Topics in Game Theory: Learning, Experimentation and Information

Juuso Välimäki

Aalto BIZ and Helsinki GSE

March 2022

1 Introduction

In these three lectures, I introduce some main concepts that are useful for all theoretical models of Bayesian learning. The first lecture provides an overview of the area from the perspective of single agent dynamic decision making.

Throughout the course, we emphasize the dual role of short run actions. They affect the distribution of immediate payoffs or rewards, but they also convey information that is useful for the rest of the dynamic decision problem. This dual role gives rise to a trade-off between *exploitation* (maximizing current payoff) and *exploration* (investing in the production of information).

In many competitive settings (say investing in the financial market), all individual traders are small enough so that they do not have an impact on the information content of the equity prices, and as a result, selecting an optimal portfolio is an exploitation exercise in optimizing the returns.

A (monopolist) venture capitalist considering investment in a risky start-up understands that if no funds are forwarded, the start-up cannot

continue and hence gets no information about the market viability of the company. By investing in the start-up for a period, the venture capitalist learns about the quality of the start-up and may be able to use this information in future decisions.

In the first lecture, we set up a specific model of dynamic optimization where the only connection between decisions across periods is information. The reason for this choice is purely analytical, in order to learn about the effects of information, it is best to abstract from other dynamic connections (capital accumulation, savings, habit formation etc.). Even though some general results (existence and uniqueness of optimal policies) can be obtained for the general model, useful characterizations (how to find the solution, comparative statics etc.) are not possible.

To make progress on the problem, we introduce the model of *multi-armed bandits*. In a nutshell, the objective is to maximize cumulative expected discounted rewards by choosing one alternative at a time from a set of statistically independent options (the arms). This class of problems is sufficiently general to cover interesting economic interactions, yet specific enough to allow useful characterizations.

The second lecture provides a solution (Gittins Index Theorem) to the multi-armed bandit problem. Alternative approaches to the Theorem are discussed and also some alternative formulations of the bandit problem are discussed briefly.

The third lecture contains some first examples of bandits (one-armed in some cases) in economic models.

Not surprisingly, time constraints set a bound on both the breadth and the depth of coverage.

2 Bayesian Learning

2.1 Model

Consider the following simple yet quite general setting where an economic agent learns about underlying uncertainty in her economic environment:

- Time is discrete, $t = 0, 1, ...$
- In each period t, an action a_t is taken in a finite set $A = \{a^1, ..., a^K\}$.
- A random variable X_t is observed in each period. To avoid complicated and unhelpful discussions around measurability, we take the realizations of this random variable to be in a finite set $X = \{x^1, ..., x^M\}$.
- The decision maker receives a reward $r(a_t, x_t)$ in each period t.
- The distribution of X_t depends on the action a_t and on a parameter θ controlling the uncertainty in the model. For simplicity, assume that θ is in a finite set $\Theta = \{\theta^1, ..., \theta^L\}$. Let $p(x \mid a, \theta)$ be the conditional probability mass function on X given (a, θ).
- The parameter is initially unknown and the decision maker has a prior probability $\mu_0(\theta)$ on Θ.
- The objective of the decision maker is to maximize the expected discounted sum of rewards, where the discount factor $\delta < 1$:

$$\max_{(a_t)_{t=0}^{\infty}} \mathbb{E} \sum_{t=0}^{\infty} \delta^t r(a_t, x_t).$$

2.2 Analysis of the Model

At t, the information that the decision maker has is given by $(a_0, x_0, ..., a_{t-1}, x_{t-1})$. Given that the decision problem from t onward does not depend on the

past values (a_s, x_s) for $s < t$, the only payoff relevant state variable for the dynamic problem is the posterior μ_t on Θ computed by Bayes' Rule:

$$\mu_{t+1}(\theta \mid a_t, x_t) = \frac{\mu_t(\theta)p(x_t \mid a_t, \theta)}{\sum_{\theta\in\Theta} \mu_t(\theta)p(x_t \mid a_t, \theta)}.$$

We can write $\mu_{t+1} \sim B(\mu_t; a_t)$ for the stochastic process for the posterior obtained from Bayes' rule (random since the realized posterior depends also on x_t).

With this, we may write the objective function explicitly as

$$\sum_{\Theta}\sum_{X} \delta^t r(a_t, x_t)p(x_t \mid a_t, \theta)\mu_t(\theta).$$

By defining

$$u(a_t, \mu_t) := \sum_{\Theta}\sum_{X} r(a_t, x_t)p(x_t \mid a_t, \theta)\mu_t(\theta),$$

we may write the problem as

$$\max_{(a_t)_{t=0}^{\infty}} \mathbb{E}\sum_{t=0}^{\infty} \delta^t u(a_t, \mu_t) \tag{1}$$

$$\text{subject to:} \quad \mu_{t+1} \sim B(\mu_t, a_t). \tag{2}$$

It should be noted that $u(a_t, \mu_t)$ is linear in μ_t.

Proposition 2.1. The sequence problem 1 has a solution a_t^*.

Proof. Since A is finite, the choice set A^∞ is compact in product topology. The payoff is continuous in product topology so Weierstrass' Theorem guarantees the existence of an optimal policy. For a proof based on the equivalent dynamic programming problem, see Blackwell (1965), Theorem 7(b). □

Since the A, X, Θ are all assumed finite, the the sequence problem above may be solved by dynamic programming. Let $V(\mu)$ be the value function

to the sequence problem starting at prior μ_t. Bellman equation for the problem is then:

$$V(\mu) = \max_{a_t} \mathbb{E}r(a_t, \mu_t) + \delta V(\mu_{t+1}), \tag{3}$$

$$\text{subject to:} \quad \mu_{t+1} \sim B(\mu_t, a_t). \tag{4}$$

Standard results on dynamic programming guarantee that the set of maximizers $A(\mu)$ is an upper-hemicontinuous correspondence and therefore a measurable selection $a(\mu)$ exists (Stokey, Lucas and Prescott (1989), Theorem 7.6). This means that we can write the process of posterior beliefs μ_t as a Markov process on probability distributions on Θ. The most important property of this process is that μ_t is a martingale with respect to the information contained in $(A_0, X_0, ..., A_{t-1}, X_{t-1})$ (formally the σ -algebra $\{\mathcal{F}_t\}$ generated by $(A_0, X_0, ..., A_{t-1}, X_{t-1})$.

Definition 2.1. A sequence of random variables $\{Y_t\}$ on a probability space $(\mu, \Omega, \mathcal{F})$ is called a *martingale* with respect to information contained in the increasing sequence of σ-algebras $\{\mathcal{F}_t\}$ if (for almost all w.r.t. $\mu\,\omega$),

$$E[Y_{t+1} \mid \mathcal{F}_t] = Y_t.$$

Proposition 2.2. $\{\mu_t(B)\}$ is a martingale for all $B \subset \Theta$ with respect to the σ-algebra generated by the observables $(a_0, x_0, ..., a_{t-1}, x_{t-1})$.

Proof. Using Bayes' rule from above, we have for all $a(\mu_t), x$:

$$E[\mu_{t+1}(B) \mid \mu_t] = \sum_{x \in X} \frac{\sum_{\theta \in B} \mu_t(\theta) p(x \mid a(\mu_t), \theta)}{\sum_{\theta \in \Theta} \mu_t(\theta) p(x_t \mid a(\mu_t), \theta)} \Pr\{X_t = x \mid a(\mu_t)\}$$

$$= \sum_{x \in X} \sum_{\theta \in B} \mu_t(\theta) p(x \mid a(\mu_t), \theta) = \sum_{\theta \in B} \mu_t(\theta) \sum_{x \in X} p(x \mid a(\mu_t), \theta) = \mu_t(B).$$

□

We turn next to the question of convergence of these posterior probabilities μ_t. In this quest, we use one of the most famous theorems in the theory of stochastic processes.

Theorem 2.1 (Martingale Convergence Theorem, Doob)**.** Let $\{Y_t\}$ be a martingale with respect to $\{\mathcal{F}_t\}$ which satisfies

$$\sup_t \mathbb{E} \mid Y_t \mid < \infty$$

Then the limit $Y_\infty := \lim_t Y_t$ exists and is finite, almost surely.

Remark. It may be useful to have some examples of martingales in mind.

1. Random walk. Let $X_t = 1$ w.p. $\frac{1}{2}$ and $X_t = -1$ w.p. $\frac{1}{2}$. Put $Y_t = \sum_{s=1}^{t} X_s$. Then Y_t is a martingale with respect to the information generated by Y_t and $\mathbb{E}[Y_{t+1} \mid Y_t] = Y_t$.

2. Random product. Put $X_0 = 1$ and let $X_t = 2$ w.p. $\frac{1}{2}$ and $X_t = 0$ w.p. $\frac{1}{2}$. Set $Y_t = \prod_{s=1}^{t} X_s$. Then Y_t is a martingale with respect to the information generated by Y_t and $\mathbb{E}[Y_{t+1} \mid Y_t] = Y_t$.

3. Polya's Urn. An urn contains one black and one white ball at $t = 0$. At each $t > 1$ one ball is drawn at random and returned to the urn together with another ball of the same color. Let N_t denote the number of white balls at t. Put $Y_t = \frac{N_t}{t}$. Then $\mathbb{E}[Y_{t+1} \mid Y_t] = Y_t$.

4. For each of the these cases, check if the conditions of the martingale convergence theorem are satisfied. If yes, what is the limiting random variable Y_∞? (For Polya's Urn, this is not an easy problem).

5. Let $\Theta = \{0, 1\}$ and let $p_0 = \frac{1}{2}$ be the prior probability that $\theta = 1$. At each $t > 0$ a realization X from the conditional distribution $F(x \mid \theta)$ on $[0, 1]$ is observed. Formulate Bayes' rule for the case where $F(\cdot)$ has densities
$$f(x \mid 0) = \frac{2 + \gamma}{2} - \gamma x,$$

$$f(x \mid 1) = \frac{2-\gamma}{2} + \gamma x.$$

. Denote the posterior on $\theta = 1$ by p_1 and compute the unconditional distribution of p_1 as a function of $\gamma \in [0, 2]$.

6. Consider an independent sample $X_1, ..., X_t$ from $F(\cdot \mid \theta)$, where $\theta \in \{0, 1\}$. Let p_t denote the posterior of $\theta = 1$. Using Bayes' rule, show that the likelihood ratio $\frac{1-p_t}{p_t}$ is a martingale if $\theta = 1$.

7. Let X_t be a sequence of conditionally i.i.d. draws from normal distribution with mean θ and variance σ_x. Let θ be drawn from a normal prior distribution with mean 0 and variance σ_θ. How would you compute the posterior distribution of θ after observing n draws of the X_t? (Hint: is it normal?) Can you show that the mean of the posterior is a martingale with respect to the information contained in the X_t?

An excellent source for more material on martingales is: David Williams, Probability with martingales, Cambridge Mathematical Textbooks, Cambridge University Press, 1991.

Martingale Convergence Theorem implies that along the optimally chosen sequence of actions, there exists a random variable μ_∞ such that with probability 1,

$$\mu_t(B) \to \mu_\infty(B) \text{ for all } B \subset \Theta.$$

This means that with probability 1, $\mu_t(\theta)$ converges to a constant for all $\theta \in \Theta$. When is this possible? We say that there is complete learning if μ_t converges to a point mass on the correct θ. There are two long-run possibilities for the process of beliefs. i) Either the parameter is learned or ii) no new information is generated along the optimal sequence. The latter case implies that θ cannot be identified from the distribution of X_t at the optimal action $a(\mu_\infty)$.

Remark. Let me say a few words on how the analysis above generalizes. The mathematical analysis of the dynamic optimization problem becomes hard if one assumes that X is uncountable, e.g. $X \subset \mathbb{R}$ and/or the parameter set Θ is uncountable. Defining conditional probability measures is somewhat more tricky in this case. The second difficulty concerns the continuity of the Bayes operator $B(\mu_t; a_t)$. Stokey, Lucas and Prescott contains a discussion on the continuity requirements required for a rigorous dynamic programming approach. Aghion et al. (1991) cited in the readings is quite careful about such matters.

These are technical concerns and as long as you are willing to impose continuity assumptions on the model, they can be handled and they will not cause insurmountable problems. For example, various types of normally distributed X_t with an unknown mean θ sampled from a normal prior distribution can be accommodated.

Martingale convergence theorem is a very important result for economic theory models. It applies in a much more general setting and it leads to some interesting connections to diverse other areas of probability theory. Williams' book is a bit demanding, but very entertaining (for a math book).

2.3 Results

- Asymptotically, we have that $a(\mu_\infty) \in \arg\max_{a \in A} u(a, \mu_\infty$. In words, since the beliefs have converged there are no more dynamic considerations in the model and only exploitation motive remains. See Aghion, Bolton, Harris, and Jullien (1991) and Easley and Kiefer (1988) for details and additional material.

- A simple corollary of this is that if complete learning takes place, then asymptotically a full information optimal action is chosen.

- As $\delta \to 1$, full information optimal payoffs can be approximated arbitrarily closely.

- When $A \subset \mathbb{R}$ and the reward function is a deterministic (but unknown) analytic function of (a, θ), then asymptotically full learning is achieved. Is this a reasonable economic setting?

- If the reward is deterministic, differentiable and quasi-concave, then the full learning results. What is the economic significance of this and how does this relate to the previous observation.

- The value function $V(\mu)$ is convex. To see this, consider a binary experiment resulting in posterior μ' with probability λ and in μ'' with probability $(1 - \lambda)$. By the martingale property of posteriors, we have $\mu = \lambda\mu' + (1 - \lambda)\mu''$. Since the policy $\mathbf{a}(\mu)$ (sequence of actions contingent on the realizations of a_t, x_t)) is also feasible starting with prior μ' or μ'', and therefore

$$\lambda V(\mu') + (1 - \lambda)V(\mu'') \geq V(\mu).$$

  This just states the obvious fact that not reacting to the result of the experiment is a feasible strategy. What can you say about strict convexity of V?

- Budd, Harris, and Vickers (1993) considers models with almost myopic decision makers. Moscarini and Smith (2002) analyzes demand for information near certainty and Moscarini and Smith (2001) provides a nice connection to the sequential statistical decision problem of Wald (1949) (i.e. when to accept a hypothesis vs. collect more observations).

2.4 Examples

1. Learning the bias in a coin. Suppose a coin is tossed independently. The result of each toss X_t is independent of the result of other tosses conditional on the bias θ of the coin. The bias just gives the probability of observing heads in a single toss. Assume that the prior on Θ is

the uniform distribution on $[0,1]$. Compute the posterior after a first observation of heads (H). or tails (T) as follows:

$$\mu(\theta;H)=\frac{\theta}{\int_0^1\theta d\theta}=2\theta,\mu(\theta;T)=\frac{1-\theta}{\int_0^1(1-\theta)d\theta}=2-2\theta.$$

More generally, recall that Beta function $B(\alpha,\beta)$ is defined as:

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx,$$

and Beta -distribution with parameters α,β for $\theta\in[0,1]$ is:

$$p(\theta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}.$$

The probability of getting n heads in t trials is:

$$\binom{t}{n}\theta^n(1-\theta)^{t-n}.$$

By plugging into Bayes' rule, you can verify that if the prior is distributed as Beta distribution with parameters α,β and n heads are observed in t trials, then the posterior is a Beta -distribution with parameters $\alpha+n,\beta+(t-n)$. Note that the uniform distribution is a Beta. -distribution with parameters $1,1$ and the two posteriors computed above comply with the general formula.

2. Learning via pricing the (common) valuation of a sequence of customers. Consider next the optimal pricing problem of a monopolist with zero cost facing a sequence of customers with unit demand and with a common willingness to pay θ for the good. Assume that the monopolist's prior belief on θ is that it is uniformly distributed on $[0,1]$. In each t, the monopolist sets a price a_t and a sale is realized, i.e. $x_t=1$ if $\theta\geq a_t$ and $x_t=0$ if $a_t>\theta$. The stage payoff $r(x_t,a_t)=x_ta_t$. Let $\underline{\theta}_t=\max\{a_s\mid x_s=1,s<t\}$, and $\overline{\theta}_t=\min\{a_s\mid x_s=0,s<t\}$.

Then the monopolist's posterior on Θ before setting the price in period t is uniform on $[\underline{\theta}_t, \overline{\theta}_t]$. Consider then the following Bellman equation:

$$V(\underline{\theta}, \overline{\theta}) = \max_{\underline{\theta} \le a \le \overline{\theta}} (1-\delta)\frac{\overline{\theta} - a}{\overline{\theta} - \underline{\theta}} a + \delta \frac{\overline{\theta} - a}{\overline{\theta} - \underline{\theta}} V(a, \overline{\theta}) + \delta \frac{a - \underline{\theta}}{\overline{\theta} - \underline{\theta}} V(\underline{\theta}, a).$$

It is a very good exercise to show formally that this Bellman equation has a unique solution and that the solution is continuous and convex and that an optimal pricing strategy exists. I invite you to solve his problem and in particular, you should determine if complete learning results.

3. Monopolist learning the slope of the demand curve. Consider next a monopolist with zero cost facing uncertain noisy demand. The demand depends on a binary parameter $\theta \in \{0, 1\}$, and market probability of making a sale at price a_t is given by

$$\pi_t = \alpha_\theta - \beta_\theta p_t,$$

over a suitable range of prices (to make probabilities well defined). The monopolist maximizes expected profit $\pi_t p_t$. See McLennan (1984) how full learning may fail in this setting for low enough δ and also see Harrison, Keskin, and Zeevi (2012) and Loertscher and McLennan (2020) for more recent developments and variations on related topics.

References

AGHION, P., P. BOLTON, C. HARRIS, AND B. JULLIEN (1991): "Optimal Learning by Experimentation," *Review of Economic Studies*, 58, 621–654.

BUDD, C., C. HARRIS, AND J. VICKERS (1993): "A Model of the Evolution of Duopoly: Does the Asymmetry Between Firms Tend to Increase or Decrease?," *Review of Economic Studies*, 60, 543–573.

EASLEY, D., AND N. KIEFER (1988): "Controlling a Stochastic Process with Unknown Parameters," *Econometrica*, 56, 1045–1064.

HARRISON, J. M., N. B. KESKIN, AND A. ZEEVI (2012): "Bayesian Dynamic Pricing Policies: Learning and Earning Under a Binary Prior Distribution," *Management Science*, 58(3), 570–586.

LOERTSCHER, S., AND A. MCLENNAN (2020): "Some People Never Learn, Rationally: Multidimensional Learning Traps and Smooth Solutions to Dynamic Programming," Discussion paper.

MCLENNAN, A. (1984): "Price Dispersion and Incomplete Learning in the Long-Run," *Journal of Economic Dynamics and Control*, 7, 331–347.

MOSCARINI, G., AND L. SMITH (2001): "The Optimal Level of Experimentation," *Econometrica*, 69(6), 1629–1644.

——— (2002): "The Law of Large Demand for Information," *Econometrica*, 70(6), 2351–2366.