Best of both worlds

Yifang Chen, Xiujun Li, Zhaoqi Li, Lianhui Qin, Zhihan Xiong
{yifangc, xiujun, zli9, lianhuiq, zhihanx}@uw.edu

1 Motivation and preliminary

In this section, we review stochastic and adversarial settings for both multi-armed bandits and semi-bandits; for corruption setting, we leave it to Thursday's section.

1.1 Multi-Armed Bandits (MAB) and Semi-bandits

Multi-Armed Bandits Setting
for $t = 1, \dots, T$
- The learner selects an arm $i_t \in [n]$
- Simultaneously, environment selects a loss vector $\ell_t \in [0,1]^d$
- The learner suffers and observes $\ell^t_{i_t} \in [0,1]$.

Semi-bandit problem is a natural generation of multi-armed bandits, where the agent has to pick a subset of arms (called combinatorial actions or macro-arm[1]), but can still observe the loss for each arm. Notice that in a more generalized setting linear bandits we talked in the class, one can only observes $\langle X_t, \ell_t \rangle$.

Semi-Bandit Setting
for $t = 1, \dots, T$
- The learner selects a combinatorial action $X_t \in \mathcal{X}$, where $\mathcal{X} \subset \{0,1\}^d$.
- Simultaneously, environment selects a loss vector $\ell_t \in [-1,1]^d$
- The learner suffers the loss $\langle X_t, \ell_t \rangle$
- The learner observes the (semi-bandit) feedback $o_t = X_t \circ \ell_t \in [-1,1]^d$, where $\circ$ is element-wise multiplication.

It is easy to see that, when $\mathcal{X} = \{\mathbf{e}_1, \dots, \mathbf{e}_d\}$, the problem reduces to the general MAB.

Our target regret form: The performance of a learner is measured by pseudo-regret:

$$\overline{\text{Reg}}_T := \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t - x^*, \ell_t \rangle\right]$$

where $x^* = \text{argmin}_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \langle x, \ell_t \rangle\right]$ is the best action in hindsight and the expectation is with respect to the randomness of both the learner and the environment.

1.2 Intro to both worlds and the key motivation of this work

Since MAB is a special case of semi-bandits, so in the following we will just use the semi-bandits notation. First we introduce the traditional stochastic and adversarial settings, which the "both worlds" refers to.

1. **Stochastic bandits**: $\ell_1, \ell_2, \dots, \ell_T$ are i.i.d loss vector drawn from some fixed unknown distribution ν and therefore, for each arm X, we can calculate its gap $\Delta_X = \langle X - x^*, \nu \rangle$. One near-optimal algorithm is Elimination-style algorithm (in the lecture 2), which guarantees a

[1]In hierarchical reinforcement learning, a macro-action is a "multi-step" action, here you can think a macro-arm is a subset of primitive arms.

gap-dependent (or sometimes called instance-dependent) regret bound,[2]

$$\text{Reg} \le \mathcal{O}\left(\min\left\{\frac{\max_{X\in\mathcal{X}}\|X\|d\log(T)}{\min_{X\in\mathcal{X}}\Delta_X}, \sqrt{\max_{X\in\mathcal{X}}\|X\|d\log(T)T}\right\}\right)$$

2. **Adversarial bandits**: ℓ_T is chosen in an arbitrary way based on the history $\ell_1, X_1, \ell_2, X_2, \ldots, \ell_{T-1}, X_{T-1}$ and possibly an internal randomization by the environment. One near-optimal algorithm is EXP3 (in the lecture 3), which guarantees

$$\text{Reg} \le \mathcal{O}\left(\sqrt{\max_{X\in\mathcal{X}}\|X\|dT}\right)$$

Key motivation: It is easy to see that, when given a prior knowledge of whether the environment is stochastic or adverserial, we can always choose a proper algorithm to get a near-optimal regret. **But can we achieve the optimal regret without knowing the property of environment in advance?** This paper says yes !

1.3 Intro to some mediate setting between two worlds

Besides the i.i.d setting and the purely adversarial settings we present above, there are also some mediate settings between these two.

1. **Stochastically constrained adversarial setting**: $\ell_1, \ell_2, \ldots, \ell_T$ are drawn from a sequence of unknown distributions $\nu_1, \nu_2, \ldots, \nu_T$ that satisfies the following:

$$\mathbb{E}_{l_t\sim\nu_t}[\langle X_t - x^*, \ell_t\rangle] \ge \Delta_X, \forall t$$

2. **Corruption setting:** Instead of observing $o_t = X_t \circ \ell_t \in [-1,1]^d$, the learner observes $o_t = X_t \circ \left(\ell_t \in [-1,1]^d + c_t\right)$. Here $c_t \in \mathbb{R}^d$ is an arbitrary corruption that determined by environment. More constraints on corruption will be discussed in later section. It is easy to see that, it recovers the stochastic setting when all $c_t = 0$ and it recover the adversarial setting when all c_t are chosen adversarially.

If we can achieve the near-optimal result simultaneously in the stochastic and adversarial world, it is also natural to ask: **Can we also achieve near-optimal result simultaneously in these mediate settings?** Fortunately, the answer is yes. And we will discuss that in later sections.

2 An overview of best-of-both world result

- Use OMD/FTRL framework with proper regularizer (this paper, only works for MAB and semi-bandits)
- Continuously update the estimated gaps and sample according to the estimated gaps (next week, works for linear bandits)
- Start with stochastic and switch to adversarial via some testing (old, usually deprecated)

3 A general analysis process for FTRL algorithm

- For any potential function, we can divide that into stability and penalty term Zimmert et al. [2019]
- A revisit to entropy regularizer and exp3, explain the intuition why it cannot work well
- An introduction to the class of Tsallis-entropy and three special case Zimmert and Seldin [2021]
- Hybrid regularizer (just a brief mentioning here) Zimmert et al. [2019], Masoudian and Seldin [2021]

Before going to the details, we want to clarify that the FTRL presented below can also be analyzed under the OMD framework because these two are closely related, and sometimes equivalent.**?** Here we use the FTRL framework by following the convention of Zimmert et al. [2019]. People interested in OMD version analysis can refer to Zimmert and Seldin [2021], Masoudian and Seldin [2021]. But the basic ideas behind these two version are essentially the same.

[2]Actually, the existing algorithm can achieve a better regret, but for simplicity we present a close one here.

3.1 A general analysis framework for FTRL

The algorithm we will talk about is based on the general Follow-The-Regularized-Leader (FTRL) framework, with regularization function $\Psi(\cdot)$, loss estimator $\hat{\ell}_t$ and learning rate η_t, that will be specified later.

Algorithm 1 General FTRL for Semi-bandit

1: **input:** A regularization function $\Psi(x)$, a time varying learning rate η_t
2: **initialize:** $\hat{L}_0 = (0, 0, \ldots, 0)$
3: **for** t= 1,2,... **do**
4: Compute $x_t = \text{argmin}_{x \in \text{Conv}(\mathcal{X})} \left\langle x, \hat{L}_{t-1} \right\rangle + \eta_t^{-1} \Psi(x)$
5: Sample $X \sim P(x_t)$ according to some sample scheme P satisfying

$$\mathbb{E}_{X \sim P(x)}[X] = x$$

6: Observe $o_t = X_t \circ \ell_t$ and suffer loss $\langle X_t, \ell_t \rangle$
7: Construct estimator $\hat{\ell}_t$ such that $\mathbb{E}[\hat{\ell}_t] = \ell_t$
8: Update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
9: **end for**

For any regularization function $\Psi(x)$, learning rate η_t and estimator $\hat{l}_t$, we can always get a standard analysis of the regret by dividing it into stability term (Reg_{stan}) and the penalty term (Reg_{pen}),

$$\overline{\text{Reg}}_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, l_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{\text{Reg}_{stab}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \langle x^*, \ell_t \rangle\right]}_{\text{Reg}_{pen}}$$

where potential function $\Phi_t(\cdot) = \max_{x \in \text{Conv}(\mathcal{X})} \{\langle x, \cdot \rangle - \Psi_t(x)\}$ and $\Psi_t(\cdot) = \eta_t^{-1}\Psi(\cdot)$. Here, Φ_t is the convex conjugate function of $\Psi_t + \mathcal{I}_{\text{Conv}(\mathcal{X})}$.

At the first glance, people may think this as a simply arrangement and introduce nothing new. Here we provide more intuition by connecting this with something we have already learnt: Recall the OMD algorithm we did in class shown in **Figure 1** (Note OMD is equivalent to FTRL in some cases). The theorem it shows is not exactly the same as what we did in lecture, but a direct consequence of one of our intermediate steps. Specifically, in previous lecture, we have the following intermediate step at Eq. (1).[3]

$$\begin{aligned}
\text{Reg}_T &\leq \frac{\sup_{a \in \mathcal{A} F(a) - f(a_1)}}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} D_F\left(a_t, \tilde{a}_{t+1}\right) && (1) \\
&= \frac{\sup_{a \in \mathcal{A} F(a) - f(a_1)}}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} D_{F^*}\left(\nabla F(\tilde{a}_{t+1}, \nabla F(a_t))\right) \\
& && (D_f(x,y) = D_{f^*}(\nabla f(y), \nabla f(x))) \\
&= \frac{\sup_{a \in \mathcal{A} F(a) - f(a_1)}}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} D_{F^*}\left(\nabla F(a) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t)\right) \\
& && (\nabla F(\nabla F^*(x)) = x)
\end{aligned}$$

First we can show that, regardless of what regularization function and estimator we choose, we can always bound the penalty term as (By Lemma 3 in Zimmert et al. [2019])

$$\text{Reg}_{pen} \leq \mathbb{E}\left[\frac{-\Psi(x_1) + \Psi(x^*)}{\eta_1} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1})(-\Psi(x_t) + \Psi(x^*))\right]$$

[3]We use notations in Figure 1.

For any open convex set $\mathcal{D} \subset \mathbb{R}^d$ and its closure denoted $\bar{\mathcal{D}}$, for any Legendre F on $\bar{\mathcal{D}}$ define $F^*(x) := \sup_{y \in \bar{\mathcal{D}}} x^\top y - F(y)$.

Define $D_F(x, y) = F(x) - F(y) - (x - y)^\top \nabla F(y)$.

Let the Mirror Descent iterations satisfy, $a_1 = \arg\min_{a \in \mathcal{A}} F(a)$ then

$$\widetilde{a}_{t+1} = \nabla F^*(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t)) \tag{6.2}$$

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} D_F(a, \widetilde{a}_{t+1}) \tag{6.3}$$

where we have assumed the iterates exist.

Theorem 17 (Online Mirror Descent)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ be a closed convex action set, ℓ a subdifferentiable loss, and F a Legendre function defined on $\mathcal{A} \subset \bar{\mathcal{D}}$, such that $\nabla F(a) - \eta z \in dom(\nabla F(\mathcal{D}))$ for all $(a, z) \in \mathcal{A} \times \mathcal{Z}$ is satisfied. Then OMD satisfies*

$$\sup_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a_t, z_t) - \ell(a, z_t) \leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} D_{F^*}\left(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t)\right).$$

Online Mirror Descent with Linear Losses

Input: Time horizon T, convex action set $\mathcal{A} \subset \mathbb{R}^d$.

Initialize: Player sets $a_1 = \arg\min_{a \in \mathcal{A}} F(a)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0,1]^d$.

for: $t = 1, \cdots, T$

Player suffers (and observes) loss $\ell(a_t, z_t) = a_t^\top z_t$

Player observes z_t

Update mirror descent iterates:

$$\widetilde{a}_{t+1} = \nabla F^*(\nabla F(a_t) - \eta z_t)$$

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} D_F(a, \widetilde{a}_{t+1})$$

Figure 1: OMD and its analysis from Kevin's class notes

So now, if η_t is a constant over time, presented as η as shown in **Figure 1**, then the second term in upper bound goes to 0 and we get an upper bound $\mathbb{E}\left[\frac{-\Psi(x_1)+\Psi(x^*)}{\eta}\right]$, which is very similar to the bias term $\frac{\sup_{a \in A} F(s) - F(a_1)}{\eta}$. $(F = \Psi)$

Secondly we can show that, as long as the loss estimator is *unbiased* and the regularizer $\Psi(\cdot)$ is *convex*, for any $t_0 \geq 0$, we can always bound the stability term as

$$\text{Reg}_{stab} \leq \mathbb{E}\left[\sum_{t=t_0}^{T} D_{\Psi_t^*}\left(\nabla \Psi_t(x_t) - \hat{\ell}_t, \nabla \Psi_t(x_t)\right)\right] + 2\sum_{t=1}^{t_0} \|\nabla \Phi_t(-\hat{\ell}_t)\|_1$$

(Partially shown in proof of Lemma 2 Zimmert et al. [2019])

Recall that $\Phi_t(\cdot) = \max_{x \in \text{Conv}(\mathcal{X})} \{\langle x, \cdot \rangle - \Psi_t(x)\}$, the property of conjugate function tells us $\nabla \Phi_t(\cdot) = \text{argmax}_{x \in \text{Conv}(\mathcal{X})} \{\langle x, \cdot \rangle - \Psi_t(x)\}$ (see section 5.4.1 in Bertsekas [2009]). Since we assume $\mathcal{X}$ is bounded, $\left\|\nabla \Phi_t(-\hat{\ell}_t)\right\|_1$ is upper bounded by some constant m. If we further have η_t being a constant, we have

$$\text{Reg}_{stab} \leq \mathbb{E}\left[\sum_{t=t_0}^{T} \frac{1}{\eta} \cdot D_{\Psi_t^*}\left(\nabla \Psi(x_t) - \eta \hat{\ell}_t, \nabla \Psi(x_t)\right)\right] + 2t_0 m,$$

which is exactly similar to $\frac{1}{\eta}\sum_{t=1}^{T} D_{F^*}(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t))$ in the lecture notes.

3.2 A closer look at the regularization function

While the general framework of FTRL is fixed, the key for this class of algorithm is always *the choice of proper regularization function* as well as the corresponding learning rate and estimator.

EXP3: Exponential Weights for Exploration Exploitation

Input: Time horizon T, $\mathcal{A} = \{x \in \mathbb{R}^d : x_i \geq 0, \sum_{i=1}^d = 1\}$.
Initialize: Player sets $a_1 = (1/d, \ldots, 1/d)$. Adversary chooses $\{z_t\}_{t=1}^T \subset [0,1]^d$.
for: $t = 1, \cdots, T$

Player draws $I_t \sim a_t$ and suffers (and observes) loss $\ell(\mathbf{e}_{I_t}, z_t) = z_{t,I_t}$

Player sets $\widehat{z}_{t,i} = \frac{\mathbf{1}\{I_t = i\} z_{t,i}}{a_{t,i}}$

Update mirror descent iterates:

$$\widetilde{a}_{t+1,i} = \exp(-\eta \sum_{s=1}^{t} \widehat{z}_{s,i}) \qquad a_{t,i} = \widetilde{a}_{t+1,i} / \sum_{j=1}^{d} \widetilde{a}_{t+1,j}.$$

Corollary 6 (EXP3). *Under the conditions of Example 1 where the player can only play $\mathbf{e}_i$ for $i \in \{1, \ldots, d\}$ with $\ell(a, z) = a^\top z$ and only observe bandit feedback, the EXP3 algorithm satisfies*

$$\begin{aligned}\sup_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \ell(A_t, z_t) - \ell(a, z_t)\right] &\leq \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d} a_{t,i}\widehat{z}_{t,i}^2\right] \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta T d}{2} \leq \sqrt{2dT\log(d)}\end{aligned}$$

Proof. We can directly apply Corollary 2:

$$\begin{aligned}\mathbb{E}\left[\sup_{a \in \mathcal{A}}\sum_{t=1}^{T}(A_t - a)^\top z_t\right] &\leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T}\mathbb{E}\left[D_{F^*}\left(\nabla F(a_t) - \eta\widehat{z}_t, \nabla F(a_t)\right)\right] \\ &= \frac{\log(d)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d} a_{t,i}(\exp(-\eta\widehat{z}_{t,i}) - 1 + \eta\widehat{z}_{t,i})\right] \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d} a_{t,i}\widehat{z}_{t,i}^2\right] \\ &= \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{d} a_{t,i}\frac{\mathbf{1}\{I_t = i\} z_{t,i}^2}{a_{t,i}^2}\right] \\ &= \frac{\log(d)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} z_{t,i}^2 \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta d T}{2}\end{aligned}$$

Figure 2: Exp 3 algorithm and its analysis

Before we going to the regularization function chosen in this paper, that's first revisit the most common regularizer – the negative Shannon entropy regularizer previously talked in class as shown in **Figure 2**. Note that in the figure it said Exp3 because Exp3 is equivalent to FTRL with entropy regularizer in most cases. As we stated in Section 1, this algorithm works well for adversarial setting and achieves $\mathcal{O}(\sqrt{dT\log(d)})$, but it is not sufficient for getting a tighter gap-dependent bound for the stochastic setting. *How should we modify it ?*

By carefully examine the proof, we notice that the $\sqrt{T}$ actually comes from the stability term $\mathbb{E}\left[\sum_{i=1}^{d} a_{t,i}\frac{\mathbf{1}\{I_t = i\} z_{t,i}^2}{a_{t.i}^2}\right]$ as shown in the red circle. Written this using our notation, it is equivalent to say the $\sqrt{T}$ comes from the fact that

$$\left[\sum_{i=1}^{d} x_{t,i}\frac{\mathbf{1}\{X_{t,i} = 1\} l_{t,i}^2}{x_{t,i}^2}\right] = o(1)$$

So the idea is to find a better regularizer that decrease this stability term. For example, Can we find an regularizer that change the stability term into something close to

$$\left[\sum_{i=1}^{d} x_{t,i}\frac{\mathbf{1}\{X_{t,i} = 1\} l_{t,i}^2}{x_{t,i}^{\frac{3}{2}}}\right] = o\left(\sum_{i=1}^{d}\sqrt{x_{t,i}}\right) \textbf{?}$$

As a trade of, we need to slightly increase the penalty term, for example, from $\frac{\log(d)}{\eta}$ to $\frac{d}{\eta}$. But such increment exactly matches the lower bound of stochastic setting, so it is totally ok. With this in mind, we introduce a class of regularization functions called Tsallis entropy regularization.

P.S. To more rigorously illustrate why this form is good requires more analysis as well as the choice of learning rate. So we left that in the later sections. Here just try to give an intuition why we need better regularizer.

Intro to α-Tsallis-INF entropy The original α-Tsallis entropy is defined as $H_\alpha(x) := \frac{1}{1-\alpha}\left(1 - \sum_i x_i^\alpha\right)$. Here by a little bit of modification, we define the α-Tsallis-INF entropy as

$$\Psi(x) := -\sum_i \frac{w_i^\alpha - \alpha w_i - (1-\alpha)}{\alpha(1-\alpha)\xi_i}$$

For now, you can think $\xi_i = 1$. In Zimmert and Seldin [2021], they actually provide some analysis on how the choice of ξ_i can be influence by gap. But we will not expend this complicated analysis here.

The magic thing for this regularizer is, by taking $\alpha \to 1$, we can exactly recover the negative Shannon entropy since

$$\lim_{\alpha\to 1} \Psi(x) = \sum_i \xi_i^{-1}(x_i \log(x_i) - x_i + 1).$$

It also gives interesting results in many cases by taking $\alpha \to 0$. Please refer to Wei and Luo [2018] for details if you are interested and we will not go into details here.

In this week's literature, we will focus on Tsallis-INF entropy with $\alpha = \frac{1}{2}$. That is,

$$\Psi(x) = -\sum_i \sqrt{x}_i$$

The other terms can be ignored because we always consider x to be some discrete probability distribution. We also ignore the leading constant because it can be incorporated into learning rate.

Intro to Hybrid regularizer Now by considering from a class α-Tsallis-INF entropy, we are very close to the final entropy we want. In fact, for the simple MAB case, this $\Psi(x) = -\sum_i \sqrt{x}_i$ can already render a good result as shown in Zimmert and Seldin [2021], Masoudian and Seldin [2021]. For more complicated structure, like semi-bandits, we need to consider a hybrid regularizer, which is the combination of several kind of α-Tsallis-INF entropy with different choice of α. In Zimmert et al. [2019], the authors use

$$\Psi(x) = \sum_{i=1}^{d} -\sqrt{x}_i + \gamma(1 - x_i)\log(1 - x_i) \tag{2}$$

where γ is some constant parameter. This is a hybrid regularizer with a combination of negative Shannon entropy and $\frac{1}{2}$-Tsallis-INF entropy. The reason for choosing such regularizer will be explained in the following section.

4 A detailed analysis on semi-bandits result

4.1 Algorithm and result

The major novelty of this work is its hybrid regularizer as shown in Eq. (2). To fully specify the Algorithm 1, the authors use learning rate $\eta_t = 1/\sqrt{t}$ and loss estimator

$$\hat{\ell}_t = \frac{(o_{ti} + 1)\mathbb{1}\{X_{ti} = 1\}}{x_{ti}} - 1. \tag{3}$$

It is clear $\mathbb{E}[\hat{\ell}_t] = \ell_t$ and the shift by 1 is introduced in order to ensure $\hat{\ell}_t \geq -1$, which will be used in later proof.

Before presenting its major theorem, several new notations need to be introduced. First, for any concrete instance of semi-bandit with action set $\mathcal{X}$ and optimal action x^*, we define two functions $f, g : \mathcal{X} \mapsto \mathbb{R}$ as

$$f(x) = \sum_{i:x_i^*=0} \sqrt{x_i}$$

$$g(x) = \sum_{i:x_i^*=1} \left(\gamma^{-1} - \gamma \log(1 - x_i)\right)(1 - x_i).$$

For stochastic environment, we also define the instantaneous regret function $r : [0, \infty)^{|\mathcal{X}|} \mapsto \mathbb{R}$ as

$$r(\alpha) = \sum_{x \in \mathcal{X} \setminus \{x^*\}} \alpha_x \Delta_x.$$

Further, for any $\alpha \in [0, \infty)^{|\mathcal{X}|}$, we define $\overline{\alpha} = \sum_{x \in \mathcal{X}} \alpha_x x$ and let $\Delta(\mathcal{X})$ denote the simplex of distribution over $\mathcal{X}$. Then, the major theorem of this paper is summarized as the following

Theorem 1. *For any $\gamma \in (0, 1]$, by applying Algorithm 1 with regularizer $\Psi(\cdot)$ in Eq.* (2)*, estimator $\hat{\ell}_t$ in Eq.* (3) *and learning rate $\eta_t = 1/\sqrt{t}$, the pseudo-regret is upper bounded by*

$$\overline{\text{Reg}}_T \leq \mathcal{O}\left(C_{\text{sto}} \log(T)\right) + \mathcal{O}\left(C_{\text{add}}\right)$$

in the stochastic case and

$$\overline{\text{Reg}}_T \leq \mathcal{O}\left(C_{\text{adv}} \sqrt{T}\right)$$

in the adversarial case. Here, C_{sto}, C_{add} and C_{adv} are defined as

$$\begin{aligned} C_{\text{sto}} &= \max_{\alpha \in [0,\infty)^{|\mathcal{X}|}} \left\{ f(\overline{\alpha}) - r(\alpha) \right\}, \\ C_{\text{add}} &= \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(\mathcal{X})} \left(\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha) \right), \\ C_{\text{adv}} &= \max_{x \in \text{Conv}(\mathcal{X})} \left\{ f(x) + g(x) \right\}. \end{aligned}$$

Moreover, by defining $m = \max_{x \in \mathcal{X}} \|x\|_1$ and $\Delta_{\min} = \min_{x \in \mathcal{X} \setminus \{x^*\}} \Delta_x$, it always holds that $C_{\text{sto}} = \mathcal{O}\left(\frac{md}{\Delta_{\min}}\right)$, $C_{\text{add}} = \mathcal{O}\left(\frac{m^2}{\gamma^2 \Delta_{\min}}\right)$ and $C_{\text{adv}} = \mathcal{O}\left(\frac{\sqrt{md}}{\gamma}\right)$.

Here, we will provide a proof overview. For full details, please refer to Zimmert et al. [2019].

4.2 A proof overview

Recall that in Section 3.1, we state that for all the FTRL framework, its regret can be divided into the stability and penalty terms:

$$\begin{aligned} \overline{\text{Reg}}_T &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, l_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{\text{Reg}_{stab}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \langle x^*, \ell_t \rangle\right]}_{\text{Reg}_{pen}} \\ &\leq \mathbb{E}\left[\frac{-\Psi(x_1) + \Psi(x^*)}{\eta_1} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_t^{-1})(-\Psi(x_t) + \Psi(x^*))\right] \\ &\quad + \mathbb{E}\left[\sum_{t=t_0}^{T} D_{\Psi_t^*}\left(\nabla \Psi_t(x_t) - \hat{\ell}_t, \nabla \Psi_t(x_t)\right)\right] + 2t_0 m \end{aligned}$$

So **the first step** is to show that both terms can be upper bounded by

$$\text{Reg}_{pen} \leq \sum_{t=1}^{T} \frac{3}{2\sqrt{t}} \left(\sum_{i: x_i^* = 0} \sqrt{\mathbb{E}[x_{ti}]} - \sum_{i: x_i^* = 1} \gamma (1 - \mathbb{E}[x_{ti}]) \log(1 - \mathbb{E}[x_{ti}]) \right) \tag{4}$$

$$\text{Reg}_{stab} \leq \sum_{t=1}^{T} \frac{16\sqrt{2}}{\sqrt{t}} \left(\sum_{i: x_i^* = 0} \sqrt{\mathbb{E}[x_{ti}]} + \sum_{i: x_i^* = 1} \gamma^{-1} (1 - \mathbb{E}[x_{ti}]) \right) \tag{5}$$

Notice that this $\sum_{i: x_i^*=0} \sqrt{\mathbb{E}[x_{ti}]}$ term is the key to achieve best-of-both-worlds, while by using the negative Shannon entropy (Exp3) we can only hope to get some $o(1)$ upper bound.

The key difficulty here is, when bound the $D_{\Psi_t^*}\left(\nabla \Psi_t(x_t) - \hat{\ell}_t, \nabla \Psi_t(x_t)\right)$, you need to bound $\nabla \Psi_t^*(\nabla \Psi_t(x) - \hat{\ell}_t)$ in terms of x_i.

> ***Lemma 1:*** *Consider a strategy whose regret with respect to the optimal action i^* is upper bounded by*
>
> $$C \sum_{t=1}^{T} \sum_{i \neq i^*} \sqrt{\frac{x_{i,t}}{t}} . \quad (2)$$
>
> *(Recall that for multi-armed bandit one selects a probability distribution x_t over the actions, so $x_{i,t}$ denote here the probability of playing action i at time t.) Then one has that the regret is in fact bounded by $2C\sqrt{KT}$ (this follows trivially by Jensen on the i sum), and moreover if the environment is stochastic one has that the regret is in fact bounded by C^2 times (1).*

Proof: Assuming that the environment is stochastic we can write the regret as $\sum_{i,t} \Delta_i x_{i,t}$, so by assumption and using that $C\sqrt{\frac{x_{i,t}}{t}} \leq \frac{1}{2}\left(\Delta_i x_{i,t} + \frac{C^2}{t\Delta_i}\right)$ one has:

$$\sum_{i \neq i^*, t} \Delta_i x_{i,t} \leq \frac{1}{2} \sum_{i \neq i^*, t} \left(\Delta_i x_{i,t} + \frac{C^2}{t \Delta_i} \right) ,$$

which means that the left hand side is smaller than $\sum_{i \neq i^*, t} \frac{C^2}{t\Delta_i}$ which is indeed smaller than C^2 times (1).

The second step is to show that as long as the above inequalities are satisfied, we can always using the *self-bounding technique* to get a desired stochastic bound. Of course, we also show that we can get a desired adversarial bound which is very direct.

Here we copy and paste a MAB version from Sebastian Bubeck's blog to give you an intuition on the *self-bounding technique*

4.3 Step 1

4.3.1 Penalty Term

In this section, we aim to show that the regularization penalty term is upper bounded by (4). We first cite the following standard result from FTRL.

Lemma 1. *The penalty term can be upper bounded as follows:*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \langle x^*, \ell_t \rangle\right] \\
&\leq \mathbb{E}\left[\frac{-\Psi(x_1) + \Psi(x^*)}{\eta_1} + \sum_{t=2}^{T} \left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\left(-\Psi(x_t) + \Psi(x^*)\right)\right].
\end{aligned}$$

With this result, we are able to prove the upper bound.

Proof. We directly plug into Lemma 1 the learning rate of $\eta_t = \frac{1}{\sqrt{t}}$ and the regularizer as (2). Since $\gamma \leq 1$ and $-(1-x)\log(1-x) \leq \frac{\sqrt{x}}{2}$ for $x \in [0,1]$, we get

$$
\begin{aligned}
-\Psi\left(x_t\right)+\Psi\left(x^*\right) &= \sum_{i=1}^{d} \sqrt{x_{ti}}-\gamma\left(1-x_{ti}\right)\log\left(1-x_{ti}\right)-\sum_{i:x_i^*=1}\sqrt{1} \\
&\leq \sum_{i:x_i^*=0}\frac{3}{2}\sqrt{x_{ti}}-\sum_{i:x_i^*=1}\gamma\left(1-x_{ti}\right)\log\left(1-x_{ti}\right) \\
&\leq \frac{3}{2}\left(\sum_{i:x_i^*=0}\sqrt{x_{ti}}-\sum_{i:x_i^*=1}\gamma\left(1-x_{ti}\right)\log\left(1-x_{ti}\right)\right).
\end{aligned}
$$

It further holds that $\eta_1=1=\eta_1^{-1}$ and

$$
\eta_t^{-1}-\eta_{t-1}^{-1}=\sqrt{t}-\sqrt{t-1}\leq\frac{1}{2\sqrt{t-1}}\leq\frac{1}{\sqrt{t}}=\eta_t.
$$

Insert everything into Lemma 1 yields

$$
\begin{aligned}
\mathrm{Reg}_{\text{pen}} &\leq \mathbb{E}\left[\frac{-\Psi\left(x_1\right)+\Psi\left(x^*\right)}{\eta_1}+\sum_{t=2}^{T}\left(\eta_t^{-1}-\eta_{t-1}^{-1}\right)\left(-\Psi\left(x_t\right)+\Psi\left(x^*\right)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\eta_t\left(-\Psi\left(x_t\right)+\Psi\left(x^*\right)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{3}{2\sqrt{t}}\left(\sum_{i:x_i^*=0}\sqrt{x_{ti}}-\sum_{i:x_i^*=1}\gamma\left(1-x_{ti}\right)\log\left(1-x_{ti}\right)\right)\right] \\
&\leq \sum_{t=1}^{T}\frac{3}{2\sqrt{t}}\left(\sum_{i:x_i^*=0}\sqrt{\mathbb{E}\left[x_{ti}\right]}-\sum_{i:x_i^*=1}\gamma\left(1-\mathbb{E}\left[x_{ti}\right]\right)\log\left(1-\mathbb{E}\left[x_{ti}\right]\right)\right)
\end{aligned}
$$

where the last step follows from Jensen's inequality and the concavity of functions $\sqrt{x}$ and $-(1-x)\log(1-x)$. ☐

4.3.2 Stability Term

We first observe that the stability term is essentially the Bregman divergence. Usually it is very complicated to directly calculate the value of $D_{\Psi_t^*}(x,y)$. A common trick is to using the following property, which might be more intuitive to people

Lemma 2. *For any $x,y\in\mathbb{R}^f$, then*

$$
D_{\Psi_t^*}(x,y)\leq\frac{1}{2}\max_{x\in\mathcal{A}}\|x-y\|^2_{\nabla^2\Psi_t^{-1}(x)},
$$

where $\mathcal{A}=\bigotimes_{i=1}^{d}[x_i,y_i]$

Proof. By Talyor theorem, for any $x,y\in\mathbb{R}^f$ there exists a $z\in Conv(\{x,y\})$ such that

$$
D_{\Psi_t^*}(x,y)=\frac{1}{2}\|x-y\|^2_{\nabla^2\Psi_t^*(z)}=\frac{1}{2}\|x-y\|^2_{\nabla^2\Psi_t^{-1}(z)}
$$

where Ψ_t^* is the dual function of Ψ_t. Finally you just apply the coordinate wise monotonicity to get the final bound. ☐

Equip with this property, we can upper bound the stability term as

$$
\mathbb{E}\left[\sum_{t=t_0}^{T}D_{\Psi_t^*}\left(\nabla\Psi_t(x_t)-\hat{\ell}_t,\nabla\Psi_t(x_t)\right)\right]\leq\frac{1}{2}\mathbb{E}\left[\sum_{t=t_0}^{T}\max_{x\in\mathcal{A}_t}\|\hat{\ell}_t\|^2_{\nabla^2\Psi_t^{-1}(x)}\right]
$$

where $\tilde{x}_t = \nabla\Psi^*(\nabla\Psi(x_t) - \hat{\ell}_t)$ and $\mathcal{A}_t = \bigotimes_{i=1}^d [x_{ti}, \tilde{x}_{ti}]$.

So the key step here is to find $\text{argmax}_{x\in\mathcal{A}_t} \|x-y\|^2_{\nabla^2\Psi_t^{-1}(x)}$. In fact, if we are using the Shannon entropy, then this upper bound simply goes to

$$\frac{1}{2}\mathbb{E}\left[\sum_{t=t_0}^{T} \|\hat{l}_t\|^2_{\nabla^2\Psi_t^{-1}(x_t)}\right]. \text{ (you can prove it yourself)}$$

But it is not clear whether the hybrid bound we presented here has a similar property. So in the following we show the key lemma saying $\tilde{x}_t$ is always close to $\tilde{x}$.

Lemma 3. *If $\eta_t \leq \min\left\{\frac{\sqrt{2}-1}{2}, \frac{\gamma\log(2)}{4}\right\}$, then for any $x \in (0,1)^d$ and $\hat{\ell}$ such that $-1 \leq \hat{\ell}_i \leq \frac{2}{x_i}$ for all i, we have*

$$2x_i - 1 \leq \nabla\Psi_t^*\left(\nabla\Psi_t(x) - \hat{\ell}\right)_i \leq 2x_i.$$

Proof sketch. The functions $\nabla\Psi_t$ and $\nabla\Psi_t^*$ are symmetric and independent in each dimension. Therefore it is sufficient to consider $d = 1$ and drop the index i.

Note that we have $\nabla\Psi_t = (\nabla\Psi_t^*)^{-1}$ by definition of conjugate function, we have

$$\hat{\ell} = \nabla\Psi_t(x) - \nabla\Psi_t\left(\nabla\Psi_t^*\left(\nabla\Psi_t(x) - \hat{\ell}\right)\right) \in \left[-1, \frac{2}{x}\right].$$

Recall that

$$\nabla\Psi_t(x) = \eta_t^{-1}\left(-\frac{1}{2\sqrt{x_i}} - \gamma\log(1-x_i) - \gamma\right)_{i=1,\ldots,d},$$

it is a strictly increasing function with finite derivative for $x_i \in (0,1)$. Therefore, the difference in the argument cannot be too large, which shows that $\nabla\Psi_t^*\left(\nabla\Psi_t(x) - \hat{\ell}\right)$ is close to x. With some technical computation omitted here the statement is proved.

□

With some technical work omitted here, we get the final bound for the stability term as

$$\text{Reg}_{\text{stab}} \leq \sum_{t=1}^{T} \frac{16\sqrt{2}}{\sqrt{t}}\left(\sum_{i:x_i^*=0} \sqrt{\mathbb{E}\left[x_{ti}\right]} + \sum_{i:x_i^*=1} \gamma^{-1}\left(1-\mathbb{E}\left[x_{ti}\right]\right)\right) + c$$

where $c = 58\gamma^{-2}\max_{x\in\mathcal{X}} \|x\|_1$.

4.4 Step 2: Towards best of both worlds regret bounds

We first show the bound in adversarial case. By simply adding the two upper bounds in Eq. (4) and (5), we can have

$$\begin{aligned}
\overline{\text{Reg}}_T &\leq \sum_{t=1}^{T} \frac{16\sqrt{2}+1.5}{\sqrt{t}}\left(\sum_{i:x_i^*=0}\sqrt{\mathbb{E}[x_{ti}]} + \sum_{i:x_i^*=1}\left(\gamma^{-1} - \gamma\log\left(1-\mathbb{E}[x_{ti}]\right)\right)\left(1-\mathbb{E}[x_{ti}]\right)\right) + c \\
&= \sum_{t=1}^{T} \frac{25}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c \\
&\leq 50\sqrt{T}\max_{x\in\text{Conv}(\mathcal{X})}\{f(x)+g(x)\} + c \\
&\leq \mathcal{O}\left(C_{\text{adv}}\sqrt{T}\right).
\end{aligned}$$

For stochastic case, let $p_{x,t} = P(\mathbb{E}[x_t])_x$. Since $P(\cdot)$ is designed to be an unbiased sampling probability, we have $\sum_{x\in\mathcal{X}} p_{x,t}x = \mathbb{E}[x_t]$. Therefore, we have

$$\begin{aligned} r(P(\mathbb{E}[x_t])) &= \sum_{x\in\mathcal{X}\setminus\{x^*\}} p_{x,t}\Delta_x \le \sum_{x\in\mathcal{X}} p_{x,t}\mathbb{E}\left[\langle x - x^*, \ell_t\rangle\right] \\ &= \mathbb{E}\left[\left\langle \sum_{x\in\mathcal{X}} p_{x,t}x - x^*, \ell_t \right\rangle\right] = \mathbb{E}\left[\langle \mathbb{E}[x_t] - x^*, \ell_t\rangle\right]. \end{aligned}$$

As a result, we have

$$\begin{aligned} &\overline{\mathrm{Reg}}_T = \mathbb{E}\left[\sum_{t=1}^T \langle \mathbb{E}[x_t] - x^*, \ell_t\rangle\right] \ge \sum_{t=1}^T r(P(\mathbb{E}[x_t])) \\ &\implies \sum_{t=1}^T \frac{25}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c - \sum_{t=1}^T r(P(\mathbb{E}[x_t])) \ge 0 \end{aligned}$$

$$\begin{aligned} \implies \overline{\mathrm{Reg}}_T &\le \sum_{t=1}^T \left[\frac{50}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) - r(P(\mathbb{E}[x_t]))\right] + 2c \\ &= \underbrace{\sum_{t=1}^T \left[\frac{50}{\sqrt{t}} f(\mathbb{E}[x_t]) - \frac{1}{2} r(P(\mathbb{E}[x_t]))\right]}_{\text{term A}} + \underbrace{\sum_{t=1}^T \left[\frac{50}{\sqrt{t}} g(\mathbb{E}[x_t]) - \frac{1}{2} r(P(\mathbb{E}[x_t]))\right]}_{\text{term B}} + 2c. \end{aligned}$$

We can then bound term A and B separately. Specifically,

$$\begin{aligned} \text{term A} &\le \sum_{t=1}^T \max_{\alpha\in\Delta(\mathcal{X})} \left\{\frac{50}{\sqrt{t}} f(\overline{\alpha}) - \frac{1}{2} r(\alpha)\right\} \\ &\le \sum_{t=1}^T \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} \left\{\frac{50}{\sqrt{t}} f\left(\frac{10^4}{t}\overline{\alpha}\right) - \frac{1}{2} r\left(\frac{10^4}{t}\alpha\right)\right\} \\ &= \sum_{t=1}^T \frac{10^4}{2t} \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} \left\{f(\overline{\alpha}) - r(\alpha)\right\} \qquad \text{(Since } f(ax) = \sqrt{a}f(x) \text{ and } r(ax) = ar(x)\text{)} \\ &= \mathcal{O}\left(C_{\mathrm{sto}} \log(T)\right). \end{aligned}$$

$$\begin{aligned} \text{term B} &\le \frac{1}{2}\sum_{t=1}^T \max_{\alpha\in\Delta(\mathcal{X})} \left\{\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha)\right\} \\ &\le \frac{1}{2}\sum_{t=1}^\infty \max_{\alpha\in\Delta(\mathcal{X})} \left\{\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha)\right\} \\ &= \mathcal{O}\left(C_{\mathrm{add}}\right). \end{aligned}$$

The last inequality above holds because for $\alpha = \mathbf{e}_{x^*}$, $r(\mathbf{e}_{x^*}) = 0$, which implies that $\max_{\alpha\in\Delta(\mathcal{X})} \left\{\frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha)\right\} \ge 0$ for any $t \ge 1$.

By plugging the bounds for term A and B back, we can get

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(C_{\mathrm{sto}}\log(T)\right) + \mathcal{O}\left(C_{\mathrm{add}}\right).$$

4.5 Bounds for time-independent constants

Finally, to conclude the theorem, we briefly explain why C_{adv}, C_{sto} and C_{add} have the claimed magnitudes. For C_{adv}, it is bounded as the following

Bounding $\mathrm{C}_{\mathrm{adv}}$:

$$
\begin{aligned}
C_{adv} &= \max_{x\in\mathrm{Conv}(\mathcal{X})} \sum_{i:x_i^*=0} \sqrt{x_i} + \sum_{i:x_i^*=1} (\gamma^{-1} - \gamma\log(1-x_i))(1-x_i) \\
&\le \max_{x\in\mathrm{Conv}(\mathcal{X})} \sum_{i:x_i^*=0} \sqrt{x_i} + \sum_{i:x_i^*=1} \gamma\sqrt{1-x_i} + \sum_{i:x_i^*=1} \gamma^{-1}(1-x_i) && (-y\log y \le \sqrt{y} \text{ for } y\in[0,1]) \\
&\le \max_{x\in\mathrm{Conv}(\mathcal{X})} \sqrt{\left(\sum_{i:x_i^*=0} 1\right)\left(\sum_{i:x_i^*=0} x_i\right)} + \gamma\sqrt{\left(\sum_{i:x_i^*=1} 1\right)\left(\sum_{i:x_i^*=1} (1-x_i)\right)} + \gamma^{-1}m && \text{(Cauchy-Schwarz)} \\
&\le \sqrt{dm} + \gamma m + \gamma^{-1} m \\
&\le \mathcal{O}\left(\gamma^{-1}\sqrt{md}\right).
\end{aligned}
$$

For C_{sto}, by using Cauchy-Schwartz inequality, we have

$$
f(\overline{\alpha}) = \sum_{i:i^*=0} \sqrt{\sum_{x\in\mathcal{X}} \alpha_x x_i} \le \sqrt{dm \sum_{x\in\mathcal{X}\setminus\{x^*\}} \alpha_x}.
$$

Meanwhile, we have $r(\alpha) \ge \Delta_{\min} \sum_{x\in\mathcal{X}\setminus\{x^*\}} \alpha_x$. Therefore, it holds that

$$
\begin{aligned}
C_{\mathrm{sto}} &\le \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} \left\{ \sqrt{dm \sum_{x\in\mathcal{X}\setminus\{x^*\}} \alpha_x} - \Delta_{\min} \sum_{x\in\mathcal{X}\setminus\{x^*\}} \alpha_x \right\} \\
&\le \max_{A\ge 0} \sqrt{dmA} - \Delta_{\min} A \\
&= \frac{dm}{4\Delta_{\min}}. && \text{(By solving this univariate optimization problem)}
\end{aligned}
$$

Finally, for C_{add}, we will only give a brief argument and please refer to section A.3 in Zimmert et al. [2019] for the full proof. First, to give an upper bound for $g(\overline{\alpha})$, we need the following three facts:

- For any $y\in\mathbb{R}^N$, $\sum_{i=1}^N y_i \log\frac{1}{y_i} \le \|y\|_1 \log\frac{1}{\|y\|_1}$.
- $y \mapsto y\left(\gamma^{-1} + \gamma\log\frac{m}{y}\right)$ is increasing in y if $y\in[0,m]$.
- $\sum_{i:x_i^*=1}(1-\overline{\alpha}_i) \le m\left(\sum_{x\in\mathcal{X}\setminus\{x^*\}}\alpha_x\right)$.

With these facts, we can have

$$
g(\overline{\alpha}) \le m\left(\sum_{x\in\mathcal{X}\setminus\{x^*\}}\alpha_x\right)\left(\gamma^{-1} + \log\frac{1}{\sum_{x\in\mathcal{X}\setminus\{x^*\}}\alpha_x}\right).
$$

By using the same bound for $r(\alpha)$, we have

$$
C_{\mathrm{add}} \le \sum_{t=1}^{\infty} \max_{A\in[0,1]} \left\{ \frac{100}{\sqrt{t}} mA\left(\gamma^{-1} + \gamma\log\frac{1}{A}\right) - \Delta_{\min}A \right\}
$$

The final bound can be obtained by using the two following results:

$$
\begin{aligned}
&\sum_{t=1}^{\infty} \max_{A\in[0,1]} \left\{ \frac{100}{\sqrt{t}} mA\gamma^{-1} - \frac{1}{2}\Delta_{\min}A \right\} \le \mathcal{O}\left(\frac{m^2}{\gamma^2\Delta_{\min}}\right), \\
&\sum_{t=1}^{\infty} \max_{A\in[0,1]} \left\{ \frac{100}{\sqrt{t}} mA\gamma\log\frac{1}{A} - \frac{1}{2}\Delta_{\min}A \right\} \le \mathcal{O}\left(\frac{m^2\gamma^2}{\Delta_{\min}}\right).
\end{aligned}
$$

In brief, the first bound is obtained by noticing that as t increases, it will eventually become a finite sum; the second bound is obtained by first solving the univariate optimization problem and then approximating the infinite sum by an integral.

5 Corruption

Corruption model. The setting is a simplification from Lykouris et al. [2018]. Consider a stochastic bandit with K arms, and an adversary who can corrupt some of the stochastic rewards. More formally, the protocol is as follows. At each round $t = 1, \cdots, T$:

1. The learner picks a distribution w_t over the K arms.
2. The stochastic loss $\ell_t^s(a)$ are drawn from each arm.
3. The stochastic loss $\ell_t^s(a)$ as well as the choice of the learner in previous steps a_{t-1} is observed by the adversary and returns a (possibly) corrupted reward $\ell_t^c(a) \in [0, 1]$.
4. The learner draws arm $a_t \sim w_t$ and observes $\ell_t^c(a_t)$.

We call the instance C-corrupted if the adversary can corrupt at most C of the total T rounds, i.e.

$$\sum_{t=1}^{T} \mathbb{1}\left\{\ell_t^c \neq \ell_t^s\right\} \leq C.$$

Theorem 2. *For any $\gamma \in (0, 1]$, by applying Algorithm 1 with the same setting as in Theorem 1 to a C-corrupted environment, the pseudo-regret is upper bounded by*

$$\overline{\mathrm{Reg}}_T \leq \mathcal{O}\left(C_{\mathrm{adv}}\sqrt{C} + C_{\mathrm{sto}}\log(T) + C_{\mathrm{add}}\right).$$

Proof. Recall from our previous analysis, regardless of the environment, it holds that

$$\begin{aligned}
\overline{\mathrm{Reg}}_T &\leq \sum_{t=1}^{T} \frac{25}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c \\
&= \sum_{t:\ell_t \text{ corrupted}} \frac{25}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + \sum_{t:\ell_t \text{ stochastic}} \frac{25}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c \\
&\leq \mathcal{O}\left(C_{\mathrm{adv}}\sqrt{C} + C_{\mathrm{sto}}\log(T) + C_{\mathrm{add}}\right).
\end{aligned}$$

The term $\mathcal{O}\left(C_{\mathrm{adv}}\sqrt{C}\right)$ appears for obvious reason. Meanwhile, the term $\mathcal{O}\left(C_{\mathrm{sto}}\log(T) + C_{\mathrm{add}}\right)$ can be obtained by using exactly the same analysis as in step 2 by simply replacing "$\sum_{t=1}^{T}$" by "$\sum_{t:\text{ stochastic}}$". ☐

References

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits, 2021.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of robustness of the tsallis-inf algorithm to adversarial corruptions in stochastic multiarmed bandits, 2021.

Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits, 2018.

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