Errata for The Art of Reinforcement Learning by Michael Hu (First Edition)

Last updated: July 6, 2024

Acknowledgement

Special thanks to Professor Ercan Atam for providing the feedback and support for this book.

Chapter 1

Page 5, last paragraph, last line — *Contributed by Ercan Atam*

Incorrect: one of the greatest Go player

Correct: one of the greatest Go players

Page 9, 7th paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: the agent may also has its internal state

Correct: the agent may also have its internal state

Page 10, 5th paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: probabilities of chose different actions

Correct: probabilities of choosing different actions

Page 17, 5th paragraph, 5th line — *Contributed by Ercan Atam*

Incorrect: as the agent may learn to exploit a loophole by simply bouncing the ball back and forth on the same side of the screen without actually clearing any bricks.

Correct: as the agent may not learn to clear bricks efficiently or prioritize efficient ball movements.

Chapter 2

Page 36, 1st paragraph, line 3-4 — *Contributed by Ercan Atam*

Incorrect: so a more accurate estimate that only includes legal actions is $3^5 = 243$.

Correct: so the actually number of valid deterministic policies may be much smaller.

Page 37, Eq. 2.15 — *Contributed by Ercan Atam*

Correct:

$$R(s,a) + \gamma \mathbb{E}_{\pi} \left[Q_{\pi}(S_{t+1}, A_{t+1}) \;\middle|\; S_t = s, A_t = a \right]$$

Chapter 3

Page 53, last paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: Eq. (3.3)

Correct: Eq. (3.4)

Page 57, 3rd paragraph, 6th line — *Contributed by Ercan Atam*

Incorrect: maximizes the expected value of the next state

Correct: maximizes the expected value of immediate reward plus the value of its next state

Page 57, 3rd paragraph, 7th line — *Contributed by Ercan Atam*

Incorrect: repeating this process until the end of the episode

Correct: repeating this process for all states in the state space

Page 58, 7th paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: we select the action that yields the highest value of its successor state

Correct: we select the action that yields the highest value of immediate reward plus the value of its successor state

Page 58, 7th paragraph 4th line, and 8th paragraph 5th line — *Contributed by Ercan Atam*

Incorrect: highest expected reward

Correct: highest expected return

Page 59, 2nd paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: However, for larger problems,

Correct: However, for problems where the true model of the MDPs are unknown,

Chapter 4

Page 66, first paragraph under 4.2, 2nd line — *Contributed by Ercan Atam*

Incorrect: The idea behind incremental updates is to update an estimate in small steps, rather than updating it all at once

Correct: The idea behind incremental updates is to estimate and update a value over multiple small steps, rather than updating it all at once.

Page 71, 6th paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: For example, we might use a linear schedule where we start with a high value of ϵ and decrease it by a fixed amount of episodes or time steps until it reaches a minimum value.

Correct: For example, we might use a linear schedule where we start with a high value of ϵ and decrease it over a large number of episodes (or time steps) until it reaches a pre-defined minimum value. Mathematically, this can be expressed as $\epsilon_t = \max(\epsilon_{\text{min}}, \epsilon_{\text{start}} - \Delta\epsilon \cdot t)$. Where ϵ_t is the value of ϵ at time step t, ϵ_{start} is the initial high value of ϵ (e.g., 1.0), $\Delta\epsilon$ is the fixed amount by which ϵ is decreased at each episode (or time step), and ϵ_{min} is the minimum value that ϵ can reach (e.g., 0.05).

Page 72, 4th paragraph, 5th line — *Contributed by Ercan Atam*

Incorrect: The new policy π' is then computed as the greedy policy with respect to Q_π, that is, for each state we choose the action with the highest estimated state-action value.

Correct: The new ϵ-greedy policy π' is then computed with respect to Q_π, that is, with probability 1 - ϵ, we choose the action with the highest estimated state-action value, and with probability ϵ, we would chose one of the non-optimal actions randomly.

Page 73, last paragraph, 3rd line — *Contributed by Ercan Atam*

Incorrect: the agent has never visited the state that often during the learning process.

Correct: the agent has barely visited the state during the learning process.

Page 74, 2nd paragraph, last line — *Contributed by Ercan Atam*

Incorrect: updating of the value function after each episode.

Correct: updating of the value function.

Page 74, last paragraph, 4rd line — *Contributed by Ercan Atam*

Incorrect: by estimating the value of each state under the current policy

Correct: by estimating the value of each state-action pair under the current policy

Chapter 5

Page 75, 2nd paragraph, 6th line — *Contributed by Ercan Atam*

Incorrect: there is a terminal state that marks the end of the episode

Correct: where a terminal state that marks the end of the episode

Page 80, second-to-last paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: we run episodes using the policy

Correct: we run steps either over some episodes or continuously using the policy

Page 86, 4th paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: There is a special case when the denominator

Correct: There is a special case when the numerator

Page 86, last paragraph, 3rd line — *Contributed by Ercan Atam*

Incorrect: Specifically, the TD target is weighted

Correct: Specifically, the TD error is weighted

Page 89, 3rd paragraph, 7th line — *Contributed by Ercan Atam*

Incorrect: the importance sampling ratio becomes zero as well. This means that the value of the sequence will also become zero and be discarded

Correct: the importance sampling ratio becomes zero or undefined as well. This means that the value of the sequence will also become zero or undefined and be discarded

Page 90, 2nd paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: is significantly different from the target policy that the agent is currently following

Correct: is significantly different from the target policy that the agent is trying to learn

Page 90, 2nd paragraph under 5.2, 3rd line — *Contributed by Ercan Atam*

Incorrect: Although we use a much simpler version of the ϵ-greedy policy that skips the step to compute a better deterministic policy, the general idea of policy iteration still applies to SARSA.

Correct: Although we use a much simpler version of the ϵ-greedy policy, the general idea of policy iteration still applies to SARSA.

Page 92, 2nd paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: we use the target policy to evaluate the action $a' = \max_{a'} Q(S_{t+1}, a')$

Correct: we use the target policy to evaluate $Q(S_t, A_t)$

Page 92, 3rd paragraph, 7th line — *Contributed by Ercan Atam*

Incorrect: the ratio of probabilities of selecting a and a' under the behavior policy

Correct: the ratio of probabilities of selecting a' under the behavior policy

Chapter 6

Page 114, last paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: the value of each state

Correct: the value of each state-action pair

Page 116, first paragraph under 6.3, 6th line — *Contributed by Ercan Atam*

Incorrect: the slope of the tangent line

Correct: the slope of the tangent line (or tangent plane for multivariable Calculus)

Page 116, last paragraph, last line — *Contributed by Ercan Atam*

Incorrect: the direction of the direction to the bottom

Correct: the direction to the bottom

Page 119, last paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: The intuition behind this equation is to ... partial gradients of each parameter $\nabla_{\boldsymbol{w}}\hat{V}(S;\boldsymbol{w})$ and the

Correct: The intuition behind this equation is to adjust the parameters in $\boldsymbol{w}$ to reduce the error between the true value $V_\pi(S)$ and the current predicted value $\hat{V}(S;\boldsymbol{w})$. The magnitude of the change is determined by the gradients $\nabla_{\boldsymbol{w}}\hat{V}(S;\boldsymbol{w})$ and the

Page 120, 2nd paragraph, 6th line — *Contributed by Ercan Atam*

Incorrect: due the computation efficient reasons

Correct: due to computation efficient reasons

Page 120, 3rd paragraph, last line — *Contributed by Ercan Atam*

Incorrect: to update the model's parameters.

Correct: to update the model's parameters. To keep things simple, Eq. (6.9) only shows value update rule using SGD for a single transition, not a mini-batch, since there are not expectation sign $\mathbb{E}$ in front of it.

Page 122, third-to-last paragraph, 3rd line — *Contributed by Ercan Atam*

Incorrect: rewards received after visiting a state t until the end of the episode

Correct: rewards received after visiting state S_t until the end of the episode

Page 122, last paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: the context of large-scale MDPs where almost all states are unique

Correct: the context of MDPs with a large discrete state space, or a continuous state space where nearly all states are unique

Page 124, Eq. 6.17 — *Contributed by Ercan Atam*

Correct:

$$\begin{aligned}\boldsymbol{w} &= \boldsymbol{w} + \alpha\left[\Big(R_t + \gamma\hat{Q}(S_{t+1}, A_{t+1}; \boldsymbol{w}) - \hat{Q}(S_t, A_t; \boldsymbol{w})\Big)\boldsymbol{x}(S_t, A_t)\right] \\ &= \boldsymbol{w} + \alpha\left[\Big(R_t + \gamma\boldsymbol{x}(S_{t+1}, A_{t+1})^T\boldsymbol{w} - \boldsymbol{x}(S_t, A_t)^T\boldsymbol{w}\Big)\boldsymbol{x}(S_t, A_t)\right]\end{aligned}$$

Page 125, 1st paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: to approximate the state-action value function.

Correct: to approximate the state-action value function $\boldsymbol{w}$.

Page 126, Algorithm 5 line 2, and Algorithm 6 lines 2, 6 — *Contributed by Ercan Atam*

Incorrect: ϵ-greedy policy w.r.t. $\boldsymbol{w}$

Correct: ϵ-greedy policy w.r.t. Q

Page 127, Algorithm 7 line 2 — *Contributed by Ercan Atam*

Incorrect: ϵ-greedy policy w.r.t. $\boldsymbol{w}$

Correct: ϵ-greedy policy w.r.t. Q

Page 127, second-to-last paragraph — *Contributed by Ercan Atam*

Incorrect: In practice, we often use a separate simulation environment to run the evaluation process, often with different exploration rate ϵ and seed.

Correct: In practice, we often use a separate simulation environment to run the evaluation process, often with different exploration rate ϵ and seed. This setup during evaluation ensures a fair, unbiased, and comprehensive assessment of the RL agent's performance. It helps in understanding the agent's generalization abilities, robustness, and true capabilities independent of the training specifics.

Page 128, Fig. 6.8 — *Contributed by Ercan Atam*

Correct: The y-axis label should be "Mean training episode return"

Page 128, 1st paragraph, 4th line — *Contributed by Ercan Atam*

Incorrect: or specific domain-specific metrics

Correct: or domain-specific metrics

Chapter 7

Page 131, 2nd paragraph, 4th line — *Contributed by Ercan Atam*

Incorrect: taking the environment state

Correct: taking the environment state or state-action pair

Page 132, second-to-last paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: Here, $\boldsymbol{W}$ is a weight matrix with dimensions m x n

Correct: Here, $\boldsymbol{W}$ is a weight matrix with dimensions n x m

Page 132, Eq. 7.2 — *Contributed by Ercan Atam*

Correct:

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W} + \mathbf{b}$$

Page 140, 3rd paragraph, 6th line — *Contributed by Ercan Atam*

Incorrect: the parameters for these different layers for these different layers

Correct: the parameters for these different layers

Page 145, 1st paragraph, 6th line — *Contributed by Ercan Atam*

Incorrect: for the given mni-batch:

Correct: for the given mini-batch. Please note, to keep things simple, Eq. (7.4) only shows value update for a single transition tuple, not mini-batch.

Page 146, second-to-last paragraph, 9th line — *Contributed by Ercan Atam*

Incorrect: the actual reward received

Correct: the actual return received

Page 146, last paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: we ran 20,000 evaluation environment steps (approximately 40 episodes) on a separate evaluation environment with an exploration epsilon of 0.05 after every 20,000 training environment steps

Correct: we use a separate evaluation environment with an exploration epsilon of 0.05 and different seed to ensures a fair, unbiased assessment of the agent's performance. Specifically, for every 20,000 training environment steps, we would set the agent to ran 20,000 evaluation environment steps (approximately 40 evaluation episodes) on the evaluation environment.

Page 147, 151, Fig. 7.12, Fig. 7.14 — *Contributed by Ercan Atam*

Correct: The y-axis label should be "Mean evaluation episode return"

Page 151, last paragraph, 2nd line — *Contributed by Ercan Atam*

Incorrect: we ran 20,000 evaluation environment steps (approximately 40 episodes) on a separate evaluation environment with an exploration epsilon of 0.05 after every 20,000 training environment steps.

Correct: we use a separate evaluation environment with an exploration epsilon of 0.05, similar to the cart pole experiment. Specifically, for every 20,000 training environment steps, we would set the agent to ran 20,000 evaluation environment steps (approximately 100 evaluation episodes) on the evaluation environment.

Page 152, 1st paragraph — *Contributed by Ercan Atam*

Incorrect: We use learning rate $\alpha = 0.00025$, discount rate $\gamma = 0.99$, batch size = 32, and replay capacity = 50,000 in .... (if any) every 200 updates.

Correct: This repeats previous work, and some parameters are also inconsistent, and this whole paragraph should be removed.

Page 153, 1st paragraph under Environment Preprocessing, 6th line — *Contributed by Ercan Atam*

Incorrect: small red block in the middle left of the screen

Correct: small red block in the middle right of the screen

Page 154, 1st paragraph, second-to-last line — *Contributed by Ercan Atam*

Incorrect: small red block in the middle left of the screenskips processing $(k-1)/k$ percent of the frames

Correct: skips processing $(k-1)/k$ fraction of the frames

Page 155, last paragraph, 5th line — *Contributed by Ercan Atam*

Incorrect: Another method is to randomizing the initial state the initial state of each episode by applying some no-ops action for some steps can also be effective in improving the agent's learning progress.

Correct: Another method is to randomize the initial state of each episode by applying some no-op actions. Instead of moving left or right, the agent performs no-op actions, which do nothing. This allows the environment's dynamics to change the state, ensuring the agent encounters a variety of initial states.

Page 156, second-to-last paragraph, second-to-last line — *Contributed by Ercan Atam*

Incorrect: output a vector contains

Correct: output a vector that contains

Page 157, 1st paragraph, 1st line — *Contributed by Ercan Atam*

Incorrect: number (s, a)

Correct: number $\hat{Q}(s, a)$

Page 157, 2nd paragraph, 7th line — *Contributed by Ercan Atam*

Incorrect: the agent since since it has lesser time to generate more transitions

Correct: the agent, since it has to spend more time and resource on updating the neural network, this means lesser time and resource to generating new transitions for learning

Page 157, last paragraph, 3rd line — *Contributed by Ercan Atam*

Incorrect: To evaluate the agent's performance, we ran 200,000 evaluation steps on a separate testing environment with an ϵ-greedy policy and a fixed exploration epsilon ($\epsilon = 0.05$) at the end of each training iteration, which consisted of 250,000 training steps or 1 million frames,

Correct: To evaluate the agent's performance, we adapt similar settings from previous experiments, where we use a separate evaluation environment with a fixed exploration epsilon ($\epsilon = 0.05$). Specifically, at the end of each training iteration (every 250,000 training steps or 1 million training frames), we ran 200,000 evaluation steps on the evaluation environment,

Page 159, Fig. 7.21, Fig. 7.22 — *Contributed by Ercan Atam*

Correct: The y-axis label should be "Mean evaluation episode return"

Page 161, 3rd paragraph, 4th line — *Contributed by Ercan Atam*

Incorrect: target networks provide a fixed target for

Correct: target networks provide a more stable target value for