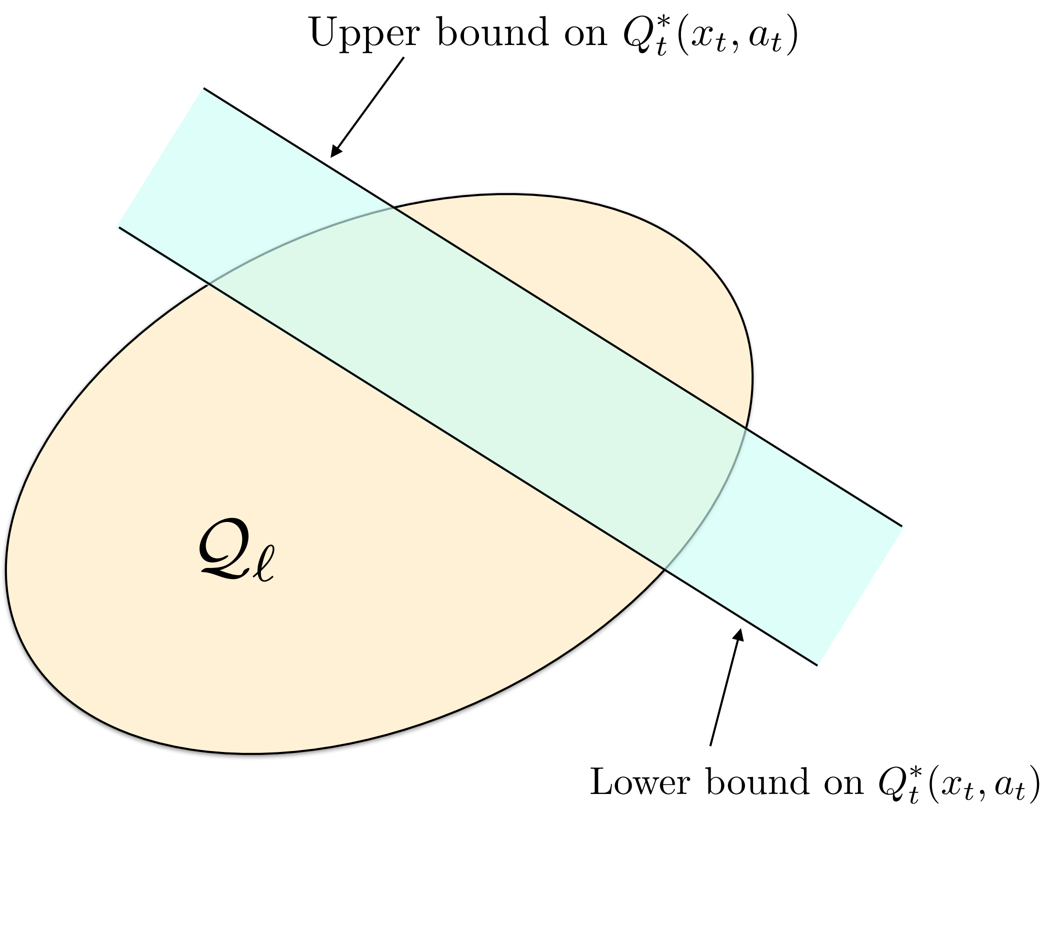
Optimistic Constraint Propagation

- Value function learning
  - Based on UCB
  - For deterministic MDPs

- Statistically plausible set = possible set
- Data from each state transition supplies two constraints



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 set of possible value functions $\mathcal{Q}_0$
  - Act according to optimistic values
  - After each episode, further constrain $\mathcal{Q}_\ell$

$$Q_t^*(x_t, a_t) \leq \sup_{Q' \in \mathcal{Q}_\ell} \left(R_t(x_t, a_t) + \sup_{a \in \mathcal{A}} Q'_{t+1}(x_{t+1}, a) \right)$$

$$Q_t^*(x_t, a_t) \geq \inf_{Q' \in \mathcal{Q}_\ell} \left(R_t(x_t, a_t) + \sup_{a \in \mathcal{A}} Q'_{t+1}(x_{t+1}, a) \right)$$

- Regret bound $\mathrm{Regret}(L) \leq H \dim_E(\mathcal{Q}_0)$
  - Depends on eluder dimension

Linear Combination of Features

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

- Set of possible functions

$$\mathcal{Q}_0 = \left\{ \sum_{k=1}^{K} \theta_k \phi_k : \theta \in \Re^K \right\}^H$$

- Eluder dimension

$$\dim_E(\mathcal{Q}_0) \leq KH$$

- Regret bound

$$\mathrm{Regret}(L) \leq KH^2$$

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