Announcements

- **Reading week next week**
  - *No office hours with Adam. I am taking vacation*
- **Midterm when you come back from reading week**
  - practice midterm on eClass

Worksheet day

Assume we are given a fixed policy π. Recall that the mean-squared value error is

Q1

$$\overline{\mathrm{VE}}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s)(v_\pi(s) - \hat{v}(s, \mathbf{w}))^2.$$

Recall that we can use TD with linear function approximation to learn parameters $\mathbf{w}$:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha[R_{t+1} + \gamma \mathbf{w}^\top \mathbf{x}(S_{t+1}) - \mathbf{w}^\top \mathbf{x}(S_t)]\mathbf{x}(S_t).$$

When linear TD converges, it converges to what we call the TD fixed-point. Let's denote the weight vector found by linear TD at convergence as $\mathbf{w}_{\mathrm{TD}}$. We denote the estimated value as $\hat{v}(s, \mathbf{w}_{\mathrm{TD}}) = \mathbf{w}_{\mathrm{TD}}^\top \mathbf{x}(s)$. At the TD fixed point, we know that the VE is within a bounded expansion of the lowest possible error:

$$\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$$

(a) Recall that $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$ is the minimal value error you can achieve under this value function parameterization. If we have a tabular parameterization, then what might $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$ be? What if the parameterization is a state aggregation?

Assume we are given a fixed policy π. Recall that the mean-squared value error is

Q1

$$\overline{\mathrm{VE}}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s)(v_\pi(s) - \hat{v}(s, \mathbf{w}))^2.$$

Recall that we can use TD with linear function approximation to learn parameters $\mathbf{w}$:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha[R_{t+1} + \gamma \mathbf{w}^\top \mathbf{x}(S_{t+1}) - \mathbf{w}^\top \mathbf{x}(S_t)]\mathbf{x}(S_t).$$

When linear TD converges, it converges to what we call the TD fixed-point. Let's denote the weight vector found by linear TD at convergence as $\mathbf{w}_{\mathrm{TD}}$. We denote the estimated value as $\hat{v}(s, \mathbf{w}_{\mathrm{TD}}) = \mathbf{w}_{\mathrm{TD}}^\top \mathbf{x}(s)$. At the TD fixed point, we know that the VE is within a bounded expansion of the lowest possible error:

$$\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$$

(b) If $\gamma = 0.9$ and $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 0$, then what is the minimum and maximum value of $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}})$?

Assume we are given a fixed policy π. Recall that the mean-squared value error is

Q1

$$\overline{\mathrm{VE}}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s)(v_\pi(s) - \hat{v}(s, \mathbf{w}))^2.$$

Recall that we can use TD with linear function approximation to learn parameters $\mathbf{w}$:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha[R_{t+1} + \gamma \mathbf{w}^\top \mathbf{x}(S_{t+1}) - \mathbf{w}^\top \mathbf{x}(S_t)]\mathbf{x}(S_t).$$

When linear TD converges, it converges to what we call the TD fixed-point. Let's denote the weight vector found by linear TD at convergence as $\mathbf{w}_{\mathrm{TD}}$. We denote the estimated value as $\hat{v}(s, \mathbf{w}_{\mathrm{TD}}) = \mathbf{w}_{\mathrm{TD}}^\top \mathbf{x}(s)$. At the TD fixed point, we know that the VE is within a bounded expansion of the lowest possible error:

$$\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$$

(c) If $\gamma = 0.9$ and $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 1$, then what is the minimum and maximum value of $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}})$?

(d) If $\gamma = 0.99$ and $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 1$, then what is the minimum and maximum value of $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}})$?

Assume we are given a fixed policy π. Recall that the mean-squared value error is

Q1

$$\overline{\mathrm{VE}}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s)(v_\pi(s) - \hat{v}(s, \mathbf{w}))^2.$$

Recall that we can use TD with linear function approximation to learn parameters $\mathbf{w}$:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha[R_{t+1} + \gamma \mathbf{w}^\top \mathbf{x}(S_{t+1}) - \mathbf{w}^\top \mathbf{x}(S_t)]\mathbf{x}(S_t).$$

When linear TD converges, it converges to what we call the TD fixed-point. Let's denote the weight vector found by linear TD at convergence as $\mathbf{w}_{\mathrm{TD}}$. We denote the estimated value as $\hat{v}(s, \mathbf{w}_{\mathrm{TD}}) = \mathbf{w}_{\mathrm{TD}}^\top \mathbf{x}(s)$. At the TD fixed point, we know that the VE is within a bounded expansion of the lowest possible error:

$$\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$$

(e) We have seen that if we can perfectly represent the value function, then $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 0$. How about the other direction: if $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 0$, then does that mean we can represent true value function? Consider **two cases**: (1) where $\mu(s)$ can be zero for some states, and (2) where $\mu(s) > 0$.

Q1

Assume we are given a fixed policy π. Recall that the mean-squared value error is

$$\overline{\mathrm{VE}}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s)(v_\pi(s) - \hat{v}(s, \mathbf{w}))^2.$$

Recall that we can use TD with linear function approximation to learn parameters $\mathbf{w}$:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha[R_{t+1} + \gamma \mathbf{w}^\top \mathbf{x}(S_{t+1}) - \mathbf{w}^\top \mathbf{x}(S_t)]\mathbf{x}(S_t).$$

When linear TD converges, it converges to what we call the TD fixed-point. Let's denote the weight vector found by linear TD at convergence as $\mathbf{w}_{\mathrm{TD}}$. We denote the estimated value as $\hat{v}(s, \mathbf{w}_{\mathrm{TD}}) = \mathbf{w}_{\mathrm{TD}}^\top \mathbf{x}(s)$. At the TD fixed point, we know that the VE is within a bounded expansion of the lowest possible error:

$$\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$$

(f) **Challenge Question:** Imagine instead we obtained the asymptotic solution under the Monte carlo update; let us call this $\mathbf{w}_{\mathrm{MC}}$. This is the solution we would obtain after updating with many many pairs of states and sampled returns, with s sampled proportionally to μ. How is $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{MC}})$ and $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$ related? Hint: Try to write down the objective for Monte carlo and reason about the solution to this objective.

(f) **Challenge Question:** Imagine instead we obtained the asymptotic solution under the Monte carlo update; let us call this $\mathbf{w}_{\mathrm{MC}}$. This is the solution we would obtain after updating with many many pairs of states and sampled returns, with s sampled proportionally to μ. How is $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{MC}})$ and $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w})$ related? Hint: Try to write down the objective for Monte carlo and reason about the solution to this objective.

- Monte Carlo objective:
$$\overline{\mathrm{RE}}(\mathbf{w}) = \sum_s \mu(s)\mathbb{E}_\pi[(G_t - \hat{v}(S_t, \mathbf{w}))^2 \mid S_t = s]$$

- Take the gradient and simplify
$$\nabla_{\mathbf{w}}\overline{\mathrm{RE}}(\mathbf{w}) = \nabla_{\mathbf{w}} \sum_s \mu(s)\mathbb{E}_\pi[(G_t - \hat{v}(S_t, \mathbf{w}))^2 \mid S_t = s]$$

Q2

Challenge Question: Recall that the mean-squared value error is

$$\overline{\mathrm{VE}}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s)\left(v_\pi(s) - \hat{v}(s, \mathbf{w})\right)^2.$$

We discussed one choice for μ, which is to use the state visitation under the behavior policy. Namely, as the policy is executed in the state, the weighting $\mu(s)$ is proportional to how frequently the agent is in state s. What are some other weightings could be used instead?

Q3

(*Exercise 9.1 S&B*) Show that tabular methods such as presented in Course 2 of the MOOC (and Part I of the book) are a special case of linear function approximation. What would the feature vectors be?

Matrix cookbook, very handy:
https://www2.imm.dtu.dk/pubdb/edoc/imm3274.pdf

Q4

1. Let $f(x, y) = (x + y)^2 + e^{xy}$. Recall that the gradient is composed of the partial derivatives for each variable

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix}$$

where $\frac{\partial f(x,y)}{\partial x}$ is the derivative of $f(x, y)$ w.r.t. x assuming that y is fixed.

(a) What is $\nabla f(x, y)$ for the f defined above? Hint: Recall that the derivative of e^z is e^z.

(b) What is $\nabla f(0, 1)$?

Q5

2. Find the gradient of f

(a) if $f(x, y, z) = \frac{y^z}{x}$

(b) if $f(x) = e^{x^2+5}$

(c) if $f(\beta) = \beta^T \mathbf{x}$ where β is a vector in $\mathbb{R}^N$ and $\mathbf{x}$ is a vector of constants in $\mathbb{R}^N$

(d) if $f(\mathbf{x}) = (\mathbf{x}^T \beta - y)^2$ where $\mathbf{x}$ is a vector in $\mathbb{R}^N$, β is a vector of constants in $\mathbb{R}^N$, and y is a scalar in $\mathbb{R}$