Factored MDPs

- Vector state/action

$$\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_N \qquad \mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$$

- Each component has *scope* $Z_n \subseteq \{1, \ldots, N\}$

- Transitions for *n*th component of state

$$(x_{Z_n,t}, a_{Z_n,t}) \Rightarrow x_{n,t+1}$$

- *n*th reward term

$$(x_{Z_n,t}, a_{Z_n,t}) \Rightarrow r_{n,t}$$

- *HN* tables, each with $\prod_{m \in Z_n} |\mathcal{S}_m \times \mathcal{A}_m|$ entries
  - Versus H tables, each with $|\mathcal{S} \times \mathcal{A}| = \prod_{m=1}^{N} |\mathcal{S}_m \times \mathcal{A}_m|$
  - Also…dimensionality of table entries

Thompson Sampling

- Priors for each table entry
  - Dirichlet over probability vectors
  - Normal-gamma over reward distributions

- Algorithm
  - Sample factored MDP from distribution
  - Apply optimal policy for one episode
  - Update distribution
  - Repeat

- Regret bounds
  - Depend on number of parameters rather than numbers of states and actions

- Intractable MDP

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