Nonstochastic Bandits and Experts with Arm-Dependent Delays

Dirk van der Hoeven
Dept. of Computer Science
Università degli Studi di Milano, Italy

Nicolò Cesa-Bianchi
DSRC & Dept. of Computer Science
Università degli Studi di Milano, Italy

Abstract

We study nonstochastic bandits and experts in a delayed setting where delays depend on both time and arms. While the setting in which delays only depend on time has been extensively studied, the arm-dependent delay setting better captures real-world applications at the cost of introducing new technical challenges. In the full information (experts) setting, we design an algorithm with a first-order regret bound that reveals an interesting trade-off between delays and losses. We prove a similar first-order regret bound also for the bandit setting, when the learner is allowed to observe how many losses are missing. These are the first bounds in the delayed setting that depend on the losses and delays of the best arm only. When in the bandit setting no information other than the losses is observed, we still manage to prove a regret bound through a modification to the algorithm of Zimmert and Seldin (2020). Our analyses hinge on a novel bound on the drift, measuring how much better an algorithm can perform when given a look-ahead of one round.

1 INTRODUCTION

Delayed feedback is a common challenge in many online learning problems. For example, suppose you want to sell several shares of a company on the stock market, but you do not know yet at what price you want to sell them. To learn a good price, you could do a piecemeal sale and track the price at which shares sell. However, shares on sale at a low price will most likely be bought quickly, while shares on sale at a high price could be bought only much later in time. Hence, depending on the price you set, a sale may be completed sooner or later. Similarly, when deciding which advertisement to show on a web page, one might receive feedback at different points in time depending on which advertisement you show. If you show an ad of an expensive car, it might take some time for the person who saw the advertisement to make the decision to buy the car Vice versa, showing an advertisement of something significantly cheaper than a car might not induce a long delay between impression and conversion.

In this paper we work in a general setting of sequential decision-making that captures these problems. We consider two forms of feedback: full information feedback, in which after each prediction the learner observes the losses of all actions, and bandit feedback, in which the learner only observes the loss of the chosen action. For both full information and bandit setting we assume an oblivious adversary, which is to say that both losses and delays are adversarially generated before the start of the first round.

Formally, the full information setting works as follows. In each round $t = 1, \dots, T$ the learner issues a prediction $\boldsymbol{q}_t \in \triangle$ and suffers loss $\langle \boldsymbol{q}_t, \boldsymbol{\ell}_t \rangle$, where $\triangle = \left\{ \boldsymbol{q} : q(i) \geq 0, \sum_{i=1}^K q(i) = 1 \right\}$ is the K-dimensional simplex. The loss $\ell_t(i) \in [0,1]$ of arm i in round t is revealed to the learner at the end of round $t + d_t(i)$, where $d_t(i) \geq 0$. In the full information setting, the goal is to have small regret $\mathcal{R}_T(\boldsymbol{u})$ against any $\boldsymbol{u} \in \triangle$ after any number T of rounds, where

$$\mathcal{R}_T(\boldsymbol{u}) = \sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle$$

In the bandit setting, the learner issues prediction $i_t \in [K]$ and suffers loss $\ell_t(i_t)$. We consider two variants of the bandit setting. In the first variant, the learner receives the loss $\ell_t(i_t)$ as well as the number of missing observations $\rho_t(i_t) = |\{s : s < t, s + d_s(i_t) \geq t\}|$ for arm i_t at the end of round $t + d_t(i_t)$, when $\ell_t(i_t)$ is observed. We refer to this variant as the partially-concealed bandit setting. In the second variant, at the end of round $t + d_t(i_t)$ the learner just receives $\ell_t(i_t)$.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

We refer to the second variant as the concealed bandit setting. In both variants, the goal is to have small expected regret: $\mathbb{E}[\mathcal{R}_T(\boldsymbol{u})] = \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(i_t) - \langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle)\right]$, where the expectation is with respect to the randomness in the learner's actions.

The reason why we differentiate between the partially-concealed and concealed bandit setting is the following. Due to arm-dependent delays, we face an additional challenge in the bandit setting. In the full information setting, $\rho_t(i)$ may be used to tune the algorithm, and is readily available to the learner who can keep track of the missing losses for each arm. In the bandit setting, however, the learner only knows whether or not a loss is missing for an arm if that arm was played. Therefore, the learner cannot compute $\rho_t(i)$ or have knowledge of $d_t(i)$ if arm i was not played in round t. Assuming that the learner also observes $\rho_t(i_t)$ when $\ell_t(i_t)$ is observed is a relatively mild assumption slightly simplifying the aforementioned challenge. Other assumptions in the literature are stronger than ours. For example, a common assumption is that delays are the same in each round (Weinberger and Ordentlich, 2002; Neu et al., 2010, 2014; Cesa-Bianchi et al., 2016) or known before issuing a prediction (Thune et al., 2019; Bistritz et al., 2019)[1]. When all arms suffer the same delay, $\rho_t(i_t)$ can be readily computed by simply counting the number of missing observations at prediction time. Regardless, we also study the case in which the learner does not observe $\rho_t(i_t)$, and derive algorithms for both the partially concealed and the concealed bandit settings.

Related Works In the full information setting, Weinberger and Ordentlich (2002) were the first to study online learning with delayed feedback. They assumed full information feedback and a constant and known delay. Subsequent works by Zinkevich et al. (2009); Joulani et al. (2013); Quanrud and Khashabi (2015) relaxed the assumption on the delay, and worked in the more general setting of online convex optimization. Joulani et al. (2016) also work with a convex domain and provide a reduction for deriving algorithms with a gradient-dependent regret bound using a strongly convex regularizer. Recently, Flaspohler et al. (2021) analysed the role of optimism in online convex optimization with delays, and developed several adaptive algorithms for the arm-independent setting.

Other works consider the distributed or parallel online convex optimization setting, see (Agarwal and Duchi, 2012; McMahan and Streeter, 2014; Sra et al., 2015; Hsieh et al., 2020; Van der Hoeven et al., 2021) and the references therein. In these works, delays arise due to communication between agents, gradient evaluations, or distance between agents in a network of agents.

All previous works in the nonstochastic bandit setting considered delayed feedback assuming equal delays for each arm. The impact of delay under bandit feedback was first studied by Neu et al. (2010, 2014), who assumed that the delay was constant and known. Cesa-Bianchi et al. (2019) stated a lower bound for nonstochastic bandits of order $\max\{\sqrt{KT}, \sqrt{\ln(K)D}\}$, where D is the total delay. Thune et al. (2019); Bistritz et al. (2019, 2021) all provided algorithms based on EXP3 that matched the lower bound up to a factor $\ln(K)$ by utilizing either a priori knowledge of D or knowledge of d_t before issuing a prediction. Zimmert and Seldin (2020) provide an algorithm matching the lower bound without using any additional assumptions on d_t. György and Joulani (2021) present an improved analysis of a variant of EXP3 under delayed feedback, and provide a delay adaptive regret bound, a data and delay adaptive regret bound, and a high-probability regret bound.

Another related setting is the nonstochastic bandits with composite anonymous feedback setting (Cesa-Bianchi et al., 2018). In this setting, the learner does not observe the identity of the losses, which is to say that the learner does not see which round a given loss is from. To complicate matters further, losses from several rounds may be composed without the possibility to distinguish which rounds or actions are in the composed losses. As long as a uniform upper bound on the delays $d^\star$ is known, the algorithm by Cesa-Bianchi et al. (2018) also works in our setting, but using this algorithm would lead to a suboptimal regret bound of order $\sqrt{d^\star KT\ln(K)}$.

A stochastic version of the concealed bandit setting has been considered by Gael et al. (2020); Lancewicki et al. (2021). They study i.i.d. rewards (rather than losses) and assume a slightly more general feedback model in which the learner does not always see the reward for actions. Other works in the stochastic setting include (Chapelle and Li, 2011; Dudík et al., 2011; Joulani et al., 2013; Desautels et al., 2014; Chapelle, 2014; Mandel et al., 2015; Vernade et al., 2017, 2018; Pike-Burke et al., 2018), all with varying assumptions on the delay, but all with uniform delay over the arms.

Contributions In the full information setting, we provide an algorithm that satisfies, for any $\boldsymbol{u} \in \triangle$, a

[1] Although Bistritz et al. (2019) also claim a result that does not use a-priori knowledge of delay, there appears to be an error in their analysis: see the discussion by György and Joulani (2021).

first-order bound of the form

$$\mathcal{R}_T(\boldsymbol{u}) = \widetilde{O}\left(\sqrt{\big(\mathrm{KL}(\boldsymbol{u},\pi) + \ln(T)\big)\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho\rangle}\right) \tag{1}$$

where π is a prior set by the learner, $L_T(i) = \sum_{t=1}^T \ell_t(i)$, $L_T^\rho(i) = \sum_{t=1}^T \ell_t(i)\rho_t(i)$, and $\widetilde{O}$ hides logarithmic factors. This bound takes advantage of scenarios when there is an arm with consistently small losses, or when the delays of a given arm tend to be small. In the worst case, bound (1) is $\widetilde{O}\big(\sqrt{T(1+\rho_T^{\max})\ln(K)}\big)$ for a uniform prior π. While several first-order regret bounds are known in the literature—see for example (Hutter and Poland, 2004; Kalai and Vempala, 2005; Cesa-Bianchi and Lugosi, 2006; Cesa-Bianchi et al., 2007; Van Erven et al., 2014; Chen et al., 2021)—to the best of our knowledge there are no first-order regret bounds for the delayed feedback setting. The regret bound in (1) reveals an interesting trade-off. When competing with $\boldsymbol{e}_{i^\star} = \arg\min_{\boldsymbol{u}\in\triangle}\langle \boldsymbol{u}, \boldsymbol{L}_T\rangle$ we could suffer a large penalty for delay if $i^\star$ always suffers large delays (and thus incur large regret). On the other hand, if we try to balance delay and losses by competing with $\arg\min_{\boldsymbol{u}\in\triangle}\{\langle \boldsymbol{u}, \boldsymbol{L}_T\rangle + \sqrt{\mathrm{KL}(\boldsymbol{u},\pi)\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho\rangle}\}$, we could incur a smaller penalty for the delay at the cost of moving away from the minimizer of the loss. In our first motivating example, these scenarios correspond to, respectively, the optimal price of a share always incurring a small loss but a large delay, and a price at which a share always sells quickly while possibly incurring a larger loss.

In the partially concealed bandit setting we also provide an algorithm with a trade-off between losses and delays: our algorithm guarantees that the regret $\mathbb{E}\left[\mathcal{R}_T(\boldsymbol{u})\right]$ is bounded, for any $\boldsymbol{u} \in \triangle$, by

$$\widetilde{O}\Big(\sqrt{K\langle \boldsymbol{u}, \boldsymbol{L}_T\rangle} + \sqrt{\big(\mathrm{KL}(\boldsymbol{u},\pi) + \ln(T)\big)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle}\Big)$$

In the bandit literature there are several algorithms known to achieve first-order regret bounds. For example, the Green algorithm (Allenberg et al., 2006), FPL-TrIX (Neu, 2015b), or FTRL with the log-barrier (Foster et al., 2016) all guarantee a first-order regret bound. In the delayed bandit literature, György and Joulani (2021) provide an algorithm with a regret bound of order $\sqrt{d^\star\langle \boldsymbol{u}, \boldsymbol{L}_T\rangle} + \sqrt{\ln(K)\sum_{i=1}^K L_T(i)}$ for $d^\star = \max_{i,t} d_t(i)$, which—to the best of our knowledge—is the one closest to a first-order regret bound. While in the worst case this bound is a factor $\ln(T)$ better than our bound, unlike our bound it does not capture the trade-off between delays and losses.

In the concealed bandit setting, we circumvent the need of knowing $\rho_t(i_t)$ by modifying the algorithm of Zimmert and Seldin (2020). Our variant guarantees an expected regret bound of order

$$\sqrt{KT} + \sqrt{\ln(K)\,\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^K q_t(i)\rho_t(i)\right]} + \rho^\star$$

where $\rho^\star \geq \max_{i,t}\rho_t(i)$ is the maximum number of missing observations and , we recall it, $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_T$ are the learner's predictions. While our bound is stated in terms of $\sum_{i=1}^K q_t(i)\rho_t(i)$, a hardly interpretable quantity which depends on the actions of the algorithm, we may also recover a bound in terms of $\sum_{i=1}^K q_t(i)\rho_t(i) \leq \max_i \rho_t(i)$, the maximum number of missing observations per round. Note that $\sum_{i=1}^K q_t(i)\rho_t(i)$ or even $\max_i \rho_t(i)$ are not known in the concealed bandit setting, making adaptive tuning necessary. Zimmert and Seldin (2020) consider a special case of our setting in which delays are equal across arms, which we call the arm-independent delay setting. In the arm-independent delay setting $\sum_{i=1}^K q_t(i)\rho_t(i) = \rho_t$ is known and may be used for tuning, which Zimmert and Seldin (2020) do to recover the optimal regret bound. However, in our arm-dependent delay setting we do not know $\sum_{i=1}^K q_t(i)\rho_t(i)$. We address this challenge by modifying the algorithm of Zimmert and Seldin (2020), with which we recover the results of Zimmert and Seldin (2020) up to the additive $\rho^\star$ in the arm-independent setting as well as providing the bound in (1) in the arm-dependent delay setting. The data-adaptive algorithm of György and Joulani (2021) could also be applied in the concealed bandit setting, as they refrain from using any information not available to the learner in the tuning of their algorithm. However, this would yield an additional factor $\ln(K)$ and an additive $(\rho^\star)^2$ in the regret bound, rather than the additive $\rho^\star$ term which we obtain.

As for the technical contributions, our algorithm is an instance of Follow the Regularized Leader (FTRL) with corrections. Hazan and Kale (2010); Steinhardt and Liang (2014); Wei and Luo (2018); Chen et al. (2021) also use FTRL with corrections, but solely in the non-delayed setting. Our contribution here is the analysis of the FTRL with corrections algorithm under delayed feedback. In particular, we provide a powerful tool for deriving regret bounds in both the full information and bandit settings. In our analysis we build upon the framework of György and Joulani (2021), who split the regret into a cheating regret term and a drift term. The cheating regret, which is the regret of an omniscient algorithm knowing all past losses (including the loss of the current round), can be bounded by standard techniques. The drift is the difference between the loss of the algorithm and the loss of the omniscient predictor. While György and Joulani (2021) only provide bounds on the drift term for a version of exponential weights, we provide a general bound for

FTRL algorithms under mild conditions on the regularizer, thus significantly simplifying the subsequent analysis. In the full information setting, we use this new bound to simultaneously learn the optimal learning rate and the best arm using a particular version of the negative entropy as regularizer. In the partially concealed bandit setting, we develop a new hybrid regularizer that allows us to simultaneously adapt to delay and loss of the best arm. In the concealed bandit setting, we show the power of our new bound by recovering the nontuned results of Zimmert and Seldin (2020), and subsequently adapting the tuning to suit our purposes.

The remainder of the paper is organised as follows. In Section 2 we introduce the algorithm and the corresponding analysis, which we use to derive our results. In Section 3 we specify the regularizer used in the aforementioned algorithm and use the analysis from Section 2 to prove a regret bound. In Section 4 we describe and analyze the regularizers used to derive the results for the concealed and partially concealed bandit settings. Finally, in section 5 we discuss some open problems.

2 ALGORITHM

One of the main challenges in designing an adaptive algorithm that can handle delayed feedback is that tuning the learning rate may involve unknown losses due to the delay. In order to resolve this issue, we learn the best learning rate by specifying multiple learning rates for each expert. This means that we create several pseudo-experts and the algorithm computes distribution a $\boldsymbol{p}_t^{av}$ over the pseudo-experts experts in each round. To compute the actual predictions we sum the weights over pseudo-experts that share an expert, i.e., we sum over the learning rates. A precise definition of our predictions can be found in the subsequent sections.

Our version of FTRL with corrections is given in Algorithm 1. Note that both the correction $a_t(i')$ and loss (estimate) $\widetilde{\ell}_t(i')$ are left unspecified. $\widetilde{\ell}_t(i')$ has different definitions in the bandit and the full information setting, as only in the bandit setting we need to estimate the loss. Similarly, $a_t(i')$ will be specified in the relevant sections as different choices for $a_t(i')$ lead to different regret bounds. We use this flexibility to its full extent by choosing different corrections for the full information, partially concealed bandit, and concealed bandit settings.

Our analysis builds on the following simple but useful decomposition due to György and Joulani (2021):

Algorithm 1 Delayed FTRL with Corrections

Require: Regularizers Φ_t and Ψ_t, priors $\boldsymbol{p}_1^{\Phi}$ and $\boldsymbol{p}_1^{\Psi}$.

for $t = 1 \dots T$ **do**

  Using R_t defined in (5), compute

$$\boldsymbol{p}_t^{av} = \underset{\boldsymbol{p} \in \triangle'}{\arg\min} \langle \boldsymbol{p}, \widehat{\boldsymbol{L}}_t^{av} \rangle + R_t(\boldsymbol{p}) \tag{2}$$

  for $i' \in K'$ **do**

    for $s \in A_t(i') \equiv \{s : s + d_s(i') = t\}$ **do**

      Get loss $\widetilde{\ell}_s(i')$ and correction $a_s(i')$

      Set $\widehat{\ell}_s(i') = \widetilde{\ell}_s(i') + a_s(i')$

    end for

    Update $\widehat{L}_{t+1}^{av}(i') = \widehat{L}_t^{av}(i') + \sum_{s \in A_t(i')} \widehat{\ell}_s(i')$

  end for

end for

$$\begin{aligned}&\sum\nolimits_{t=1}^{T} \langle \boldsymbol{p}_t^{av} - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t \rangle \\ &= \underbrace{\sum\nolimits_{t=1}^{T} \langle \boldsymbol{p}_{t+1} - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t \rangle}_{\text{Cheating regret}} + \underbrace{\sum\nolimits_{t=1}^{T} \langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t \rangle}_{\text{Drift}}\end{aligned} \tag{3}$$

Here $\boldsymbol{p}_{t+1}$ is the FTRL distribution when all losses up to and including the loss at round t are known:

$$\boldsymbol{p}_{t+1} = \arg\min\nolimits_{\boldsymbol{p} \in \triangle'} \langle \boldsymbol{p}, \widehat{\boldsymbol{L}}_t \rangle + R_t(\boldsymbol{p}) \tag{4}$$

We let $\widehat{\boldsymbol{L}}_t = \sum_{s=1}^{t} \widehat{\boldsymbol{\ell}}_s$ and also let $\triangle'$ be the enlarged simplex $\{\boldsymbol{p} : p(i') \geq 0, \sum_{i'=1}^{K'} p(i') = 1\}$ over $K' \geq K$ coordinates. Note our notational convention here: $\widehat{\boldsymbol{L}}_t^{av}$ contains all the information available to the learner *before* issuing a prediction in round t and is used to compute $\boldsymbol{p}_t^{av}$, whereas $\widehat{\boldsymbol{L}}_t$ contains all the losses up to and *including* $\widehat{\boldsymbol{\ell}}_t$.

For some Ψ_t and Φ_t to be specified later, we always use regularizers of the form

$$R_t(\boldsymbol{p}) = B_{\Psi_t}(\boldsymbol{p}, \boldsymbol{p}_1^{\Psi}) + B_{\Phi_t}(\boldsymbol{p}, \boldsymbol{p}_1^{\Phi}) \tag{5}$$

where $B_F(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}) - F(\boldsymbol{y}) - \langle \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$ is the Bregman divergence between $\boldsymbol{x}$ and $\boldsymbol{y}$ generated by F. We require that $F_t = \Psi_t + \Phi_t$ be a Legendre function—see, for example, (Cesa-Bianchi and Lugosi, 2006, Chapter 11.2) for a definition—with a positive definite Hessian on $\mathrm{dom}(F_t)$. Throughout the paper, $\mathrm{dom}(F_t)$ is the positive cone. The motivation for using a hybrid regularizer comes from Zimmert and Seldin (2020), who show how to obtain optimal bounds in the nonstochastic bandit setting by using a regularizer to control the variance of the loss estimates and a different regularizer to control the impact of delay.

As discussed by György and Joulani (2021), the challenge for most delayed (bandit) algorithms is to control the drift in terms of $\boldsymbol{p}_t^{av}$. The following Lemma, whose proof can be found in Appendix A, provides a general bound on the drift when the Hessian of R_t at $\boldsymbol{p}_t^{av}$ satisfies some mild conditions. In the following, we use $\boldsymbol{u} \leq \boldsymbol{v}$ to denote $u_i \leq v_i$ for all i and we use $\boldsymbol{u} < \boldsymbol{v}$ if $\boldsymbol{u} \leq \boldsymbol{v}$ and $u_i < v_i$ for at least one i. We define $G_t(\boldsymbol{y}, \boldsymbol{z}) = \langle \nabla F_t^\star(\boldsymbol{y}), \boldsymbol{z}\rangle$ and $W_t(\boldsymbol{x}, \boldsymbol{z}) = \langle \nabla F_t(\boldsymbol{x}), \boldsymbol{z}\rangle$ for any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

Lemma 1. *For any $\boldsymbol{z} > \boldsymbol{0}$, if $W_t(\cdot, \boldsymbol{z})$ is concave, $W_t(\boldsymbol{x}_1, \boldsymbol{z}) \leq W_t(\boldsymbol{x}_2, \boldsymbol{z})$ iff $\boldsymbol{x}_1 \leq \boldsymbol{x}_2$, and $W_t(\boldsymbol{x}_1, \boldsymbol{1}) < W_t(\boldsymbol{x}_2, \boldsymbol{1})$ iff $\boldsymbol{x}_1 < \boldsymbol{x}_2$, then*

$$\langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle \leq (\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t$$

Denote by $\mathcal{S}_t(i') = \{s : s + d_s(i') < t\}$ the indices of available losses for arm i' at the beginning of round t, and by $m_t(i') = \{s : s < t, s + d_s(i') \geq t\}$ the set of missing losses for arm i' at the beginning of round t, i.e., the losses that in a non-delayed full information setting would have been available to the learner for prediction at round t but are not available due to delay. Note that $\rho_t(i') = |m_t(i')|$. The quantity $\widehat{L}_t(i') - \widehat{L}_t^{av}(i')$ is a central quantity and is given by

$$\begin{aligned} \widehat{L}_t(i') - \widehat{L}_t^{av}(i') &= \sum_{s \leq t} \widehat{\ell}_s(i') - \sum_{s \in \mathcal{S}_t(i')} \widehat{\ell}_s(i') \\ &= \sum_{s \in [t] \setminus \mathcal{S}_t(i')} \widehat{\ell}_s(i') = \widehat{\ell}_t(i') + \sum_{s \in m_t(i')} \widehat{\ell}_s(i') \end{aligned} \tag{6}$$

The challenge and the main difference with respect to the standard delayed setting comes from the $m_t(i')$ term. While in the standard delayed setting the $m_t(i')$ term is the same for all arms, in our setting the set of missing losses can be different for each arm. This challenge is addressed in the next sections. To bound the cheating regret we will use the following standard lemma—see, for example, (Joulani et al., 2020):

Lemma 2. *Suppose $R_t(\boldsymbol{p}) \geq R_{t-1}(\boldsymbol{p})$ for all $t \geq 1$ and $\boldsymbol{p} \in \triangle'$, where $R_0 = R_1$. Let $\boldsymbol{p}_0 = \arg\min_{\boldsymbol{p} \in \triangle'} R_1(\boldsymbol{p})$. Then, for any $\boldsymbol{u} \in \triangle'$ and for $\boldsymbol{p}_{t+1}$ defined in (4), $\sum_{t=1}^T \langle \boldsymbol{p}_{t+1} - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t\rangle \leq R_T(\boldsymbol{u})$.*

For completeness we provide the proof of Lemma 2 in Appendix A. Since $\widehat{\boldsymbol{\ell}}_t = \widetilde{\boldsymbol{\ell}}_t + \boldsymbol{a}_t$, to recover a regret bound with respect to $\widetilde{\boldsymbol{\ell}}_t$ we simply subtract $\sum_{t=1}^T \langle \boldsymbol{p}_t^{av} - \boldsymbol{u}, \boldsymbol{a}_t\rangle$ from both sides of (3):

$$\begin{aligned} \sum_{t=1}^T \langle \boldsymbol{p}_t^{av} - \boldsymbol{u}, \widetilde{\boldsymbol{\ell}}_t\rangle &= \sum_{t=1}^T \langle \boldsymbol{u}, \boldsymbol{a}_t\rangle + \sum_{t=1}^T \langle \boldsymbol{p}_{t+1} - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t\rangle \\ &\quad + \sum_{t=1}^T \langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle - \sum_{t=1}^T \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t\rangle \end{aligned} \tag{7}$$

The $\sum_{t=1}^T \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t\rangle$ term is used to control the drift. The $\sum_{t=1}^T \langle \boldsymbol{u}, \boldsymbol{a}_t\rangle$ term is what allows us to derive bounds that depend on $\boldsymbol{u}$. Applying Lemmas 1 and 2 to (7), we can prove the following result.

Lemma 3. *Under the assumptions of Lemmas 1 and 2, Algorithm 1 satisfies:*

$$\begin{aligned} \sum_{t=1}^T \langle \boldsymbol{p}_t^{av} - \boldsymbol{u}, \widetilde{\boldsymbol{\ell}}_t\rangle &\leq \sum_{t=1}^T \langle \boldsymbol{u}, \boldsymbol{a}_t\rangle + R_T(\boldsymbol{u}) \\ &+ \sum_{t=1}^T \Big((\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t - \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t\rangle \Big) \end{aligned}$$

The remaining challenge is to choose Ψ_t, Φ_t, and a_t.

3 FULL INFORMATION SETTING

In the full information setting we use $\Phi_t \equiv 0$ and

$$\Psi_t(\boldsymbol{p}) = \sum_{i=1}^K \sum_{\eta \in H_t} \frac{p(i, \eta)}{\eta} \ln(p(i, \eta)) \tag{8}$$

where H_t is a multiset of learning rates we define below. Similar regularizers have been used by Bubeck et al. (2017) and Chen et al. (2021) to obtain multiscale algorithms. In particular, the algorithm of Chen et al. (2021) is related to ours, as they also use corrections. However, we are the first ones to analyze this regularizer in a delayed feedback setting.

As mentioned already, we simultaneously learn to compete with arbitrary $\boldsymbol{u} \in \triangle$ and learn the best learning rate. We do so by using several learning rates for each expert. Denote by H_t the multiset of learning rates available in round t and let $i' = (i, \eta)$. We have that $|H_t| = J$ for all t, where we set $J = \lceil \log_2(\sqrt{T}) \rceil$. Then $\boldsymbol{p}_t^{av}$ is a distribution over $[K] \times [J]$. The probability the algorithm assigns to expert i is the marginal:

$$q_t(i) = \textstyle\sum_{\eta \in H_t} p_t^{av}(i, \eta) \tag{9}$$

The two missing inputs to Algorithm 1 are $\widetilde{\ell}_t(i')$ and $a_t(i')$. In the full information setting we do not need to estimate $\ell_t(i)$, and can simply set $\widetilde{\ell}_t(i') = \widetilde{\ell}_t(i, \eta) = \ell_t(i)$. As we mentioned above, the correction term is chosen to control the drift

$$a_t(i') = a_t(i, \eta) = 4\eta \ell_t(i)(1 + \rho_t(i)) \tag{10}$$

The multiset H_t of learning rates in round t contains

$$\min\left\{\frac{1}{4(1 + \rho_t^{\max})}, \frac{\sqrt{\ln(K) + 2(\ln(T) + 1)}}{4\sqrt{1 + \rho_t^{\max}} 2^j}\right\} \tag{11}$$

for all $j = 1, \ldots, J$, where $\rho_t^{\max} = \max_{i \leq K, s \leq t} \rho_s(i)$. Hence for each j there is a (not necessarily distinct) $\eta \in H_t$, which we use to define the prior:

$$p_1^\Psi(i') = p_1^\Psi(i, \eta) = \frac{2^{-2j}}{\sum_{j'=1}^J 2^{-2j'}} \pi(i) \tag{12}$$

Note that this is similar to the prior used by Chen et al. (2021). Next, we state the main result of this section in asymptotic form (see Theorem 4 in Appendix B for the same bound with explicit constants).

Theorem 1. *If we run Algorithm 1 with regularizer $\Phi_t \equiv 0$, Ψ_t defined by* (8)*, corrections a_t defined by* (10)*, and prior defined by* (12)*, then for any $\boldsymbol{u} \in \triangle$ the predictions $\boldsymbol{q}_t$ defined by* (9) *using $\boldsymbol{p}_t^{av}$ computed by the algorithm satisfy*

$$\sum\nolimits_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle$$
$$= O\left(\rho_T^{\max} + \sqrt{\big(\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T)\big)\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^{\rho} \rangle}\right)$$

Sketch proof of Theorem 1. Denote by $\eta^\star \in H_T$ the target learning rate and let $u'(i') = u'(i, \eta) = 0$ if $\eta \neq \eta^\star$ and $u'(i, \eta) = u(i)$ otherwise. First, observe that by definition of $\boldsymbol{q}_t$, $\boldsymbol{u}'$, and $\widetilde{\boldsymbol{\ell}}$ we have that $\sum_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle = \sum_{t=1}^{T} \langle \boldsymbol{p}_t^{av} - \boldsymbol{u}', \widetilde{\boldsymbol{\ell}}_t \rangle$, where the inner product on the left-hand side is on $\mathbb{R}^K$ and the inner products on the right-hand side are on $\mathbb{R}^{K \times J}$. Since R_t can be shown to verify the conditions of Lemma 3, we obtain

$$\sum\nolimits_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle \leq R_T(\boldsymbol{u}') + \sum\nolimits_{t=1}^{T} \langle \boldsymbol{u}', \boldsymbol{a}_t \rangle$$
$$+ \sum\nolimits_{t=1}^{T} \left((\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t - \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t \rangle \right)$$

With some work we may then prove a bound on $R_T(u')$ of order $O\big((\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T))/\eta^\star\big)$. Using that $\widehat{\ell}_t(i') \leq 2\widetilde{\ell}_t(i') = 2\ell_t(i)$ due to the restrictions on η, and that $\nabla^2 R_t(\boldsymbol{p}_t^{av})$ is a diagonal matrix with components equal to $1/\eta p_t^{av}(i, \eta)$, we can see that

$$(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t$$
$$\leq \sum\nolimits_{i=1}^{K} \sum\nolimits_{\eta \in H_t} 4\eta p_t^{av}(i, \eta) \widetilde{\ell}_t(i, \eta) \big(1 + \rho_t(i)\big)$$

Thus, by using the definition of $\boldsymbol{u}'$ and that $a_t(i, \eta) = 4\eta \widetilde{\ell}_t(i, \eta)(1 + \rho_t(i))$, we derive a bound on $\sum_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle$ of order

$$\min_{\eta \in H_T} \left(\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T)}{\eta} + \eta \sum_{t=1}^{T} \ell_t(i)(1 + \rho_t(i)) \right)$$

To complete the proof one has to show that there is a suitable $\eta^\star \in H_T$ approximating the minimizer of the above expression (see Appendix B). ☐

Interestingly, the result in Theorem 1 holds for all $\boldsymbol{u}$ simultaneously, which implies that the algorithm automatically performs a trade-off between losses and delays. This interplay between losses and delays also occurs when we run the algorithm: the loss $\widehat{\ell}_t(i, \eta) = \ell_t(i)\big(1 + 4\eta(1 + \rho_t(i))\big)$ we feed to the algorithm penalizes arms for suffering large delay, as when the delay is large the amount of missing feedback is also large.

4 BANDIT SETTING

In the bandit setting we face two additional challenges compared to the full information setting. The first challenge is to estimate $\boldsymbol{\ell}_t$. For this we use the implicit exploration technique (Neu, 2015a): sample $i_t \sim \boldsymbol{q}_t$, observe $\ell_t(i_t)$, and estimate $\ell_t(i)$ for any i by $\mathbb{1}[i_t = i](q_t(i) + \epsilon_t)^{-1} \ell_t(i)$, where $\epsilon_t \geq 0$ is a user-specified parameter which we will either set to 0 or to $O(t^{-1/2})$.

The second challenge is due to arm-dependent delays. In Section 3 we used $\rho_t(i)$ in the correction term. While ideally we would like to do the same in the bandit setting, we do not know how many observations are missing from each arm: if we do *not* pull arm i in round t we will not know if $\ell_t(i)$ is missing in subsequent rounds. Recall that in the partially concealed bandit setting we assume that the learner observes $\rho_t(i_t)$ whenever the loss $\ell_t(i_t)$ is observed. Note that in the arm-independent delay setting, $\rho_t(i)$ can always be computed by simply counting the number of missing losses. In the concealed bandit setting, we do not make this additional assumption. The fact that we do not know $\rho_t(i_t)$ at prediction time complicates learning $\rho_T^{\max}$. While in the full information setting we could adjust the grid of learning rates *before* issuing a prediction, we can not use this trick in the bandit setting. Instead, we use a fixed grid of learning rates. For this reason we assume that the learner has preliminary access to an upper bound on $\rho_T^{\max}$, denoted by $\rho^\star$.

4.1 Partially Concealed Bandit Setting

As in the full information setting, we simultaneously learn the best learning rate and expert. For each expert i we create J pseudo-experts, each having its own personal learning rate $\gamma \in \Gamma$. Thus, as in the full information setting, we write $p(i') = p(i, \gamma)$. Unlike before, we use Algorithm 1 with a double regularizer:

$$\Psi_t(\boldsymbol{p}) = -\eta_t^{-1} \sum\nolimits_{i=1}^{K} \ln\Big(\sum\nolimits_{\gamma \in \Gamma} p(i, \gamma)\Big) \tag{13}$$

$$\Phi_t(\boldsymbol{p}) = \sum\nolimits_{i=1}^{K} \sum\nolimits_{\gamma \in \Gamma} \gamma^{-1} p(i, \gamma) \ln\big(p(i, \gamma)\big) \tag{14}$$

The role of (13) is to control the variance $\mathbb{E}\left[\widehat{\boldsymbol{\ell}}_t^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t\right]$ and (14) is used to control $\mathbb{E}\left[\big(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{\ell}}_t - \widehat{\boldsymbol{L}}_t^{av}\big)^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t\right]$, a quantity associated with the additional regret due to delayed feedback. The predictions i_t are sampled from $\boldsymbol{q}_t$ given by

$$q_t(i) = \sum\nolimits_{\gamma \in \Gamma} p_t^{av}(i, \gamma) \tag{15}$$

Our regularizer has a somewhat unusual structure. The log barrier regularizer (13) controls the variance of the loss estimates, and regularizes $q_t(i)$ rather than

$p_t(i,\gamma)$. This implies that, as far as the log barrier is concerned, there are only K arms rather than $K' = O\big(K\ln(T)\big)$ arms. To develop some intuition as to why this matters, let us consider the regret bound of FTRL run with the log barrier, a constant learning rate η, and N arms. Without delays, this bound is $O\big(N\ln(T)\eta^{-1} + \eta T\big)$ (Foster et al., 2016). Therefore, using K rather than K' experts reduces the regret of FTRL with the log barrier by a factor of $\sqrt{\ln(T)}$.

We continue by specifying loss estimates $\widetilde{\ell}_t(i') = \widetilde{\ell}_t(i,\gamma) = \mathbb{1}[i_t = i]q_t(i)^{-1}\ell_t(i)$ (i.e., we set $\epsilon_t = 0$) and the corrections

$$a_t(i,\gamma) = 4\gamma\widetilde{\ell}_t(i,\gamma)\rho_t(i) \tag{16}$$

Recall that in the partially concealed bandit setting the learner observes $\rho(i_t)$ whenever the loss $\ell_t(i_t)$ is observed, making (16) a valid choice for corrections. The grid Γ of learning rates we use in this section contains all

$$\min\left\{(4\rho^\star)^{-1}, \frac{\sqrt{\ln(K)+\ln(T)+1}}{4\sqrt{\rho^\star}2^j}\right\} \tag{17}$$

for $j = 1,\ldots,J = \lceil\log_2(\sqrt{T})\rceil$, where $\rho^\star \geq \rho_T^{\max}$. The learning rate for the log barrier is

$$\eta_t = \sqrt{K\ln(T)\left(4(1+\rho^\star) + 4\sum\nolimits_{s\in\mathcal{S}_t}\ell_s(i_s)\right)^{-1}} \tag{18}$$

where $\mathcal{S}_t = \{s : s + d_s(i_s) < t\}$ is the set of available losses at the beginning of round t in the bandit setting.

Theorem 2. *Let $\boldsymbol{p}_1^\Psi \equiv \frac{1}{KJ}$ and let $\boldsymbol{p}_1^\Phi$ be given by (12). Assume $\pi_1(i) \geq \frac{1}{T^2}$ for all $i \in [K]$. If we run Algorithm 1 using regularizers (13) and (14) with corresponding learning rates (17) and (18), then the predictions $i_t \sim \boldsymbol{q}_t$, with $\boldsymbol{q}_t$ as in (15), satisfy*

$$\mathbb{E}\left[\sum_{t=1}^T(\ell_t(i_t) - \langle\boldsymbol{u},\boldsymbol{\ell}_t\rangle)\right] = O\Bigg(K\ln(T) + \rho^\star\ln(T) \sqrt{K\ln(T)\langle\boldsymbol{u},\boldsymbol{L}_T\rangle} + \sqrt{\big(\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)\big)\langle\boldsymbol{u},\boldsymbol{L}_T^\rho\rangle}\Bigg)$$

The proof of Theorem 2 follows from Theorem 5 in Appendix C. As in the full information setting, we use Lemma 3 to bound the regret. The main difference with the full information setting is how $\mathbb{E}\left[(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top\big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t\right]$ is bounded. The unusual structure of the regularizer presents us with a challenge here, as $\nabla^2 R_t(\boldsymbol{p}_t^{av})$ is a complicated block-diagonal matrix. We carefully analyse the inverse of $\nabla^2 R_t(\boldsymbol{p}_t^{av})$ in Lemmas 4 and 5:

Lemma 4. *For all t, the function $F_t = \Psi_t + \Phi_t$, where Ψ_t is defined in (13) and Φ_t is defined in (14), is Legendre and satisfies the conditions of Lemma 1. Moreover,*

$$(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top\big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t \leq \eta_t 4\ell_t(i_t)^2 + 4\sum\nolimits_{\gamma\in\Gamma}\gamma\sum\nolimits_{s\in m_t(i_t)}p_t^{av}(i_t,\gamma)\widetilde{\ell}_t(i_t,\gamma)\widetilde{\ell}_s(i_t,\gamma)$$

Lemma 5. *With regularizers (13) and (14) we have that*

$$\mathbb{E}\left[(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top\big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t\right] - \mathbb{E}\left[\langle\boldsymbol{p}_t^{av},\boldsymbol{a}_t\rangle\right] \leq \mathbb{E}\left[\eta_t 4\ell_t(i_t)^2\right]$$

Proof. Since $\nabla^2 R_t(\boldsymbol{p}_t^{av}) = \nabla^2 F_t(\boldsymbol{p}_t^{av})$ and $\sum_{s\in m_t(i_t)}\widetilde{\ell}_t(i_t)f_s = \sum_{i=1}^K\sum_{s=1}^T\mathbb{1}[s\in m_t(i)]\widetilde{\ell}_t(i)f_s$ for arbitrary f_s we can use Lemma 4 to show that

$$\mathbb{E}\left[(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top\big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t\right] \leq \mathbb{E}\left[\eta_t 4\ell_t(i_t)^2\right] + \mathbb{E}\left[4\sum_{\gamma\in\Gamma}\sum_{i=1}^K\sum_{s=1}^T\mathbb{1}[s\in m_t(i)]p_t^{av}(i,\gamma)\gamma\widetilde{\ell}_t(i,\gamma)\widetilde{\ell}_s(i,\gamma)\right] \tag{19}$$

We proceed by studying the expectation of the sum in the above equation. Denote by $\mathcal{B}_s \equiv \{i_r : r\in\mathcal{S}_s\}$ all i_r with $r \in \mathcal{S}_s \equiv \{r : r + d_r(i_r) < s\}$. Recall that $m_t(i) \equiv \{s : s < t, s + d_s(i) \geq t\}$ identifies the set of losses that can not be used for prediction in round t due to delay. We may write

$$\begin{aligned}
&\mathbb{E}\left[p_t^{av}(i,\gamma)\widetilde{\ell}_t(i,\gamma)\mathbb{1}[s\in m_t(i)]\widetilde{\ell}_s(i,\gamma)\middle|\mathcal{B}_s\right] \leq \\
&\mathbb{E}\left[p_t^{av}(i,\gamma)\widetilde{\ell}_t(i,\gamma)\frac{\mathbb{1}[i=i_s]}{q_s(i)}\middle|\mathcal{B}_s\right]\mathbb{1}[s\in m_t(i)] = \qquad (20)\\
&\mathbb{E}\left[p_t^{av}(i,\gamma)\widetilde{\ell}_t(i,\gamma)\middle|\mathcal{B}_s\right]\mathbb{E}\left[\frac{\mathbb{1}[i=i_s]}{q_s(i)}\middle|\mathcal{B}_s\right]\mathbb{1}[s\in m_t(i)]\\
&= \mathbb{E}\left[p_t^{av}(i,\gamma)\widetilde{\ell}_t(i,\gamma)\middle|\mathcal{B}_s\right]\mathbb{1}[s\in m_t(i)]
\end{aligned}$$

The first step is true because $s \in m_t(i)$ is a deterministic event for all $s \leq t$ and i and simply uses $\ell_s(i) \leq 1$. Equality (20) vacuously holds when $s \notin m_t(i)$. If $s \in m_t(i)$, then $\widetilde{\ell}_s(i,\gamma)$ is not used for computing the prediction in round t because it is not available at the beginning of round t. Hence, when $i = i_s$, $p_t^{av}(i,\gamma)\widetilde{\ell}_t(i,\gamma)$ does not depend on i_s and (20) holds; if $i \neq i_s$ then both sides of the equality are again zero and (20) vacuously holds. The above implies that we can further bound equation (19) as

$$\mathbb{E}\left[(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top\big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t\right] \leq \mathbb{E}\left[\eta_t 4\ell_t(i_t)^2\right] + \mathbb{E}\left[4\sum\nolimits_{\gamma\in\Gamma}p_t^{av}(i_t,\gamma)\gamma\widetilde{\ell}_t(i_t,\gamma)\rho_t(i_t)\right]$$

which completes the proof after we subtract

$$\mathbb{E}\left[\langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t\rangle\right] = \mathbb{E}\left[4\sum_{\gamma\in\Gamma} p_t^{av}(i_t,\gamma)\gamma\widetilde{\ell}_t(i_t,\gamma)\rho_t(i_t)\right] \quad \square$$

Lemma 5 combined with the bound on $R_T(\boldsymbol{u}')$ in Lemma 7 and showing that there is a $\gamma \in \Gamma$ that is close to optimal are the essential parts of the proof of Theorem 5, which can be found in Appendix C.

4.2 Concealed Bandit Setting

In this section we design an algorithm that does not require knowledge of $\rho_t(i)$ for any arm. To see why not knowing $\rho_t(i)$ poses a problem for standard algorithms, we analyse the optimal algorithm of Zimmert and Seldin (2020). This corresponds to Algorithm 1 with the following setup: The algorithm uses $i' = i$, which is to say that it samples its actions from $q_t(i) = p_t^{av}(i)$ where $p_t^{av}(i)$ are computed by Algorithm 1. Furthermore, it uses the following regularizers:

$$\Psi_t(\boldsymbol{p}) = \sum_{i=1}^K -\eta_t^{-1}\sqrt{p(i)} \tag{21}$$

$$\Phi_t(\boldsymbol{p}) = \sum_{i=1}^K \gamma_t^{-1}p(i)\ln\big(p(i)\big) \tag{22}$$

Zimmert and Seldin (2020) use $a_t(i) = 0$ and we postpone specifying the loss estimates until later. Recall that $F_t = \Psi_t + \Phi_t$ and observe that $\big(\nabla F_t(\boldsymbol{x})\big)_i = -\frac{1}{2\eta_t\sqrt{x(i)}} + \frac{\ln(x(i))+1}{\gamma_t}$ and that $\nabla^2 F_t(\boldsymbol{x})$ is a diagonal matrix with $\big(\nabla^2 F_t(\boldsymbol{x})\big)_{i,i} = -\frac{1}{4\eta_t x(i)^{3/2}} + \frac{1}{\gamma_t x(i)}$. As the conditions of Lemma 3 are satisfied, we may write:

$$\mathbb{E}\left[\sum_{t=1}^T\langle \boldsymbol{q}_t - \boldsymbol{u}, \widetilde{\boldsymbol{\ell}}_t\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^K 4\eta_t q_t(i)^{3/2}\widetilde{\ell}_t(i)^2\right] + \mathbb{E}\left[R_T(\boldsymbol{u})\right] + \mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^K \gamma_t q_t(i)\widetilde{\ell}_t(i)\sum_{s\in m_t(i)}\widetilde{\ell}_s(i)\right] \tag{23}$$

where we used that $\left(\frac{1}{4\eta_t q_t(i)^{3/2}} + \frac{1}{\gamma_t q_t(i)}\right)^{-1} \le \min\{4\eta_t q_t(i)^{3/2}, \gamma_t q_t(i)\}$. Following the proof of Lemma 5 to bound the expectation, we see that

$$\mathbb{E}\left[\sum_{i=1}^K q_t(i)\widetilde{\ell}_t(i)\sum_{s\in m_t(i)}\widetilde{\ell}_s(i)\right] \le \mathbb{E}\left[\sum_{i=1}^K q_t(i)\rho_t(i)\right]$$

Since Zimmert and Seldin (2020) assume that the delays are equal for each arm, implying that $\rho_t(i) = \rho_t$ for all arms, they can use an unbiased loss estimator combined with $\gamma_t = \sqrt{\frac{\ln(K)}{\sum_{s=1}^t \rho_t}}$ to achieve the optimal regret bound. In an ideal scenario, in our setting we would set $\gamma_t = \sqrt{\frac{\ln(K)}{\sum_{s=1}^t\sum_{i=1}^K \gamma_t q_t(i)\rho_t(i)}}$. Unfortunately, we do not know $\rho_t(i)$, meaning we have to resort to tuning γ_t in terms of (a suitable upper bound on) $\sum_{i=1}^K \gamma_t q_t(i)\widetilde{\ell}_t(i)\sum_{s\in m_t(i)}\widetilde{\ell}_s(i)$. This brings us an additional challenge in the form of having to control the loss estimates, which we do by using the implicit exploration of Neu (2015a), as is also done by György and Joulani (2021), i.e., we set $\widetilde{\ell}_t(i) = \mathbb{1}[i_t = i]\ell_t(i)(q_t(i)+\epsilon_t)^{-1}$. The role of ϵ_t is to control the range of $\widetilde{\ell}(i)$, which is $[0, \epsilon_t^{-1}]$. The learning rate for Φ_t will be set as

$$\gamma_t = \sqrt{\frac{\ln(K)}{\frac{\rho^\star}{\epsilon_t} + \sum_{s=1}^{t-1}\sum_{i=1}^K \mathbb{1}[i_s = i]\sum_{s'\in m_s(i)}\frac{\mathbb{1}[i_{s'}=i]}{q_{s'}(i)}}} \tag{24}$$

Note that while we do not know $m_s(i) \equiv \{s : s < t, s + d_s(i) > t\}$, the set of missing losses for arm i at time t, we can indeed compute $\sum_{s\in m_s(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}$, as we only add $\frac{1}{q_s(i)}$ to the sum whenever we pulled arm i in round s and have not yet observed $\ell_s(i_s)$ by round t, which we do know. The final result can be found in Theorem 3, of which the proof can be found in Appendix D.

Theorem 3. *Let $\boldsymbol{q}_t$ be computed by Algorithm 1 with $a_t(i) = 0$, $\eta_t = \frac{1}{\sqrt{4t}}$, $\epsilon_t = \frac{1}{\sqrt{t}}$, γ_t as in (24), $p_1^\Psi(i) = p_1^\Phi(i) = \frac{1}{K}$, $\widetilde{\ell}_t(i) = \frac{\mathbb{1}[i_t=i]\ell_t(i)}{q_t(i)+\epsilon_t}$, and regularizers specified in (21) and (22). Then actions $i_t \sim \boldsymbol{q}_t$ guarantee*

$$\mathbb{E}\left[\sum_{t=1}^T(\ell_t(i_t) - \langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right]) \le 9\sqrt{KT} + \tfrac{1}{2}\rho^\star + 3\sqrt{\ln(K)\sum_{t=1}^T\mathbb{E}\left[\sum_{i=1}^K q_t(i)\rho_t(i)\right]}$$

5 FUTURE WORK

Our analysis of Lemma 3 (and of Lemma 1 in particular) hinges on the non-negativity of the losses. While for first-order bounds non-negativity is a standard assumption, for other types of bounds a typical assumption is $\ell_t(i) \in [-1, 1]$. Since most works in the delayed feedback setting also assume and use non-negativity of the losses, we think that deriving an algorithm and a corresponding analysis without relying on the non-negativity of the losses would be an interesting contribution. Another interesting direction to pursue is developing a skipping procedure for algorithms that are adaptive to both data and delay. While several authors have proposed different procedures for skipping rounds with a large delay (Thune et al., 2019; Zimmert and Seldin, 2020; György and Joulani, 2021), it is not clear how to apply these techniques to the arm-dependent delay setting. With regard to a skipping technique for data and delay adaptive algorithms, as György and Joulani (2021) note, applying standard skipping techniques induces a complicated dependence on past actions, which significantly complicates the analysis.

Acknowledgements

Van der Hoeven and Cesa-Bianchi are supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and by the EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence).

References

Agarwal, A. and Duchi, J. C. (2012). Distributed delayed stochastic optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5451–5452.

Allenberg, C., Auer, P., Györfi, L., and Ottucsák, G. (2006). Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 229–243.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. (2021). No discounted-regret learning in adversarial bandits with delays. *arXiv preprint arXiv:2103.04550*.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. H. (2019). Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pages 11345–11354.

Bubeck, S., Devanur, N. R., Huang, Z., and Niazadeh, R. (2017). Online auctions and multi-scale online learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 497–514.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2018). Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pages 750–773.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2019). Delay and cooperation in nonstochastic bandits. *J. Mach. Learn. Res.*, 20:17:1–17:38.

Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. (2016). Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

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

Chapelle, O. (2014). Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in Neural Information Processing Systems*, pages 2249–2257.

Chen, L., Luo, H., and Wei, C. (2021). Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pages 1216–1259.

Desautels, T., Krause, A., and Burdick, J. W. (2014). Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923.

Dudík, M., Hsu, D. J., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178.

Van Erven, T., Kotłowski, W., and Warmuth, M. K. (2014). Follow the leader with dropout perturbations. In *Conference on Learning Theory*, pages 949–974.

Flaspohler, G., Orabona, F., Cohen, J., Mouatadid, S., Oprescu, M., Orenstein, P., and Mackey, L. (2021). Online learning with optimism and delay. In *International Conference on Machine Learning*, volume 139, pages 3363–3373.

Foster, D. J., Li, Z., Lykouris, T., Sridharan, K., and Tardos, E. (2016). Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, pages 4734–4742.

Gael, M. A., Vernade, C., Carpentier, A., and Valko, M. (2020). Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356.

György, A. and Joulani, P. (2021). Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pages 3988–3997.

Hazan, E. and Kale, S. (2010). Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188.

Van der Hoeven, D., Hadiji, H., and Van Erven, T. (2021). Distributed online learning for joint regret with communication constraints. *arXiv preprint arXiv:2102.07521*.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2020). Multi-agent online optimization with delays: Asynchronicity, adaptivity, and optimism. *arXiv preprint arXiv:2012.11579*.

Hutter, M. and Poland, J. (2004). Prediction with expert advice by following the perturbed leader for general weights. In *International Conference on Algorithmic Learning Theory*, pages 279–293.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461.

Joulani, P., György, A., and Szepesvári, C. (2016). Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 1744–1750.

Joulani, P., György, A., and Szepesvári, C. (2020). A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808:108–138.

Kalai, A. and Vempala, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. (2021). Stochastic multi-armed bandits with unrestricted delay distributions. *arXiv preprint arXiv:2106.02436*.

Mandel, T., Liu, Y.-E., Brunskill, E., and Popović, Z. (2015). The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

McMahan, B. and Streeter, M. (2014). Delay-tolerant algorithms for asynchronous distributed online learning. *Advances in Neural Information Processing Systems*, pages 2915–2923.

Neu, G. (2015a). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, pages 3168–3176.

Neu, G. (2015b). First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375.

Neu, G., György, A., Szepesvári, C., and Antos, A. (2010). Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*.

Neu, G., György, A., Szepesvári, C., and Antos, A. (2014). Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. (2018). Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113.

Quanrud, K. and Khashabi, D. (2015). Online learning with adversarial delays. In *Advances in Neural Information Processing Systems*, pages 1270–1278.

Sra, S., Yu, A. W., Li, M., and Smola, A. J. (2015). Adadelay: Delay adaptive distributed stochastic convex optimization. *arXiv preprint arXiv:1508.05003*.

Steinhardt, J. and Liang, P. (2014). Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pages 1593–1601.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. (2019). Nonstochastic multiarmed bandits with unrestricted delays. *Advances in Neural Information Processing Systems*, pages 6541–6550.

Vernade, C., Cappé, O., and Perchet, V. (2017). Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*.

Vernade, C., Carpentier, A., Zappella, G., Ermis, B., and Brueckner, M. (2018). Contextual bandits under delayed feedback. *arXiv preprint arXiv:1807.02089*.

Wei, C.-Y. and Luo, H. (2018). More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291.

Weinberger, M. J. and Ordentlich, E. (2002). On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976.

Zimmert, J. and Seldin, Y. (2020). An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3285–3294.

Zinkevich, M., Smola, A. J., and Langford, J. (2009). Slow learners are fast. In *Advances in Neural Information Processing Systems*, pages 2331–2339.

A DETAILS OF SECTION 2

We start with some definitions and a technical Lemma. Let $\boldsymbol{L}_0^t = \nabla\Psi_t(\boldsymbol{p}_1^\Psi) + \nabla\Phi_t(\boldsymbol{p}_1^\Phi)$ and

$$\overset{\triangle}{F}{}_t^\star = \min_{\boldsymbol{p}\in\triangle'} \langle \boldsymbol{p}, \cdot\rangle + F_t(\boldsymbol{p})$$

Lemma 6. *Under the assumptions of Lemma 1, for any* $\boldsymbol{p}^{\boldsymbol{L}} = \arg\min_{\boldsymbol{p}\in\triangle'}\langle \boldsymbol{p}, \boldsymbol{L}\rangle + R_t(\boldsymbol{p})$ *there exists a* $c \in \mathbb{R}$ *such that*

$$\boldsymbol{p}^{\boldsymbol{L}} = \nabla \overset{\triangle}{F}{}_t^\star(-\boldsymbol{L} - \boldsymbol{L}_0^t) = \nabla F_t^\star(-\boldsymbol{L} - \boldsymbol{L}_0^t + c\mathbf{1})$$

Furthermore, for any vector $\boldsymbol{z} > \mathbf{0}$*,* $G_t(\cdot, \boldsymbol{z})$ *is a convex function,* $G_t(\boldsymbol{y}_1, \boldsymbol{z}) \leq G_t(\boldsymbol{y}_2, \boldsymbol{z})$ *if* $\boldsymbol{y}_1 \leq \boldsymbol{y}_2$*, and* $G_t(\boldsymbol{y}_1, \boldsymbol{z}) < G_t(\boldsymbol{y}_2, \boldsymbol{z})$ *if* $\boldsymbol{y}_1 < \boldsymbol{y}_2$*.*

Proof. Recall that $F_t = \Psi_t + \Phi_t$. To prove that $\boldsymbol{p}^{\boldsymbol{L}} = \nabla \overset{\triangle}{F}{}_t^\star(-\boldsymbol{L} - \boldsymbol{L}_0) = \nabla F_t^\star(-\boldsymbol{L} - \boldsymbol{L}_0^t + c\mathbf{1})$, observe that we may rewrite

$$\begin{aligned}
\boldsymbol{p}^{\boldsymbol{L}} &= \arg\min_{\boldsymbol{p}\in\triangle'} \langle \boldsymbol{p}, \boldsymbol{L}\rangle + R_t(\boldsymbol{p}) \\
&= \arg\min_{\boldsymbol{p}\in\triangle'} \langle \boldsymbol{p}, \boldsymbol{L}\rangle + F_t(\boldsymbol{p}) - \Psi_t(\boldsymbol{p}_1^\Psi) - \Phi_t(\boldsymbol{p}_1^\Phi) \\
&- \langle \nabla\Psi_t(\boldsymbol{p}_1^\Psi), \boldsymbol{p} - \boldsymbol{p}_1^\Psi\rangle - \langle \nabla\Phi_t(\boldsymbol{p}_1^\Phi), \boldsymbol{p} - \boldsymbol{p}_1^\Phi\rangle \\
&= \arg\min_{\boldsymbol{p}\in\triangle'} \langle \boldsymbol{p}, \boldsymbol{L} + \boldsymbol{L}_0^t\rangle + F_t(\boldsymbol{p})
\end{aligned}$$

Following Zimmert and Seldin (2020, Fact 3), by the KKT conditions there exists a $c \in \mathbb{R}$ such that $\boldsymbol{p}^{\boldsymbol{L}}$ satisfies $\nabla F_t(\boldsymbol{p}^{\boldsymbol{L}}) = -\boldsymbol{L} - \boldsymbol{L}_0^t + c\mathbf{1}$. Using that $\nabla F_t = (\nabla F_t^\star)^{-1}$, due to F_t being Legendre, completes the proof of the first property.

In order to prove the second part, suppose now $\boldsymbol{y}_1 < \boldsymbol{y}_2$. Since F_t is Legendre, $\boldsymbol{y}_1 = \nabla F_t\big(\nabla F_t^\star(\boldsymbol{y}_1)\big)$ and we have that

$$\langle \nabla F_t(\nabla F_t^\star(\boldsymbol{y}_1)), \mathbf{1}\rangle < \langle \nabla F_t(\nabla F_t^\star(\boldsymbol{y}_2)), \mathbf{1}\rangle$$

Since $\langle \nabla F_t(\boldsymbol{x}_1), \mathbf{1}\rangle < \langle \nabla F_t(\boldsymbol{x}_2), \mathbf{1}\rangle$ iff $\boldsymbol{x}_1 < \boldsymbol{x}_2$ holds by hypothesis, the above implies $\nabla F_t^\star(\boldsymbol{y}_1) < \nabla F_t^\star(\boldsymbol{y}_2)$, which is equivalent to $G_t(\boldsymbol{y}_1, \mathbf{1}) < G_t(\boldsymbol{y}_2, \mathbf{1})$ if $\boldsymbol{y}_1 < \boldsymbol{y}_2$. Using the same proof technique, replacing the strict inequality with an inequality, one can prove that $G_t(\boldsymbol{y}_1, \boldsymbol{z}) \leq G_t(\boldsymbol{y}_2, \boldsymbol{z})$ if $\boldsymbol{y}_1 \leq \boldsymbol{y}_2$

As for the convexity of $G_t(\cdot, \boldsymbol{z}) = \langle \nabla F_t^\star(\cdot), \boldsymbol{z}\rangle$, let $\boldsymbol{a} = \nabla F_t^\star(p\,\boldsymbol{y}_1 + (1-p)\boldsymbol{y}_2)$ and $\boldsymbol{b} = p\nabla F_t^\star(\boldsymbol{y}_1) + (1-p)\nabla F_t^\star(\boldsymbol{y}_2)$ for $0 \leq p \leq 1$. By the concavity of W_t, we have that

$$\begin{aligned}
&W_t(\boldsymbol{a}, \boldsymbol{z}) = \langle \nabla F_t(\boldsymbol{a}), \boldsymbol{z}\rangle \\
&= \langle p\,\boldsymbol{y}_1 + (1-p)\boldsymbol{y}_2, \boldsymbol{z}\rangle \\
&= \langle p\nabla F_t(\nabla F_t^\star(\boldsymbol{y}_1)) + (1-p)\nabla F_t(\nabla F_t^\star(\boldsymbol{y}_2)), \boldsymbol{z}\rangle \\
&= pW_t(\nabla F_t^\star(\boldsymbol{y}_1)), \boldsymbol{z}) + (1-p)W_t(\nabla F_t^\star(\boldsymbol{y}_2)), \boldsymbol{z}) \\
&\leq W_t(p\nabla F_t^\star(\boldsymbol{y}_1)) + (1-p)\nabla F_t^\star(\boldsymbol{y}_2)), \boldsymbol{z}) \\
&= W_t(\boldsymbol{b}, \boldsymbol{z})
\end{aligned}$$

Since $W_t(\boldsymbol{x}_1, \boldsymbol{z}) \leq W_t(\boldsymbol{x}_2, \boldsymbol{z})$ iff $\boldsymbol{x}_1 \leq \boldsymbol{x}_2$, we must have that $\boldsymbol{a} \leq \boldsymbol{b}$, and thus

$$p\,G_t(\boldsymbol{y}_1, \boldsymbol{z}) + (1-p)G_t(\boldsymbol{y}_2, \boldsymbol{z}) \geq G_t(p\,\boldsymbol{y}_1 + (1-p)\boldsymbol{y}_2, \boldsymbol{z})$$

which completes the proof. □

Lemma 1. *For any* $\boldsymbol{z} > \mathbf{0}$*, if* $W_t(\cdot, \boldsymbol{z})$ *is concave,* $W_t(\boldsymbol{x}_1, \boldsymbol{z}) \leq W_t(\boldsymbol{x}_2, \boldsymbol{z})$ *iff* $\boldsymbol{x}_1 \leq \boldsymbol{x}_2$*, and* $W_t(\boldsymbol{x}_1, \mathbf{1}) < W_t(\boldsymbol{x}_2, \mathbf{1})$ *iff* $\boldsymbol{x}_1 < \boldsymbol{x}_2$*, then*

$$\langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle \leq (\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t$$

Proof. We start by using Lemma 6:

$$\begin{aligned}\langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle &= \langle \nabla F_t^\star(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}) - \nabla F_t^\star(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t\mathbf{1}), \widehat{\boldsymbol{\ell}}\rangle \\ &= G_t(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}, \widehat{\boldsymbol{\ell}}_t) - G_t(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t\mathbf{1}, \widehat{\boldsymbol{\ell}}_t)\end{aligned}$$

Now, we claim that $c_t^{av} \le c_t$, which we prove by contradiction. First, note that in the case where $\widehat{\boldsymbol{L}}_t^{av} = \widehat{\boldsymbol{L}}_t$ we have that $c_t^{av} = c_t$, so we only need to prove that $c_t^{av} \le c_t$ when $\widehat{\boldsymbol{L}}_t^{av} \ne \widehat{\boldsymbol{L}}_t$, in which case $\widehat{\boldsymbol{L}}_t^{av} < \widehat{\boldsymbol{L}}_t$ must hold since $\widehat{\ell}_s(i) \ge 0$. Suppose that $c_t^{av} > c_t$. We have that

$$\begin{aligned}1 &= \langle \boldsymbol{p}_t^{av}, \mathbf{1}\rangle \\ &= G_t(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}, \mathbf{1}) \\ &> G_t(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t\mathbf{1}, \mathbf{1}) \\ &= \langle \nabla F_t^\star(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t\mathbf{1}), \mathbf{1}\rangle \\ &= \langle \boldsymbol{p}_{t+1}, \mathbf{1}\rangle = 1\end{aligned}$$

where in the first inequality we used Lemma 6, the assumption that $c_t^{av} > c_t$, and the fact that $\widehat{\boldsymbol{L}}_t > \widehat{\boldsymbol{L}}_t^{av}$. For the penultimate equality we used that $\nabla F_t^\star(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t\mathbf{1}) = \boldsymbol{p}_{t+1}$ by Lemma 6. The last equation provides a contradiction, allowing us to conclude that $c_t^{av} \le c_t$. Using Lemma 6 we can then write

$$\begin{aligned}\langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}\rangle &= G_t(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}, \widehat{\boldsymbol{\ell}}_t) - G_t(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t\mathbf{1}, \widehat{\boldsymbol{\ell}}_t) \\ &\le G_t(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}, \widehat{\boldsymbol{\ell}}_t) - G_t(-\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}, \widehat{\boldsymbol{\ell}}_t) \\ &\le \left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top \nabla G_t(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1}, \widehat{\boldsymbol{\ell}}_t)\end{aligned}$$

where in the last step we used the convexity of G_t in its first argument (Lemma 6). Recalling the definition of G_t, we have $\nabla G_t(\boldsymbol{y}, \boldsymbol{z}) = \nabla^2 F_t^\star(\boldsymbol{y})\boldsymbol{z}$, implying

$$\langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}\rangle \le \left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top \nabla^2 F_t^\star(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1})\widehat{\boldsymbol{\ell}}_t$$

Since F is Legendre, $\nabla F^\star$ is the inverse function of ∇F and we can use the inverse function theorem stating that $\nabla^2 F_t^\star(\boldsymbol{x}) = \left(\nabla^2 F_t(\nabla F_t^\star(\boldsymbol{x}))\right)^{-1}$. Hence

$$\begin{aligned}\langle \boldsymbol{p}_t^{av} - \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}\rangle &\le \left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top \nabla^2 F_t^\star(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1})\widehat{\boldsymbol{\ell}}_t \\ &= \left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top \left(\nabla^2 F_t\big(F_t^\star(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1})\big)\right)^{-1}\widehat{\boldsymbol{\ell}}_t \\ &= \left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top \left(\nabla^2 F_t(\boldsymbol{p}_t^{av})\right)^{-1}\widehat{\boldsymbol{\ell}}_t\end{aligned}$$

where in the last step we used $\boldsymbol{p}_t^{av} = \nabla F_t^\star(-\widehat{\boldsymbol{L}}_t^{av} - \boldsymbol{L}_0^t + c_t^{av}\mathbf{1})$. The proof is concluded by observing that $\nabla^2 F_t = \nabla^2 R_t$. $\square$

Lemma 2. *Suppose $R_t(\boldsymbol{p}) \ge R_{t-1}(\boldsymbol{p})$ for all $t \ge 1$ and $\boldsymbol{p} \in \triangle'$, where $R_0 = R_1$. Let $\boldsymbol{p}_0 = \arg\min_{\boldsymbol{p}\in\triangle'} R_1(\boldsymbol{p})$. Then, for any $\boldsymbol{u} \in \triangle'$ and for $\boldsymbol{p}_{t+1}$ defined in (4), $\sum_{t=1}^T \langle \boldsymbol{p}_{t+1} - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t\rangle \le R_T(\boldsymbol{u})$.*

Proof. We start by proving by induction that $\sum_{t=1}^T \langle \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle + R_1(\boldsymbol{p}_0) \le \sum_{t=1}^T \langle \boldsymbol{p}, \widehat{\boldsymbol{\ell}}_t\rangle + R_T(\boldsymbol{p})$ holds for any $\boldsymbol{p} \in \triangle$ and T. In the base case $T = 0$, for which the inequality holds by definition of $\boldsymbol{p}_0$, as for $T = 0$ the sum is empty and $R_1 = R_0$. To prove the induction step, we fix $T > 0$ and assume that for any $\boldsymbol{p} \in \triangle'$

$$\sum_{t=1}^{T-1} \langle \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle + R_1(\boldsymbol{p}_0) \le \sum_{t=1}^{T-1} \langle \boldsymbol{p}, \widehat{\boldsymbol{\ell}}_t\rangle + R_{T-1}(\boldsymbol{p})$$

Adding $\langle \boldsymbol{p}_{T+1}, \widehat{\boldsymbol{\ell}}_T\rangle$ to both sides of the equation above we find

$$\langle \boldsymbol{p}_{T+1}, \widehat{\boldsymbol{\ell}}_T\rangle + \sum_{t=1}^{T-1} \langle \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t\rangle + R_1(\boldsymbol{p}_0) \le \langle \boldsymbol{p}_{T+1}, \widehat{\boldsymbol{\ell}}_T\rangle + \sum_{t=1}^{T-1} \langle \boldsymbol{p}, \widehat{\boldsymbol{\ell}}_t\rangle + R_{T-1}(\boldsymbol{p}).$$

Choosing $\boldsymbol{p} = \boldsymbol{p}_{T+1}$, using that $R_{T-1}(\boldsymbol{p}) \leq R_T(\boldsymbol{p})$, and using the definition of $\boldsymbol{p}_{T+1}$ we see that

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t \rangle + R_1(\boldsymbol{p}_0) &\leq \sum_{t=1}^{T} \langle \boldsymbol{p}_{T+1}, \widehat{\boldsymbol{\ell}}_t \rangle + R_{T-1}(\boldsymbol{p}_{T+1}) \\
&\leq \sum_{t=1}^{T} \langle \boldsymbol{p}_{T+1}, \widehat{\boldsymbol{\ell}}_t \rangle + R_T(\boldsymbol{p}_{T+1}) \\
&= \min_{\boldsymbol{p} \in \triangle} \left\{ \sum_{t=1}^{T} \langle \boldsymbol{p}, \widehat{\boldsymbol{\ell}}_t \rangle + R_T(\boldsymbol{p}) \right\}
\end{aligned}$$

which proves the inductive step.

The above implies that

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_{t+1}, \widehat{\boldsymbol{\ell}}_t \rangle + R_1(\boldsymbol{p}_0) \leq \sum_{t=1}^{T} \langle \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t \rangle + R_T(\boldsymbol{u})$$

Finally, by reordering the equation above we find

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_{t+1} - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t \rangle \leq R_T(\boldsymbol{u}) - R_1(\boldsymbol{p}_0)$$

concluding the proof after observing that $R_1(\boldsymbol{p}_0)$ since the Bregman divergence is non-negative. □

B DETAILS OF SECTION 3

Theorem 4. *If we run Algorithm 1 with regularizer* $\Phi_t \equiv 0$, Ψ_t *defined by* (8), *corrections* a_t *defined by* (10), *and prior defined by* (12), *then the predictions* $\boldsymbol{q}_t$ *defined by* (9) *using* $\boldsymbol{p}_t^{av}$ *returned by the algorithm satisfy*

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle \leq\ & 4(1 + \rho_T^{\max}) + 12\sqrt{1 + \rho_T^{\max}} + 8\sqrt{\big(\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)\big)\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^{\rho} \rangle} \\
&+ 8\big(\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)\big)(1 + \rho_T^{\max}) + 16\sqrt{(1 + \rho_T^{\max})\big(\ln(K) + 2(1 + \ln(T))\big)}
\end{aligned}$$

for any $\boldsymbol{u} \in \triangle$.

Proof of Theorem 4. We consider the enlarged simplex Δ' over $K' = K \times J$ coordinates. Fix any $\boldsymbol{u}' \in \Delta'$. First, observe that by definition of $\boldsymbol{q}_t$, $\boldsymbol{u}'$, and $\widetilde{\boldsymbol{\ell}}$ we have that

$$\sum_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle = \sum_{t=1}^{T} \langle \boldsymbol{p}_t^{av} - \boldsymbol{u}', \widetilde{\boldsymbol{\ell}}_t \rangle$$

where the inner product on the left-hand side is on $\mathbb{R}^K$ and the inner products on the right-hand side are on $\mathbb{R}^{K \times J}$.

It is relatively easy to verify that, for any t,

$$R_T(\boldsymbol{u}') = \sum_{i=1}^{K} \sum_{\eta \in H_T} \left(\frac{1}{\eta} u'(i,\eta) \ln\left(\frac{u'(i,\eta)}{p_1^{\Psi}(i,\eta)} \right) + \frac{1}{\eta}\big(p_1^{\Psi}(i,\eta) - u'(i,\eta)\big) \right)$$

verifies the conditions of Lemmas 1 and 2. Indeed,

$$W_t(\boldsymbol{u}', \boldsymbol{z}) = \langle \nabla F_t(\boldsymbol{u}'), \boldsymbol{z} \rangle = \sum_{i=1}^{K} \sum_{\eta \in H_T} \left(\frac{z(i,\eta)}{\eta} \ln\big(u'(i,\eta)\big) + 1 \right)$$

is concave and strictly monotone in the coefficients $u'(i,\eta)$ as required by Lemma 1. To see that F_t is Legendre, note that $\nabla^2 F(\boldsymbol{x})$ is positive definite and thus F_t is strictly convex. Now pick any sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ in the interior of the domain of F_t converging to a boundary point. $\|\nabla F_t(\boldsymbol{x}_n)\| \to \infty$ because $(\nabla F_t(\boldsymbol{x}_n))_{i'}$ is increasing in $x_n(i')$. Moreover, the learning rates defined in (11) are decreasing in t and the Bregman divergence is non-negative, implying $R_t(\boldsymbol{u}') \geq R_{t-1}(\boldsymbol{u}')$ as required by Lemma 2. Therefore we can apply Lemma 3 to R_t and obtain:

$$\sum_{t=1}^{T}\langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle \leq R_T(\boldsymbol{u}') + \sum_{t=1}^{T}\langle \boldsymbol{u}', \boldsymbol{a}_t\rangle + \sum_{t=1}^{T}\left((\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \left(\nabla^2 R_t(\boldsymbol{p}_t^{av})\right)^{-1}\widehat{\boldsymbol{\ell}}_t - \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t\rangle\right) \tag{25}$$

We continue by bounding R_T. We have that

$$\begin{aligned} R_T(\boldsymbol{u}') &= \sum_{i=1}^{K}\sum_{\eta\in H_T}\left(\frac{1}{\eta}u'(i,\eta)\ln\left(\frac{u'(i,\eta)}{p_1^{\Psi}(i,\eta)}\right) + \frac{1}{\eta}\left(p_1^{\Psi}(i,\eta) - u'(i,\eta)\right)\right) \\ &\leq \sum_{i=1}^{K}\sum_{\eta\in H_T}\frac{1}{\eta}u'(i,\eta)\ln\left(\frac{u'(i,\eta)}{p_1^{\Psi}(i,\eta)}\right) + \sum_{i=1}^{K}\sum_{\eta\in H_T}\frac{1}{\eta}p_1^{\Psi}(i,\eta) \end{aligned} \tag{26}$$

We proceed by bounding the first sum on the right-hand side of equation (26). Denote by $\eta^\star \in H_T$ a target learning rate and let $j^\star$ be the corresponding index of $\eta^\star$. Let $\boldsymbol{u}' \in \Delta'$ be defined by $u'(i') = u'(i,\eta) = 0$ if $\eta \neq \eta^\star$ and $u'(i,\eta) = u(i)$ otherwise. Then

$$\begin{aligned} \sum_{i=1}^{K}\sum_{\eta\in H_T}\frac{1}{\eta}u'(i,\eta)\ln\left(\frac{u'(i,\eta)}{p_1^{\Psi}(i,\eta)}\right) &= \frac{1}{\eta^\star}\sum_{i=1}^{K}u(i)\ln\left(\frac{u(i)}{\pi(i)\frac{2^{-2j^\star}}{\sum_{j'=1}^{J}2^{-2j'}}}\right) \\ &= \frac{1}{\eta^\star}\left(\sum_{i=1}^{K}u(i)\ln\left(\frac{u(i)}{\pi(i)}\right) + \sum_{i=1}^{K}u(i)\ln\left(\frac{\sum_{j'=1}^{J}2^{-2j'}}{2^{-2j^\star}}\right)\right) \\ &= \frac{1}{\eta^\star}\left(\mathrm{KL}(\boldsymbol{u},\pi) + 2j^\star + \ln\left(\sum_{j'=1}^{J}2^{-2j'}\right)\right) \\ &\leq \frac{\mathrm{KL}(\boldsymbol{u},\pi) + J}{\eta^\star} \end{aligned}$$

where we used that $\sum_{j=1}^{J}2^{-2j} = \frac{1}{3} - \frac{4^{-J}}{3} \leq 1$. We now bound the final sum on the right-hand side of equation (26) by using that $\frac{1}{\min\{a,b\}} \leq \frac{1}{a} + \frac{1}{b}$ for $a, b > 0$:

$$\begin{aligned} \sum_{i=1}^{K}\sum_{\eta\in H_T}\frac{1}{\eta}p_1^{\Psi}(i,\eta) &= \sum_{i=1}^{K}\sum_{j=1}^{J}\frac{\frac{2^{-2j}}{\sum_{j'\in[J]}2^{-2j'}}\pi(i)}{\min\left\{\frac{1}{4(1+\rho_t^{\max})}, \frac{\sqrt{\ln(K)+2(\ln(T)+1)}}{4\sqrt{1+\rho_t^{\max}}}2^{-j}\right\}} \\ &\leq 4(1+\rho_T^{\max}) + \frac{4\sqrt{1+\rho_T^{\max}}}{\sqrt{\ln(K)+2(\ln(T)+1)}}\frac{\sum_{j=1}^{J}2^{-j}}{\sum_{j=1}^{J}2^{-2j}} \\ &= 4(1+\rho_T^{\max}) + \frac{12\sqrt{1+\rho_T^{\max}}}{\sqrt{\ln(K)+2(\ln(T)+1)}}\frac{1-2^{-J}}{1-4^{-J}} \\ &\leq 4(1+\rho_T^{\max}) + 12\sqrt{1+\rho_T^{\max}} \end{aligned} \tag{27}$$

where in the last equality we used again that $\sum_{j=1}^{J}2^{-2j} = \frac{1}{3} - \frac{4^{-J}}{3}$ and that $\sum_{j=1}^{J}2^{-j} = 1 - 2^{-J}$. We continue by further bounding the drift term using the fact that $\nabla^2 R_t(\boldsymbol{p}_t^{av})$ is a diagonal matrix with diagonal

$$\left(\nabla^2 R_t(\boldsymbol{p}_t^{av})\right)_{i'i'} = \frac{1}{\eta p_t^{av}(i,\eta)}$$

for $i' = (i, \eta)$. We have

$$(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 R_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t = \sum_{i=1}^K \sum_{\eta \in H_t} \big(\widehat{L}_t(i,\eta) - \widehat{L}_t^{av}(i,\eta)\big)\eta p_t^{av}(i,\eta)\widehat{\ell}_t(i,\eta)$$
$$\leq \sum_{i=1}^K \sum_{\eta \in H_t} \big(1 + \rho_t(i,\eta)\big) p_t^{av}(i,\eta)\widetilde{\ell}_t(i,\eta) = \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t \rangle$$

where in the last step we used (6), the inequality $\widehat{\ell}_t(i') \leq 2\widetilde{\ell}_t(i') = 2\ell_t(i)$, and the assumption $\ell_t(i) \in [0,1]$.

Combining the above and using that $\sum_{t=1}^T \langle \boldsymbol{u}', \boldsymbol{a}_t \rangle = \eta^\star \sum_{i=1}^K u(i) \sum_{t=1}^T 4\ell_t(i)(1 + \rho_t(i)) = 4\eta^\star \langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle$ and $J \leq \ln(T) + 1$, we continue from (25):

$$\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle \leq 4(1 + \rho_T^{\max}) + 12\sqrt{1 + \rho_T^{\max}} + \frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{\eta^\star} + 4\eta^\star \langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle \tag{28}$$

Ideally, we would set $\eta^\star$ to

$$\sqrt{\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle}} \tag{29}$$

However, we can only pick a learning rate from H_T. We now show that H_T contains a good approximation of the quantity in (29). We split the remainder of the analysis into two cases. First, assume

$$\max_{\eta \in H_T} \eta \leq \sqrt{\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle}}$$

and thus

$$4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle \leq \frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{(\max_{\eta \in H_T} \eta)^2}$$

Using the above in equation (28) and choosing $\eta^\star = \max_{\eta \in H_T} \eta$ we obtain

$$\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle \leq 4(1 + \rho_T^{\max}) + 12\sqrt{1 + \rho_T^{\max}} + \frac{2(\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1))}{\eta^\star}$$
$$\leq 4(1 + \rho_T^{\max}) + 12\sqrt{1 + \rho_T^{\max}} + 8(\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1))(1 + \rho_T^{\max})$$
$$+ 16\sqrt{(1 + \rho_T^{\max})(\ln(K) + 2(1 + \ln(T)))}$$

where in the last step we used

$$\frac{1}{\max_{\eta \in H_T} \eta} \leq 4(1 + \rho_T^{\max}) + 8\sqrt{\frac{1 + \rho_T^{\max}}{\ln(K) + 2(\ln(T) + 1)}}$$

In the second case

$$\max_{\eta \in H_T} \eta > \sqrt{\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle}}$$

This implies that there is an $\eta \in H_T$ that is within a factor 2 of $\sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+2(\ln(T)+1)}{\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle}}$ as

$$\min_{\eta \in H_T} \eta \leq \sqrt{\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{4(1 + \rho_T^{\max})T}} \leq \sqrt{\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle}} \leq \max_{\eta \in H_T} \eta$$

and thus

$$\frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{\eta^\star} + \eta^\star 4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle \leq 8\sqrt{(\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1))\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho \rangle}.$$

Therefore, we have that in the second case

$$\sum_{t=1}^{T} \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle \leq 4(1+\rho_T^{\max}) + 12\sqrt{1+\rho_T^{\max}} + 8\sqrt{(\mathrm{KL}(\boldsymbol{u},\pi) + 2(\ln(T)+1))\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^{\rho}\rangle}.$$

Combining the first and second case completes the proof. □

C DETAILS OF SECTION 4.1

Lemma 4. *For all t, the function $F_t = \Psi_t + \Phi_t$, where Ψ_t is defined in* (13) *and Φ_t is defined in* (14)*, is Legendre and satisfies the conditions of Lemma 1. Moreover,*

$$\begin{aligned}(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t &\leq \eta_t 4 \ell_t(i_t)^2 \\ &+ 4\sum\nolimits_{\gamma\in\Gamma} \gamma \sum\nolimits_{s\in m_t(i_t)} p_t^{av}(i_t,\gamma)\widetilde{\ell}_t(i_t,\gamma)\widetilde{\ell}_s(i_t,\gamma)\end{aligned}$$

Proof. We first show that F_t is Legendre. Note that F_t is strictly convex because Φ_t is strictly convex and Ψ_t is convex. Now pick any sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ in the interior of the domain of F_t converging to a boundary point. We have that $\|\nabla F_t(\boldsymbol{x}_n)\| \to \infty$, because each component

$$\big(\nabla F_t(\boldsymbol{x})\big)_{(i,\gamma)} = -\frac{1}{\eta_t \sum_{\gamma\in\Gamma} x(i,\gamma)} + \frac{1}{\gamma}\ln\big(x(i,\gamma)\big) + \frac{1}{\gamma}$$

of the gradient of F_t is increasing in each coordinate of $\boldsymbol{x}$. The same observation, together with the fact that both $\langle \nabla\Psi_t(\cdot), \boldsymbol{z}\rangle$ and $\langle \nabla\Phi_t(\cdot), \boldsymbol{z}\rangle$ are concave, shows that $W_t(\cdot, \boldsymbol{z}) = \langle \nabla F_t(\cdot), \boldsymbol{z}\rangle$ satisfies the conditions of Lemma 1.

Observe that $\nabla^2 F_t(\boldsymbol{p}_t^{av})$ is a block-diagonal matrix with blocks $i = 1, \ldots, K$

$$B_t(i) = \frac{\boldsymbol{1}\boldsymbol{1}^\top}{\eta_t\,(q_t(i))^2} + \mathrm{diag}\left(\frac{1}{\gamma_1 p_t^{av}(i,\gamma_1)}, \ldots, \frac{1}{\gamma_J p_t^{av}(i,\gamma_J)}\right)$$

of size $J \times J$ each. Denote by $V_t = \mathrm{diag}(\boldsymbol{v}_t)$, where $\boldsymbol{v}_t = \big(\gamma_1 p_t^{av}(i,\gamma_1), \ldots, \gamma_J p_t^{av}(i,\gamma_J)\big)$. The inverse of $B_t(i)$ can be computed by employing the Sherman-Morrison formula:

$$B_t(i)^{-1} = V_t - \frac{\eta_t^{-1} q_t(i)^{-2} V_t \boldsymbol{1}\boldsymbol{1}^\top V_t}{1 + \eta_t^{-1} q_t(i)^{-2}\langle \boldsymbol{v}_t, \boldsymbol{1}\rangle} = V_t - \frac{\boldsymbol{v}_t \boldsymbol{v}_t^\top}{\eta_t q_t(i)^2 + \langle \boldsymbol{v}_t, \boldsymbol{1}\rangle} \tag{30}$$

Note that $\widehat{\boldsymbol{\ell}}_t$ is only non-zero in the J coordinates of the form (i_t, γ) for $\gamma \in \Gamma$. Let $\boldsymbol{h}$ be the J-vector including only these non-zero elements of $\widehat{\boldsymbol{\ell}}_t$, and let $\boldsymbol{b} = q_t(i_t)\boldsymbol{h}$ so that $b(j) = \ell_t(i_t) + q_t(i_t) a_t(i_t,\gamma_j) = \ell_t(i_t) + 4\gamma_j \ell_t(i_t)\rho_t(i)$. Denote by $\boldsymbol{0}_{K\times(J-1)}$ a vector of zeros of length $K \times (J-1)$. Since the block-diagonal structure is preserved when taking inverses, $\big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1}$ is a block-diagonal matrix with blocks $B_t(i)^{-1}$ and $\big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t = \big(\boldsymbol{0}_{K\times(J-1)}, B_t(i_t)^{-1}\boldsymbol{h}\big)$. Next, we write $\boldsymbol{b} = Y\boldsymbol{1}\ell_t(i_t)$, where

$$Y = \boldsymbol{I}_J + 4\rho_t(i)\mathrm{diag}(\gamma_1, \ldots, \gamma_J)$$

and $\boldsymbol{I}_J$ is the $J \times J$ identity matrix. Using (30) and $Y \preceq 2\boldsymbol{I}_J$ because $\max_j \gamma_j \leq (4\rho^\star)^{-1}$, we continue with

$$\begin{aligned}\widehat{\boldsymbol{\ell}}_t^\top \big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1}\widehat{\boldsymbol{\ell}}_t &= \frac{\ell_t(i_t)^2}{q_t(i_t)^2}\boldsymbol{1}^\top Y B_t(i_t)^{-1} Y \boldsymbol{1} \\ &\leq \frac{4\ell_t(i_t)^2}{q_t(i_t)^2}\boldsymbol{1}^\top B_t(i_t)^{-1}\boldsymbol{1} \\ &= \frac{4\ell_t(i_t)^2}{q_t(i_t)^2}\left(\langle \boldsymbol{v}_t, \boldsymbol{1}\rangle - \frac{\langle \boldsymbol{v}_t, \boldsymbol{1}\rangle^2}{\eta_t q(i_t)^2 + \langle \boldsymbol{v}_t, \boldsymbol{1}\rangle}\right) \\ &= \frac{4\ell_t(i_t)^2}{q_t(i_t)^2}\frac{\eta_t q(i_t)^2 \langle \boldsymbol{v}_t, \boldsymbol{1}\rangle}{\eta_t q(i_t)^2 + \langle \boldsymbol{v}_t, \boldsymbol{1}\rangle} \\ &\leq 4\eta_t \ell_t(i_t)^2\end{aligned} \tag{31}$$

where in the last step we used $\boldsymbol{v}_t \geq 0$. Denote by $\boldsymbol{L}^{\text{miss}} = \widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{\ell}}_t - \widehat{\boldsymbol{L}}_t^{av}$ and denote by $\boldsymbol{y}$ the components of $\boldsymbol{L}^{\text{miss}}$ corresponding to i_t. Namely,

$$y(i_t, \gamma) = \sum_{s \in m_t(i_t)} \widehat{\ell}_t(i_t, \gamma)$$

Since $\langle \boldsymbol{y}, \boldsymbol{v}_t \rangle \langle \boldsymbol{v}_t, \boldsymbol{h} \rangle \geq 0$ (these are all nonnegative vectors), we have that

$$\begin{aligned}
\boldsymbol{L}^{\text{miss}} \big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t &= \boldsymbol{y}^\top B_t(i_t)^{-1} \boldsymbol{h} \\
&= \boldsymbol{y}^\top V_t \boldsymbol{h} - \frac{\langle \boldsymbol{y}, \boldsymbol{v}_t \rangle \langle \boldsymbol{v}_t, \boldsymbol{h} \rangle}{\eta_t q(i_t)^2 + \langle \boldsymbol{v}_t, \boldsymbol{1} \rangle} \\
&\leq \boldsymbol{y}^\top V_t \boldsymbol{h} \\
&= \sum_{\gamma \in \Gamma} \sum_{s \in m_t(i_t)} p_t^{av}(i_t, \gamma) \gamma \widehat{\ell}_t(i_t, \gamma) \widehat{\ell}_s(i_t, \gamma) \\
&\leq 4 \sum_{\gamma \in \Gamma} \sum_{s \in m_t(i_t)} p_t^{av}(i, \gamma) \gamma \widetilde{\ell}_t(i_t, \gamma) \widetilde{\ell}_s(i_t, \gamma)
\end{aligned} \tag{32}$$

where in the last step we used $\widehat{\ell}_t(i') \leq 2\widetilde{\ell}_t(i')$. Combining equations (31) and (32) we obtain

$$(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top \big(\nabla^2 F_t(\boldsymbol{p}_t^{av})\big)^{-1} \widehat{\boldsymbol{\ell}}_t \leq \eta_t 4 \ell_t(i_t)^2 + 4 \sum_{\gamma \in \Gamma} \sum_{s \in m_t(i_t)} p_t^{av}(i_t, \gamma) \gamma \widetilde{\ell}_t(i_t, \gamma) \widetilde{\ell}_s(i_t, \gamma)$$

which completes the proof. □

Lemma 7. *Let $\boldsymbol{p}_1^\Psi \equiv \frac{1}{KJ}$, let $\boldsymbol{p}_1^\Phi$ be set as in* (12) *and Γ be set as in* (17). *If $\boldsymbol{p}^\star = (1-\alpha)\boldsymbol{u}' + \alpha \frac{\boldsymbol{1}}{K'}$ with $\alpha = \frac{1}{T}$ and $\pi_1(i) \geq \frac{1}{T^2}$, then $R_t(\boldsymbol{p}) \geq R_{t-1}(\boldsymbol{p})$ for all $t \geq 1$ and $\boldsymbol{p} \in \triangle'$. Furthermore*

$$\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{p}^\star - \boldsymbol{u}', \widehat{\boldsymbol{\ell}}_t \rangle\right] + R_T(\boldsymbol{p}^\star) \leq 50\rho^\star + \frac{K \ln(T)}{\eta_T} + \frac{1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)}{\gamma^\star}$$

Proof. Let $\phi_\gamma(x) = \frac{x}{\gamma} \ln(x)$ and $\psi(x) = -\ln(x)$. To see that $R_t(\boldsymbol{p}) \geq R_{t-1}(\boldsymbol{p})$ observe that

$$\begin{aligned}
R_t(\boldsymbol{p}) &= B_{\Psi_t}(\boldsymbol{p}, \boldsymbol{p}_1^\Psi) + B_{\Phi_t}(\boldsymbol{p}, \boldsymbol{p}_1^\Phi) \\
&= \sum_{i=1}^K \frac{1}{\eta_t} B_\psi \left(\sum_{\gamma \in \Gamma} p(i, \gamma), \sum_{\gamma \in \Gamma} p_1^\Psi(i, \gamma)\right) \\
&+ \sum_{i=1}^K \sum_{\gamma \in \Gamma} B_{\phi_\gamma}\big(p(i, \gamma), p_1^\Phi(i, \gamma)\big) \\
&\geq \sum_{i=1}^K \frac{1}{\eta_{t-1}} B_\psi \left(\sum_{\gamma \in \Gamma} p(i, \gamma), \sum_{\gamma \in \Gamma} p_1^\Psi(i, \gamma)\right) \\
&+ \sum_{i=1}^K \sum_{\gamma \in \Gamma} B_{\phi_\gamma}\big(p(i, \gamma), p_1^\Phi(i, \gamma)\big) \\
&= B_{\Psi_{t-1}}(\boldsymbol{p}, \boldsymbol{p}_1^\Psi) + B_{\Phi_{t-1}}(\boldsymbol{p}, \boldsymbol{p}_1^\Phi) = R_{t-1}(\boldsymbol{p})
\end{aligned}$$

where the inequality is due to the non-negativity of the Bregman divergence and the fact that $\eta_t \leq \eta_{t-1}$.

Now $a_t(i') = 4\widetilde{\ell}_t(i')\gamma(i')\rho_t \leq \widetilde{\ell}_t(i)$ because $\max_j \gamma_j \leq \big(4\rho^\star\big)^{-1}$, and thus

$$\sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{p}^\star - \boldsymbol{u}', \widehat{\boldsymbol{\ell}}_t \rangle\right] \leq \sum_{t=1}^T \mathbb{E}\left[\langle \frac{\alpha}{K'} \boldsymbol{1}, \widehat{\boldsymbol{\ell}}_t \rangle\right] \leq 2\alpha T$$

For any $\boldsymbol{p}^\star$ we have that:

$$R_T(\boldsymbol{p}^\star) = -\frac{1}{\eta_T}\sum_{i=1}^K \ln\left(K\sum_{\gamma\in\Gamma} p^\star(i,\gamma)\right) + \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1}{\gamma}\left(p^\star(i,\gamma)\ln\left(\frac{p^\star(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + p_1^\Phi(i,\gamma) - p^\star(i,\gamma)\right) \tag{33}$$

We continue by bounding the first sum on the right hand side of equation (33):

$$\begin{aligned}
-\sum_{i=1}^K \ln\left(K\sum_{\gamma\in\Gamma} p^\star(i,\gamma)\right) &= -\sum_{i=1}^K \ln\left(\sum_{\gamma\in\Gamma}(1-\alpha)u'(i,\gamma)K + \alpha\right)\\
&\leq -K\ln(\alpha)
\end{aligned}$$

Next, we bound the second sum on the right hand side of (33):

$$\begin{aligned}
&\sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1}{\gamma}\left(p^\star(i,\gamma)\ln\left(\frac{p^\star(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + p_1^\Phi(i,\gamma) - p^\star(i,\gamma)\right)\\
&\quad\leq \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1}{\gamma}\left(p^\star(i,\gamma)\ln\left(\frac{p^\star(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + p_1^\Phi(i,\gamma)\right)\\
&\quad\leq \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1}{\gamma}p^\star(i,\gamma)\ln\left(\frac{p^\star(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + 4\rho^\star + \frac{12\sqrt{\rho^\star}}{\sqrt{\ln(K)+\ln(T)+1}}\\
&\quad\leq \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1}{\gamma}p^\star(i,\gamma)\ln\left(\frac{p^\star(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + 4\rho^\star + 12\sqrt{\rho^\star}
\end{aligned}$$

where we used that $\sum_{j=1}^J 2^{-2j} = 1/3 - \frac{4^{-J}}{3}$ and that that $\frac{1}{\min\{a,b\}} \leq \frac{1}{a} + \frac{1}{b}$ for $a, b > 0$, see also (27). Now, to bound the double sum on the right-hand side of the last inequality we use Jensen's inequality, which we may use due the convexity of $f(x) = x\ln(xa)$ for $a > 0$:

$$\begin{aligned}
\sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1}{\gamma}p^\star(i,\gamma)\ln\left(\frac{p^\star(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) &\leq \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1-\alpha}{\gamma}u'(i,\gamma)\ln\left(\frac{u'(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{\alpha/K'}{\gamma}\ln\left(\frac{1/K'}{p_1^\Phi(i,\gamma)}\right)\\
&\leq \sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1-\alpha}{\gamma}u'(i,\gamma)\ln\left(\frac{u'(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) + \alpha\max_{i,\gamma}\left\{\frac{1}{\gamma}\left(-\ln\left(p_1^\Phi(i,\gamma)\right) - \ln(K')\right)\right\}
\end{aligned}$$

Since

$$-\ln\left(p_1^\Phi(i,\gamma)\right) \leq -\ln(\pi(i)) + 2J + \ln\left(\sum_{j=1}^J 2^{-2j}\right) \leq -\ln(\pi(i)) + 2J$$

and $J \leq \ln(T) + 1$, $1/\gamma \leq 4\rho^\star\sqrt{T}$, and $\pi_1(i) \geq \frac{1}{T^2}$ we have that

$$\alpha\max_{i,\gamma}\left\{\frac{1}{\gamma}\left(-\ln\left(p_1^\Phi(i,\gamma)\right) - \ln(K')\right)\right\} \leq 16\alpha\rho^\star\sqrt{T}(1+\ln(T)) \leq 32\alpha\rho^\star T$$

where we used that $\ln(x) \leq \sqrt{x}$. Now, setting $u'(i,\gamma) = u(i)$ if $\gamma = \gamma^\star$ and 0 otherwise, we obtain

$$\begin{aligned}
\sum_{i=1}^K\sum_{\gamma\in\Gamma}\frac{1-\alpha}{\gamma}u'(i,\gamma)\ln\left(\frac{u'(i,\gamma)}{p_1^\Phi(i,\gamma)}\right) &= \frac{1-\alpha}{\gamma^\star}\left(-\ln\left(\frac{2^{-2j}}{\sum_{j\in[J]}2^{-2J}}\right) + \sum_{i=1}^K u(i)\ln\left(\frac{u(i)}{\pi(i)}\right)\right)\\
&\leq \frac{1}{\gamma^\star}\left(2J + \mathrm{KL}(\boldsymbol{u},\pi)\right)\\
&\leq \frac{1}{\gamma^\star}\left(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u},\pi)\right)
\end{aligned}$$

Combining the above and setting $\alpha = \frac{1}{T}$ we find

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{p}^\star - \boldsymbol{u}', \widehat{\boldsymbol{\ell}}_t\rangle\right] + R_T(\boldsymbol{p}^\star) &\leq 32\alpha\rho^\star T + \mathbb{E}\left[-\frac{1}{\eta_T}K\ln(\alpha)\right] + 2\alpha T + 4\rho^\star + 12\sqrt{\rho^\star} \\ &\quad + \frac{1}{\gamma^\star}\left(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\right) \\ &\leq 50\rho^\star + \mathbb{E}\left[\frac{K\ln(T)}{\eta_T}\right] + \frac{1}{\gamma^\star}\left(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\right)\end{aligned}$$

which completes the proof. □

Theorem 5. *Let $\boldsymbol{p}_1^\Psi \equiv \frac{1}{KJ}$ and let $\boldsymbol{p}_1^\Phi$ be given by* (12)*. Assume $\pi_1(i) \geq \frac{1}{T^2}$ for all $i \in [K]$. If we run Algorithm 1 with regularizers* (13) *and* (14)*, corresponding learning rates* (17) *and* (18)*, then the predictions $i_t \sim \boldsymbol{q}_t$, with $\boldsymbol{q}_t$ as in* (15)*, satisfy*

$$\begin{aligned}\mathbb{E}\Big[\sum_{t=1}^{T}(\ell_t(i_t) - \langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle)\Big] &\leq 12\sqrt{K\ln(T)\langle \boldsymbol{u}, \boldsymbol{L}_T\rangle} + 16\sqrt{\big(\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T) + 1\big)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle} \\ &\quad + 48\big(5 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\big)\rho^\star + 42K\ln(T)\end{aligned}$$

Proof. Let $\boldsymbol{p}^\star = (1-\alpha)\boldsymbol{u}' + \alpha\boldsymbol{1}\frac{1}{K'}$, where $\alpha = \frac{1}{T}$. We start by applying Lemma 3 to bound the expected regret against $\boldsymbol{p}^\star$, which we may do because the conditions of Lemma 3 are satisfied as per Lemma 4:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] &\leq \sum_{t=1}^{T}\mathbb{E}\left[\langle \boldsymbol{u}, \boldsymbol{a}_t\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{p}^\star - \boldsymbol{u}', \widehat{\ell}_t\rangle\right] + \mathbb{E}\left[R_T(\boldsymbol{p}^\star)\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T}\left(\left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top\left(\nabla^2 R_t(\boldsymbol{p}_t^{av})\right)^{-1}\widehat{\boldsymbol{\ell}}_t - \langle \boldsymbol{p}_t^{av}, \boldsymbol{a}_t\rangle\right)\right]\end{aligned}$$

Let $u'(i') = u'(i, \gamma) = u(i)$ if $\gamma = \gamma^\star$ and 0 otherwise and recall that $L_T^d(i) = \sum_{t=1}^{T}\ell_t(i)(1 + \rho_t(i))$. Since $\mathbb{E}[\ell_t(i)\mathbb{1}[i = i_t]q_t(i)^{-1}] = \ell_t(i)$ we have that $\sum_{t=1}^{T}\mathbb{E}\left[\langle \boldsymbol{u}, \boldsymbol{a}_t\rangle\right] = \sum_{t=1}^{T}4\gamma^\star\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle$. By applying Lemmas 7 and 5 we may further bound the expected regret:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \leq & 50\rho^\star + \mathbb{E}\left[\sum_{t=1}^{T}\eta_t 4\ell_t(i_t)^2\right] + \mathbb{E}\left[\frac{1}{\eta_T}K\ln(T)\right] \\ & + \sum_{t=1}^{T}4\gamma^\star\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle + \frac{1}{\gamma^\star}\big(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\big)\end{aligned} \tag{34}$$

Ideally, we would have that $\gamma^\star = \sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1}{4\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle}}$. However, we have a grid of learning rates Γ from which we have to choose an appropriate learning rate. Instead, we will show that Γ contains a close approximation of the optimal learning rate. As in the proof of Theorem 4 we will split the analysis into two case. In the first case $\max_{\gamma\in\Gamma}\gamma \leq \sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1}{4\langle \boldsymbol{u}, \boldsymbol{L}_T + \boldsymbol{L}_T^\rho\rangle}}$ and thus

$$4\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle \leq \frac{\mathrm{KL}(\boldsymbol{u}, \pi) + 2(\ln(T) + 1)}{(\max_{\gamma\in\Gamma}\gamma)^2}$$

Thus, choosing $\gamma^\star = \max_{\gamma\in\Gamma}\gamma$ we obtain

$$\begin{aligned}4\gamma^\star\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle + \frac{1}{\gamma^\star}\big(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\big) &\leq \frac{2}{\gamma^\star}\big(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\big) \\ &\leq 8\big(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\big)\big(\rho^\star + 2\sqrt{\rho^\star}\big)\end{aligned}$$

In the second case $\max_{\gamma\in\Gamma}\gamma > \sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1}{4\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle}}$. This implies that there is an $\gamma \in \Gamma$ that is within a factor 2 of $\sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1}{\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle}}$ as

$$\min_{\gamma\in\Gamma}\gamma \leq \sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1}{4(1+\rho_T^{\max})T}} \leq \sqrt{\frac{\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1}{4\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle}} \leq \max_{\gamma\in\Gamma}\gamma$$

and thus

$$4\gamma^\star\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle + \frac{1}{\gamma^\star}\big(+ \ln(T) + \mathrm{KL}(\boldsymbol{u},\pi)\big) \leq 8\sqrt{\big(\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1\big)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle}$$

Therefore, combining the two cases we find that

$$\begin{aligned}4\gamma^\star\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle + \frac{1}{\gamma^\star}\left(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u},\pi)\right) \leq & 8\sqrt{(\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle} \\ & + 8\left(1 + \ln(T) + \mathrm{KL}(\boldsymbol{u},\pi)\right)\left(\rho^\star + 2\sqrt{\rho^\star}\right)\end{aligned} \tag{35}$$

As for the terms involving η_t, we have that

$$\eta_t 4\ell_t(i_t) = 4\ell_t(i_t)\sqrt{\frac{K\ln(T)}{4(1+\rho^\star) + 4\sum_{s\in\mathcal{S}_t}\ell_s(i_s)}} \leq 4\ell_t(i_t)\sqrt{\frac{K\ln(T)}{4\sum_{s\leq t}\ell_s(i_s)}}$$

Summing over t and using that $\sum_{t=1}^t \frac{x_t}{\sum_{s\leq t} x_t} \leq 2\sqrt{\sum_{t=1}^T x_t}$ we find that

$$\sum_{t=1}^T \eta_t 4\ell_t(i_t) \leq 4\sqrt{K\ln(T)\sum_{t=1}^T \ell_t(i_t)}$$

which means that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \eta_t 4\ell_t(i_t)^2 + \frac{1}{\eta_T}K\ln(T)\right] &\leq 6\,\mathbb{E}\left[\sqrt{K\ln(T)(1+\rho^\star) + K\ln(T)\sum_{t=1}^T \ell_t(i_t)}\right] \\ &\leq 6\,\mathbb{E}\left[\sqrt{K\ln(T)\sum_{t=1}^T \ell_t(i_t)}\right] + 6\sqrt{K\ln(T)(1+\rho^\star)} \\ &\leq 6\sqrt{K\ln(T)\sum_{t=1}^T \mathbb{E}\left[\ell_t(i_t)\right]} + 6\sqrt{K\ln(T)(1+\rho^\star)}\end{aligned} \tag{36}$$

where we used that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$ and Jensen's inequality. Combining equations (34), (35), and (36) we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T\langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] \leq & 50\rho^\star + 6\sqrt{K\ln(T)\sum_{t=1}^T \mathbb{E}\left[\ell_t(i_t)\right]} + 6\sqrt{K\ln(T)(1+\rho^\star)} + 8\sqrt{(\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle} \\ & + 8\left(1+\ln(T)+\mathrm{KL}(\boldsymbol{u},\pi)\right)\left(\rho^\star + 2\sqrt{\rho^\star}\right) \\ \leq & 6\sqrt{K\ln(T)\sum_{t=1}^T \mathbb{E}\left[\ell_t(i_t)\right]} + 6\sqrt{K\ln(T)(1+\rho^\star)} + 8\sqrt{(\mathrm{KL}(\boldsymbol{u},\pi)+\ln(T)+1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho\rangle} \\ & + 24\big(4+\ln(T)+\mathrm{KL}(\boldsymbol{u},\pi)\big)\rho^\star\end{aligned}$$

To complete the proof observe that $\mathbb{E}\left[\ell_t(i_t)\right] = \mathbb{E}\left[\langle \boldsymbol{q}_t, \boldsymbol{\ell}_t \rangle\right] = \mathbb{E}\left[\langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right] + \langle \boldsymbol{u}, \boldsymbol{\ell}_t \rangle$. Using that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and using that $\sqrt{xy} = \frac{1}{2}\inf_{\zeta > 0} \frac{x}{\zeta} + \zeta y$ for $x, y \geq 0$ we find

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right] \leq & 6\sqrt{K\ln(T)\left(\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right] + \langle \boldsymbol{u}, \boldsymbol{L}_T \rangle\right)} + 6\sqrt{K\ln(T)(1+\rho^\star)} \\
& + 8\sqrt{(\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T) + 1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho \rangle} + 24\left(4 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\right)\rho^\star \\
\leq & 6\sqrt{K\ln(T)\langle \boldsymbol{u}, \boldsymbol{L}_T \rangle} + 6\sqrt{K\ln(T)(1+\rho^\star)} + 8\sqrt{(\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T) + 1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho \rangle} \\
& + 24\left(4 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\right)\rho^\star + \tfrac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right] + 18K\ln(T) \\
\leq & 6\sqrt{K\ln(T)\langle \boldsymbol{u}, \boldsymbol{L}_T \rangle} + 8\sqrt{(\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T) + 1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho \rangle} \\
& + 24\left(5 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\right)\rho^\star + \tfrac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right] + 21K\ln(T)
\end{aligned}$$

After reordering the above equation gives us

$$\begin{aligned}
\tfrac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right] \leq & 21K\ln(T) + 6\sqrt{K\ln(T)\langle \boldsymbol{u}, \boldsymbol{L}_T \rangle} \\
& + 8\sqrt{(\mathrm{KL}(\boldsymbol{u}, \pi) + \ln(T) + 1)\langle \boldsymbol{u}, \boldsymbol{L}_T^\rho \rangle} + 24\left(5 + \ln(T) + \mathrm{KL}(\boldsymbol{u}, \pi)\right)\rho^\star
\end{aligned}$$

which completes the proof after multiplying both sides by 2 □

D DETAILS OF SECTION 4.2

Theorem 3. *Let $\boldsymbol{q}_t$ be computed by Algorithm 1 with $a_t(i) = 0$, $\eta_t = \frac{1}{\sqrt{4t}}$, $\epsilon_t = \frac{1}{\sqrt{t}}$, γ_t as in (24), $p_1^\Psi(i) = p_1^\Phi(i) = \frac{1}{K}$, $\widetilde{\ell}_t(i) = \frac{\mathbb{1}[i_t = i]\ell_t(i)}{q_t(i) + \epsilon_t}$, and regularizers specified in (21) and (22). Then actions $i_t \sim \boldsymbol{q}_t$ guarantee*

$$\begin{aligned}
\mathbb{E}\left[\sum\nolimits_{t=1}^T (\ell_t(i_t) - \langle \boldsymbol{u}, \boldsymbol{\ell}_t \rangle\right]) \leq & 9\sqrt{KT} + \tfrac{1}{2}\rho^\star \\
& + 3\sqrt{\ln(K)\sum\nolimits_{t=1}^T \mathbb{E}\left[\sum\nolimits_{i=1}^K q_t(i)\rho_t(i)\right]}
\end{aligned}$$

Proof. Fix t and let $F(\cdot) = \eta_t F_t(\cdot)$. To verify the conditions of Lemma 2, recall that a Bregman divergence is non-negative, and that since $\eta_t \leq \eta_{t-1}$ we have that

$$B_{F_t}(\boldsymbol{p}, \boldsymbol{p}_1^\phi) = \frac{1}{\eta_t} B_F(\boldsymbol{p}, \boldsymbol{p}_1^\phi) \geq \frac{1}{\eta_{t-1}} B_F(\boldsymbol{p}, \boldsymbol{p}_1^\phi) = B_{F_{t-1}}(\boldsymbol{p}, \boldsymbol{p}_1^\phi)$$

To verify the conditions of Lemma 1 observe that $\left(\nabla F_t(\boldsymbol{x})\right)_i$ is increasing and concave in $x(i)$ and that $\nabla^2 F_t(\boldsymbol{x})$ positive definite, which implies that F_t is Legendre—see also (Zimmert and Seldin, 2020).

Thus, we may use Lemma 3 to bound

$$\mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{q}_t - \boldsymbol{u}, \widetilde{\boldsymbol{\ell}}_t \rangle\right] \leq \mathbb{E}\left[R_T(\boldsymbol{u})\right] + \mathbb{E}\left[\sum_{t=1}^T (\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av})^\top (\nabla^2 R_t(\boldsymbol{q}_t))^{-1} \widehat{\boldsymbol{\ell}}_t\right]$$

We continue by analysing the cost of implicit exploration. First, we rewrite the loss of the learner by using (Neu, 2015a, equation (5))

$$\sum_{i=1}^K q_t(i)\widetilde{\ell}_t(i) = \ell_t(i_t) - \epsilon_t \sum_{i=1}^K \widetilde{\ell}_t(i)$$

Since $\widetilde{\ell}_t(i) \leq \frac{\mathbb{1}[i_t=i]\ell_t(i)}{q_t(i)}$ we have that $\mathbb{E}\left[\widetilde{\ell}_t(i)\right] \leq 1$. This in turn implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}_t - \boldsymbol{u}, \widetilde{\boldsymbol{\ell}}_t\rangle\right] \geq \mathbb{E}\left[\mathcal{R}_T(\boldsymbol{u})\right] - \sum_{t=1}^{T}\epsilon_t$$

Similarly to (23), we use that $\left(\frac{1}{a}+\frac{1}{b}\right)^{-1} \leq \min\{a,b\}$ for $a, b > 0$. Also, we use (6) and $\ell_t(i) \leq 1$ to write

$$\begin{aligned}
\left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^{\top}\left(\nabla^2 R_t(\boldsymbol{q}_t)\right)^{-1}\widehat{\boldsymbol{\ell}}_t &= \sum_{i=1}^{K}\left(\widehat{L}_t(i) - \widehat{L}_t^{av}(i)\right)\left(\frac{1}{4\eta_t q_t(i)^{3/2}} + \frac{1}{\gamma_t q_t(i)}\right)^{-1}\widehat{\ell}_t(i) \\
&\leq \sum_{i=1}^{K}\left(4\eta_t q_t(i)^{3/2}\widetilde{\ell}_t(i)^2 + \gamma_t q_t(i)\widetilde{\ell}_t(i)\sum_{s\in m_t(i)}\widetilde{\ell}_s(i)\right) \\
&\leq \sum_{i=1}^{K}\left(4\eta_t q_t(i)^{1/2}\widetilde{\ell}_t(i) + \gamma_t\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}\right)
\end{aligned}$$

Let us now study the expectation of the first sum on the right-hand side of the last equation:

$$\mathbb{E}\left[\sum_{i=1}^{K}4\eta_t q_t(i)^{1/2}\widetilde{\ell}_t(i)\right] \leq \mathbb{E}\left[\sum_{i=1}^{K}4\eta_t q_t(i)^{1/2}\right] \leq \eta_t 4\sqrt{K}$$

where we used that $\max_{\boldsymbol{p}\in\triangle}\sum_{i=1}^{K}\sqrt{p(i)} = \sqrt{K}$ (Zimmert and Seldin, 2020).

Since $\epsilon_s \geq \epsilon_t$ for $s \leq t$ we have that $\sum_{s\in m_t(i)}\widetilde{\ell}_t(i) \leq \frac{\rho^\star}{\epsilon_t}$ and thus by using that $\sum_{t=1}^{T}\frac{y_t}{\sqrt{\sum_{s=1}^{t}y_s}} \leq 2\sqrt{\sum_{t=1}^{T}y_t}$ for $y_1 > 0$ and $y_2, \ldots, y_T \geq 0$ we find that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{K}\gamma_t\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}\right] &= \mathbb{E}\left[\frac{\sqrt{\ln(K)}\sum_{i=1}^{K}\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}}{\sqrt{\rho^\star\epsilon_t^{-1} + \sum_{s=1}^{t-1}\sum_{i=1}^{K}\mathbb{1}[i_s=i]\sum_{s'\in m_s(i)}\frac{\mathbb{1}[i_{s'}=i]}{q_{s'}(i)}}}\right] \\
&\leq \mathbb{E}\left[\frac{\sqrt{\ln(K)}\sum_{i=1}^{K}\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}}{\sqrt{\sum_{s=1}^{t}\sum_{i=1}^{K}\mathbb{1}[i_s=i]\sum_{s'\in m_s(i)}\frac{\mathbb{1}[i_{s'}=i]}{q_{s'}(i)}}}\right] \\
&\leq \mathbb{E}\left[2\sqrt{\ln(K)\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}}\right] \\
&\leq 2\sqrt{\ln(K)\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}\right]}
\end{aligned}$$

where the final inequality is due to Jensen's inequality. Following the logic used to bound (19), we can evaluate the expectation:

$$\mathbb{E}\left[\mathbb{1}[i_t=i]\sum_{s\in m_t(i)}\frac{\mathbb{1}[i_s=i]}{q_s(i)}\right] = \mathbb{E}\left[\sum_{i=1}^{K}q_t(i)\rho_t(i)\right]$$

Let us now bound $R_T(\boldsymbol{u})$:

$$R_T(\boldsymbol{u}) \leq \frac{\sqrt{K} - \sum_{i=1}^{K}u(i)}{\eta_T} + \frac{\ln(K)}{\gamma_T} \leq \frac{\sqrt{K}}{\eta_T} + \frac{\ln(K)}{\gamma_T}$$

To complete the proof we combine the above. Let

$$A_T = \sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{K}q_t(i)\rho_t(i)\right]$$

We have:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{q}_t - \boldsymbol{u}, \boldsymbol{\ell}_t\rangle\right] &\leq \mathbb{E}\left[R_T(\boldsymbol{u})\right] + \sum_{t=1}^{T}\epsilon_t + \mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{\boldsymbol{L}}_t - \widehat{\boldsymbol{L}}_t^{av}\right)^\top \left(\nabla^2 R_t(\boldsymbol{q}_t)\right)^{-1}\widehat{\boldsymbol{\ell}}_t\right] \\
&\leq 6\sqrt{KT} + 2\sqrt{\ln(K)A_T} + 2\sqrt{T} + \sqrt{\ln(K)\left(\rho^\star\sqrt{T} + A_T\right)} \\
&\leq 6\sqrt{KT} + 2\sqrt{\ln(K)A_T} + 2\sqrt{T} + \sqrt{\rho^\star\sqrt{T}\ln(K)} + \sqrt{\ln(K)A_T} \\
&\leq 9\sqrt{KT} + 3\sqrt{\ln(K)A_T} + \tfrac{1}{2}\rho^\star
\end{aligned}
$$

where we used that $\sqrt{ab} \leq \frac{1}{2}a + \frac{1}{2}b$ for $a, b > 0$, and that $\ln(K) \leq \sqrt{K}$. □