# Deep Exploration via Randomized Value Functions

Presenter: Yingru Li

The Chinese University of Hong Kong, Shenzhen, China

January 30, 2021

Mainly based on:

Osband, I., Van Roy, B., Russo, D. J., & Wen, Z. (2019). Deep Exploration via Randomized Value Functions. Journal of Machine Learning Research, 20(124), 1-62.

Osband, I. (2016). Deep Exploration via Randomized Value Functions (Doctoral dissertation, Stanford University). —— Won **2017 INFORMS George Dantzig Dissertation Award**

# Outline

## Exploration in Online RL

- Active knowledge acquisition is a key feature of intelligence.
- Exploration is one of the central challenges in reinforcement learning (RL).
- Exploration is also a key engine for data efficiency problem when applying RL in real-world problem.

# Motivating example

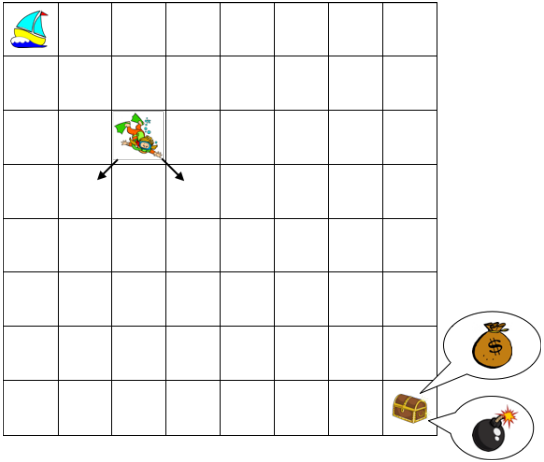


**Figure:** Deep-sea exploration: a simple example where deep exploration is critical.

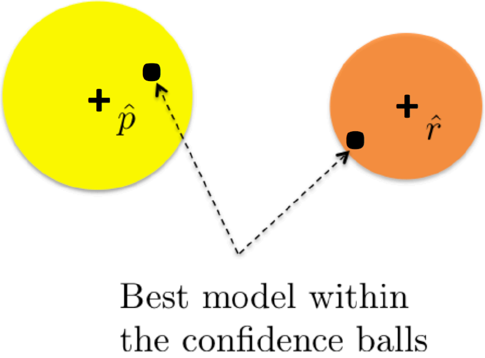
## Importance of Deep Exploration

- myopic: acquire high $\hat{Q}_L$ given the data before episode $L$ or explore immediate information (e.g. action associations).
- dithering: random perturbed the action selected by $\hat{Q}_L$-greedy e.g. $\epsilon$-greedy or boltzmann exploration.
- Deep exploration: the agent needs to consider how actions influence downstream learning opportunities even if expected to no values or immediate information.
- Optimistic algorithm serves as one guiding principle for deep exploration

| exploration method | expected episodes to learn |
| --- | --- |
| optimal | $\Theta(N)$ |
| myopic | $\infty$ |
| dithering | $\Theta\left(2^N\right)$ |
| optimistic | $\Theta(N)$ |

# Review: Optimism in the Face of Uncertainty

"Select the policy which would obtain the best possible rewards in the best (statistically) plausible environment."



## Review: Optimism in the Face of Uncertainty

- In episode $k$, form a uncertainty set $\mathbb{M}_k$ of all statistically plausible models from historical data $\mathcal{H}_{k-1}$, s.t. $M^* \in \mathbb{M}_k$ w.h.p
- Double maximization:

$$(\pi_k, M_k) = \arg\max_{\pi} \max_{M \in \mathbb{M}_k} V(\pi, M)$$

  such that $V(\pi_k, M_k) \geq V(\pi^*, M^*) = V^*$ w.h.p.
- Generally difficult optimization problem

# Review: Optimism in the Face of Uncertainty

- Instead of directly solving double maximization
- OFU principle approximates the benefits of exploration by assigning an optimistic bonus to poorly understood states and actions.
- Value based approach: add UCB bonus to reward function and backward update value function, such that directly ensure,

$$V_k \geq V^*, \quad w.h.p.$$

## Review: Optimism in the Face of Unceratinty

- UCB bonus should be carefully design specialized to particular problem.
- The performance of a UCB algorithm depends critically on the choice of UCBs.
- For tabular MDP: e.g. (Azar et al. '17)

$$b(x,a) = 7HL\sqrt{\frac{1}{N_k(x,a)}}$$

or

$$b(x,a) = \sqrt{\frac{8L\,\mathrm{Var}_{Y\sim\widehat{P}_k(\cdot|x,a)}\left(V_{k,h+1}(Y)\right)}{N_k(x,a)}} + \frac{14HL}{3N_k(x,a)} + \sqrt{\frac{8\sum_y \widehat{P}_k(y \mid x,a)\left[\min\left(\frac{100^2 H^3 S^2 A L^2}{N'_{k,h+1}(y)}, H^2\right)\right]}{N_k(x,a)}}$$

## Review: Optimism in the Face of Uncertainty

- LSVI with Exploration Bonus (e.g., Jin et al '20) for $t = H, \ldots, 1$,

$$\bar{\theta}_t = \left(\sum_{i=1}^{k} \phi_{ti}\phi_{ti}^\top\right)^{-1} \sum_{i=1}^{k} \phi_{ti} \left[r_{ti} + \max_{a^+} \left(\phi\left(s_{t+1}^+, a^+\right)^\top \bar{\theta}_{t+1} + \sqrt{\beta} \left\|\phi\left(s_{t+1}^+, a^+\right)\right\|_{\Sigma_{t+1}^{-1}}\right)\right]$$

- Globally Optimistic LSVI (Zanette et al '20)

$$\max_{\xi_1,\ldots,\xi_H} \max_{a^+} \phi\left(s_{1k}, a^+\right)^\top \bar{\theta}_1$$

s.t. $\quad \|\xi_t\|_{\Sigma_t} \leq \sqrt{\alpha}$

for $t = H, \ldots, 1 \quad \bar{\theta}_t = \left(\sum_{i=1}^{k} \phi_{ti}\phi_{ti}^\top\right)^{-1} \sum_{i=1}^{k} \phi_{ti} \left[r_{ti} + \max_{a^+} \phi\left(s_{t+1}^+, a^+\right)^\top \bar{\theta}_{t+1}\right] + \xi_t$

# Optimistic (UCB-based) algorithms are hard to scale up

- Overwhelmingly, this literature focuses on optimistic algorithms, with most algorithms explicitly maintaining uncertainty sets that are likely to contain the true MDP or constructing UCBs.
- It has been difficult to adapt UCB-based algorithms to the more complex problems investigated in the applied RL literature.
  - Some progress in linear function approximation.
  - No principled solution but some heuristics based on OFU for deep network approximation.

## $\epsilon$-greedy still dominates in applied RL literature

- Most applied papers seem to generate exploration through $\epsilon$-greedy or Boltzmann exploration.
- Those simple methods are compatible with practical value function learning algorithms, which use parametric approximations to value/policy/transition functions to generalize across high dimensional state spaces.
- Unfortunately, such exploration algorithms can fail catastrophically in simple finite state MDPs (e.g. Deep-sea exploration example).

- Today's topic inspired by the search for principled exploration algorithms that both

(1) are compatible with practical function learning algorithms and

(2) provide robust performance (provable guarantee), at least when specialized to simple benchmarks like tabular MDPs.

# Deep Exploration via Radomized Value Functions

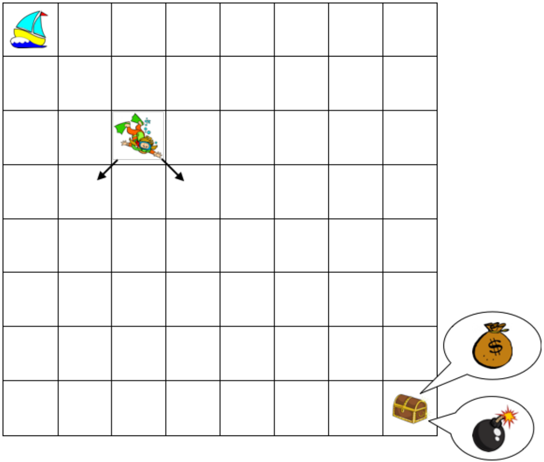


**Figure:** Deep-sea exploration: a simple example where deep exploration is critical.

# Radnomized Exploration is Deep Exploration

| exploration method | expected episodes to learn |
| --- | --- |
| optimal | $\Theta(N)$ |
| myopic | $\infty$ |
| dithering | $\Theta\left(2^N\right)$ |
| optimistic | $\Theta(N)$ |
| randomized | $\Theta(N)$ |

# Outline

# Bayesian linear regression

- Estimate $\theta \in \mathbb{R}^D$ with $N(\bar{\theta})$ prior. (MAP)
- Data $\mathcal{D} = ((x_n, y_n) : n = 1, \ldots, N)$
- "Feature vector" $x_n \in \mathbb{R}^D$ is a row vector, together is $X \in \mathbb{R}^{N \times D}$
- Target $y_n$ is generated from $y_n = x_n\theta + w_n$, where $w_n \overset{i.i.d}{\sim} N(0, v)$, together is $y \in \mathbb{R}^N$.
- Conditioned on $\mathcal{D}$, $\theta$ is Gaussian with

$$\mathbb{E}[\theta \mid \mathcal{D}] = \operatorname*{argmin}_{\theta \in \mathbb{R}^D} \left( \frac{1}{v} \sum_{n=1}^{N} (y_n - x_n\theta)^2 + \frac{1}{\lambda} \|\bar{\theta} - \theta\|^2 \right) = \left( \frac{1}{v} X^\top X + \frac{1}{\lambda} I \right)^{-1} \left( \frac{1}{v} X^\top y + \frac{1}{\lambda} \bar{\theta} \right)$$

and

$$\mathrm{Cov}[\theta \mid \mathcal{D}] = \left( \frac{1}{v} X^\top X + \frac{1}{\lambda} I \right)^{-1}$$

## Randomization via Gaussian noise

- One way of generating a random sample $\widetilde{\theta}_1$ with the same condition distribution as $\theta$ is simply sample from $\widetilde{\theta}_1 \sim N(\mathbb{E}[\theta \mid \mathcal{D}], \mathrm{Cov}[\theta \mid \mathcal{D}])$.
- An alternative construction is given by injecting noise $\hat{\theta} \sim N(\bar{\theta}, \lambda I)$ and $z_n \overset{i.i.d}{\sim} N(0, v)$

$$\widetilde{\theta} \leftarrow \underset{\theta \in \mathbb{R}^D}{\operatorname{argmin}} \left( \frac{1}{v} \sum_{n=1}^{N} (y_n + z_n - x_n \theta)^2 + \frac{1}{\lambda} \|\hat{\theta} - \theta\|^2 \right) \tag{1}$$

$$= \left( \frac{1}{v} X^\top X + \frac{1}{\lambda} I \right)^{-1} \left( \frac{1}{v} X^\top (y + z) + \frac{1}{\lambda} \hat{\theta} \right) \tag{2}$$

- First notice this $\widetilde{\theta}$ is Gaussian.
- We will see why the above $\widetilde{\theta}$ has the same conditional distribution as $\widetilde{\theta}_1$.
- Pointer to Bellman operator of RLSVI (Page 41)

# Randomization via Gaussian noise

- Same conditional expectation

$$\mathbb{E}[\tilde{\theta} \mid \mathcal{D}] = \left(\frac{1}{v}X^\top X + \frac{1}{\lambda}I\right)^{-1} \left(\frac{1}{v}X^\top(y + \mathbb{E}[z \mid \mathcal{D}]) + \frac{1}{\lambda}\mathbb{E}[\hat{\theta} \mid \mathcal{D}]\right) = \mathbb{E}[\theta \mid \mathcal{D}]$$

- Same conditional variance

$$\begin{aligned}
\mathrm{Cov}[\tilde{\theta} \mid \mathcal{D}] &= \left(\frac{1}{v}X^\top X + \frac{1}{\lambda}I\right)^{-1} \left(\frac{1}{v^2}X^\top \mathbb{E}\left[zz^\top \mid \mathcal{D}\right] X + \frac{1}{\lambda^2}\mathbb{E}\left[\hat{\theta}\hat{\theta}^\top \mid \mathcal{D}\right]\right) \left(\frac{1}{v}X^\top X + \frac{1}{\lambda}I\right)^{-1} \\
&= \left(\frac{1}{v}X^\top X + \frac{1}{\lambda}I\right)^{-1} \left(\frac{1}{v}X^\top X + \frac{1}{\lambda}I\right) \left(\frac{1}{v}X^\top X + \frac{1}{\lambda}I\right)^{-1} \\
&= \mathrm{Cov}[\theta \mid \mathcal{D}].
\end{aligned}$$

# Randomization via Gaussian noise

- Randomized least square provides a key understanding to Bayesian linear regression through a purely computational perspective.
- For the linear setting, we see that training a least-squares solution on perturbed versions of the data $\widetilde{\mathcal{D}} = ((x_n, y_n + z_n), n = 1, \ldots, N)$ is equivalent to conjugate Bayesian posterior.

# Gaussian RLSVI = Thompson Sampling in Linear Bandit

## Bayesian Linear Regression

- Prior $\theta \sim N(\mu_0, \lambda^2 I)$
- Observe $\mathcal{H}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$
  - $Y_i = X_i^\top \theta + N(0, \sigma^2)$
- Update posterior $\theta \mid \mathcal{H}_n \sim N(\mu_n, \Sigma_n)$

## Thompson Sampling

- Sample $\hat{\theta} \sim N(\mu_n, \Sigma_n)$
- Play $\operatorname{argmax}_{x \in \mathcal{X}} x^\top \hat{\theta}$

## Posterior Sampling is Equivalent to fitting to perturbed data

*Sample noise to inject:*

1. $\tilde{\theta} \sim N(\mu_0, \lambda^2 I)$
2. $\xi_1, \dots, \xi_n \sim N(0, \sigma^2)$

*Regularized Least-squares on perturbed loss:*

$$\hat{\theta} = \operatorname{argmin}_\theta \sigma^{-2} \sum_{i=1}^{n} \left(x_i^\top \theta - y_i - \xi_i\right)^2 + \lambda^{-2} \|\theta - \tilde{\theta}\|$$

Then

$$\hat{\theta} \mid \mathcal{H}_n \sim N(\mu_n, \Sigma_n)$$

## Implication on scalable approximation for PSRL

- We can think of posterior sampling reinforcement learning (PSRL) as
  - Sample from a posterior of MDPs, then optimize
  - Sample from a posterior over policies, then apply
  - Sample from a posterior over $Q^*$, then argmax
- We can generalize across states/actions via
  - Parametrized models
  - Parameterized policies
  - Parameterized value functions
- In order to generate **approximate** posterior samples for $Q^*$, we can replace the least-square value iteration to an alternative value iteration that trains on randomly perturbed versions of the data.

# RL Notations

- MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho)$.
- $\mathcal{S}$ state space, $\mathcal{A}$ action space, $\mathcal{R}$ reward model, $\mathcal{P}$ transition model, and $\rho$ initial state distribution.

# Value Iteration

**Algorithm 1:** `vi`

**Input:** $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho)$ MDP
$H \in \mathbb{N}$ planning horizon

$Q^*_H \leftarrow 0$ ;

**for** $h$ in $(0, \ldots, H-1)$ **do**

$\quad Q^*_{h+1}(s,a) \leftarrow \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,a}(s') \left( \int r \mathcal{R}_{s,a,s'}(dr) + \max_{a' \in \mathcal{A}} Q^*_h(s',a') \right) \; \forall s,a \in \mathcal{S} \times \mathcal{A}$ ;

**end**

# Value function learning

- Value function family $\mathcal{Q} = \{\mathcal{Q}_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$, e.g. linear function $\mathcal{Q}_\theta(s,a) = \phi(s,a)^\top \theta$
- Observed data $\mathcal{D} = \{(s_t, a_t, r_t, s'_t, t)\}$
- Target parameters $\theta^-$
- we define the empirical temporal difference (TD) loss

$$\mathcal{L}\left(\theta; \theta^-, \mathcal{D}\right) := \sum_{t \in \mathcal{D}} \left( \underbrace{r_t + \max_{a' \in \mathcal{A}} \mathcal{Q}_{\theta^-}\left(s'_t, a'\right)}_{y_t} - \mathcal{Q}_\theta\left(s_t, a_t\right) \right)^2$$

- and Regularizer

$$\mathcal{R}\left(\theta; \theta^p\right) := \frac{v}{\lambda} \left\|\theta^p - \theta\right\|_2^2$$

# Least Square Value Iteration

**Algorithm 2:** `learn_lsvi`

| | | |
|---|---|---|
| **Agent:** | $\mathcal{L}(\theta{=}\cdot\,;\theta^{-}{=}\cdot\,,\mathcal{D}{=}\cdot\,)$ | TD error loss function |
| | $\mathcal{R}(\theta{=}\cdot\,;\theta^{p}{=}\cdot\,)$ | regularization function |
| | `buffer` | memory buffer of observations |
| | `prior` | prior distribution of $\theta$ |
| | $H \in \mathbb{N}$ | planning horizon |

$\tilde{\theta}_H \leftarrow \mathbf{null}$, Data $\tilde{\mathcal{D}} \leftarrow$ `buffer.data()` ;

Prior parameter $\tilde{\theta}^p \leftarrow$ `prior.mean()` ;

**for** $h$ in $(0, \ldots, H-1)$ **do**

$\quad \tilde{\theta}_h \leftarrow \arg\min_{\theta \in \mathbb{R}^D} \left( \mathcal{L}(\theta; \tilde{\theta}_{h+1}, \tilde{\mathcal{D}}) + \mathcal{R}(\theta; \tilde{\theta}^p) \right)$

**end**

# Randomized Least Square Value Iteration

**Algorithm 3:** `learn_rlsvi`

**Agent:**

| | |
|---|---|
| $\mathcal{L}(\theta{=}\cdot\,;\theta^{-}{=}\cdot\,,\mathcal{D}{=}\cdot\,)$ | TD error loss function |
| $\mathcal{R}(\theta{=}\cdot\,;\theta^{p}{=}\cdot\,)$ | regularization function |
| `buffer` | memory buffer of observations |
| `prior` | prior distribution of parameters |
| $H \in \mathbb{N}$ | planning horizon |

$\tilde{\theta}_H \leftarrow$ **null**, Data $\tilde{\mathcal{D}} \leftarrow$ `buffer.sample_perturbed_data()` ;

/* $[(s_t, a_t, r_t + z_t, s'_t, t)\,, \forall t \in \texttt{buffer}, z_t \sim N(0, v)]$ */

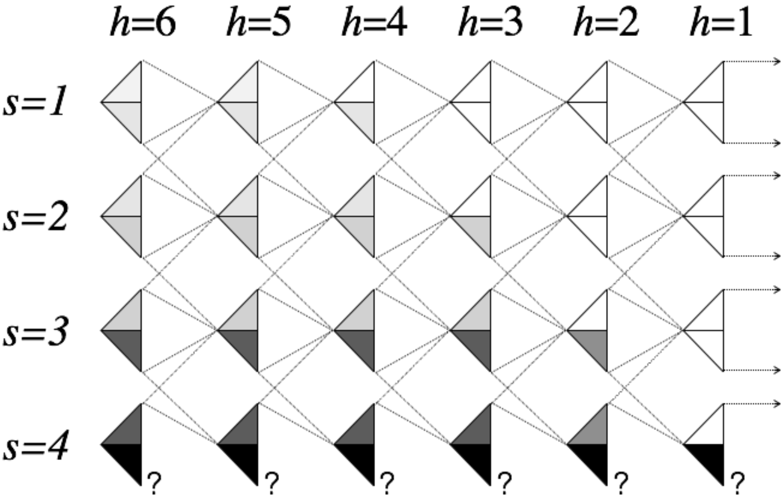
**for** $h$ in $(0, \ldots, H-1)$ **do**

  Prior parameter $\tilde{\theta}^p \leftarrow$ `prior.sample()` ;

  $\tilde{\theta}_h \leftarrow \arg\min_{\theta \in \mathbb{R}^D} \left(\mathcal{L}(\theta; \tilde{\theta}_{h+1}, \tilde{\mathcal{D}}) + \mathcal{R}(\theta; \tilde{\theta}^p)\right)$

**end**

# Illustration of how RLSVI achieves deep exploration



# Outline

# Outline

Background

RLSVI

Theoretical Analysis

Bayesian Regret Bound

# Notations for finite-horizon inhomogeneous MDP

- This can be formulated as a special case the paper's general formulation as follows.
- Assume $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_{H-1}$
- The state always advances from some state $s_t \in \mathcal{S}_t$ to $s_{t+1} \in \mathcal{S}_{t+1}$
- The process terminates w.p. 1 in period $H$.
- Assume each set $\mathcal{S}_0, \ldots, \mathcal{S}_{H-1}$ contains an equal number of elements.
- The sequence of observations made during episode $\ell$ is

$$\mathcal{O}_\ell = \left(s_0^\ell, a_0^\ell, r_1^\ell, s_1^\ell, a_1^\ell, \ldots, s_{H-1}^\ell, a_{H-1}^\ell, r_H^\ell\right)$$

- History observations before episode $\ell$,

$$\mathcal{H}_{\ell-1} = (\mathcal{O}_1, \ldots, \mathcal{O}_{\ell-1})$$

# Notations for finite-horizon inhomogeneous MDP

- $s \in \mathcal{S}_t$ can be written as a pair $s = (t, x)$ where $t \in \{0, \ldots, H-1\}$ and $x \in \mathcal{X} = \{1, \ldots, |\mathcal{S}_0|\}$.
- Similarly, a policy $\pi : \mathcal{S} \to \mathcal{A}$ can be viewed as a sequence $\pi = (\pi_0, \ldots, \pi_{H-1})$ where $\pi_t : x \mapsto \pi((t, x))$.
- Transition probabilities as $\mathcal{P}_{t,x,a}(x') \equiv \mathcal{P}_{(t,x),a}((t+1, x'))$,
- Reward probabilities as $\mathcal{R}_{t,x,a,x'}(r) \equiv \mathcal{R}_{(t,x),a,(t+1,x')}(r)$.
- Value $V^{\pi}_{\mathcal{M},t}(x) \equiv V^{\pi}_{\mathcal{M}}((t,x)) = \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{h=t+1}^{H} r_h \mid s_t = (t,x)\right]$ and Optimal Value $V^{*}_{\mathcal{M},t}(x) := \max_{\pi} V^{\pi}_{\mathcal{M},t}(x)$
- State-action value function $Q^{\pi}_{\mathcal{M},t}(x,a) = \mathbb{E}\left[r_{t+1} + V^{\pi}_{\mathcal{M},t+1}(x_{t+1}) \mid \mathcal{M}, x_t = x, a_t = a\right]$ and similar definition for optimal one.

## Beyasian Regret

- **Regret** of algorithm $\mathrm{alg}$ over $L$ episodes on underlying MDP $\mathcal{M}$:

$$\mathrm{Regret}(\mathcal{M}, \mathrm{alg}, L) = \sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{M},\mathrm{alg}} \left[ V^{*}\left(s_0^{\ell}\right) - V^{\pi^{\ell}}\left(s_0^{\ell}\right)\right]$$

- **Bayesian Regret** with a prior (representative distribution) over MDPs $\mathbb{P}(\mathcal{M} \in \mathbb{M})$:

$$\mathrm{BayesRegret}(\mathrm{alg}, L) = \mathbb{E}[\mathrm{Regret}(\mathcal{M}, \mathrm{alg}, L)]$$

- **Frequentist (worst-case) Regret** holds for any MDP instance $\mathcal{M} \in \mathbb{M}$

$$\mathrm{WorstRegret}(\mathrm{alg}, L) = \sup_{\mathcal{M} \in \mathbb{M}} \mathrm{Regret}(\mathcal{M}, \mathrm{alg}, L)$$

# Assumptions for Bayesian Regret Analysis

- **Outcome of the decision:** $o = (x', r)$

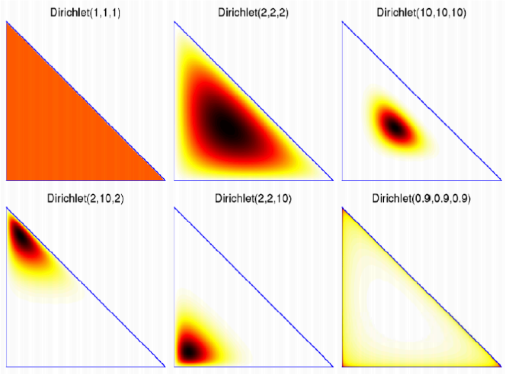
**Assumption 1 (Independent Dirichlet prior for Outcomes).**

*Rewards take values in {0,1} and so the cardinality of the outcome space is $|\mathcal{X} \times \{0,1\}| = 2|\mathcal{X}|$. For each, $(t, x, a) \in \{0, \ldots, H-2\} \times \mathcal{X} \times \mathcal{A}$, the outcome distribution is drawn from a Dirichlet prior*

$$\mathcal{P}^{O}_{t,x,a}(\cdot) \sim \mathrm{Dirichlet}\left(\alpha_{0,t,x,a}\right)$$

*for $\alpha_{0,t,x,a} \in \mathbb{R}_{+}^{2|\mathcal{X}|}$ and each $\mathcal{P}^{O}_{t,x,a}$ is drawn independently across $(t, x, a)$. Assume there is $\beta \geq 3$ such that $\mathbf{1}^{\top}\alpha_{0,t,x,a} = \beta$ for all $(t, x, a)$.*

*Remark: Dirichlet prior is the conjugate prior for multinomial distribution. Dirichlet-multinomial is a generalization of Beta-bernoulli.*




- Assumption in the paper assume $\beta \geq 3$ to avoid extreme distributions (right-down figure).
- As more data gathered, $\mathbf{1}^\top \alpha_{\ell,t,x,a} \to \infty$, Dirichlet posterior distribution concentrates.

## Empirical and posterior distribution of Outcomes

- $D_{\ell-1}(t,x,a) = \left\{\left(r_{t+1}^k, x_{t+1}^k\right) : k < \ell, x_t^k = x, a_t^k = a\right\}$
- $n_\ell(t,x,a) = |D_{\ell-1}(t,x,a)|$
- $\hat{P}_{\ell,y}^O(r',x')$: the empirical distribution over outcomes $(r',x')$ in the dataset $D_{\ell-1}(y)$
- Under Dirichlet prior assumption, the posterior transition probabilities are distributed as

$$\mathcal{P}_y^O(\cdot) \mid \mathcal{H}_{\ell-1} \sim \mathrm{Dirichlet}\left(\alpha_{\ell,y}\right) \text{ where } \alpha_{\ell,y} = \alpha_{0,y} + n_\ell(y)\hat{P}_{\ell,y}^O \in \mathbb{R}^{2|\mathcal{X}|}$$

  for any triple $y = (t,x,a)$.
- The posterior mean of $\mathcal{P}_y^O$ as a weighted linear combination of the prior and the empirical observations:

$$\mathbb{E}\left[\mathcal{P}_y^O \mid \mathcal{H}_{\ell-1}\right] = \frac{\alpha_{0,y} + n_\ell(y)\hat{P}_{\ell,y}^O}{\beta + n_\ell(y)}$$

## Bayesian regret bound for RLSVI in tabular setting

- Tabular representation: $\mathcal{Q}_\theta = \theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $\mathcal{Q}_\theta(s,a) = \theta_{s,a}$ and $\phi(s,a) = \mathbf{1}_{(s,a)}$ is a one-hot vector.

**Theorem 1 (Bayesian regret bound for RLSVI).**

*Consider an RLSVI agent with an infinite buffer, greedy actions and with tabular representation. Under Independent Dirichlet Prior assumption with $\beta \geq 3$, if this version of $RLSVI$ is applied with planning horizon $H$, and parameters $v = 3H^2, \bar{\theta} = H\mathbf{1}$ and $v/\lambda = \beta$, then for all $L \in \mathbb{N}$,*

$$\textit{BayesRegret}\left(\mathrm{RLSVI}_{\bar{\theta},v,\lambda}, L\right) \leq 6H^2\sqrt{\beta|\mathcal{X}||\mathcal{A}|L\log_+(1+|\mathcal{X}||\mathcal{A}|HL)}\log_+\left(1+\frac{L}{|\mathcal{X}||\mathcal{A}|}\right) \quad (3)$$

$$\textit{BayesRegret}\left(\mathrm{RLSVI}_{\bar{\theta},v,\lambda}, L\right) \leq 5\beta H^3|\mathcal{X}|||\mathcal{A}|\sqrt{\log_+(1+|\mathcal{X}||\mathcal{A}|HL)}\log_+\left(1+\frac{L}{|\mathcal{X}||\mathcal{A}|}\right) \quad (4)$$

$$+ 2H^2\sqrt{6|\mathcal{X}||\mathcal{A}|L\log(|\mathcal{X}||\mathcal{A}|)}$$

$$\textit{where } \log_+(x) = \max\{1, \log(x)\}$$

# Comments on the Bayesian regret bound

1. The parameter $\beta$ governs the relative strength of prior mean $\bar{\theta}$ in the $Q$-functions sampled by RLSVI, typically a constant.
2. When $L$ large, second term dominates.

- In both case, this regret bound is $\tilde{O}\left(H^2\sqrt{|\mathcal{X}||\mathcal{A}|L}\right)$ where $\tilde{O}$ ignores poly-logarithmic factors.

$$\text{BayesRegret}\left(\mathrm{RLSVI}_{\bar{\theta},v,\lambda}, L\right) = \tilde{O}(H\sqrt{H|\mathcal{X}||\mathcal{A}|T})$$

# Comments on the Bayesian regert bound

- BayesRegret $\left(\mathrm{RLSVI}_{\bar{\theta},v,\lambda}, L\right) = \tilde{O}(H\sqrt{H|\mathcal{X}||\mathcal{A}|T})$
- Minimax lower bound (Not apple-to-apple comparison)

$$\inf_{\mathrm{alg}} \sup_{\mathcal{M}} \mathsf{Regret}(\mathcal{M}, \mathrm{alg}, L) = \Omega(H\sqrt{|\mathcal{X}||\mathcal{A}|T})$$

- Can we prove the following by Sion's minimax theorem? (Open question?)

$$\sup_{\mathrm{prior}(\mathcal{M})} \inf_{\mathrm{alg}} \mathsf{BayesRegret}\left(\mathrm{alg}, L\right) = \inf_{\mathrm{alg}} \sup_{\mathcal{M}} \mathsf{Regret}(\mathcal{M}, \mathrm{alg}, L)$$

## Stochastic Bellman Operator

- **Induced value function:** For a state-action value function $Q \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ define the corresponding value function $V_Q \in \mathbb{R}^{2|\mathcal{X}|}$ over outcomes by

$$V_Q\left(r, x'\right) := r + \max_{a \in \mathcal{A}} Q\left(x', a\right) \quad \forall (x', r)$$

- **True Bellman Operator.** For $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ the true Bellman operator at timestep $t$ applied to $Q$ is defined by

$$\begin{aligned} F_{\mathcal{M},t} Q(x,a) &= \mathbb{E}\left[r_{t+1} + \max_{a' \in \mathcal{A}} Q(x_{t+1}, a') \mid \mathcal{M}, x_t = x, a_t = a\right] \\ &= \mathbb{E}\left[V_Q(r_{t+1}, x_{t+1}) \mid \mathcal{M}, x_t = x, a_t = a)\right] \\ &= V_Q^\top \mathcal{P}_{t,x,a}^O \end{aligned}$$

Remark: True Bellman Operator is also random variable related to Dirichlet.

## Stochastic Bellman Operator

- **Bellman Operator of RLSVI** Equation 1 (Gaussian)

$$F_{\ell,t}Q(x,a) := \sigma_\ell^2(t,x,a)\left(\frac{\bar{\theta}_{t,x,a}}{\lambda} + \frac{1}{v}\left(\sum_{(r,x')\in\mathcal{D}_{\ell-1}(t,x,a)} r + \max_{a'\in\mathcal{A}} Q\left(x',a'\right)\right)\right) + w_\ell(t,x,a)$$

where $w_\ell(t,x,a) \mid \mathcal{H}_{\ell-1} \sim N\left(0,\sigma_\ell^2(t,x,a)\right)$ and

$$\sigma_\ell^2(t,x,a) = \left(\frac{1}{\lambda} + \frac{n_\ell(t,x,a)}{v}\right)^{-1} = \frac{v}{n_\ell(t,x,a) + v/\lambda}$$

- $w_\ell(y)/\sigma_\ell(y) \sim N(0,1)$ is drawn independently across episodes $\ell$ and triples $y = (t,x,a)$
- In episode $\ell$, RLSVI generates $Q_{\ell,1},\ldots,Q_{\ell,H}$ where $Q_{\ell,H} = 0 \in \mathbb{R}^{|\mathcal{X}\|\mathcal{A}|}$ and for all $t < H$,

$$Q_{\ell,t} = F_{\ell,t}Q_{\ell,t+1}.$$

## Stochastic Bellman Operator

- Connection between RLSVI Bellman operator and the empirical distribution $\hat{P}^{O}_{\ell,y}$ from $y=(t,x,a)$

$$\sum_{(r,x')\in D_{\ell-1}(y)} \left( r + \max_{a'\in\mathcal{A}} Q\left(x',a'\right) \right) = n_\ell(y) V_Q^T \hat{P}^{O}_{\ell,y}$$

- From direct calculation,

$$F_{\ell,t}Q(x,a) = \frac{(v/\lambda)\bar{\theta}_y + n_\ell(y)V_Q^T \hat{P}^{O}_{\ell,y}}{(v/\lambda)+n_\ell(y)} + w_\ell(y) \quad \forall y=(t,x,a)$$

- Bellman update of RLSVI differs from the empirical Bellman update $V_Q^T \hat{P}^{O}_{\ell,y}$ in two ways:
  1. slight regularization toward the prior mean $\bar{\theta}$,
  2. adds independent Gaussian noise to each update.

## Optimism and regret decompositions

- Regret decomposition in one episode,

$$V^*_{\mathcal{M},0}(x) - V^\pi_{\mathcal{M},0}(x) = \underbrace{\left(\max_{a\in\mathcal{A}} Q^*_{\mathcal{M},0}(x,a) - \max_{a\in\mathcal{A}} Q_0(x,a)\right)}_{pessimism} + \underbrace{\left(\max_{a\in\mathcal{A}} Q_0(x,a) - V^\pi_{\mathcal{M},0}(x)\right)}_{prediction/planning\ error}$$

**Lemma 1 (Planning Error to On Policy Bellman Error).**

*Let* $Q_0, Q_1, Q_2, \ldots, Q_H \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ *be any sequence with* $Q_H = 0 \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ *and take* $\pi = (\pi_0, \pi_1, \ldots, \pi_{H-1})$ *to be a policy with* $\pi_t(x) \in \arg\max_{a\in\mathcal{A}} Q_t(x,a)$ *for all* $x$*. Then for any MDP* $\mathcal{M}$ *and initial state* $x \in \mathcal{X}$*,*

$$Q_0\left(x, \pi_0(x)\right) - V^\pi_{\mathcal{M},0}(x) = \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{t=0}^{H-1} \left(Q_t - F_{\mathcal{M},t}Q_{t+1}\right)(x_t, a_t) \mid x_0 = x\right]$$

- The sequence $(Q_{\ell,0}, \ldots, Q_{\ell,H})$ generated by RLSVI in some episode $\ell$. On policy Bellman error can be simplified further by plugging in $Q_{\ell,t} = F_{\ell,t} Q_{\ell,t+1}$.

**Corollary 2 (Optimistic regret bounds).**

*For any episode* $\ell \in \mathbb{N}$, *if*

$$\mathbb{E}\left[\max_{a \in \mathcal{A}} Q_{\ell,0}\left(x_0^\ell, a\right)\right] \geq \mathbb{E}\left[\max_{a \in \mathcal{A}} Q^*_{\mathcal{M},0}\left(x_0^\ell, a\right)\right],$$

*then*

$$\mathbb{E}\left[V^*_{\mathcal{M},0}\left(x_0^\ell\right) - V^{\pi_\ell}_{\mathcal{M},0}\left(x_0^\ell\right)\right] \leq \mathbb{E}\left[\sum_{t=0}^{H} \left(F_{\ell,t} Q_{\ell,t+1} - F_{\mathcal{M},t} Q_{\ell,t+1}\right)\left(x_t^\ell, a_t^\ell\right)\right]$$

# Stochastic Optimism

- Goal: prove the stronger condition under Dirichlet prior assumption with appropriately chosen parameters $\lambda, v, \bar{\theta}$

$$\mathbb{E}\left[\max_{a \in \mathcal{A}} Q_{\ell,0}\left(x_0^\ell, a\right) \mid \mathcal{H}_{\ell-1}\right] \geq \mathbb{E}\left[\max_{a \in \mathcal{A}} Q^*_{\mathcal{M},0}\left(x_0^\ell, a\right) \mid \mathcal{H}_{\ell-1}\right]$$

# Stochastic Optimism

**Definition 3 (Stochastic optimism).**

A random variable $X$ is stochastically optimistic with respect to another random variable $Y$, written $X \succeq_{SO} Y$, if for all convex increasing functions $u : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)] \tag{5}$$

**Example 4 (Stochastic optimism in Gaussian random variables).**

If $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$ then $X \succeq_{SO} Y$ if and only if $\mu_X \geq \mu_Y$ and $\sigma_X^2 \geq \sigma_Y^2$.

**Remark 1.**

*Our goal then is to show if RLSVI is applied with appropriate parameters, it generates iterates that are larger and noisier than the true $Q^*$.*

# Stochastic Optimism

- This definition of $SO$ closely mirrors that of "second order stochastic dominance", which is widely used in decision theory (Hadar and Russell, 1969).
- A random payout $X$ is second order stochastically dominant with respect to $Y$ if (5) holds for all concave increasing function $u$.
- This means that any rational risk-averse agent prefers $X$ to $Y$,
- while $X \succeq_{SO} Y$ implies that any rational risk-loving agent prefers $X$ to $Y$.

## Stochastic Optimism

**Lemma 5 (Preservation of optimism under convex operations).**

*For any two collections* $(X_1, \ldots, X_n)$ *and* $(Y_1, \ldots, Y_n)$ *of independent random variables with* $X_i \succeq_{SO} Y_i$ *for each* $i \in \{1, \ldots n\}$ *and any convex increasing function* $f : \mathbb{R}^n \to \mathbb{R}$

$$f(X_1, \ldots, X_n) \succeq_{SO} f(Y_1 \ldots, Y_n)$$

- If $X \succeq_{SO} Y$ we can conclude $X + Z \succeq_{SO} Y + Z$
- For two pairs of independent random variables $(X_1, X_2)$ and $(Y_1, Y_2)$ with $X_1 \succeq_{SO} Y_1$ and $X_2 \succeq_{SO} Y_2$,

$$\max\{X_1, X_2\} \succeq_{SO} \max\{Y_1, Y_2\}$$

# Stochastic Optimism

**Lemma 6 (Monotonicity).**

*Fix two random $Q$ functions $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$. Suppose that conditioned on $\mathcal{H}_{\ell-1}$, for each $i = 1, 2$ the entries of $Q_i(x, a)$ are drawn independently across $x, a$, and drawn independently of the RLSVI noise terms $w_\ell(t, x, a)$. Then*

$$Q_1(x,a)\,|\mathcal{H}_{\ell-1} \succeq_{SO} Q_2(x,a)|\,\mathcal{H}_{\ell-1} \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A}$$

*implies*

$$F_{\ell,t}Q_1(x,a)\,|\mathcal{H}_{\ell-1} \succeq_{SO} F_{\ell,t}Q_2(x,a)|\,\mathcal{H}_{\ell-1} \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A}, t \in \{0, \ldots, H-1\}$$

# Stochastic Optimism

**Lemma 7 (Gaussian vs Dirichlet optimism).**

*Let* $Y = P^T V$ *for* $V \in \mathbb{R}^n$ *fixed and* $P \sim \mathrm{Dirichlet}(\alpha)$ *with* $\alpha \in \mathbb{R}^n_+$ *and* $\sum_{i=1}^n \alpha_i \geq 3$. *Let* $X \sim N\left(\mu, \sigma^2\right)$ *with* $\mu \geq \frac{\sum_{i=1}^n \alpha_i V_i}{\sum_{i-1}^n \alpha_i}, \sigma^2 \geq 3\left(\sum_{i=1}^n \alpha_i\right)^{-1} \mathrm{Span}(V)^2$, *then* $X \succeq_{SO} Y$

**Lemma 8 (Stochastically optimistic operators).**

*Suppose Dirichlet prior assumption holds and* $RLSVI$ *is applied with parameters* $(\bar{\theta}, v, \lambda)$ *satisfying* $(v/\lambda) = \beta$. *Then for any episode* $\ell$ *with history* $\mathcal{H}_{\ell-1}$, *time* $t \in \{0, \ldots, H-1\}$, *and pair* $(x, a) \in \mathcal{X} \times \mathcal{A}$

$$F_{\ell,t} Q(x,a) \,|\mathcal{H}_{\ell-1} \succeq_{SO} F_{\mathcal{M},t} Q(x,a)|\, \mathcal{H}_{\ell-1}$$

*for any fixed* $Q \in \mathbb{R}^{|\mathcal{X}\|\mathcal{A}|}$ *such that* $v \geq 3\,\mathrm{Span}\left(V_Q\right)^2$ *and* $\max_{x \in \mathcal{X}} V_Q(x) \leq \min_{t,x,a} \bar{\theta}_{t,x,a}$

## Corollary 9.

*If Dirichlet prior assumption holds and RLSVI is applied with parameters* $(\bar{\theta}, v, \lambda)$ *satisfying* $(v/\lambda) = \beta, v \geq 3H^2$ *and* $\min_y \bar{\theta}_y \geq H$

$$Q_{\ell,0}(x,a) \,\middle|\mathcal{H}_{\ell-1} \succeq_{SO} Q^*_{\mathcal{M},0}(x,a)\middle|\, \mathcal{H}_{\ell-1}$$

*for any history* $\mathcal{H}_{\ell-1}$ *and state-action pair* $(x,a) \in \mathcal{X} \times \mathcal{A}$

- $(F_{1,H-1}0)\,(x,a) \succeq_{SO} (F_{\mathcal{M},H-1}0)\,(x,a) \quad \forall x, a$
- Proceeding by induction, suppose for some $t \leq H-1$

$$(F_{1,t+1}F_{1,t+2}\cdots F_{1,H-1}0)\,(x,a) \succeq_{SO} (F_{\mathcal{M},t+1}F_{\mathcal{M},t+2}\cdots F_{\mathcal{M},H-1}0)\,(x,a) \quad \forall x, a$$

- $$\begin{aligned} F_{1,t}\,(F_{1,t+1}F_{1,t+2}\cdots F_{1,H-1}0)\,(x,a) &\succeq_{SO} \quad F_{1,t}\,(F_{\mathcal{M},t+1}F_{\mathcal{M},t+2}\cdots F_{\mathcal{M},H-1}0)\,(x,a) \\ &\succeq_{SO} \quad F_{\mathcal{M},t}\,(F_{\mathcal{M},t+1}F_{\mathcal{M},t+2}\cdots F_{\mathcal{M},H-1}0)\,(x,a) \end{aligned}$$

## Analysis of on-policy Bellman error

- Denote $\Delta_\ell = V^*_{\mathcal{M},0}\left(x_0^\ell\right) - V^{\pi_\ell}_{\mathcal{M},0}\left(x_0^\ell\right)$, by Corollary 2 and Corollary 9,

$$\mathbb{E}\left[\sum_{\ell=1}^{L}\Delta_\ell\right] \leq \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{t=0}^{H-1}\left(F_{\ell,t}Q_{\ell,t+1} - F_{\mathcal{M},t}Q_{\ell,t+1}\right)\left(x_t^\ell, a_t^\ell\right)\right]$$
$$= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{t=0}^{H-1}\left(\left(F_{\ell,t}Q_{\ell,t+1}\right)\left(x_t^\ell, a_t^\ell\right) - \mathbb{E}\left[F_{\mathcal{M},t}Q_{\ell,t+1}\left(x_t^\ell, a_t^\ell\right) \mid \mathcal{H}_{\ell-1}\right]\right)\right]$$

- Recall the definition of two Bellman Operators:

$$\mathbb{E}\left[F_{\mathcal{M},t}Q(x,a) \mid \mathcal{H}_{\ell-1}\right] = \frac{V_Q^T\alpha_{0,y} + n_\ell(y)V_Q^T\hat{P}^O_{\ell,y}}{\beta + n_\ell(y)} \geq \frac{-\beta\left\|V_Q\right\|_\infty}{\beta + n_\ell(y)} + \frac{n_\ell(y)V_Q^T\hat{P}^O_{\ell,y}}{\beta + n_\ell(y)}$$
$$F_{\ell,t}Q(x,a) = \frac{\beta\bar{\theta}_y + n_\ell(y)V_Q^T\hat{P}^O_{\ell,y}}{\beta + n_\ell(y)} + w_\ell(y), \text{ where } y = (t,x,a)$$

# Analysis of on-policy Bellman error

- $Q_{\ell,t+1}$ and $F_{\mathcal{M},t}$ are independent conditioned on $\mathcal{H}_{\ell t}$

$$\left(F_{\ell,t}Q_{\ell,t+1}\right)\left(x_t^\ell, a_t^\ell\right) - \mathbb{E}\left[F_{\mathcal{M},t}Q_{\ell,t+1}\left(x_t^\ell, a_t^\ell\right) \mid \mathcal{H}_{\ell-1}\right] = \frac{\beta\left(\bar{\theta}_{t,x_t^\ell,a_t^\ell} + \left\|V_{Q_{\ell,t+1}}\right\|_\infty\right)}{\beta + n_\ell(t, x_t^\ell, a_t^\ell)} + w_\ell(t, x_t^\ell, a_t^\ell)$$

$$\mathbb{E}\sum_{\ell=1}^{L}\Delta_\ell \leq \mathbb{E}\left[\beta\left(\underbrace{\|\bar{\theta}\|_\infty}_{=H} + \underbrace{\max_{\ell\leq L, t<H}\left\|V_{Q_{\ell,t+1}}\right\|_\infty}_{\text{Lemma 10}}\right)\underbrace{\sum_{t<H,\ell\leq L}\frac{1}{\beta + n_\ell\left(t, x_t^\ell, a_t^\ell\right)}}_{\text{Integral inequality}} + \underbrace{\sum_{\ell\leq L, t\leq H} w_\ell\left(t, x_t^\ell, a_t^\ell\right)}_{\text{lemma 10}}\right]$$

**Lemma 10 (Proposition 1 and 8 in Russo and Zhou, IEEE Transactions on Information Theory (Volume: 66, Issue: 1, Jan. 2020) ).**

*Let $(X, J)$ be jointly distributed random variables where $X \in \mathbb{R}^n$ follows a multivariate Gaussian distribution with $X_j \sim N\left(0, \sigma_j^2\right)$ and $J \in \{1, \dots n\}$ is a random index. Then*

$$\mathbb{E}\left[X_J\right] \leq \sqrt{2I(J;X)\mathbb{E}\left[\sigma_J^2\right]} \leq \sqrt{2\log(n)\mathbb{E}\left[\sigma_J^2\right]}$$

- $\mathbb{E}\left[w_\ell\left(t, x_t, a_t\right)\right] \leq \sqrt{2\log(|\mathcal{A}||\mathcal{X}|)\mathbb{E}\left[\sigma_\ell\left(t, x_t, a_t\right)^2\right]}$
- $\mathbb{E}\left[\max_{\ell \leq L, t < H}\left\|V_{Q_{\ell,t+1}}\right\|_\infty\right] \leq 2H + H^2\sqrt{2\log(1 + |\mathcal{X}||\mathcal{A}|HL)}$ when $v/\lambda = \beta \geq 3, v = 3H^2$ and $\bar{\theta} = H\mathbf{1}$.

# Proof of Lemma 10

**Fact 11 (Donsker-Varadhan representation, Theorem 4.1 in Stanford Stat311/EE377 Lecture notes, Duchi J.).**

*Let $P$ and $Q$ be distributions on a common space $\mathcal{X}$. Then*

$$D_{\mathrm{kl}}(P\|Q) = \sup_g \left\{ \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q \left[ e^{g(X)} \right] \right\}$$

*where the supremum is taken over measurable functions $g : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q \left[ e^{g(X)} \right] < \infty$.*

- Applying the Fact with $P = \mathbb{P}\left(X_j = \cdot \mid J = j\right)$ and $Q = \mathbb{P}\left(X_j = \cdot\right)$, since $\{\lambda X_j : \lambda \in \mathbb{R}\}$ is a measurable function class, we have

$$D_{\mathrm{kl}}\left(\mathbb{P}\left(X_j = \cdot \mid J = j\right) \mid \mathbb{P}\left(X_j = \cdot\right)\right) \geq \sup_\lambda \lambda \mathbb{E}\left[X_j | J = j\right] - \lambda^2 \sigma_j^2 / 2$$

  where the optimizer is $\lambda = \mathbb{E}\left[X_j | J = j\right] / \sigma_j^2$.

# Proof of Lemma 10

- Taking $\lambda$ to be the optimizer, we have $\mathbb{E}\left[X_j \mid J=j\right] \leq \sigma_j \sqrt{2 D_{\mathrm{kl}}\left(\mathbb{P}\left(X_j=\cdot \mid J=j\right) \mid \mathbb{P}\left(X_j=\cdot\right)\right)}$ and then

$$
\begin{aligned}
\mathbb{E}\left[X_J\right] &= \sum_{J=j} \mathbb{P}(J=j) \mathbb{E}\left[X_J \mid J=j\right] \\
&\leq \sum_j \mathbb{P}(J=j) \sigma_j \sqrt{2 D_{\mathrm{kl}}\left(\mathbb{P}\left(X_j=\cdot \mid J=j\right) \mid \mathbb{P}\left(X_j=\cdot\right)\right)} \\
&\overset{\text{CS}}{\leq} \sqrt{\sum_j \mathbb{P}(J=j) \sigma_j^2} \sqrt{2 \sum_j \mathbb{P}(J=j) D_{\mathrm{kl}}\left(\mathbb{P}\left(X_j=\cdot \mid J=j\right) \mid \mathbb{P}\left(X_j=\cdot\right)\right)} \\
&\overset{\text{DP}}{\leq} \sqrt{\mathbb{E}\left[\sigma_J^2\right]} \sqrt{2 \sum_j \mathbb{P}(J=j) D_{\mathrm{kl}}(\mathbb{P}(X=\cdot \mid J=j) \mid \mathbb{P}(X=\cdot))} \\
&= \sqrt{2 \mathbb{E}\left[\sigma_J^2\right] I(J ; X)}
\end{aligned}
$$