Information Directed Sampling and Bandits with Heteroscedastic Noise

Johannes Kirschner JKIRSCHNER@INF.ETHZ.CH

Andreas Krause KRAUSEA@INF.ETHZ.CH

Department of Computer Science
ETH Zurich

Abstract

In the stochastic bandit problem, the goal is to maximize an unknown function via a sequence of noisy evaluations. Typically, the observation noise is assumed to be independent of the evaluation point and to satisfy a tail bound uniformly on the domain; a restrictive assumption for many applications. In this work, we consider bandits with heteroscedastic noise, where we explicitly allow the noise distribution to depend on the evaluation point. We show that this leads to new trade-offs for information and regret, which are not taken into account by existing approaches like upper confidence bound algorithms (UCB) or Thompson Sampling. To address these shortcomings, we introduce a frequentist regret analysis framework, that is similar to the Bayesian framework of Russo and Van Roy (2014), and we prove a new high-probability regret bound for general, possibly randomized policies, which depends on a quantity we refer to as *regret-information ratio*. From this bound, we define a frequentist version of Information Directed Sampling (IDS) to minimize the regret-information ratio over all possible action sampling distributions. This further relies on concentration inequalities for online least squares regression in separable Hilbert spaces, which we generalize to the case of heteroscedastic noise. We then formulate several variants of IDS for linear and reproducing kernel Hilbert space response functions, yielding novel algorithms for Bayesian optimization. We also prove frequentist regret bounds, which in the homoscedastic case recover known bounds for UCB, but can be much better when the noise is heteroscedastic. Empirically, we demonstrate in a linear setting with heteroscedastic noise, that some of our methods can outperform UCB and Thompson Sampling, while staying competitive when the noise is homoscedastic.

1. Introduction

In the stochastic bandit problem one seeks to maximize an unknown function through a sequence of noisy evaluations. At each iteration, an algorithm chooses a point of the domain, trying to maximize the sum of rewards obtained. This objective creates a *dilemma of exploration and exploitation*, as one needs to balance acquiring more information about the unknown function and choosing points which lead to a high reward. To make the problem tractable in cases where the domain is infinite or continuous, the class of possible reward functions has to be restricted, for example to linear functions or a reproducing kernel Hilbert space (RKHS). Further, an assumption commonly used to get concentration results, is that the observation noise is (conditionally) independent and satisfies a tail bound. In most of the existing literature, the noise bound is assumed to be known and independent of the evaluation point. For many applications, this is arguably restrictive in several respects, as for instance the noise distribution might indeed depend on the evaluation point, or the noise in the relevant part of the domain is much lower than the uniform upper bound. To name just

© J. Kirschner & A. Krause.

a few examples, consider reinforcement learning applications, where data is generated via roll-outs, and noise inherently dependents on the exploration policy and the part of the environment explored. We also point to hyper-parameter tuning on physical systems, where the measurement noise often largely varies with the chosen parameter settings.

In this work, we introduce *stochastic bandits with heteroscedastic noise*, where we explicitly allow the noise distribution to depend on the evaluation point. With heteroscedastic noise, *two key challenges* arise, which are not present in the homoscedastic case. First, to *estimate* the unknown function, we would like to use the noise dependency to construct a more accurate estimator. Intuitively, one should assign more weight to observations which are less noisy, and indeed, for least squares estimation this leads to a variant known as *weighted least squares*. Second, regarding *exploration and exploitation*, an algorithm should in general prefer actions with less noise, as the observations obtained then result in a better estimate of the unknown function. This calls for methods which efficiently trade off information and regret, something of importance beyond the setting we discuss here. By introducing a novel frequentist regret analysis framework, we derive several algorithms which are able to balance the expected regret against the *informativeness* of the action taken at each time step.

1.1. Related Work

For a general introduction to the multi-armed bandit problem we refer the reader to the survey of Bubeck and Cesa-Bianchi (2012) and references therein, and we only point out a few key references below. The well-known upper confidence bound strategy (UCB) was first introduced for the multi-armed bandit problem by Auer et al. (2002), and later extended to linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011) and Gaussian processes (Srinivas et al., 2010; Abbasi-Yadkori, 2012; Chowdhury and Gopalan, 2017). Another popular algorithm is Thompson Sampling (Thompson, 1933), where frequentist regret guarantees are known in the linear case (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017) and for RKHS functions (Chowdhury and Gopalan, 2017).

A Bayesian analysis of Thompson Sampling was given by Russo and Van Roy (2016), and this work also motivated *Information Directed Sampling* (IDS, Russo and Van Roy, 2014). Bayesian IDS has recently been shown to achieve state-of-the-art performance in the sequential matrix completion problem (Marsden and Bacallado, 2017).

For the case of finite, linear bandits, Lattimore and Szepesvari (2017) provide an asymptotically instance-optimal algorithm, and showed that both UCB and Thompson Sampling can be arbitrarily far away from the lower bound. Our work adds to this line of research insofar as we identify another setting, where standard approaches like UCB fail to appropriately trade off information and regret.

We are aware of only a few publications addressing heteroscedastic noise in the bandit setting. Most notably is the work of Cowan et al. (2015) for the case of finite, independent arms with Gaussian noise and unknown variances, where they show both finite time bounds and an asymptotically instance optimal algorithm. Some connections to the resource allocation problem have been pointed out by Lattimore et al. (2015), and heavy-tailed noise distributions are discussed by Lattimore (2017). While Bayesian IDS handles the complications introduced in our setting in principle, there are no frequentist regret guarantees known for this algorithm. In different context, the heteroscedastic noise problem has been studied before; for example in linear least squares regression (Aitken, 1936), Gaussian process regression (Goldberg et al., 1998; Kersting et al., 2007), in active learning (Antos et al., 2010; Chaudhuri et al., 2017) and Bayesian optimization (Assael et al., 2014).

1.2. Contributions

In this work, we formalize the stochastic bandit problem with heteroscedastic noise and address the challenges we have outlined in the introduction.

- We show in Example 1, that the widely used UCB or Thompson Sampling strategies do not appropriately account for heteroscedasticity, and can fail to resolve even simple examples.
- In Section 3, we show a new high-probability regret bound for randomized policies, that depends on the choice of an *information gain function*, and a quantity we call *regret-information ratio*. One can view this derivation as a frequentist analog of the Bayesian framework of Russo and Van Roy (2014); or as a generalization of the standard analysis of UCB and its variants (Abbasi-Yadkori et al., 2011; Srinivas et al., 2010). We also introduce a new concentration inequality for supermartingales (Theorem 11), which might be of independent interest.
- Our framework motivates a frequentist version of *Information Directed Sampling* (IDS), introduced in Section 4. This strategy is designed to have a small *regret-information ratio*, as it appears in our main result. Opposed to the Bayesian setting, where the information ratio can be directly calculated using the prior distribution, we instead define IDS to minimize a frequentist surrogate of the regret-information ratio over all sampling distributions. Moreover, we prove several properties of the regret-information ratio, which were previously only known for finite action sets. We also propose a deterministic variant of IDS, which is computationally cheaper while retaining some of the advantages of the randomized version.
- We extend known concentration bounds for online least squares regression (Abbasi-Yadkori, 2012) to heteroscedastic noise (Section 5), which allows us to derive several variants of IDS in Section 6. Further, we prove regret bounds for IDS, which in the homoscedastic case are comparable to known bounds for UCB, but can be much better when the noise is heteroscedastic.
- In Section 7, we empirically demonstrate that some of our approaches can outperform UCB and Thompson Sampling in a linear setting with heteroscedastic noise.

Finally, we emphasize that, even though in this work we primarily focus on bandits with heteroscedastic noise, *Information Directed Sampling* and the surrounding framework are formulated more generally, and could be used to derive new algorithms for different settings as well.

2. Problem Statement

Let $f : \mathcal{X} \to \mathbb{R}$ be a continuous function, that maps from a compact, metric space $\mathcal{X}$ to the real numbers. Using the terminology of the bandit literature, we refer to $\mathcal{X}$ as the action set, and call an element $x \in \mathcal{X}$ an action (or arm). Denote by $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $\mathcal{X}$. Define a policy $\pi = (\pi_t)_{t\geq 1}$ to be a sequence of mappings $\pi_t : (\mathcal{X} \times \mathbb{R})^{t-1} \to \mathcal{P}(\mathcal{X})$ from histories to distributions on $\mathcal{X}$. We say a policy is deterministic, if π_t is always supported on a single action. Further, let $(\epsilon_t)_{t\geq 1}$ be a real-valued noise process. Then we define a stochastic sequence $(x_t, y_t)_{t\geq 1}$ with natural filtration $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_t, y_t)$, such that $x_t \sim \pi_t(x_1, y_1, \dots, x_{t-1}, y_{t-1})$ is the action chosen by the policy and $y_t = f(x_t) + \epsilon_t$ is a noisy evaluation of $f(x_t)$. The goal is to find a policy, which maximizes the cumulative reward $\sum_{t=1}^T f(x_t)$ or, equivalently, minimizes the regret $R_T = \max_{x\in\mathcal{X}} \sum_{t=1}^T \big(f(x) - f(x_t)\big)$. By introducing the sub-optimality gaps $\Delta(x) = \max_{x'\in\mathcal{X}} f(x') - f(x)$, we can write $R_T = \sum_{t=1}^T \Delta(x_t)$, and we define $S = \max_{x\in\mathcal{X}} \Delta(x)$.

In order to derive concentration inequalities for least squares estimators, we make use of a tail bound on the noise distribution. In contrast to existing work, we allow this bound to depend on the evaluation point $x \in \mathcal{X}$. Formally, let $\rho : \mathcal{X} \to \mathbb{R}_{>0}$ be a continuous, positive function, such that ϵ_t is conditionally $\rho(x_t)$-subgaussian, that is for all $t \geq 1$ and $\rho_t = \rho(x_t)$,

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[e^{\lambda \epsilon_t} | \mathcal{F}_{t-1}, x_t] \leq \exp\left(\frac{\lambda^2 \rho_t^2}{2}\right) . \tag{1}$$

Note that this condition implies that the noise has zero mean, and common examples include Gaussian, Rademacher, and uniform random variables. At this point, we leave open the kind of knowledge the policy has about the noise function ρ; one can think of different settings, where either ρ is given, $\rho(x_t)$ is observed at time t, or is estimated from the observations.

In what follows, we also rely on a frequentist notion of *information gain*. We assume that we are given a sequence $(I_t)_{t\geq 1}$ of non-negative, continuous functions $I_t : (\mathcal{X} \times \mathbb{R})^{t-1} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, and we think of $I_t(x) = I_t(x_1, y_1, \dots, x_{t-1}, y_{t-1}, x)$ as the information gain of evaluating $x \in \mathcal{X}$ at time t. For a given policy π, we define the *maximum information gain at time* T,

$$\gamma_T = \operatorname*{ess\,sup}_{\mathcal{F}_T} \sum_{t=1}^{T} I_t(x_t) , \tag{2}$$

where the *essential supremum* is taken over the stochastic process $(x_t, y_t)_{t=1}^T$, up to sets of measure zero. A specific example, that we explore in greater detail later, is $I_t(x) = \log(1 + \sigma_t(x)^2/\rho(x)^2)$, which in the Bayesian setting with posterior variance $\sigma_t(x)^2$ corresponds to the mutual information $\mathbb{I}(x; f|\mathcal{F}_{t-1})$. In this case our definition of γ_T relates to the notion from Srinivas et al. (2010).

Example 1 (Choice between low and high noise) *To start with, we take any bandit instance with homoscedastic noise and duplicate the set of actions. Then, we increase the amount of observation noise on each copy, such that for each action, one has the choice between a low and a high noise version. As we will see shortly, standard algorithms fail to resolve this seemingly simple example.*

Formally, let $\mathcal{S}$ be a compact set, and define $\mathcal{X} = \mathcal{S} \times \{\rho_1, \rho_2\}$ for $0 < \rho_1 < \rho_2$. We denote $(s, \rho') \in \mathcal{X}$ for $s \in \mathcal{S}$ and $\rho' \in \{\rho_1, \rho_2\}$. Further, let $f = f_{\theta^*} : \mathcal{X} \to \mathbb{R}$ be a parametrized function, such that $f_{\theta^*}(s, \rho_1) = f_{\theta^*}(s, \rho_2)$ for all $s \in \mathcal{S}$. We use the last coordinate to specify the noise bound $\rho(s, \rho') = \rho'$. Naively, any regret guarantee using a uniform noise bound degenerates as $\rho_2 \to \infty$. This seems unnecessary, as choosing (s, ρ_1) over (s, ρ_2) clearly yields a more efficient solution. Unfortunately, this is not the default choice. For standard algorithms, consider the widely used UCB strategy. Here, a confidence set C_t is constructed (possibly depending on ρ), containing the true parameter θ^* with high probability. UCB then evaluates $x_t^{UCB} \in \arg\max_{x \in \mathcal{X}} \max_{\theta \in C_t} f_\theta(x)$. Note that the acquisition function is independent of $\rho(x_t^{UCB})$, therefore UCB has no preference of ρ_1 over ρ_2. The same is true for Thompson Sampling, where $x_t^{TS} \in \arg\max_{x \in \mathcal{X}} f_{\tilde{\theta}}(x)$ is determined using a sample $\tilde{\theta}$ from the posterior distribution of θ.

Beyond this simple example, one can easily come up with more complicated noise structures, which can be exploited by a learning algorithm. As another extreme case, take a linear bandit problem in $\mathbb{R}^d$, and add a set of d basis vectors to the action set, such that observations from these actions have no (or infinitesimal small) observation noise. By scaling the new basis vectors, one can also achieve that none of the new actions is optimal. Still, if these actions are played first, one can have vanishing regret after d steps, but again UCB and Thompson Sampling fail to take these actions; compare also Example 2 of Russo and Van Roy (2014) and Example 4 of Lattimore and Szepesvari (2017).

3. A Frequentist Regret Framework

In this section we introduce a novel high-probability regret bound for general, possibly randomized policies. The only assumption we will use is that $S = \max_{x\in\mathcal{X}} \Delta(x)$ is a bound on the sub-optimality gaps, and we emphasize that the results of this section are independent of any assumptions on the noise process. The framework can be seen as a frequentist analog of the Bayesian analysis of Russo and Van Roy (2014). Similarly, as their bound is stated in terms of the entropy of the optimal action, our bounds depend on the total information gain γ_T, defined through the choice of information gain functions I_t. Alternatively, our framework can be understood as a generalization of the standard regret analysis of UCB and its variants (Abbasi-Yadkori et al., 2011; Srinivas et al., 2010), and in fact, our framework recovers the usual guarantees for UCB. Similar as in the analysis for UCB, we apply Cauchy-Schwarz's inequality to separate the expected information gain from the expected instantaneous regret. Analogously to Russo and Van Roy (2014), this motivates the definition of the *regret-information ratio* at time t,

$$\Psi_t : \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}, \quad \mu \mapsto \frac{\mathbb{E}_\mu[\Delta(x)|\mathcal{F}_{t-1}]^2}{\mathbb{E}_\mu[I_t(x)|\mathcal{F}_{t-1}]} \,. \tag{3}$$

For a fixed policy π with sampling distribution π_t, we write $\Psi_t = \Psi_t(\pi_t)$, and with slight abuse of notation, we define $\Psi_t(x) = \Psi_t(\delta_x)$ as the regret-information ratio of $x \in \mathcal{X}$, where δ_x denotes the Dirac delta distribution at x. In the next theorem, we introduce a regret bound for general, randomized policies, which depends on the regret-information ratio Ψ_t and the total information gain γ_T.

Theorem 1 (Regret Bound for Randomized Policies) *Let $S = \max_{x\in\mathcal{X}} \Delta(x)$, and let Ψ_t, γ_T as defined above. Then, for any policy, with probability at least $1-\delta$, at any time $T \geq 1$, it holds that*

$$R_T \leq \tfrac{5}{4}\sqrt{\sum_{t=1}^T \Psi_t \left(2\gamma_T + 4\gamma_T \log\frac{2}{\delta} + 8\gamma_T\log(4\gamma_T) + 1\right)} + 4S\log\left(\tfrac{8\pi^2T^2}{3\delta}(\log(T)+1)\right) \,.$$

If also $I_t(x_t) \leq 1$ holds for all $t \geq 1$, then, with probability at least $1-\delta$, at any time $T \geq 1$,

$$R_T \leq \tfrac{5}{4}\sqrt{\sum_{t=1}^T \Psi_t \left(2\gamma_T + 4\log\frac{2}{\delta} + 8\log(4) + 1\right)} + 4S\log\left(\tfrac{8\pi^2T^2}{3\delta}(\log(T)+1)\right) \,.$$

We remark that the assumption $I_t \leq 1$ can be viewed as a continuity assumption on the information gain, and is not very restrictive. In fact, for the information gain functions we use later, γ_T is sublinear in T, which in turn implies $I_t \to 0$, and $I_t \leq 1$ is trivially satisfied after some finite time.

The main difficulty in the proof of the general theorem is to deal with randomized policies. Indeed, for deterministic policies, we have the following, simpler version of the previous theorem.

Theorem 2 (Regret Bound for Deterministic Policies) *Let the sequence $(x_t)_{t=1}^T$ be generated by any deterministic policy. Then, at any time $T \geq 1$, the regret is bounded by $R_T \leq \sqrt{\sum_{t=1}^T \Psi_t \, \gamma_T}$.*

Proof Note that for deterministic policies, x_t is predictable with respect to the filtration $\mathcal{F}_t$, and therefore $\Psi_t = \frac{\Delta(x_t)^2}{I_t(x_t)}$. Hence we can write $R_T = \sum_{t=1}^T \Delta(x_t) = \sum_{t=1}^T \sqrt{\Psi_t I_t(x_t)}$ and apply Cauchy-Schwarz to find $R_T \leq \sqrt{\sum_{t=1}^T \Psi_t \cdot \sum_{t=1}^T I_t(x_t)} \leq \sqrt{\sum_{t=1}^T \Psi_t\gamma_T}$. □

We outline the proof of the general theorem, but have to defer details to Appendix B. For brevity, we denote $\Delta_t = \Delta(x_t)$ and $I_t = I_t(x_t)$. The central step in the proof, that allows us to apply the definition of Ψ_t and Cauchy-Schwarz as in the deterministic case, is

$$\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] = \sum_{t=1}^T \sqrt{\Psi_t}\sqrt{\mathbb{E}[I_t|\mathcal{F}_{t-1}]} \leq \sqrt{\sum_{t=1}^T \Psi_t \cdot \sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}]} \,.$$

To get there, we have to bound the sum of martingale differences $\Delta_t - \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]$, such that we can express the regret in terms of $\mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]$. We attain this at the expense of an additive factor of order $\mathcal{O}\left(S \log \frac{T}{\delta}\right)$ using a 'peeling' argument on Freedman's inequality and the union bound on T.

The more challenging part of the proof is to bound the sum of expected information gains $\sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}]$ by γ_T. Again, we need to bound a martingale difference sequence $\mathbb{E}[I_t|\mathcal{F}_{t-1}] - I_t$, such that we can make use of $\sum_{t=1}^T I_t \leq \gamma_T$. It is possible to apply Freedman's inequality as in the first step, or even an Azuma-Hoeffding type inequality like for example Corollary 2.7 of Fan et al. (2015). However, both ways would introduce a multiplicative $\sqrt{\log(T)}$ factor in our leading term through the union bound over T.

To avoid this, note that $g_t = \mathbb{E}[I_t|\mathcal{F}_{t-1}]$ is a predictable, non-negative process, which upper bounds the martingale differences $g_t - I_t \leq g_t$. Using the method of mixtures, similar as done by Abbasi-Yadkori et al. (2011), and Corollary 2.7 in (Fan et al., 2015), we prove a new anytime concentration inequality for bounded supermartingales (Theorem 11 in Appendix A). As a consequence, we get the following lemma, showing that, for any non-negative stochastic process X_t, with high-probability the sum of conditional means $\sum_{t=1}^T \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ is not much larger than $\sum_{t=1}^T X_t$.

Lemma 3 (Concentration of conditional mean) *Let X_t be any non-negative stochastic process adapted to a filtration $\{\mathcal{F}_t\}$, and define $m_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ and $M_T = \sum_{t=1}^T m_t$. Further assume that $X_t \leq b_t$ for a fixed, non-decreasing sequence $(b_t)_{t\geq 1}$ and let $(l_t)_{t\geq 1}$ be any fixed, positive sequence. Then, with probability at least $1-\delta$, for any $T \geq 1$,*

$$\sum_{t=1}^T m_t - X_t \leq \sqrt{2(b_T M_T + l_T)\log\left(\frac{1}{\delta}\frac{(b_T M_T + l_T)^{1/2}}{l_T^{1/2}}\right)} \,.$$

Further, if $b_T \geq 1$, with probability at least $1-\delta$ for any $T \geq 1$ it holds that,

$$\sum_{t=1}^T m_t \leq 2\sum_{t=1}^T X_t + 4 b_T \log\frac{1}{\delta} + 8 b_T \log(4 b_T) + 1$$

To prove the first inequality, we use the Bhatia-Davis inequality (Lemma 18, Appendix G) to upper bound the sum of conditional variances $\sum_{t=1}^T \mathrm{Var}[X_t|\mathcal{F}_{t-1}] \leq \sum_{t=1}^T m_t(b_t - m_t) \leq b_T \sum_{t=1}^T m_t$, and for the second part of the lemma, note that the right-hand side of the first inequality depends only sublinearly on M_T. All details on the concentration results can be found in Appendix A.

Applied to the sequence I_t, the last equation then gives the main result for the case $I_t \leq 1$, and the unrestricted case follows by using a more conservative bound $I_t \leq \gamma_t$ instead. We also remark that the $\log \frac{1}{\delta}$ factor introduced in the last step cannot be completely avoided, as one would expect for general randomized policies (see also Example 2 in Appendix B).

4. Frequentist Information Directed Sampling

As the regret bound given in the previous section primarily depends on $\sum_{t=1}^T \Psi_t$, we look for policies such that the regret-information ratio is as small as possible. Unfortunately, Ψ_t cannot be directly controlled, as it depends on $\Delta(x_t)$ and therefore on the unknown value $\max_{x\in\mathcal{X}} f(x)$. However, if a confidence band $[l_t(x), u_t(x)]$ is available, containing the true function values $f(x)$ with probability $1-\delta$, one can construct an upper bound $\Delta_t^+(x) = \max_{x'} u_t(x') - l_t(x)$, such that $\Delta(x) \leq \Delta_t^+$ also holds with probability $1-\delta$. Such confidence bounds are known for different estimators and function classes, and we will discuss some variants in the next section. For now we rely on Δ_t^+ as being given to define a *surrogate of the regret-information ratio*,

$$\Psi_t^+(\mu) : \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}, \quad \mu \mapsto \frac{\mathbb{E}_\mu[\Delta_t^+(x)|\mathcal{F}_{t-1}]^2}{\mathbb{E}_\mu[I_t(x)|\mathcal{F}_{t-1}]} .$$

Clearly, by our assumption on Δ_t^+, $\Psi_t(\mu) \leq \Psi_t^+(\mu)$ holds with probability $1-\delta$ for any $\mu \in \mathcal{P}(\mathcal{X})$. We define *Information Directed Sampling* (IDS) to be a policy π_{IDS}, which depends on the choice of information functions $(I_t)_{t\geq 1}$, such that at any time t,

$$\pi_t^{\text{IDS}} \in \arg\min_{\mu \in \mathcal{P}(\mathcal{X})} \Psi_t^+(\mu) . \tag{4}$$

Using Prokhorov's Theorem (Prokhorov, 1956), we show in Lemma 16 in Appendix C, that a minimizer $\mu \in \mathcal{P}(\mathcal{X})$ of $\Psi_t^+(\mu)$ always exists for compact $\mathcal{X}$, assuring that IDS is well-defined. Like the Bayesian counterpart in (Russo and Van Roy, 2014), the regret-information ratio $\Psi_t(\mu)$ and its surrogate $\Psi_t^+(\mu)$ are convex functions of μ, which is a direct consequence of the fact that $h(x,y) = \frac{x^2}{y}$ is convex on $\mathbb{R} \times \mathbb{R}_{>0}$ (see Lemma 17). Moreover, we have the following result, which might also have far-reaching algorithmic implications.

Lemma 4 (Compare Proposition 6 in (Russo and Van Roy, 2014)) *There exists a minimizing distribution $\mu^* \in \arg\min_{\mu\in\mathcal{P}(\mathcal{X})} \Psi_t^+(\mu)$, which is supported on at most two actions.*

The lemma is proven for finite action sets in (Russo and Van Roy, 2014). We address the technical challenges of having a compact set $\mathcal{X}$ in Appendix C, and show that even for continuous action sets, a simple parameterization of the sampling distribution suffices to find a minimizer of Ψ_t^+. In settings with finitely many actions, this allows to minimize $\Psi_t^+(\mu)$ by iterating over all pairs of actions, or one can use convexity of Ψ_t^+ to employ gradient descent on the probability simplex.

By the previous lemma, IDS can be chosen in such a way, that the sampling distribution π_t^{IDS} is supported on at most two actions at any time. That this is necessary in general can easily be seen in a simple example. However, one can still directly minimize the regret-information ratio over distributions supported on a single action. We define *Deterministic Information Directed Sampling* (DIDS) to be a deterministic policy which at time t chooses an action

$$x_t^{\text{DIDS}} \in \arg\min_{x\in\mathcal{X}} \Psi_t^+(x) . \tag{5}$$

As the minimization happens over a subset of $\mathcal{P}(\mathcal{X})$, clearly $\min_{\mu\in\mathcal{P}(\mathcal{X})} \Psi_t^+(\mu) \leq \min_{x\in\mathcal{X}} \Psi_t^+(x)$. Empirically, we found that which version has lower regret strongly dependents on the information function used to define IDS. We think that DIDS could oftentimes yield a computational cheaper alternative to IDS with similar or sometimes better performance. Understanding the potential benefits of the randomized version is a task for future work.

5. Online Least Squares Estimation with Heteroscedastic Noise

So far we have defined *Information Directed Sampling* in terms of a surrogate Ψ_t^+ of the regret-information ration, which in turn relies on a high-probability upper bound Δ_t^+ on the sub-optimality gaps. To construct this upper bound in the next section, we briefly discuss variants of classical least squares estimation for observations with heteroscedastic noise (see also Aitken, 1936; Rasmussen and Williams, 2006), and generalize the concentration inequalities of Abbasi-Yadkori (2012) for least squares estimators in $\mathbb{R}^d$ and separable Hilbert spaces to the setting with heteroscedastic noise. Throughout this section, we assume that $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and that the noise satisfies the subgaussian condition (1), which we summarize in a diagonal *noise matrix* $\Sigma = \Sigma_T = \text{diag}(\rho_1^2, \dots, \rho_T^2)$.

5.1. Linear Least Squares with Heteroscedastic Noise

Assume that the data $\{x_t, y_t\}_{t=1}^T$ is generated from a linear function $f(x) = x^\top \theta^*$ with unknown parameter $\theta^* \in \mathbb{R}^d$. As common, we define the design matrix $X = X_T = (x_1, \dots, x_T)^\top$ and the observation vector $y = y_T = (y_1, \dots, y_T)^\top$, where we drop the subscript T in favor of readability.

With heteroscedastic noise, the ordinary least squares estimator is known to be inefficient in general. Aitken (1936) was first to consider an alternative known as *weighted linear least squares*, which is defined to minimize the weighted norm $\|X\theta - y\|_{\Sigma^{-1}}^2$. This amounts to applying ordinary least squares to normalized data $X' = \Sigma^{-1/2} X$ and $y' = \Sigma^{-1/2} y$, and has the wanted effect of putting less weight on noisier observation. Moreover, the Gauss-Markov theorem asserts that the estimator has the smallest possible variance among all linear and unbiased estimators, given a fixed design X and Gaussian noise. Adding a Tikhonov regularization term depending a positive definite matrix V_0, we set $\hat{\theta}_T = \arg\min_\theta \|y - \theta^\top X\|_{\Sigma^{-1}}^2 + \|\theta\|_{V_0}^2$, and with $V_T = X^\top \Sigma^{-1} X + V_0$, a closed form solution is given by $\hat{\theta}_T = V_T^{-1} X^\top \Sigma^{-1} y$. The estimator $\hat{\theta}_T$ can also be motivated in the Bayesian setting with prior $p(\theta) \sim \mathcal{N}(0, V_0^{-1})$ and likelihood $p(y|X, \theta) \sim \mathcal{N}(X\theta, \Sigma)$, where $\hat{\theta}$ appears as mean of the posterior distribution $p(\theta|X, y) = \mathcal{N}(\hat{\theta}_T, V_T^{-1})$.

In the next theorem, we extend the elegant concentration inequality for $\hat{\theta}$ from Abbasi-Yadkori et al. (2011) to the setting with heteroscedastic noise.

Lemma 5 (Theorem 2 in (Abbasi-Yadkori et al., 2011) for heteroscedastic noise) *Let $(x_t, y_t)_{t\geq 1}$ be any stochastic process in $\mathbb{R}^d \times \mathbb{R}$ such that $y_t = x_t^\top \theta^* + \epsilon_t$ and ϵ_t satisfies noise assumption (1). Then, for $\hat{\theta}_T$ as defined above, with probability at least $1 - \delta$, for any $T \geq 1$, it holds that*

$$\|\hat{\theta}_T - \theta^*\|_{V_T} \leq \sqrt{2 \log\left(\frac{1}{\delta} \frac{\det(V_T)^{1/2}}{\det(V_0)^{1/2}}\right)} + \|\theta^*\|_{V_0} .$$

Proof Theorem 2 in (Abbasi-Yadkori et al., 2011) proves the case $\rho_t = R$. We show that this immediately translates to the general case. For any $t \geq 1$, define $x_t' = \frac{x_t}{\rho_t}$, and corresponding observations $y_t' = \frac{y_t}{\rho_t}$. Now we apply the case $R = 1$ to the process $\{x_t', y_t'\}$ and the corresponding estimator $\hat{\theta}_T' = V_T'^{-1} X' y'$. Observe that $\hat{\theta}_T = \hat{\theta}_T'$ and $V_T = V_T'$ to complete the proof. □

Note that by Cauchy-Schwarz, $|x^\top \hat{\theta}_T - x^\top \theta^*| \leq \|\hat{\theta}_T - \theta^*\|_{V_T} \|x\|_{V_T^{-1}}$ for all $x \in \mathcal{X}$, and in combination with the previous lemma, this yields confidence intervals $\hat{f}_T(x) \pm \beta_T \sigma_T(x)$ that contain the true function value $f(x)$ with probability at least $1 - \delta$ at any time T, where $\hat{f}_T(x) = x^\top \hat{\theta}_T$, $\beta_T = \sqrt{2 \log\left(\frac{1}{\delta} \frac{\det(V_T)^{1/2}}{\det(V_0)^{1/2}}\right)} + \|\theta^*\|_{V_0}$ and $\sigma_t(x) = \|x\|_{V_T^{-1}}$.

5.2. Least Squares Estimation in Separable Hilbert Spaces

Let $\mathcal{H}$ be a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and corresponding norm $\| \cdot \|_{\mathcal{H}}$. For a positive definite operator $V_0 : \mathcal{H} \to \mathcal{H}$, we denote the inner product $\langle \cdot, \cdot \rangle_{V_0} := \langle \cdot, V_0 \cdot \rangle$ with corresponding norm $\| \cdot \|_{V_0}$. Let $f^* \in \mathcal{H}$ be the unknown function used to generate the data $\{v_t, y_t\}_{t=1}^T$, such that $y_t = \langle v_t, f^* \rangle_{\mathcal{H}} + \epsilon_t$. As substitute for the design matrix, define the operator $M = M_T : \mathcal{H} \to \mathbb{R}^T$, such that for all $v \in \mathcal{H}$ and $t = 1, \dots, T$, $(Mv)_t = \langle v_t, v \rangle$, and denote its adjoint by $M^* : \mathbb{R}^T \to \mathcal{H}$. Taking the noise matrix $\Sigma = \Sigma_T = \mathrm{diag}(\rho_1^2, \dots, \rho_T^2)$ into account, we define the regularized least squares estimator $\hat{\mu}_T = \arg\min_{f \in \mathcal{H}} \|Mf - y\|_{\Sigma^{-1}}^2 + \|f\|_{V_0}^2$. Again, a closed form is given by $\hat{\mu}_T = V_T^{-1} M^* \Sigma^{-1} y$, where $V_T = M^* \Sigma^{-1} M + V_0$.

We now turn to concentration inequalities for $\hat{\mu}_T$ in separable Hilbert spaces as first introduced by Abbasi-Yadkori (2012). The same inequality for the special case of a RKHS was recently re-derived by Chowdhury and Gopalan (2017).

Lemma 6 (Theorem 3.11 in (Abbasi-Yadkori, 2012) for heteroscedastic noise) *Let $(v_t, y_t)_{t \geq 1}$ be a stochastic process in $\mathcal{H} \times \mathbb{R}$, such that $y_t = \langle v_t, f^* \rangle_{\mathcal{H}} + \epsilon_t$ satisfies noise assumption (1). Then with probability at least $1 - \delta$, for any $v \in \mathcal{H}$ and for all $T \geq 1$,*

$$|\langle v, \hat{\mu}_T \rangle - \langle v, f^* \rangle| \leq \left(\sqrt{2 \log \left(\frac{1}{\delta} \frac{\det(\Sigma_T + M_T V_0^{-1} M_T^*)^{1/2}}{\det(\Sigma_T)^{-1/2}} \right)} + \|f^*\|_{V_0} \right) \|v\|_{V_T^{-1}} .$$

Proof Theorem 3.11 in Abbasi-Yadkori (2012) proves the case $\rho_t = R$. Define $y_t' = y_t / \rho_t$ and $v_t' = v_t / \rho_t$, and apply the case $R = 1$ to $\hat{\mu}_T'$. Note that $\hat{\mu}_T = \hat{\mu}_T'$ to complete the proof. □

Least Squares Estimation in RKHS Consider the important special case, where $\mathcal{H}$ is a RKHS over $\mathbb{R}^d$ with kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and canonical embeddings $k_x = k(x, \cdot)$ of $x \in \mathbb{R}^d$. Let $v_t = k_{x_t} \in \mathcal{H}$ be the embedding of $x_t \in \mathbb{R}^d$, such that $\langle v_t, f_* \rangle_{\mathcal{H}} = \langle k_{x_t}, f_* \rangle_{\mathcal{H}} = f^*(x_t)$. Notably, for $V_0 = \lambda I$, the representer theorem yields the following, tractable form of $\hat{\mu}_T$,

$$\hat{\mu}_T(x) = \langle \hat{\mu}_T, k_x \rangle_{\mathcal{H}} = k_T(x)^\top (K_T + \lambda \Sigma_T)^{-1} y ,$$

where we defined $(K_T)_{i,j} = k(x_i, x_j)$ and $k_T(x)_t = k(x_t, x)$. Details of this derivation are given in Appendix D. Similarly, we get $\|v\|_{V_T^{-1}}^2 = \frac{1}{\lambda} \left(k(x,x) - k_T(x)^\top (K_T + \lambda \Sigma_T)^{-1} k_T(x) \right) =: \sigma_T(x)^2$. Also note that $\hat{\mu}_T$ and σ_T can be iteratively updated using Schur's complement (Chowdhury and Gopalan, 2017, Appendix F), avoiding the computation of a $T \times T$ matrix inverse at every step. In the next lemma, we specialize the previous result to the RKHS setting.

Lemma 7 *Let $\mathcal{H}$ be a RKHS, and $\hat{\mu}_T$ be the estimator defined above with $V_0 = \lambda I$ for $\lambda > 0$. Assume that the process $(x_t, y_t)_{t \geq 1}$ satisfies the noise assmption (1). Then, the following holds with probability at least $1 - \delta$, for all $T \geq 1$,*

$$|\hat{\mu}_T(x) - f^*(x)| \leq \left(\sqrt{2 \log \left(\frac{1}{\delta} \frac{\det(\lambda \Sigma_T + K_T)^{1/2}}{\det(\lambda \Sigma_T)^{1/2}} \right)} + \sqrt{\lambda} \|f^*\|_{\mathcal{H}} \right) \sigma_T(x) .$$

We again draw the connection to Bayesian optimization. Let $f^* \sim \mathrm{GP}(0, \lambda^{-1} k)$ be a sample from a Gaussian process, where $\lambda > 0$ is a parameter trading off the magnitude of the prior variance, and assume that the observation noise ϵ_t independent and $\mathcal{N}(0, \rho_t^2)$ distributed. Then the posterior distribution of f^* is also a Gaussian process with mean $\hat{\mu}_T$ and covariance kernel $k_T(x, y) = \frac{1}{\lambda} \left(k(x, y) - k_T(x)(K_T + \lambda \Sigma_T)^{-1} k_T(y) \right)$, and $\sigma_T(x)^2$ is the posterior variance at x.

6. Regret Bounds for Information Directed Sampling on Heteroscedastic Bandits

We now define Δ_t^+ for linear and RKHS functions f, using the regularized least squares estimator and the corresponding confidence intervals of Lemma 5 and 7. In this case, we get estimators $\hat{f}_t : \mathcal{X} \to \mathbb{R}$ and a symmetric confidence interval $|f(x) - \hat{f}_t(x)| \leq \beta_t^\delta \sigma_t(x)$ at level $1 - \delta$ for non-decreasing scale factor $\beta_t = \beta_t^\delta$ and confidence width $\sigma_t(x)$. We define the surrogate

$$\Delta_t^+(x) = \max_{x' \in \mathcal{X}} \left(\hat{f}_t(x') + \beta_t \sigma_t(x')\right) - \left(\hat{f}_t(x) - \beta_t \sigma_t(x)\right) .$$

Note that $x_t^{\text{UCB}} \in \arg\max_{x \in \mathcal{X}} \left(\hat{f}_t(x) + \beta_t \sigma_t(x)\right)$ is the UCB action and $\Delta_t^+(x_t^{\text{UCB}}) = 2\beta_t \sigma_t(x_t^{\text{UCB}})$. For simplicity, we now make the assumption that the noise function ρ is known, and introduce two information gain functions $I_t^{\text{F}}(x)$ and $I_t^{\text{UCB}}(x)$, which will naturally lead to regret bounds for IDS.

IDS-F The first information function we propose is $I_t^{\text{F}} = \log\left(1 + \frac{\sigma_t(x)^2}{\rho(x)^2}\right)$. We motivate this in the Bayesian setting with Gaussian prior and likelihood, where I_t^{F} is up to a constant factor equal to the *conditional mutual information* $\mathbb{I}(f; x|\mathcal{F}_{t-1})$. In the linear case, this follows by writing $\mathbb{I}(\theta; x_t, y_t|\mathcal{F}_{t-1}) = \mathbb{H}(\theta|\mathcal{F}_{t-1}) - \mathbb{H}(\theta|\mathcal{F}_{t-1}, x_t, y_t)$, and then using the matrix determinant lemma,

$$2\,\mathbb{I}(\theta; x_t, y_t|\mathcal{F}_{t-1}) = \log\left((2\pi e)^d \det(V_t^{-1})\right) - \log\left((2\pi e)^d \det(V_{t+1}^{-1})\right) = \log\left(1 + \frac{\sigma_t(x_t)^2}{\rho(x_t)^2}\right) .$$

Note that the information gain $I_t(x)$ is diminishing as the uncertainty $\sigma_t(x)$ in f decreases, allowing to bound γ_T. Importantly, $I_t(x)$ also characterizes the information gain depending on the noise, such that $I_t(x)$ decreases for larger $\rho(x)$. By compactness of $\mathcal{X}$, we can also ensure that $I_t(x) \leq 1$.

IDS-UCB Let $I_t^{\text{UCB}}(x) = \log\left(\frac{\sigma_t(x_t^{\text{UCB}})^2}{\sigma_t(x_t^{\text{UCB}}|x)^2}\right)$, where we define $\sigma_t(x_t^{\text{UCB}}|x)$ as the confidence width at x_t^{UCB} after x has been evaluated. For the least squares estimators that we use in Section 5, this can be computed without knowing the outcome of evaluating at x and we explicitly calculate this quantity in Appendix E . Again, I_t^{UCB} can be motivated in the Bayesian setting, where it corresponds to the mutual information $\mathbb{I}(x; f(x_t^{\text{UCB}})|\mathcal{F}_{t-1})$. From the data processing inequality (Cover and Thomas, 2012), it further follows that $I_t^{\text{UCB}}(x) \leq I_t^{\text{F}}(x)$ for all $x \in \mathcal{X}$. The resulting IDS-UCB policy collects information about the UCB action, but is not restricted to playing the same, if a more informative alternative action is available. Note also that both DIDS-F and DIDS-UCB are not more expensive to compute than standard UCB. We discuss further variants of IDS in Appendix E.

To bound the regret of IDS-F and IDS-UCB, we will need the following lemma, which relates the regret-information ratio of IDS to the regret-information ratio of UCB.

Lemma 8 *At any time t, denote the sampling distribution of IDS with μ_t^{IDS}, the action chosen by DIDS and UCB with x_t^{DIDS} and x_t^{UCB} respectively. Then, $\Psi_t^+(\mu_{IDS}) \leq \Psi_t^+(x_{DIDS}) \leq \Psi_t^+(x_{UCB})$ holds. Moreover, if I_t^F or I_t^{UCB} is used, $\Psi_t^+(x_{UCB}) = \frac{8\beta_t^2 \sigma_t(x_{UCB})^2}{\log(1+\sigma_t(x_{UCB})^2/\rho(x_{UCB})^2)}$.*

If in this case also $I_t(x) \leq 1$ holds, then $\Psi_t^+(x_{UCB}) \leq 8\beta_t^2 \rho(x_{UCB})^2$, and consequently with $R = \max_{x \in \mathcal{X}} \rho(x)$, for any of these strategies, $\sum_{t=1}^T \Psi_t \leq 8\beta_T^2 R^2 T$.

Proof The sequence of inequalities follows directly from the respective definitions. The formula for $\Psi_t^+(x_{\text{UCB}})$ follows by noting that $I_t^F(x_t^{\text{UCB}}) = I_t^{\text{UCB}}(x_t^{\text{UCB}}) = \log\left(1 + \frac{\sigma_t(x_{\text{UCB}})^2}{\rho(x_{\text{UCB}})^2}\right)$. If $I_t(x) \leq 1$ for all $x \in \mathcal{X}$, we can use $2\log\left(1 + \frac{\sigma_t(x)^2}{\rho(x)^2}\right) \geq \frac{\sigma_t(x)^2}{\rho(x)^2}$, which gives$\Psi_t^+(x_{\text{UCB}}) \leq 8\beta_t^2 \rho(x_{\text{UCB}})^2$. The last inequality uses that β_t is non-decreasing, and uniformly bounds $\rho(x) \leq R$. □

f	β_T^δ	γ_T	R_T
linear	$\mathcal{O}\Big(\sqrt{d\log(T)+\log(\frac{1}{\delta})}\Big)$	$\mathcal{O}\big(d\log(T)\big)$	$\mathcal{O}\left((d\log(T)+\log\frac{1}{\delta})\sqrt{T}\right)$
RKHS with RBF kernel	$\mathcal{O}\left(\sqrt{\log(T)^d+\log(\frac{1}{\delta})}\right)$	$\mathcal{O}\left(\log(T)^d\right)$	$\mathcal{O}\left((\log(T)^d+\log(\frac{1}{\delta}))\sqrt{T}\right)$

Table 1: Summary of worst case regret bounds for IDS-F/IDS-UCB. For the bounds on β_T^δ and γ_T, refer to (Abbasi-Yadkori et al., 2011; Srinivas et al., 2010), including bounds for further kernels.

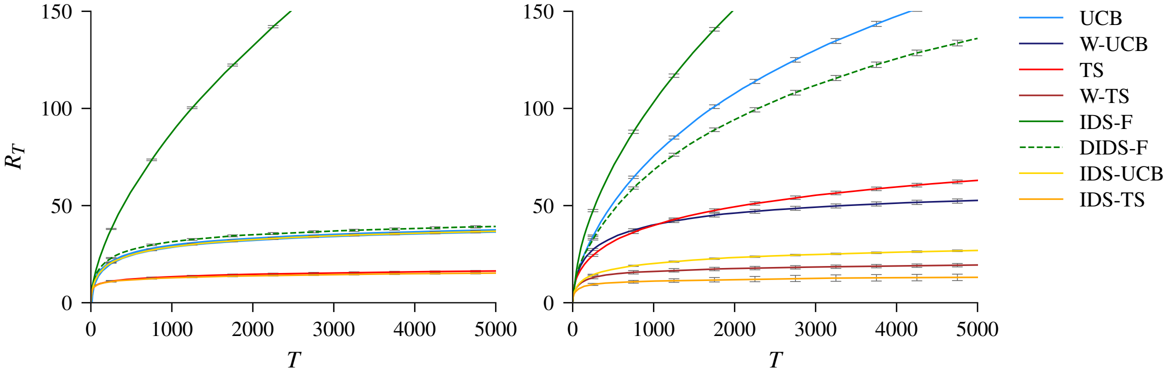
Finally, we obtain regret bounds for IDS and DIDS using the previous lemma, Theorem 1 or 2, and the concentration inequalities in Lemma 5 and 7.

Corollary 9 *Denote $\mathcal{X}_t^{UCB}$ the set of UCB actions at time t as observed during the run of IDS and DIDS respectively, and set $R = \sup_{t\geq 1} \min_{x\in\mathcal{X}_t^{UCB}} \rho(x)$. For estimators $\hat{f}_t$ with confidence bounds $\hat{f}_t \pm \beta_t^\delta \sigma_t(x)$, with probability at least $1-\delta$, the regret of IDS-F and IDS-UCB is bounded by $R_T = \mathcal{O}\left(R\beta_T^\delta\sqrt{T(\gamma_T+\log\frac{1}{\delta})}\right)$, and of DIDS-F and DIDS-UCB by $R_T = \mathcal{O}\Big(R\beta_T^\delta\sqrt{T\gamma_T}\Big)$.*

We summarize the resulting regret bounds in Table 1. Note that in the homoscedastic setting, we recover known bounds for UCB via Lemma 8. In settings with heteroscedastic noise, our bounds can be much better, for example as depicted in Example 1, where the noise constant R, as defined in the Corollary, is smaller for IDS than for UCB. To obtain bounds adaptive to more general heteroscedastic settings, one would need to bound the sum $\sum_{t=1}^T \Psi_t$ directly in relation to ρ and the policy; and general instance depending bounds are a significant undertaking even for linear functions (Lattimore and Szepesvari, 2017). As our bounds are *worst-case guarantees*, the actual instance-dependent regret can be rather different; see also the experiments in the next section.

7. Experiments

We show a simple, synthetic experiment, where we fix a linear function in $\mathbb{R}^3$ and for each experiment randomly generated a set of 30 actions contained in the Euclidean unit ball. In the homoscedastic setting (Figure 1a) we use Gaussian noise $\mathcal{N}(0,\rho^2)$ with constant noise bound $\rho = 0.5$, and compare to UCB and TS. We define TS directly using the posterior, and do not employ oversampling as needed for the frequentist guarantees in Agrawal and Goyal (2013). In the heteroscedastic setting (Figure 1b), we chose the bound for each action uniformly in $[0.1, 1]$, and compare to UCB and TS, which use 1 as an upper bound on the noise, and to W-UCB and W-TS, which use the weighted least squares estimator and Lemma 5. For our methods, we show D/IDS-F, IDS-UCB and IDS-TS, where the latter policy is defined by $I_t^{\text{TS}}(x) = \mathbb{I}(x; f(x_t^{\text{TS}})|\mathcal{F}_{t-1})$ with respect to a *proposal action* x_t^{TS} from Thomson Sampling; but we currently do not have regret bounds for IDS-TS. We omit the regret curve of DIDS-UCB and DIDS-TS as it was too close to the respective randomized version. For all methods, we calculate confidence bounds from Lemma 5 directly using the determinant and the matrix V_T, and we fix $\delta = 0.01$, $\lambda = 1$. Error bars show 2 times standard error of 500 trials. We observe that our methods stay competitive with UCB/TS in the homoscedastic setting, expect for IDS-F, which is over-exploring, likely because I_t^F captures global information about f irrespective of the location of the optimum. With heteroscedastic noise, IDS-UCB/IDS-TS show an advantage over UCB/TS, demonstrating the importance of using the noise both for estimation and exploration. Further experiments, including a simulation of Example 1, can be found in Appendix F.



(a) Homoscedastic Noise

(b) Heteroscedastic Noise

Figure 1: With homoscedastic noise, IDS-UCB and UCB achieve almost the same regret, and the same holds for IDS-TS and TS. Interestingly, while DIDS-F performs similar to IDS-UCB, randomization allows IDS-F to explore more; numerically, this regret curve is close to the theoretical bound $\mathcal{O}(\sqrt{T \log T})$. In the heteroscedastic experiment, IDS-UCB is the best method with known frequentist guarantees, outperforming UCB/W-UCB, and similarly IDS-TS improves on TS/W-TS. This demonstrates the importance of using the noise bound both for estimation and exploration.

8. Conclusion and Future Work

We introduced a frequentist version of IDS, and use it to address bandits with heteroscedastic noise. However, we believe that the generality of our analysis opens up many promising directions for future work, some of which we outline below. As termed by Russo and Van Roy (2014), *Information Directed Sampling* should be rather seen as a *design principle* than an algorithm, as it depends on the user-defined choices of I_t and Δ_t^+. Empirically, we observed that the performance of IDS strongly depends on the choice of I_t, hence better understanding and deriving new information functions is an important task for future work. For instance, Russo and Van Roy (2014) use $I_t(x) = \mathbb{I}(x; \arg\max_{x \in \mathcal{X}} f(x) | \mathcal{F}_{t-1})$, but this is in general more expensive to compute and one might have to use approximation techniques, as for example done by Wang and Jegelka (2017) for $I_t(x) = \mathbb{I}(x; \max_{x \in \mathcal{X}} f(x) | \mathcal{F}_{t-1})$. The choice of I_t can also be seen as introducing prior information on the problem structure to the algorithm, which could be useful in other settings as well.

Regarding bandits with heteroscedastic noise, we assumed that the heteroscedastic noise bound ρ is known. For applications one might instead want to estimate the variance of the noise directly from the data, as done by Cowan et al. (2015) for Gaussian multi-armed bandits. We also point out a striking similarity of IDS-F and an instance-dependent lower bound, due to Burnetas and Katehakis (1996), see equation (8) in (Cowan et al., 2015). Finally, it needs to be seen if the notion of regret-information ratio is useful beyond the algorithms we study here. For example, it might be possible to bound the regret-information ratio of Thompson Sampling, where regret bounds in the frequentist setting remain notoriously harder to achieve than in the Bayesian setting.

Acknowledgments

We thank Kfir Levy for helpful discussions on this paper and for pointing out Lemma 10. This research was supported by SNSF grant 200020_159557.

References

Yasin Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, 2012.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320. 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In *AISTATS 2017-20th International Conference on Artificial Intelligence and Statistics*, 2017.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Alexander C Aitken. IV. on least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29):2712–2728, 2010.

John-Alexander M Assael, Ziyu Wang, Bobak Shahriari, and Nando de Freitas. Heteroscedastic treed bayesian optimisation. *arXiv preprint arXiv:1410.7172*, 2014.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Kamalika Chaudhuri, Prateek Jain, and Nagarajan Natarajan. Active Heteroscedastic Regression. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 694–702. PMLR, August 2017.

Sayak Ray Chowdhury and Aditya Gopalan. On Kernelized Multi-armed Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, August 2017.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Wesley Cowan, Junya Honda, and Michael N. Katehakis. Normal Bandits of Unknown Means and Variances: Asymptotic Optimality, Finite Horizon Regret Bounds, and a Solution to an Open Problem. *arXiv:1504.05823 [cs, stat]*, April 2015.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, pages 355–366. Omnipress, 2008.

Rick Durrett. *Probability: Theory and Examples*. Cambridge university press, 2010.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Paul Goldberg, Christopher K. I. Williams, and Christopher M. Bishop. *Regression with Input-Dependent Noise: A Gaussian Process Treatment*. 1998.

Sham M Kakade and Ambuj Tewari. On the Generalization Ability of Online Strongly Convex Programming Algorithms. In *Advances in Neural Information Processing Systems 21*, pages 801–808. 2009.

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most Likely Heteroscedastic Gaussian Process Regression. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 393–400. ACM, 2007.

Tor Lattimore. A Scale Free Algorithm for Stochastic Bandits with Bounded Kurtosis. In *Advances in Neural Information Processing Systems 30*, pages 1583–1592. 2017.

Tor Lattimore and Csaba Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017.

Tor Lattimore, Koby Crammer, and Csaba Szepesvari. Linear Multi-Resource Allocation with Semi-Bandit Feedback. In *Advances in Neural Information Processing Systems 28*, pages 964–972. Curran Associates, Inc., 2015.

Annie Marsden and Sergio Bacallado. Sequential Matrix Completion. *arXiv:1710.08045 [cs, stat]*, October 2017.

Y. Prokhorov. Convergence of Random Processes and Limit Theorems in Probability Theory. *Theory of Probability & Its Applications*, 1(2):157–214, January 1956.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*, volume 1. 2006.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling. In *Advances in Neural Information Processing Systems 27*, pages 1583–1591. 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Niranjan Srinivas, Andreas Krause, Matthias Seeger, and Sham M Kakade. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933.

Zi Wang and Stefanie Jegelka. Max-value Entropy Search for Efficient Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3627–3635. PMLR, August 2017.

Appendix A. Concentration Inequalities for Martingales

We first state the following consequence of Freedman's inequality.

Lemma 10 (Similar to Lemma 3 in (Kakade and Tewari, 2009)) *Let $X_1, \dots, X_T$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_t$ such that $X_t \leq b$ holds for all $t = 1, \dots, T$. Denote the corresponding martingale by $M_T = \sum_{t=1}^T X_t$ and the sum of conditional variances by $V_T = \sum_{t=1}^T \mathbb{E}[X_t^2 | \mathcal{F}_{t-1}]$. Then, for any $\beta > 1$, $l = \left\lceil \frac{\log(T\lambda^{-2})}{\log \beta} \right\rceil$, $\lambda \geq 0$,*

$$\mathbb{P}[M_T \geq \lambda \max\{\lambda b, \sqrt{V_T}\}] \leq (l+1) \exp\left(-\frac{\lambda^2}{2\beta + \frac{2}{3}}\right) ,$$

and consequently, with probability at least $1 - \delta$,

$$M_T \leq \max\left\{4b \log \frac{2T+2}{\delta}, 2\sqrt{V_T \log \frac{2T+2}{\delta}}\right\} \tag{6}$$

Proof Our proof uses the same ideas as in the proof of Lemma 3 of Kakade and Tewari (2009) and is included for completeness; along the way we also get rid of the assumptions $\delta < 1/e$ and $T \geq 3$ used in the original lemma. Define $l = \left\lceil \frac{\log T\lambda^{-2}}{\log \beta} \right\rceil$, and $\alpha_i = \lambda^2 b^2 \beta^i$ for $i = 0, \dots, l$, further set $\alpha_{-1} = 0$ for notational convenience. Note that since $V_T \leq Tb^2$, our choice of l implies $\alpha_l \geq V_T$. Then,

$$\begin{aligned}
&\mathbb{P}[M_T \geq \lambda \max\{\lambda b, \sqrt{V_T}\}] \\
&= \mathbb{P}[M_T \geq \lambda \max\{\lambda b, \sqrt{V_T}\},\ \alpha_{i-1} \leq V_T \leq \alpha_i \text{ for } i = 0, \dots, l] \\
&\leq \sum_{i=0}^{l} \mathbb{P}[M_T \geq \lambda \max\{\lambda b, \sqrt{V_T}\},\ \alpha_{i-1} \leq V_T \leq \alpha_i] \\
&\leq \sum_{i=0}^{l} \mathbb{P}[M_T \geq \lambda \max\{\lambda b, \sqrt{\alpha_{i-1}}\},\ V_T \leq \alpha_i] \\
&\overset{\text{(i)}}{\leq} \sum_{i=0}^{l} \exp\left(-\frac{\lambda^2 \max\{\lambda^2 b^2, \alpha_{i-1}\}}{2\alpha_i + \frac{2}{3} b\lambda \max\{\lambda b, \sqrt{\alpha_{i-1}}\}}\right) \\
&= \exp\left(-\frac{\lambda^2}{2 + \frac{2}{3}}\right) + \sum_{i=1}^{l} \exp\left(-\frac{\lambda^2}{2\beta + \frac{2}{3}\sqrt{\frac{1}{\beta^{i-1}}}}\right) \\
&\overset{\text{(ii)}}{\leq} (l+1) \exp\left(-\frac{\lambda^2}{2\beta + \frac{2}{3}}\right) .
\end{aligned}$$

Here, (i) is Freedman's inequality (Freedman, 1975), and (ii) uses that $\beta > 1$. This shows the first part of the lemma. For the second part, note that for $T = 1$, the claim is trivially true by $X_1 \leq b$, hence we assume $T \geq 2$. If also $\lambda \geq 1$, we can upper bound $l \leq \left\lceil \frac{\log T}{\log \beta} \right\rceil =: l'$. We choose $\beta = 5/3$, and set $\lambda = \sqrt{\log \frac{l'+1}{\delta}(2\beta + \frac{2}{3})} = 2\sqrt{\log \frac{l'+1}{\delta}}$, such that indeed $\lambda \geq 2\sqrt{\log\left(\frac{\log(2)}{\log(5/3)} + 1\right)} \geq 1$ for $T \geq 2$, hence proving the claim. □

The next concentration inequality for supermartingales is an anytime variant of Corollary 2.7 by Fan et al. (2015), and might be of independent interest.

Theorem 11 *Let $S_T = \sum_{t=1}^T \xi_t$ be a sum of supermartingale differences ξ_t with filtration $\{\mathcal{F}_t\}$. Further, let U_t be a non-negative predictable process, such that $\xi_t \leq U_t$ holds for all $t \geq 1$. Define*

$$C_t^2 = \begin{cases} \mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}], & \textit{if } \mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] \geq U_t^2 \ , \\ \frac{1}{4}\left(U_t + \frac{\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}]}{U_t}\right)^2 & \textit{otherwise.} \end{cases}$$

Further denote $A_T = \sum_{t=1}^T C_t^2$. Then, for any fixed positive sequence $(l_t)_{t=1}^T$, with probability at least $1-\delta$,

$$\forall\, T \geq 1, \quad S_T \leq \sqrt{2(A_T + l_T)\log\left(\frac{1}{\delta}\frac{(A_T + l_T)^{1/2}}{l_T^{1/2}}\right)} \, .$$

The proof combines Corollary 2.7 in (Fan et al., 2015), with the *method of mixtures* as used in the proof of Theorem 1 in (Abbasi-Yadkori et al., 2011). We start with the following lemma.

Lemma 12 *Let ξ_t, C_t as in Theorem 11, and define for $\lambda \geq 0$, $t \geq 1$,*

$$M_t^\lambda = \exp\left(\sum_{s=1}^t \lambda \xi_s - \frac{\lambda^2}{2} C_s^2\right) \, . \tag{7}$$

Further, let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}$. Then M_t^λ is a supermartingale, M_τ^λ is almost surely well-defined, and $\mathbb{E}[M_\tau^\lambda] \leq 1$.

Proof The proof is along the lines of Lemma 8 in (Abbasi-Yadkori et al., 2011), where we replace the subgaussian condition by the suitable analog to showing that M_t^λ is a supermartingale. Let

$$D_s^\lambda = \exp\left(\lambda \xi_s - \frac{\lambda}{2} C_s^2\right) \, .$$

By Corollary 2.6 of Fan et al. (2015), we have that $\mathbb{E}[e^{\lambda \xi_s}|\mathcal{F}_{s-1}] \leq \exp\left(\frac{\lambda^2}{2} C_s^2\right)$ for all $\lambda > 0$, and consequently, $\mathbb{E}[D_s|\mathcal{F}_{s-1}] \leq 1$. Therefore, $\mathbb{E}[M_t^\lambda|\mathcal{F}_{t-1}] = \mathbb{E}[D_t^\lambda|\mathcal{F}_{t-1}]M_{t-1}^\lambda \leq M_{t-1}^\lambda$, which shows that M_t^λ is a supermartingale such that $\mathbb{E}[M_t^\lambda] \leq 1$ for all $t \geq 1$. The rest of the argument is standard (compare for example Theorem 5.7.6 in Durrett, 2010); by the convergence theorem for non-negative supermartingales, M_τ^λ is well defined for any stopping time $\tau \leq \infty$, then using Fatou's lemma it follows that $M_\tau^\lambda \leq \liminf_{t\to\infty} M_{\tau \wedge t}^\lambda \leq 1$. □

To prove Theorem 11, we now use the *method of mixtures*, similar to Theorem 1 in (Abbasi-Yadkori et al., 2011). The main difference is that the supermartingale M_t^λ from the the previous lemma is only defined for $\lambda \geq 0$, which requires to choose a mixing density supported on $[0, \infty)$.

Proof of Theorem 11 Remember that $S_t = \sum_{s=1}^t \xi_s$, $A_t = \sum_{s=1}^t C_s^2$ and $M_t^\lambda = \exp\left(\lambda S_t - \frac{\lambda^2}{2} A_t\right)$. Further, let $\Lambda = (\Lambda_t)_{t\geq 1}$ be a sequence of independent Gaussian random variable truncated to $\mathbb{R}^+ = [0, \infty)$ with densities $f_{\Lambda_t}(\lambda) = c(l_t) \exp\left(-\frac{1}{2}\lambda^2 l_t\right) \mathbb{1}\{\lambda \geq 0\}$ where $c(A) = \sqrt{\frac{2A}{\pi}}$ is a normalizing constant. Using Λ as a *mixing distribution* we define

$$M_t = \mathbb{E}[M_t^{\Lambda_t} | \mathcal{F}_\infty] \,, \tag{8}$$

where $\mathcal{F}_\infty = \sigma\left(\cup_{t=1}^\infty \mathcal{F}_t\right)$ is the tail σ-algebra of the filtration $\mathcal{F}_t$. In particular, using Fubini's theorem, we still get $\mathbb{E}[M_\tau] = \mathbb{E}[\mathbb{E}[M_\tau^{\Lambda_t} | \Lambda]] \leq 1$. In the next step, we explicitly calculate M_t for any $t \geq 1$,

$$\begin{aligned}
M_t &= \int_{\mathbb{R}^+} \exp\left(\lambda S_t - \tfrac{\lambda^2}{2} A_t\right) f_{\Lambda_t}(\lambda)\, d\lambda \\
&= \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left(\lambda - \frac{S_t}{A_t}\right)^2 A_t + \frac{1}{2}\frac{S_t^2}{A_t}\right) f_{\Lambda_t}(\lambda)\, d\lambda \\
&= \exp\left(\frac{1}{2}\frac{S_t^2}{A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left(\lambda - \frac{S_t}{A_t}\right)^2 A_t\right) f_{\Lambda_t}(\lambda)\, d\lambda \\
&= c(l_t) \exp\left(\frac{1}{2}\frac{S_t^2}{A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left((\lambda - S_t/A_t)^2 A_t + \lambda^2 l_t\right)\right)\, d\lambda \,.
\end{aligned}$$

Completing the square yields

$$\left(\lambda - \frac{S_t}{A_t}\right)^2 A_t + \lambda^2 A = \left(\lambda - \frac{S_t}{l_t + A_t}\right)^2 (l_t + A_t) + \frac{S_t^2}{A_t} - \frac{S_t^2}{l_t + A_t} \,,$$

and with the previous equation,

$$\begin{aligned}
M_t &= c(l_t) \exp\left(\frac{1}{2}\frac{S_t^2}{l_t + A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left((\lambda - S_t/(l_t + A_t))^2 (l_t + A_t)\right)\right)\, d\lambda \\
&\overset{\text{(i)}}{\geq} \mathbb{1}\{S_t \geq 0\} c(l_t) \exp\left(\frac{1}{2}\frac{S_t^2}{l_t + A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left((\lambda - S_t/(l_t + A_t))^2 (l_t + A_t)\right)\right)\, d\lambda \\
&\overset{\text{(ii)}}{\geq} \mathbb{1}\{S_t \geq 0\} c(l_t) \exp\left(\frac{1}{2}\frac{S_t^2}{l_t + A_t}\right) \int_{\mathbb{R}^+} \exp\left(-\frac{1}{2}\left(\lambda^2 (l_t + A_t)\right)\right)\, d\lambda \\
&= \mathbb{1}\{S_t \geq 0\} \frac{c(l_t)}{c(l_t + A_t)} \exp\left(\frac{1}{2}\frac{S_t^2}{l_t + A_t}\right) \,.
\end{aligned}$$

In (i) we introduced an indicator function and used that all other terms are positive. To get (ii), we first applied a change of variables $\lambda' = \lambda - S_t/(l_t + A_t)$ and then made use of $S_t \geq 0$ to reduce the integration range (and again that the integrand is positive).

A final application of Markov's inequality yields

$$\begin{aligned}
&\mathbb{P}\left[S_\tau \geq \sqrt{2(l_\tau + A_\tau)\log\left(\frac{1}{\delta}\frac{(l_\tau + A_\tau)^{1/2}}{l_\tau^{1/2}}\right)}\right] \\
&= \mathbb{P}\left[\frac{c(l_\tau)}{c(l_\tau + A_\tau)}\exp\left(\frac{1}{2}\frac{S_\tau^2}{l_\tau + A_\tau}\right) \geq \frac{1}{\delta}, S_\tau \geq 0\right] \\
&\leq \delta \cdot \mathbb{E}\left[\mathbb{1}\{S_\tau \geq 0\}\frac{c(l_\tau)}{c(l_\tau + A_\tau)}\exp\left(\frac{1}{2}\frac{S_\tau^2}{l_\tau + A_\tau}\right)\right] \\
&\overset{\text{(i)}}{\leq} \delta \cdot \mathbb{E}[M_\tau] \overset{\text{(ii)}}{\leq} \delta\,,
\end{aligned}$$

where (i) uses the inequality for M_t derived above, and (ii) follows from Lemma 12.

To get the anytime result as stated in the Theorem, we use the same argument as in (Abbasi-Yadkori et al., 2011) on the stopping time

$$\tau = \min\left\{t \geq 1 \mid S_t \geq \sqrt{2(l_t + A_t)\log\left(\frac{1}{\delta}\frac{(l_t + A_t)^{1/2}}{l_t^{1/2}}\right)}\right\}.$$

Expressing the quantity of interest in terms of τ, and applying the previous inequality yields

$$\begin{aligned}
&\mathbb{P}\left[S_t \geq \sqrt{2(l_t + A_t)\log\left(\frac{1}{\delta}\frac{(l_t + A_t)^{1/2}}{l_t^{1/2}}\right)} \text{ for any } t \geq 1\right] \\
&= \mathbb{P}[\tau < \infty] \\
&= \mathbb{P}\left[\tau < \infty,\, S_\tau \geq \sqrt{2(l_\tau + A_\tau)\log\left(\frac{1}{\delta}\frac{(l_\tau + A_\tau)^{1/2}}{l_\tau^{1/2}}\right)}\right] \\
&\leq \mathbb{P}\left[S_\tau \geq \sqrt{2(l_\tau + A_\tau)\log\left(\frac{1}{\delta}\frac{(l_\tau + A_\tau)^{1/2}}{l_\tau^{1/2}}\right)}\right] \\
&\leq \delta\,.
\end{aligned}$$

This completes the proof. □

As a consequence of the previous result, we have the following lemma.

Lemma 3 *(in Section 3) Let X_t be a non-negative stochastic process adapted to a filtration $\{\mathcal{F}_t\}$. Assume that $X_t \leq b_t$ for a fixed, non-decreasing sequence $\{b_t\}$. Define $m_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}]$ and $M_T = \sum_{t=1}^T m_t$, and let $(l_t)_{t\geq 1}$ be any fixed, positive sequence. Then, with probability at least $1-\delta$,*

$$\forall\, T \in \mathbb{N}, \quad \sum_{t=1}^T m_t - X_t \leq \sqrt{2(b_T M_T + l_T)\log\left(\frac{1}{\delta}\frac{(b_T M_T + l_T)^{1/2}}{l_T^{1/2}}\right)}.$$

Further, if $b_T \geq 1$, with probability at least $1-\delta$ for any $T \geq 1$ it holds that,

$$\sum_{t=1}^T m_t \leq 2\sum_{t=1}^T X_t + 4b_T \log\frac{1}{\delta} + 8b_T\log(4b_T) + 1\,.$$

Proof Clearly, $\xi_t = m_t - X_t$ is a martingale difference sequence such that $\xi_t \leq m_t$, and m_t is a predictable process. This allows us to apply Theorem 11. Remember that

$$C_t^2 = \begin{cases} \mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}], & \text{if } \mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] \geq m_t^2\,, \\ \frac{1}{4}\left(m_t + \frac{\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}]}{m_t}\right)^2 & \text{otherwise,} \end{cases}$$

and in particular $C_t^2 \leq b_t m_t$. To see this, note that the Bhatia-Davis inequality (Lemma 18, Appendix G) implies $\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] = \mathrm{Var}(X_t|\mathcal{F}_{t-1}) \leq m_t(b_t - m_t) \leq m_t b_t$, and if $\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] \leq m_t^2$, and further, if $m_t \leq b_t$ hold, $\frac{1}{4}\left(m_t + \frac{\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}]}{m_t}\right)^2 \leq m_t^2 \leq m_t b_t$. Consequently, $A_T = \sum_{t=1}^T C_t^2 \leq b_T \sum_{t=1}^T m_t = b_T M_T$. A direct application of Theorem 11 shows the first inequality.

For the second part, we use a trade-off parameter $\eta > 1$, and set $l_t = b_t$ to find

$$\begin{aligned} &\sqrt{2(b_T M_T + b_T)\log\left(\frac{1}{\delta}\frac{(b_T M_T + b_T)^{1/2}}{b_T^{1/2}}\right)} \\ &\overset{\text{(i)}}{\leq} \sqrt{2b_T(M_T+1)\log\left(\frac{1}{\delta}\right)} + \sqrt{b_T(M_T+1)\log(M_T+1)} \\ &\overset{\text{(ii)}}{\leq} \frac{M_T}{2\eta} + \frac{1}{2\eta} + \eta b_T \log\left(\frac{1}{\delta}\right) + \frac{M_T}{2\eta} + \frac{1}{2\eta} + 2\eta b_T\log(2\eta b_T) \end{aligned}$$

where for (i) we used the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and for (ii) we used $2\sqrt{ab} \leq a+b$ (Lemma 19) on the first term, and $\sqrt{ab\log(ab)} \leq a + b\log(b)$ (Lemma 20) on the second term with the condition $b_T \geq 1$. We set $\eta = 2$ and combine the previous inequalities. Solving for $M_T = \sum_{t=1}^T m_t$ yields

$$\sum_{t=1}^T m_t \leq 2\sum_{t=1}^T X_t + 4b_T\log\left(\frac{1}{\delta}\right) + 8b_T\log(4b_T) + 1\,.$$

This completes the proof. □

Appendix B. Proof of Theorem 1

We will write $\Delta_t = \Delta(x_t)$ and $I_t = I_t(x_t)$ for notational convenience.

Step 1) Bounding the regret by the expected regret. We first use Lemma 10 to bound the regret in terms of the sum of expected instantaneous regret $\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]$.

Lemma 13 *For any policy π we have with probability at least $1-\delta$,*

$$R_T \leq \tfrac{5}{4}\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] + 4S\log\left(\tfrac{2\log(T)+2}{\delta}\right) .$$

Further, with probability at least $1-\delta$, at any time $T \geq 1$,

$$R_T \leq \tfrac{5}{4}\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] + 4S\log\left(\tfrac{4\pi^2T^2}{3\delta}(\log(T)+1)\right) .$$

Proof First, we rewrite the regret

$$R_T = \sum_{t=1}^T \Delta_t = \sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] + \sum_{t=1}^T (\Delta_t - \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]) \ ,$$

such that we only need to bound the martingale difference sequence $D_t = \Delta_t - \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]$. Note that by assumption $0 \leq \Delta_t \leq S$, hence, $D_t \leq S$. Further, by the Bhatia-Davis inequality,

$$\mathrm{Var}[D_t|\mathcal{F}_{t-1}] = \mathrm{Var}[\Delta_t|\mathcal{F}_{t-1}] \leq \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}](S - \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]) \ ,$$

and consequently,

$$V_T = \sum_{t=1}^T \mathrm{Var}[D_t|\mathcal{F}_{t-1}] \leq S\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] \, .$$

Now we apply Lemma 10 to find that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^T D_t &\leq \max\left\{4S\log\tfrac{2\log(T)+2}{\delta}, 2\sqrt{V_T\log\tfrac{2\log(T)+2}{\delta}}\right\}\\ &\leq \max\left\{4S\log\tfrac{2\log(T)+2}{\delta}, 2\sqrt{S\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}]\log\tfrac{2\log(T)+2}{\delta}}\right\}\\ &\leq \frac{1}{4}\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] + 4S\log\tfrac{2\log(T)+2}{\delta}\end{aligned}$$

where for the last inequality we have used $2\sqrt{ab} \leq a+b$ for $a,b \geq 0$. This proves the first part of the lemma. Taking the union bound over all time steps $T \geq 1$, effectively replacing δ by $\frac{6\delta}{\pi^2T^2}$, gives the anytime bound. □

Step 2) Application of Cauchy-Schwarz. In the next lemma, we bound the sum of expected instantaneous regret using the information-regret ratio and the expected information gain.

Lemma 14

$$\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_t] = \sum_{t=1}^T \sqrt{\Psi_t}\sqrt{\mathbb{E}[I_t|\mathcal{F}_{t-1}]} \leq \sqrt{\sum_{t=1}^T \Psi_t \cdot \sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}]} \, .$$

Proof First use the definition of Ψ_t and then apply Cauchy-Schwarz. □

Step 3) Bounding the expected information gain. From the last step, we are left with the sum of expected information gains, $\sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}]$. Remember that $\sum_{t=1}^T I_t \leq \gamma_T$. Using Lemma 3, we show that the sum of expected information gains $\sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}]$ is similarly bounded.

Lemma 15 *With probability at least $1-\delta$,*

$$\forall\, T \in \mathbb{N}, \quad \sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}] \leq 2\gamma_T + 4\gamma_T \log\frac{1}{\delta} + 8\gamma_T\log(4\gamma_T) + 1\,,$$

and if further $I_t \leq 1$ holds, then, with probability at least $1-\delta$,

$$\forall\, T \in \mathbb{N}, \quad \sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}] \leq 2\gamma_T + 4\log\frac{1}{\delta} + 8\log(4) + 1\,.$$

Proof of Lemma 15. Remember that $\sum_{t=1}^T I_t \leq \gamma_T$ by definition. The proof of the lemma is a direct application of Lemma 3, first with the conservative bound $I_t \leq \sum_{s=1}^t I_s \leq \gamma_t$, and then with the assumption $I_t \leq 1$. □

Step 4) Completing the proof. For the proof of the theorem, we take the union bound over the events such that Lemma 13 and Lemma 15 simultaneously hold, effectively replacing δ by $\frac{\delta}{2}$ in respective statements. Applying Lemma 13, 14 and 15 in order, we find

$$\begin{aligned}R_T &\leq \tfrac{5}{4}\sum_{t=1}^T \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] + 4S\log\left(\tfrac{8\pi^2T^2}{3\delta}(\log(T)+1)\right)\\ &\leq \tfrac{5}{4}\sqrt{\sum_{t=1}^T \Psi_t \sum_{t=1}^T \mathbb{E}[I_t|\mathcal{F}_{t-1}]} + 4S\log\left(\tfrac{8\pi^2T^2}{3\delta}(\log(T)+1)\right)\\ &\leq \tfrac{5}{4}\sqrt{\sum_{t=1}^T \Psi_t \left(2\gamma_T + 4\gamma_T\log\frac{2}{\delta} + 8\gamma_T\log(4\gamma_T)+1\right)} + 4S\log\left(\tfrac{8\pi^2T^2}{3\delta}(\log(T)+1)\right)\,,\end{aligned}$$

and if $I_t \leq 1$ holds, the last step can be strengthend to

$$R_T \leq \tfrac{5}{4}\sqrt{\sum_{t=1}^T \Psi_t \left(2\gamma_T + 4\log\frac{2}{\delta} + 8\log(4)+1\right)} + 4S\log\left(\tfrac{8\pi^2T^2}{3\delta}(\log(T)+1)\right)\,.$$

This completes the proof. □

Dropping all constants, we find

$$R_T \leq \mathcal{O}\left(\sqrt{\sum_{t=1}^T \Psi_t \left(\gamma_T\log\frac{1}{\delta} + \gamma_T\log(\gamma_T)\right)} + S\log\left(\tfrac{T\log(T)}{\delta}\right)\right)\,,$$

and if $I_t \leq 1$ holds

$$R_T \leq \mathcal{O}\left(\sqrt{\sum_{t=1}^{T} \Psi_t \, \left(\gamma_T + \log\frac{1}{\delta}\right)} + S \log\left(\tfrac{T\log(T)}{\delta}\right)\right) .$$

We remark that in Step 3) in our proof, a dependence on $\log\frac{1}{\delta}$ is unavoidable, as shown by the following example.

Example 2 *Define a stochastic process I_t, such that $I_t \sim Ber(\frac{1}{2})$, if $I_1 = 0, \dots, I_{t-1} = 0$, and $I_t = 0$ else. Then $\sum_{t=1}^{T} \mathbb{E}_\mu[I_t|\mathcal{F}_{t-1}] = \Omega(T)$ with positive probability, but $\sum_{t=1}^{T} I_t \leq 1$ always holds.*

Appendix C. Properties of the Information Regret Ratio

Here we present details on the properties of the regret-information ratio and its surrogate. In the next lemmas, we show that a minimizer of $\Psi_t^+(\mu)$ always exists, and in fact can be chosen to be supported on two actions, and further, that the regret-information ratio is a convex function of $\mu \in \mathcal{P}(\mathcal{X})$. Formally, we assume that $\mathcal{X}$ is a compact, metric space and $\Delta : \mathcal{X} \to \mathbb{R}_{\geq 0}$, $g : \mathcal{X} \to \mathbb{R}_{\geq 0}$ are continuous, bounded functions, such that g is not zero everywhere. In the following, we let $\Psi : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be any function of the form $\Psi(\mu) = \frac{\mathbb{E}_\mu[\Delta(x)]^2}{\mathbb{E}_\mu[g(x)]}$, which includes both the regret-information ratio and its surrogate. Note however, that minimizing the actual regret-information ratio is a trivial matter (assuming that Δ is known), as for $x^* \in \arg\max_{x\in\mathcal{X}} f(x)$, we have by definition $\Delta(x^*) = 0$, and therefore $\min_{x\in\mathcal{X}} \Psi_t(x) = 0$; but this is of course not true in general for the surrogate Ψ_t^+.

Lemma 16 *There exists a $\mu^* \in \mathcal{P}(\mathcal{X})$ such that $\Psi(\mu^*) = \inf_{\mu\in\mathcal{P}(\mathcal{X})} \Psi(\mu)$.*

Proof Clearly $\Psi(\mu) \geq 0$, and if there is a point $x \in \mathcal{X}$ with $\Delta(x) = 0$, we have $\Psi(\mu^*) = 0$ for $\mu^* = \delta_x$, with the convention that $\frac{0}{0} = 0$. Therefore, without loss of generality, we can assume that $\Delta(x) \geq \epsilon > 0$ for all $x \in \mathcal{X}$.

Define $h : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (x, y) \mapsto x^2/y$, such that $\Psi(\mu) = h(\mathbb{E}_\mu[\Delta], \mathbb{E}_\mu[g])$. Denote by $\Psi^* = \inf_{\mu\in\mathcal{P}(\mathcal{X})} \Psi(\mu)$ and pick a sequence μ_n such that $\Psi(\mu_n) \to \Psi^*$ as $n \to \infty$. By the assumption that $\Delta \geq \epsilon$ and g is not zero everywhere, $\Psi^* < \infty$, and in particular $\mathbb{E}_{\mu_n}[g] > 0$ for n large enough. Since μ_n is a probability distribution over a compact, metric space, the sequence is trivially *tight*. We apply Prokhorov's theorem (Prokhorov, 1956), which guarantees the existence of a subsequence μ_{n_i} converging weakly to some $\mu^* \in \mathcal{P}(\mathcal{X})$ (in fact, $\mathcal{P}(\mathcal{X})$ is compact in the weak* topology). By definition of weak convergence (or weak* convergence from the point of view of functional analysis), $\int s(x)\, d\mu_{n_i} \to \int s(x) d\mu$ for every continuous bounded function $s : \mathcal{X} \to \mathbb{R}$. Hence $\mathbb{E}_{\mu_{n_i}}[\Delta] \to \mathbb{E}_{\mu^*}[\Delta]$, and the same for $\mathbb{E}_{\mu_{n_i}}[g] \to \mathbb{E}_{\mu^*}[g]$. Finally, continuity of h gives

$$\Psi^* = \lim_{i\to\infty} h(\mathbb{E}_{\mu_{n_i}}[\Delta], \mathbb{E}_{\mu_{n_i}}[g]) = h(\mathbb{E}_{\mu^*}[\Delta], \mathbb{E}_{\mu^*}[g]) = \Psi(\mu^*) ,$$

which proves the claim. □

In the next two lemmas, we extend Proposition 6 of Russo and Van Roy (2014) to the case of compact action sets $\mathcal{X}$.

Lemma 17 *$\Psi(\mu)$ is convex in μ.*

Proof The proof is completely analogs to the proof of Proposition 6 of Russo and Van Roy (2014). Define $h : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (x, y) \mapsto x^2/y$. As shown in Chapter 3 of Boyd and Vandenberghe (2004), h is convex on its domain. Further, $f : \mathcal{P}(X) \to \mathbb{R}^2, \mu \mapsto (\mathbb{E}_\mu[\Delta], \mathbb{E}_\mu[g])$ is an affine function. Since $\Psi(\mu) = h(f(\mu))$ is a composition of a convex and an affine function, Ψ is convex, too. □

Lemma 4 (in Section 4) *There exists a measure $\mu^* \in \arg\min_{\mu\in\mathcal{P}(\mathcal{X})} \Psi(\mu)$ with $|supp(\mu^*)| \leq 2$, i.e. a measure which achieves the minimum value of Ψ and is supported on at most two actions.*

Proof The proof is an extension of Proposition 6 in (Russo and Van Roy, 2014), where we deal with the technicalities of having a compact action set.

Define $\Psi^* = \min_{\mu\in\mathcal{P}(\mathcal{X})} \Psi(\mu)$. Again, without loss of generality, we can assume that $\Delta(x) \geq \epsilon > 0$ for all $x \in \mathcal{X}$, as otherwise we immediately get a measure supported on a single action, such that $\Psi^* = 0$. In particular, for what follows, we can assume that $\Psi^* > 0$. Define $\eta(\mu) = \mathbb{E}_\mu[\Delta]^2 - \mathbb{E}_\mu[g]\Psi^*$. As a first step, we show that η has the same minimizers as Ψ. Note that $\eta(\mu) \geq 0$, and if $\Psi(\mu) = \Psi^*$, then $\eta(\mu) = 0$, which shows one direction. For the converse, assume that μ minimizes η, i.e. $\eta(\mu) = 0$. This immediately gives $\Psi(\mu) = \Psi^*$.

Now, let μ be any minimizer of Ψ, and therefore also a minimizer of η. Remember that by definition of the support, for any $x \in \text{supp}(\mu)$ and any neighborhood N_x of x, $\mu(N_x) > 0$. We claim that the function

$$s : \mathcal{X} \to \mathbb{R}, \quad s(x) = 2\, \mathbb{E}_\mu[\Delta]\Delta(x) - \Psi^* g(x)$$

is constant on $\text{supp}(\mu)$. Suppose there exist $x, y \in \text{supp}(\mu)$ such that $s(x) < c < s(y)$ for some $c \in \mathbb{R}$. Then, by continuity of s, we can find disjoint neighborhoods N_x and N_y of x and y respectively, such that for all $x' \in N_x$ and $y' \in N_y$, $s(x') < c < s(y')$.

Let μ_x, μ_y and $\mu_\perp$ be the measures obtained by restricting μ to N_x, N_y and $(N_x \cup N_y)^c$ respectively, and let $\alpha_x = \mu_x(N_x)$ and $\alpha_y = \mu_y(N_y)$. By the definition of the support, $\alpha_x > 0$ and $\alpha_y > 0$ (in particular, μ_x/α_x and μ_y/α_y are the conditional probability distributions of μ given N_x and N_y respectively). Note that for our choice of x and y,

$$\frac{1}{\alpha_x}\int_{N_x} s(x)\, d\mu_x < c < \frac{1}{\alpha_y}\int_{N_y} s(y)\, d\mu_y\,. \tag{9}$$

Define for $\lambda \in [-\alpha_x^{-1}, \alpha_y^{-1}]$ a parametrized measure

$$\mu_\lambda = \mu_\perp + (1 - \lambda\alpha_y)\mu_x + (1 + \lambda\alpha_x)\mu_y\,.$$

Clearly $\mu_\lambda(\mathcal{X}) = 1$, hence μ_λ defines a probability measure for $-\alpha_x^{-1} \leq \lambda \leq \alpha_y^{-1}$ such that $\mu_0 = \mu$. Intuitively, by varying λ we transfer probability mass from N_x to N_y. With a slight abuse of notation, we write $\eta(\lambda) = \eta(\mu_\lambda)$. Because μ is a minimizer of η, it must be that $\frac{d}{d\lambda}\eta(\lambda)|_{\lambda=0} = 0$. We have that

$$\begin{aligned}\eta(\lambda) &= \mathbb{E}_{\mu_\lambda}[\Delta]^2 - \Psi^*\, \mathbb{E}_{\mu_\lambda}[g]\\ &= \mathbb{E}_{\mu_\perp+(1-\lambda\alpha_y)\mu_x+(1+\lambda\alpha_x)\mu_y}[\Delta]^2 - \Psi^*\, \mathbb{E}_{\mu_\perp+(1-t\alpha_y)\mu_x+(1+\lambda\alpha_x)\mu_y}[g]\,,\end{aligned}$$

and we calculate the derivative with respect to λ,

$$\frac{d}{d\lambda}\eta(\lambda) = 2\mathbb{E}_{\mu_\lambda}[\Delta] \int_{\mathcal{X}} \Delta \; (\alpha_x d\mu_y - \alpha_y d\mu_x) + \Psi^* \int_{\mathcal{X}} g \; (\alpha_x d\mu_y - \alpha_y \mu_x) \; .$$

Evaluating the previous expression at $\lambda = 0$ yields

$$0 = 2\mathbb{E}_{\mu}[\Delta] \int_{\mathcal{X}} \Delta \; (\alpha_x d\mu_y - \alpha_y d\mu_x) + \Psi^* \int_{\mathcal{X}} g \; (\alpha_x d\mu_y - \alpha_y d\mu_x),$$

which contradicts (9). This shows that $d^* := s(x)$ is constant for $x \in \text{supp}(\mu)$, and in particular for all $x \in \text{supp}(\mu)$,

$$g(x) = -\frac{d^*}{\Psi^*} + \frac{2\mathbb{E}_\mu[\Delta]}{\Psi^*}\Delta(x) \; .$$

To complete the proof, note that supp(μ) is compact because the support is always closed and $\mathcal{X}$ compact; hence we can pick $x_{\min} \in \arg\min_{x \in \text{supp}(\mu)} g(x)$ and $x_{\max} \in \arg\max_{x \in \text{supp}(\mu)} g(x)$. Let $0 \leq \beta \leq 1$, such that

$$\mathbb{E}_\mu[g] = \beta g(x_{\min}) + (1-\beta) g(x_{\max}) \; .$$

Finally, $\mu^* = \beta\delta_{x_{\min}} + (1-\beta)\delta_{x_{\max}}$ is a distribution supported on $\{x_{\min}, x_{\max}\}$. From the last two equations, it follows that $\mathbb{E}_\mu[g] = \mathbb{E}_{\mu^*}[g]$, and also $\mathbb{E}_\mu[\Delta] = \mathbb{E}_{\mu^*}[\Delta]$. Consequently, μ^* achieves the same objective value on η as μ, and therefore is a minimizer of both η and Ψ^*. □

Appendix D. Least Squares Estimation with Heteroscedastic Noise

D.1. Least Squares Estimator for RKHS

Note that for any linear operator $A : \mathcal{H} \to \mathbb{R}^T$,

$$(A^*A + \lambda I_{\mathcal{H}})^{-1}A^* = A^*(AA^* + \lambda I_{\mathbb{R}^T})^{-1} \; .$$

Remember that we defined $(K_T)_{i,j} := (MM^*)_{i,j} = k(x_i, x_j)$ and $k_T(x)_i := (Mk_x)_i = \langle x_i, x\rangle = k(x_i, x)$. Using the reproducing property and the equation above with $A = \Sigma^{-1}M$, we calculate $\hat{\mu}_T(x)$ for any $x \in \mathbb{R}^d$,

$$\begin{aligned}
\hat{\mu}_T(x) &= \langle \hat{\mu}_T, k_x\rangle_{\mathcal{H}} \\
&= \langle (M^*\Sigma^{-1}M + \lambda I_{\mathcal{H}})^{-1}M^*\Sigma^{-1}y, k_x\rangle_{\mathcal{H}} \\
&= \langle M^*\Sigma^{-1/2}(\Sigma^{-1/2}MM^*\Sigma^{-1/2} + \lambda I_{\mathbb{R}^n})^{-1}\Sigma^{-1/2}y, k_x\rangle_{\mathcal{H}} \\
&= \langle M^*(MM^* + \lambda\Sigma)^{-1}y, k_x\rangle_{\mathcal{H}} \\
&= \langle (MM^* + \lambda\Sigma)^{-1}y, Mk_x\rangle_{\mathbb{R}^T} \\
&= k_T(x)^\top (K_T + \lambda\Sigma)^{-1}y \; .
\end{aligned}$$

Further, observe that for $A = \Sigma^{-1/2}M$,

$$\lambda(A^*A + \lambda I)^{-1} = I - A^*(AA^* + \lambda I)^{-1}A \; .$$

In particular, we have $\lambda \langle k_x, k_y \rangle_{V_T^{-1}} = k(x,y) - k_T(x)^\top (K_T + \lambda \Sigma_T)^{-1} k_T(y)$, hence using $v = k_x = k_y$, we can calculate $\|v\|_{V_T^{-1}}^2$ as follows,

$$\|v\|_{V_T^{-1}}^2 = \frac{1}{\lambda}\Big(k(x,x) - k_T(x)^\top (K + \lambda\Sigma)^{-1} k_T(x)\Big) = \sigma_T(x)^2 \, .$$

Appendix E. Summary of Information Gain Functions

IDS-F In the main text, we have defined $I_t^{\mathrm{F}}(x) = \log\left(1 + \frac{\sigma_t(x)^2}{\rho(x)^2}\right)$, which in the Bayesian setting, with Gaussian prior and likelihood and posterior variance $\sigma_t(x)$, corresponds to the mutual information $\mathbb{I}(x; f|\mathcal{F}_{t-1})$.

IDS-UCB Remember that we defined $\sigma_t(y|x)^2$ as the confidence width at $y \in \mathcal{X}$ if we were to evaluate $x \in \mathcal{X}$ at time t. Then IDS-UCB is defined through

$$I_t^{\mathrm{UCB}}(x) = \log\left(\frac{\sigma_t(x_t^{\mathrm{UCB}})^2}{\sigma_t(x_t^{\mathrm{UCB}}|x)^2}\right) \, .$$

Using the Sherman-Morrison formula, we compute for the linear case, denoting $\sigma_t(x) = \|x\|_{V_t^{-1}}$,

$$\sigma_t(x_t^{\mathrm{UCB}}|x)^2 = \sigma(x_t^{\mathrm{UCB}})^2 - \frac{(x^\top V_t^{-1} x_t^{\mathrm{UCB}})^2}{\rho(x)^2 + \sigma_t(x)^2} \, .,$$

Using Schur's complement, for RKHS functions we get

$$\sigma_t(x_t^{\mathrm{UCB}}|x)^2 = \sigma(x_t^{\mathrm{UCB}})^2 - \frac{k_t(x, x_t^{\mathrm{UCB}})^2}{\rho(x)^2 + \sigma_t(x)^2} \, ,$$

compare Appendix F in (Chaudhuri et al., 2017). From the last expression, it is also easy to see that

$$I_t^{\mathrm{UCB}}(x_t^{\mathrm{UCB}}) = I_t^{\mathrm{F}}(x_t^{\mathrm{UCB}}) = \log\left(1 + \frac{\sigma(x_t^{\mathrm{UCB}})^2}{\rho(x_t^{\mathrm{UCB}})^2}\right) \, .$$

IDS-TS Similar as IDS-UCB, we define IDS-TS by using the information function

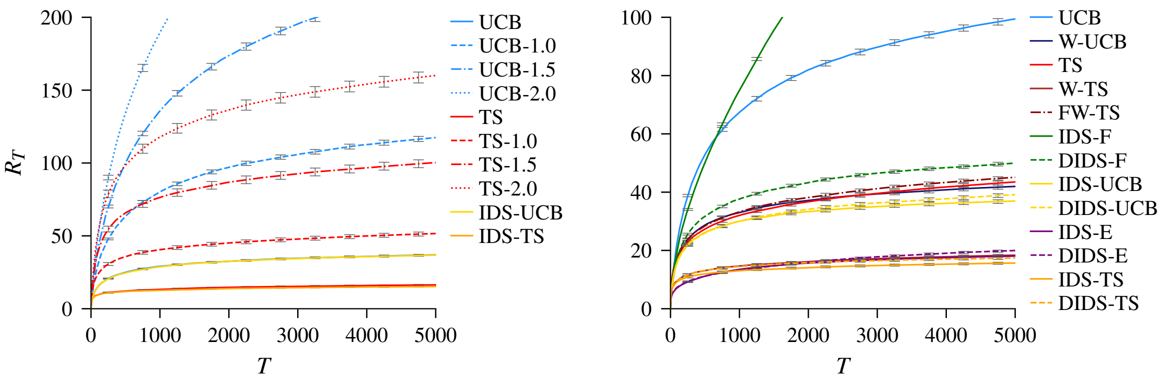
$$I_t^{\mathrm{TS}}(x) = \log\left(\frac{\sigma_t(x_t^{\mathrm{TS}})^2}{\sigma_t(x_t^{\mathrm{TS}}|x)^2}\right) \, ,$$

where x_t^{TS} is a *proposal action* by the Thompson Sampling policy, that is, x_t^{TS} maximizes a sample $\tilde{f}$ from the posterior distribution of f. Intuitively, we use Thompson Sampling to identify an action to gather information on, but the resulting IDS-TS is not restricted to playing x_t^{TS}. Though currently we do not have a bound for the regret-information ratio of this policy, we point out that if the regret-information ratio of Thompson Sampling is bounded, then by a similar argument as in Lemma 8, one can also bound the regret-information ratio of IDS-TS.

IDS-E We also propose the use of the averaged information gain

$$I_t^E(x) = \frac{1}{m}\sum_{i=1}^m \log\left(\frac{\sigma_t(x_{t,i}^{\mathrm{TS}})^2}{\sigma_t(x_{t,i}^{\mathrm{TS}}|x)^2}\right) \, ,$$

calculated using m proposals $x_{t,1}^{\mathrm{TS}}, \dots, x_{t,m}^{\mathrm{TS}}$ from the Thompson Sampling policy.



(a) Simulation of Example 1

(b) Heteroscedastic Noise

Figure 2: Left: We show a simulation of Example 1 with pairs of correlated actions. For each pair, one of the action yields low observation noise $\mathcal{N}(0, \sigma_0^2)$ with $\sigma_0 = 0.5$, whereas the second action has increased observation noise $\mathcal{N}(0, \sigma^2)$ for $\sigma \in \{1, 1.5, 2\}$. Without the second action UCB and IDS-UCB perform similarly, and the same holds true for TS/IDS-TS. Assuming that UCB/TS always commit to the noisier action among each pair, UCB-σ and TS-σ show the regret achieved given increased noise variance σ^2. As expected, the performance quickly deteriorates, because the confidence intervals concentrate at a slower rate. On the other hand, IDS-UCB/IDS-TS are guaranteed to always pick the less noisy action among each pair.
Right: We show another instance of our heteroscedastic experiment, including also further variants of IDS. We observed that the randomized versions of IDS-TS, IDS-E and IDS-UCB perform very similar to the respective deterministic DIDS-TS, DIDS-E and DIDS-UCB. Only for IDS-F and DIDS-F the performance is drastically different, where as noted before IDS-F is over-exploring. Again, one can see that W-UCB outperforms UCB, and W-TS outperforms TS, which shows the benefit of using the weighted least squares estimator in the heteroscedastic setting. In this case, W-TS is actually competitive with IDS-TS (which can hardly be seen due to overlapping lines). Finally, IDS-UCB clearly outperforms W-UCB, and IDS-TS achieves the lowest regret among all policies after 5000 steps.

Appendix F. Further Numerical Experiments

We simulated the setting of Example 1, and also evaluated the IDS variants proposed in Appendix E in the same basic setup as described in Section 7, but with a different sample of randomly generated actions (Figure 2a, 2b). Further, in the heteroscedastic case, we instead choose the noise bound for each action uniformly in $[0.4, 0.6]$, a setting closer to the homoscedastic case. We also included an additional variant of Thomson Sampling (FW-TS), which engages in the oversampling as needed for the frequentist guarantees of Agrawal and Goyal (2013) and Abeille and Lazaric (2017), in addition to using the weighted least squares estimator. IDS-E is averaging over 10 proposals of Thompson Sampling.

We stress that for the strategies TS, WTS, D/IDS-TS and D/IDS-E there are currently no frequentist regret bounds known, still they perform well in the experiments. The difference between the IDS and DIDS versions is small except for IDS-F and DIDS-F, where as previously pointed out, the randomized version seems to engage in over-exploration along the theoretical upper bound.

Similar, as IDS-F is much more efficient than DIDS-F in the sense that it achieves a higher information gain in terms if I_t^F while still maintaining the worst-case regret guarantee, we conjecture that there are settings and information functions, where the randomized IDS has lower regret than the deterministic version. As usual, all experiments have to be taken with a grain of salt; of course it is possible to degrade the performance of information directed strategies by setting the noise of a suboptimal action to a very small value; and the other way round, if the optimal action has very little noise, our methods gain an additional advantage. To compile a more thorough experimental study, also including real-world applications, is a task for future work.

Appendix G. Useful Inequalities

The following lemma gives an upper bound on the variance of a bounded random variable.

Lemma 18 (Bhatia and Davis (2000)) *Let X be a real random variable supported in $[m, M]$. Then,*

$$\mathrm{Var}(X) \leq (M - \mathbb{E}[X])(\mathbb{E}[X] - m)\,,$$

and the bound is tight, if all mass is concentrated on the end-points of the interval.

The next two lemmas are two technical inequalities.

Lemma 19 *Let $a, b \geq 0$. Then $2\sqrt{ab} \leq a + b$.*

Proof Squaring both sides gives $4ab \leq a^2 + b^2 + 2ab$, or equivalently, $0 \leq a^2 + b^2 - 2ab = (a - b)^2$, which is clearly true. □

Lemma 20 *Let $a, b \geq 1$. Then $\sqrt{ab\log(ab)} \leq a + b\log(b)$.*

Proof By $a, b \geq 1$, we can equivalently show that $ab\log(ab) \leq (a + b\log(b))^2 = a^2 + 2ab\log(b) + (b\log(b))^2$. Assume first that $b\log(a) \leq a$. Then

$$ab\log(ab) = ab\log(a) + ab\log(b) \leq a^2 + ab\log(b) \leq (a + b\log(b))^2\,.$$

Assume now the opposite, i.e. $a \leq b\log(a)$. With the inequality $\log(a) \leq \sqrt{a}$, this implies that $a \leq b^2$. Consequently,

$$ab\log(ab) \leq ab\log(b^3) = 3ab\log(b) < 4ab\log(b)$$

Hence, by Lemma 19,

$$\sqrt{ab\log(ab)} \leq 2\sqrt{ab\log(b)} \leq a + b\log(b)\,.$$

This completes the proof. □