On Convergence and Optimality of Best-Response Learning with Policy Types in Multiagent Systems

—

Appendix

Stefano V. Albrecht
The University of Edinburgh
Edinburgh, United Kingdom
s.v.albrecht@sms.ed.ac.uk

Subramanian Ramamoorthy
The University of Edinburgh
Edinburgh, United Kingdom
s.ramamoorthy@ed.ac.uk

Proof of Theorem 1

Theorem 1 (restated from [1])**.** Let Γ be a SBG with a pure type distribution Δ. If HBA uses a product posterior and if the prior probabilities P_j are positive ($\forall \theta_j^* \in \Theta_j^* : P_j(\theta_j^*) > 0$), then:
for any $\epsilon > 0$, there is a time t from which ($\tau \geq t$)

$$P_{\text{Pr}}(H^\tau, H^\infty)(1-\epsilon) \leq P_\Delta(H^\tau, H^\infty) \leq (1+\epsilon) P_{\text{Pr}}(H^\tau, H^\infty) \tag{1}$$

for all H^∞ with $P_\Delta(H^\tau, H^\infty) > 0$.

Proof. Kalai and Lehrer [2] studied a model which can be equivalently described as a single-state SBG (i.e. $|S| = 1$) with a pure type distribution and product posterior. They showed that, if the player's assessment of future play is *absolutely continuous* with respect to the true probabilities of future play (i.e. any event that has true positive probability is assigned positive probability by the player), then (1) must hold. In our case, absolute continuity always holds by Assumption 5 and the fact that the prior probabilities P_j are positive, as well as the fact that the type distribution is pure (from which we can infer that the true types always have positive posterior probability).

In this proof, we seek to extend the convergence result of Kalai and Lehrer (henceforth [2]) to multi-state SBGs with pure type distributions. Our strategy is to translate a SBG Γ into a *modified SBG* $\hat{\Gamma}$ which is equivalent to Γ in the sense that the players behave identically, and which is compatible to the model used in [2] in the sense that the informational assumptions of [2] ignore the differences. We achieve this by introducing a new player *nature*, denoted ξ, which emulates the transitions of Γ in $\hat{\Gamma}$.

Given a SBG $\Gamma = (S, s^0, \bar{S}, N, A_i, \Theta_i, u_i, \pi_i, T, \Delta)$, we define the modified SBG $\hat{\Gamma}$ as follows: Firstly, $\hat{\Gamma}$ has only one state, which can be arbitrary since it has no effect. The players in $\hat{\Gamma}$ are $\hat{N} = N \cup \{\xi\}$ where $i \in N$ have the same actions and types as in Γ (i.e. A_i and Θ_i), and where we define the actions and types of ξ to be $A_\xi = \Theta_\xi = S$ (i.e. nature's actions and types correspond to the states of Γ). The payoffs of ξ are always zero and the strategy of ξ at time t is defined as

$$\pi_\xi^t(H^\tau, a_\xi, \theta_\xi) = \begin{cases} 0 & \tau = t, a_\xi \not\equiv \theta_\xi \\ 1 & \tau = t, a_\xi \equiv \theta_\xi \\ T(a_\xi^{\tau-1}, (a_i^{\tau-1})_{i \in N}, a_\xi) & \tau > t \end{cases}$$

where H^τ is any history of length $\tau \geq t$. (H^τ allows the players $i \in N$ to use π_ξ^t for future predictions about ξ's actions. This will be necessary to establish equivalence of $\hat{\Gamma}$ and Γ.)

The purpose of ξ is to emulate the state transitions of Γ. Therefore, the modified strategies $\hat{\pi}_i$ and payoffs $\hat{u}_i$ of $i \in N$ are now defined with respect to the actions and types (since the current type of ξ determines its next action) of ξ. Formally, $\hat{\pi}_i(H^t, a_i, \theta_i) = \pi_i(\bar{H}^t, a_i, \theta_i)$ where

$$\bar{H}^t = (\theta_\xi^0, (a_i^0)_{i \in N}, \theta_\xi^1, (a_i^1)_{i \in N}, ..., \theta_\xi^t)$$

and $\hat{u}_i(s, a^t, \theta_i^t) = u_i(\theta_\xi^t, (a_j^t)_{j \in N}, \theta_i^t)$, where s is the only state of $\hat{\Gamma}$ and $a^t \in \times_{i \in \hat{N}} A_i$.

Finally, $\hat{\Gamma}$ uses two type distributions, Δ and Δ_ξ, where Δ is the type distribution of Γ and Δ_ξ is defined as $\Delta_\xi(H^t, \theta_\xi) = T(a_\xi^{t-1}, (a_i^{t-1})_{i \in N}, \theta_\xi)$. If s^0 is the initial state of Γ, then $\Delta_\xi(H^0, \theta_\xi) = 1$ for $\theta_\xi \equiv s^0$.

The modified SBG $\hat{\Gamma}$ proceeds as the original SBG Γ, except for the following changes: *(a)* Δ is used to sample the types for $i \in N$ (as usual) while Δ_ξ is used to sample the types for ξ; *(b)* Each player is informed about its own type *and* the type of ξ. This completes the definition of $\hat{\Gamma}$.

The modified SBG $\hat{\Gamma}$ is equivalent to the original SBG Γ in the sense that the players $i \in N$ have identical behaviour in both SBGs. Since the players always know the type of ξ, they also know the next action of ξ, which corresponds to knowing the current state of the game. Furthermore, note that the strategy of ξ uses two time indices, t and τ, which allow it to distinguish between the current time ($\tau = t$) and a future time ($\tau > t$). This means that π_ξ^t can be used to compute expected payoffs in $\hat{\Gamma}$ in the same way as T is used to compute expected payoffs in Γ. In other words, the formulas (2) and (3) can be modified in a straightforward manner by replacing the original components of Γ with the modified components of $\hat{\Gamma}$, yielding the same results. Finally, since $\hat{\Gamma}$ uses the same type distribution as Γ to sample types for $i \in N$, there are no differences in their payoffs and strategies.

To complete the proof, we note that *(a)* and *(b)* are the only procedural differences between the modified SBG and the model used in [2]. However, since we specify that the players always know the type of ξ, there is no need to learn the type distribution Δ_ξ, hence *(a)* and *(b)* have no effect in [2]. The important point is that [2] assume a model in which the players only interact with other players, but not with an environment. Since we eliminated the environment by replacing it with a player ξ, this is precisely what happens in the modified SBG. Therefore, the convergence result of [2] carries over to multi-state SBGs with pure type distributions. □

Corrigendum

The original paper published in the *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence* contained an incorrect Theorem 1. The theorem claimed that, under a pure type distribution and using a product posterior, HBA would eventually learn the true type distribution of the game. In fact, the proof was exactly the same as the one given in this appendix. However, the error was in the implicit assumption that the convergence result of Kalai and Lehrer [2] would imply knowledge of the true type distribution. Unfortunately, there is a subtle but important asymmetry between making correct future predictions and knowing the true type distribution: while the latter implies the former, examples can be created to show that the reverse is not generally true. Therefore, while HBA is guaranteed to make correct future predictions after some time, it is not guaranteed to learn the type distribution of the game. This has now been corrected.

References

[1] S. V. Albrecht and S. Ramamoorthy. On convergence and optimality of best-response learning with policy types in multiagent systems. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, Quebec City, Canada, July 2014.

[2] E. Kalai and E. Lehrer. Rational learning leads to Nash equilibrium. *Econometrica*, pages 1019–1045, 1993.