A Recommendation System Model

- Consider recommending movies
  - *N* movies
  - Sequence of *H* recommendations for each customer
  - Customer accepts/rejects each
  - Goal: high acceptance rate

- MDP formulation
  - state: $r_t \in \{0, 1\}$
  - action: $s_t \in \{-1, 0, 1\}^N$
  - reward: $a_t \in \{1, \ldots, N\}$

- Parameterization

$$\mathbb{E}[r_t = 1 | s_t, a_t] = \begin{cases} \frac{\exp(\theta_{a_t}^\top s_t)}{1+\exp(\theta_{a_t}^\top s_t)} & \text{if } s_{a_t,t} = 0 \\ 0 & \text{otherwise} \end{cases}$$

Thompson Sampling

- Independent priors over parameters
  - Possibly finite support

- Algorithm
  - Sample parameters from posterior via Gibbs sampling
  - Apply optimal policy for one episode
  - Repeat

- Gibbs sampling
  - Sample parameters from priors
  - Iterate over components
    - Fix all other components θ_a
    - Sample component from one-dimensional distribution

$$\prod_{n=1}^{N} p_n(\theta_{an}) \prod_{k:a^k=a} \frac{\exp(\theta_a^\top s^k)}{1+\exp(\theta_a^\top s^k)}$$

- Regret bound?

- Intractable MDP

Value Function Learning

- Online optimization
  - Optimize optimistic estimates of immediate value
- Reinforcement learning
  - Optimize optimistic estimates of net present value
- UCB approach
  - Maintain set of plausible value functions
  - Select actions that optimize optimistic state-action values
- Thompson sampling
  - Maintain distribution over plausible value functions
  - Before each episode, sample a value function
  - Select actions that optimize sampled values

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 $\mathcal{Q}$
  - Act according to optimistic values
  - After each episode, further constrain $\mathcal{Q}$

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

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

- Regret bound $\mathrm{Regret}(T) \leq 2H \dim_E(\mathcal{Q})$
  - Example: linear combination of features