CSCI 145 Problem 23: An Interest Rate on Reward

July 25, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 23: An Interest Rate on Reward

Would you rather have \$100 today or \$100 next year? Anyone with a bank account says today: money in hand earns interest, so at a rate of 11%, next year's \$100 is worth only $100/1.11 \approx \$90$ now. Lecture's discount factor is the same arithmetic with $\gamma = 1/1.11 \approx 0.9$: a reward that arrives one step later is worth γ times as much, and the return $G_t = \sum_{k \geq 0} \gamma^k r_{t+k}$ is a stream of payments quoted in today's currency. In class we verified the recursion $G_t = r_t + \gamma G_{t+1}$ and the ceiling $1/(1-\gamma)$ numerically, on a four-step episode and on CartPole's all-ones rewards. This problem proves both in general and then puts them to work. The recursion becomes the *Bellman consistency equation*, the workhorse identity of reinforcement learning, and the ceiling becomes one half of a bias–variance tradeoff, the course's variance-reduction arc arriving at its final unit.

Part A: The Return Is Bounded and Recursive

Fix $\gamma \in [0, 1)$ and let $r_t, r_{t+1}, r_{t+2}, \ldots$ be any infinite stream of rewards with $0 \leq r_{t+k} \leq r_{\max}$ for all $k \geq 0$.

First, show that the return is finite no matter how long the stream runs:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \leq \frac{r_{\max}}{1-\gamma}.$$

(Bound each reward by $r_{\max}$ and sum the geometric series. This is why the infinite sum defining G_t is a well-defined number at all.) Sanity-check the bound against the demo: with CartPole's $r_{\max} = 1$ and $\gamma = 0.99$, what ceiling does it give, and what discounted return did the demo's 500-step balanced episode earn?

Next, show that the return satisfies the one-step recursion

$$G_t = r_t + \gamma\, G_{t+1}$$

for every t, by splitting off the $k = 0$ term of the sum and reindexing what remains. The in-class exercise checked this on a four-step episode; your derivation covers every reward stream at once, including infinite ones. Then say in one sentence what the recursion buys: which quantity does it let you compute from which, and why is that useful when the sum runs forever?

Part B: The Bellman Consistency Equation

Now let the stream be random: an agent follows a policy π in a Markov decision process, so the states s_t, the rewards r_t, and hence the returns G_t are all random variables. Define the **value function** of the policy,

$$V^\pi(s) = \mathbb{E}\left[G_t \mid s_t = s\right],$$

the expected return from state s onward. (The Markov property is what makes this depend only on s: the future looks the same from a state no matter when or how the agent arrived there.)

Take the conditional expectation of Part A's recursion given $s_t = s$ and derive the **Bellman consistency equation**:

$$V^\pi(s) = \mathbb{E}\left[r_t + \gamma\, V^\pi(s_{t+1}) \mid s_t = s\right].$$

Proceed in two steps. First apply linearity of expectation (lecture one, still undefeated) to split the reward term from the future term. Then handle $\mathbb{E}[G_{t+1} \mid s_t = s]$ by splitting the expectation over the possible next states s', the same move as lecture one splitting $\Pr(B)$ over whether A occurred, and using the Markov property to argue that $\mathbb{E}[G_{t+1} \mid s_{t+1} = s', s_t = s] = V^\pi(s')$. Then read the equation back in one sentence of plain English: what does it say the value of a state is made of?

Part C: γ Is a Bias–Variance Knob

How would an agent *measure* a value like $V^\pi(s)$? By Monte Carlo, of course: roll out one trajectory from s and compute its discounted return. To make everything computable, model the rewards along the trajectory as independent random variables $r_0, r_1, r_2, \ldots$, each with mean $\mu > 0$ and variance σ^2. Suppose the quantity the agent truly cares about is far-sighted value at $\gamma^* = 0.99$, but its estimator uses a knob $\gamma \le \gamma^*$ it is free to turn down:

$$\widehat{G}_\gamma = \sum_{k=0}^{\infty} \gamma^k r_k.$$

1. **Bias.** Use linearity of expectation to show $\mathbb{E}[\widehat{G}_\gamma] = \mu/(1-\gamma)$, so the target is $\mu/(1-\gamma^*) = 100\mu$ and the bias is $\mathbb{E}[\widehat{G}_\gamma] - 100\mu$. Evaluate the bias at $\gamma = 0.99$ and at $\gamma = 0.9$.

2. **Variance.** Independent contributions have additive variances, and constants come out of variances squared (both from lecture one). Show

   $$\mathrm{Var}\big(\widehat{G}_\gamma\big) = \sigma^2 \sum_{k=0}^{\infty} \gamma^{2k} = \frac{\sigma^2}{1-\gamma^2},$$

   and evaluate it at $\gamma = 0.99$ and at $\gamma = 0.9$. By what factor does turning the knob from 0.99 down to 0.9 shrink the variance?

3. Conclude in one sentence, with your numbers: what does shrinking γ buy, and what does it cost? (Your effective horizons from the in-class exercise, 10 steps versus 100, are the plain-language version of the same tradeoff.)

Where this goes next. Part C brings the variance-reduction theme into the final unit: a Monte Carlo estimator of value is unbiased only if it is far-sighted, and far-sighted only if it is noisy. The next lecture builds the policy-gradient estimator and finds it noisy in a far worse way. The fix, in the course's closing lecture, is to subtract a baseline from the return before using it. The (nearly) optimal baseline turns out to be exactly the value function V^π you defined in Part B, doing for trajectories what Problem 2's control variate did for samples: same target, far less noise.