Parameter-Free Multi-Armed Bandit Algorithms with Hybrid Data-Dependent Regret Bounds

Shinji Ito I-SHINJI@NEC.COM
NEC Corporation

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

This paper presents multi-armed bandit (MAB) algorithms that work well in adversarial environments and that offer improved performance by exploiting inherent structures in such environments, as stochastic generative models, as well as small variations in loss vectors. The fundamental aim of this work is to overcome the limitation of worst-case analyses in MAB contexts. There can be found two basic approaches achieving this purpose: best-of-both-worlds algorithms that work well for both stochastic and adversarial settings, and data-dependent regret bounds that work well depending on certain difficulty indicators w.r.g. loss sequences. One remarkable study w.r.t. the best-of-both-worlds approach deals with the Tsallis-INF algorithm (Zimmert and Seldin, 2021), which achieves nearly optimal regret bounds up to small constants in both settings, though such bounds have remained unproven for a special case of a stochastic setting with multiple optimal arms.

This paper offers two particular contributions: (i) We show that the Tsallis-INF algorithm enjoys a regret bound of a logarithmic order in the number of rounds for stochastic environments, even if the best arm is not unique. (ii) We provide a new algorithm with a new *hybrid* regret bound that implies logarithmic regret in the stochastic regime and multiple data-dependent regret bounds in the adversarial regime, including bounds dependent on cumulative loss, total variation, and loss-sequence path-length. Both our proposed algorithm and the Tsallis-INF algorithm are based on a follow-the-regularized-leader (FTRL) framework with a time-varying regularizer. The analyses in this paper rely on *skewed Bregman divergence*, which provides simple expressions of regret bounds for FTRL with a time-varying regularizer.

Keywords: multi-armed bandit, Tsallis-INF, optimistic follow the regularized leader, best-of-both-worlds, path-length regret bound

1. Introduction

This paper considers the multi-armed bandit (MAB) problem in which a player is given a set of actions $[K] = \{1, 2, \dots, K\}$, sequentially chooses actions $i^t \in [K]$, and then observes loss $\ell^t_{i^t} \in [0, 1]$ at each time step $t \in [T]$, where $\ell^t = (\ell^t_1, \ell^t_2, \dots, \ell^t_K)^\top \in [0, 1]^K$ is the *loss vector*, of which i-th element represents the loss for choosing the i-th action in the t-th round. The goal of the player is to minimize regret R^T, defined as

$$R^T_{i^*} = \mathbf{E}\left[\sum_{t=1}^{T} \ell^t_{i^t} - \sum_{t=1}^{T} \ell^t_{i^*}\right], \quad R^T = \max_{i^* \in [K]} R^T_{i^*} \tag{1}$$

where $\mathbf{E}[\cdot]$ refers to the expectation taken w.r.t. the loss sequence ℓ^t and algorithms' internal randomness.[1]

1. R^T defined here is also called pseudo-regret or expected regret in some literature.

Studies on MAB can generally be divided into two categories: those for *stochastic settings* and those for *adversarial settings*. In a stochastic setting, loss ℓ^t is assumed to follow an unknown distribution i.i.d. for all t, i.e., the environment is assumed to be time-invariant. There are algorithms achieving $O\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right)$-regret for this setting, where Δ_i stands for the *suboptimality gap* of the i-th action (Lai and Robbins, 1985; Auer et al., 2002a). In an adversarial setting, no generative models for ℓ^t are assumed but ℓ^t may behave adversarially against the player. For this setting, the tight regret bound is $\Theta(\sqrt{KT})$ (Audibert and Bubeck, 2009; Auer et al., 2002b).

Though the adversarial model captures broader problem settings, the worst-case optimality in the adversarial regime (i.e., $O(\sqrt{KT})$-regret) does not necessarily imply practical advantages over other algorithms in many applications, as, essentially, "worst-case" scenarios are quite rare. For example, if the environment is time-invariant, i.e., if losses follow i.i.d. distributions, $O(\sqrt{KT})$-regret will not be as good as the $O(\sum_{i\Delta_i>0} \frac{\log T}{\Delta_i})$-regret achieved by stochastic MAB algorithms. On the other hand, many stochastic MAB algorithms may perform poorly in the adversarial setting.

Two types of approaches have been studied for the purpose of overcoming the limitations of worst-case optimality. One is to design *best-of-both-worlds* algorithms (Zimmert and Seldin, 2021; Jin and Luo, 2020; Seldin and Lugosi, 2017; Seldin and Slivkins, 2014; Bubeck and Slivkins, 2012; Auer and Chiang, 2016; Zimmert et al., 2019; Jin and Luo, 2020; Mourtada and Gaïffas, 2019) that work (nearly) optimally both for the stochastic setting and for the adversarial counterpart. The other is to develop adversarial MAB algorithms with *data-dependent regret bounds* (Hazan and Kale, 2011; Allenberg et al., 2006; Bubeck et al., 2019; Wei and Luo, 2018; Lee et al., 2020; Wei et al., 2020), which show improved performance for environments with some "benign" properties, e.g., little variation in the loss sequence or small cumulative loss.

One remarkable study w.r.t. the best-of-both-worlds approach deals with the Tsallis-INF algorithm (Zimmert and Seldin, 2021). This algorithm has optimal regret bounds up to small constant factors in stochastic settings as well as in adversarial settings, which hold *anytime*, i.e., are valid for unknown time horizons. Even more surprisingly, this algorithm works well for *mixed* environments of stochastic and adversarial models. More precisely, the algorithm enjoys regret bounds for an *adversarial regime with a self-bounding constraint*, which implies, e.g., an $O\left(\sum_{i\neq i^*} \frac{\log T}{\Delta_i} + \sqrt{T' \sum_{i\neq i^*} \frac{\log T}{\Delta_i}}\right)$-regret bound in a stochastic setting with T'-rounds of adversarial corruptions. One remaining issue regarding this algorithm is the regret bound for a stochastic setting in which the optimal arm is not unique. In such cases with multiple optimal arms, it has remained unproven whether or not the algorithm has $O(\log T)$-regret bounds, while empirical evaluation suggests $O(\log T)$-regret.

Designing algorithms with data-dependent regret bounds is another approach to overcoming the limitations of the worst-case analysis, one for which performance may improve depending on the *difficulty indicators* of loss sequences. Examples of difficulty indicators include the following:

- Cumulative loss for optimal arm: $L_{i^*} = \sum_{t=1}^T \ell_{i^*}^t = \min_{i\in[K]} \sum_{t=1}^T \ell_i^t$.
- Total variation in losses: $Q_\infty = \sum_{t=1}^T \|\ell^t - \bar{\ell}\|_\infty^2$, $Q_2 = \sum_{t=1}^T \|\ell^t - \bar{\ell}\|_2^2$, $Q_{i^*} = \sum_{t=1}^T (\ell_{i^*}^t - \bar{\ell}_i^*)^2$, where $\bar{\ell} = \frac{1}{T}\sum_{t=1}^T \ell^t$.
- Path length of losses: $V_\infty = \sum_{t=1}^{T-1} \|\ell^t - \ell^{t+1}\|_\infty$, $V_1 = \sum_{t=1}^{T-1} \|\ell^t - \ell^{t+1}\|_1$

As shown in Table 1, there are existing algorithms with regret bounds that depend on a difficulty indicator rather than on T. Note that the above examples $L_{i^*}, Q_\infty, V_\infty$ are at most T, and

Table 1: Regret bounds for multi-armed bandit. Δ_i is the suboptimality gap defined in Section 2. $\Delta^* = \min_{i \in [K] \setminus \{i^*\}} \Delta_i$. Contributions of this work are highlighted with gray boxes.

Algorithm	Regime	Regret	Param. free?
BROAD, Option I (Wei and Luo, 2018)	Adv.	$O(\sqrt{KQ_{i^*} \log T})$	Require Q_{i^*}
BROAD, Option II (Wei and Luo, 2018)	Sto. Adv.	$O\left(\frac{K \log T}{\Delta_*}\right)$ $O(\sqrt{KL_{i^*} \log T})$	Yes
(Bubeck et al., 2018)	Adv.	$O\left(\sqrt{Q_2 \log K}\right)$	Require Q_2
(Bubeck et al., 2019)	Adv.	$O\left(\sqrt{KV_\infty \log T}\right)$	Require V_∞
Exp3++ (Seldin and Lugosi, 2017)	Sto. Adv.	$O\left(\sum_{i:\Delta_i>0} \frac{(\log T)^2}{\Delta_i}\right)$ $O(\sqrt{KT \log K})$	Yes
Tsallis-INF (Zimmert and Seldin, 2021)	Sto. Adv.	$O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right)$ $O\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right)$ $O(\sqrt{KT})$	Yes
Algorithm 1 (This work)	Sto. Adv.	$O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right)$ $O\left(\sqrt{KL_{i^*} \log T}\right)$ $O\left(\sqrt{KQ_\infty \log T}\right)$ $O\left(\sqrt{KV_1 \log T}\right)$	Yes

therefore, e.g., $O(\sqrt{KQ_{i^*} \log T})$-regret bounds imply an $O(\sqrt{KT \log T})$-bound. Some of the existing algorithms in Table 1 rely on prior knowledge of a difficulty indicator. For example, the $O(\sqrt{KV_\infty \log T})$-regret algorithm (Bubeck et al., 2019) requires a constant-factor approximation of V_∞ as an input parameter.

1.1. Contributions of this work

The contributions of this work are two-fold: First, we show that the Tsallis-INF algorithm has an $O\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right)$-regret bound in the stochastic setting, even when there are multiple optimal arms. This resolves an open question posed in the literature by Zimmert and Seldin (2021). Second, we propose an algorithm (Algorithm 1) that achieves best-of-both-worlds and has multiple data-dependent regret bounds *simultaneously*. The regret bounds achieved by Algorithm 1 are shown in Table 1.

One main contribution in this paper is proof of the fact that the Tsallis-INF algorithm has an $O(\log T)$-regret bound in the stochastic setting, even when there are multiple optimal arms. A more specific claim can be stated as follows:

Definition 1 (Adversarial regime with a self-bounding constraint (Zimmert and Seldin, 2021)) If the regret for any algorithm satisfies

$$R^T \geq \sum_{i \in [K]} \Delta_i \cdot \sum_{t=1}^{T} \mathrm{Prob}\left[i^t = i\right] - C \tag{2}$$

for some $C \geq 0$, T and $\Delta \in [0,1]^K$, we say that the environment is in an *adversarial regime with a* (Δ, C, T) *self-bounding constraint*.

Theorem 2 *Suppose that* (2) *holds with* Δ *satisfying* $\Delta_i > 0$ *for all* $i \in V$*, where* $V \subsetneq [K]$ *is an arbitrary nonempty set of actions. Suppose that there exists* $D \geq 0$ *such that*

$$\mathbf{E}\left[\sum_{t=1}^{T} \max_{i \in [K] \setminus V} \mathbf{E}^t\left[\ell_i^t - \ell_{i^*}^t\right]\right] \leq D, \tag{3}$$

where $\mathbf{E}^t[\cdot]$ *represents the conditional expectation given the history* $h^t = \{(\ell^j, i^j)\}_{j=1}^{t-1}$*. The regret for the Tsallis-INF with IW estimators (Zimmert and Seldin, 2021) will then be bounded as*

$$R_{i^*}^T \leq O\left(\sum_{i \in V} \frac{\log T}{\Delta_i} + \sqrt{C \sum_{i \in V} \frac{\log T}{\Delta_i}} + D + K\right). \tag{4}$$

This theorem captures stochastic settings with multiple optimal arms. In fact, in a stochastic setting with suboptimality gap $\Delta \in \mathbb{R}_{\geq 0}$, (2) and (3) hold with $C = D = 0$, for $V = \{i \in [K] \mid \Delta_i > 0\}$. Note that, as shown in (Zimmert and Seldin, 2021), the problem instances with (2) include stochastic settings, as well as the *stochastic bandits with adversarial corruptions* (Lykouris et al., 2018).

To prove Theorem 2, this paper introduces new techniques for analyzing FTRL with a time-varying regularizer. A core tool in the analysis is *skewed Bregman divergence*, which is defined with two different regularizers, in contrast to standard Bregman divergence, which is defined with a single regularizer. This tool helps with a refined insight into the FTRL with a time-varying regularizer and is especially powerful for analyzing a kind of *dynamic regret* (Zinkevich, 2003) that plays a central role in the proof of Theorem 2.

Another contribution of this work is to develop a new algorithm (Algorithm 1) that achieves best-of-both-worlds and has multiple data-dependent regret bounds. Our proposed algorithm enjoys the following regret bounds:

Theorem 3 *If* i^t *are chosen by Algorithm 1, for any* $i^* \in [K]$*, the regret is bounded as*

$$R^T = O\left(\sum_{i \in [K] \setminus i^*} \sqrt{\log T \cdot \sum_{t=1}^{T} \mathrm{Prob}\left[i^t = i\right]} + K \log T\right). \tag{5}$$

Simultaneously, for arbitrary sequence $\{u^t\}_{t=1}^T \subseteq [0,1]^K$*,*[2] *the regret is bounded as*

$$R^T = O\left(\sqrt{K \log T \cdot \mathbf{E}\left[\sum_{t=1}^{T} \left(\ell_{i^t}^t - u_{i^t}^t\right)^2 + \sum_{t=1}^{T-1} \|u^t - u^{t+1}\|_1\right]} + K \log T\right). \tag{6}$$

2. The algorithm does *not* require $\{u^t\}$ as an input. In an analysis, we can choose $\{u^t\}$ *arbitrarily*, which may depend on ℓ^t and i^t.

From the bound of (5), we obtain an $O(\log T)$ regret bound in the stochastic regime. Indeed, in the adversarial regime with a self-bounding constraint given in Definition 1, we have the following regret bound:

Corollary 4 (Regret bound in the adversarial regime with a self-bounding constraint) *Suppose that there exist $i^* \in [K]$ and $\Delta \in [0,1]^K$ satisfying $\Delta_i > 0$ for all $i \in [K] \setminus \{i^*\}$ for which (2) holds. Then, the regret for Algorithm 1 is bounded as*

$$R_{i^*}^T = O\left(\sum_{i \in [K] \setminus \{i^*\}} \frac{\log T}{\Delta_i} + \sqrt{C \sum_{i \in [K] \setminus \{i^*\}} \frac{\log T}{\Delta_i}} \right). \tag{7}$$

In contrast to Theorem 2, Corollary 4 requires the assumption that there exists a unique optimal arm i^*, i.e., $\Delta_i > 0$ for all $i \in [K] \setminus \{i^*\}$.

As (6) holds for *arbitrary* sequence $\{u^t\}$, from some specific choices of $\{u^t\}$, we have the following data-dependent regret bounds:

Corollary 5 (Data-dependent regret bound in the adversarial regime) *For any $\{\ell^t\} \subseteq [0,1]^K$, and for any $i^* \in [K]$ and $\bar{\ell} \in [0,1]$, the regret for Algorithm 1 is bounded as*

$$R^T = O\left(\sqrt{K \log T \cdot \min\left\{ \sum_{t=1}^T \ell_{i^*}^t,\ \sum_{t=1}^T \left(\ell_{i^t}^t - \bar{\ell}_{i^t}\right)^2,\ \sum_{t=1}^{T-1} \|\ell^t - \ell^{t+1}\|_1 \right\}} + K \log T \right). \tag{8}$$

Note that the algorithm is parameter-free, i.e., does not require any prior knowledge regarding $\{\ell^t\}$ except for the assumption of boundedness: $\ell^t \in [0,1]^K$.

Remark 6 Though Algorithm 1 requires the time horizon T as an input, we can avoid this requirement, i.e., we can obtain an *anytime* regret bound with just an additional constant factor, by modifying the algorithm. Details regarding this can be found in Section 5.3.

Remark 7 (On constant factors hidden within $O(\cdot)$ notation) The constant factors in the regret bounds for the Tsallis-INF shown in this paper (Thoerem 2) are not as small as those presented in (Zimmert and Seldin, 2021). For our proposed algorithm, constant factors hidden within $O(\cdot)$ notations in Thoerem 3 and Corollaries 4 and 5 are stated in the proofs of them given in Subsection 5.2. We believe that a more sophisticated analysis would improve the constant factors in regret bounds given in Theorems 2, 3 and Corollaries 4, 5.

The data-dependent regret bounds of the proposed algorithm are not necessarily better than existing ones. For example, the variation dependent bound of $O(\sqrt{KQ_\infty \log T})$ of Algorithm 1 is not superior to an $O(\sqrt{KQ_{i^*} \log T})$-bound in (Wei and Luo, 2018) or an $O(\sqrt{Q_2 \log T})$-bound by (Bubeck et al., 2018). Similarly, an $O(\sqrt{KV_1 \log T})$-bound of Algorithm 1 is not always better than the regret bounds of $O(\sqrt{KV_\infty \log T})$ and $O\left(K^{1/3}\sqrt{V_1^{2/3} T^{1/3} \log T}\right)$ shown by Bubeck et al. (2019). On the other hand, the proposed algorithm has the advantage of being parameter-free and enjoys several different data-dependent bounds simultaneously.

The proposed algorithm is based on the follow-the-regularized-leader (FTRL) framework with a time-varying regularizer, similarly to Tsallis-INF (Zimmert and Seldin, 2021). Unlike Tsallis-INF, which employs Tsallis entropy, the proposed algorithm uses a log-barrier regularizer. A time-varying log-barrier regularizer has been used in an MAB context in a study by Wei and Luo (2018). One difference between our work and theirs is that their algorithms are based on an online mirror descent (OMD) framework. Though OMD coincides with FTRL in a special case of a (time-invariant) fixed regularizer (McMahan, 2011, 2010), they offer different output if a time-varying regularizer is used. By combining an FTRL approach and a novel update rule for the regularizer, the proposed algorithm achieves an $O\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i}\right)$-regret bound in a stochastic setting, as shown in Table 1. This is an improvement over an $O\left(\frac{K\log T}{\Delta^*}\right)$ bound by (Wei and Luo, 2018), which depends only on the smallest suboptimality gap $\Delta^* = \min_{i\in[K]\setminus\{i^*\}}\Delta_i$.

One benefit of using a log-barrier rather than Tsallis entropy is that an *optimistic online learning* framework (Rakhlin and Sridharan, 2013a,b) can be used effectively,[3] as it was in (Wei and Luo, 2018). In this optimistic online learning framework, the algorithm is given (or maintains with a certain strategy) *hint vectors* m^t before choosing an action. We can see that the better that hints approximate true losses the more that regret improves. This paper shows that a surprisingly simple update rule for hints m^t leads to the data-dependent regret bounds in (6) and (8).

1.2. Related work

For MAB contexts, there can be found a variety of approaches to designing best-of-both-worlds algorithms. The pioneering work by Bubeck and Slivkins (2012); Auer and Chiang (2016) adopts the approach of selecting an appropriate mode by determining whether the environment is i.i.d. or not. Seldin and Slivkins (2014) and Seldin and Lugosi (2017) consider modifying the well-known adversarial MAB algorithm EXP3 (Auer et al., 2002b) so that it achieves improved regret bounds in stochastic settings. The approach adopted in Wei and Luo (2018); Zimmert and Seldin (2021); Pogodin and Lattimore (2020) is the most relevant to this paper. These studies lead to improved regret bounds for stochastic settings on the basis of a lower bound for the regret, such as in (2), referred to as self-bounding constraints.

Data-dependent regret bounds are defined with many different difficulty indicators. Auer et al. (2002b) have presented an $O(\sqrt{G_{\max}K\log K})$-regret bound for the adversarial MAB problem with rewords (maximization problem), where $G_{\max}$ represents the total return of the best action. Allenberg et al. (2006) have provided an MAB algorithm with a regret bound depending on cumulative loss L_{i^*} rather than on T. Hazan and Kale (2011) proposed an algorithm with a regret bound depending on the total variation Q_2 of the loss sequences, which can be applied to linear bandits, a general model including MAB. This regret bound was improved by Bubeck et al. (2018), in which $O(\sqrt{Q_2\log K})$-regret was achieved as shown in Table 1. For general linear bandits, an algorithm by Ito et al. (2020) achieves a cumulative-loss-dependent regret bound and a variation-dependent regret bound simultaneously. The sparsity of loss vectors (Kwon and Perchet, 2016; Bubeck et al., 2018) is an example of difficulty indicators that are not well addressed in this paper.

For the purpose of achieving hybrid regret bounds, one may consider combining multiple bandit algorithms. Such approaches have been studied for a wide range of bandit problems, e.g., by

3. Although we here consider algorithms using Tsallis entropy and an optimistic online learning framework, it is currently unclear to the authors if such an approach in MAB offers non-trivial results, e.g., the data-dependent regret bounds in Table 1.

Agarwal et al. (2017); Pacchiano et al. (2020); Arora et al. (2021). These studies, however, do not seem to directly offer hybrid regret bound as in Corollary 5, due to the overhead incurred when merging multiple bandit algorithms. Challenges in combining bandit algorithms are well discussed in (Agarwal et al., 2017).

2. Problem Setting

In each time step t, the player chooses $i^t \in [K]$ and then observes the losses ℓ^t_{ti}, where the loss vector $\ell^t = (\ell^t_1, \ell^t_2, \dots, \ell^t_K)^\top \in [0,1]^K$ is chosen by the environment and is assumed to be bounded in $[0,1]^K$. The goal of the player is to minimize the regret defined in (1).

In an *adversarial setting*, the environment chooses an ℓ^t depending on the history $h^t = \{(\ell^j, i^j)\}_{j=1}^{t-1}$ of losses and actions selected up to the $(t-1)$-th round. An *adversarial regime with a self-bounding constraint* (Zimmert and Seldin, 2021) is defined with parameters $\Delta \in \mathbb{R}^K_{\geq 0}$, $C \geq 0$ and T. In this regime, the losses may be chosen in an adversarial way, but the regret is required to satisfy the condition of (2). This regime includes a *stochastic setting*, in which ℓ^t are assumed to follow probability distributions with fixed means $\mu \in [0,1]^K$, i.e., $\mathbf{E}^t[\ell^t] = \mu$ holds for all $t \in [T]$. Indeed, if we set $\Delta_i = \mu_i - \mu_{i^*}$, where $i^* \in \arg\min_{i \in [K]} \mu_i$, (2) holds with $C = 0$. As shown in (Zimmert and Seldin, 2021), this regime includes a *stochastic setting with adversarial corruptions* (Lykouris et al., 2018), in which an adversary creates an amount of corruption of at most C in a stochastic environment with a suboptimality gap Δ. For this setting, we can easily confirm that (3) holds with $D = C$. See, e.g., Section 5.1 of (Zimmert and Seldin, 2021) for more details.

3. (Optimistic) Follow the Regularized Leader

In this section, we introduce an online linear optimization framework. For it, we formulate algorithms referred to as (optimistic) follow the regularized leader algorithms, and provide regret analyses for them. The regret bounds in this section are used in the analysis of Algorithm 1 in Section 5 and Tsallis-INF in section 4.

In the online linear optimization, the player is given a bounded convex action set $\Omega \subseteq \mathbb{R}^d$ before the game start. In each round t, the player is given a *hint vector* $m^t \in \mathbb{R}^d$, which is a prediction of a cost vector, and then chooses an action $x^t \in \Omega$. The player then observes the cost vector $c^t \in \mathbb{R}^d$ and suffers a loss of $\langle c^t, x^t \rangle$.

Follow the regularized leader (FTRL) and optimistic follow the regularized leader (OFTRL) methods are defined with *regularization functions* $\psi^t : \mathcal{D} \to \mathbb{R}^d$, convex functions of the Legendre type (Cesa-Bianchi and Lugosi, 2006; Rockafellar, 1970). We assume here that Ω is an affine subset of $\mathcal{D}$, i.e, Ω can be expressed as $\Omega = \{x \in \mathcal{D} \mid Ax = b\}$ for a matrix $A \in \mathbb{R}^{k \times d}$ and a vector $b \in \mathbb{R}^k$. Define $\Psi^t : \mathbb{R}^d \to \Omega$ by

$$\Psi^t(L) = \arg\min_{x \in \Omega} \left\{ \langle L, x \rangle + \psi^t(x) \right\}. \tag{9}$$

The player's actions given by FTRL and OFTRL, denoted by z^t and $\tilde{z}^t$, can be expressed as follows:

$$z^t = \Psi^t\left(\sum_{j=1}^{t-1} c^j\right), \quad \tilde{z}^t = \Psi^t\left(\sum_{j=1}^{t-1} c^j + m^t\right). \tag{10}$$

To analyze regret bounds for FTRL and OFTRL, we define *skewed Bregman divergence* $D^{s,t}(x, y)$ for $x \in \mathcal{D}$ and $y \in \mathrm{int}(\mathcal{D})$ as

$$D^{s,t}(x, y) = \psi^s(x) - \psi^t(y) - \left\langle \nabla\psi^t(y), x - y \right\rangle. \tag{11}$$

We denote $D^t(x, y) = D^{t,t}(x, y)$, which is the standard Bregman divergence associated with ψ^t. As can clearly be seen, the skewed Bregman divergence can be expressed as $D^{s,t}(x, y) = D^t(x, y) + \psi^s(x) - \psi^t(x)$. Note that the values of skewed Bregman divergence are not always nonnegative while those with the standard Bregman divergence are always nonnegative. The regret for FTRL and OFTRL can be bounded as follows:

Proposition 8 *Suppose that z^t and $\tilde{z}^t$ are given by (10). For any $u \in \Omega$, we have*

$$\sum_{t=1}^{T} \left\langle c^t, z^t - u \right\rangle \le D^1(u, z^1) + \sum_{t=1}^{T} D^{t,t+1}(z^t, z^{t+1}) + \psi^{T+1}(u) - \psi^1(u), \tag{12}$$

$$\sum_{t=1}^{T} \left\langle c^t, \tilde{z}^t - u \right\rangle \le D^1(u, z^1) + \sum_{t=1}^{T} D^{t,t+1}(\tilde{z}^t, z^{t+1}) + \psi^{T+1}(u) - \psi^1(u). \tag{13}$$

Let Ω^* be an affine subspace of Ω and define $u^t = \arg\min_{x \in \Omega^*} D^t(x, z^t)$ for all t. We then have

$$\sum_{t=1}^{T} \left\langle c^t, z^t - u^t \right\rangle \le D^1(u^1, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) - D^{t,t+1}(u^t, u^{t+1}) \right). \tag{14}$$

All proofs omitted for convenience here can be found in the appendix. The bounds in Proposition 8 can be derived via standard analysis techniques for FTRL. Note that (12) follows immediately from the inequality of Exercise 28.12 in the book by Lattimore and Szepesvári (2020) and the assumption that each ψ^t is a Legendre function. The difference between OMD and FTRL can be seen in Exercises 28.11 and 28.12 in this literature (Lattimore and Szepesvári, 2020). Further, Amir et al. (2020) have pointed out that FTRL is strictly superior to OMD in the problem of prediction with corrupted expert advice, a full-information online decision problem in the adversarial regime with a self-bounding constraint.

4. Refined Analysis of Tsallis-INF

The purpose of this section is to prove Theorem 2. The Tsallis-INF algorithm (with IW estimators) (Zimmert and Seldin, 2021) is given as FTRL with $\psi^t(p)$ and $\hat{\ell}^t$ defined as follows:

$$\psi^t(p) = -2\gamma^t \sum_{i=1}^{K} \sqrt{p_i}, \quad \hat{\ell}^t = \frac{\ell_{i^t}^t}{p_{i^t}^t} \chi_{i^t}, \quad p^t = \Psi^t\left(\sum_{j=1}^{t-1} \hat{\ell}^j \right) = \arg\min_{p \in \Delta^n} \left\{ \left\langle \sum_{j=1}^{t-1} \hat{\ell}^j, p \right\rangle + \psi^t(p) \right\}, \tag{15}$$

where $\gamma^t = \sqrt{t}$. Let $V \subsetneq [K]$ be an arbitrary nonempty subset of $[K]$. Denote $U = [K] \setminus V$. Our analysis uses $q^t \in \Delta^U = \{p \in \Delta^K \mid p_i = 0 \ (i \in V)\}$ defined as follows:

$$q^t = \arg\min_{p \in \Delta^U} \left\{ \left\langle \sum_{j=1}^{t-1} \hat{\ell}^j, p \right\rangle + \psi^t(p) \right\} = \arg\min_{p \in \Delta^U} D^t(p, p^t). \tag{16}$$

From the assumption of (3), the regret can be bounded with q^t as follows:

Lemma 9 *Suppose ℓ^t satisfies (3) and q^t is defined as (16). The regret is then bounded as $R_{i^*}^T \leq \mathbf{E}\left[\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - q^t \right\rangle\right] + D$.*

This lemma can be shown via $-\mathbf{E}[\sum_{t=1}^T \ell_{i^*}^t] \leq -\mathbf{E}[\sum_{t=1}^T \langle \ell^t, q_t \rangle] + D$, which follows from (3) and $q_t \in \Delta^U$, and that $\hat{\ell}^t$ is an unbiased estimator of ℓ^t. Further, from (14) in Proposition 8, $\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - q^t \right\rangle$ can be bounded as

$$\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - q^t \right\rangle \leq D^1(q^1, p^1) + \sum_{t=1}^T \left(D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1}) \right), \tag{17}$$

where $D_U^{t,t+1}$ denotes the skewed Bregman divergence associated with $\psi_U^t(p) = -2\gamma^t \sum_{i \in U} \sqrt{p_i}$.

From (17), we will show

$$\mathbf{E}\left[\left\langle \hat{\ell}^t, p^t - q^t \right\rangle\right] = O\left(\sum_{t=1}^T \frac{1}{\sqrt{t}} \sum_{i \in V} \sqrt{p_i^t} + K \right), \tag{18}$$

which leads to Theorem 2 via an argument similar to that by Zimmert and Seldin (2021). We can show (18) on the basis of the intuition that the terms regarding optimal arms $i \in U$ will be canceled out in (17). We can show that each term in (17) is bounded as $\mathbf{E}[D^{t,t+1}(p^t, p^{t+1})] = O(\frac{1}{\sqrt{t}} \sum_{i=1}^K \sqrt{p_i^t})$ and $\mathbf{E}[D_U^{t,t+1}(q^t, q^{t+1})] = O(\frac{1}{\sqrt{t}} \sum_{i \in U} \sqrt{q_i^t})$ for sufficiently large t, where U is the set of optimal arms (which follows from Lemma 19 in the appendix). As we have $q_i^t \geq p_i^t$ for $i \in U$ from their definitions, we may expect that the terms regarding optimal arms $i \in U$ are canceled out, which leads to (18). To prove this rigorously, we need more precise evaluations for $D^{t,t+1}(p^t, p^{t+1})$ and $D_U^{t,t+1}(q^t, q^{t+1})$, details of which are given in the appendix. Consequently, we can show that each term in (17) can be bounded as follows:

Lemma 10 *Suppose that $t \geq 4K$ holds. We then have $D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1}) \leq \frac{1}{\sqrt{t}} \left(\frac{\mathbf{1}[i^t \in V]}{\sqrt{p_{i^t}^t}} + \frac{2 \cdot \mathbf{1}[i^t \in U] \cdot p_{i^t}^t \sum_{i \in V} (p_i^t)^{3/2}}{\sum_{i \in U} (p_i^t)^{3/2} \sum_{i=1}^K (p_i^t)^{3/2}} + 2 \sum_{i \in V} \sqrt{p_i^{t+1}} \right) + \frac{\sqrt{|U|}}{t^{3/2}}$.*

Combining this lemma with $\mathbf{E}^t\left[\frac{\mathbf{1}[i^t \in V]}{\sqrt{p_{i^t}^t}}\right] = \sum_{i \in V} \sqrt{p_i^t}$ and $\mathbf{E}^t\left[\frac{\mathbf{1}[i^t \in U] \cdot p_{i^t}^t \sum_{i \in V} (p_i^t)^{\frac{3}{2}}}{\sum_{i \in U} (p_i^t)^{\frac{3}{2}} \sum_{i=1}^K (p_i^t)^{\frac{3}{2}}}\right] \leq \sqrt{\sum_{i \in V} p_i^t}$ (Lemma 23 in the appendix), we obtain $\mathbf{E}\left[D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1})\right] \leq \frac{5}{\sqrt{t}} \mathbf{E}\left[\sum_{i \in V} \sqrt{p_i^t}\right] + \frac{\sqrt{K}}{t^{3/2}}$ for $t \geq 4K$. For $t < 4K$, we use $\mathbf{E}\left[D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1})\right] \leq 2\frac{\sqrt{K}}{\sqrt{t}}$ (Lemma 24 in the appendix). Hence, by combining these inequalities, Lemma 9, 10 and (17), we have $R_{i^*}^T \leq \sum_{t=1}^T \frac{5}{\sqrt{t}} \mathbf{E}\left[\sum_{i \in V} \sqrt{p_i^t}\right] + 10K + D$. From this and (2), via an argument similar to that in the proof of Theorem 1 in (Zimmert and Seldin, 2021), we obtain the regret bound of Theorem 2. The complete proof of Theorem 2 can be found in the appendix.

5. MAB Algorithm based on OFTRL with Adaptive Log-Barrier Regularizer

In this section, we provide an MAB algorithm that achieves the regret bound of Theorem 3.

5.1. Algorithm description

In the proposed algorithm, we maintain distributions $p^t \in \Delta^K := \{p \in [0,1]^K \mid \|p\|_1 = 1\}$ with OFTRL and pick i^t from p^t in each round. Similarly to (Wei and Luo, 2018), from the bandit feedback $\ell^t_{i^t}$ for the chosen action i^t, we construct an unbiased estimator $\hat{\ell}^t$ of ℓ^t as follows:

$$\hat{\ell}^t = m^t + \frac{\ell^t_{i^t} - m^t_{i^t}}{p^t_{i^t}} \chi_{i^t}, \tag{19}$$

where $m^t \in [0,1]^K$ represents any hint vectors fixed before i^t is chosen, and $\chi_i \in \{0,1\}^K$ represents the indicator vector of $i \in [K]$.

The distribution p^t is computed with OFTRL with $c^t = \hat{\ell}^t$. Let $\mathcal{D} = \mathbb{R}^K_{>0}$ and define $\psi^t : \mathcal{D} \to \mathbb{R}$ by

$$\psi^t(p) = -\sum_{i=1}^K \gamma^t_i \log(p_i) \tag{20}$$

where $\gamma^t_i \geq 2$ is defined later so that $\gamma^1_i = 2$ and $\gamma^{t+1}_i \geq \gamma^t_i$ for all $i \in [K]$ and t. Define $\Omega = \Delta^K \cap \mathcal{D} = \{p \in \mathbb{R}^K_{>0} \mid \sum_{i=1}^K p_i = 1\}$. Using the unbiased estimators $\hat{\ell}^t$ defined as (19), we set p^t with OFTRL as follows:

$$p^t \in \operatorname*{arg\,min}_{p \in \Omega} \left\{ \left\langle \sum_{j=1}^{t-1} \hat{\ell}^j + m^t, p \right\rangle + \psi^t(p) \right\} = \Psi^t \left(\sum_{j=1}^{t-1} \hat{\ell}^j + m^t \right). \tag{21}$$

In the proposed algorithm, γ^t is updated as follows:

$$\gamma^1_i = 2, \quad \gamma^{t+1}_i = \gamma^t_i + \frac{B \cdot \nu^t_i}{2\gamma^t_i}, \quad \text{where} \quad \nu^t_i = (\ell^t_{i^t} - m^t_{i^t})^2 \cdot \begin{cases} \left(1 - p^t_i\right)^2 & \text{if } i = i^t \\ \left(p^t_i\right)^2 & \text{if } i \neq i^t \end{cases}, \tag{22}$$

where $B \in (0,1]$ is a parameter. In (22), the values of ν^t_i are defined so that a certain term of (skewed) Bregman divergences associated with ψ^t and ψ^{t+1} can be bounded by $\sum_{i=1}^K \frac{(p_i \ell_i)^2}{\gamma^t_i}$, as shown in Lemma 25 in the appendix. The values of γ^t_i are defined so that $\gamma^t_i \leq \sqrt{B \sum_{j=1}^{t-1} \nu^j_i} + 2$ holds as shown in Lemma 26. Note that γ^t_i defined by $\gamma^t_i = \sqrt{B \sum_{j=1}^{t-1} \nu^j_i} + 2$ works as well as those defined by (22), which can be seen from the proof of Lemma 26. We update m_t as follows:

$$m^1_i = \frac{1}{2}, \quad m^{t+1}_i = \begin{cases} (1-\eta) m^t_i + \eta \ell^t_i & \text{if } i = i^t \\ m^t_i & \text{if } i \neq i^t \end{cases}, \tag{23}$$

where $\eta \in [0,1]$ is a parameter. Note that setting $\eta = 0$, i.e., setting $m^t_i = 1/2$ for all $t \in [T]$ and $i \in [K]$, leads to an unbiased estimator $\hat{\ell}_t$ that is close to reduced-variance (RV) loss estimators in the Tsallis-INF algorithm (Zimmert and Seldin, 2021), though they are not exactly the same. As we will mention in Remark 15 (5) in Theorem 3 and Corollary 4 hold even if m^t are chosen in a different way than in (23), as long as $m^t \in [0,1]^K$. We show that the regret bounds in Theorem 3 are achieved for the parameter setting of $B = (\log T)^{-1}$ and $\eta = 1/4$. The proposed algorithm can be summarized in Algorithm 1.

Algorithm 1 OFTRL with adaptive log-barrier regularization

Require: The number K of actions, time horizon T.

1: Initialize $\eta = \frac{1}{4}$, $B = (\log T)^{-1}$, $\gamma^1 = 2 \cdot \mathbf{1} \in \mathbb{R}^K$ and $m^1 = \frac{1}{2} \cdot \mathbf{1} \in \mathbb{R}^K$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute p^t defined by (20) and (21).
4: Choose i^t according to p^t, i.e., so that $\mathrm{Prob}[i^t = i] = p_i^t$, and get feedback of $\ell_{i^t}^t$.
5: Compute $\hat{\ell}^t$ by (19) as an unbiased estimator of ℓ^t.
6: Update γ^t and m^t as in (22) and (23), respectively.
7: **end for**

5.2. Regret analysis

To demonstrate Theorem 3, we start with the following proposition:

Proposition 11 *Suppose that p^t and i^t are chosen by (20), (21) and (22) with arbitrary $m^t \in [0,1]^K$ and $B \in (0,1]$. We then have $R^T \leq \left(\frac{2}{\sqrt{B}} + \sqrt{B}\log\frac{K}{\varepsilon}\right)\mathbf{E}\left[\sum_{i=1}^K \sqrt{\sum_{t=1}^T \nu_i^t}\right] + 2K\log\frac{K}{\varepsilon} + \varepsilon T$ for any $\varepsilon \in (0,1]$ and T.*

Proof We here provide a proof sketch. A complete proof can be found in the appendix. Set p^* by $p^* = (1-\varepsilon)\cdot\chi_{i^*} + \frac{\varepsilon}{K}\cdot\mathbf{1}$. Since $\hat{\ell}^t$ is an unbiased estimator, as shown in (Wei and Luo, 2018), we have $R_{i^*}^T \leq \mathbf{E}\left[\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - p^*\right\rangle\right] + \varepsilon T$. From (13) in Proposition 8, for $r^t := \Psi^t\left(\sum_{j=1}^{t-1}\hat{\ell}^j\right)$, we can bound the term of $\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - p^*\right\rangle$ as follows:

$$\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - p^*\right\rangle \leq \sum_{t=1}^T D^{t,t+1}(p^t, r^{t+1}) + D^1(p^*, r^1) + \psi^{T+1}(p^*) - \psi^1(p^*). \tag{24}$$

As we have $D^{t,t+1}(p^t, r^{t+1}) \leq \sum_{i=1}^K \frac{\nu_i^t}{\gamma_i^t}$ (see Lemma 25 in the appendix), the first term of (24) can be bounded as $\sum_{t=1}^T D^{t,t+1}(p^t, r^{t+1}) \leq \sum_{i=1}^K \sum_{t=1}^T \frac{\nu_i^t}{\gamma_i^t} = \frac{2}{B}\sum_{i=1}^K \left(\gamma_i^{T+1} - \gamma_i^1\right)$, where the equality follows from the update rule of γ_i^t given in (22). The remaining part of (24) can be bounded as $D^1(p^*, p^1) + \psi^{T+1}(p^*) - \psi^1(p^*) = \psi^{T+1}(p^*) - \psi^1(r^1) - \left\langle \nabla\psi^1(r^1), p^* - r^1\right\rangle \leq \psi^{T+1}(p^*) \leq -\sum_{i=1}^K \gamma_i^{T+1}\log(p_i^*) \leq \log\frac{K}{\varepsilon}\sum_{i=1}^K \gamma_i^{T+1}$, where the first inequality holds since $\psi^1(r^1) \geq 0$ and $\nabla\psi^1(r^1) = -2K\cdot\mathbf{1}$ follow from the definition of ψ^t in (20) and $\gamma_i^1 = 2$. Combining the above inequalities with (24), we have $\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - p^*\right\rangle \leq \sum_{i=1}^K \left(\frac{2}{B}\left(\gamma_i^{T+1} - \gamma_i^1\right) + \gamma_i^{T+1}\log\frac{K}{\varepsilon}\right)$. From the definition (22) of γ_i^t, we can show $\gamma_i^t \leq \sqrt{B\sum_{j=1}^{t-1}\nu_i^j} + 2$ in induction in t. We hence have $\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - p^*\right\rangle \leq \left(\frac{2}{\sqrt{B}} + \sqrt{B}\log\frac{K}{\varepsilon}\right)\sum_{i=1}^K \sqrt{\sum_{t=1}^T \nu_i^t} + 2K\log\frac{K}{\varepsilon}$. By combining this with $R^T \leq \mathbf{E}\left[\sum_{t=1}^T \left\langle \hat{\ell}^t, p^t - p^*\right\rangle\right] + \varepsilon T$, we obtain the regret bound in Proposition 11. ■

The term of $\sum_{i=1}^K \sqrt{\sum_{t=1}^T \nu_i^t}$ in Proposition 11 can be bounded as follows.

Lemma 12 *If ν_i^t is defined as in (22), we have $\mathbf{E}\left[\sum_{i=1}^K \sqrt{\sum_{t=1}^T \nu_i^t}\right] \leq 2\sum_{i\neq i^*}\sqrt{\mathbf{E}\left[\sum_{t=1}^T p_i^t\right]}$ and $\sum_{i=1}^K \sqrt{\sum_{t=1}^T \nu_i^t} \leq \sqrt{2K\sum_{t=1}^T (\ell_{i^t}^t - m_{i^t}^t)^2}$.*

These can be shown via simple calculation. By combining the first part of this lemma and Proposition 11, we can obtain (5) in Theorem 3. The other part, (6), comes from the second part of Lemma 12, Proposition 11, and the following:

Proposition 13 *Suppose m^t is updated by (23) with $\eta \in (0, \frac{1}{2})$. It will then hold for any sequence $\{u^t\}_{t=1}^T \subseteq [0,1]^K$ that $\sum_{t=1}^T (\ell_{i^t}^t - m_{i^t}^t)^2 \leq \frac{1}{1-2\eta} \sum_{t=1}^T (\ell_{i^t}^t - u_{i^t}^t)^2 + \frac{2}{\eta(1-2\eta)} \left(\frac{K}{8} + \sum_{t=1}^{T-1} \|u^t - u^{t+1}\|_1 \right)$.*

Remark 14 The update rule of m^t in (23) can be seen as a gradient descent method for the convex objective $f^t(m) = (\ell_{i^t}^t - m_{i^t})^2$. Hence, we may apply tracking-regret analysis, e.g., that by Herbster and Warmuth (2001); Cesa-Bianchi and Lugosi (2006). For example, Theorem 11.4 in (Cesa-Bianchi and Lugosi, 2006) with $p = q = 2$ implies that $\sum_{t=1}^T (\ell_{i^t}^t - m_{i^t}^t)^2 \leq \frac{1}{1-2\eta} \sum_{t=1}^T (\ell_{i^t}^t - u_{i^t}^t)^2 + \frac{2}{\eta(1-2\eta)} \left(\frac{K}{2} + \sqrt{K} \sum_{t=1}^{T-1} \|u^t - u^{t+1}\|_2 \right)$. Proposition 13 can be regarded as a variant of this.

Proof of Theorem 3 We set $B = (\log T)^{-1}$, $\eta = 1/4$ as in Algorithm 1, and set $\varepsilon = K/T$. By combining Proposition 11 and the first part of Lemma 12 we obtain $R^T \leq 6\sqrt{\log T} \sum_{i \neq i^*} \sqrt{\mathbf{E}\left[\sum_{t=1}^T p_i^t\right]} + 3K \log T$, which means that (5) holds. Similarly, from Proposition 11, the second part of Lemma 12, and Proposition 13, we have $R^T \leq 6\sqrt{K \log T \cdot \mathbf{E}\left[\sum_{t=1}^T (\ell_{i^t}^t - u_{i^t}^t)^2 + K + 8 \sum_{t=1}^{T-1} \|u^t - u^{t+1}\|_1\right]} + 3K \log T$ for arbitrary $\{u^t\}_{t=1}^T \subseteq [0,1]^K$, which means that (6) holds. ■

Using Theorem 3, we can demonstrate Corollaries 4 and 5 as follows:

Proof of Corollary 4 From (5) and (2), for any $\lambda > 0$, we have $R_{i^*}^T = (1+\lambda) R_{i^*}^T - \lambda R_{i^*}^T \leq (1+\lambda) \cdot \left(6 \sum_{i \neq i^*} \sqrt{\log T \cdot \mathbf{E}\left[\sum_{t=1}^T p_i^t\right]} + 3K \log T \right) - \lambda \cdot \left(\sum_{i \neq i^*} \Delta_i \, \mathbf{E}\left[\sum_{t=1}^T p_i^t\right] - C \right) \leq \sum_{i \neq i^*} \left(6(1+\lambda) \sqrt{\log T \cdot \mathbf{E}\left[\sum_{t=1}^T p_i^t\right]} - \lambda \cdot \Delta_i \, \mathbf{E}\left[\sum_{t=1}^T p_i^t\right] \right) + 3(1+\lambda) \cdot K \log T + \lambda C \leq \frac{9(1+\lambda)^2}{\lambda} \sum_{i \neq i^*} \frac{\log T}{\Delta_i} + 3(1+\lambda) K \log T + \lambda C$, where the last inequality follows from the fact that $2bx - ax^2 \leq \frac{b^2}{a}$ holds for any $b, x \in \mathbb{R}$ and $a > 0$. Similarly to the proof of Theorem 1 in (Zimmert and Seldin, 2021), by choosing $\lambda = \left(1 + \frac{C + 3K \log T}{9 \sum_{i \neq i^*} \frac{\log T}{\Delta_i}} \right)^{-\frac{1}{2}}$, we obtain

$$
\begin{aligned}
R^T &\leq 36 \sum_{i \in [K] \setminus \{i^*\}} \frac{\log T}{\Delta_i} + 6 \sqrt{\sum_{i \in [K] \setminus \{i^*\}} \frac{(C + 3K \log T) \log T}{\Delta_i}} + 3K \log T \\
&= O\left(\sum_{i \in [K] \setminus \{i^*\}} \frac{\log T}{\Delta_i} + \sqrt{C \sum_{i \in [K] \setminus \{i^*\}} \frac{\log T}{\Delta_i}} \right),
\end{aligned}
$$

where the equality follows from $\frac{1}{\Delta_i} \geq 1$ for $i \in [K] \setminus i^*$. ■

Remark 15 As can be seen in the proofs, Corollary 4 and (5) in Theorem 3 are still valid even when m^t is chosen in a different way than in (23). More precisely, we can obtain the gap-dependent regret bound in Corollary 4 as long as $m^t \in [0,1]^K$ and m^t are independent of i^t given p^t. In

addition, if we choose $m_i^t = \frac{1}{2}$ for all $t \in [T]$ and $i \in [N]$, the constant factor in the gap-dependent regret bound will be improved. Indeed, as we have $\mathbf{E}[\nu_i^t] \leq \frac{1}{4}\mathbf{E}[p_i^t(1-p_i^t)]$ for $m^t = \frac{1}{2}\cdot\mathbf{1}$, we obtain $R^T \leq 9\sum_{i\in[K]\setminus\{i^*\}}\frac{\log T}{\Delta_i} + 3\sqrt{\sum_{i\in[K]\setminus\{i^*\}}\frac{(C+3K\log T)\log T}{\Delta_i}} + 3K\log T$ via an argument similar to that in the proof of Corollary 4. In this case, however, we do not have the bounds in Corollary 5.

Proof of Corollary 5 From (6) with $u^t = 0$ for all t, we have $R_{i^*}^T \leq 6\sqrt{K\log T\cdot\mathbf{E}\left[\sum_{t=1}^T(\ell_{i^t}^t)^2\right]} + 9K\log T \leq 6\sqrt{K\log T\cdot\mathbf{E}\left[\sum_{t=1}^T\ell_{i^t}^t\right]} + 9K\log T \leq 6\sqrt{K\log T\cdot\left(R_{i^*}^T + \mathbf{E}\left[\sum_{t=1}^T\ell_{i^*}^t\right]\right)} + 9K\log T$, where the second inequality follows from $\ell_i^t \in [0,1]$, and the last inequality follows from the definition (1) of $R_{i^*}^T$. As $R \leq a\sqrt{R+L}+b$ implies $R \leq a\sqrt{L}+a^2+2b$ for $a,b,L \geq 0$ (see Lemma 27 in the appendix), we have $R_{i^*}^T \leq 6\sqrt{K\log T\cdot\mathbf{E}\left[\sum_{t=1}^T\ell_{i^*}^t\right]} + 54K\log T$. Similarly, $R^T \leq 6\sqrt{K\log T\cdot\mathbf{E}\left[\sum_{t=1}^T(\ell_{i^t}^t-\bar{\ell}_{i^t})^2\right]} + 9K\log T$ follows from (6) with $u^t = \bar{\ell}$ (fixed for all t) and $R^T \leq 6\sqrt{K\log T\cdot\mathbf{E}\left[8\sum_{t=1}^{T-1}\|\ell^t-\ell^{t+1}\|_1\right]} + 9K\log T$ follows from (6) with $u^t = \ell^t$. ■

5.3. Extension to problems with unknown time horizons

Though the time horizon T is assumed to be given in Algorithm 1, we can eliminate this assumption, i.e., we can obtain an *anytime regret bound* as in Theorem 3, by modifying the algorithm. One approach is to apply the doubling trick w.r.t. $\log T$ (not w.r.t. T). For example, consider dividing all the rounds into segments $\{\mathcal{T}_k\}_{k\in\mathbb{N}}$ so that $\mathbb{N} = \cup_{k\in\mathbb{N}}\mathcal{T}_k$, where $\mathcal{T}_k = \{S_k+1, S_k+2, \ldots, S_k+T_k\}$ with $T_k = 2^{2^k}$ and $S_k = \sum_{j=1}^{k-1}T_j$. The meta-algorithm for unknown horizons starts by applying Algorithm 1 with $T = T_1$. Whenever the round reaches $t = S_k+1$ for some k, the meta-algorithm restarts Algorithm 1 with $T = T_k$. When restarting, we refresh all parameters except for m^t. For any T' (unknown time horizon), we denote $\mathcal{T}' = [T'] \cap \mathcal{T}$ and define $k(T')$ to be the integer k such that $T' \in \mathcal{T}_k$. We then have, $R^{T'} \leq \sum_{k=1}^{k(T')}\left(3\sum_{i=1}^K\sqrt{\log T_k\sum_{t\in\mathcal{T}_k'}\nu_i^t} + 6K\log T_k\right) \leq 3\sum_{i=1}^K\sqrt{\sum_{k=1}^{k(T')}\log T_k}\sqrt{\sum_{k=1}^{k(T')}\sum_{t\in\mathcal{T}_k'}\nu_i^t} + 6K\sum_{k=1}^{k(T')}\log T_k$, where the second inequality follows from the Cauchy-Schwarz inequality. As $\log T_k$ can be expressed as $\log T_k = 2^k\log 2$ from the definition of T_k, we have $\sum_{k=1}^{k(T')}\log T_k = \log 2\sum_{k=1}^{k(T')}2^k \leq \log 2\cdot 2^{k(T')+1} = 4\log T_{k(T')-1} \leq 4\log T'$, where the last inequality follows from $T' \geq S_{k(T')} \geq T_{k(T')-1}$. Hence, the meta-algorithm can achieve $R^{T'} \leq 6\sum_{i=1}^K\sqrt{\log T'\sum_{t=1}^T\nu_i^t} + 24K\log T'$ for any T'. This, combined with Lemmas 12 and 13, implies that the regret bound stated in (3) can be achieved anytime, without prior knowledge regarding the time horizon.

6. Conclusion

In this work, we have presented a newly developed MAB algorithm (Algorithm 1) that achieves best-of-both-worlds and multiple data-dependent regret bounds. Further, we have proved that the Tsallis-INF algorithm has a logarithmic regret bound for stochastic environments, even when there

are multiple optimal arms. One area for future work is to clarify if Algorithm 1 has a logarithmic regret for a stochastic regime with multiple optimal arms, similarly to that found in the Tsallis-INF algorithm. Investigating constant factors in the regret bounds would be an important area for future work as well.

Acknowledgments

The author thanks anonymous reviewers for their many comments and suggestions that improved the presentation of the manuscript. The author was supported by JST, ACT-I, Grant Number JPMJPR18U5, Japan.

References

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

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 229–243. Springer, 2006.

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

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, pages 217–226, 2009.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.

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

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

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127, 2018.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference on Learning Theory*, pages 508–528, 2019.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge university press, 2006.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Mark Herbster and Manfred K Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1(281-309):10–1162, 2001.

Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *The Journal of Machine Learning Research*, 17(1):8106–8137, 2016.

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

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

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, 2020.

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.

Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 525–533. JMLR Workshop and Conference Proceedings, 2011.

H Brendan McMahan. A unified view of regularized dual averaging and mirror descent with implicit updates. *arXiv preprint arXiv:1009.3240*, 2010.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20:1–28, 2019.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Roman Pogodin and Tor Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Uncertainty in Artificial Intelligence*, pages 894–904, 2020.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013b.

R Tyrrell Rockafellar. *Convex Analysis*, volume 36. Princeton university press, 1970.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295, 2014.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291, 2018.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pages 3583–3634. PMLR, 2020.

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.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A. Proof of Proposition 8

To prove Proposition 8, we introduce a lemma for single-round regret:

Lemma 16 *Suppose that z^t and $\tilde{z}^t$ are given by (10). For any $u \in \Omega$, we have*

$$\left\langle c^t, z^t - u \right\rangle = D^{t,t+1}(z^t, z^{t+1}) + D^t(u, z^t) - D^{t,t+1}(u, z^{t+1}), \tag{25}$$

$$\left\langle c^t, \tilde{z}^t - u \right\rangle = D^{t,t+1}(\tilde{z}^t, z^{t+1}) - D^t(\tilde{z}^t, z^t) + D^t(u, z^t) - D^{t,t+1}(u, z^{t+1}). \tag{26}$$

Proof Remember that $\Omega \subseteq \mathcal{D}$ can be expressed as $\Omega = \{x \in \mathcal{D} \mid Ax = b\}$ for a matrix A and vector b. From the assumptions on ψ^t and the first-order necessary conditions for the optimization problem of $\arg\min_{z \in \Omega}(\langle L, z \rangle + \psi^t(z))$, for $z^* = \Psi^t(L)$, there exists $v \in \mathbb{R}^k$ such that

$$L + \nabla \psi^t(z^*) = A^\top v. \tag{27}$$

Hence, from (10), there exists $v \in \mathbb{R}^k$ such that

$$\sum_{j=1}^{t-1} c^j + \nabla \psi^t(z^t) - \sum_{j=1}^{t} c^j - \nabla \psi^{t+1}(z^{t+1}) = \nabla \psi^t(z^t) - \nabla \psi^{t+1}(z^{t+1}) - c^t = A^\top v. \tag{28}$$

Hence, we have

$$\begin{aligned}
\left\langle c^t, z^t - u \right\rangle &= \left\langle \nabla \psi^t(z^t) - \nabla \psi^{t+1}(z^{t+1}) - A^\top v, z^t - u \right\rangle \\
&= \left\langle \nabla \psi^t(z^t) - \nabla \psi^{t+1}(z^{t+1}), z^t - u \right\rangle - \left\langle v, A(z^t - u) \right\rangle \\
&= \left\langle \nabla \psi^t(z^t) - \nabla \psi^{t+1}(z^{t+1}), z^t - u \right\rangle,
\end{aligned} \tag{29}$$

where the last equality follows from $z^t, u \in \Omega$, which implies $Az^t = Au = b$. On the other hand, from the definition (20) of $D^{s,t}$, the right-hand side of (25) can be expressed as

$$\begin{aligned}
&D^{t,t+1}(z^t, z^{t+1}) + D^t(u, z^t) - D^{t,t+1}(u, z^{t+1}) \\
&= -\left\langle \nabla\psi^{t+1}(z^{t+1}), z^t - z^{t+1} \right\rangle - \left\langle \nabla\psi^t(z^t), u - z^t \right\rangle + \left\langle \nabla\psi^{t+1}(z^{t+1}), u - z^{t+1} \right\rangle \\
&= \left\langle \nabla\psi^t(z^t) - \nabla\psi^{t+1}(z^{t+1}), z^t - u \right\rangle,
\end{aligned}$$

which, combined with (29), implies that (25) holds. Similarly, we have

$$\left\langle c^t, \tilde{z}^t - u \right\rangle = \left\langle \nabla\psi^t(z^t) - \nabla\psi^{t+1}(z^{t+1}), \tilde{z}^t - u \right\rangle$$

as well as

$$\begin{aligned}
&D^{t,t+1}(\tilde{z}^t, z^{t+1}) - D^t(\tilde{z}^t, z^t) + D^t(u, z^t) - D^{t,t+1}(u, z^{t+1}) \\
&= -\left\langle \nabla\psi^{t+1}(z^{t+1}), \tilde{z}^t - z^{t+1} \right\rangle + \left\langle \nabla\psi^t(z^t), \tilde{z}^t - z^t \right\rangle - \left\langle \nabla\psi^t(z^t), u - z^t \right\rangle + \left\langle \nabla\psi^{t+1}(z^{t+1}), u - z^{t+1} \right\rangle \\
&= \left\langle \nabla\psi^t(z^t) - \nabla\psi^{t+1}(z^{t+1}), \tilde{z}^t - u \right\rangle,
\end{aligned}$$

which lead to (26). ■

Lemma 17 *For any $t, t' \in \mathbb{N}$ and $L, L' \in \mathbb{R}^d$, denote $z = \Psi^t(L), z' = \Psi^{t'}(L')$. We then have*

$$-\left\langle L' - L, z' - z \right\rangle = D^{t,t'}(z, z') + D^{t',t}(z', z). \tag{30}$$

Proof In a similar way to (29), we can see that

$$-\left\langle L' - L, z' - z \right\rangle = \left\langle \nabla\psi^{t'}(z') - \nabla\psi^t(z), z' - z \right\rangle.$$

On the other hand, from the definition (20) of $D^{s,t}$, the right-hand side of (30) is expressed as

$$\begin{aligned}
&D^{t,t'}(z, z') + D^{t',t}(z', z) \\
&= \phi^t(z) - \phi^{t'}(z') - \left\langle \nabla\psi^{t'}(z'), z - z' \right\rangle + \phi^{t'}(z') - \phi^t(z) - \left\langle \nabla\psi^t(z), z' - z \right\rangle \\
&= \left\langle \nabla\psi^{t'}(z') - \nabla\psi^t(z), z' - z \right\rangle.
\end{aligned}$$

The above two equalities lead to (30). ■

Lemma 18 *Let t, t' be any natural numbers and let $\Omega' \subseteq \Omega$ be an affine subset of Ω such that $\Omega' \cap \mathrm{int}(\mathcal{D}) \neq \emptyset$. We then have*

$$D^{t,t'}(x, y) = D^{t,t'}(x, \pi^{t'}_{\Omega'}(y)) + D^{t'}(\pi^{t'}_{\Omega'}(y), y), \quad \textit{where} \quad \pi^{t'}_{\Omega'}(y) = \operatorname*{arg\,min}_{\omega \in \Omega'} D^{t'}(\omega, y). \tag{31}$$

Proof This lemma immediately follows from $D^{t,t'}(x, y) = \psi^t(x) - \psi^{t'}(x) + D^{t'}(x, y)$ and from the fact that the generalized Pythagorean inequality holds with equality for the standard Bregman

projection on a hyperplane (see, e.g., Lemma 11.3 and Exercise 11.2 in the book by (Cesa-Bianchi and Lugosi, 2006)). ■

Proof of Proposition 8 By taking the summation of (25) for $t \in [T]$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \langle c^t, z^t - u \rangle \\
&= \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) + D^t(u, z^t) - D^{t,t+1}(u, z^{t+1}) \right) \\
&= D^1(u, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) + D^{t+1}(u, z^{t+1}) - D^{t,t+1}(u, z^{t+1}) \right) - D^{T+1}(u, z^{T+1}) \\
&\leq D^1(u, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) + D^{t+1}(u, z^{t+1}) - D^{t,t+1}(u, z^{t+1}) \right) \\
&= D^1(u, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) + \psi^{t+1}(u) - \psi^t(u) \right) \\
&= D^1(u, z^1) + \sum_{t=1}^{T} D^{t,t+1}(z^t, z^{t+1}) + \psi^{T+1}(u) - \psi^1(u),
\end{aligned}$$

where the inequality follows from $D^{T+1}(u, z^{T+1}) \geq 0$. Hence, (12) holds. Similarly, by taking the summation of (26) for $t \in [T]$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \langle c^t, \tilde{z}^t - u \rangle &= \sum_{t=1}^{T} \left(D^{t,t+1}(\tilde{z}^t, z^{t+1}) - D^t(\tilde{z}^t, z^t) + D^t(u, z^t) - D^{t,t+1}(u, z^{t+1}) \right) \\
&\leq D^1(u, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(\tilde{z}^t, z^{t+1}) - D^t(\tilde{z}^t, z^t) + \psi^{t+1}(u) - \psi^t(u) \right) \\
&\leq D^1(u, z^1) + \sum_{t=1}^{T} D^{t,t+1}(\tilde{z}^t, z^{t+1}) + \psi^{T+1}(u) - \psi^1(u),
\end{aligned}$$

where the first inequality follows from $D^{T+1}(u, z^{T+1}) \geq 0$ and the second inequality follows from $D^t(\tilde{z}^t, z^t) \geq 0$. Hence, (13) holds. The inequality of (14) can be shown as follows:

$$\begin{aligned}
\sum_{t=1}^{T} \langle c^t, z^t - u^t \rangle &= \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) + D^t(u^t, z^t) - D^{t,t+1}(u^t, z^{t+1}) \right) \\
&\leq D^1(u^1, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) + D^{t+1}(u^{t+1}, z^{t+1}) - D^{t,t+1}(u^t, z^{t+1}) \right) \\
&= D^1(u^1, z^1) + \sum_{t=1}^{T} \left(D^{t,t+1}(z^t, z^{t+1}) - D^{t,t+1}(u^t, u^{t+1}) \right),
\end{aligned}$$

where the first equality follows from (25) and the second equality follows from Lemma 18. ■

Appendix B. Proof of Lemma 9

From the assumption of (3), and since $q^t \in \Delta^U$, we have

$$\mathbf{E}\left[\sum_{t=1}^{T}\left\langle \ell^t, q^t\right\rangle - \sum_{t=1}^{T}\ell^t_{i^*}\right] \le \mathbf{E}\left[\sum_{t=1}^{T}\max_{i\in U}\mathbf{E}^t\left[\ell^t_i - \ell^t_{i^*}\right]\right] \le D. \tag{32}$$

Hence, we have

$$\begin{aligned} R^T_{i^*} &= \mathbf{E}\left[\sum_{t=1}^{T}\ell^t_{i^t} - \sum_{t=1}^{T}\ell^t_{i^*}\right] = \mathbf{E}\left[\sum_{t=1}^{T}\ell^t_{i^t} - \sum_{t=1}^{T}\left\langle \ell^t, q^t\right\rangle + \sum_{t=1}^{T}\left\langle \ell^t, q^t\right\rangle - \sum_{t=1}^{T}\ell^t_{i^*}\right] \\ &\le \mathbf{E}\left[\sum_{t=1}^{T}\ell^t_{i^t} - \sum_{t=1}^{T}\left\langle \ell^t, q^t\right\rangle\right] + D = \mathbf{E}\left[\sum_{t=1}^{T}\left\langle \hat{\ell}^t, p^t - q^t\right\rangle\right] + D. \end{aligned}$$

where the inequality follows from (32) and the last equality follows from the definition of $\hat{\ell}^t$ in (15) and that $\mathbf{E}^t\left[\hat{\ell}^t\right] = \ell^t$.

Appendix C. Proof of Lemma 10

In this section, for any vector $x = (x_1, x_2, \ldots, x_K) \in \mathbb{R}^K$ and for any subset $U \subseteq [K]$ we denote $x_U = (x_i)_{i\in U} \in \mathbb{R}^U$. We start with the following lemma for evaluating the skewed Bregman divergence given by (15):

Lemma 19 *Suppose that ψ^t is defined by (15). For $p \in \mathbb{R}^K_{>0}$ and $\ell \in \mathbb{R}^K$ such that $\frac{\sqrt{p_i}\ell_i}{\gamma^t} \ge -\frac{1}{2}$ holds for all $i \in [K]$, let $q \in \mathbb{R}^K_{>0}$ be a vector such that*

$$\nabla\psi^{t+1}(q) = \nabla\psi^t(p) - \ell. \tag{33}$$

We then have

$$D^{t,t+1}(p,q) = \gamma^t\sum_{i=1}^{K}\sqrt{p_i}\cdot g\left(\frac{\sqrt{p_i}\ell_i}{\gamma^t}\right) + \frac{(\gamma^{t+1})^2 - (\gamma^t)^2}{\gamma^{t+1}}\sum_{i=1}^{K}\sqrt{q_i}, \tag{34}$$

where $g(x)$ is defined as $g(x) = \frac{x^2}{1+x}$. Consequently, we have

$$D^{t,t+1}(p,q) \le \frac{1}{\gamma^t}\sum_{i=1}^{K}(1 + \mathbf{1}[\ell_i < 0])\cdot(p_i)^{3/2}\cdot(\ell_i)^2 + \frac{(\gamma^{t+1})^2 - (\gamma^t)^2}{\gamma^{t+1}}\sum_{i=1}^{K}\sqrt{q_i}. \tag{35}$$

Proof As $\nabla\psi^t(p) = -\gamma^t\left(\frac{1}{\sqrt{p_i}}\right)_{i=1}^K$ from the definition of ψ^t in (15), (33) implies that

$$-\gamma^{t+1}\frac{1}{\sqrt{q_i}} = -\gamma^t\frac{1}{\sqrt{p_i}} - \ell_i \tag{36}$$

for all $i \in [K]$. From this, we have

$$g\left(\frac{\sqrt{p_i}\ell_i}{\gamma^t}\right) = \frac{(\sqrt{p_i}\ell_i)^2}{(\gamma^t)^2}\cdot\frac{1}{1+\frac{\sqrt{p_i}\ell_i}{\gamma^t}} = \frac{(\sqrt{p_i}\ell_i)^2}{(\gamma^t)^2}\frac{\gamma^t\sqrt{q_i}}{\gamma^{t+1}\sqrt{p_i}} = \frac{\sqrt{p_iq_i}}{\gamma^t\gamma^{t+1}}(\ell_i)^2. \tag{37}$$

From the definition (11) of $D^{t,t+1}$, we have

$$\begin{aligned}
D^{t,t+1}(p,q) &= \sum_{i=1}^{K}\left(-2\gamma^t\sqrt{p_i}+2\gamma^{t+1}\sqrt{q_i}+\gamma^{t+1}\frac{1}{\sqrt{q_i}}(p_i-q_i)\right)\\
&= \sum_{i=1}^{K}\left(-2\gamma^t\sqrt{p_i}+\gamma^{t+1}\sqrt{q_i}+\gamma^{t+1}\frac{p_i}{\sqrt{q_i}}\right)\\
&= \frac{1}{\gamma^{t+1}}\sum_{i=1}^{K}p_i\sqrt{q_i}\left(-2\gamma^t\gamma^{t+1}\frac{1}{\sqrt{p_iq_i}}+(\gamma^{t+1})^2\frac{1}{p_i}+(\gamma^{t+1})^2\frac{1}{q_i}\right)\\
&= \frac{1}{\gamma^{t+1}}\sum_{i=1}^{K}p_i\sqrt{q_i}\left(\left(\frac{\gamma^{t+1}}{\sqrt{q_i}}-\frac{\gamma^t}{\sqrt{p_i}}\right)^2+\frac{(\gamma^{t+1})^2-(\gamma^t)^2}{p_i}\right)\\
&= \frac{1}{\gamma^{t+1}}\sum_{i=1}^{K}p_i\sqrt{q_i}(\ell_i)^2+\frac{(\gamma^{t+1})^2-(\gamma^t)^2}{\gamma^{t+1}}\sum_{i=1}^{K}\sqrt{q^i}\\
&= \gamma^t\sum_{i=1}^{K}\sqrt{p_i}g\left(\frac{\sqrt{p_i}\ell_i}{\gamma^t}\right)+\frac{(\gamma^{t+1})^2-(\gamma^t)^2}{\gamma^{t+1}}\sum_{i=1}^{K}\sqrt{q^i}.
\end{aligned}$$

■

Note that we can use this lemma to evaluate $D^t(p,q)$ as well, by substituting $\gamma^{t+1}=\gamma^t$. In addition, the value of $D_U^{t,t+1}(p,q)$ can be expressed in a similar form, i.e., if $\nabla\psi_U^{t+1}(q)=\nabla\psi_U^t(p)-\ell_U$ then

$$D_U^{t,t+1}(p,q)=\gamma^t\sum_{i\in U}\sqrt{p_i}\cdot g\left(\frac{\sqrt{p_i}\ell_i}{\gamma^t}\right)+\frac{(\gamma^{t+1})^2-(\gamma^t)^2}{\gamma^{t+1}}\sum_{i\in U}\sqrt{q_i} \tag{38}$$

for any $U\in[K]$.

We define $\bar{p}^{t+1}\in\mathbb{R}_{>0}^U$ as

$$\bar{p}^{t+1}=\mathop{\arg\min}_{p\in\mathbb{R}_{>0}^U,\sum_{i\in U}p_i=\|p^t\|_U} D_U^{t+1}(p,p_U^{t+1}), \tag{39}$$

where we denote $\|p^t\|_U=\sum_{i\in U}p_i^t$. From the definition of $\bar{p}^{t+1}$, there exists $\alpha\in\mathbb{R}$ such that

$$\nabla\psi_U^{t+1}(\bar{p}^{t+1})=\nabla\psi_U^t(p^t)-\hat{\ell}_U^t+\alpha\cdot\mathbf{1}. \tag{40}$$

Further, for $\beta<\min_{i\in[K]}\left\{\frac{\gamma^t}{\sqrt{p_i^t}}\right\}$, we define $p^{t+1}(\beta)\in\mathbb{R}_{>0}^K$ to be the vector satisfying

$$\nabla\psi^{t+1}(p^{t+1}(\beta))=\nabla\psi^t(p^t)-\hat{\ell}^t+\beta\cdot\mathbf{1}. \tag{41}$$

From the definition of p^t and p^{t+1}, there exists β^* such that $p^{t+1}(\beta^*)=p^{t+1}$. In addition, for any β, we have

$$p^{t+1}=\mathop{\arg\min}_{p\in\Delta^K} D^{t+1}(p,p^{t+1}(\beta)), \tag{42}$$

$$\bar{p}^{t+1}=\mathop{\arg\min}_{p\in\mathbb{R}_{>0}^U,\sum_{i\in U}p_i=\|p^t\|_U} D_U^{t+1}(p,p^{t+1}(\beta)). \tag{43}$$

We then have

Lemma 20 *For any β, we have*

$$D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1}) \leq D_V^{t,t+1}(p^t, p^{t+1}(\beta)) + D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\beta)). \tag{44}$$

Proof From Lemma 18, (42), and the fact that $p^t \in \Delta^K$, we have

$$D^{t,t+1}(p^t, p^{t+1}) = D^{t,t+1}(p^t, p^{t+1}(\beta)) - D^{t+1}(p^{t+1}, p^{t+1}(\beta)) \leq D^{t,t+1}(p^t, p^{t+1}(\beta)). \tag{45}$$

Similarly, from Lemma 18 and (43), we have

$$D_U^{t,t+1}(p^t, p^{t+1}(\beta)) = D_U^{t,t+1}(p^t, \bar{p}^{t+1}) + D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\beta)) \tag{46}$$

Combining the above two inequalities, we obtain

$$\begin{aligned}
&D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1}) \leq D^{t,t+1}(p^t, p^{t+1}(\beta)) - D_U^{t,t+1}(q^t, q^{t+1}) \\
&= D_U^{t,t+1}(p^t, p^{t+1}(\beta)) + D_V^{t,t+1}(p^t, p^{t+1}(\beta)) - D_U^{t,t+1}(q^t, q^{t+1}) \\
&= D_U^{t,t+1}(p^t, \bar{p}^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1}) + D_V^{t,t+1}(p^t, p^{t+1}(\beta)) + D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\beta)).
\end{aligned} \tag{47}$$

In the following, we show $D_U^{t,t+1}(p^t, \bar{p}^{t+1}) \leq D_U^{t,t+1}(q^t, q^{t+1})$. From the definition of q^t, there exists $\xi \in \mathbb{R}$ such that

$$\nabla \psi_U^{t+1}(q^{t+1}) = \nabla \psi_U^t(q^t) - \hat{\ell}_U^t + \xi \cdot \mathbf{1}. \tag{48}$$

From Lemma 18 and (43) with $\beta = \xi$, we have

$$D_U^{t,t+1}(p^t, \bar{p}^{t+1}) = D_U^{t,t+1}(p^t, p^{t+1}(\xi)) - D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\xi)) \leq D_U^{t,t+1}(p^t, p^{t+1}(\xi)). \tag{49}$$

From (41), $p^{t+1}(\xi)$ satisfies

$$\nabla \psi_U^{t+1}(p^{t+1}(\xi)) = \nabla \psi_U^t(\bar{p}^t) - \hat{\ell}_U^t + \xi \cdot \mathbf{1}. \tag{50}$$

As $p_i^t \leq q_i^t$ holds for any $i \in U$, and the value of (34) is monotone-increasing w.r.t. p for any fixed ℓ, from Lemma 9, (48) and (50), we have $D_U^{t,t+1}(p^t, p^{t+1}(\xi)) \leq D_U^{t,t+1}(q^t, q^{t+1})$. Combining this with (49) and (47), we obtain (44). ■

Lemma 21 *The value of α defined in (40) is bounded as*

$$\alpha \leq \frac{\sum_{i \in U} (p_i^t)^{3/2} \hat{\ell}_i^t}{\sum_{i \in U} (p_i^t)^{3/2}} \leq \frac{\mathbf{1}[i^t \in U] \sqrt{p_{i^t}^t}}{\sum_{i \in U} (p_i^t)^{3/2}}, \tag{51}$$

$$-\alpha \leq \frac{\sqrt{t+1}}{2t} \frac{\sum_{i \in U} \bar{p}_i^{t+1}}{\sum_{i \in U} (\bar{p}_i^{t+1})^{3/2}} \leq \frac{\sqrt{t+1}}{2t} \frac{|U|}{\sqrt{\bar{p}_j^{t+1}}} \tag{52}$$

for any $j \in U$.

Proof From (40), for any $i \in U$, we have

$$-\frac{\sqrt{t+1}}{\sqrt{\bar{p}_i^{t+1}}} = -\frac{\sqrt{t}}{\sqrt{p_i^t}} - \hat{\ell}_i^t + \alpha \tag{53}$$

which implies that

$$\frac{\bar{p}_i^{t+1}}{t+1} = \frac{1}{\left(\frac{\sqrt{t}}{\sqrt{p_i^t}} + \hat{\ell}_i^t - \alpha\right)^2} \geq \frac{p_i^t}{t} - 2\left(\frac{p_i^t}{t}\right)^{\frac{3}{2}} \cdot (\hat{\ell}_i^t - \alpha), \tag{54}$$

where the inequality follows from the fact that the function $1/x^2$ is convex in $x \in \mathbb{R}_{>0}$, and that its derivative is $-2/x^3$. From (54) and $\sum_{i\in U} \bar{p}_i^{t+1} = \sum_{i\in U} p_i^t$ (this follows from (39)), we have

$$\alpha \leq \frac{\sum_{i\in U}(p_i^t)^{3/2}\hat{\ell}_i^t}{\sum_{i\in U}(p_i^t)^{3/2}} = \frac{\mathbf{1}[i^t \in U]\ell_{i^t}^t\sqrt{p_{i^t}^t}}{\sum_{i\in U}(p_i^t)^{3/2}} \leq \frac{\mathbf{1}[i^t \in U]\sqrt{p_{i^t}^t}}{\sum_{i\in U}(p_i^t)^{3/2}}, \tag{55}$$

where the equality follows from the definition of $\hat{\ell}^t$ in (15) Similarly, from (53), we have

$$\frac{p_i^t}{t} = \frac{1}{\left(\frac{\sqrt{t+1}}{\sqrt{\bar{p}_i^{t+1}}} - \hat{\ell}_i^t + \alpha\right)^2} \geq \frac{\bar{p}_i^{t+1}}{t+1} - 2\left(\frac{\bar{p}_i^{t+1}}{t+1}\right)^{\frac{3}{2}} \cdot (-\hat{\ell}_i^t + \alpha), \tag{56}$$

which implies

$$\begin{aligned}
-\alpha &\leq \frac{1}{2\sum_{t\in U}\left(\frac{\bar{p}_i^{t+1}}{t+1}\right)^{\frac{3}{2}}} \cdot \left(\frac{1}{t}\sum_{i\in U} p_i^t - \frac{1}{t+1}\sum_{i\in U}\bar{p}_i^{t+1}\right) = \frac{\sqrt{t+1}}{2t}\frac{\sum_{i\in U}\bar{p}_i^{t+1}}{\sum_{i\in U}(\bar{p}_i^{t+1})^{3/2}} \\
&\leq \frac{\sqrt{t+1}}{2t} \cdot \frac{|U|\bar{p}_{i^*}^{t+1}}{\sum_{i\in U}\left(\bar{p}_i^{t+1}\right)^{\frac{3}{2}}} \leq \frac{\sqrt{t+1}}{2t} \cdot \frac{|U|\bar{p}_{i^*}^{t+1}}{\left(\bar{p}_{i^*}^{t+1}\right)^{\frac{3}{2}}} = \frac{\sqrt{t+1}}{2t} \cdot \frac{|U|}{\sqrt{\bar{p}_{i^*}^{t+1}}} \leq \frac{\sqrt{t+1}}{2t} \cdot \frac{|U|}{\sqrt{\bar{p}_j^{t+1}}}
\end{aligned}$$

where we set $i^* \in \arg\max_{i\in U} \bar{p}_i^{t+1}$. Hence, we obtain (52). ■

Lemma 22 *Suppose that $\bar{p}^{t+1}$ is given by (39) for $t \geq 4$. We then have*

$$\sum_{i\in U}(\bar{p}_i^{t+1})^{3/2} \leq \sqrt{2}\left(1 + \frac{1}{t}\right)^{3/2}\sum_{i\in U}(p_i^t)^{3/2} \leq 2 \cdot \sum_{i\in U}(p_i^t)^{3/2}. \tag{57}$$

Proof Suppose $\alpha \leq 0$. Then, from (53), we have

$$\frac{\sqrt{t+1}}{\sqrt{\bar{p}_i^{t+1}}} = \frac{\sqrt{t}}{\sqrt{p_i^t}} + \hat{\ell}_i^t - \alpha \geq \frac{\sqrt{t}}{\sqrt{p_i^t}}, \tag{58}$$

which implies $\bar{p}^{t+1} \leq \frac{t+1}{t} p^t_i$. Hence, (57) holds in the case of $\alpha \leq 0$. We next consider the case of $\alpha \geq 0$. For all $i \in U \setminus \{i^t\}$, we have $\frac{\sqrt{t+1}}{\sqrt{\bar{p}^{t+1}_i}} \leq \frac{\sqrt{t}}{\sqrt{p^t_i}}$, therefore $\frac{\bar{p}^{t+1}_i}{t+1} \geq \frac{p^t_i}{t}$. We define $r^t_i = \frac{p^t_i}{\sum_{i \in U} p^t_i}$ and $\bar{r}^{t+1}_i = \frac{t}{t+1} \cdot \frac{\bar{p}^{t+1}_i}{\sum_{i \in U} p^t_i}$. Then $\bar{r}^{t+1}_i \geq r^t_i$ for $i \in U \setminus \{i^t\}$ and $\sum_{i \in U} \bar{r}^{t+1}_i \leq \sum_{i \in U} r^t_i = 1$. We have

$$\frac{\sum_{i \in U} (\bar{p}^{t+1}_i)^{3/2}}{\sum_{i \in U} (p^t_i)^{3/2}} = \left(1 + \frac{1}{t}\right)^{3/2} \frac{\sum_{i \in U} (\bar{r}^{t+1}_i)^{3/2}}{\sum_{i \in U} (r^t_i)^{3/2}}. \tag{59}$$

The maximum of $\frac{\sum_{i \in U} (\bar{r}^{t+1}_i)^{3/2}}{\sum_{i \in U} (r^t_i)^{3/2}}$ subject to $r^t_i, \bar{r}^{t+1}_i \geq 0$ for $i \in U$, $\bar{r}^{t+1}_i \geq r^t_i$ for $i \in U \setminus \{i^t\}$ and $\sum_{i \in U} \bar{r}^{t+1}_i \leq \sum_{i \in U} r^t_i = 1$ is at most $\sqrt{2}$, which follows from Lemma 28. Hence, from (59), we have $\frac{\sum_{i \in U} (\bar{p}^{t+1}_i)^{3/2}}{\sum_{i \in U} (p^t_i)^{3/2}} = \sqrt{2} \left(1 + \frac{1}{t}\right)^{3/2}$, which implies (57). ■

In the following, we set

$$\beta = \max \left\{ 0, \frac{\sum_{i \in U} (p^t_i)^{3/2}}{\sum_{i=1}^K (p^t_i)^{3/2}} \alpha \right\}. \tag{60}$$

Then, it follows from (51) that

$$\beta \leq \frac{\mathbf{1}[i^t \in U] \sqrt{p^t_{i^t}}}{\sum_{i=1}^K (p^t_i)^{3/2}} \leq \sqrt{K} \leq \frac{\sqrt{t}}{2} \tag{61}$$

under the assumption of $t \geq 4K$, which implies that

$$\frac{\sqrt{t+1}}{\sqrt{p^{t+1}_i(\beta)}} = \frac{\sqrt{t}}{\sqrt{p^t_i}} + \hat{\ell}^t_i - \beta \geq \frac{\sqrt{t}}{\sqrt{p^t_i}} - \beta \geq \frac{\sqrt{t}}{2\sqrt{p^t_i}}. \tag{62}$$

Hence, we have

$$\sqrt{p^{t+1}_i(\beta)} \leq 2 \sqrt{\frac{t+1}{t}} \sqrt{p^t_i} \tag{63}$$

for all $i \in [K]$. We are now ready to prove Lemma 10.

Proof of Lemma 10 We consider evaluating the right-hand side of (44) using Lemma 19. From (40) and (41), p^t, $p^{t+1}(\beta)$ and $\bar{p}^{t+1}$ satisfies

$$\nabla \psi^{t+1}_V (p^{t+1}(\beta)) = \nabla \psi^t_V (p^t) - \hat{\ell}^t_V + \beta \cdot \mathbf{1}, \tag{64}$$

$$\nabla \psi^{t+1}_U (p^{t+1}(\beta)) = \nabla \psi^{t+1}_U (\bar{p}^{t+1}) - (\alpha - \beta) \cdot \mathbf{1}. \tag{65}$$

We use Lemma 19 with $\ell = \hat{\ell}^t_V - \beta \cdot \mathbf{1}$ and $\ell = (\alpha - \beta) \cdot \mathbf{1}$ to bound the right-hand side of (44). We can confirm that the assumption of $\frac{\sqrt{p_i} \ell_i}{\gamma^t} \geq -1/2$ in Lemma 19 is satisfied as $\hat{\ell}^t_i \geq 0$ and β is bounded as in (61).

Suppose that $i^t \in V$. We then have $\alpha \leq 0$ from (51) and $\beta = 0$ from (60). Hence, from (64), (65) and Lemma 19, we have

$$D_V^{t,t+1}(p^t, p^{t+1}(\beta)) \leq \frac{1}{\sqrt{t}} \sum_{i \in V} (p_i^t)^{3/2} (\hat{\ell}_i^t)^2 + \frac{1}{\sqrt{t+1}} \sum_{i \in V} \sqrt{p_i^{t+1}(\beta)}$$
$$\leq \frac{1}{\sqrt{t}} \left(\frac{1}{\sqrt{p_{i^t}^t}} + 2 \sum_{i \in V} \sqrt{p_i^t} \right), \tag{66}$$
$$D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\beta)) \leq \frac{2}{\sqrt{t}} \sum_{i \in U} (\bar{p}_i^{t+1})^{3/2} \cdot \mathbf{1}[\alpha < 0] \cdot (\alpha)^2, \tag{67}$$

where the second inequality follows from (62). On the other hand, if $i^t \in U$, we have $\alpha - \beta \geq 0$, $\beta \geq 0$, and Lemma 19 with (64) and (65) implies the following:

$$D_V^{t,t+1}(p^t, p^{t+1}(\beta)) \leq \frac{2}{\sqrt{t}} \sum_{i \in V} (p_i^t)^{3/2} \cdot (\beta)^2 + \frac{1}{\sqrt{t+1}} \sum_{i \in V} \sqrt{p_i^{t+1}(\beta)}$$
$$\leq \frac{2}{\sqrt{t}} \sum_{i \in V} (p_i^t)^{3/2} \cdot (\beta)^2 + \frac{2}{\sqrt{t}} \sum_{i \in V} \sqrt{p_i^t}, \tag{68}$$
$$D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\beta)) \leq \frac{1}{\sqrt{t}} \sum_{i \in U} (\bar{p}_i^{t+1})^{3/2} \left(\mathbf{1}[\alpha > 0] \cdot (\alpha - \beta)^2 + 2 \cdot \mathbf{1}[\alpha < 0] \cdot (\alpha)^2 \right)$$
$$\leq \frac{2}{\sqrt{t}} \left(\mathbf{1}[\alpha > 0] \cdot (\alpha - \beta)^2 \cdot \sum_{i \in U} (p_i^t)^{3/2} + \mathbf{1}[\alpha < 0] \cdot (\alpha)^2 \cdot \sum_{i \in U} (\bar{p}_i^{t+1})^{3/2} \right), \tag{69}$$

where the second inequality follows from (62) and the last inequality follows from (57). We can show the first inequality above by applying Lemma 19 with $\ell = -\beta \cdot \mathbf{1}$. More precisely, as the condition $i^t \in U$ implies $\hat{\ell}_V^t = 0$, from (64), we can apply Lemma 19 with $\ell = -\beta \cdot \mathbf{1}$. Similarly, we obtain the third inequality by applying Lemma 19 with $\ell = (\alpha - \beta) \cdot \mathbf{1}$. Combining (68), (66), (69) and (67), we obtain

$$D_V^{t,t+1}(p^t, p^{t+1}(\beta)) + D_U^{t+1}(\bar{p}^{t+1}, p^{t+1}(\beta)) \leq \frac{1}{\sqrt{t}} \left(2 \sum_{i \in V} \sqrt{p_i^{t+1}} + \mathbf{1}[i^t \in V] \frac{1}{\sqrt{p_{i^t}^t}} \right) \tag{70}$$
$$+ \frac{2}{\sqrt{t}} \left(\mathbf{1}[\alpha > 0] \left((\beta)^2 \sum_{i \in V} (p_i^t)^{3/2} + (\alpha - \beta)^2 \sum_{i \in U} (p_i^t)^{3/2} \right) + \mathbf{1}[\alpha < 0] (\alpha)^2 \sum_{i \in U} (\bar{p}_i^{t+1})^{3/2} \right).$$

From (51) and (60), we have

$$\mathbf{1}[\alpha > 0] \left((\beta)^2 \sum_{i \in V} (p_i^t)^{3/2} + (\alpha - \beta)^2 \sum_{i \in U} (p_i^t)^{3/2} \right)$$
$$= \mathbf{1}[\alpha > 0] \cdot \frac{(\alpha)^2}{\sum_{i=1}^K (p_i^t)^{3/2}} \cdot \sum_{i \in V} (p_i^t)^{3/2} \cdot \sum_{i \in U} (p_i^t)^{3/2} \leq \frac{\mathbf{1}[i^t \in U] \cdot p_{i^t}^t \cdot \sum_{i \in V} (p_i^t)^{3/2}}{\sum_{i=1}^K (p_i^t)^{3/2} \cdot \sum_{i \in U} (p_i^t)^{3/2}}, \tag{71}$$

where the equality follows from (60) and the inequality follows from (51).

From (52), we have

$$\mathbf{1}[\alpha<0]\cdot(\alpha)^2\cdot\sum_{i\in U}(\bar{p}_i^{t+1})^{3/2}\le\frac{t+1}{4t^2}\sum_{i\in U}(\bar{p}_i^{t+1})^{3/2}\left(\frac{\sum_{i\in U}\bar{p}_i^{t+1}}{\sum_{i\in U}(\bar{p}_i^{t+1})^{3/2}}\right)^2$$
$$=\frac{t+1}{4t^2}\cdot\frac{\left(\sum_{i\in U}\bar{p}_i^{t+1}\right)^2}{\sum_{i\in U}(\bar{p}_i^{t+1})^{3/2}}=\frac{t+1}{4t^2}\cdot\frac{\sqrt{\sum_{i\in U}\bar{p}_i^{t+1}}}{\sum_{i\in U}\left(\frac{\bar{p}_i^{t+1}}{\sum_{j\in U}\bar{p}_j^{t+1}}\right)^{3/2}}\le\frac{(t+1)\sqrt{U}}{4t^2},\tag{72}$$

where the first inequality follows from (52) and the last inequality follows from $\sum_{i\in U}\bar{p}_i^{t+1}\le 1$ and $\sum_{i\in U}\left(\frac{\bar{p}_i^{t+1}}{\sum_{j\in U}\bar{p}_j^{t+1}}\right)^{3/2}\ge\frac{1}{\sqrt{|U|}}$. Combining (70), (71) and (72), we obtain the bound of Lemma 10. ■

Appendix D. Proof of Theorem 2

Lemma 23 *We have*

$$\mathbf{E}^t\left[\frac{\mathbf{1}[i^t\in U]\cdot p_{i^t}^t\cdot\sum_{i\in V}(p_i^t)^{\frac{3}{2}}}{\sum_{i\in U}(p_i^t)^{\frac{3}{2}}\sum_{i=1}^K(p_i^t)^{\frac{3}{2}}}\right]\le\sqrt{\sum_{i\in V}p_i^t}.\tag{73}$$

Proof As we have $\mathbf{E}^t\left[\mathbf{1}[i^t\in U]\cdot p_{i^t}^t\right]=\sum_{i\in U}(p_i^t)^2$, we have

$$\mathbf{E}^t\left[\frac{\mathbf{1}[i^t\in U]\cdot p_{i^t}^t\sum_{i\in V}(p_i^t)^{\frac{3}{2}}}{\sum_{i\in U}(p_i^t)^{\frac{3}{2}}\sum_{i=1}^K(p_i^t)^{\frac{3}{2}}}\right]=\frac{\sum_{i\in U}(p_i^t)^2\sum_{i\in V}(p_i^t)^{\frac{3}{2}}}{\sum_{i\in U}(p_i^t)^{\frac{3}{2}}\sum_{i=1}^K(p_i^t)^{\frac{3}{2}}}$$
$$\le\frac{\sum_{i\in U}(p_i^t)^2\sum_{i\in V}(p_i^t)^{\frac{3}{2}}}{\sum_{i\in U}(p_i^t)^{\frac{3}{2}}\cdot 3^{\frac{1}{3}}\cdot\left(\frac{3}{2}\right)^{\frac{2}{3}}\left(\sum_{i\in U}(p_i^t)^{\frac{3}{2}}\right)^{\frac{1}{3}}\left(\sum_{i\in V}(p_i^t)^{\frac{3}{2}}\right)^{\frac{2}{3}}}$$
$$=\frac{2^{\frac{2}{3}}\sum_{i\in U}(p_i^t)^2\left(\sum_{i\in V}(p_i^t)^{\frac{3}{2}}\right)^{\frac{1}{3}}}{3\left(\sum_{i\in U}(p_i^t)^{\frac{3}{2}}\right)^{\frac{4}{3}}}\le\left(\sum_{i\in V}(p_i^t)^{\frac{3}{2}}\right)^{\frac{1}{3}}\le\sqrt{\sum_{i\in V}p_i^t}$$

where the first inequality follows from $\frac{1}{3}a+\frac{2}{3}b\ge a^{\frac{1}{3}}b^{\frac{2}{3}}$ for $a,b\ge 0$, the second inequality follows from $\|x\|_2\le\|x\|_{\frac{3}{2}}$, and the last inequality follows from $\sum_{i\in V}(p_i^t)^{\frac{3}{2}}\le\left(\sum_{i\in V}p_i^t\right)^{\frac{3}{2}}$. ■

Lemma 24 *Suppose that p^t and q^t are defined by (15) and (16). For any $t\ge 1$, we have*

$$\mathbf{E}\left[D^{t,t+1}(p^t,p^{t+1})-D_U^{t,t+1}(q^t,q^{t+1})\right]\le 2\frac{\sqrt{K}}{\sqrt{t}}.\tag{74}$$

Proof As we have $\psi_U^{t+1}(p) \leq \psi_U^t(p)$ for any U and p from the definition of ψ_U^t, we have

$$D_U^{t,t+1}(q^t, q^{t+1}) = D_U^{t+1}(q^t, q^{t+1}) + \psi_U^t(q^t) - \psi_U^{t+1}(q^t) \geq 0. \tag{75}$$

From (45) with $\beta = 0$, we have

$$D^{t,t+1}(p^t, p^{t+1}) \leq D^{t,t+1}(p^t, p^{t+1}(0)), \tag{76}$$

where $p^{t+1}(0)$ is define by (41) with $\beta = 0$. From Lemma 19 with $\ell = \hat{\ell}^t$, we have

$$\begin{aligned} D^{t,t+1}(p^t, p^{t+1}(0)) &\leq \frac{1}{\sqrt{t}} \sum_{i=1}^K (p_i^t)^{3/2} (\hat{\ell}_i^t)^2 + \frac{1}{\sqrt{t+1}} \sum_{i=1}^K \sqrt{p_i^{t+1}(0)} \\ &\leq \frac{1}{\sqrt{t}} \frac{1}{\sqrt{p_{i^t}^t}} + \frac{1}{\sqrt{t}} \sum_{i=1}^K \sqrt{p_i^t}. \end{aligned}$$

where the second inequality follows from the definition of $\hat{\ell}^t$ in (15) and from (62) with $\beta = 0$. By taking the expectation w.r.t. i^t, we obtain

$$\mathbf{E}\left[D^{t,t+1}(p^t, p^{t+1}(0))\right] \leq \frac{2}{\sqrt{t}} \sum_{i=1}^K \sqrt{p_i^t} \leq 2\frac{\sqrt{K}}{\sqrt{t}}. \tag{77}$$

By combining (75), (76) and (77), we obtain (74). ■

From Lemmas 10, 23 and 24, we have

$$\begin{aligned} &\sum_{t=1}^T \mathbf{E}\left[D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1})\right] \\ &= \sum_{t=1}^{4K} \mathbf{E}\left[D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1})\right] + \sum_{t=4K+1}^T \mathbf{E}\left[D^{t,t+1}(p^t, p^{t+1}) - D_U^{t,t+1}(q^t, q^{t+1})\right] \\ &\leq 2\sum_{t=1}^{4K} \frac{\sqrt{K}}{\sqrt{t}} + \sum_{t=4K+1}^T \left(\frac{5}{\sqrt{t}} \sum_{i \in V} \sqrt{\mathbf{E}\left[p_i^t\right]} + \frac{\sqrt{K}}{t^{3/2}}\right) \\ &\leq 4\sqrt{K} \cdot \sum_{t=1}^{4K} \left(\sqrt{t} - \sqrt{t-1}\right) + \sum_{t=4K+1}^T \left(\frac{5}{\sqrt{t}} \sum_{i \in V} \sqrt{\mathbf{E}\left[p_i^t\right]}\right) + \frac{\sqrt{K}}{2} \sum_{t=4K+1}^T \left(\frac{1}{\sqrt{t-1}} - \frac{1}{\sqrt{t}}\right) \\ &\leq 9K + \sum_{t=1}^T \left(\frac{5}{\sqrt{t}} \sum_{i \in V} \sqrt{\mathbf{E}\left[p_i^t\right]}\right) \end{aligned}$$

Combining this with Lemma 9, (17), and $D^1(q^1, p^1) \leq K$ we have

$$R_{i^*}^T \leq \sum_{t=1}^T \left(\frac{5}{\sqrt{t}} \sum_{i \in V} \sqrt{\mathbf{E}\left[p_i^t\right]}\right) + 10K + D. \tag{78}$$

From this and (2), it holds for any $\lambda > 0$ that

$$\begin{aligned}
R_{i^*}^T &= (1+\lambda)R_{i^*}^T - \lambda R_{i^*}^T \\
&\leq (1-\lambda)\cdot\left(\sum_{t=1}^T\left(\frac{5}{\sqrt{t}}\sum_{i\in V}\sqrt{\mathbf{E}\left[p_i^t\right]}\right)+10K+D\right)-\lambda\cdot\left(\sum_{t=1}^T\sum_{i\in V}\Delta_i\,\mathbf{E}\left[p_i^t\right]-C\right) \\
&= \sum_{t=1}^T\sum_{i\in V}\left(\frac{5(1+\lambda)}{\sqrt{t}}\sqrt{\mathbf{E}\left[p_i^t\right]}-\lambda\cdot\Delta_i\,\mathbf{E}\left[p_i^t\right]\right)+(1+\lambda)\cdot(K+D)+\lambda C \\
&\leq \frac{25(1+\lambda)^2}{4\lambda}\sum_{t=1}^T\frac{1}{t}\sum_{i\in V}\frac{1}{\Delta_i}+(1+\lambda)\cdot(K+D)+\lambda C \\
&\leq \frac{25(1+\lambda)^2\log(T+1)}{4\lambda}\sum_{i\in V}\frac{1}{\Delta_i}+(1+\lambda)\cdot(K+D)+\lambda C.
\end{aligned}$$

By choosing $\lambda = \min\left\{1, \frac{5\sqrt{\log(T+1)\sum_{i\in V}\frac{1}{\Delta_i}}}{\sqrt{C}}\right\}$, we have

$$R_{i^*}^T \leq 25\log(T+1)\sum_{i\in V}\frac{1}{\Delta_i}+10\sqrt{C\log(T+1)\sum_{i\in V}\frac{1}{\Delta_i}}+2(K+D), \tag{79}$$

which proves Theorem 2. ■

Appendix E. Proof of Proposition 11

Set p^* by $p^* = (1-\varepsilon)\cdot\chi_{i^*}+\frac{\varepsilon}{K}\cdot\mathbf{1}$. As $\hat{\ell}^t$ is an unbiased estimator, the regret can be expressed as

$$\begin{aligned}
R_{i^*}^T &= \mathbf{E}\left[\sum_{t=1}^T\left(\ell_{i^t}^t-\ell_{i^*}^t\right)\right] = \mathbf{E}\left[\sum_{t=1}^T\left(\left\langle\ell^t,p^t-p^*\right\rangle\right)+\sum_{t=1}^T\left(\left\langle\ell^t,p^*-\chi_{i^*}\right\rangle\right)\right] \\
&= \mathbf{E}\left[\sum_{t=1}^T\left\langle\hat{\ell}^t,p^t-p^*\right\rangle+\varepsilon\sum_{t=1}^T\left\langle\ell^t,\frac{1}{K}\mathbf{1}-\chi_{i^*}\right\rangle\right] \leq \mathbf{E}\left[\sum_{t=1}^T\left\langle\hat{\ell}^t,p^t-p^*\right\rangle\right]+\varepsilon T.
\end{aligned} \tag{80}$$

Define r^t by $r^t = \Psi^t\left(\sum_{j=1}^{t-1}\hat{\ell}^j\right)$. From (13) in Proposition 8, we have

$$\sum_{t=1}^T\left\langle\hat{\ell}^t,p^t-p^*\right\rangle \leq D^1(p^*,p^1)+\sum_{t=1}^T D^{t,t+1}(p^t,r^{t+1})+\psi^{T+1}(p^*)-\psi^1(p^*). \tag{81}$$

The value of $D^{t,t+1}(p^t,r^{t+1})$ can be bounded via the following lemma:

Lemma 25 *Suppose ψ^t is defined by (20) and $\gamma_i^{t+1}\geq\gamma_i^t$ for all $i\in[K]$. Suppose that $p,q\in\Omega$ are expressed as $p=\Psi^t(L), q=\Psi^{t+1}(L+\ell)$ by $L,\ell\in\mathbb{R}^K$ such that $|\frac{p_i\ell_i}{\gamma_i^t}|\leq\frac{1}{2}$ for all $i\in[K]$. We then have*

$$D^{t,t+1}(p,q)\leq\sum_{i=1}^K\gamma_i^t\cdot g\left(\frac{p_i\ell_i}{\gamma_i^t}\right)\leq\sum_{i=1}^K\left(\frac{(p_i\ell_i)^2}{2\gamma_i^t}+\frac{|p_i\ell_i|^3}{(\gamma_i^t)^2}\right)\leq\sum_{i=1}^K\frac{(p_i\ell_i)^2}{\gamma_i^t} \tag{82}$$

where $g(x)$ is defined as $g(x)=-\log(x+1)+x$.

Proof From Lemma 17, we have

$$D^{t,t+1}(p,q) = \langle \ell, p-q\rangle - D^{t+1,t}(q,p) = \langle \ell, p-q\rangle - D^t(q,p) - \psi^{t+1}(q) + \psi^t(q) \\ \leq \langle \ell, p-q\rangle - D^t(q,p), \tag{83}$$

where the first equality follows from Lemma 17, the second equality follows from the definition (11) of the skewed Bregman divergence, and tha inequality follows from the definition of ψ^t in (20) and the assumption of $\gamma_i^{t+1} \geq \gamma_i^t$. Consider the following maximum value:

$$\max_{q\in\mathbb{R}^K_{>0}} \{f(q)\}, \quad \text{where} \quad f(q) = \langle \ell, p-q\rangle - D^t(q,p). \tag{84}$$

As the function $f(q)$ is concave in $q \in \mathbb{R}^K_{>0}$ for any fixed p, the maximum is attained by q^* satisfying

$$\nabla f(q^*) = -\ell - \nabla\psi^t(q^*) + \nabla\psi^t(p) = 0. \tag{85}$$

A vector $q^* \in \mathbb{R}^K_{>0}$ satisfying (85) indeed exists under the condition of $|\frac{p_i\ell_i}{\gamma_i^t}| \leq \frac{1}{2}$. As $\nabla\psi^t(p) = \left(-\frac{\gamma_i^t}{p_i}\right)_{i=1}^K$, (85) implies

$$\frac{\ell_i p_i}{\gamma_i^t} = \frac{p_i}{\gamma_i^t}\cdot\left(\frac{\gamma_i^t}{q_i^*} - \frac{\gamma_i^t}{p_i}\right) = \frac{p_i}{q_i^*} - 1. \tag{86}$$

For this q^*, we have

$$\begin{aligned} f(q^*) &= \langle \ell, p-q^*\rangle - D^t(q^*,p) = \left\langle \nabla\psi^t(q^*) - \nabla\psi^t(p), p-q^*\right\rangle - D^t(q^*,p) \\ &= \psi^t(p) - \psi^t(q^*) - \left\langle \nabla\psi^t(q^*), p-q^*\right\rangle \\ &= \sum_{i=1}^K \gamma_i^t\cdot\left(-\log\frac{p_i}{q_i^*} + \frac{p_i-q_i^*}{q_i^*}\right) \\ &= \sum_{i=1}^K \gamma_i^t\cdot\left(-\log\left(1+\frac{\ell_i p_i}{\gamma_i^t}\right) + \frac{\ell_i p_i}{\gamma_i^t}\right) = \sum_{i=1}^K \gamma_i^t\cdot g\left(\frac{\ell_i p_i}{\gamma_i^t}\right), \end{aligned} \tag{87}$$

which means that the first inequality in (82) holds. The other inequality in (82) follows from the definition of $g(x) = -\log(x+1)+x$ and the assumption of $|\frac{p_i\ell_i}{\gamma_i^t}| \leq \frac{1}{2}$. In fact, as we have $g(0) = 0$, $g'(0) = 0$ and $g''(x) = \frac{1}{(1+x)^2} \leq 1 + 6|x|$ for $x \geq -\frac{1}{2}$, we have $g(x) \leq \frac{x^2}{2} + |x|^3 = x^2(\frac{1}{2} + |x|)$. Hence, for $x \in [-\frac{1}{2}, \frac{1}{2}]$, we have $g(x) \leq x^2(\frac{1}{2} + |x|) \leq x^2$. Combining these with (87), we obtain (82). ■

From the definition of the algorithm, p^t and r^{t+1} can be expressed as $p^t = \Psi^{t+1}(L)$ and $r^{t+1} = \Psi^{t+1}(L+\ell)$ where $L = \sum_{j=1}^{t-1}\hat{\ell}^j + m^t$ and $\ell = \hat{\ell}^t - m^t + \alpha\mathbf{1}$ for any $\alpha \in \mathbb{R}$. We choose $\alpha = -(\ell_{i^t}^t - m_{i^t}^t)$ in the following. Then ℓ is expressed as $\ell = (\ell_{i^t}^t - m_{i^t}^t)(\frac{1}{p_{i^t}^t}\chi_{i^t} - \mathbf{1})$. As $|\frac{p_i^t\ell_i}{\gamma_i^t}| \leq \frac{1}{2}$ follows from the conditions that $\ell_{i^t}^t, m_{i^t}^t \in [0,1]$ and $\gamma_i^t \geq 2$, we can apply Lemma 25 to p^t and r^{t+1} to obtain

$$D^{t,t+1}(p^t, r^{t+1}) \leq \sum_{i=1}^K \frac{(p_i^t\ell_i)^2}{\gamma_i^t} = (\ell_{i^t}^t - m_{i^t}^t)^2\left(\frac{(1-p_{i^t}^t)^2}{\gamma_{i^t}^t} + \sum_{i\in[K]\setminus\{i^t\}}\frac{(p_i^t)^2}{\gamma_i^t}\right) \leq \sum_{i=1}^K \frac{\nu_i^t}{\gamma_i^t},$$

where ν_i^t is defined in (22). From this and the definition of γ_i^t in (22), we have

$$\sum_{t=1}^{T} D^{t,t+1}(p^t, r^{t+1}) \leq \sum_{i=1}^{K}\sum_{t=1}^{T} \frac{\nu_i^t}{\gamma_i^t} = \sum_{i=1}^{K}\sum_{t=1}^{T} \frac{2}{B}\left(\gamma_i^{t+1} - \gamma_i^t\right) = \frac{2}{B}\sum_{i=1}^{K}\left(\gamma_i^{T+1} - \gamma_i^1\right). \tag{88}$$

Further, we have

$$\begin{aligned} &D^1(p^*, r^1) + \psi^{T+1}(p^*) - \psi^1(p^*) = \psi^{T+1}(p^*) - \psi^1(r^1) - \left\langle \nabla\psi^1(r^1), p^* - r^1 \right\rangle \\ &\leq \psi^{T+1}(p^*) = -\sum_{i=1}^{K} \gamma_i^{T+1} \log(p_i^*) \leq \log\frac{K}{\varepsilon}\sum_{i=1}^{K}\gamma_i^{T+1}, \end{aligned} \tag{89}$$

where the first equality follows from the definition (11) of the Bregman divergence, the first inequality follows from $\psi^1(r^1) \geq 0$ and that $r^1 = \frac{1}{K}\cdot\mathbf{1}$ implies $\left\langle \nabla^1\psi^1(r^1), p^* - r^1\right\rangle = -2K\cdot\left\langle \mathbf{1}, p^* - r^1\right\rangle = 0$, and the last inequality follows from $p_i^* \geq \frac{\varepsilon}{K}$ for all $i \in [K]$. Combining (81), (88) and (89), we obtain

$$\sum_{t=1}^{T}\left\langle \hat{\ell}^t, p^t - p^*\right\rangle \leq \sum_{i=1}^{K}\left(\frac{2}{B}\left(\gamma_i^{T+1} - \gamma_i^1\right) + \log\frac{K}{\varepsilon}\gamma_i^{T+1}\right). \tag{90}$$

We can bound this using ν_i^t via the following lemma:

Lemma 26 *Suppose that γ_i^t is defined as in (22) with $\nu_i^t \in [0,1]$ and $B \in [0,1]$. We then have*

$$\gamma_i^t \leq \sqrt{B\sum_{j=1}^{t-1}\nu_i^j} + 2 \tag{91}$$

for all $t \geq 1$.

Proof We show this lemma by induction in t. For $t = 1$, (91) holds as $\gamma_i^1 = 2$. Suppose (91) holds for a fixed t. Denote $\bar{\gamma}_i^t = \sqrt{B\sum_{j=1}^{t-1}\nu_i^j} + 2$. We then have

$$\begin{aligned} \bar{\gamma}_i^{t+1} - \bar{\gamma}_i^t &= \sqrt{B\sum_{j=1}^{t}\nu_i^j} - \sqrt{B\sum_{j=1}^{t-1}\nu_i^j} = \frac{B\nu_i^t}{\sqrt{B\sum_{j=1}^{t}\nu_i^j} + \sqrt{B\sum_{j=1}^{t-1}\nu_i^j}} \\ &\geq \frac{B\nu_i^t}{2\sqrt{B\sum_{j=1}^{t-1}\nu_i^j + 1}} \geq \frac{B\nu_i^t}{2\bar{\gamma}_i^t}, \end{aligned}$$

where the first inequality follows from $B\nu_i^t \in [0,1]$ and $\sqrt{x+1} \leq \sqrt{x} + 1$ for $x \geq 0$. From this, we have

$$\bar{\gamma}_i^{t+1} \geq \bar{\gamma}_i^t + \frac{B\nu_i^t}{2\bar{\gamma}_i^t} \geq \gamma_i^t + \frac{B\nu_i^t}{2\gamma_i^t} = \gamma_i^{t+1},$$

where the second inequality follows from the inductive hypothesis of $\gamma_i^t \leq \bar{\gamma}_i^t$ and the fact that $x + c/x$ is monotone increasing in $x > \sqrt{c}$, for any fixed $c \geq 0$. Hence, by induction in t, (91) is shown for all $t \geq 1$. ■

Combining this lemma, (90) and $\gamma_i^1 = 2$, we have

$$\sum_{t=1}^{T} \left\langle \hat{\ell}^t, p^t - p^* \right\rangle \leq \left(\frac{2}{B} + \log \frac{K}{\varepsilon} \right) \sum_{i=1}^{K} \sqrt{B \sum_{t=1}^{T} \nu_i^t + 2K \log \frac{K}{\varepsilon}}$$
$$= \left(\frac{2}{\sqrt{B}} + \sqrt{B} \log \frac{K}{\varepsilon} \right) \sum_{i=1}^{K} \sqrt{\sum_{t=1}^{T} \nu_i^t + 2K \log \frac{K}{\varepsilon}}.$$

By combining this and (80), we obtain the regret bound in Proposition 11.

Appendix F. Proof of Lemma 12

From Jensen's inequality, for any $i \in [K]$, we have

$$\mathbf{E}\left[\sqrt{\sum_{t=1}^{T} \nu_i^t} \right] \leq \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} \nu_i^t \right]} \leq \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} p_i^t (1 - p_i^t) \right]} \leq \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} p_i^t \right]},$$

where the second inequality follows from the definition of ν_i^t in (22). Further, for $i = i^*$, we have

$$\mathbf{E}\left[\sqrt{\sum_{t=1}^{T} \nu_{i^*}^t} \right] \leq \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} p_{i^*}^t (1 - p_{i^*}^t) \right]} \leq \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} p_i^t \right]} \leq \sum_{i \neq i^*} \sqrt{\mathbf{E}\left[\sum_{t=1}^{T} p_i^t \right]}$$

Combining these two inequalities, we obtain

$$\mathbf{E}\left[\sum_{i=1}^{K} \sqrt{\sum_{t=1}^{T} \nu_{i^*}^t} \right] \leq 2 \sum_{i \neq i^*} \sqrt{\sum_{t=1}^{T} \mathbf{E}\left[p_i^t \right]},$$

which means that the first part of Lemma 12 holds. The second part can be shown as follows:

$$\sum_{i=1}^{K} \sqrt{\sum_{t=1}^{T} \nu_i^t} \leq \sqrt{K \sum_{t=1}^{T} \sum_{i=1}^{K} \nu_i^t} = \sqrt{K \sum_{t=1}^{T} (\ell_{i^t}^t - m_{i^t}^t)^2 \left((1 - p_{i^t}^t)^2 + \sum_{i \neq i^t} (p_i^t)^2 \right)}$$
$$\leq \sqrt{2K \sum_{t=1}^{T} (\ell_{i^t}^t - m_{i^t}^t)^2},$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the equality follows from the definition of ν_i^t in (22). This means that the second part of Lemma 12 holds.

Appendix G. Proof of Proposition 13

For any $\{u_t\}_{t=1}^T \subseteq [0,1]^K$, from (23), we have

$$\begin{aligned}
(\ell_{i^t}^t - m_{i^t}^t)^2 - (\ell_{i^t}^t - u_{i^t}^t)^2 &\le 2(\ell_{i^t}^t - m_{i^t}^t)(u_{i^t}^t - m_{i^t}^t) \\
&= 2(\ell_{i^t}^t - m_{i^t}^t)(m_{i^t}^{t+1} - m_{i^t}^t) + 2(\ell_{i^t}^t - m_{i^t}^t)(u_{i^t}^t - m_{i^t}^{t+1}) \\
&= 2\eta(\ell_{i^t}^t - m_{i^t}^t)^2 + \frac{2}{\eta}(m_{i^t}^{t+1} - m_{i^t}^t)(u_{i^t}^t - m_{i^t}^{t+1}) \\
&\le 2\eta(\ell_{i^t}^t - m_{i^t}^t)^2 + \frac{1}{\eta}((u_{i^t}^t - m_{i^t}^t)^2 - (u_{i^t}^t - m_{i^t}^{t+1})^2) \\
&= 2\eta(\ell_{i^t}^t - m_{i^t}^t)^2 + \frac{1}{\eta}(\|u^t - m^t\|_2^2 - \|u^t - m^{t+1}\|_2^2),
\end{aligned}$$

where the inequalities follows from $y^2 - x^2 = 2y(y-x) - (x-y)^2 \le 2y(y-x)$ that holds for any $x, y \in \mathbb{R}$, and the last inequality holds since (23) implies $(u_i^t - m_i^t)^2 = (u_i^t - m_i^{t+1})^2$ for $i \in [K] \setminus \{i^t\}$. Hence, we have

$$(\ell_{i^t}^t - m_{i^t}^t)^2 \le \frac{1}{1-2\eta}\left((\ell_{i^t}^t - u_{i^t}^t)^2 + \frac{1}{\eta}(\|u^t - m^t\|_2^2 - \|u^t - m^{t+1}\|_2^2)\right). \tag{92}$$

Taking the summation for $t \in [T]$ and by telescoping, we obtain

$$\begin{aligned}
&\sum_{t=1}^T (\ell_{i^t}^t - m_{i^t}^t)^2 \\
&\le \frac{1}{1-2\eta}\sum_{t=1}^T (\ell_{i^t}^t - u_{i^t}^t)^2 + \frac{1}{\eta(1-2\eta)}\left(\|u^1 - m^1\|_2^2 + \sum_{t=1}^{T-1}(\|u^{t+1} - m^{t+1}\|_2^2 - \|u^t - m^{t+1}\|_2^2)\right) \\
&\le \frac{1}{1-2\eta}\sum_{t=1}^T (\ell_{i^t}^t - u_{i^t}^t)^2 + \frac{1}{\eta(1-2\eta)}\left(\frac{K}{4} + 2\sum_{t=1}^{T-1}(u^{t+1} - m^{t+1})^\top (u^{t+1} - u^t)\right) \\
&\le \frac{1}{1-2\eta}\sum_{t=1}^T (\ell_{i^t}^t - u_{i^t}^t)^2 + \frac{1}{\eta(1-2\eta)}\left(\frac{K}{4} + 2\sum_{t=1}^{T-1}\|u^{t+1} - u^t\|_1\right),
\end{aligned}$$

where the second inequality follows from $m^1 = \frac{1}{2}\mathbf{1}$ and the convexity of $x \mapsto \|x\|_2^2$, and the last inequality follows from $\|u^{t+1} - m^{t+1}\|_\infty \le 1$.

Appendix H. Other Lemmas

Lemma 27 *Suppose that $R \ge 0$ satisfies $R \le a\sqrt{R+L} + b$ for $a, b, L \ge 0$. We then have $R \le a\sqrt{L} + a^2 + 2b$.*

Proof We assume $R - b \ge 0$ because the condition of $R - b < 0$ immediately implies $R \le a\sqrt{L} + a^2 + 2b$. From this and the assumption of $R \le a\sqrt{R+L} + b$, we have

$$0 \ge (R-b)^2 - (a\sqrt{R+L})^2 = R^2 - 2bR + b^2 - a^2(R+L) = R^2 - (2b + a^2)R + b^2 - a^2 L.$$

By solving this quadratic inequation in R, we obtain

$$\begin{aligned} R &\leq \frac{1}{2}\left(2b + a^2 + \sqrt{(2b+a^2)^2 + 4(a^2 L - b^2)}\right) \\ &\leq \frac{1}{2}\left(2b + a^2 + \sqrt{(2b+a^2)^2} + \sqrt{4(a^2 L - b^2)}\right) \leq 2b + a^2 + a\sqrt{L}, \end{aligned}$$

which completes the proof. ■

Lemma 28 *Suppose that $p, q \in \Delta^K$ satisfy $q_K \leq p_K$ and $q_i \geq p_i$ for all $i \in [K] \setminus \{K\} = [K-1]$. We then have*

$$\frac{\sum_{i=1}^K q_i^{3/2}}{\sum_{i=1}^K p_i^{3/2}} \leq \sqrt{2}. \tag{93}$$

Proof Define g by $g(p) = \sum_{i=1}^K p_i^{3/2}$. The left-hand side of (93) can be expressed as $\frac{g(q)}{g(p)}$. We consider minimizing $g(p)$ subject to the constraints for fixed q. As g is a convex function, from the first-order optimality condition, we have the following: Case (i): if $q_K \geq \max_{i\in[K-1]} q_i$, the minimum of $g(p)$ is attained by $p = q$, which means $\frac{g(q)}{g(p)} \leq 1$. Case (ii): if $q_K < \max_{i\in[K-1]} q_i$, there exists $c \in [q_K, \max_{i\in[K-1]} q_i]$ such that the minimum of $g(p)$ is attained when $p_K = \max\{q_K, c\}$ and $p_i = \min\{q_i, c\}$ for all $i \in [K-1]$. For such p, denote $W = \{i \in [K] : p_i = c\}$ and $W' = \{i \in [K-1] : p_i = c\} = W \setminus \{K\}$. As we have $p_i = q_i$ for $i \in [K] \setminus W$ and $\frac{g(q)}{g(p)} \geq 1$, we have $\frac{g(q)}{g(p)} = \frac{\sum_{i\in W} q_i^{3/2} + \sum_{i\in[K]\setminus W} p_i^{3/2}}{|W|c^{3/2} + \sum_{i\in[K]\setminus W} p_i^{3/2}} \leq \frac{\sum_{i\in W} q_i^{3/2}}{|W|c^{3/2}} =: h(q)$. Noting that $c = \sum_{i\in W} q_i/|W|$, $q_K \leq c$, and $q_i \geq c$ for $i \in W'$, we can show that $h(q)$ is maximized when $|W| = 2$ and $q_K = 0$, and then $h(q) = \sqrt{2}$. Consequently, we have $\frac{g(q)}{g(p)} \leq \sqrt{2}$, which means that (93) holds. ■