Linear Value Function Learning

- Hypothesis class
  - Features $\phi_k : \mathcal{S} \times \mathcal{A} \mapsto \Re \qquad k = 1, \ldots, K$
  - Function class

$$Q_t^* \in \mathcal{Q}_t = \left\{ \sum_{k=1}^{K} \theta_k \phi_k : \theta \in \Re^K \right\}$$

  - Number of parameters = KH

- Parameterization

$$\tilde{Q}^\theta(s, a) = \sum_{k=1}^{K} \theta_k \phi_k(s, a)$$

- Before each episode
  - Estimate value function from data

$$\left\{ (s_t^\ell, a_t^\ell, r_t^\ell, s_{t+1}^\ell) : \begin{array}{l} t = 0, \ldots, H-1 \\ \ell = 1, \ldots, L \end{array} \right\}$$

  - Select actions using value function

Least-Squares Value Iteration

- Data available after episode L

$$\left\{ (s_t^\ell, a_t^\ell, r_t^\ell, s_{t+1}^\ell) : \begin{array}{l} t = 0, \ldots, H-1 \\ \ell = 1, \ldots, L \end{array} \right\}$$

- Estimate coefficients

$$\hat{\theta}_0 \leftarrow \hat{\theta}_1 \leftarrow \cdots \leftarrow \hat{\theta}_{H-2} \leftarrow \hat{\theta}_{H-1}$$

$$\min_{\hat{\theta}_t} \sum_{\ell=1}^{L} \left(\tilde{Q}^{\hat{\theta}_t}(s_t^\ell, a_t^\ell) - y_t^\ell \right)^2 + \lambda \|\hat{\theta}_t\|_2^2$$

$$y_{H-1}^\ell = r_{H-1}^\ell \qquad\qquad y_t^\ell = r_t^\ell + \max_{a \in \mathcal{A}} \tilde{Q}^{\hat{\theta}_{t+1}}(s_{t+1}^\ell, a)$$

- Linear algebra version

$$\hat{\theta}_t = (A^\top A + \lambda I)^{-1} A^\top b$$

$$A = \begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_K \\ | & & | \end{bmatrix} \qquad\qquad b = \begin{bmatrix} y_t^1 \\ \vdots \\ y_t^L \end{bmatrix}$$

UC-LSVI

- Confidence set propagation

$$\mathcal{Q}_t^L = \left\{ Q^\theta : (\theta - \hat{\theta}_t)^\top \hat{\Sigma}_t^{-1} (\theta - \hat{\theta}) \leq \gamma \right\}$$

- LSVI, but with different target samples

$$y_t^\ell = r_t^\ell + \max_{\tilde{Q}^\theta \in \mathcal{Q}_{t+1}^L} \max_{a \in \mathcal{A}} \tilde{Q}^\theta (s_{t+1}^\ell, a)$$

- Point estimate $\hat{\theta}_t = (A^\top A + \lambda I)^{-1} A^\top b$

- Error covariance $\hat{\Sigma}_t = \left(\frac{1}{\sigma^2} A^\top A + \lambda I \right)^{-1}$

- Action $a_t^{L+1} \in \arg\max_{a \in \mathcal{A}} \max_{\tilde{Q}^\theta \in \mathcal{Q}_t^L} \tilde{Q}^\theta (s_{t+1}^L, a)$

- Bayesian interpretation of regression
  - Prior distribution $\theta_t \sim N(0, \lambda I)$
  - Pretend that data samples are Gaussian

$$y_t^\ell = Q_t^*(s_t^\ell, a_t^\ell) + w_t^\ell \qquad w_t^\ell \sim N(0, \sigma^2)$$

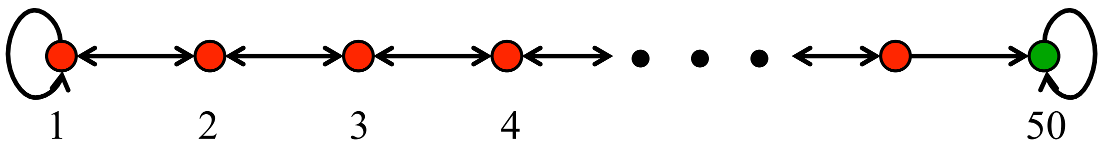
Thompson Sampling Approach

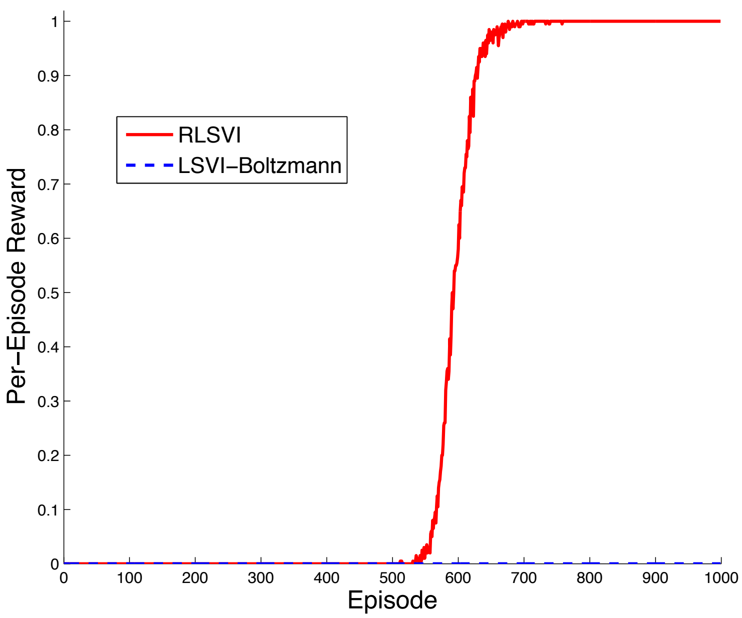
- Carry spirit over to value function learning?
- Main idea
  - Begin with pseudo-prior over value functions
  - Before each episode, sample $\tilde{\theta}_t \sim \mathbb{P}(\theta_t \mid \mathrm{data})$
  - Act $a_t^{L+1} \in \max_{a \in \mathcal{A}} \tilde{Q}_t^{\tilde{\theta}_t}(x_t^{L+1}, a)$

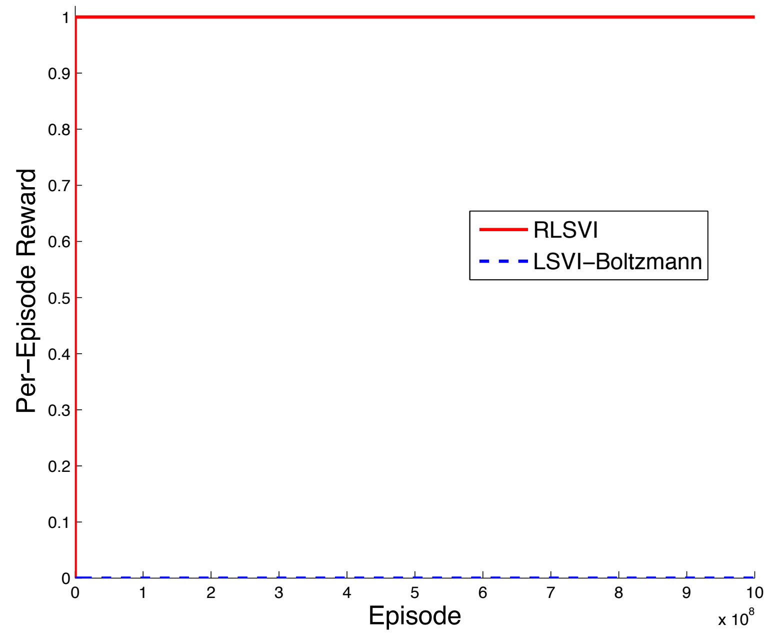
One Sampling Method: RLSVI

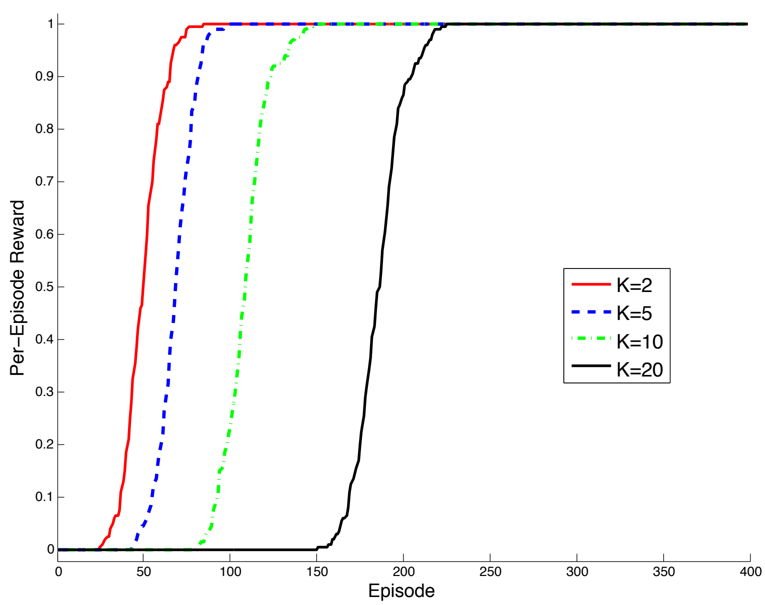
- LSVI, but with different target samples
    - Point estimate $\hat{\theta}_t = (A^\top A + \lambda I)^{-1} A^\top b$
    - Covariance $\hat{\Sigma}_t = \left(\frac{1}{\sigma^2} A^\top A + \lambda I \right)^{-1}$
    - Randomized estimate $\tilde{\theta}_t \sim N(\hat{\theta}_t, \hat{\Sigma}_t)$
    - Target samples $y_t^\ell = r_t^\ell + \max_{a \in \mathcal{A}} \tilde{Q}^{\tilde{\theta}_t}(s_{t+1}^\ell, a)$
- Computationally efficient
- Conjecture: statistically much more efficient that UC-LSVI

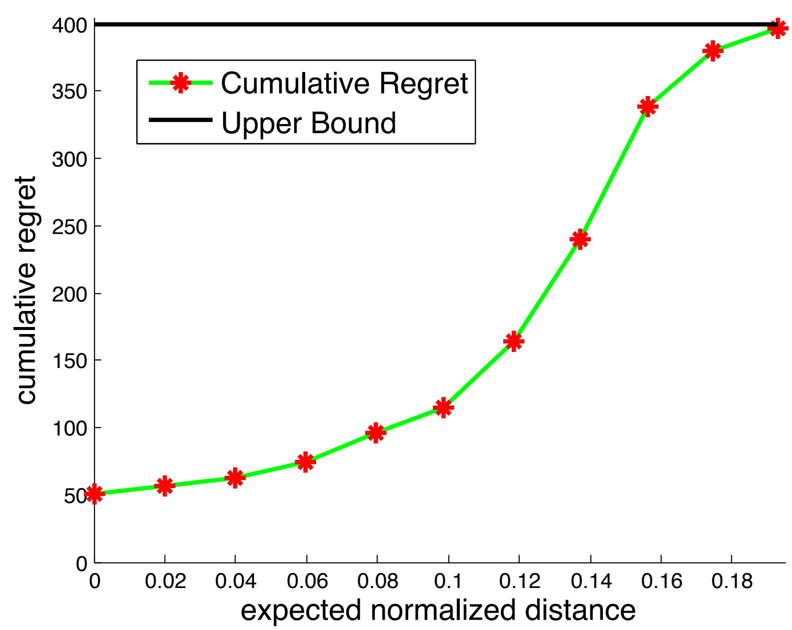
Application to Chain Problem











Sampling via Bootstrap

- Sample L past episodes with replacement
- Apply LSVI to the sampled data set
- This produces a randomized model
- Applicable to nonlinear parameterizations
- Does this serve our exploration needs?

Closing Remarks

generalization

exploration

\+ delayed consequences

---

intelligent learning system

- Methods are emerging to effectively combine these three elements
  - The area is still in flux — great research opportunities!
  - Enormous potential value to practice