Summer Research Internship Project

Convex Optimization and its Connections to Markov Decision Processes

Aritrabha Majumdar

Under the guidance of Dr. Somnath Pradhan

Indian Institute of Science Education and Research, Bhopal

Contents

1 Foundations of Convex Optimization

This part of the project develops the analytical and geometric foundations of convex optimization that will later be used in the linear programming formulation of MDPs and their dual interpretation.

1.1 Convex Sets: Geometry and Separation

Definition 1.1 (Convex Combination)**.** *Let $x_1, \dots, x_k \in \mathbb{R}^n$. A point*

$$x = \sum_{i=1}^{k} \theta_i x_i$$

is a convex combination if $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$.

Definition 1.2 (Convex Set)**.** *A set $C \subset \mathbb{R}^n$ is convex if for all $x, y \in C$ and $\theta \in [0, 1]$,*

$$\theta x + (1 - \theta) y \in C.$$

Equivalently, C contains all convex combinations of its elements.

Definition 1.3 (Affine and Convex Hull)**.** *The affine hull of $S \subset \mathbb{R}^n$ is*

$$\operatorname{aff}(S) = \left\{ \sum_{i=1}^{k} \theta_i x_i \ : \ x_i \in S, \sum_{i=1}^{k} \theta_i = 1 \right\}.$$

The convex hull is

$$\operatorname{conv}(S) = \left\{ \sum_{i=1}^{k} \theta_i x_i \ : \ x_i \in S, \theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1 \right\}.$$

Proposition 1.1. *Intersections of convex sets are convex. Linear images of convex sets are convex.*

Proof. If C_i are convex and $x, y \in \cap_i C_i$, then for any θ, $\theta x + (1-\theta) y \in C_i$ for all i, hence in the intersection.

If C is convex and A linear, then for $x, y \in C$,

$$A(\theta x + (1 - \theta) y) = \theta A x + (1 - \theta) A y.$$

□

Hyperplanes and Separation

Definition 1.4 (Hyperplane)**.** *A hyperplane is a set of the form*

$$H = \{x \in \mathbb{R}^n : a^\top x = b\},$$

where $a \neq 0$.

Theorem 1.1 (Separating Hyperplane Theorem)**.** *Let* $C \subset \mathbb{R}^n$ *be nonempty, closed, and convex, and let* $x_0 \notin C$*. Then there exists* $a \neq 0$ *and* α *such that*

$$a^\top x \leq \alpha < a^\top x_0 \quad \forall x \in C.$$

Proof. Let us define $y = \text{proj}_C x_0$. Now we consider any $x \in C$. Now we define $y_t = y + t(x - y)$. As per the definition of projection, we know that

$$\begin{aligned}
&||x_0 - y||^2 \leq ||x_0 - y_t||^2 \quad \forall t \in [0,1] \\
\implies &||x_0 - y||^2 \leq ||x_0 - y - t(x-y)||^2 \\
\implies &||a||^2 \leq ||a - tb||^2 \quad (\text{take } a = x_0 - y, b = x - y) \\
\implies &2ta^\top b \leq t^2 ||b||^2 \\
\implies &2a^\top b \leq t||b||^2 \\
\implies &a^\top b \leq 0 \quad (\text{take } t \downarrow 0)
\end{aligned}$$

Therefore we have $\langle x_0 - y, x - y \rangle \leq 0$. Take $a = x_0 - y$ and $\alpha = a^\top y$. This proves the left part of the inequality prescribed.

For the right side, just consider $||x_0 - y|| \geq 0 \implies \langle a, x_0 - y \rangle \geq 0 \implies a^\top x_0 \geq \alpha$ □

This theorem is fundamental for duality theory and underlies strong duality in convex optimization.

1.2 Convex Functions and Analytical Characterizations

Definition 1.5 (Convex Function)**.** *Let* C *be convex. A function* $f : C \to \mathbb{R}$ *is convex if*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

Definition 1.6 (Strict and Strong Convexity)**.** f *is strictly convex if the inequality is strict for* $x \neq y$*.* f *is* μ*-strongly convex if*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

First-Order Characterization

Theorem 1.2. *Let f be differentiable on open convex C. Then f is convex iff*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in C.$$

Proof. ($\Rightarrow$) Define $\phi(t) = f(x + t(y - x))$. Convexity of f implies convexity of ϕ. Therefore, the *secant* slope of ϕ is *non-decreasing*. Then

$$\phi'(0) = \lim_{h \downarrow 0} \frac{\phi(h) - \phi(0)}{h} \leq \frac{\phi(t) - \phi(0)}{t}$$

Putting $t = 1$ and rearranging, we get

$$\phi(1) \geq \phi(0) + \phi'(0).$$

Now,

$$\begin{aligned} \phi'(0) &= \left[\left(\nabla f(x + t(y - x))^\top \right) \left(\frac{d}{dt}(x + t(y - x)) \right) \right]_{t=0} = \nabla f(x)^\top (y - x) \\ \implies f(y) &\geq f(x) + \nabla f(x)^\top (y - x) \end{aligned}$$

($\Leftarrow$) Let $z = \theta x + (1 - \theta) y$, $x, y \in C$. Therefore $z \in C$ as C is convex. Now we can write

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z)$$

$$f(y) \geq f(z) + \nabla f(z)^\top (y - z)$$

Now, we have

$$\begin{aligned} \theta f(x) + (1 - \theta) f(y) &\geq f(z) + \nabla f(z)^\top [\theta(x - z) + (1 - \theta)(y - z)] \\ &= f(z) + \nabla f(z)^\top [(\theta x + (1 - \theta) y) - z] \\ &= f(z) \end{aligned}$$

Hence f is *convex*. □

Second-Order Characterization

Theorem 1.3. *If f is twice differentiable, then f is convex iff*

$$\nabla^2 f(x) \succeq 0 \quad \forall x.$$

Proof. We fix $x \in C$ and $v \in \mathbb{R}^n$. For t sufficiently small, we have $x + tv \in C$. Then we take $\phi(t) = f(x + tv)$. As f is convex and twice differentiable, so is ϕ. Therefore

$$\phi''(0) \geq 0 \implies v^\top \nabla^2 f(x) v \geq 0$$

As it holds for all $v \in \mathbb{R}^n$ and selection of $x \in C$ was arbitrary,

$$\nabla^2 f(x) \succeq 0 \quad \forall x.$$

For the other direction, fix $x, y \in C$, and then take $\phi(t) = f(x + t(y - x))$, $t \in [0, 1]$. We also take $d = x - y$. Then we have $\phi''(t) = d^\top \nabla^2 f(x + td)d$. As $\nabla^2 f(x) \succeq 0$, we put $t = 0$ and obtain $\phi''(0) \geq 0$, i.e., ϕ is convex. Then using the argument used in previous proof, we get

$$\phi(1) \geq \phi(0) + \phi'(0) \implies f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

Hence, f is convex. □

1.3 Subgradients and Nonsmooth Analysis

Definition 1.7 (Subgradient)**.** *For convex f, a vector g is a subgradient at x if*

$$f(y) \geq f(x) + g^\top (y - x), \quad \forall y.$$

Definition 1.8 (Subdifferential)**.**

$$\partial f(x) = \{g : g \text{ is a subgradient at } x\}.$$

Proposition 1.2. *x^* minimizes convex f iff $0 \in \partial f(x^*)$.*

Proof. If x^* is a minimizer, then if we fix $g = 0$, then $f(y) \geq f(x^*)$ $\forall y$.

Also, if $0 \in \partial f(x^*)$, then $f(y) \geq f(x^*)$ $\forall y$. □

Convex Conjugate

Definition 1.9 (Fenchel Conjugate)**.**

$$f^*(y) = \sup_x \{y^\top x - f(x)\}.$$

Theorem 1.4 (Fenchel–Young Inequality)**.**

$$f(x) + f^*(y) \geq x^\top y.$$

Equality holds iff $y \in \partial f(x)$.

Proof. According to definition

$$\begin{aligned} & f^*(y) = \sup_x \{y^\top x - f(x)\} \\ \implies & f^*(y) \geq x^\top y - f(x) \\ \implies & f(x) + f^*(y) \geq x^\top y \end{aligned}$$

Equality in the Fenchel–Young inequality holds if and only if

$$f^*(y) = y^\top x - f(x).$$

Since

$$f^*(y) = \sup_z \{y^\top z - f(z)\},$$

this means that the supremum is attained at $z = x$. Hence

$$y^\top z - f(z) \leq y^\top x - f(x), \qquad \forall z \in \mathbb{R}^n.$$

Rearranging,

$$f(z) \geq f(x) + y^\top (z - x), \qquad \forall z \in \mathbb{R}^n.$$

By the definition of a subgradient, this is equivalent to

$$y \in \partial f(x).$$

Conversely, if $y \in \partial f(x)$, then

$$f(z) \geq f(x) + y^\top (z - x), \qquad \forall z \in \mathbb{R}^n,$$

which implies

$$y^\top z - f(z) \leq y^\top x - f(x), \qquad \forall z \in \mathbb{R}^n.$$

Therefore,

$$f^*(y) = \sup_z \{y^\top z - f(z)\} = y^\top x - f(x),$$

and hence

$$f(x) + f^*(y) = x^\top y.$$

Thus,

$$f(x) + f^*(y) = x^\top y \quad \iff \quad y \in \partial f(x).$$

□

Conjugacy will later appear in entropy-regularized MDPs.

1.4 Convex Optimization Problems

Standard Form

$$\min_x f(x) \quad \text{s.t. } g_i(x) \leq 0,\ Ax = b.$$

Assume f, g_i convex and constraints affine.

Lagrangian and Dual Function

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_i \lambda_i g_i(x) + \nu^\top (Ax - b),$$

with $\lambda_i \geq 0$.

$$q(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu).$$

Theorem 1.5 (Weak Duality)**.**

$$\sup_{\lambda \geq 0, \nu} q(\lambda, \nu) \leq \inf_{x\ \textit{feasible}} f(x).$$

Proof. For each i, because $\lambda_i \geq 0$. Hence

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) \leq f(x).$$

Here, $h_j(x)$s corresponds to the columns of $(Ax - b)$.

By definition of the dual function,

$$q(\lambda, \nu) = \inf_z L(z, \lambda, \nu) \leq L(x, \lambda, \nu).$$

Combining the above inequalities yields

$$q(\lambda, \nu) \leq L(x, \lambda, \nu) \leq f(x).$$

Since this holds for every feasible x,

$$q(\lambda, \nu) \leq \inf_{x \text{ feasible}} f(x) = p^*.$$

Finally, taking the supremum over all dual-feasible multipliers (λ, ν) with $\lambda \geq 0$, we obtain

$$\sup_{\lambda \geq 0, \nu} q(\lambda, \nu) \leq p^*.$$

Thus

$$d^* = \sup_{\lambda \geq 0, \nu} q(\lambda, \nu) \leq \inf_{x \text{ feasible}} f(x) = p^*.$$

□

Strong Duality

Definition 1.10 (Slater Condition)**.** *There exists x such that $g_i(x) < 0$ and $Ax = b$.*

Theorem 1.6 (Strong Duality)**.** *If the problem is convex and Slater's condition holds, then*

$$p^* = d^*.$$

Proof. By weak duality,

$$d^* \leq p^*.$$

It therefore suffices to prove the reverse inequality.

Define the set

$$\mathcal{A} = \Big\{(u, v, t) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} : \exists x,\ g(x) \leq u,\ Ax - b = v,\ f(x) \leq t\Big\},$$

where

$$g(x) = (g_1(x), \ldots, g_m(x)).$$

Since f and the functions g_i are convex, the set $\mathcal{A}$ is convex.

Fix $\alpha < p^*$. We claim that

$$(0, 0, \alpha) \notin \mathcal{A}.$$

Indeed, if $(0, 0, \alpha) \in \mathcal{A}$, then there exists x satisfying

$$g_i(x) \leq 0, \qquad Ax = b, \qquad f(x) \leq \alpha.$$

Hence x is feasible and

$$p^* \le f(x) \le \alpha,$$

contradicting $\alpha < p^*$.

Because Slater's condition holds, $\mathcal{A}$ has nonempty interior. Therefore, by the strong separating hyperplane theorem, there exist

$$\lambda \in \mathbb{R}^m, \qquad \nu \in \mathbb{R}^p, \qquad \mu \in \mathbb{R},$$

not all zero, such that

$$\lambda^\top u + \nu^\top v + \mu t \ge \mu\alpha$$

for every $(u, v, t) \in \mathcal{A}$.

We first show that $\mu > 0$. If $\mu = 0$, then

$$\lambda^\top u + \nu^\top v \ge 0$$

for all $(u, v, t) \in \mathcal{A}$. Since u can be increased arbitrarily while remaining in $\mathcal{A}$, this forces $\lambda = 0$. Since v ranges over an affine space, we also obtain $\nu = 0$, contradicting the fact that $(\lambda, \nu, \mu) \neq 0$. Thus $\mu > 0$.

Dividing the separating inequality by μ, define

$$\tilde{\lambda} = \frac{\lambda}{\mu}, \qquad \tilde{\nu} = \frac{\nu}{\mu}.$$

Then

$$t + \tilde{\lambda}^\top u + \tilde{\nu}^\top v \ge \alpha$$

for all $(u, v, t) \in \mathcal{A}$.

Now let x be arbitrary. Since

$$(g(x), Ax - b, f(x)) \in \mathcal{A},$$

we obtain

$$f(x) + \tilde{\lambda}^\top g(x) + \tilde{\nu}^\top (Ax - b) \ge \alpha.$$

That is,

$$L(x, \tilde{\lambda}, \tilde{\nu}) \ge \alpha, \qquad \forall x.$$

Taking the infimum over x,

$$q(\tilde{\lambda}, \tilde{\nu}) = \inf_x L(x, \tilde{\lambda}, \tilde{\nu}) \ge \alpha.$$

Next, we show that $\tilde{\lambda} \ge 0$. Let e_i denote the i-th standard basis vector of $\mathbb{R}^m$. Since $\mathcal{A}$ is monotone in the u-coordinates, for every $s > 0$,

$$(u + se_i, v, t) \in \mathcal{A} \qquad \text{whenever} \qquad (u, v, t) \in \mathcal{A}.$$

Applying the separating inequality and letting $s \to \infty$ yields

$$\tilde{\lambda}_i \ge 0.$$

Hence

$$\tilde{\lambda} \geq 0.$$

Therefore $(\tilde{\lambda}, \tilde{\nu})$ is dual feasible, and

$$d^* = \sup_{\lambda \geq 0, \nu} q(\lambda, \nu) \geq q(\tilde{\lambda}, \tilde{\nu}) \geq \alpha.$$

Since the above holds for every $\alpha < p^*$, letting $\alpha \uparrow p^*$ gives

$$d^* \geq p^*.$$

Combining this with weak duality,

$$d^* \leq p^*,$$

we conclude that

$$d^* = p^*.$$

Thus strong duality holds. □

KKT Conditions

If strong duality holds, x^* optimal iff there exist λ^*, ν^*:

$$\begin{aligned} \nabla f(x^*) + \sum_i \lambda_i^* \nabla g_i(x^*) + A^\top \nu^* &= 0, \\ g_i(x^*) &\leq 0, \\ \lambda_i^* &\geq 0, \\ \lambda_i^* g_i(x^*) &= 0. \end{aligned}$$

1. Primal Feasibility. The condition

$$g_i(x^*) \leq 0, \qquad Ax^* = b$$

simply states that x^* satisfies the original constraints of the optimization problem. In other words, x^* belongs to the feasible region.

Geometrically, the feasible set is the region obtained by intersecting the sets

$$\{x : g_i(x) \leq 0\}$$

and the affine subspace

$$\{x : Ax = b\}.$$

Any candidate optimum must lie inside this region. Thus primal feasibility merely says that x^* is an admissible point.

2. Dual Feasibility. The condition

$$\lambda_i^* \geq 0$$

requires the Lagrange multipliers associated with inequality constraints to be nonnegative.

This condition originates from weak duality. Indeed, for a feasible point x,

$$g_i(x) \leq 0.$$

Hence

$$\lambda_i g_i(x) \leq 0$$

whenever $\lambda_i \geq 0$. This guarantees that the Lagrangian

$$L(x, \lambda, \nu) = f(x) + \sum_i \lambda_i g_i(x) + \nu^\top (Ax - b)$$

remains a lower bound on the primal objective when evaluated at feasible points.

Geometrically, the multiplier λ_i measures how strongly the corresponding constraint influences the optimum. The nonnegativity requirement means that constraints can only *push back* against the objective and cannot create an artificial improvement by rewarding constraint violations.

3. Stationarity. The stationarity condition is

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + A^\top \nu^* = 0.$$

This condition is the constrained analogue of the familiar first-order condition

$$\nabla f(x^*) = 0$$

from unconstrained optimization.

To understand its meaning, recall that $\nabla f(x^*)$ points in the direction of greatest increase of the objective function. If there were no constraints, the optimum would occur at a point where this gradient vanishes. In the presence of constraints, however, the optimum may occur on the boundary of the feasible set. At such a point, the gradient of the objective need not be zero because moving in the direction of steepest descent may violate the constraints. The vectors

$$\nabla g_i(x^*)$$

are normal vectors to the constraint surfaces

$$g_i(x) = 0,$$

while the vectors in the range of $A^\top$ are normal to the affine equality constraints.

The stationarity condition therefore states that the negative objective gradient can be represented as a linear combination of the normals to the active constraints:

$$-\nabla f(x^*) = \sum_i \lambda_i^* \nabla g_i(x^*) + A^\top \nu^*.$$

Geometrically, the forces generated by the constraints exactly balance the tendency of the objective function to decrease. Consequently there is no feasible first-order direction along which the objective can be improved.

4. Complementary Slackness. The complementary slackness condition is

$$\lambda_i^* g_i(x^*) = 0, \qquad i = 1, \ldots, m.$$

Since

$$\lambda_i^* \geq 0$$

and

$$g_i(x^*) \leq 0,$$

their product can vanish only if at least one of the factors is zero.

Thus for each constraint exactly one of the following situations occurs:

1. The constraint is *inactive*:
$$g_i(x^*) < 0.$$
In this case
$$\lambda_i^* = 0.$$
The constraint does not influence the optimum.

2. The constraint is *active*:
$$g_i(x^*) = 0.$$
In this case the multiplier may be positive, indicating that the constraint is restricting the optimizer.

Complementary slackness will later correspond to Bellman optimality conditions in MDP LP formulations.

1.5 Linear Programming

A linear program:

$$\min_x c^\top x \quad \text{s.t. } Ax \geq b.$$

Dual:

$$\max_y b^\top y \quad \text{s.t. } A^\top y = c,\ y \geq 0.$$

Theorem 1.7 (Strong Duality for LP)**.** *If either primal or dual has a finite optimal solution, so does the other, and optimal values coincide.*

Proof. Follows from proof of **Theorem 1.6** □

Theorem 1.8 (Complementary Slackness for LP)**.** *At optimality:*

$$y_i^*(Ax^* - b)_i = 0.$$

Proof. Clear from the description of ***KKT Conditions*** for *convex optimization.* □

This structure will be central in identifying occupancy measures as dual variables in the MDP linear program.

2 Foundations of Markov Decision Processes

This section develops the theory of finite discounted Markov Decision Processes (MDPs) in a mathematically rigorous manner. We establish the Bellman equations, contraction properties, existence of optimal stationary policies, and convergence of classical dynamic programming algorithms.

2.1 Finite Markov Decision Processes

Definition 2.1 (Finite Discounted MDP)**.** *A finite discounted MDP is a tuple*

$$(\mathcal{S}, \mathcal{A}, P, r, \gamma),$$

where:

- $\mathcal{S}$ *is a finite state space with* $|\mathcal{S}| = n$*,*
- $\mathcal{A}$ *is a finite action space,*
- $P(s'|s,a)$ *are transition probabilities satisfying*

$$P(s'|s,a) \geq 0, \quad \sum_{s'} P(s'|s,a) = 1,$$

- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is a bounded reward function,*
- $\gamma \in (0,1)$ *is a discount factor.*

We assume rewards are uniformly bounded:

$$|r(s,a)| \leq R_{\max}.$$

2.2 Policies

Definition 2.2. *A* policy π *defines the agent's behavior. It can be one of:*

- ***Stochastic policy:*** *A mapping $\pi(a \mid x)$ that gives a probability distribution over actions $a \in A$ for each state x.*
- ***Stationary policy:*** *The policy is time-independent, i.e. the action choice depends only on the current state, not on t. Stationary policies can be deterministic or stochastic. In other words, a stationary (possibly randomized) policy is a mapping*

$$\pi(a|s) \geq 0, \quad \sum_a \pi(a|s) = 1.$$

- ***Deterministic policy:*** *A mapping $\pi : X \to A$ that assigns a single action $a = \pi(x)$ to each state x. In a nutshell, a deterministic policy satisfies $\pi(a|s) \in \{0, 1\}$.*
- ***Non-stationary (time-dependent) policy:*** *The policy may change over time, denoted $\pi_t(a \mid x)$ or $\pi_t(x)$ for each time step t.*

Under a fixed policy π, the induced transition matrix is

$$P^\pi(s'|s) = \sum_a \pi(a|s) P(s'|s, a),$$

and the expected one-step reward:

$$r^\pi(s) = \sum_a \pi(a|s) r(s, a).$$

In most analysis, we focus on stationary policies, since for a discounted infinite-horizon MDP there always exists an optimal deterministic stationary policy. The set of all (stationary) policies is denoted Π. For a given policy π, the MDP becomes a Markov chain with transition probabilities $P_\pi(y \mid x) = P(y \mid x, \pi(x))$ (in the deterministic case) and expected rewards accordingly.

2.3 Value Functions

Definition 2.3 (Value Function)**.** *For policy π, define*

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right].$$

Since rewards are bounded by R_{max} and $\gamma < 1$, the series converges absolutely:

$$|V^\pi(s)| \leq \frac{R_{\max}}{1 - \gamma}.$$

2.4 Bellman Equation for a Fixed Policy

Proposition 2.1 (Policy Evaluation Equation)**.** *For any stationary policy π,*

$$V^\pi = r^\pi + \gamma P^\pi V^\pi.$$

Proof. Condition on the first step:

$$V^\pi(s) = \mathbb{E}^\pi[r(s, a_0) + \gamma V^\pi(s_1)] = r^\pi(s) + \gamma \sum_{s'} P^\pi(s'|s) V^\pi(s') = r^\pi + \gamma P^\pi V^\pi$$

□

Theorem 2.1.

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

Proof. We have

$$V^\pi (I - \gamma P^\pi) = r^\pi$$

Now, P^π being a *stochastic* matrix, we have $|\lambda| \leq 1$, for all eigenvalues λ of P^π. Now, eigenvalues of $(I - \gamma P^\pi)$ are of the form $1 - \gamma\lambda$. Say, there exists an eigenvalue λ of P^π such that

$$1 - \gamma\lambda = 0 \implies \lambda = \frac{1}{\gamma} > 1 \Rightarrow\Leftarrow$$

Therefore all the eigenvalues of the matrix $(I - \gamma P^\pi)$ are non-zero. Therefore the matrix $(I - \gamma P^\pi)$ is invertible, and we have

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$$

□

2.5 Optimal Value Function

Definition 2.4 (Optimal Value Function)**.**

$$V^*(s) = \sup_\pi V^\pi(s).$$

Definition 2.5 (Bellman Optimality Operator)**.**

$$(TV)(s) = \max_a \left[r(s, a) + \gamma \sum_{s'} P(s'|s, a) V(s') \right].$$

Theorem 2.2 (Bellman Optimality Equation)**.**

$$V^* = TV^*.$$

Proof. For any policy π,

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s').$$

Taking supremum over policies yields $V^* \leq TV^*$.

Conversely, fix a state s. Choose a greedy action

$$a^*(s) \in \arg\max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right\}.$$

Then, by definition of the Bellman optimality operator,

$$(TV^*)(s) = r(s, a^*(s)) + \gamma \sum_{s'} P(s'|s, a^*(s)) V^*(s').$$

Now consider the following strategy: At time 0, take action $a^*(s)$; after the next state S_1 is observed, follow an optimal policy from S_1 onward. The expected return of this strategy is

$$r(s, a^*(s)) + \gamma \sum_{s'} P(s'|s, a^*(s)) V^*(s'),$$

which is exactly $(TV^*)(s)$. On the other hand,

$$V^*(s) = \sup_{\pi} V^{\pi}(s)$$

is the optimal value achievable from state s, i.e., the largest expected return over all admissible strategies.

Since the strategy described above is one particular admissible strategy, its value cannot exceed the optimal value. Therefore,

$$(TV^*)(s) \leq V^*(s).$$

Since s was arbitrary, we conclude that

$$TV^* \leq V^*.$$

Combining this with the previously established inequality we obtain

$$V^* = TV^*.$$

□

2.6 Contraction Mapping Theorem

Let $\mathbb{R}^{|\mathcal{S}|}$ be equipped with $\|\cdot\|_\infty$.

Theorem 2.3. *The Bellman operator T is a γ-contraction:*

$$\|TV - TW\|_\infty \leq \gamma \|V - W\|_\infty.$$

Proof. For each s,

$$|TV(s) - TW(s)| \leq \gamma \max_a \sum_{s'} P(s'|s,a)|V(s') - W(s')| \leq \gamma \|V - W\|_\infty.$$

□

Corollary 2.1 (Due to Banach Fixed Point Theorem)**.** *T admits a unique fixed point V^*.*

2.7 Value Iteration

Value iteration is an iterative dynamic programming method to compute V^*. It starts with an initial guess $V_0(x)$ (often 0) and repeatedly applies the *Bellman optimality backup*. In equation form:

$$V_{k+1}(x) = \max_{a \in A} \Big\{ r(x,a) + \gamma \sum_y P(y \mid x,a)\, V_k(y) \Big\}, \quad \forall x.$$

where $V_{k+1}(x)$ is the updated value, $r(x,a)$ is the immediate reward, γ is the discount factor, $P(y \mid x,a)$ is the transition probability, $V_k(y)$ is the value at the previous iteration, and A is the set of actions.

As $k \to \infty$, $V_k(x) \to V^*(x)$ for all x, under standard conditions. One can stop when the change $\|V_{k+1} - V_k\|_\infty$ is below a tolerance. After convergence, an optimal policy is $\pi^*(x) = \arg\max_a \big[r(x,a) + \gamma \sum_y P(y \mid x,a) V_k(y)\big]$.

Algorithm 1 Value Iteration

Initialize $V_0(x)$ for all $x \in \mathcal{X}$ (e.g., $V_0(x) = 0$).
Set $k \leftarrow 0$.
repeat
    for all $x \in \mathcal{X}$ **do**

$$V_{k+1}(x) \leftarrow \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{y \in \mathcal{X}} P(y \mid x,a)\, V_k(y) \right\}.$$

    end for
    $k \leftarrow k+1$.
until

$$\max_{x \in \mathcal{X}} |V_k(x) - V_{k-1}(x)| < \varepsilon$$

Output V_k as an approximation of V^*.
for all $x \in \mathcal{X}$ **do**

$$\pi^*(x) \leftarrow \arg\max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{y \in \mathcal{X}} P(y \mid x,a)\, V_k(y) \right\}.$$

end for
Output policy π^*.

Theorem 2.4 (Convergence Rate)**.**

$$\|V_k - V^*\|_\infty \leq \gamma^k \|V_0 - V^*\|_\infty.$$

Proof.

$$||V_k - V^*||_\infty = ||TV_{k-1} - TV^*||_\infty \leq \gamma ||V_{k-1} - V^*||_\infty = \cdots \leq \gamma^k ||V_0 - V^*||_\infty$$

□

2.8 Policy Iteration

Definition 2.6. *Policy iteration consists of:*

1. *Policy evaluation: solve* V^{π_k}.
2. *Policy improvement:*

$$\pi_{k+1}(s) \in \arg\max_a \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi_k}(s') \right].$$

Theorem 2.5 (Policy Improvement Theorem)**.**

$$V^{\pi_{k+1}} \geq V^{\pi_k}.$$

Proof. Let π' be greedy w.r.t. V^π. Then

$$T^{\pi'}V^\pi \geq T^\pi V^\pi = V^\pi.$$

Monotonicity of $T^{\pi'}$ implies $V^{\pi'} \geq V^\pi$. □

Theorem 2.6. *Policy iteration converges in finitely many steps to an optimal deterministic policy.*

Proof. There are finitely many deterministic policies. Each improvement strictly increases value unless optimal. □

2.9 Existence of Optimal Deterministic Policies

Theorem 2.7. *In finite discounted MDPs, there exists an optimal deterministic stationary policy.*

Proof. Let V^* denote the optimal value function. By the Bellman optimality theorem, V^* satisfies

$$V^*(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)V^*(s') \right\}, \qquad \forall s \in \mathcal{S}.$$

Since the action set is finite, the maximum is attained. Therefore, for each state s, we may select an action

$$\pi^*(s) \in \arg\max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)V^*(s') \right\}.$$

This defines a deterministic stationary policy π^*.

By construction,

$$V^*(s) = r(s,\pi^*(s)) + \gamma \sum_{s'} P(s'|s,\pi^*(s))V^*(s'), \qquad \forall s.$$

In vector form,

$$V^* = r^{\pi^*} + \gamma P^{\pi^*} V^*.$$

On the other hand, the value function of the policy π^* is the unique solution of the Bellman equation

$$V^{\pi^*} = r^{\pi^*} + \gamma P^{\pi^*} V^{\pi^*}.$$

Subtracting the two equations gives

$$V^* - V^{\pi^*} = \gamma P^{\pi^*}\big(V^* - V^{\pi^*}\big).$$

Rearranging,

$$(I - \gamma P^{\pi^*})\big(V^* - V^{\pi^*}\big) = 0.$$

Since $0 < \gamma < 1$, the matrix

$$I - \gamma P^{\pi^*}$$

is invertible. Therefore,

$$V^* - V^{\pi^*} = 0,$$

and hence

$$V^{\pi^*} = V^*.$$

Thus the deterministic stationary policy π^* achieves the optimal value function in every state. Consequently,

$$V^{\pi^*}(s) = V^*(s) = \sup_{\pi} V^{\pi}(s), \qquad \forall s \in \mathcal{S}.$$

Therefore π^* is an optimal deterministic stationary policy. □

These results prepare the ground for reformulating MDPs as linear programs and connecting them to convex duality theory in the subsequent weeks.

3 Linear Programming Formulation of MDPs

In this section we reformulate finite discounted MDPs as linear programs (LPs), derive their duals rigorously, interpret dual variables as discounted occupancy measures, and establish equivalence with Bellman optimality conditions via complementary slackness.

Throughout, assume a finite discounted MDP

$$(\mathcal{S}, \mathcal{A}, P, r, \gamma), \quad \gamma \in (0, 1),$$

with $|\mathcal{S}| = n$, $|\mathcal{A}| = m$, and bounded rewards.

3.1 Bellman Inequalities and Primal Linear Program

Recall the Bellman optimality equation:

$$V^*(s) = \max_a \left[r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^*(s') \right].$$

Equivalently, V^* is the smallest function satisfying the Bellman inequalities.

Definition 3.1 (Bellman Feasible Set)**.** *Define*

$$\mathcal{V} = \left\{ V \in \mathbb{R}^n : V(s) \geq r(s, a) + \gamma \sum_{s'} P(s'|s, a) V(s'),\ \forall s, a \right\}.$$

Note: $\mathcal{V}$ is a subset of real valued function defined on $\mathcal{S}$. This vector space is isomorphic to $\mathbb{R}^n$, and the basis of the original vector spaces can be thought as $\{\varphi_i\}_{i=1}^n$ with the property $\varphi_i(e_j) = \delta_{ij}$, where e_i's are the *canonical bases* of $\mathbb{R}^n$ and δ_{ij} is the *Kronecker delta function.*

Proposition 3.1. *$\mathcal{V}$ is a nonempty closed convex polyhedron.*

Proof. For each $(s, a) \in S \times A$, define

$$c_{s,a}(i) = \delta_{si} - \gamma P(i|s, a), \qquad i \in S,$$

where δ_{si} is the *Kronecker delta function.* Then

$$c_{s,a}^\top V = V(s) - \gamma \sum_{s'} P(s'|s, a)V(s').$$

Hence the Bellman inequality can be written as

$$c_{s,a}^\top V \geq r(s, a).$$

Therefore

$$\mathcal{V} = \bigcap_{(s,a)\in S\times A} \left\{ V \in \mathbb{R}^n : c_{s,a}^\top V \geq r(s, a) \right\}.$$

Each set

$$H_{s,a} = \left\{ V \in \mathbb{R}^n : c_{s,a}^\top V \geq r(s, a) \right\}$$

is a closed half-space.

Let $V_1, V_2 \in \mathcal{V}$ and let $\lambda \in [0, 1]$. Since $V_1, V_2 \in \mathcal{V}$,

$$c_{s,a}^\top V_1 \geq r(s, a), \qquad c_{s,a}^\top V_2 \geq r(s, a)$$

for every (s, a). Multiplying by λ and $1 - \lambda$, respectively, and adding yields

$$c_{s,a}^\top \big(\lambda V_1 + (1 - \lambda)V_2\big) = \lambda c_{s,a}^\top V_1 + (1 - \lambda)c_{s,a}^\top V_2 \geq r(s, a).$$

Thus

$$\lambda V_1 + (1 - \lambda)V_2 \in \mathcal{V},$$

and therefore $\mathcal{V}$ is convex.

For each (s, a), the map

$$V \mapsto c_{s,a}^\top V$$

is continuous on $\mathbb{R}^n$. Since $[r(s, a), \infty)$ is closed,

$$H_{s,a} = (c_{s,a}^\top)^{-1}\big([r(s, a), \infty)\big)$$

is closed. Because S and A are finite,

$$\mathcal{V} = \bigcap_{(s,a)\in S\times A} H_{s,a}$$

is a finite intersection of closed sets and hence is closed.

Finally, $\mathcal{V}$ is the intersection of finitely many affine half-spaces, namely $|S||A|$ half-spaces. Therefore $\mathcal{V}$ is a polyhedron. Consequently, $\mathcal{V}$ is a closed convex polyhedron. □

Note: In case of *Borel state space* and *compact action space*, we consider $\mathcal{V}$ to be a subset of the *appropriate* Banach Space under sup norm. The set is *closed* and *convex*, but not necessarily a *polyhedron*.

3.2 Primal LP Formulation

Fix any strictly positive distribution $\mu \in \mathbb{R}^n$ with $\mu(s) > 0$.

Consider the linear program:

$$\min_{V \in \mathbb{R}^n} \sum_s \mu(s) V(s) \tag{P}$$

subject to,

$$V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s'), \quad \forall s, a.$$

Theorem 3.1. *The optimal solution of* (P) *is* V^*.

Proof. Step 1: $V^* \in \mathcal{V}$ since it satisfies Bellman optimality equality.

Step 2: Minimality. Let $V \in \mathcal{V}$. Then for all s,

$$V(s) \geq \max_a \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \right] = (TV)(s).$$

Thus $V \geq TV$ componentwise.

Applying T repeatedly:

$$V \geq TV \geq T^2 V \geq \dots$$

Since $T^k V \to V^*$ (contraction property), taking limits yields

$$V \geq V^*.$$

Therefore

$$\sum_s \mu(s) V(s) \geq \sum_s \mu(s) V^*(s).$$

Since V^* is feasible, it is optimal. □

Thus V^* is the minimal element of $\mathcal{V}$ under the partial order induced by $\mathbb{R}^n_+$.

3.3 Derivation of the Dual Program

Rewrite constraints in standard LP form:

For each (s, a),

$$V(s) - \gamma \sum_{s'} P(s'|s, a)V(s') \geq r(s, a).$$

Introduce dual variables $d(s, a) \geq 0$ for each constraint.

The Lagrangian is

$$\mathcal{L}(V, d) = \sum_{s} \mu(s)V(s) + \sum_{s,a} d(s, a) \left(r(s, a) - V(s) + \gamma \sum_{s'} P(s'|s, a)V(s') \right).$$

Rearranging terms in V:

$$\mathcal{L}(V, d) = \sum_{s,a} d(s, a)r(s, a) + \sum_{s} V(s) \left[\mu(s) - \sum_{a} d(s, a) + \gamma \sum_{s',a'} d(s', a')P(s|s', a') \right].$$

The dual function is finite only if the coefficient of $V(s)$ vanishes:

$$\mu(s) - \sum_{a} d(s, a) + \gamma \sum_{s',a'} d(s', a')P(s|s', a') = 0.$$

Thus the dual is:

$$\max_{d \geq 0} \sum_{s,a} d(s, a)r(s, a) \tag{D}$$

subject to

$$\sum_{a} d(s, a) = \mu(s) + \gamma \sum_{s',a'} d(s', a')P(s|s', a'), \quad \forall s.$$

Theorem 3.2 (Strong Duality)**.** *Since (P) is feasible and bounded, strong duality holds:*

$$\min(P) = \max(D).$$

3.4 Discounted Occupancy Measures

Definition 3.2 (Discounted Occupancy Measure)**.** *For policy π, define*

$$d^{\pi}(s, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^{\pi}(s_t = s, a_t = a).$$

This basically gives us the expected number of visits to the state-action pair (s, a).

Proposition 3.2. *d^π satisfies*

$$\sum_a d^\pi(s,a) = (1-\gamma)\mu_\pi(s),$$

where μ_π is the discounted state distribution.

Theorem 3.3. *d^π satisfies the dual feasibility constraints with μ equal to the initial state distribution scaled by $(1-\gamma)$.*

Proof. Using the Markov property,

$$\Pr(s_t = s) = \sum_{s',a'} \Pr(s_{t-1} = s', a_{t-1} = a')P(s|s',a').$$

Multiply by $(1-\gamma)\gamma^t$ and sum over t to obtain the flow balance equation:

$$\begin{aligned}
&\sum_{t=0}^{\infty}(1-\gamma)\gamma^t \Pr(s_t = s) = (1-\gamma)\sum_{t=0}^{\infty}\sum_{s',a'}\gamma^t \Pr(s_{t-1}=s', a_{t-1}=a)P(s \mid s', a') \\
\implies &\Pr(s_t = s) = (1-\gamma)P(s_0) + (1-\gamma)\gamma\sum_{t=1}^{\infty}\sum_{s',a'}\gamma^{t-1}\Pr(s_{t-1}=s', a_{t-1}=a)P(s \mid s', a') \\
\implies &\sum_a d^\pi(s,a) = (1-\gamma)\mu_0(s) + \gamma\sum_{s',a'} d^\pi(s',a')P(s \mid s', a')
\end{aligned}$$

□

Thus feasible dual variables correspond exactly to valid discounted state-action flows.

3.5 Policy Recovery from Dual Variables

Theorem 3.4. *Let d^* be an optimal solution of (D). Define*

$$\pi^*(a|s) = \frac{d^*(s,a)}{\sum_{a'} d^*(s,a')}.$$

Then π^* is optimal.

Proof. Complementary slackness implies:

$$d^*(s,a)\left[V^*(s) - r(s,a) - \gamma\sum_{s'} P(s'|s,a)V^*(s')\right] = 0.$$

Thus if $d^*(s,a) > 0$, the corresponding action attains the Bellman maximum. Hence π^* is greedy w.r.t. V^* and therefore optimal. □

3.6 Geometric Interpretation

The primal feasible set $\mathcal{V}$ is a polyhedron. The dual feasible set is a polytope of discounted flows.

Optimality corresponds to a saddle point of the Lagrangian:

$$\mathcal{L}(V^*, d) \leq \mathcal{L}(V^*, d^*) \leq \mathcal{L}(V, d^*).$$

Thus the MDP solution is characterized by a primal-dual equilibrium.

4 Extensions

In this final part of the project we study extensions of the linear programming and convex-analytic perspective on MDPs. We treat:

- Constrained Markov Decision Processes (CMDPs),
- Entropy-regularized (soft) MDPs,
- Convex-analytic and saddle-point formulations,
- Mirror descent and primal-dual interpretations.

Throughout we assume a finite discounted MDP with $\gamma \in (0, 1)$.

4.1 Constrained Markov Decision Processes

We consider additional performance constraints beyond reward maximization.

Problem Formulation

Let $c_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $i = 1, \ldots, k$, be bounded cost functions and $d_i \in \mathbb{R}$ given thresholds.

We define the discounted reward and costs:

$$J_r(\pi) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \quad J_{c_i}(\pi) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) \right].$$

The constrained MDP is:

$$\max_{\pi} J_r(\pi) \quad \text{s.t.} \quad J_{c_i}(\pi) \leq d_i, \quad i = 1, \ldots, k. \tag{CMDP}$$

4.2 Occupancy Measure Reformulation

Let $d^\pi(s, a)$ denote the discounted occupancy measure (as derived earlier). Then

$$J_r(\pi) = \frac{1}{1-\gamma} \sum_{s,a} d^\pi(s, a) r(s, a),$$

and similarly

$$J_{c_i}(\pi) = \frac{1}{1-\gamma} \sum_{s,a} d^\pi(s, a) c_i(s, a).$$

Hence CMDP becomes a linear program over occupancy measures:

$$\max_{d \geq 0} \sum_{s,a} d(s, a) r(s, a)$$

subject to

$$\sum_{a} d(s, a) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} d(s', a') P(s|s', a'), \quad \forall s,$$

$$\sum_{s,a} d(s, a) c_i(s, a) \leq (1-\gamma) d_i, \quad i = 1, \ldots, k.$$

Theorem 4.1. *The CMDP is equivalent to a linear program in d.*

Proof. Every stationary policy induces a feasible d^π satisfying flow constraints. Conversely, any feasible d defines a stationary policy

$$\pi(a|s) = \frac{d(s, a)}{\sum_{a'} d(s, a')}.$$

The reward and constraint values coincide via the occupancy representation. □

4.3 Lagrangian Formulation

Define Lagrange multipliers $\lambda_i \geq 0$ and Lagrangian:

$$\mathcal{L}(\pi, \lambda) = J_r(\pi) - \sum_{i=1}^{k} \lambda_i \left(J_{c_i}(\pi) - d_i \right).$$

Define the dual function:

$$g(\lambda) = \sup_{\pi} \mathcal{L}(\pi, \lambda).$$

Theorem 4.2. *Strong duality holds under Slater-type feasibility assumptions.*

Proof. By Theorem 4.1, the CMDP is equivalent to the occupancy-measure linear program

Since the state and action spaces are finite, it is a finite-dimensional linear program. The Slater condition implies that the feasible region is nonempty and contains a point satisfying all inequality constraints strictly. Hence the primal LP is feasible.

Furthermore, because rewards are bounded and discounted occupancy measures satisfy

$$\sum_{s,a} d(s,a) = 1,$$

the objective function is bounded above on the feasible region. Therefore the primal LP has a finite optimal value.

Introduce Lagrange multipliers $\lambda_i \geq 0$ for the inequality constraints. The Lagrangian is

$$\mathcal{L}(d, \lambda) = \sum_{s,a} d(s,a) r(s,a) - \sum_{i=1}^{k} \lambda_i \left(\sum_{s,a} d(s,a) c_i(s,a) - (1-\gamma) d_i \right).$$

Rearranging terms yields

$$\mathcal{L}(d, \lambda) = \sum_{s,a} d(s,a) \left(r(s,a) - \sum_{i=1}^{k} \lambda_i c_i(s,a) \right) + (1-\gamma) \sum_{i=1}^{k} \lambda_i d_i.$$

The associated dual function is

$$g(\lambda) = \sup_{d \text{ satisfying flow constraints}} \mathcal{L}(d, \lambda).$$

Using the correspondence between occupancy measures and stationary policies, this can be written as

$$g(\lambda) = \sup_{\pi} \left[J_r(\pi) - \sum_{i=1}^{k} \lambda_i \big(J_{c_i}(\pi) - d_i \big) \right] = \sup_{\pi} L(\pi, \lambda).$$

Since the primal problem is a feasible finite-dimensional linear program with finite optimal value, the strong duality theorem of linear programming implies that the optimal values of the primal and dual problems coincide. Consequently,

$$\sup_{\pi} \Big\{ J_r(\pi) : J_{c_i}(\pi) \leq d_i,\ i = 1, \ldots, k \Big\} = \inf_{\lambda \geq 0} g(\lambda).$$

Therefore,

$$\max_{\pi} \min_{\lambda \geq 0} L(\pi, \lambda) = \min_{\lambda \geq 0} \max_{\pi} L(\pi, \lambda).$$

This proves the theorem. □

4.4 Entropy-Regularized (Soft) MDPs

We now regularize the control problem via entropy.

Regularized Objective

Define the Shannon entropy at state s:

$$H(\pi(\cdot|s)) = -\sum_{a} \pi(a|s) \log \pi(a|s).$$

The entropy-regularized objective:

$$J_\tau(\pi) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \tau H(\pi(\cdot|s_t)) \right) \right],$$

where $\tau > 0$.

4.5 Soft Bellman Operator

Define:

$$(T_\tau V)(s) = \tau \log \sum_a \exp\left(\frac{1}{\tau}\left[r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')\right]\right).$$

Theorem 4.3. *T_τ is a γ-contraction under $\|\cdot\|_\infty$.*

Proof. The $\log \sum \exp$ function is 1-Lipschitz with respect to the ℓ_∞ norm. Hence

$$|T_\tau V(s) - T_\tau W(s)| \leq \gamma \|V - W\|_\infty.$$

□

Corollary 4.1. *There exists a unique V_τ^* satisfying*

$$V_\tau^* = T_\tau V_\tau^*.$$

4.6 Convex Dual Interpretation

The soft Bellman operator arises from convex conjugacy.

Recall the Fenchel conjugate of negative entropy:

$$\left(\tau \sum_a \pi(a) \log \pi(a)\right)^* = \tau \log \sum_a e^{x_a/\tau}.$$

Thus the soft Bellman update equals:

$$\max_{\pi(\cdot|s)} \left\{ \sum_a \pi(a|s) Q_V(s,a) + \tau H(\pi(\cdot|s)) \right\},$$

where

$$Q_V(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s').$$

Theorem 4.4. *The maximizing policy is Gibbs (Boltzmann):*

$$\pi_\tau^*(a|s) = \frac{\exp(Q_{V_\tau^*}(s,a)/\tau)}{\sum_{a'} \exp(Q_{V_\tau^*}(s,a')/\tau)}.$$

Proof. Fix a state s and define

$$q_a := Q_V(s,a), \qquad \pi_a := \pi(a|s).$$

The optimization problem becomes

$$\max_{\pi} \left\{ \sum_a \pi_a q_a - \tau \sum_a \pi_a \log \pi_a \right\}$$

subject to

$$\sum_a \pi_a = 1, \qquad \pi_a \geq 0.$$

The objective is strictly concave since the entropy term is strictly concave on the probability simplex. Hence any stationary point is the unique maximizer.

Introduce a Lagrange multiplier λ for the normalization constraint $\sum_a \pi_a = 1$. The Lagrangian is

$$\mathcal{L}(\pi, \lambda) = \sum_a \pi_a q_a - \tau \sum_a \pi_a \log \pi_a + \lambda \left(\sum_a \pi_a - 1 \right).$$

For each action a at optimum,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi_a} &= q_a - \tau(\log \pi_a + 1) + \lambda = 0 \\ \implies & q_a - \tau(\log \pi_a + 1) + \lambda = 0 \\ \implies & \log \pi_a = \frac{q_a}{\tau} + \frac{\lambda - \tau}{\tau} \\ \implies & \pi_a = \exp\left(\frac{\lambda - \tau}{\tau} \right) \exp\left(\frac{q_a}{\tau} \right). \end{aligned}$$

Let

$$C = \exp\left(\frac{\lambda - \tau}{\tau} \right).$$

Then

$$\pi_a = C\, e^{q_a/\tau}.$$

Using the normalization condition,

$$1 = \sum_a \pi_a = C \sum_a e^{q_a/\tau},$$

so

$$C = \frac{1}{\sum_a e^{q_a/\tau}}.$$

Substituting back,

$$\pi_a = \frac{e^{q_a/\tau}}{\sum_{a'} e^{q_{a'}/\tau}}.$$

Recalling that $q_a = Q_V(s,a)$, we obtain

$$\pi_\tau^*(a|s) = \frac{\exp(Q_V(s,a)/\tau)}{\sum_{a'} \exp(Q_V(s,a')/\tau)}.$$

Plugging the value of $\pi_\tau^*(\cdot|s)$ in the Lagrangian, we get the *Fenchel conjugate* of the *negative entropy*.

Since the objective is strictly concave, this stationary point is the unique maximizer. □

4.7 Saddle-Point and Mirror Descent Interpretation

Define policy space Π (product of simplices). This can be done as $\pi(a \mid s) \geq 0 \; \forall a, s$ and $\sum_a \pi(a \mid s) = 1$. The regularized problem can be written:

$$\max_{\pi \in \Pi} \min_{V} \mathcal{L}_\tau(\pi, V),$$

where

$$\mathcal{L}_\tau(\pi, V) = \sum_{s,a} d^\pi(s,a) \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') - V(s) \right] + \tau \sum_s H(\pi(\cdot|s)).$$

This is a convex-concave saddle problem.

Mirror Descent

Define KL divergence:

$$D_{\mathrm{KL}}(\pi \| \pi_k) = \sum_s \sum_a \pi(a|s) \log \frac{\pi(a|s)}{\pi_k(a|s)}.$$

Mirror descent update:

$$\pi_{k+1} = \arg\max_{\pi \in \Pi} \left\{ \langle g_k, \pi \rangle - \frac{1}{\eta} D_{\mathrm{KL}}(\pi \| \pi_k) \right\}.$$

The choice of *mirror descent update* is quite natural because, if our update rule is

$$\pi_{k+1} = \pi_k + \eta g_k$$

Then the iteration might leave the simplex. So, we would like to move towards action with a larger gradient (due to the first term), but we would also like to stay in the vicinity of our old policy (due to the second term).

Theorem 4.5. *The mirror descent update yields exponential weights:*

$$\pi_{k+1}(a|s) \propto \pi_k(a|s) \exp(\eta g_k(s,a)).$$

Proof. Since the optimization separates across states, fix a state s and write

$$\pi_a := \pi(a|s), \qquad \pi_a^k := \pi_k(a|s), \qquad g_a := g_k(s,a).$$

The mirror descent update becomes

$$\max_{\pi} \left\{ \sum_a \pi_a g_a - \frac{1}{\eta} \sum_a \pi_a \log \frac{\pi_a}{\pi_a^k} \right\}$$

subject to

$$\sum_a \pi_a = 1, \qquad \pi_a \geq 0.$$

We introduce a Lagrange multiplier λ for the normalization constraint. The Lagrangian is

$$\mathcal{L}(\pi, \lambda) = \sum_a \pi_a g_a - \frac{1}{\eta} \sum_a \pi_a \log \frac{\pi_a}{\pi_a^k} + \lambda \left(\sum_a \pi_a - 1 \right).$$

Differentiating with respect to π_a,

$$\frac{\partial \mathcal{L}}{\partial \pi_a} = g_a - \frac{1}{\eta} \left(\log \frac{\pi_a}{\pi_a^k} + 1 \right) + \lambda.$$

At an optimum,

$$\frac{\partial \mathcal{L}}{\partial \pi_a} = 0,$$

which yields

$$g_a - \frac{1}{\eta} \left(\log \frac{\pi_a}{\pi_a^k} + 1 \right) + \lambda = 0.$$

Multiplying by η and rearranging,

$$\log \frac{\pi_a}{\pi_a^k} = \eta(g_a + \lambda) - 1.$$

Exponentiating both sides,

$$\pi_a = \pi_a^k e^{(\eta(g_a+\lambda)-1)}$$

Since $\eta(g_a + \lambda) - 1 = \eta g_a + (\eta\lambda - 1)$, we may write

$$\pi_a = C\, \pi_a^k e^{\eta g_a},$$

where

$$C := e^{\eta\lambda - 1}$$

is independent of a.

Using the constraint $\sum_a \pi_a = 1$,

$$1 = C \sum_a \pi_a^k e^{\eta g_a},$$

and therefore

$$C = \frac{1}{\sum_a \pi_a^k e^{\eta g_a}}.$$

Substituting back,

$$\pi_a = \frac{\pi_a^k e^{\eta g_a}}{\sum_b \pi_b^k e^{\eta g_b}}.$$

Returning to the original notation,

$$\pi_{k+1}(a|s) = \frac{\pi_k(a|s) \exp(\eta g_k(s,a))}{\sum_{a'} \pi_k(a'|s) \exp(\eta g_k(s,a'))}.$$

This proves the result. □

Thus entropy-regularized policy optimization corresponds to mirror descent in the space of policies.

5 Q-Learning

Q-learning is a stochastic approximation algorithm used for fully observed finite space Markov Decision Processes (MDPs). Notably, it does not require prior knowledge of the transition kernel or the cost (or reward) function for its implementation. In this algorithm, the per-stage cost is observed via simulation along a single sample path. For finite state and action spaces, and under mild assumptions—such as visiting all state-action pairs infinitely often the algorithm is known to converge to the optimal cost function.

5.1 Q-Function over MDPs

In many Markov Decision Problems (MDPs), the true transition kernel and the cost or reward structures may not be known. In such scenarios, one often relies on historical data to obtain an asymptotically optimal solution, meaning that the solution approaches optimality as more data becomes available. This learning based approach can also serve as an efficient numerical method for computing approximate optimal solutions. Depending on the available knowledge about the system, one may use prior probabilistic information to guide the learning process.

5.2 Q-Function and Bellman Equation

Definition 5.1 (Q-Function)**.** *The Q-function is defined as:*

$$Q(x,a) = r(x,a) + \gamma \sum_{y \in X} P(y \mid x,a) \min_{a'} Q(y,a'). \tag{1}$$

The optimal value function is then given by $V^*(x) = \min_a Q(x,a)$*, and the optimal policy is:*

$$\pi^*(x) = \arg\min_a Q(x,a). \tag{2}$$

5.3 Q-Learning Update Rule

Given a sample transition (x_k, a_k, r_k, x_{k+1}), the Q-learning update is:

$$Q_{k+1}(x_k,a_k) = Q_k(x_k,a_k) + \alpha_k \left[r_k + \gamma \min_{a'} Q_k(x_{k+1},a') - Q_k(x_k,a_k) \right],$$

where $Q_k(x_k,a_k)$ is the Q-value at iteration k, α_k is the learning rate, r_k is the reward, γ is the discount factor, $Q_k(x_{k+1},a')$ is the Q-value for the next state and action, and a' ranges over all possible actions.

5.4 Deterministic Q-Learning

In deterministic environments (where transitions are deterministic functions of state and action), the update simplifies to:

$$Q_{k+1}(x_k,a_k) = r(x_k,a_k) + \gamma \min_{a'} Q_k(f(x_k,a_k),a').$$

5.5 Proof for Convergence of Q Learning Algorithm

Let F be an operator acting on the Q-functions, defined as:

$$F(Q)(x,u) = c(x,u) + \beta \sum_{x'} T(x' \mid x,u) \min_v Q(x',v) \tag{3}$$

The corresponding fixed-point equation is:

$$Q^*(x,u) = F(Q^*)(x,u) = c(x,u) + \beta \sum_{x'} T(x' \mid x,u) \min_v Q^*(x',v) \tag{4}$$

6 Contraction Mapping and Q-Learning Convergence

Definition 6.1 (Contraction Mapping)**.** *Let $F : X \to X$ be a function on a normed space $(X, \|\cdot\|)$. F is called a* contraction mapping *if there exists a constant $\gamma \in [0, 1)$ such that*

$$\|F(x) - F(y)\| \leq \gamma \cdot \|x - y\|, \quad \text{for all } x, y \in X.$$

Here, γ is the contraction constant. This implies F brings points closer together.

Remark 6.1. *A contraction mapping is a function that always brings points closer together. This means that if you keep applying it, you will eventually end up at a single point, no matter where you start. This is why contraction mappings guarantee a unique fixed point.*

Theorem 6.1 (Banach Fixed Point Theorem)**.** *Let $(X, \|\cdot\|)$ be a non-empty complete normed vector space (Banach space), and $F : X \to X$ be a contraction mapping with contraction constant $\gamma \in [0, 1)$. Then:*

1. *There exists a unique fixed point $x^* \in X$ such that $F(x^*) = x^*$.*
2. *For any initial point x_0, the sequence $x_{n+1} = F(x_n)$ converges to x^*.*
3. *The convergence is at least linear:*

$$\|x_n - x^*\| \leq \frac{\gamma^n}{1 - \gamma} \|x_1 - x_0\|.$$

Review 6.1. *The Banach Fixed Point Theorem is a powerful tool in mathematics and is the backbone of many algorithms in reinforcement learning. It tells us that if we have a contraction mapping, we are guaranteed to find a unique solution by repeatedly applying the mapping.*

Theorem 6.2 (Bellman Operator is a Contraction)**.** *The Bellman operator F is a contraction in the sup norm:*

$$|F(Q_t)(x, u) - F(Q^*)(x, u)| \leq \beta \|Q_t - Q^*\|_\infty := \beta \max_{x,u} |Q_t(x, u) - Q^*(x, u)|.$$

Proof. Recall:

$$F(Q)(x, u) = c(x, u) + \beta \sum_{x'} T(x'|x, u) \min_v Q(x', v).$$

Subtract:

$$|F(Q_t)(x, u) - F(Q^*)(x, u)| = \beta \left| \sum_{x'} T(x'|x, u) (\min_v Q_t(x', v) - \min_v Q^*(x', v)) \right|.$$

Apply triangle inequality and inequality:

$$|\min_v f(v) - \min_v g(v)| \leq \max_v |f(v) - g(v)| \leq \|Q_t - Q^*\|_\infty.$$

This gives:

$$|F(Q_t)(x, u) - F(Q^*)(x, u)| \leq \beta \|Q_t - Q^*\|_\infty.$$

□

Q-learning Update Equation:

$$Q_{t+1}(x,u) = Q_t(x,u) + \alpha_t(x,u)\left[F(Q_t)(x,u) - Q_t(x,u) + \omega_t\right], \tag{5}$$

where:

$$\omega_t := c(x_t,u_t) + \beta \min_v Q_t(x_{t+1},v) - F(Q_t)(x_t,u_t)$$

is zero-mean noise.

Zero-Mean Nature of the Noise: Let H_t be the history at time t, including the current state-action pair (x_t,u_t) and the current iterate Q_t. Then

$$\begin{aligned}
\mathbb{E}(\omega_t \mid H_t) &= \mathbb{E}\left[c(x_t,u_t) + \beta \min_v Q_t(x_{t+1},v) - F(Q_t)(x_t,u_t) \mid H_t\right] \\
&= c(x_t,u_t) + \beta \mathbb{E}\left[\min_v Q_t(x_{t+1},v) \mid H_t\right] - F(Q_t)(x_t,u_t) \\
&= \beta\left[\mathbb{E}\left[\min_v Q_t(x_{t+1},v) \mid H_t\right] - \sum_{x'} T(x' \mid x_t,u_t) \min_v Q_t(x',v)\right] \\
&= 0.
\end{aligned}$$

The last equality follows from the definition of the transition probabilities, since conditional on H_t the next state has law $T(\cdot \mid x_t,u_t)$.

Alternative Form:

$$Q_{t+1}(x,u) = (1-\alpha_t(x,u))Q_t(x,u) + \alpha_t(x,u)[F(Q_t)(x,u) + \omega_t]. \tag{6}$$

Let $S_t = Q_t - Q^*$:

$$S_{t+1}(x,u) = (1-\alpha_t(x,u))S_t(x,u) + \alpha_t(x,u)[F(Q_t)(x,u) - F(Q^*)(x,u) + \omega_t]. \tag{7}$$

Split as:

$$S^a_{t+1}(x,u) = (1-\alpha_t(x,u))S^a_t(x,u) + \alpha_t(x,u)\omega_t, \tag{8}$$

$$S^b_{t+1}(x,u) = (1-\alpha_t(x,u))S^b_t(x,u) + \alpha_t(x,u)(F(Q_t)(x,u) - F(Q^*)(x,u)). \tag{9}$$

So:

$$S_{t+1}(x,u) = S^a_{t+1}(x,u) + S^b_{t+1}(x,u).$$

Bellman Error Bound:

$$\delta_t(x,u) = F(Q_t)(x,u) - F(Q^*)(x,u), \tag{10}$$

$$\|\delta_t\|_\infty \le \beta(\|S^a_t\|_\infty + \|S^b_t\|_\infty). \tag{11}$$

Assume $\|S_t^a\|_\infty \leq \epsilon/M$, then:

$$\|S_{t+1}^b(x,u)\| \leq (1-\alpha_t(x,u))\|S_t^b(x,u)\| + \alpha_t(x,u)\beta'\|S_t^b\|_\infty, \tag{12}$$

$$= [1-\alpha_t(x,u)(1-\beta')]\|S_t^b\|_\infty. \tag{13}$$

Let $\gamma = 1-\beta'$, then:

$$\|S_{t+1}^b\|_\infty \leq (1-\gamma\alpha_t)\|S_t^b\|_\infty.$$

Convergence via Grönwall Argument

$$\|S_t^b\|_\infty \leq \left(\prod_{s=0}^{t-1}(1-\gamma\alpha_s)\right)\|S_0^b\|_\infty,$$

$$\log(\cdot) \leq -\gamma\sum_{s=0}^{t-1}\alpha_s \to -\infty \Rightarrow \prod(\cdot) \to 0.$$

So $\|S_t^b\|_\infty \to 0$.

Martingale Argument for the Noise Term

We now show that the noise component S_t^a converges to zero. For notational simplicity, fix a state-action pair and write $\alpha_t = \alpha_t(x,u)$. Define

$$P_t := \prod_{k=0}^{t-1}(1-\alpha_k), \qquad P_0 := 1.$$

Since $P_{t+1} = (1-\alpha_t)P_t$, the recursion for S_t^a can be written as

$$S_{t+1}^a = \frac{P_{t+1}}{P_t}S_t^a + \alpha_t\omega_t.$$

Dividing by P_{t+1} gives

$$\frac{S_{t+1}^a}{P_{t+1}} = \frac{S_t^a}{P_t} + \frac{\alpha_t}{P_{t+1}}\omega_t.$$

Let

$$Y_t := \frac{S_t^a}{P_t}.$$

Then

$$Y_{t+1} = Y_t + \frac{\alpha_t}{P_{t+1}}\omega_t,$$

and iterating yields

$$Y_t = Y_0 + \sum_{j=0}^{t-1}\frac{\alpha_j}{P_{j+1}}\omega_j.$$

Thus

$$S_t^a = P_t S_0^a + P_t \sum_{j=0}^{t-1} \frac{\alpha_j}{P_{j+1}} \omega_j.$$

Define

$$M_t := \sum_{j=0}^{t-1} \frac{\alpha_j}{P_{j+1}} \omega_j.$$

Then

$$S_t^a = P_t S_0^a + P_t M_t.$$

First, $P_t \to 0$. Indeed,

$$\log P_t = \sum_{k=0}^{t-1} \log(1 - \alpha_k).$$

Using $\log(1 - x) \leq -x$ for $0 < x < 1$,

$$\log P_t \leq -\sum_{k=0}^{t-1} \alpha_k.$$

Since $\sum_{k=0}^{\infty} \alpha_k = \infty$, it follows that $\log P_t \to -\infty$ and hence $P_t \to 0$. Consequently, $P_t S_0^a \to 0$.

Next, $\{M_t\}$ is a martingale with respect to the natural filtration $\mathcal{F}_t = H_t$, since

$$M_{t+1} - M_t = \frac{\alpha_t}{P_{t+1}} \omega_t$$

and

$$\mathbb{E}[M_{t+1} - M_t \mid \mathcal{F}_t] = \frac{\alpha_t}{P_{t+1}} \mathbb{E}[\omega_t \mid \mathcal{F}_t] = 0.$$

Assume also that the conditional second moments of the noise are uniformly bounded: $\mathbb{E}[\omega_t^2 \mid \mathcal{F}_t] \leq C$. Then

$$\mathbb{E}\left[(M_{t+1} - M_t)^2 \mid \mathcal{F}_t\right] = \frac{\alpha_t^2}{P_{t+1}^2} \mathbb{E}[\omega_t^2 \mid \mathcal{F}_t] \leq C \frac{\alpha_t^2}{P_{t+1}^2}.$$

Therefore

$$\mathbb{E}[M_t^2] \leq C \sum_{j=0}^{t-1} \frac{\alpha_j^2}{P_{j+1}^2}.$$

Using $S_t^a = P_t S_0^a + P_t M_t$ and $(a + b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\mathbb{E}[(S_t^a)^2] \leq 2P_t^2 (S_0^a)^2 + 2CP_t^2 \sum_{j=0}^{t-1} \frac{\alpha_j^2}{P_{j+1}^2}.$$

Lemma 6.1. *Let*

$$P_t = \prod_{k=0}^{t-1} (1 - \alpha_k),$$

where

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2 < \infty.$$

Then

$$P_t^2 \sum_{j=0}^{t-1} \frac{\alpha_j^2}{P_{j+1}^2} \longrightarrow 0.$$

Proof. Define $B_t := 1/P_t^2$. Since $P_t \to 0$, we have $B_t \to \infty$. Moreover,

$$B_{t+1} - B_t = \frac{1}{P_{t+1}^2} - \frac{1}{P_t^2} = \frac{2\alpha_t - \alpha_t^2}{P_{t+1}^2}.$$

Thus

$$\frac{\alpha_t^2}{P_{t+1}^2} = \frac{\alpha_t}{2 - \alpha_t}(B_{t+1} - B_t).$$

Let $c_t := \alpha_t/(2 - \alpha_t)$. Since $\sum_t \alpha_t^2 < \infty$, we have $\alpha_t \to 0$, and hence $c_t \to 0$. Therefore

$$P_t^2 \sum_{j=0}^{t-1} \frac{\alpha_j^2}{P_{j+1}^2} = \frac{1}{B_t} \sum_{j=0}^{t-1} c_j (B_{j+1} - B_j).$$

By Kronecker's lemma, since $B_t \uparrow \infty$ and $c_t \to 0$,

$$\frac{1}{B_t} \sum_{j=0}^{t-1} c_j (B_{j+1} - B_j) \longrightarrow 0.$$

This proves the claim. □

The lemma implies that the second term in the bound for $\mathbb{E}[(S_t^a)^2]$ converges to zero, while the first term also converges to zero because $P_t \to 0$. Hence

$$\mathbb{E}[(S_t^a)^2] \to 0,$$

so $S_t^a \to 0$ in L^2.

Theorem 6.3 (Robbins–Siegmund Almost-Supermartingale Theorem)**.** *Let $(X_t, \mathcal{F}_t)$ be a nonnegative adapted stochastic process with $\mathbb{E}[X_t] < \infty$ for all t. Suppose there exist nonnegative deterministic sequences $\{a_t\}$, $\{b_t\}$, and $\{c_t\}$ such that*

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] \leq (1 + a_t)X_t - b_t + c_t$$

almost surely for all t. If

$$\sum_{t=0}^{\infty} a_t < \infty, \qquad \sum_{t=0}^{\infty} c_t < \infty,$$

then X_t converges almost surely to a finite random variable and

$$\sum_{t=0}^{\infty} b_t < \infty$$

almost surely.

Proof. Define

$$p_t := \prod_{k=0}^{t-1}(1 + a_k), \qquad p_0 := 1.$$

Since $\sum_t a_t < \infty$, the sequence p_t converges to a finite positive limit. Dividing the assumed inequality by $p_{t+1} = p_t(1 + a_t)$ gives

$$\mathbb{E}\left[\frac{X_{t+1}}{p_{t+1}} \mid \mathcal{F}_t\right] \le \frac{X_t}{p_t} - \frac{b_t}{p_{t+1}} + \frac{c_t}{p_{t+1}}.$$

Let

$$Z_t := \frac{X_t}{p_t} + \sum_{k=0}^{t-1} \frac{b_k}{p_{k+1}} + \sum_{k=t}^{\infty} \frac{c_k}{p_{k+1}}.$$

The final sum is finite because $p_{k+1} \ge 1$ and $\sum_k c_k < \infty$. Moreover, $Z_t \ge 0$ and

$$\begin{aligned}
\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] &= \mathbb{E}\left[\frac{X_{t+1}}{p_{t+1}} \mid \mathcal{F}_t\right] + \sum_{k=0}^{t} \frac{b_k}{p_{k+1}} + \sum_{k=t+1}^{\infty} \frac{c_k}{p_{k+1}} \\
&\le \frac{X_t}{p_t} - \frac{b_t}{p_{t+1}} + \frac{c_t}{p_{t+1}} + \sum_{k=0}^{t} \frac{b_k}{p_{k+1}} + \sum_{k=t+1}^{\infty} \frac{c_k}{p_{k+1}} \\
&= \frac{X_t}{p_t} + \sum_{k=0}^{t-1} \frac{b_k}{p_{k+1}} + \sum_{k=t}^{\infty} \frac{c_k}{p_{k+1}} = Z_t.
\end{aligned}$$

Thus $\{Z_t\}$ is a nonnegative supermartingale, and hence it converges almost surely to a finite random variable. Since

$$\sum_{k=0}^{t-1} \frac{b_k}{p_{k+1}} \le Z_t,$$

the weighted series $\sum_k b_k/p_{k+1}$ is finite almost surely. Because p_{k+1} is bounded above and bounded away from zero, this implies $\sum_k b_k < \infty$ almost surely.

Finally, the tail term $\sum_{k=t}^{\infty} c_k/p_{k+1}$ converges to zero, and the partial sums $\sum_{k=0}^{t-1} b_k/p_{k+1}$ converge to a finite limit. Since Z_t also converges, X_t/p_t converges almost surely. As p_t has a finite positive limit, X_t itself converges almost surely to a finite random variable. □

Applying the Robbins–Siegmund theorem to the squared noise recursion gives that $\{S_t^a\}$ converges almost surely. Since its L^2 limit is 0, the almost sure limit must also be 0. Therefore

$$S_t^a \to 0 \qquad \text{almost surely.}$$

Conclusion:

- $\|S_t^a\|_\infty \to 0$ (via martingale convergence),
- $\|S_t^b\|_\infty \to 0$ (via exponential decay),

Thus:

$$\boxed{\|Q_t - Q^*\|_\infty \to 0.}$$

this indicates that Q-Value converges to some point at some level of recursion.

7 Q-Approximation

In large finite-state problems, a tabular representation of $Q(x, u)$ may be too expensive. Instead, we approximate the action-value function by a parametric class. In the linear case,

$$Q_\theta(x, u) = \phi(x, u)^\top \theta,$$

where $\phi(x, u) \in \mathbb{R}^K$ is a feature vector and $\theta \in \mathbb{R}^K$ is the parameter vector. If there are n state-action pairs, define the feature matrix

$$\Phi = \begin{bmatrix} \phi(x_1, u_1)^\top \\ \vdots \\ \phi(x_n, u_n)^\top \end{bmatrix}, \qquad Q_\theta = \Phi\theta.$$

Thus the approximation class

$$\mathcal{Q} = \{\Phi\theta : \theta \in \mathbb{R}^K\}$$

is a K-dimensional subspace of $\mathbb{R}^n$.

7.1 Why Projection is Needed

In the tabular case, the optimal Q-function satisfies the Bellman fixed-point equation

$$Q^* = TQ^*,$$

where T is the Bellman operator. However, if $Q = \Phi\theta$ belongs to the approximation space, the Bellman update $T(\Phi\theta)$ generally need not belong to $\mathcal{Q}$. Therefore, after applying the Bellman operator, one projects the result back onto the approximation space.

7.2 Projected Bellman Equation

Let $W \succ 0$ be a positive definite weighting matrix and define

$$\|z\|_W^2 = z^\top W z.$$

The W-weighted projection onto the feature space is

$$\Pi_W y = \arg\min_{v \in \mathcal{Q}} \|v - y\|_W^2.$$

The projected Bellman equation is

$$Q = \Pi_W TQ.$$

This replaces the exact Bellman equation $Q = TQ$ when the exact fixed point may not lie in the approximation class.

At iteration t, the projected Bellman update can be written as the weighted least-squares problem

$$\hat{\theta}_{t+1} = \arg\min_{\theta} \frac{1}{2} \left\| \Phi\theta - T(\Phi\hat{\theta}_t) \right\|_W^2 .$$

Differentiating the objective gives the normal equation

$$\Phi^\top W \big(\Phi\theta - T(\Phi\hat{\theta}_t) \big) = 0.$$

Assuming $\Phi^\top W \Phi$ is nonsingular, the explicit solution is

$$\hat{\theta}_{t+1} = (\Phi^\top W \Phi)^{-1} \Phi^\top W T(\Phi\hat{\theta}_t).$$

Equivalently,

$$\theta_{t+1} = AT(\Phi\theta_t), \qquad A = (\Phi^\top W \Phi)^{-1} \Phi^\top W.$$

Geometrically, the Bellman step $Q_t \mapsto TQ_t$ usually leaves the feature subspace, and the projection step maps TQ_t back to $\Pi_W TQ_t$. Thus the approximate iteration is

$$Q_{t+1} = \Pi_W TQ_t.$$

7.3 Why Convergence Depends on the Weighting Matrix

The Bellman operator is a contraction in the sup norm:

$$\|TV - TU\|_\infty \le \alpha \|V - U\|_\infty, \qquad 0 < \alpha < 1.$$

However, the composed operator $\Pi_W T$ need not be a contraction. Hence projected value iteration may diverge for poorly chosen weights. The reason is that projection can amplify errors by more than the Bellman operator contracts them.

A particularly important choice is stationary-distribution weighting. If

$$W = \mathrm{diag}(\pi),$$

where π is the stationary distribution of the Markov chain induced by a fixed policy, then under standard policy-evaluation assumptions one can show

$$\|\Pi_\pi TV - \Pi_\pi TU\|_\pi \le \alpha \|V - U\|_\pi.$$

Therefore $\Pi_\pi T$ is a contraction, and by Banach's fixed-point theorem the projected Bellman equation has a unique solution and the iteration converges.

Remark 7.1. *Stationary-distribution weighting emphasizes states that are visited often under the policy. Errors in rarely visited states matter less, so the approximation is aligned with the long-run dynamics of the Markov chain.*

7.4 ODE Method for Linear Approximation

When the model is not known exactly, the least-squares problem may be replaced by an incremental stochastic approximation recursion

$$\theta_{t+1} = \theta_t + \epsilon_t H(\theta_t, Z_t),$$

where Z_t denotes sampled transition data. Define the mean field

$$f(\theta) = \mathbb{E}[H(\theta, Z)].$$

Ignoring noise, the recursion behaves like

$$\theta_{t+1} \approx \theta_t - \epsilon_t f(\theta_t).$$

As $\epsilon_t \to 0$, this tracks the limiting ODE

$$\dot{\theta} = -f(\theta).$$

This is the same form as an Euler discretization of the differential equation. If $f(\theta^*) = 0$ and

$$(\theta - \theta^*)^\top f(\theta) > 0, \qquad \theta \neq \theta^*,$$

then trajectories move toward θ^*, so the ODE is globally stable. Under standard stochastic approximation assumptions, the stochastic recursion then converges to θ^*.

For policy evaluation with linear features, the temporal-difference error has the form

$$\delta = c(x) + \alpha\phi(x')^\top\theta - \phi(x)^\top\theta.$$

The associated mean field can be written as

$$f(\theta) = -\mathbb{E}[\delta\phi(x)].$$

Thus the equilibrium condition $f(\theta) = 0$ becomes

$$\mathbb{E}[\delta\phi(x)] = 0,$$

which is precisely the projected Bellman equation in moment form.

Remark 7.2. *The ODE method studies the continuous-time path that the noisy discrete-time algorithm tracks. If the ODE has a unique globally stable equilibrium, then the stochastic approximation is pulled toward the same point.*

7.5 Actor-Critic Methods

Actor-critic methods separate learning into two coupled parts.

- **Critic:** Estimates Q^π or V^π, usually through TD learning and possibly function approximation.

- **Actor:** Improves the policy using the critic, for example by taking a greedy step

$$\pi_{k+1} = \text{Greedy}(Q_k),$$

or by a policy-gradient update

$$\theta_{k+1} = \theta_k + \eta_k \nabla J(\theta_k).$$

The critic asks, "How good is the current policy?" while the actor asks, "How can the policy be improved?" In convergence analyses, one often uses time-scale separation:

$$\eta_k \ll \epsilon_k.$$

The critic then learns on the faster time scale, while the actor changes slowly. This allows the critic to nearly equilibrate before the actor makes a significant policy update.

7.6 Q-Learning with Function Approximation

The tabular Q-learning recursion is

$$Q_{t+1}(x_t, u_t) = Q_t(x_t, u_t) + \epsilon_t \left[c_t + \alpha \min_v Q_t(x_{t+1}, v) - Q_t(x_t, u_t) \right].$$

With linear approximation, replace $Q_t(x, u)$ by $\phi(x, u)^\top \theta_t$. The parameter update becomes

$$\theta_{t+1} = \theta_t + \epsilon_t \delta_t \phi(x_t, u_t),$$

where

$$\delta_t = c_t + \alpha \min_v \phi(x_{t+1}, v)^\top \theta_t - \phi(x_t, u_t)^\top \theta_t.$$

This update is not guaranteed to converge. The Bellman optimality operator is nonlinear because of the minimization over actions, and projection and minimization do not generally commute. Consequently, ΠT may fail to be a contraction, and the update can move away from the desired solution.

Remark 7.3. ***Deadly triad:*** *Baird's counterexample shows that divergence can occur even with linear features, diminishing step sizes, and off-policy sampling. The combination of function approximation, bootstrapping, and off-policy learning is therefore often called the deadly triad.*

8 Model-Free Policy Evaluation Methods

Now suppose a policy μ is fixed. The goal is no longer policy optimization, but evaluation of the value function J^μ.

8.1 Monte Carlo Policy Evaluation

For one trajectory $x_0, x_1, \ldots$ generated under policy μ, define the discounted return

$$G = \sum_{k=0}^{\infty} \alpha^k c(x_k, u_k).$$

Then

$$J^\mu(i) = \mathbb{E}[G \mid x_0 = i].$$

Given N independent episodes starting from state i, with observed returns $G_1, \ldots, G_N$, the Monte Carlo estimator is

$$\hat{J}^\mu(i) = \frac{1}{N} \sum_{m=1}^{N} G_m.$$

By the law of large numbers,

$$\hat{J}^\mu(i) \to J^\mu(i)$$

as $N \to \infty$. Monte Carlo estimates are unbiased, but they require complete episodes and may have high variance.

8.2 Temporal-Difference Learning

TD methods use the Bellman equation without waiting for a complete episode. For a fixed policy μ,

$$J^\mu(i) = c(i, \mu(i)) + \alpha \sum_j P_{ij}^{\mu(i)} J^\mu(j).$$

After observing one transition $x_k \to x_{k+1}$, define the TD error

$$\delta_k = c(x_k, \mu(x_k)) + \alpha \hat{J}_k(x_{k+1}) - \hat{J}_k(x_k).$$

The TD(0) update is

$$\hat{J}_{k+1}(x_k) = \hat{J}_k(x_k) + \epsilon_k \delta_k.$$

Monte Carlo uses the full target

$$G_k = c_k + \alpha c_{k+1} + \alpha^2 c_{k+2} + \cdots,$$

whereas TD(0) uses the bootstrapped target

$$c_k + \alpha \hat{J}_k(x_{k+1}).$$

Thus Monte Carlo is unbiased but high variance, while TD is biased but often lower variance.

8.3 TD(λ)

TD(0) uses one-step information. TD(λ) combines n-step returns with geometric weights. If $G^{(n)}$ denotes the n-step return, define

$$G^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G^{(n)}.$$

The special cases are

$$\lambda = 0 \Rightarrow \text{TD}(0), \qquad \lambda = 1 \Rightarrow \text{Monte Carlo}.$$

Therefore TD(λ) interpolates between one-step bootstrapping and full-return Monte Carlo estimation.

Review 8.1. *TD(0) updates the value estimate using only the next step, making it fast but sometimes noisy. TD(λ) uses information from multiple future steps, balancing speed and accuracy. Monte Carlo methods use the entire trajectory, which can be accurate but slow to update.*

9 Synchronous Q-learning Convergence via ODE

The synchronous Q-learning recursion can be written as

$$\hat{Q}_{k+1} = \hat{Q}_k + \epsilon_k\big(T(\hat{Q}_k) - \hat{Q}_k\big) + M_k.$$

Here T is the Bellman operator. The term M_k is the sampling noise:

$$M_k = \text{sample target} - \mathbb{E}[\text{sample target} \mid \mathcal{F}_k].$$

Hence

$$\mathbb{E}[M_k \mid \mathcal{F}_k] = 0,$$

so M_k is a martingale-difference term.

Ignoring the noise gives the limiting ODE

$$\dot{Q} = T(Q) - Q.$$

Let Q^* be the Bellman fixed point, so $Q^* = TQ^*$. Define $e(t) = Q(t) - Q^*$. Then

$$\dot{e} = T(Q) - T(Q^*) - e.$$

Using the Bellman contraction property,

$$\|T(Q) - T(Q^*)\| \le \alpha\|e\|, \qquad 0 < \alpha < 1.$$

Therefore the error decays at rate at least $1-\alpha$:

$$\frac{d}{dt}\|e(t)\| \le -(1-\alpha)\|e(t)\|.$$

Consequently,

$$\|e(t)\| \le e^{-(1-\alpha)t}\|e(0)\|,$$

and hence $Q(t) \to Q^*$. Thus the equilibrium Q^* is globally asymptotically stable, and under standard stochastic approximation assumptions the noisy synchronous Q-learning recursion converges to Q^*.