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, $J(\theta_\star, \pi_1) - J(\theta_\star, \pi_2) = J(\theta_\star, \pi^n) - J(\theta_\star, \pi^0) = \sum_{t=1}^n J(\theta_\star, \pi^t) - J(\theta_\star, \pi^{t-1})$. $\square$

We will use Lemma 7 with π_2 as the policy $\pi^\star_{\mathrm{M}}$ 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 $\pi^\star_{\mathrm{M}}$ as X'_t. By Lemma 7 and equation (4) we have,

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

where we have defined

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

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 $\pi^{\mathrm{PS}}_{\mathrm{M}}$ at time instant t with D_{t-1} be X_t and the action taken by $\pi^\star_{\mathrm{M}}$ be X'_t. Then,*

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

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

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,

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

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,

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

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,

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

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

Here, $F_{n,i}$ is the data collected by $\pi^\star_{\mathrm{M}}$ 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,

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

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, $\mathrm{I}(\theta_\star; D_n) \leq \Psi_n$. □

B.3. Proof of Theorem 3

In this section, we will let $D^{\star\star}_m$ be the data collected $\pi^\star_{\mathrm{G}}$ in m steps and $D^\star_n$ be the data collected by $\pi^\star_{\mathrm{M}}$ 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 $\theta_\star \in \Theta$,*

$$\mathbb{E}_Y[\lambda(\theta_\star, D^\star_n)] \geq (1 - e^{-n/m}) \mathbb{E}_Y[\lambda(\theta_\star, D^{\star\star}_m)]$$

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 $\pi^\star_{\mathrm{G}}$ under suitable conditions.

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

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

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,

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

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 $1/2$. $\square$

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,*

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

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. $\square$

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 $z_1, z_2, \dots$. 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 $z_{n+1} = \mathrm{argmax}_{\|z\|=1}\det(\mathrm{P}_n^H + \frac{1}{\eta_\star^2}zz^\top)$. We therefore have,

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

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,

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

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). $\square$

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 75$^{\text{th}}$ percentile when we randomly sampled the function value at several thousand points.