The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

On Adaptivity in Information-Constrained Online Learning

Siddharth Mitra,[1] **Aditya Gopalan**[2]
[1]Chennai Mathematical Institute, smitra@cmi.ac.in
[2]Indian Institute of Science, aditya@iisc.ac.in

Abstract

We study how to adapt to smoothly-varying ('easy') 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 (partial) 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. We then consider revealing action-partial monitoring games – a version of label efficient prediction with additive information costs – which in general are known to lie in the *hard* class of games having minimax regret of order $T^{2/3}$. We provide a strategy with an $\mathcal{O}((Q^*T)^{1/3})$ bound for revealing action games, along with one with a $\mathcal{O}((QT)^{1/3})$ bound for the full class of hard partial monitoring games, both being strict improvements over current bounds.

1 Introduction

Online learning is a branch of machine learning that is concerned with the problem of dynamically optimizing utility (or loss) over time in the face of uncertainty, and gives valuable principles to reason about acting under uncertainty. 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 from an average-case point of view, such as i.i.d. multi-armed bandits (see, e.g., the survey of (Bubeck, Cesa-Bianchi, and others 2012)), online learning in Markov decision processes (Jaksch, Ortner, and Auer 2010; Azar, Osband, and Munos 2017), stochastic partial monitoring (Bartók, Pál, and Szepesvári 2011), etc., 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 – prediction with experts, bandits and partial monitoring problems to name a few (Cesa-Bianchi and Lugosi 2006) – 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. These include the derivation of first-order regret bounds or adapting to loss sequences with low loss of the best action (Allenberg et al. 2006), second-order bounds or adapting to loss sequences with low variation in prediction with experts (Rakhlin and Sridharan 2012; Steinhardt and Liang 2014) and 'benign' multi-armed bandits (Hazan and Kale 2011; Bubeck et al. 2019; Bubeck, Cohen, and Li 2017; Wei and Luo 2018).

In this regard, this paper is an attempt to extend our understanding of adaptivity to low variation in several standard online learning problems where information comes at a cost, namely label efficient prediction (Cesa-Bianchi, Lugosi, and Stoltz 2005), label efficient bandits (Cesa-Bianchi and Lugosi 2006) and classes of partial monitoring problems (Bartók et al. 2014). In the process, we uncover new ways of using existing online learning techniques within the Online Mirror Descent (OMD) family, and partially make progress towards a program of studying the impact of 'easy' (i.e., slowly-varying) environments in information-constrained online learning and partial monitoring problems. Our specific contributions are:

1. For the label efficient prediction game with expert advice, we give a learning algorithm with a regret bound of $\mathcal{O}\big(\sqrt{(Q^*T\log K)/n}\big)$ where Q^* is the quadratic variation of the best expert, T is the time horizon of the game, K is the number of experts and n is the bound on label queries; the bound holds for all regimes except when ${}^{nQ^*}/_{T} = \widetilde{\mathcal{O}}(K^2)$. We follow this up with an algorithm with an unconditional

regret guarantee of $\mathcal{O}(\sqrt{(QT\log K)/n})$ that holds for any label query budget n and total quadratic variation Q. Our algorithms are based on the optimistic OMD framework, but with new combinations of the negative entropy and log-barrier regularization that are best suited to the label efficient game's information structure.

2. We generalize the results to label efficient bandits where one receives bandit (i.e., for only the chosen expert) feedback at only up to n chosen time instants, and obtain $\mathcal{O}\big(\sqrt{(Q^*TK)/n}\big)$ regret. We also show that our upper bounds on regret for label efficient prediction and label efficient bandits are tight in their dependence on Q and n by demonstrating variation-dependent fundamental lower bounds on regret.

3. We show that adapting to low variation is also possible in the class of *hard* partial monitoring games as per the taxonomy of partial monitoring problems by (Bartók et al. 2014), where we show an algorithm that achieves $\mathcal{O}((QTK)^{1/3})$ regret. To the best of our knowledge, this is the first algorithm exhibiting instance-dependent bounds for partial monitoring.

Problem Setup and Notation A label efficient prediction game 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

$$\max_{i^* \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i_t} - \sum_{t=1}^{T} \ell_{t,i^*}\right],$$

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 $D_{\mathcal{R}}(x,y) \triangleq \mathcal{R}(x) - \mathcal{R}(y) - \langle \nabla\mathcal{R}(y), x-y\rangle \ \forall x, y \in \Omega$. 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 $\|u\|_{x,*} = \sqrt{u^\top \nabla^{-2} \mathcal{R}(x) u}$. 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) (Hazan 2016, e.g.). 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. Throughout the paper, $\tilde{\ell}_t$ 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 (Wei and Luo 2018), whereas its Follow-the-Regularized-Leader (FTRL) counterpart was introduced earlier by (Steinhardt and Liang 2014). Both of these approaches build upon the general optimistic OMD framework of (Rakhlin and Sridharan 2012) and (Chiang et al. 2012). 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 $\approx \eta\epsilon^2(\tilde{\ell}_t - m_t)^2$. It is worthwhile to note that this is explicitly different from the $\eta\epsilon(\tilde{\ell}_t - m_t)^2$ that one may expect in light of the analysis done in (Wei and Luo 2018), or the $\eta(\tilde{\ell}_t - m_t)^2$ one would anticipate when following (Steinhardt and Liang 2014). 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 (Steinhardt and Liang 2014) or for bandits (Wei and Luo 2018). 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 We briefly discuss the challenges posed by label efficient prediction and how our choice of regularizer addresses these. When shifting away from the classical prediction with expert advice problem to any *limited* feedback (i.e., over experts or arms) information structure, one usually works with importance-weighted estimates of the loss vectors constructed using the observed (limited) feedback (called inverse propensity weighting estimation). This is indeed the case with label

[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.

REFERENCE	FEEDBACK	NEGENTROPY:LOG-BARRIER REGULARIZER RATIO USED
(Bubeck, Cohen, and Li 2017)	Bandit	$1 : 2\eta$
(Wei and Luo 2018)	Bandit	$0 : 1$
(Bubeck et al. 2019)	Bandit	$K/\eta : 1/\eta = K : 1$
(Steinhardt and Liang 2014)	Full Information	$1 : 0$
This work	Label Efficient – Full Information	$1/\eta : 1/K\eta = K : 1$
This work	Label Efficient – Bandit Feedback	$0 : 1$

Table 1: Choice of regularization (negative entropy vs. logarithmic barrier) in OMD for exploiting regularity

efficient prediction, however, the probabilities in the denominator remain fixed at ϵ, 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 (Steinhardt and Liang 2014), 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 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[2] , key in obtaining OMD regret bounds, in such situations. The log-barrier regularizer, used by (Wei and Luo 2018) 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 place 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:

$$\mathcal{R} = 1/\eta \sum_{i=1}^{K} x_i \log x_i - 1/(K\eta) \sum_{i=1}^{K} \log x_i$$

This regularizer has been of recent interest due to the work of (Bubeck et al. 2019), and (Bubeck, Cohen, and Li 2017), 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. 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.

[2]In the sense of successive points being sufficiently close to each other. Please refer to Lemma 14 in https://arxiv.org/abs/1910.08805.

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 (Hazan and Kale 2011), 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 $\text{Var}[\tilde{\mu}_t] = Q/kt$. It does so by maintaining a carefully constructed reservoir S of size k, the elements from which are then averaged to output the estimate of the mean. Our message m_t at any time t is the average of the vectors contained in the reservoir S. We specify the reservoir sampling algorithm in Algorithm 1.

Algorithm 1 RESERVOIR SAMPLING

Input: Reservoir S, Reservoir size k, Stream $\ell_1, \ell_2, \ldots$
for $t = 1, 2, \ldots, k$ **do**
  Include ℓ_t in S
end for
for $t = k+1, \ldots$ **do**
  $b_t \sim \text{Bern}(k/t)$
  if $b_t = 1$ **then**
    Include ℓ_t in S by replacing it with a uniformly at random chosen element of S
  end if
end for

2.1 Main Algorithm

Algorithm 2 builds upon the preceding ideas and as stated, is specifically for the label efficient prediction problem discussed thus far. The algorithms required for the extensions we provide in section 4 are based upon Algorithm 2 with a few minor differences. Also, in the interest of brevity, we have excluded the explicit mentioning of the reservoir sampling steps in Algorithm 2. Before we proceed, we would like to cleanly state our choice of messages, loss estimates, and second order corrections used and this is done in Table 2. Our messages, for all the sections will be $m_{t,i} = \tilde{\mu}_{t-1,i}$. Note that throughout the paper, the random variable $d_t = 1$ signifies that we ask for feedback at time t, and is 0 otherwise. Additionally, note that we consider not exceeding the budget of n in expectation, however, there is a standard reduction

PROBLEM	SECTION	$\tilde{\ell}_{t,i} - m_{t,i}$	a_t	REGRET BOUND
Label Efficient Prediction	2.1, 3	$\frac{(\ell_{t,i}-m_{t,i})}{\epsilon}\mathbb{1}_{\{d_t=1\}}$	$6\eta\epsilon^2(\tilde{\ell}_t - m_t)^2$	$\widetilde{\mathcal{O}}\left(\sqrt{Q^*T/n}\right)$
Label Efficient Bandits	4.1	$\frac{(\ell_{t,i}-m_{t,i})}{\epsilon x_{t,i}}\mathbb{1}_{\{d_t=1,i_t=i\}}$	$6\eta\epsilon^2 x_{t,i}(\tilde{\ell}_t - m_t)^2$	$\widetilde{\mathcal{O}}\left(\sqrt{Q^*TK/n}\right)$
Revealing Action Games	4.2	$\frac{(\ell_{t,i}-m_{t,i})\mathbb{1}_{\{d_t=1\}}}{\alpha}\mathbb{1}_{\{d_t=1\}}$	$6\eta\alpha^2(\tilde{\ell}_t - m_t)^2$	$\widetilde{\mathcal{O}}\left((Q^*T)^{1/3}\right)$
Hard Partial Monitoring	4.2	$\frac{(\ell_{t,i}-m_{t,i})}{x_{t,j}}\mathbb{1}_{\{i_t=j\}}$	0	$\mathcal{O}\left((QTK)^{1/3}\right)$

Table 2: Overview of loss estimates, second order corrections, and the corresponding upper bounds on regret

to get a high probability guarantee which can be found in (Cesa-Bianchi and Lugosi 2006).

Algorithm 2 ADAPTIVE LABEL EFFICIENT PREDICTION

Input: $\mathcal{R} = {}^1\!/\!_\eta \sum_{i=1}^K x_i \log x_i - {}^1\!/\!_{K\eta} \sum_{i=1}^K \log x_i$,
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 for

3 Results and Analysis

We now give a general regret result for the OMD updates (1) and (2). It spells out the condition we must maintain to ultimately enable best arm dependent bounds while also demonstrating the price of limited information on regret, which is the additional ${}^1\!/\!_\epsilon$ factor. The proofs for all results in this section and subsequent sections are available in the full version of this paper[3].

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

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

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

$$\langle x_t - u, \tilde{\ell}_t\rangle \le \frac{1}{\epsilon}\left(D_{\mathcal{R}}(u, x'_t) - D_{\mathcal{R}}(u, x'_{t+1}) + \langle u, a_t\rangle - P_t\right), \qquad (4)$$

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

Note that when $a_t = 0$ is employed in the updates (1)-(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 ($a_t \ge 0$), 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 (Wei and Luo 2018)). We now state our main result for adaptive label efficient prediction which bounds the regret of Algorithm 2.

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 $\eta \le {}^1\!/\!_{162K}$ where the sequence of messages m_t are generated using the reservoir sampling scheme, the expected regret of Algorithm 2 satisfies the following:*

$$\mathbb{E}[R_T] \le \frac{\log K + \log T}{\epsilon\eta} + 18\eta Q^* .$$

Furthermore, if $\epsilon Q^* \ge 1458K^2 \log KT$, then $\mathbb{E}[R_T] = \mathcal{O}\left(\sqrt{\frac{Q^* T \log K}{n}}\right)$ 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 as given in (Cesa-Bianchi, Lugosi, and Stoltz 2005), we would get a regret bound of $\mathcal{O}(T^{3/4})$; following an FTRL with negative entropy[4]-based strategy would be inapplicable in this setting due to the constraint we highlight in section 2, however, Algorithm 2 would incur $\sqrt{T}$ regret – a marked improvement. Also, note that because of the full vector feedback, it is not required to allocate any rounds *exclusively* for reservoir sampling. This fact is reflected in not having to incur any additive penalty for reservoir sampling.

Proof sketch of Theorem 2 The result of Theorem 2 follows rather straightforwardly from (4). The key part of the proof lies in showing that the choice of messages, second-order

[3]The full version is available at https://arxiv.org/abs/1910.08805

[4]As done in (Steinhardt and Liang 2014) for prediction with experts

corrections, and loss estimators satisfy (3). To show that (3) i.e. $\langle x_t - x'_{t+1}, \epsilon(\tilde{\ell}_t - m_t) + a_t \rangle \leq \langle x_t, a_t \rangle$ is satisfied, we upper bound the left hand side by $\|x_t - x'_{t+1}\|_{x_t} \|\epsilon(\tilde{\ell}_t - m_t) + a_t\|_{x_t,*}$, and show that this is indeed upper bounded by $\langle x_t, a_t \rangle$. We then relate both $\|x_t - x'_{t+1}\|_{x_t}$ and $\|\epsilon(\tilde{\ell}_t - m_t) + a_t\|_{x_t,*}$ and show that our choice of estimator, messages, and second-order corrections guarantee that $\|\epsilon(\tilde{\ell}_t - m_t) + a_t\|_{x_t,*}$ is 'small'.

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. We now proceed to discuss what can be said without any constraint on ϵQ^*. Specifically, we will provide an algorithm obtaining $\mathcal{O}(\sqrt{(QT \log K)/n})$ regret under *all* scenarios, the trade-off however being that we will be penalized by Q instead of Q^*. In settings where the ϵQ^* condition does not hold and incurring regret in terms of Q is not unfavourable (as an extreme example, consider constant variation on all arms, with very limited feedback) the strategy below will certainly be of use. The algorithm, again based on OMD, foregoes second order corrections and has updates defined by:

$$x_t = \operatorname*{argmin}_{x \in \Omega} \{\langle x, \epsilon m_t \rangle + D_{\mathcal{R}}(x, x'_t)\} \tag{5}$$

$$x'_{t+1} = \operatorname*{argmin}_{x \in \Omega} \{\langle x, \epsilon \tilde{\ell}_t \rangle + D_{\mathcal{R}}(x, x'_t)\} \tag{6}$$

Without second order corrections, the ϵ term can be folded into the regularizer and the updates reduce to the ones studied in (Rakhlin and Sridharan 2012). For updates (5) and (6), 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 (Rakhlin and Sridharan 2012).

Lemma 3. *For any $u \in \Omega$, updates (5) and (6) guarantee that:*

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

Theorem 4. *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 2 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.

Trying to deeper understand how the constraint of Theorem 2 can be sidestepped to yield a universal algorithm dependent on Q^* remains a direction of future interest.

Parameter-Free Algorithms Note that we have assumed knowledge of T, Q and Q^* when optimising for the fixed step size η in the above discussion. This is often not possible and we now briefly discuss the extent to which we can obtain parameter-free algorithms. In Theorem 5 we claim that we can choose η adaptively for the Q dependent bound we present in Theorem 4[5]. 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 3, 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 will write T_α for the first time instance in epoch α.

Algorithm 3 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 $t \leq T$ **do**
    Draw $d_t \sim \mathrm{Bern}(\epsilon)$, update x_t according to (5)
    Play $i_t \sim x_t$ and if $d_t = 1$, observe ℓ_t
    Update x'_{t+1} according to (6)
    if $\sum_{s=T_\alpha}^{t} \sum_{i=1}^{K} (\tilde{\ell}_{s,i} - m_{s,i})^2 \geq \frac{2 \log K}{\epsilon^2 \eta_{\alpha-1}^2}$ **then**
      $\eta \leftarrow \eta/2, T_{\alpha+1} \leftarrow t, t \leftarrow t + 1$
      break
    end if
    $t \leftarrow t + 1$
  end while
end for

Theorem 5. *For the conditions mentioned in Theorem 4, Algorithm 3 (a parameter free algorithm) achieves:*

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

4 Adapting to Slowly Varying Losses in Other Information-constrained Games

We will now investigate exploiting the regularity of losses in a variety of other settings with implicit/explicit information constraints. We will first focus on bandit feedback, following which we will briefly discuss partial monitoring.

4.1 Label Efficient Bandits

The change here is in the feedback information the learner receives when asking for information. Instead of receiving

[5]Note that similarly to (Hazan and Kale 2011) we still assume knowledge of T, but this can be circumvented using standard tricks.

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 will continue to use the same update rules (1) and (2) here. What will change most importantly is the regularizer which will now solely be the log barrier regularizer $\mathcal{R} = \frac{1}{\eta}\sum_{i=1}^{K} \log \frac{1}{x_i}$. Note that the coefficient of log barrier is also $1/\eta$ instead of the earlier $1/K\eta$. The loss estimates and second order corrections will also change and these are all mentioned in Table 2. We will now state the main theorem for label efficient bandits.

Theorem 6. *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* $\eta \leq 1/162K$ *where the sequence of messages* m_t *are given by reservoir sampling, the regret of Algorithm 2 modified for label efficient bandits satisfies:*

$$\mathbb{E}\left[R_T\right] \leq \frac{K\log T}{\epsilon\eta} + 18\eta Q^* + K(\log T)^2 \,.$$

Note that since we are in the bandit feedback setting, we now reserve certain rounds solely for reservoir sampling. This is reflected in the additive $K(\log T)^2$ term in regret. There are now $(\log T)^2$ rounds allotted to each of the K arms, hence the term. There will also be a few minor changes in the algorithm primarily corresponding to the appropriate execution of reservoir sampling for bandit feedback.

4.2 Partial Monitoring

We will now discuss adaptivity in partial monitoring games. A partial monitoring game $G = (L, H)$ is defined by a pair L and H of $K \times N$ matrices. Both matrices are visible to the learner and the adversary. At each time t, the learner selects a row (or arm, action) $i_t \in [K]$ and the opponent chooses a column $y_t \in [N]$. The learner then incurs a loss of $\ell(i_t, y_t)$ and observes feedback $h(i_t, y_t)$ [6]. When clear from context, we will denote by $\ell(i,t)$ the loss of arm i at time t and by $h(i,t)$ the feedback of arm i at time t. The expected regret here is:

$$\max_{i^*\in[K]} \mathbb{E}\left[\sum_{t=1}^{T}\ell(i_t, y_t) - \sum_{t=1}^{T}\ell(i^*, y_t)\right]$$

Revealing Action Partial Monitoring First consider the class of partial monitoring games with a *revealing action* – that is, suppose H has a row with N distinct elements. It is clear that if the learner plays this row, they can receive full information regarding which column the adversary has chosen. The cost of playing this row very well defines which *class* this game falls into (see for example the spam game discussed in (Lattimore and Szepesvári 2019)), but in general, the minimax regret of these games scales as $T^{2/3}$ and these games therefore fall in the *hard* class of games. Revealing action games and label efficient prediction differ in the way they *charge* the learner for information. For label efficient prediction, we have seen that there is a fixed number of times (budget) one can obtain information, but there is no additional cost of doing so. In revealing action games however, there is a loss associated to each time the learner asks for information. We will now show a reduction from this class of games to the standard label efficient prediction we discussed in sections 2 and 3.

Let the cost of playing the revealing action be $c = \max_{b\in[N]} L(a,b)$ where $a \in [K]$ is the revealing action row of L. Suppose α is the probability with which we play the revealing action at each round. α here corresponds to the ϵ from earlier sections, however α is now a free parameter[7]. We will still run reservoir sampling in the background as before to obtain the optimistic messages m_t. Now, in this light, the following theorem can be seen to follow from Theorem 2.

Theorem 7. *For* $a_t = 6\eta\alpha^2(\tilde{\ell}_t - m_t)^2$, $\tilde{\ell}_t = \frac{(\ell_t - m_t)}{\alpha}\mathbb{1}_{\{d_t=1\}} + m_t$, $\alpha \leq 1$ *and* $\eta \leq 1/162K$ *where the sequence of messages* m_t *are generated using reservoir sampling, the expected regret of Algorithm 2 modified for revealing action partial monitoring games with loss entries in* $[0,1]$ *satisfies the following:*

$$\mathbb{E}\left[R_T\right] \leq \frac{\log K + \log T}{\alpha\eta} + 18\eta Q^* + \alpha T c + (\log T)^2 \,.$$

Optimising the parameters η *and* γ *yields a bound of* $\mathcal{O}\Big((Q^* T \log K)^{1/3}\Big)$.

Note that now, we will again have to allocate rounds specifically for reservoir sampling as was the case with bandits, hence the additive $(\log T)^2$ term. The added $\alpha T c$ corresponds to the cost paid for playing the revealing action.

Hard Partial Monitoring Games We now turn to the *hard* class of partial monitoring games. As mentioned in (Piccolboni and Schindelhauer 2008) and (Cesa-Bianchi and Lugosi 2006), we will assume that there exists a matrix W such that $L = WH$. This is not an unreasonable assumption, as if this does not hold for the given L and H, one can suitably modify (see (Piccolboni and Schindelhauer 2008)) L and H to ensure $L' = W'H'$, and if this condition continues to fail after appropriate modifications, (Piccolboni and Schindelhauer 2008) show that sublinear regret is not possible for the original $G = (L, H)$. Observe that $L = WH$ will allow us to write $\ell(i,t) = \sum_{j\in[K]} w(i,j)h(j,t)$. Therefore:

$$\tilde{\ell}(i,t) = \frac{\left(\sum_{j\in[K]} w(i,j)h(j,t) - m_{t,i}\right)\mathbb{1}_{\{i_t=j\}}}{x_{t,j}} + m_{t,i}$$

is now an unbiased estimate of $\ell(i,t)$. m_t is still the optimistic messages where $m_{t,i}$ corresponds to an estimate of the average loss incurred by arm i till time t. These will still be obtained using reservoir sampling and we will maintain a separate reservoir for each arm $i \in [K]$. Note that since $\ell(i,t) = \sum_{j\in[K]} w(i,j)h(j,t)$ and the matrices L, W, and H are all visible to the learner, playing action r at time t for example will allow the learner to observe the rth component $w(i,r)h(r,t)$ of the loss for each action $i \in [K]$. Therefore,

[6] We are considering oblivious adversarial opponents as before and further take entries of H to be in $[0,1]$. The assumption on the entries is not major since the learner can always appropriately encode the original entries by numbers.

[7] Note that the update rules (1) and (2) will now also have α in place of ϵ

by maintaining an estimate (reservoir) for each *component*, we will be able to maintain an estimate for each arm.

Now, for these games we will use optimistic OMD without second order corrections (Rakhlin and Sridharan 2012; Chiang et al. 2012). The update rules are the same as equations (5) and (6) without the ϵ term. Additionally, the arm we play will be sampled from w_t where $w_t = (1-\gamma)x_t + \gamma\mathbf{1}$. The forced exploration is necessary to allow a minimum mass on all arms. Note that the structure defined by $\ell(i,t) = \sum_{j\in[K]} w(i,j)h(j,t)$ says that we potentially have to play *all* arms to maintain unbiased estimates of *any* arm. This forced exploration is unavoidable (see (Cesa-Bianchi and Lugosi 2006)).

Theorem 8. *Given $G = (L, H)$ with loss entries in $[0,1]$, a matrix W such that $L = WH$, $\eta > 0$ and $\mathcal{R} = {}^1\!/\!_\eta \sum_{i=1}^K x_i \log x_i$, the update rules (5) and (6) (omitting the ϵ) mixed with γ forced exploration satisfies: $\mathbb{E}[R_T] \leq \frac{\log K}{\eta} + \frac{KQ\eta}{2\gamma} + \gamma T$. Optimising for η and γ gives us a regret of $\mathcal{O}((QTK)^{1/3})$.*

Note here the strong dependence on K which is an outcome of each $\ell(i,t)$ being dependent on potentially all (K) other actions.

5 Lower Bounds

We now prove explicit quadratic variation-based lower bounds for (standard) label efficient prediction and label efficient bandits. By capturing both the constraint on information as well as the quadratic variation of the loss sequence, our lower bounds 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).

Our bounds will be proven in a 2-step manner similar to that in (Gerchinovitz and Lattimore 2016). 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 (defined below) which also incurs regret of the same order. Recall the quadratic variation for a given loss sequence: $Q = \sum_{t=1}^T \|\ell_t - \mu_T\|_2^2 \leq {}^{TK}\!/\!_4$. Now, for $\alpha \in [0, {}^1\!/\!_4]$ define an α-variation ball as: $\mathcal{V}_\alpha \triangleq \{\{\ell_t\}_{t=1}^T : {}^Q\!/\!_{TK} \leq \alpha\}$.

Theorems 10 and 12, after incorporating $Q \leq \alpha TK$ 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 $\mathcal{O}(\sqrt{{}^{QTK}\!/\!_n})$.[8] 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 (Gerchinovitz and Lattimore 2016) for the specific case of the multi-armed bandit problem, and was closed recently in (Bubeck, Cohen, and Li 2017). Barring the easy to see $\sqrt{{}^{(Q\log K)}\!/\!_K}$ lower bound for prediction with expert advice (which is also what Theorem 10 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 $\mathcal{O}(\sqrt{Q\log K})$ ((Hazan and Kale 2010), (Steinhardt and Liang 2014) 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.

[8] 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 $Q^* \leq Q$.

Label Efficient Prediction (Full Information) As mentioned previously, the main difference here from the standard label efficient prediction lower bound proof (Cesa-Bianchi, Lugosi, and Stoltz 2005) is that of centering the Bernoulli random variables around a parameter α which is responsible for ultimately bringing out the quadratic variation of the sequence. Our main statements for label efficient prediction are as follows.

Lemma 9. *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 $c = {}^{\sqrt{e}}\!/\!_{\sqrt{5(1+e)}}$.*

Theorem 10. *Let $K \geq 2$, $T \geq n \geq \max\{32\log(K-1),\ 256\log T\}$ 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, $\max_{\{\ell_t\}\in v_\alpha} \mathbb{E}[R_T] \geq 0.36T\sqrt{\frac{\alpha\log(K-1)}{n}}$.*

Label Efficient Bandits The main difference here 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 $t-1$).

Lemma 11. *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 $\mathbb{E}[R_T] \geq \frac{T}{8}\sqrt{\alpha(1-\alpha)K/n}$.*

Theorem 12. *Let $K \geq 2$, $T \geq n \geq \max\{32K,\ 384\log T\}$ and $\alpha \in \left[\max\left\{\frac{2c\log T}{n}, \frac{8K}{n}\right\}, \frac{1}{4}\right]$ with $c = (4/9)^2(3\sqrt{5}+1)^2 \leq 12$. Then, for any randomized strategy for the label efficient bandit problem, $\max_{\{\ell_t\}\in v_\alpha} \mathbb{E}[R_T] \geq 0.04T\sqrt{\frac{\alpha K}{n}}$.*

6 Conclusion

We consider problems lying at the intersection of 2 relevant questions in online learning – how does one adapt to slowly varying data, and what best can be done with a constraint on information. As far as we know, the proposed algorithms

are the first to jointly address both of these questions. There remain plenty of open problems in the area. Seeing to what extent universal Q^* dependent algorithms can be obtained in starved information settings is a direction of future interest, as is closing the gap in K highlighted in Section 5. Moreover, extending the notion of adaptivity to partial monitoring games to consider locally observable games and even more interestingly, locally observable sub-games within hard games also remain open. Higher order lower bounds for partial monitoring games have also not been studied and one wonders to what extent adaptivity can help in partial monitoring.

7 Acknowledgments

Siddharth Mitra would like to thank GD Raghava and Prathamesh Mayekar for many helpful discussions.

References

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*, 229–243. Springer.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 263–272. JMLR. org.

Bartók, G.; Foster, D. P.; Pál, D.; Rakhlin, A.; and Szepesvári, C. 2014. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research* 39(4):967–997.

Bartók, G.; Pál, D.; and Szepesvári, C. 2011. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, 133–154.

Bubeck, S.; Li, Y.; Luo, H.; and Wei, C.-Y. 2019. Improved path-length regret bounds for bandits. *CoRR* abs/1901.10604.

Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.

Bubeck, S.; Cohen, M. B.; and Li, Y. 2017. Sparsity, variance and curvature in multi-armed bandits. In *ALT*.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press.

Cesa-Bianchi, N.; Lugosi, G.; and Stoltz, G. 2005. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory* 51(6):2152–2162.

Chiang, C.; Yang, T.; Lee, C.; Mahdavi, M.; Lu, C.; Jin, R.; and Zhu, S. 2012. Online optimization with gradual variations. In *COLT*, volume 23 of *JMLR Proceedings*, 6.1–6.20. JMLR.org.

Gerchinovitz, S., and Lattimore, T. 2016. Refined lower bounds for adversarial bandits. In *NIPS*, 1190–1198.

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

Hazan, E., and Kale, S. 2011. Better algorithms for benign bandits. *J. Mach. Learn. Res.* 12:1287–1311.

Hazan, E. 2016. Introduction to online convex optimization. *Found. Trends Optim.* 2(3-4):157–325.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(Apr):1563–1600.

Lattimore, T., and Szepesvári, C. 2019. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *ALT*, volume 98 of *Proceedings of Machine Learning Research*, 529–556. PMLR.

Piccolboni, A., and Schindelhauer, C. 2008. Discrete prediction games with arbitrary feedback and loss (extended abstract). 208–223.

Rakhlin, A., and Sridharan, K. 2012. Online learning with predictable sequences. *arXiv preprint arXiv:1208.3728*.

Steinhardt, J., and Liang, P. S. 2014. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *ICML*.

Wei, C.-Y., and Luo, H. 2018. More adaptive algorithms for adversarial bandits.