1) Model-Based RL in Query model
(MBRL)

Query model: any (s,a) we can observe sample
$s' \sim P(s,a)$ (or $s' = f(s,a,w)$ w/ $w \sim D$)

Black-box
- games
- Simulator

Sample complexity: How many samples are required for near-optimal performance?

Meta-Algorithm (MBRL)

1) For $i = 1, \dots, N$
sample $s_i' \sim P(\underbrace{s_i, a_i}_{\text{specify}})$ record (s_i', s_i, a_i)

2) $\underbrace{\text{Fit}}_{\text{specify}}$ transition model $\hat{P}$ using $\{(s_i', s_i, a_i)\}_{i=1}^{N}$

3) $\underbrace{\text{Design}}_{\text{specify}}$ $\hat{\pi}$ using $\hat{P}$

2) Tabular Setting

1) Samples (s_i, a_i) evenly: $\frac{N}{SA}$ $(N > SA)$ times each

2) Fit transition model

$$\hat{P}(s' \mid \underline{s}, \underline{a}) = \frac{\sum_{i=1}^{N} \mathbb{1}\{s_i = s \ \& \ a_i = a\} \mathbb{1}\{s_i' = s'\}}{\sum_{i=1}^{N} \mathbb{1}\{s_i = s, a_i = a\}}$$

3) Design $\hat{\pi}$ Policy Iteration $\hat{\pi} = PI(\hat{P}, r)$

Recall: $PI(P, r)$

Initialize π^0

For $t = 1, \dots, T$

$$V^{\pi^t} = (I - \gamma P^{\pi})^{-1} R^{\pi}$$

$$Q^{\pi^t}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} V^{\pi}(s')$$

$$Q^{\pi^t} = \text{Policy } \underline{\text{Eval}}(\pi^t; P, r)$$

$$V^{\pi}(s) = Q^{\pi}(s, \pi(s))$$

$$\rightarrow \pi^{t+1}(s) = \arg\max_{a \in \mathcal{A}} \underline{Q^{\pi_t}(s,a)} \quad \forall s$$

Goal: compare π^* vs. $\hat{\pi}$

strategy: i) compare $\hat{P}$ vs. P

ii) Translate $\hat{P}$ vs. P to $\hat{V}$ vs. V

iii) translate $\hat{V}$ vs. V to $\hat{\pi}$ vs. π_*

1) Model Estimation

$\underline{\text{Lemma}}$: With probability $1-\delta$, $\forall s, a$

$$\underbrace{\sum_{s' \in S} |\hat{P}(s'|s,a) - P(s'|s,a)|}_{\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1} \leq \sqrt{\frac{S^2 A \log(2\overbrace{SA}^{SA \cdot S}/\delta)}{\underline{N}}}$$

Proof is out of scope

ii) Value Functions.

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \begin{matrix} s_0 = s \\ P \\ \pi \end{matrix}\right]$$

$$\hat{V}^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \begin{matrix} s_0 = s \\ \hat{P} \\ \pi \end{matrix}\right]$$

Recall: Discounted State-Action Distribution

$$d^{\pi}_{s_0}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \underbrace{P^{\pi}_t(s,a;s_0)}_{\text{prob. of } (s,a) \text{ @ } t \text{ given } s_0 \,\&\, \pi, P}$$

Simulation Lemma: $0 \le r(s,a) \le 1$

$$\left|\hat{V}^{\pi}(s_0) - V^{\pi}(s_0)\right| \le \frac{\gamma}{(1-\gamma)^2} \mathop{\mathbb{E}}_{s,a\sim d^{\pi}_{s_0}}\left[\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1\right] \quad \star$$

$\forall s_0$

under P

disagreement $\hat{P}$ & P

Proof: claim:

$$\hat{V}^{\pi}(s_0) - V^{\pi}(s_0) = \frac{\gamma}{1-\gamma}\mathop{\mathbb{E}}_{s,a\sim d^{\pi}_{s_0}}\Big[\underbrace{\mathop{\mathbb{E}}_{s'\sim \hat{P}(s,a)}(\hat{V}^{\pi}(s')) - \mathop{\mathbb{E}}_{s'\sim P(s,a)}(\hat{V}^{\pi}(s'))}_{\star}\Big]$$

$$\star = \sum_{s'\sim S}\left[\hat{P}(s'|s,a) - P(s'|s,a)\right]\hat{V}^{\pi}(s')$$

using $r \le 1$ $\quad \hat{V}^{\pi}(s') \le \frac{1}{1-\gamma}$

$$\le \underbrace{\frac{1}{1-\gamma}\sum_{s'\sim S}|\hat{P}(s'|s,a) - P(s'|s,a)|}$$

III) Policy Iteration $\quad \hat{\pi} = PI(\hat{P}, r)$ $\quad \hat{\pi}$ is optimal for $\hat{P} \rightarrow$ no approx error on PI

$$V^{*}(s_0) - V^{\hat{\pi}}(s_0) \le \overset{V^{\pi^*}}{V^{*}(s_0)} - \hat{V}^{\pi^*}(s_0) + \underbrace{\overset{\ge 0}{\hat{V}^{\pi^*}(s_0) + \hat{V}^{\hat{\pi}}(s_0)}}_{\hat{\pi} \text{ is optimal on } \hat{P},\; \hat{V}^{\hat{\pi}}(s) \ge \hat{V}^{\pi}(s)\ \forall \pi} - V^{\hat{\pi}}(s_0)$$

$$\le \frac{1}{(1-\gamma)^2}\left(\mathop{\mathbb{E}}_{s,a\sim d^{\pi^*}_{s_0}}\left[\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1\right] + \mathop{\mathbb{E}}_{s,a\sim d^{\hat{\pi}}_{s_0}}\left[\|\hat{P}(\cdot|s,a) - P(s,a)\|_1\right]\right.$$

$$\rightarrow \leq \underbrace{\frac{2}{(1-\gamma)^2}\sqrt{\frac{S^2 A \log(2SA/\delta)}{N}}} = \varepsilon$$

<u>Theorem:</u> $N = \dfrac{4S^2 A \log\left(\frac{2SA}{\delta}\right)}{\varepsilon^2}$, then $V^*(s_0) - V^{\hat{\pi}}(s_0) \leq \varepsilon$ w.p $\geq 1-\delta$

3) LQR $\min \mathbb{E}_{\omega \sim \mathcal{N}(0,\sigma^2 I)}\left[\sum s_t^T Q s_t + a_t^T R a_t \mid s_{t+1} = A s_t + B a_t + w_t\right]$

MBRL:

1) i.i.d. $s_i \sim \mathcal{N}(0, \sigma^2 I_{n_s})$ $a_i \sim \mathcal{N}(0, \sigma^2 I_{n_a})$

2) estimate

$$(\hat{A}, \hat{B}) = \operatorname{argmin} \sum_{i=1}^N \| s_i' - A s_i - B a_i \|_2^2$$

3) $\hat{K} = LQR(\hat{A}, \hat{B}, Q, R)$