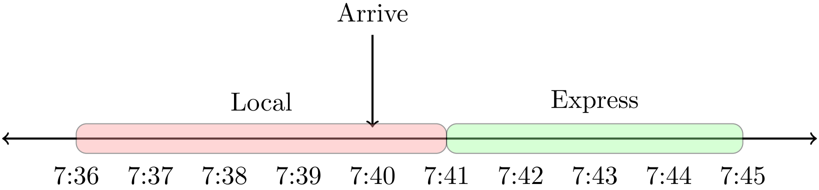
Distributional Reinforcement Learning

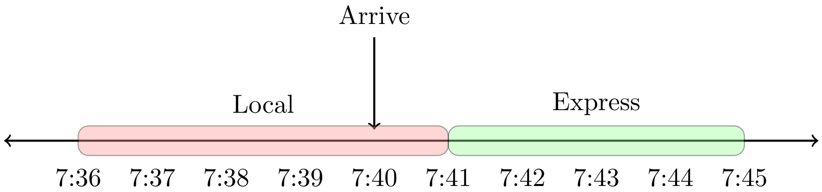
John Martin Jr.

March 16, 2018

My Commute

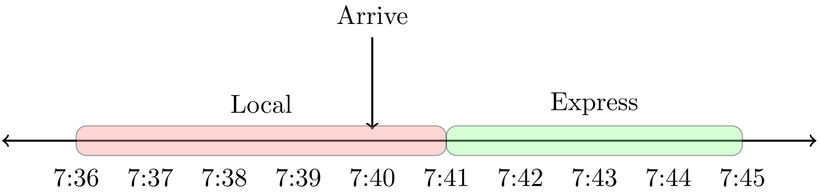


My Commute

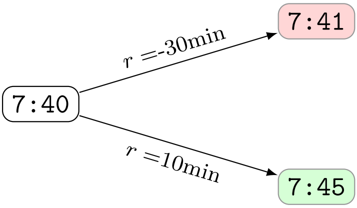


- Local is 5 min late 25% of the time

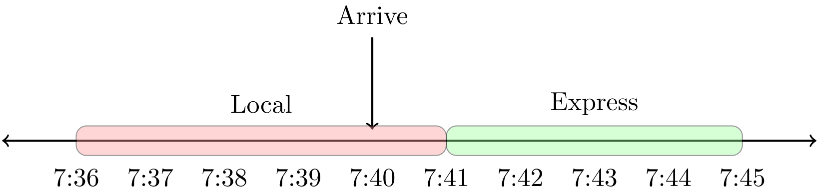
My Commute



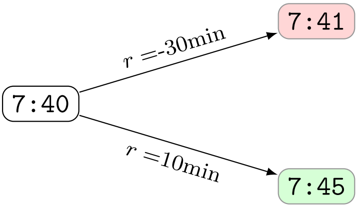
- Local is 5 min late 25% of the time



My Commute

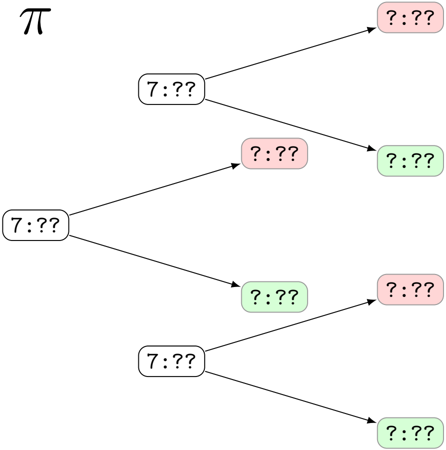


- Local is 5 min late 25% of the time



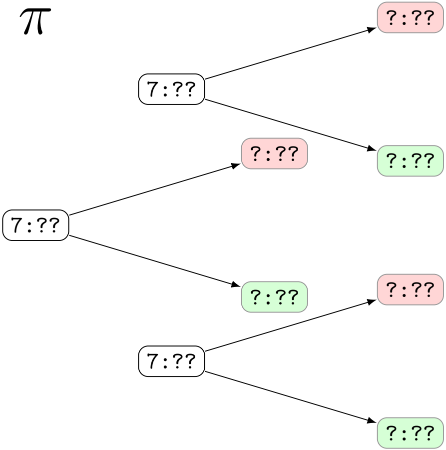
$$\mathbf{E}[R] = \frac{1}{4} \times -30 + \frac{3}{4} \times 10 = 0$$

My Commute



My Commute

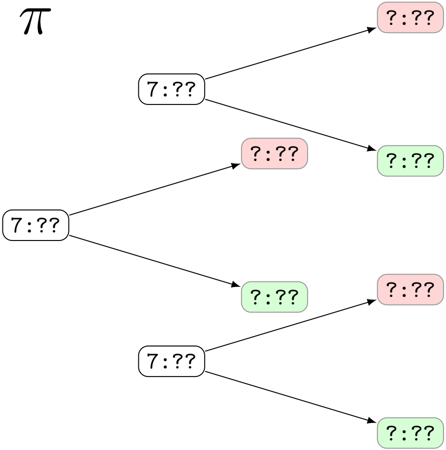
π



Reward sequence

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots$$

My Commute



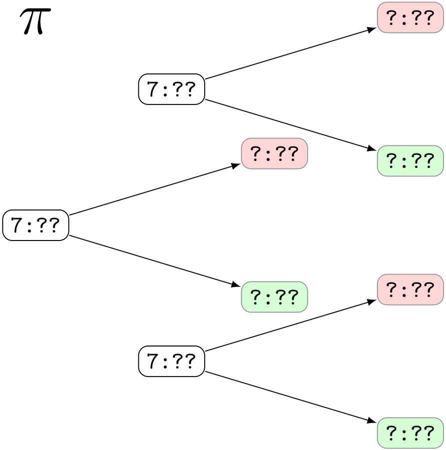
Reward sequence

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots$$

The Bellman Equation

$$V^\pi(s) = \mathbf{E}[R(s)] + \gamma \mathop{\mathbf{E}}_{s' \sim P^\pi}[V^\pi(s')]$$

My Commute



Reward sequence

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots$$

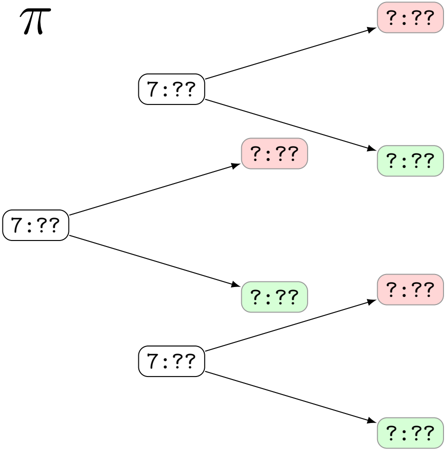
The Bellman Equation

$$V^\pi(s) = \mathbf{E}[R(s)] + \gamma \mathop{\mathbf{E}}_{s' \sim P^\pi}[V^\pi(s')]$$

Mean process

$$\texttt{7:40} \xrightarrow{\mathbf{E}[R] = 0\text{min}} \texttt{7:43}$$

My Commute



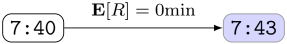
Reward sequence

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots$$

The Bellman Equation

$$V^\pi(s) = \mathbf{E}[R(s)] + \gamma \mathop{\mathbf{E}}_{s' \sim P^\pi}[V^\pi(s')]$$

Mean process



This is never realized!

Distributional Reinforcement Learning

- Central quantity is the random return Z_π

$$V^\pi(s) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s\right] = \mathbf{E}\left[Z_\pi(S_0) \middle| S_0 = s\right]$$

Distributional Reinforcement Learning

- Central quantity is the random return Z_π

$$V^\pi(s) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s\right] = \mathbf{E}\left[Z_\pi(S_0) \middle| S_0 = s\right]$$

- Peel back the expectations

$$Z_\pi(s) \stackrel{D}{=} R(s, a) + \gamma Z_\pi(S')$$

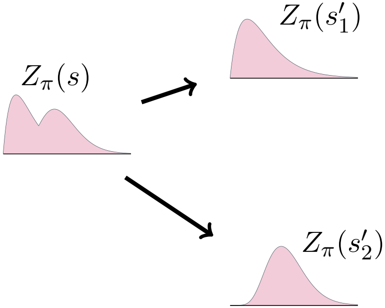
Distributional Reinforcement Learning

- Central quantity is the random return Z_π

$$V^\pi(s) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s\right] = \mathbf{E}\left[Z_\pi(S_0) \middle| S_0 = s\right]$$

- Peel back the expectations

$$Z_\pi(s) \overset{D}{=} R(s,a) + \gamma Z_\pi(S')$$



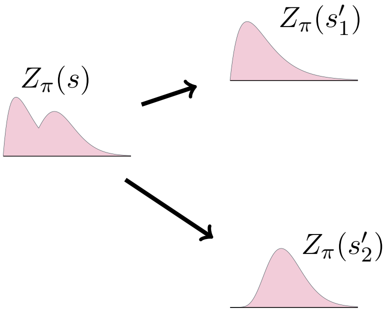
Distributional Reinforcement Learning

- Central quantity is the random return Z_π

$$V^\pi(s) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s\right] = \mathbf{E}\left[Z_\pi(S_0) \middle| S_0 = s\right]$$

- Peel back the expectations

$$Z_\pi(s) \stackrel{D}{=} R(s, a) + \gamma Z_\pi(S')$$



Intrinsic randomness

- Rewards
- Transitions
- Next state return

Distributional Reinforcement Learning

- The distributional Bellman operator

$$\mathcal{T}^{\pi} Z_{\pi}(s,a) = R(s,a) + \gamma Z_{\pi}(S', A'),\ \ S' \sim P^{\pi}(\cdot|s,a), A' \sim \pi(\cdot|S')$$

- Distributional Bellman equation: $\mathcal{T}^{\pi} Z_{\pi} = Z_{\pi}$
- Contraction: $\bar{d}_p(\mathcal{T}^{\pi} Z_{\pi}, \mathcal{T}^{\pi} Z'_{\pi}) \leq \gamma \bar{d}_p(Z_{\pi}, Z'_{\pi})$

Distributional Reinforcement Learning

- The distributional Bellman operator

$$\mathcal{T}^\pi Z_\pi(s,a) = R(s,a) + \gamma Z_\pi(S', A'),\ S' \sim P^\pi(\cdot|s,a), A' \sim \pi(\cdot|S')$$

- Distributional Bellman equation: $\mathcal{T}^\pi Z_\pi = Z_\pi$
- Contraction: $\bar{d}_p(\mathcal{T}^\pi Z_\pi, \mathcal{T}^\pi Z'_\pi) \leq \gamma \bar{d}_p(Z_\pi, Z'_\pi)$

Current results

- $\mathcal{T}^\pi$ is a contraction in Wass_p (Bellemare et al. 2017)
- $\Pi_{\text{KL}}\mathcal{T}^\pi$ is not a contraction in Wass_p (Rowland et al. 2017)
- $\Pi_{\text{KL}}\mathcal{T}^\pi$ is a contraction in Cram_2 (Rowland et al. 2017)
- Existence and optimality (Rowland et al. 2017)

Distributional Reinforcement Learning

- The distributional Bellman operator

$$\mathcal{T}^\pi Z_\pi(s,a) = R(s,a) + \gamma Z_\pi(S',A'),\ S' \sim P^\pi(\cdot|s,a), A' \sim \pi(\cdot|S')$$

- Distributional Bellman equation: $\mathcal{T}^\pi Z_\pi = Z_\pi$
- Contraction: $\bar{d}_p(\mathcal{T}^\pi Z_\pi, \mathcal{T}^\pi Z'_\pi) \leq \gamma \bar{d}_p(Z_\pi, Z'_\pi)$

Current results

- $\mathcal{T}^\pi$ is a contraction in Wass_p (Bellemare et al. 2017)
- $\Pi_{\mathrm{KL}}\mathcal{T}^\pi$ is not a contraction in Wass_p (Rowland et al. 2017)
- $\Pi_{\mathrm{KL}}\mathcal{T}^\pi$ is a contraction in Cram_2 (Rowland et al. 2017)
- Existence and optimality (Rowland et al. 2017)

Observations

- A more powerful theory to study RL?
- Wasserstein is useful in RL!

References

- Rowland et al. (2018), An Analysis of Categorical Distributional Reinforcement Learning
- Bellemare et al. (2017), A Distributional Perspective on Reinforcement Learning
- Dabney et al. (2017), Distributional Reinforcement Learning with Quantile Regression
- Bellemare (2017), Symposium IA Montreal
- Tamar et al. (2016), Learning the Variance of the Reward-To-Go
- Veness et al. (2014), Compress and Control
- Morimura et al. (2010), Nonparametric Return Distribution Approximation for Reinforcement Learning
- Engel et al. (2005), Reinforcement learning with Gaussian processes

Questions