Optimistic Constraint Propagation

- Context
  - Deterministic episodic MDP
  - Coherent reinforcement learning
  - Rewards in [0,1]

- Bellman's equation

$$Q_t^*(x_t, a_t) = R_t(x_t, a_t) + \sup_{a \in \mathcal{A}} Q_{t+1}^*(x_{t+1}, a)$$

- Algorithm
  - Begin with hypothesis class $\mathcal{Q} = \mathcal{Q}_0 \times \cdots \times \mathcal{Q}_{H-1}$
  - Act according to optimistic values
  - After each episode, further constrain $\mathcal{Q}$

- Regret bound $\mathrm{Regret}(L) \leq H \dim_E(\mathcal{Q})$
  - Depends on eluder dimension

Linear Combination of Features

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

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

- Eluder dimension $\dim_E(\mathcal{Q}) \leq KH$
- Regret bound $\text{Regret}(L) \leq KH^2$
- Observations:
  - Number of parameters = *KH*
  - Regret bound allows for *KH* "throw-away" episodes

Constraint Propagation Algorithm

- Data available after episode L

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

- Propagate constraints to produce new sets

$$\mathcal{Q}_0^L \leftarrow \mathcal{Q}_1^L \leftarrow \cdots \leftarrow \mathcal{Q}_{H-1}^L$$

  - All are polytopes
- Each data point generates bounds

linear inequalities

$$L_t^\ell \leq Q_t^*(s_t^\ell, a_t^\ell) \leq U_t^\ell$$

$$U_t^\ell \leftarrow r_t^\ell + \max_{Q_{t+1} \in \mathcal{Q}_{t+1}^L} \max_{a \in \mathcal{A}} Q_{t+1}(s_{t+1}^\ell, a)$$

$$L_t^\ell \leftarrow r_t^\ell + \min_{Q_{t+1} \in \mathcal{Q}_{t+1}^L} \max_{a \in \mathcal{A}} Q_{t+1}(s_{t+1}^\ell, a)$$

  - Upper and lower bounds computed via LP
- New hypothesis class

$$\mathcal{Q}_t^L = \left\{ Q_t \in \mathcal{Q}_t : \begin{array}{l} L_t^\ell \leq Q_t(s_t^\ell, a_t^\ell) \leq U_t^\ell \\ \forall \ell = 1, \ldots, L \end{array} \right\}$$

Beyond OCP

- Shortcomings of OCP
  - Does not accommodate agnostic learning
    - Slight misspecification can make regret explode
  - Does not accommodate stochastic MDPs

- Will develop TS-based approach
  - Accommodates stochastic MDPs
  - Seems to work in agnostic setting