@zblu Reinforcement Learning Theory

INF8250AE Course Notes

RL 3:Bellman Operators: Properties and Consequences

Monotonicity · Contraction Mapping · Banach Fixed Point
Fixed-Point Uniqueness · Error Bounds · Optimality

Key Topics: Bellman Equations, Bellman Operators, Greedy Policy, Banach Fixed Point Theorem, Contraction & Monotonicity
Reference: `https://amfarahmand.github.io/IntroRL/`

Notes on Reinforcement Learning – Sep 22, 2025

POLYTECHNIQUE MONTRÉAL

Mila

Reinforcement Learning 3: Bellman Operators: Properties and Consequences

Zhuobie

2025 年 10 月 6 日

1. Supremum Norms

Definition 1.1 (Supremum Norm) *In RL we use the sup-norm for bounded functions:* $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|, \|V\|_\infty = \sup_{x\in\mathcal{X}} |V(x)|$ *and* $\|Q\|_\infty = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |Q(x,a)|$.

2. Contraction Mapping & Banach Fixed Point

Definition 2.1 (Contraction Mapping) $L : Z \to Z$ *is a contraction if* $\exists\, 0 \le a < 1$ *s.t.* $d(Lz_1, Lz_2) \le a\, d(z_1, z_2)$ *for all* z_1, z_2.

Theorem 2.2 (Banach Fixed Point) *If* L *is a contraction, then there is a unique fixed point* $z^\star$ *with* $Lz^\star = z^\star$, *and for any* z_0, *the iteration* $z_{k+1} = Lz_k$ *satisfies* $d(z_k, z^\star) \le a^k d(z_0, z^\star) \to 0$.

Linear warm-up: $L : z \mapsto az + b$ with $|a| < 1$: $z^\star = \frac{b}{1-a}$; $z_k \to z^\star$ geometrically.

3. Monotonicity and Contraction — Full Proofs

Lemma 3.1 (Monotonicity of T^π and T^*) *If* $V_1 \le V_2$ *pointwise, then* $T^\pi V_1 \le T^\pi V_2$ *and* $T^* V_1 \le T^* V_2$.

Proof

For T^π: P^π is linear and positive, so $P^\pi V_1 \le P^\pi V_2 \Rightarrow T^\pi V_1 \le T^\pi V_2$.
For T^*: for each a, the map $V \mapsto r(x,a) + \gamma \int V(x') P(dx'|x,a)$ is monotone, hence the pointwise maximum over a preserves monotonicity.

Lemma 3.2 (Contraction) *For any Bellman operator* T^π *or* T^* *and any discount factor* $0 \le \gamma < 1$, *the operators are* γ*-contractions under the supremum norm:*

$$\|T^\pi Q_1 - T^\pi Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty, \qquad \|T^* Q_1 - T^* Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty.$$

Proof

Let $Q_1, Q_2 \in \mathcal{B}(\mathcal{X} \times \mathcal{A})$, and for any $(x, a) \in \mathcal{X} \times \mathcal{A}$:

$$(T^\pi Q)(x, a) = r(x, a) + \gamma \int_{\mathcal{X}} \int_{\mathcal{A}} Q(x', a')\, \pi(da'|x')\, P(dx'|x, a).$$

Then

$$\begin{aligned} &|(T^\pi Q_1)(x, a) - (T^\pi Q_2)(x, a)| \\ &= \gamma \left| \int_{\mathcal{X}} \int_{\mathcal{A}} \big(Q_1(x', a') - Q_2(x', a')\big)\, \pi(da'|x')\, P(dx'|x, a) \right|. \end{aligned}$$

By linearity of integration,

$$\begin{aligned} &\leq \gamma \int_{\mathcal{X}} \int_{\mathcal{A}} |Q_1(x', a') - Q_2(x', a')|\, \pi(da'|x')\, P(dx'|x, a) \quad \text{(triangle inequality)} \\ &\leq \gamma \|Q_1 - Q_2\|_\infty \int_{\mathcal{X}} \int_{\mathcal{A}} \pi(da'|x')\, P(dx'|x, a) = \gamma \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Math trick: For any probability measure $p(x)$ and bounded f,

$$\left| \int p(x) f(x)\, dx \right| \leq \int |p(x) f(x)|\, dx = \int p(x)\, |f(x)|\, dx \leq \sup_x |f(x)| \int p(x)\, dx = \|f\|_\infty.$$

Hence taking $\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}}$ gives

$$\|T^\pi Q_1 - T^\pi Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty.$$

Similarly, for the optimality operator:

$$(T^* Q)(x, a) = r(x, a) + \gamma \int_{\mathcal{X}} \max_{a'} Q(x', a')\, P(dx'|x, a),$$

and using $\left| \max_i u_i - \max_i v_i \right| \leq \max_i |u_i - v_i|$, we get

$$\begin{aligned} |(T^* Q_1)(x, a) - (T^* Q_2)(x, a)| &\leq \gamma \int_{\mathcal{X}} \max_{a'} |Q_1(x', a') - Q_2(x', a')| P(dx'|x, a) \\ &\leq \gamma \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Therefore, both T^π and T^* are γ-contractions.

4. Properties's Consequences

4.1 Uniqueness of Fixed Points

Proposition 4.1 (Uniqueness of Fixed Point) *$V^\pi = T^\pi V^\pi$ and $V^* = T^* V^*$ have unique solution. Let $V^* = T^* V^*$ be the unique solution of the fixed-point equation. For any initial $V_0 \in \mathcal{B}(\mathcal{X})$, define*

$$V_{k+1} = T^* V_k, \qquad k = 0, 1, 2, \ldots$$

Then

$$\lim_{k\to\infty} \|V_k - V^*\|_\infty = 0.$$

Proof

Let $f(V) = T^*V - V$. We want to find V such that $f(V) = 0$.
Because T^* is a γ-contraction, the fixed-point iteration

$$V_{k+1} = T^*V_k$$

converges to the unique V^* satisfying $f(V^*) = 0$.
Hence $T^*V^* = V^*$, and V^* is the only fixed point.

Proposition 4.2 (Value of Greedy Policy of V^* is V^*) *If $T^*V^* = T^{\pi^*}V^*$, then and only then $V^{\pi^*} = V^*$.*

$$T^{\pi_g(V^*)}V^* = T^{\pi^*}V^* = T^*V^* = V^*.$$

4.2 Error Bounds via Bellman Residuals

Definition 4.3 (Bellman residual) $\mathrm{BR}^*(V) = V - T^*V$ *and* $\mathrm{BR}^\pi(V) = V - T^\pi V$.

Proposition 4.4 (Bellman Error Bounds (Full Derivation)) *For any V,* $\|V - V^*\|_\infty \le \dfrac{\|V - T^*V\|_\infty}{1-\gamma}$, $\quad \|V - V^\pi\|_\infty \le \dfrac{\|V - T^\pi V\|_\infty}{1-\gamma}$.

Proof

Contraction: We know that T^* is a γ-contraction, i.e.

$$\|T^*V - T^*V^*\|_\infty \le \gamma\|V - V^*\|_\infty, \quad \text{and } T^*V^* = V^*.$$

Then,

$$\begin{aligned}\|V - V^*\|_\infty &= \|(V - T^*V) + (T^*V - V^*)\|_\infty \\ &\le \|V - T^*V\|_\infty + \|T^*V - V^*\|_\infty \\ &\le \|V - T^*V\|_\infty + \gamma\|V - V^*\|_\infty.\end{aligned}$$

Rearranging the inequality gives

$$(1-\gamma)\|V - V^*\|_\infty \le \|V - T^*V\|_\infty,$$

hence

$$\|V - V^*\|_\infty \le \frac{\|V - T^*V\|_\infty}{1-\gamma}.$$

Interpretation: $\|V - T^*V\|_\infty$ is the *Bellman error*. Therefore, the difference $\|V - V^*\|_\infty$ can be bounded by the Bellman error scaled by $\frac{1}{1-\gamma}$.

We just can do approximated value:

$$\|V - V^*\|_\infty \le \frac{\|T^*V - V\|_\infty}{1-\gamma} \quad \Rightarrow \quad \text{error bound of } V \text{ depends on its Bellman error.}$$

4.3 Small Numeric Illustrations

Example 4.5 *Consider a two-state system with transition matrix*

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}, \qquad \gamma = \frac{1}{2}.$$

Let

$$V_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad V_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We want to compute

$$\frac{\|TV_1 - TV_2\|_\infty}{\|V_1 - V_2\|_\infty}.$$

Since $TV = r + \gamma PV$, let $r = 0$ for simplicity. Then

$$TV_1 = \frac{1}{2}PV_1, \qquad TV_2 = \frac{1}{2}PV_2.$$

Compute:

$$PV_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \end{bmatrix}, \qquad PV_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \end{bmatrix}.$$

Hence

$$TV_1 = \frac{1}{2}\begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{6} \end{bmatrix}, \qquad TV_2 = \frac{1}{2}\begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{3} \end{bmatrix}.$$

Now compute the difference:

$$TV_1 - TV_2 = \begin{bmatrix} 0 \\ -\frac{1}{6} \end{bmatrix}, \quad \|TV_1 - TV_2\|_\infty = \frac{1}{6}.$$

Meanwhile,

$$V_1 - V_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \|V_1 - V_2\|_\infty = 1.$$

Therefore,

$$\frac{\|TV_1 - TV_2\|_\infty}{\|V_1 - V_2\|_\infty} = \frac{1}{6}$$

5. Composition of Bellman Operators

Definition 5.1 (Operator Pf) *For any bounded measurable function $f : \mathcal{X} \to \mathbb{R}$, define the operator*

$$(Pf)(x) = \mathbb{E}_{X' \sim P(\cdot|x)}[\, f(X')\,] = \int_{\mathcal{X}} f(x')\, P(dx'|x).$$

Remark (Composition of transition kernels). *For the composed transition probability P^{π},*

$$(P^{\pi} f)(x) = \mathbb{E}_{X' \sim P^{\pi}(\cdot|x)}[\, f(X')\,] = \int_{\mathcal{A}} \int_{\mathcal{X}} f(x')\, P(dx'|x,a)\, \pi(da|x).$$

Characterization. *For any measurable set $A \subseteq \mathcal{X}$,*

$$P^{\pi_1:\pi_m}(A|x) = \int_{\mathcal{X}} P^{\pi_m}(A|y)\, P^{\pi_{1:(m-1)}}(dy|x).$$

Equivalently,

$$P^{\pi_{1:m}}(A|x,a) = \int_{\mathcal{X}} P^{\pi_m}(A|y)\, P^{\pi_{1:(m-1)}}(dy|x,a).$$

Takeaways: Bellman operators are monotone and γ-contractive $\Rightarrow$ unique fixed points and value-iteration convergence; greedy policy from V^* is optimal; Bellman residual controls the value error.