OVERTAKING EQUILIBRIA FOR ZERO–SUM MARKOV GAMES

Onésimo Hernández–Lerma
Mathematics Department
CINVESTAV–IPN
México City
<ohernand@math.cinvestav.mx>

Abstract. *Overtaking optimality* (also known as *catching–up optimality*) is a concept that can be traced back to a paper by Frank P. Ramsey (1928) in the context of economic growth. At present, however, we use a weaker form introduced independently by Atsumi (1965) and von Weizäcker (1965). The apparently different concept of long–run expected *average payoff* (a.k.a. *ergodic payoff*) was introduced by Richard Bellman (1957). In this talk we make a description of how these concepts are related to other optimality criteria, such as *bias optimality* and *canonical strategies*. In fact, we show that

$$\Pi_{00} \subset \Pi_{bias} \subset \Pi_{ca} \subset \Pi_{A0}$$

We do this for a class of (discrete–or continuous–time) Markov games,

- **Part 1:** Control problems
- **Part 2:** Markov games

PART 1. OPTIMAL CONTROL PROBLEMS

An **optimal control problem** has three main components:

1. A "controllable" dynamical system. Examples:

• discrete time:

$$x_{t+1} = F(x_t, a_t, \xi_t) \; \forall \; t = 0, 1, \ldots, \tau \leq \infty$$

• continuous time: diffusion processes, say,

$$dx_t = F(x_t, a_t)dt + \sigma(x_t, a_t)dW_t \quad \forall \; 0 \leq t \leq \tau \leq \infty;$$

continuous–time controlled Markov chains; ...

2. A family Π of admissible control policies (or strategies) $\pi = \{\pi_t\}$.

3. A performance index (or objective function) $V : \Pi \times \mathrm{X} \to \mathbb{R}$,

$$(\pi, x) \mapsto V(\pi, x).$$

The **optimal control problem** is then, for every initial state $x_0 = x$,

$$\text{optimize} \quad \pi \mapsto V(\pi, x) \quad \text{over} \quad \Pi.$$

Notation and terminology: Suppose "optimize" means "maximize". Let

$$V^*(x) := \sup_{\pi \in \Pi} V(\pi, x) \quad \forall \; x_0 = x,$$

be the control problem's **value function**. If there exists $\pi^* \in \Pi$ such that

$$V(\pi^*, x) = V^*(x) \quad \forall\, x \in \mathrm{X},$$

then π^* is said to be an **optimal** control policy (or strategy).

EXAMPLES OF OBJECTIVE FUNCTIONS

• **Finite–horizon** $T > 0$:

$$J_T(\pi, x) := E_x^\pi \left[\sum_{t=0}^{T-1} r(x_t, a_t) \right].$$

• **Discounted reward:** given $\alpha > 0$,

$$V_\alpha(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t r(x_t, a_t) \right].$$

This is, in fact, a medium–term reward criterion because, if $V_\alpha(\pi, x)$ is finite, then

$$E_x^\pi \left[\sum_{t=T}^{\infty} \alpha^t r(x_t, a_t) \right] \to 0 \quad \text{as} \quad T \to \infty.$$

• **Long–run expected average** (or **ergodic) reward**:

$$\begin{aligned} J(\pi, x) &:= \liminf_{T\to\infty} \frac{1}{T}\, J_T(\pi, x) \\ &= \liminf_{T\to\infty} \frac{1}{T}\, E_x^\pi \left[\sum_{t=0}^{T-1} r(x_t, a_t) \right]. \end{aligned}$$

This criterion was introduced by Richard Bellman (1957), motivated by the control of a manufacturing process. The terminology in Bellman's work originated the term **Markov decision problem**.

Bellman, R. (1957). A Markovian decision problem. *J. Math. Mech.* **6**, pp. 679–684.

Typical applications of the average reward criterion

- Queueing systems
- Telecommunication networks (e.g., computer networks, satellite networks, telephone networks, ...)
- Manufacturing processes
- Control of a satellite's attitude

Remark 1. The average reward criterion, why is it called an **ergodic criterion**? In general, an "ergodic" result refers to convergence of averages, either **pathwise averages** (as in the *Law of Large Numbers* or in *Boltzmann's ergodic hypothesis*)

$$\frac{1}{T}\sum_{t=0}^{T-1} r_t \to \int_\Omega R(\omega)\mathrm{P}(d\omega) \equiv E(R) \ \ \text{w.p.}1, \tag{1}$$

or **expected averages**

$$\frac{1}{T} E\left[\sum_{t=0}^{T-1} r_t\right] \to E(R) \tag{2}$$

Sometimes, "ergodic" means something stronger that (1) or (2), for instance, as $t \to \infty$:

$$r_t \to E(R) \;\text{ w.p.}1 \quad \text{or} \quad E(r_t) \to E(R).$$

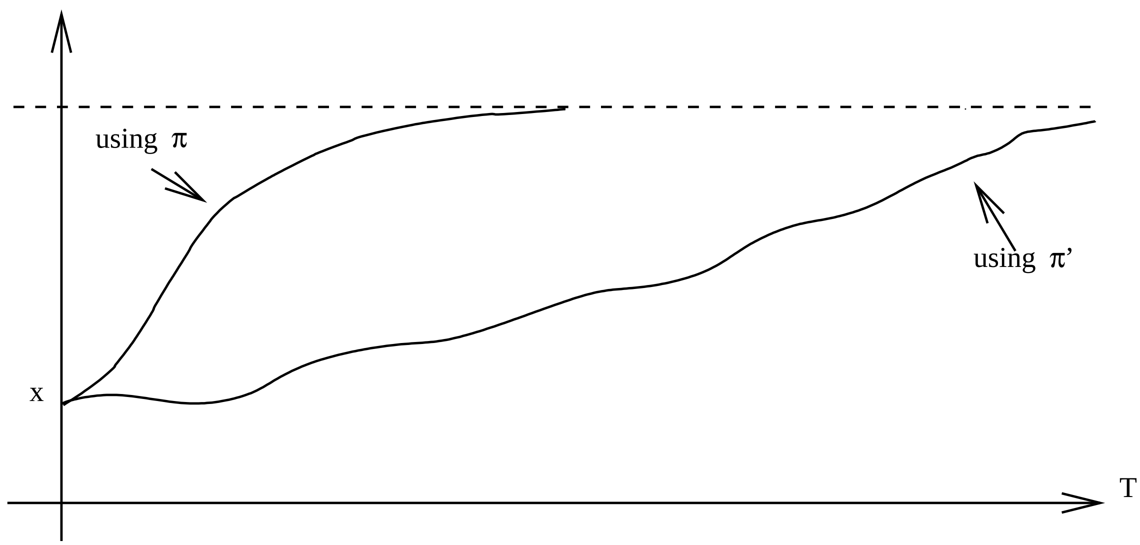

Figure 1

Remark 2. The average criterion is extremely **underselective**, in the sense that it ignores what happens in a finite horizon T, **for every** $T > 0$. For instance, one can have policies π and π', and $\gamma \in (0,1)$, such that

$$J_T(\pi, x) = J_T(\pi', x) + T^\gamma \quad \forall\, T > 0.$$

Therefore,

- $J_T(\pi, x) - J_T(\pi', x) \to \infty$ as $T \to \infty$; however,
- π and π' *have the same long–run average reward*: $J(\pi, x) = J(\pi', x)$.

Problem in financial engineering. For some class Π of portfolios (or investment strategies) determine the "benchmark"

$$\rho^* := \sup_{\pi \in \Pi} J(\pi, x) \quad \forall\, x \in \mathrm{X}.$$

Let Π_{A0} be the family of average optimal portfolios, and suppose Π_{A0} is nonempty.

Problem: Find $\pi^* \in \Pi_{A0}$ with the **fastest growth rate**.

OVERTAKING OPTIMALITY

Ramsey, F.P. (1928). A mathematical theory of saving. The Economic Journal **38**, pp. 543–559.

A policy π^* **overtakes** (or **catches–up**) π if, for every $x \in \mathrm{X}$, there exists $\tau(x, \pi^*, \pi)$ such that

$$J_T(\pi^*, x) \geq J_T(\pi, x) \quad \forall\, T \geq \tau(x, \pi^*, \pi).$$

Here we will use a weaker notion introduced independently by several authors in the 1960s.

We will restrict ourselves to **stationary strategies** $\pi \in \Pi_s$, that is, functions $\pi : \mathrm{X} \to A,\ x_t \to \pi(x_t) \in A$. (Sometimes we will consider **Markov strategies** $(t, x_t) \mapsto \pi(t, x_t) \in A$.)

Definition [Atsumi 1965, von Weiszäcker 1965, ...] A stationary strategy $\pi^* \in \Pi_s$ is **overtaking optimal** (in Π_s) if, for every $\pi \in \Pi_s$ and $x \in \mathrm{X}$,

$$\liminf_{T\to\infty}[J_T(\pi^*, x) - J_T(\pi, x)] \geq 0;$$

equivalently, for every $\pi \in \Pi_s,\ x \in \mathrm{X}$, and $\varepsilon > 0$ there exists

$$T_\varepsilon = T_\varepsilon(\pi^*, \pi, x, \varepsilon)$$

such that

$$J_T(\pi^*, x) \geq J_T(\pi, x) - \varepsilon \quad \forall\, T \geq T_\varepsilon. \qquad (*).$$

Remark. (a) Observe that in overtaking optimality there is no "objective function" to be optimized.

(b) If $(*)$ holds, then the average reward $J(\pi^*, x) \geq J(\pi, x)$ for every $\pi \in \Pi_s$ and $x \in \mathrm{X}$. Therefore

$$\boxed{\text{overtaking optimality} \Longrightarrow \text{average optimality,}}$$

i.e.

$$\boxed{\Pi_{00} \subset \Pi_{A0}.}$$

(c) By $(*)$ again, if π^* is overtaking optimal, then it has the **fastest growth rate**.

How do we find $\boldsymbol{\pi}^*$?

BIAS OPTIMALITY

Suppose that, for each $\pi \in \Pi_{A0}$, the **bias function**

$$b(\pi, x) := E_x^\pi \sum_{t=0}^{\infty} [r(x_t, a_t) - \rho^*]$$

is well defined, where $\rho^* := \sup_{\pi \in \Pi_s} J(\pi, x)$ for all $x \in \mathrm{X}$. Then, for every $T > 0$,

$$J_T(\pi, x) = T \cdot \rho^* + b(\pi, x) + e_T(\pi, x)$$

such that $e_T(\pi, x) \to 0$ as $T \to \infty$.

• If π and π^* are in Π_{A0}, then for every $T > 0$

$$J_T(\pi^*, x) - J_T(\pi, x) = b(\pi^*, x) - b(\pi, x) + e_T(\pi^*, x) - e_T(\pi, x).$$

Definition. $\pi^* \in \Pi_s$ is **bias optimal** if

(a) π^* is in Π_{A0}, and

(b) π^* maximizes the bias, i.e.

$$b(\pi^*, x) = \sup_{\pi \in \Pi_{A0}} b(\pi, x) =: \hat{b}(x) \quad \forall\, x \in \mathrm{X}.$$

Observe that bias optimality is a **lexicographical** optimality criterion.

Theorem. Under some assumptions, the following statements are equivalent for $\pi^* \in \Pi_s$:

(a) π^* is overtaking optimal.

(b) π^* is bias optimal.

(c) There is a constant ρ^* and a function h that satisfy, for all $x \in \mathrm{X}$,

$$\rho^* + h(x) = \max_{a \in A(x)} \left[r(x,a) + \int_{\mathrm{X}} h(y) \mathrm{P}(dy|x,a) \right], \qquad (3)$$

and π^* attains the maximum in (3), i.e.

$$\rho^* + h(x) = r(x, \pi^*(x)) + \int_{\mathrm{X}} h(y) \mathrm{P}(dy|x, \pi^*(x)), \qquad (4)$$

and in addition

$$\int_{\mathrm{X}} \hat{b}(x) \mu_{\pi^*}(dx) = 0.$$

A policy π^* that satisfies (3) and (4) is called **canonical**. In brief, we have

$$\Pi_{A0} \supset \Pi_{ca} \supset \Pi_{bias} = \Pi_{00}.$$

For proofs and examples see, for instance: [5,7,11]. (The theorem is **not** true for games [10].)

PART 2. ZERO–SUM MARKOV GAMES

Consider a two–person **Markov game**, for instance:

- discrete–time: $x_{t+1} = F(x_t, a_t, b_t, \xi_t) \ \forall\, t = 0, 1, \ldots, \tau \leq \infty$;
- stochastic differential game:

$$dx_t = F(x_t, a_t, b_t)dt + \sigma(x_t)dW_t \quad \forall\, 0 \leq t \leq \tau \leq \infty;$$

- jump Markov game with a countable state space;...

Let A (resp. B) be the action space of player 1 (resp. player 2). For $i = 1, 2$, we denote by Π_s^i the family of (randomized) stationary strategies π^i for player i.

Let $r : \mathrm{X} \times A \times B \to \mathbb{R}$ be a measurable function (representing the reward function for player 1, and the cost function for player 2), and define

$$J_T(\pi^1, \pi^2, x) := E_x^{\pi^1,\pi^2}\left[\sum_{t=0}^{T-1} r(x_t, a_t, b_t)\right].$$

The **long–run expected average** (or **ergodic**) **payoff** is:

$$J(\pi^1, \pi^2, x) := \liminf_{T\to\infty} \frac{1}{T}\, J_T(\pi^1, \pi^2, x)$$

Assumption: The ergodic game has a **value** $V(\cdot)$ that is, the lower value

$$L(x) := \sup_{\pi^1} \inf_{\pi^2} J(\pi^1, \pi^2, x)$$

and the upper value

$$U(x) := \inf_{\pi^2} \sup_{\pi^1} J(\pi^1, \pi^2, x)$$

coincide: $L(\cdot) = U(\cdot) \equiv V(\cdot)$.

AVERAGE OPTIMALITY

Definition. A pair $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$ is a pair of **average optimal** strategies if

$$\inf_{\pi^2} J(\pi_*^1, \pi^2, x) = V(x) \quad \forall\, x \in \mathrm{X},$$

and

$$\sup_{\pi^1} J(\pi^1, \pi_*^2, x) = V(x) \quad \forall\, x \in \mathrm{X}.$$

Equivalently, (π_*^1, π_*^2) is a **saddle point,** i.e.

$$J(\pi^1, \pi_*^2, x) \leq J(\pi_*^1, \pi_*^2, x) \leq J(\pi_*^1, \pi^2, x)$$

for every $x \in \mathrm{X}$ and every $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$.

OVERTAKING OPTIMALITY

Definition [Rubinstein 1979]. A pair $(\pi_*^1, \pi_*^2) \in \Pi_s^1 \times \Pi_s^2$ is **overtaking optimal** (in $\Pi_s^1 \times \Pi_s^2$) if, for every $x \in \mathrm{X}$ and every pair $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$, we have

$$\liminf_{T \to \infty} [J_T(\pi_*^1, \pi_*^2, x) - J_T(\pi^1, \pi_*^2, x)] \geq 0$$

and

$$\limsup_{T\to\infty}[J_T(\pi^1_*, \pi^2_*, x) - J_T(\pi^1_*, \pi^2, x)] \leq 0.$$

Under some conditions,

$$\boxed{\Pi_{00} \subset \Pi_{A0}.}$$

Question. Can we characterize Π_{00}?

CANONICAL PAIRS

Definition. A pair $(\pi^1_*, \pi^2_*) \in \Pi^1_s \times \Pi^2_s$ is said to be **canonical** if there is a number $\rho^* \in \mathbb{R}$ and a function $h : \mathrm{X} \to \mathbb{R}$ such that

$$\begin{aligned}
\rho^* + h(x) &= r(x, \pi^1_*, \pi^2_*) + \int_{\mathrm{X}} h(y) P(dy|x, \pi^1_*, \pi^2_*) \\
&= \max_{\pi^1} \left[r(x, \pi^1, \pi^2_*) + \int_{\mathrm{X}} h(y) P(dy|x, \pi^1, \pi^2_*) \right] \\
&= \min_{\pi^2} \left[r(x, \pi^1_*, \pi^2) + \int_{\mathrm{X}} h(y) P(dy|x, \pi^1_*, \pi^2) \right]
\end{aligned}$$

Under some conditions,

$$\boxed{\Pi_{00} \subset \Pi_{ca} \subset \Pi_{A0}.}$$

BIAS OPTIMALITY

Under some conditions, for every pair $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$ there exists a probability measure μ^{π^1,π^2} on X such that

$$J(\pi^1, \pi^2, x) = \int_{\mathrm{X}} r(x, \pi^1, \pi^2)\mu^{\pi^1,\pi^2}(dx) =: \rho(\pi^1, \pi^2) \quad \forall\, x \in \mathrm{X}.$$

Moreover, define the **bias** of (π^1, π^2) as

$$b(\pi^1, \pi^2, x) := E_x^{\pi^1,\pi^2} \sum_{t=0}^{\infty} \left[r(x_t, a_t, b_t) - \rho(\pi^1, \pi^2) \right].$$

Definition. A pair $(\pi_*^1, \pi_*^2) \in \Pi_s^1 \times \Pi_s^2$ is said to be **bias optimal** if it is in Π_{A0} and, in addition,

$$b(\pi^1, \pi_*^2, x) \leq b(\pi_*^1, \pi_*^2, x) \leq b(\pi_*^1, \pi^2, x)$$

for every $x \in \mathrm{X}$ and every pair (π^1, π^2) in Π_{A0}.

$$\boxed{\Pi_{00} \subset \Pi_{bias} \subset \Pi_{ca} \subset \Pi_{A0}.}$$

Partial converse: If (π_*^1, π_*^2) is in Π_{bias}, then it is overtaking optimal in Π_{A0}.

References.

1. H. Atsumi (1965). Neoclassical growth and the efficient program of capital accumulation. *Rev. Econ. Stud.* **32**, 127–136.

2. R. Bellman (1957). A Markovian decision problem. *J. Math. Mech.* **6**, 679–684.

3. D. Carlson, A. Haurie (1996). A tumpike theory for infinite horizon open–loop competitive processes. *SIAM J. Control Optim.* **34**, 1405–1419.

4. B.A. Escobedo–Trujillo, O. Hernández–Lerma, J.D. López--Barrientos (2010). Overtaking equilibria for zero–sum stochastic differential games. In preparation.

5. O. Hernández–Lerma, J.B. Lasserre (1999). *Further Topics on Discrete–Time Markov Control Processes*. Springer–Verlag, New York. Chapter.

6. O. Hernández–Lerma, N. Hilgert (2003). "Bias optimality versus strong 0-discount optimality in Markov control processes with unbounded costs". *Acta Appl. Math.* **77**, 215–235.

7. H. Jasso–Fuentes, O. Hernández–Lerma (2008). "Characterizations of overtaking optimality for controlled diffusion processes". *Appl. Math. Optim.* **21**, 349–369.

8. —, — (2009). "Ergodic control, bias and sensitive discount optimality for Markov diffusion processes". *Stoch. Anal. Appl.* **27** (2009), 363–385.

9. A.S. Nowak (2008). Equilibrium in a dynamic game of capital accumulation with the overtaking criterion. *Econom. Lett.* **99**, 233–237.

10. T. Prieto–Rumeau, O. Hernández–Lerma (2005). "Bias and overtaking equilibria for zero–sum continuous–time Markov games". *Math. Meth. Oper. Res.* **61**, 437–454.

11. —, — (2006). "Bias optimality for continuous–time controlled Markov chains". *SIAM J. Control Optim.* **45**, 51–73.

12. —, — (2008). Ergodic control of continuous–time Markov chains with pathwise constraints. *SIAM J. Control Optim.* **47**, 1888–1908.

13. F.P. Ramsey (1928). A mathematical theory of saving. *Econom. J.* **38**, 545–559.

14. A. Rubinstein (1979). Equilibrium in supergames with the overtaking criterion. *J. Econom. Theory* **21**, 1–9.

15. C.C. von Weizsacker (1965). Existence of optimal programs of accumulation for an infinite horizon. *Rev. Econ. Stud.* **32**, 85–104.