Policy-Gradient-Like Derivation for Shared-Parameter Agents

Setting Two agents A and B interact synchronously with a shared data stream D. At each step t (Moore timing), each agent emits a message based on its internal state, then observes (D_t, partner's message at t) to update its state. Each agent $X \in \{A, B\}$ has parameters θ_X, *shared* between its policy and predictor: the policy π_{θ_X} samples messages and the predictor P_{θ_X} assigns likelihoods to D given the partner's message. Let $\theta = (\theta_A, \theta_B)$.

Objective (infinite horizon) Define the per-step cooperative reward

$$r_t \;=\; \ln P_{\theta_B}(D_t \mid m^t_{A\to B}) \;+\; \ln P_{\theta_A}(D_t \mid m^t_{B\to A}), \qquad t = 1, 2, \ldots$$

Two standard infinite-horizon objectives are

$$J_{\text{avg}}(\theta) := \lim_{T\to\infty} \mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} r_t\Big],$$

$$J_\gamma(\theta) := \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^{\,t-1}\, r_t\Big], \qquad \gamma \in (0, 1).$$

When r_t is bounded and suitable ergodicity holds, $(1-\gamma)\, J_\gamma(\theta) \to J_{\text{avg}}(\theta)$ as $\gamma \uparrow 1$.

Equivalently, maximizing J_γ is maximizing a discounted product of the data likelihoods assigned by the predictors across time (the exogenous $P(D_t)$ drops out since it is θ-independent):

$$J_\gamma(\theta) = \mathbb{E}\Big[\sum_{t\ge 1} \gamma^{\,t-1} \ln P_{\theta_B}(D_t \mid m^t_{A\to B}) + \gamma^{\,t-1} \ln P_{\theta_A}(D_t \mid m^t_{B\to A})\Big] \tag{1}$$

$$\iff \text{ maximize } \mathbb{E}\Big[\prod_{t\ge 1} \big(P_{\theta_B}(D_t \mid m^t_{A\to B})\, P_{\theta_A}(D_t \mid m^t_{B\to A})\big)^{\gamma^{t-1}}\Big]. \tag{2}$$

Trajectory distribution Under Moore timing and deterministic state updates, an infinite trajectory $\tau = (D_t, m^t_{A\to B}, m^t_{B\to A})_{t\ge 1}$ has likelihood

$$P_\theta(\tau) \;=\; \prod_{t\ge 1} P(D_t)\, \pi_{\theta^t_A}(m^t_{A\to B})\, \pi_{\theta^t_B}(m^t_{B\to A}),$$

where $P(D_t)$ is exogenous and θ^t_X denotes the (possibly updated) parameters of agent X at time t.

Gradient decomposition (discounted infinite horizon) Let the discounted return-from-time-t be $G_t = \sum_{k=t}^{\infty} \gamma^{\,k-t}\, r_k$. Using the likelihood-ratio identity and exchanging sums (justified by bounded r_t), the per-agent gradients are

$$\nabla_{\theta_A} J_\gamma(\theta) = \mathbb{E}\left[\sum_{t\ge 1} \nabla_{\theta_A} \ln \pi_{\theta_A}(m^t_{A\to B})\; G_t\right] \;+\; \mathbb{E}\left[\sum_{t\ge 1} \gamma^{\,t-1}\, \nabla_{\theta_A} \ln P_{\theta_A}(D_t \mid m^t_{B\to A})\right],$$

$$\nabla_{\theta_B} J_\gamma(\theta) = \mathbb{E}\left[\sum_{t\ge 1} \nabla_{\theta_B} \ln \pi_{\theta_B}(m^t_{B\to A})\; G_t\right] \;+\; \mathbb{E}\left[\sum_{t\ge 1} \gamma^{\,t-1}\, \nabla_{\theta_B} \ln P_{\theta_B}(D_t \mid m^t_{A\to B})\right].$$

The first term is the usual REINFORCE policy gradient with return G_t; the second arises because r_t depends on θ through the predictors.

Time-varying parameters without meta-parameters Treat the per-time parameters θ_X^t as the variables we update online. Consider the discounted data log-likelihood along the trajectory

$$\mathcal{L}_\gamma^{\mathrm{data}}(\tau;\theta^{1:\infty}) = \sum_{t\geq 1} \gamma^{t-1} \Big(\ln P_{\theta_B^t}(D_t \mid m_{A\to B}^t) + \ln P_{\theta_A^t}(D_t \mid m_{B\to A}^t) \Big),$$

where $\ln P(D_t)$ is omitted as θ-independent. Maximizing $\mathbb{E}[\mathcal{L}_\gamma^{\mathrm{data}}]$ by steepest ascent with respect to each θ_X^t yields the online MLE-style updates

$$\theta_X^{t+1} = \theta_X^t + \eta_t \Big(\underbrace{\nabla_{\theta_X^t} \ln \pi_{\theta_X^t}(m_X^t)\, G_t}_{\text{policy (REINFORCE) term}} + \underbrace{\gamma^{t-1} \nabla_{\theta_X^t} \ln P_{\theta_X^t}(D_t \mid m_{\bar{X}\to X}^t)}_{\text{predictor (supervised) term}} \Big),$$

which are exactly the gradients of $\mathbb{E}[\mathcal{L}_\gamma^{\mathrm{data}}]$ w.r.t. θ_X^t. A baseline b_t can be subtracted from G_t without bias. This ties the update rule directly to maximizing the likelihood of the entire (discounted) data sequence, accounting for the exogenous D_t that θ cannot influence.

Single-step Monte Carlo and online updates At time t, let $\widehat{G}_t$ be any causal estimator with $\mathbb{E}[\widehat{G}_t \mid \mathcal{F}_t] = G_t$ (e.g., sampled discounted return from t onward, or a bootstrapped critic). With any baseline b_t independent of m_X^t given $\mathcal{F}_t$, the ascent-style online updates for *time-varying* parameters are

$$\theta_A \leftarrow \theta_A \;+\; \eta_t \Big[(\widehat{G}_t - b_t)\, \nabla_{\theta_A} \ln \pi_{\theta_A}(m_{A\to B}^t) \;+\; \gamma^{t-1}\, \nabla_{\theta_A} \ln P_{\theta_A}(D_t \mid m_{B\to A}^t) \Big],$$
$$\theta_B \leftarrow \theta_B \;+\; \eta_t \Big[(\widehat{G}_t - b_t)\, \nabla_{\theta_B} \ln \pi_{\theta_B}(m_{B\to A}^t) \;+\; \gamma^{t-1}\, \nabla_{\theta_B} \ln P_{\theta_B}(D_t \mid m_{A\to B}^t) \Big].$$

Why unbiased? Conditioning on the filtration $\mathcal{F}_t$ up to emission and treating θ^t as fixed at time t (since emission precedes the update),

$$\mathbb{E}\big[(\widehat{G}_t - b_t)\, \nabla \ln \pi_{\theta_X}(m_X^t) \mid \mathcal{F}_t\big] = \big(G_t - b_t\big)\, \mathbb{E}[\nabla \ln \pi_{\theta_X}(m_X^t) \mid \mathcal{F}_t] = G_t\, \nabla \ln \pi_{\theta_X}(m_X^t),$$

since $\mathbb{E}[\nabla \ln \pi_{\theta_X^t}(m_X^t) \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[b_t\, \nabla \ln \pi_{\theta_X^t}(m_X^t) \mid \mathcal{F}_t] = 0$. Taking full expectations and summing over t recovers the policy-gradient terms in ∇J_γ evaluated at θ^t. The supervised terms are already unbiased for the predictor gradients. Thus $\mathbb{E}[\Delta\theta_X^t] \propto \nabla_{\theta_X} J_\gamma(\theta^t)$.

With diminishing step sizes $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$ (Robbins–Monro), the iterates converge to stationary points under standard regularity assumptions. With small constant η_t, the updates ascend J_γ in expectation.

Average-reward variant Under ergodicity and bounded rewards, replacing G_t with differential returns and taking $\gamma \uparrow 1$ yields estimators for $\nabla J_{\mathrm{avg}}(\theta)$. In practice one uses large γ (e.g., 0.99–0.999) as a low-variance surrogate for average reward.