Learning with Linear Bellman Completion & Generative Model

CS 6789: Foundations of Reinforcement Learning

Announcements

1. HW1 is going to be out Thursday.

2. Wen's office hour: after lectures

Recap: Generative model + Tabular

1. Generative model assumption:

At any (s, a), we can sample $s' \sim P(\cdot \mid s, a)$

Recap: Generative model + Tabular

1. Generative model assumption:

At any (s, a), we can sample $s' \sim P(\cdot \mid s, a)$

Q: why this could be a strong assumption in practice?

Recap: Generative model + Tabular

Algorithm:

1. For each (s, a), i.i.d sample N next states, $s_i' \sim P(\cdot \mid s, a)$

Recap: Generative model + Tabular

Algorithm:

1. For each (s, a), i.i.d sample N next states, $s_i' \sim P(\cdot \mid s, a)$

2. For each (s, a, s'), construct $\hat{P}(s' \mid s, a) = \dfrac{\sum_{i=1}^{N} \mathbf{1}(s_i' = s')}{N}$

Recap: Generative model + Tabular

Algorithm:

1. For each (s, a), i.i.d sample N next states, $s'_i \sim P(\cdot \mid s, a)$

2. For each (s, a, s'), construct $\hat{P}(s' \mid s, a) = \frac{\sum_{i=1}^{N} \mathbf{1}(s'_i = s')}{N}$

3. Find optimal policy under $\hat{P}$, i.e., $\hat{\pi}^\star = \mathrm{PI}(\hat{P}, r)$

Recap: Generative model + Tabular

Result:

When $N \geq \frac{\ln(SA/\delta)}{\epsilon^2(1-\gamma)^6}$, then w/ prob $1-\delta$, we will learn a $\hat{\pi}^\star$, such that $\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \leq \epsilon$

Remarks:

Recap: Generative model + Tabular

Result:

When $N \geq \frac{\ln(SA/\delta)}{\epsilon^2(1-\gamma)^6}$, then w/ prob $1-\delta$, we will learn a $\hat{\pi}^\star$, such that $\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \leq \epsilon$

Remarks:

1. Horizon factor is not tight at all (Ch2 in AJKS optimizes it to $1/(1-\gamma)^5$)

Recap: Generative model + Tabular

Result:

When $N \geq \frac{\ln(SA/\delta)}{\epsilon^2(1-\gamma)^6}$, then w/ prob $1-\delta$, we will learn a $\hat{\pi}^\star$, such that $\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \leq \epsilon$

Remarks:

1. Horizon factor is not tight at all (Ch2 in AJKS optimizes it to $1/(1-\gamma)^5$)

2. Remarkably, our learned model $\hat{P}$ in this case is not necessarily accurate at all

Today: Generative model + linear function approximation

Key question: what happens when state-action space is large or even continuous?

Outline:

1. The Linear Bellman Completion Condition

2. The Least Square Value Iteration Algorithm

3. Guarantee and the proof sketch

Finite Horizon MDPs and DP

$$\mathcal{M} = \{S, A, P_h, r, H\}$$

$$P_h : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1]$$

Compute $\pi^\star$ via DP (backward in time):

Finite Horizon MDPs and DP

$$\mathscr{M} = \{S, A, P_h, r, H\}$$

$$P_h : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1]$$

Compute $\pi^\star$ via DP (backward in time):

1. set $Q^\star_{H-1}(s,a) = r(s,a)$, $\pi^\star_{H-1}(s) = \arg\max_a Q^\star_{H-1}(s,a)$, $V^\star_{H-1}(s) = \max_a Q^\star_{H-1}(s,a)$

Finite Horizon MDPs and DP

$$\mathscr{M} = \{S, A, P_h, r, H\}$$

$$P_h : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1]$$

Compute $\pi^\star$ via DP (backward in time):

1. set $Q^\star_{H-1}(s,a) = r(s,a)$, $\pi^\star_{H-1}(s) = \arg\max_a Q^\star_{H-1}(s,a)$, $V^\star_{H-1}(s) = \max_a Q^\star_{H-1}(s,a)$

2. At h, set $Q^\star_h(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V^\star_{h+1}(s')$,
$\pi^\star_h(s) = \arg\max_a Q^\star_h(s,a)$, $V^\star_h(s) = \max_a Q^\star_h(s,a)$

Recall Error amplification

1. **Bellman optimality:** $\|Q - \mathcal{T} Q\|_\infty = 0$, then $Q = Q^\star$

Recall Error amplification

1. **Bellman optimality:** $\|Q - \mathcal{T}Q\|_\infty = 0$, then $Q = Q^\star$

2. If nearly Bellman-consistent, i.e., $\|Q - \mathcal{T}Q\|_\infty \le \epsilon$,

Recall Error amplification

1. **Bellman optimality:** $\|Q - \mathcal{T}Q\|_\infty = 0$, then $Q = Q^\star$

2. If nearly Bellman-consistent, i.e., $\|Q - \mathcal{T}Q\|_\infty \leq \epsilon$,

Then we have error amplification:

$\|Q - Q^\star\|_\infty \leq \epsilon/(1-\gamma)$, => $V^\star - V^{\hat{\pi}} \leq \epsilon/(1-\gamma)^2$

Recall Error amplification

1. **Bellman optimality:** $\|Q - \mathcal{T}Q\|_\infty = 0$, then $Q = Q^\star$

2. If nearly Bellman-consistent, i.e., $\|Q - \mathcal{T}Q\|_\infty \le \epsilon$,

Then we have error amplification:

$\|Q - Q^\star\|_\infty \le \epsilon/(1-\gamma)$, => $V^\star - V^{\hat{\pi}} \le \epsilon/(1-\gamma)^2$

Similar results hold in finite horizon, with the effective horizon $1/(1-\gamma)$ being replaced by H

Linear Bellman Completion

Given feature ϕ, take any linear function $w^\top \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t.., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

Linear Bellman Completion

Given feature ϕ, take any linear function $w^\top \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

This is a function of (s,a), and it's linear in $\phi(s,a)$

Linear Bellman Completion

Given feature ϕ, take any linear function $w^\top \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

This is a function of (s,a), and it's linear in $\phi(s,a)$

Notation: we will denote such $\theta := \mathcal{T}_h(w)$, where $\mathcal{T}_h : \mathbb{R}^d \mapsto \mathbb{R}^d$

What does Linear Bellman completion imply

Given feature ϕ, take any linear function $w^\top \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

What does Linear Bellman completion imply

Given feature ϕ, take any linear function $w^\top \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

It implies that $Q_h^\star$ is linear in ϕ:

$$Q_h^\star = (\theta^\star)^\top \phi, \forall h$$

Why?

What does Linear Bellman completion imply

Given feature ϕ, take any linear function $w^\top \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s', a'), \forall s, a$$

It implies that $Q_h^\star$ is linear in ϕ:

$$Q_h^\star = (\theta^\star)^\top \phi, \forall h$$

Why?

reward $r(s,a)$ is linear in ϕ, i.e., $Q_{H-1}^\star(s,a)$ is linear,

now recursively show that $Q_h^\star$ is linear

Why this is a reasonable assumption?

It captures at least two special cases: tabular MDP and linear dynamical systems

1. Tabular MDP:

Set $\phi(s, a)$ to be a one-hot encoding vector in $\mathbb{R}^{SA}$, i.e., $\phi(s, a) = [0,\ldots,0,1,0,\ldots 0]^\top$

Why this is a reasonable assumption?

It captures at least two special cases: tabular MDP and linear dynamical systems

1. Tabular MDP:

Set $\phi(s, a)$ to be a one-hot encoding vector in $\mathbb{R}^{SA}$, i.e., $\phi(s, a) = [0, \ldots, 0, 1, 0, \ldots 0]^\top$

2. Linear System with Quadratic feature ϕ

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s, a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

Why this is a reasonable assumption?

It captures at least two special cases: tabular MDP and linear dynamical systems

1. Tabular MDP:

Set $\phi(s,a)$ to be a one-hot encoding vector in $\mathbb{R}^{SA}$, i.e., $\phi(s,a) = [0,\ldots,0,1,0,\ldots 0]^\top$

2. Linear System with Quadratic feature ϕ

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s,a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

$$\phi(s,a) = [s_1, s_2, s_1^2, s_2^2, s_1 s_2, s_1 a, s_2 a, a, a^2, 1]^\top$$

Why this is a reasonable assumption?

It captures at least two special cases: tabular MDP and linear dynamical systems

1. Tabular MDP:

Set $\phi(s, a)$ to be a one-hot encoding vector in $\mathbb{R}^{SA}$, i.e., $\phi(s, a) = [0,\ldots,0,1,0,\ldots0]^{\top}$

2. Linear System with Quadratic feature ϕ

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s, a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

$$\phi(s, a) = [s_1, s_2, s_1^2, s_2^2, s_1 s_2, s_1 a, s_2 a, a, a^2, 1]^{\top}$$

Claim: $r(s, a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} w^T \phi(s', a')$ is a linear function in ϕ

Why this is a strong assumption?

Assume the given feature ϕ has linear Bellman completion, i.e.,

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

Why this is a strong assumption?

Assume the given feature ϕ has linear Bellman completion, i.e.,

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

Adding additional elements to ϕ can break the condition!

Why this is a strong assumption?

Assume the given feature ϕ has linear Bellman completion, i.e.,

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s',a'), \forall s, a$$

Adding additional elements to ϕ can break the condition!

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s, a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

Why this is a strong assumption?

Assume the given feature ϕ has linear Bellman completion, i.e.,

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s', a'), \forall s, a$$

Adding additional elements to ϕ can break the condition!

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s,a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

$$\phi(s,a) = [s_1, s_2, s_1^2, s_2^2, s_1 s_2, s_1 a, s_2 a, a, a^2, 1, \mathbf{s_1^3}]^\top$$

Why this is a strong assumption?

Assume the given feature ϕ has linear Bellman completion, i.e.,

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s', a'), \forall s, a$$

Adding additional elements to ϕ can break the condition!

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s, a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

$$\phi(s,a) = [s_1, s_2, s_1^2, s_2^2, s_1 s_2, s_1 a, s_2 a, a, a^2, 1, \mathbf{s_1^3}]^\top$$

Linear Bellman completion breaks!

Why this is a strong assumption?

Assume the given feature ϕ has linear Bellman completion, i.e.,

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} w^\top \phi(s', a'), \forall s, a$$

Adding additional elements to ϕ can break the condition!

$$s \in \mathbb{R}^2, a \in \mathbb{R}, P_h(\cdot \mid s, a) = \mathcal{N}\left(As + ba, \sigma^2 I\right)$$

$$\phi(s,a) = [s_1, s_2, s_1^2, s_2^2, s_1 s_2, s_1 a, s_2 a, a, a^2, 1, \mathbf{s_1^3}]^\top$$

Linear Bellman completion breaks!

This is counter-intuitive: in SL (e.g., linear regression), adding elements to features is ok!

Can we just assume $Q^\star$ being linear?

No! There are lower bounds (even under generative model):

Can we just assume $Q^\star$ being linear?

No! There are lower bounds (even under generative model):

For any RL algorithm, there exist MDPs with $Q^\star_h(s, a)$ is linear in $\phi(s, a)$ (known), such that in order to find a policy π with $V^\pi(s_1) \geq V^\star(s_1) - 0.05$, it requires at least $\min\{2^d, 2^H\}$ many samples!

Can we just assume $Q^\star$ being linear?

No! There are lower bounds (even under generative model):

For any RL algorithm, there exist MDPs with $Q_h^\star(s, a)$ is linear in $\phi(s, a)$ (known), such that in order to find a policy π with $V^\pi(s_1) \geq V^\star(s_1) - 0.05$, it requires at least $\min\{2^d, 2^H\}$ many samples!

i.e., polynomial bound poly(d, H) is not possible for linear $Q^\star$ (Ch5 AJKS)

What we will show today:

1. Generative Model

(i.e., we can reset system to any (s, a), query $r(s, a), s' \sim P(\,.\,|\,s, a))$

$$+$$

2. Linear Bellman Completion

$$=$$

Sample efficient Learning
(poly time)

Outline:

1. The Linear Bellman Completion Condition

2. Learning: The Least Square Value Iteration Algorithm

3. Guarantee and the proof sketch

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

Given datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/
$\mathcal{D}_h = \{s, a, r, s'\}, r = r(s,a), s' \sim P_h(\cdot \mid s, a)$

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

Given datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/
$\mathcal{D}_h = \{s, a, r, s'\}, r = r(s,a), s' \sim P_h(\cdot \mid s, a)$

Let's simulate the DP process w/
linear function to approximate $Q^\star$

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

Set $V_H(s) = 0, \forall s$

Given datasets $\mathscr{D}_0, \ldots, \mathscr{D}_{H-1}$, w/
$\mathscr{D}_h = \{s, a, r, s'\}, r = r(s,a), s' \sim P_h(\cdot \mid s, a)$

Let's simulate the DP process w/
linear function to approximate $Q^\star$

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_\theta \sum_{\mathcal{D}_h} \left(\theta^T \phi(s,a) - \left(r + V_{h+1}(s')\right)\right)^2$$

Given datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/

$\mathcal{D}_h = \{s, a, r, s'\}, r = r(s,a), s' \sim P_h(\cdot \mid s, a)$

Let's simulate the DP process w/ linear function to approximate $Q^\star$

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_\theta \sum_{\mathcal{D}_h} \left(\theta^T \phi(s,a) - \left(r + V_{h+1}(s')\right)\right)^2$$

Set $V_h(s) := \max_a \theta_h^\top \phi(s,a), \forall s$

Given datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ $\mathcal{D}_h = \{s, a, r, s'\}, r = r(s,a), s' \sim P_h(\cdot \mid s, a)$

Let's simulate the DP process w/ linear function to approximate $Q^\star$

LSVI: Least-Square Value Iteration

Recall linear bellman-completion implies $Q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a), \forall s, a, h$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_\theta \sum_{\mathcal{D}_h} \left(\theta^T \phi(s,a) - \left(r + V_{h+1}(s')\right)\right)^2$$

Set $V_h(s) := \max_a \theta_h^\top \phi(s,a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^\top \phi(s,a), \forall h$

Given datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/

$\mathcal{D}_h = \{s, a, r, s'\}, r = r(s,a), s' \sim P_h(\cdot \mid s, a)$

Let's simulate the DP process w/ linear function to approximate $Q^\star$

Why LSVI may work?

When we do linear regression at step h:

$$x := \phi(s, a), \quad y := r + V_{h+1}(s')$$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_{\theta} \sum_{\mathcal{D}_h} \left(\theta^T \phi(s, a) - \left(r + V_{h+1}(s')\right) \right)^2$$

Set $V_h(s) := \max_a \theta_h^\top \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^\top \phi(s, a), \forall h$

Why LSVI may work?

When we do linear regression at step h:

$$x := \phi(s, a), \quad y := r + V_{h+1}(s')$$

We note that:

$$\mathbb{E}[y \mid x] = \underbrace{r(s, a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta_{h+1}^{\top} \phi(s', a')}_{\mathcal{T}_h(\theta_{h+1})^{\top}\phi(s,a) \text{ due to Linear BC}}$$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_{\theta} \sum_{\mathcal{D}_h} \left(\theta^T \phi(s, a) - \left(r + V_{h+1}(s') \right) \right)^2$$

Set $V_h(s) := \max_a \theta_h^{\top} \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^{\top} \phi(s, a), \forall h$

Why LSVI may work?

When we do linear regression at step h:

$$x := \phi(s, a), \quad y := r + V_{h+1}(s')$$

We note that:

$$\mathbb{E}[y \mid x] = \underbrace{r(s, a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta_{h+1}^\top \phi(s', a')}_{\mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) \text{ due to Linear BC}}$$

i.e., our regression target is indeed linear in ϕ, and it is close to $Q_h^\star$ if

$$V_{h+1} \approx V_{h+1}^\star$$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_\theta \sum_{\mathcal{D}_h} \left(\theta^T \phi(s, a) - \left(r + V_{h+1}(s') \right) \right)^2$$

Set $V_h(s) := \max_a \theta_h^\top \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^\top \phi(s, a), \forall h$

Why LSVI may work?

When we do linear regression at step h:

$$x := \phi(s,a), \quad y := r + V_{h+1}(s')$$

We note that:

$$\mathbb{E}[y \mid x] = \underbrace{r(s,a) + \mathbb{E}_{s'\sim P_h(s,a)} \max_{a'} \theta_{h+1}^\top \phi(s',a')}_{\mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) \text{ due to Linear BC}}$$

i.e., our regression target is indeed linear in ϕ, and it is close to $Q_h^\star$ if $V_{h+1} \approx V_{h+1}^\star$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_\theta \sum_{\mathcal{D}_h} \left(\theta^T \phi(s,a) - \left(r + V_{h+1}(s') \right) \right)^2$$

Set $V_h(s) := \max_a \theta_h^\top \phi(s,a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^\top \phi(s,a), \forall h$

If $V_{h+1} \approx V_{h+1}^\star$, and linear regression succeeds (e.g., $\theta_h \approx \mathcal{T}_h(\theta_{h+1})$),

Why LSVI may work?

When we do linear regression at step h:

$$x := \phi(s,a), \quad y := r + V_{h+1}(s')$$

We note that:

$$\mathbb{E}[y \mid x] = \underbrace{r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta_{h+1}^\top \phi(s',a')}_{\mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) \text{ due to Linear BC}}$$

i.e., our regression target is indeed linear in ϕ, and it is close to $Q_h^\star$ if $V_{h+1} \approx V_{h+1}^\star$

Set $V_H(s) = 0, \forall s$

For h = H-1 to 0:

$$\theta_h = \arg\min_\theta \sum_{\mathcal{D}_h} \left(\theta^T \phi(s,a) - \left(r + V_{h+1}(s') \right) \right)^2$$

Set $V_h(s) := \max_a \theta_h^\top \phi(s,a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^\top \phi(s,a), \forall h$

If $V_{h+1} \approx V_{h+1}^\star$, and linear regression succeeds (e.g., $\theta_h \approx \mathcal{T}_h(\theta_{h+1})$),

Then we should hope $\theta_h^\top \phi(s,a) \approx Q_h^\star(s,a)$

Outline:

1. The Linear Bellman Completion Condition

2. Learning: The Least Square Value Iteration Algorithm

3. Guarantee and the proof sketch

Sample complexity of LSVI

Theorem: There exists a way to construct datasets $\{\mathscr{D}_h\}_{h=0}^{H-1}$, such that with probability at least $1-\delta$, we have:

$$V^{\hat{\pi}} - V^{\star} \leq \epsilon$$

w/ total number of samples in these datasets scaling $\widetilde{O}\left(d^2 + H^6 d^2/\epsilon^2\right)$

Sample complexity of LSVI

Theorem: There exists a way to construct datasets $\{\mathcal{D}_h\}_{h=0}^{H-1}$, such that with probability at least $1-\delta$, we have:

$$V^{\hat{\pi}} - V^\star \leq \epsilon$$

w/ total number of samples in these datasets scaling $\widetilde{O}\left(d^2 + H^6 d^2/\epsilon^2\right)$

Plans: (1) OLS and D-optimal design; (2) construct $\mathcal{D}_h$ using D-optimal design; (3) transfer regression error to $\|\theta_h^\top \phi - Q_h^\star\|_\infty$

Detour: Ordinary Linear Squares

Consider a dataset $\{x_i, y_i\}_{i=1}^{N}$, where $y_i = (\theta^\star)^\top x_i + \epsilon_i, \quad \mathbb{E}[\epsilon_i \mid x_i] = 0$, ϵ_i are independent with $|\epsilon_i| \leq \sigma$, assume $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$ is full rank;

Detour: Ordinary Linear Squares

Consider a dataset $\{x_i, y_i\}_{i=1}^N$, where $y_i = (\theta^\star)^\top x_i + \epsilon_i, \quad \mathbb{E}[\epsilon_i \,|\, x_i] = 0$, ϵ_i are independent with $|\epsilon_i| \le \sigma$, assume $\Lambda = \sum_{i=1}^N x_i x_i^\top / N$ is full rank;

$$\text{OLS} : \hat{\theta} = \arg\min_\theta \sum_{i=1}^N (\theta^\top x_i - y_i)^2$$

Detour: Ordinary Linear Squares

Consider a dataset $\{x_i, y_i\}_{i=1}^N$, where $y_i = (\theta^\star)^\top x_i + \epsilon_i, \quad \mathbb{E}[\epsilon_i | x_i] = 0$, ϵ_i are independent with $|\epsilon_i| \leq \sigma$, assume $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$ is full rank;

$$\text{OLS} : \hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} (\theta^\top x_i - y_i)^2$$

Standard OLS guarantee: with probability at least $1 - \delta$, we have:

$$(\hat{\theta} - \theta^\star)^\top \Lambda (\hat{\theta} - \theta^\star) \leq O\left(\frac{\sigma^2 d \ln(1/\delta)}{N} \right)$$

Detour: Issues in Ordinary Linear Squares

Recall $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$;

With probability at least $1 - \delta$:

$$(\hat{\theta} - \theta^\star)^\top \Lambda (\hat{\theta} - \theta^\star) \leq O\left(\frac{\sigma^2 d \ln(1/\delta)}{N}\right)$$

Detour: Issues in Ordinary Linear Squares

Recall $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$;

With probability at least $1 - \delta$:

$$(\hat{\theta} - \theta^\star)^\top \Lambda (\hat{\theta} - \theta^\star) \leq O\left(\frac{\sigma^2 d \ln(1/\delta)}{N}\right)$$

If the test point x is not covered by the training data, i.e., $x^\top \Lambda^{-1} x$ is huge, then we cannot guarantee $\hat{\theta}^\top x$ is close to $(\theta^\star)^\top x$

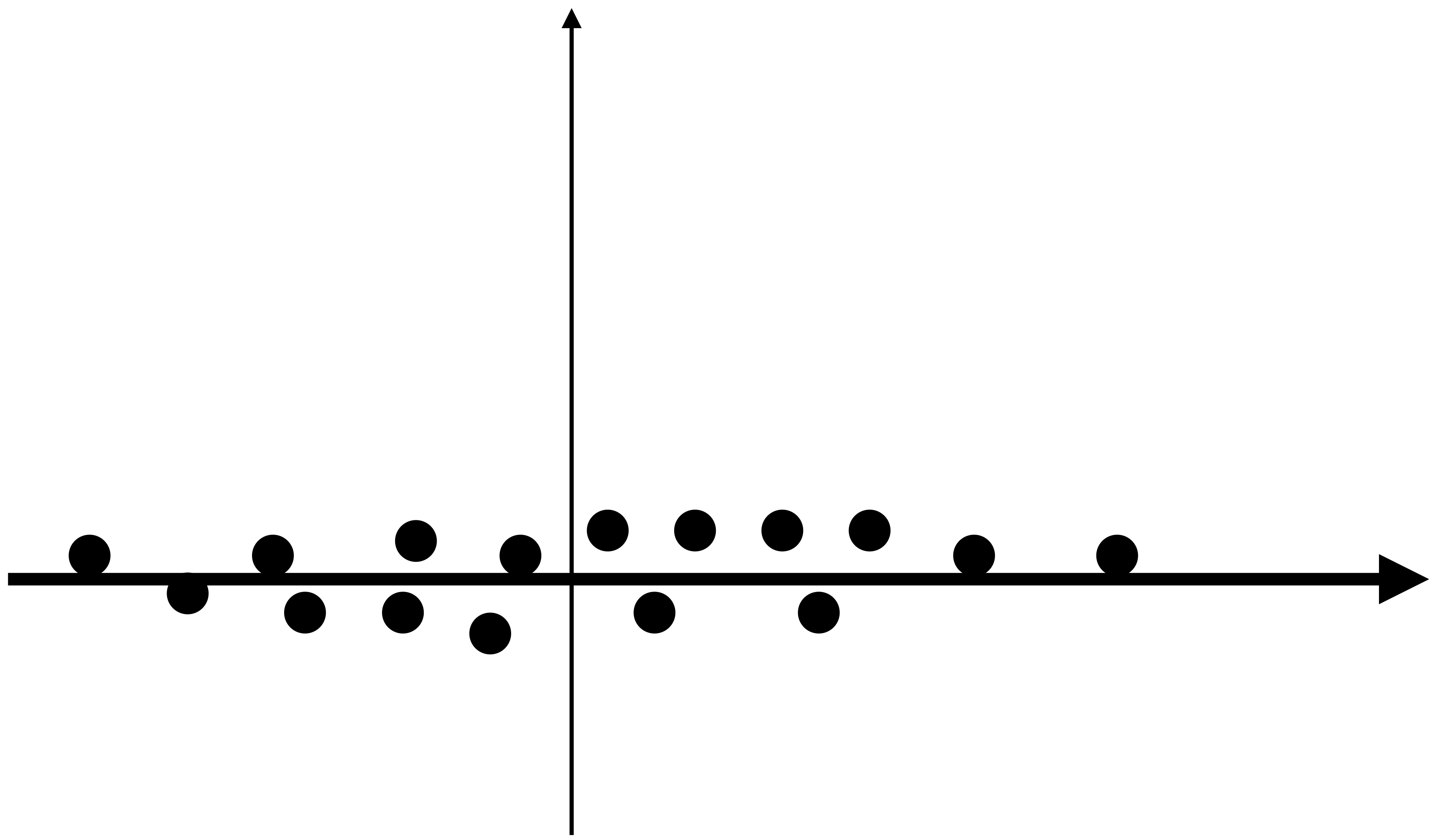
Detour: Issues in Ordinary Linear Squares

Recall $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$;

With probability at least $1 - \delta$:

$$(\hat{\theta} - \theta^\star)^\top \Lambda (\hat{\theta} - \theta^\star) \leq O\left(\frac{\sigma^2 d \ln(1/\delta)}{N}\right)$$

If the test point x is not covered by the training data, i.e., $x^\top \Lambda^{-1} x$ is huge, then we cannot guarantee $\hat{\theta}^\top x$ is close to $(\theta^\star)^\top x$



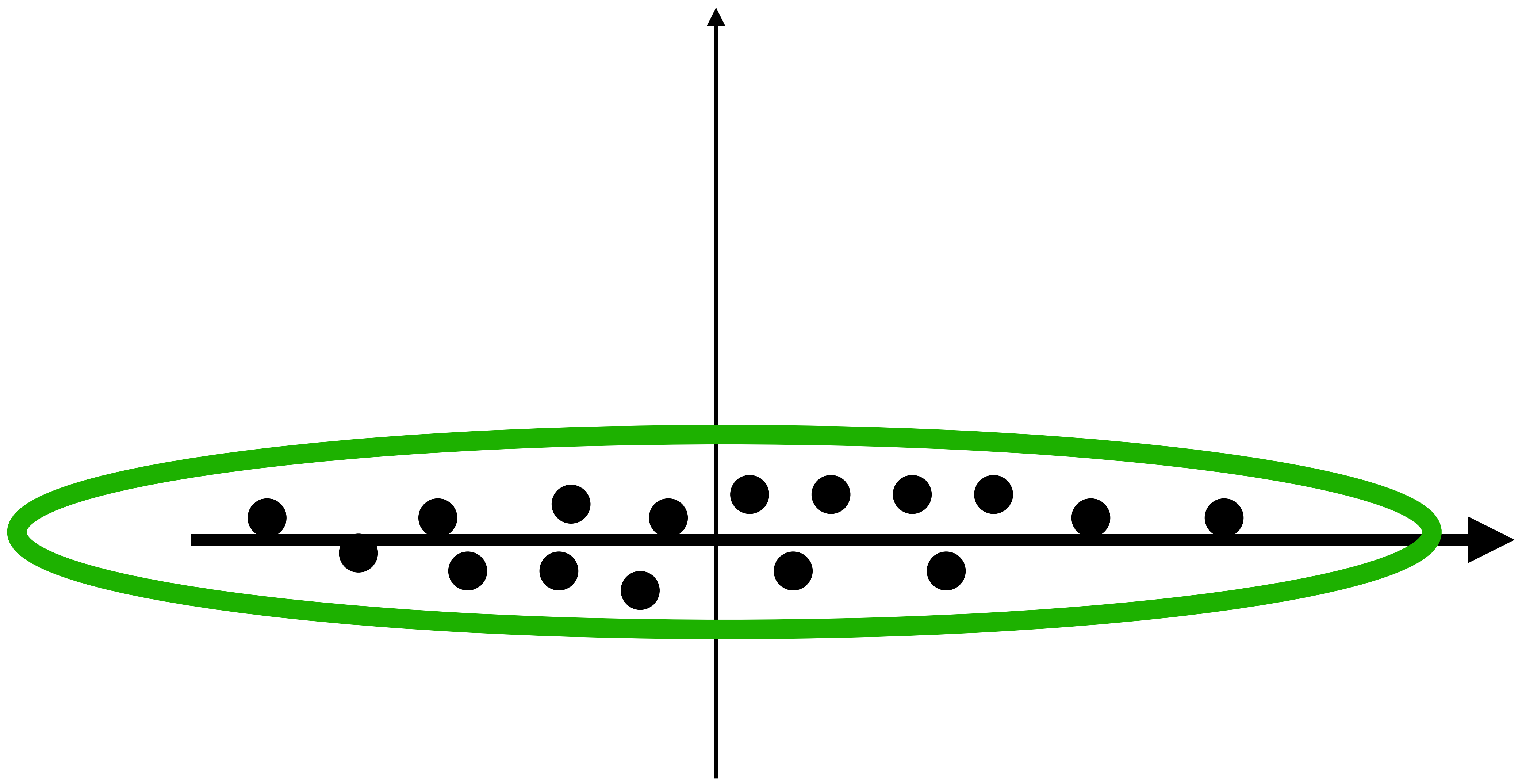
Detour: Issues in Ordinary Linear Squares

Recall $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$;

With probability at least $1 - \delta$:

$$(\hat{\theta} - \theta^\star)^\top \Lambda (\hat{\theta} - \theta^\star) \leq O\left(\frac{\sigma^2 d \ln(1/\delta)}{N}\right)$$

If the test point x is not covered by the training data, i.e., $x^\top \Lambda^{-1} x$ is huge, then we cannot guarantee $\hat{\theta}^\top x$ is close to $(\theta^\star)^\top x$



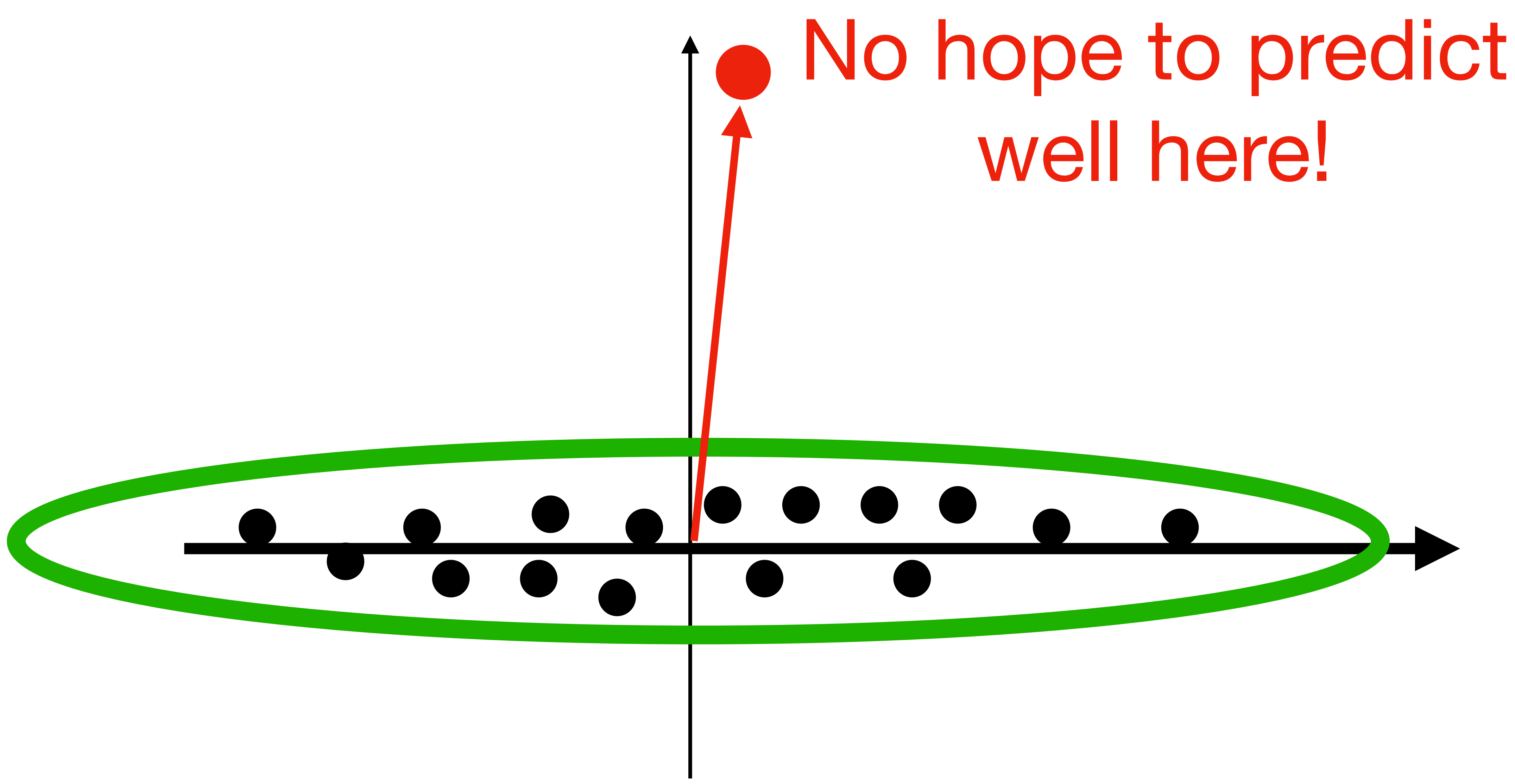
Detour: Issues in Ordinary Linear Squares

Recall $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$;

With probability at least $1 - \delta$:

$$(\hat{\theta} - \theta^\star)^\top \Lambda (\hat{\theta} - \theta^\star) \leq O\left(\frac{\sigma^2 d \ln(1/\delta)}{N}\right)$$

If the test point x is not covered by the training data, i.e., $x^\top \Lambda^{-1} x$ is huge, then we cannot guarantee $\hat{\theta}^\top x$ is close to $(\theta^\star)^\top x$



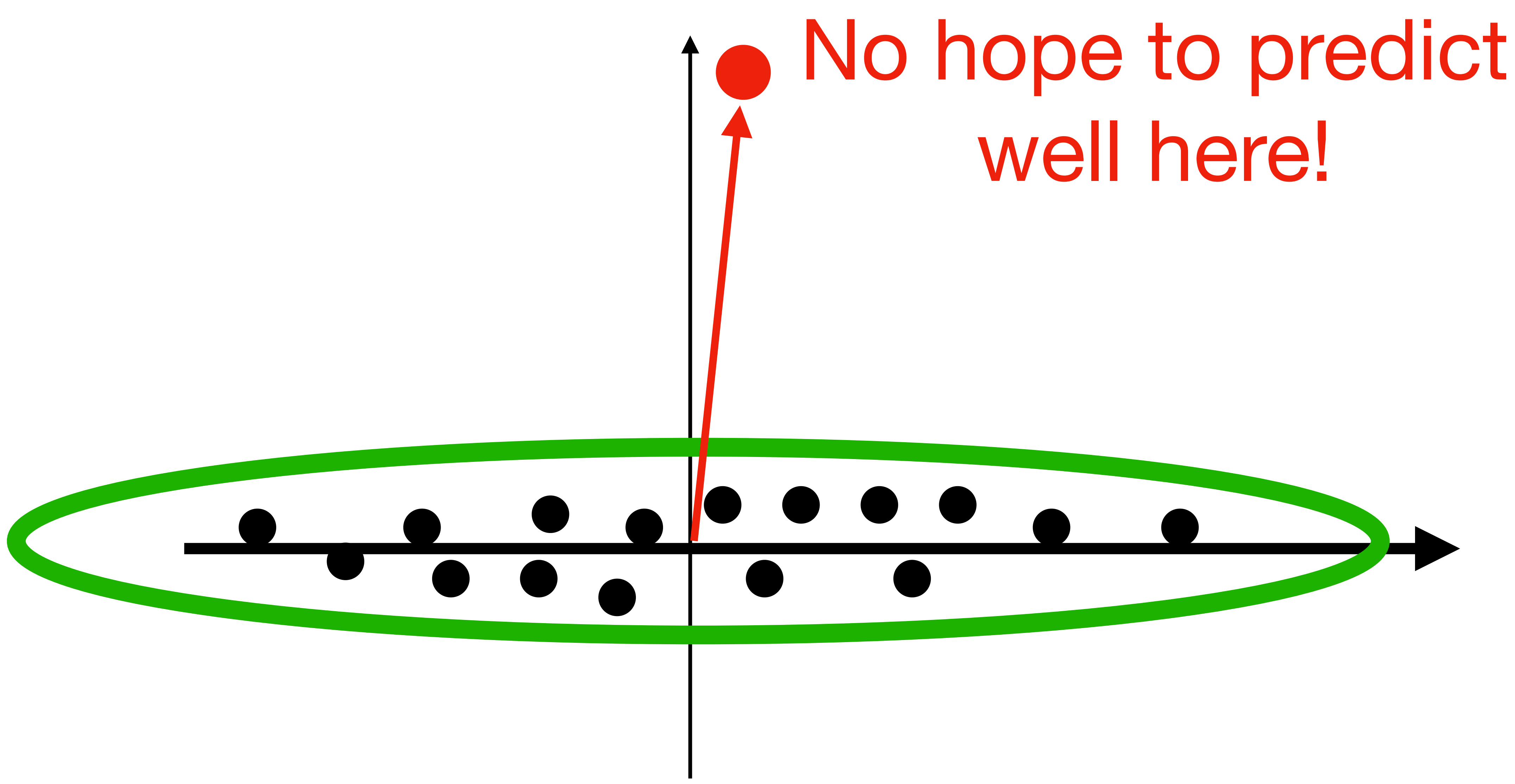
Detour: Issues in Ordinary Linear Squares

Recall $\Lambda = \sum_{i=1}^{N} x_i x_i^\top / N$;

With probability at least $1 - \delta$:

$$(\hat\theta - \theta^\star)^\top \Lambda (\hat\theta - \theta^\star) \le O\left(\frac{\sigma^2 d \ln(1/\delta)}{N}\right)$$

If the test point x is not covered by the training data, i.e., $x^\top \Lambda^{-1} x$ is huge, then we cannot guarantee $\hat\theta^\top x$ is close to $(\theta^\star)^\top x$



Let's actively design a diverse dataset !
(D-optimal Design)

Detour: D-optimal Design

Consider a compact space $\mathscr{X} \subset \mathbb{R}^d$ (without loss of generality, assume span($\mathscr{X}$) $= \mathbb{R}^d$)

Detour: D-optimal Design

Consider a compact space $\mathscr{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathscr{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathscr{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathscr{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

Detour: D-optimal Design

Consider a compact space $\mathscr{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathscr{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathscr{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathscr{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

Properties of the D-optimal Design:

support$(\rho^\star) \leq d(d+1)/2$

Detour: D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

Properties of the D-optimal Design:

support$(\rho^\star) \leq d(d+1)/2$

$$\max_{y \in \mathcal{X}} y^\top \left[\mathbb{E}_{x\sim\rho^\star} xx^\top\right]^{-1} y \leq d$$

Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

We **actively** construct a dataset $\mathcal{D}$, which contains $\lceil \rho(x)N \rceil$ many copies of x

Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

We **actively** construct a dataset $\mathcal{D}$, which contains $\lceil \rho(x) N \rceil$ many copies of x

For each $x \in \mathcal{D}$, query y (noisy measure);

Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume span$(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

We **actively** construct a dataset $\mathcal{D}$, which contains $\lceil \rho(x) N \rceil$ many copies of x

For each $x \in \mathcal{D}$, query y (noisy measure);

The OLS solution $\hat\theta$ on $\mathcal{D}$ has the following point-wise guarantee: w/ prob $1-\delta$

$$\max_{x\in\mathcal{X}} \left| \langle \hat\theta - \theta^\star, x\rangle \right| \leq \frac{\sigma d \ln(1/\delta)}{\sqrt{N}}$$

Summary so far on OLS & D-optimal Design

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

Summary so far on OLS & D-optimal Design

D-optimal Design $\rho^\star \in \Delta(\mathcal{X})$**:** $\rho^\star = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln\det\left(\mathbb{E}_{x\sim\rho}\left[xx^\top\right]\right)$

D-optimal design allows us to **actively** construct a dataset $\mathcal{D} = \{x, y\}$, such that OLS solution is **POINT-WISE** accurate:

$$\max_{x \in \mathcal{X}} \left| \langle \hat{\theta} - \theta^\star, x \rangle \right| \leq \frac{\sigma d \ln(1/\delta)}{\sqrt{N}}$$