Myopic Posterior Sampling for Adaptive Goal Oriented Design of Experiments

Kirthevasan Kandasamy [1] Willie Neiswanger [1] Reed Zhang [1] Akshay Krishnamurthy [2]
Jeff Schneider [1] Barnabás Póczos [1]

Abstract

Bayesian methods for adaptive decision-making, such as Bayesian optimisation, active learning, and active search have seen great success in relevant applications. However, real world data collection tasks are more broad and complex, as we may need to achieve a combination of the above goals and/or application specific goals. In such scenarios, specialised methods have limited applicability. In this work, we design a new myopic strategy for a wide class of adaptive design of experiment (DOE) problems, where we wish to collect data in order to fulfil a given goal. Our approach, Myopic Posterior Sampling (MPS), which is inspired by the classical posterior sampling algorithm for multi-armed bandits, enables us to address a broad suite of DOE tasks where a practitioner may incorporate domain expertise about the system and specify her desired goal via a reward function. Empirically, this general-purpose strategy is competitive with more specialised methods in a wide array of synthetic and real world DOE tasks. More importantly, it enables addressing complex DOE goals where no existing method seems applicable. On the theoretical side, we leverage ideas from adaptive submodularity and reinforcement learning to derive conditions under which MPS achieves sublinear regret against natural benchmark policies.

1. Introduction

Many problems in adaptive decision-making under uncertainty fall into the design of experiments (DOE) framework, where one wishes to design a sequence of experiments and collect data so as to achieve a desired goal. For example, in electrolyte design for batteries, a chemist would like to conduct experiments that measure battery conductivity in order to identify an electrolyte design that maximises conductivity. On a different day, she would like to experiment with different designs to learn how the viscosity of the electrolyte changes with the design. These two tasks, black-box optimisation and active learning, fall under the umbrella of DOE and are pervasive in industrial and scientific applications.

While several methods exist for specific DOE tasks, real world problems are broad and complex, and specialised methods have limited applicability. Continuing with the electrolyte example, the chemist can typically measure both conductivity and viscosity with a single experiment. Since such experiments are expensive, it is wasteful to first perform a set of experiments to optimise conductivity and then a fresh set to learn viscosity. Rather, it is desirable to design a single set of experiments that simultaneously achieves both goals. A second example is metallurgy, where one wishes to identify phase transitions in an alloy by carefully selecting a sequence of X-ray diffraction experiments (Bunn et al., 2016). Here and elsewhere, the goal of the experimenter is application specific and cannot be simply shoehorned into standard DOE formulations such as black-box optimisation, active learning, etc. In addition, domain knowledge about the problem may need to be considered in selecting experiments, as it may significantly reduce the number of experiments needed to achieve the desired goal.

To address these desiderata, we develop a general and flexible framework for *goal oriented* DOE, where a practitioner may specify her desired goal via a reward function λ. λ can depend on the data collected during the DOE process and unknown system characteristics, and hence cannot be directly computed by a decision-maker. We then develop an *adaptive* myopic strategy for DOE, inspired by posterior (Thompson) sampling for multi-armed bandits (Thompson, 1933), which uses results from past experiments to plan future experiments and achieve the goal, i.e. maximise λ. Our approach has two key advantages. First, our Bayesian formulation allows one to straightforwardly specify domain expertise. Moreover, modern tools for probabilistic programming enable pratitioners to apply a Bayesian algorithm such as ours in a fairly straightforward manner. Second, our myopic strategy is simple and computationally attractive in comparison with policies that engage in long-term planning. Nevertheless, borrowing ideas from submodular optimisation and reinforcement learning, we derive natural

[1] Carnegie Mellon University [2] Microsoft Research. Correspondence to: Kirthevasan Kandasamy <kandasamy@cs.cmu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

conditions under which our myopic policy is competitive with the globally optimal one. Our contributions are:

1. We propose a flexible framework for DOE that allows a practitioner to describe their system via a probabilistic model, and specify their goal via a reward function. We derive a general purpose algorithm, Myopic Posterior Sampling (MPS), for this setting.
2. Empirically, we demonstrate MPS performs favourably in a variety of synthetic and real world DOE tasks. Despite its generality, MPS is competitive with specialised methods designed for particular settings. More importantly, it enables DOE in non-standard application-specific settings. Our implementation and experiments are available at `github.com/kirthevasank/mps`.
3. In our theoretical analysis, we explore conditions under which MPS, which learns about the system over time, is competitive with myopic and globally optimal strategies that have full knowledge of the system.

Related Work: The term DOE has been used to refer to different settings in the literature. Classically, the focus has been on learning an unknown system, and as such, the objective has been framed as maximising some notion of information gathered about the system. We will refer to these tasks as L-DOE problems to differentiate it from our setting, which subsumes L-DOE. Classical L-DOE focuses on discrete settings (Chernoff, 1972; Robbins, 1952) or linear models (Allen-Zhu et al., 2017; Fedorov, 1972). Recently, there have been several Bayesian approaches for L-DOE that adopt probabilistic programming in more complex models (Ouyang et al., 2016; Rainforth, 2017). However, L-DOE approaches may not be efficient or appropriate for an arbitrary user-specified reward λ. Moreover, many of these approaches are non-adaptive, aiming to find an optimal batch of experiments beforehand without incorporating feedback from completed experiments. While some do explore adaptive approaches for L-DOE, they aim for globally optimal policies (e.g. Rainforth (2017)), which can be computationally prohibitive, except in the most simple cases.

We focus on posterior sampling (PS) (Thompson, 1933) as the bandit algorithm, since it has proven to be quite general and admits a clean Bayesian analysis (Russo and Van Roy, 2016a). PS has been studied in a number of bandit settings (Gopalan et al., 2014; Kandasamy et al., 2018; Kawale et al., 2015), and some episodic RL problems (Gopalan and Mannor, 2015; Osband and Van Roy, 2014; Osband et al., 2013), where the agent is allowed to restart. In contrast, here we study PS on a single long trajectory with no restarts.

Myopic/greedy policies, while computationally simple, are known to be near-optimal for sequential decision making problems with *adaptive submodularity* (Golovin and Krause, 2011), which generalises submodularity (Nemhauser et al., 1978) and formalises a diminishing returns property. Adaptive submodularity has been used for several adaptive DOE setups (Chen and Krause, 2013; Chen et al., 2014; 2017; Golovin et al., 2010). However, in these work, the reward only depends on the data collected and can be directly computed by the decision-maker. In our setting, this translates to the agent knowing the system characteristics. As such, these results are complementary to ours: adaptive submodularity controls the approximation error (the difference between myopic- and globally-optimal strategies, both of which know the system), while we control the estimation error (how close our policy which needs to learn about the system is to the myopic optimal policy that knows the system). As we show in Theorem 3, with adaptive submodularity, MPS can also compete with the globally optimal policy. In a similar vein, Frazier et al. (2008); Wang and Powell (2018) use knowledge gradient approaches for information collection tasks which are framed as myopic adaptive submodular set maximisation problems; but as before, the system is known to the decision-maker. Prior results for learning in submodular environments are episodic and allow restarts (Gabillon et al., 2013; 2014), which is unnatural in the DOE setup. In addition to the above, several papers have developed Bayesian methods for specific DOE applications such as black-box optimisation (Frazier, 2018), active search (Jiang et al., 2018), level set estimation (Gotovos et al., 2013a) and more (Kandasamy et al., 2015; Osborne et al., 2012).

Our theoretical analysis leverages ideas from reinforcement learning (RL) since at each round the agent makes a decision (what experiment to perform) with the goal of maximising a long-term reward. In that light, one goal of our work is to understand when myopic "bandit-like" strategies perform well in RL environments with long-term temporal dependencies. There are two main differences with prior work (Jaksch et al., 2010; Kearns and Singh, 2002; Liu and Brunskill, 2018; Osband and Van Roy, 2014; Strehl et al., 2009). First, we make no explicit assumptions about the complexity of the state and action space, instead placing assumptions on the reward structure and optimal policy, which is a better fit for our applications. Crucially, in our setup, the true reward is never revealed to the agent, and instead it receives side-observations that provide information about an underlying parameter governing the environment. Secondly, our focus is on understanding when myopic strategies have reasonable performance rather than on achieving global optimality.

2. Set up and Method

Let Θ denote a parameter space, $\mathcal{X}$ an action space, and $\mathcal{Y}$ an outcome space. We consider a Bayesian setting where a *true parameter* $\theta_\star \in \Theta$ is drawn from a prior distribution ρ_0. A decision maker repeatedly chooses an action $X \in \mathcal{X}$, conducts an experiment at X, and observes the outcome $Y_X \in \mathcal{Y}$. We assume Y_X is drawn from a *like-*

lihood $\mathbb{P}(\cdot|X, \theta_\star)$, with known distributional form. This process proceeds for n rounds, resulting in a *data sequence* $D_n = \{(X_j, Y_{X_j})\}_{j=1}^n$, which is an ordered multi-set of action-observation pairs. Unlike, classical formalisms for DOE, we study a setting where we intend to achieve a desired goal, specified via a *reward function* $\lambda : \Theta \times \mathcal{D} \to \mathbb{R}$, that we wish to maximise. Here, $\mathcal{D}$ denotes the set of all possible data sequences. In particular, after n rounds, we focus on the following two criteria, depending on the application:

$$\textbf{(a)}\ \ \Lambda(\theta_\star, D_n) = \sum_{t=1}^{n} \lambda(\theta_\star, D_t) \qquad \textbf{(b)}\ \ \lambda(\theta_\star, D_n), \qquad (1)$$

Here, $D_t = \{(X_j, Y_{X_j})\}_{j=1}^t$ denotes the *prefix* of length t of the data sequence D_n collected by the decision maker. The former notion is the cumulative sum of all rewards, while the latter corresponds to the reward once all experiments are complete. Since λ depends on the unknown true parameter $\theta_\star$, the decision maker cannot compute the reward during the data collection process, and instead must infer the reward from observations in order to maximise it. This is a key distinction from existing work on reinforcement learning and sequential optimisation, and one of the new challenges in our setting.

Example 1. *A motivating example is* Bayesian active learning *(Chen et al., 2017). Here, actions X correspond to data points while Y_X is the label and $\mathbb{P}(y|x, \theta)$ specifies an assumed discriminative model. We may set $\lambda(\theta, D_n) = -\|\beta(\theta) - \hat{\beta}(D_n)\|_2^2$ where β is a parameter of interest and $\hat{\beta}$ is a predetermined estimator (e.g. via maximum likelihood). The true reward $\lambda(\theta_\star, D_n)$ is not available to the decision maker since it requires knowing $\beta(\theta_\star)$.*

Notation: For each $t \in \mathbb{N}$, let $\mathcal{D}_t = \{(X_j, Y_{X_j})\}_{j=1}^t : X_j \in \mathcal{X}, Y_{X_j} \in \mathcal{Y}\}$ denote the set of all data sequences of length t, so that $\mathcal{D} = \bigcup_{t \in \mathbb{N}} \mathcal{D}_t$. Let $D \uplus D'$ denote the concatenation of two sequences. $D \prec D'$ and $D' \succ D$ both equivalently denote that D is a prefix of D'. Given a data sequence D_t, we use $D_{t'}$ for $t' < t$ to denote the prefix of the first t' action-observation pairs.

A *policy* for experiment design chooses a sequence of actions $\{X_j\}_{j \in \mathbb{N}}$ based on past actions and observations. In particular, for a *randomised* policy $\pi = \{\pi_j\}_{j \in \mathbb{N}}$, at time t, an action is drawn from $\pi_t(D_{t-1}) = \mathbb{P}(X_t \in \cdot|D_{t-1})$. Two policies that will appear frequently in the sequel are $\pi_{\mathrm{M}}^\star$ and $\pi_{\mathrm{G}}^\star$, both of which operate with knowledge of $\theta_\star$. $\pi_{\mathrm{M}}^\star$ is the myopic optimal policy, which, from every data sequence D_t chooses the action X maximising the expected reward at the next step: $\mathbb{E}[\lambda(\theta_\star, D_t \uplus \{(X, Y_X)\})|\theta_\star, D_t]$. On the other hand $\pi_{\mathrm{G}}^\star$ is the non-myopic, globally optimal adaptive policy, which in state D_t with $n - t$ steps to go chooses the action to maximise the expected long-term reward: $\mathbb{E}[\lambda(\theta_\star, D_t \uplus \{(X, Y_X)\} \uplus D_{t+2:n}) \mid \pi_{\mathrm{G}}^\star, \theta_\star, D_t]$. $\pi_{\mathrm{G}}^\star$ may depend on the time horizon n while $\pi_{\mathrm{M}}^\star$ does not.

Design of Experiments via Posterior Sampling

We present a simple and intuitive myopic strategy that aims to maximise λ based on the posterior of the data collected so far. For this, first define the expected look-ahead reward $\lambda^+ : \Theta \times \mathcal{D} \times \mathcal{X} \to [0, 1]$, such that $\lambda^+(\theta, D, x)$ is the expected reward at the next time step if $\theta \in \Theta$ were the true parameter, D was the current data sequence collected, and we were to take action $x \in \mathcal{X}$. Precisely,

$$\lambda^+(\theta, D, x) = \mathbb{E}_{Y_x \sim \mathbb{P}(Y|x,\theta)}\Big[\lambda\big(\theta,\ D \uplus \{(x, Y_x)\}\big)\Big]. \quad (2)$$

The proposed policy, presented in Algorithm 1, is called MPS (Myopic Posterior Sampling) and is denoted $\pi_{\mathrm{M}}^{\mathrm{PS}}$. At time step t, it first samples a parameter value θ from the posterior for $\theta_\star$ conditioned on the data, i.e. $\theta \sim \mathbb{P}(\theta_\star|D_{t-1})$. Then, it chooses the action X_t that is expected to maximise the reward λ by pretending that θ was the true parameter. It performs the experiment at X_t, collects the observation Y_{X_t}, and proceeds to the next time step.

Algorithm 1 MPS ($\pi_{\mathrm{M}}^{\mathrm{PS}}$)

```
Require: Prior $\rho_0$ for $\theta_\star$, Conditional $\mathbb{P}(Y|X, \theta)$.
1: $D_0 \leftarrow \varnothing$.
2: for $t = 1, 2, \ldots$ do
3:     Sample $\theta \sim \rho_{t-1} \equiv \mathbb{P}(\theta_\star|D_{t-1})$.
4:     Choose $X_t = \mathrm{argmax}_{x \in \mathcal{X}}\ \lambda_{t-1}^+(\theta, D_{t-1}, x)$.
5:     $Y_{X_t} \leftarrow$ conduct experiment at $X_t$.
6:     Set $D_t \leftarrow D_{t-1} \cup \{(X_t, Y_{X_t})\}$.
7: end for
```

A natural question that may arise is the need to sample from the posterior ρ_{t-1} for $\theta_\star$, instead of taking an expectation of λ^+ over ρ_{t-1}. In fact, many policies for non-adaptive L-DOE take an expectation over the posterior (Ouyang et al., 2016). However, in adaptive settings where $\theta_\star$ is unknown, taking the expectation may fail as it may not explore sufficiently. For example, in bandit problems, which is a special case of our setting (Section 3.5), this amounts to choosing the maximum of the posterior mean of the payoff function, which is known to fail spectacularly.

Computational considerations: It is worth pointing out some computational considerations in Algorithm 1. First, sampling from the posterior in step 3 might be difficult, especially in complex Bayesian models. Fortunately however, the field of Bayesian inference has made great strides in the recent past with the development of fast techniques for approximate inference methods such as MCMC or variational inference (Hensman et al., 2012; Neiswanger et al., 2015). Moreover, today we have efficient probabilistic programming tools (Bingham et al., 2018; Carpenter et al., 2017;

Tran et al., 2017) that allow a practitioner to intuitively incorporate domain expertise via a prior and obtain the posterior given data. Secondly, the maximisation of the look ahead reward in step 4 can also be non-trivial, especially since it might involve empirically computing the expectation in (2). This is similar to existing work in Bayesian optimisation which assume access to such an optimisation oracle (Bull, 2011; Srinivas et al., 2010). That said, in many practical settings where experiments can cost significant time and money, these considerations are less critical.

Despite these concerns, it is worth mentioning that myopic strategies are still computationally far more attractive than policies which try to behave globally optimally. For example, extending MPS to a k step look-ahead might involve an optimisation over $\mathcal{X}^k$ in step 4 of Algorithm 1 which might be impractical for large values of k. Moreover, in many problems where system characteristics $\theta_\star$ are known to the decision maker, myopic policies can be competitive with globally optimal policies (Golovin and Krause, 2011; Nemhauser et al., 1978; Wei et al., 2015). In Section 4, we identify conditions where $\pi_{\rm M}^{\rm PS}$ can be competitive with the globally optimal policy $\pi_{\rm G}^\star$ which knows $\theta_\star$.

3. Examples & Experiments

We now describe some concrete examples of DOE problems that can be specified by a reward function λ and present experimental results. We compare $\pi_{\rm M}^{\rm PS}$ to random sampling (RAND), the myopically optimal policy $\pi_{\rm M}^\star$ which has access to $\theta_\star$, and to specialised methods for the particular problem, when available. In the interest of aligning our experiments with our theoretical analysis, we compare methods on both criteria in (1), although in these applications, the final reward $\lambda(\theta_\star, D_n)$ is more relevant than the cumulative one $\Lambda(\theta_\star, D_n)$. In all cases, except Experiments 2 and 4 which have conjugate priors, We use variational inference (VI) in Edward (Tran et al., 2017) to approximate the posterior $\mathbb{P}(\theta_\star|D_t)$. While VI is known to underestimate the variance in practice, it worked well in our experiments. For better visualisation, we plot the negative reward in a semilog plot. We defer some experimental details to Appendix D.

High-level Takeaways: Despite being a quite general, $\pi_{\rm M}^{\rm PS}$ outperforms, or performs as well as, specialised methods. $\pi_{\rm M}^{\rm PS}$ is competitive, but typically worse than the non-realisable $\pi_{\rm M}^\star$. Finally $\pi_{\rm M}^{\rm PS}$ enables effective DOE in complex settings where no prior methods seem applicable.

3.1. Active Learning

Problem: As described previously, we wish to learn some parameter $\beta_\star = \beta(\theta_\star)$ which is a function of the true parameter $\theta_\star$. Each time we query some $X \in \mathcal{X}$, we observe a label $Y \sim \mathbb{P}(Y|X, \theta_\star)$. We conduct two synthetic experiments in this setting. We use $\lambda(\theta_\star) \overset{\Delta}{=} -\|\beta_\star - \hat{\beta}(D_n)\|_2^2$ as the reward where $\hat{\beta}$ is a regularised maximum likelihood estimator. In addition to RAND and $\pi_{\rm M}^\star$, we compare $\pi_{\rm M}^{\rm PS}$ to ActiveSetSelect of Chaudhuri et al. (2015).

Experiment 1: We use the following parametric model: $Y_x|x, \theta \sim \mathcal{N}(f_\theta(x), \eta^2)$ where $f_\theta(x) = \frac{a}{1+e^{b(x-c)}}$ is a logistic function. The true parameter is $\theta_\star = (a, b, c, \eta^2)$ and our goal is to estimate $\beta_\star = (a, b, c)$. The MLE is computed via gradient ascent on the log likelihood. In our experiments, we used $a = 2.1, b = 7, c = 6$ and $\eta^2 = 0.01$ as $\theta_\star$. We used normal priors $\mathcal{N}(2, 1), \mathcal{N}(5, 3)$ and $\mathcal{N}(5, 3)$ for a, b, c respectively and an inverse gamma $\mathrm{IG}(20, 1)$ prior for η^2. As the action space, we used $\mathcal{X} = [0, 10]$. For variational inference, we used a normal approximation for the posterior for a, b, c and an inverse gamma approximation for η^2. The results are given in the first column of Figure 1.

Experiment 2: In the second example, we use the following linear regression model: $Y_x|x, \theta \sim \mathcal{N}(f_\theta(x), 0.01)$ where $f_\theta(x) = \sum_{i=1}^{16} \theta_{*i}\phi(x - c_i)$. Here, $\phi(v) = \frac{1}{\sqrt{0.2\pi}} e^{-5\|v\|_2^2}$ and the points $c_1, \ldots, c_{16}$ were arranged in a 4×4 grid within $[0, 1]^2$. We set $\theta_{*i} = g(c_i)$, with $g(v) = \sin(3.9\pi((v_1 - 0.1)^2 + v_2 + 0.1))$. Our goal is to estimate $\beta_\star = \theta_\star$. As the action space, we used $\mathcal{X} = [0, 1]^2$. The posterior for $\theta_\star$ was calculated in closed form using a normal distribution $\mathcal{N}(0, I_{16})$ as the prior. The results are given in the second column of Figure 1.

Alternative Problem Formalism: A common formalism for parameter estimation in discriminative models (Chaudhuri et al., 2015; Frostig et al., 2015) is to maximise the expected likelihood of the data for a given sampling distribution Γ on $\mathcal{X}$. Here, one wishes to maximise $\lambda(\theta_\star, D_n) \overset{\Delta}{=} \mathbb{E}_{X\sim\Gamma, Y\sim\mathbb{P}(Y|X,\theta_\star)}[\log \mathbb{P}(Y|X, \hat{\theta})]$, where $\hat{\theta}$ is an estimator for θ obtained from D_n.

Experiments 3 & 4: We use the same models as in Experiment 1 & 2 but with the above reward function. We let Γ be the uniform distribution on the respective domains and $\hat{\theta}$ be the maximum likelihood estimator for θ. The results are given in the third and fourth columns of Figure 1.

3.2. Posterior Estimation & Active Regression

Problem: Consider estimating a non-parametric function $f_{\theta_\star}$, which is known to be uniformly smooth. An action $x \in \mathcal{X}$ queries $f_{\theta_\star}$, upon which we observe $Y_x = f_{\theta_\star}(x) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$. If the goal is to learn $f_{\theta_\star}$ uniformly well in L^2 error, i.e. with reward $-\|f_{\theta_\star} - \hat{f}(D_n)\|^2$, adaptive techniques may not perform significantly better than non-adaptive ones (Willett et al., 2006). However, if our reward was $\lambda(\theta_\star, D_n) \overset{\Delta}{=} -\|\sigma(f_{\theta_\star}) - \sigma(\hat{f}(D_n))\|^2$ for some monotone super-linear transformation σ, then adaptive techniques may do better by requesting more evaluations at regions

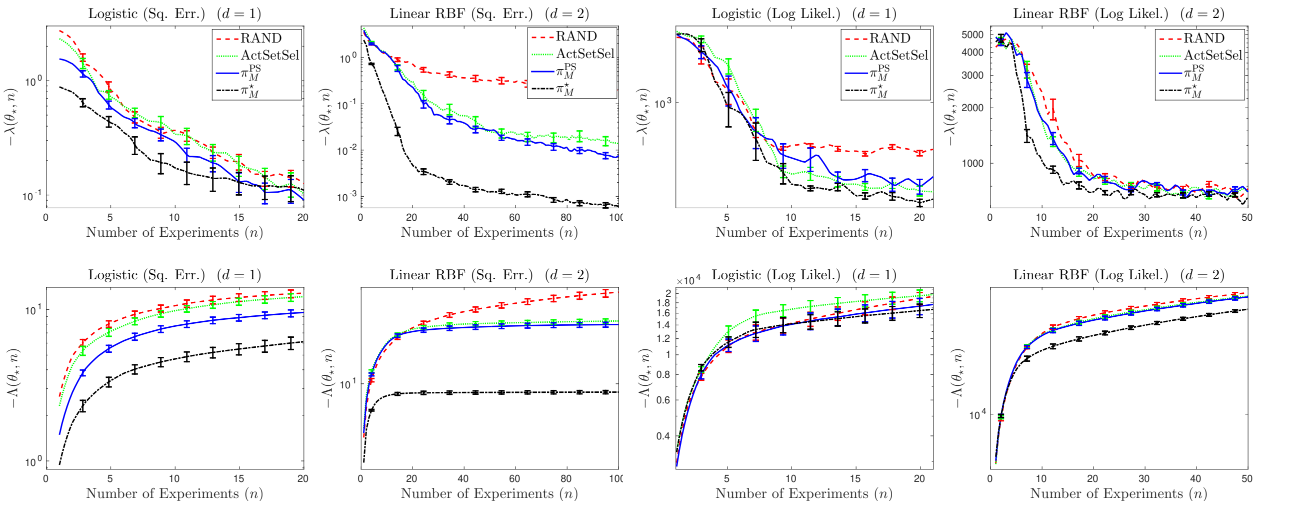

Figure 1: Results on the synthetic active learning experiments in Section 3.1. The title states the model and the dimensionality. In all figures, the x axis is the number of experiments n. In the top four figures, the y axis is the final negative reward $-\lambda(\theta_\star, n)$ at the n^{th} iteration. In the bottom figures, it is the corresponding negative cumulative reward $-\Lambda(\theta_\star, n)$. Lower is better in both cases. The legend for all plots is given in the top left figure. All curves were averaged over 20 runs, and error bars indicate one standard error.

with high $f_{\theta_\star}$ value. This is because, $\lambda(\theta_\star, D_n)$ is more sensitive to such regions due to the transformation σ.

A particularly pertinent instance of this formulation arises in astrophysical applications where one wishes to estimate the posterior distribution of cosmological parameters, given some astronomical data Q (Parkinson et al., 2006). Here, an astrophysicist specifies a prior Ξ over the cosmological parameters $Z \in \mathcal{X}$, and the likelihood of the data for a given choice of the cosmological parameters $x \in \mathcal{X}$ is computed via an expensive astrophysical simulation. The prior and the likelihood gives rise to an unknown log joint density[1] $f_{\theta_\star}$ defined on $\mathcal{X}$, and the goal is to estimate the the joint density $p(Z = x, Q) = \exp(f_{\theta_\star}(x))$ so that we can perform posterior inference. Adopting assumptions from Kandasamy et al. (2015), we model $f_{\theta_\star}$ as a Gaussian process, which is reasonable since we expect a log density to be smoother than the density itself. As we wish to estimate the joint density, λ takes the above form with $\sigma = \exp$.

Experiment 5: We use data on Type I-a supernova from Davis et al (2007). We wish to estimate the posterior over the Hubble constant $H \in (60, 80)$, the dark matter fraction $\Omega_M \in (0, 1)$ and the dark energy fraction $\Omega_E \in (0, 1)$, which constitute our three dimensional action space $\mathcal{X}$. The likelihood is computed via the Robertson-Walker metric. In addition to $\pi_{\mathrm{M}}^{\star}$ and RAND, we compare $\pi_{\mathrm{M}}^{\mathrm{PS}}$ to Gaussian process based exponentiated variance reduction (GP-EVR) (Kandasamy et al., 2015) designed for this setting. We evaluate the reward via numerical integration. The results are presented in the first column of Figure 2.

3.3. Level Set Estimation

Problem: In active level set estimation (LSE), one wishes to determine which regions of a space $\mathcal{X}$ fall above or below a given level set of an expensive to evaluate function $f_{\theta_\star}$. An experiment evaluates this function and returns $Y_x = f_{\theta_\star}(x) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$. We adopt the setting of Gotovos et al. (2013a), where a method for LSE returns its predictions for being above/below the threshold on a pre-specified set of discrete points $\mathcal{X}' \subset \mathcal{X}$. The reward function λ is set to be average prediction accuracy.

Experiment 6: We used data on luminous red galaxies (LRGs) to compute the galaxy power spectrum of 9 cosmological parameters including the spatial curvature, cold dark matter density, and baryonic density. We wish to find regions of the cosmological parameter space, where the power spectrum is larger than a pre-specified threshold. Software and data were taken from Tegmark et al (2006). We compare $\pi_{\mathrm{M}}^{\mathrm{PS}}$ to random search, $\pi_{\mathrm{M}}^{\star}$, and the Gaussian process based level set estimation (GP-LSE) method of Gotovos et al. (2013a). Following Gotovos et al. (2013a), we model the power spectrum as a GP, and define the reward function as described above where $\mathcal{X}'$ is a set of $\sim 20K$ points. The results are presented in the second column of Figure 2.

3.4. Combined and Customised Objectives

Problem: In many real world problems, one needs to design experiments with multiple goals. For example, an experiment might evaluate multiple objectives, and the task might be to optimise some of them, while learning the parameters for another. Classical methods specifically designed for active learning or optimisation may not be suitable in such

[1] One should not conflate the prior over $\mathcal{X}$ specified with the astrophysics model, with prior over Θ assumed in our set up.

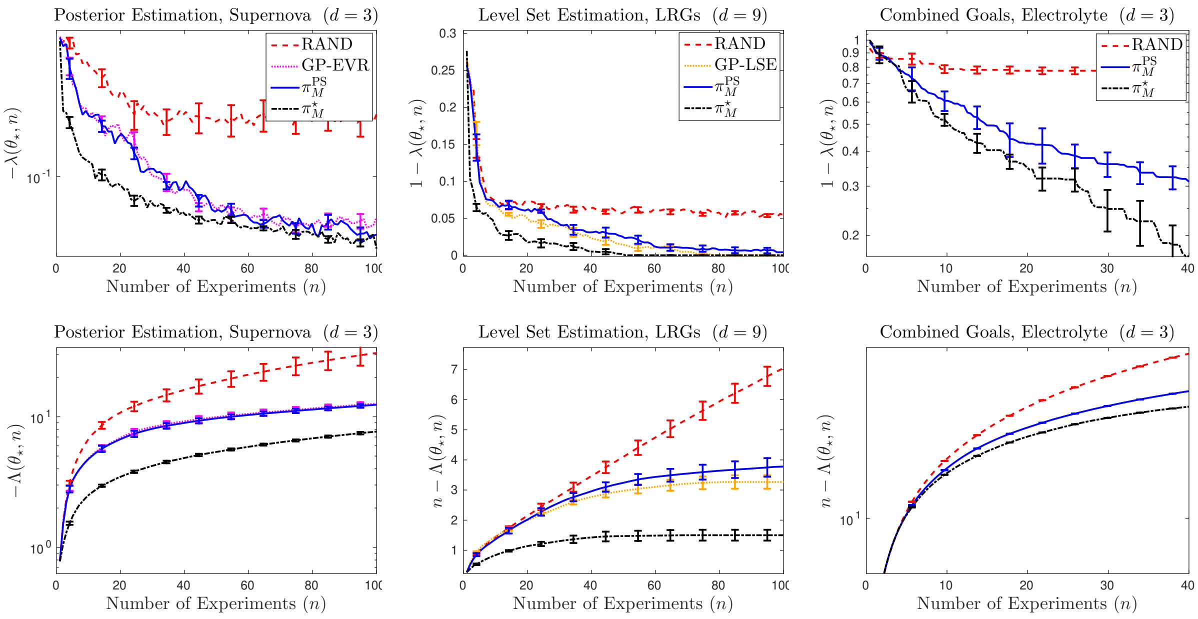

Figure 2: Results on the real experiments. The first column is for the posterior estimation problem in Sec. 3.2, the second column is for the level set estimation problem in Sec. 3.3, and the third column is for the combined objective problem in Sec. 3.2, In the top figures, the y axis is the negative reward $-\lambda(\theta_\star, D_n)$ and in the bottom figures, it is the negative cumulative reward $-\Lambda(\theta_\star, D_n)$ for the corresponding experiment. See caption under Figure 1 for more details.

settings. One advantage to the proposed framework is that it allows us to combine multiple goals in the form of a reward function. For instance, if an experiment measures two functions $f_{\theta_\star,1}, f_{\theta_\star,2}$ and we wish to learn f_1 while optimising f_2, we can define the reward as $\lambda(\theta_\star, D_n) = -\|f_{\theta_\star,1} - \hat{f}_1(D_n)\|^2 + \max_{X_t, t\le n}\left(f_{\theta_\star,2}(X_t) - \max_x f_{\theta_\star,2}(x)\right)$. Here $\hat{f}_1$ is an estimate for $f_{\theta_\star,1}$ obtained from the data, $\|\cdot\|$ is the L_2 norm and $\max_{X_t,t\le n} f_{\theta_\star,2}(X_t)$ is the maximum point of $f_{\theta_\star,2}$ we have evaluated so far. Below, we demonstrate one such application.

Experiment 7: In battery electrolyte design, one tests an electrolyte composition under various physical conditions. On an experiment at $x \in \mathcal{X}$, we obtain measurements $Y_x = (Y_{x,\mathrm{sol}}, Y_{x,\mathrm{vis}}, Y_{x,\mathrm{con}})$ which are noisy measurements of the solvation energy f_{sol}, the viscosity f_{vis} and the specific conductivity f_{con}. Our goal is to estimate f_{sol} and f_{vis} while optimising f_{con}. Hence,

$$\begin{aligned}\lambda(\theta_\star, D_n) \;=\; & \alpha\big(\max_{X_t, t\le n} f_{\mathrm{con}}(X_t) - \max_{x\in\mathcal{X}} f_{\mathrm{con}}(x)\big) \\ & - \beta\|f_{\mathrm{sol}} - \hat{f}_{\mathrm{sol}}(D_n)\|^2 - \gamma\|f_{\mathrm{vis}} - \hat{f}_{\mathrm{vis}}(D_n)\|^2,\end{aligned}$$

where, the parameters α, β, γ were chosen so as to scale each objective and ensure that none of them dominate the reward. In our experiment, we use the dataset from Gering (2006). Our action space $\mathcal{X}$ is parametrised by the following three variables: $Q \in (0,1)$ measures the proportion of two solvents EC and EMC in the electrolyte, $S \in (0, 3.5)$ is the molarity of the salt LiPF_6 and $T \in (-20, 50)$ is the temperature in Celsius. We use the following prior which is based off a physical understanding of the interaction of these variables. $f_{\mathrm{con}} : \mathcal{X} \to \mathbb{R}$ is sampled from a Gaussian process (GP), $f_{\mathrm{vis}}(Q,S,T) = \exp(-aT)g_{\mathrm{vis}}(Q,S)$ where g_{vis} is sampled from a GP, and $f_{\mathrm{sol}}(Q,S,T) = b + \exp(cQ - dS - eT)$. We use inverse gamma priors for a, b, d, e and a normal prior for c. For variational inference, we used inverse gamma approximations for a, b, d, e, a normal approximation for c, and GP approximations for f_{con} and g_{vis}. We use the posterior mean of f_{sol} and f_{vis} under this prior as the estimates $\hat{f}_{\mathrm{sol}}, \hat{f}_{\mathrm{vis}}$. We present the results in the third column of Figure 2 where we compare RAND, $\pi_{\mathrm{M}}^{\mathrm{PS}}$ and $\pi_{\mathrm{M}}^{\star}$. This is an example of a customised DOE problem for which no prior method seems directly applicable.

3.5. Bandits & Bayesian Optimisation

Bandits and Bayesian optimisation are self-evident special cases of our formulation. Here, $\theta_\star$ specifies a function $f_{\theta_\star} : \mathcal{X} \to \mathbb{R}$. When we choose a point $X \in \mathcal{X}$ to evaluate the function, we observe $Y_X = f_{\theta_\star}(X) + \epsilon$ where $\mathbb{E}[\epsilon] = 0$. In the bandit framework, the reward is the instantaneous regret $\lambda(\theta_\star, D_n) = f_{\theta_\star}(X_n) - \max_{x\in\mathcal{X}} f_{\theta_\star}(x)$. In Bayesian optimisation, one is interested in simply finding a single value close to the optimum and hence $\lambda(\theta_\star, D_n) = \max_{t\le n} f_{\theta_\star}(X_t) - \max_{x\in\mathcal{X}} f_{\theta_\star}(x)$. In either case, $\pi_{\mathrm{M}}^{\mathrm{PS}}$ reduces to the Thompson sampling procedure as $\mathrm{argmax}_{x\in\mathcal{X}}\, \lambda^+(\theta_\star, D_{t-1}, x) = \mathrm{argmax}_{x\in\mathcal{X}}\, f_\theta(x)$, where f_θ is a random function drawn from the posterior. Since prior work has demonstrated that TS performs empirically well in several real world optimisation tasks (Chapelle and Li, 2011; Hernández-Lobato et al., 2017; Kandasamy et al., 2018), we omit experimental results for this example. One can also cast other variants of Bayesian optimisation, including multi-objective optimisation (Hernández-Lobato et al., 2016; Paria et al., 2018) and constrained optimization (Gardner et al., 2014), in our general formulation.

4. Theoretical Analysis

In this section we derive theoretical guarantees for $\pi^{\rm PS}_{\rm M}$. Our emphasis is on understanding conditions under which myopic algorithms which need to learn $\theta_\star$ can perform competitively with the myopic optimal and the globally optimal oracles $\pi^\star_{\rm M}, \pi^\star_{\rm G}$ which know $\theta_\star$ (see Section 2). Going forward, to simplify the exposition, we will assume that λ is bounded, i.e. $\lambda : \Theta \times \mathcal{X} \to [0, 1]$. Moreover, w.l.o.g, we will assume for all $\theta \in \Theta$, $\sup_{D\in\mathcal{D}} \lambda(\theta, D) = 1$. This condition is for free since for any bounded reward λ', we can set $\lambda(\theta, D) \triangleq 1 + \lambda'(\theta, D) - \sup_{D\in\mathcal{D}} \lambda'(\theta, D)$.

For criterion (**a**), we are interested in upper bounding $\mathbb{E}[\Lambda(\theta_\star, D_n)|D_n \sim \pi^{\rm PS}_{\rm M}]$ in terms of $\mathbb{E}[\Lambda(\theta_\star, D_n)|D_n \sim \pi^\star_{\rm M}]$, which yields a *cumulative regret bound*, and for criterion (**b**), we wish to bound $\mathbb{E}[\lambda(\theta_\star, D_n) \mid D_n \sim \pi^{\rm PS}_{\rm M}]$ in terms of the analogous quantities for $\pi^\star_{\rm M}, \pi^\star_{\rm G}$, which serves as a *final regret bound*. Note that a comparison with $\pi^\star_{\rm G}$ on (**a**) is meaningless since it might take low reward actions in the early stages in order to do well in the long run. In fact, our bounds for (**a**) will hold when $\lambda(\theta_\star, D)$ is an ordered multi-set function in D, but for (**b**) when $\lambda(\theta_\star, D)$ is a multi-set function, i.e. the ordering does not matter. Our bounds will hold in expectation over $\theta_\star \sim \rho_0$.

The following proposition shows that without further assumptions, a non-trivial regret bound is impossible. Such results are common in the RL literature, and necessitate several structural assumptions (Dann and Brunskill, 2015; Jaksch et al., 2010; Kearns and Singh, 2002). Its proof is given in Appendix B.4.

Proposition 1. *For all policies π which do not know $\theta_\star$, there exists a DOE problem where $\mathbb{E}_{\theta_\star\sim\rho_0}[\lambda(\theta_\star, D^\star_n) - \lambda(\theta_\star, D_n)|D^\star_n \sim \pi^\star_{\rm M}, D_n \sim \pi] \geq 1/2$ for all $n \geq 1$.*

Motivated by this lower bound, we impose the following condition on the parameter space and reward structure, under which a policy can achieve sub-linear regret. For this, first note we can assume that, at all time steps, the observations $Y \sim \mathbb{P}(\cdot|x, \theta_\star)$ are generated for all $x \in \mathcal{X}$, but we only observe those for the chosen X_t. With this in consideration, let $\mathbb{E}_{Y,t+1:|\theta}$ denote expectation over all observations generated from time $t + 1$ onwards when $\theta_\star = \theta$.

Condition 1. *Let θ, θ' denote parameter values in Θ and $\pi^\theta_{\rm M}, \pi^{\theta'}_{\rm M}$ be the myopically optimal policies when $\theta_\star = \theta$, and $\theta_\star = \theta'$ respectively. Let H denote a data sequence and D_n, D'_n be the data sequences collected by $\pi^\theta_{\rm M}$ and $\pi^{\theta'}_{\rm M}$ respectively when starting from H when $\theta_\star = \theta$ and $\theta_\star = \theta'$ respectively, i.e. the myopically optimal policies operate in their respective environments. Then, there exists sequences $\{\epsilon_n\}_{n\geq1}$, $\{\tau_n\}_{n\geq1}$ such that the following hold.*

1. *$\pi^\theta_{\rm M}$ achieves asymptotically similar reward $\forall\ \theta \in \Theta$. That is,*
$$\sup_{\theta,\theta'\in\Theta}\ \sup_{H\in\mathcal{D}} \Big\{\mathbb{E}_{Y,|H|+1:|\theta}\, \lambda(\theta, H \uplus D_n) - \mathbb{E}_{Y,|H|+1:|\theta'}\, \lambda(\theta', H \uplus D'_n)\Big\} \ \leq\ \epsilon_n.$$
2. *The rate of convergence is better than $\mathcal{O}(1/\sqrt{n})$. That is, letting $\sqrt{\tau_n} = 1 + \sum_{j=1}^n \epsilon_j$, we have $\tau_n \in o(n)$.*

The condition states that when we execute $\pi^\star_{\rm M}$, the myopically optimal policy which knows and depends on the value of $\theta_\star$, from any prefix H, it achieves asymptotically similar λ for all values of $\theta_\star$. It is worth emphasising that the condition involves executing $\pi^\theta_{\rm M}$ in the environment where $\theta_\star = \theta$. A condition of the above form seems necessary for any myopic algorithm π that does not know $\theta_\star$ for the following reason. Assume that the myopic $\pi^\star_{\rm M}$ can quickly achieve large λ value when $\theta_\star \in \Theta_g$ but is slow when $\theta_\star \in \Theta_b$. Since π does not know $\theta_\star$ it needs to hedge against the "bad" situation, i.e. $\theta_\star \in \Theta_b$. However, in doing so, it will necessarily perform poorly against $\pi^\star_{\rm M}$ when $\theta_\star \in \Theta_g$ as $\pi^\star_{\rm M}$ can quickly achieve large λ. Condition 1 prevents such situations. As we will see shortly, the regret for $\pi^{\rm PS}_{\rm M}$ will depend on τ_n which dictates how differently $\pi^\star_{\rm M}$ can behave for different values of $\theta_\star$. In particular, sublinearity of τ_n is necessary for sublinear regret with $\pi^\star_{\rm M}$.

In Appendix C we provide a more interpretable sufficient condition which implies Condition 1, and demonstrate that it is satisfied with $\tau_n \in \mathcal{O}(1)$ for bandit and black-box optimisation problems and $\tau_n \in \mathcal{O}(\log n)$ for an active learning problem. We also consider a setting where λ has "state-like" structure; under assumptions similar to standard assumptions in reinforcement learning with ergodic Markov decision processes, we are able to show that Condition 1 holds. Finally we mention that if Condition 1 holds for two reward functions λ_1, λ_2, it is also true for the sum, $\lambda_1 + \lambda_2$, and the product, $\lambda_1 \cdot \lambda_2$, and can thus be applied to combined objective settings such as in Section 3.4.

Before stating the main theorem, we introduce the maximum information gain, Ψ_n, which captures the statistical difficulty of the learning problem.

$$\Psi_n = \max_{D_n\subset\mathcal{D}_n} {\rm I}(\theta_\star; D_n). \tag{3}$$

Here ${\rm I}(\cdot;\cdot)$ is the Shannon mutual information. Ψ_n measures the maximum information a set of n action-observation pairs can tell us about the true parameter $\theta_\star$.

The quantity appears as a statistical complexity measure in many Bayesian adaptive data analysis settings (Gotovos et al., 2013b; Ma et al., 2015; Srinivas et al., 2010). Below, we list some examples of common models which demonstrate that Ψ_n is typically sublinear in n.

Example 2. *We have the following bounds on Ψ_n for common models (Srinivas et al., 2010):*

1. ***Finite sets:*** *If Θ is finite, $\Psi_n \leq \log(|\Theta|)$ for all n.*
2. ***Linear models:*** *Let $\mathcal{X} \subset \mathbb{R}^d$, $\theta \in \mathbb{R}^d$, and $Y_x|x, \theta \sim \mathcal{N}(\theta^\top x, \eta^2)$. For a multi-variate Gaussian prior on $\theta_\star$, $\Psi_n \in \mathcal{O}(d\log(n))$.*
3. ***Gaussian process:*** *For a Gaussian process prior with RBF kernel over a compact domain $\mathcal{X} \subset \mathbb{R}^d$, and with Gaussian likelihood, we have $\Psi_n \in \mathcal{O}(\log(n)^{d+1})$.*

We now state our main theorem for finite action spaces $\mathcal{X}$.

Theorem 2. *Let $\mathcal{X}$ be finite and assume Condition 1 holds. Let τ_n be as defined in Condition 1. Then,*

$$\mathbb{E}[\Lambda(\theta_\star, \pi^\star_{\mathrm{M}}) - \Lambda(\theta_\star, \pi^{\mathrm{PS}}_{\mathrm{M}})] \leq \sqrt{\frac{|\mathcal{X}|n\tau_n\Psi_n}{2}}.$$

Theorem 2 establishes a sublinear regret bound for $\pi^{\mathrm{PS}}_{\mathrm{M}}$ against $\pi^\star_{\mathrm{M}}$ when $\tau_n\Psi_n \in o(n)$. The $|\mathcal{X}|$ term captures the complexity of our action space, Ψ_n captures the complexity of the prior on $\theta_\star$. The $\sqrt{n}$ dependence is in agreement with prior results for Thompson sampling (Kaufmann et al., 2012; Russo and Van Roy, 2016b). Thus, under Condition 1, $\pi^{\mathrm{PS}}_{\mathrm{M}}$ is competitive with the myopic optimal policy $\pi^\star_{\mathrm{M}}$, with average regret tending to 0.

We now compare $\pi^{\mathrm{PS}}_{\mathrm{M}}$ to the globally optimal policy $\pi^\star_{\mathrm{G}}$, when λ is a multi-set function, i.e. the ordering in D_n does not matter. For this, we first introduce the notions of *monotonicity* and *adaptive submodularity*.

Condition 2. *(Monotonicity and Adaptive Submodularity (Golovin and Krause, 2011)) Let $\mathbb{E}_{Y_x}$ denote the expectation over the likelihood $Y_x \sim \mathbb{P}(\cdot|x, \theta_\star)$. The following two statements are true for all $\theta \in \Theta$, $D, D' \in \mathcal{D}$, $D \prec D'$, and $x \in \mathcal{X}$. λ is a monotone, meaning that $\mathbb{E}_{Y_x}[\lambda(\theta, D \uplus \{(x, Y_x)\})] \geq \lambda(\theta, D)$. Moreover, λ is adaptive submodular, meaning that,*

$$\begin{aligned}&\mathbb{E}_{Y_x}[\lambda(\theta, D \uplus \{(x, Y_x)\})] - \lambda(\theta_\star, D)\\&\quad \geq \mathbb{E}_{Y_x}[\lambda(\theta, D' \uplus \{(x, Y_x)\})] - \lambda(\theta_\star, D').\end{aligned}$$

Monotonicity states that adding more data increases the reward in expectation, while adaptive submodularity formalises a notion of diminishing returns. That is, performing the same action is more beneficial when we have less data. It is easy to see that some assumption is needed here, since even in simple problems $\pi^\star_{\mathrm{M}}$ can be arbitrarily worse than $\pi^\star_{\mathrm{G}}$. We now state the second main result of this paper.

Theorem 3. *Assume that λ satisfies conditions 1 and 2. Let τ_n be as defined in Theorem 2. Then, for all $\gamma < 1$, we have*

$$\begin{aligned}&\mathbb{E}[\lambda(\theta_\star, D_n)|D_n \sim \pi^{\mathrm{PS}}_{\mathrm{M}}] \geq\\&\quad (1-\gamma)\mathbb{E}[\lambda(\theta_\star, D^\star_{\gamma n})|D^\star_{\gamma n} \sim \pi^\star_{\mathrm{G}}] - \sqrt{\frac{|\mathcal{X}|\tau_n\Psi_n}{2n}}.\end{aligned}$$

The theorem states that $\pi^{\mathrm{PS}}_{\mathrm{M}}$ in n steps is guaranteed to perform up to a $1-\gamma$ factor as well as $\pi^\star_{\mathrm{G}}$ executed for $\gamma n < n$ steps, up to an additive $\sqrt{\tau_n\Psi_n/n}$ term. The result captures both approximation and estimation errors, in the sense that we are using a myopic policy to approximate a globally optimal one, and we are learning a good myopic policy from data. In comparison, prior works on adaptive submodular optimisation focus on approximation errors and typically achieve $1-1/e$ approximation ratios against the n steps of $\pi^\star_{\mathrm{G}}$. Our bound is quantitatively worse, but focusing on a much more difficult task, and we view the results as complementary. Observe that an analogous bound holds against $\pi^\star_{\mathrm{M}}$, since it is necessarily worse that $\pi^\star_{\mathrm{G}}$.

Finally, we believe that the above results can be generalised to large or infinite action spaces under additional structure on λ. For example, when $\mathcal{X} \subset \mathbb{R}^d$, and the expected rewards are linear in the actions taken, we expect an $\mathcal{O}(d)$ dependence similar to linear bandit settings (Agrawal and Goyal, 2013; Russo and Van Roy, 2016b). Algorithm 1 can be applied as is when we can execute multiple experiments in parallel. We expect that similar results to Theorems 2 and 3 should hold, with mild dependence on the number of workers, using similar analyses to Kandasamy et al. (2018).

5. Conclusion

We study settings for adaptive goal oriented DOE problems in a Bayesian setting. Our formulation is quite general, allowing practitioners to incorporate domain knowledge via a probabilistic model, and specify their goal via a reward function that may depend on system characteristics. We focus on myopic policies due to their computational simplicity. Yet, our empirical results demonstrate that MPS has broad applicability, performing favourably with more specialised methods, and enabling complex DOE tasks where existing methods are not applicable. Our theoretical results establish conditions under which a myopic algorithm based on posterior sampling is competitive with myopic and globally optimal policies, both of which know the underlying system parameters. One interesting avenue for future theoretical work is to relax and/or find other conditions under which myopic strategies can do well. For instance, we believe that Condition 1 is stronger than necessary, and that it is sufficient if $\pi^\star_{\mathrm{M}}$ is able to do well in its own environment.

Acknowledgements

This research is partly funded by DOE grant DESC0011114, NSF grant IIS1563887, the Darpa D3M program, AFRL, and Toyota Research Institute, Accelerated Materials Design & Discovery (AMDD) program. KK is supported by a Facebook fellowship and a Siebel scholarship. AK is supported in part by NSF Award IIS-1763618. KK and AK would like to thank La Flamenca, Lanzarote for their delicious vegetarian paellas, which fueled the initial ideas for our theoretical analysis.

References

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *International Conference on Machine Learning*, pages 126–135, 2017.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *arXiv preprint arXiv:1810.09538*, 2018.

Adam D. Bull. Convergence Rates of Efficient Global Optimization Algorithms. *JMLR*, 2011.

Jonathan Kenneth Bunn, Jianjun Hu, and Jason R Hattrick-Simpers. Semi-supervised approach to phase identification from combinatorial sample diffraction patterns. *JOM*, 68(8):2116–2125, 2016.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015.

Yuxin Chen and Andreas Krause. Near-optimal batch mode active learning and adaptive submodular optimization. *ICML (1)*, 28:160–168, 2013.

Yuxin Chen, Hiroaki Shioi, Cesar Fuentes Montesinos, Lian Pin Koh, Serge Wich, and Andreas Krause. Active detection via adaptive submodularity. In *ICML*, pages 55–63, 2014.

Yuxin Chen, S Hamed Hassani, Andreas Krause, et al. Near-optimal bayesian active learning with correlated and noisy tests. *Electronic Journal of Statistics*, 11(2):4969–5017, 2017.

Herman Chernoff. *Sequential analysis and optimal design*. Siam, 1972.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

T. M. Davis et al. Scrutinizing Exotic Cosmological Models Using ESSENCE Supernova Data Combined with Other Cosmological Probes. *Astrophysical Journal*, 2007.

Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Peter I Frazier, Warren B Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47 (5):2410–2439, 2008.

Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on learning theory*, pages 728–763, 2015.

Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson, and S Muthukrishnan. Adaptive submodular maximization in bandit setting. In *Advances in Neural Information Processing Systems*, pages 2697–2705, 2013.

Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson, and S Muthukrishnan. Large-scale optimistic adaptive submodularity. In *AAAI*, pages 1816–1823, 2014.

Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *ICML*, pages 937–945, 2014.

Kevin L Gering. Prediction of electrolyte viscosity for aqueous and non-aqueous systems: Results from a molecular model based on ion solvation and a chemical physics framework. *Electrochimica Acta*, 51(15):3125–3138, 2006.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, pages 766–774, 2010.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898, 2015.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108, 2014.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active Learning for Level Set Estimation. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2013a.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *IJCAI*, pages 1344–1350, 2013b.

James Hensman, Magnus Rattray, and Neil D Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in neural information processing systems*, pages 2888–2896, 2012.

Daniel Hernández-Lobato, Jose Hernandez-Lobato, Amar Shah, and Ryan Adams. Predictive entropy search for multi-objective bayesian optimization. In *International Conference on Machine Learning*, pages 1492–1501, 2016.

José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. *arXiv preprint arXiv:1706.01825*, 2017.

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

Shali Jiang, Gustavo Malkomes, Matthew Abbott, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic batch active search. In *Advances in Neural Information Processing Systems*, pages 1107–1117, 2018.

Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. Bayesian Active Learning for Posterior Estimation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Jeff Schneider, and Barnabás Póczos. Parallelised bayesian optimisation via thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 133–142, 2018.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient thompson sampling for online matrix factorization recommendation. In *Advances in neural information processing systems*, pages 1297–1305, 2015.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49 (2-3):209–232, 2002.

Yao Liu and Emma Brunskill. When simple exploration is sample efficient: Identifying sufficient conditions for random exploration to yield pac rl algorithms. *arXiv:1805.09045*, 2018.

Yifei Ma, Tzu-Kuo Huang, and Jeff G Schneider. Active search and bandits on graphs using sigma-optimality. In *UAI*, pages 542–551, 2015.

Willie Neiswanger, Chong Wang, and Eric Xing. Embarrassingly parallel variational inference in nonconjugate models. *arXiv preprint arXiv:1510.04163*, 2015.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Michael Osborne, Roman Garnett, Zoubin Ghahramani, David K Duvenaud, Stephen J Roberts, and Carl E Rasmussen. Active learning of model evidence using bayesian quadrature. In *Advances in neural information processing systems*, pages 46–54, 2012.

Long Ouyang, Michael Henry Tessler, Daniel Ly, and Noah Goodman. Practical optimal experiment design with probabilistic programs. *arXiv preprint arXiv:1608.05046*, 2016.

Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible multi-objective bayesian optimization approach using random scalarizations. *arXiv preprint arXiv:1805.12168*, 2018.

D. Parkinson, P. Mukherjee, and A.. R Liddle. A Bayesian model selection analysis of WMAP3. *Physical Review*, 2006.

Tom Rainforth. *Automating Inference, Learning, and Design using Probabilistic Programming*. PhD thesis, PhD thesis, 2017.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

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

Daniel Russo and Benjamin Van Roy. An Information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research (JMLR)*, 2016b.

J. Snoek, H. Larochelle, and R. P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *NIPS*, 2012.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *ICML*, 2010.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

M. Tegmark et al. Cosmological Constraints from the SDSS Luminous Red Galaxies. *Physical Review*, December 2006.

W. R. Thompson. On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 1933.

Dustin Tran, Matthew D Hoffman, Rif A Saurous, Eugene Brevdo, Kevin Murphy, and David M Blei. Deep probabilistic programming. *arXiv preprint arXiv:1701.03757*, 2017.

Yingfei Wang and Warren B Powell. Finite-time analysis for the knowledge-gradient policy. *SIAM Journal on Control and Optimization*, 56(2):1105–1129, 2018.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963, 2015.

Rebecca Willett, Robert Nowak, and Rui M Castro. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems*, pages 179–186, 2006.

A. Some Information Theoretic Results

We will need the following technical results for our analysis. The first is a version of Pinsker's inequality.

Lemma 4 (Pinsker's inequality)**.** *Let $X, Z \in \mathcal{X}$ be random quantities and* $\sup f - \inf f \leq B$*. Then,* $\big|\mathbb{E}[f(X)] - \mathbb{E}[f(Z)]\big| \leq B\sqrt{\frac{1}{2}\mathrm{KL}(P(X)\|P(Z))}$.

The next, taken from Russo and Van Roy (2016b), relates the KL divergence to the mutual information for two random quantities X, Y.

Lemma 5 (Russo and Van Roy (2016b), Fact 6)**.** *For random quantities $X, Z \in \mathcal{X}$, $I(X; Z) = \mathbb{E}_X[\mathrm{KL}(P(Y|X)\|P(Y))]$.*

The next result is a property of the Shannon mutual information.

Lemma 6. *Let X, Y, Z be random quantities such that Y is a deterministic function of X. Then, $I(Y; Z) \leq I(X; Z)$.*

Proof. Let Y' capture the remaining randomness in X so that $X = Y \cup Y'$. Since conditioning reduces entropy, $I(Y; Z) = H(Z) - H(Z|Y) \leq H(Z) - H(Z|Y \cup Y') = I(X; Z)$. □

B. Proofs

B.1. Notation and Set up

In this subsection, we will introduce some notation, prove some basic lemmas, and in general, lay the groundwork for our analysis. $\mathbb{P}, \mathbb{E}$ denote probabilities and expectations. $\mathbb{P}_t, \mathbb{E}_t$ denote probabilities and expectations when conditioned on the actions and observations up to and including time t, e.g. for any event E, $\mathbb{P}_t(E) = \mathbb{P}(E|D_t)$. For two data sequences A, B, $A \uplus B$ denotes the concatenation of the two sequences. When $x \in \mathcal{X}$, Y_x will denote the random observation from $\mathbb{P}(Y|x, \theta)$.

Let $J_n(\theta_\star, \pi)$ denote the expected sum of cumulative rewards for fixed policy π after n evaluations under $\theta_\star$, i.e. $J_n(\theta_\star, \pi) = \mathbb{E}[\Lambda(\theta_\star, D_n)|\theta_\star, D_n \sim \pi]$ (Recall (1)). Let $D_t \in \mathcal{D}_t$ be a data sequence of length t. Then, $Q^\pi(D_t, x, y)$ will denote the expected sum of future rewards when, having collected the data sequence D_n, we take action $x \in \mathcal{X}$, observe $y \in \mathcal{Y}$ and then execute policy π for the remaining $n - t - 1$ steps. That is,

$$Q^\pi(D_t, x, y) = \lambda(\theta_\star, D_j \uplus \{(x, y)\}) + \mathbb{E}_{F_{t+2:n}}\left[\sum_{j=t+2}^{n} \lambda(\theta_\star, D_j \uplus \{(x, y)\} \uplus F_{t+2:j})\right]. \tag{4}$$

Here, the action-observation pairs collected by π from steps $t + 2$ to n are $F_{t+2:n}$. The expectation is over the observations and any randomness in π. While we have omitted for conciseness, Q^π is a function of the true parameter $\theta_\star$. Let d_π^t denote the distribution of D_t when following a policy π for the first t steps. We then have, for all $t \leq n$,

$$J_n(\theta_\star, \pi) = \mathbb{E}_{D_t \sim d_\pi^t}\left[\sum_{j=1}^{t} \lambda(\theta_\star, D_j)\right] + \mathbb{E}_{D_t \sim d_\pi^t}\left[\mathbb{E}_{X \sim \pi(D_t)}[Q^\pi(D_t, X, Y_X)]\right], \tag{5}$$

where, recall, Y_X is drawn from $\mathbb{P}(Y|X, \theta_\star)$. The following Lemma decomposes the regret $J_n(\theta_\star, \pi_{\mathrm{M}}^\star) - J_n(\theta_\star, \pi)$ as a sum of terms which are convenient to analyse. The proof is adapted from Lemma 4.3 in Ross and Bagnell (2014).

Lemma 7. *For any two policies π_1, π_2,*

$$J_n(\theta_\star, \pi_2) - J_n(\theta_\star, \pi_1) \;=\; \sum_{t=1}^{n} \mathbb{E}_{D_{t-1} \sim d_{\pi_1}^{t-1}}\left[\mathbb{E}_{X \sim \pi_1(D_{t-1})}\left[Q^{\pi_2}(D_{t-1}, X, Y_X)\right] - \mathbb{E}_{X \sim \pi_2(D_{t-1})}\left[Q^{\pi_2}(D_{t-1}, X, Y_X)\right]\right]$$

Proof. Let π^t be the policy that follows π_1 from time step 1 to t, and then executes policy π_2 from $t+1$ to n. Hence, by (5),

$$J_n(\theta_\star, \pi^t) = \mathbb{E}_{D_{t-1}\sim d_\pi^{t-1}}\left[\sum_{j=1}^{t-1}\lambda(\theta_\star, D_j)\right] + \mathbb{E}_{D_{t-1}\sim d_{\pi_1}^{t-1}}\left[\mathbb{E}_{X\sim\pi_1(D_{t-1})}[Q^{\pi_2}(D_{t-1}, X, Y_X)]\right],$$

$$J_n(\theta_\star, \pi^{t-1}) = \mathbb{E}_{D_{t-1}\sim d_\pi^{t-1}}\left[\sum_{j=1}^{t-1}\lambda(\theta_\star, D_j)\right] + \mathbb{E}_{D_{t-1}\sim d_{\pi_1}^{t-1}}\left[\mathbb{E}_{X\sim\pi_2(D_{t-1})}[Q^{\pi_2}(D_{t-1}, X, Y_X)]\right].$$

The claim follows from the observation, $J(\theta_\star, \pi_1) - J(\theta_\star, \pi_2) = J(\theta_\star, \pi^n) - J(\theta_\star, \pi^0) = \sum_{t=1}^n J(\theta_\star, \pi^t) - J(\theta_\star, \pi^{t-1})$. □

We will use Lemma 7 with π_2 as the policy $\pi_{\mathrm{M}}^\star$ which knows $\theta_\star$ and with π_1 as the policy π whose regret we wish to bound. For this, denote the action chosen by π when it has seen data D_{t-1} as X_t and that taken by $\pi_{\mathrm{M}}^\star$ as X_t'. By Lemma 7 and equation (4) we have,

$$\begin{aligned}\mathbb{E}_{\theta_\star}[J_n(\theta_\star, \pi_{\mathrm{M}}^\star) - J_n(\theta_\star, \pi)] &= \sum_{t=1}^n \mathbb{E}_{D_{t-1}}\left[\mathbb{E}_{t-1}\left[Q^{\pi_{\mathrm{M}}^\star}(D_{t-1}, X_t', Y_{X_t'}) - Q^{\pi_{\mathrm{M}}^\star}(D_{t-1}, X_t, Y_{X_t})\right]\right]\\ &= \mathbb{E}\sum_{t=1}^n \mathbb{E}_{t-1}\left[q_t(\theta_\star, X_t', Y_{X_t'}) - q_t(\theta_\star, X_t, Y_{X_t})\right],\end{aligned} \tag{6}$$

where we have defined

$$q_t(\theta_\star, x, y) = Q^{\pi_{\mathrm{M}}^\star}(D_{t-1}, x, y). \tag{7}$$

Note that the randomness in q_t stems from its dependence on $\theta_\star$ and future observations.

B.2. Proof of Theorem 2

We will let $\tilde{\mathbb{P}}_{t-1}$ denote the distribution of X_t given D_{t-1}; i.e. $\tilde{\mathbb{P}}_{t-1}(\cdot) = \mathbb{P}_{t-1}(X_t = \cdot)$. The density (Radon-Nikodym derivative) $\tilde{p}_{t-1}$ of $\tilde{\mathbb{P}}_{t-1}$ can be expressed as $\tilde{p}_{t-1}(x) = \int_\Theta p_\star(x|\theta_\star = \theta)p(\theta_\star = \theta|D_{t-1})\mathrm{d}\theta$ where $p_\star(x|\theta_\star = \theta)$ is the density of the maximiser of λ given $\theta_\star = \theta$ and $p(\theta_\star = \cdot|D_{t-1})$ is the posterior density of $\theta_\star$ conditoned on D_{t-1}. Note that $p_\star(x|\theta_\star = \theta)$ puts all its mass at the maximiser of $\lambda^+(\theta, D_{t-1}, x)$. Hence, X_t has the same distribution as X_t'; i.e. $\mathbb{P}_{t-1}(X_t' = \cdot) = \tilde{\mathbb{P}}_{t-1}(\cdot)$. This will form a key intuition in our analysis. To this end, we begin with a technical result, whose proof is adapted from Russo and Van Roy (2016b). We will denote by $\mathrm{I}_{t-1}(A; B)$ the mutual information between two variables A, B under the posterior measure after having seen D_{t-1}; i.e. $\mathrm{I}_{t-1}(A; B) = \mathrm{KL}(\mathbb{P}_{t-1}(A, B)\|\mathbb{P}_{t-1}(A)\cdot\mathbb{P}_{t-1}(B))$.

Lemma 8. *Assume that we have collected a data sequence D_{t-1}. Let the action taken by $\pi_{\mathrm{M}}^{\mathrm{PS}}$ at time instant t with D_{t-1} be X_t and the action taken by $\pi_{\mathrm{M}}^\star$ be X_t'. Then,*

$$\mathbb{E}_{t-1}[q_t(\theta_\star, X_t', Y_{X_t'}) - q_t(\theta_\star, X_t, Y_{X_t})] = \sum_{x\in\mathcal{X}}\big(\mathbb{E}_{t-1}[q_t(\theta_\star, x, Y_x)|X_t' = x] - \mathbb{E}_{t-1}[q_t(\theta_\star, x, Y_x)]\big)\tilde{\mathbb{P}}_{t-1}(x)$$

$$\mathrm{I}_{t-1}(X_t'; (X_t, Y_{X_t})) = \sum_{x_1, x_2\in\mathcal{X}}\mathrm{KL}(\mathbb{P}_{t-1}(Y_{x_1}|X_t' = x_2)\|\mathbb{P}_{t-1}(Y_{x_1}))\,\tilde{\mathbb{P}}_{t-1}(x_1)\tilde{\mathbb{P}}_{t-1}(x_2)$$

Proof. The proof for both results uses the fact that $\mathbb{P}_{t-1}(X_t = x) = \mathbb{P}_{t-1}(X_t' = x) = \tilde{\mathbb{P}}_{t-1}(x)$. For the first result,

$$\begin{aligned}&\mathbb{E}_{t-1}[q_t(\theta_\star, X_t', Y_{X_t'}) - q_t(\theta_\star, X_t, Y_{X_t})]\\ &= \sum_{x\in\mathcal{X}}\mathbb{P}_{t-1}(X_t' = x)\mathbb{E}_{t-1}[q_t(\theta_\star, X_t', Y_{X_t'})|X_t' = x] - \sum_{x\in\mathcal{X}}\mathbb{P}_{t-1}(X_t = x)\mathbb{E}_{t-1}[q_t(\theta_\star, X_t, Y_{X_t})|X_t = x]\\ &= \sum_{x\in\mathcal{X}}\mathbb{P}_{t-1}(X_t' = x)\mathbb{E}_{t-1}[q_t(\theta_\star, x, Y_x)|X_t' = x] - \sum_{x\in\mathcal{X}}\mathbb{P}_{t-1}(X_t = x)\mathbb{E}_{t-1}[q_t(\theta_\star, x, Y_x)]\\ &= \sum_{x\in\mathcal{X}}\big(\mathbb{E}_{t-1}[q_t(\theta_\star, x, Y_x)|X_t' = x] - \mathbb{E}_{t-1}[q_t(\theta_\star, x, Y_x)]\big)\tilde{\mathbb{P}}_{t-1}(x)\,.\end{aligned}$$

The second step uses that the observation Y_x does not depend on the fact that x may have been chosen by $\pi_{\mathrm{M}}^{\mathrm{PS}}$; this is because $\pi_{\mathrm{M}}^{\mathrm{PS}}$ makes its decisions based on past data D_{t-1} and is independent of $\theta_\star$ given D_{t-1}. Y_x however can depend on the fact that x may have been the action chosen by $\pi_{\mathrm{M}}^{\star}$ which knows $\theta_\star$. For the second result,

$$\begin{aligned}
&\mathrm{I}_{t-1}(X'_t;(X_t,Y_{X_t})) = \mathrm{I}_{t-1}(X'_t;X_t) + \mathrm{I}_{t-1}(X'_t;Y_{X_t}|X_t) = \mathrm{I}_{t-1}(X'_t;Y_{X_t}|X_t)\\
&\quad= \sum_{x_1\in\mathcal{X}} \mathbb{P}_{t-1}(X_t=x_1)\,\mathrm{I}_{t-1}(X_t;Y_{X_t}|X_t=x) = \sum_{x_1\in\mathcal{X}} \tilde{\mathbb{P}}_{t-1}(x_1)\,\mathrm{I}_{t-1}(X'_t;Y_{x_1})\\
&\quad= \sum_{x_1\in\mathcal{X}} \tilde{\mathbb{P}}_{t-1}(x_1) \sum_{x_2\in\mathcal{X}} \mathbb{P}_{t-1}(X'_t=x_2)\,\mathrm{KL}(\mathbb{P}_{t-1}(Y_{x_1}|X'_t=x_2)\|\mathbb{P}_{t-1}(Y_{x_1}))\\
&\quad= \sum_{x_1,x_2\in\mathcal{X}} \mathrm{KL}(\mathbb{P}_{t-1}(Y_{x_1}|X'_t=x_2)\|\mathbb{P}_{t-1}(Y_{x_1}))\,\tilde{\mathbb{P}}_{t-1}(x_1)\tilde{\mathbb{P}}_{t-1}(x_2)
\end{aligned}$$

The first step uses the chain rule for mutual information. The second step uses that X_t is chosen based on an external source of randomness and D_{t-1}; therefore, it is independent of $\theta_\star$ and hence X'_t given D_{t-1}. The fourth step uses that Y_{x_1} is independent of X_t. The fifth step uses lemma 5 in Appendix A. □

We are now ready to prove theorem 2.

Proof of Theorem 2: Using the first result of Lemma 8, we have,

$$\begin{aligned}
&\mathbb{E}_{t-1}[q_t(\theta_\star,X'_t,Y_{X'_t}) - q_t(\theta_\star,X_t,Y_{X_t})]^2\\
&\quad= \Big(\sum_{x\in\mathcal{X}} \tilde{\mathbb{P}}_{t-1}(x)\big(\mathbb{E}_{t-1}[q_t(\theta_\star,x,Y_x)|X'_t=x] - \mathbb{E}_{t-1}[q_t(\theta_\star,x,Y_x)]\big)\Big)^2\\
&\quad\overset{(a)}{\leq} |\mathcal{X}| \sum_{x\in\mathcal{X}} \tilde{\mathbb{P}}_{t-1}(x)^2\big(\mathbb{E}_{t-1}[q_t(\theta_\star,x,Y_x)|X'_t=x] - \mathbb{E}_{t-1}[q_t(\theta_\star,x,Y_x)]\big)^2\\
&\quad\overset{(b)}{\leq} |\mathcal{X}| \sum_{x_1,x_2\in\mathcal{X}} \tilde{\mathbb{P}}_{t-1}(x_1)\tilde{\mathbb{P}}_{t-1}(x_2)\big(\mathbb{E}_{t-1}[q_t(\theta_\star,x_1,Y_{x_1})] - \mathbb{E}_{t-1}[q_t(\theta_\star,x_1,Y_{x_1})|X'_t=x_2]\big)^2\\
&\quad\overset{(c)}{\leq} |\mathcal{X}| \sum_{x_1,x_2\in\mathcal{X}} \tilde{\mathbb{P}}_{t-1}(x_1)\tilde{\mathbb{P}}_{t-1}(x_2)\mathbb{E}_{Y_{x_1}}\Big[\big(\mathbb{E}_{t-1}[q_t(\theta_\star,x_1,y)|Y_{x_1}=y] - \mathbb{E}_{t-1}[q_t(\theta_\star,x_1,y)|X'_t=x_2,Y_{x_1}=y]\big)^2\Big] \qquad (8)\\
&\quad\overset{(d)}{\leq} \frac{|\mathcal{X}|}{2} \sum_{x_1,x_2\in\mathcal{X}} \tau_{n-t}\tilde{\mathbb{P}}_{t-1}(x_1)\tilde{\mathbb{P}}_{t-1}(x_2)\mathbb{E}_{Y_{x_1}}\left[\mathrm{KL}(\mathbb{P}_{t-1}(Y_{x_1}|X'_t=x_2,Y_{x_1}=y)\|\mathbb{P}_{t-1}(Y_{x_1}|Y_{x_1}=y))\right]\\
&\quad\overset{(e)}{\leq} \frac{|\mathcal{X}|}{2} \sum_{x_1,x_2\in\mathcal{X}} \tau_{n-t}\tilde{\mathbb{P}}_{t-1}(x_1)\tilde{\mathbb{P}}_{t-1}(x_2)\mathrm{KL}(\mathbb{P}_{t-1}(Y_{x_1}|X'_t=x_2)\|\mathbb{P}_{t-1}(Y_{x_1}))\\
&\quad\overset{(f)}{=} \frac{1}{2}|\mathcal{X}|\tau_n\mathrm{I}_{t-1}(X'_t;(X_t,Y_{X_t})) \overset{(g)}{\leq} \frac{1}{2}|\mathcal{X}|\tau_n\mathrm{I}_{t-1}(\theta_\star;(X_t,Y_{X_t}))
\end{aligned}$$

Here, step (a) uses the Cauchy-Schwarz inequality and step (b) uses the fact that the previous line can be viewed as the diagonal terms in a sum over x_1, x_2. Step (c) conditions on $Y_{x_1} = y$ and applies Jensen's inequality. Step (e) uses the definition of conditional KL divergence. Step (f) uses the second result of Lemma 8, and step (g) uses Lemma 6 and the fact that X'_t is a deterministic function of $\theta_\star$ given D_{t-1}. For step (d), we use the version of Pinsker's inequality given in Lemma 4 in conjunction with Condition 1. Precisely, we let H in Condition 1 to be $D_{t-1} \uplus \{(x,y)\}$. Now using (7) and (4), and the fact that $\pi_{\mathrm{M}}^{\star}$ is deterministic, we can write,

$$\begin{aligned}
&q_t(\theta_1,x,y) - q_t(\theta_2,x,y)\\
&\quad= \lambda(\theta_1,D_{t-1}\uplus\{(x,y)\}) - \lambda(\theta_2,D_{t-1}\uplus\{(x,y)\}) +\\
&\qquad\quad \sum_{j=1}^{n} \mathbb{E}_{Y,t+1:n|\theta_1}\left[\lambda(\theta_1,D_{t-1}\uplus\{(x,y)\}\uplus F_{j,1})\right] - \mathbb{E}_{Y,t+1:n|\theta_2}\left[\lambda(\theta_2,D_{t-1}\uplus\{(x,y)\}\uplus F_{j,2})\right]
\end{aligned}$$

$$\le 1 \;+\; \sum_{t=1}^{n} \epsilon_t \le \sqrt{\tau_{n-t}}.$$

Here, $F_{n,i}$ is the data collected by $\pi_{\mathrm{M}}^{\star}$ when $\theta_\star = \theta_i$, having observed H, and $F_{j,i}$ is its prefix of length j. The last step uses Condtion 1. Hence, by Lemma 4, the term with the squared paranthesis in (8) can be bounded by $\tau_{n-t}\mathrm{KL}(\mathbb{P}_{t-1}(Y_{x_1}|X'_t = x_2)\|\mathbb{P}_{t-1}(Y_{x_1}))$.

Now, using (6) and the Cauchy-Schwarz inequality we have,

$$\mathbb{E}[J_n(\theta_\star, \pi_{\mathrm{M}}^{\star}) - J_n(\theta_\star, \pi_{\mathrm{M}}^{\mathrm{PS}})]^2 \le n \sum_{t=1}^{n} \frac{1}{2}|\mathcal{X}|\tau_n \mathrm{I}_{t-1}(\theta_\star; (X_t, Y_{X_t})) = \frac{1}{2}|\mathcal{X}|\tau_n \mathrm{I}(\theta_\star; D_n)$$

Here the last step uses the chain rule of mutual information in the following form,

$$\sum_t \mathrm{I}_{t-1}(\theta_\star; (X_t, Y_{X_t})) = \sum_t \mathrm{I}(\theta_\star; (X_t, Y_{X_t})|\{(X_j, Y_{X_j})\}_{j=1}^{t-1}) = \mathrm{I}(\theta_\star; \{(X_j, Y_{X_j})\}_{j=1}^{n}).$$

The claim follows from the observation, $\mathrm{I}(\theta_\star; D_n) \le \Psi_n$. □

B.3. Proof of Theorem 3

In this section, we will let $D_m^{\star\star}$ be the data collected $\pi_{\mathrm{G}}^{\star}$ in m steps and $D_n^{\star}$ be the data collected by $\pi_{\mathrm{M}}^{\star}$ in n steps. We will use the following result on adaptive submodular maximisation from (Golovin and Krause, 2011).

Lemma 9. *(Theorem 38 in Golovin and Krause (2011), modified) Under condition 2, we have for all $\theta_\star \in \Theta$,*

$$\mathbb{E}_Y[\lambda(\theta_\star, D_n^{\star})] \ge (1 - e^{-n/m})\mathbb{E}_Y[\lambda(\theta_\star, D_m^{\star\star})]$$

Lemma 10 controls the approximation error when we approximate the globally optimal policy which knows $\theta_\star$ with the myopic policy which knows $\theta_\star$. Our proof of theorem 3, combines the above result with Theorem 2, to show that MPS can approximate $\pi_{\mathrm{G}}^{\star}$ under suitable conditions.

Proof of Theorem 3. Let D_n be the data collected by $\pi_{\mathrm{M}}^{\mathrm{PS}}$. By monotonicity of λ, and the fact that the maximum is larger than the average we have $\mathbb{E}[\lambda(\theta_\star, D_n)] \ge \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[\lambda(\theta_\star, D_t)] = \frac{1}{n}\mathbb{E}[\Lambda(\theta_\star, D_n)]$. Using theorem 2 the following holds for all m,

$$\begin{aligned}
\mathbb{E}[\lambda(\theta_\star, D_n)] &\ge \frac{1}{n}\left(\mathbb{E}\left[\Lambda(\theta_\star, D_n^{\star})\right] - \sqrt{\frac{|\mathcal{X}|\tau_n n \Psi_n}{2}}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\theta_\star}\left[\mathbb{E}_Y[\lambda(\theta_\star, D_t^{\star})]\right] - \sqrt{\frac{|\mathcal{X}|\tau_n \Psi_n}{2n}} \\
&\ge \mathbb{E}[\lambda(\theta_\star, D_m^{\star\star})]\frac{1}{n}\sum_{t=1}^{n}(1 - e^{-t/m}) - \sqrt{\frac{|\mathcal{X}|\tau_n \Psi_n}{2n}} \\
&\ge \mathbb{E}[\lambda(\theta_\star, D_m^{\star\star})](1 - \frac{m}{n}e^{-1/m} - \frac{1}{n}e^{-1/m}) - \sqrt{\frac{|\mathcal{X}|\tau_n \Psi_n}{2n}}.
\end{aligned}$$

Here, the first step uses Theorem 2, the second step rearranges the expectations noting that λ takes the expectation over the observations. The third step uses Lemma 9 for each t. The last step bounds the sum by an integral as follows,

$$\sum_{t=1}^{n} e^{-t/m} \le e^{-1/m} + \int_1^\infty e^{-t/m}\mathrm{d}t \le e^{-1/m} + me^{-1/m}.$$

The result follows by using $m = \gamma n$. □

B.4. Proof of Lower Bound (Proposition 1)

Consider a setting with uniform prior over two parameters θ_0, θ_1 with two actions X_0, X_1. Set $\lambda(\theta_i, D) = \mathbf{1}\{X_i \notin D\}$. If $\theta_\star = \theta_0$, then $\pi^\star_{\mathrm{M}}$ will repeatedly choose X_1 and achieve reward 1 on every time step, and similarly when $\theta_\star = \theta_1$. On the other hand, conditioned on any randomness of the decision maker (which is external to the randomness of the prior and the observations), the first decision for the decision maker must be the same for both choices of $\theta_\star$. Hence, for one of the two choices for $\theta_\star$, $\lambda(\theta_\star, D_n) = 0$ for all n. Since the prior is equal on both θ_0, θ_1, the average instantaneous regret is at least $1/2$. □

C. On Condition 1

The following proposition shows that when the myopic policy has value 1, and achieves this at a fast enough rate, for all values of θ, we satisfy Condition 1. For this, let $\theta, \theta', \pi^\theta_{\mathrm{M}}, \pi^{\theta'}_{\mathrm{M}}, D_n, D'_n, \mathbb{E}_{Y,t+1:}$ be as defined in Condition 1.

Proposition 10. *($\pi^\star_{\mathrm{M}}$ has value 1). Let π^θ_{M} denote the myopically optimal policy when $\theta_\star = \theta$. Assume there exists a sequence $\{\epsilon'_n\}_{n\geq 1}$ such that,*

$$\sup_{\theta\in\Theta}\sup_{H\in\mathcal{D}} \big(1 - \mathbb{E}_{Y,|H|+1}[\lambda(\theta, H \uplus D_n)]\big) \leq \epsilon'_n.$$

Then, Condition 1 is satisfied with $\epsilon_n = \epsilon'_n$.

Proof. Let $H \in \mathcal{D}$ and $\theta, \theta' \in \Theta$. Then,

$$\begin{aligned}
&\mathbb{E}_{Y,|H|+1|\theta}\lambda(\theta, H \uplus D_n) - \mathbb{E}_{Y,|H|+1|\theta'}\lambda(\theta', H \uplus D'_n) \\
&\quad= \Big(\mathbb{E}_{Y,|H|+1|\theta}\lambda(\theta, H \uplus D_n) - 1\Big) + \Big(1 - \mathbb{E}_{Y,|H|+1|\theta'}\lambda(\theta', H \uplus D'_n)\Big) \;\leq\; \epsilon'_n,
\end{aligned}$$

since the first term is always negative. □

We next show two examples of DOE problems where the condition in Proposition 10 is satisfied.

C.1. Bandits & Bayesian Optimisation

In both settings, the parameter $\theta_\star$ specifies a function $f_{\theta_\star} : \mathcal{X} \to \mathbb{R}$. When we choose a point $X \in \mathcal{X}$ to evaluate the function, we observe $Y_X = f_{\theta_\star}(X) + \epsilon$ where $\mathbb{E}[\epsilon] = 0$. In the bandit framework, we can define the reward to be $\lambda(\theta_\star, D_n) = 1 + f_{\theta_\star}(X_n) - \max_{x\in\mathcal{X}} f_{\theta_\star}(x)$ which is equivalent to maximising the instantaneous reward. In Bayesian optimisation, one is interested in simply finding a single value close to the optimum and hence $\lambda(\theta_\star, D_n) = 1 + \max_{t\leq n} f_{\theta_\star}(X_t) - \max_{x\in\mathcal{X}} f_{\theta_\star}(x)$.

In both cases, since $\pi^\star_{\mathrm{M}}$ knows it will always choose $\mathrm{argmax}_{x\in\mathcal{X}} f_{\theta_\star}(x)$ achieving reward 1. Thus Proposition 10 is satisfied with $\epsilon_n = 0$ and $\tau_n = 1$.

C.2. An Active Learning Example

We describe an active learning task on a Bayesian linear regression problem, and outline how it can be formulated to satisfy the conditions in Section 4.

In this example, our parameter space is $\Theta = \{\theta = (\beta, \eta^2) | \beta \in \mathbb{R}^k, \eta^2 \in [a, b]\}$ for some positive numbers $b > a > 0$. We will assume the following prior on $\theta_\star = (\beta_\star, \eta^2_\star)$,

$$\beta_\star \sim \mathcal{N}(\mathbf{0}_k, \mathrm{P}_0^{-1}), \quad \eta^2_\star \sim \mathrm{Unif}(a, b),$$

where $\mathrm{P}_0 \in \mathbb{R}^{k\times k}$ is the non-singular precision matrix of the Gaussian prior for $\beta_\star$. Our domain $\mathcal{X} = \{x \in \mathbb{R}^k; \|x\|_2 \leq 1\}$ is the unit ball in $\mathbb{R}^k$ and $\mathcal{Y} = \mathbb{R}$. When we query the model at $x \in \mathcal{X}$, we observe $Y_x = \beta^\top x + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \eta^2)$. Our goal in DOE is to choose a sequence of experiments $\{X_t\}_t \subset \mathcal{X}$ so as to estimate β well.

Given a dataset $D_n = \{(x_j, y_j)\}_{j=1}^n$, a natural quantity to characterise how well we have estimated $\beta_\star$ in the Bayesian setting is via the entropy of the posterior for β. This ensures that the data is sampled also considering the uncertainty in the prior. For example, if the prior covariance is small along certain directions, an active learning agent is incentivised to collect data so as to minimise the variance along other directions. Specifically, in this example, we wish to minimise $\mathrm{H}(\beta_\star | D_n = D_n, \eta_\star^2 = \eta_\star^2)$, the entropy of $\beta_\star$ assuming we have collected data D_n and the true $\eta_\star^2$ value were to be revealed at the end. It is straightforward to see that, $\mathbb{P}(\beta_\star|\eta_\star^2, D_n) = \mathcal{N}(\mu_n, \mathrm{P}_n^{-1})$, where,

$$\mathrm{P}_n = \mathrm{P}_0 + \frac{1}{\eta_\star^2}\sum_{j=1}^n x_j x_j^\top, \qquad \mu_n = \mathrm{P}_n \sum_{j=1}^n y_j x_j.$$

The entropy of this posterior is

$$\mathrm{H}(\beta_\star|D_n = D_n, \eta_\star^2 = \eta_\star^2) \;= \frac{1}{2}\log\det(2\pi e \mathrm{P}_n^{-1}) \;= \frac{k}{2}\log(2\pi e) - \frac{1}{2}\log\det \mathrm{P}_n.$$

Minimising the posterior entropy can be equivalently formulated as maximising the following reward function,

$$\lambda(\theta_\star, D_n) = 1 - \frac{1}{\det \mathrm{P}_n} = 1 - \frac{1}{\det\left(\mathrm{P}_0 + \frac{1}{\eta_\star^2}\sum_{j=1}^n x_j x_j^\top\right)}. \tag{9}$$

The reward depends on $\theta_\star$ due to the $\eta_\star^2$ term, and an adaptive policy can be expected to do better than a non-adaptive one since the observations $\{y_j\}_{j=1}^n$ can inform us about the true value of $\eta_\star^2$.

Note that since $\lambda(\theta_\star, D_n)$ is a multi-set function, D_n can be viewed as a (non-ordered) mulit-set and the $\uplus$ operator is simply the union operator. We will now demonstrate that λ satisfies the two conditions set out in Section 4.

Condition 1: We will show that it satisfies the condition in Proposition 10. Let c be the smallest eigenvalue of P_0. For a given data set $H = \{(x_j, y_j)\}_{j=1}^m$ of size m, denote $\mathrm{P}_0^H = \mathrm{P}_0 + \frac{1}{\eta_\star^2}\sum_{i=1}^m x_j x_j^\top$. Moreover, assume that the points chosen by $\pi_{\mathrm{M}}^\star$ in $\mathcal{X}$ are $z_1, z_2, \ldots$. Note that this is a deterministic sequence since $\pi_{\mathrm{M}}^\star$ knows $\eta_\star^2$ and the reward does not depend on the observations.

Let $\mathrm{P}_n^H = \mathrm{P}_0^H + \frac{1}{\eta_\star^2}\sum_{i=1}^n z_j z_j^\top$ and denote its eigenvalues by $\sigma_1 > \sigma_2 > \cdots > \sigma_k$. Note that since the myopic policy chooses actions to maximise the reward at the next step, it will choose $z_{n+1} = \mathrm{argmax}_{\|z\|=1} \det(\mathrm{P}_n^H + \frac{1}{\eta_\star^2} z z^\top)$. We therefore have,

$$\det \mathrm{P}_{n+1}^H = \max_{\|z\|=1} \det\left(\mathrm{P}_n^H + \frac{1}{\eta_\star^2} z z^\top\right) \geq \left(\sigma_1 + \frac{1}{\eta_\star^2}\right)\prod_{j=2}^k \sigma_j$$

Noting that $\mathrm{P}_0^H - cI_k$ is positive definite, we have, via an inductive argument $\det \mathrm{P}_n^H \geq c^{k-1}(c + n\eta_\star^{-2})$. Letting $D_n^\star$ be the data collected by $\pi_{\mathrm{M}}^\star$, we have

$$1 - \lambda(\theta_\star, D_n^\star) \leq \frac{1}{c^{k-1}(c + nb)} \triangleq \epsilon_n',$$

as $\eta_\star^2 \leq b$. This leads to $\epsilon_n', \epsilon_n \in \mathcal{O}(1/n)$ and hence $\tau_n \in \mathcal{O}(\log n)$ in Proposition 10 and Condition 1. We next look at the adaptive submodularity condition.

Condition 2 (Adaptive Submodularity): Let $D_n = \{(x_j, y_j)\}_{j=1}^n$ $D_m = \{(x_j, y_j)\}_{j=1}^m$ be two data sets such that $D_m \subset D_n$ and $m < n$. Let $Q_m = \mathrm{P}_0 + \frac{1}{\eta_\star^2}\sum_{j=1}^n x_j x_j^\top$ and $Q_n = \mathrm{P}_0 + \frac{1}{\eta_\star^2}\sum_{j=1}^m x_j x_j^\top = Q_m + \frac{1}{\eta_\star^2}\sum_{j=m+1}^n x_j x_j^\top$. Let (x, Y_x) be a new observation. We then have,

$$\begin{aligned}\mathbb{E}[\lambda(\theta_\star, D_n \uplus \{(x, Y_x)\})] - \lambda(\theta_\star, D_n) &= \frac{1}{\det(Q_n)} - \frac{1}{\det(Q_n + xx^\top)} \\ &= \frac{\det(Q_n + xx^\top) - \det(Q_n)}{\det(Q_n)\det(Q_n + xx^\top)} \;= \frac{x^\top Q_n^{-1} x}{\det(Q_n + xx^\top)},\end{aligned}$$

and similarly for Q_m. Here the last step uses the identity $\det(A + uv^\top) = \det(A)(1 + v^\top A^{-1} u)$. Submodularity follows by observing that Q_m, Q_n are positive definite and $Q_n - Q_m$ is positive semidefinite. Hence,

$$\frac{1 + x^\top Q_m^{-1} x}{\det(Q_m + xx^\top)} \geq \frac{1 + x^\top Q_n^{-1} x}{\det(Q_n + xx^\top)}.$$

C.3. Rewards with State-like structure

Here, we will show that $\pi_{\mathrm{M}}^{\mathrm{PS}}$ can achieve sublinear regret with respect to $\pi_{\mathrm{M}}^{\star}$, when there is additional structure in the rewards. In particular, we will assume that there exists a set of "states" $\mathcal{S}$ and a mapping $\sigma : \Theta \times \mathcal{D} \to \mathcal{S}$ from parameter, data sequence pairs to states. Moreover, λ takes the form $\lambda(\theta_\star, D) = \lambda_S(\theta_\star, \sigma(\theta_\star, D))$ for some known function $\lambda_S : \Theta \times \mathcal{S} \to [0, 1]$. We will also assume that the state transitions are Markovian, in that for any $S \in \mathcal{S}$, let $D_S = \{D \in \mathcal{D} : \sigma(\theta_\star, D) = S\}$. Then, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $D, D' \in D_S$, $\sigma(\theta_\star, D \cup \{(x, y)\}) = \sigma(\theta_\star, D' \cup \{(x, y)\})$.

Now, for any policy π, define,

$$\begin{aligned} V_n(\pi, D; \theta) &= \frac{1}{n}\mathbb{E}\left[\sum_{j=1}^{n} \lambda(\theta, D \uplus D_j) \,\middle|\, \theta_\star = \theta, D, D_n \sim \pi \right] \\ V(\pi, D; \theta) &= \lim_{n \to \infty} V_n(\pi, D; \theta) \end{aligned}$$

V_n is the expected sum of future rewards in n steps for a policy π when $\theta_\star = \theta$, and it starts from a prefix D. The expectation is over the observations and any randomness in π. V is the limit of V_n. A common condition used in reinforcement learning is that the associated Markov chain mixes when starting from any state $S \in \mathcal{S}$. Under this condition, V does not depend on the prefix D and we will simply denote it by $V(\pi; \theta)$. We have the following result.

Proposition 11. *Assume that there exists a sequence $\{\nu_n\}_{n \geq 1}$, such that $\nu_n \in o(1/\sqrt{n})$, and the following two statements are true.*

1. $V(\pi_{\mathrm{M}}^{\theta}; \theta) = V(\pi_{\mathrm{M}}^{\theta'}; \theta')$ *for all $\theta, \theta' \in \Theta$.*
2. *For all θ, and all data sequences H, H', $|V_n(\pi_{\mathrm{M}}^{\theta}, H; \theta) - V(\pi_{\mathrm{M}}^{\theta}; \theta)| \leq \nu_n$.*

Then Theorem 2 holds with $\sqrt{\tau_n} = 1 + 2n\nu_n$.

The second condition is similar to the requirements in Definition 5 in (Kearns and Singh, 2002). However, while they only use a thresholding behaviour, we assume a uniform rate of convergence, where our bounds depend on this rate. However, while results for non-episodic RL settings are given in terms of the mixing characteristics of the globally optimal policy, our results are in terms of the myopic policy.

Proof of Proposition 11. We will turn to our proof of Theorem 2, where we need to bound $q_t(\theta_1, x, y) - q_t(\theta_2, x, y)$. We will use Proposition 11 with $H = D_{t-1} \uplus \{(x, y)\}$ and have,

$$\begin{aligned} &q_t(\theta_1, x, y) - q_t(\theta_2, x, y) \\ &\quad = \lambda(\theta_1, D_{t-1} \uplus \{(x, y)\}) - \lambda(\theta_2, D_{t-1} \uplus \{(x, y)\}) + \\ &\qquad\quad \sum_{j=1}^{n} \mathbb{E}_{Y,t+1:n|\theta_1}\big[\lambda(\theta_1, D_{t-1} \uplus \{(x, y)\} \uplus F_{j,1})\big] - \mathbb{E}_{Y,t+1:n|\theta_2}\big[\lambda(\theta_2, D_{t-1} \uplus \{(x, y)\} \uplus F_{j,2})\big] \\ &\quad \leq 1 + (n - t)\left(V_n(\pi_{\mathrm{M}}^{\theta}, D_{t-1} \uplus \{(x, y)\}; \theta) - V_{n-t}(\pi_{\mathrm{M}}^{\theta'}, D_{t-1} \uplus \{(x, y)\}; \theta')\right) \\ &\quad \leq 1 + (n - t)\left(|V_{n-t}(\pi_{\mathrm{M}}^{\theta}, D_{t-1} \uplus \{(x, y)\}; \theta) - V(\pi_{\mathrm{M}}^{\theta}; \theta')| + |V_{n-t}(\pi_{\mathrm{M}}^{\theta'}, D_{t-1} \uplus \{(x, y)\}; \theta') - V(\pi_{\mathrm{M}}^{\theta'}; \theta')|\right) \\ &\quad \leq 1 + 2(n - t)\nu_{n-t} = \sqrt{\tau_{n-1}} \end{aligned}$$

Here, the second step uses that λ is bounded in $[0, 1]$, the third step simply uses the first condition in Proposition 11 along with the triangle inequality, and the fourth step uses the second condition. The remainder of the proof carries through by applying Pinksker's inequality with this bound in (8). □

Conditions of the above form are necessary in non-episodic undiscounted settings for RL (Kearns and Singh, 2002), and we show that under similar conditions, $\pi_{\mathrm{M}}^{\mathrm{PS}}$ achieves sublinear regret with $\pi_{\mathrm{M}}^{\star}$.

D. Some Experimental Details

Specification of the prior: In our experiments, we use a fixed prior in all our applications. In real world applications, the prior could be specified by a domain expert with knowledge of the given DOE problem. In some instances, the expert may only be able to specify the relations between the various variables involved. In such cases, one can specify the parametric form for the prior, and learn the parameters of the prior in an adaptive data dependent fashion using maximum likelihood and/or maximum a posteriori techniques (Snoek et al., 2012).

Computing the posterior: Experiments 2 and 4 which use a Bayesian linear regression model admit analytical computation of the posterior. So do experiments 5 and 6 which use a Gaussian process model. For experiments 1, 3, and 7 we use the Edward probabilistic programming framework (Tran et al., 2017) for a variational approximation of the posterior. The sample in step 3 is drawn from this approximation.

Optimising λ^+: In all our experiments, the look-ahead reward (2) is computed empirically by drawing 50 samples from $Y|X, \theta$ for the sampled θ and a given $x \in \mathcal{X}$. For experiments 1 and 3 which are one dimensional, we maximise λ^+ by evaluating it on a fine grid of size 100 and choosing the maximum. Similarly, for experiments 2 and 4 which have two dimensional domains, we use a grid of size 2500 and for experiments 5 and 7 which are three dimensional, we use a grid of size 8000. Since experiment 6 is in nine dimensions, on each iteration, we sample 4000 points randomly from the domain and choose the maximum.

Synthetic Active Learning Experiments: In all 4 experiments, the observations are generated from the true model. In the log likelihood formalism of Experiments 3 and 4, in order to compute the reward λ, we evaluate the expecation over $X \sim \Gamma, Y \sim \mathbb{P}(\cdot|X, \theta)$ empirically by drawing 1000 (x, y) pairs; we first sample 1000 x values uniformly at random and then draw y from the likelihood for the given θ value.

Level Set Estimation on LRGs: Here we used data on Luminous Red Galaxies (LRGs) to compute the galaxy power spectrum of 9 cosmological parameters: spatial curvature $\Omega_k \in (-1, 0.9)$, dark energy fraction $\Omega_\Lambda \in (0, 1)$, cold dark matter density $\omega_c \in (0, 1.2)$, baryonic density $\omega_B \in (0.001, 0.25)$, scalar spectral index $n_s \in (0.5, 1.7)$, scalar fluctuation amplitude $A_s \in (0.65, 0.75)$, running of spectral index $\alpha \in (-0.1, 0.1)$ and galaxy bias $b \in (0, 3)$. Following Gotovos et al. (2013a), we model the function as a Gaussian process. The function values vary from approximately -1×10^{18} and -1×10^{15}. We set the threshold to -3×10^{16} which is approximately the 75th percentile when we randomly sampled the function value at several thousand points.