Perfect Bayesian Equilibrium

Pierpaolo Battigalli
Bocconi University
Game Theory: Analysis of Strategic Thinking

May 6, 2026

Abstract

Much like Nash equilibria of multistage games with complete information, Bayesian equilibria of multistage Bayesian games may allow for non-best-reply continuation strategies starting from histories that are not supposed to occur in equilibrium. Standard game theory fixes the problem in multistage games with complete information (and observed actions) by imposing a subgame-perfection requirement. The same kind of fix has been pursued for multistage Bayesian games, giving rise to notions of "perfect Bayesian equilibrium." Here we consider the most general one among those satisfying a minimal Bayes-consistency requirement.

[These slides summarize and, in part, complement Section 15.7 of Chapter 15 of GT-AST.]

Multistage Bayesian Games

A **simple multistage Bayesian game** (with observed actions) is a *finite* structure

$$\Gamma = \left\langle I, \left(\Theta_i, A_i, \mathcal{A}_i(\cdot), u_i, (p_i(\cdot|\theta_i))_{\theta_i \in \Theta_i}\right)_{i \in I}\right\rangle \text{ where}$$

- $\langle I, (\Theta_i, A_i, \mathcal{A}_i(\cdot), u_i)_{i \in I}\rangle$ is a *finite* game with payoff uncertainty;
- For each $i \in I$ and $\theta_i \in \Theta_i$, $p_i(\cdot|\theta_i) \in \Delta(\Theta_{-i})$ is the *initial* exogenous belief of type θ_i of player i, also called **interim belief**.
- Indeed, w.l.o.g., we may posit **prior** beliefs $P_i \in \Delta(\Theta)$ such that $P_i(\theta_i) := P_i(\{\theta_i\} \times \Theta_{-i}) > 0$ for each θ_i and

$$\forall \theta_{-i} \in \Theta_{-i},\ p_i(\theta_{-i}|\theta_i) = \frac{P_i(\theta_i, \theta_{-i})}{P_i(\theta_i)}$$

  (this is w.l.o.g. if we allow for heterogeneous priors).
- We say *"simple"* because, compared with the analysis of static Bayesian games, here *we assume that Harsanyi types coincide with information types*: $T_i \cong \Theta_i$ for each $i \in I$.

Bayesian equilibrium

- If we define Γ using priors, we can characterize Bayesian equilibrium as the Nash equilibrium of the ex ante strategic form

$$\mathcal{AS}(\Gamma) = \left\langle I, (\Sigma_i, U_i)_{i \in I} \right\rangle,$$

where

  - $\Sigma_i = S_i^{\Theta_i} = \left(\times_{h \in H} \mathcal{A}_i(h)\right)^{\Theta_i}$;
  - $\forall \sigma \in \times_{i \in} \Sigma_i,\ U_i(\sigma) = \sum_{\theta \in \Theta} P_i(\theta) u_i(\theta, \zeta(\sigma(\theta)))$.

- Of course, Bayesian equilibrium suffers from the same problem as Nash eq.: an equilibrium σ^* may be s.t., for some i, θ_i, and h with $\mathrm{P}^{\sigma^*}(h) = 0$, $\sigma_i^*(\theta_i)$ is not a best reply in the h-continuation.
- We try to fix this with an *extension of the subgame perfect equilibrium idea*. Players' behavior is described by randomized decision rules

$$(\beta_i(\cdot|\theta_i, h))_{\theta_i \in \Theta_i, h \in H} \in B_i^{\Theta_i} = \left(\times_{h \in H} \Delta(\mathcal{A}_i(h))\right)^{\Theta_i} \quad (i \in I),$$

that we call "**extended behavior strategies**" ($\beta_i(\cdot|\theta_i, \cdot)$ is the behavior strategy of type θ_i).

Perfect Bayesian equilibrium

General ideas

- Beside the candidate profile $\beta = (\beta_i)_{i \in I} \in \times_{i \in I} B_i^{\Theta_i}$, we need to specify candidate beliefs $(\mu_i(\cdot|\theta_i, h))_{i \in I, \theta_i \in \Theta_i}$ about θ_{-i} at each $h \in H$, otherwise we cannot compute conditional expected payoffs:

Definition

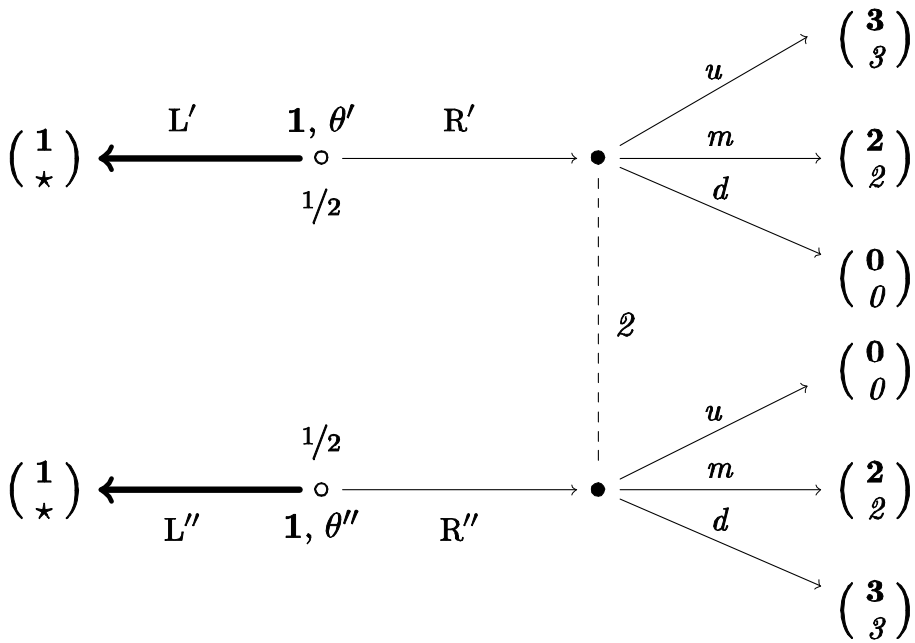
A **system of beliefs** (for Γ) is a conditional belief profile $\mu = (\mu_i)_{i \in I}$, where

$$\mu_i = (\mu_i(\cdot|\theta_i, h))_{\theta_i \in \Theta_i, h \in H} \in \Delta(\Theta_{-i})^{\Theta_i \times H}.$$

$\mu_i(\theta_{-i}|\theta_i, h)$=prob. assigned by type θ_i of i to θ_{-i} conditional on h, and $\mu_i(\cdot|\theta_i, \varnothing) = p_i(\cdot|\theta_i)$. A pair (β, μ) is called **assessment**.

- Clearly, beliefs μ *must be related to* β and are therefore **endogenous** [except for $\mu_i(\cdot|\theta_i, \varnothing)$]. Thus, a candidate *equilibrium* is not just an extended behavior strategy profile, but a *whole assessment* (β, μ).

Example



- $L'.L''$ cannot be part of an equilibrium. *Three equilibria:*
  - $(R'.R'', m) \Rightarrow \mu_2\left(\theta'|R\right) = p_2\left(\theta'\right) = \frac{1}{2}$,
  - $(R'.L'', u) \Rightarrow \mu_2\left(\theta'|R\right) = 1$,
  - $(L'.R'', d) \Rightarrow \mu_2\left(\theta'|R\right) = 0$.

Perfect Bayesian equilibrium

General ideas: connection to rational planning

- Each type θ_i has a **personal system of beliefs** $\mu_i(\cdot|\theta_i,\cdot) \in \Delta(\Theta_{-i})^H$ about others' types, and a **conjecture**

$$\beta^i = \left(\beta^i(\cdot|\theta_{-i},h)\right)_{\theta_{-i}\in\Theta_{-i},h\in H} \in \left(\times_{h\in H}\Delta(\mathcal{A}_{-i}(h))\right)^{\Theta_{-i}}$$

about others' behavior as a function of their types. *In* an *equilibrium* (β,μ), β^i corresponds to β_{-i}:

$$\beta^i(a_{-i}|\theta_{-i},h) = \prod_{j\neq i}\beta_j(a_j|\theta_j,h).$$

- Pair $\left(\mu_i(\cdot|\theta_i,\cdot),\beta_{-i}\right)$ is a **personal assessment** and it yields a **subjective decision tree**.
- **Personal Bayes consistency:** $\mu_i(\cdot|\theta_i,(h,(a_i,a_{-i})))$ is derived from $\mu_i(\cdot|\theta_i,h)$ and $\beta^i(\cdot|\theta_i,h)$ *via* Bayes rule when possible, and is independent of a_i (no unjustified change in beliefs).
- Under personal Bayes consistency the *OD principle holds.*

Perfect Bayesian equilibrium (PBE)

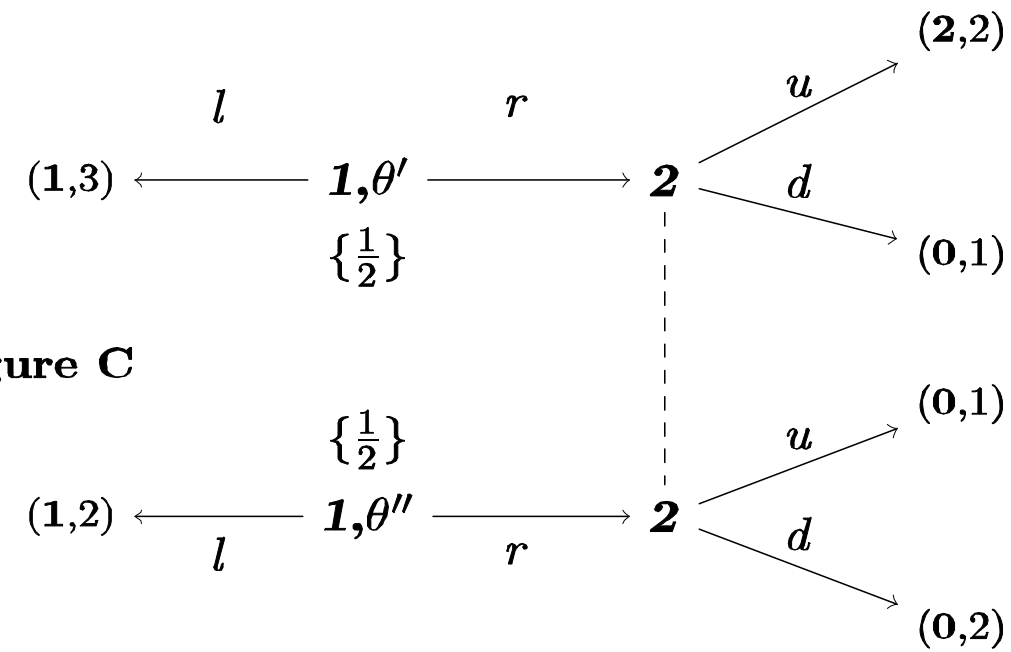
General ideas

- Fix a candidate equilibrium assessment (β, μ); when can we say that (β, μ) is a PBE?
- *Note:* for each $h \in H$, beliefs $(\mu_i(\cdot|\theta_i, h))_{i \in I, \theta_i \in \Theta_i}$ define a (h, μ)-**continuation game** $\Gamma(h, \mu)$ [consider the set of feasible continuations of h, $\{h' \in A^{\leq \mathbb{N}} : (h, h') \in \bar{H}\}$, the resulting θ-dependent payoffs and the interactive beliefs $(\mu_i(\cdot|\theta_i, h))_{i \in I, \theta_i \in \Theta_i}$].
- In a PBE (β, μ), for all i and θ_i, β_{i,θ_i} must be sequentially optimal given personal assessment $\left(\mu_i\left(\cdot|\theta_i, \cdot\right), \beta^i\right)$, where β^i corresponds to β_{-i} (correct conjecture, in a two-person game, $\beta^i = \beta_{-i}$). Thus:
  - **(Interpersonal) Bayes consistency:** the initial beliefs $(p_{\theta_i})_{i \in I, \theta_i \in \Theta_i}$, the system of beliefs μ and the behavior strategies β must be related to each other *via Bayes rule* (when possible).
  - **Continuation equilibrium** (often called—*misleadingly*—"sequential rationality"): for each $h \in H$, β induces a Bayesian equilibrium of the (h, μ)-continuation game $\Gamma(h, \mu)$.

Perfect Bayesian equilibrium: the state of the art

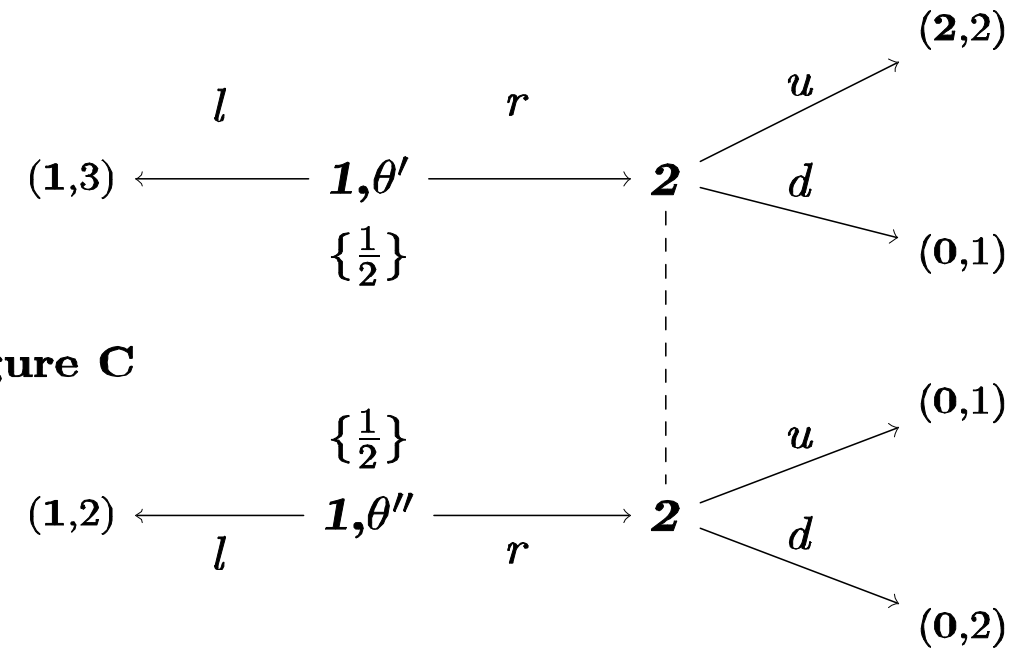
- Yet, *it is not obvious how to define "interpersonal Bayes consistency*."
- Game theorists have proposed different definitions. The reason is that, on top of mere consistency with Bayes rule (which often they failed to express well), they wanted to incorporate additional assumptions in the spirit of the Bayesian-Nash equilibrium analysis, such as
  1. players "update in the same way" (*differences in beliefs are only due to differences in information and priors*), and
  2. beliefs satisfy *independence across opponents*.
- Unfortunately, it was not even very clear which additional assumptions one was trying to incorporate in the PBE concept, nor how to exactly express them. Appeals to intuition and references to particular examples dominated the analysis.
- **Hence the mess:** *There is no universally accepted notion of PBE* despite the widespread application of the (fuzzily defined) "PBE"!

PBE: an example, 1



- $\left(\ell'.\ell'', d, \mu\left(\theta''|r\right) \geq 1/2\right)$ is a (set of) PBE(s):
  - d is a best reply to $\mu\left(\theta''|r\right) \geq 1/2$;
  - ℓ is a best reply to d for both θ' and θ'';
  - $\mu\left(\cdot|r\right)$ is not determined by $p\left(\cdot\right)$ and β_1, because $\beta_1\left(r|\theta'\right) = \beta_1\left(r|\theta''\right) = 0$.

PBE: an example, 2



- $\left(r'.\ell'', u, \mu\left(\theta''|r\right) = 0\right)$ is another PBE (the only one consistent with strong belief in rationality!):
  - u is a best reply to $\mu\left(\theta''|r\right) = 0$ (i.e., $\mu\left(\theta'|r\right) = 1$);
  - r is a best reply to u for θ' (and ℓ is dominant for θ'');
  - $\beta_1\left(r|\theta'\right) = 1 = \beta_1\left(\ell|\theta''\right)$ implies $\mu\left(\theta''|r\right) = 0$.

Bayes Consistency of Personal Assessments

- Recall, if a personal assessement $\left(\beta^i, \mu_i\right)$ is derived from a CPS $\bar{\mu}^i$, then it has to be **Bayes consistent:** For all $h \in H$, $a_{-i} \in \mathcal{A}_{-i}(h)$, θ_{-i}, write
  - $\mathrm{P}^{\beta^i}\left(a_{-i}|\theta_{-i}, h\right) := \beta^i\left(a_{-i}|\theta_{-i}, h\right)$, $\mathrm{P}^{\mu_i}\left(\theta_{-i}|h\right) := \mu_i\left(\theta_{-i}|h\right)$,
  - $\mathrm{P}^{\beta^i,\mu_i}\left(\theta_{-i}, a_{-i}|h\right) := \beta^i\left(a_{-i}|\theta_{-i}, h\right)\mu_i\left(\theta_{-i}|h\right)$,
  - $\mathrm{P}^{\beta^i,\mu_i}\left(a_{-i}|h\right) = \sum_{\theta'_{-i}} \mathrm{P}^{\beta^i,\mu_i}\left(\theta'_{-i}, a_{-i}|h\right) = \sum_{\theta'_{-i}} \beta^i\left(a_{-i}|\theta'_{-i}, h\right)\mu_i\left(\theta'_{-i}|h\right)$.
  - If $\mathrm{P}^{\beta^i,\mu_i}\left(a_{-i}|h\right) > 0$, write $\mu_i\left(\theta_{-i}|h, a_{-i}\right) := \frac{\mathrm{P}^{\beta^i,\mu_i}(\theta_{-i},a_{-i}|h)}{\mathrm{P}^{\beta^i,\mu_i}(a_{-i}|h)}$ $= \frac{\beta^i(a_{-i}|\theta_{-i},h)\mu_i(\theta_{-i}|h)}{\sum_{\theta'_{-i}}\beta^i\left(a_{-i}|\theta'_{-i},h\right)\mu_i\left(\theta'_{-i}|h\right)}$ (BR).
  - **Bayes consistency**: for all $h \in H$ s.t. $L\left(\hat{\Gamma}(h)\right) > 1$, $a_i \in \mathcal{A}_i(h)$, $a_{-i} \in \mathcal{A}_{-i}(h)$, and θ_{-i}

$$\mu_i\left(\theta_{-i}|h, (a_i, a_{-i})\right) = \mu_i\left(\theta_{-i}|h, a_{-i}\right),$$

    where $\mu_i\left(\theta_{-i}|h, a_{-i}\right)$ satisfies (BR) whenever possible. (Hence, $\mu_i\left(\cdot|h, (a_i, a_{-i})\right)$ is *independent of own-action* a_i.)

PBE: A Minimalistic, but Rigorous Approach

Fix $i \in I$, a profile $\beta_{-i} = (\beta_j)_{j \neq i}$ and a conjecture β^i. We say that β^i **corresponds to** β_{-i} if, for all $h \in H$, $a_{-i} \in \mathcal{A}_{-i}(h)$, and $\theta_{-i} \in \Theta_{-i}$, $\beta^i(a_{-i}|\theta_{-i}, h) = \prod_{j \neq i} \beta_j(a_j|\theta_j, h)$.

Definition

Assessment (β, μ) is **Bayes consistent** if, for every $i \in I$ and $\theta_i \in \Theta_i$, the personal assessment $\left(\beta^i, \mu_i(\cdot|\theta_i, \cdot)\right)$ where β^i corresponds to β_{-i} is Bayes consistent.

Definition

Assessment (β, μ) is a **perfect Bayesian equilibrium (PBE)** if it is Bayes consistent and, for every $i \in I$ and $\theta_i \in \Theta_i$, behavior strategy $\beta_i(\cdot|\theta_i, \cdot)$ is sequentially optimal given $\left(\beta^i, \mu_i(\cdot|\theta_i, \cdot)\right)$, where β^i corresponds to β_{-i}.

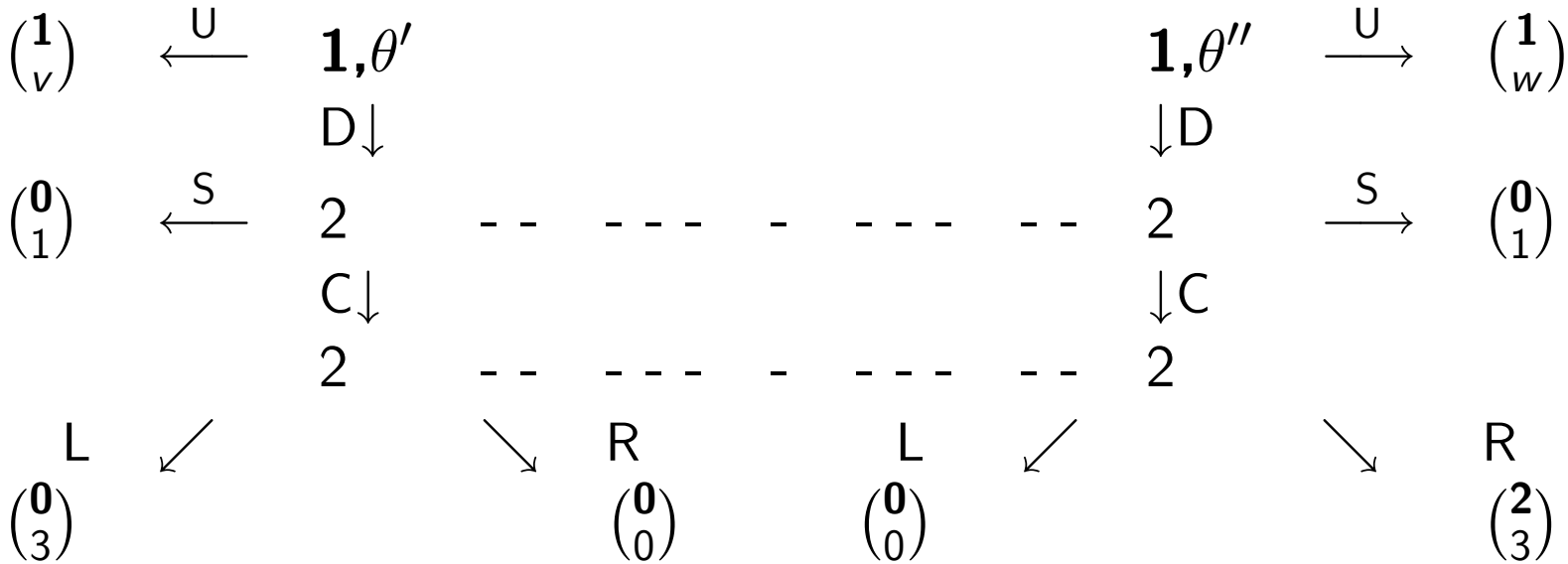
OD Principle for PBE

- From the *OD principle* for personal assessments we obtain the OD Principle for PBE:

Corollary

An assessment (β, μ) *satisfying Bayes consistency is a PBE if and only if, for every* $i \in I$ *and* $\theta_i \in \Theta_i$*, behavior strategy* $\beta_i(\cdot|\theta_i, \cdot)$ *is one-step optimal given* $\left(\beta^i, \mu_i(\cdot|\theta_i, \cdot)\right)$*, where* β^i *corresponds to* β_{-i}*.*

Two-Person Example



- Bayes consistency implies $\mu_2\left(\theta'|\mathrm{D}\right) = \mu_2\left(\theta'|\mathrm{D},\mathrm{C}\right)$. With this:
  - if $\mu_2\left(\theta'|\mathrm{D},\mathrm{C}\right) > \frac{1}{2}$, $\beta_2\left(\mathrm{L}|\mathrm{D},\mathrm{C}\right) = 1$, $\beta_2\left(\mathrm{C}|\mathrm{D}\right) = 1$, $\beta_1\left(\mathrm{U}|\theta'\right) = \beta_1\left(\mathrm{U}|\theta''\right) = 1$ gives an equilibrium (BR ok);
  - if $\mu_2\left(\theta'|\mathrm{D},\mathrm{C}\right) = \frac{1}{2}$, $\beta_2\left(\mathrm{L}|\mathrm{D},\mathrm{C}\right) \geq \frac{1}{2}$, $\beta_2\left(\mathrm{C}|\mathrm{D}\right) = 1$, $\beta_1\left(\mathrm{U}|\theta'\right) = \beta_1\left(\mathrm{U}|\theta''\right) = 1$ also gives an eq (BR ok);
  - if $\mu_2\left(\theta'|\mathrm{D},\mathrm{C}\right) < \frac{1}{2}$, $\beta_2\left(\mathrm{R}|\mathrm{D},\mathrm{C}\right) = 1$, $\beta_2\left(\mathrm{C}|\mathrm{D}\right) = 1$, $\beta_1\left(\mathrm{U}|\theta'\right) = \beta_1\left(\mathrm{D}|\theta''\right) = 1$, $\mu_2\left(\theta''|\mathrm{D}\right) = \mu_2\left(\theta'|\mathrm{D},\mathrm{C}\right) = 1$ also eq.

Three-Person Example

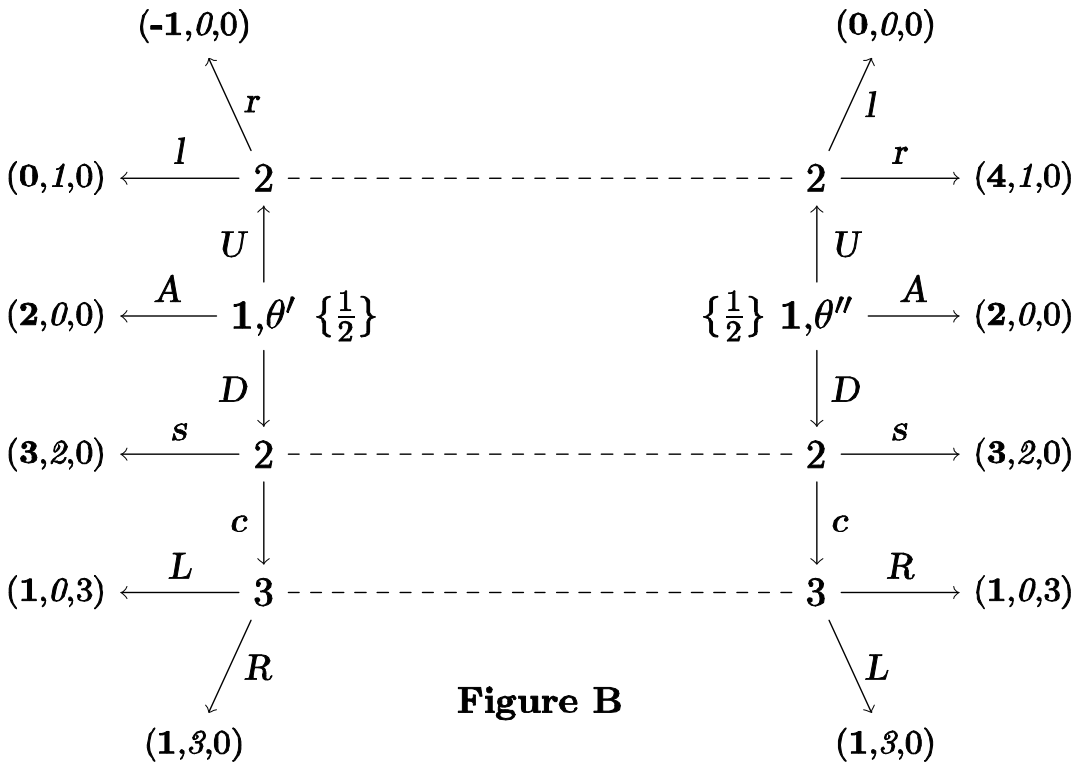


Figure B

- We may have
  - (despite the common prior) $\mu_2(\cdot|D) \neq \mu_3(\cdot|D)$ IF $\beta_1(D|\theta') = \beta_1(D|\theta'') = 0$,
  - $\mu_3(\cdot|D) \neq \mu_3(\cdot|D, c)$ IF $\beta_2(c|D) = 0$.
- Stronger notions of PBE do not allow this (see the references).

Three-Person Example: Additional Requirements

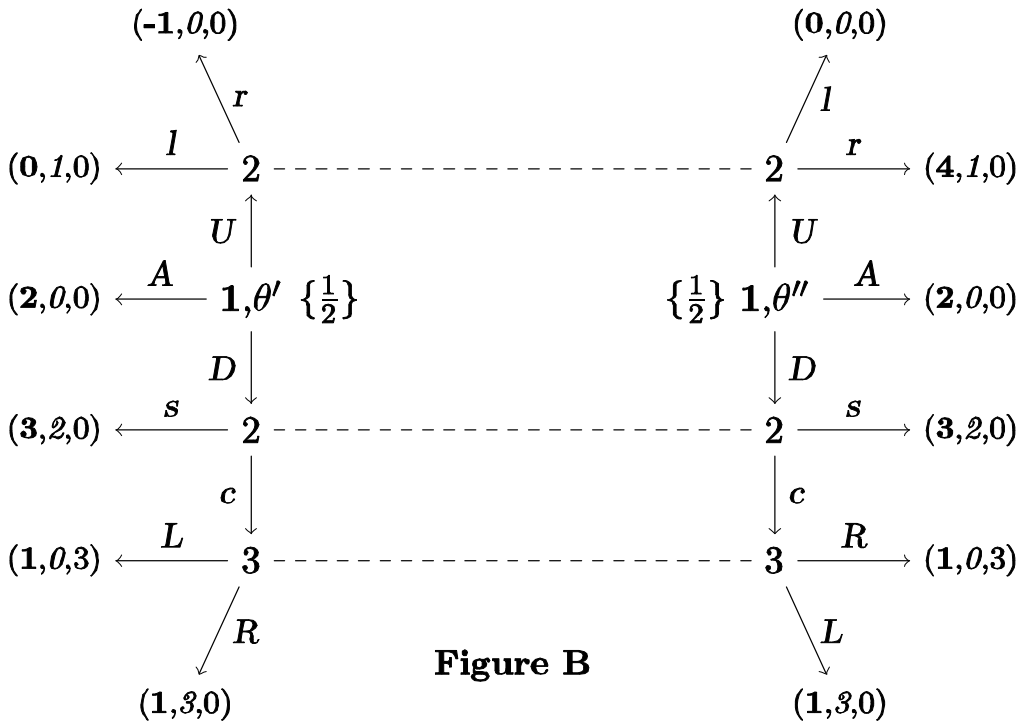


Figure B

- Additional requirements *on top of Bayes consistency* yield stronger notions of PBE:
  - (C) *Common Information* $\Rightarrow$ *Common Beliefs*: $\mu_2(\cdot|D) = \mu_3(\cdot|D)$ [even if D is "surprising", i.e., even if $\beta_1(D|\theta') = \beta_1(D|\theta'') = 0$].
  - (I) *Independent Updating*: 2's action cannot signal 1's private information $\Rightarrow \mu_3(\cdot|D) = \mu_3(\cdot|D, c)$ [even if $\beta_2(c|D) = 0$].

References

- Battigalli, P., E. Catonini, & N. De Vito (2025): *Game Theory: Analysis of Strategic Thinking*. Typescript, Bocconi University.
- Battigalli, P. (2025): *Mathematical Language and Game Theory.* Typescript, Bocconi University.
- Battigalli, P. (1996): "Strategic Independence and Perfect Bayesian Equilibria," *Journal of Economic Theory*, 70, 201-234.
- Fudenberg, D., and J. Tirole (1991): "Perfect Bayesian Equilibria," *Journal of Economic Theory,* 53, 236-260.