Coalgebraic Prequential Compression and Online Updates

1 Coalgebraic setup

Let $F(S) = \mathsf{Out} \times S^{\mathsf{In}}$ denote the Moore functor. An agent $X \in \{A, B\}$ is an F-coalgebra (S_X, α_X) with

$$\alpha_X(s) = (\mathrm{out}_X(s), \mathrm{upd}_X(s)) \in \mathsf{Out}_X \times S_X^{\mathsf{In}_X}.$$

We use Moore timing: at step t each agent emits from its current state, then consumes inputs to update the next state. Let $S = S_A \times S_B$ and $\alpha = \alpha_A \times \alpha_B$ be the product coalgebra. The environment provides an exogenous data stream $(D_t)_{t \geq 1}$ that does not depend on the agents' parameters.

For each X, we factor its output into

$$\mathrm{out}_X(s_X^t) = m_X^t, \qquad m_X^t \sim \pi_{\theta_X^t}(\cdot \mid s_X^t),$$

and it carries a predictor $P_{\theta_X^t}(\cdot \mid m_{\bar{X}}^t)$ for the shared datum D_t, where $\bar{X}$ denotes the partner of X. The update is deterministic

$$s_X^{t+1} = U_X\big(s_X^t,\, D_t,\, m_{\bar{X}}^t\big), \qquad \theta_X^{t+1} \subset s_X^{t+1}.$$

Thus, given (s_A^1, s_B^1) and $(D_t)_{t \geq 1}$, the product coalgebra induces a unique infinite stream of states and outputs.

2 Likely = compressible via prequential code

By the coding theorem, higher probability corresponds to shorter codelengths under a (semi-)measure. Operationally we use *prequential* (online) coding: at step t, the system assigns a probability to the next symbol conditioned on its current state, and the codelength is the negative log-probability. We focus on compressing the *data* sequence using the agents' internal predictors (messages act as an emergent code used by the partner).

Definition 1 (Per-step discounted codelength)**.** *For $\gamma \in (0, 1)$, define*

$$\ell_t \;=\; -\ln P_{\theta_B^t}(D_t \mid m_A^t) \;-\; \ln P_{\theta_A^t}(D_t \mid m_B^t),$$

and the discounted prequential codelength

$$L_\gamma \;=\; \mathbb{E}\Big[\sum_{t \geq 1} \gamma^{t-1}\, \ell_t\Big].$$

Since $P(D_t)$ is exogenous, it contributes no parameter-dependent codelength. Minimizing L_γ is equivalent to maximizing the expected discounted predictive log-likelihood assigned to the realized data stream.

Optional: compressing messages as well If desired, include $-\ln \pi_{\theta_A^t}(m_A^t)$ and $-\ln \pi_{\theta_B^t}(m_B^t)$ inside ℓ_t to compress the full trace (messages + data). We focus on data-only codelengths below.

3 Online gradient and update law

Let the per-step reward be $r_t = \ln P_{\theta_B^t}(D_t \mid m_A^t) + \ln P_{\theta_A^t}(D_t \mid m_B^t)$ and define the discounted return-from-t:

$$G_t \;=\; \sum_{k=t}^{\infty} \gamma^{k-t}\, r_k.$$

Then $L_\gamma = -\,\mathbb{E}[\sum_{t\geq 1} \gamma^{t-1} r_t]$ up to an additive constant. With Moore timing (emission precedes update) and time-varying parameters θ_X^t, the stochastic gradients at time t decompose as

$$\nabla_{\theta_A^t}\big(-\mathbb{E}[L_\gamma]\big) = \mathbb{E}\Big[\nabla_{\theta_A^t} \ln \pi_{\theta_A^t}(m_A^t)\; G_t\Big] \;+\; \mathbb{E}\Big[\gamma^{t-1}\, \nabla_{\theta_A^t} \ln P_{\theta_A^t}(D_t \mid m_B^t)\Big],$$
$$\nabla_{\theta_B^t}\big(-\mathbb{E}[L_\gamma]\big) = \mathbb{E}\Big[\nabla_{\theta_B^t} \ln \pi_{\theta_B^t}(m_B^t)\; G_t\Big] \;+\; \mathbb{E}\Big[\gamma^{t-1}\, \nabla_{\theta_B^t} \ln P_{\theta_B^t}(D_t \mid m_A^t)\Big].$$

Therefore the online steepest-ascent updates that *minimize* the expected codelength L_γ (equivalently, maximize $-L_\gamma$) take the form

$$\theta_A^{t+1} = \theta_A^t + \eta_t \Big[(\widehat{G}_t - b_t)\, \nabla_{\theta_A^t} \ln \pi_{\theta_A^t}(m_A^t) + \gamma^{t-1}\, \nabla_{\theta_A^t} \ln P_{\theta_A^t}(D_t \mid m_B^t)\Big],$$
$$\theta_B^{t+1} = \theta_B^t + \eta_t \Big[(\widehat{G}_t - b_t)\, \nabla_{\theta_B^t} \ln \pi_{\theta_B^t}(m_B^t) + \gamma^{t-1}\, \nabla_{\theta_B^t} \ln P_{\theta_B^t}(D_t \mid m_A^t)\Big],$$

where $\widehat{G}_t$ is any causal estimator with $\mathbb{E}[\widehat{G}_t \mid \mathcal{F}_t] = G_t$, and b_t is a baseline independent of m_X^t given the emission filtration $\mathcal{F}_t$.

Proposition 1 (Unbiasedness and ascent-in-expectation)**.** *Conditioning on $\mathcal{F}_t$ and treating θ^t as fixed at emission,*

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

and the supervised terms are already unbiased. Hence $\mathbb{E}[\Delta\theta_X^t]$ is a (scaled) stochastic gradient step on $-\,\mathbb{E}[L_\gamma]$. With diminishing step sizes ($\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$), the iterates converge to stationary points under standard regularity conditions.

Interpretation The REINFORCE terms capture how the agent's message at time t influences *all* future discounted codelengths through the partner's predictive success, while the supervised terms capture the immediate predictive codelength at time t. Minimizing L_γ makes the behavior of the product coalgebra asymptotically as compressible as possible relative to the exogenous data stream.

4 Model description length (optional)

In a two-part MDL view, include a model cost for the initial joint state (s_A^1, s_B^1) and the deterministic update rules U_A, U_B. Given U, state transitions impose no residual codelength. Thus, the online updates above specialize to minimizing the prequential data codelength, while model complexity can be controlled by priors or explicit penalties.

5 Variants

- Compressing the full trace: add $-\ln \pi_{\theta_X^t}(m_X^t)$ to ℓ_t. - Entropy regularization: add $\beta\, \nabla_{\theta_X^t}\, \mathcal{H}(\pi_{\theta_X^t}(\cdot \mid s_X^t))$ to encourage exploration. - Average-reward limit: recover average codelength per step by letting $\gamma \uparrow 1$ under ergodicity.