Optimistic Information Directed Sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of online learning in contextual bandit problems where the loss function is assumed to belong to a known parametric function class. We propose a new analytic framework for this setting that bridges the Bayesian theory of information-directed sampling due to Russo and Van Roy [2018] and the worst-case theory of Foster, Kakade, Qian, and Rakhlin [2021] based on the decision-estimation coefficient. Drawing from both lines of work, we propose an algorithmic template called Optimistic Information-Directed Sampling and show that it can achieve instance-dependent regret guarantees similar to the ones achievable by the classic Bayesian IDS method, but with the major advantage of not requiring any Bayesian assumptions. The key technical innovation of our analysis is introducing an optimistic surrogate model for the regret and using it to define a frequentist version of the Information Ratio of Russo and Van Roy [2018], and a less conservative version of the Decision Estimation Coefficient of Foster et al. [2021].

1 Introduction

We present a framework for the analysis of a family of sequential decision-making algorithms known as Information-Directed Sampling (**IDS**). First proposed by Russo and Van Roy [2018], **IDS** is a Bayesian algorithm that selects its policies by optimizing a measure called the *information-ratio*, which measures the tradeoff between instantaneous regret and information gain about the problem instance at hand. In a Bayesian setup, both components of the information ratio are explicit functions of the posterior distribution over models, and can thus be explicitly calculated. As shown by Russo and Van Roy [2018], the resulting algorithm can guarantee massive statistical gains over more common approaches like Thompson sampling [Thompson, 1933] or optimistic exploration methods [Lai and Robbins, 1985], and in particular can take advantage of the structure of the problem instance much more effectively. Realizing the same gains in a non-Bayesian setup (which we will sometimes call *frequentist*, for lack of a better word) is hard for multiple reasons, the most severe obstacle being that the true model is entirely unknown and Bayesian posteriors cannot be used to quantify the uncertainty about the model in a meaningful way. As such, defining appropriate notions of information gain and information ratio is not straightforward. This is the problem we address in this paper.

Our main contribution is constructing a version of information-directed sampling that is implementable without Bayesian assumptions, and yields frequentist versions of the same problem-dependent guarantees as the ones achieved by the original **IDS** method in a Bayesian setup. The key element in our approach is the introduction of a *surrogate model* that allows for a meaningful definition of the information ratio that is amenable to a frequentist analysis. This surrogate model is the function of an optimistically adjusted posterior distribution inspired by the "feel-good Thompson sampling" algorithm of Zhang [2022], and is used to estimate the components of the information

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

ratio: the regret and the information gain. With these components, it becomes possible to define an information ratio that is an explicit function of the optimistic posterior, which can then be optimized to yield a decision-making rule that we call "optimistic information-directed sampling" (**OIDS**).

For the sake of concreteness, we focus on the problem of contextual bandits and show that **OIDS** can not only recover worst-case optimal regret bounds in this case, but also satisfies problem-dependent guarantees that are commonly referred to as *first-order bounds* [Cesa-Bianchi et al., 2005, Agarwal et al., 2017, Allen-Zhu et al., 2018, Foster and Krishnamurthy, 2021]. Besides these general guarantees, we also provide some illustrative examples that show that **OIDS** can reproduce the expedited learning behavior of **IDS** on easy problems, but without requiring Bayesian assumptions.

Our methodology also draws inspiration from the analytic framework of Foster, Kakade, Qian, and Rakhlin [2021], developed for a very general range of sequential decision-making problems. Their analysis revolves around the notion of the *decision-estimation coefficient* (DEC), which quantifies the tradeoffs that need to be made between achieving low regret and gaining information about the true model in a way that is similar to the information ratio of Russo and Van Roy [2016]. The main contribution of Foster et al. [2021] is showing that the minimax regret in any sequential decision-making problem can be lower bounded in terms of the DEC, and they also show that nearly matching upper bounds can be achieved via a simple algorithm they call *estimation to decisions* (**E2D**). Unlike the information ratio, the DEC does not make use of a Bayesian posterior to quantify uncertainty, but is rather defined as a worst-case notion, and as such provides frequentist guarantees that hold uniformly for all problem instances. However, the worst-case nature of the DEC can also be seen as an inherent limitation of their framework. In particular, the **E2D** algorithm is also based on the same conservative notion of regret-information tradeoff, and thus all known guarantees for this algorithm (and its variants such as the ones proposed by Chen et al., 2022, Foster et al., 2023a,b, Kirschner et al., 2023) fail to take advantage of problem structures that may facilitate fast learning.

Our own framework unifies the advantages of the two threads of literature described above: unlike **E2D**, it is able to achieve instance-dependent guarantees and learn faster in problems with more structure, and, unlike standard **IDS**, it can do so without relying on Bayesian assumptions. Our analysis draws on elements of both lines of work, and also on the techniques introduced by Zhang [2022], as mentioned above.

We are not the first to attempt the generalization of **IDS** beyond the Bayesian setting. Kirschner and Krause [2018] proposed a frequentist alternative to the information ratio for the special case of loss functions that are linear in some unknown parameter, and constructed an appropriate version of **IDS** that is able to take advantage of certain problem structures and obtain guarantees that improve upon the minimax rates. Their approach has inspired a line of work aiming to prove tighter and tighter problem-dependent bounds for a range of sequential decision-making problems, but so far all of these results remained limited to linearly structured losses and observations [Kirschner et al., 2020, 2021, Hao et al., 2022]. In contrast, our notion of information ratio does not require any specific problem structure like linearity, and arguably constitutes a more universal generalization of **IDS** beyond the Bayesian setting.

Notation. The squared Hellinger distance between two probability distributions P and P' (with a common dominating measure Q) is defined as $\mathcal{D}_H^2\left(P, P'\right) = \frac{1}{2}\int\left(\sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}} - \sqrt{\frac{\mathrm{d}P'}{\mathrm{d}Q}}\right)^2 \mathrm{d}Q$, and the relative entropy (or Kullback–Leibler divergence) as $\mathcal{D}_{\mathrm{KL}}\left(P \| P'\right) = \int \log \frac{\mathrm{d}P}{\mathrm{d}P'} \mathrm{d}P$.

2 Preliminaries

We study contextual bandit problems with finite action spaces and parametric loss functions. The sequential interaction scheme between the *learner* and the *environment* consists of the following steps being repeated for a sequence of rounds $t = 1, 2, \ldots, T$:

- The environment picks a context $X_t \in \mathcal{X}$, possibly using randomization and taking into account the history of actions, losses and contexts,
- the learner observes X_t and picks an action $A_t \in \mathcal{A}$, possibly using randomization and taking into account the history of actions, losses and contexts,

- the learner incurs a loss L_t, drawn independently of the past from a fixed distribution that depends on X_t, A_t.

We denote the sigma-algebra generated by the interaction history between the learner and the environment up to the end of round t as $\mathcal{F}_t = \sigma(X_1, A_1, L_1, \dots, X_t, A_t, L_t)$, and the probabilities and expectations conditioned on the history as $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-1}, X_t]$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}, X_t]$.

We will suppose that the action space is finite with cardinality $|\mathcal{A}| = K$, and that the loss function belongs to a known parametric class, but is otherwise unknown to the learner. Specifically, we assume that there is a known parameter space Θ that parametrizes a class of loss functions $\ell : \Theta \times \mathcal{X} \times \mathcal{A} \to \boldsymbol{R}$, and a true parameter $\theta_0 \in \Theta$ such that $\mathbb{E}_t[L_t| X_t, A_t] = \ell(\theta_0, X_t, A_t)$. We will refer to this condition as *realizability*. The distribution of random losses under parameter θ generated in response to taking action a in context x will be denoted by $p(\theta, x, a)$, and we will write $p(\cdot|\theta, x, a)$ to designate the corresponding density with respect to a reference measure (usually the counting measure or Lebesgue measure). Unless stated otherwise, we will assume that the loss distribution is fully supported on the interval $[0, 1]$ for all parameters θ. Furthermore, we will often abbreviate $\ell(\theta, X_t, a)$ as $\ell_t(\theta, a)$ and $p(\theta, X_t, a)$ as $p_t(\theta, a)$ to lighten our notation. Our formulation will make central use of *policies* which prescribe randomized behavior rules for the learning agent. Precisely, a policy $\pi : \mathcal{X} \to \Delta_{\mathcal{A}}$ maps each context x to a distribution over actions denoted as $\pi(\cdot|x)$. Since we will mostly work with action distributions conditioned on the fixed contexts X_t, we will mostly represent policies as distributions over actions, and use the same notation $\pi \in \Delta_{\mathcal{A}}$ for this purpose. We will focus on learning algorithms that, in each round t, select a randomized policy $\pi_t \in \Delta_{\mathcal{A}}$ based on the interaction history $\mathcal{F}_{t-1}$ and X_t. We also define the *optimal loss* in round t under model parameter θ as $\ell_t^*(\theta) = \min_a \ell_t(\theta, a)$. The agent aims to make its decisions in a way in that minimizes the expected sum of losses, and in particular aims to incur nearly as little loss as the true optimal policy. The extent to which the learner succeeds in achieving this goal is measured by the (total expected) *regret* defined as

$$R_T(\theta_0) = \mathbb{E}\left[\sum_{t=1}^{T} (\ell_t(\theta_0, A_t) - \ell_t^*(\theta_0))\right]. \tag{1}$$

The expectation is over all sources of randomness: the agent's randomization over actions, the adversary's randomization over contexts and the randomness of the realization of the losses. We also define *instantaneous regret* of an action a under parameter θ for each t as

$$r_t(a; \theta) = \ell_t(\theta, a) - \ell_t^*(\theta),$$

and the instantaneous regret of policy π as $r_t(\pi; \theta) = \sum_a \pi(a) r_t(a; \theta)$. With this notation, the regret of the online learning algorithm can be written as $R_T(\theta_0) = \mathbb{E}\left[\sum_{t=1}^{T} r_t(\pi_t; \theta_0)\right]$.

3 Two competing theories of sequential decision making

Our work connects two well-established analytic frameworks for sequential decision making: the Bayesian framework of Russo and Van Roy [2018] and the worst-case framework of Foster, Kakade, Qian, and Rakhlin [2021]. We review the two in some detail below, highlighting some of their merits and limitations that we address in this paper.

3.1 The information ratio and Bayesian information-directed sampling

The influential work of Russo and Van Roy [2016, 2018] set forth an analytic framework based on a Bayesian learning paradigm where the true model parameter θ_0 is supposed to be sampled from a known prior distribution $Q_0 \in \Delta_\Theta$, and the performance of the learner is measured on expectation with respect to this random choice of instance. We refer to the expected regret under this prior as the *Bayesian regret*. Their work has established that the Bayesian regret of any algorithm can be upper bounded in terms of a quantity called the *Information Ratio* (IR). For the sake of exposition, we will follow the setup and notation of Neu et al. [2022], who study the Bayesian version of our contextual bandit setting, and define the information ratio of policy π in the t-th round of interaction as

$$\rho_t(\pi) = \frac{\left(\mathbb{E}_{\theta_0 \sim Q_t}[r_t(\pi; \theta_0)]\right)^2}{\mathrm{IG}_t(\pi)}. \tag{2}$$

In the above expression, both the numerator and the denominator are functions of the *posterior distribution* Q_t of the parameter θ_0, computed based on all information available to the learner up to the beginning of round t. Specifically, the numerator is the squared expected regret in round t, where the expectation is taken under the posterior distribution Q_t, and the denominator is an appropriately defined measure of *information gain* that serves to quantify the amount of new information revealed about θ_0 after having observed the latest loss L_t. The information gain is formally defined as

$$\mathrm{IG}_t(\pi) = \sum_a \pi(a) \int \mathcal{D}_{\mathrm{KL}}\left(p_t(\theta, a) \| \overline{p}_t(a)\right) \mathrm{d}Q_t(\theta), \tag{3}$$

where $\overline{p}_t(a) = \int p_t(\theta, a)\mathrm{d}Q_t(\theta)$ is the posterior predictive distribution of the loss L_t given that action a is played in context X_t. In other words, the information gain is the *mutual information* between the posterior-sample parameter $\theta_t \sim Q_t$ and a randomly sampled loss $\widehat{L}_t \sim p_t(\theta_t, a)$.

Given the above definitions, Russo and Van Roy [2016, 2018] show that the Bayesian regret of *any* algorithm can be upper bounded as follows:

$$\mathbb{E}_{\theta_0 \sim Q_0}\left[R_T(\theta_0)\right] \leq \sqrt{\mathbb{E}\left[\sum_{t=1}^T \rho_t(\pi_t)\right] \cdot \mathbb{E}\left[\sum_{t=1}^T \mathrm{IG}(\pi_t)\right]}. \tag{4}$$

The second sum above can be upper bounded by the entropy of θ_0 under the prior distribution, regardless of what algorithm is used to select the sequence of policies. This suggests that one can achieve low regret by picking the sequence of policies in a way that minimizes the information ratio: $\pi_t = \arg\min_\pi \rho_t(\pi)$. This algorithm is called *information-directed sampling* (**IDS**), and has been shown to achieve regret guarantees that often improve significantly over worst-case bounds achieved by more traditional methods based on posterior sampling or optimistic exploration methods. In particular, for the contextual bandit setting we study in this paper, the works of Neu et al. [2022] and Min and Russo [2023] have shown that the information ratio of **IDS** is bounded by the number of actions K. When the parameter space is finite with cardinality N, this result implies that the algorithm achieves the minimax optimal regret bound of $\mathcal{O}(\sqrt{KT \log N})$ for this Bayesian setting.

Despite their appealing properties, **IDS**-style methods have however remained largely limited to the Bayesian setting, as there appears to be no universal way of defining an algorithmically useful information ratio without Bayesian assumptions. In particular, the instantaneous regret $r_t(\pi; \theta_0)$ cannot be computed without knowledge of θ_0, and there is no reason to believe that the information gain defined in terms of a Bayesian posterior would meaningfully measure the reduction in uncertainty about θ_0 in this more general setting.

3.2 The decision-estimation coefficient and the estimations-to-decisions algorithm

The fundamental work of Foster et al. [2021] provides a general theory of sequential decision making, providing a range of upper and lower bounds depending on a quantity they call the *decision-estimation coefficient* (DEC). With a little deviation from their notation and terminology, the DEC associated with a policy π, a model class Θ and a "reference model" $\widehat{p}_t : \mathcal{A} \to \Delta_{\mathbb{R}}$ is defined as

$$\mathrm{DEC}_{\gamma,t}(\pi; \Theta, \widehat{p}) = \sup_{\theta \in \Theta} \sum_a \pi(a) \left(\ell(\theta, X_t, a) - \ell(\theta, X_t, \pi_\theta) - \gamma \mathcal{D}_H^2\left(p_t(\theta, a), \widehat{p}_t(a)\right)\right), \tag{5}$$

where $\gamma > 0$ is a trade-off parameter. With this notation, Foster et al. [2021] define the decision-estimation coefficient associated with the model class Θ as

$$\mathrm{DEC}_\gamma(\Theta) = \sup_t \sup_{\widehat{p} \in \Delta_{\mathbb{R}}} \inf_{\pi \in \Delta_{\mathcal{A}}} \mathrm{DEC}_{\gamma,t}(\pi; \Theta, \widehat{p}).$$

Besides the remarkable feat of showing that the minimax regret can be lower bounded in terms of the above quantity, they also show that nearly matching upper bounds can be achieved via a simple algorithm they call *estimation to decisions* (**E2D**). In each round t, **E2D** takes as input a reference model $\widehat{p}_t$ and outputs the policy achieving the minimum in the definition of the DEC: $\pi_t = \arg\min_\pi \mathrm{DEC}_{\gamma,t}(\pi; \Theta, \widehat{p}_t)$. They show that the regret of this method can be upper bounded in terms of the DEC as follows:

$$R_T(\theta_0) \leq \mathrm{DEC}_\gamma(\Theta) \cdot T + \gamma \sum_{t=1}^T \mathcal{D}_H^2\left(p_t(\theta_0, a), \widehat{p}_t(a)\right).$$

This shows that the regret of **E2D** can be upper bounded as the sum of the DEC of the model class Θ and the total *estimation error* associated with the sequence of predictions $\widehat{p}_t$ (measured in terms of Hellinger distance). For the contextual bandit setting with finite parameter class of size N, they show that the total estimation error can be upper bounded by $\gamma \log N$ (under an appropriate choice of the predictions $\widehat{p}_t$), and that the DEC is upper bounded by K/γ, which once again recovers the minimax optimal rate of order $\mathcal{O}(\sqrt{KT \log N})$ when γ is tuned correctly.

A significant problem with the approach outlined above is that the DEC is an inherently worst-case measure of complexity due to the supremum taken over θ in its definition (5). Since the **E2D** algorithm itself is based on this possibly loose bound on the regret-to-information gap, this looseness may not only affect the bound but also the actual performance of the algorithm. Intuitively, one may hope to be able to do better by replacing the supremum over model parameters by only considering models that are still "statistically plausible" in an appropriate sense. In what follows, we provide an algorithm that realizes this potential.

4 Optimistic information-directed sampling

Our approach solves the issues outlined in the previous sections with both the Bayesian information ratio and the decision estimation coefficient. In particular, our method will extend Bayesian **IDS** by being able to provide non-Bayesian performance guarantees, and will be able to address the over-conservative nature of the DEC and provide strong instance-dependent guarantees.

Following Zhang [2022], we start by defining the *optimistic posterior* $Q_t^+ \in \Delta_\Theta$ via the following recursive update rule (starting from an arbitrary prior $Q_1^+(\theta) \in \Delta_\Theta$):

$$\frac{dQ_{t+1}^+}{dQ_t^+}(\theta) \propto (p_t(L_t|\theta, A_t))^\eta \cdot \exp(-\lambda \cdot \ell_t^*(\theta)). \tag{6}$$

Here, η and λ are positive constants that will be specified later. For now, we will only say that η should be thought of as a "large" constant of order 1, and λ as a "small" parameter of order $1/\sqrt{T}$ in the worst case. To proceed, we define the *optimistic posterior predictive distribution* of the loss for each t and a as the mixture $\overline{p}_t(a) = \int p_t(\theta, a) dQ_t^+(\theta)$, and the *surrogate loss function* and *surrogate optimal loss function* respectively as

$$\bar{\ell}_t(a) = \int \ell_t(\theta, a) \mathrm{d}Q_t^+(\theta) \qquad \text{and} \qquad \bar{\ell}_t^* = \int \ell_t^*(\theta) \mathrm{d}Q_t^+(\theta). \tag{7}$$

In words, these quantities are averages with respect to a mixture model over all contextual bandit instances with mixture weights given by the optimistic posterior Q_t^+. Notably, they are *improper* estimators of the true likelihood, loss, and optimal loss functions respectively, as there may be no single $\theta \in \Theta$ that corresponds to these exact functions (unless one assumes certain convexity properties of the relevant objects). With these notations, we define the *surrogate regret* of policy π in round t as $\overline{r}_t(\pi) = \bar{\ell}_t(\pi) - \bar{\ell}_t^*$. As we will see in the analysis, the optimistic posterior plays a key role in ensuring that the surrogate regret does not overestimate the true regret by too much on average, which makes it a sensible target for minimization.

It remains to define our notion of information gain that we will call *surrogate information gain*. Formally, this quantity is defined for each policy π as follows:

$$\overline{\mathrm{IG}}_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a) \int \mathcal{D}_H^2\big(p_t(\theta, a), \overline{p}_t(a)\big) \mathrm{d}Q_t^+(\theta). \tag{8}$$

Notably, this definition matches the original notion of information gain used by Russo and Van Roy [2016, 2018], up to the differences that the divergence being used is the squared Hellinger divergence instead of Shannon's relative entropy, and that the expectation is taken over the optimistic posterior instead of the plain Bayesian posterior. We will sometimes write $\overline{r}_t(\pi; Q_t^+)$ and $\overline{\mathrm{IG}}_t(\pi; Q_t^+)$ to emphasize that these are functions of the optimistic posterior Q_t^+. With the above definitions, we are now ready to introduce the central quantity of our algorithmic framework and our analysis: the *surrogate information ratio* defined for each policy π as

$$\overline{\mathrm{IR}}_t(\pi) = \frac{\big(\overline{r}_t(\pi)\big)^2}{\overline{\mathrm{IG}}_t(\pi)} = \frac{\Big(\sum_{a \in \mathcal{A}} \pi(a) \int \big(\ell_t(\theta, a) - \bar{\ell}_t^*(\theta)\big) \mathrm{d}Q_t^+(\theta)\Big)^2}{\sum_{a \in \mathcal{A}} \pi(a) \int \mathcal{D}_H^2\big(\overline{p}_t(a), p_t(\theta, a)\big) \mathrm{d}Q_t^+(\theta)}. \tag{9}$$

Importantly, computing the surrogate information ratio does not require knowledge of θ_0: both its denominator and numerator can be expressed in terms of the optimistic posterior Q_t^+. To emphasize this fact, we will sometimes write $\overline{\mathrm{IR}}_t(\pi; Q_t^+)$ for $\overline{\mathrm{IR}}_t(\pi)$.

We will also define the "offset" counterpart of the surrogate information ratio that is more closely related to the decision-estimation coefficient of Foster et al. [2021]. Following the terminology introduced in Section 3.2, we introduce the *averaged decision-estimation coefficient* (ADEC) of policy π for each $\mu > 0$ as

$$\begin{aligned}\overline{\mathrm{DEC}}_{\mu,t}(\pi) &= \bar{r}_t(\pi) - \mu \cdot \overline{\mathrm{IG}}_t(\pi) \\ &= \sum_a \pi(a) \int \left(\ell_t(\theta, \pi) - \ell_t^*(\theta) - \mu \mathcal{D}_H^2\left(\ell_t(\theta, \pi), \bar{\ell}_t(\pi)\right)\right) \mathrm{d}Q_t^+(\theta).\end{aligned} \tag{10}$$

Once again, we also define the notation $\overline{\mathrm{DEC}}_{\mu,t}(\pi; Q_t^+) = \overline{\mathrm{DEC}}_{\mu,t}(\pi)$ to emphasize the dependence of the ADEC on the posterior distribution Q_t^+. This definition departs from the classic DEC in that, instead of taking a supremum over model parameters, it is defined via an expectation with respect to the optimistic posterior, thus preventing overly conservative choices of θ. It should be clear from this definition that the ADEC is always smaller than its original counterpart defined by Foster et al. [2021], as long the latter uses the optimistic posterior predictive distribution as its reference model: $\overline{\mathrm{DEC}}_{\mu,t}(\pi; Q_t^+) \leq \mathrm{DEC}_{\mu,t}(\pi; \overline{p}_t, \Theta)$.

The surrogate information ratio and the ADEC are related to each other by the inequality

$$\overline{\mathrm{DEC}}_{\mu,t}(\pi) \leq \frac{\overline{\mathrm{IR}}_t(\pi)}{4\mu} \tag{11}$$

that holds for all $\mu > 0$. Conversely, it can be seen that

$$\overline{\mathrm{IR}}_t(\pi) = \inf\left\{C > 0 : \overline{\mathrm{DEC}}_{\mu,t}(\pi) \leq \frac{C}{4\mu} \quad (\forall \mu > 0)\right\}. \tag{12}$$

These are both direct consequences of the inequality of arithmetic and geometric means. That is, whenever the ADEC behaves as C_t/μ for all μ, the surrogate information ratio succinctly summarizes its behavior at all levels μ. We will dedicate special attention to this case below, but we also note that there are several important cases where the ADEC behaves differently, and the information ratio is a less appropriate notion of complexity. We defer further discussion of this to Section 7.

With the above notions, we are now ready to define the algorithmic framework we study in this paper, with two separate versions depending on whether we consider the surrogate information ratio or the average DEC as the basis of decision making. Both versions are referred to as *optimistic information-directed sampling* (optimistic **IDS** or **OIDS**). Following the terminology of Hao and Lattimore [2022], we call the first variant which selects its policies as $\pi_t = \arg\min_\pi \overline{\mathrm{IR}}(\pi; Q_t^+)$ *vanilla optimistic information-directed sampling* (**VOIDS**), and the second variant that selects $\pi_t = \arg\min_\pi \overline{\mathrm{DEC}}_\mu(\pi; Q_t^+)$ *regularized optimistic information-directed sampling* (**ROIDS**). We provide the pseudocode for these methods for quick reference in Appendix A.

5 Main results

We now present our main results regarding the two varieties of our optimistic **IDS** algorithm. We first show a general worst-case regret bound stated in terms of the time horizon T and the information ratio. More importantly, we also show instance-dependent guarantees on the performance of **OIDS** that replace the scaling with T in the upper bounds by the total loss of the best policy after T steps. For simplicity of exposition and easy comparison with existing results, we will present our main results assuming that the parameter space Θ is finite with cardinality N, and that the losses are almost surely bounded in the interval $[0, 1]$. We extend these results to compact metric parameter spaces in Section 5.3, and provide an extension to subgaussian losses in Section 5.4. Besides these general results, we also present several examples where **OIDS** can achieve very low regret by exploiting various flavors of problem structure, in Appendix B.

5.1 Worst-case bounds

We start by stating a general worst-case regret bound that relates the regret of any algorithm to its surrogate information ratio. This result is the non-Bayesian counterpart of the bounds stated in Russo and Van Roy [2018], Hao and Lattimore [2022] and Neu et al. [2022] in that it basically says that any algorithm with bounded information ratio will enjoy bounded regret.

Theorem 1. *Assume $|\Theta| = N < \infty$ and let $\lambda > 0$ be arbitrary. Then, for any choice of prior $Q_1 \in \Delta_\Theta$, the regret of any algorithm satisfies the following bound:*

$$\begin{aligned}\mathbb{E}\left[R_T(\theta_0)\right] &\leq \frac{\log \frac{1}{Q_1(\theta_0)}}{\lambda} + \lambda T \cdot \left(\frac{\sum_{t=1}^T \mathbb{E}\left[\overline{\mathrm{DEC}}_{1/10\lambda,t}(\pi_t; Q_t^+)\right]}{\lambda T} + \frac{21}{4}\right) \\ &\leq \frac{\log \frac{1}{Q_1(\theta_0)}}{\lambda} + \lambda T \cdot \left(10 \cdot \frac{\sum_{t=1}^T \mathbb{E}\left[\overline{\mathrm{IR}}_t(\pi_t; Q_t^+)\right]}{T} + \frac{21}{4}\right).\end{aligned} \tag{13}$$

We provide a proof sketch, with pointers to the full technical proof details, in Section 6.1. As is common in the information directed sampling literature, we will turn this guarantee into a more concrete bound on the regret of **OIDS** by exhibiting a "forerunner" algorithm that is able to control the surrogate information ratio and is relatively easier to analyze. Indeed, this will certify a regret bound for **OIDS**, since the latter precisely minimizes the surrogate information ratio at every round, and as such is guaranteed to achieve the same or a better bound. In particular, we use the *feel-good Thompson sampling* (**FGTS**) algorithm of Zhang [2022] as our forerunner, which samples a parameter θ_t from the optimistic posterior and then plays the policy $\pi_t = \arg\max_\pi \sum_a \pi(a)\ell_t(\theta_t, a)$.

Lemma 1. *The surrogate information ratio and averaged decision-to-estimation-coefficient of **VOIDS** and **ROIDS** satisfy for any $\mu \geq 0$*

$$4\mu\overline{\mathrm{DEC}}_{\mu,t}(\textbf{ROIDS}) \leq 4\mu\overline{\mathrm{DEC}}_{\mu,t}(\textbf{VOIDS}) \leq \overline{\mathrm{IR}}_t(\textbf{VOIDS}) \leq \overline{\mathrm{IR}}_t(\textbf{FGTS}) \leq 8K. \tag{14}$$

We note that the above result is more of a property of the posterior sampling policy than **FGTS** itself, as the bound holds for any distribution that is handed to **OIDS**. This result is not especially new: similar statements have been proven in a variety of papers including Russo and Van Roy [2016, 2018], Zhang [2022], Foster et al. [2021], Neu et al. [2022]. We provide a proof in Appendix E.4.1. Putting the two previous results together, we get the following upper bound on the regret of **OIDS**:

Corollary 1. *Assume $|\Theta| = N < \infty$, and let $\lambda = \sqrt{\frac{\log N}{(80K + \frac{21}{4})T}}$. Then, the regret of **ROIDS** with input parameter $\mu = \frac{1}{10\lambda}$ and **VOIDS** both satisfy*

$$\mathbb{E}\left[R_T\right] \leq \sqrt{(320K + 21)\, T \log N}. \tag{15}$$

In particular, this recovers the minimax optimal rate of $\mathcal{O}(\sqrt{KT \log N})$ for this problem.

5.2 First-order bounds

We now present a more interesting result that replaces the dependence on T in the previous bound by the cumulative loss of the best policy—constituting an instance-dependent guarantee that is often called *first-order regret bound*. In particular, in the important class of "noiseless" problems where the optimal loss is zero, the result implies that **OIDS** achieves constant regret.

Theorem 2. *Assume $|\Theta| = N < \infty$, let L^* be such that $\mathbb{E}\left[\sum_{t=1}^T \ell_t^*(\theta_0)\right] \leq L^*$, and let $\lambda = \sqrt{\frac{5 \log N}{(500K+108)L^*}} \wedge \frac{1}{250K+54}$. Then the regret of **ROIDS** with input parameter $\mu = \frac{1}{10\lambda}$ and **VOIDS** both satisfy*

$$\mathbb{E}\left[R_T\right] \leq \sqrt{(2500K + 540) \log N L^*} + (1250K + 270) \log N. \tag{16}$$

We provide a proof in Appendix D.1.

5.3 Infinite parameter spaces

We extend the result of Theorem 1 to work for infinite parameter spaces. For simplicity, we focus on the case in which Θ is a bounded subset of a finite-dimensional vector space.

Theorem 3. *Assume $\Theta \subset \mathbb{R}^d$, $max_{x,y\in\Theta} \|x - y\| = 2R < \infty$. Assume that for all $x \in X, a \in \mathcal{A}$, and $L \in [0, 1]$, the log-likelihood of the losses $p(\cdot, x, a, L)$ is C-Lipschitz. Assume that a ball of* *radius $\frac{1}{CT}$ containing θ_0 is included in Θ and set $\lambda = \sqrt{\frac{2d\log(RCT)}{(20K+\frac{21}{4})T}}$ and Q_1 a uniform prior on Θ. Then the regret of **ROIDS** with input parameter $\mu = \frac{1}{10\lambda}$ and **VOIDS** both satisfy*

$$\mathbb{E}\left[R_T\right] \leq \sqrt{(160K+42)dT\log(CRT)} + 1 = \mathcal{O}(\sqrt{dKT\log(CRT)}). \tag{17}$$

We provide a proof in Appendix D.2.

5.4 Subgaussian losses

We also extend the basic result of Theorem 1 to work for a more general family of losses. In particular, we drop the assumption that the likelihood model is well-specified and allow the losses to be sub-Gaussian. As the following result shows, we can still recover our regret bound of $\mathcal{O}(\sqrt{KT\log N})$ with some minor tweaks of the algorithm and the analysis. The resulting method is called **OIDS-SG**, and is presented in Appendix D.3 in full detail, along with the proof of the theorem below.

Theorem 4. *Assume that the losses are v-sub-Gaussian, that $|\Theta| = N < \infty$ and set $\lambda = \sqrt{\frac{\log N}{\left(\frac{1}{4}+20(v\wedge 1)(1+K)\right)T}}$. Then the regret of **ROIDS-SG** with input parameter $\mu = \frac{1}{80\lambda(v\wedge 1)}$ and **VOIDS-SG** both satisfy*

$$\mathbb{E}\left[R_T\right] \leq \sqrt{(1+80(v\vee 1)(1+K))T\log N} = \mathcal{O}(\sqrt{KT\log N}). \tag{18}$$

6 Analysis

This section provides an outline of the proofs of our main results. We first give a high-level overview of the key ideas that are shared in all proofs, and then fill in provide further technical details that are required to prove Theorems 1. Theorems 2, 3 and 4 are proved in Appendices D.1, D.2 and D.3.

The core of our analysis is the following decomposition of the instantaneous regret in round t:

$$\begin{aligned}\mathbb{E}\left[r_t(\pi_t;\theta_0)\right] &= \mathbb{E}\left[\overline{r}_t(\pi_t)\right] + \mathbb{E}\left[r_t(\pi_t;\theta_0) - \overline{r}_t(\pi_t)\right] \\ &= \mathbb{E}\left[\overline{r}_t(\pi_t)\right] + \mathbb{E}\left[\mathbb{E}_t\left[\ell_t(\theta_0,A_t) - \overline{\ell}_t(A_t)\right]\right] + \mathbb{E}\left[\overline{\ell}_t^* - \ell_t^*(\theta_0)\right] \\ &= \mathbb{E}\left[\overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + \mu\overline{\mathrm{IG}}_t(\pi_t) + \mathrm{UE}_t + \mathrm{OG}_t\right].\end{aligned} \tag{19}$$

Here, in the last line we have introduced the notations $\mathrm{UE}_t = \mathbb{E}_t\left[\ell_t(\theta_0,A_t) - \overline{\ell}_t(A_t)\right]$ to denote the *underestimation error* of the losses incurred by our own policy π_t, and $\mathrm{OG}_t = \overline{\ell}_t^* - \ell_t^*(\theta_0)$ as the *optimatily gap* between the best loss possible in our mixture of models and the optimal loss attainable under the true parameter. The first term is small if the mixture model accurately evaluates the losses seen during learning (which is generally easy to ensure on average), and the second term is small if the model remains optimistic about the best attainable performance (which is facilitated by the optimistic adjustment to the posterior updates). An important quantity in the analysis is the (true) *information gain* of policy π defined as

$$\mathrm{IG}_t(\pi) = \sum_{a\in\mathcal{A}} \pi(a) \int \mathcal{D}_H^2\left(p_t(\theta_0,a,\cdot), p_t(\theta,a,\cdot)\right) \mathrm{d}Q_t^+(\theta). \tag{20}$$

This quantity is closely related to the surrogate information gain that is optimized by our algorithm, and plays a key role in bounding the underestimation errors. In particular, the following simple lemma establishes a connection between the true and surrogate information gains:

Lemma 2. *For any t and policy π, the information gain satisfies $\overline{\mathrm{IG}}_t(\pi) \leq 4\mathrm{IG}_t(\pi)$.*

The proof can be found in Appendix E.2.1. Notably, the proof makes critical use of properties of the squared Hellinger distance, and is the main reason that the surrogate information gain is defined the way it is. In particular, the proof uses the fact that the Hellinger distance is a metric and as such it satisfies the triangle inequality—which is the reason that we were not able to go with the otherwise more natural choice of relative entropy in our definition of the information gain.

6.1 The proof of Theorem 1

We first use the following worst-case bound on the underestimation error:

Lemma 3. *For any t and $\gamma > 0$, the underestimation error is bounded as* $|\mathrm{UE}_t| \leq \frac{\gamma}{2} + \frac{\mathrm{IG}_t(\pi_t)}{\gamma}$.

The proof is relegated to Appendix E.1.1. Putting this bound together with the previous derivations, we get a regret bound that only depends on the averaged Decision-to-Estimation-Coefficient, the information gain and the optimality gap:

$$\mathbb{E}\left[r_t\right] \leq \mathbb{E}\left[\overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + \left(4\mu + \frac{1}{\gamma}\right)\mathrm{IG}_t(\pi_t) + \mathrm{OG}_t\right] + \frac{\gamma}{2}. \tag{21}$$

Following the terminology of Foster et al. [2023b], we will refer to the sum $\left(4\mu + \frac{1}{\gamma}\right)\mathrm{IG}_t(\pi_t) + \mathrm{OG}_t$ as the *optimistic estimation error*. The following result establishes that the optimistic posterior updates can effectively control a quantity that is closely related to this term.

Lemma 4. *Let $0 < \eta < \frac{1}{2}$, $\lambda > 0$, and $\beta = \frac{1}{1-2\eta}$. Then, the following inequality holds :*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{2\eta}{\lambda}\cdot\mathrm{IG}_t(\pi_t) + \mathrm{OG}_t\right)\right] \leq \frac{\log\frac{1}{Q_1(\theta_0)}}{\lambda} + \frac{\lambda\beta T}{8}. \tag{22}$$

See Appendix E.3.1 for the proof. It remains to pick the hyperparameters in a way that the left-hand side matches the total optimistic estimation error, which is achieved when setting way that $\frac{2\eta}{\lambda} = 4\mu + \frac{1}{\gamma}$. To make sure that this holds while minimizing the final constant, we choose $\eta = \frac{1}{4}$, $\beta = 2$, and $\gamma = \frac{1}{\mu} = 10\lambda$. Plugging these constants into the bound above, and putting the result together with the bound of Equation (21) completes the proof of Theorem 1.

7 Conclusion

We have proposed a new analysis framework that bridges the concepts of information ratio and decision-estimation coefficient, and unifies the advantages of both frameworks. We conclude by discussing some directions of future work. We expand our discussion of related works and open problems in Appendix C.

A very important open question is whether our notion of averaged DEC can also serve as a lower bound on the minimax regret like its original version proposed by Foster et al. [2021]. Since the ADEC is a lower bound on the DEC under a special choice of nominal model, we conjecture that it can also be used to lower bound the minimax regret in the same "low-probability" fashion as the original results of Foster et al. [2021]. On the same note, we remark that it seems unlikely that our DEC variant can be reconciled with the "constrained DEC" of Foster et al. [2023a], which has so far yielded the tightest lower bounds on the regret within this family of complexity notions. Whether or not the averaging idea we advocate for in this paper will turn out to be useful for fully characterizing the minimax regret in sequential decision making remains to be seen.

It is interesting to observe that the optimistic posterior updates used by our method simplify drastically in the special case of "noiseless" problems where $\ell^*(\theta, X_t) = 0$ holds for all θ. This condition holds in two of the examples discussed in Appendix B, and more broadly in all problems where the optimal policy is guaranteed to achieve zero loss under all candidate parameters θ. As a more concrete example, we highlight the problem of bandit linear classification with surrogate losses, which satisfies this condition if the data is separable with a margin [Kakade et al., 2008, Beygelzimer et al., 2017, 2019]. In such noise-free problems, the optimistic posterior update collapses to $\frac{dQ_{t+1}^+}{dQ_t^+}(\theta) \propto (p_t(L_t|\theta, A_t))^\eta$, which is closer to the standard Bayesian update up to the important difference that it involves the "stepsize" parameter η. Interestingly, such "generalized" or "safe" Bayesian updates have been studied extensively in the context of statistical learning under misspecified models—see, e.g., Zhang [2006a,b], Grünwald [2012], de Heide et al. [2020]. This connection leads to a multitude of questions that we cannot hope to address in this short discussion, so we close with mentioning only one aspect that we find to be particularly exciting. Specifically, we wonder if the techniques established in these works could be useful for addressing misspecification in the context of sequential decision making under uncertainty, where this issue has been notoriously hard to formalize and handle [Du et al., 2019, Lattimore et al., 2020, Weisz et al., 2021]. We leave this exciting question open for future research.

References

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, pages 3–11, 1999.

A. Agarwal, A. Krishnamurthy, J. Langford, and H. Luo. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, pages 4–7, 2017.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, pages 186–194, 2018.

A. Beygelzimer, F. Orabona, and C. Zhang. Efficient online bandit multiclass learning with $\widetilde{O}(\sqrt{T})$ regret. In *International Conference on Machine Learning*, pages 488–497, 2017.

A. Beygelzimer, D. Pál, B. Szörényi, D. Thiruvenkatachari, C.-Y. Wei, and C. Zhang. Bandit multiclass linear classification: Efficient algorithms for the separable case. In *International Conference on Machine Learning*, pages 624–633, 2019.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

S. Bubeck and M. Sellke. First-order Bayesian regret analysis of Thompson sampling. In *Algorithmic Learning Theory*, pages 196–233, 2020.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. In *Conference on Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 217–232. Springer, 2005.

F. Chen, S. Mei, and Y. Bai. Unified Algorithms for RL with Decision-Estimation Coefficients: No-Regret, PAC, and Reward-Free Learning, 2022.

R. de Heide, A. Kirichenko, P. Grünwald, and N. Mehta. Safe-bayesian generalized linear regression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 2623–2633, 2020.

S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019.

D. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210, 2020.

D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.

D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The Statistical Complexity of Interactive Decision Making, 2021.

D. J. Foster, N. Golowich, and Y. Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023a.

D. J. Foster, N. Golowich, J. Qian, A. Rakhlin, and A. Sekhari. Model-free reinforcement learning with the decision-estimation coefficient. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

P. Grünwald. The safe Bayesian: learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*, pages 169–183, 2012.

B. Hao and T. Lattimore. Regret bounds for information-directed reinforcement learning. *Advances in Neural Information Processing Systems*, 35:28575–28587, 2022.

B. Hao, T. Lattimore, and C. Qin. Contextual information-directed sampling. In *International Conference on Machine Learning*, pages 8446–8464, 2022.

S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447, 2008.

J. Kirschner and A. Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384, 2018.

J. Kirschner, T. Lattimore, and A. Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369, 2020.

J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821, 2021.

J. Kirschner, A. Bakhtiari, K. Chandak, V. Tkachuk, and C. Szepesvári. Regret minimization via saddle point optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

T. Lattimore and A. György. Mirror Descent and the Information Ratio. In *Conference on Learning Theory*, volume 134, pages 2965–2992, 2021.

T. Lattimore, C. Szepesvári, and G. Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670, 2020.

S. Min and D. Russo. An information-theoretic analysis of nonstationary bandit learning. In *International Conference on Machine Learning*, 2023.

G. Neu, J. Olkhovskaya, M. Papini, and L. Schwartz. Lifting the information ratio: An information-theoretic analysis of Thompson sampling for contextual bandits. *Advances in Neural Information Processing Systems*, 35:9486–9498, 2022.

J. Olkhovskaya, J. Mayo, T. van Erven, G. Neu, and C.-Y. Wei. First-and second-order bounds for adversarial linear contextual bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *J. Mach. Learn. Res.*, 17:68:1–68:30, 2016.

D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Oper. Res.*, 66 (1):230–252, 2018.

S. Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. Hebrew University, 2007.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, 25:285–294, 1933.

G. Weisz, P. Amortila, and C. Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264, 2021.

T. Zhang. From $epsilon$-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5), 2006a.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.

T. Zhang. Feel-good Thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

A Pseudocode of OIDS

We provide the pseudocode for OIDS in Algorithm 1 below.

Algorithm 1 Optimistic Information Directed Sampling (**OIDS**)

Input: prior Q_1^+, parameters η, λ, μ.
For $t = 1, \ldots, T$, **repeat**:

1. Observe context X_t,

2a. **VOIDS**: play policy $\pi_t = \arg\min_{\pi \in \Delta(\mathcal{A})} \overline{\mathrm{IR}}_t(\pi, Q_t^+)$,

2b. **ROIDS**: play policy $\pi_t = \arg\min_{\pi \in \Delta(\mathcal{A})} \overline{\mathrm{DEC}}_t(\pi, Q_t^+, \mu)$,

3. incur loss L_t,
4. update optimistic prior, $Q_{t+1}^+(\cdot) \propto Q_t^+(\cdot)(p_t(\cdot, A_t, L_t))^\eta \exp(-\lambda \ell_t^*(\cdot))$.

B Examples

The most appealing property of **IDS** in the Bayesian setting is that it can take advantage of the structure of the problem at hand to achieve extremely good performance that is otherwise not achievable by methods like Thompson sampling or UCB. Indeed, unlike these methods, **IDS** has the ability to pick actions that are not optimal under any statistically plausible model, but can reveal useful information about the problem. Russo and Van Roy [2018] demonstrate several examples of situations where **IDS** provably achieves massive speedups via such queries. It is not clear that such speedups are achievable without Bayesian assumptions, although some evidence was offered by the work of Kirschner and Krause [2018] in the case of linear rewards. In this section, we demonstrate that our version of **IDS** can fully reproduce the fast learning behavior of Bayesian **IDS** on the original examples of Russo and Van Roy [2018], thus suggesting that **OIDS** may inherit many more good properties of its Bayesian counterpart than what our main theoretical results show. We also provide an additional example on which we demonstrate that **OIDS** can outperform DEC-based methods by addressing the over-conservatism encoded in the definition of the DEC.

B.1 Revealing action

We first adapt the "revealing actions" example of the original work of Russo and Van Roy [2018]. This example features the action set $\mathcal{A} = \{0, 1, \ldots, K\}$, the set of parameters $\Theta = \{1, \ldots, K\}$, and the loss function $\ell(\theta, a) = \mathbb{I}_{\{a>0,\, a\neq\theta\}} + \mathbb{I}_{\{a=0\}}(1 - \frac{1}{2^\theta})$. The losses are deterministic and the agent gets loss 0 by picking the action corresponding to the unknown parameter θ_0. Action 0 is special, it results in a large loss that however encodes the identity of the optimal action. Thus, the optimal exploration strategy is to pick this revealing action once, read out the identity of the optimal action, and play that action until the end of time. Russo and Van Roy [2018] show that **IDS** follows this exact strategy, and here we show that **OIDS** does the same when taking as input a (completely noninformative) uniform prior over the parameters.

To show this, we will compute for any action the surrogate reward and surrogate information gain under the optimistic posterior (which is identical to the uniform prior, given that we are in the first round). For $a \neq 0$, the surrogate regret is written as

$$\bar{r}_1(a) = \int_\Theta \ell(\theta, a) - \ell(\theta)\, dQ_0(\theta) = \frac{1}{K}\sum_{\theta=1}^{K}(1 - \mathbb{I}_{\{a=\theta\}}) = 1 - \frac{1}{K},$$

while for the revealing action, the surrogate regret is

$$\bar{r}_1(0) = 1 - \frac{1}{K} + \frac{2^{-K}}{K}.$$

In particular $\bar{r}_t(0) > \bar{r}_t(a)$ so the action 0 has the worst expected reward under our model. As for the information gain, we an explicit computation of the Hellinger distance for $a \neq 0$ shows

$$\mathrm{IG}_t(a) = \frac{1}{K} \cdot \left(1 - \sqrt{\frac{1}{K}}\right) + \frac{K-1}{K} \cdot \left(1 - \sqrt{\frac{K-1}{K}}\right) = \mathcal{O}\left(\frac{1}{K}\right).$$

Meanwhile, for action 0 we have

$$\mathrm{IG}_t(0) = 1 - \sqrt{\frac{1}{K}} = \Theta(1).$$

B.2 Sparse linear model

Our second example is a linear bandit problem where the action space corresponds to a finite subset of the Euclidean unit ball $\mathcal{A} = \{\frac{x}{\|x\|_1} : x \in \{0,1\}^d,\, x \neq 0\}$, the parameter space consists of the set of coordinate vectors $\Theta = \{\theta' \in \{0,1\}^d, \|\theta'\|_1 = 1\}$, and the loss function is $\ell(\theta, a) = 1 - \langle a, \theta \rangle$. As in the previous example, the losses are again deterministic. This is a linear bandit problem where the parameter θ is known to be 1-sparse. In particular, the optimal action under the model θ consists in only selecting action $a = \theta$ so any Thompson Samling based algorithm will only select one coordinate at a time and will take up to d steps to determine the true parameter θ_0. In contrast, the optimal exploration policy will perform binary search on the action space and find the optimal action exponentially faster.

To investigate the behaviour of **OIDS** on this problem, we will compute the surrogate regret and surrogate information gain of an action a. Since our prior is uniform, we have

$$\bar{r}_1(a) = \bar{\ell}_1(a) = \mathbb{P}\left[\langle \theta_0, a\rangle > 0\right] \cdot \frac{1}{\|a\|_1} = \frac{\|a\|_1}{d} \cdot \frac{1}{\|a\|_1} = \frac{1}{d}$$

and

$$\begin{aligned}\mathrm{IG}_1(a) &= \frac{\|a\|_1}{d} \cdot \left(1 - \sqrt{\frac{\|a\|_1}{d}}\right) + \frac{d - \|a\|_1}{d} \cdot \left(1 - \sqrt{\frac{d - \|a\|_1}{d}}\right) \\ &= 1 - \left(\frac{\|a\|_1}{d}\right)^{\frac{3}{2}} - \left(1 - \frac{\|a\|_1}{d}\right)^{\frac{3}{2}}\end{aligned}$$

Thus, the expected reward of all actions is the same, and the information gain is maximized for actions with norm $\|a\|_1 = \frac{d}{2}$. **IDS** thus picks an action A_1 uniformly at random and updates the posterior as follows. If the observed loss is 1, all parameters with $\langle \theta, A_1 \rangle > 0$ will be eliminated by the posterior update. If the observed loss is smaller than 1, all parameters satisfying $\langle \theta, A_1 \rangle = 0$ are excluded. The posterior is thus set as uniform over all surviving parameters and the process repeats. Continuing along the same lines, we can see that both versions of **OIDS** will continue performing binary search and identify the true parameter in $\log_2 d$ time steps.

B.3 Bandits with a revelatory zero

Our final example is a multi-armed bandit problem where the losses keep looking exactly the same until a low-probability event happens that reveals the optimal action perfectly. In this setup (vaguely inspired by Example 3.3 of Foster et al., 2021), $\Theta = [K]$, and the losses are defined as uniformly distributed random variables in $[0,1]$ for all actions except $a = \theta$. For this special action, the loss is defined as $B_t U_t$, with U_t uniform on $[0,1]$, and B_t is Bernoulli with mean $1 - 2\Delta \in [0,1]$. The mean loss for this action is $\frac{1}{2} - \Delta$. For this model, there is essentially no way for any algorithm to discover the optimal action until the first time that a loss of zero is observed. In this case, the (optimistic) posterior immediately collapses on θ_0. Consequently, **OIDS** keeps drawing uniform random actions until the first zero is observed, and plays the optimal action in all remaining rounds. The number of time steps spent with uniform exploration are geometrically distributed with mean $\frac{K}{2\Delta}$, thus making for a total regret of approximately $\frac{K}{2}$. Note that in this instance, the optimistic adjustment to the posterior is not necessary as the optimal loss of all models are the same, so the performance of the algorithm is unaffected by the choice of λ or μ.

Interestingly, the **E2D** algorithm of Foster et al. [2021] cannot take advantage of the structure of this problem so effectively. When using the posterior predictive distribution $\overline{p}_t$ as the nominal model, the Hellinger distance will approximately behave as $\mathcal{D}_H^2\left(p(\theta, a), \widehat{p}_t(a)\right) \approx \mathbb{I}_{\{\theta \neq \theta_0\}}$ after observing the first zero. Thus, the worst-case DEC associated with policy π is written as

$$\mathrm{DEC}_\gamma(\pi;\overline{p}_t,\Theta) = \sup_\theta \left\{\ell(\theta,\pi) - \ell(\theta,a_\theta) - \gamma \mathbb{I}_{\{\theta\neq\theta_0\}}\right\} = \sup_\theta \left\{\Delta \sum_{a\neq\theta} \pi(a) - \gamma \mathbb{I}_{\{\theta\neq\theta_0\}}\right\}$$
$$= \sup_\theta \left\{\Delta(1-\pi(\theta)) - \gamma \mathbb{I}_{\{\theta\neq\theta_0\}}\right\}.$$

When $\gamma \geq \Delta$, the expression in the supremum can be positive for certain policies π and parameters $\theta \neq \theta_0$, and thus the θ player will prefer picking $\theta \neq \theta_0$ for some policies. More precisely, the DEC for any policy will be given as

$$\mathrm{DEC}(\pi;\Theta,\widehat{p}_t) = \max\left\{\Delta(1-\min_{a\neq\theta_0}\pi(a)) - \gamma,\, \Delta(1-\pi(\theta_0))\right\}.$$

In the extreme case $\gamma = 0$, the policy achieving maximum value is approximately uniform, and it approximates the optimal policy π^* gradually as γ increases. When γ is large enough, the alternative $\theta \neq \theta_0$ stops being attractive to the max player and **E2D** starts outputting π^*. This happens at the threshold $\gamma > \Delta$ at the latest. This observation matches the discussion of Foster et al. [2021, Example 3.3] and Foster et al. [2023a, p. 8], who demonstrate the same threshold behavior of the DEC and point out that this leads to tight lower bounds, without discussing the potential shortcomings of **E2D** that prevents it from obtaining tight upper bounds. It is easy to see that **E2D** fails because of the over-conservative definition of the DEC: while there is sufficient evidence to reject all alternative parameters, **E2D** still computes its optimization objective by taking a supremum over *all* model parameters θ, including ones that have already been ruled out by the observations. This clearly demonstrates the advantage of the surrogate model used by **OIDS**, which computes its objective with the help of the optimistic posterior distribution that allows faster elimination of unlikely parameters.

C Further discussion

In this section, we expand our discussion of related works and open questions.

General bounded losses. At the surface level, it may seem that our results only apply to well-specified models where the likelihood model correctly captures the distribution of the random losses. This is of course a very restrictive assumption. However, it is easy to see that our framework can tackle arbitrary bounded losses via a standard binarization trick [Agrawal and Goyal, 2013]: supposing that the losses are bounded in $[0, 1]$, they can be randomly rounded to $\{0, 1\}$ to apply **OIDS** with a Bernoulli likelihood. It is easy to see that the regret bounds for these post-processed losses continue to hold for the original losses as well. We presume that our approach can be generalized beyond such sub-Bernoulli and sub-Gaussian losses to more general sub-exponential-family losses, but we leave the investigation of this generalization open for future work.

Beyond contextual bandits. For the sake of simplicity, we have presented our results within the relatively modest framework of contextual bandits. That said, it is clear that our framework can be generalized to the much broader setting of "decision making with structured observations" studied by Foster et al. [2021], and that it can be used to prove regret bounds of the form of Theorem 1 straightforwardly in said setting. However, so far we could only prove quantitative improvements over the DEC for contextual bandits, and thus we decided not to let down the reader by introducing a very general setting and then only providing interesting results in a narrow special case. Nevertheless, our results demonstrate that our framework can achieve strictly superior upper bounds on the regret in a highly nontrivial setting that has been studied extensively (see, e.g., Agarwal et al., 2017, Allen-Zhu et al., 2018, Foster and Krishnamurthy, 2021, Bubeck and Sellke, 2020, Olkhovskaya et al., 2023).

Multiplicative or additive tradeoff? All of our results are stated in terms of both the surrogate information ratio, which measures the regret-to-information tradeoff multiplicatively, and the averaged DEC, which does so in an additive fashion. Based on these results, it is not immediately clear which of the two notions is more useful. Equations (11) and (12) suggest that the ADEC is always smaller than the information ratio, which may suggest that it may yield better guarantees. To a certain degree, Russo and Van Roy [2018] have already addressed this question: their Proposition 11 shows that measuring the regret-information tradeoff additively results in strictly *worse* regret for a range of hyperparameter choices. While at the surface, this seems to defy the intuition provided our results, in reality their additive tradeoff is only vaguely related to the one we consider, and the regularization range for which the result holds does not seem to be practical in the first place. On the other hand, Foster et al. [2021] make a more robust argument against the information ratio in comparison with the DEC, showing that there are some hard problems for which the information ratio is infinite but the DEC remains finite (see their Section 9.3). Besides the fact that their information ratio is defined in an unorthodox way via the same conservative supremum as what appears in the definition of the DEC, this claim seems to miss some important follow-up work on **IDS** that has already addressed this issue. Specifically, Lattimore and György [2021] have pointed out that the information ratio is only suitable for problems where the minimax regret is of the order $\sqrt{T}$ (which one can already notice by inspecting the general bound of Equation 4), and studying harder games with larger minimax regret may be done by introducing a generalized notion of information ratio that features a different power of the regret in the denominator. In the present paper, we decided to stay impartial and state our results for both flavors of optimistic **IDS**, and we hope that this debate will progress productively in the future.

Connection with the Bayesian DEC. The attentive reader may have noticed that a notion closely related to our averaged DEC has already been mentioned in the original work of Foster et al. [2021]. Indeed, their Section 4.2 proposes a Bayesian version of the **E2D** algorithm that optimizes $\overline{\mathrm{DEC}}_{\gamma,t}(\cdot; Q_t)$, where Q_t is the exact Bayesian posterior over the model parameters. They show that the resulting algorithm enjoys essentially the same guarantees on the Bayesian regret as the worst-case guarantees obtained by the standard **E2D** method. Our approach effectively considers the same optimization objective, with the important change that the standard Bayesian posterior is replaced with the optimistic posterior of Zhang [2022]. This not only strengthens the mentioned results of Foster et al. [2021] by removing the Bayesian assumption necessary for its analysis, but also allows us to obtain instance-dependent guarantees as well. We believe that the same instance-dependent improvements (and more) should be directly provable for the Bayesian **E2D** method of Foster et al. [2021], but we did not pursue this direction as we preferred to focus on pointwise regret guarantees this time.

D Proofs of the main results

We now give the complete proofs of our main results. We relegate most of the technical content into Appendix E and only provide the main arguments here for better readability.

D.1 The proof of Theorem 2

We start our analysis from the regret decomposition of Equation (19) and apply Lemma 2 to obtain

$$\mathbb{E}\left[r_t\right] \le \mathbb{E}\left[\overline{\mathrm{DEC}}_{\mu,t}(\pi_t) + 4\mu \mathrm{IG}_t(\pi_t) + \mathrm{UE}_t + \mathrm{OG}_t\right].$$

As before, we can control the ADEC of **OIDS** by producing a suitable forerunner. In particular, we use the *inverse-gap weighting* **IGW** algorithm of Foster and Krishnamurthy [2021]

Lemma 5. *The surrogate information ratio and averaged decision-to-estimation-coefficient of* ***VOIDS*** *and* ***ROIDS*** *satisfy for any* $\mu \ge 0$

$$4\mu\overline{\mathrm{DEC}}_{\mu,t}(\textbf{ROIDS}) \le 4\mu\overline{\mathrm{DEC}}_{\mu,t}(\textbf{VOIDS}) \le \overline{\mathrm{IR}}_t(\textbf{VOIDS}) \le \overline{\mathrm{IR}}_t(\textbf{IGW}) \le 40K \min_{a\in\mathcal{A}} \bar{\ell}_t(a). \quad (23)$$

See Appendix E.4.2 for a definition of the (**IGW**) algorithm and the proof. The term on the right-hand side can be further bounded as

$$\overline{\mathrm{DEC}}_{\mu,t}(\pi_t) \le \frac{10K}{\mu}\min_a \bar{\ell}_t(a) \le \frac{10K}{\mu}\left(\mathbb{E}_t\left[\bar{\ell}_t(A_t)\right]\right) = \frac{10K}{\mu}\left(\mathbb{E}_t\left[\ell_t(\theta_0, A_t)\right] - \mathrm{UE}_t\right)$$

The final tool is a refined version of Lemma 3 that controls the underestimation error in terms of the information gain and the current estimate of the loss.

Lemma 6. *For any* t *and* $\gamma > 0$*, the underestimation error is bounded as*

$$\mathrm{UE}_t \le \frac{\mathrm{IG}_t(\pi_t)}{\gamma} + 2\gamma \mathbb{E}_t\left[\ell_t(\theta_0, A_t)\right]. \quad (24)$$

See Appendix E.1.2 for the proof. Putting this together with the previous regret decomposition, as long as $\frac{10K}{\mu} \leq 1$, we get:

$$\mathbb{E}\left[r_t\right] \leq \mathbb{E}\left[\left(4\mu + \frac{1}{\gamma}\cdot\left(1-\frac{10K}{\mu}\right)\right)\mathrm{IG}_t(\pi_t) + \mathrm{OG}_t + \left(2\gamma\left(1-\frac{10K}{\mu}\right)+\frac{10K}{\mu}\right)\ell_t(\theta_0, A_t)\right], \tag{25}$$

As before, we will regard the term $\left(4\mu + \frac{1}{\gamma}\cdot\left(1-\frac{10K}{\mu}\right)\right)\mathrm{IG}_t + \mathrm{OG}_t$ as the optimistic estimation error, and adapt Lemma 4 to provide a refined bound on this quantity:

Lemma 7. *Let* $0 < \eta < \frac{1}{2}$, $\lambda > 0$, *and* $\beta = \frac{1}{1-2\eta}$. *Then, the optimistic estimation error satisfies*

$$\sum_{t=1}^{T}\left(\frac{2\eta}{\lambda}\cdot\mathrm{IG}_t(\pi_t) + \left(1-\frac{\lambda\beta}{2}\right)\mathrm{OG}_t\right) \leq \frac{\log N}{\lambda} + \frac{\lambda\beta}{2}\sum_{t=1}^{T}\ell_t^*(\theta_0). \tag{26}$$

See Appendix E.3.2 for the proof. The claim of the theorem is then proved by tuning the hyperparameters in a way that the quantity bounded in the previous Lemma matches the optimistic estimation error.

Under the condition $\frac{10K}{\mu} \leq 1$, the following holds

$$\begin{aligned}\mathbb{E}\left[r_t\right] &\leq \mathbb{E}\left[\left(4\mu + \frac{1}{\gamma}\cdot\left(1-\frac{10K}{\mu}\right)\right)\mathrm{IG}_t(\pi_t) + \mathrm{OG}_t + \left(2\gamma\left(1-\frac{10K}{\mu}\right)+\frac{10K}{\mu}\right)\ell_t(\theta_0, A_t)\right] \\ &\leq \mathbb{E}\left[\left(4\mu + \frac{1}{\gamma}\right)\mathrm{IG}_t(\pi_t) + \mathrm{OG}_t + \left(2\gamma+\frac{10K}{\mu}\right)\ell_t(\theta_0, A_t)\right],\end{aligned}$$

where in the last line we also used that IG_t and $\ell_t(\theta_0, A_t)$ are nonnegative to upper bound $1-\frac{10K}{\mu} \leq 1$. In order to apply Lemma 7, we would like to manipulate the above expression so that the coefficients of IG_t and OG_t match. To this end, we use the condition that $\frac{\lambda\beta}{2} \leq \frac{1}{5}$, which ensures that $1 \leq \frac{1}{1-\frac{\lambda\beta}{2}} \leq \frac{5}{4}$ and thus we can continue the above bound as

$$\mathbb{E}\left[r_t\right] \leq \mathbb{E}\left[\frac{5}{4}\cdot\left(\left(4\mu+\frac{1}{\gamma}\right)\mathrm{IG}_t(\pi_t) + \left(1-\frac{\lambda\beta}{2}\right)\mathrm{OG}_t + \left(2\gamma+\frac{10K}{\mu}\right)\ell_t(\theta_0, A_t)\right)\right].$$

To apply Lemma 7, we choose $\eta = \frac{1}{4}, \beta = 2, \gamma = \frac{1}{\mu} = 10\lambda$, and sum over all rounds to obtain

$$\begin{aligned}\mathbb{E}\left[R_T\right] &\leq \mathbb{E}\left[\frac{5}{4}\cdot\frac{\log N}{\lambda} + \frac{5\lambda}{4}\sum_{t=1}^{T}\ell_t^*(\theta_0) + (125K+25)\lambda\sum_{t=1}^{T}\ell_t(\theta_0, A_t)\right] \\ &\leq \mathbb{E}\left[\frac{5}{4}\cdot\frac{\log N}{\lambda} + (125K+27)\lambda\sum_{t=1}^{T}\ell_t(\theta_0, A_t)\right],\end{aligned}$$

where we upper-bounded the optimal loss $\frac{5\lambda}{4}\ell_t^*(\theta_0)$ by $2\lambda\ell_t(\theta_0, A_t)$ in the last step. Introducing the notation $\widehat{L}_T = \sum_{t=1}^{T}\ell_t(\theta*, A_t)$ and $L_t^* = \sum_{t=1}^{T}\ell_t^*(\theta_0)$, the two sides of the equation can be rewritten as

$$R_T = \widehat{L}_T - L_t^* \leq \mathbb{E}\left[\frac{5}{4}\cdot\frac{\log N}{\lambda} + (125K+27)\lambda\widehat{L}_T\right].$$

Hence, after some reordering we arrive at

$$\mathbb{E}\left[R_T\right]\cdot(1-(125K+27)\lambda) \leq \mathbb{E}\left[\frac{5}{4}\cdot\frac{\log N}{\lambda} + (125K+27)\lambda L_T^*\right].$$

If $\lambda < \frac{1}{2(125K+30)}$, we can divide both sides of the inequality by $(1-(125K+27)\lambda)$ to obtain

$$\mathbb{E}\left[R_T\right] \leq \mathbb{E}\left[\frac{5}{2}\cdot\frac{\log N}{\lambda} + (250K+54)\lambda L^*\right],$$

where L^* is an upper bound on $\mathbb{E}\left[L_T^*\right]$. Finally, we plug the value $\lambda = \sqrt{\frac{5\log N}{(500K+108)L^*}} \wedge \frac{1}{250K+54}$ to get the regret bound of Theorem 2.

D.2 The proof of Theorem 3

The only difference with the finite parameter space analysis is in the control of the optimistic estimation error. In particular, we only need to adapt our analysis of the optimistic posterior and Lemma 4 to get the regret bound claimed in Theorem 3. We do this with the following lemma.

Lemma 8. *Let $0 < \eta < \frac{1}{2}$, $\lambda > 0$, and $\beta = \frac{1}{1-2\eta}$, assume the hypothesis of Theorem 3 hold. Then, the following inequality holds :*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{2\eta}{\lambda}\cdot \mathrm{IG}_t(\pi_t) + \mathrm{OG}_t\right)\right] \leq \frac{d\log\frac{R}{\epsilon}}{\lambda} + \frac{\lambda\beta T}{8} + \left(\frac{\eta}{\lambda}+1\right)\cdot CT\epsilon. \tag{27}$$

The proof is found in Appendix E.3.4. We can now put this together with the regret decomposition of Equation (21). As in the proof of Theorem 1, we need to pick the hyperparameters such that the optimistic estimation error matches the left hand side of the previous lemma. The same choice of hyperparameters $\eta = \frac{1}{4}, \beta = 2$, and $\gamma = \frac{1}{\mu} = 10\lambda$ combined with Lemma 1 gives us the following bound

$$\mathbb{E}\left[R_T\right] \leq \lambda T(20K + \frac{1}{4} + 5) + \frac{d\log\frac{R}{\epsilon}}{\lambda} + \left(\frac{1}{4\lambda}+1\right)\cdot CT\epsilon. \tag{28}$$

Picking $\epsilon = 1/(CT)$ gives us

$$\mathbb{E}\left[R_T\right] \leq \frac{2d\log RCT}{\lambda} + \lambda T\left(20K + \frac{21}{4}\right) + 1, \tag{29}$$

where we used $\frac{1}{4} \leq d\log RCT$. Finally picking $\lambda = \sqrt{\frac{2d\log(RCT)}{T\left(20K+\frac{21}{4}\right)}}$ recovers the claim of Theorem 3.

D.3 The proof of Theorem 4

One of the appeals of our approach is that with minor tweaking, we can extend the previous guarantees so subgaussian losses. To do that, we consider the following family of likelihoods:

$$p(c|\theta, x, a) \propto \exp\left(-\frac{(c-\ell(\theta,x,a))^2}{2}\right).$$

We also readjust our definition of information gain for this setting by replacing the squared Hellinger distance by the square loss. In particular, the *Gaussian surrogate information gain* is defined as

$$\overline{\mathrm{IG}}_t^{\mathcal{G}}(\pi) = \sum_{a\in\mathcal{A}}\pi(a)\int\left(\ell_t(\theta,a) - \overline{\ell}_t(a)\right)^2\,\mathrm{d}Q_t^+(\theta)$$

and the *(true) Gaussian information gain* as

$$\mathrm{IG}_t^{\mathcal{G}}(\pi) = \sum_{a\in\mathcal{A}}\pi(a)\int\left(\ell_t(\theta,a) - \ell_t(\theta_0,a)\right)^2\,\mathrm{d}Q_t^+(\theta).$$

The surrogate information ratio and averaged DEC are adapted as any policy π

$$\overline{\mathrm{IR}}_t^{\mathcal{G}}(\pi) = \frac{\overline{r}_t(\pi)}{\overline{\mathrm{IG}}_t^{\mathcal{G}}(\pi)} \qquad\text{and}\qquad \overline{\mathrm{DEC}}_{\mu,t}^{\mathcal{G}}(\pi) = \overline{r}_t(\pi) - \mu\cdot\overline{\mathrm{IG}}_t^{\mathcal{G}}(\pi). \tag{30}$$

Then, we define the corresponding algorithm template (called Optimistic Information Directed Sampling for subgaussian losses, **OIDS-SG**) as the method that either picks π_t as the minimizer of $\overline{\mathrm{IR}}_t^{\mathcal{G}}$ or $\overline{\mathrm{DEC}}_T^{\mathcal{G}}$. The two varieties are referred to as **VOIDS-SG** and **ROIDS-SG**.

Replacing the surrogate information gain by its Gaussian counterpart, the regret decomposition of Equation (19) is still valid:

$$\mathbb{E}\left[r_t\right] = \mathbb{E}\left[\overline{\mathrm{DEC}}_t^{\mathcal{G}}(\pi_t,\mu) + \mu\overline{\mathrm{IG}}_t^{\mathcal{G}}(\pi_t) + \mathrm{UE}_t + \mathrm{OG}_t\right].$$

The surrogate and true information gains are related to each other by the following lemma:

Lemma 9. *For any t and policy π, the information gain for Gaussians satisfies $\overline{\mathrm{IG}}_t^{\mathcal{G}}(\pi) \leq 4\mathrm{IG}_t^{\mathcal{G}}(\pi)$.*

See Appendix E.2.2 for the proof. We also relate the underestimation error to the information gain through the following lemma

Lemma 10. *For any t and $\gamma > 0$, the underestimation error is bounded as*

$$|\mathrm{UE}_t| \leq \frac{\gamma}{4} + \frac{\mathrm{IG}_t^{\mathcal{G}}(\pi_t)}{\gamma}.$$

The proof is presented in Appendix E.1.3. Putting these together, we get a regret bound that only depends on the average DEC, the information gain and optimality gap:

$$\mathbb{E}\left[r_t\right] \leq \mathbb{E}\left[\overline{\mathrm{DEC}}_{\mu,t}^{\mathcal{G}}(\pi_t) + \left(4\mu + \frac{1}{\gamma}\right)\mathrm{IG}_t^{\mathcal{G}}(\pi_t) + \mathrm{OG}_t + \frac{\gamma}{4}\right]. \tag{31}$$

We again refer to the sum $\left(4\mu + \frac{1}{\gamma}\right)\mathrm{IG}_t^{\mathcal{G}}(\pi_t)$ as the optimistic estimation error and will control it through an analysis of the optimistic posterior adapted to the sub-Gaussianity of the losses. This is done in the following lemma, whose proof we relegate to Appendix E.3.3.

Lemma 11. *Assume that the losses are v sub-Gaussian and that for all $\theta \in \Theta, x \in X, a \in A$, $\ell(\theta, x, a) \in [0,1]$, then setting $\eta = \frac{1+\sqrt{1-1\wedge v}}{2v}$ the following inequality holds :*

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{16\lambda(v \vee 1)} \cdot \mathrm{IG}_t^{\mathcal{G}}(\pi_t) + \mathrm{OG}_t\right] \leq \frac{\log N}{\lambda} + \frac{\lambda T}{4}. \tag{32}$$

Now we pick $\mu = \frac{1}{\gamma} = \frac{1}{80\lambda(v\vee 1)}$ and apply the previous lemma to obtain the bound

$$\mathbb{E}\left[R_T\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{DEC}}_{\frac{1}{80\lambda(v\vee 1)},t}^{\mathcal{G}}(\pi_t)\right] + \frac{\log N}{\lambda} + \lambda T\left(\frac{1}{4} + 20(v \vee 1)\right). \tag{33}$$

It remains to bound the ADEC. We do this by exhibiting a "forerunner" algorithm that is able to control the *Surrogate Information Ratio*. In particular, we use again the feel-good Thompson sampling (**`FGTS`**) algorithm of Zhang [2022] for this purpose.

Lemma 12. *The surrogate information and averaged decision-to-estimation-coefficient of **`OIDS`** and **`VOIDS`** satisfy the following bound for any $\mu > 0$:*

$$4\mu\overline{\mathrm{DEC}}_{\mu,t}^{\mathcal{G}}(\textbf{\texttt{ROIDS-SG}}) \leq 4\mu\overline{\mathrm{DEC}}_{\mu,t}^{\mathcal{G}}(\textbf{\texttt{VOIDS-SG}}) \leq \overline{\mathrm{IR}}_t^{\mathcal{G}}(\textbf{\texttt{VOIDS-SG}}) \leq \overline{\mathrm{IR}}_t^{\mathcal{G}}(\textbf{\texttt{FGTS}}) = K \tag{34}$$

Putting everything together, we obtain the bound

$$\mathbb{E}\left[R_T\right] \leq \frac{\log N}{\lambda} + \lambda T\left(\frac{1}{4} + 20(v \vee 1)(1+K)\right), \tag{35}$$

from which the bound claimed in Theorem 4 follows by picking the optimal choice of λ.

E Technical proofs

This section presents the more technical parts of the analysis, along with detailed proofs. The content is organized into four main parts: Appendix E.1 presents techniques for bounding the underestimation error, Appendix E.2 provides techniques for relating the surrogate information gain to the true information gain, Appendix E.3 presents the analysis of the optimistic posterior updates to control the optimistic estimation error, and Appendix E.4 provides bounds on the surrogate information ratio and the ADEC. All subsections include a variety of results, stated respectively for the worst-case bounds, first-order bounds, and subgaussian losses.

E.1 Analysis of the Underestimation error

E.1.1 Worst case analysis: The proof of Lemma 3

We define the total variation distance between two distributions P, Q sharing a common dominating measure λ as

$$\mathrm{TV}(P, Q) = \frac{1}{2}\int |p(x) - q(x)|\, d\lambda(x),$$

where p, q are their densities with respect to λ. The total variation distance can be upper bounded by the Hellinger distance as follows:

$$\begin{aligned}
\mathrm{TV}(P, Q) &= \frac{1}{2}\int \left|(\sqrt{p(x)} - \sqrt{q(x)}) \cdot (\sqrt{p(x)} + \sqrt{q(x)})\right| d\lambda(x) \\
&\leq \frac{1}{2}\sqrt{\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\lambda(x) \cdot \int \left(\sqrt{p(x)} + \sqrt{q(x)}\right)^2 d\lambda(x)} \\
&\leq \frac{1}{2}\sqrt{2\mathcal{D}_H^2(P, Q) \cdot 2\int (p(x) + q(x))\, d\lambda(x)} \\
&= \sqrt{2\mathcal{D}_H^2(P, Q)} \\
&\leq \frac{\gamma}{2} + \frac{\mathcal{D}_H^2(P, Q)}{\gamma}.
\end{aligned}$$

Here, the first two inequalities follow from applying Cauchy–Schwarz, and the last one from the inequality of arithmetic and geometric means. Thus, we proceed as

$$\begin{aligned}
|\mathrm{UE}_t| &= \left|\sum_a \pi_t(a) \int \ell_t(\theta_0, a) - \ell_t(\theta, a)\, \mathrm{d}Q_t^+(\theta)\right| \\
&\leq \sum_a \pi_t(a) \int \left|\ell_t(\theta_0, a) - \ell_t(\theta, a)\right| \mathrm{d}Q_t^+(\theta) \\
&= \sum_a \pi_t(a) \int \mathrm{TV}\big(\mathrm{Ber}(\ell_t(\theta_0, a)), \mathrm{Ber}(\ell_t(\theta, a))\big)\, \mathrm{d}Q_t^+(\theta) \\
&\leq \sum_a \pi_t(a) \int \mathrm{TV}\big(p_t(\theta_0, a), p_t(\theta, a)\big)\, \mathrm{d}Q_t^+(\theta) \\
&\leq \frac{\gamma}{2} + \frac{\sum_a \pi_t(a) \int \mathcal{D}_H^2\big(p_t(\theta_0, a), p_t(\theta, a)\big)\, \mathrm{d}Q_t^+(\theta)}{\gamma} \\
&= \frac{\gamma}{2} + \frac{IG_t}{\gamma}.
\end{aligned}$$

The first inequality above uses the boundedness of the losses in $[0, 1]$, the second inequality is the data-processing inequality for the total variation distance (applied on the noisy channel $X \to Y$ that randomly rounds $X \in [0, 1]$ to $Y \in \{0, 1\}$), and the last one is the inequality we have just proved above. This concludes the proof.

E.1.2 Instance-dependent analysis: The proof of Lemma 6

This proof requires a more sophisticated technique based on careful specialized handling of the "underestimated" and "overestimated" actions. The argument is vaguely inspired by the techniques of Bubeck and Sellke [2020] and Foster and Krishnamurthy [2021]. Specifically, for a parameter θ, we define $\mathcal{A}_\theta^- = \{a \in \mathcal{A} : \ell_t(\theta, a) < \ell_t(\theta_0, a)\}$ as the set of actions where $\ell_t(\theta, a)$ underestimates

$\ell_t(\theta_0, a)$. With this notation, we write

$$\begin{aligned}
\mathrm{UE}_t &= \sum_a \pi_t(a)(\ell_t(\theta_0, a) - \bar{\ell}_t(a)) \\
&= \int \sum_a \pi_t(a)\big(\ell_t(\theta_0, a) - \ell_t(\theta, a))\big) \,\mathrm{d}Q_t^+(\theta) \\
&\leq \int \sum_{a \in \mathcal{A}_\theta^-} \pi_t(a)\big(\ell_t(\theta_0, a) - \ell_t(\theta, a)\big) \,\mathrm{d}Q_t^+(\theta) \\
&= \int \sum_{a \in \mathcal{A}_\theta^-} \pi_t(a) \cdot \frac{\sqrt{\gamma(\ell_t(\theta_0, a) + \ell_t(\theta, a))}}{\sqrt{\gamma(\ell_t(\theta_0, a) + \ell_t(\theta, a))}} (\ell_t(\theta_0, a) - \ell_t(\theta, a)) \,\mathrm{d}Q_t^+(\theta),
\end{aligned}$$

where the inequality follows by dropping the negative terms of the sum. Now, the inequality of arithmetic and geometric means implies that for any $x, y \geq 0$, $xy \leq \frac{x^2+y^2}{2}$. We apply it to $x = 2\sqrt{\gamma(\ell_t(\theta_0, a) + \ell_t(\theta, a))}$ and $y = \frac{(\ell_t(\theta_0,a) - \ell_t(\theta,a))}{2\sqrt{\gamma(\ell_t(\theta_0,a)+\ell_t(\theta,a))}}$ to obtain

$$\mathrm{UE}_t \leq \int \left(\gamma \sum_{a \in \mathcal{A}_\theta^-} \pi_t(a) \cdot \big(\ell_t(\theta_0, a) + \ell_t(\theta, a)\big) + \frac{1}{4\gamma} \sum_{a \in \mathcal{A}_\theta^-} \pi_t(a) \frac{\big(\ell_t(\theta_0, a) - \ell_t(\theta, a)\big)^2}{\ell_t(\theta_0, a) + \ell_t(\theta, a)} \right) \mathrm{d}Q_t^+(\theta).$$

To proceed, we use the inequality $\frac{(\ell_t(\theta_0,a) - \ell_t(a))^2}{\ell_t(\theta_0,a)+\ell_t(\theta,a)} \leq 4\mathcal{D}_H^2(p_t(\theta, a), p_t(\theta_0, a))$ that holds for all a and θ, and is proved separately as Lemma 23. Hence,

$$\begin{aligned}
\mathrm{UE}_t &\leq 2\gamma \sum_a \pi_t(a)\ell_t(\theta_0, a) + \frac{1}{\gamma} \int \sum_a \mathcal{D}_H^2(p_t(\theta, a), p_t(\theta_0, a)) \,dQ_t^+(\theta) \\
&\leq 2\gamma \sum_a \pi_t(a)\ell_t(\theta_0, a) + \frac{\mathrm{IG}_t}{\gamma},
\end{aligned}$$

which concludes the proof.

E.1.3 Subgaussian analysis: The proof of Lemma 10

The claim follows from the following calculations:

$$\begin{aligned}
|\mathrm{UE}_t| &= \left| \sum_a \pi_t(a) \int \ell(\theta_0, a) - \bar{\ell}_t(a) \,dQ_t^+(\theta) \right| \\
&\leq \sum_a \pi_t(a) \int \left|\ell(\theta_0, a) - \bar{\ell}_t(a)\right| \,dQ_t^+(\theta) \\
&\leq \sqrt{\sum_a \pi_t(a) \int \big(\ell(\theta_0, a) - \bar{\ell}_t(a)\big)^2 \,dQ_t^+(\theta)} \\
&= \sqrt{\mathrm{IG}_t^{\mathcal{G}}(\pi_t)} \\
&\leq \frac{\gamma}{4} + \frac{\mathrm{IG}_t^{\mathcal{G}}(\pi_t)}{\gamma}.
\end{aligned}$$

Here, the second inequality is Cauchy–Schwarz and the last one is the inequality of arithmetic and geometric means.

E.2 Analysis of the Surrogate Information Gain and the True Information Gain

E.2.1 Bounded losses: The proof of Lemma 2

The claim is proved as

$$\begin{aligned}
\overline{\mathrm{IG}}_t(\pi) &= \sum_a \pi(a) \int \mathcal{D}_H^2\left(\overline{\ell}_t(a), \ell_t(\theta, a)\right) \, \mathrm{d}Q_t^+(\theta) \\
&\leq 2 \cdot \sum_a \pi(a) \int \mathcal{D}_H^2\left(\overline{\ell}_t(a), \ell_t(\theta_0, a)\right) \, \mathrm{d}Q_t^+(\theta) \\
&\qquad\qquad + 2 \cdot \sum_a \pi(a) \int \mathcal{D}_H^2\left(\ell_t(\theta_0, a), \ell_t(\theta, a)\right) \, \mathrm{d}Q_t^+(\theta) \\
&\leq 4 \cdot \sum_a \pi(a) \int \mathcal{D}_H^2\big(\ell_t(\theta_0, a), \ell_t(\theta, a)\big) \, \mathrm{d}Q_t^+(\theta) \\
&= 4\mathrm{IG}_t(\pi),
\end{aligned}$$

where the first inequality critically uses that the Hellinger distance is a metric and as such it satisfies the triangle inequality, and thus $\mathcal{D}_H^2(P, P') \leq 2\mathcal{D}_H^2(P, Q) + 2\mathcal{D}_H^2(Q, P')$ holds for any P, P' and Q by an additional application of Cauchy–Schwarz. The final inequality then uses the convexity of the Hellinger distance and Jensen's inequality.

E.2.2 Subgaussian losses: The proof of Lemma 9

The claims follows from writing

$$\begin{aligned}
\overline{\mathrm{IG}}_t^{\mathcal{G}}(\pi) &= \sum_a \pi(a) \int \left(\overline{\ell}_t(a) - \ell(\theta, a)\right)^2 \, \mathrm{d}Q_t^+(\theta) \\
&\leq 2 \cdot \sum_a \pi(a) \int \left(\overline{\ell}_t(a) - \ell(\theta_0, a)\right)^2 \, \mathrm{d}Q_t^+(\theta) \\
&\qquad\qquad + 2 \cdot \sum_a \pi(a) \int \left(\ell(\theta_0, a) - \ell(\theta, a)\right)^2 \, \mathrm{d}Q_t^+(\theta) \\
&\leq 4 \cdot \sum_a \pi(a) \int \left(\ell(\theta_0, a) - \ell(\theta, a)\right)^2 \, \mathrm{d}Q_t^+(\theta) \\
&= 4\mathrm{IG}_t^{\mathcal{G}}(\pi),
\end{aligned}$$

where the first inequality comes an application of the triangle inequality and Cauchy–Schwarz, and the second one comes from the convexity of the squared loss and Jensen's inequality.

E.3 Analysis of the Optimistic Posterior

We start by providing a general statement about the properties of the optimistic posterior updates, which will then prove useful for bounding the optimistic estimation error.

Lemma 13. *Consider the optimistic posterior defined recursively by*

$$\frac{\mathrm{d}Q_{t+1}^+}{\mathrm{d}Q_t^+}(\theta) = \frac{\exp\left(-\eta \log\left(\frac{1}{p_t(L_t|\theta, A_t)}\right) - \lambda \ell_t^*(\theta)\right)}{\int \exp\left(-\eta \log\left(\frac{1}{p_t(L_t|\theta', A_t)}\right) - \lambda \ell_t^*(\theta')\right) \, \mathrm{d}Q_t^+(\theta')}, \tag{36}$$

where $Q_1^+ = Q_1$ is some prior distribution on Θ and $p_t(\cdot|\theta, a) \in \Delta_{\mathbb{R}^+}$ is the density the loss distribution associated with parameter θ. For any $T > 0$, for any $\alpha, \beta > 0$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, for any distribution $Q^* \in \Delta(\Theta)$, and for any sequence of actions $A_1, \ldots, A_T$ and losses $L_1, \ldots, L_T$, *the following inequality holds:*

$$-\frac{1}{\lambda\alpha}\sum_{t=1}^{T}\log\int p_t(\theta,A_t,L_t)^{\eta\alpha}\mathrm{d}Q_t^+(\theta)-\frac{1}{\lambda\beta}\sum_{t=1}^{T}\log\int\exp\left(-\lambda\beta\ell_t^*(\theta)\right)\mathrm{d}Q_t^+(\theta)$$
$$\leq\int\left(\frac{1}{\lambda\alpha}\cdot\sum_{t=1}^{T}\log\frac{1}{p_t(\theta,A_t,L_t)^{\eta\alpha}}+\sum_{t=1}^{T}\ell_t^*(\theta)\right)\mathrm{d}Q^*(\theta)+\frac{1}{\lambda}\cdot\mathcal{D}_{KL}\left(Q^*\|Q_1\right). \tag{37}$$

Proof. We study the potential function Φ defined for all $c\in\mathbb{R}^\Theta$ as

$$\Phi(c)=\frac{1}{\lambda}\log\int_\Theta\exp(-\lambda c(\theta))\mathrm{d}Q_1(\theta).$$

We define $c_t(\theta)=\frac{\eta}{\lambda}\log\frac{1}{p_t(\theta,A_t,L_t)}+\ell_t^*(\theta)$ and evaluate $\Phi\left(\sum_{t=1}^T c_t\right)$:

$$\Phi\left(\sum_{t=1}^{T}c_t\right)=\frac{1}{\lambda}\log\int_\Theta\exp\left(-\lambda\sum_{t=1}^{T}c_t(\theta)\right)\mathrm{d}Q_1(\theta)\geq-\int_\Theta\sum_{t=1}^{T}c_t(\theta)\mathrm{d}Q^*(\theta)-\frac{\mathcal{D}_{\mathrm{KL}}\left(Q^*\|Q_1\right)}{\lambda}.$$

where the inequality is the Donsker-Varadhan variational formula [cf. Section 4.9 in Boucheron et al., 2013]. We also have

$$\begin{aligned}\Phi\left(\sum_{t=1}^{T}c_t\right)&=\sum_{t=1}^{T}\left(\Phi\left(\sum_{k=1}^{t}c_k\right)-\Phi\left(\sum_{k=1}^{t-1}c_k\right)\right)\\&=\sum_{t=1}^{T}\frac{1}{\lambda}\log\frac{\int_\Theta\exp\left(-\lambda\sum_{k=1}^{t}c_k(\theta)\right)\mathrm{d}Q_1(\theta)}{\int_\Theta\exp\left(-\lambda\sum_{k=1}^{t-1}c_k(\theta)\right)\mathrm{d}Q_1(\theta)}\\&=\sum_{t=1}^{T}\frac{1}{\lambda}\log\int_\Theta\frac{\exp\left(-\lambda\sum_{k=1}^{t-1}c_k(\theta)\right)}{\int_\Theta\exp\left(-\lambda\sum_{k=1}^{t-1}c_k(\theta)\right)\mathrm{d}Q_1(\theta)}\cdot\exp\left(-\lambda c_t(\theta)\right)\mathrm{d}Q_1(\theta)\\&=\sum_{t=1}^{T}\frac{1}{\lambda}\log\int_\Theta\exp\left(-\lambda c_t(\theta)\right)\mathrm{d}Q_t^+(\theta)\\&=\sum_{t=1}^{T}\frac{1}{\lambda}\log\int_\Theta p_t(\theta,A_t,L_t)^\eta\cdot\exp\left(-\lambda\ell_t^*(\theta)\right)\mathrm{d}Q_t^+(\theta),\end{aligned}$$

where the fourth equality is by definition of Q_t^+ and c_t.

We can now apply Hölder's inequality with $\alpha,\beta>0$ such that $\frac{1}{\alpha}+\frac{1}{\beta}=1$, obtaining

$$\Phi\left(\sum_{t=1}^{T}c_t\right)\leq\frac{1}{\lambda}\cdot\sum_{t=1}^{T}\left(\frac{1}{\alpha}\log\int_\Theta p_t(\theta,A_t,L_t)^{\eta\alpha}\mathrm{d}Q_t^+(\theta)+\frac{1}{\beta}\log\int_\Theta\exp\left(-\lambda\beta\ell_t^*(\theta)\right)\mathrm{d}Q_t^+(\theta)\right).$$

Plugging both bounds together, we get the claim of the lemma. □

The following statement will be useful for turning the above guarantee into a bound on the information gain:

Lemma 14. *For any $t\geq1$ and any policy $\pi_t\in\Delta(\mathcal{A})$, the following inequality holds:*

$$\mathbb{E}\left[\mathrm{IG}_t(\pi_t)\right]\leq\mathbb{E}\left[-\log\int_\Theta\left(\frac{p_t(\theta,A_t,L_t)}{p_t(\theta_0,A_t,L_t)}\right)^{\frac{1}{2}}\mathrm{d}Q_t^+(\theta)\right]. \tag{38}$$

Proof. Let τ be the dominating measure used to define the densities $p_t(\cdot|\theta,a)$. We write:

$$\mathbb{E}\left[\mathrm{IG}_t(\pi_t)\right]=\mathbb{E}\left[\int_\Theta\sum_a\pi_t(a)\mathcal{D}_H^2\left(p_t(\theta_0,a),p_t(\theta,a)\right)\,\mathrm{d}Q_t^+(\theta)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\int_{\Theta} \sum_a \pi_t(a) \left(1 - \int_{\mathbb{R}} \left(p_t(c|\theta, A_t) p_t(c|\theta_0, A_t)\right)^{\frac{1}{2}} \, \mathrm{d}\tau(c)\right) \mathrm{d}Q_t^+(\theta)\right] \\
&= \mathbb{E}\left[\int_{\Theta} \mathbb{E}_t\left[\int_{\mathbb{R}} \left(1 - \left(\frac{p_t(c|\theta, A_t)}{p_t(c|\theta_0, A_t)}\right)^{\frac{1}{2}}\right) p_t(c|\theta_0, A_t) \mathrm{d}\tau(c)\right] \mathrm{d}Q_t^+(\theta)\right] \\
&= \mathbb{E}\left[\int_{\Theta} \mathbb{E}_t\left[\int_{\mathbb{R}} \left(1 - \left(\frac{p_t(L_t|\theta, A_t)}{p_t(L_t|\theta_0, A_t)}\right)^{\frac{1}{2}}\right) p_t(L_t|\theta_0, A_t)\right] \mathrm{d}Q_t^+(\theta)\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_t\left[-\log \int \left(\frac{p_t(L_t|\theta, A_t)}{p_t(L_t|\theta_0, A_t)}\right)^{\frac{1}{2}} \mathrm{d}Q_t^+(\theta)\right]\right] \\
&= \mathbb{E}\left[-\log \int_{\Theta} \left(\frac{p_t(\theta, A_t, L_t)}{p_t(\theta_0, A_t, L_t)}\right)^{\frac{1}{2}} \mathrm{d}Q_t^+(\theta)\right].
\end{aligned}
$$

Here, we used the tower rule of expectation several times, and also the elementary inequality $\log(x) \leq x - 1$ that holds for all x. This concludes the proof. □

E.3.1 Worst case analysis: The proof of Lemma 4

Lemma 15. *For any $t \geq 1$, $\beta, \lambda > 0$, as long as $\ell_t^*(\theta) \in [0, 1]$ for all values of θ, the following inequality holds*

$$
\frac{1}{\lambda\beta} \log \int_{\Theta} \exp\left(-\lambda\beta\ell_t^*(\theta)\right) \mathrm{d}Q_t^+(\theta) \leq -\bar{\ell}_t^* + \frac{\lambda\beta}{8}. \tag{39}
$$

Proof. This is a direct consequence of Hoeffding's lemma for bounded random variables, see for example Section 2.3 of Boucheron et al. [2013]. □

The proof of Lemma 4 then follows directly by applying Lemma 13 with η, α such that $\eta\alpha = \frac{1}{2}$ (which means $\beta = 1/(1 - 2\eta)$) and with Q^* a dirac distribution in θ_0, and combining the result with Lemmas 14 and 15 above.

E.3.2 Instance dependent analysis and proof of Lemma 7

Lemma 16. *For any $t \geq 1$, $\beta, \lambda > 0$, as long as $\ell_t^*(\theta) \in [0, 1]$ for all values of θ, the following inequality holds*

$$
\frac{1}{\lambda\beta} \log \int_{\Theta} \exp\big((-\lambda\beta\ell_t^*(\theta))\big) \mathrm{d}Q_t^+(\theta) \leq -\bar{\ell}_t^* \left(1 - \frac{\lambda\beta}{2}\right). \tag{40}
$$

Proof. We use the two elementary inequalities $\log(x) \leq x - 1$ that holds for all $x \in \mathbb{R}$ and $e^{-x} \leq 1 - x + \frac{x^2}{2}$ that holds for all $x \geq 0$ to show

$$
\begin{aligned}
\frac{1}{\lambda\beta} \log \int_{\Theta} \exp\big(-\lambda\beta\ell_t^*(\theta)\big) \mathrm{d}Q_t^+(\theta) &\leq \frac{1}{\lambda\beta} \left(\int_{\Theta} 1 - \lambda\beta\ell_t^*(\theta) + \left(\frac{\lambda\beta}{2}\ell_t^*(\theta)\right)^2 \mathrm{d}Q_t^+(\theta) - 1\right) \\
&\leq \frac{1}{\lambda\beta} \left(\int_{\Theta} -\lambda\beta\ell_t^*(\theta) + \left(\frac{\lambda\beta}{2}\right)^2 \ell_t^*(\theta) \mathrm{d}Q_t^+(\theta)\right) \\
&= -\bar{\ell}_t^* \left(1 - \frac{\lambda\beta}{2}\right),
\end{aligned}
$$

where we used the fact that for all $\theta \in \Theta$, we have $\ell_t^*(\theta) \in [0, 1]$ and thus $\ell_t^*(\theta)^2 \leq \ell_t^*(\theta)$. □

We use again Lemma 13 with η, α such that $\eta\alpha = 1/2$ and with Q^* a dirac distribution in θ_0. Then we apply Lemma 16 and Lemma 14 to conclude the proof of Lemma 7.

E.3.3 Subgaussian analysis: The proof of Lemma 11

Lemma 17. *Assume that the losses are v sub-Gaussian and that for all $\theta \in \Theta, x \in \mathcal{X}, a \in \mathcal{A}, \ell(\theta, x, a) \in [0,1]$. For any $t \geq 1, \eta, \alpha \geq 0$ such that $\delta = \frac{\eta\alpha}{2}\left(1 - \frac{\eta\alpha v}{2}\right) \geq 0$ and any policy $\pi_t \in \Delta(\mathcal{A})$, the following inequality holds*

$$\delta(1-2\delta) \cdot \mathbb{E}\left[\mathrm{IG}_t^{\mathcal{G}}(\pi_t)\right] \leq \mathbb{E}\left[-\log \int_\Theta \left(\frac{p_t(\theta, A_t, L_t)}{p_t(\theta_0, A_t, L_t)}\right)^{\eta\alpha} \mathrm{d}Q_t^+(\theta)\right]. \tag{41}$$

Proof. We remind the reader than $\mathcal{F}_t = \theta(X_1, A_1, L_1, \ldots, X_{t-1}, A_{t-1}, L_{t-1})$ is the σ-algebra generated by the interaction history between the learner and the environment up to the end of round t. By the tower rule of expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[-\log \int_\Theta \left(\frac{p_t(\theta, A_t, L_t)}{p_t(\theta_0, A_t, L_t)}\right)^{\eta\alpha} \mathrm{d}Q_t^+(\theta)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[-\log \int_\Theta \left(\frac{p_t(\theta, A_t, L_t)}{p_t(\theta_0, A_t, L_t)}\right)^{\eta\alpha} \mathrm{d}Q_t^+(\theta) \middle| \mathcal{F}_{t-1}, X_t, A_t\right]\right] \\
&\leq \mathbb{E}\left[-\log \mathbb{E}\left[\int_\Theta \left(\frac{p_t(\theta, A_t, L_t)}{p_t(\theta_0, A_t, L_t)}\right)^{\eta\alpha} \mathrm{d}Q_t^+(\theta) \middle| \mathcal{F}_{t-1}, X_t, A_t\right]\right] \\
&= \mathbb{E}\left[-\log \int_\Theta \int_{\mathbb{R}} \left(\frac{p_t(\theta, A_t, L)}{p_t(\theta_0, A_t, L)}\right)^{\eta\alpha} d\mathcal{P}_{L_t|X_t,A_t}(L) \mathrm{d}Q_t^+(\theta)\right].
\end{aligned} \tag{42}$$

Where the first inequality comes from Jensen's Inequality applied to the logarithm and $\mathcal{P}_{L_t|X_t,A_t}$ is the conditional law of L_t given X_t and A_t. We fix $\theta \in \Theta$, drop the subscripts for simplicity and define $\ell = \ell_t(A_t)$, $\ell_0 = \ell_t(\theta_0, A_t)$ and $\mathcal{P}_t = \mathcal{P}_{L_t|X_t,A_t}$. Using the definition of the likelihood p_t, we get

$$\begin{aligned}
&\int \left(\frac{p_t(\theta, A_t, L)}{p_t(\theta_0, A_t, L)}\right)^{\eta\alpha} d\mathcal{P}_t(L) \\
&= \int \exp\left(-\eta\alpha\left(\frac{(L - \ell_t(\theta, A_t))^2}{2} + \frac{(L - \ell(\theta_0, A_t))^2}{2}\right)\right) d\mathcal{P}_t(L) \\
&= \int \exp\left(\frac{\eta\alpha}{2}(2L - \ell - \ell_0) \cdot (\ell - \ell_0)\right) d\mathcal{P}_t(L) \\
&= \exp\left(-\frac{\eta\alpha}{2}(\ell + \ell_0) \cdot (\ell - \ell_0)\right) \cdot \int \exp\left(\eta\alpha L(\ell - \ell_0)\right) d\mathcal{P}_t(L) \\
&= \exp\left(\frac{\eta\alpha}{2}(\ell_0^{*2} - \ell^2)\right) \cdot \int \exp\left(\eta\alpha L(\ell - \ell_0)\right) d\mathcal{P}_t(L) \\
&\leq \exp\left(\frac{\eta\alpha}{2}(\ell_0^2 - \ell^2)\right) \cdot \exp\left(\eta\alpha\ell_0 \cdot (\ell - \ell_0)\right) \exp\left(\frac{\eta^2\alpha^2 v}{2}(\ell - \ell_0)^2\right) \\
&\leq \exp\left(-(\ell - \ell_0)^2 \cdot \frac{\eta\alpha}{2}\left(1 - \frac{\eta\alpha v}{2}\right)\right).
\end{aligned}$$

Now we define $\delta = \frac{\eta\alpha}{2}(1 - \frac{\eta\alpha v}{2})$ we have :

$$\begin{aligned}
&\int \left(\frac{p_t(\theta, A_t, L)}{p_t(\theta_0, A_t, L)}\right)^{\eta\alpha} d\mathcal{P}_t(L) \\
&\leq \exp\left(-(\ell - \ell_0)^2 \cdot \delta\right) \\
&\leq 1 - \delta(\ell - \ell_0)^2 + \frac{\delta^2}{2}(\ell - \ell_0)^4 \\
&\leq 1 - \delta(\ell - \ell_0)^2 + 4\delta^2(\ell - \ell_0)^2 \\
&\leq 1 - \delta(1 - 2\delta)(\ell - \ell_0)^2.
\end{aligned}$$

Where we use that $|\ell - \ell_0| \leq 2$. Finally using that for any $x > 0$, $\log x \leq x - 1$ and equation 42, we get the claim of the Lemma. $\square$

It remains to pick the best values for η, α and β and apply Lemma 13 with Q^* a dirac distribution in θ_0 and Lemma 15. To finish the proof of Lemma 17, we combine the previous Lemma (17) with Lemma 15 and Lemma 13. We want the quantity $\delta(1-2\delta)$ to be as big as possible, this happens when $\delta = \frac{1}{4}$. This is only possible if $v \leq 1$ and $\frac{\eta\alpha}{2} = \frac{1+\sqrt{1-v}}{2v}$. If $v > 1$, our best choice of $\frac{\eta\alpha}{2}$ is $\frac{1}{2v}$ and in that case $\delta(1-2\delta) = \frac{1}{4v}\left(1 - \frac{1}{2v}\right) \geq \frac{1}{8v}$. Finally, uniting both cases, we set $\alpha = \beta = 2$, $\eta = \frac{1+\sqrt{1-v\wedge 1}}{2v}$ and we have that $\delta(1-2\delta) \geq \frac{1}{8(1\vee v)}$.

E.3.4 Metric Parameter Analysis : the proof of Lemma 8

We start by a technical lemma on the Lipschtzness of the losses and the optimal losses.

Lemma 18. *For any x, θ, a, $\ell_t(\cdot, x, a)$ and $\ell_t^*(\cdot, x)$ are C-Lipschitz.*

Proof. Let τ be the measure against which the densities $p(\cdot|\theta, x, a)$ are defined. Without loss of generality, we can assume that $\int_{[0,1]} \mathrm{d}\tau(c) = 1$. Letting $\theta_1, \theta_2 \in \Theta$, we have

$$\begin{aligned}
|\ell(\theta_1, x, a) - \ell(\theta_2, x, a)| &= \left| \int_{[0,1]} c(p(c|\theta_1, x, a) - p(c|\theta_2, x, a))\mathrm{d}\tau(c) \right| \\
&\leq \int_{[0,1]} |(p(c|\theta_1, x, a) - p(c|\theta_2, x, a))| \,\mathrm{d}\tau(c) \\
&= \int_{[0,1]} |\exp\left(\log(p(c|\theta_1, x, a))\right) - \exp\left(\log(p(c|\theta_2, x, a))\right)| \,\mathrm{d}\tau(c) \\
&\leq \int_{[0,1]} C \|\theta_1 - \theta_2\| \,\mathrm{d}\tau(c) \\
&= C \|\theta_1 - \theta_2\|,
\end{aligned}$$

where the second inequality comes from the C-Lipschtzness of the composition of the exponential that is 1-Lipschitz on the negative numbers and the log likelihood that is C-Lipschitz. This proves the C-Lipschtzness of $\ell_t(\cdot, x, a)$. Now it easily follows that $\ell^*(\cdot, x)$ is also C-Lipschitz, being an infimum of a family of C-Lipschitz functions. □

Now we introduce two further lemmas related to Lemma 13 when Q^* is chosen as a uniform distribution on a ball of radius ϵ.

Lemma 19. *Fix $\theta_0 \in \Theta$, and $\epsilon > 0$, and assume that a ball including θ_0 with radius ϵ is contained in Θ. Letting Q^* be the uniform distribution on such a ball, we have*

$$\mathcal{D}_{KL}\left(Q^* \| Q_1\right) = d \log\left(\frac{R}{\epsilon}\right). \tag{43}$$

Proof. Since both Q^* and Q_1 are uniform, the ratio of their density is equal to the ratio of the volume of Θ and the volume of a ball of radius ϵ. Since Θ is included in a ball of radius R, this ratio is bounded by $(\frac{R}{\epsilon})^d$. Finally

$$\mathcal{D}_{\mathrm{KL}}\left(Q^* \| Q_1\right) = \int_\Theta \frac{\mathrm{d}Q^*}{\mathrm{d}Q_1}(\theta) \log\left(\frac{\mathrm{d}Q^*}{\mathrm{d}Q_1}(\theta)\right) \mathrm{d}Q_1(\theta) \leq \log\left(\frac{R}{\epsilon}\right)^d \int_\Theta \mathrm{d}Q^*(\theta) = d\log\left(\frac{R}{\epsilon}\right).$$

□

Lemma 20. *Under the same conditions as Lemma 19, we have*

$$\left| \int \left(\frac{1}{\lambda\alpha} \cdot \sum_{t=1}^T \log \frac{p_t(\theta_0, A_t, L_t)^{\eta\alpha}}{p_t(\theta, A_t, L_t)^{\eta\alpha}} + \sum_{t=1}^T \left(\ell_t^*(\theta) - \ell_t^*(\theta_0)\right) \right) \mathrm{d}Q^*(\theta) \right| \leq \left(\frac{\eta}{\lambda} + 1\right) \cdot CT\epsilon. \tag{44}$$

Proof. This is a direct consequence of the Lipschitzness of the log-likelihood and Lemma 18. □

Putting Lemma 13 together with this choice of Q^* and with Lemma 14 and Lemma 15, we finish the proof of Lemma 8

E.4 Upper bounds on the averaged DEC and the Surrogate Information ratio

Here we provide the technical tools to bound the surrogate information ratio and the averaged DEC for some appropriately chosen forerunner algorithms.

E.4.1 Worst-case analysis: The proof of Lemmas 1 and 12

Here we study the performance of Thompson sampling as the forerunner algorithm, which will certify a bound on the surrogate information ratio of **OIDS**. The Thompson sampling policy π_t works by sampling θ_t according to the posterior Q_t^+ and then playing the action $A_t \in \arg\min_a \ell_t(\theta_t, a)$. To facilitate the derivations below, we define $a_t^* : \Theta \to \mathcal{A}$ the greedy action selector by $a_t^*(\theta) = \arg\min_a \ell_t(\theta, a)$ (with ties broken arbitrarily). By definition of the policy, sampling according to π_t is the same as sampling according to $\mathrm{d}Q_t^+$ and then applying the greedy action selector. More formally, for any measurable function f, we have

$$\int_\Theta f(a_t^*(\theta))\mathrm{d}Q_t^+(\theta) = \sum_a \pi_t(a) f(a).$$

Moreover, we have that $\bar{\ell}_t^* = \int_\Theta \ell_t^*(\theta)\mathrm{d}Q_t^+(\theta) = \int_\Theta \ell_t(\theta, a_t^*(\theta))\mathrm{d}Q_t^+(\theta)$. Putting these observations together, we can write the surrogate regret as

$$\bar{r}_t(\pi_t) = \sum_a \pi_t(a)\big(\bar{\ell}_t(a) - \bar{\ell}_t^*\big) = \int_\Theta \bar{\ell}_t(a_t^*(\theta)) - \ell_t(\theta, a_t^*(\theta)). \tag{45}$$

Observe that the regret is the difference of the expectation of the same function under the joint distribution of θ_t and A_t and their product distribution, and thus measures the extent to which the two are "coupled". We will analyze this quantity by a decoupling argument inspired by Zhang [2022] and Neu et al. [2022].

For setting up the decoupling analysis, we first need some technical lemmas. We start by a corollary of the Fenchel–Young inequality for strongly convex functions that will come handy.

Lemma 21. *Let I be an interval on the real line and let $\mathcal{D} : I^2 \to \mathbb{R}$ be a convex function satisfying the following conditions:*

- *For any $y \in I$, the function $x \to \mathcal{D}(x, y)$ is proper, closed and C-strongly convex.*
- *For any $x \in I$, $\mathcal{D}(x, x) = 0$.*

Then for any $x, y \in I$ and any $\mu \in \mathbb{R}$ we have

$$(x - y)u \le \mathcal{D}(x, y) + \frac{u^2}{2C}. \tag{46}$$

Proof. Let $y \in I$. We compute the Legendre–Fenchel conjugate of $x \to \mathcal{D}(x, y)$, defined for any $u \in R$ as

$$\mathcal{D}^*(u, y) = \sup_{x \in I} \{xu - f(y)\}.$$

Since y is a minimum of $x \to \mathcal{D}(x, y)$ and $\mathcal{D}(y, y) = 0$, we have that $\mathcal{D}^*(0, y) = 0$. Moreover using Lemma 15 of Shalev-Shwartz [2007], we directly have that $\mathcal{D}^*$ is $\frac{1}{C}$ smooth in its first coordinate and that $\frac{\partial \mathcal{D}^*}{\partial u}(0, y) = y$, so that for any $u \in \mathbb{R}$ we have

$$\mathcal{D}^*(u, y) \le \mathcal{D}^*(0, y) + u\frac{\partial \mathcal{D}^*}{\partial u}(0, y) + \frac{u^2}{2C} \le yu + \frac{u^2}{2C}.$$

Then, by the Fenchel–Young inequality, this implies the following for any $x \in I$ and any $u \in \mathbb{R}$:

$$x \cdot \mu \le \mathcal{D}(x, y) + \mathcal{D}^*(u, y) \le y \cdot u + \frac{u^2}{2C}.$$

This proves the statement. □

We use this inequality to prove the following general decoupling lemma that can handle arbitrary joint distributions of random variables.

Lemma 22. *Let $\mathcal{D} : [0,1]^2 \to \mathbb{R}$ be C-strongly convex and satisfy the same hypothesis as for the previous lemma. Let $Q \in \Delta(\Theta)$, $f : \Theta \times \mathcal{A} \to [0,1]$ and $a^* : \Theta \to \mathcal{A}$. Assume f and a^* are measurable. We define $\pi \in \Delta(\mathcal{A})$ by $\pi(a) = \int_\Theta \mathbb{I}_{\{a^*(\theta)=a\}} \mathrm{d}Q(\theta)$ and $\bar{f}(a) = \int_\Theta f(\theta, a)\mathrm{d}Q(\theta)$. Then for any $\mu > 0$ the following holds*

$$\int_\Theta \bar{f}(a^*(\theta)) - f(\theta, a^*(\theta))\mathrm{d}Q(\theta) \le \mu \int_\Theta \sum_a \pi(a)\mathcal{D}(\bar{f}(a), f(\theta, a))\mathrm{d}Q(\theta) + \frac{K}{2\mu C} \tag{47}$$

Proof. We start by writing

$$\begin{aligned}\int_\Theta \bar{f}(a^*(\theta)) - f(\theta, a^*(\theta)) &= \int_\Theta \sum_a \frac{\mu\pi(a)}{\mu\pi(a)} \mathbb{I}_{\{a^*(\theta)=a\}} \left(\bar{f}(a) - f(\theta, a)\right) \mathrm{d}Q(\theta) \\ &= \int_\Theta \sum_a \mu\pi(a) \left(\frac{\mathbb{I}_{\{a^*(\theta)=a\}}}{\mu\pi(a)} \left(\bar{f}(a) - f(\theta, a)\right)\right) \mathrm{d}Q(\theta) \\ &\le \int_\Theta \sum_a \mu\pi(a) \left(\mathcal{D}(\bar{f}(a), f(\theta, a)) + \frac{\mathbb{I}_{\{a^*(\theta)=a\}}}{2C\mu^2\pi(a)^2}\right) \mathrm{d}Q(\theta),\end{aligned}$$

where we used Lemma 21 with $u = \frac{\mathbb{I}_{\{a^*(\theta)=a\}}}{\mu\pi(a)}$ in the last line. Finally, we have

$$\begin{aligned}\int_\Theta \bar{f}(a^*(\theta)) - f(\theta, a^*(\theta)) &\le \mu \int_\Theta \sum_a \pi(a)\mathcal{D}(\bar{f}(a), f(\theta, a))\mathrm{d}Q(\theta) + \frac{1}{2\mu C} \sum_a \int_\Theta \frac{\mathbb{I}_{a^*(\theta)=a}}{\pi(a)} \mathrm{d}Q(\theta) \\ &\le \mu \int_\Theta \sum_a \pi(a)\mathcal{D}(\bar{f}(a), f(\theta, a))\mathrm{d}Q(\theta) + \frac{K}{2\mu C},\end{aligned}$$

where we used $\pi(a) = \int_\Theta \mathbb{I}_{a^*(\theta)=a}\mathrm{d}Q(\theta)$ in the last line. □

To prove Lemma 1, we use the above result with $Q = Q_t^+$, $f = \ell_t$ and $a^* = aj_t$ and $\mathcal{D}$ chosen as the squared Hellinger distance $\mathcal{D}_H^2$, which is $\frac{1}{4}$-strongly convex in its first argument by Lemma 24 provided in Appendix E.5. Thus, applying Lemma 22 we get for any $\mu > 0$ that

$$\bar{r}_t(\pi_t) \le \mu \int_\Theta \sum_a \pi_t(a)\mathcal{D}_H^2(\bar{\ell}_t(a), \ell_t(\theta, a))\mathrm{d}Q_t^+(\theta) + \frac{2K}{\mu}.$$

This concludes the proof of Lemma 1.

Lemma 12 is proved by choosing $\mathcal{D}(x, y) = (x - y)^2$ that is 2-strongly convex in its first argument, which yields the advertised result as

$$\bar{r}_t(\pi_t) \le \mu \int_\Theta \sum_a \pi_t(a)(\bar{\ell}_t(a) - \ell_t(\theta, a))^2\mathrm{d}Q_t^+(\theta) + \frac{K}{4\mu}.$$

E.4.2 Instance-dependent analysis: The proof of Lemma 5

This analysis uses the so-called *inverse-gap weighting* algorithm of Abe and Long [1999] as forerunner—see also the works of Foster and Rakhlin [2020] and Foster and Krishnamurthy [2021] that reignited interest in this method. Our analysis below is especially inspired by the latter work.

We define the inverse gap weighting policy with scale parameter γ and with respect to a nominal loss function $f : \mathcal{A} \to \mathbb{R}^+$ as

$$\pi^{(\mathrm{IGW})}_{\gamma,f}(a) = \begin{cases} \frac{f(b)}{Kf(b)+\gamma(f(b)-f(a))} & \text{if } a \ne b \\ 1 - \sum_{a \ne b} \pi^{(\mathrm{IGW})}_{\gamma,f}(a) & \text{if } a = b \end{cases}$$

where $b \in \arg\min_a f(a)$ is fixed (with ties broken arbitrarily). We fix θ and apply Lemma 4 of Foster and Krishnamurthy [2021] with nominal loss $\bar{\ell}_t : A \to \mathbb{R}$ and true loss $\ell_t(\theta) : A \to \mathbb{R}$ to get

$$\bar{\ell}_t(b) - \ell_t(\theta, a_t^*(\theta)) \le \frac{K}{4\gamma}\bar{\ell}_t(b) + 2\gamma \cdot \pi^{(\mathrm{IGW})}_{\gamma,\bar{\ell}_t}(a^*(\theta)) \frac{(\bar{\ell}_t(a^*(\theta)) - \ell_t(\theta, a^*(\theta)))^2}{\bar{\ell}_t(a_t^*(\theta)) + \ell_t(\theta, a_t^*(\theta))}$$

$$\leq \frac{K}{4\gamma}\bar{\ell}_t(b) + 2\gamma \cdot \sum_a \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a) \frac{(\bar{\ell}_t(a) - \ell_t(\theta, a))^2}{\bar{\ell}_t(a) + \ell_t(\theta, a)},$$

where $b \in \arg\min_a \bar{\ell}_t(a)$ and $a_t^*(\theta) \in \arg\min_a \ell_t(\theta, a)$. To proceed, we use that for any $p, q \in [0,1]$, we have $\frac{(p-q)^2}{p+q} \leq 4 \cdot \mathcal{D}_H^2(\text{Ber}(p), \text{Ber}(q))$ (cf. Lemma 23 in Appendix E.5). We combine this with the data processing inequality for f-divergences to obtain

$$\begin{aligned} \bar{\ell}_t(b) - \ell_t(\theta, a_t^*(\theta))) &\leq \frac{K}{4\gamma}\bar{\ell}_t(b) + 8\gamma \cdot \sum_a \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a) \mathcal{D}_H^2(\text{Ber}(\bar{\ell}_t(a)), \text{Ber}(\ell_t(\theta, a))) \\ &\leq \frac{K}{4\gamma}\bar{\ell}_t(b) + 8\gamma \cdot \sum_a \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a) \mathcal{D}_H^2(\bar{p}_t(a, \cdot), p_t(\theta, a, \cdot)). \end{aligned} \tag{48}$$

On the other hand, we can rewrite the surrogate regret of the inverse gap weighting policy as

$$\begin{aligned} \bar{r}_t(\pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}) &= \int \sum_a \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a)(\bar{\ell}_t(a) - \ell_t^*(\theta)) \, \mathrm{d}Q_t^+(\theta) \\ &= \int \sum_a \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a)(\bar{\ell}_t(a) - \ell_t(\theta, a_t^*(\theta))) \, \mathrm{d}Q_t^+(\theta) \\ &= \int \sum_{a \neq b} \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a)(\bar{\ell}_t(a) - \bar{\ell}_t(b)) \, \mathrm{d}Q_t^+(\theta) + \int \sum_a \pi(a)(\bar{\ell}_t(b) - \ell_t(\theta, a_t^*(\theta))) \, \mathrm{d}Q_t^+(\theta). \end{aligned}$$

The second term in the above decomposition can be bounded using Equation (48). As for the first term, we can exploit the definition of the policy to write

$$\sum_{a \neq b} \pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}(a)(\bar{\ell}_t(a) - \bar{\ell}_t(b)) = \sum_{a \neq b} \frac{\bar{\ell}_t(b)(\bar{\ell}_t(a) - \bar{\ell}_t(b))}{K\bar{\ell}_t(b) + \gamma(\bar{\ell}_t(a) - \bar{\ell}_t(b))} \leq \frac{K\bar{\ell}_t(b)}{\gamma}.$$

Putting these bounds together gives

$$\begin{aligned} \bar{r}_t(\pi^{\text{(IGW)}}_{\gamma,\bar{\ell}_t}) &\leq \frac{K\bar{\ell}_t(b)}{\gamma} + \frac{K\bar{\ell}_t(b)}{4\gamma} + 8\gamma \cdot \int \sum_a \mathcal{D}_H^2(\bar{p}_t(a, \cdot), p_t(\theta, a, \cdot)) \, \mathrm{d}Q_t^+(\theta) \\ &\leq \frac{5K\bar{\ell}_t(b)}{4\gamma} + 8\gamma \cdot \overline{\text{IG}}_t, \end{aligned}$$

Optimizing over γ, we get the claim of Lemma 5.

E.5 Auxiliary results

Lemma 23 (Proposition 3 Foster and Krishnamurthy, 2021)**.** *For any $p, q \in [0,1]$, we have*

$$\frac{(p-q)^2}{p+q} \leq 4\mathcal{D}_H^2(\text{Ber}(p), \text{Ber}(q)).$$

Proof. The statement follows from the simple calculation

$$\mathcal{D}_H^2(p,q) \geq \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 = \frac{1}{2}\left(\frac{(\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})}{\sqrt{p} + \sqrt{q}}\right)^2 = \frac{1}{2}\frac{(p-q)^2}{(\sqrt{p} + \sqrt{q})^2} \geq \frac{1}{4}\frac{(p-q)^2}{p+q},$$

where the last step uses the elementary inequality $(x+y)^2 \leq 2(x^2 + y^2)$ that holds for any x, y. □

Lemma 24. *For any fixed $q \in [0,1]$, the function $p \mapsto \mathcal{D}_H^2(\text{Ber}(p), \text{Ber}(q))$ is $\frac{1}{4}$-strongly convex.*

Proof. The proof is based on showing that the second derivative of the function of interest is uniformly lower bounded by a positive constant. This follows from calculating the first derivative as

$$\frac{\partial \mathcal{D}_H^2(p,q)}{\partial p} = \frac{1}{2}\left(-\sqrt{\frac{q}{p}} + \sqrt{\frac{1-q}{1-p}}\right),$$

and then lower-bounding the second derivative as

$$\frac{\partial^2 \mathcal{D}_H^2(p,q)}{\partial^2 p} = \frac{1}{4}\left(\sqrt{\frac{q}{p^3}} + \sqrt{\frac{1-q}{(1-p)^3}}\right) \geq \frac{1}{4}\left(\sqrt{q} + \sqrt{1-q}\right) \geq \frac{1}{4}.$$

This inequality is tight when $q = 0$ or $q = 1$. □