# Chapter 5: Statistical Limits of Generalization

Hui Shen, Yangjianchen Xu

The University of North Carolina at Chapel Hill

April 28, 2023

# Table of Contents

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning

2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# Introduction

Connection with previous chapters:

- Chapter 3&4: sufficient conditions under which sample complexity results do not explicitly depend on the size of the state (or action) space.
- Chapter 5: necessary conditions.

# Introduction

Two most basic settings in supervised learning:

- agnostic learning (i.e. finding the best classifier or hypothesis in some class)
- learning with linear models (i.e. learning the best linear regressor or the best linear classifier).

# Introduction

Two most basic settings in supervised learning:

- agnostic learning (i.e. finding the best classifier or hypothesis in some class)
- learning with linear models (i.e. learning the best linear regressor or the best linear classifier).

This chapter: focus on lower bounds under these two settings for reinforcement learning.

- finite horizon MDPs
- the episodic setting and the generative model setting

# Outline

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning

2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# Introduction

Settings in Agnostic Learning:

- A hypothesis class $\mathcal{H}$ (either finite or infinite)
- Each $f \in \mathcal{H}$ has an associated policy $\pi_f : \mathcal{S} \to \mathcal{A}$
- Each policy is deterministic.

# Introduction

Settings in Agnostic Learning:

- A hypothesis class $\mathcal{H}$ (either finite or infinite)
- Each $f \in \mathcal{H}$ has an associated policy $\pi_f : \mathcal{S} \to \mathcal{A}$
- Each policy is deterministic.
- Examples of $\mathcal{H}$:
    - $\mathcal{H}$ itself is a class of policies.

# Introduction

Settings in Agnostic Learning:

- A hypothesis class $\mathcal{H}$ (either finite or infinite)
- Each $f \in \mathcal{H}$ has an associated policy $\pi_f : \mathcal{S} \rightarrow \mathcal{A}$
- Each policy is deterministic.
- Examples of $\mathcal{H}$:
  - $\mathcal{H}$ itself is a class of policies.
  - $\mathcal{H}$ is a set of state-action values; a greedy policy $\pi_f(s, h) = \text{argmax}_a f_h(s, a)$ for $f$.

# Introduction

Settings in Agnostic Learning:

- A hypothesis class $\mathcal{H}$ (either finite or infinite)
- Each $f \in \mathcal{H}$ has an associated policy $\pi_f : \mathcal{S} \rightarrow \mathcal{A}$
- Each policy is deterministic.
- Examples of $\mathcal{H}$:
  - $\mathcal{H}$ itself is a class of policies.
  - $\mathcal{H}$ is a set of state-action values; a greedy policy $\pi_f(s, h) = \text{argmax}_a f_h(s, a)$ for $f$.
  - $\mathcal{H}$ could be a class of models (i.e. each $f \in \mathcal{H}$ is an MDP itself); the optimal policy for $f$.

# Introduction

Settings in Agnostic Learning:

- A hypothesis class $\mathcal{H}$ (either finite or infinite)
- Each $f \in \mathcal{H}$ has an associated policy $\pi_f : \mathcal{S} \to \mathcal{A}$
- Each policy is deterministic.
- Examples of $\mathcal{H}$:
  - $\mathcal{H}$ itself is a class of policies.
  - $\mathcal{H}$ is a set of state-action values; a greedy policy $\pi_f(s, h) = \text{argmax}_a f_h(s, a)$ for $f$.
  - $\mathcal{H}$ could be a class of models (i.e. each $f \in \mathcal{H}$ is an MDP itself); the optimal policy for $f$.
- $\Pi = \{\pi_f \mid f \in \mathcal{H}\}$.

# Introduction

Settings in Agnostic Learning:

- A hypothesis class $\mathcal{H}$ (either finite or infinite)
- Each $f \in \mathcal{H}$ has an associated policy $\pi_f : \mathcal{S} \rightarrow \mathcal{A}$
- Each policy is deterministic.
- Examples of $\mathcal{H}$:
    - $\mathcal{H}$ itself is a class of policies.
    - $\mathcal{H}$ is a set of state-action values; a greedy policy $\pi_f(s, h) = \text{argmax}_a f_h(s, a)$ for $f$.
    - $\mathcal{H}$ could be a class of models (i.e. each $f \in \mathcal{H}$ is an MDP itself); the optimal policy for $f$.
- $\Pi = \{\pi_f \mid f \in \mathcal{H}\}$.

The goal of agnostic learning:

$$\max_{\pi \in \Pi} \mathbb{E}_{s_0 \sim \mu} V^{\pi}(s_0)$$

# Outline

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning

2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# Binary Classification

Binary classification as a RL problem:

- MDP with $H = 1$
- $|\mathcal{A}| = 2$
- $r(s, a) = \mathbf{1}(\text{label}(s) = a)$

# Binary Classification

Binary classification as a RL problem:

- MDP with $H = 1$
- $|\mathcal{A}| = 2$
- $r(s, a) = \mathbf{1}(\text{label}(s) = a)$

Setting in binary classification:

- $N$ samples $(x_i, y_i)_{i=1}^{N}$
- a set $\mathcal{H}$ of binary classifiers: $h : \mathcal{X} \to \{0, 1\}$ for $h \in \mathcal{H}$
- $(x_i, y_i) \overset{i.i.d}{\sim} D$

# Binary Classification

Define the empirical error and the true error as:

$$\widehat{\mathrm{err}}(h) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(h\left(x_i\right) \neq y_i\right), \quad \mathrm{err}(h) = \mathbb{E}_{(X,Y)\sim D}\mathbf{1}(h(X) \neq Y)$$

# Binary Classification

Define the empirical error and the true error as:

$$\widehat{\text{err}}(h) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(h\left(x_i\right) \neq y_i\right), \quad \text{err}(h) = \mathbb{E}_{(X,Y)\sim D}\mathbf{1}(h(X) \neq Y)$$

For a given $h \in \mathcal{H}$, Hoeffding's inequality implies that with probability at least $1-\delta$ :

$$|\,\text{err}(h) - \widehat{\text{err}}(h)| \leqslant \sqrt{\frac{2}{N}\log\frac{2}{\delta}}$$

# Binary Classification

Hoeffding inequality + union bound give the following result.

## Proposition 5.1. (The "Occam's razor" bound)

*Suppose $\mathcal{H}$ is finite. Let $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{err}(h)$. With probability at least $1 - \delta$ :*

$$\widehat{\text{err}}(\widehat{h}) \leqslant \min_{h \in \mathcal{H}} \text{err}(h) + \sqrt{\frac{2}{N} \log \frac{2|\mathcal{H}|}{\delta}}.$$

# Binary Classification

Hoeffding inequality + union bound give the following result.

### Proposition 5.1. (The “Occam’s razor” bound)

*Suppose $\mathcal{H}$ is finite. Let* $\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{err}(h)$. *With probability at least* $1-\delta$ *:*

$$\widehat{\mathrm{err}}(\widehat{h}) \leqslant \min_{h\in\mathcal{H}} \mathrm{err}(h) + \sqrt{\frac{2}{N}\log\frac{2|\mathcal{H}|}{\delta}}.$$

Hence, provided that

$$N \geqslant \frac{2\log\frac{2|\mathcal{H}|}{\delta}}{\epsilon^2},$$

then with probability at least $1-\delta$, we have that:

$$\widehat{\mathrm{err}}(\widehat{h}) \leqslant \min_{h\in\mathcal{H}} \mathrm{err}(h) + \epsilon.$$

**Key observation**: the regret has no dependence on the size of $\mathcal{S}$.

# Outline

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning

2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# An Occam's Razor Bound for RL

Consider the episodic setting where we collect $N$ trajectories using a uniform policy $\mathrm{Unif}_{\mathcal{A}}$.

### Lemma 5.2. (Unbiased estimation of $V_0^\pi(\mu)$)

*Let $\pi$ be any deterministic policy. We have that:*

$$V_0^\pi(\mu) = |\mathcal{A}|^H \cdot \mathbb{E}_{\tau \sim \Pr_{\mathrm{Unif}_{\mathcal{A}}}} \left[ \mathbf{1}\left(\pi\left(s_0\right) = a_0, \ldots, \pi\left(s_{H-1}\right) = a_{H-1}\right) \sum_{h=0}^{H-1} r\left(s_h, a_h\right) \right]$$

*where $\Pr_{\mathrm{Unif}_{\mathcal{A}}}$ specifies the distribution over trajectories $\tau = (s_0, a_0, r_0, \ldots s_{H-1}, a_{H-1}, r_{H-1})$ under the policy $\mathrm{Unif}_{\mathcal{A}}$.*

# Proof of Lemma 5.2.

From a standard importance sampling argument, we have

$$
\begin{aligned}
V_0^{\pi}(\mu) &= \mathbb{E}_{\tau \sim \mathsf{Pr}_{\pi}} \left[ \sum_{h=0}^{H-1} r_h \right] \\
&= \mathbb{E}_{\tau \sim \mathsf{Pr}_{\mathrm{Unif}_{\mathcal{A}}}} \left[ \frac{\mathsf{Pr}_{\pi}(\tau)}{\mathsf{Pr}_{\mathrm{Unif}_{\mathcal{A}}}(\tau)} \sum_{h=0}^{H-1} r_h \right] \\
&= |\mathcal{A}|^H \cdot \mathbb{E}_{\tau \sim \mathsf{Pr}_{\mathrm{Unif}_{\mathcal{A}}}} \left[ \mathbf{1}\left(\pi\left(s_0\right) = a_0, \ldots, \pi\left(s_{H-1}\right) = a_{H-1}\right) \sum_{h=0}^{H-1} r_h \right]
\end{aligned}
$$

# An Occam's Razor Bound for RL

Obtain $N$ trajectories under $\mathrm{Unif}_{\mathcal{A}}$ with $n$-th sampled trajectory as

$$\left(s_0^n, a_0^n, r_1^n, s_1^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n\right).$$

# An Occam's Razor Bound for RL

Obtain $N$ trajectories under $\mathrm{Unif}_{\mathcal{A}}$ with $n$-th sampled trajectory as

$$\left(s_0^n, a_0^n, r_1^n, s_1^n, \dots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n\right).$$

Estimate the finite horizon reward of any given policy $\pi$ via

$$\widehat{V}_0^{\pi}(\mu) = \frac{|\mathcal{A}|^H}{N} \sum_{n=1}^{N} \mathbf{1}\left(\pi\left(s_0^n\right) = a_0^n, \dots \pi\left(s_{H-1}^n\right) = a_{H-1}^n\right) \sum_{t=0}^{H-1} r\left(s_t^n, a_t^n\right)$$

# An Occam's Razor Bound for RL

## Proposition 5.3. (An "Occam's razor bound" for *RL* )

*Let* $\delta \geqslant 0$. *Suppose* $\Pi$ *is a finite and suppose we use the aforementioned estimator,* $\widehat{V}_0^{\pi}(\mu)$, *to estimate the value of every* $\pi \in \Pi$. *Let* $\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{V}_0^{\pi}(\mu)$. *We have that with probability at least* $1 - \delta$,

$$V_0^{\widehat{\pi}}(\mu) \geqslant \max_{\pi \in \Pi} V_0^{\pi}(\mu) - H|\mathcal{A}|^H \sqrt{\frac{2}{N} \log \frac{2|\Pi|}{\delta}}$$

# An Occam's Razor Bound for RL

## Proposition 5.3. (An "Occam's razor bound" for *RL* )

*Let* $\delta \geqslant 0$*. Suppose* $\Pi$ *is a finite and suppose we use the aforementioned estimator,* $\widehat{V}_0^{\pi}(\mu)$*, to estimate the value of every* $\pi \in \Pi$*. Let* $\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{V}_0^{\pi}(\mu)$*. We have that with probability at least* $1 - \delta$*,*

$$V_0^{\widehat{\pi}}(\mu) \geqslant \max_{\pi \in \Pi} V_0^{\pi}(\mu) - H|\mathcal{A}|^H \sqrt{\frac{2}{N} \log \frac{2|\Pi|}{\delta}}$$

**Proof**:

- $|\mathcal{A}|^H \mathbf{1}\left(\pi\left(s_0^n\right) = a_0^n, \ldots \pi\left(s_{H-1}^n\right) = a_{H-1}^n\right) \sum_{t=0}^{H-1} r\left(s_t^n, a_t^n\right) \leqslant H|\mathcal{A}|^H,$
- Hoeffding + union bound

# An Occam's Razor Bound for RL

## Proposition 5.3. (An "Occam's razor bound" for *RL* )

*Let* $\delta \geqslant 0$*. Suppose* $\Pi$ *is a finite and suppose we use the aforementioned estimator,* $\widehat{V}_0^{\pi}(\mu)$*, to estimate the value of every* $\pi \in \Pi$*. Let* $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{V}_0^{\pi}(\mu)$*. We have that with probability at least* $1-\delta$*,*

$$V_0^{\widehat{\pi}}(\mu) \geqslant \max_{\pi\in\Pi} V_0^{\pi}(\mu) - H|\mathcal{A}|^H \sqrt{\frac{2}{N}\log\frac{2|\Pi|}{\delta}}$$

**Proof**:

- $|\mathcal{A}|^H \mathbf{1}\left(\pi\left(s_0^n\right) = a_0^n, \ldots \pi\left(s_{H-1}^n\right) = a_{H-1}^n\right) \sum_{t=0}^{H-1} r\left(s_t^n, a_t^n\right) \leqslant H|\mathcal{A}|^H,$
- Hoeffding + union bound

Hence, provided that

$$N \geqslant H|\mathcal{A}|^H \frac{2\log(2|\Pi|/\delta)}{\epsilon^2}$$

then with probability at least $1-\delta$, we have that:

$$V_0^{\widehat{\pi}}\left(s_0\right) \geqslant \max_{\pi\in\Pi} V_0^{\pi}\left(s_0\right) - \epsilon$$

# Lower Bounds

## Proposition 5.4. (Lower Bound with a Generative Model)

*Suppose algorithm $\mathcal{A}$ has access to a generative model. There exists a policy class $\Pi$, where $|\Pi| = |\mathcal{A}|^H$ such that if algorithm $\mathcal{A}$ returns any policy $\pi$ (not necessarily in $\Pi$ ) such that*

$$V_0^\pi(\mu) \geqslant \max_{\pi \in \Pi} V_0^\pi(\mu) - 0.5.$$

*with probability greater than $1/2$, then $\mathcal{A}$ must make a number of number calls $N$ to the generative model where:*

$$N \geqslant c|\mathcal{A}|^H$$

*(where $c$ is an absolute constant).*

**Proof Sketch**:

- Consider a $|\mathcal{A}|$-ary balanced tree, with $|\mathcal{A}|^H$ states and $|\mathcal{A}|$ actions.
- States correspond nodes and actions correspond to edges; actions always move the agent from the root towards a leaf node.
- Make only one leaf node rewarding, which is unknown to the algorithm.

# Lower Bounds

## Proposition 5.4. (Lower Bound with a Generative Model)

*Suppose algorithm $\mathcal{A}$ has access to a generative model. There exists a policy class $\Pi$, where $|\Pi| = |\mathcal{A}|^H$ such that if algorithm $\mathcal{A}$ returns any policy $\pi$ (not necessarily in $\Pi$ ) such that*

$$V_0^{\pi}(\mu) \geqslant \max_{\pi \in \Pi} V_0^{\pi}(\mu) - 0.5.$$

*with probability greater than $1/2$, then $\mathcal{A}$ must make a number of number calls $N$ to the generative model where:*

$$N \geqslant c|\mathcal{A}|^H$$

*(where $c$ is an absolute constant).*

**Proof Sketch**:

- Consider the policy class to be all $|\mathcal{A}|^H$ policies.
- The theorem immediately follows since the algorithm gains no knowledge of the rewarding leaf node unless it queries that node.

# Outline

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning
2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# Linear Realizability

- In supervised learning, two of the most widely studied settings are those of linear regression and binary classification with halfspaces.
- In both settings, we are able to obtain sample complexity results that are polynomial in the feature dimension.
- We now consider the analogue of these assumptions for RL, we may hope that linearly realizability assumptions may permit a more sample efficient approach.

# Outline

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning

2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# Notation

We start with linear realizability on $Q^{\pi}$ and consider the offline policy evaluation problem.

- Data distributions: $\{\mu_h\}_{h=0}^{H-1}$ where for each $h \in [H], \mu_h \in \Delta\left(\mathcal{S}_h \times \mathcal{A}\right)$
- Inputs: $\{D_h\}_{h=0}^{H-1}$, and for each $h \in [H], D_h$ consists i.i.d. samples of the form $(s, a, r, s') \in \mathcal{S}_h \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}_{h+1}$ tuples
- Policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
- Feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$
- Goal: output an accurate estimate of the value of $\pi$ (i.e., $V^{\pi}$ ) approximately, using the collected datasets $\{D_h\}_{h=0}^{H-1}$, with as few samples as possible.

# Linear Realizability Assumption

## Assumption 5.5 (Realizable Linear Function Approximation)

*For every policy* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, *there exists* $\theta_0^\pi, \ldots \theta_{H-1}^\pi \in \mathbb{R}^d$ *such that for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$ *and* $h \in [H]$,

$$Q_h^\pi(s, a) = (\theta_h^\pi)^\top \phi(s, a).$$

- $\phi(s, a)$ is either hand-crafted or from a pre-trained neural network and transforms a state-action pair to a $d$-dimensional embedding
- $Q$-functions can be predicted by linear functions of the features

# Linear Realizability Assumption

## Assumption 5.6 (Coverage)

*For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, assume our feature map is bounded such that $\|\phi(s, a)\|_2 \leqslant 1$. Furthermore, suppose for each $h \in [H]$, the data distributions $\mu_h$ satisfies the following:*

$$\mathbb{E}_{(s,a)\sim\mu_h}\left[\phi(s, a)\phi(s, a)^\top\right] = \frac{1}{d} I.$$

- This distribution satisfies the D-optimal design property introduced before.

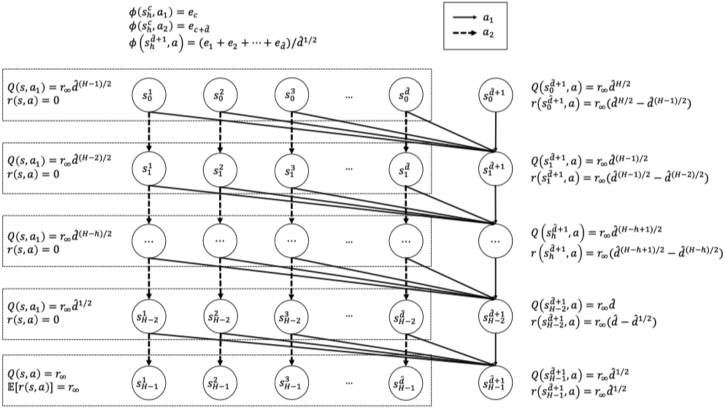
# Hardness result

The following theorem shows these assumptions are not sufficient for offline policy evaluation for long horizon problems.

## Theorem 5.7.

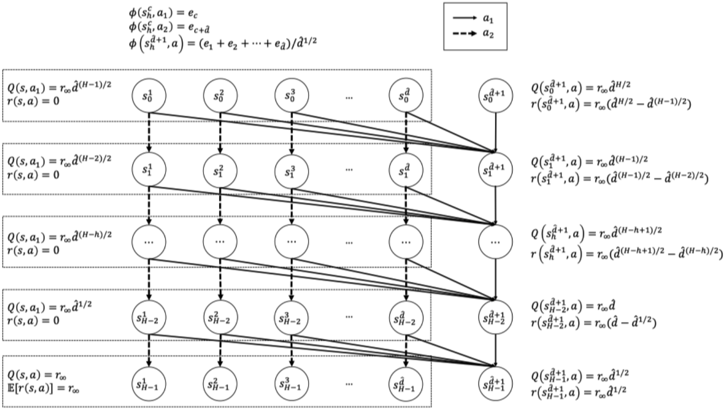
*Suppose Assumption 5.6 holds. Fix an algorithm that takes as input both a policy and a feature mapping. There exists a (deterministic) MDP satisfying Assumption 5.5 such that for any policy* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$*, the algorithm requires* $\Omega\left((d/2)^H\right)$ *samples to output the value of* $\pi$ *up to constant additive approximation error with probability at least 0.9.*

# Proof of Theorem 5.7


$\phi(s_h^c, a_1) = e_c$
$\phi(s_h^c, a_2) = e_{c+\hat{d}}$
$\phi\left(s_h^{\hat{d}+1}, a\right) = (e_1 + e_2 + \cdots + e_{\hat{d}})/\hat{d}^{1/2}$
$a_1$
$a_2$
$Q(s, a_1) = r_\infty \hat{d}^{(H-1)/2}$
$r(s, a) = 0$
$s_0^1$
$s_0^2$
$s_0^3$
...
$s_0^{\hat{d}}$
$s_0^{\hat{d}+1}$
$Q(s_0^{\hat{d}+1}, a) = r_\infty \hat{d}^{H/2}$
$r(s_0^{\hat{d}+1}, a) = r_\infty(\hat{d}^{H/2} - \hat{d}^{(H-1)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-2)/2}$
$r(s, a) = 0$
$s_1^1$
$s_1^2$
$s_1^3$
...
$s_1^{\hat{d}}$
$s_1^{\hat{d}+1}$
$Q(s_1^{\hat{d}+1}, a) = r_\infty \hat{d}^{(H-1)/2}$
$r(s_1^{\hat{d}+1}, a) = r_\infty(\hat{d}^{(H-1)/2} - \hat{d}^{(H-2)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-h)/2}$
$r(s, a) = 0$
...
$Q\left(s_h^{\hat{d}+1}, a\right) = r_\infty \hat{d}^{(H-h+1)/2}$
$r\left(s_h^{\hat{d}+1}, a\right) = r_\infty(\hat{d}^{(H-h+1)/2} - \hat{d}^{(H-h)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{1/2}$
$r(s, a) = 0$
$s_{H-2}^1$
$s_{H-2}^2$
$s_{H-2}^3$
...
$s_{H-2}^{\hat{d}}$
$s_{H-2}^{\hat{d}+1}$
$Q(s_{H-2}^{\hat{d}+1}, a) = r_\infty \hat{d}$
$r(s_{H-2}^{\hat{d}+1}, a) = r_\infty(\hat{d} - \hat{d}^{1/2})$
$Q(s, a) = r_\infty$
$\mathbb{E}[r(s, a)] = r_\infty$
$s_{H-1}^1$
$s_{H-1}^2$
$s_{H-1}^3$
...
$s_{H-1}^{\hat{d}}$
$s_{H-1}^{\hat{d}+1}$
$Q(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$
$r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$

- Assume $d$ is even for simplicity
- $\hat{d} = d/2$
- A hard instance is constructed

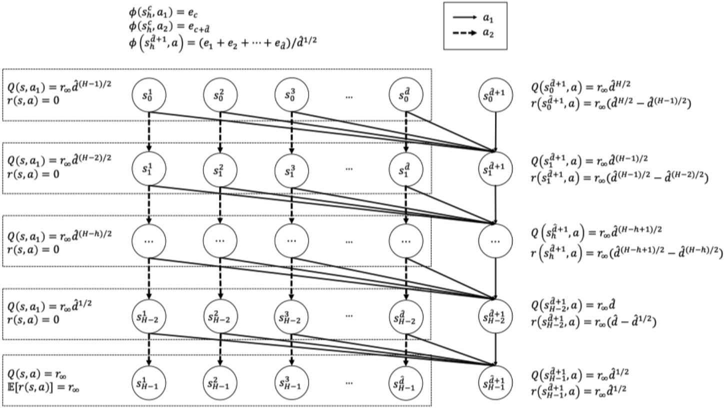
# State Space and Action Space



$\phi(s_h^c, a_1) = e_c$
$\phi(s_h^c, a_2) = e_{c+\hat{d}}$
$\phi\left(s_h^{\hat{d}+1}, a\right) = (e_1 + e_2 + \cdots + e_{\hat{d}})/\hat{d}^{1/2}$
$a_1$
$a_2$
$Q(s, a_1) = r_\infty \hat{d}^{(H-1)/2}$
$r(s,a) = 0$
$s_0^1$ $s_0^2$ $s_0^3$ ... $s_0^{\hat{d}}$ $s_0^{\hat{d}+1}$
$Q(s_0^{\hat{d}+1}, a) = r_\infty \hat{d}^{H/2}$
$r(s_0^{\hat{d}+1}, a) = r_\infty(\hat{d}^{H/2} - \hat{d}^{(H-1)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-2)/2}$
$r(s,a) = 0$
$s_1^1$ $s_1^2$ $s_1^3$ ... $s_1^{\hat{d}}$ $s_1^{\hat{d}+1}$
$Q(s_1^{\hat{d}+1}, a) = r_\infty \hat{d}^{(H-1)/2}$
$r(s_1^{\hat{d}+1}, a) = r_\infty(\hat{d}^{(H-1)/2} - \hat{d}^{(H-2)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-h)/2}$
$r(s,a) = 0$
... ... ... ... ... ...
$Q\left(s_h^{\hat{d}+1}, a\right) = r_\infty \hat{d}^{(H-h+1)/2}$
$r\left(s_h^{\hat{d}+1}, a\right) = r_\infty(\hat{d}^{(H-h+1)/2} - \hat{d}^{(H-h)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{1/2}$
$r(s,a) = 0$
$s_{H-2}^1$ $s_{H-2}^2$ $s_{H-2}^3$ ... $s_{H-2}^{\hat{d}}$ $s_{H-2}^{\hat{d}+1}$
$Q(s_{H-2}^{\hat{d}+1}, a) = r_\infty \hat{d}$
$r(s_{H-2}^{\hat{d}+1}, a) = r_\infty(\hat{d} - \hat{d}^{1/2})$
$Q(s, a) = r_\infty$
$\mathbb{E}[r(s,a)] = r_\infty$
$s_{H-1}^1$ $s_{H-1}^2$ $s_{H-1}^3$ ... $s_{H-1}^{\hat{d}}$ $s_{H-1}^{\hat{d}+1}$
$Q(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$
$r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$

- The action space $\mathcal{A} = \{a_1, a_2\}$.
- For each $h \in [H], \mathcal{S}_h$ contains $\hat{d} + 1$ states $s_h^1, s_h^2, \ldots, s_h^{\hat{d}}$ and $s_h^{\hat{d}+1}$.

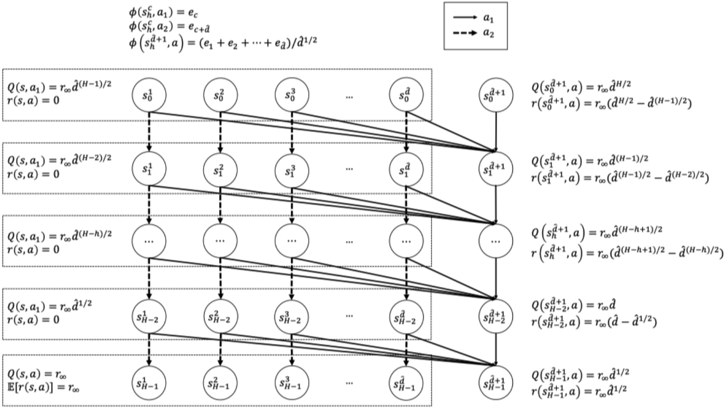
# Transition Operator


$\phi(s_h^c, a_1) = e_c$
$\phi(s_h^c, a_2) = e_{c+\hat{d}}$
$\phi\left(s_h^{\hat{d}+1}, a\right) = (e_1 + e_2 + \cdots + e_{\hat{d}})/\hat{d}^{1/2}$
$a_1$
$a_2$
$Q(s, a_1) = r_\infty \hat{d}^{(H-1)/2}$
$r(s, a) = 0$
$s_0^1$
$s_0^2$
$s_0^3$
...
$s_0^{\hat{d}}$
$s_0^{\hat{d}+1}$
$Q(s_0^{\hat{d}+1}, a) = r_\infty \hat{d}^{H/2}$
$r(s_0^{\hat{d}+1}, a) = r_\infty(\hat{d}^{H/2} - \hat{d}^{(H-1)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-2)/2}$
$r(s, a) = 0$
$s_1^1$
$s_1^2$
$s_1^3$
...
$s_1^{\hat{d}}$
$s_1^{\hat{d}+1}$
$Q(s_1^{\hat{d}+1}, a) = r_\infty \hat{d}^{(H-1)/2}$
$r(s_1^{\hat{d}+1}, a) = r_\infty(\hat{d}^{(H-1)/2} - \hat{d}^{(H-2)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-h)/2}$
$r(s, a) = 0$
...
$Q\left(s_h^{\hat{d}+1}, a\right) = r_\infty \hat{d}^{(H-h+1)/2}$
$r\left(s_h^{\hat{d}+1}, a\right) = r_\infty(\hat{d}^{(H-h+1)/2} - \hat{d}^{(H-h)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{1/2}$
$r(s, a) = 0$
$s_{H-2}^1$
$s_{H-2}^2$
$s_{H-2}^3$
...
$s_{H-2}^{\hat{d}}$
$s_{H-2}^{\hat{d}+1}$
$Q(s_{H-2}^{\hat{d}+1}, a) = r_\infty \hat{d}$
$r(s_{H-2}^{\hat{d}+1}, a) = r_\infty(\hat{d} - \hat{d}^{1/2})$
$Q(s, a) = r_\infty$
$\mathbb{E}[r(s, a)] = r_\infty$
$s_{H-1}^1$
$s_{H-1}^2$
$s_{H-1}^3$
...
$s_{H-1}^{\hat{d}}$
$s_{H-1}^{\hat{d}+1}$
$Q(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$
$r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$

- For each $h \in \{0, 1, \ldots, H-2\}$, for each $c \in \{1, 2, \ldots, \hat{d}+1\}$, we have

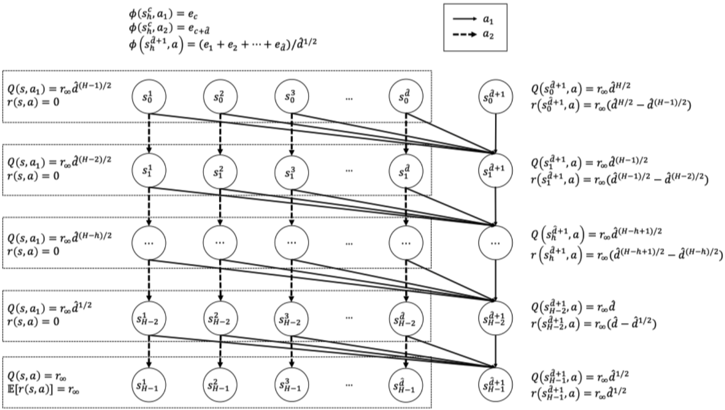
$$P\left(s \mid s_h^c, a\right) = \begin{cases} 1 & s = s_{h+1}^{\hat{d}+1}, a = a_1 \\ 1 & s = s_{h+1}^c, a = a_2 \\ 0 & \text{else} \end{cases}$$

# Reward Distributions


$\phi(s_h^c, a_1) = e_c$
$\phi(s_h^c, a_2) = e_{c+\hat{d}}$
$\phi\left(s_h^{\hat{d}+1}, a\right) = (e_1 + e_2 + \cdots + e_{\hat{d}})/\hat{d}^{1/2}$
$a_1$
$a_2$
$Q(s, a_1) = r_\infty \hat{d}^{(H-1)/2}$
$r(s, a) = 0$
$s_0^1$
$s_0^2$
$s_0^3$
...
$s_0^{\hat{d}}$
$s_0^{\hat{d}+1}$
$Q(s_0^{\hat{d}+1}, a) = r_\infty \hat{d}^{H/2}$
$r(s_0^{\hat{d}+1}, a) = r_\infty(\hat{d}^{H/2} - \hat{d}^{(H-1)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-2)/2}$
$r(s, a) = 0$
$s_1^1$
$s_1^2$
$s_1^3$
...
$s_1^{\hat{d}}$
$s_1^{\hat{d}+1}$
$Q(s_1^{\hat{d}+1}, a) = r_\infty \hat{d}^{(H-1)/2}$
$r(s_1^{\hat{d}+1}, a) = r_\infty(\hat{d}^{(H-1)/2} - \hat{d}^{(H-2)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-h)/2}$
$r(s, a) = 0$
...
$Q\left(s_h^{\hat{d}+1}, a\right) = r_\infty \hat{d}^{(H-h+1)/2}$
$r\left(s_h^{\hat{d}+1}, a\right) = r_\infty(\hat{d}^{(H-h+1)/2} - \hat{d}^{(H-h)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{1/2}$
$r(s, a) = 0$
$s_{H-2}^1$
$s_{H-2}^2$
$s_{H-2}^3$
...
$s_{H-2}^{\hat{d}}$
$s_{H-2}^{\hat{d}+1}$
$Q(s_{H-2}^{\hat{d}+1}, a) = r_\infty \hat{d}$
$r(s_{H-2}^{\hat{d}+1}, a) = r_\infty(\hat{d} - \hat{d}^{1/2})$
$Q(s, a) = r_\infty$
$\mathbb{E}[r(s, a)] = r_\infty$
$s_{H-1}^1$
$s_{H-1}^2$
$s_{H-1}^3$
...
$s_{H-1}^{\hat{d}}$
$s_{H-1}^{\hat{d}+1}$
$Q(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$
$r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$

- Let $0 \leqslant r_\infty \leqslant \hat{d}^{-H/2}$ be a parameter to be determined.
- For each $(h, c) \in \{0, 1, \ldots, H-2\} \times [\hat{d}]$ and $a \in \mathcal{A}$, we set $r(s_h^c, a) = 0$ and $r(s_h^{\hat{d}+1}, a) = r_\infty \cdot (\hat{d}^{(H-h)/2} - \hat{d}^{(H-h-1)/2})$.

# Reward Distributions

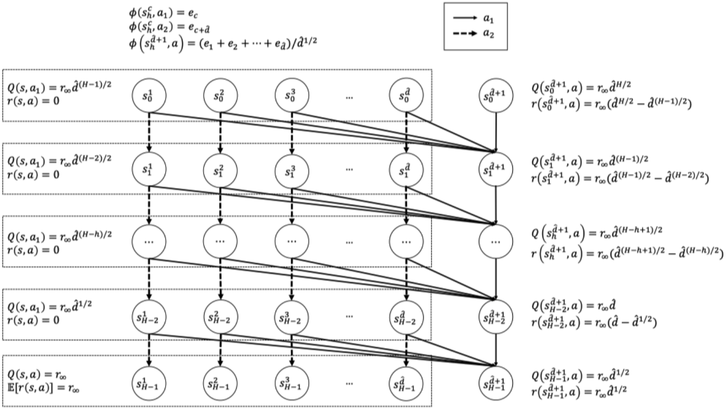

$\phi(s_h^c, a_1) = e_c$
$\phi(s_h^c, a_2) = e_{c+\hat{d}}$
$\phi\left(s_h^{\hat{d}+1}, a\right) = (e_1 + e_2 + \cdots + e_{\hat{d}})/\hat{d}^{1/2}$
$a_1$
$a_2$
$Q(s, a_1) = r_\infty \hat{d}^{(H-1)/2}$
$r(s, a) = 0$
$s_0^1$
$s_0^2$
$s_0^3$
...
$s_0^{\hat{d}}$
$s_0^{\hat{d}+1}$
$Q(s_0^{\hat{d}+1}, a) = r_\infty \hat{d}^{H/2}$
$r(s_0^{\hat{d}+1}, a) = r_\infty(\hat{d}^{H/2} - \hat{d}^{(H-1)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-2)/2}$
$r(s, a) = 0$
$s_1^1$
$s_1^2$
$s_1^3$
...
$s_1^{\hat{d}}$
$s_1^{\hat{d}+1}$
$Q(s_1^{\hat{d}+1}, a) = r_\infty \hat{d}^{(H-1)/2}$
$r(s_1^{\hat{d}+1}, a) = r_\infty(\hat{d}^{(H-1)/2} - \hat{d}^{(H-2)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-h)/2}$
$r(s, a) = 0$
...
$Q\left(s_h^{\hat{d}+1}, a\right) = r_\infty \hat{d}^{(H-h+1)/2}$
$r\left(s_h^{\hat{d}+1}, a\right) = r_\infty(\hat{d}^{(H-h+1)/2} - \hat{d}^{(H-h)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{1/2}$
$r(s, a) = 0$
$s_{H-2}^1$
$s_{H-2}^2$
$s_{H-2}^3$
...
$s_{H-2}^{\hat{d}}$
$s_{H-2}^{\hat{d}+1}$
$Q(s_{H-2}^{\hat{d}+1}, a) = r_\infty \hat{d}$
$r(s_{H-2}^{\hat{d}+1}, a) = r_\infty(\hat{d} - \hat{d}^{1/2})$
$Q(s, a) = r_\infty$
$\mathbb{E}[r(s, a)] = r_\infty$
$s_{H-1}^1$
$s_{H-1}^2$
$s_{H-1}^3$
...
$s_{H-1}^{\hat{d}}$
$s_{H-1}^{\hat{d}+1}$
$Q(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$
$r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$

- For the last level, for each $c \in [\hat{d}]$ and $a \in \mathcal{A}$, we set

$$r(s_{H-1}^c, a) = \begin{cases} 1 & \text{with probability } (1 + r_\infty)/2 \\ -1 & \text{with probability } (1 - r_\infty)/2 \end{cases}$$

- Moreover, for all actions $a \in \mathcal{A}$, $r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \cdot \hat{d}^{1/2}$.

# Feature Mapping


$\phi(s_h^c, a_1) = e_c$
$\phi(s_h^c, a_2) = e_{c+\hat{d}}$
$\phi\left(s_h^{\hat{d}+1}, a\right) = (e_1 + e_2 + \cdots + e_{\hat{d}})/\hat{d}^{1/2}$
$a_1$
$a_2$
$Q(s, a_1) = r_\infty \hat{d}^{(H-1)/2}$
$r(s, a) = 0$
$s_0^1$ $s_0^2$ $s_0^3$ ... $s_0^{\hat{d}}$ $s_0^{\hat{d}+1}$
$Q(s_0^{\hat{d}+1}, a) = r_\infty \hat{d}^{H/2}$
$r(s_0^{\hat{d}+1}, a) = r_\infty(\hat{d}^{H/2} - \hat{d}^{(H-1)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-2)/2}$
$r(s, a) = 0$
$s_1^1$ $s_1^2$ $s_1^3$ ... $s_1^{\hat{d}}$ $s_1^{\hat{d}+1}$
$Q(s_1^{\hat{d}+1}, a) = r_\infty \hat{d}^{(H-1)/2}$
$r(s_1^{\hat{d}+1}, a) = r_\infty(\hat{d}^{(H-1)/2} - \hat{d}^{(H-2)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{(H-h)/2}$
$r(s, a) = 0$
... ... ... ... ... ...
$Q\left(s_h^{\hat{d}+1}, a\right) = r_\infty \hat{d}^{(H-h+1)/2}$
$r\left(s_h^{\hat{d}+1}, a\right) = r_\infty(\hat{d}^{(H-h+1)/2} - \hat{d}^{(H-h)/2})$
$Q(s, a_1) = r_\infty \hat{d}^{1/2}$
$r(s, a) = 0$
$s_{H-2}^1$ $s_{H-2}^2$ $s_{H-2}^3$ ... $s_{H-2}^{\hat{d}}$ $s_{H-2}^{\hat{d}+1}$
$Q(s_{H-2}^{\hat{d}+1}, a) = r_\infty \hat{d}$
$r(s_{H-2}^{\hat{d}+1}, a) = r_\infty(\hat{d} - \hat{d}^{1/2})$
$Q(s, a) = r_\infty$
$\mathbb{E}[r(s, a)] = r_\infty$
$s_{H-1}^1$ $s_{H-1}^2$ $s_{H-1}^3$ ... $s_{H-1}^{\hat{d}}$ $s_{H-1}^{\hat{d}+1}$
$Q(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$
$r(s_{H-1}^{\hat{d}+1}, a) = r_\infty \hat{d}^{1/2}$

- Let $e_1, e_2, \ldots, e_d$ be a set of orthonormal vectors in $\mathbb{R}^d$.
- For each $(h, c) \in [H] \times [\hat{d}]$, we set $\phi\left(s_h^c, a_1\right) = e_c, \phi\left(s_h^c, a_2\right) = e_{c+\hat{d}}$, and

$$\phi\left(s_h^{\hat{d}+1}, a\right) = \frac{1}{\hat{d}^{1/2}} \sum_{c \in \hat{d}} e_c$$

for all $a \in \mathcal{A}$.

# Verifying Assumption 5.5.

## Lemma 5.8

*For every policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, for each $h \in [H]$, for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, we have $Q_h^\pi(s,a) = \left(\theta_h^\pi\right)^\top \phi(s,a)$ for some $\theta_h^\pi \in \mathbb{R}^d$.*

**Proof**: We first verify $Q^\pi$ is linear for the **first** $H-1$ **levels**. For each $(h,c) \in \{0,1,\ldots,H-2\} \times [\hat{d}]$, we have

$$Q_h^\pi(s_h^c, a_1) = r(s_h^c, a_1) + r(s_{h+1}^{\hat{d}+1}, a_1) + r(s_{h+2}^{\hat{d}+1}, a_1) + \ldots + r(s_{H-1}^{\hat{d}+1}, a_1) = r_\infty \cdot \hat{d}^{(H-h-1)/2}.$$

Moreover, for all $a \in \mathcal{A}$,

$$Q_h^\pi(s_h^{\hat{d}+1}, a) = r(s_h^{\hat{d}+1}, a) + r(s_{h+1}^{\hat{d}+1}, a_1) + r(s_{h+2}^{\hat{d}+1}, a_1) + \ldots + r(s_{H-1}^{\hat{d}+1}, a_1) = r_\infty \cdot \hat{d}^{(H-h)/2}.$$

Therefore, if we define

$$\theta_h^\pi = \sum_{c=1}^{\hat{d}} r_\infty \cdot \hat{d}^{(H-h-1)/2} \cdot e_c + \sum_{c=1}^{\hat{d}} Q_h^\pi(s_h^c, a_2) \cdot e_{c+\hat{d}},$$

then $Q_h^\pi(s,a) = (\theta_h^\pi)^\top \phi(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$.

# Verifying Assumption 5.5

Now we verify that the $Q$-function is linear for **the last level**. Clearly, for all $c \in [\hat{d}]$ and $a \in \mathcal{A}$,

$$Q^{\pi}_{H-1}(s^{c}_{H-1}, a) = r_{\infty}$$

and

$$Q^{\pi}_{H-1}(s^{\hat{d}+1}_{H-1}, a) = r_{\infty} \cdot \sqrt{\hat{d}}.$$

Thus, by defining $\theta^{\pi}_{H-1} = \sum_{c=1}^{d} r_{\infty} \cdot e_c$, we have $Q^{\pi}_{H-1}(s, a) = \left(\theta^{\pi}_{H-1}\right)^{\top} \phi(s, a)$ for all $(s, a) \in \mathcal{S}_{H-1} \times \mathcal{A}$.

# The Data Distributions

For each level $h \in [H]$, the data distribution $\mu_h$ is a uniform distribution over the set $\{(s_h^1, a_1), (s_h^1, a_2), (s_h^2, a_1), (s_h^2, a_2), \ldots, (s_h^{\hat{d}}, a_1), (s_h^{\hat{d}}, a_2)\}$. Notice that $(s_h^{\hat{d}+1}, a)$ is **not** in the support of $\mu_h$ for all $a \in \mathcal{A}$. It can be seen that,

$$\mathbb{E}_{(s,a)\sim\mu_h}\left[\phi(s,a)\phi(s,a)^\top\right] = \frac{1}{d}\sum_{c=1}^{d} e_c e_c^\top = \frac{1}{d} I.$$

# Proof of Theorem 5.7

- It is information-theoretically hard for any algorithm to distinguish the case $r_\infty = 0$ and $r_\infty = \hat{d}^{-H/2}$.
- Fix the initial state to be $s_0^{\hat{d}+1}$.
- When $r_\infty = 0$, the value of $\pi$ would be zero.
- When $r_\infty = \hat{d}^{-H/2}$, the value of $\pi$ would be $r_\infty \cdot \hat{d}^{H/2} = 1$.
- Thus, if the algorithm approximates the value of the policy up to an error of $1/2$, then it must distinguish the case that $r_\infty = 0$ and $r_\infty = \hat{d}^{-H/2}$.

# Proof of Theorem 5.7

- For the case $r_\infty = 0$ and $r_\infty = \hat{d}^{-H/2}$, $\{\mu_h\}_{h=0}^{H-1}$, $\phi$, $\pi$ and $P$ are the same.
- Thus, in order to distinguish the 2 cases, the only way is to query the reward distribution by using sampling taken from the data distributions.
- For all state-action pairs $(s, a)$ in the support of the data distributions of the first $H - 1$ levels, the reward distributions will be identical.
- For the case $r_\infty = 0$ and $r_\infty = \hat{d}^{-H/2}$, for all state-action pairs $(s, a)$ in the support of the data distribution of the last level,

$$r(s,a) = \begin{cases} 1 & \text{with probability } \ (1 + r_\infty)/2 \\ -1 & \text{with probability } \ (1 - r_\infty)/2 \end{cases} .$$

# Proof of Theorem 5.7

Therefore, to distinguish the case that $r_\infty = 0$ and $r_\infty = \hat{d}^{-H/2}$, the agent needs to distinguish two reward distributions

$$r^{(1)} = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

and

$$r^{(2)} = \begin{cases} 1 & \text{with probability } (1 + \hat{d}^{-H/2})/2 \\ -1 & \text{with probability } (1 - \hat{d}^{-H/2})/2 \end{cases} .$$

It is standard argument that in order to distinguish $r^{(1)}$ and $r^{(2)}$ with probability at least 0.9, any algorithm requires $\Omega(\hat{d}^H)$ samples.

# Comments

- The key in this construction is the state $s_h^{\hat{d}+1}$ in each level.
- In each level, $s_h^{\hat{d}+1}$ amplifies the $Q$-values by a $\hat{d}^{1/2}$ factor.
- After all the $H$ levels, the value will be amplified by a $\hat{d}^{H/2}$ factor.
- Since $s_h^{\hat{d}+1}$ is not in the support of the data distribution, the only way for the agent to estimate the value of the policy is to estimate the expected reward value in the last level.
- This construction forces the estimation error of the last level to be amplified exponentially and thus implies an exponential lower bound.

# Outline

1. Agnostic Learning
   - Review: Binary Classification
   - Importance Sampling and a Reduction to Supervised Learning

2. Linear Realizability
   - Offline Policy Evaluation with Linearly Realizable Values
   - Linearly Realizable $Q^*$

# Linearly Realizable $Q^*$

We consider the problem of learning with only a linearly realizability assumption on $Q^\star$ (along with access to either a generative model or sampling access in the episodic setting).

# Linear Realizability Assumption

### Assumption 5.9 (Linear $Q^*$ Realizability)

*For all $h \in [H]$, assume there exists $\theta_h^* \in \mathbb{R}^d$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$Q_h^*(s, a) = \theta_h^* \cdot \phi(s, a).$$

- The hope is that this assumption may permit a sample complexity that is polynomial in $d$ and $H$, with no explicit $|\mathcal{S}|$ or $|\mathcal{A}|$ dependence.

# Linear Realizability Assumption

### Assumption 5.10 (Constant Sub-optimality Gap)

*For any state $s \in \mathcal{S}$, $a \in \mathcal{A}$, the suboptimality gap is defined as $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a)$. We assume that*

$$\min_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}} \{\Delta_h(s,a) : \Delta_h(s,a) > 0\} \geqslant \Delta_{\min}.$$

- The hope is that with a "large gap", the identification of the optimal policy itself (as opposed to just estimating its value accurately) may be statistically easier, thus making the problem easier.

# Hardness results

**Theorem 5.11. (Linear $Q^{\star}$; Generative Model Case)**

*Consider any algorithm $\mathcal{A}$ which has access to a generative model and which takes as input the feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. There exists an MDP with a feature mapping $\phi$ satisfying Assumption 5.9 and where the size of the action space is $|\mathcal{A}| = c_1 \left\lceil \min\left\{d^{1/4}, H^{1/2}\right\}\right\rceil$ such that if $\mathcal{A}$ (when given $\phi$ as input) finds a policy $\pi$ such that*

$$\mathbb{E}_{s_1 \sim \mu} V^{\pi}(s_1) \geqslant \mathbb{E}_{s_1 \sim \mu} V^{*}(s_1) - 0.05$$

*with probability 0.1, then $\mathcal{A}$ requires $\min\left\{2^{c_2 d}, 2^{c_2 H}\right\}$ samples ($c_1$ and $c_2$ are absolute constants).*

- The implications of the above show that the linear $Q^*$ assumption, alone, is not sufficient for sample efficient RL, even with access to a generative model.

# Hardness results

### Theorem 5.12. (Linear $Q^\star$ Realizability + Gap; Episodic Setting)

*Consider any algorithm $\mathcal{A}$ which has access to the episodic sampling model and which takes as input the feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. There exists an MDP with a feature mapping $\phi$ satisfying Assumption 5.9 and Assumption 5.10 (where $\Delta_{\min}$ is an absolute constant) such that if $\mathcal{A}$ (using $\phi$ as input) finds a policy $\pi$ such that*

$$\mathbb{E}_{s_1 \sim \mu} V^{\pi}\left(s_1\right) \geqslant \mathbb{E}_{s_1 \sim \mu} V^{*}\left(s_1\right) - 0.05$$

*with probability 0.1, then $\mathcal{A}$ requires $\min\left\{2^{cd}, 2^{cH}\right\}$ samples (where $c$ is an absolute constant).*

# Proof of Theorem 5.12

We prove Theorem 5.12 by providing the construction of a hard family of MDPs where $Q^*$ is linearly realizable and has constant suboptimality gap and where it takes exponential samples to learn a near-optimal policy.

- Let $m$ be an integer to be determined.
- The state space is $\{\bar{1}, \cdots, \bar{m}, f\}$. The special state $f$ is called the terminal state.
- The action space is $\mathcal{A} = \{1, 2, \ldots, m\}$.
- Each MDP in this family is specified by an index $a^* \in \{1, 2, \ldots, m\}$ and denoted by $\mathcal{M}_{a*}$.

# Proof of Theorem 5.12

We will use the Johnson-Lindenstrauss lemma:

**Lemma 5.13 (Johnson-Lindenstrauss)**

*For any $\alpha > 0$, if $m \leqslant \exp\left(\frac{1}{8}\alpha^2 d'\right)$, there exists $m$ unit vectors $\{v_1, \cdots, v_m\}$ in $\mathbb{R}^{d'}$ such that $\forall i, j \in \{1, 2, \ldots, m\}$ such that $i \neq j, |\langle v_i, v_j \rangle| \leqslant \alpha$.*

- Set $\alpha = \frac{1}{6}$ and $m = \left\lfloor \exp\left(\frac{1}{8}\alpha^2 d\right) \right\rfloor$.
- By Lemma 5.13, we can find such a set of $d$-dimensional unit vectors $\{v_1, \cdots, v_m\}$.
- We use $v_i$ and $v(i)$ interchangeably.

# Features

The feature map, which maps state-action pairs to $d + 1$ dimensional vectors, is defined as follows.

$$\begin{aligned}
\phi(\overline{a_1}, a_2) &:= \left(0, \left(\left\langle v(a_1), v(a_2)\right\rangle + 2\alpha\right) \cdot v(a_2)\right), && (\forall a_1, a_2 \in \{1, 2, \ldots, m\}, a_1 \neq a_2) \\
\phi(\overline{a_1}, a_1) &:= \left(\frac{3}{4}\alpha, \mathbf{0}\right), && (\forall a_1 \in \{1, 2, \ldots, m\}) \\
\phi(f, 1) &= (0, \mathbf{0}), \\
\phi(f, a) &:= (-1, \mathbf{0}). && (\forall a \neq 1)
\end{aligned}$$

Here $\mathbf{0}$ is the zero vector in $\mathbb{R}^d$. Note that the feature map is independent of $a^*$ and is shared across the MDP family.

# Rewards

For $h \in \{0, \ldots, H-2\}$, the rewards are defined as

$$\begin{aligned}
r_h(\overline{a_1}, a^*) &:= \left\langle v(a_1), v(a^*) \right\rangle + 2\alpha, && (a_1 \neq a^*) \\
r_h(\overline{a_1}, a_2) &:= -2\alpha \left[ \left\langle v(a_1), v(a_2) \right\rangle + 2\alpha \right], && (a_2 \neq a^*, a_2 \neq a_1) \\
r_h(\overline{a_1}, a_1) &:= \frac{3}{4}\alpha, && (\forall a_1) \\
r_h(f, 1) &:= 0, \\
r_h(f, a) &:= -1. && (a \neq 1)
\end{aligned}$$

For $h = H - 1$, $r_{H-1}(s, a) := \langle \phi(s, a), (1, v(a^*)) \rangle$ for every state-action pair.

# Transitions

The initial state distribution $\mu$ is set as a uniform distribution over $\{\bar{1}, \cdots, \bar{m}\}$. The transition probabilities are set as follows.

$$
\begin{aligned}
\Pr[f|\overline{a_1}, a^*] &= 1, \\
\Pr[f|\overline{a_1}, a_1] &= 1, \\
\Pr[\cdot|\overline{a_1}, a_2] &= \begin{cases} \overline{a_2} : \left\langle v(a_1), v(a_2) \right\rangle + 2\alpha \\ f : 1 - \left\langle v(a_1), v(a_2) \right\rangle - 2\alpha \end{cases}, \qquad (a_2 \neq a^*, a_2 \neq a_1) \\
\Pr[f|f, \cdot] &= 1.
\end{aligned}
$$

**After taking action $a_2$, the next state is either $\overline{a_2}$ or $f$.**

# Verifying Assumption 5.9.

### Lemma 5.14. (Linear realizability)

*In the MDP $\mathcal{M}_{a*}, \forall h \in [H]$, for any state-action pair $(s, a)$, $Q_h^*(s, a) = \langle \phi(s, a), \theta^* \rangle$ with $\theta^* = (1, v(a^*))$.*

**Proof**: **We first verify the statement for the terminal state $f$.** Observe that at the terminal state $f$, the next state is always $f$ and the reward is either 0 (if action 1 is chosen) or -1 (if an action other than 1 is chosen). Hence, we have

$$Q_h^*(f, a) = \begin{cases} 0 & a = 1 \\ -1 & a \neq 1 \end{cases}$$

and

$$V_h^*(f) = 0$$

This implies $Q_h^*(f, \cdot) = \langle \phi(f, \cdot), (1, v(a^*)) \rangle$.

# Verifying Assumption 5.9.

**Proof**: We now verify realizability for other states via **induction** on $h = H - 1, \cdots, 0$. The induction hypothesis is that for all $a_1, a_2 \in \{1, 2, \ldots, m\}$, we have

$$Q_h^*\left(\overline{a_1}, a_2\right) = \begin{cases} \left(\left\langle v\left(a_1\right), v\left(a_2\right)\right\rangle + 2\alpha\right) \cdot \left\langle v\left(a_2\right), v\left(a^*\right)\right\rangle & a_1 \neq a_2 \\ \frac{3}{4}\alpha & a_1 = a_2 \end{cases} \tag{1}$$

and

$$V_h^*\left(\overline{a_1}\right) = \begin{cases} \left\langle v\left(a_1\right), v\left(a^*\right)\right\rangle + 2\alpha & a_1 \neq a^* \\ \frac{3}{4}\alpha & a_1 = a^* \end{cases}. \tag{2}$$

Note that (1) implies that realizability is satisfied. In the remaining part of the proof we verify Eq. (1) and (2).

# Verifying Assumption 5.9.

**Proof**: When $h = H - 1$, (1) holds by the definition of rewards. Next, note that for all $h \in [H]$, (2) follows from (1). This is because for all $a_1 \neq a^*$, for all $a_2 \notin \{a_1, a^*\}$.

$$Q_h^* \left(\overline{a_1}, a_2\right) = \left(\left\langle v\left(a_1\right), v\left(a_2\right)\right\rangle + 2\alpha\right) \cdot \left\langle v\left(a_2\right), v\left(a^*\right)\right\rangle \leqslant 3\alpha^2$$

Moreover, for all $a_1 \neq a^*$,

$$Q_h^* \left(\overline{a_1}, a_1\right) = \frac{3}{4}\alpha < \alpha$$

Furthermore, for all $a_1 \neq a^*$,

$$Q_h^* \left(\overline{a_1}, a^*\right) = \left\langle v\left(a_1\right), v\left(a^*\right)\right\rangle + 2\alpha \geqslant \alpha > 3\alpha^2$$

In other words, (1) implies that $a^*$ is always the optimal action for all state $\overline{a_1}$ with $a_1 \neq a^*$. Now, for state $\overline{a^*}$, for all $a \neq a^*$, we have

$$Q_h^* \left(\overline{a^*}, a\right) = \left(\left\langle v\left(a^*\right), v(a)\right\rangle + 2\alpha\right) \cdot \left\langle v\left(a^*\right), v(a)\right\rangle \leqslant 3\alpha^2 < \frac{3}{4}\alpha = Q_h^* \left(\overline{a^*}, a^*\right).$$

**Hence, (1) implies that $a^*$ is always the optimal action for all states $\bar{a}$ with $a \in \{1, 2, \ldots, m\}$.**

# Verifying Assumption 5.9.

**Proof**: **Thus, it remains to show that** (1) **holds for** $h$ **assuming** (2) **holds for** $h+1$**.** Here we only consider the case that $a_2 \neq a_1$ and $a_2 \neq a^*$, since otherwise $\Pr\left[f \mid \overline{a_1}, a_2\right] = 1$ and thus (1) holds by the definition of the rewards and the fact that $V_h^*(f) = 0$. When $a_2 \notin \{a_1, a^*\}$, we have

$$
\begin{aligned}
Q_h^*(\overline{a_1}, a_2) &= r_h(\overline{a_1}, a_2) + \mathbb{E}_{s_{h+1}}\left[V_{h+1}^*(s_{h+1}) \middle| \overline{a_1}, a_2\right] \\
&= -2\alpha\left[\left\langle v(a_1), v(a_2)\right\rangle + 2\alpha\right] + \Pr[s_{h+1} = \overline{a_2}] \cdot V_{h+1}^*(\overline{a_2}) + \Pr[s_{h+1} = f] \cdot V_{h+1}^*(f) \\
&= -2\alpha\left[\left\langle v(a_1), v(a_2)\right\rangle + 2\alpha\right] + \left[\left\langle v(a_1), v(a_2)\right\rangle + 2\alpha\right] \cdot \left(\left\langle v(a_2), v(a^*)\right\rangle + 2\alpha\right) \\
&= \left(\left\langle v(a_1), v(a_2)\right\rangle + 2\alpha\right) \cdot \left\langle v(a_2), v(a^*)\right\rangle.
\end{aligned}
$$

This is exactly (1) for $h$. Hence both (1) and (2) hold for all $h \in [H]$.

# Verifying Assumption 5.10.

## Lemma 5.15. (Constant Gap)

*Assumption 5.10 is satisfied with* $\Delta_{\min} = 1/24$.

**Proof**: From Eq. (1) and (2), it is easy to see that at state $\overline{a_1} \neq \overline{a*}$, for $a_2 \neq a^*$, the suboptimality gap is

$$\Delta_h\left(\overline{a_1}, a_2\right) := V_h^*\left(\overline{a_1}\right) - Q_h^*\left(\overline{a_1}, a_2\right) \geqslant \alpha - \max\left\{3\alpha^2, \frac{3}{4}\alpha\right\} = \frac{1}{24}.$$

Moreover, at state $\overline{a*}$, for $a \neq a^*$, the suboptimality gap is

$$\Delta_h\left(\overline{a*}, a\right) := V_h^*\left(\overline{a*}\right) - Q_h^*\left(\overline{a*}, a\right) \geqslant \frac{3}{4}\alpha - 3\alpha^2 = \frac{1}{24}$$

Finally, for the terminal state $f$, the suboptimality gap is obviously 1 . Therefore $\Delta_{\min} \geqslant \frac{1}{24}$ for all MDPs under consideration.

# Proof Sketch of Theorem 5.12

- The feature map of $\mathcal{M}_{a*}$ does not depend on $a^*$.
- For $h < H - 1$ and $a_2 \neq a^*$, the reward $r_h\left(\overline{a_1}, a_2\right)$ contains no information about $a^*$.
- The transition probabilities are also independent of $a^*$, unless the action $a^*$ is taken.
- The reward at state $f$ is always 0 .
- **Thus, to receive information about $a^*$, the agent either needs to take the action $a^*$, or be at a non-game-over state at the final time step.**
- However, note that the probability of remaining at a non-terminal state at the next layer is at most

$$\sup_{a_1 \neq a_2} \left\langle v\left(a_1\right), v\left(a_2\right)\right\rangle + 2\alpha \leqslant 3\alpha \leqslant \frac{3}{4}.$$

Thus, for any algorithm, $\Pr\left[s_{H-1} \neq f\right] \leqslant \left(\frac{3}{4}\right)^H$, which is exponentially small.

# Proof Sketch of Theorem 5.12

- In other words, any algorithm that does not know $a^*$ either needs to "be lucky" so that $s_{H-1} \neq f$, or needs to take $a^*$ "by accident".
- Since the number of actions is $m = 2^{\Theta(d)}$, either event cannot happen with constant probability unless the number of episodes is exponential in $\min\{d, H\}$.