Entropy-Regularized Stochastic Control and Mean Field Games

Güneş Akbaş

Department of Mathematics
Bilkent University

May 5, 2025

Overview

1. Quick Review for Markov Decision Process

2. Entropy Regularized Average Cost MDP

3. Entropy Regularized Mean-Field Games

What is a Markov Process?

Definition

A discrete–time *Markov process* is a sequence of random variables $(X_n)_{n\geq 0}$ on a state space S such that

$$\Pr\big(X_{n+1} \in A \,\big|\, X_0, \ldots, X_n\big) = \Pr\big(X_{n+1} \in A \,\big|\, X_n\big) \quad \forall\, A \subseteq S,\ n \geq 0.$$

The right-hand side is governed by a *transition kernel* $P(x, A) = \Pr(X_{n+1} \in A \mid X_n = x)$ that is independent of n (time-homogeneous case).

Key Ingredients

- **State space** S (finite, countable, or general Borel).
- **Initial law** $\mu_0(A) = \Pr(X_0 \in A)$.
- **Kernel** $P : S \times \mathcal{B}(S) \to [0, 1]$.
- k**-step kernel** $P^k(x, A)$ obtained by composition: $P^k = P^{k-1} * P$.

What is a Controlled Markov Process?

Setup (**Markov Decision Process**)

At each step n the controller observes the state X_n and chooses an *action* $U_n \in \mathcal{A}(X_n)$. The next state obeys

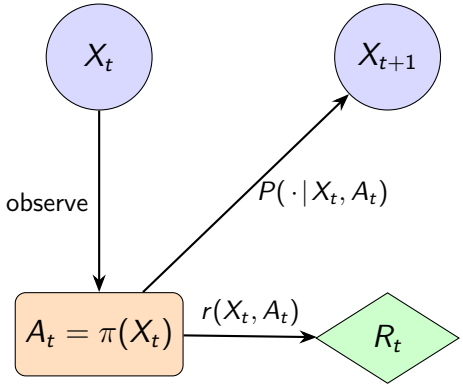
$$\Pr\big(X_{n+1} \in A \mid X_n = x, A_n = a\big) = P(x, a, A).$$

A *policy* $\pi = \{\pi_n\}_{n \geq 0}$ maps histories or states to actions (deterministically or stochastically).

Why "Controlled"?

In a pure Markov process the kernel P is fixed. Here we **control** P *through the action* u, steering the state toward desirable regions while optimising rewards or costs.

Controlled Markov Process — Schematic



State	$X_t \in S$ (finite set)
Action	$A_t \in \mathcal{A}(X_t)$ chosen by policy π
Reward	$R_t = r(X_t, A_t)$
Dynamics	$X_{t+1} \sim P(\cdot \mid X_t, A_t)$

One step: observe state → act → receive reward & transition.

Reward Criteria (Maximisation Versions)

1. Finite-Horizon Total Reward $(t = 0, \ldots, T-1)$

$$J_\pi^{(T)}(x_0) = \mathbb{E}_{x_0}^\pi\Big[\sum_{n=0}^{T-1} r(X_n, A_n) \;+\; r_T(X_T)\Big], \qquad V_T(x) = r_T(x),\; V_t(x) = \max_a\big\{r(x,a) + \mathbb{E}[V_{t+1}]\big\}.$$

2. Discounted Infinite-Horizon Reward

$$J_\pi^\beta(x) = \mathbb{E}_x^\pi\Big[\sum_{n=0}^{\infty} \beta^n\, r(X_n, A_n)\Big], \qquad V(x) = \max_a\big\{r(x,a) + \beta\,\mathbb{E}[V(x)]\big\}, \quad 0 < \beta < 1.$$

3. Long-Run Average (Per-Step) Reward

$$\rho_\pi = \limsup_{N\to\infty} \frac{1}{N}\,\mathbb{E}_x^\pi\Big[\sum_{n=0}^{N-1} r(X_n, A_n)\Big], \qquad \rho + h(x) = \max_a\big\{r(x,a) + \mathbb{E}[h(x)]\big\}.$$

ρ is the optimal steady-state reward, h the bias function

Same dynamics, different objectives—choose the criterion that matches the application's time scale and risk preference.

When Should You Buy? — A Discrete-Time Purchasing Game

The Question

You have N days to purchase a limited-stock gadget. Each morning you see its price, a fresh random number uniformly drawn from $[0, 1]$. You can either BUY immediately at that price or WAIT for tomorrow, but you must buy on the last day if you still haven't.

How do you decide when to buy so that the expected discount ($1-$price) is maximised?

Model Snapshot

- Time steps $t = 0, \ldots, N-1$.
- Action set $\mathcal{A} = \{\text{BUY}, \text{WAIT}\}$.
- Reward if you buy at price x_t: $r_t = 1 - x_t$.
- If waiting until day $N-1$, forced purchase at x_{N-1}.

Dynamic-Programming Solution

Let V_t = expected future reward at day t .

$$V_{N-1} = \frac{1}{2}, \qquad V_t = \mathbb{E}\big[\max\{1-P,\ V_{t+1}\}\big], \quad t = 2, \ldots, N.$$

Because $P \sim U[0,1]$ this yields a **buy threshold**

$$c_t = V_{t-1}, \qquad \pi_t(P) = \begin{cases} \text{BUY}, & P \le c_t, \\ \text{WAIT}, & P > c_t. \end{cases}$$

Take-away: the optimal rule is a **declining price threshold**—becoming less picky as time runs out.

Average–Cost Criterion: Key Facts & Solution Approach

Why is it Subtle?

Unlike the discounted case, the value function may diverge; we seek **steady-state performance** ρ and a *bias* (relative value) function $h(\cdot)$.

Average-Cost Optimality Equation (ACOE)

$$\rho + h(x) = \min_{u \in \mathcal{A}(x)} \Big\{ r(x,u) + \sum_{y \in S} P(x,u,y)\, h(y) \Big\}, \qquad x \in S.$$

- $\rho =$ optimal long-run cost (scalar, independent of x).

Average–Cost Operator is a Contraction in the Span Seminorm

Span Seminorm on $\mathbb{R}^{|S|}$

$$\|h\|_{\mathrm{sp}} := \max_{x \in S} h(x) - \min_{x \in S} h(x) \quad \text{(insensitive to constant shifts)}.$$

Relative–Value (Bias) Operator

For an MDP with running reward r and kernel P, fix a reference state x_0 and define, for any $h \in \mathbb{R}^{|S|}$,

$$(Th)(x) = \min_{u \in \mathcal{A}(x)} \Big\{ r(x,u) + \sum_{y \in S} P(x,u,y)\, h(y) \Big\},$$

Key Facts About The Contraction

Assumption (Doeblin / Minorisation)

$\exists$ sub-probability measure ν on $\mathcal{S}$ such that $P(y \mid x, u) \geq \nu(y), \quad \forall\, y \in \mathcal{S},$

$$\sum_{y \in \mathcal{S}} \nu(y) = \eta.$$

Contraction Property

If the MDP is communicating, then $\eta > 0$ and

$$\|Th - Tv\|_{\mathrm{sp}} \leq (1 - \eta)\, \|h - v\|_{\mathrm{sp}}, \qquad \forall\, h, v \in \mathbb{R}^{|S|}.$$

Stochastic Policies in Finite-State MDPs

Definition

$$\pi(a \mid x) = \Pr\big(A_t = a \mid X_t = x\big), \qquad \sum_{a \in \mathcal{A}(x)} \pi(a \mid x) = 1.$$

A *deterministic* policy picks a single action, $A_t = \pi(x)$; a *stochastic* (randomised) policy samples from $\pi(\cdot \mid x)$ every visit to x.

Given π, the controlled chain behaves like an *ordinary* Markov chain with

$$P_\pi(y \mid x) = \sum_{a \in \mathcal{A}(x)} \pi(a \mid x)\, P(y \mid x, a),$$

$$r_\pi(x) = \sum_{a \in \mathcal{A}(x)} \pi(a \mid x)\, r(x, a).$$

- Guarantees existence of optimal policies (finite MDPs).
- Exploration in reinforcement learning.

Entropy–Regularised ACOE (Average Reward)

Bellman Equation with Entropy Penalty

$$\rho + h(x) = \max_{\pi(\cdot|x)} \Bigg[\sum_{a \in \mathcal{A}(x)} \pi(a \mid x) \Big(r(x,a) + \sum_{y} P(y \mid x, a) h(y) \Big) + \delta \, H\big(\pi(\cdot \mid x)\big) \Bigg]$$

- $H(\pi) = -\sum_a \pi(a \mid x) \log \pi(a \mid x)$ (Shannon entropy).
- $\delta > 0$ penalises low entropy $\Rightarrow$ encourages exploration.
- $\delta \downarrow 0$ recovers the classical average-reward Bellman equation.

Soft-Max Map $\sigma : \mathbb{R}^d \to \Delta^{d-1}$

Definition

For $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$

$$\boxed{\sigma_i^{\delta}(z) = \frac{\exp(\delta z_i)}{\sum_{j=1}^{d} \exp(\delta z_j)}, \quad i = 1, \ldots, d} \quad \Longrightarrow \quad \sigma(z) \in \Delta^{d-1} \text{ (probability simplex).}$$

Classical arg max	**Softmax**
Selects the single largest entry (ties broken arbitrarily); all others receive zero weight.	Assigns every entry a positive share; larger inputs get more weight but smaller ones still contribute.
Non-differentiable, so gradients are undefined at ties and awkward for gradient-based optimisation.	Smooth and differentiable everywhere, ideal for back-propagation and policy-gradient methods.
Outputs a single value (or index).	Outputs a full probability distribution whose components sum to 1.

Soft-Max Solution via Variational Formula

Let $Q(x,a) = r(x,a) + \sum_y P(y \mid x,a) h(y)$. For each state x solve

$$\max_{\pi(\cdot|x)} \sum_a \pi(a \mid x) Q(x,a) + \delta\, H\big(\pi(\cdot \mid x)\big).$$

Optimal Policy (soft–max)

$$\pi^*(a \mid x) = \frac{\exp\big(\frac{1}{\delta}\, Q(x,a)\big)}{\sum_{a'} \exp\big(\frac{1}{\delta}\, Q(x,a')\big)}.$$

Soft-Maximum Bias Update

$$\rho + h(x) = \delta\, \log \sum_a \exp\Big(\tfrac{1}{\delta}\, Q(x,a)\Big).$$

- As $\delta \downarrow 0$ the soft-max $\to$ hard arg max.
- Relative-value and policy-iteration algorithms simply replace the hard max with this closed-form soft-max yet remain span-norm contractions under standard minorization conditions.

Soft-Max Map $\sigma : \mathbb{R}^d \to \Delta^{d-1}$

Key Properties (temperature $= \delta$)

- **Lipschitz (non-expansive) in ℓ_2 norm:**

$$\|\sigma^\delta(z) - \sigma^\delta(z')\|_2 \;\leq\; \delta\, \|z - z'\|_2,$$

Soft Bellman Operator as Log-Sum-Exp

Entropy-Regularized Bellman Softmax

$$\max_{\pi(\cdot|x)} \Big\{ \sum_a \pi(a \mid x)\, Q(x,a) + \delta\, H\big(\pi(\cdot \mid x)\big) \Big\} = \delta\, \log \sum_a \exp\!\big(\tfrac{Q(x,a)}{\delta}\big).$$

Softmax Value Definition

$$\pi^\star(\cdot \mid x) \;=\; \arg\max_{\pi(\cdot|x)} \Big[\sum_{a \in \mathcal{A}} \pi(a \mid x)\, Q(x,a) \;+\; \delta\, H\big(\pi(\cdot \mid x)\big) \Big] \;=\; \mathrm{softmax}_{\frac{1}{\delta}}\big(Q(x,\cdot)\big)$$

Notice that inverse temperature constant is $\frac{1}{\delta}$

Mean-Field Limit $N \to \infty$ and MFG Equilibrium

Exchangeability

For the joint state vector $(x_{1,t}, \ldots, x_{N,t})$ that is exchangeable,

$$m_t^N = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_{j,t}} \quad \Longrightarrow \quad m_t \qquad \text{(weakly a.s.) as } N \to \infty,$$

and, conditional on m_t, any single agent's state–action law becomes independent of the others .

Representative-Agent Problem

Fix a flow of measures $\{m_t\}_{t=0}^{T}$. The representative agent solves

$$\min_{\pi} \ \mathbb{E}\Big[\sum_{t=0}^{T-1} r(x_t, a_t, m_t) + r_T(x_T, m_T)\Big], \qquad x_{t+1} \sim P\big(\cdot \mid x_t, a_t, m_t\big).$$

When Should You Buy? — A Mean-Field Purchasing Game

The Question

A continuum of buyers has N days to purchase a limited-stock gadget. Each morning t, each buyer sees a price

$$P_t \sim \mathrm{Uniform}\big[\alpha\, m_t,\ 1\big],$$

where m_t is the fraction still waiting and $\alpha \in [0, 1]$ controls demand sensitivity. You can BUY at P_t (then exit) or WAIT — but if $t = N$ you *must* buy.

How should each buyer's policy $\pi_t(P_t, m_t)$ be chosen to maximize the expected discount $1 - P$?

Model Snapshot

- $t = 1, \ldots, N$, state = "still waiting."
- Mean-field mass m_t: fraction of buyers active at t.
- Price law: $f(p \mid m) = \frac{1}{1-\alpha m}\, \mathbf{1}_{[\alpha m, 1]}(p)$.
- Action set: $\mathcal{A} = \{\mathrm{BUY}, \mathrm{WAIT}\}$.
- Reward: if buy at p, $r = 1 - p$; else 0. Forced buy at $t = N$.

Mean-Field Games: Transition & Reward

Assumption:Time homogenous Mean Field Parameter

$$m_t = \bar{m}, \qquad \forall\, t \geq 0,$$

Mean-Field (Nash) Equilibrium

A pair $\left(\pi^*, \{m^*\}\right)$ is an MFG equilibrium iff

1. **Optimality:** π^* minimises the cost when the population distribution is m^*.
2. **Consistency (fixed point):**

$$m^* = \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} P\big(\cdot \mid x, a, m^*\big)\, \pi^*(a \mid x)\, m^*(x), \qquad t = 0, \ldots, T-1.$$

Mean-Field Average-Cost Soft Bellman Equation

Mean-Field Soft Average-Cost Bellman

For a fixed population law m and parameter $\delta > 0$:

$$\rho(m) + h(x; m)$$

$$= \max_{\pi(\cdot|x)} \Big\{ \sum_a \pi(a \mid x) \Big[r(x, a; m) + \sum_y P(y \mid x, a; m)\, h(y; m) \Big] + \delta\, H\big(\pi(\cdot \mid x)\big) \Big\}, \text{ where}$$

$H(\pi) = -\sum_a \pi(a) \log \pi(a)$.

Softmax Closed Form

$$\rho(m) + h(x; m) = \delta \, \log \sum_a \exp\Big(\tfrac{1}{\delta} \big[r(x, a; m) + \sum_y P(y \mid x, a; m)\, h(y; m) \big] \Big).$$

Optimal Stochastic Policy

$$\pi^*(a \mid x; m) = \frac{\exp\big(\frac{1}{\delta} [r(x, a; m) + \sum_y P(y \mid x, a; m)\, h(y; m)] \big)}{\sum_{a'} \exp\big(\frac{1}{\delta} [r(x, a'; m) + \sum_y P(y \mid x, a'; m)\, h(y; m)] \big)}.$$

Lipschitz Conditions for Stochastic Policies

(TK–Lip) Transition-Kernel Lipschitz

For scalar states x, x', stochastic policies π, π', and measures μ, μ':

$$\left\| P^{\pi}(\cdot \mid x, m) \;-\; P^{\pi'}(\cdot \mid x', m') \right\|_{\ell^1} \;\le\; L_x \mathbf{1}_{\{x=x'\}} \;+\; L_\pi \left\| \pi(\cdot \mid x) - \pi'(\cdot \mid x') \right\|_{\ell^1} \;+\; L_m \left\| m - m' \right\|_{\ell^1}.$$

Here $\|\pi(\cdot \mid x) - \pi'(\cdot \mid x')\|_{\ell^1} = \sum_a \left| \pi(a \mid x) - \pi'(a \mid x') \right|$.

(R–Lip) Reward Lipschitz

For the policy-averaged reward $r^{\pi}(x, m) = \sum_a \pi(a \mid x)\, r(x, a, m)$:

$$\left| r^{\pi}(x, m) \;-\; r^{\pi'}(x', m') \right| \;\le\; K_x \, \mathbf{1}_{\{x=x'\}} \;+\; K_\pi \left\| \pi(\cdot \mid x) - \pi'(\cdot \mid x') \right\|_{\ell^1} \;+\; K_m \left\| m - m' \right\|_{\ell^1}.$$

Soft-Q and Policy Lipschitz Bounds

Lemma (Soft-Q Lipschitz)

Under (TK–Lip) and (R–Lip), the optimal entropy-regularized Q-function satisfies

$$\|Q_m^* - Q_{m'}^*\|_{\ell^1} \;\le\; C_Q \, \|m - m'\|_{\ell^1}.$$

Corollary (Policy Lipschitz)

For the soft-max policy $\pi_m(a \mid x) \propto \exp\big(Q_m^*(x,a)/\tau\big)$, we have

$$\|\pi_m - \pi_{m'}\|_{\ell^1} \;\le\; C_\pi \, \|m - m'\|_{\ell^1}$$

$$Q_m(x,a) \;=\; r(x,a;m) \;+\; \sum_y P(y \mid x,a;m)\, h(y;m)$$

Fixed-Point Iteration for Mean-Field Equilibrium

Population Update Operator

$$\operatorname{softmax}_{\frac{1}{\delta}}\big(Q(x,\cdot)\big) = \arg\max_{\pi(\cdot\mid x)\in\Delta(\mathcal{A})} \Big\{ \sum_{a'\in\mathcal{A}} \pi(a' \mid x)\Big[r(x,a';m) + \sum_{y\in\mathcal{X}} P(y \mid x,a';m)\, h_m(y)\Big] + \delta\, H\big(\pi(\cdot \mid x)\big)\Big\}$$

with $H(\pi) := -\sum_a \pi(a)\log\pi(a)$ and h_m solving the average-cost Bellman equation at the fixed population measure m.

$$m'(y) = \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} P\big(y \mid x,a,m\big)\,\pi_m(a \mid x)\,m(x), \qquad y\in\mathcal{X}.$$

References

- Óscar Hernández-Lerma and Jean B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York, 1996. doi:10.1007/978-1-4612-0729-0. :contentReference[oaicite:0]index=0
- Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi, "Q-Learning in Regularized Mean-field Games," *Dynamic Games and Applications*, vol. 13, no. 1, pp. 89–117, Mar. 2023. doi:10.1007/s13235-022-00450-2. :contentReference[oaicite:1]index=1
- Bolin Gao and Lacra Pavel, "On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning," arXiv preprint arXiv:1704.00805, 2017. :contentReference[oaicite:2]index=2

The End