OPT2019: 11th Annual Workshop on Optimization for Machine Learning

On Adaptivity in Information-constrained Online Learning

Siddharth Mitra SMITRA@CMI.AC.IN
Chennai Mathematical Institute

Aditya Gopalan ADITYA@IISC.AC.IN
Indian Institute of Science

Abstract

We study how to adapt to smoothly-varying environments in well-known online learning problems where acquiring information is expensive. For the problem of label efficient prediction, which is a budgeted version of prediction with expert advice, we present an online algorithm whose regret depends optimally on the number of labels allowed and Q^* (the quadratic variation of the losses of the best action in hindsight), along with a parameter-free counterpart whose regret depends optimally on Q (the quadratic variation of the losses of all the actions). These quantities can be significantly smaller than T (the total time horizon), yielding an improvement over existing, variation-independent results for the problem. We then extend our analysis to handle label efficient prediction with bandit feedback, i.e., label efficient bandits. Our work builds upon the framework of optimistic online mirror descent, and leverages second order corrections along with a carefully designed hybrid regularizer that encodes the constrained information structure of the problem.

1. Introduction

The study of online learning has developed along two concrete lines insofar as modeling the uncertain environment is concerned. On one hand, there is a rich body of work on learning in stochastic environments which often yields performance guarantees that are strong but can closely depend on the stochastic models at hand. On the other hand, much work has been devoted to studying non-stochastic (or arbitrary or adversarial) models of environments from a worst-case point of view which naturally yields rather pessimistic guarantees. Recent efforts have focused on bridging this spectrum of modeling structure in online learning problems as arising from non-stochastic environments with loss function sequences exhibiting adequate temporal regularity [3, 4, 12–15]. In this regard, this paper is an attempt to extend our understanding of adapting to low variation in several standard online learning problems where information comes at a cost, namely label efficient prediction [6], and label efficient bandits.

Problem Setup A label efficient prediction game [6] proceeds for T rounds with $K \leq T$ arms or 'experts'. In each round (time instant) t, the learner selects an arm $i_t \in [K] := 1, 2, \ldots, K$. Simultaneously, the adversary chooses a loss vector $\ell_t \in [0,1]^K$ where $\ell_{t,i}$ is the loss of arm i at time t. At each round, the learner can additionally choose to observe the full loss vector ℓ_t, provided the number of times it has done so in the past has not exceeded a given positive integer $n \leq T$ that represents an information budget or constraint. We work in the *oblivious* adversarial setting where ℓ_t does not depend on the previous actions of the learner $i_1, i_2, \ldots, i_{t-1}$; this is akin to the adversary fixing the (worst-possible) sequence of loss vectors in advance. The learner's goal is to minimize its expected regret defined as $\max_{i^* \in [K]} \mathbb{E}\big[\sum_{t=1}^T \ell_{t,i_t} - \sum_{t=1}^T \ell_{t,i^*}\big]$, where

the expectation is taken with respect to the learner's randomness. Given a convex function $\mathcal{R}$ over Ω, we denote by $D_{\mathcal{R}}$ the Bregman divergence with respect to $\mathcal{R}$ defined as $D_{\mathcal{R}}(x,y) \triangleq \mathcal{R}(x) - \mathcal{R}(y) - \langle \nabla\mathcal{R}(y), x-y\rangle \ \forall x,y \in \Omega$. We denote by ϵ, the fraction of time we are allowed the full loss vector i.e. $\epsilon = n/T$. The ϵ can be seen as a way to model the constraint on information defined by the problem. The quadratic variation for a loss vector sequence $\ell_1, \ldots, \ell_T$ is defined by $Q = \sum_{t=1}^{T} \|\ell_t - \mu_T\|_2^2$ with $\mu_s = \frac{1}{s}\sum_{t=1}^{s} \ell_t$. Additionally, the quadratic variation of the best arm(s) is $Q^* = \sum_{t=1}^{T}(\ell_{t,i^*} - \mu_{T,i^*})^2$ where $\mu_{s,i} = \frac{1}{s}\sum_{t=1}^{s}\ell_{t,i}$ and $i^* = \operatorname{argmin}_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}$.

2. Key Ideas and Algorithms

Optimistic OMD The underlying framework behind our algorithms is that of Online Mirror Descent (OMD) (see, for example [10]). The *vanilla* update rule of (active) mirror descent can be written as: $x_t = \operatorname{argmin}_{x\in\Omega}\{\langle x, \tilde{\ell}_{t-1}\rangle + D_{\mathcal{R}}(x, x_{t-1})\}$. On the other hand, our updates are:

$$x_t = \operatorname*{argmin}_{x\in\Omega}\{\langle x, \epsilon m_t\rangle + D_{\mathcal{R}}(x, x'_t)\} \tag{1}$$

$$x'_{t+1} = \operatorname*{argmin}_{x\in\Omega}\{\langle x, \epsilon\tilde{\ell}_t + a_t\rangle + D_{\mathcal{R}}(x, x'_t)\} \tag{2}$$

where $\epsilon = n/T$, m_t corresponds to *optimistic*[1] estimates of the loss vectors (which we will also refer to as messages), and a_t denotes a second order correction that we explicitly define later. $\tilde{\ell}_t$ is used to denote an (unbiased) estimate of ℓ_t that the learner constructs at time t. Optimistic OMD with second order corrections was first studied in [15], whereas its Follow-the-Regularized-Leader (FTRL) counterpart was introduced earlier by [14]. Both of these approaches build upon the general optimistic OMD framework of [13] and [7]. We define our updates with *scaled* losses and messages, where we reiterate that the scaling factor ϵ reflects the limitation on information. This scaling also impacts our second order corrections which are $\approx \eta\epsilon^2(\tilde{\ell}_t - m_t)^2$, different from the $\eta\epsilon(\tilde{\ell}_t - m_t)^2$ one may expect in light of the analysis done in [15], or the $\eta(\tilde{\ell}_t - m_t)^2$ one would anticipate when following [14]. One may argue that our update rules are equivalent to dividing throughout by ϵ, or put differently, by merging an ϵ into the step size, and this indeed true. However, the point we would like to emphasize is that no matter how one defines the updates, the second order correction a_t can be seen to incorporate the problem dependent parameter ϵ. This tuning of the second order correction based on ϵ is different from what one observes for the full information problem [14] or for bandits [15]. The second order corrections represent a further penalty on arms which are deviating from their respective messages, and these corrections are what enable us to furnish best arm dependent bounds. As usual, the arm we play is still sampled from the distribution x_t given by equation (1).

Challenges & Our Choice of Regularization The inverse propensity weighted loss estimators for label efficient prediction have fixed probabilities of ϵ in the denominator, unlike in bandits where the $x_{t,i}$ in the denominator can be arbitrarily small. Consequently, one may be led to believe that the standard negative entropic regularizer, as is typically used for full information [14], will suffice for the more general but related label efficient prediction. However, maintaining the $|\eta\tilde{\ell}_t| \leq 1$ inequality which is standard in analyses similar to Exp3 imposes a strict bound of $\eta \leq \epsilon$. Since the low quadratic variation, on the other hand, would encourage one to set an aggressive learning rate η, this makes the applicability of the algorithm rather limited, and even then, with marginal gain. Put crisply, it

1. 'Optimistic' is used to denote the fact that we would be best off if these estimates were exactly the upcoming loss. Indeed, if m_t were ℓ_t, it would be equivalent to 1-step lookahead, known to yield low regret.

is desirable that low quadratic variation should lead an algorithm to choose an aggressive learning rate, and negative entropy fails to maintain a 'stability' property (in the sense of Lemma 7), key in obtaining OMD regret bounds, in such situations. The log-barrier regularizer, used by [15] for bandit feedback certainly guarantees this, however using log-barrier blindly translates to a $\sqrt{K}$ dependence on the number of arms K.

These challenges places label efficient prediction with slowly varying losses in a unique position, as one requires enough curvature to ensure stability, yet not let this added curvature significantly hinder exploration. Our solution is to use a hybrid regularizer, that is, a weighted sum of the negative entropic regularizer and the log-barrier regularizer: $\mathcal{R} = {}^1\!/\!_\eta \sum_{i=1}^K x_i \log x_i - {}^1\!/\!_{\eta K} \sum_{i=1}^K \log x_i$. This regularizer has been of recent interest due to the work of [4], and [3], but the weights chosen for both components is highly application-specific and tends to reflect the nature of the problem. As reported above, we only require the log-barrier to guarantee stability, and therefore associate a small (roughly ${}^1\!/\!_{K\eta}$) weight to it and a dominant mass of ${}^1\!/\!_\eta$ to negative entropy. This fact is revealed in the analysis where we use the log-barrier component solely to satisfy Lemmas 6 and 7, following which it is essentially dispensed. The additional ${}^1\!/\!_K$ factor part of the log-barrier weight is carefully chosen to exactly cancel the K in the leading $K \log T$ term generated by the log-barrier component, and consequently, not have a $\sqrt{K}$ dependence on the number of arms in the final regret bound.

Reservoir Sampling When considering quadratic variation as a measure of adaptivity, a natural message to pass is the mean of the previous loss history, that is $m_t = \mu_{t-1} = {}^1\!/\!_{t-1} \sum_{s=1}^{t-1} \ell_s$. However, the constraint on information prohibits us from having the full history, and we therefore have to settle for some estimate of the mean. Reservoir sampling, first used in [12], solves this very problem. Specifically, by allocating roughly $k(1 + \log T)$ rounds for reservoir sampling (where we choose k to be $\log T$), reservoir sampling gives us estimates $\tilde{\mu}_t$ such that $\mathbb{E}[\tilde{\mu}_t] = \mu_t$, and $\mathrm{Var}[\tilde{\mu}_t] = {}^Q\!/\!_{kt}$.

Algorithm 1 ADAPTIVE LABEL EFFICIENT PREDICTION

Input: $\mathcal{R} = {}^1\!/\!_\eta \sum_{i=1}^K x_i \log x_i - {}^1\!/\!_{\eta K} \sum_{i=1}^K \log x_i$, η , ϵ
Initialize: $x'_1 = \operatorname{argmin}_{x \in \Omega} \mathcal{R}(x)$
for $t = 1, 2, \ldots, T$ **do**
  $d_t \sim \mathrm{Bern}(\epsilon)$
  $x_t = \operatorname{argmin}_{x \in \Omega} \{\langle x, \epsilon m_t \rangle + D_{\mathcal{R}}(x, x'_t)\}$
  Play $i_t \sim x_t$, and if $d_t = 1$, observe ℓ_t
  Construct $\tilde{\ell}_t = \frac{(\ell_t - m_t)}{\epsilon} \mathbb{1}_{\{d_t = 1\}} + m_t$
  Let $a_t = 6\eta\epsilon^2(\tilde{\ell}_t - m_t)^2$
  Update: $x'_{t+1} = \operatorname{argmin}_{x \in \Omega} \left\{ \left\langle x, \epsilon\tilde{\ell}_t + a_t \right\rangle + D_{\mathcal{R}}(x, x'_t) \right\}$
end

Algorithm 1 builds upon the ideas presented above and as stated, is specifically for the label efficient prediction problem discussed thus far. The algorithm for label efficient bandits discussed in subsection 3.1 is based on Algorithm 1, although with a few minor differences which we specify later. Note that throughout the paper, the random variable $d_t = 1$ signifies that we ask for feedback at time t, and is 0 otherwise. Also, in the interest of brevity, we have excluded the explicit mentioning of the reservoir sampling steps. Additionally, note that we consider not exceeding the budget of n in expectation, however, there is a standard reduction to get a high probability guarantee which can be found in [5].

3. Results and Analysis

We now give a general regret result for the OMD updates (1) and (2). The proofs for all results in this section appear in the supplementary material.

Lemma 1 *For the update rules (1) and (2), if:*

$$\langle x_t - x'_{t+1}, \epsilon(\tilde{\ell}_t - m_t) + a_t \rangle - \langle x_t, a_t \rangle \leq 0 \tag{3}$$

then, for all $u \in \Omega$, we have:

$$\langle x_t - u, \tilde{\ell}_t \rangle \leq {}^1\!/\!_\epsilon \left(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \langle u, a_t \rangle - P_t \right), \tag{4}$$

where $P_t \triangleq D_{\mathcal{R}}(x'_{t+1}, x_t) + D_{\mathcal{R}}(x_t, x'_t) \geq 0$

When $a_t = 0$ is employed in the updates (1) and (2), i.e., no second order corrections, the first term in (3) can directly be handled using Hölder's inequality (in some norm where $\mathcal{R}$ is strongly convex). Doing so allows us to cancel the unwanted $\|x_t - x'_{t+1}\|^2$ term using the $D_{\mathcal{R}}(x'_{t+1}, x_t)$ term in P_t (which follows by strong convexity) while retaining the crucial $\|(\tilde{\ell}_t - m_t)\|^2$ variance term. However, with general second order corrections ($a_t \geq 0$), the key variance term is $\langle u, a_t \rangle$ as it corresponds to the best arm's second moment under a suitably chosen u and the responsibility of cancelling the entire first term of (3) now falls upon $\langle x_t, a_t \rangle$. Under limited information, negative entropy is unable to maintain this and we therefore have to incorporate the log barrier function (see also [15]).

Theorem 2 *For $a_t = 6\eta\epsilon^2(\tilde{\ell}_t - m_t)^2$, $\tilde{\ell}_t = \frac{(\ell_t - m_t)}{\epsilon}\mathbb{1}_{\{d_t=1\}} + m_t$, $\epsilon = {}^n\!/\!_T$ and $\eta \leq {}^1\!/\!_{162K}$ where the sequence of messages m_t are generated using the reservoir sampling scheme, the expected regret of Algorithm 1 satisfies $\mathbb{E}[R_T] \leq \frac{\log K + \log T}{\epsilon\eta} + 18\eta Q^*$. Furthermore, if $\epsilon Q^* \geq 1458K^2 \log KT$, then $\mathbb{E}[R_T] = \mathcal{O}\left(\sqrt{(Q^* T \log K)/n}\right)$ with an optimal choice of η.*

Consider a concrete example of a game played for time T, where we anticipate $Q^* \approx \sqrt{T}$ and $n \approx \sqrt{T}$. In this scenario, if we were to run the standard label efficient prediction algorithm [6] which attains $\mathcal{O}(\sqrt{(T^2 \log K)/n})$ regret, we would get a regret bound of $\mathcal{O}\left(T^{3/4}\right)$; following an FTRL with negative entropy[2]-based strategy would be inapplicable in this setting due to the constraint we highlight in section 2, however, Algorithm 1 would incur $\sqrt{T}$ regret – a marked improvement. Also note that because of the full vector feedback, we don't incur any additive penalty for reservoir sampling as we don't have to allocate any rounds *exclusively* for it.

Unconditional & Parameter-Free Algorithms Theorem 2 is slightly restricted in scope, due to the lower bound required on ϵQ^*, in its ability to attain the optimal regret scaling with quadratic variation. It also assumes prior knowledge of T, Q and Q^* when optimising for the fixed step size η. We address both of these questions and note that optimistic OMD without second order corrections– an algorithm defined by updates (1) and (2) with $a_t = 0$ obtains $\mathcal{O}(\sqrt{(QT \log K)/n})$ regret under *all* scenarios. The trade-off however being that we are now penalized by Q instead of Q^*. With $a_t = 0$, we are also able to yield a parameter-free version of the algorithm. We discuss both of these in subsection A.1 of the supplementary material.

2. As done in [14] for prediction with experts

3.1. Label Efficient Bandits

Instead of receiving the full loss vector, the learner now only receives the loss of the played arm i_t, i.e. the i_tth coordinate of ℓ_t. We continue here with updates (1) and (2) but with $\mathcal{R} = {}^1\!/\!_\eta \sum_{i=1}^K \log {}^1\!/\!_{x_i}$, and appropriately defined loss estimators and second order corrections. Also note that due to bandit feedback, we now have to reserve certain rounds solely for reservoir sampling. This is reflected in the additive $K(\log T)^2$ term in the bound below. The proof for Theorem 3 can be found in Appendix B.

Theorem 3 *For $a_{t,i} = 6\eta\epsilon^2 x_{t,i}(\tilde{\ell}_t - m_t)^2$, $\tilde{\ell}_t = \frac{\ell_t - m_t}{\epsilon x_{t,i}}\mathbb{1}_{\{d_t=1, i_t=i\}} + m_{t,i}$, $\epsilon = {}^n\!/\!_T$ and $\eta \le {}^1\!/\!_{162K}$ where the sequence of messages m_t are given by reservoir sampling, the regret of Algorithm 1 modified for label efficient bandits satisfies $\mathbb{E}[R_T] \le {}^{K\log T}\!/\!_{\epsilon\eta} + 18\eta Q^* + K(\log T)^2$.*

4. Lower Bounds

By capturing both the constraint on information as well as the quadratic variation of the loss sequence, our lower bounds for label efficient prediction and label efficient bandits generalize and improve upon existing lower bounds. We extend the lower bounds for label efficient prediction to further incorporate the quadratic variation of the loss sequence and enhance the quadratic variation dependent lower bounds for multi-armed bandits to also include the constraint on information by bringing in the number of labels the learner can observe (n). The proofs for this section can be found in Appendix C.

Recall the quadratic variation for a given loss sequence: $Q = \sum_{t=1}^T \|\ell_t - \mu_T\|_2^2 \le {}^{TK}\!/\!_4$. Now, for $\alpha \in [0, {}^1\!/\!_4]$ define an α-variation ball as: $\mathcal{V}_\alpha \triangleq \{\{\ell_t\}_{t=1}^T : {}^Q\!/\!_{TK} \le \alpha\}$. Theorems 4 and 5, after incorporating $Q \le \alpha TK$ give us lower bounds of $\Omega(\sqrt{{}^{(QT\log(K-1))}\!/\!_{Kn}})$ and $\Omega(\sqrt{{}^{QT}\!/\!_n})$ respectively. Our corresponding upper bounds are $\mathcal{O}(\sqrt{{}^{(QT\log K)}\!/\!_n})$ and $\mathcal{O}(\sqrt{{}^{QTK}\!/\!_n})$.[3] Comparing the two tells us that our strategies are optimal in their dependence on Q and on the constraint in information indicated by n. There is however a gap of $\sqrt{K}$. This gap was mentioned in [9] for the specific case of the multi-armed bandit problem, and was closed recently in [3]. Barring the easy to see $\sqrt{{}^{(Q\log K)}\!/\!_K}$ lower bound for prediction with expert advice (which is also what Theorem 4 translates to for $n = T$), we are unaware of other fundamental Q based lower bounds for prediction with expert advice. The upper bounds for prediction with expert advice however are of $\mathcal{O}(\sqrt{Q\log K})$ ([11], [14] etc.), and this again suggests the $\sqrt{K}$ gap. Closing this for prediction with expert advice, label efficient prediction and for label efficient bandits remains open, as does the question of finding Q^* dependent lower bounds.

Theorem 4 *Let $K \ge 2$, $T \ge n \ge \max\{32\log(K-1),\ 256\log T\}$ and $\alpha \in \left[\max\left\{\frac{32\log T}{n}, \frac{8\log(K-1)}{n}\right\}, \frac{1}{4}\right]$. Then, for any randomized strategy for the label efficient prediction problem, $\max_{\{\ell_t\}\in v_\alpha} \mathbb{E}[R_T] \ge 0.36T\sqrt{{}^{(\alpha\log(K-1))}\!/\!_n}$ where expectation is taken with respect to the internal randomization available to the algorithm.*

Theorem 5 *Let $K \ge 2$, $T \ge n \ge \max\{32K,\ 384\log T\}$ and $\alpha \in \left[\max\left\{{}^{2c\log T}\!/\!_n, {}^{8K}\!/\!_n\right\}, \frac{1}{4}\right]$ with $c = (4/9)^2(3\sqrt{5}+1)^2 \le 12$. Then, for any randomized strategy for the label efficient bandit problem, $\max_{\{\ell_t\}\in v_\alpha} \mathbb{E}[R_T] \ge 0.04T\sqrt{{}^{\alpha K}\!/\!_n}$ where expectation is taken with respect to the internal randomization available to the algorithm.*

3. We upper bound all of our Q^* dependent upper bounds by Q so as to consistently compare with the lower bounds. Note that Q^* and Q are in general incomparable and all that be said is that $Q^* \le Q$.

References

[1] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

[2] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities: A nonasymptotic theory of independence. 01 2013.

[3] Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *ALT*, 2017.

[4] Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. *CoRR*, abs/1901.10604, 2019.

[5] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

[6] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

[7] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT*, volume 23 of *JMLR Proceedings*, pages 6.1–6.20. JMLR.org, 2012.

[8] Yuan-Shih Chow and Henry Teicher. Probability theory : independence, interchangeability martingales / yuan shih chow, henry teicher. *SERBIULA (sistema Librum 2.0)*, 01 1980.

[9] Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *NIPS*, pages 1190–1198, 2016.

[10] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2:157–325, 01 2016. doi: 10.1561/2400000013.

[11] Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: regret bounded by variation in costs. *Machine Learning*, 80(2):165–188, Sep 2010.

[12] Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *J. Mach. Learn. Res.*, 12: 1287–1311, July 2011. ISSN 1532-4435.

[13] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. *arXiv preprint arXiv:1208.3728*, 2012.

[14] Jacob Steinhardt and Percy S. Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *ICML*, 2014.

[15] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory, COLT 2018*, 2018.

Appendix A. Label Efficient Prediction Main Proofs

We will now prove Lemma 1 and Theorem 2. Following this, we will discuss the unconditional algorithm for label efficient prediction mentioned in the main text.

Proof of Lemma 1 Let Ω be a convex compact set in $\mathbb{R}^K$, $\mathcal{R}$ be a convex function on Ω, x' be an arbitrary point in Ω, c be any point in $\mathbb{R}^K$, and $x^* = \operatorname{argmin}_{x\in\Omega}\{\langle x, c\rangle + D_{\mathcal{R}}(x, x')\}$. Then, for any $u \in \Omega$, we have (see for example [1]) :

$$\langle x^* - u, c\rangle \leq D_{\mathcal{R}}(u, x') - D_{\mathcal{R}}(u, x^*) - D_{\mathcal{R}}(x^*, x')$$

Applying this on our update rules (1) and (2) gives us:

$$\langle x_t - x'_{t+1}, \epsilon m_t\rangle \leq D_{\mathcal{R}}(x'_{t+1}, x'_t) - D_{\mathcal{R}}(x'_{t+1}, x_t) - D_{\mathcal{R}}(x_t, x'_t). \tag{5}$$

and

$$\langle x'_{t+1} - u, \epsilon\tilde{\ell}_t + a_t\rangle \leq D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) - D_{\mathcal{R}}(x'_{t+1}, x'_t); \tag{6}$$

where we chose $u = x'_{t+1}$ when applying it to update rule (1). Now observe that:

$$\begin{aligned}
&\langle x_t - u, \epsilon\tilde{\ell}_t\rangle \\
&= \langle x_t - u, \epsilon\tilde{\ell}_t + a_t\rangle - \langle x_t, a_t\rangle + \langle u, a_t\rangle \\
&= \langle x_t - x'_{t+1}, \epsilon\tilde{\ell}_t + a_t\rangle - \langle x_t, a_t\rangle + \langle x'_{t+1} - u, \epsilon\tilde{\ell}_t + a_t\rangle + \langle u, a_t\rangle \\
&= \langle x_t - x'_{t+1}, \epsilon\tilde{\ell}_t + a_t - \epsilon m_t\rangle - \langle x_t, a_t\rangle + \langle x'_{t+1} - u, \epsilon\tilde{\ell}_t + a_t\rangle + \langle x_t - x'_{t+1}, \epsilon m_t\rangle + \langle u, a_t\rangle
\end{aligned} \tag{7}$$

Combining the above inequalities with equation (3) gives us

$$\langle x_t - u, \tilde{\ell}_t\rangle \leq \frac{1}{\epsilon}\left(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \langle u, a_t\rangle - P_t\right), \tag{8}$$

where $P_t \triangleq D_{\mathcal{R}}(x'_{t+1}, x_t) + D_{\mathcal{R}}(x_t, x'_t) \geq 0$ (by non-negativity of Bregman divergence). ■

We will now proceed to prove a series of lemmas which will build towards the proof of Theorem 2. For any point $u \in \mathbb{R}^K$, we define the local norm at x with respect to $\mathcal{R}$ as $\|u\|_x = \sqrt{u^\top \nabla^2\mathcal{R}(x)u}$ and the corresponding dual norm as $\|u\|_{x,*} = \sqrt{u^\top \nabla^{-2}\mathcal{R}(x)u}$.

Lemma 6 *For some radius $r > 0$, define the ellipsoid $\mathcal{E}_x(r) = \left\{u \in \mathbb{R}^K : \|u - x\|_x \leq r\right\}$. If $x' \in \mathcal{E}_x(1)$, $\eta \leq \frac{1}{81K}$, then, for all $i \in [K]$, we have $\frac{x'_i}{x_i} \leq 10/9$. Additionally, $\|u\|_{x'} \geq \frac{9}{10}\|u\|_x$ for all $u \in \mathbb{R}^K$.*

Proof of Lemma 6 As $x' \in \mathcal{E}_x(1)$, we can say that $\sum_{i=1}^K \frac{1}{\eta}(x'_i - x_i)^2 \left(\frac{1}{x_i} + \frac{1}{Kx_i^2}\right) \leq 1$ which further implies $\sum_{i=1}^K \frac{1}{\eta K}\frac{(x'_i - x_i)^2}{x_i^2} \leq 1$. Hence, we have $\frac{|x'_i - x_i|}{x_i} \leq \sqrt{\eta K} \;\; \forall i$. Now, since $\eta \leq \frac{1}{81K}$, the first part of the lemma follows. Further observe $\|u\|_{x'} = \sqrt{\frac{1}{\eta}\sum_{i=1}^K u_i^2 \left(\frac{1}{x'_i} + \frac{1}{Kx_i'^2}\right)} \geq \frac{1}{10/9}\|u\|_x = \frac{9}{10}\|u\|_x$. ■

Lemma 7 *Let x_t and x'_t correspond to our update rules (1) and (2) and suppose $\eta \leq \frac{1}{81K}$. Then, if $\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} \leq \frac{1}{3}$, we have $x' \in \mathcal{E}_x(1)$.*

Proof of Lemma 7 Let us rewrite our update rules (1) and (2) in the following way:

$$x_t = \operatorname*{argmin}_{x\in\Omega} F_t(x) \text{ where } F_t(x) = \left\{\langle x, \epsilon m_t\rangle + D_{\mathcal{R}}(x, x'_t)\right\}$$
$$x'_{t+1} = \operatorname*{argmin}_{x\in\Omega} F'_{t+1}(x) \text{ where } F'_{t+1}(x) = \left\{\langle x, \epsilon\tilde{\ell}_t + a_t\rangle + D_{\mathcal{R}}(x, x'_t)\right\}$$

Because of the convexity of F'_t , to prove our claim, it is sufficient to show that $F'_{t+1}(u) \geq F'_{t+1}(x_t)$ for all points u on the boundary of the ellipsoid. By Taylor's theorem, we know that $\exists\, \xi$ on the line segment between u and x_t such that:

$$\begin{aligned}
F'_{t+1}(u) &= F'_{t+1}(x_t) + \langle \nabla F'_{t+1}(x_t), u - x_t\rangle + \frac{1}{2}(u - x_t)^\top \nabla^2 F'_{t+1}(\xi)(u - x_t) \\
&= F'_{t+1}(x_t) + \langle \epsilon(\tilde{\ell}_t - m_t) + a_t, u - x_t\rangle + \langle \nabla F_t(x_t), u - x_t\rangle \\
&\quad + \frac{1}{2}(u - x_t)^\top \nabla^2 \mathcal{R}(\xi)(u - x_t) \\
&\geq F'_{t+1}(x_t) + \langle \epsilon(\tilde{\ell}_t - m_t) + a_t, u - x_t\rangle + \frac{1}{2}\|u - x_t\|_\xi^2 \\
&\geq F'_{t+1}(x_t) + \langle \epsilon(\tilde{\ell}_t - m_t) + a_t, u - x_t\rangle + \frac{81}{200}\|u - x_t\|_{x_t}^2 \\
&\geq F'_{t+1}(x_t) - \left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*}\|u - x_t\|_{x_t} + \frac{1}{3}\|u - x_t\|_{x_t}^2 \\
&= F'_{t+1}(x_t) - \left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} + \frac{1}{3} \qquad (\text{as } \|u - x_t\|_{x_t} = 1) \\
&\geq F'_{t+1}(x_t).
\end{aligned}$$

Where the first inequality follows from the optimality of x_t, the second from Lemma (6), the third from Hölder's inequality, and the last by the assumption of this lemma.

■

Lemma 8 *Let x_t and x'_t be defined as in our update rules (1) and (2). Additionally, suppose $\eta \leq \frac{1}{81K}$. Then, if $\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} \leq \frac{1}{3}$, we have that $\left\|x'_{t+1} - x_t\right\|_{x_t} \leq 3\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*}$.*

Proof of Lemma 8 We will begin by defining $F_t(x)$ and $F'_{t+1}(x)$ as above. Then we have that:

$$\begin{aligned} F'_{t+1}(x_t) - F'_{t+1}(x'_{t+1}) &= \langle x_t - x'_{t+1},\ \epsilon(\tilde{\ell}_t - m_t) + a_t\rangle + F_t(x_t) - F_t(x'_{t+1}) \\ &\leq \langle x_t - x'_{t+1},\ \epsilon(\tilde{\ell}_t - m_t) + a_t\rangle \\ &\leq \left\|x_t - x'_{t+1}\right\|_{x_t}\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} \end{aligned} \tag{9}$$

By Taylor's theorem and the optimality of x'_{t+1}, we again have that,

$$\begin{aligned} F'_{t+1}(x_t) - F'_{t+1}(x'_{t+1}) &= \langle \nabla F'_{t+1}(x'_{t+1}), x_t - x'_{t+1}\rangle + \frac{1}{2}(x_t - x'_{t+1})^\top \nabla^2 F'_{t+1}(\xi)(x_t - x'_{t+1}) \\ &\geq \frac{1}{2}\left\|x_t - x'_{t+1}\right\|_\xi^2 \\ &\geq \frac{1}{3}\left\|x_t - x'_{t+1}\right\|_{x_t}^2 \end{aligned} \tag{10}$$

where the last inequality again follows using the same arguments as done in Lemma 7. Combining (9) and (10) proves the claimed result.

■

Lemma 9 *For $a_t = 6\eta\epsilon^2(\tilde{\ell}_t - m_t)^2$, $\tilde{\ell}_t = \frac{(\ell_t - m_t)}{\epsilon}\mathbb{1}_{\{d_t=1\}} + m_t$, $\eta \leq \frac{1}{162K}$ we have that $\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} \leq \frac{1}{3}$.*

Proof of Lemma 9

$$\begin{aligned} \left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*}^2 &= \sum_{i=1}^K \frac{\eta}{\frac{1}{x_i} + \frac{1}{Kx_i^2}}\left(\epsilon(\tilde{\ell}_{t,i} - m_{t,i}) + 6\eta\epsilon^2(\tilde{\ell}_{t,i} - m_{t,i})^2\right)^2 \\ &= \eta\sum_{i=1}^K \frac{\epsilon^2(\tilde{\ell}_{t,i} - m_{t,i})^2}{\frac{1}{x_i} + \frac{1}{Kx_i^2}}\left[1 + 12\eta\epsilon(\tilde{\ell}_{t,i} - m_{t,i}) + 36\eta^2\epsilon^2(\tilde{\ell}_{t,i} - m_{t,i})^2\right] \\ &\leq 2\eta\epsilon^2\sum_{i=1}^K(\tilde{\ell}_{t,i} - m_{t,i})^2 x_i \\ &\leq 2\eta \\ &\leq \frac{1}{9} \end{aligned}$$

The above inequalities follow by observing that $|\epsilon(\tilde{\ell}_{t,i} - m_{t,i})| = |(\ell_t - m_t)\mathbb{1}_{\{d_t=1\}}| \leq 1$ along with the assumption on η.

■

Proof of Theorem 2 We will first show that our choices of loss vectors, messages, and corrections obey the condition of Lemma 1. To this end, observe that:

$$\begin{aligned}\langle x_t - x'_{t+1}, \epsilon(\tilde{\ell}_t - m_t) + a_t\rangle &\le \left\|x_t - x'_{t+1}\right\|_{x_t} \left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} \\ &\le 3\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|^2_{x_t,*} \\ &\le 3\eta \sum_{i=1}^{K} \frac{\epsilon^2(\tilde{\ell}_t - m_t)^2}{\frac{1}{x_i} + \frac{1}{Kx_i^2}} \left[1 + 12\eta\epsilon(\tilde{\ell}_t - m_t) + 36\eta^2\epsilon^2(\tilde{\ell}_t - m_t)^2\right] \\ &\le 6\eta\epsilon^2 \sum_{i=1}^{K} x_{t,i}(\tilde{\ell}_{t,i} - m_{t,i})^2 = \langle x_t, a_t\rangle\end{aligned}$$

where the first inequality follows from Hölder's inequality, the second from Lemma (8), and the last 2 are as done in the proof of Lemma 9.

We can therefore proceed to sum both sides of the result of Lemma 1 over t to get:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle\right] \le \frac{1}{\epsilon}\sum_{t=1}^{T}\mathbb{E}\left[\left(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \langle u, a_t\rangle\right)\right]$$

Now we can see that the first 2 terms on the right hand side will telescope to yield a remaining term of $D_{\mathcal{R}}(u, x'_1)$. We will pick $u = (1 - \frac{1}{T})\boldsymbol{e}_{i^*} + \frac{1}{KT}\mathbf{1}$ instead of simply $\boldsymbol{e}_{i^*}$ so as to ensure that the log barrier component is well defined. Hence we will have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle\right] \le \frac{1}{\epsilon}\left(D_{\mathcal{R}}(u, x'_1) + \mathbb{E}\left[\sum_{t=1}^{T}\langle u, a_t\rangle\right]\right) \tag{11}$$

$$\begin{aligned}D_{\mathcal{R}}(u, x'_1) &= \mathcal{R}(u) - \mathcal{R}(x'_1) - \langle\nabla\mathcal{R}(x'_1), u - x'_1\rangle \\ &= \mathcal{R}(u) - \mathcal{R}(x'_1) \le \frac{\log T}{\eta} + \frac{\log K}{\eta} = \frac{\log K + \log T}{\eta}\end{aligned}$$

This choice of u will also introduce an additional term in regret of $\mathbb{E}\frac{1}{T}\sum_{t=1}^{T}\langle x'_1 - e_{i^*}, \tilde{\ell}_t + a_t\rangle$, but as can be seen in [15], this term is $\mathcal{O}(1)$.

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle u, a_t\rangle\right] = 6\eta\epsilon^2\mathbb{E}\left[\sum_{t=1}^{T}\frac{(\ell_{t,i^*} - m_{t,i^*})^2}{\epsilon^2}\mathbb{1}_{\{d_t=1\}}\right] \tag{12}$$

$$\le 18\eta\epsilon\left[\sum_{t=1}^{T}(\ell_{t,i^*} - \mu_{t,i^*})^2 + \sum_{t=1}^{T}(\mu_{t,i^*} - \mu_{t-1,i^*})^2 + \mathbb{E}\left[\sum_{t=1}^{T}(\mu_{t-1,i^*} - \tilde{\mu}_{t-1,i^*})^2\right]\right] \tag{13}$$

The first and third terms of (13) can be bounded using Lemmas 10 and 11 from [12] and are order $\mathcal{O}(Q_{T,i^*}+1)$. The middle term above is $\mathcal{O}(1)$ from Lemma 18 in [15]. Therefore, substituting everything back into (11), we have that:

$$\mathbb{E}\left[R_T\right] \leq \frac{\log K + \log T}{\epsilon\eta} + 18\eta Q^* \tag{14}$$

■

A.1. Unconditional and Parameter-Free Algorithms

The updates we consider for OMD without second order corrections, as mentioned briefly in the main text are the following:

$$x_t = \operatorname*{argmin}_{x\in\Omega}\{\langle x, \epsilon m_t\rangle + D_{\mathcal{R}}(x, x'_t)\} \tag{15}$$

$$x'_{t+1} = \operatorname*{argmin}_{x\in\Omega}\{\langle x, \epsilon\tilde{\ell}_t\rangle + D_{\mathcal{R}}(x, x'_t)\} \tag{16}$$

With $a_t = 0$, the ϵ term can be folded into the regularizer and the updates reduce to the ones studied in [13].

For updates (15) and (16), we have the following analogue of Lemma 1, and then consequently, the analogue of Theorem 2. We include these here in the interest of completeness, but equivalent statements can be found in [13].

Lemma 10 *For any $u \in \Omega$, updates (15) and (16) guarantee that:*

$$\langle x_t - u, \tilde{\ell}_t\rangle \leq \frac{1}{\epsilon}\Bigg(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) \\ + \langle x_t - x'_{t+1}, \epsilon\tilde{\ell}_t - \epsilon m_t\rangle - D_{\mathcal{R}}(x'_{t+1}, x_t) - D_{\mathcal{R}}(x_t, x'_t)\Bigg).$$

Proof of Lemma 10 We will proceed similarly to the proof of Lemma 2 and rewrite (7) with $a_t = 0$:

$$\langle x_t - u, \epsilon\tilde{\ell}_t\rangle = \langle x_t - x'_{t+1}, \epsilon\tilde{\ell}_t - \epsilon m_t\rangle + \langle x'_{t+1} - u, \epsilon\tilde{\ell}_t\rangle + \langle x_t - x'_{t+1}, \epsilon m_t\rangle$$

We will again use the inequalities (5) and (6) (with $a_t = 0$) to get:

$$\langle x_t - u, \epsilon\tilde{\ell}_t\rangle \leq \langle x_t - x'_{t+1}, \epsilon\tilde{\ell}_t - \epsilon m_t\rangle + D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) - D_{\mathcal{R}}(x'_{t+1}, x_t) - D_{\mathcal{R}}(x_t, x'_t)$$

which proves the lemma after rearranging the ϵ.

■

Theorem 11 *For $\mathcal{R} = \frac{1}{\eta}\sum_{i=1}^{K} x_i \log x_i$, $\tilde{\ell}_t = \frac{(\ell_t - m_t)}{\epsilon}\mathbb{1}_{\{d_t=1\}} + m_t$, $\epsilon = {}^{n}/_{T}$ and $\eta > 0$, where the sequence of messages are generated using the reservoir sampling scheme, Algorithm 1 with $a_t = 0$ yields:*

$$\mathbb{E}[R_T] \leq \frac{\log K}{\eta\epsilon} + \frac{\eta Q}{2}.$$

Optimally tuning η yields a $\mathcal{O}\left(\sqrt{{}^{QT\log K}/_{n}}\right)$ bound.

Proof of Theorem 11

$$\begin{aligned}\epsilon\langle x_t - u, \tilde{\ell}_t\rangle &\leq D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \langle x_t - x'_{t+1}, \epsilon\tilde{\ell}_t - \epsilon m_t\rangle - D_{\mathcal{R}}(x'_{t+1}, x_t) - D_{\mathcal{R}}(x_t, x'_t)\\ &\leq D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \frac{1}{2\eta}\|x_t - x'_{t+1}\|_2^2 + \frac{\eta}{2}\|\epsilon\tilde{\ell}_t - \epsilon m_t\|_2^2 - D_{\mathcal{R}}(x'_{t+1}, x_t)\\ &\leq D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \frac{\eta\epsilon^2}{2}\|\tilde{\ell}_t - m_t\|_2^2\end{aligned}$$

where the first inequality follows from Lemma 10, the second one follows from Hölder's inequality and the non-negativity of the Bregman divergence, and the final one from the strong convexity of negative entropy in the ℓ_2 norm. We therefore have that $\langle x_t - u, \tilde{\ell}_t\rangle \leq \frac{1}{\epsilon}\big(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \frac{\eta\epsilon^2}{2}\|\tilde{\ell}_t - m_t\|_2^2\big)$. Now summing both sides over t will yield:

$$\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle \leq \frac{D_{\mathcal{R}}(u, x'_1)}{\epsilon} + \frac{\eta\epsilon}{2}\sum_{t=1}^{T}\|\tilde{\ell}_t - m_t\|_2^2 \tag{17}$$

$$\leq \frac{\log K}{\eta\epsilon} + \frac{\eta\epsilon}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}(\tilde{\ell}_{t,i} - m_{t,i})^2 \tag{18}$$

Now, substituting the stated estimators, unravelling the right hand side similar to the analysis of (13) and taking expectation will yield the $\frac{\log K}{\eta\epsilon} + \frac{\eta Q}{2}$ upper bound.

■

Parameter-Free Algorithms Algorithm 1 along with the discussion above assumes knowledge of T, Q and Q^* when optimising for the fixed step size η. This is often not possible and we now discuss the extent to which we can obtain parameter-free algorithms. We claim that we can choose η adaptively for the Q dependent bound presented in Theorem 11, and show this in Theorem 12[4]. It remains open whether a Q^* dependent bound (or in general, any non-monotone dependent bound) can be made parameter free for even the standard prediction with expert advice problem. The challenge is essentially that our primary tool to sidestep prior knowledge of a parameter– the doubling trick is inapplicable for non-monotone quantities.

Even freeing algorithms from prior knowledge of non-decreasing arm dependent quantities, such as $\max_i Q_i$ remains open for limited information setups (i.e. anything outside prediction with expert advice) due to the lack of a clear auxiliary term one can observe.

In Algorithm 2, we proceed in epochs (or rounds) such that η remains fixed per epoch. Denote by η_α the value of η in epoch α. We write T_α to be the first time instance in epoch α.

4. Note that similarly to [12] we still assume knowledge of T, but this can be circumvented using standard tricks.

Algorithm 2 PARAMETER FREE ADAPTIVE LABEL EFFICIENT PREDICTION

Initialize: $\eta = \frac{\sqrt{2\log K}}{\epsilon}, T_1 = 1, t = 1.$
for $\alpha = 1, 2, \ldots$ **do**
  $x'_t = \operatorname{argmin}_{x\in\Omega} \mathcal{R}(x)$
  while $t \le T$ **do**
    Draw $d_t \sim \text{Bern}(\epsilon)$, update x_t according to (15)
    Play $i_t \sim x_t$ and if $d_t = 1$, observe ℓ_t
    Update x'_{t+1} according to (16)
    if $\sum_{s=T_\alpha}^{t}\sum_{i=1}^{K}(\tilde{\ell}_{s,i} - m_{s,i})^2 \ge \frac{2\log K}{\epsilon^2\eta_{\alpha-1}^2}$ **then**
      $\eta \leftarrow \eta/2, T_{\alpha+1} \leftarrow t, t \leftarrow t+1$
      break
    end
    $t \leftarrow t+1$
  end
end

Theorem 12 *For the conditions mentioned in Theorem 11, Algorithm 2 (a parameter free algorithm) achieves:*

$$\mathbb{E}[R_T] \le \mathcal{O}\left(\sqrt{(QT\log K)/n} + \sqrt{\log K}\right).$$

Proof of Theorem 12 We start from (18) and get the following for some epoch α:

$$\sum_{t=T_\alpha}^{T_{\alpha+1}-1}\langle x_t - u, \tilde{\ell}_t\rangle \le \frac{1}{\epsilon}\left[\frac{\log K}{\eta_\alpha} + \frac{\eta_\alpha\epsilon^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}(\tilde{\ell}_{t,i} - m_{t,i})^2\right]$$
$$= \mathcal{O}\left(\frac{\log K}{\epsilon\eta_\alpha}\right)$$

We can consequently write:

$$\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle \le \sum_{\alpha=0}^{\alpha^*}\mathcal{O}\left(\frac{\log K}{\epsilon\eta_\alpha}\right) \le \mathcal{O}\left(2^{\alpha^*}\sqrt{\log K}\right)$$

where α^* is the epoch at T. Now we also know that epoch $\alpha^* - 1$ has completed, hence:

$$\sum_{t=T_{\alpha^*-1}}^{T_{\alpha^*}-1}\sum_{i=1}^{K}(\tilde{\ell}_{t,i} - m_{t,i})^2 \ge \frac{2\log K}{\epsilon^2\eta_{\alpha^*-1}^2} = \Omega\left(2^{2\alpha^*}\right)$$

So, we can write the entire bound as

$$\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle \le \mathcal{O}\left(2^{\alpha^*}\sqrt{\log K}\right) \le \mathcal{O}\left(\sqrt{\log K\sum_{t=T_{\alpha^*-1}}^{T_{\alpha^*}}\sum_{i=1}^{K}(\tilde{\ell}_{t,i} - m_{t,i})^2}\right)$$

$$\leq \mathcal{O}\left(\sqrt{\log K \sum_{t=1}^{T}\sum_{i=1}^{K}(\tilde{\ell}_{t,i}-m_{t,i})^2}\right)$$

Also consider the case when $\alpha^* = 0$, where $\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle \leq \sqrt{\log K}$. Combining the above 2 cases, we get:

$$\sum_{t=1}^{T}\langle x_t - u, \tilde{\ell}_t\rangle \leq \mathcal{O}\left(\sqrt{\log K \sum_{t=1}^{T}\sum_{i=1}^{K}(\tilde{\ell}_{t,i}-m_{t,i})^2} + \sqrt{\log K}\right)$$

Taking expectation and using Jensen's inequality gives us:

$$\mathbb{E}[R_T] \leq \mathcal{O}\left(\sqrt{\log K\, \mathbb{E}\sum_{t=1}^{T}\sum_{i=1}^{K}(\tilde{\ell}_{t,i}-m_{t,i})^2} + \sqrt{\log K}\right)$$

We can now plug in the usual $\tilde{\ell}_{t,i} = \frac{(\ell_{t,i}-m_{t,i})}{\epsilon}\mathbb{1}_{\{d_t=1\}}+m_{t,i}$, and choose messages corresponding to the quadratic variation based bound (i.e. $m_t = \tilde{\mu}_t$ via reservoir sampling) to give us:

$$\mathbb{E}[R_T] \leq \mathcal{O}\left(\sqrt{(QT\log K)/n} + \sqrt{\log K}\right)$$

Note that once again, taking expectation for the above estimates and messages will have to be done carefully similarly to as it is done for (13).

■

Appendix B. Label Efficient Bandits

The sequence of lemmas for proving Theorem 5 will be very similar as that done above for Theorem 2. As mentioned in the main text, the key difference in the label efficient bandit setting is that we will have just the log barrier regularizer (instead of the hybrid regularizer). Additionally, our second order corrections are also $a_t = 6\eta\epsilon^2 x_t(\tilde{\ell}_t - m_t)^2$. Lemmas 6, 7, and 8 follow almost identically. We provide the analogue to Lemma 9 below and then prove Theorem 3.

Lemma 13 *For* $a_t = 6\eta\epsilon^2 x_{t,i}(\tilde{\ell}_{t,i}-m_{t,i})^2$, $\tilde{\ell}_{t,i} = \frac{(\ell_{t,i}-m_{t,i})}{\epsilon x_{t,i}}\mathbb{1}_{\{d_t=1,i_t=i\}} + m_{t,i}$, $\eta \leq \frac{1}{162K}$ *we have that* $\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|_{x_t,*} \leq \frac{1}{3}$.

Proof of Lemma 13

$$\begin{aligned}\left\|\epsilon(\tilde{\ell}_t - m_t) + a_t\right\|^2_{x_t,*} &= \sum_{i=1}^{K}\eta x_i^2\left(\epsilon(\tilde{\ell}_{t,i}-m_{t,i}) + 6\eta\epsilon^2 x_i(\tilde{\ell}_{t,i}-m_{t,i})^2\right)^2 \\ &= \eta\sum_{i=1}^{K}x_i^2\epsilon^2(\tilde{\ell}_{t,i}-m_{t,i})^2\times \\ &\quad\left[1 + 12\eta\epsilon x_i(\tilde{\ell}_{t,i}-m_{t,i}) + 36\eta^2\epsilon^2 x_i^2(\tilde{\ell}_{t,i}-m_{t,i})^2\right]\end{aligned}$$

$$\leq 2\eta\epsilon^2 \sum_{i=1}^{K} (\tilde{\ell}_{t,i} - m_{t,i})^2 x_i^2$$
$$\leq 2\eta$$
$$\leq \frac{1}{9}$$

The above inequalities again follow by observing that $|\epsilon x_{t,i}(\tilde{\ell}_{t,i} - m_{t,i})| = |(\ell_t - m_t)\mathbb{1}_{\{d_t=1, i_t=1\}}| \leq 1$ along with the assumption on η.

■

Proof of Theorem 3 As before, we will again show that our choices of loss estimates, messages, and corrections guarantee Lemma 1.

$$\begin{aligned}
\langle x_t - x'_{t+1}, \epsilon(\tilde{\ell}_t - m_t) + a_t \rangle &\leq \left\| x_t - x'_{t+1} \right\|_{x_t} \left\| \epsilon(\tilde{\ell}_t - m_t) + a_t \right\|_{x_t,*} \\
&\leq 3 \left\| \epsilon(\tilde{\ell}_t - m_t) + a_t \right\|_{x_t,*}^2 \\
&\leq 3\eta \sum_{i=1}^{K} \epsilon^2 (\tilde{\ell}_{t,i} - m_{t,i})^2 x_i^2 \times \\
&\quad \left[1 + 12\eta\epsilon x_i (\tilde{\ell}_{t,i} - m_{t,i}) + 36\eta^2\epsilon^2 x_i^2 (\tilde{\ell}_{t,i} - m_{t,i})^2 \right] \\
&\leq 6\eta\epsilon^2 \sum_{i=1}^{K} x_{t,i}^2 (\tilde{\ell}_{t,i} - m_{t,i})^2 = \langle x_t, a_t \rangle
\end{aligned}$$

Therefore we can again proceed to take summation over t on both sides of the result of Lemma 1.

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - u, \tilde{\ell}_t \rangle \right] \leq \frac{1}{\epsilon} \sum_{t=1}^{T} \mathbb{E}\left[\left(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \langle u, a_t \rangle \right) \right]$$

The first 2 terms on the right hand side will again telescope to yield a remaining $D_{\mathcal{R}}(u, x'_1)$, therefore giving us:

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - u, \tilde{\ell}_t \rangle \right] \leq \frac{1}{\epsilon} \left(D_{\mathcal{R}}(u, x'_1) + \mathbb{E}\left[\sum_{t=1}^{T} \langle u, a_t \rangle \right] \right)$$

$$\begin{aligned}
D_{\mathcal{R}}(u, x'_1) &= \mathcal{R}(u) - \mathcal{R}(x'_1) - \langle \nabla \mathcal{R}(x'_1), u - x'_1 \rangle \\
&= \mathcal{R}(u) - \mathcal{R}(x'_1) \leq \frac{K \log T}{\eta}
\end{aligned}$$

Note that this time, we will not have the cancellation of K as we did for Theorem 2. We will pick $u = (1 - \frac{1}{T})\boldsymbol{e}_{i^*} + \frac{1}{KT}\mathbf{1}$ as before. The rest of the proof will follow similarly to Theorem 2 to ultimately give us:

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - u, \tilde{\ell}_t \rangle \right] = \mathcal{O}\left(\frac{K \log T}{\epsilon\eta} + 18\eta Q^* + K(\log T)^2 \right) \tag{19}$$

Also note that now, we will have an added reservoir sampling cost in the final regret bound which is the $K(\log T)^2$ term.

■

Appendix C. Lower Bound Proofs

Our bounds will be proven in a 2-step manner similar to that in [9]. The main feature of step 1 (the Lemma step) is that of centering the Bernoulli random variables around a parameter α instead of $1/2$, which leads the regret bound to involve the $\alpha(1-\alpha)$ term corresponding to the variance of the Bernoulli distribution. Step 2 (the Theorem step) builds upon step 1 and shows the existence of a loss sequence belonging to an α-variation ball ($\mathcal{V}_\alpha \triangleq \{\{\ell_t\}_{t=1}^T : Q/TK \leq \alpha\}$ for $\alpha \in [0, 1/4]$) which also incurs regret of the same order. Theorem 4 follows Lemma 14 while Theorem 5 follows Lemma 15.

Lemma 14 *Let $\alpha \in (0,1)$, $K \geq 2$, $T \geq n \geq \frac{c^2 \log(K-1)}{1-\alpha}$. Then, for any randomized strategy for the label efficient prediction problem, there exists a loss sequence under which $\mathbb{E}[R_T] \geq cT\sqrt{\frac{\alpha(1-\alpha)\log(K-1)}{n}}$ for $c = \sqrt{e}/\sqrt{5(1+e)}$ where expectation is taken with respect to the internalrandomization available to the algorithm and the random loss sequence.*

Lemma 15 *Let $\alpha \in (0,1)$, $K \geq 2$, $T \geq n \geq K/(4(1-\alpha))$. Then, for any randomized strategy for the label efficient bandit problem, there exists a loss sequence under which $\mathbb{E}[R_T] \geq \frac{T}{8}\sqrt{\alpha(1-\alpha)K/n}$ where expectation is taken with respect to the internal randomization available to the algorithm and the random loss sequence.*

Proof of Lemma 14 Our proof for this lemma closely follows the proof of [6] with a few changes:

- Our Bernoulli random variables are centred at α instead of at $1/2$.
- We define our random variables a little differently to make the calculations easier. Namely, Z^* is a Bernoull(α) random variable instead of Bernoulli $(\alpha - \epsilon)$ as is done in [6] (for $\alpha = 1/2$) and Z_j is Bernoulli $(\alpha + \epsilon)$ instead of Bernoulli (α) (again for $\alpha = 1/2$).

Given $y^t \in [0,1]$, consider the first K coefficients of its unique dyadic expansion and denote these as $y_1^t, y_2^t, \ldots, y_K^t$. We will then define $\ell_{t,i} = y_i^t$ for all $i \in [K] = \{1, 2, \ldots, K\}$. We will construct a random outcome sequence $Y_1, \ldots, Y_T$, where each random variable is supported on $[0,1]$. The realizations of these random variables will then define an associated loss sequence as explained above. We will show that the expected regret of any randomized algorithm is bounded below by the claimed quantity, where we will take expectation with respect to the random outcome sequence as well as the internal/auxiliary randomness available to the algorithm. Denote by $A_1, A_2, \ldots A_T$ the internal randomization available to the strategy (associated distribution is $\mathbb{P}_A$), which we will take to be an i.i.d. sequence of uniform random variables supported on $[0,1]$. Now define K(no. of arms) joint distributions $\mathbb{P}_i \otimes \mathbb{P}_A$ where $\mathbb{P}_1, \ldots, \mathbb{P}_K$ are probability distributions over the outcome sequence which we define below. For $i \in [K]$, define by $\mathbb{Q}_i$ the distribution of:

$$Z^* 2^{-i} + \sum_{j=1,\ldots,K,\ j \neq i} Z_j 2^{-j} + 2^{-(K+1)} A$$

$A, Z^*, Z_1, \dots Z_K$ are all independent random variables. A is distributed uniformly over $[0,1]$, Z^* is a Bernoulli (α) random variable, and Z_j is distributed Bernoulli $(\alpha+\varepsilon)$ (we specify ϵ later). Now, under $\mathbb{P}_i$, the outcome sequence $Y_1, \dots, Y_T$ is i.i.d. from $\mathbb{Q}_i$. Hence, under $\mathbb{P}_i$, for all $j \in [K]$ and $t \in [T]$, $\ell_{t,j}$ are i.i.d. Bernoulli random variables. $\ell_{t,i}$ is Bernoulli (α), and $\ell_{t,j}$, for $j \neq i$ is Bernoulli $(\alpha+\varepsilon)$. Denote the cumulative loss of the strategy by $\hat{L}_T = \sum_{t=1}^T \ell_{t,i_t}$ and the cumulative loss of arm i by $L_{T,i}$. Let $\mathbb{E}_i$ be the expectation with respect to $\mathbb{P}_i$ and $\mathbb{E}_A$ the expectation with respect to $\mathbb{P}_A$. We then have that:

$$\begin{aligned}\max_{\{\ell_s\}_{s=1}^T}\left(\mathbb{E}_A\hat{L}_T - \min_{i\in[K]} L_{T,i}\right) &= \max_{\{\ell_s\}_{s=1}^T,\, i\in[K]}\left(\mathbb{E}_A\hat{L}_T - L_{T,i}\right)\\ &\geq \max_{i\in[K]}\mathbb{E}_i\left[\mathbb{E}_A\hat{L}_T - L_{T,i}\right]\end{aligned}$$

Using Lemma 16, we have that $\mathbb{P}_A\left[i_t = i|\{\ell_s\}_{s=1}^{t-1}\right] = \sum_{d=1}^D \beta_d \mathbb{1}_{\left[i_t^d=i|\{\ell_s\}_{s=1}^{t-1}\right]}$ where $\mathbb{1}_{\left[i_t^d=i|\{\ell_s\}_{s=1}^{t-1}\right]}$ is an indicator for the d-th deterministic algorithm choosing i. We therefore rewrite the regret as:

$$\begin{aligned}\max_{i\in[K]}\mathbb{E}_i\left[\mathbb{E}_A\hat{L}_T - L_{T,i}\right] &= \max_{i\in[K]}\mathbb{E}_i\left[\sum_{t=1}^T\sum_{d=1}^D\beta_d\sum_{k=1}^K \mathbb{1}_{\left[i_t^d=i|\{\ell_s\}_{s=1}^{t-1}\right]}\ell_{t,k} - L_{T,i}\right]\\ &= \max_{i\in[K]}\sum_{d=1}^D\beta_d\mathbb{E}_i\left[\sum_{t=1}^T\sum_{k=1}^K \mathbb{1}_{\left[i_t^d=i|\{\ell_s\}_{s=1}^{t-1}\right]}\ell_{t,k} - L_{T,i}\right]\\ &= \mathcal{E}\max_{i\in[K]}\sum_{d=1}^D\beta_d\sum_{t=1}^T\mathbb{P}_i\left[i_t^d\neq i\right]\\ &= \mathcal{E}T\left(1 - \min_{i\in[K]}\sum_{d=1}^D\sum_{t=1}^T\frac{\beta_d}{T}\mathbb{P}_i\left[i_t^d = i\right]\right)\end{aligned}$$

where the third equality uses the fact that the regret grows by $\mathcal{E}$ under $\mathbb{P}_i$ whenever $i_t \neq i$. Now for the d-th deterministic algorithm, let $1 \leq T_1^d \leq \dots \leq T_n^d \leq T$ be the times when the strategy asks for the n labels. Then $T_1^d, \dots, T_n^d$ correspond to the finite stopping times with respect to the i.i.d. process $Y_1, \dots Y_T$. Hence, the revealed outcomes $Y_{T_1^d}, \dots, Y_{T_n^d}$ are i.i.d. from Y_1 (see [8]). Denote by R_t^d the number of revealed labels at time t. Now, as the sub-algorithms are deterministic, R_t^d is fully determined by $Y_{T_1^d}, \dots, Y_{T_n^d}$. Hence, in general, i_t^d can be thought to be a function of $Y_{T_1^d}, \dots, Y_{T_n^d}$ instead of the revealed labels *just* till time t, which are $Y_{T_1^d}, \dots, Y_{T_{R_t^d}^d}$. As the joint distribution of $Y_{T_1^d}, \dots, Y_{T_n^d}$ under $\mathbb{P}_i$ is $\mathbb{Q}_i^n$, we have that $\mathbb{P}_i[i_t^d = i] = \mathbb{Q}_i^n[i_t^d = i]$. Hence the regret becomes:

$$\max_{i\in[K]}\mathbb{E}_i\left[\mathbb{E}_A\hat{L}_T - L_{T,i}\right] = \mathcal{E}T\left(1 - \min_{i\in[K]}\sum_{d=1}^D\sum_{t=1}^T\frac{\beta_d}{T}\mathbb{Q}_i\left[i_t^d = i\right]\right)$$

By the generalized Fano's inequality, we know that $\min_{i\in[K]}\sum_{d=1}^D\sum_{t=1}^T\frac{\beta_d}{T}\mathbb{Q}_i\left[i_t^d = i\right] \leq \max\left\{\frac{e}{1+e}, \frac{\bar{S}}{\log(K-1)}\right\}$ where $\bar{S} = \frac{1}{K-1}\sum_{i=2}^K \mathrm{KL}(\mathbb{Q}_i^n, \mathbb{Q}_1^n)$.
Now observe that:

$$\mathrm{KL}(\mathbb{Q}_i^n, \mathbb{Q}_1^n) = n\mathrm{KL}(\mathbb{Q}_i, \mathbb{Q}_1)$$

$$
\begin{aligned}
&\leq n(\mathrm{KL}(\mathrm{Bern}(\alpha), \mathrm{Bern}(\alpha+\mathcal{E})) + \mathrm{KL}(\mathrm{Bern}\,(\alpha+\mathcal{E}), \mathrm{Bern}(\alpha))) \\
&\leq n\left(\chi^2(\alpha, \alpha+\varepsilon) + \chi^2(\alpha+\varepsilon, \alpha)\right) \\
&= n\left(\frac{\varepsilon^2}{(\alpha+\varepsilon)(1-\alpha-\varepsilon)} + \frac{\varepsilon^2}{\alpha(1-\alpha)}\right) \\
&\leq \frac{5n\varepsilon^2}{\alpha(1-\alpha)}
\end{aligned}
$$

where we upper bound KL divergence by χ^2 divergence and restrict ε to $\left[0, \frac{3(1-\alpha)}{4}\right]$ (our proposed ε below doesn't exceed $3(1-\alpha)/4$ as $n \geq \log K/(1-\alpha)$). Therefore, we have that

$$
\max_{i\in[K]} \mathbb{E}_i\left[\mathbb{E}_A \hat{L}_T - L_{T,i}\right] \geq \mathcal{E}T\left(1 - \max\left\{\frac{e}{1+e}, \frac{5n\mathcal{E}^2}{\log(K-1)\alpha(1-\alpha)}\right\}\right)
$$

Choosing $\varepsilon = \sqrt{\frac{e\alpha(1-\alpha)\log(K-1)}{5n(1+e)}}$ reveals the claimed bound. ■

Lemma 16 *(Lemma 3 from [6]) For any randomized strategy, there exists D deterministic strategies and a probability vector $\beta = (\beta_1, \ldots, \beta_D)$ such that for every t and every possible outcome sequence $\{\ell_s\}_{s=1}^{t-1}$,*

$$
\mathbb{P}_A\left[i_t = i | \{\ell_s\}_{s=1}^{t-1}\right] = \sum_{d=1}^{D} \beta_d \mathbb{1}_{\left[i_t^d = i | \{\ell_s\}_{s=1}^{t-1}\right]}
$$

Proof of Theorem 4 We will begin by applying Lemma 14 with $\alpha/2$ and with the constant $c = 0.36$ (out of convenience) which is indeed lesser than the one we have proven the above lemma for. Note that there is some $j \in [K]$, for which

$$
\mathbb{E}_j[R_T] \geq 0.36\sqrt{\frac{\alpha}{2}(1-\frac{\alpha}{2})T\log K\frac{T}{n}} \geq 0.09\sqrt{7\alpha T\log K\frac{T}{n}} \text{ (as } \alpha \leq 1/4) \tag{20}
$$

We will now show that under $\mathbb{P}_j$, the probability that $Q \geq \alpha TK$ is less than $\frac{9}{100T}$. Recall that $\mu_T = \frac{1}{T}\sum_{t=1}^{T} \ell_t$ and $Q = \sum_{t=1}^{T} \|\ell_t - \mu_T\|_2^2 = \sum_{i=1}^{K} v_{\alpha,i}$ where $v_{\alpha,i} = \sum_{t=1}^{T}(\ell_{t,i} - \mu_{T,i})^2$. Noting that $\ell_{t,i} \in \{0,1\}$, we have $v_{\alpha,i} = T\mu_{T,i}(1-\mu_{T,i}) \leq T\mu_{T,i} = \sum_{t=1}^{T} \ell_{t,i}$. Applying Bernstein's inequality (refer to Theorem 2.10 in [2] with $b = 1,\ v = T(\alpha/2)(1-\alpha/2),\ c = b/3 = 1/3$) along with a union bound gives us that for all $\delta \in (0,1)$, under $\mathbb{P}_j$, with probability at least $1-\delta$, we have:

$$
\sum_{t=1}^{T} \ell_{t,i} \leq T\left(\frac{\alpha}{2} + \epsilon\right) + \sqrt{2T\left(\frac{\alpha}{2} + \epsilon\right)\log\frac{K}{\delta}} + \frac{1}{3}\log\frac{K}{\delta}\,. \tag{21}
$$

for all $i \in \{1, \ldots, K\}$. Now note that by definition of $\epsilon = 0.36\sqrt{(\alpha/2)(1-\alpha/2)\log(K-1)/n}$ and by the assumption $n \geq 8\log(K-1)/\alpha$,

$$
\frac{\alpha}{2} + \epsilon = \frac{\alpha}{2} + 0.36\sqrt{\frac{\alpha}{2}\left(1-\frac{\alpha}{2}\right)\frac{\log(K-1)}{n}} \leq 0.59\alpha
$$

Substituting this in (21) above, we get:

$$\sum_{t=1}^{T} \ell_{t,i} \leq 0.59T\alpha + \sqrt{2T(0.59\alpha)\log\frac{K}{\delta}} + \frac{1}{3}\log\frac{K}{\delta} \tag{22}$$

Now we claim that $T\alpha \geq 16\log\frac{K}{\delta}$ holds for $\delta = \frac{9}{100T}$. This follows from our assumptions that $\alpha \geq \frac{32\log T}{T}$ and $T \geq \frac{100}{9}K$. Substituting this back into (22), we can see that $\sum_{t=1}^{T}\ell_{t,i} \leq T\alpha$. Hence, this gives us that $Q \leq \alpha TK$.

Now we will show that there exists a sequence of losses with $Q \leq \alpha TK$ and $\mathbb{E}[R_T] \geq 0.045\sqrt{7\alpha T\log K\frac{T}{n}}$ where the expectation is taken with respect to the internal randomisation of the strategy. Suppose this were not true, then we would have that $\mathbb{1}_{\{Q\leq\alpha TK\}}\mathbb{E}_j[R_T|\{\ell_t\}_{t=1}^T] \leq 0.045\sqrt{7\alpha T\log K\frac{T}{n}}$ (since $\mathbb{P}_j$ is independent of the internal randomisation). Then we would consequently have:

$$\begin{aligned}
\mathbb{E}_j[R_T] &= \mathbb{E}_j\left[R_T\mathbb{1}_{\{Q\leq\alpha TK\}}\right] + \mathbb{E}_j\left[R_T\mathbb{1}_{\{Q>\alpha TK\}}\right] \\
&\leq 0.045\sqrt{7\alpha T\log K\frac{T}{n}} + T\cdot\mathbb{P}_j(Q>\alpha TK) \\
&\leq 0.045\sqrt{7\alpha T\log K\frac{T}{n}} + 0.09 < 0.09\sqrt{7\alpha T\log K\frac{T}{n}}
\end{aligned}$$

which contradicts equation (20). Hence, $\mathbb{E}[R_T] \geq 0.09\sqrt{7\alpha T\log K\frac{T}{n}}$.

■

The main difference for label efficient bandits from standard bandit proofs is that now, the total number of revealed labels (each label is now a single loss vector entry) cannot exceed n. Hence, the $\sum_{i\in[K]}N_i(t-1)$ term which appears in the analysis is upper bounded by n (where $N_i(t-1)$ denotes the pulls of arm i up till time $t-1$).

Proof of Lemma 15 As mentioned, the key difference here from standard bandit lower bounds is that $\sum_{i\in[K]}N_i(t-1)$ (the sum of all revealed labels till time $t-1$) is upper bounded by n. Barring this, the proof follows almost identically as that done in [9] but we mention it here for completeness. Consider the following K probability distributions used to construct the stochastic losses. For $i\in[K]$, let $\mathbb{Q}_i$ be a distributions such that under $\mathbb{Q}_i$, $\ell_{t,i}$ is drawn Bernoulli (α) for all $t\in\{1,2,\ldots,T\}$, and $\ell_{t,j}$ is drawn Bernoulli $(\alpha+\varepsilon)$ for all $t\in\{1,2,\ldots,T\},\ j\in[K],\ j\neq i$ (we specify ε later). Additionally, let $\mathbb{Q}_0$ be the joint distribution under which all $\ell_{t,i}$ are i.i.d Bernoulli $(\alpha+\varepsilon)$ random variables for $t\in\{1,2,\ldots,T\}$ and $i\in[K]$. Also define $\bar{\mathbb{Q}} = \frac{1}{K}\sum_{i=1}^{K}\mathbb{Q}_i$, the distribution our losses will finally be drawn from. As before, let $\mathbb{E}_i$ denote the expectation taken with respect to $\mathbb{Q}_i$. Under (each) $\mathbb{Q}_i$ we have the following:

$$\begin{aligned}
\mathbb{E}_i\left[\hat{L}_T - \min_{j\in[K]}L_{T,j}\right] &\geq \mathbb{E}_i\left[\hat{L}_T\right] - \min_{j\in[K]}\mathbb{E}_i[L_{T,j}] = \mathbb{E}_i\left[\sum_{t=1}^{T}\ell_{t,i_t}\right] - \min_{j\in[K]}\mathbb{E}_i\left[\sum_{t=1}^{T}\ell_{t,j}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_i\left[\alpha+\varepsilon-\varepsilon\mathbb{1}_{\{i_t=i\}}\right] - T\alpha
\end{aligned}$$

$$= T\varepsilon \left(1 - \frac{1}{T}\sum_{t=1}^{T} \mathbb{Q}_i(i_t = i)\right) , \tag{23}$$

Now, we can further lower bound the above expression by appealing to Pinsker's inequality which tells us that $\mathbb{Q}_i(i_t = i) \leq \mathbb{Q}_0(i_t = i) + (\mathrm{KL}(\mathbb{Q}_0^{i_t}, \mathbb{Q}_i^{i_t})/2)^{1/2}$ [5] for all $t \in \{1, 2, \ldots, T\}$ and all $i \in [K]$. We substitute this in (23), average over $i \in [K]$ in order to bound the regret under $\bar{\mathbb{Q}}$, and use the concavity of the square root to yield:

$$\mathbb{E}_{\bar{\mathbb{Q}}}\left[\hat{L}_T - \min_{j\in[K]} L_{T,j}\right] \geq T\varepsilon \left(1 - \frac{1}{K} - \sqrt{\frac{1}{2T}\sum_{t=1}^{T}\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}\big(\mathbb{Q}_0^{i_t}, \mathbb{Q}_i^{i_t}\big)}\right) \tag{24}$$

Now we will upper bound the KL divergence terms:

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{Q}_0^{i_t}, \mathbb{Q}_i^{i_t}\Big) &\leq \mathrm{KL}\left(\mathbb{Q}_0^{(h_t,i_t)}, \mathbb{Q}_i^{(h_t,i_t)}\right) = \mathbb{E}_{\mathbb{Q}_0}\big[N_i(t-1)\big]\,\mathrm{KL}\big(\mathrm{Bern}(\alpha+\epsilon), \mathrm{Bern}(\alpha)\big) \\ &\leq \mathbb{E}_{\mathbb{Q}_0}\big[N_i(t-1)\big]\,\frac{\varepsilon^2}{\alpha(1-\alpha)} ,\end{aligned}$$

where the first inequality follows from the Data Processing Inequality and second by upper bounding the KL divergence by the χ^2 divergence. h_t denotes the history available at time t and $N_i(t-1)$ refers to the number of pulls of arm i till time $t-1$. We now average the above quantity over $i \in [K]$ and $t \in \{1, 2, \ldots, T\}$ to yield:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{j=1}^{K}\mathrm{KL}\big(\mathbb{Q}_0^{i_t}, \mathbb{Q}_i^{i_t}\big) \leq \frac{1}{T}\sum_{t=1}^{T}\frac{n\varepsilon^2}{K\alpha(1-\alpha)} \leq \frac{n\varepsilon^2}{K\alpha(1-\alpha)} .$$

The above equation incorporates the strict restriction on the revealed labels as $\sum_{i\in[K]} N_i(t-1)$ is upper bounded by n. Plugging the above inequality into (24) and substituting $\varepsilon = (1/2\sqrt{2})\sqrt{\alpha(1-\alpha)K/n}$ gives us the claimed bound.

■

Proof of Theorem 5 The proof follows almost identically as that of Theorem 4. ■

5. $\mathbb{Q}_i^{i_t}$ denotes the probability measure of i_t under $\mathbb{Q}_i$