“For even the very wise cannot see all ends.”

J.R.R. Tolkien, The Fellowship of the Ring

# CMPUT 365 Introduction to RL

Marlos C. Machado

Class 25/35

# Coursera Reminder

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks.

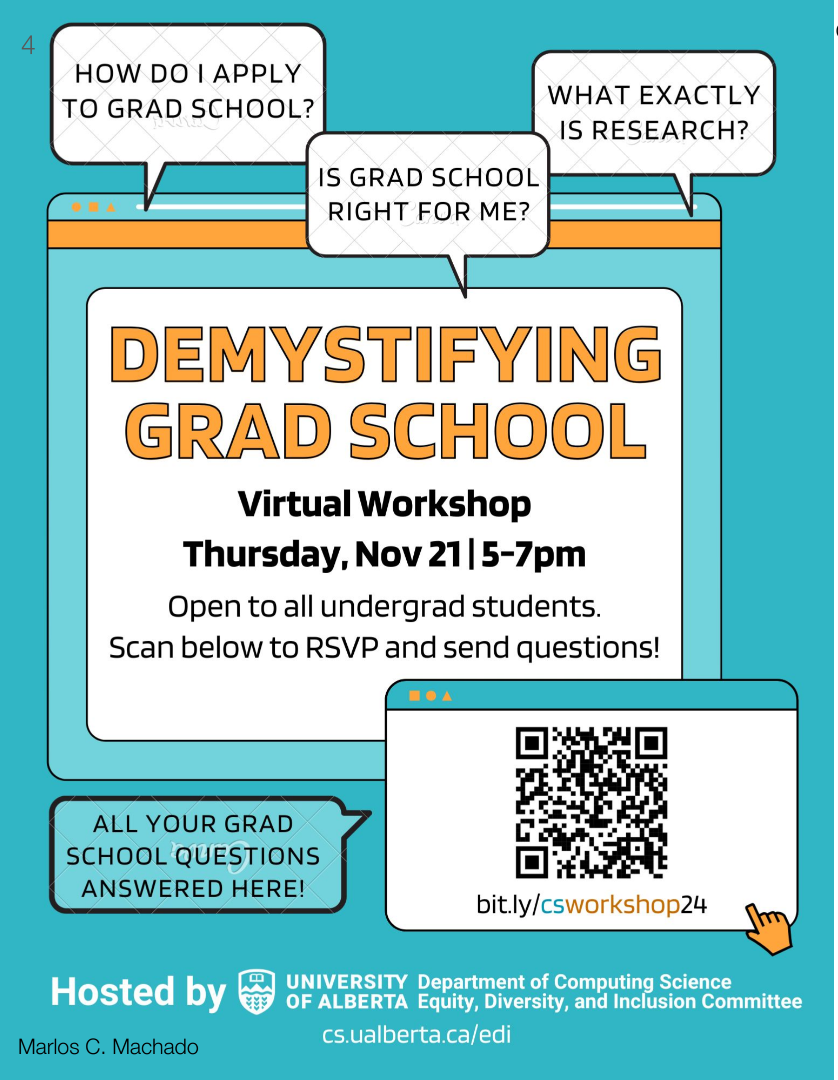
You **need** to **check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

The deadlines in the public session **do not align** with the deadlines in Coursera.

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

# Reminders and Notes

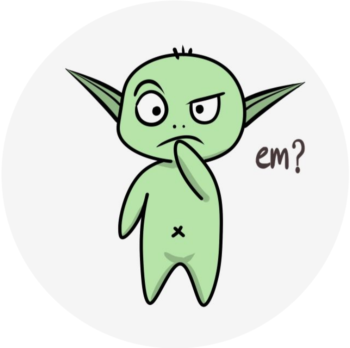
- Next week is reading week.
  - There won't be office hours next week.
- Our final exam will indeed be on December 17th, 1pm, at CCIS 1-440.




RSVP form (not required, but appreciated):
https://docs.google.com/forms/d/11odJJgO3kgJ_XFDq9v nEz4FABjlNKx7-iL6AkCJ67ZQ/edit

Direct link to the zoom:
https://ualberta-ca.zoom.us/j/93282952849?pwd=eqE7h m46hwMJS02EZoqjw5GOngtWkK.1

# Please, interrupt me at any time!





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

# Last Class: A More Realistic Update

- Let $U_t$ denote the *t*-th training example, $S_t \mapsto v_\pi(S_t)$, of some (possibly random), approximation to the true value.

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\Big[U_t - \hat{v}(S_t,\mathbf{w}_t)\Big]\nabla\hat{v}(S_t,\mathbf{w}_t)$$

**Gradient Monte Carlo Algorithm for Estimating $\hat{v} \approx v_\pi$**

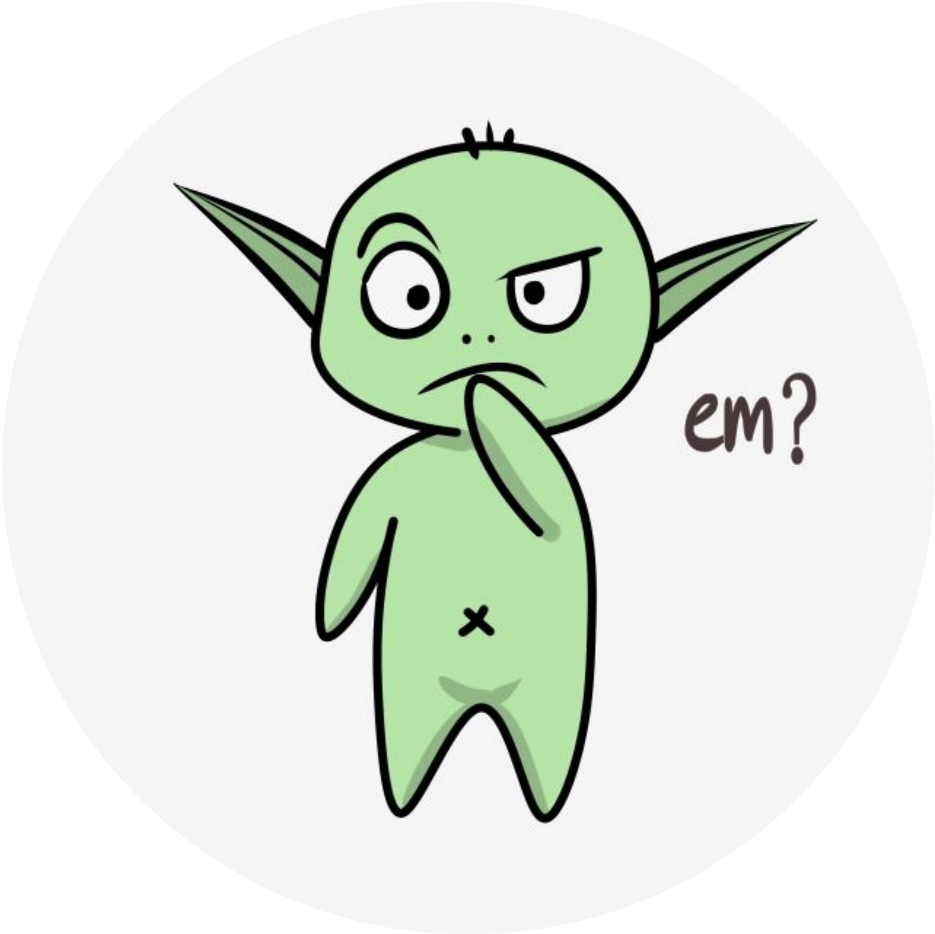
Input: the policy $\pi$ to be evaluated
Input: a differentiable function $\hat{v} : \mathcal{S} \times \mathbb{R}^d \rightarrow \mathbb{R}$
Algorithm parameter: step size $\alpha > 0$
Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, S_1, A_1, \ldots, R_T, S_T$ using $\pi$
  Loop for each step of episode, $t = 0, 1, \ldots, T-1$:
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[G_t - \hat{v}(S_t,\mathbf{w})\big]\nabla\hat{v}(S_t,\mathbf{w})$

# A Clearer Instantiation — Linear Function Approximation

- Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top\mathbf{w}$. We have $\nabla_{\mathbf{w}}\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}(s)$.
- Thus, $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\,[U_t - \hat{v}(\mathbf{x}, \mathbf{w})]\,\nabla_{\mathbf{w}}\hat{v}(\mathbf{x}, \mathbf{w})$ becomes:

  $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\,[U_t - \hat{v}(\mathbf{x}, \mathbf{w})]\mathbf{x}$.






https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

# Semi-gradient TD

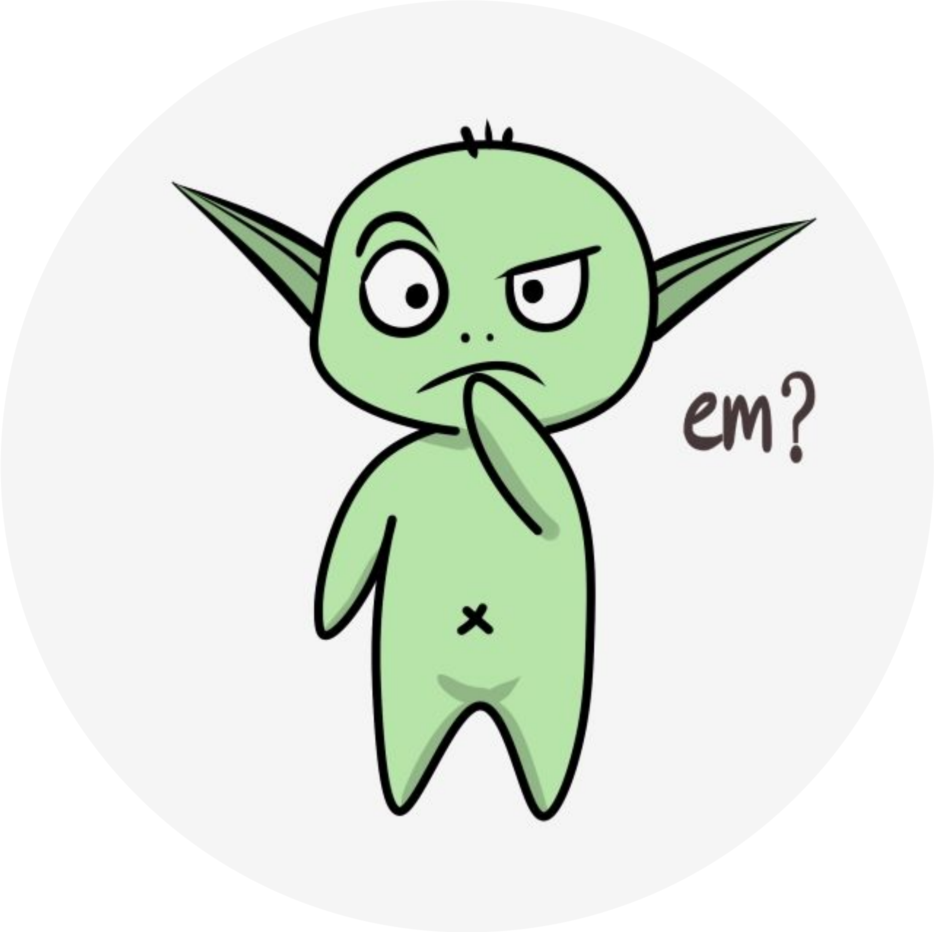
- What if $U_t \doteq R_{t+1} + \gamma \, \hat{\mathbf{v}}(S_{t+1}, \mathbf{w}_t)$?
- We lose several guarantees when we use a bootstrapping estimate as target.
  - The target now also depends on the value of $\mathbf{w}_t$, so the target is not independent of $\mathbf{w}_t$.
- Bootstrapping are not instances of true gradient descent. They take into account the effect of changing the weight vector $\mathbf{w}_t$ on the estimate, but ignore its effect on the target. Thus, they are a *semi-gradient method*.
- Regardless of the theoretical guarantees, we use them all the time ¯_(ツ)_/¯

# Semi-gradient TD(0)

**Semi-gradient TD(0) for estimating $\hat{v} \approx v_\pi$**

Input: the policy $\pi$ to be evaluated
Input: a differentiable function $\hat{v} : \mathcal{S}^+ \times \mathbb{R}^d \to \mathbb{R}$ such that $\hat{v}(\text{terminal},\cdot) = 0$
Algorithm parameter: step size $\alpha > 0$
Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)

Loop for each episode:
  Initialize $S$
  Loop for each step of episode:
    Choose $A \sim \pi(\cdot|S)$
    Take action $A$, observe $R, S'$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[R + \gamma\hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})\big]\nabla\hat{v}(S,\mathbf{w})$
    $S \leftarrow S'$
  until $S$ is terminal





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

# TD Fixed Point with Linear Function Approximation

- We do have convergence results for linear function approximation.

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\Big(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t\Big)\mathbf{x}_t$$
$$= \mathbf{w}_t + \alpha\Big(R_{t+1}\mathbf{x}_t - \mathbf{x}_t\big(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\big)^\top \mathbf{w}_t\Big)$$

# TD Fixed Point with Linear Function Approximation

- We do have convergence results for linear function approximation.

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\Big(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t\Big)\mathbf{x}_t$$
$$= \mathbf{w}_t + \alpha\Big(R_{t+1}\mathbf{x}_t - \mathbf{x}_t\big(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\big)^\top \mathbf{w}_t\Big)$$

In a steady state, for any given $\mathbf{w}_t$, the expected next weight vector can be written

$$\mathbb{E}[\mathbf{w}_{t+1}|\mathbf{w}_t] = \mathbf{w}_t + \alpha(\mathbf{b} - \mathbf{A}\mathbf{w}_t)$$

where $\mathbf{b} \doteq \mathbb{E}[R_{t+1}\mathbf{x}_t] \in \mathbb{R}^d$ and $\mathbf{A} \doteq \mathbb{E}\Big[\mathbf{x}_t\big(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\big)^\top\Big] \in \mathbb{R}^{d \times d}$

# TD Fixed Point with Linear Function Approximation

- We do have convergence results for linear function approximation.

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\Big(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t\Big)\mathbf{x}_t$$
$$= \mathbf{w}_t + \alpha\Big(R_{t+1}\mathbf{x}_t - \mathbf{x}_t\big(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\big)^\top \mathbf{w}_t\Big)$$

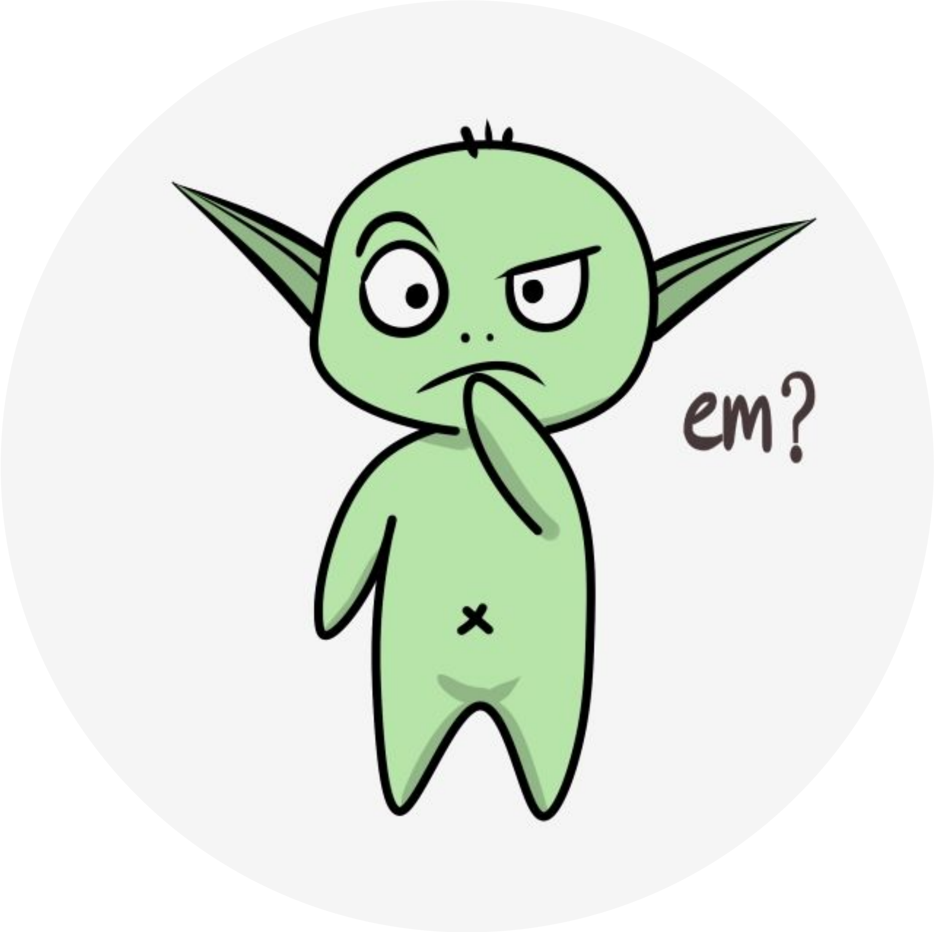
In a steady state, for any given $\mathbf{w}_t$, the expected next weight vector can be written

$$\mathbb{E}[\mathbf{w}_{t+1}|\mathbf{w}_t] = \mathbf{w}_t + \alpha(\mathbf{b} - \mathbf{A}\mathbf{w}_t)$$

where $\mathbf{b} \doteq \mathbb{E}[R_{t+1}\mathbf{x}_t] \in \mathbb{R}^d$ and $\mathbf{A} \doteq \mathbb{E}\Big[\mathbf{x}_t\big(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\big)^\top\Big] \in \mathbb{R}^{d \times d}$

It converges to:

$$\mathbf{b} - \mathbf{A}\mathbf{w}_{\text{TD}} = \mathbf{0}$$
$$\Rightarrow \quad \mathbf{b} = \mathbf{A}\mathbf{w}_{\text{TD}}$$
$$\Rightarrow \quad \mathbf{w}_{\text{TD}} \doteq \mathbf{A}^{-1}\mathbf{b}.$$






https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

# Example / Exercise

Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top\mathbf{w}$, and consider the update rule: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\,[G_t - \mathbf{x}_t^\top\mathbf{w}_t]\mathbf{x}_t$.
Let $\alpha = 0.1$, and consider $\mathbf{w}_5 = [1.0, 0.5, 3.0]^\top$ and $\mathbf{x}_5 = [0, 2, -1]^\top$.

What's $\hat{v}(\mathbf{x}_5, \mathbf{w}_5)$? What's $\mathbf{w}_6$ when applying the update rule above for $G_5 = 10$?
What do we observe from this process?

# Example / Exercise

# Example / Exercise

Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top\mathbf{w}$, and consider the update rule: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\, [G_t - \mathbf{x}_t^\top\mathbf{w}_t]\mathbf{x}_t$.
Let $\alpha = 0.1$, and consider $\mathbf{w}_5 = [1.0, 0.5, 3.0]^\top$ and $\mathbf{x}_5 = [0, 2, -1]^\top$.

What's $\hat{v}(\mathbf{x}_5, \mathbf{w}_5)$? What's $\mathbf{w}_6$ when applying the update rule above for $G_5 = 10$? What do we observe from this process?

$$\hat{v}(\mathbf{x}_5, \mathbf{w}_5) = \mathbf{x}_5^\top\mathbf{w}_5 = \begin{bmatrix}0, & 2, & -1\end{bmatrix}\begin{bmatrix}1.0\\0.5\\3.0\end{bmatrix} = 0 \times 1.0 + 2 \times 0.5 - 1 \times 3.0 = 0 + 1 - 3 = -2$$

# Example / Exercise

Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top\mathbf{w}$, and consider the update rule: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\, [G_t - \mathbf{x}_t^\top\mathbf{w}_t]\mathbf{x}_t$.
Let $\alpha = 0.1$, and consider $\mathbf{w}_5 = [1.0, 0.5, 3.0]^\top$ and $\mathbf{x}_5 = [0, 2, -1]^\top$.

What's $\hat{v}(\mathbf{x}_5, \mathbf{w}_5)$? What's $\mathbf{w}_6$ when applying the update rule above for $G_5 = 10$? What do we observe from this process?

$\hat{v}(\mathbf{x}_5, \mathbf{w}_5) = -2$

$$\mathbf{w}_6 \leftarrow \mathbf{w}_5 + \alpha\, [G_5 - \hat{v}(\mathbf{x}_5, \mathbf{w}_5)]\mathbf{x}_5$$

$$\mathbf{w}_6 \leftarrow \begin{pmatrix} 1.0 \\ 0.5 \\ 3.0 \end{pmatrix} + 0.1\ [10 - -2] \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}$$

# Example / Exercise

Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top \mathbf{w}$, and consider the update rule: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha [G_t - \mathbf{x}_t^\top \mathbf{w}_t]\mathbf{x}_t$.
Let $\alpha = 0.1$, and consider $\mathbf{w}_5 = [1.0, 0.5, 3.0]^\top$ and $\mathbf{x}_5 = [0, 2, -1]^\top$.

What's $\hat{v}(\mathbf{x}_5, \mathbf{w}_5)$? What's $\mathbf{w}_6$ when applying the update rule above for $G_5 = 10$? What do we observe from this process?

$\hat{v}(\mathbf{x}_5, \mathbf{w}_5) = -2$

$$\mathbf{w}_6 \leftarrow \mathbf{w}_5 + \alpha [G_5 - \hat{v}(\mathbf{x}_5, \mathbf{w}_5)]\mathbf{x}_5$$

$$\mathbf{w}_6 \leftarrow \begin{pmatrix} 1.0 \\ 0.5 \\ 3.0 \end{pmatrix} + 0.1\ [10 - -2] \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} \qquad \mathbf{w}_6 \leftarrow \begin{pmatrix} 1.0 \\ 0.5 \\ 3.0 \end{pmatrix} + 1.2 \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}$$

# Example / Exercise

Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top\mathbf{w}$, and consider the update rule: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\, [G_t - \mathbf{x}_t^\top\mathbf{w}_t]\mathbf{x}_t$.
Let $\alpha = 0.1$, and consider $\mathbf{w}_5 = [1.0, 0.5, 3.0]^\top$ and $\mathbf{x}_5 = [0, 2, -1]^\top$.

What's $\hat{v}(\mathbf{x}_5, \mathbf{w}_5)$? What's $\mathbf{w}_6$ when applying the update rule above for $G_5 = 10$? What do we observe from this process?

$\hat{v}(\mathbf{x}_5, \mathbf{w}_5) = -2$

$$\mathbf{w}_6 \leftarrow \mathbf{w}_5 + \alpha\, [G_5 - \hat{v}(\mathbf{x}_5, \mathbf{w}_5)]\mathbf{x}_5$$

$$\mathbf{w}_6 \leftarrow \begin{pmatrix} 1.0 \\ 0.5 \\ 3.0 \end{pmatrix} + 0.1\ [10 - -2] \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} \qquad \mathbf{w}_6 \leftarrow \begin{pmatrix} 1.0 \\ 0.5 \\ 3.0 \end{pmatrix} + 1.2 \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} \qquad \mathbf{w}_6 \leftarrow \begin{pmatrix} 1.0 \\ 0.5 \\ 3.0 \end{pmatrix} + \begin{pmatrix} 0 \\ 2.4 \\ -1.2 \end{pmatrix}$$

# Example / Exercise

Let $\hat{v}(\mathbf{x}, \mathbf{w}) = \mathbf{x}^\top\mathbf{w}$, and consider the update rule: $\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha\,[G_t - \mathbf{x}_t^\top\mathbf{w}_t]\mathbf{x}_t$.
Let $\alpha = 0.1$, and consider $\mathbf{w}_5 = [1.0, 0.5, 3.0]^\top$ and $\mathbf{x}_5 = [0, 2, -1]^\top$.

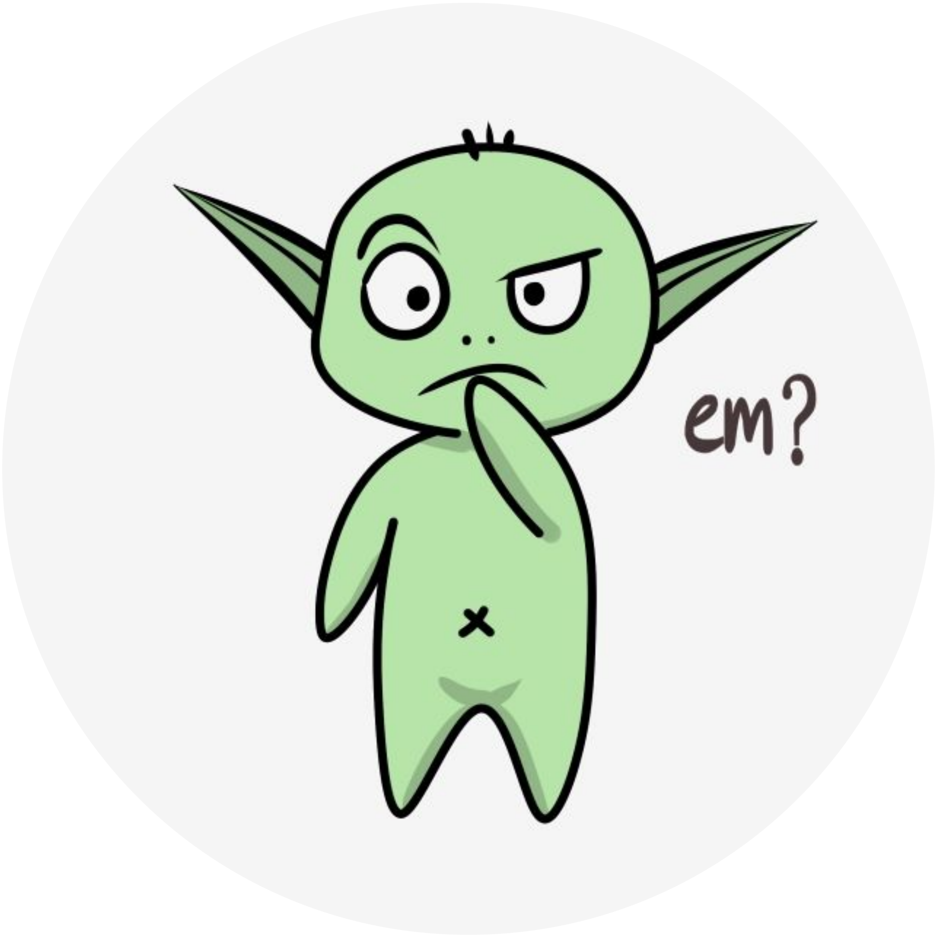
What's $\hat{v}(\mathbf{x}_5, \mathbf{w}_5)$? What's $\mathbf{w}_6$ when applying the update rule above for $G_5 = 10$? What do we observe from this process?

$\hat{v}(\mathbf{x}_5, \mathbf{w}_5) = -2$

$$\mathbf{w}_6 \leftarrow \mathbf{w}_5 + \alpha\,[G_5 - \hat{v}(\mathbf{x}_5, \mathbf{w}_5)]\mathbf{x}_5$$

$$\mathbf{w}_6 \leftarrow \begin{pmatrix}1.0\\0.5\\3.0\end{pmatrix} + 0.1\,[10 - -2]\begin{pmatrix}0\\2\\-1\end{pmatrix} \qquad \mathbf{w}_6 \leftarrow \begin{pmatrix}1.0\\0.5\\3.0\end{pmatrix} + 1.2\begin{pmatrix}0\\2\\-1\end{pmatrix} \qquad \mathbf{w}_6 \leftarrow \begin{pmatrix}1.0\\0.5\\3.0\end{pmatrix} + \begin{pmatrix}0\\2.4\\-1.2\end{pmatrix}$$

$$\mathbf{w}_6 = \begin{bmatrix}1.0, 2.4, 1.8\end{bmatrix}$$






https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht