Need for Generalization

- Curse of dimensionality
  - State spaces grow exponentially
  - Example: queueing system

- *Tabula rasa* learning learns exponentially many parameters

- *Tabula rasa* regret bounds

$$\tilde{O}(H\mathcal{S}\sqrt{\mathcal{A}HL})$$

- How many episodes before we do well?

$$\tilde{O}(H^3\mathcal{S}^2\mathcal{A})$$

Approaches to Generalization

- Model learning
  - Learn MDP (P, R)
  - Parameterized model (P^θ, R^θ)

- Value function learning
  - Learn value function Q^*
  - Parameterized value function Q^θ

- Policy learning
  - Learn policy μ^*
  - Parameterized policy μ^θ

- Coherent versus agnostic learning
  - Parametric versus nonparametric representations

Factored MDPs

- State-action pair is a vector

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

- Each component has *scope*

$$Z_n \subseteq \{1, \ldots, N\}$$

- Scope constrains model

$$\mathbb{P}(s_{t+1} = s | x_t) = \prod_{n=1}^{N} \mathbb{P}(s_{n,t+1} = s_n | x_{Z_n,t})$$

$$\mathbb{E}[r_t | x_t] = \sum_{n=1}^{N} \mathbb{E}[r_{n,t} | x_{Z_n,t}]$$

- How many parameters to learn?
  - Exponential in N ?
  - Exponential in $|Z_n|$?

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

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

$$s_{a_t,t+1} \leftarrow r_t$$