1) Types of Feedback in RL

1) "control feedback" - inner loop "reaction"
   ex- thermostat regulating temperature
   Unit 1

2) "data feedback" - "outer" loop "adaptation"
   ex- Smart thermostat learn preferences
   Unit 2

From now on, transition/dynamics are unknown
   \( P(\cdot, \cdot) \) \( f(\cdot, \cdot, \cdot) \)
   (often \( R(\cdot, \cdot) \) also unknown)

2) Supervised Learning (SL)

\[ \mathcal{N}(x_i, y_i) \]

\( \hat{f}: \) \text{model}
\[ \hat{y} = \hat{f}(x) \] \text{prediction}
\text{new features}

\( x \) - image (pixels)
\( y \) - cat vs. dog

\text{classification of images}
\text{regression}
\text{financial history}
\text{probability of loan repayment}

\( \mathcal{N} \) - dataset
Analogy between SL and special case RL

\[
\min_{f} \mathbb{E}\left[u(y, \hat{y}) \mid \hat{y} = f(x)\right]
\]

- **features/states**
- **loss/cost**
- **predictions/actions**
- **distribution**
- **transition probabilities**
- **model/policy dependence**
- **offline/online**
- **no impact**

What should we learn to solve (i.e. find near-optimal policy)?

\[
\mathcal{M} = \{S, \mathcal{A}, \mathcal{U}\} \times \mathcal{X}^3
\]

- transition \( P \), reward \( r \)
- value function \( V^P \), Q function \( Q^P \)
- optimal value/Q function \( V^*, Q^* \)
- optimal policy \( P^* \)

3) Estimation & Prediction

A) Tabular Setting:

Suppose \( x_i \overset{i.i.d.}{\sim} \mathcal{D}, x_i \in \mathcal{X}, p(x) = P_D(x_i = x) \)

\[
\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} 1_{\exists i: x_i = x}
\]

How good is \( \hat{p} \)?

**Lemma:** (consistency)

\[
\mathbb{E}\left[\hat{p}(x)\right] = p(x)
\]

Supervision

- \( s_{t+1} \sim P(s_t, a_t) \)
- \( r_t = r(s_t, a_t) \)
- \( \sum_t r_t \) after delay \( \sum_0^t \sum r_t \)
- \( \times \) exception: imitation learning
Proof. \[ E[\hat{p}(x)] = \frac{1}{N} \sum_{i=1}^{N} E[\prod \delta_{x_i=x}] = \frac{1}{N} \sum_{i=1}^{N} P(x_i=x) = p(x) \]

**Theorem (concentration)**

For all \( x \in \mathcal{X} \) with probability \( 1-\delta \)

\[
\left| \hat{p}(x) - p(x) \right| \leq \sqrt{\frac{2 \log(2 \mathcal{X} \mathcal{Y} \delta)}{N}} \sim O\left(\frac{1}{\sqrt{N}}\right)
\]

Proof out of scope (Hoeffding's inequality)

\[
x, y \sim D \quad \hat{f}(x, y) \sim \sum_{i=1}^{N} (x_i, y_i)
\]

By similar approach,

\[
\hat{f}(x) = \frac{\sum_{i=1}^{N} y_i \prod \delta_{x=x_i}^x}{\sum_{i=1}^{N} \prod \delta_{x=x_i}^x}
\]

Details out of scope, but if \( y = f^*(x) + w \)

\( \forall x \in \mathcal{X} \), w.p. \( 1-\delta \)

\[
\left| \hat{f}(x) - f^*(x) \right| \leq \sqrt{\frac{1N \log(1/\delta)}{N}}
\]

applies when \( |\mathcal{X}| < N \)

B) **Non-tabular:**

**Empirical Risk Minimization**

\[
f = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, f(x_i))
\]

\( l \) function class loss label prediction
1) Estimation (Parametric) \( \mathcal{F} = \{ \mathcal{f}_\Theta(x) \mid \Theta \in \mathbb{R}^d \} \) \( \Theta \in \mathbb{R}^d \), e.g., neural network \( \Theta \)-weights \( f_\Theta(x) = \hat{\Theta}^T \mathbf{\phi}(x) \) fixed

Suppose \( y = f_{\Theta^*}(x) + \mathbf{w} = \text{iid noise} \)

Estimation Error: \( \| \Theta^* - \hat{\Theta} \| \)

Details out of scope, estimation error bounded

Parametric \( \| \Theta^* - \hat{\Theta} \| \leq \sqrt{\frac{d \log(C/\delta)}{N}} \)

Useful when \( N > d \)

Prediction

Expected prediction error on \( (x, y) \sim \mathcal{D} \)

\[ \mathbb{E}_{x, y \sim \mathcal{D}} \left[ e(\hat{f}(x), y) \right] \]

Assume \( \mathcal{D}: x \sim \mathcal{D}_x \) \( \text{noise} \) and \( f^* \in \mathcal{F} \)

Details out of scope, but

\[ \mathbb{E}_{x, y \sim \mathcal{D}} \left[ e(\hat{f}(x), y) \right] \leq \sqrt{\frac{\log(C/\delta)}{N}} \]

Prediction error is about average case and fixed distribution