Meta-Learning Adversarial Bandits

Maria-Florina Balcan NINAMF@CS.CMU.EDU
Keegan Harris KEEGANH@CS.CMU.EDU
Mikhail Khodak KHODAK@CMU.EDU
Zhiwei Steven Wu ZSTEVENWU@CMU.EDU
Carnegie Mellon University School of Computer Science

Abstract

We study online learning with bandit feedback across multiple tasks, with the goal of improving average performance across tasks if they are similar according to some natural task-similarity measure. As the first to target the adversarial setting, we design a unified meta-algorithm that yields setting-specific guarantees for two important cases: multi-armed bandits (MAB) and bandit linear optimization (BLO). For MAB, the meta-algorithm tunes the initialization, step-size, and entropy parameter of the Tsallis-entropy generalization of the well-known Exp3 method, with the task-averaged regret provably improving if the entropy of the distribution over estimated optima-in-hindsight is small. For BLO, we learn the initialization, step-size, and boundary-offset of online mirror descent (OMD) with self-concordant barrier regularizers, showing that task-averaged regret varies directly with a measure induced by these functions on the interior of the action space. Our adaptive guarantees rely on proving that unregularized follow-the-leader combined with multiplicative weights is enough to online learn a non-smooth and non-convex sequence of affine functions of Bregman divergences that upper-bound the regret of OMD.

Keywords: Meta-Learning, Multi-Armed Bandits, Bandit Linear Optimization

1. Introduction

Meta-learning (Thrun and Pratt, 1998) is a popular approach to studying multi-task learning whose goal is to leverage information from previously-seen tasks in order to achieve better performance on unseen tasks. While most meta-learning algorithms are designed for tasks with full information feedback, there is a growing amount of work aiming to design meta-learning algorithms capable of operating under bandit feedback (c.f. Appendix A). While this literature has focused on *stochastic feedback*, where feedback is sampled i.i.d. from some distribution, we are the first to theoretically study meta-learning under *adversarial* bandit feedback, where it is chosen by an adversary possibly trying to harm the learner.

Specifically, we target low regret *on average* across a sequence of bandit tasks; this regret should be no worse than the single-task setting in general, and much better when tasks are related. We design a meta-algorithm based on initializing and tuning bandit methods based on online mirror descent (OMD), e.g. Exp3 (Auer et al., 2002). Our algorithm is applicable to both multi-armed bandits (MAB) and bandit linear optimization (BLO), and yields new meta-learning algorithms with provable guarantees for both. For MAB, its average m-round regret across T tasks is

$$o_T(1) + 2 \min_{\beta \in (0,1]} \sqrt{\hat{H}_\beta d^\beta m / \beta} \tag{1}$$

where d is the number of actions and $\hat{H}_\beta$ is the Tsallis entropy (Tsallis, 1988) of the empirical distribution over the estimated optimal actions across tasks. At $\beta = 1$ the Tsallis entropy reduces to the Shannon entropy; both are small if most tasks are estimated to be solved by the same few arms and large if all are used roughly the same amount, making it a natural task-similarity notion. The bound of $\log d \geq \hat{H}_1$ means that the bound (1) recovers Exp3's guarantee in the worst-case of dissimilar tasks. In the important case of $s \ll d$ arms always being estimated to be optimal we have $\hat{H}_\beta = \mathcal{O}(s)$, so using $\beta = \frac{1}{\log d}$ in bound (1) yields a task-averaged regret of $\mathcal{O}(\sqrt{sm \log d})$ as $T \to \infty$. For $s = \mathcal{O}_d(1)$ this beats the single-task lower bound of $\Omega(\sqrt{dm})$ (Audibert et al., 2011). We also obtain natural task-averaged regret bounds for BLO, albeit with different setting-specific notions of task similarity.

Our main technical contributions are as follows:

1. We design a unified meta-algorithm to initialize and tune OMD when using regularizers used by different bandit algorithms (Algorithm 1). Apart from strong guarantees and generality, our approach is notable for its adaptivity: we do not need to know anything about the task-similarity—e.g. the size of the subset of optimal arms—to adapt to similar tasks.
2. We apply our meta-approach to obtain a meta-learning algorithm for adversarial MAB. In particular, we use OMD with the Tsallis regularizer (Abernethy et al., 2015) as our within-task algorithm to achieve bounds on task-averaged regret that depend on a natural notion of task similarity: the Tsallis entropy of the estimated optima-in-hindsight.
3. We adapt Algorithm 1 to the adversarial BLO problem by setting the regularizer to be a self-concordant barrier function, as in Abernethy et al. (2008). As in MAB, we obtain task-averaged regret bounds which depend on a natural notion of task similarity based on the constraints defining the convex action space. We instantiate the BLO result in two settings: linear bandits over the sphere and an application to the bandit shortest-path problem (Takimoto and Warmuth, 2003; Kalai and Vempala, 2005).

2. Learning the regularizers of bandit algorithms

We consider the problem of meta-learning across bandit tasks $t = 1, \dots, T$ over some fixed set $\mathcal{K} \subset \mathbb{R}^d$. On each round $i = 1, \dots, m$ of task t we play action $\mathbf{x}_{t,i} \in \mathcal{K}$ and receive feedback $\ell_{t,i}(\mathbf{x}_{t,i})$ for some function $\ell_{t,i} : \mathcal{K} \mapsto [-1, 1]$. Note that all functions we consider will be linear and so we will also write $\ell_{t,i}(\mathbf{x}) = \langle \ell_{t,i}, \mathbf{x} \rangle$. Additionally, we allow each $\ell_{t,i}$ to be chosen by an *oblivious adversary*, i.e. an adversary with knowledge of the algorithm that must select $\ell_{t,i}$ independent of $\mathbf{x}_{t,i}$. We will also denote $\mathbf{x}(a)$ to be the ath element of the vector $\mathbf{x} \in \mathbb{R}^d$, $\overline{\mathcal{K}}$ to be the convex hull of $\mathcal{K}$, and $\triangle_n$ to be the simplex on n elements. Finally, note that all proofs can be found in the Appendix.

In online learning, the goal on a single task t is to play actions $\mathbf{x}_{t,1}, \dots \mathbf{x}_{t,m}$ that minimize the regret $\sum_{i=1}^m \ell_{t,i}(\mathbf{x}_{t,i}) - \ell_{t,i}(\mathbf{x}_t^*)$, where $\mathbf{x}_t^* \in \arg\min_{\mathbf{x} \in \mathcal{K}} \sum_{i=1}^m \ell_{t,i}(\mathbf{x})$. Lifting this to the meta-learning setting, our goal as in past work (Khodak et al., 2019; Balcan et al., 2021) will be to minimize the **task-averaged regret** $\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^m \ell_{t,i}(\mathbf{x}_{i,t}) - \ell_{t,i}(\mathbf{x}_t^*)$. In-particular, we hope to use multi-task data in order to improve average performance as the number of tasks $T \to \infty$, e.g. by attaining a task-averaged regret of $o_T(1) + \tilde{\mathcal{O}}(V\sqrt{m})$, where $V \in \mathbb{R}_{\geq 0}$ is a measure of task-similarity that is small if the tasks are similar but still yields the worst-case single-task performance if they are not.

In meta-learning we are commonly interested in learning a within-task algorithm or **base-learner**, a parameterized method that we run on each task t. A popular approach (Finn et al., 2017; Nichol et al., 2018) is to learn the initialization of a gradient-based method such as stochastic gradient descent. The hope is that optimal parameters for each task are close to each other and thus a meta-learned initialization will result in a strong model after only a few steps. In this paper we take a similar approach applied to online mirror descent, which given a strictly convex **regularizer** $\phi : \overline{\mathcal{K}} \mapsto \mathbb{R}$ and step-size $\eta > 0$ performs the update

$$\mathbf{x}_{t,i+1} = \underset{\mathbf{x} \in \overline{\mathcal{K}}}{\arg\min}\, B_\phi(\mathbf{x}||\mathbf{x}_{t,1}) + \eta \sum_{j<i} \langle \nabla \ell_{t,j}(\mathbf{x}_{t,j}), \mathbf{x} \rangle \tag{2}$$

where $B_\phi(\mathbf{x}||\mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla\phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ is the **Bregman divergence** of ϕ. OMD recovers online gradient descent when $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$; another example is **exponentiated gradient**, for which ϕ is the negative Shannon entropy on probability vectors and B_ϕ is the KL-divergence (Shalev-Shwartz, 2011). Mirror descent using **loss estimators** $\hat{\ell}_{t,i}$ constructed using bandit feedback $\ell_{t,i}(\mathbf{x}_{t,i})$ forms an important class of methods for bandit settings (Abernethy et al., 2008; Neu, 2015; Abernethy et al., 2015), including Exp3 (Auer et al., 2002).

Following the average regret-upper-bound analysis (ARUBA) framework of Khodak et al. (2019), in this paper we learn to initialize and tune OMD by online learning a sequence of losses $U_t(\mathbf{x}, \theta)$, each of which is a hyperparameter θ-dependent affine function of a Bregman divergence from an initialization $\mathbf{x} \in \overline{\mathcal{K}}$ to some known fixed point in $\overline{\mathcal{K}}$. We are interested in learning such functions because the regret after m rounds of OMD initialized at $\mathbf{x}$ with step-size η is usually upper-bounded by $\frac{1}{\eta}B_\phi(\mathbf{x}_t^*||\mathbf{x}) + \mathcal{O}(\eta m)$ for $\mathbf{x}_t^*$, the optimum-in-hindsight on task t (Shalev-Shwartz, 2011). Unlike past work, we use a parameter $\varepsilon > 0$ to constrain this optimum to lie in a convex subset $\mathcal{K}_\varepsilon \subset \overline{\mathcal{K}}$ whose boundary is ε-away from that of $\overline{\mathcal{K}}$ and which satisfies $\mathcal{K}_\varepsilon \subset \mathcal{K}_{\varepsilon'}$ whenever $\varepsilon \leq \varepsilon'$; for example, we use $\mathcal{K}_\varepsilon = \{\mathbf{x} \in \triangle_d : \min_a \mathbf{x}(a) \geq \varepsilon/d\}$ for the simplex. Thus, unlike with full-information, the feedback we receive from the within-task algorithm will be the minimizer $\mathrm{OPT}_\varepsilon(\hat{\ell}_t) = \arg\min_{\mathbf{x} \in \mathcal{K}_\varepsilon} \langle \hat{\ell}_t, \mathbf{x} \rangle$ of the estimated loss $\hat{\ell}_t = \sum_{i=1}^m \hat{\ell}_{t,i}$ over the ε-constrained subset. This allows us to handle regularizers that diverge near the boundary. Thus in full generality and for constants $G_\beta \geq 1, C \geq 0$ the upper bounds of interest are the following functions of the initialization $\mathbf{x}$ and three parameters: the step-size $\eta > 0$, a parameter β of the regularizer ϕ_β, and the boundary offset $\varepsilon > 0$.

$$U_t(\mathbf{x}, (\eta, \beta, \varepsilon)) = \frac{B_{\phi_\beta}(\mathrm{OPT}_\varepsilon(\hat{\ell}_t)||\mathbf{x})}{\eta} + (\eta G_\beta^2 + C\varepsilon)m \tag{3}$$

The reason to optimize this sequence of upper bounds is because the resulting average regret directly bounds the task-averaged regret. Furthermore, an affine sum over Bregman divergences is minimized at the average optimum in hindsight, which leads to natural and problem specific task-similarity measures V (Khodak et al., 2019); specifically, V is the square root of the average divergence between optima in hindsight and their mean, which is small if the tasks are optimized by similar parameters. At a high-level, our meta-algorithm for online learning these upper bounds learns the initialization by taking the mean of $\mathcal{K}_\varepsilon$-constrained estimated optima-in-hindsight—i.e. follow-the-leader over the Bregman divergences in (3)—while simultaneously tuning OMD via multiplicative weights over a discrete grid Θ over $\theta = (\eta, \beta, \varepsilon)$. We provide a more detailed description, pseudo-code, and a structural result showing that such an algorithm can learn the sequence of upper bounds U_t in the Appendix.

3. Multi-armed bandits

We now turn to our first application: the multi-armed bandits problem. Here at each round i of task t we take action $a_{t,i} \in [d]$ and observe loss $\ell_{t,i}(a_{t,i}) \in [0,1]$. We use as a base-learner a generalization of Exp3 (Auer et al., 2002), which runs multiplicative weights over unbiased loss estimators. The first generalization is of the regularizer, where we use the negative Tsallis entropy $\phi_\beta(\mathbf{p}) = \frac{1-\sum_{a=1}^d \mathbf{p}^\beta(a)}{1-\beta}$ for $\beta \in [0,1]$, which obtains a better dependence on dimension (Abernethy et al., 2015). Note that ϕ_β recovers the Shannon entropy in the limit $\beta \to 1$, and also that $B_{\phi_\beta}(\mathbf{x}||\cdot)$ is non-convex in the second argument, making ours the first known application of the online learnability of non-convex Bregman divergences. The second generalization is in the loss estimators; for $\gamma > 0$ we employ $\hat{\ell}_{t,i}(a) = \frac{\ell_{t,i}(a)1_{a_{t,i}=a}}{\mathbf{x}_{t,i}(a)+\gamma}$, where $\mathbf{x}_{t,i}(a)$ is the probability of sampling a on round i of task t, which is critical for high-probability bounds (Neu, 2015).

As ϕ_β is non-smooth at the boundary, learning Tsallis divergences requires the tools developed previously for initializing `OMD` in the interior of $\overline{\mathcal{K}}$. We set $\mathcal{K}_\varepsilon = \{\mathbf{x} \in \triangle_d : \min_a \mathbf{x}(a) \geq \varepsilon/d\}$, so that the offset optimum $\hat{\mathbf{x}}_t^{(\theta)}$ then has the very simple form $\mathrm{OPT}_\varepsilon(\hat{\ell}_t) = (1-\varepsilon)\hat{\mathbf{x}}_t + \varepsilon\mathbf{1}_d/d$, i.e. it is the mixture of the estimated optimum $\hat{\mathbf{x}}_t$ with the uniform distribution. Note that for MAB we will *not* need to learn ε and can just set it assuming knowledge of the number of tasks. Thus the method can be summarized as doing the following at each task $t > 1$:

1. sample $\theta_t = (\eta_t, \beta_t)$ from a distribution $\mathbf{p}_t$ over the discretization Θ
2. run $\texttt{OMD}_{\beta_t,\eta_t}$ using the initialization $\mathbf{x}_{t,1} = \frac{1}{t-1}\sum_{s<t} \hat{\mathbf{x}}_t^{(\theta_t)} = \frac{\varepsilon}{d}\mathbf{1}_d + \frac{1-\varepsilon}{t-1}\sum_{s<t} \hat{\mathbf{x}}_t$
3. update $\mathbf{p}_{t+1}$ using multiplicative weights over losses $\frac{B_{\phi_{\beta_t}}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x}_{t,1})+\rho^2D^2}{\eta_t} + \frac{\eta_t d_t^\beta m}{\beta_t}$

The latter regret-upper-bound is derived from the regret of OMD with the Tsallis regularizer, which is then offset by a factor $\rho^2 D^2 \geq 0$ to handle non-Lipschitzness near $\eta = 0$, as described in the Appendix. This procedure achieves the following guarantees on the task-averaged regret:

Theorem 3.1. *Suppose $\texttt{OMD}_{\eta,\beta}$ is online mirror descent with the Tsallis entropy regularizer ϕ_β over γ-offset loss estimators. For each of the following regimes of $\underline{\beta}$ we can specify $\varepsilon > 0$, $\rho^2 D^2 \geq 0$, integer $k = \tilde{\mathcal{O}}(\lceil d^4\sqrt{mT}\log\frac{1}{\varepsilon}\rceil)$, and $\alpha, \underline{\eta}, \overline{\eta} \in (0,\infty)$ such that running the above procedure (Algorithm 1) with Θ the product of uniform grids of size k over each non-singleton dimension of $[\underline{\eta}, \overline{\eta}] \times [\max\{\underline{\beta}, 1/\log d\}, 1] \times \{\varepsilon\}$ and α the meta-step-size of multiplicative weights yields w.p. at least $1-\delta$ the listed task-averaged regret.*

$$\underline{\beta} = 1 \qquad \tilde{\mathcal{O}}\left(\frac{d^{\frac{3}{2}} + \sqrt{m}}{\sqrt[4]{T}}\sqrt{m\log\frac{4}{\delta}}\right) + 2\sqrt{H_1(\hat{\bar{\mathbf{x}}})dm} \tag{4}$$

$$\underline{\beta} = \frac{1}{2} \qquad \tilde{\mathcal{O}}\left(\frac{dm\sqrt{d\log\frac{4}{\delta}}}{\sqrt[4]{T}}\right) + 2\sqrt{d} + 2\min_{\beta\in\left[\frac{1}{2},\frac{\log d-1}{\log d}\right]}\sqrt{H_\beta(\hat{\bar{\mathbf{x}}})d^\beta m/\beta} \tag{5}$$

$$\underline{\beta} = \frac{1}{\log d} \qquad \tilde{\mathcal{O}}\left(\frac{d^{\frac{3}{2}} + \sqrt{m}}{\sqrt[6]{T}}\sqrt{m\log\frac{4}{\delta}}\right) + 2\min_{\beta\in(0,1]}\sqrt{H_\beta(\hat{\bar{\mathbf{x}}})d^\beta m/\beta} + \sqrt{\frac{d1_{\beta<1}}{\beta(1-\beta)mT^{\frac{\beta}{3}}}} \tag{6}$$

Here $H_\beta = -\phi_\beta$ is the Tsallis entropy and $\hat{\bar{\mathbf{x}}}$ is the mean of the estimated optima $\hat{\mathbf{x}}_t$. Note that below $\underline{\beta} = \frac{1}{\log d}$ the upper bound always worsens and so it does not make sense to try $\beta < \frac{1}{\log d}$.

These results show that for all three settings of $\underline{\beta}$, as the meta-learner sees more tasks the average regret depends directly on the entropy of the estimated optima-in-hindsight, a natural notion of task-similarity since it is small if most tasks are estimated to be solved by the same arms and large if all arms are used roughly the same amount. It also demonstrates how our algorithm's automatic tuning of the step-size η allows us to set the asymptotic rate optimally depending on the entropy. The algorithm's tuning of the entropy itself via β also enables adaptation to similar tasks; specifically, a smaller β weights the $H_\beta(\hat{\hat{\mathbf{x}}})/\eta$ term higher and is thus beneficial if tasks are similar. As a natural example, suppose a constant $s \ll d$ actions are always minimizers, i.e. $\hat{\hat{\mathbf{x}}}$ is s-sparse. Then the last bound (6) implies that Algorithm 1 can achieve task-averaged regret $o_T(1) + \mathcal{O}(\sqrt{sm \log d})$, albeit at the cost of slow convergence. In-general, for the case of tuning over all $\beta \geq 1/\log d$ the speed of the convergence depends on the optimal β; the algorithm will converge very slowly at rate $\tilde{\mathcal{O}}(1/\sqrt[6\log d]{T})$ if the optimal β is around $1/\log d$, but for β near 1 the rate will be $\tilde{\mathcal{O}}(1/\sqrt[4]{T})$. Note that we show in the intermediate case of tuning only as low as $\beta = 1/2$ that we can still achieve $\tilde{\mathcal{O}}(1/\sqrt[4]{T})$ at the cost of a fast $2\sqrt{d}$ term per-task. Finally, note that because the entropy is bounded by $d^{1-\beta}$ we do asymptotically recover worst-case guarantees in all three cases if the tasks are dissimilar.

To put these results in context, we can compare them to Azizi et al. (2022), who achieve task-averaged regret bounds of the form $\tilde{\mathcal{O}}(1/\sqrt{T} + \sqrt{sm})$ in the *stochastic* MAB setting, where s is an unknown subset of optimal actions. Unlike their result, we study the harder adversarial setting and do *not* place restrictions on how the tasks are related; despite this greater generality, our bounds are asymptotically comparable if the estimated and true optima-in-hindsight are roughly equivalent, as we also have $\tilde{\mathcal{O}}(\sqrt{sm})$ average regret as $T \to \infty$. On the other hand, the rate in the number of tasks of Azizi et al. (2022) is much better, albeit at a cost of runtime exponential in s. Apart from generality, we believe a great strength of our result is its adaptiveness; unlike this work, we do not need to know how many optimal arms there are or their entropy in order to improve task-averaged regret with task-similarity.

4. Bandit linear optimization

Our second general application is to bandit linear optimization, in which at each round i of task t we play a vector $\mathbf{x}_{t,i} \in \mathcal{K}$ for some convex set $\mathcal{K}$ and observe loss $\langle \ell_{t,i}, \mathbf{x}_{t,i} \rangle \in [-1, 1]$. We will again use a variant of mirror descent on top of estimated losses, this time setting ϕ to be a self-concordant barrier function with specialized loss estimators as in Abernethy et al. (2008). This is generally applicable to any convex domain $\mathcal{K}$ via the construction of such barriers and its regret has optimal dependence on the number of rounds m. Note that our ability to handle non-smooth regularizers is even more important here, as the barrier functions are infinite at the boundaries. Indeed, in this section we will no longer learn a β parameterizing the regularizer and instead focus on learning an offset $\varepsilon > 0$ away from the boundary. For each such offset define $\mathcal{K}_\varepsilon = \{\mathbf{x} \in \mathbb{R}^d : \pi_{\mathbf{x}_{1,1}}(\mathbf{x}) \leq 1/(1+\varepsilon)\} \subset \mathcal{K}$, where $\mathbf{x}_{1,1} = \arg\min_{\mathbf{x} \in \mathcal{K}} \phi(\mathbf{x})$ and $\pi_{\mathbf{x}_{1,1}}(\mathbf{x}) = \inf_{\lambda \geq 0, \mathbf{x}_{1,1} + (\mathbf{x} - \mathbf{x}_{1,1})/\lambda \in \mathcal{K}} \lambda$ is the Minkowski function. As before we obtain the ε-restricted optima-in-hindsight via the primitive $\text{OPT}_\varepsilon(\hat{\ell}_t) = \arg\min_{\mathbf{x} \in \mathcal{K}_\varepsilon} \langle \hat{\ell}_t, \mathbf{x} \rangle$. With this specified, we can again adapt our meta-approach, roughly summarized for BLO as doing the following at each task $t > 1$:

1. sample $\theta_t = (\eta_t, \varepsilon_t)$ from a distribution $\mathbf{p}_t$ over the discretization Θ
2. run $\texttt{OMD}_{\eta_t}$ using the initialization $\mathbf{x}_{t,1} = \frac{1}{t-1}\sum_{s<t}\hat{\mathbf{x}}_t^{(\theta_t)} = \frac{1}{t-1}\sum_{s<t}\mathrm{OPT}_{\varepsilon_t}(\hat{\ell}_t)$
3. update $\mathbf{p}_{t+1}$ using multiplicative weights with losses $\frac{B_\phi(\hat{\mathbf{x}}_t^{(\varepsilon_t)}||\mathbf{x}_{t,1})+\rho^2D^2}{\eta_t} + (32d^2\eta_t + \varepsilon_t)m$

Note that this algorithm is very similar to that for MAB, with both being special cases of the meta-algorithm for optimizing upper bounds (3), with the main difference being the different upper bound passed to multiplicative weights. The procedure has the following guarantee:

Theorem 4.1. *Suppose $\texttt{OMD}_{\eta,\beta}$ is online mirror descent with a self-concordant barrier ϕ as a regularizer and loss estimators specified as in Abernethy et al. (2008). Then for every $\overline{\varepsilon} \in [1/m, 1/\sqrt{m}]$ and $\underline{\varepsilon} \in [1/m, \overline{\varepsilon}]$ there exists an integer $k = \mathcal{O}(dm\lceil\sqrt{mT}\rceil)$ and $\alpha, \underline{\eta}, \overline{\eta} \in (0, \infty)$ such that the above procedure with Θ the product of uniform grids of size k over each dimension of $[\underline{\eta}, \overline{\eta}] \times [\underline{\epsilon}, \overline{\epsilon}]$ and α the meta-step-size yields the expected task-averaged regret*

$$\mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{m}\langle \ell_{t,i}, \mathbf{x}_{t,i} - \mathbf{x}_t^*\rangle \leq \tilde{\mathcal{O}}\left(\frac{dm^2}{\sqrt[4]{T}}\right) + \min_{\frac{1}{m}\leq\varepsilon\leq\frac{1}{\sqrt{m}}} 4d\hat{V}_\varepsilon\sqrt{2m} + \varepsilon m \tag{7}$$

where we call $\hat{V}_\varepsilon^2 = \min_{\mathbf{x}\in\mathcal{K}}\mathbb{E}\frac{1}{T}\sum_{t=1}^{T} B_\phi(\mathrm{OPT}_\varepsilon(\hat{\ell}_t)||\mathbf{x})$ the **barrier-divergence at level** *ε.*

As before, this shows that as the number of tasks $T \to \infty$ the average regret improves with a notion of task-similarity $\hat{V}_\varepsilon$ that decreases if the estimated task-optima are close together. Roughly speaking, if tasks have barrier-divergence $\hat{V}_\varepsilon$ then the average regret will be $\mathcal{O}(\hat{V}_\varepsilon\sqrt{m} + \varepsilon m)$, which can be a significant improvement over the single-task case, e.g. if $\hat{V}_{\frac{1}{m}}$ is small. In-particular, our analysis removes explicit dependence on the square root of the self-concordance constant of ϕ in the single-task case (Abernethy et al., 2008); as an example, this constant is equal to the number of constraints if $\mathcal{K}$ is defined by linear inequalities, as in the bandit shortest-path application below. Note that the use of ε-constrained optima is necessary for this problem due to the regularizers being infinite at the boundaries, where all true optima lie.

To make the above result and task-similarity notion more concrete, consider the following corollary for BLO over the unit sphere $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$:

Corollary 4.1. *Let $\mathcal{K}$ be the unit sphere with the self-concordant barrier $\phi(\mathbf{x}) = -\log(1 - \|\mathbf{x}\|_2^2)$. Then the above procedure attains expected task-averaged regret bounded by*

$$\tilde{\mathcal{O}}\left(\frac{dm^2}{\sqrt[4]{T}}\right) + \min_{\frac{1}{m}\leq\varepsilon\leq\frac{1}{\sqrt{m}}} 4d\mathbb{E}\sqrt{2m\log\left(\frac{1 - \|\hat{\bar{\ell}}^{(\varepsilon)}\|_2^2}{2\varepsilon - \varepsilon^2}\right)} + \varepsilon m \tag{8}$$

for $\hat{\bar{\ell}}^{(\varepsilon)} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{OPT}_\varepsilon(\hat{\ell}_t) = \frac{\varepsilon-1}{T}\sum_{t=1}^{T}\frac{\hat{\ell}_t}{\|\hat{\ell}_t\|_2}$ the average over normalized estimated optima.

Thus in this setting if all tasks have similar estimated losses then $\hat{\bar{\ell}}^{(\varepsilon)}$ will be an average over similar vectors and thus have large Euclidean norm close to $1 - \varepsilon$, making the term in the logarithm above close to 1. In this case $\hat{V}_\varepsilon$ is close to zero and so the average regret is εm as $T \to \infty$; setting $\varepsilon = 1/m$ yields constant asymptotic averaged regret. This demonstrates the usefulness of the barrier-divergence as a measure of task-similarity. We show another novel notion that it yields for sets defined by linear constraints in the Appendix, where we apply our meta-BLO result to the shortest-path problem in online optimization (Takimoto and Warmuth, 2003; Kalai and Vempala, 2005).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the International Conference on Computational Learning Theory*, 2008.

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

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the International Conference on Computational Learning Theory*, 2011.

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

Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, 2004.

MohammadJavad Azizi, Thang Duong, Yasin Abbasi-Yadkori, András György, Claire Vernade, and Mohammad Ghavamzadeh. Non-stationary bandits and meta-learning with a small set of optimal arms. *arXiv preprint arXiv:2202.13001*, 2022.

Maria-Florina Balcan, Mikhail Khodak, Dravyansh Sharma, and Ameet Talwalkar. Learning-to-learn non-convex piecewise-Lipschitz functions. In *Advances in Neural Information Processing Systems*, 2021.

Soumya Basu, Branislav Kveton, Manzil Zaheer, and Csaba Szepesvári. No regrets for learning the prior in bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, pages 1360–1370. PMLR, 2020.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

Varsha Dani, Thomas Hayes, and Sham Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, 2008.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Online-within-online meta-learning. In *Advances in Neural Information Processing Systems*, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Elad Hazan. Introduction to online convex optimization. In *Foundations and Trends in Optimization*, volume 2, pages 157–325. now Publishers Inc., 2015.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, 2019.

Mikhail Khodak, Renbo Tu, Tian Li, Liam Li, Maria-Florina Balcan, Virginia Smith, and Ameet Talwalkar. Federated hyperparameter tuning: Challenges, baselines, and connections to weight-sharing. In *Advances in Neural Information Processing Systems*, 2021.

Mikhail Khodak, Maria-Florina Balcan, Ameet Talwalkar, and Sergei Vassilvitskii. Learning predictions for algorithms with predictions. arXiv, 2022.

Branislav Kveton, Martin Mladenov, Chih-Wei Hsu, Manzil Zaheer, Csaba Szepesvari, and Craig Boutilier. Meta-learning bandit policies by gradient ascent. *arXiv preprint arXiv:2006.05094*, 2020.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.

Alessandro Lazaric, Emma Brunskill, et al. Sequential transfer in multi-armed bandit with finite set of models. *Advances in Neural Information Processing Systems*, 26, 2013.

Haipeng Luo. Lecture 13. 2017. URL `https://haipeng-luo.net/courses/CSCI699/lecture13.pdf`.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, Cambridge, UK, 2021.

Ahmadreza Moradipari, Mohammad Ghavamzadeh, Taha Rajabzadeh, Christos Thrampoulidis, and Mahnoosh Alizadeh. Multi-environment meta-learning in stochastic linear bandits. *arXiv preprint arXiv:2205.06326*, 2022.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. arXiv, 2018.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Amr Sharaf and III Hal Daumé. Meta-learning effective exploration strategies for contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9541–9548, 2021.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel J Hsu, Thodoris Lykouris, Miro Dudik, and Robert E Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Eiji Takimoto and Manfred K. Warmuth. Path kernels and multiplicative updates. *Journal of Machine Learning Research*, 4:773–818, 2003.

Sebastian Thrun and Lorien Pratt. *Learning to Learn*. Springer Science & Business Media, 1998.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

Takuya Yamano. Some properties of q-logarithm and q-exponential functions in Tsallis statistics. *Physica A: Statistical Mechanics and its Applications*, 305:486–496, 2002.

Appendix A. Related work

While we are the first to consider meta-learning under adversarial bandit feedback, many have studied meta-learning in various *stochastic* bandit settings (Sharaf and Hal Daumé, 2021; Simchowitz et al., 2021; Kveton et al., 2020; Cella et al., 2020; Kveton et al., 2021; Basu et al., 2021; Azizi et al., 2022; Lazaric et al., 2013). Kveton et al. (2021), Basu et al. (2021), and Simchowitz et al. (2021) study meta-learning algorithms for the Bayesian bandit setting. Kveton et al. (2020) and Sharaf and Hal Daumé (2021) consider meta-learning for contextual bandits, although they allow their algorithms to have offline access to a set of training tasks for which full feedback is available. Cella et al. (2020) and Moradipari et al. (2022) provide algorithms based on OFUL (Abbasi-Yadkori et al., 2011) for meta-learning in stochastic linear bandits under various assumptions on how the bandit learning tasks are generated. Azizi et al. (2022) study a setting in which a meta-learner faces a sequence of stochastic multi-armed bandit tasks. While the sequence of tasks may be adversarially designed, the adversary is constrained to choose the optimal arm for each task from a smaller but unknown subset of arms. In contrast to Cella et al. (2020); Moradipari et al. (2022); Azizi et al. (2022), we make no assumptions about how the sequence of tasks is generated and our guarantees adapt to a natural measure of similarity between tasks.

Theoretically our analysis draws on the average regret-upper-bound analysis (ARUBA) framework of Khodak et al. (2019), which was designed for meta-learning under full information. While the general approach is not restricted by convexity (Balcan et al., 2021) and has been combined with bandit algorithms on the meta-level (Khodak et al., 2021), the existing results cannot be applied to OMD methods for within-task learning under bandit feedback because the associated regularizers are non-Lipschitz or sometimes even unbounded near the boundaries of the action space. We thus require a specialized analysis for the bandit setting. Denevi et al. (2019) also study an OMD-based algorithm for meta-learning in the online setting, but their results are also only applicable in the full information setting.

Appendix B. Description and guarantee for the meta-algorithm

The algorithm assumes two primitives: (1) the base-learner $\mathtt{OMD}_{\eta,\beta}$ that outputs an estimated cumulative loss $\hat{\ell}_t \in \mathbb{R}^d$ after running online mirror descent over the m losses $\ell_{t,1}, \dots, \ell_{t,m}$ of task t, and (2) an optimizer OPT_ε that, given a vector $\mathbf{c} \in \mathbb{R}^d$, finds the minimum of $\langle \mathbf{c}, \cdot \rangle$ over $\mathcal{K}_\varepsilon$. Algorithm 1 maintains a categorical distribution $\mathbf{p}_t$ over a finite set $\Theta \subset \mathbb{R}^3$ containing triples $\theta = (\eta, \beta, \varepsilon)$, each with its own associated initialization $\mathbf{x}_t^{(\theta)}$; at each task t it samples $\theta_t = (\eta_t, \beta_t, \varepsilon_t)$ from Θ using $\mathbf{p}_t$ and runs $\mathtt{OMD}_{\eta_t,\beta_t}$ from initialization $\mathbf{x}_t^{(\theta_t)}$, obtaining a loss estimate $\hat{\ell}_t$. Then for each $\theta = (\eta, \beta, \varepsilon)$ in Θ the method updates the corresponding initialization $\mathbf{x}_t^{(\theta)}$ by taking the average of the ε-constrained optima-in-hindsight $\mathrm{OPT}_\varepsilon(\hat{\ell}_1), \dots, \mathrm{OPT}_\varepsilon(\hat{\ell}_t)$ seen so far. Finally, the algorithm updates the distribution $\mathbf{p}_t$ using multiplicative weights over the following modification of the regret-upper-bound (3) above for some $\rho > 0$:

$$U_t^{(\rho)}(\mathbf{x}, \theta) = \frac{B_{\phi_\beta}(\hat{\mathbf{x}}_t^{(\theta)}||\mathbf{x}) + \rho^2 D^2}{\eta} + (\eta G_\beta^2 + C\varepsilon)m \tag{9}$$

Note that given $\rho > 0$ this function is fully defined after running $\mathtt{OMD}_{\eta_t,\beta_t}$ on task t to obtain loss estimates $\hat{\ell}_t$ and then computing the ε-constrained optimum-in-hindsight $\hat{\mathbf{x}}_t^{(\theta)} = \mathrm{OPT}_\varepsilon(\hat{\ell}_t)$ for each $\theta = (\eta, \beta, \varepsilon)$. This allows us to use full-information multiplicative weights for θ. $\rho > 0$ is necessary for learning η, as if its optimum is near zero then U_t will not be Lipschitz near the optimum. Theorem B.1 shows a sublinear regret guarantee for Algorithm 1 over the unmodified regret-upper-bounds (9) w.r.t. all elements in $\overline{\mathcal{K}}$ and in a continous set of hyperparameters $\Theta^* \subset \mathbb{R}^3$.

Theorem B.1. *Let $\Theta^* = (0, \infty) \times [\underline{\beta}, \overline{\beta}] \times [\underline{\varepsilon}, \overline{\varepsilon}]$ for $0 \le \underline{\beta} \le \overline{\beta} \le 1$ and $0 \le \underline{\varepsilon} \le \overline{\varepsilon} \le 1$ be the set of hyperparameters $(\eta, \beta, \varepsilon)$ of interest. Then there exists integer $k = \mathcal{O}(\lceil \sqrt{mT} \rceil)$ and $\alpha, \underline{\eta}, \overline{\eta} \in (0, \infty)$ such that running Algorithm 1 with Θ the product of uniform grids of size k over each non-singleton dimension of $[\underline{\eta}, \overline{\eta}] \times [\underline{\beta}, \overline{\beta}] \times [\underline{\varepsilon}, \overline{\varepsilon}]$ and α the meta-step-size yields regret*

$$\begin{aligned}
&\mathbb{E} \sum_{t=1}^T U_t(\mathbf{x}_t^{(\theta_t)}, \theta_t) - \min_{\mathbf{x} \in \overline{\mathcal{K}}, \theta \in \Theta^*} \sum_{t=1}^T U_t(\mathbf{x}, \theta) \\
&\le \left(C\sqrt{m} + 2DG\left(\frac{1}{\rho} + M\right)\right) \sqrt{6mT \log k} + \frac{8SK^2G\sqrt{m}}{\rho D}(1 + \log T) + \rho DGT\sqrt{m}
\end{aligned} \tag{10}$$

for $G = \max_\beta G_\beta \ge 1$, $M = \frac{G}{\min_\beta G_\beta}$, $D^2 = \max_{\beta,\varepsilon,\mathbf{x},\mathbf{y} \in \mathcal{K}_\varepsilon} B_{\phi_\beta}(\mathbf{x}||\mathbf{y}) \ge 1$, L the maximum Lipschitz constant of $\phi_\beta(\mathrm{OPT}_\varepsilon(\ell))$ w.r.t. (β, ε) over $\ell \in \mathbb{R}^d$, $S = \max_{\beta,\varepsilon,\mathbf{x} \in \mathcal{K}_\varepsilon} \|\nabla^2 \phi_\beta(\mathbf{x})\|_2$, $K = \max_{\mathbf{x},\mathbf{y} \in \mathcal{K}} \|\mathbf{x} - \mathbf{y}\|_2$, and the expectation is over sampling $\theta_t \sim \mathbf{p}_t$. The result without the expectation holds w.p. $1 - \delta$ at the cost of an additional $\left(C\sqrt{m} + 2DG\left(\frac{1}{\rho} + M\right)\right)\sqrt{\frac{T}{2} \log \frac{1}{\delta}}$ term.

Algorithm 1: Algorithm for tuning an online mirror descent (OMD) base-learner $\texttt{OMD}_{\eta,\beta}$ with parameterized regularizer $\phi_\beta : \overline{\mathcal{K}} \mapsto \mathbb{R}$ and step-size $\eta > 0$ that runs OMD on loss estimators $\hat{\ell}_{t,1}, \dots, \hat{\ell}_{t,m}$ from an initialization $\mathbf{x} \in \overline{\mathcal{K}}$ and returns estimated loss $\hat{\ell}_t = \sum_{i=1}^m \hat{\ell}_{t,i} \in \mathbb{R}^d$. Then for every $\varepsilon > 0$ the constrained optimizer $\mathrm{OPT}_\varepsilon(\hat{\ell}) = \arg\min_{\mathbf{x}\in\mathcal{K}_\varepsilon} \langle \hat{\ell}, \mathbf{x} \rangle$ returns the minimizer of the estimated loss over the constrained subset $\mathcal{K}_\varepsilon \subset \overline{\mathcal{K}}$ (set $\mathrm{OPT}_\varepsilon(\mathbf{0}_d) = \arg\min_{\mathbf{x}\in\overline{\mathcal{K}}} \phi(\mathbf{x})$).

```
Input: compact $\overline{\mathcal{K}} \subset \mathbb{R}^d$, meta-hyperparameters $\alpha, \rho > 0$, finite $\Theta \subset \mathbb{R}^3$ over $(\eta, \beta, \varepsilon)$,
       base-learner $\texttt{OMD}_{\eta,\beta} : \overline{\mathcal{K}} \mapsto \mathbb{R}^d$, constrained linear minimizer $\mathrm{OPT}_\varepsilon : \mathbb{R}^d \mapsto \mathcal{K}_\varepsilon$
for $\theta = (\eta, \beta, \varepsilon) \in \Theta$ do
    $\mathbf{x}_1^{(\theta)} \leftarrow \arg\min_{\mathbf{x}\in\overline{\mathcal{K}}} \phi(\mathbf{x})$      // maintain an initialization for each $\theta \in \Theta$
$\mathbf{p}_1 \leftarrow \mathbf{1}_{|\Theta|}/|\Theta|$                    // multiplicative weights (MW) initialization
for task $t = 1, \dots, T$ do
    sample $\theta_t = (\eta_t, \beta_t, \varepsilon_t) \sim \mathbf{p}_t$ from $\Theta$
    $\hat{\ell}_t \leftarrow \texttt{OMD}_{\eta_t,\beta_t}(\mathbf{x}^{(\theta_t)})$                        // run bandit OMD within-task
    for $\theta = (\eta, \beta, \varepsilon) \in \Theta$ do
        $\mathbf{x}_{t+1}^{(\theta)} \leftarrow \frac{1}{t} \sum_{s=1}^t \mathrm{OPT}_\varepsilon(\hat{\ell}_s)$                 // update all initializations
        $\mathbf{p}_{t+1}(\theta) \leftarrow \mathbf{p}_{t+1}(\theta) \exp\left(-\alpha U_t^{(\rho)}(\mathbf{x}_t^{(\theta)}, \theta)\right)$  // MW update using loss in (9)
    $\mathbf{p}_{t+1} \leftarrow \mathbf{p}_{t+1}/\|\mathbf{p}_{t+1}\|_1$
```

Note that we keep details of the dependence on values like Lipschitz constants because they are important in applying this result; however, in general setting $\rho = 1/\sqrt[4]{T}$ in (10) yields $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$-regret. While a slow rate, note that Algorithm 1 is learning a sequence of affine functions of Bregman divergences that are non-smooth and non-convex in-general. Theorem B.1 is an important structural result; our main contributions to multi-armed and linear bandits follow by applying its instantiations for specific regularizers ϕ and hyperparameter sets Θ^*. We also believe Theorem B.1 may be of independent interest as it holds for any choice of Bregman divergence beyond those we consider, and unlike past work (Khodak et al., 2019) allows for explicit control of non-smooth regularizers near the boundaries. The theorem allows tuning the hyperparameters over user-specified intervals for β and ε and over an infinite interval for the step-size $\eta > 0$. Note that a similar result is straightforward to show for β outside $[0, 1]$ or for discrete rather than continuous set of hyperparameters.

Appendix C. Learning to parameterize a bandit-shortest-path algorithm

As a final application, we apply our meta-BLO result to the shortest-path problem in online optimization (Takimoto and Warmuth, 2003; Kalai and Vempala, 2005). In its bandit variant (Awerbuch and Kleinberg, 2004; Dani et al., 2008), at each time step $i = 1, \dots, m$ the player must choose a path p_i from a fixed source $u \in V$ to a fixed sink $v \in V$ in a directed graph $G(V, E)$. At the same time the adversary chooses edge weights $\ell_i \in \mathbb{R}^{|E|}$ and the player suffers the sum $\sum_{e \in p_t} \ell_i(e)$ of the weights in their chosen path p_t. This can be transformed into BLO over vectors $\mathbf{x}$ in a convex set $\mathcal{K} \subset [0,1]^{|E|}$ defined by a set $\mathcal{C}$ of $\mathcal{O}(|E|)$ linear constraints $(\mathbf{a}, b)$ s.t. $\langle \mathbf{a}, \mathbf{x} \rangle \leq b$ enforcing flows from u to v; paths from u to v can then be sampled from any $\mathbf{x} \in \mathcal{K}$ in an unbiased manner (Abernethy et al., 2008, Proposition 1). In the single-task case the BLO method of Abernethy et al. (2008) yields an $\mathcal{O}(|E|^{\frac{3}{2}}\sqrt{m})$-regret algorithm for this problem.

In the multi-task case consider a sequence of $t = 1, \dots, T$ shortest path instances, each consisting of m edge loss vectors $\ell_{t,i}$ selected by an adversary. The goal is to minimize average regret across instances. Note that our setup may be viewed as learning a prediction of the optimal path in a manner similar to the algorithms with predictions paradigm in beyond-worst-case-analysis (Mitzenmacher and Vassilvitskii, 2021); in-particular, we have incorporated predictions into the algorithm of Abernethy et al. (2008) via the meta-initialization approach and now present the learning-theoretic result for an end-to-end guarantee (Khodak et al., 2022).

Corollary C.1. *Let $\mathcal{K} = \{\mathbf{x} \in [0,1]^{|E|} : \langle \mathbf{a}, \mathbf{x} \rangle \leq b \ \forall \ (\mathbf{a}, b) \in \mathcal{C}\}$ be the set of flows from u to v on a graph $G(V, E)$, where $\mathcal{C} \subset \mathbb{R}^{|E|} \times \mathbb{R}$ is a set of $\mathcal{O}(|E|)$ linear constraints. Suppose we see T instances of the bandit online shortest path problem with m timesteps each. Then sampling from probability distributions over paths from u to v returned by running Algorithm 1 with regularizer $\phi(\mathbf{x}) = -\sum_{\mathbf{a},b \in \mathcal{C}} \log(b - \langle \mathbf{a}, \mathbf{x} \rangle)$ attains the following expected average regret across instances:*

$$\tilde{\mathcal{O}}\left(\frac{|E|m^2}{\sqrt[4]{T}}\right) + \min_{\frac{1}{m} \leq \varepsilon \leq \frac{1}{\sqrt{m}}} 4|E|\mathbb{E}\sqrt{2m \sum_{\mathbf{a},b \in \mathcal{C}} \log\left(\frac{\frac{1}{T}\sum_{t=1}^{T} b - \langle \mathbf{a}, \hat{\mathbf{x}}_t^{(\varepsilon)} \rangle}{\sqrt[T]{\prod_{t=1}^{T} b - \langle \mathbf{a}, \hat{\mathbf{x}}_t^{(\varepsilon)} \rangle}}\right)} + \varepsilon m \tag{11}$$

Here $\hat{\mathbf{x}}_t^{(\varepsilon)} = \mathrm{OPT}_\varepsilon(\hat{\ell}_t)$ is the ε-constrained estimated optimal flow for instance t.

Corollary C.1 shows that the average regret on the T bandit shortest-path problems scales with the sum across all constraints $\mathbf{a}, b \in \mathcal{C}$ of the log of the ratio between the arithmetic and geometric mean of the distances $b - \langle \mathbf{a}, \hat{\mathbf{x}}_t^{(\varepsilon)} \rangle$ from the estimated optimum flow $\hat{\mathbf{x}}_t^{(\varepsilon)}$ to the constraint boundary. Since the arithmetic and geometric mean are equal exactly when all entries are equal—and otherwise the former is larger—this means that the regret is small when the estimated optimal flows $\hat{\mathbf{x}}_t^{(\varepsilon)}$ for each task are at similar distances from the constraints.

Appendix D. Proof of Theorem B.1

Proof We define $\underline{\eta} = \frac{\rho D}{G\sqrt{m}}$, $\overline{\eta} = \frac{2DM}{G\sqrt{m}}$, number of grid points $k = \Omega(\lceil (4D^2M^2LG + C)\sqrt{mT}\rceil)$, $\Theta = \left\{\underline{\eta} + \frac{j}{k}(\overline{\eta} - \underline{\eta})\right\}_{j=0}^{k} \times \{\underline{\beta} + \frac{j}{k}(\overline{\beta} - \underline{\beta})\}_{j=0}^{k1_{\overline{\beta} > \underline{\beta}}} \times \{\underline{\varepsilon} + \frac{j}{k}(\overline{\varepsilon} - \underline{\varepsilon})\}_{j=0}^{k1_{\overline{\varepsilon} > \underline{\varepsilon}}}$, and meta-step-size $\alpha = \frac{1}{DG/\rho + 2DMG + C\sqrt{m}}\sqrt{\frac{3\log k}{2Tm}}$. Note that $\frac{\rho D}{G\sqrt{m}} \le \arg\min_{\eta > 0} \min_{\mathbf{x} \in \mathcal{K}_{\underline{\varepsilon}}, \beta, \varepsilon} \sum_{t=1}^{T} \tilde{U}_t(\mathbf{x}, (\eta, \beta, \varepsilon)) \le \frac{DM}{G}\sqrt{\frac{1+\rho^2}{m}} \le \frac{2DM}{G\sqrt{m}}$ so $\max_{t \in [T]} U_t^{(\rho)}(\mathbf{x}_t^{(\theta_t)}, \theta_t) \le \frac{DG\sqrt{m}}{\rho} + 2DMG\sqrt{m} + Cm$. Therefore applying the regret guarantee for exponentiated gradient (Shalev-Shwartz, 2011, Corollary 2.14) followed by the regret of follow-the-leader on a sequent of Bregman divergences (Lemma D.1) yields

$$
\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T} U_t(\mathbf{x}_t^{(\theta_t)}, \theta_t) \\
&\le \mathbb{E}\sum_{t=1}^{T} U_t^{(\rho)}(\mathbf{x}_t^{(\theta_t)}, \theta_t) \\
&\le \left(C\sqrt{m} + DG\left(\frac{1}{\rho} + 2M\right)\right)\sqrt{2mT\log|\Theta|} + \min_{\theta \in \Theta} \mathbb{E}\sum_{t=1}^{T} U_t^{(\rho)}(\mathbf{x}_t^{(\theta)}, \theta) \\
&\le \left(C\sqrt{m} + DG\left(\frac{1}{\rho} + 2M\right)\right)\sqrt{2mT\log|\Theta|} \\
&\qquad + \min_{(\eta,\beta,\varepsilon) \in \Theta} \frac{8SK^2}{\eta}(1 + \log T) + \min_{\mathbf{x} \in \mathcal{K}_\varepsilon} \mathbb{E}\sum_{t=1}^{T} \frac{B_{\phi_\beta}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x}) + \rho^2 D^2}{\eta} + (\eta G_\beta^2 + C\varepsilon)m \\
&\le \left(C\sqrt{m} + DG\left(\frac{1}{\rho} + 2M\right)\right)\sqrt{2mT\log|\Theta|} + \frac{8SK^2\overline{G}\sqrt{m}}{\rho D}(1 + \log T) \\
&\qquad + \min_{(\eta,\beta,\varepsilon) \in \Theta} \eta G_\beta^2 mT + C\varepsilon mT + \frac{\rho^2 D^2 T}{\eta} + \mathbb{E}\sum_{t=1}^{T} \frac{\phi_\beta(\hat{\mathbf{x}}_t^{(\varepsilon)}) - \phi_\beta(\hat{\bar{\mathbf{x}}}^{(\varepsilon)})}{\eta} \\
&\le \left(C\sqrt{m} + DG\left(\frac{1}{\rho} + 2M\right)\right)\sqrt{2mT\log|\Theta|} + \frac{8SK^2G\sqrt{m}}{\rho D}(1 + \log T) + \rho DGT\sqrt{m} \\
&\qquad + \left(4D^2M^2 + \left(C\sqrt{m} + \frac{2LG}{\rho D}\right)(\overline{\varepsilon} - \underline{\varepsilon})\sqrt{m} + 2\left(\frac{2M}{G} + \frac{G}{\rho}\right)DL\sqrt{m}(\overline{\beta} - \underline{\beta})\right)\frac{T}{k} \\
&\qquad + \min_{(\eta,\beta,\varepsilon) \in \overline{\Theta}} \eta G_\beta^2 mT + C\varepsilon mT + \mathbb{E}\sum_{t=1}^{T} \frac{\phi_\beta(\hat{\mathbf{x}}_t^{(\varepsilon)}) - \phi_\beta(\hat{\bar{\mathbf{x}}}^{(\varepsilon)})}{\eta} \\
&\le \left(C\sqrt{m} + DG\left(\frac{1}{\rho} + 2M\right)\right)\sqrt{6mT\log k} + (4D^2M^2LG + C)\frac{Tm}{\rho k} \\
&\qquad + \frac{8SK^2G\sqrt{m}}{\rho D}(1 + \log T) + \rho DGT\sqrt{m} + \min_{\mathbf{x} \in \mathcal{K}, \theta \in \Theta^*} \mathbb{E}\sum_{t=1}^{T} U_t(\mathbf{x}, \theta)
\end{aligned}
\tag{12}
$$

where the fourth inequality follows by Claim D.1, the fifth by Lipschitzness of $1/\eta$ on $\eta \ge \frac{\rho D}{G\sqrt{m}}$ and of $\phi_\beta(\hat{\mathbf{x}}_t^{(\varepsilon)}||\cdot)$ on $\mathcal{K}_{\underline{\varepsilon}}$, and the sixth by simplifying and substituting the lower bound for k. The w.h.p. version of the bound follows by applying Cesa-Bianchi and Lugosi (2006, Lemma 4.1) when obtaining the second inequality.

Lemma D.1. *Let* $\phi : \mathcal{K} \mapsto \mathbb{R}_{\geq 0}$ *be a strictly-convex function with* $\max_{\mathbf{x}\in\mathcal{K}} \|\nabla^2 \phi(\mathbf{x})\|_2 \leq S$ *over a convex set* $\mathcal{K} \subset \mathbb{R}^d$ *with* $\max_{\mathbf{x}\in\mathcal{K}} \|\mathbf{x}\|_2 \leq K$*. Then for any points* $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathcal{K}$ *the actions* $\mathbf{y}_1 = \arg\min_{\mathbf{x}\in\mathcal{K}} \phi(\mathbf{x})$ *and* $\mathbf{y}_t = \frac{1}{t-1}\sum_{s<t} \mathbf{x}_s$ *have regret*

$$\sum_{t=1}^T B_\phi(\mathbf{x}_t||\mathbf{y}_t) - B_\phi(\mathbf{x}_t||\mathbf{y}_{T+1}) \leq \sum_{t=1}^T \frac{8SK^2}{2t-1} \leq 8SK^2(1+\log T) \tag{13}$$

Proof Note that

$$\nabla_{\mathbf{y}} B_\phi(\mathbf{x}||\mathbf{y}) = -\nabla\phi(\mathbf{y}) - \nabla_{\mathbf{y}}\langle \nabla\phi(\mathbf{y}), \mathbf{x}\rangle + \nabla_{\mathbf{y}}\langle\nabla\phi(\mathbf{y}), \mathbf{y}\rangle = \operatorname{diag}(\nabla^2\phi(\mathbf{y}))(\mathbf{y}-\mathbf{x}) \tag{14}$$

so $B_\phi(\mathbf{x}_t||\mathbf{y})$ is $2SK$-Lipschitz w.r.t. the Euclidean norm. Applying Khodak et al. (2019, Proposition B.1) yields the result.

Claim D.1. *Let* $\phi : \mathcal{K} \mapsto \mathbb{R}$ *be a strictly-convex function over a convex set* $\mathcal{K} \subset \mathbb{R}^d$ *containing points* $\mathbf{x}_1, \dots, \mathbf{x}_T$*. Then their mean* $\bar{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^T \mathbf{x}_t$ *satisfies*

$$\sum_{t=1}^T B_\phi(\mathbf{x}_t||\bar{\mathbf{x}}) = \sum_{t=1}^T \phi(\mathbf{x}_t) - \phi(\bar{\mathbf{x}}) \tag{15}$$

Proof

$$\begin{aligned}\sum_{t=1}^T B_\phi(\mathbf{x}_t||\bar{\mathbf{x}}) &= \sum_{t=1}^T \phi(\mathbf{x}_t) - \phi(\bar{\mathbf{x}}) - \langle\nabla\phi(\bar{\mathbf{x}}), \mathbf{x}_t - \bar{\mathbf{x}}\rangle \\ &= \sum_{t=1}^T \phi(\mathbf{x}_t) - \phi(\bar{\mathbf{x}}) - \langle\nabla\phi(\bar{\mathbf{x}}), \sum_{t=1}^T \mathbf{x}_t - \bar{\mathbf{x}}\rangle \\ &= \sum_{t=1}^T \phi(\mathbf{x}_t) - \phi(\bar{\mathbf{x}})\end{aligned} \tag{16}$$

Appendix E. Proof of Theorem 3.1

Proof Since ε is constant we use the shorthand $\hat{\mathbf{x}}_t^{(\varepsilon)} = \hat{\mathbf{x}}_t^{(\theta)}\ \forall\ \theta \in \Theta$. Note that we use search space $\Theta = \left[\frac{\rho}{\sqrt{m}}, 2\sqrt{\frac{d\log d}{em}}\right] \times \left[\max\left\{\frac{1}{\log d}, \underline{\beta}\right\}, 1\right] \times \{\varepsilon\}$. We have the constants $D \leq \sqrt{d}$, $G \leq \sqrt{d}$, $M \leq \sqrt{\frac{d\log d}{e}}$, $S \leq \left(\frac{d}{\varepsilon}\right)^{2-\underline{\beta}}$, and $K = 1$. Note that the second term $d^{1-\beta}m/\beta$ is decreasing on $\beta < 1/\log d$, so since ϕ_β is always increasing in β we know that the optimal β is in $[1/\log d, 1]$. Note that by Lemma E.2 we have that $L = d\log\frac{d}{\varepsilon}$. We thus have

$$
\begin{aligned}
&\sum_{t=1}^T \sum_{i=1}^m \ell_{t,i}(a_{t,i}) - \ell_{t,i}(a_t^*) \\
&\qquad \leq \sum_{t=1}^T \sum_{i=1}^m \langle \hat{\ell}_{t,i}, \mathbf{x}_{t,i} - \ell_{t,i}(a_t^*)\rangle + \gamma \sum_{a=1}^d \hat{\ell}_{t,i}(a) \\
&\qquad \leq \sum_{t=1}^T \frac{B_{\phi_{\beta_t}}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x}_{t,1})}{\eta_t} + \sum_{i=1}^m \langle \hat{\ell}_{t,i}, \hat{\mathbf{x}}_{t,i}^{(\varepsilon)}\rangle - \ell_{t,i}(a_t^*) + \frac{\eta_t}{\beta_t}\sum_{a=1}^d \mathbf{x}_{t,i}^{2-\beta_t}(a)\hat{\ell}_{t,i}^2(a) + \gamma \sum_{a=1}^d \hat{\ell}_{t,i}(a) \\
&\qquad \leq \frac{\varepsilon m T}{\gamma d} + \sum_{t=1}^T \frac{B_{\phi_{\beta_t}}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x}_{t,1})}{\eta_t} + \sum_{i=1}^m \hat{\ell}_{t,i}(a_t^*) - \ell_{t,i}(a_t^*) \\
&\qquad\qquad + \sum_{t=1}^T \frac{\eta_t}{\beta_t}\sum_{i=1}^m \sum_{a=1}^d \mathbf{x}_{t,i}^{1-\beta_t}(a)\hat{\ell}_{t,i}(a) + \gamma \sum_{a=1}^d \hat{\ell}_{t,i}(a) \\
&\qquad \leq \frac{\varepsilon m T}{\gamma d} + \frac{1 + \frac{\overline{\eta}}{\underline{\beta}} + \gamma}{2\gamma}\log\frac{4}{\delta} + \sum_{t=1}^T \frac{B_{\phi_{\beta_t}}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x}_{t,1})}{\eta_t} \\
&\qquad\qquad + \sum_{t=1}^T \frac{\eta_t}{\beta_t}\sum_{i=1}^m \sum_{a=1}^d \mathbf{x}_{t,i}^{1-\beta_t}(a)\ell_{t,i}(a) + \gamma \sum_{a=1}^d \ell_{t,i}(a) \\
&\qquad \leq \frac{\varepsilon m T}{\gamma d} + \frac{1 + \sqrt{\frac{d\log^3 d}{em}}}{\gamma}\log\frac{4}{\delta} + \gamma d m T + \sum_{t=1}^T \frac{B_{\phi_{\beta_t}}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x}_{t,1})}{\eta_t} + \frac{\eta_t d^{\beta_t} m}{\beta_t} \\
&\qquad \leq \frac{\varepsilon m T}{\gamma d} + \frac{1 + \sqrt{\frac{d\log^3 d}{em}}}{\gamma}\log\frac{4}{\delta} + \gamma d m T + \min_{\mathbf{x}\in\triangle_d, \eta>0, \beta\in[\underline{\beta},1]} \sum_{t=1}^T \frac{B_{\phi_\beta}(\hat{\mathbf{x}}_t^{(\varepsilon)}||\mathbf{x})}{\eta} + \frac{\eta d^\beta m}{\beta} \\
&\qquad\qquad + 2d\left(\frac{1}{\rho} + \sqrt{\frac{d\log d}{e}}\right)\sqrt{6mT\log\frac{4k}{\delta}} + \frac{8d^{2-\underline{\beta}}\sqrt{m}}{\rho\varepsilon^{2-\underline{\beta}}}(1 + \log T) + \rho d T\sqrt{m} \\
&\qquad \leq \left(\frac{\varepsilon}{\gamma d} + \gamma d\right) m T + \frac{1 + \sqrt{\frac{d\log^3 d}{em}}}{\gamma}\log\frac{4}{\delta} + T \min_{\eta>0, \beta\in[\underline{\beta},1]} \frac{H_\beta(\hat{\hat{\mathbf{x}}})}{\eta} + \frac{\eta d^\beta m}{\beta} + \frac{\varepsilon^\beta d^{1-\beta} 1_{\beta<1}}{(1-\beta)\eta} \\
&\qquad\qquad + 2d\left(\frac{1}{\rho} + \sqrt{\frac{d\log d}{e}}\right)\sqrt{6mT\log\frac{4k}{\delta}} + \frac{8d^{2-\underline{\beta}}\sqrt{m}}{\rho\varepsilon^{2-\underline{\beta}}}(1 + \log T) + \rho d T\sqrt{m}
\end{aligned}
\tag{17}
$$

where the second inequality follows by Lemma E.1, the third by Hölder's inequality and the definitions $\hat{\ell}_{t,i}$ and $\hat{\mathbf{x}}_{t,i}^{(\varepsilon)}$, the fourth by Neu (2015, Lemma 1), the fifth by the definition of $\ell_{t,i}$, the sixth by Theorem B.1, and the last by the derivation below for $\beta < 1$ (otherwise it holds by joint convexity of the KL-divergence) followed by Claim D.1 combined with the fact that the entropy of optima-in-hindsight is zero.

$$\begin{aligned}
-\phi_\beta((1-\varepsilon)\mathbf{x} + \varepsilon \mathbf{1}_d/d) &= \frac{\sum_{a=1}^d ((1-\varepsilon)\mathbf{x}(a) + \varepsilon/d)^\beta - 1}{1-\beta} \\
&\le \frac{\varepsilon^\beta d^{1-\beta} + (1-\varepsilon)^\beta \sum_{a=1}^d \mathbf{x}^\beta(a) - 1}{1-\beta} \le \frac{\varepsilon^\beta d^{1-\beta}}{1-\beta}
\end{aligned} \tag{18}$$

Lemma E.1. *Suppose we play* $\texttt{OMD}_{\beta,\eta}$ *with regularizer* ϕ_β *the negative Tsallis entropy and initialization* $\mathbf{x}_1 \in \triangle_d$ *on the sequence of linear loss functions* $\ell_1, \dots, \ell_T \in [0,1]^d$*. Then for any* $\mathbf{x}^* \in \triangle_d$ *we have*

$$\sum_{t=1}^T \langle \ell_t, \mathbf{x}_t - \mathbf{x}^* \rangle \le \frac{B_{\phi_\beta}(\mathbf{x}^* || \mathbf{x}_1)}{\eta} + \frac{\eta}{\beta} \sum_{a=1}^d \mathbf{x}_t^{2-\beta}(a) \ell_t^2(a) \tag{19}$$

Proof Note that the following proof follows parts of the course notes by Luo (2017), which we reproduce for completeness. The OMD update at each step t involves the following two steps: set $\mathbf{y}_{t+1} \in \triangle_d$ s.t. $\nabla\phi_\beta(\mathbf{y}_{t+1}) = \nabla\phi_\beta(\mathbf{x}_t) - \eta\ell_t$ and then set $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\triangle_d} B_{\phi_\beta}(\mathbf{x}, \mathbf{y}_{t+1})$ (Hazan, 2015, Algorithm 14). Note that by Hazan (2015, Equation 5.3) and nonnegativity of the Bregman divergence we have

$$\sum_{t=1}^T \langle \ell_t, \mathbf{x}_t - \mathbf{x}^* \rangle \le \frac{B_{\phi_\beta}(\mathbf{x}^* || \mathbf{x}_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T B_{\phi_\beta}(\mathbf{x}_t || \mathbf{y}_{t+1}) \tag{20}$$

To bound the second term, note that when ϕ_β is the negative Tsallis entropy we have

$$\begin{aligned}
&B_{\phi_\beta}(\mathbf{x}_t || \mathbf{y}_{t+1}) \\
&\quad= \frac{1}{1-\beta} \sum_{a=1}^d \left(\mathbf{y}_{t+1}^\beta(a) - \mathbf{x}_t^\beta(a) + \frac{\beta}{\mathbf{y}_{t+1}^{1-\beta}(a)} (\mathbf{x}_t(a) - \mathbf{y}_{t+1}(a)) \right) \\
&\quad= \frac{1}{1-\beta} \sum_{a=1}^d \left((1-\beta)\mathbf{y}_{t+1}^\beta(a) - \mathbf{x}_t^\beta(a) + \beta \left(\frac{1}{\mathbf{x}_t^{1-\beta}(a)} + \frac{1-\beta}{\beta} \eta \ell_t(a) \right) \mathbf{x}_t(a) \right) \\
&\quad= \sum_{a=1}^d \left(\mathbf{y}_{t+1}^\beta(a) - \mathbf{x}_t^\beta(a) + \eta \mathbf{x}_t(a) \ell_t(a) \right)
\end{aligned} \tag{21}$$

Plugging the following result, which follows from $(1+x)^\alpha \leq 1+\alpha x+\alpha(\alpha-1)x^2 \;\forall\; x \geq 0, \alpha < 0$, into the above yields the desired bound.

$$\begin{aligned}
\mathbf{y}_{t+1}^\beta(a) = \mathbf{x}_t^\beta(a)\left(\frac{\mathbf{y}_{t+1}^{\beta-1}(a)}{\mathbf{x}_t^{\beta-1}(a)}\right)^{\frac{\beta}{\beta-1}} &= \mathbf{x}_t^\beta(a)\left(1+\frac{1-\beta}{\beta}\eta\mathbf{x}_t^{1-\beta}(a)\ell_t(a)\right)^{\frac{\beta}{\beta-1}} \\
&\leq \mathbf{x}_t^\beta(a)\left(1-\eta\mathbf{x}_t^{1-\beta}(a)\ell_t(a)+\frac{\eta^2}{\beta}\mathbf{x}_t^{2-2\beta}(a)\ell_t(a)^2\right) \qquad (22) \\
&= \mathbf{x}_t^\beta(a)-\eta\mathbf{x}_t(a)\ell_t(a)+\frac{\eta^2}{\beta}\mathbf{x}_t^{2-\beta}(a)\ell_t(a)^2
\end{aligned}$$

Lemma E.2. *For any $\rho \in (0, 1/d]$ and $\mathbf{x} \in \triangle_d$ s.t. $\mathbf{x}(a) \geq \rho \;\forall\; a \in [d]$ the β-Tsallis entropy $H_\beta(\mathbf{x}) = -\frac{1-\sum_{a=1}^d \mathbf{x}^\beta(a)}{1-\beta}$ is $d\log\frac{1}{\rho}$-Lipschitz w.r.t. $\beta \in [0,1]$.*

Proof Let $\log_\beta x = \frac{x^{1-\beta}-1}{1-\beta}$ be the β-logarithm function and note that by Yamano (2002, Equation 6) we have $\log_\beta x - \log x = (1-\beta)(\partial_b \log_\beta x + \log_\beta x \log x) \geq 0 \;\forall\; \beta \in [0,1]$. Then we have for $\beta \in [0,1)$ that

$$\begin{aligned}
|\partial_\beta H_\beta(\mathbf{x})| &= \left|\frac{-H_\beta(\mathbf{x}) - \sum_{a=1}^d \mathbf{x}^\beta(a)\log\mathbf{x}(a)}{1-\beta}\right| \\
&= \frac{1}{1-\beta}\left|\sum_{a=1}^d \mathbf{x}^\beta(a)(\log_\beta \mathbf{x}(a) - \log\mathbf{x}(a))\right| \\
&= \frac{1}{1-\beta}\sum_{a=1}^d \mathbf{x}^\beta(a)(\log_\beta \mathbf{x}(a) - \log\mathbf{x}(a)) \\
&\leq \frac{1}{1-\beta}\left(\sum_{a=1}^d \mathbf{x}(a)\right)^\beta \left(\sum_{a=1}^d (\log_\beta \mathbf{x}(a) - \log\mathbf{x}(a))^{\frac{1}{1-\beta}}\right)^{1-\beta} \qquad (23) \\
&\leq \frac{1}{1-\beta}\sum_{a=1}^d \log_\beta \mathbf{x}(a) - \log\mathbf{x}(a) \\
&\leq \frac{d}{1-\beta}(\log_\beta \rho - \log\rho) \\
&\leq -d\log\rho
\end{aligned}$$

where the fourth line follows by Hölder's inequality, the fifth by subadditivity of x^a for $a \in (0,1]$, the sixth by the fact that $\partial_x(\log_\beta x - \log x) = x^{-\beta} - 1/x \leq 0 \;\forall\; \beta, x \in [0,1)$, and the last line by substituting $\beta = 0$ since $\partial_\beta\left(\frac{\log_\beta \rho - \log\rho}{1-\beta}\right) = \frac{2(\rho-\rho^\beta)-(1-\beta)(\rho^\beta+\rho)\log\rho}{\rho^\beta(1-\beta)^3} \leq 0 \;\forall\; \beta \in [0,1), \rho \in (0, 1/d]$. For $\beta = 1$, applying L'Hôpital's rule yields

$$\lim_{\beta\to1}\partial_\beta H_\beta(\mathbf{x}) = -\frac{1}{2}\lim_{\beta\to1}\sum_{a=1}^d \mathbf{x}^\beta(a)\log^2\mathbf{x}(a)(1-(1-\beta)\log\mathbf{x}(a)) = -\frac{1}{2}\sum_{a=1}^d \mathbf{x}(a)\log^2\mathbf{x}(a) \qquad (24)$$

which is bounded on $[-2d/e^2, 0]$.

Appendix F. Proof of Theorem 4.1

Proof Applying Theorem B.1 with constants $D = D_{\underline{\varepsilon}}$, $G = 4d\sqrt{2}$, $M = 1$, $S = S_{\underline{\varepsilon}}$, and $K = K$ yields

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\sum_{i=1}^{m}\langle \ell_{t,i}, \mathbf{x}_{t,i} - \mathbf{x}_t^*\rangle \\
&\leq \mathbb{E}\sum_{t=1}^{T} \varepsilon_t m + \sum_{i=1}^{m}\langle \ell_{t,i}, \mathbf{x}_{t,i} - \mathrm{OPT}_{\varepsilon_t}(\ell_t)\rangle \\
&= \mathbb{E}\sum_{t=1}^{T} \varepsilon_t m + \sum_{i=1}^{m}\langle \hat{\ell}_{t,i}, \mathbf{x}_{t,i} - \mathrm{OPT}_{\varepsilon_t}(\ell_t)\rangle \\
&\leq \mathbb{E}\sum_{t=1}^{T} \varepsilon_t m + \sum_{i=1}^{m}\langle \hat{\ell}_{t,i}, \mathbf{x}_{t,i} - \hat{\mathbf{x}}_t^{(\theta_t)}\rangle \\
&\leq \mathbb{E}\sum_{t=1}^{T} \frac{B_\phi(\hat{\mathbf{x}}_t^{(\theta_t)}||\mathbf{x}_{t,1}^{(\theta_t)})}{\eta_t} + (\eta_t G^2 + \varepsilon_t)m \qquad (25) \\
&\leq \left(\sqrt{m} + \frac{4D_{\underline{\varepsilon}}G}{\rho}\right)\sqrt{6mT\log k} + \frac{8S_{\underline{\varepsilon}}K^2G\sqrt{m}}{\rho D_{\underline{\varepsilon}}}(1+\log T) + \rho D_{\underline{\varepsilon}}GT\sqrt{m} \\
&\qquad + \min_{\mathbf{x}\in\mathcal{K},\eta>0,\varepsilon\in[\underline{\varepsilon},\bar{\varepsilon}]} \mathbb{E}\sum_{t=1}^{T}\frac{B_\phi(\mathrm{OPT}_\varepsilon(\hat{\ell}_t)||\mathbf{x})}{\eta} + (\eta G^2 + \varepsilon)m \\
&\leq 72d\sqrt{m}\sqrt[4]{T}\left(D_{\underline{\varepsilon}}\sqrt{mT\log k} + \frac{S_{\underline{\varepsilon}}K^2}{D_{\underline{\varepsilon}}}(1+\log T)\right) \\
&\qquad + \min_{\mathbf{x}\in\mathcal{K},\eta>0,\varepsilon\in[\underline{\varepsilon},\bar{\varepsilon}]} \mathbb{E}\sum_{t=1}^{T}\frac{B_\phi(\mathrm{OPT}_\varepsilon(\hat{\ell}_t)||\mathbf{x})}{\eta} + (32\eta d^2 + \varepsilon)m
\end{aligned}$$

where the third inequality follows from Lemma F.1.

Lemma F.1. *Let $\overline{\mathcal{K}} \subset \mathbb{R}^d$ be a convex set and ϕ be a self-concordant barrier. Suppose $\ell_1, \dots, \ell_T$ are a sequence of loss functions satisfying $|\langle \ell_t, \mathbf{x}\rangle| \leq 1 \ \forall\ \mathbf{x} \in \mathcal{K}$. Then if we run OMD with step-size $\eta > 0$ as in Abernethy et al. (2008, Algorithm 1) on the sequence of estimators $\hat{\ell}_t$ our estimated regret w.r.t. any $\mathbf{x}^* \in \mathcal{K}_\varepsilon$ for $\varepsilon > 0$ will satisfy*

$$\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{x}_t - \mathbf{x}^*\rangle \leq \frac{B_\phi(\mathbf{x}^*||\mathbf{x}_1)}{\eta} + 32d^2\eta T \qquad (26)$$

Proof The result follows from Abernethy et al. (2008) by stopping the derivation on the second inequality below Equation 10.