eta_t\xi_t}{\sqrt{\pi_{t,i}}}\right] \\
=&\mathbb{E}_t\left[\left(\left(\ell_{t,i} - \frac{\eta_t(\xi_t+\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\mathbb{I}(A_t=i)}{\pi_{t,i}}\right)^2\right]
\end{aligned}$$

Assumption 1. *Assume that for all $t \in [T], i \in [K]$ it holds that $|\eta_t\sqrt{\pi_{t,i}\hat{L}_{t,i}}| \leq \frac{1}{6}$.*

Lemma 13. First for non-negative ℓ we show that $\epsilon \in [0, \ell]$. This follows by observing that $1/(1+3/2x)^2 \leq 1-2x$ for $x \in [0, 1/6]$ and $1/(1+x)^2 \geq 1-2x$ for $x \geq 0$ so by the Intermediate Value theorem there exists an $\epsilon \in [0, \ell]$ such that equality is obtained. Next, for $\ell < 0$ for $\epsilon = 0$ we have $1/(1-x)^2 \geq 1+2x$ for $x \geq 0$ and for $\epsilon = \ell/2$ we have $1/(1-x/2)^2 \leq 1-2x$ for $x \in [0, 1/4]$ and so we have $\epsilon = [-\frac{|\ell|}{2}, \frac{|\ell|}{2}]$.

For the second part of the lemma using the Taylor expansion around 0 of $1/(1+x)^2$ implies that

$$\frac{1}{(1+\eta\sqrt{\pi}\tilde{\ell})^2} \leq 1 - 2\eta\sqrt{\pi}\tilde{\ell} + 3(\eta\sqrt{\pi}\tilde{\ell})^2,$$

and so

$$\begin{aligned}
1 - 2\eta\sqrt{\pi}(\tilde{\ell}-\epsilon) \leq 1 - 2\eta\sqrt{\pi}\tilde{\ell} + 3(\eta\sqrt{\pi}\tilde{\ell})^2 &\iff \\
2\eta\sqrt{\pi}\epsilon \leq 3(\eta\sqrt{\pi}\tilde{\ell})^2 &\iff \\
2\eta\sqrt{\pi}\epsilon \leq 3(\eta\sqrt{\pi}(\ell+\epsilon))^2 &\implies \\
\epsilon \leq 4\eta\sqrt{\pi}\ell^2.&
\end{aligned}$$

□

Stochastic bound. Let $D_t^{TS}(u,w) = \frac{1}{\eta_t}D_{TS}(u,v)$ and let

$$\tilde{\pi}_{t+1,i} = \frac{\pi_{t,i}}{(1+\eta_{t+1}\sqrt{\pi_{t,i}}(\hat{\ell}_{t,i}+\epsilon_{t,i}-\lambda_t))^2},$$

where $\lambda_t = \ell_{t,A_t}$. We note that π_{t+1} is now the projection of $\tilde{\pi}_{t+1}$ onto the simplex. Further by the 3-point rule for Bregman divergence we have that

$$\begin{aligned}
\langle\hat{L}_t, \pi_t - u\rangle &= D_t^{TS}(u,\pi_t) - D_t^{TS}(u,\tilde{\pi}_{t+1}) + D_t^{TS}(\pi_t,\tilde{\pi}_{t+1}) \\
&\leq D_t^{TS}(u,\pi_t) - D_t^{TS}(u,\pi_{t+1}) + D_t^{TS}(\pi_t,\tilde{\pi}_{t+1}).
\end{aligned}$$

Penalty term.

Lemma 16 (Lemma 14)**.** *For stochastic losses the penalty term is bounded as follows*

$$\mathbb{E}[D_{t+1}^{TS}(u,\pi_{t+1}) - D_t^{TS}(u,\pi_{t+1})] \leq O\left(\frac{\mathbb{E}\left[\left(\sum_{i\neq i^*}\pi_{t+1,i}\right)^2\right]\sqrt{K}\log(t)}{\sqrt{t}}\right),$$

where $D_t^{TS}(u,v) = \frac{1}{\eta_t}D_{TS}(u,v)$ and $\eta_t = \frac{1}{\sqrt{t}}$.

https://doi.org/10.52202/079017-2710

Proof. for $u = e_{i^*}$:

$$\begin{aligned}
D_{t+1}^{TS}(u, \pi_{t+1}) - D_t^{TS}(u, \pi_{t+1}) &= -\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \\
&+ \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\left(\frac{1}{\sqrt{\pi_{t+1,i^*}}} - 1\right) \\
&+ \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\sum_{i=1}^{K}\sqrt{\pi_{t+1,i}} \\
&= \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\frac{1 - \sqrt{\pi_{t+1,i^*}} + \pi_{t+1,i^*}}{\sqrt{\pi_{t+1,i^*}}} \\
&+ \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\sum_{i\neq i^*}\sqrt{\pi_{t+1,i}} \\
&= \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\frac{(1 - \sqrt{\pi_{t+1,i^*}})^2}{\sqrt{\pi_{t+1,i^*}}} \\
&+ \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\sum_{i\neq i^*}\sqrt{\pi_{t+1,i}}.
\end{aligned}$$

From the update in Equation 4 we have

$$\frac{1}{\sqrt{\pi_{t+1,i^*}}} = \sqrt{K} + \sum_{s=1}^{t}\eta_s(\hat{L}_{s,i^*} + \epsilon_{s,i^*}),$$

which implies

$$\frac{(1 - \sqrt{\pi_{t+1,i^*}})^2}{\sqrt{\pi_{t+1,i^*}}} \leq \sqrt{K}(1 - \sqrt{\pi_{t+1,i^*}})^2 + (1 - \sqrt{\pi_{t+1,i^*}})^2\sum_{s=1}^{t}\eta_s(\hat{L}_{s,i^*} + \epsilon_{s,i^*})$$

First we bound $(1 - \sqrt{\pi_{t+1,i^*}})^2$:

$$\begin{aligned}
(1 - \sqrt{\pi_{t+1,i^*}})^2 &= \left(1 - \sqrt{1 - \sum_{i\neq i^*}\pi_{t+1,i}}\right)^2 = \left(\frac{\sum_{i\neq i^*}\pi_{t+1,i}}{1 + \sqrt{1 - \sum_{i\neq i^*}\pi_{t+1,i}}}\right)^2 \\
&\leq \left(\sum_{i\neq i^*}\pi_{t+1,i}\right)^2.
\end{aligned}$$

In the stochastic case WLOG we can take $\mathbb{E}[\ell_{t,i^*}] = 0, \forall t \in [T]$. We first control $\hat{L}_{t,i^*}$. We have

$$\hat{L}_{t,i^*} = \left(\ell_{t,i^*} - \frac{\eta_t\gamma(1 - \pi_{t,i^*})}{\sqrt{\pi_{t,i^*}}}\right)\frac{\mathbb{I}(A_t = i^*)}{\pi_{t,i^*}} - \frac{\eta_t\xi_t}{\sqrt{\pi_{t,i^*}}} - \sum_{i=1}^{K}\frac{\pi_{t,i}^{\frac{3}{2}}\hat{\ell}_{t,i}}{\sum_{j=1}^{K}\pi_{t,j}^{\frac{3}{2}}}$$

The first term, $\ell_{t,i^*}\frac{\mathbb{I}(A_t=i^*)}{\pi_{t,i^*}}$ is 0 in expectation. The second term, $-\frac{\eta_t\gamma(1-\pi_{t,i^*})}{\sqrt{\pi_{t,i^*}}}\frac{\mathbb{I}(A_t=i^*)}{\pi_{t,i^*}}$, will be used to cancel out the contribution from the perturbation ϵ_{t,i^*}. The third term is there to help with the adversarial setting analysis. Next, we decompose the fourth term as

$$\sum_{i=1}^{K}\frac{\pi_{t,i}^{\frac{3}{2}}\hat{\ell}_{t,i}}{\sum_{j=1}^{K}\pi_{t,j}^{\frac{3}{2}}} = \frac{1}{\sum_{j=1}^{K}\pi_{t,j}^{3/2}}\sum_{i=1}^{K}\sqrt{\pi_{t,i}}\ell_{t,i}\mathbb{I}(A_t = i) - \eta_t\gamma(1 - \pi_{t,i})\mathbb{I}(A_t = i) - \pi_{t,i}\eta_t\xi_t.$$

The first part of the above has a non-positive contribution to $\hat{L}_{t,i^*}$ in expectation. The only non-negative contribution now comes from

$$\frac{1}{\sum_{j=1}^{K}\pi_{t,j}^{3/2}}\sum_{i=1}^{K}\eta_t(\xi_t + \gamma(1 - \pi_{t,i}))\mathbb{I}(A_t = i) \leq \sqrt{K}(\eta_t\xi_t + \eta_t\gamma) \leq 2\sqrt{K}\eta_t\gamma$$

and so we bound

$$\sum_{s=1}^{t} -\eta_s \sum_{i=1}^{K} \frac{\pi_{s,i}^{\frac{3}{2}} \hat{\ell}_{s,i}}{\sum_{j=1}^{K} \pi_{s,j}^{\frac{3}{2}}} \leq \sum_{s=1}^{t} 2\eta_s \sqrt{K} \eta_s \gamma \leq 64\sqrt{K} \log(t).$$

Next we are going to bound $\mathbb{E}[\epsilon_{t,i^*}]$ using Lemma 10 together with Lemma 13:

$$\mathbb{E}[\epsilon_{t,i^*}] \leq \mathbb{E}[\eta_t \sqrt{\pi_{t,i^*}} \hat{L}_{t,i^*}^2] \leq \frac{13}{2} \mathbb{E}[\eta_t (1 - \pi_{t,i^*}) / \sqrt{\pi_{t,i^*}}].$$

This term is exactly canceled out by $\frac{\eta_t \gamma (1 - \pi_{t,i^*})}{\sqrt{\pi_{t,i^*}}}$ as $\gamma = \frac{13}{2}$. For the final bound we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \frac{(1 - \sqrt{\pi_{t+1,i^*}})^2}{\sqrt{\pi_{t+1,i^*}}}\right] \\
&\leq \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \mathbb{E}\left[\left(\sum_{i \neq i^*} \pi_{t+1,i}\right)^2\right] + 4\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \mathbb{E}\left[\left(\sum_{i \neq i^*} \pi_{t+1,i}\right)^2 \sum_{s=1}^{t} \eta_s (\hat{L}_{s,i^*} + \epsilon_{s,i^*})\right] \\
&\leq \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \mathbb{E}\left[\left(\sum_{i \neq i^*} \pi_{t+1,i}\right)^2\right] + \mathbb{E}\left[\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) (1 - \sqrt{\pi_{t+1,i^*}})^2 \sum_{s=1}^{t} -\eta_s \sum_{i=1}^{K} \frac{\pi_{s,i}^{\frac{3}{2}} \hat{\ell}_{s,i}}{\sum_{j=1}^{K} \pi_{s,j}^{\frac{3}{2}}}\right] \\
&\quad + \mathbb{E}\left[\sum_{s=1}^{t} \eta_s \left(\epsilon_{s,i^*} - \frac{\gamma \eta_s (1 - \pi_{s,i^*})}{\sqrt{\pi_{s,i^*}}}\right)\right] \\
&\leq 32 \frac{\mathbb{E}\left[\left(\sum_{i \neq i^*} \pi_{t+1,i}\right)^2\right] \sqrt{K} \log(t)}{\sqrt{t}}.
\end{aligned}$$

□

Stability term. Recall that the stability term is $D_t^{TS}(\pi_t, \tilde{\pi}_{t+1})$. This term is bounded in a standard way. We proceed to do so as follows for any $t \geq 4\sqrt{K}$:

Lemma 17 (Lemma 15)**.** *For stochastic losses the stability term is bounded as follows*

$$\mathbb{E}[D_t^{TS}(\pi_t, \tilde{\pi}_{t+1})] \leq O\left(\frac{1}{\sqrt{t}} \sum_{i=1}^{K} \sqrt{\pi_{t,i}}(1 - \pi_{t,i}) + \frac{1}{t}\right),$$

where $D_t^{TS}(u, v) = \frac{1}{\eta_t} D_{TS}(u, v)$.

 https://doi.org/10.52202/079017-2710

Proof. We have the following

$$
\begin{aligned}
D_t^{TS}(\pi_t, \tilde{\pi}_{t+1}) &= \frac{1}{\eta_t}\sum_{i=1}^{K} 2\sqrt{\pi_{t+1,i}} - 2\sqrt{\pi_{t,i}} - \frac{1}{\sqrt{\pi_{t+1,i}}}(\pi_{t+1,i} - \pi_{t,i}) \\
&= \frac{1}{\eta_t}\sum_{i=1}^{K} \sqrt{\pi_{t+1,i}} - 2\sqrt{\pi_{t,i}} + \sqrt{\pi_{t,i}}\left(1 + \eta_{t+1}\sqrt{\pi_{t,i}}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t)\right) \\
&= \frac{1}{\eta_t}\sum_{i=1}^{K} \sqrt{\pi_{t+1,i}} - \sqrt{\pi_{t,i}} + \eta_{t+1}\pi_{t,i}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t) \\
&= \frac{1}{\eta_t}\sum_{i=1}^{K} \sqrt{\pi_{t,i}}\left(\frac{1}{1 + \eta_{t+1}\sqrt{\pi_{t,i}}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t)} - 1\right) + \eta_{t+1}\pi_{t,i}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t) \\
&\leq \frac{1}{\eta_t}\sum_{i=1}^{K} \sqrt{\pi_{t,i}}\left(-\eta_{t+1}\sqrt{\pi_{t,i}}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t) + 2\eta_{t+1}^2\pi_{t,i}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t)^2\right) \\
&+ \eta_{t+1}\pi_{t,i}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t) \qquad \left(\tfrac{1}{1+x} \leq 1 - x + 2x^2 \text{ for } x \geq -\tfrac{1}{2}\right) \\
&\leq 2\eta_t \sum_{i=1}^{K} \pi_{t,i}^{3/2}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t)^2,
\end{aligned}
$$

where for the second to last inequality we only need to check $\eta_{t+1}\sqrt{\pi_{t,i}}(\epsilon_{t,i} - \lambda_t) \geq -\frac{1}{2}$. We have $\eta_{t+1}\sqrt{\pi_{t,i}}\lambda_t \geq -\frac{1}{\sqrt{t}}$ and Lemma 13 implies

$$
\mathbb{E}[\eta_{t+1}\sqrt{\pi_{t,i}}\epsilon_{t,i}] \geq -\eta_t^2\,\mathbb{E}[\pi_{t,i}|\hat{L}_{t,i}|^2] \geq -\Omega(\frac{1}{t}).
$$

We bound $\mathbb{E}[\pi_{t,i}^{3/2}(\hat{\ell}_{t,i} + \epsilon_{t,i} - \lambda_t)^2] \leq 2\,\mathbb{E}[\pi_{t,i}^{3/2}\epsilon_{t,i}^2] + 2\,\mathbb{E}[\eta_{t+1}^2\pi_{t,i}^{3/2}(\hat{\ell}_{t,i} - \lambda_t)^2]$. For the first term we have

$$
2\,\mathbb{E}[\eta_t^2\pi_{t,i}^{5/2}|\hat{L}_{t,i}|^4] \leq O\left(\frac{1}{\sqrt{t}}\right).
$$

For the second term we use Lemma 10 to get $\mathbb{E}[\eta_{t+1}^2\pi_{t,i}^{3/2}(\hat{\ell}_{t,i} - \lambda_t)^2] \leq \frac{13}{2}\,\mathbb{E}[\eta_{t+1}^2\sqrt{\pi_{t,i}}(1 - \pi_{t,i})]$. □

Self-bounding the regret for stochastic losses.

Theorem 4, stochastic losses. Combining the bound in Lemma 14 and Lemma 15 together with the adversarial bound we have that the total regret is bounded as follows

$$
\begin{aligned}
&+O\left(\sum_{t=T_0}^{T}\sum_{i\neq i^*}\sqrt{\pi_{t,i}}\left(\frac{1}{\sqrt{t}} - \sqrt{\pi_{t,i}}\left(\Delta_i - \frac{\sqrt{K}\log(t)}{\sqrt{t}}\right)\right) + \sum_{t=1}^{T_0-1}\frac{\pi_{t,i}\sqrt{K}\log(t)}{\sqrt{t}}\right) \\
&+O\left(K^{3/2}\right)
\end{aligned}
$$

In the above we bound the lower order term from the stability as $\sum_{t=1}^{T}\sum_{i=1}^{K}\gamma_t^2\sqrt{\pi_{t,i}} = O(K^{3/2})$ and decompose the regret into four parts. The first and second line correspond to the two terms from the penalty bound. Each of the two lines are decomposed into two terms. The first term is the result of the self-bounding trick and the second term is the additional regret for the initial number of rounds before the self-bounding trick can be applied.

We repeatedly use the following inequality $2a\sqrt{x} - bx \leq \frac{a^2}{b}$, which holds for $a, b \geq 0$. For the first line of the decomposition we take $T_0 = 8\frac{K\log^2(K/\Delta_{min})}{\Delta_{min}}$, where Δ_{min} is the smallest non-zero expected loss. We note that

$$
\frac{\sqrt{K}\log(T_0)}{\sqrt{T_0}} = \Delta_{min}\frac{\log(8K\log^2(1/\Delta_{min}))}{8\log(K/\Delta_{min})} = \Delta_{min}\left(\frac{\log(K)}{8\log(K/\Delta)} + \frac{\log(16\log(1/\Delta_{\min})}{8\log(K/\Delta_{min})}\right) \leq \frac{\Delta_{min}}{2},
$$

for any $\Delta_{min} \leq \frac{1}{2024}$. If $\Delta_{min} > \frac{1}{2024}$, we take $T_0 = 8\frac{K\log^2(2024)}{\Delta_{min}}$. The above implies that $\sqrt{\pi_{t,i}}\left(\frac{1}{\sqrt{t}} - \sqrt{\pi_{t,i}}\left(\Delta_i - \frac{\sqrt{K}\log(t)}{\sqrt{t}}\right)\right) \leq \frac{2}{t\Delta_i}$ and further $\sum_{i\neq i^*}\sum_{t=1}^{T_0-1}\frac{\pi_{t,i}\sqrt{K}\log(t)}{\sqrt{t}} = O(\frac{K\log^2(1/\Delta_{\min})}{\Delta_{\min}})$. The final regret bound is

$$O\left(\sum_{i\neq i^*}\frac{\log(T)}{\Delta_i} + \frac{K\log^2(1/\Delta_{\min})}{\Delta_{\min}} + K^{3/2}\right).$$

□

Adversarial losses. We now present the argument for the regret bound in the adversarial setting.

Theorem 4, adversarial losses. Let $\tilde{\ell}_{t,i} = \left(\ell_{t,i} - \frac{\eta_t\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\mathbb{I}(A_t=i)}{\pi_{t,i}}$ and $\chi_t = \frac{\eta_t}{\eta_{t-1}}$ We recall the update for Algorithm 2 in Fang et al. [2022]

$$\begin{aligned}
\hat{w}_{t+1} &= \nabla\Phi(\pi_t) - \eta_{t-1}(\tilde{\ell}_t + \epsilon_t),\\
\hat{y}_{t+1} &= \chi_t\hat{w}_{t+1} + (1-\chi_t)\nabla\Phi(\pi_1),\\
\pi_{t+1} &= \nabla\Phi^*(\hat{y}_{t+1}),
\end{aligned}$$

where Φ is the $1/2$-Tsallis potential plus the indicator function for the probability simplex Δ^{K-1}. Since π_1 is uniform the second step of the update is equivalent to $\hat{y}_{t+1} = \chi_t\hat{w}_{t+1}$. Re-writing the first step of the update we have

$$\begin{aligned}
-\hat{y}_{t+1,i} &= \frac{\chi_t}{\sqrt{\pi_{t,i}}} + \eta_t(\tilde{\ell}_{t,i} + \epsilon_{t,i}) = \frac{1}{\sqrt{\pi_{t,i}}} + \eta_{t+1}(\tilde{\ell}_{t,i} + \epsilon_{t,i}) - \eta_t\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\frac{1}{\sqrt{\pi_{t,i}}}\\
&= \frac{1}{\sqrt{\pi_{t,i}}} + \eta_t\left(\left(\ell_{t,i} - \frac{\eta_t\gamma(1-\pi_{t,i})}{\sqrt{\pi_{t,i}}}\right)\frac{\mathbb{I}(A_t=i)}{\pi_{t,i}} + \epsilon_{t,i} - \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\frac{1}{\sqrt{\pi_{t,i}}}\right)\\
&= \frac{1}{\sqrt{\pi_{t,i}}} + \eta_t(\hat{\ell}_{t,i} + \lambda_t).
\end{aligned}$$

Since $\nabla\Phi^*$ is invariant under constant vector perturbations we finally have

$$\begin{aligned}
\pi_{t+1,i} &= \nabla\Phi^*(\hat{y}_{t+1})_i = \nabla\Phi^*\left(-\frac{1}{\sqrt{\pi_t}} - \eta_t(\hat{\ell}_{t+1} + \epsilon_t)\right) = \nabla\Phi^*\left(-\frac{1}{\sqrt{\pi_t}} - \eta_t(\hat{L}_{t+1} + \epsilon_t)\right)\\
&= \frac{1}{(1/\sqrt{\pi_{t,i}} + \eta_t(\hat{L}_{t,i} + \epsilon_{t,i}))^2}\\
&= \frac{\pi_{t,i}}{(1 + \eta_t\sqrt{\pi_{t,i}}(\hat{L}_{t,i} + \epsilon_{t,i}))^2}.
\end{aligned}$$

And so the update in Algorithm 2 of Fang et al. [2022] is equivalent to the perturbed OMD update which we have shown enjoys an optimistic regret guarantee. The regret guarantee in the adversarial setting is now recovered from Theorem 3 in Fang et al. [2022]. In particular the theorem guarantees that the regret is bounded as

$$\sum_{t=1}^{T} D_t^{TS}(\pi_t, \nabla F^*(\nabla F(\pi_t) - \eta_t(\hat{\ell}_t + \epsilon_t))) + \sqrt{KT}.$$

Every term in the sum is bounded in the same way as the stability terms, that is $D_t^{TS}(\pi_t, \nabla F^*(\nabla F(\pi_t) - \eta_t(\hat{\ell}_t + \epsilon_t))) \leq \sum_{i=1}^{K}\pi_{t,i}^{3/2}(\hat{\ell}_{t,i}^2 + \epsilon_{t,i})$. We can now use the bound in the proof of Lemma 15 to complete proof of the adversarial bound and the proof of Theorem 4. □

https://doi.org/10.52202/079017-2710

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All results are formalized in Theorems which we proof in the supplementary material additionally to selective proof outlines in the main body.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: This is purely theoretical work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are stated in the Theorem statements and the proof can be found in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: No experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

 https://doi.org/10.52202/079017-2710

Answer: [NA]

Justification: No experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Theoretical work without direct societal impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Theoretical work without societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-2710

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Theoretical work without experiments or data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Theoretical work without existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Theoretical work without new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Theoretical work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Theoretical work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2710