Exploration–Exploitation in RL:
Calibrated Optimism in the Face of Uncertainty

Debabrota Basu
debabrota-basu.github.io

Spring School on Control Theory and Reinforcement Learning
CWI Amsterdam, 2025

PART 0

Revisiting MDPs and RL Algorithms

Markov Decision Processes

A Markov Decision Process (MDP) is a tuple $\mathcal{M} \triangleq \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- **State:** $s \in \mathcal{S} \subseteq \mathbb{R}^d$
- **Action/Intervention/Control/Input:** $a \in \mathcal{A} \subseteq \mathbb{R}^d$
- **Transition function/dynamics:** $\mathcal{P}(.|s, a)$ induces a distribution over s_{t+1} for s_t, a_t (previously f)
- **Reward Function:** $\mathcal{R}(.|s, a)$ induces a distribution over $\mathbb{R}$ measuring goodness of an action a at state s (negative of cost function)
- **Policy:** A deterministic or stochastic map $\pi(\cdot|s_t)$ from present state s_t to actions
- **How good or bas is your policy?** Value Function (Negative of cost of control)

$$V_\pi(s_0) \triangleq \sum_{t=0}^{\infty} \gamma^t \mathcal{R}_t(s_t, \pi(s_t))$$

Goal: Find an optimal policy π^* maximising $V_\pi(s_0)$.

A Generic Template of RL Algorithms: Generalised Policy Iteration

Algorithm Generalised Policy Iteration

1: **Input:** Initial Policy π_0
2: **for** episode $k = 1, 2, \ldots$ **do**
3: Observe an initial state s_0^k
4: **Rollouts:** Collect trajectory data $\{r(s_h^k, a_h^k), s_{h+1}^k\}_{h=0}^H$ and state s_{h+1}^k by playing policy π_k
5: **Policy Evaluation:** Compute the value function of the policy $V_{\pi_k}(s_0^k)$
6: **Policy Optimisation:** Use $V_{\pi_k}(s_0^k)$ to compute a better policy π_{k+1}
7: **end for**
8: **return** policy π_K.

Four Key Challenges in RL

- *Planning in Large Spaces (Curse of Dimensionality)*
  – How to optimise the policy when the number of reachable states and decidable actions are big?

- *Succinct Representation of Information*
  – How to succinctly represent the available information regarding states, actions, dynamics and policies?

- *Exploration–Exploitation Trade-off (Effect of Incomplete Information)*
  – Should you try out new decisions which may prove to be beneficial or play as best as you can with your existing knowledge?

- *Planning under Incomplete Information (Exploration + Planning)*
  – How to estimate the effect of an action and how to predict the future state reached from a state through the action?

PART 1

Optimism in the Face of Uncertainty
A Frequentist's Approach to Exploration–Exploitation Trade-off

Multi-armed Bandits

Modelling the Cost of Sequential Acquisition of Information

Sequential Decision Making

Medicine 1
$p_1^{\text{cured}} = 0.75$

Medicine 2
$p_2^{\text{cured}} = 0.95$

Medicine 3
$p_3^{\text{cured}} = 0.90$

$\cdots$

Medicine A
$p_A^{\text{cured}} = 0.5$

Sequential Decision Making

under Incomplete Information: Multi-armed Bandits [Thompson, 1933] + 92 years

Medicine 1
$p_1^{\text{cured}} = ?$

Medicine 2
$p_2^{\text{cured}} = ?$

Medicine 3
$p_3^{\text{cured}} = ?$

$\cdots$

Medicine A
$p_A^{\text{cured}} = ?$

Facing these unknowns

1. What would you do?
2. What would be a reasonable goal?

Medicine 1
$p_1^{\text{cured}} = ?$

Medicine 2
$p_2^{\text{cured}} = ?$

Medicine 3
$p_3^{\text{cured}} = ?$

$\dots$

Medicine A
$p_A^{\text{cured}} = ?$

For the t-th patient in the study

1. the doctor π chooses a Medicine A_t,
2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

Goal: Maximise the number of patients cured $\sum_{t=1}^{T} R_t$.

Maximise cumulative reward $\displaystyle\sum_{t=1}^{T} R_t$

$\underset{\text{Randomness}}{\approx}$ Maximise expected cumulative reward $\underbrace{V_T^\pi \triangleq \mathbb{E}\left[\sum_{t=0}^{T} R_t \mid A_t \sim \pi\right]}_{\text{Value of } \pi}$

$\underset{\text{Incomplete Information}}{\iff}$ Minimise expected regret $V_T^{\mathrm{OPT}} - V_T^\pi = \mathbb{E}\left[R(a^*)\right] T - V_T^\pi$

Regret $\mathrm{Reg}_\pi(T) \triangleq$ Value of Optimal Algorithm with Full Information
$-$ Value of Algorithm π with Incomplete Information

What Do We Ask next?

Can we design an algorithm that achieves zero-regret for *any set of distributions and any* T?

What Do We Ask next?

Can we design an algorithm that achieves zero-regret for *any set of distributions and any* T?

No, we cannot have zero regret under partial information! :(

What Do We Ask next?

Can we design an algorithm that achieves zero-regret for *any set of distributions and any* T?

No, we cannot have zero regret under partial information! :(

1. Why?
2. What's the minimum regret?

What Do We Ask next?

Can we design an algorithm that achieves zero-regret for *any set of distributions and any* T?

No, we cannot have zero regret under partial information! :(

1. Why?
   – Concentration of measure can happen only at a certain speed! (What's that?)
2. What's the minimum regret?

   – Minimum regret achievable by $\pi = \Omega\left(\sum_a \underbrace{(\mu^* - \mu_a)}_{\text{Suboptimality Gap}} \underbrace{\frac{\log T}{D_{\mathrm{KL}}(P_a, P_{a^*})}}_{\text{Distinguishability Gap}}\right) \approx \Omega\left(\sum_a \frac{\overbrace{\sigma_a^2}^{\text{Variance of a}} \log T}{\underbrace{\Delta_a}_{\text{Suboptimality Gap}}}\right)$.

   [Lai and Robbins, 1985]

What Do We Ask next?

Can we design an algorithm that achieves zero-regret for *any set of distributions and any* T?

No, we cannot have zero regret under partial information! :(

1. Why?
   – Concentration of measure can happen only at a certain speed! (What's that?)
2. What's the minimum regret?
   – Minimum regret achievable by $\pi = \Omega\left(\sum_a \underbrace{(\mu^* - \mu_a)}_{\text{Suboptimality Gap}} \underbrace{\frac{\log T}{D_{\mathrm{KL}}(P_a, P_{a^*})}}_{\text{Distinguishability Gap}}\right) \approx \Omega\left(\sum_a \frac{\overbrace{\sigma_a^2}^{\text{Variance of a}} \log T}{\underbrace{\Delta_a}_{\text{Suboptimality Gap}}}\right).$

Can we design an algorithm that achieves the regret lower bound for some sets of distributions?

Distribution-Independent (Minimax) Regret Bounds

Why Distribution Independent Regret?

The exact distributional form might not be known *a priori*, and you want to design an algorithm that is "good" for any distribution with a bounded range of outputs.

Minimax Regret

Let $\mathcal{F}$ be a family of distributions with output in $[0, R_{\max}]$.

$$\begin{aligned}\mathrm{Reg}(T;\mathcal{F}) &= \min_{\pi}\max_{\boldsymbol{\mu}\in\mathcal{F}} \mathrm{Reg}_{\pi}(T;\boldsymbol{\mu}) \\ &= \min_{\pi}\max_{\boldsymbol{\mu}\in\mathcal{F}} \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T}(\mu^{\star} - \mu_{A_t\sim\pi})\right]\end{aligned}$$

Distribution-Independent (Minimax) Regret Bounds

Why Distribution Independent Regret?

The exact distributional form might not be known *a priori*, and you want to design an algorithm that is "good" for any distribution with a bounded range of outputs.

Minimax Regret

Let $\mathcal{F}$ be a family of distributions with output in $[0, R_{\max}]$.

$$\begin{aligned}\operatorname{Reg}(T; \mathcal{F}) &= \min_{\pi} \max_{\boldsymbol{\mu} \in \mathcal{F}} \operatorname{Reg}_{\pi}(T; \boldsymbol{\mu}) \\ &= \min_{\pi} \max_{\boldsymbol{\mu} \in \mathcal{F}} \mathbb{E}_{\boldsymbol{\mu}} \left[\sum_{t=1}^{T} (\mu^{\star} - \mu_{A_t \sim \pi}) \right]\end{aligned}$$

Minimax Regret Lower Bound

The minimum achievable minimax regret is $\Omega\left(\sqrt{AT}\right)$.

Two Sides of a Bandit: Exploration and Exploitation

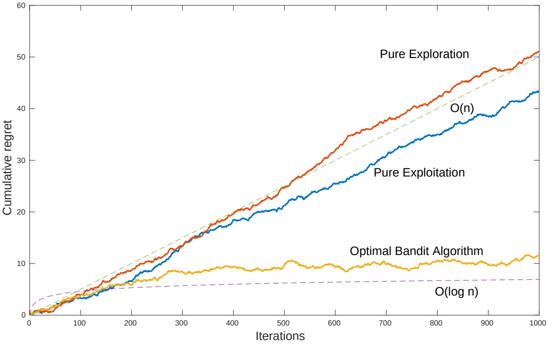
Pure Exploration
Take each decision equally randomly, and accumulate knowledge about all of them.

Pure Exploitation
Take the decision with maximum observed reward as per the present knowledge.

Two Sides of a Bandit: Exploration and Exploitation

Pure Exploration
Take each decision equally randomly, and accumulate knowledge about all of them.

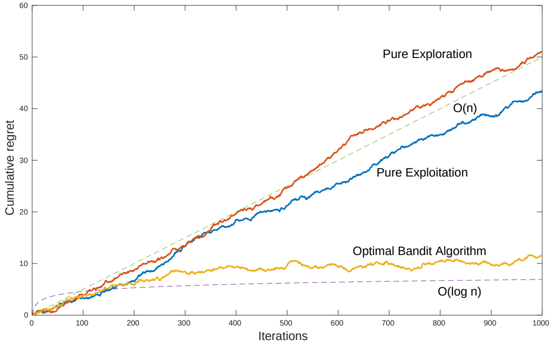
Pure Exploitation
Take the decision with maximum observed reward as per the present knowledge.



Two Sides of a Bandit: Exploration and Exploitation

Pure Exploration
Take each decision equally randomly, and accumulate knowledge about all of them.

Pure Exploitation
Take the decision with maximum observed reward as per the present knowledge.



The Exploration–exploitation Trade-off

Exploration and exploitation should be adapted on-the-go to achieve the lowest regret.

The Trade-off

Should you try out new decisions to fetch information, or play the best with your existing knowledge?

The Trade-off

Should you try out new decisions to fetch information, or play the best with your existing knowledge?

Strategy: Calibrated Optimism in the Face of Uncertainty (OFU)

Estimate an upper confidence bound on the empirical mean of the observed rewards and use it as an 'optimistic' index to choose the best arm to play.

For the t-th patient in the study

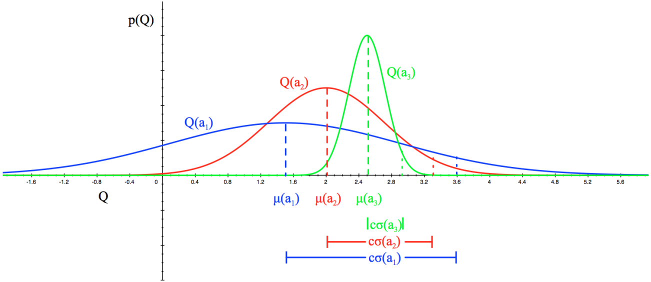
1.a. the optimistic doctor π computes optimistic indexes $I_a(t)$ for each medicine given the history

1.b. the optimistic doctor π chooses a Medicine A_t with maximum $I_a(t)$,

2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

How to Quantify Optimism?
At Any Point of Interaction

Step 1 : Construct a set of statistically plausible models for each arm from observations

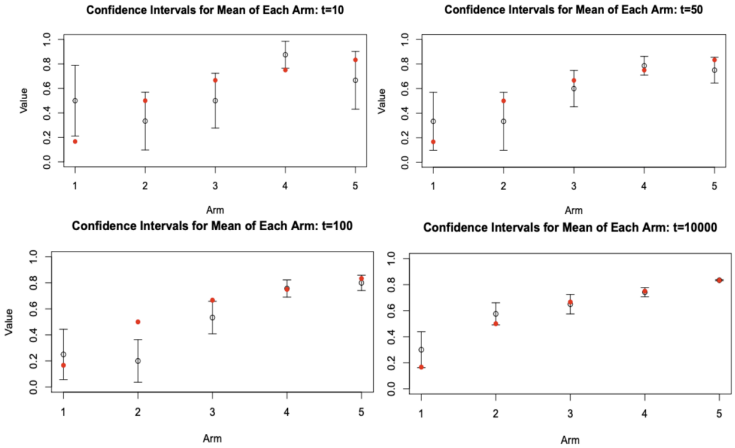


Step 2: Act as if the best possible model were the true model $\rightarrow$ Optimism (with probabilit $1-\delta$)

$$I_a(t) = \underbrace{\hat{\mu}_{a,t}}_{\text{Average reward of } a} + \sqrt{\frac{2\sigma_a^2 \log t/\delta}{\text{\# Selections of a till t}}}$$

How to Quantify Optimism?

Over Increasing Number of Interactions



For UCB, the regret is bounded by $\mathcal{O}\left(\sum_a \frac{\log T}{\mu^*-\mu_a}\right) = \mathcal{O}\left(\sum_a \frac{\log T}{\Delta_a}\right)$: reaches lower bound up to factors.

How to Quantify Optimism?

Under Noisy Observations: Aiming for Optimality

Index	UCB (Known Noise)	UCBV (Unknown Noise Variance)
$I_a(t)$	$\underbrace{\hat{\mu}_{a,t}}_{\text{Average reward of } a} + \sqrt{\frac{2\sigma_a^2 \log t}{\text{\# Selections of a till t}}}$	$\underbrace{\hat{\mu}_{a,t}}_{\text{Average reward of } a} + \underbrace{\hat{\sigma}_{a,t}}_{\sqrt{\text{Variance of rewards of } a}} \sqrt{\frac{2 \log t}{\text{\# Selections of a till t}}} + \frac{3 \times \text{range of noise} \times \log t}{\text{\# Selections of a till t}}$

- For UCB, the regret upper bound is $\mathcal{O}\left(\sum_a \Delta_a + \frac{\log T}{\Delta_a}\right)$.
- For UCBV, the regret upper bound is $\mathcal{O}\left(\sum_a \Delta_a + \left(\text{range of noise } + \frac{\sigma_a^2}{\Delta_a}\right) \log T\right)$.
- To obtain KL in the denominator, directly optimise KL to compute the optimistic index $\rightarrow$ KL-UCB [Garivier and Cappé, 2011]/IMED [Honda and Takemura, 2015]/BelMan [Basu et al., 2019]

IMED: $\quad I_a(t) = \text{\# Selections of a till t} \times \mathrm{KL}_{\inf}(\hat{\mu}_{a,t} \| \hat{\mu}_t^*) + \log(\text{\# Selections of a till t})$

Why Does it Work?

A Glimpse of Concentration of Measures [Boucheron et al., 2013]

Intuition

If you collect "enough" IID (Independent and Identically Distributed) samples from a distribution, the empirical estimates of mean and variance converges to their true values .

Hoeffding's Ineuqality

If you collect n IID samples from a distribution ν with bounded support $[a, b]$, we get

$$\mathbb{P}\left[\mid \hat{\mu}_n - \mu_\nu \mid \leq \varepsilon\right] \geq 1 - 2\exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right) \quad \text{for any } \epsilon > 0\,.$$

Why Does it Work?

A Glimpse of Concentration of Measures [Boucheron et al., 2013]

Intuition

If you collect "enough" IID (Independent and Identically Distributed) samples from a distribution, the empirical estimates of mean and variance converges to their true values.

Hoeffding's Ineuqality

If you collect n IID samples from a distribution ν with bounded support $[a, b]$, we get

$$\mathbb{P}\left[\mid \hat{\mu}_n - \mu_\nu \mid \leq \varepsilon \right] \geq 1 - 2\exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right) \quad \text{for any } \epsilon > 0\,.$$

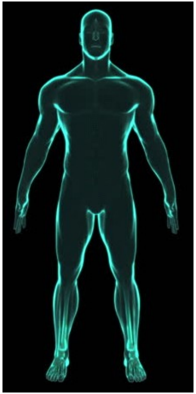
$\implies$ If we collect n IID samples, then the empirical mean satisfies with probability $1-\delta$,

$$\underbrace{\hat{\mu}_n - (b-a)\sqrt{\frac{\ln(2/\delta)}{n}}}_{\mathrm{LCB}_n} \leq \mu_\nu \leq \underbrace{\hat{\mu}_n + (b-a)\sqrt{\frac{\ln(2/\delta)}{n}}}_{\mathrm{UCB}_n}\,.$$

Markov Decision Processes

Modelling the Cost of Planning with Sequential Acquisition of Information

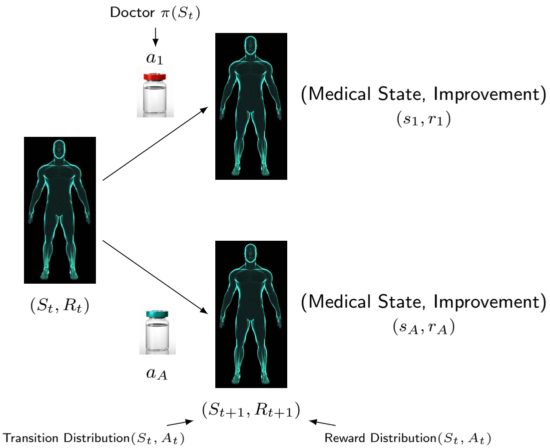
1	Fever
2	Pressure
...	...
D	Respiration



Medical State
S_t

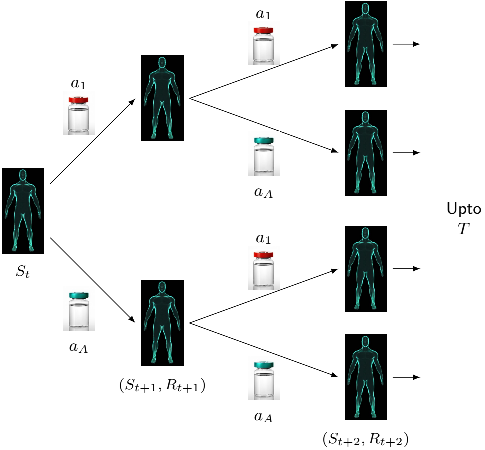
Decision Making under Incomplete Information

with Multiple States: Markov Decision Process (MDP) [Cayley, 1875; Massé, 1947] + 77 years



Decision Making under Incomplete Information

with Multiple States: Markov Decision Process (MDP) [Cayley, 1875; Massé, 1947] + 77 years



Goal: Maximise the total improvement of the patient $\sum_{t=1}^{T} R_t$

Finite-Horizon Episodic Markov Decision Process (MDP)

An MDP $\mathcal{M}$ is a model of iterative decision making under uncertainty containing

- A *state space* $\mathcal{S}$, and an *action space* $\mathcal{A}$.
- A *transition kernel* $\mathcal{P}$ dictating the probable next state given the present state and action.
- A *reward function* $\mathcal{R}$ dictating the *utility* of taking an action at a state.

Aim of doing Reinforcement Learning in an unknown MDP

Compute a policy π that *maximises* the *expected cumulative reward* over a time horizon H from any initial state s_0:

$$V_\pi(s) \triangleq \mathbb{E}\left[\sum_{t=0}^{H} \underbrace{\mathcal{R}(s_t, a_t)}_{\text{Rewards for each state-action}} \,\middle|\, \underbrace{a_t \sim \pi(s_t)}_{\text{Actions from a policy}}, \underbrace{s_t \sim \mathcal{P}(s_{t-1}, a_{t-1})}_{\text{Transitions to next state}}\right]$$

One can further treat $\mathcal{P}$ and r as functions of time $t \in \{1, \ldots, H\}$.

Minimum regret achievable by any policy π over K episodes:

Distribution-dependent bound: $\Omega\left(K_{\mathcal{M}} \log K\right)$

Distribution-independent bound: $\boxed{\Omega\left(\sqrt{H^3 SAK}\right)}$

Here, the constant charecterising hardness is the optimal trade-off between

1. minimising suboptimality gaps over the visited state-actions,
2. maximising information gain over the whole MDP. [Burnetas and Katehakis, 1997; Tirinzoni et al., 2021].

$$K_{\mathcal{M}}: \quad \inf_{\eta} \sum_{s,a,h} \eta_h(s,a) \Delta_{\mathcal{M},h}(s,a), \quad \text{such that} \ \inf_{M' \neq \mathcal{M}} \sum_{s,a,h} \eta_h(s,a) \left[\mathrm{KL}\left(\mathcal{M} || M'\right)\right]_{s,a,h} \geq 1$$

The First Try: From MDP to Bandits

Intuition

Finding optimal policy is equivalent to finding the optimal arm among the families of unique policies.

Let's play bandits with policies!

The First Try: From MDP to Bandits

Intuition

Finding optimal policy is equivalent to finding the optimal arm among the families of unique policies.

Let's play bandits with policies!

The Bad News

There are $(A^S)^H$ unique policies for a tabular episodic MDP.
Thus, running UCB or other bandit algorithms with yield $\mathcal{O}(\sqrt{(A^S)^H K})$ regret.

The Good News

Policies are not independent. They share structure and information between them.
$\rightarrow$ Leverage the information and MDP structure to achieve lower regret.

The Next Try: From Tabular MDP to Model Estimation and Optimistic Planning

Intuition

Learn estimates of rewards and transitions from the data and use it to plan further.

In tabular MDPs, what are the optimistic estimators of mean rewards and transitions?

Intuition

Learn estimates of rewards and transitions from the data and use it to plan further.

In tabular MDPs, what are the optimistic estimators of mean rewards and transitions?

Data : $$\mathcal{D}_{H,k} = \{\{s_{h,i}, a_{h,i}, r_{h,i}, s_{h+1,i}\}_{h=1}^{H}\}_{i=1}^{k}$$

Intuition

Learn estimates of rewards and transitions from the data and use it to plan further.

In tabular MDPs, what are the optimistic estimators of mean rewards and transitions?

Data : $\mathcal{D}_{H,k} = \{\{s_{h,i}, a_{h,i}, r_{h,i}, s_{h+1,i}\}_{h=1}^{H}\}_{i=1}^{k}$

Estimates of transitions : $\hat{\mathcal{P}}_k(s,a) = \dfrac{\#\text{ visits to } (s,a,s')}{\#\text{ visits to } (s,a)}$

Estimates of rewards : $\hat{\mathcal{R}}_k(s,a) = \dfrac{\sum_{i=1}^{k} r_{h,i}(s,a)}{\#\text{ visits to } (s,a)}$

In tabular MDPs, what are the optimistic estimators of mean rewards and transitions?

Data : $\mathcal{D}_{H,k} = \{\{s_{h,i}, a_{h,i}, r_{h,i}, s_{h+1,i}\}_{h=1}^{H}\}_{i=1}^{k}$

Optimistic Estimates of Transitions : $\tilde{\mathcal{P}}_k(s,a) = \frac{\#\text{ visits to }(s,a,s')}{\#\text{ visits to }(s,a)} + c_1\sqrt{\frac{H^2\log(SAT/\delta)}{\#\text{ Selections of (s,a)}}}$

Optimistic Estimates of Rewards : $\tilde{\mathcal{R}}_k(s,a) = \frac{\sum_{i=1}^{k} r_{h,i}(s,a)}{\#\text{ visits to }(s,a)} + \underbrace{c_2\sqrt{\frac{R_{\max}^2 H^2\log(SAT/\delta)}{\#\text{ Selections of (s,a)}}}}_{\mathrm{CB}_h(s,a)}$

In tabular MDPs, how to plan with the estimates of rewards and transitions?

Algorithm UCB-Value Iteration (UCB-VI) [Azar et al., 2017]

1: **Input:** Steps K, initial policy π_0
2: **Initialise:** a
3: **for** episodes $k = 1, 2, \dots, K$ **do**
4: Data Collection: Play π_{k-1} to collect $\mathcal{D}_k = \mathcal{D}_{k-1} \cup \{s_{h,k}, a_{h,k}, r_{h,k}, s_{h+1,k}\}_{h=1}^{H}$
5: Model Estimation: **For all** (s, a)**:** Compute empirical estimate of transitions $\hat{\mathcal{P}}_k$ and optimistic estimates of rewards $\tilde{\mathcal{R}}_k$
6: Planning: $\pi_k = \mathrm{ValueIteration}(\hat{\mathcal{P}}_k, \tilde{\mathcal{R}}_k)$.
7: **end for**
8: **return** π_K

Step 3: Value Iteration with Optimistic Estimates of Mean Rewards and/or Transitions

Algorithm UCB-Value Iteration (UCB-VI) [Azar et al., 2017]

1: **Input:** Steps K, initial policy π_0
2: **Initialise:** a
3: **for** episodes $k = 1, 2, \ldots, K$ **do**
4: Data Collection: Play π_{k-1} to collect $\mathcal{D}_k = \mathcal{D}_{k-1} \cup \{s_{h,k}, a_{h,k}, r_{h,k}, s_{h+1,k}\}_{h=1}^{H}$
5: Model Estimation: **For all** (s, a)**:** Compute empirical estimate of transitions $\hat{\mathcal{P}}_k$ and optimistic estimates of rewards $\tilde{\mathcal{R}}_k$
6: Planning: $\pi_k = \text{ValueIteration}(\hat{\mathcal{P}}_k, \tilde{\mathcal{R}}_k)$.
7: **end for**
8: **return** π_K

Why using optimistic estimates in rewards is enough?

Regret Analysis of UCB-VI

Goal

Minimise regret, i.e. the cost of sequential information w.r.t. the optimum value:

$$\mathrm{Reg}_{\text{UCB-VI}}(K) \triangleq \sum_{k=1}^{K} \left(V^{\pi^\star}_{\mathcal{M},1}(s^k_0) - V^{\pi_k}_{\mathcal{M},1}(s^k_0) \right).$$

$$\begin{aligned}
\mathrm{Reg}_{\text{UCB-VI}}(K) &= \sum_{k=1}^{K} \left(\boxed{V^{\pi^\star}_{\mathcal{M},1}(s^k_0)} - V^{\pi_k}_{\mathcal{M},1}(s^k_0) \right) \\
&\leq \sum_{k=1}^{K} \left(\boxed{\tilde{V}^{\pi_k}_{\mathcal{M},1}(s^k_0)} - V^{\pi_k}_{\mathcal{M},1}(s^k_0) \right)
\end{aligned}$$

Optimistic Value with Large Enough UCB

Regret Analysis of UCB-VI

Goal

Minimise regret, i.e. the cost of sequential information w.r.t. the optimum value:

$$\mathrm{Reg}_\pi(K) \triangleq \textstyle\sum_{k=1}^{K} \left(V^{\pi^\star}_{\mathcal{M},1}(s_0^k) - V^{\pi_k}_{\mathcal{M},1}(s_0^k) \right).$$

$$\begin{aligned}
\mathrm{Reg}_{\text{UCB-VI}}(K) &= \sum_{k=1}^{K} \left(V^{\pi^\star}_{\mathcal{M},1}(s_0^k) - V^{\pi_k}_{\mathcal{M},1}(s_0^k) \right) \\
&\leq \sum_{k=1}^{K} \left(\tilde{V}^{\pi_k}_{\mathcal{M},1}(s_0^k) - V^{\pi_k}_{\mathcal{M},1}(s_0^k) \right) && \text{Optimistic Value with Large Enough UCB} \\
&= \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{s,a \sim \mathcal{M}\pi_{k,h}} \left[\tilde{\mathcal{R}}(s,a) + \hat{\mathcal{P}}(\cdot|s,a)^\top V_h - \hat{\mathcal{R}}(s,a) - \mathcal{P}(\cdot|s,a)^\top V_h \right] && \text{Bellman Eq.} \\
&\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{s,a \sim \mathcal{M}\pi_{k,h}} \left[\mathrm{CB}_h(s,a) + \left(\hat{\mathcal{P}}(\cdot|s,a) - \mathcal{P}(\cdot|s,a) \right)^\top V_h \right] && \text{Optimistic bonus} \\
&= \mathcal{O}\left(\sqrt{H^4 SAK} \right)
\end{aligned}$$

Optimism Out of Tabular MDPs: Linear, Kernel, Neural Tangent Kernels, ...

Optimism with Functions

We can use the same principles of optimism as in tabular MDP and merge them with the function approximations that we have learned.

Technical Challenges

1. Designing tight confidence bounds for each of the functional forms.

2. Running regression with enough samples and good optimizer to retain the statistical guarantees of estimation and approximation.

Some Resources:

1. Chapter 2 of `https://rltheorybook.github.io/rltheorybook_AJKS.pdf` for linear & bilinear MDPs,
2. [Chowdhury and Gopalan, 2019, Ouhamma et al., 2022, Vakili and Olkhovskaya, 2023] for kernel MDPs.

PART 2

Posterior Sampling for RL (PSRL)
A Bayesian's Approach to Exploration–Exploitation Trade-off

Medicine 1
$p_1^{\text{cured}} = ?$

Medicine 2
$p_2^{\text{cured}} = ?$

Medicine 3
$p_3^{\text{cured}} = ?$

$\dots$

Medicine A
$p_A^{\text{cured}} = ?$

For the t-th patient in the study

1. the doctor π chooses a Medicine A_t,
2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

Bandits: Frequentist vs. Bayesian

Unknown reward distributions: $\{\mathbb{P}(\mu_a, \sigma_a^2)\}_{a=1}^{A}$ such that $\mathcal{R}(a,t) \sim \mathbb{P}(\mu_a, \sigma_a^2)$.

Frequentist

- **Model:** Means of reward distributions are unknown parameters $\mu_1, \ldots, \mu_A \in \mathbb{R}$

Bayesian

- **Model:** Means of reward distributions are sampled from a prior distribution: $(\mu_1, \ldots, \mu_A) \sim \mathbb{P}_0$

Bandits: Frequentist vs. Bayesian

Unknown reward distributions: $\{\mathbb{P}(\mu_a, \sigma_a^2)\}_{a=1}^{A}$ such that $\mathcal{R}(a,t) \sim \mathbb{P}(\mu_a, \sigma_a^2)$.

Frequentist

- **Model:** Means of reward distributions are unknown parameters $\mu_1, \ldots, \mu_A \in \mathbb{R}$
- **Frequentist Regret:** Distribution-dependent and Minimax
  $$\mathrm{Reg}(T; \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right]$$

Bayesian

- **Model:** Means of reward distributions are sampled from a prior distribution: $(\mu_1, \ldots, \mu_A) \sim \mathbb{P}_0$
- **Bayesian Regret:** Prior-dependent and Minimax
  $$\mathrm{BR}(T; \mathbb{P}_0) = \mathbb{E}_{\mathbb{P}_0}\left[\mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right]\right]$$

Bandits: Frequentist vs. Bayesian

Unknown reward distributions: $\{\mathbb{P}(\mu_a, \sigma_a^2)\}_{a=1}^{A}$ such that $\mathcal{R}(a,t) \sim \mathbb{P}(\mu_a, \sigma_a^2)$.

Frequentist

- **Model:** Means of reward distributions are unknown parameters
  $\mu_1, \ldots, \mu_A \in \mathbb{R}$
- **Frequentist Regret:** Distribution-dependent and Minimax
  $$\mathrm{Reg}(T; \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right]$$
- **Approach:**
  1. Create (max. likelihood) estimators of mean
  2. Create optimistic confidence bounds of the estimates and go greedy

Bayesian

- **Model:** Means of reward distributions are sampled from a prior distribution:
  $(\mu_1, \ldots, \mu_A) \sim \mathbb{P}_0$
- **Bayesian Regret:** Prior-dependent and Minimax
  $$\mathrm{BR}(T; \mathbb{P}_0) = \mathbb{E}_{\mathbb{P}_0}\left[\mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right]\right]$$
- **Approach:**
  1. Create posterior distributions of means
  2. Sample a vector of means from it and go greedy

An Example of Prior and Posterior Distributions

Gaussian Bandits

- **Reward Distributions:** $\{\mathcal{N}(\mu_a, \sigma^2)\}_{a=1}^{A}$
- **Prior Distribution:** $\mu_a \underset{\text{I.I.D.}}{\sim} \mathcal{N}(0, \sigma_0^2)$
- **Posterior Distributions:**

$$
\begin{aligned}
\mathbb{P}_t(\mu_a) &= \mathbb{P}[\mu_a \mid r_1, \ldots, r_t; \mathbb{P}_0] \\
&= \mathcal{N}\left(\frac{Z_a(t)}{N_a(t) + \frac{\sigma^2}{\sigma_0^2}}, \frac{\sigma^2(t)}{N_a(t) + \frac{\sigma^2}{\sigma_0^2}}\right)
\end{aligned}
$$

An Example of Prior and Posterior Distributions

Gaussian Bandits

- **Reward Distributions:** $\{\mathcal{N}(\mu_a, \sigma^2)\}_{a=1}^{A}$
- **Prior Distribution:** $\mu_a \underset{\text{I.I.D.}}{\sim} \mathcal{N}(0, \sigma_0^2)$
- **Posterior Distributions:**

$$\begin{aligned}\mathbb{P}_t(\mu_a) &= \mathbb{P}[\mu_a \mid r_1, \ldots, r_t; \mathbb{P}_0] \\ &= \mathcal{N}\left(\frac{Z_a(t)}{N_a(t) + \frac{\sigma^2}{\sigma_0^2}}, \frac{\sigma^2(t)}{N_a(t) + \frac{\sigma^2}{\sigma_0^2}}\right)\end{aligned}$$

Bernoulli Bandits

- **Reward Distributions:** $\{\mathcal{B}(\mu_a)\}_{a=1}^{A}$
- **Prior Distribution:** $\mu_a \underset{\text{I.I.D.}}{\sim} \mathit{Unif}([0,1])$
- **Posterior Distributions:**

$$\begin{aligned}\mathbb{P}_t(\mu_a) &= \mathbb{P}[\mu_a \mid r_1, \ldots, r_t; \mathbb{P}_0] \\ &= \mathit{Beta}\left(Z_a(t) + 1, N_a(t) - Z_a(t) + 1\right)\end{aligned}$$

Here, $N_a(t) = \#$ pulls of arm a till time t

$Z_a(t) =$ total sum of rewards obtained from arm a by time t

Resources: `https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions`

Thompson Sampling: The First Bandit Algorithm

Algorithm Thompson Sampling [Thompson, 1933]

1: **Input**: Prior $\mathbb{P}_0(\boldsymbol{\mu})$
2: **for** steps $t = 1, 2, \ldots$ **do**
3: Sample from the posterior $\boldsymbol{\mu}_t \sim \mathbb{P}_t(\boldsymbol{\mu} \mid R_1, \ldots, R_t; \mathbb{P}_0) = \prod_{a=1}^{A} \mathbb{P}_t(\mu_a \mid R_1, \ldots, R_t; \mathbb{P}_0)$
4: Planning: Play $A_t = \arg\max_a \mu_{a,t}$
5: Data Collection and posterior update: Observe reward $R_t \sim \mathbb{P}_{A_t}$ and update the posterior to $\mathbb{P}_{t+1}(\boldsymbol{\mu})$
6: **end for**

Visualisation: `https://en.wikipedia.org/wiki/Thompson_sampling`

Thompson Sampling: The First Bandit Algorithm

Algorithm Thompson Sampling [Thompson, 1933]

1: **Input**: Prior $\mathbb{P}_0(\boldsymbol{\mu})$
2: **for** steps $t = 1, 2, \ldots$ **do**
3: Sample from the posterior $\boldsymbol{\mu}_t \sim \mathbb{P}_t(\boldsymbol{\mu} \mid R_1, \ldots, R_t; \mathbb{P}_0) = \prod_{a=1}^{A} \mathbb{P}_t(\mu_a \mid R_1, \ldots, R_t; \mathbb{P}_0)$
4: Planning: Play $A_t = \arg\max_a \mu_{a,t}$
5: Data Collection and posterior update: Observe reward $R_t \sim \mathbb{P}_{A_t}$ and update the posterior to $\mathbb{P}_{t+1}(\boldsymbol{\mu})$
6: **end for**

Intuition

It is equivalent to
(a) sample an arm according to its probability of being the optimal one

(b) sample a possible bandit environment from the posterior distribution and act optimally in this sampled environment

Algorithm Thompson Sampling [Thompson, 1933]

1: **Input**: Prior $\mathbb{P}_0(\boldsymbol{\mu})$
2: **for** steps $t = 1, 2, \ldots$ **do**
3: Sample from the posterior $\boldsymbol{\mu}_t \sim \mathbb{P}_t(\boldsymbol{\mu} \mid R_1, \ldots, R_t; \mathbb{P}_0) = \prod_{a=1}^{A} \mathbb{P}_t(\mu_a \mid R_1, \ldots, R_t; \mathbb{P}_0)$
4: Planning: Play $A_t = \arg\max_a \mu_{a,t}$
5: Data Collection and posterior update: Observe reward $R_t \sim \mathbb{P}_{A_t}$ and update the posterior to $\mathbb{P}_{t+1}(\boldsymbol{\mu})$
6: **end for**

Upper Bounds on Regret

Frequentist (Distribution-dependent): For exponential family of distributions ([Kaufmann et al., 2012], and so on...)

$$\mathrm{Reg}(T; \boldsymbol{\mu}) = \mathcal{O}\left(\sum_a \frac{\Delta_a}{\mathrm{KL}\left(\mu_a || \mu^*\right)} \log T\right)$$

Bayesian: [Russo and Van Roy, 2014], and so on...

$$\mathrm{BR}(T) = \mathcal{O}(\sqrt{AT} + A)$$

Thompson Sampling for MDPs: PSRL

Algorithm PSRL [Osband et al., 2013]

Input: Prior $\mathbb{P}_0(M)$, Likelihood function $\mathcal{L}((s,a,s')|M)$,
for episode $k = 1, 2, \ldots$ **do**
    Sample $M_k \sim \mathbb{P}_k(M \mid \mathcal{H}_k)$
    Planning: Find $\pi^*(M_k)$ with Value Iteration/Policy Iteration on M_k
    Data Collection: Play $\pi^*(M_k)$ till horizon H to obtain $\{(s_i, a_i, s'_i)\}_{i=H(k-1)}^{Hk}$
    Update posterior: $\mathbb{P}_{k+1} \leftarrow \mathbb{P}(M|\mathcal{H}_{k+1})$, where $\mathcal{H}_{k+1} \leftarrow \mathcal{H}_k \cup \{x_i\}_{i=H(k-1)+1}^{Hk}$
end for

Intuition

It is equivalent to
(a) sample a policy according to its probability of being the optimal one

(b) sample a possible MDP from the posterior distribution and act optimally in this sampled MDP

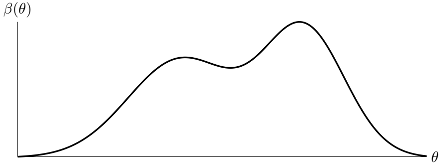
An Overview of PSRL-type Algorithms: Exact Posterior

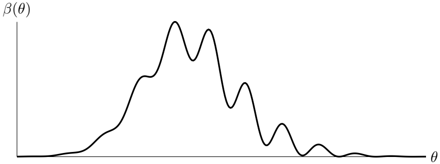
PSRL	Bayesian Regret $\mathrm{BR}(T)$	Assumptions
[Osband et al., 2013]	$\widetilde{O}(H^{1.5}S\sqrt{AK})$	Tabular MDP
[Moradipari et al., 2023]	$\widetilde{O}(H^{2}\sqrt{SAK})$	Tabular MDP
[Fan and Ming, 2021]	$\widetilde{O}(dH^{2}\sqrt{K})$	Linear MDP
[Chowdhury and Gopalan, 2019]	$\widetilde{O}(\sqrt{d}H\sqrt{K})$	Kernel MDP
[Jorge et al., 2024]	$\widetilde{O}(H^{1.75}K^{0.75})$	LSI $\mathcal{L}$ with constant α
[Jorge et al., 2024]	$\widetilde{O}(\sqrt{d}H\sqrt{K})$	LSI $\beta(M)$, linear growth on α

Theory-to-Practice Gap: Appearance of Isoperimetry

The RL Theory-to-Practice Gap in PSRL

In theory, we have provable guarantees for only exponential family distributions and log-concave distributions. But neural networks are often not log-concave.





Can we develop a more general theory?

Can we remove explicit dependencies of regret analysis on the exact parametric form?

Theory-to-Practice Gap: Appearance of Isoperimetry

Can we remove explicit dependencies of regret analysis on the exact parametric form?

Let's go for the isoperimetric distributions (e.g. mixture and perturbed log-concaves and more): one of the most general family of distributions where concentration of measure is provable and controllable.

Isoperimetric Inequality: Log-Sobolev

A distribution ν satisfies the Log-Sobolev Inequality (LSI) with a constant α if, for all smooth distributions $\rho : \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{KL}(\rho \,\|\, \nu) \leq \frac{1}{\alpha}\mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right]$$

Isoperimetric Inequality: Poincaré

A distribution ν satisfies the Poincaré Inequality (PI) with a constant α_P if, for all smooth functions $g : \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{Var}_\rho(g) \leq \frac{1}{\alpha_P}\mathbb{E}_\rho\left[\|\nabla g\|^2\right]$$

Approximate Posterior Sampling Algorithms in RL

What if You cannot Track the True Posterior?

If the posterior is too high dimensional to track and update, or does not have a closed analytical form, can we still have guarantees of PSRL.

What if You cannot Track the True Posterior?

If the posterior is too high dimensional to track and update, or does not have a closed analytical form, can we still have guarantees of PSRL.

Solution: Approximate Samplers

Design or use Approximate Samplers that takes multiple steps to find a sample which is ϵ-close to sampling from true distribution (e.g. Langevin MCMC methods, Langevin Gradient Descent methods etc.).

Algorithm Langevin PSRL (LaPSRL) [Jorge et al., 2024]

Input: Likelihood $\mathcal{L}(x|\mathcal{M})$, Prior $\mathbb{P}_0(\mathcal{M})$, Horizon H, total episodes K.
for episodes $k = 1, \ldots, K$ **do**
  $\epsilon_{\text{post},k} = \frac{H}{k\Delta^2_{\text{max}}}$
  if Chained sampling, $\rho_0 = \theta_{k-1}$ # Reuse last sample from previous iteration.
  else $\rho_0 \sim \mathbb{P}_0(\mathcal{M})$ # Resample from prior.
  Approximate Sampling: Sample $\mathcal{M}_k = \textsc{Langevin sample}(\mathcal{L}(x \mid \mathcal{M}), \mathbb{P}_0(\mathcal{M}), \mathcal{H}_k, \epsilon_{\text{post},k}, \rho_0)$
  Planning: Play $\pi^\star(\mathcal{M}_k)$ and play until horizon H obtaining data $\mathcal{H}_{k+1} = \mathcal{H}_k \cup \{z_i\}_{i=H(k-1)}^{Hk}$.
end for

An Overview of Approximate Posterior Sampling Algorithms

Algorithms	Bayesian Regret $\mathrm{BR}(T)$	Assumptions	Total gradient complexity
PSRL	$\widetilde{O}(\sqrt{dHT})$	LSI $\mathbb{P}_t(M)$, linear growth on α, exact posterior	-
[Xu et al., 2022]	$\widetilde{O}(d^{1.5}\sqrt{T})$	Lin. Bandits with cond. number κ, sub-Gaussian	$\widetilde{O}(\kappa T^2)$
[Kuang et al., 2023]	$\widetilde{O}(d^{1.5}H^{1.5}\sqrt{T})$	Linear MDP, episodic delay	$\widetilde{O}(T^2)$
[Haque et al., 2024]	$\widetilde{O}(H^{1.5}d\sqrt{T})$	Linear MDP	$\widetilde{O}(T^2/\sqrt{d})$
[Karbasi et al., 2023]	$\widetilde{O}(d\mathfrak{s}\sqrt{T})$	Infinite horizon with span $\mathfrak{s}$ $d \ll \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$, strongly log-concave	$\widetilde{O}(1)$ (due to log-conc.)
LaPSRL	$\widetilde{O}(\sqrt{dHT})$	LSI $\beta(M)$, linear growth on α	$\widetilde{O}(T\tau + T^{1.5}\tau/d)$
LaPSRL	$\widetilde{O}(\sqrt{T}g(\cdot))$	LSI $\beta(M)$	$\widetilde{O}\left(\sum_{k=1}^{K} \frac{H^3k^3}{\alpha_k^2} + \frac{dH^{4.5}k^{3.5}}{\alpha_k^2 g(\cdot)^2}\right)$
[Jorge et al., 2024]		policy with $\mathrm{BR}(T) = \widetilde{O}(\sqrt{T}g(\cdot))$ for exact post.	

PART 3

Randomised Least Square Value Iteration
Perturbing the Estimates as an Alternative to Posterior Sampling

- Bilinear Exponential Family (BEF) model of dynamics and observation:

$$\mathbb{P}(\tilde{s} \mid s, a) \propto \exp\left(\psi(\tilde{s})^\top M_{\theta^p} \varphi(s, a)\right)$$
$$\mathbb{P}(r \mid s, a) \propto \exp\left(r\, B^\top M_{\theta^r} \varphi(s, a)\right)$$

  Here, $M_{\theta^p} = \sum_i \theta_i^p A_i$ and $M_{\theta^r} = \sum_i \theta_i^r A_i$.

- Minimise regret, i.e. the cost of sequential information w.r.t. the optimum value:

$$\mathcal{R}(K) \triangleq \sum_{k=1}^{K} \left(V_{\theta,1}^{\pi^\star}(s_1^k) - V_{\theta,1}^{\pi^t}(s_1^k)\right).$$

- Bilinear Exponential Family (BEF) model of dynamics and observation:

$$\mathbb{P}(\tilde{s} \mid s, a) \propto \exp\left(\psi(\tilde{s})^\top M_{\theta^p} \varphi(s, a)\right)$$
$$\mathbb{P}(r \mid s, a) \propto \exp\left(r\, B^\top M_{\theta^r} \varphi(s, a)\right)$$

  Here, $M_{\theta^p} = \sum_i \theta_i^p A_i$ and $M_{\theta^r} = \sum_i \theta_i^r A_i$.

- Minimise regret, i.e. the cost of sequential information w.r.t. the optimum value:

$$\mathcal{R}(K) \triangleq \sum_{k=1}^{K} \left(V_{\theta,1}^{\pi^\star}(s_1^k) - V_{\theta,1}^{\pi^t}(s_1^k)\right).$$

Why Interesting?

BEF can model Linear Quadratic Regulators (LQRs), Hamiltonians controlling Schrödinger's equations, Linear Gaussian Regulators (LQGs), Block structured MDPs

Algorithm BEF-RLSVI: An Optimistic and Tractable RL Algorithm

Input: failure rate δ, constants α^p, η and $(x_k)_{k\in[K]} \propto dH^2$
for episode $k = 1, 2, \dots$ **do**
  Observe initial state s_1^k
  Explore with perturbation: $\tilde{\theta}^r(k) = \hat{\theta}^r(k) + \boxed{\xi_k}$ with $\boxed{\xi_k \sim \mathcal{N}\left(0, x_k (G^p)^{-1}\right)}$
  Planning: Compute $(\tilde{Q}_h^k)_{h\in[H]}$ via Bellman-backtracking
  for $h = 1, \dots, H$ **do**
    Pull action $a_h^k = \arg\max_a \tilde{Q}_h^k(s_h^k, a)$, observe reward $r(s_h^k, a_h^k)$ and state s_{h+1}^k.
  end for
  Update the parameters with penalised MLE $\hat{\theta}^p(k), \hat{\theta}^r(k)$
end for

This method achieves tractable planning and exploration for LQRs, LQGs, Factored MDPs etc.

A Framework for Bilinear Exponential Families [Ouhamma et al., 2022]

Linearity in Infinite Dimension

For an MDP of the BEF, we can write the state-action value function linearly, at step h:

$$\tilde{Q}^{\pi}_{h}(s,a) = \mathbb{E}^{\tilde{\theta}^{r}}[r(s,a)] + \left\langle \phi^{p}(s,a), \int_{\mathcal{S}} \mu^{p}(\tilde{s})\tilde{V}h + 1^{\pi}(\tilde{s})d\tilde{s} \right\rangle.$$

Random Fourier Transform for Finite-dimensional Approximation

Using Random Fourier Transform entails $\mathcal{O}(pH^2K\log(HK))$ dimensional approximations of ϕ^p and ψ^p leading to polynomial complexity of planning.

Linearity in Infinite Dimension

For an MDP of the BEF, we can write the state-action value function linearly, at step h:

$$\tilde{Q}_h^{\pi}(s,a) = \mathbb{E}^{\tilde{\theta}^r}[r(s,a)] + \left\langle \phi^p(s,a), \int_{\mathcal{S}} \mu^p(\tilde{s})\tilde{V}h + 1^{\pi}(\tilde{s})d\tilde{s} \right\rangle.$$

Near-Optimal Performance

For decision space with bounded curvature and bounded parameters, with probability $1-\delta$,

$$\mathcal{R}(K) = \mathcal{O}\left(\sqrt{d^3H^3K}\ln(\tfrac{1}{\delta})\right).$$

Key Changes in Regret Analysis of BEF-RLSVI

Optimism: Key reasons for choosing `RLSVI`-type algorithms:

- Perturbing the reward estimation guarantees optimism with a constant probability
- A constant probability of optimism is enough to control the value function approximation error

Transportation: Using transportation inequalities instead of the simulation lemma reduces a $\sqrt{H}$ factor

Elliptical lemma:

- Leveraging the boundedness of the true value function enables using an improved elliptical lemma ($\sqrt{H}$ less than [Chowdhury et al., 2021])
- The norm of features can only be large $\mathcal{O}(d)$ times, thus, we can omit clipping and reduce the regret by $\sqrt{d}$ compared to [Zanette et al., 2020].

Approximate planning:

- To guarantee a tractable planning, we approximate the transition with $(1/\sqrt{H^2K})$-error. Using mis-specification style analysis, we show that the approximation does not hinder the regret bound.
- Using a Linear-RL algorithm directly on top of the approximation would lead to a linear regret.

Overview of RLSVI-type Algorithms

Algorithm	Regret	Tractable exploration	Tractable planning	Free of clipping	Model, assumptions
Thompson sampling [Ren et al., 2022]	$\sqrt{d^2 H^3 K}$ (Bayesian)	✗	✓	N.A	Gaussian $\mathcal{P}$ Known rewards
$\mathcal{F}$−`PHE-LSVI` [Ishfaq et al., 2021]	$\mathrm{poly}(d_E H)\sqrt{KH}$	✓	✗	✗	Eluder dimension, Tabular
`PHE-LSVI` [Ishfaq et al., 2021]	$\sqrt{d^3 H^4 K}$	✓	✗	✗	Anti-concentration, linear transitions
`OPT-RLSVI` [Zanette et al., 2020]	$\sqrt{d^4 H^5 K}$	✓	✓	✗	Linear V
BEF-RLSVI [Ouhamma et al., 2022]	$\sqrt{d^3 H^3 K}$	✓	✓	✓	Bilinear Exp Family
Open Problem	$\sqrt{d_{\mathcal{F}}^3 H^3 K}$	✓	✓	✓	Any function class $\mathcal{F}$

Overview of RLSVI-type Algorithms

Algorithm	Regret	Tractable exploration	Tractable planning	Free of clipping	Model, assumptions
Thompson sampling [Ren et al., 2022]	$\sqrt{d^2 H^3 K}$ (Bayesian)	✗	✓	N.A	Gaussian $\mathcal{P}$ Known rewards
$\mathcal{F}-$`PHE-LSVI` [Ishfaq et al., 2021]	$\mathrm{poly}(d_E H)\sqrt{KH}$	✓	✗	✗	Eluder dimension, Tabular
`PHE-LSVI` [Ishfaq et al., 2021]	$\sqrt{d^3 H^4 K}$	✓	✗	✗	Anti-concentration, linear transitions
`OPT-RLSVI` [Zanette et al., 2020]	$\sqrt{d^4 H^5 K}$	✓	✓	✗	Linear V
BEF-RLSVI [Ouhamma et al., 2022]	$\sqrt{d^3 H^3 K}$	✓	✓	✓	Bilinear Exp Family
Open Problem	$\sqrt{d_{\mathcal{F}}^3 H^3 K}$	✓	✓	✓	Any function class $\mathcal{F}$

Limitations

We need to assume specific parametric forms to design RLSVI algorithms with provable regret guarantees.

Part 4

The Coda
Where Have We Reached and What's Ahead?

- **Model:** Reward and transition distributions with unknown parameters

- **Frequentist Regret:** Distribution-dependent and Minimax

$$\mathrm{Reg}_\pi(K, \mathcal{M}) \triangleq \sum_{k=1}^{K} \left(V^{\pi^\star}_{\mathcal{M},1}(s_0^k) - V^{\pi_k}_{\mathcal{M},1}(s_0^k) \right)$$

- **Approach:** Model-based
  1. Create (max. likelihood) estimators of mean $\mathcal{R}$ and/or $\mathcal{P}$ with/without functional form
  2. Create optimistic confidence bounds of the estimates
  3. Plan with optimistic rewards and/or transitions

- **Alternative Approach:** Model-free
  1. Create (max. likelihood) estimators of Q-values with/without functional form (e.g. regression)
  2. Create optimistic confidence bounds of Q-estimates
  3. Plan with optimistic Q-values

- **Model:** Reward and transition distributions (aka MDPs) are sampled from a prior distribution

- **Bayesian Regret:** Prior-dependent and Minimax

$$\mathrm{BR}(T; \mathbb{P}_0) = \mathbb{E}_{\mathcal{M}\sim\mathbb{P}_0}\left[\mathrm{Reg}_\pi(K)\right] = \mathbb{E}_{\mathcal{M}\sim\mathbb{P}_0}\left[\sum_{k=1}^{K}\left(V^{\pi^\star}_{\mathcal{M},1}(s_0^k) - V^{\pi_k}_{\mathcal{M},1}(s_0^k)\right)\right]$$

- **Approach:** PSRL-type
  1. Create posterior distributions of transitions and rewards/Q-values
  2. Sample a vector of (transitions,rewards)/Q-values from it
  3. Plan with sampled MDP/Q-values

- **Alternative Approach:** RLSVI-type
  1. Create (maximum likelihood) estimators of Q-values/transitions and rewards with/without functional form (e.g. regression)
  2. Perturb the model parameters with calibrated Gaussian noise
  3. Plan with perturbed Q-values

What's Out There?

What we Didn't Learn?

- What are the theoretical guarantees of RL algorithms? How to derive them?
  $\rightarrow$ Sample-complexity bounds
- How to understand generalisation ability of the function approximators and corresponding RL policies?
  $\rightarrow$ Learning theory and generalisation errors meet RL
- How to explore under robust and safe?
  $\rightarrow$ Safe Exploration in RL and regret in robust MDPs

What's the Big Bump Ahead?

Bridging the theory-to-practice gap in RL.

"There is a crack, a crack in everything, that's how the light gets in." -Leonard Cohen

Thanks to our collaborators, teachers, and the audience!

Questions?

References I

[Azar et al., 2017] Azar, M. G., Osband, I., and Munos, R. (2017).
Minimax regret bounds for reinforcement learning.
In *International conference on machine learning*, pages 263–272. PMLR.

[Basu et al., 2019] Basu, D., Senellart, P., and Bressan, S. (2019).
BelMan: Information geometric approach to stcohastic bandits.
In *ECML-PKDD*.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013).
Concentration Inequalities: A Nonasymptotic Theory of Independence.
Oxford University Press.

[Chowdhury and Gopalan, 2019] Chowdhury, S. R. and Gopalan, A. (2019).
Online learning in kernelized markov decision processes.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3197–3205. PMLR.

[Chowdhury et al., 2021] Chowdhury, S. R., Gopalan, A., and Maillard, O.-A. (2021).
Reinforcement learning in parametric mdps with exponential families.
In *AISTATS*. PMLR.

[Fan and Ming, 2021] Fan, Y. and Ming, Y. (2021).
Model-based reinforcement learning for continuous control with posterior sampling.
In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3078–3087. PMLR.

[Garivier and Cappé, 2011] Garivier, A. and Cappé, O. (2011).
The kl-ucb algorithm for bounded stochastic bandits and beyond.
In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 359–376.

[Haque et al., 2024] Haque, I., Tan, Y., Yang, Y., Lan, Q., Lu, J., Mahmood, A. R., Precup, D., and Xu, P. (2024).
More efficient randomized exploration for reinforcement learning via approximate sampling.
Reinforcement Learning Journal, 3(1).

[Honda and Takemura, 2015] Honda, J. and Takemura, A. (2015).
Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards.
J. Mach. Learn. Res., 16:3721–3756.

[Ishfaq et al., 2021] Ishfaq, H., Cui, Q., Nguyen, V., Ayoub, A., Yang, Z., Wang, Z., Precup, D., and Yang, L. (2021).
Randomized exploration in reinforcement learning with general value function approximation.
In *International Conference on Machine Learning*, pages 4607–4616. PMLR.

[Jorge et al., 2024] Jorge, E., Dimitrakakis, C., and Basu, D. (2024).
Isoperimetry is all we need: Langevin posterior sampling for RL.
In *Seventeenth European Workshop on Reinforcement Learning*.

[Karbasi et al., 2023] Karbasi, A., Kuang, N. L., Ma, Y., and Mitra, S. (2023).
Langevin thompson sampling with logarithmic communication: Bandits and reinforcement learning.
In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15828–15860. PMLR.

[Kuang et al., 2023] Kuang, N. L., Yin, M., Wang, M., Wang, Y.-X., and Ma, Y. (2023).
Posterior sampling with delayed feedback for reinforcement learning with linear function approximation.
In *Thirty-seventh Conference on Neural Information Processing Systems*.

[Lai and Robbins, 1985] Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1):4–22.

References III

[Moradipari et al., 2023] Moradipari, A., Pedramfar, M., Zini, M. S., and Aggarwal, V. (2023).
Improved bayesian regret bounds for thompson sampling in reinforcement learning.
In *Thirty-seventh Conference on Neural Information Processing Systems*.

[Osband et al., 2013] Osband, I., Russo, D., and Van Roy, B. (2013).
(more) efficient reinforcement learning via posterior sampling.
In *Advances in Neural Information Processing Systems*, pages 3003–3011.

[Ouhamma et al., 2022] Ouhamma, R., Basu, D., and Maillard, O.-A. (2022).
Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration and planning.
arXiv preprint arXiv:2210.02087.

[Ren et al., 2022] Ren, T., Zhang, T., Szepesvári, C., and Dai, B. (2022).
A free lunch from the noise: Provable and practical exploration for representation learning.
In *Uncertainty in Artificial Intelligence*. PMLR.

[Robbins, 1952] Robbins, H. (1952).
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematical Society, 58(5):527–535.

[Thompson, 1933] Thompson, W. (1933).
On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of two Samples.
Biometrika, 25(3-4):285–294.

[Vakili and Olkhovskaya, 2023] Vakili, S. and Olkhovskaya, J. (2023).
Kernelized reinforcement learning with order optimal regret bounds.
Advances in Neural Information Processing Systems, 36:4225–4247.

References IV

[Xu et al., 2022] Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. (2022).
Langevin Monte Carlo for contextual bandits.
In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24830–24850. PMLR.

[Zanette et al., 2020] Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020).
Frequentist regret bounds for randomized least-squares value iteration.
In *AISTATS*. PMLR.