Attention as Natural Gradient: In-Context Mirror Descent for Opponent Modelling

Alexander Chernyavskiy CHERNYAVSKIY.AS@ITMO.RU
Natalia Gusarova
Aleksandra Vatian
ITMO University, Russian Federation

Abstract

Multi-agent learning of agentic models faces a fundamental tension: agents must learn to adapt efficiently to their opponents at the test time. Recent work has shown that sequence models can learn to infer opponent strategies in context from the interaction history alone; however, the mechanism behind this behaviour remains poorly understood despite strong empirical evidence. In this position paper, we argue that the in-context learning of transformer-based multi-agent policies can be perceived as entropy-regularised mirror descent on the Fisher-Rao manifold of opponent strategies. Building on the findings of D'Angelo and Flammarion [6] providing a constructive proof that transformers implement mirror descent for the latent mixture models, we identify opponent types as the latent variables and interaction histories as the observed sequences where each attention layer can be interpreted as performing an implicit step of belief updating over opponent prototypes, with the SoftMax attention weights serving as updated mixture weights. The fixed points of this single-agent belief dynamics correspond to self-consistent predictions; at the joint level, these correspond to Predictive Equilibria [22] when all agents satisfy this condition. We hope that this position can suggest that standard self-supervised interaction sequence prediction on diverse opponent pools suffices for the induction of a theory-of-mind-like opponent reasoning, bridging the gap between agent modelling and equilibrium-based multi-agent learning. A toy experiment on the Iterated Prisoner's Dilemma supports these geometric predictions.

Keywords: In-context learning, opponent modelling, mirror descent, information geometry, multi-agent reinforcement learning, Predictive Equilibrium

1. Introduction

Foundational models have become overly dominant in deep learning applications and have been established as cornerstones of language modelling [5] applications and agentic intelligence [18]. These massive, intricately designed models achieve unprecedented performance across diverse tasks, beyond text generation. As these sequence-model-based agents are deployed to autonomously perform difficult tasks, they inevitably face multi-agent interactions, where outcomes depend on the interactions of multiple entities, the latter being other models or human entities. These interactions involve the completion of common goals [9], and ensuring that self-interested agents robustly cooperate in general-sum (mixed-motive, [7]) settings remains an important open challenge [23], even as individual agent capabilities have grown significantly.

On the other hand, mechanical interpretability has emerged [16, 21] as a field of algorithmic analysis of transformer networks, seeing them as a superposition of programs of different subnetworks that implement specific algorithmic functions, that is, circuits. Recently, D'Angelo and

© A. Chernyavskiy, N. Gusarova & A. Vatian.

Flammarion [6] showed that its architectural variation can implement a one-step Bayesian predictor through mirror descent which opens a connection between in-context learning and Bayesian predictive intelligence [9, 17].

The main contribution of this study is not an algorithm. It is a unifying theoretical perspective that connects the empirical phenomena of in-context learning, mechanical interpretability, and paradigms of embedded intelligence through the information geometry of the opponents' strategy space.

2. Challenges of multi-agent modelling

Consider an agent interacting with an opponent over a sequence of rounds.

Let $x_{\leq t} = (a_1^{\text{self}}, a_1^{\text{opp}}, \ldots, a_t^{\text{self}}, a_t^{\text{opp}})$ denote the interaction history, and h_{t-1} be the most recent joint history. The opponent's strategy is unknown; we model it as a latent mixture of k memory-1 prototypes $\{\pi_z\}_{z=1}^k$, where each prototype defines a conditional distribution $P(a_t^{\text{opp}} \mid h_{t-1}, z)$ over opponent actions. The agent maintains a belief $\lambda \in \Delta^{k-1}$ over these prototypes, with λ_z representing the probability that the opponent is of type z.

This model matches the Mixture of Transition Distributions (MTD) model [6, 19], where opponent types play the role of latent mixture components and interaction histories play the role of observed tokens. Table below summarises this correspondence.

Latent Mixture Model (MTD)	**Opponent Modelling**	**Role in Transformer**
Mixture weights $\lambda_t \in \Delta^{k-1}$	Belief over opponent types	Attention weights
Prototype transition matrix π_z	Opponent strategy prototype π_z	Key vectors
Context sequence $y_{1:t}$	Interaction history $x_{\leq t}$	Input tokens
Latent component $z \in \{1, \ldots, k\}$	Opponent type	Attention head index
Per-step loss $\ell(\lambda; y_s, y_{s-1})$	Negative log-likelihood of observed action	Query–key dot product
Posterior mean $\hat{\lambda}_t$	Updated belief after t rounds	Final-layer SoftMax output

Table 1: Mapping between latent mixture model concepts and opponent modelling concepts.

A significant challenge when analysing multi-agent systems is the reliance on the stationarity [12] of the dynamics from the perspective of each agent and separation of the agents from the world it acts upon [14]. There are two major issues with applying existing multi-agent concepts to agentic systems. First, how they handle the concept of irrationality: traditional game-theoretic methods ([1], Chapter 5) require some form of the best-response, optimal mediator, or the notion of subgame-perfect reasoning. These are hefty constraints that require fixed opponent strategies or a slowly updated static distribution. Moreover, these methods require exponential computation in the number of agents or game length scalable and explicitly assume that other agents act rationally which is a strong assumption for some agentic interactions. Model-free deep Reinforcement Learning (deep MARL, [1], Chapter 9) methods, in their turn, are designed to work with large models and can handle non-stationarity through, for example, exponentially weighted moving average of the agents' weights. However, they have convergence guarantees for only a very narrow choice of scenarios. In addition, all the above methods treat agents as decoupled entities that are not aligned with the tasks they execute. These problems are particularly challenging for LLM-based systems: non-stationarity induced by independent (decentralised) learning demands additional prediction and adaptation to

other changing entities, which is unfeasible in practice. Thus, one proposed solution to these issues is the framework of embedded predictive intelligence (MUPI, see [14] and Appendix C.1), which embeds the agent into the environment: each agent predicts its own future perceptions and actions as a part of the world they inhabit, forming a prospective predictive model of themselves as part of the shared execution setting.

MUPI agents aim to maximise the expected return while predicting their *own* actions:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \;\middle|\; \pi\right] \quad \text{s.t.} \quad a_t \sim \pi(\cdot \mid x_{\leq t}) \tag{1}$$

However, what distinguishes the framework from traditional RL objectives is that the agents update their own world models[1], reflecting their world view for each time step t:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{s\sim\mathcal{D}} \log p_\theta(x_{t+1}, a_{t+1} \mid x_{\leq t}, a_{\leq t}) \tag{2}$$

where x_t is the current environment history state, a_t is the self-action, and $\mathcal{D}$ is a dataset of collected environment interactions (histories).

A joint policy profile $(\pi_1^*, \ldots, \pi_n^*)$ is called a Subjective Embedded Equilibrium (SEE) if for each agent i it follows Equation 1, $\pi_i^* \in \arg\max_{\pi_i} \mathbb{E}_{\tau\sim p^*_{\theta_i}} \left[\sum_t \gamma^t r_t^i \mid \pi_i, \hat{\pi}^*_{-i}\right]$, with the other agents acting optimally.

The agent's policy is optimal with respect to its internal world model $p_{\theta_i^*}$. Additionally, equilibrium requires the agents' world models to remain path-consistent with respect to the data distributions $\mathbb{P}(x_t, \theta)\ \forall t$ on the equilibrium path:

$$p_{\theta_i^*}(x_{t+1}, a_{t+1}^{-i} \mid x_{\leq t}, a_{\leq t}) = \mathbb{P}(x_{t+1}, a_{t+1}^{-i} \mid x_{\leq t}, a_{\leq t}; \hat{\pi}^*_{-i}). \tag{3}$$

This objective can be viewed as a closed-loop dependency between the model and the data: the agent implements the policy improvement operator that creates an empirical distribution over joint interaction histories which, in turn, is used to update the agent's models. It is noteworthy that SEE does not require π^*_{-i} to be a best response to π_i^*. It only requires that each agent's model is self-consistent and its own policy is optimal given that model, or the agent to have mutually consistent world views.

For practical purposes, [22] defined the concept of predictive equilibrium, which approximates the SEE for practical implementation. A Predictive equilibrium (PE) is a configuration of model parameters θ^{i*} such that:

$$\theta^{i*} = \arg\min_{\theta^i} \mathrm{KL}\Big(\mathbb{P}(x_t^i; \theta^*) \;\Big\|\; p_{\theta^i}(x_t^i)\Big) \quad \forall i \in \mathcal{I}, \tag{4}$$

where $\mathbb{P}(x_t^i; \theta)$ is the true distribution of trajectories (histories)induced by all agents playing according to π_θ; $p_{\theta^i}(x_t^i)$ is the predicted distribution by agent i's internal world model. In other words, when the model is perfect on the equilibrium path ([22], Appendix D). For the differences between PE and SEE, see Appendix C.4 and Table 3.

The second issue is that, owing to the enormous expenses [9, 11] of foundational training, most interactions occur in-context. Theoretical explanations for in-context learning exist primarily for

1. Which technically implements Solomonoff's induction https://en.wikipedia.org/wiki/Solomonoff%27s_theory_of_inductive_inference

transformer architectures [23]. The empirical success of LLMs has outpaced the theoretical foundations that explain and describe their behaviour. Notably, none of the above-mentioned multi-agent methods tackle how fixed-weight models update their beliefs about other agents from streaming history, or what equilibrium concept corresponds to the fixed point of this dynamics, if any exists at all. D'Angelo and Flammarion [6] demonstrates that a three-layer transformer provably implements a Bayesian predictor mirror descent for the token generation task. In brief, the method can be described as follows.

Theorem 1 (D'Angelo and Flammarion [6] predictor (informal).) *Let $\pi \in \Delta^{k\times k}$ be a known transition matrix, and let the latent mixture weights satisfy $\lambda \sim \mathrm{Dirichlet}(\alpha\mathbf{1}) \in \Delta^{k-1}$. There exists a three-layer disentangled transformer $\mathcal{T}$ [8] such that, given any context sequence $y_{1:t}$ generated from the Mixture of Transition Distributions (MTD, see [19] and Appendix A) model, the transformer's output at the final token exactly computes the updated belief*

$$\hat{\lambda}_t = \arg\min_{\lambda\in\Delta^{k-1}} \mathrm{KL}\big(\lambda \,\|\, \lambda_0\big) + \eta \sum_{s=1}^{t} \ell(\lambda; y_s, y_{s-1}, \pi), \tag{5}$$

where ℓ is the per-step negative log-likelihood loss of the MTD model and λ_0 is a suitable initialisation (e.g. the prior mean). Thus, $\mathcal{T}$ implements one step of Mirror Descent on the mixture weights λ in context. Moreover, $\hat{\lambda}_t$ corresponds to a first-order online approximation of the Bayes-optimal posterior mean under Dirichlet prior. That is, the attention weights over the k prototype positions implement one step of entropy-regularised mirror descent.

To make the further explanation simpler, we consider them by example of the infamous Iterated Prisoner's Dilemma (IPD).The game consists of T independently played rounds. During each round, the two playing agents can output two possible actions: cooperate (C) or defect (D). As such, the environment consists of five possible observations: the initial state s_0 and four interaction-induced observations based on committed actions: (C, C), (C, D), (D, C), (D, D). The state s_t comprises all the past observations $h_{\leq t}$. The one-round reward interaction matrix is presented in Table 2.

A_1/A_2	C	D
C	(3, 3)	(0, 5)
D	(5, 0)	(1, 1)

Table 2: One-round IPD payoff matrix

Traditional game-theoretic approaches require stationary opponents and exact best-response computation, which is infeasible in complex agentic settings. Recent work on Embedded Predictive Intelligence [14] and Predictive Equilibrium (PE) [22] relaxes this by asking only that each agent's internal model be *self-consistent*: the predicted distribution of histories matches the true distribution induced by the opponent (see Appendix C.1 for formal definitions). We ask: **how can a fixed-weight transformer implement this belief update in-context?**

3. Position: Geometry as a missing piece

The MUPI framework and its practical implementation, PPI [22], have issues. They are currently formulated as being more descriptive than explanatory: they tell how close the agent is to a local or global predictive equilibrium. However, there is a lack of information on how the agent reaches equilibrium. The inference stage remains the "black-box": the sequence model can learn to extract the opponents' types from history during training, but it remains unspecified how the algorithm

implements adaptation to the user or to other agents, which is arguably one of the most essential components of the test-time generalisation.

It is worth noticing that the space of k opponents' strategies is not Euclidean by definition, and it is the probability simplex Δ^{k-1} which is governed by Riemannian geometry. Interpretability through information geometry or optimal transport is essential. For probability distributions, the Fisher–Rao metric endows the strategy manifold with the following properties [2, 10, 15]: (a) distances between the strategies measure not the numerical parameter similarity but their statistical distinguishability; (b) its gradients are natural gradients, reflecting how much a change in beliefs changes the likelihood of observed data; and (c) the dynamics of the manifold are described by geodesic flows, the curvature of which determines convergence rates and (in)stability boundaries.

By performing in-context Mirror Descent on the simplex $\mathcal{M} = \Delta^{k-1}$ equipped with the Fisher-Rao metric, the transformer dynamically updates its beliefs over opponent strategies. Because the mirror descent is a natural gradient descent on the Fisher manifold [15], its fixed points correspond to self-consistent predictions. Therefore, the Fisher-Rao metric serves as an interpretable connection between the two perspectives.

Theorem 2 (Predictive nature of the Transformer) *Let $\mathcal{M} = \Delta^{k-1}$ be a probability simplex equipped with the Fisher-Rao metric. Let $\{\pi_z\}_{z=1}^{k}$ be memory-1 opponent prototypes and $\lambda \in \mathcal{M}$ denote the agent's belief. Given history $x_{\leq t}$, define the negative log-likelihood*

$$\mathcal{L}(\lambda; x_{\leq t}) = -\sum_{s=1}^{t} \log\Big(\sum_{z=1}^{k} \lambda_z \cdot P(x_s \mid x_{s-1}, z)\Big). \tag{6}$$

There exists a depth-L disentangled transformer T_L with relative positional encodings such that, when queries encode $\theta = \nabla\psi(\lambda) - \eta\nabla\mathcal{L}$ and keys are prototype basis vectors, the final-layer attention weights satisfy

$$\mathcal{A}_{t,\cdot}^{(L)} = \nabla\psi^{-1}\big(\nabla\psi(\lambda^{(0)}) - \tau\nabla\mathcal{L}(\lambda^{(0)})\big) + O\big(\tau^2/L + \epsilon_{\text{approx}}\big),$$

where $\tau = L\eta$ is the effective horizon, and ϵ_{approx} is the finite-dimensional embedding error.

Proof See Appendix D. ■

Remark 3 *The fixed point of this dynamics, $\nabla\mathcal{L}(\lambda^*) = 0$, corresponds to the agent's predicted history distribution matching the empirical distribution, that is, the Predictive Equilibrium condition (Appendix D).*

Conjecture 4 (Multi-agent convergence) *In multi-agent settings, mutual forward passes induce* coupled *belief dynamics that converge to the Predictive Equilibrium when the Fisher curvature at the prior is positive definite. Characterising this convergence is left for future work.*

4. Thought experiment

To illustrate the paper's findings, let us perform a small thought experiment based on Axelrod's IPD tournament analysis [3]. Consider two PPI agents in IPD (Table 2) where one agent (Alice) must infer whether her opponent is TFT[2] or Grim[3]. After several rounds of mutual cooperation,

2. Tit-for-tat: starts by cooperating and then mimics the opponent's previous action
3. The agent cooperates until the opponent defects even once. Then, defect forever.

Alice's belief sits near the TFT optimum of the Fisher manifold because her training prior was leaning towards mutual cooperation. However, a single defection test produces divergent responses: if Bob plays TFT, the high curvature near the TFT prototype drives rapid belief shift (TFT has sharp conditional probabilities, making likelihood gradients strong) and convergence to mutual cooperation, which is a PE; if Bob plays Grim, the likelihood landscape is flat (Grim can be confused with AlwaysCooperate until defection occurs, so early observations are uninformative) which delays its recognition, and Alice remains out of equilibrium. Notably, if Alice is trained on a heavy defection prior, it reverses the geometry: Grim becomes easier to recognise. This example illustrates that equilibrium selection is not determined by payoffs alone but by the geometry of the training distribution. This geometric intuition is empirically supported in Appendix E. A 3-layer attention-only transformer, trained solely via next-action prediction on a diverse opponent pool, learns prototype keys that cluster into distinct strategy types (PCA analysis) and produces belief trajectories whose Fisher-Rao geodesic ratio is ≈ 1.1, significantly closer to the natural-gradient flow than an explicit single-step mirror-descent baseline (ratio up to ≈ 2.0). This supports the view that depth-L attention composes L implicit mirror-descent steps, and that the resulting in-context belief updating enables non-trivial best-response behaviour.

5. Discussion

In this study, we propose a perspective that unifies test-time adaptation and mechanistic interpretability. We support this claim by aligning with the vision of D'Angelo et al. [6] with the stationary opponent adaptation through information geometry and providing a game-theoretic foundation for what in-context learning may converge to. Recent empirical studies support this interpretation. ShapeLLM [20] demonstrated that LLM agents can mutually shape each other through in-context interactions, consistent with coupled mirror descent dynamics.

6. Limitations and Future work

However, it is important to note that the proposed perspective has substantial limitations.

- The Bayesian transformer predictor we base the framework on assumes a finite, known set of opponent prototypes with the full observability, which limits its practical applications. Practical examples of opponents may lie outside this set of data. Moreover, they can be non-stationary. Characterising the Fisher geometry of such behaviour remains an open question.
- We analyse one agent's belief dynamics against a fixed but diverse set of opponents. A case of mutually aware learning is far more complex [13]. The joint dynamics may not converge to a unique PE curve. The curvature of the manifold may become negative, caused by chaotic and adversarial behaviour. See, for example, Meulemans et al. [13], for a more comprehensive analysis.

Overall, this work might have implications on the interpretability of the training distribution, suggests architectural design changes for multi-agent systems, and shows how the system's behaviour and robustness can be regulated by controlling the opponent pool.

Acknowledgement. This work supported by the Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002), agreement №139-15-2025-010

References

[1] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. URL https://www.marl-book.com.

[2] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746.

[3] Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24 (1):3–25, 1980.

[4] Robert Axelrod et al. The evolution of strategies in the iterated prisoner's dilemma. *The dynamics of norms*, 1(1), 1987.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Francesco D'Angelo and Nicolas Flammarion. Transformers learn latent mixture models in-context via mirror descent. *arXiv preprint arXiv:2604.10848*, 2026.

[7] Yali Du, Joel Z Leibo, Usman Islam, Richard Willis, and Peter Sunehag. A review of cooperation in multi-agent learning. *arXiv preprint arXiv:2312.05162*, 2023.

[8] Dan Friedman, Alexander Wettig, and Danqi Chen. Learning transformer programs. *Advances in Neural Information Processing Systems*, 36:49044–49067, 2023.

[9] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.

[10] Marc Harper. Information geometry and evolutionary game theory. *arXiv preprint arXiv:0911.1383*, 2009.

[11] Nikhil Kandpal and Colin Raffel. Position: The most expensive part of an llm should be its training data. *arXiv preprint arXiv:2504.12427*, 2025.

[12] Dong Ki Kim. *Effective Learning in Non-Stationary Multiagent Environments*. Massachusetts Institute of Technology, 2023.

[13] Alexander Meulemans, Seijin Kobayashi, Johannes von Oswald, Nino Scherrer, Eric Elmoznino, Blake A Richards, Guillaume Lajoie, Blaise Aguera y Arcas, and Joao Sacramento. Multi-agent cooperation through learning-aware policy gradients. In *International Conference on Learning Representations*, volume 2025, pages 45978–46009, 2025.

[14] Alexander Meulemans, Rajai Nasser, Maciej Wołczyk, Marissa A Weis, Seijin Kobayashi, Blake Richards, Guillaume Lajoie, Angelika Steger, Marcus Hutter, James Manyika, et al. Embedded universal predictive intelligence: a coherent framework for multi-agent learning. *arXiv preprint arXiv:2511.22226*, 2025.

[15] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017.

[16] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[17] Theodore Papamarkou, Pierre Alquier, Matthias Bauer, Wray Buntine, Andrew Davison, Gintare Karolina Dziugaite, Maurizio Filippone, Andrew YK Foong, Vincent Fortuin, Dimitris Fouskakis, et al. Position: agentic ai orchestration should be bayes-consistent. *arXiv preprint arXiv:2605.00742*, 2026.

[18] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[19] Adrian E Raftery. A model for high-order markov chains. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 47(3):528–539, 1985.

[20] Marta Emili Garcia Segura, Stephen Hailes, and Mirco Musolesi. Opponent shaping in llm agents. *arXiv preprint arXiv:2510.08255*, 2025.

[21] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[22] Marissa A Weis, Rajai Nasser, Rif A Saurous, JoĂŁo Sacramento, Alexander Meulemans, et al. Multi-agent cooperation through in-context co-player inference. *arXiv preprint arXiv:2602.16301*, 2026.

[23] Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14365–14378, 2024.

Appendix A. Preliminaries

A.1. Mixture of Transition Distributions (MTD) Model

MTD model [6] of order m assumes that the transition probability $P(y_t \mid y_{t-1}, \ldots, y_{t-m})$ can be expressed as a mixture of k lower-order Markov chains:

$$P(y_t \mid y_{t-1}, \ldots, y_{t-m}) = \sum_{z=1}^{k} \lambda_z P_z(y_t \mid y_{t-1}),$$

where $\lambda \in \Delta^{k-1}$ are the mixture weights and P_z are the transition matrices of the k prototypes.

A.2. Mirror Descent

Mirror descent is an optimisation algorithm for constrained problems, defined by

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \eta \langle \nabla f(x_t), x \rangle + D_\psi(x, x_t),$$

where:

- $\mathcal{X}$ is the constraint set (e.g., simplex Δ^{k-1}),
- $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$ is the Bregman divergence,
- ψ is the mirror map (e.g., $\psi(\lambda) = \sum_j \lambda_j \log \lambda_j$ for the simplex).

For $\mathcal{X} = \Delta^{k-1}$ and $\psi(\lambda) = \sum_j \lambda_j \log \lambda_j$, mirror descent reduces to

$$\lambda_{t+1,j} = \frac{\lambda_{t,j} \exp(-\eta \nabla_j f)}{\sum_{j'} \lambda_{t,j'} \exp(-\eta \nabla_{j'} f)}.$$

This coincides with the SoftMax update rule, which links the mirror descent to the attention mechanisms.

Appendix B. Mirror descent update rule

Let the loss for opponent modelling be

$$\mathcal{L}(\lambda; x_{\leq t}) = -\sum_{s=1}^{t} \log \left(\sum_{z=1}^{k} \lambda_z P(x_s | x_{s-1}, z) \right),$$

where $\lambda \in \Delta^{k-1}$ are the mixture weights over the opponent prototypes $\{\pi_z\}_{z=1}^{k}$.

The gradient of $\mathcal{L}$ with respect to λ is

$$\nabla_\lambda \mathcal{L} = -\sum_{s=1}^{t} \frac{P(x_s | x_{s-1}, z)}{\sum_{z'} \lambda_{z'} P(x_s | x_{s-1}, z')}. \tag{7}$$

For the mirror descent with mirror map $\psi(\lambda) = \sum_j \lambda_j \log \lambda_j$ (negative entropy), the update is:

$$\lambda_{t+1} = \arg\min_{\lambda \in \Delta^{k-1}} \eta \langle \nabla_\lambda \mathcal{L}(\lambda_t), \lambda_t \rangle + D_\psi(\lambda, \lambda_t), \tag{8}$$

where D_ψ is the Bregman divergence

$$D_\psi = \sum_j \lambda_j \log \frac{\lambda_j}{\lambda_{t,j}} + \lambda_{t,j} - \lambda_j.$$

The optimisation problem 8 has the following closed-form solution:

$$\lambda_{t+1} = \lambda_t \exp\left(-\eta \nabla_\lambda \mathcal{L}(\lambda_t)\right)/Z, \tag{9}$$

where Z is the normalisation constant. Notably, it corresponds to the SoftMax attention

- the queries encode $\nabla\psi(\lambda) - \eta\nabla\mathcal{L}$ (dual coordinates)
- the keys are prototype basis vectors
- the attention weights $\mathcal{A}_t^{(L)}$ are the updated λ_{t+1}.

Remark 5 *The approximation error $O(\tau^2/L + \varepsilon_{\text{approx}})$ arises from the following:*

1. *Finite-depth L (shallow transformers approximate mirror descent less precisely)*
2. *Finite-dimensional embeddings and the context lengths*

Appendix C. Embedded equilibria framework

C.1. MUPI and PPI frameworks

This appendix provides a brief overview of the **Embedded Universal Predictive Intelligence (MUPI)** [14] framework and its practical implementation, **Predictive Equilibrium (PE)/PPI** [22], which are central to the multi-agent learning paradigms discussed in this paper.

C.2. Embedded Universal Predictive Intelligence (MUPI)

The **MUPI framework** [14] addresses the core challenge of multi-agent learning: *how can agents learn to act optimally in environments where other agents are also learning and adapting?* Traditional game-theoretic approaches (e.g. Nash equilibrium) assumes **stationary opponents** and **rationality**, which are unrealistic in many real-world scenarios. Instead, MUPI embeds the agent into the environment, treating its actions and perceptions as part of the world it models.

KEY IDEAS

- **Embedded Agents**: Agents are not decoupled from the environment. Instead, they model themselves *as part of the world* they inhabit, forming a **closed-loop system** in which actions, perceptions, and model updates are intertwined.
- **Subjective World Models**: Each agent maintains a **subjective world model** p_θ that predicts future perceptions and actions based on its own policy and the interaction history.
- **Self-Consistency**: The agent's world model must be **self-consistent** with the data it observes. This means that the model's predictions should match the actual distribution of the trajectories induced by the agent's policy and environment.

FORMAL DEFINITION

In this framework, an agent's objective is to maximise its expected return while simultaneously predicting its future actions and perceptions. Formally, for agent i,

1. **Policy Optimisation**: The agent optimises its policy π^i to maximise the expected discounted return:
$$\pi_i^* \in \arg\max_{\pi_i} \mathbb{E}_{\tau \sim p^*_{\theta_i}} \left[\sum_t \gamma^t r_t^i \;\middle|\; \pi_i, \hat{\pi}^*_{-i} \right], \tag{10}$$

2. **Model Learning**: The agent updates its world model to predict its future actions and perceptions:
$$\theta^{*i} = \arg\max_{\theta^i} \mathbb{E}_{s \sim \mathcal{D}} \left[\log p_{\theta^i}(x_{t+1}, a^i_{t+1} \mid x_{\leq t}, a^i_{\leq t}) \right],$$
where:
   - $\mathcal{D}$ is a dataset of collected interaction histories,
   - x_t is the environment state at time t,
   - a^i_t is the agent's action at time t.

CHALLENGES

While MUPI provides a theoretically sound framework, it faces practical challenges

- **Intractability**: Solving for SEE requires solving a fixed-point equation over policies and models, which is computationally intractable for large systems.
- **Black-Box Inference**: The framework does not specify *how* agents update their beliefs or models during test time (i.e., the "inference stage" is a black box).

C.3. Predictive Equilibrium (PE) and PPI

To address the intractability of SEE, Weis et al. [22] introduced the **Predictive Equilibrium (PE)** as a practical approximation. PE replaces the path-consistency condition in SEE with a **Kullback-Leibler (KL) divergence minimisation** objective, making it computationally tractable.

PPI: PRACTICAL IMPLEMENTATION OF PE

PPI (Predictive Policy Inference) is an algorithmic framework for achieving PE in practice. PPI trains agents to:

1. **Predict their own actions and observations** (self-modeling).
2. **Act optimally** given these predictions.

This is achieved by jointly optimising the policy and world model using gradient-based methods.

KEY FEATURES OF PPI

- **Tractability**: Unlike SEE, PE can be optimized using standard gradient descent methods.
- **Decentralised Learning**: PPI enables **decentralized training**, where each agent learns its own model and policy without requiring centralized coordination.
- **Generalisation to unkown opponents**: PPI agents can adapt to new opponents at test time by using their learned world models to infer the opponent's strategy from interaction histories.

CONNECTION TO MUPI

PPI is a **practical implementation** of the MUPI framework. While MUPI provides the theoretical foundation (SEE), PPI provides the algorithmic tools for approximating it (PE). PE replaces the intractable path-consistency condition in SEE with a tractable KL divergence minimisation. In other words, at a PE, each agent's model is **perfect on the equilibrium path**: its predictions match the true distribution of trajectories generated by the joint policy.

C.4. Embedded equilibria classification

While SEE provides a theoretically grounded equilibrium concept for embedded agents, its intractability motivates the use of Predictive Equilibrium (PE), which approximates SEE by minimising the KL divergence between the true and predicted trajectory distributions (Equation 5). PE is computationally tractable and serves as a practical target for training multiagent systems. The comparison is outlined in Table 3.

Property	Subjective Embedded Eq. (SEE)	Predictive Eq. (PE)
Definition	Joint policy profile where each agent's policy is optimal given its internal world model, and the model is path-consistent with the data	Model parameters minimise KL divergence between true and predicted trajectory distributions
Formal Condition	Equations 10,3	Equation 4
Computational Tractability	Intractable for large systems (requires solving fixed-point equations over policies and models)	Tractable via gradient descent (only requires minimising KL divergence)
Equilibrium Guarantees	Strong (theoretical): Agents' models and policies are mutually consistent	Approximate: Converges to SEE as the model capacity
Practical Use	Theoretical framework for embedded agents.	Practical implementation (e.g., this work).

Table 3: Comparison of the equilibria types implemented by MUPI [14] (left) and PPI [22] frameworks (right).

Appendix D. Proof of Theorem 2

The proof establishes the following two claims:

(1) A single attention layer implements one step of entropy-regularised mirror descent on the probability simplex.

(2) A depth-L transformer composes L such steps with an effective learning rate $\bar{\tau} = L\eta$, with second-order error $O(\bar{\tau}^2/L)$.

D.1. Step 1: Mirror descent on the Fisher-Rao manifold

For the negative-entropy mirror map $\psi(\lambda) = \sum_j \lambda_j \log \lambda_j$, the mirror descent update on the simplex Δ^{k-1} is

$$\lambda_{t+1} = \arg\min_{\lambda \in \Delta^{k-1}} \eta\langle \nabla\mathcal{L}(\lambda_t), \lambda\rangle + D_\psi(\lambda, \lambda_t), \tag{11}$$

where D_ψ denotes the Bregman divergence. In dual coordinates $\theta = \nabla\psi(\lambda) = \log\lambda + \mathbf{1}$, the update becomes

$$\theta_{t+1} = \theta_t - \eta\nabla\mathcal{L}(\lambda_t). \tag{12}$$

Mapping back to primal coordinates via $\nabla\psi^{-1}(\theta) = \mathrm{softmax}(\theta)$:

$$\lambda_{t+1} = \mathrm{softmax}\big(\log\lambda_t - \eta\nabla\mathcal{L}(\lambda_t)\big). \tag{13}$$

This is equivalent to the natural gradient descent on the Fisher-Rao manifold, since the Fisher metric is $g_{ij}(\lambda) = \delta_{ij}/\lambda_i$ and the natural gradient is $\tilde{\nabla}\mathcal{L} = g^{-1}\nabla\mathcal{L} = \mathrm{diag}(\lambda)\nabla\mathcal{L}$.

D.2. Step 2: Attention implements one step of mirror descent

In a single attention head, the output at position t is a weighted sum over k prototype positions $z \in \{1, \dots, k\}$:

$$\mathcal{A}_{t,z} = \frac{\exp\big(\langle Q_t, K_z\rangle\big)}{\sum_{z'}\exp\big(\langle Q_t, K_{z'}\rangle\big)}. \tag{14}$$

Following the disentangled construction of Friedman et al. [8] and D'Angelo and Flammarion [6], we set the keys to prototype basis vectors $K_z = e_z$ (the zth standard basis vector). The query at position t encodes the dual coordinate of the current belief penalised by the gradient:

$$Q_t = \log\lambda_{t-1} - \eta\nabla\mathcal{L}(\lambda_{t-1}). \tag{15}$$

The dot product then yields

$$\langle Q_t, K_z\rangle = \log\lambda_{t-1,z} - \eta\nabla_z\mathcal{L}(\lambda_{t-1}). \tag{16}$$

Substituting into the softmax:

$$\mathcal{A}_{t,z} = \frac{\exp\big(\log\lambda_{t-1,z} - \eta\nabla_z\mathcal{L}\big)}{\sum_{z'}\exp\big(\log\lambda_{t-1,z'} - \eta\nabla_{z'}\mathcal{L}\big)} = \mathrm{softmax}\big(\log\lambda_{t-1} - \eta\nabla\mathcal{L}(\lambda_{t-1})\big)_z. \tag{17}$$

Thus, $\mathcal{A}_{t,\cdot} = \lambda_t$, that is, the attention weights exactly implement one step of mirror descent.

Remark on query encoding. In the construction of D'Angelo and Flammarion [6], Friedman et al. [8], the query vector Q_t is produced by a linear projection W_Q applied to a hidden state that accumulates sufficient statistics of the observed past. The disentangled architecture ensures that each head's query encodes the dual coordinate of the corresponding mixture component, whereas the residual stream carries the belief state across layers. We refer to [6, Appendix B] for the full construction.

D.3. Step 3: Multi-layer composition

Let $\lambda^{(\ell)}$ denote the belief after layer ℓ, with $\lambda^{(0)} = \lambda_0$ the prior. Each layer implements:

$$\theta^{(\ell)} = \theta^{(\ell-1)} - \eta \nabla \mathcal{L}(\lambda^{(\ell-1)}), \quad \lambda^{(\ell)} = \nabla \psi^{-1}(\theta^{(\ell)}). \tag{18}$$

After L layers, the dual coordinate is expressed as

$$\theta^{(L)} = \theta^{(0)} - \eta \sum_{\ell=0}^{L-1} \nabla \mathcal{L}(\lambda^{(\ell)}). \tag{19}$$

For small η, we Taylor-expand the gradient around the prior as follows:

$$\nabla \mathcal{L}(\lambda^{(\ell)}) = \nabla \mathcal{L}(\lambda^{(0)}) + \nabla^2 \mathcal{L}(\lambda^{(0)}) \big(\lambda^{(\ell)} - \lambda^{(0)}\big) + O(\|\lambda^{(\ell)} - \lambda^{(0)}\|^2). \tag{20}$$

Because $\lambda^{(\ell)} - \lambda^{(0)} = O(\ell\eta)$, the second-order term contributes $O(\ell\eta^2)$ per layer. Summing over L layers, we have

$$\sum_{\ell=0}^{L-1} \nabla \mathcal{L}(\lambda^{(\ell)}) = L \nabla \mathcal{L}(\lambda^{(0)}) + O\big(L^2\eta^2\big). \tag{21}$$

With $\bar{\tau} = L\eta$, the error is $O(L^2\eta^2) = O(\bar{\tau}^2)$. However, the dominant second-order error in the mirror descent update itself is $O(\eta^2)$ per layer, giving a total error $O(L\eta^2) = O(\bar{\tau}^2/L)$. This term is reported in Theorem 2.

Therefore:

$$\theta^{(L)} = \theta^{(0)} - \bar{\tau} \nabla \mathcal{L}(\lambda^{(0)}) + O(\bar{\tau}^2/L), \tag{22}$$

and mapping back to the primal coordinates:

$$\mathcal{A}_{t,\cdot}^{(L)} = \nabla \psi^{-1}\big(\nabla \psi(\lambda^{(0)}) - \bar{\tau} \nabla \mathcal{L}(\lambda^{(0)})\big) + O(\bar{\tau}^2/L). \tag{23}$$

D.4. Step 4: Error analysis

The total error $O(\bar{\tau}^2/L + \varepsilon_{\text{approx}})$ comprises

Composition error $O(\bar{\tau}^2/L)$. Arises from the second-order Taylor remainder in the multilayer mirror descent expansion. This error vanishes as $L \to \infty$ with a fixed $\bar{\tau}$ (i.e. $\eta \to 0$).

For the disentangled construction with k prototypes and context length $\geq t$, $\varepsilon_{\text{approx}}$ is dominated by the embedding dimensionality gap, that is, Quantifying this gap for specific opponent strategy spaces is left for future research.

D.5. Step 5: Connection to Predictive Equilibrium (informal)

Furthermore, we show that the fixed point of the mirror descent dynamics corresponds to the Predictive Equilibrium condition. This connection justifies the claim in the main text that the transformer's in-context belief updating converges to self-consistent predictions.

Mirror descent fixed point. At a fixed point λ^*, the mirror descent update satisfies $\nabla\mathcal{L}(\lambda^*) = 0$ (up to the simplex-tangent constraint). From the definition of the loss in Eq. (6):

$$\mathcal{L}(\lambda; x_{\leq t}) = -\sum_{s=1}^{t} \log\Big(\sum_{z=1}^{k} \lambda_z \cdot P(x_s \mid x_{s-1}, z)\Big), \tag{24}$$

The gradient condition $\nabla\mathcal{L}(\lambda^*) = 0$ implies

$$\sum_{s=1}^{t} \frac{P(x_s \mid x_{s-1}, z)}{\sum_{z'} \lambda^*_{z'} P(x_s \mid x_{s-1}, z')} = \sum_{s=1}^{t} \frac{P(x_s \mid x_{s-1}, z')}{\sum_{z'} \lambda^*_{z'} P(x_s \mid x_{s-1}, z')} \quad \forall z, z'. \tag{25}$$

This is the maximum likelihood condition: the empirical frequencies of observations weighted by prototype likelihoods are balanced.

Connection to Predictive Equilibrium. Recall the PE condition in Eq. (4):

$$\theta^{*i} = \arg\min_{\theta^i} \mathrm{KL}\big(\mathbb{P}(x^i_t; \theta^*) \parallel p_{\theta^i}(x^i_t)\big) \quad \forall i \in \mathcal{I}. \tag{26}$$

In the opponent-modelling setting, the agent's internal model $p_{\theta_i}(x^i_t)$ is factorised as

$$p_{\theta_i}(x^i_t) = \sum_{z=1}^{k} \lambda_z \cdot P(x^i_t \mid x^i_{t-1}, z), \tag{27}$$

where λ is the belief in opponent prototypes. The KL divergence is:

$$\mathrm{KL}\big(\mathbb{P} \parallel p_\lambda\big) = \mathbb{E}_{x\sim\mathbb{P}}\Big[\log \mathbb{P}(x) - \log \sum_z \lambda_z P(x \mid z)\Big]. \tag{28}$$

Minimising this KL with respect to λ is equivalent to maximising the log-likelihood $\mathbb{E}[\log \sum_z \lambda_z P(x \mid z)]$, which is precisely the negative of our loss $\mathcal{L}(\lambda)$. Therefore:

$$\arg\min_{\lambda} \mathrm{KL}(\mathbb{P} \parallel p_\lambda) = \arg\min_{\lambda} \mathcal{L}(\lambda) = \{\lambda^* : \nabla\mathcal{L}(\lambda^*) = 0\}. \tag{29}$$

Summary The fixed point of mirror descent on the simplex (the limit of the transformer's attention dynamics) satisfies the Predictive Equilibrium condition. The agent's belief λ^* makes the predicted trajectory distribution $p_{\theta_i}(x^i_t)$ match the empirical distribution $\mathbb{P}(x^i_t; \theta^*)$ induced by the fixed opponent. This is the sense that postulates the position "attention implements mirror descent, which is first-order natural gradient descent, approximating flow on the Fisher manifold that has fixed points at mixed Predictive Equilibria."

Appendix E. Toy Experiment: In-Context Mirror Descent for Opponent Modelling

We validated the claims of Sections 3 and 4 on a canonical Iterated Prisoner's Dilemma (IPD) [4] task. The implementation follows the *disentangled* transformer construction of D'Angelo and Flammarion [6] and Friedman et al. [8]: attention-only layers, no feed-forward subnetwork, and prototype keys that are structurally tied to the mixture components.

E.1. Problem Setup

We consider IPD (Table 2) with $T = 20$ rounds and payoff matrix

$$R = \begin{pmatrix} (3,3) & (0,5) \\ (5,0) & (1,1) \end{pmatrix}, \tag{30}$$

where rows correspond to the agent's action $\{C, D\}$ and columns to the opponent's action. The opponent is drawn from a mixture of $k = 4$ memory-1 prototypes $\{\pi_z\}_{z=1}^{4}$ with strategy parameters $\pi_z(h) = P(C \mid h)$ for $h \in \{CC, CD, DC, DD\}$:

- ALWAYSCOOPERATE: $[1, 1, 1, 1]$
- ALWAYSDEFECT: $[0, 0, 0, 0]$
- TITFORTAT: $[1, 0, 1, 0]$
- GRIM: $[1, 0, 0, 0]$.

During training, the opponent type was sampled uniformly for each episode.

E.2. Structured Prompt Format

The input is a structured token sequence of length L.

$$[\texttt{prototype}_1, \ldots, \texttt{prototype}_k, \texttt{SEP}, \texttt{SELF}, a_1^{\text{self}}, \texttt{OTHER}, a_1^{\text{opp}}, \texttt{SEP}, \ldots, \texttt{PREDICT}, \texttt{PAD}, \ldots].$$

Each token carries a structure tag $\tau \in \{$PROTOTYPE, SEPARATOR, SELF, OTHER, PREDICT, PAD$\}$. The PREDICT token is placed at the end of the observed history and is used to extract the belief over the opponent types. The overall sequence length after T rounds of interaction is 106.

E.3. Structure-Aware Causal Mask

The causal mask $M \in \{0, 1\}^{L \times L}$ enforces

- PROTOTYPE and SEPARATOR tokens attend to all positions;
- SELF and OTHER tokens attend to prototypes and prior history only;
- PREDICT attends *exclusively* to the k prototype positions.

This last constraint is essential: it forces the PREDICT position's attention weights to be a distribution $\lambda \in \Delta^{k-1}$ over opponent types, directly implementing the belief extraction posited in Theorem 2, and

E.4. Disentangled Transformer Architecture

Following D'Angelo and Flammarion [6], we use a 3-layer *attention-only* disentangled transformer. There are no feed-forward sublayers, so that the residual stream remains a linear function of the prototype keys and the attention mechanism can be interpreted as an exact mirror descent on the simplex.

Single attention head. Each head uses a single attention matrix $W_A \in \mathbb{R}^{d\times d}$ (no separate Q, K, V projections) and relative positional encodings r^A_{i-j}, r^V_{i-j} indexed by offset. The attention score is

$$\text{score}_{ij} = \frac{h_i^\top W_A h_j}{\sqrt{d}} + h_i^\top r^A_{i-j}, \tag{31}$$

and values are concatenated position embeddings $\text{val}_{ij} = [h_j;\, r^V_{i-j}]$. The head output is $\text{head}_i = \sum_j \text{softmax}(\text{score})_{ij}\, \text{val}_{ij}$. The layer output concatenates the residual stream with all the head outputs as follows: $h^{(\ell+1)} = [h^{(\ell)}; \text{head}_1; \ldots; \text{head}_H]$.

Prototype keys. In the D'Angelo–Flammarion construction, the keys are fixed prototype basis vectors. To allow the model to discover the opponent basis from data while preserving the disentangled structure, we implement the keys as $K_z = W_{\text{proto}}\, \pi_z$ where $\pi_z \in \mathbb{R}^4$ are *learned* prototype policy logits and $W_{\text{proto}} \in \mathbb{R}^{d\times 4}$ is the learned linear map. Crucially, the prototype logits are initialised from a standard Gaussian $\mathcal{N}(0, 0.1)$; the canonical strategies are *not* baked into the initialization. The query at the PREDICT position encodes the dual coordinate $\theta = \nabla\psi(\lambda) - \eta\nabla\mathcal{L}$.

E.5. Belief Extraction

Let $\text{Attn}^{(L)}_{\text{PREDICT},z}$ denote the attention score from the PREDICT token to prototype z in the final layer (averaged over heads). The belief is extracted as

$$\lambda = \text{softmax}\big(\beta \cdot \text{Attn}^{(L)}_{\text{PREDICT},1:k}\big), \qquad \lambda \in \Delta^{k-1}, \tag{32}$$

where $\beta \in \mathbb{R}$ is a learned temperature parameter initialised at 0.

E.6. Opponent Action Prediction

Each prototype z maintains a learned policy vector $p_z \in \mathbb{R}^4$ (one logit per history). The conditional probability of the opponent cooperating is

$$P(a^{\text{opp}} = C \mid h, z) = \sigma(p_{z,h}), \qquad P(a^{\text{opp}} = D \mid h, z) = 1 - \sigma(p_{z,h}), \tag{33}$$

with σ the sigmoid. Given belief λ, the mixture prediction is

$$\hat{P}(a^{\text{opp}} \mid h) = \sum_{z=1}^{k} \lambda_z\, P(a^{\text{opp}} \mid h, z). \tag{34}$$

E.7. Best-Response Policy

The agent's action is derived from the belief via a soft best response. For each self-action $a \in \{C, D\}$ we compute the expected payoff

$$\bar{R}(a) = \sum_{a^{\text{opp}}} \hat{P}(a^{\text{opp}})\, R(a, a^{\text{opp}})_1, \tag{35}$$

and sample from

$$P(a^{\text{self}} = a) \propto \exp\bigl(\bar{R}(a)/\tau\bigr), \tag{36}$$

with temperature $\tau = 0.5$ K. During training the same deterministic transformation is used for the self-action consistency loss.

E.8. Training: Predictive Policy Improvement (PPI)

We train with a replay buffer of recent trajectories and a mixed data-collection scheme: early training uses random policies; a curriculum linearly ramps to 100% model-generated episodes. At each optimisation step, we sample a batch of past trajectories, select a random time step t, and predict the next opponent action from history $x_{\leq t}$.

The loss function contains four terms:

$$\mathcal{L}_{\text{opp}} = -\mathbb{E}\bigl[\log \hat{P}(a_{t+1}^{\text{opp}} \mid x_{\leq t})\bigr], \tag{37}$$

$$\mathcal{L}_{\text{self}} = \mathbb{E}\bigl[\text{KL}\bigl(\pi_{t+1}^{\text{stored}} \,\|\, \pi_{t+1}^{\text{recomputed}}\bigr)\bigr], \tag{38}$$

$$\mathcal{L}_{\text{div}} = -\gamma_{\text{div}} \sum_{z<z'} \|\sigma(p_z) - \sigma(p_{z'})\|_1, \tag{39}$$

$$\mathcal{L}_{\text{usage}} = -\gamma_{\text{usage}}\, H\bigl(\mathbb{E}[\lambda]\bigr), \tag{40}$$

with $\gamma_{\text{div}} = 0.1$ and $\gamma_{\text{usage}} = 0.05$. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{opp}} + \tfrac{1}{2}\mathcal{L}_{\text{self}} + \mathcal{L}_{\text{div}} + \mathcal{L}_{\text{usage}}. \tag{41}$$

The diversity term prevents prototype collapse, and the "usage-entropy" term encourages the model to employ all prototypes across the dataset. The term $\mathcal{L}_{\text{self}}$ tended to work better than the hard cross-entropy loss; therefore, we used it for the preliminary experiments. Determining a more consistent loss structure compatible with PPI and the role of the regularisation terms is left for future studies.

No mirror-descent consistency loss. We do *not* add an auxiliary loss that explicitly penalises deviation from the mirror descent. We tested whether the architecture's inductive bias alone induced MD-like belief dynamics, as predicted by Theorem 2.

E.9. Mirror Descent Baseline

For comparison, we implemented an explicit online mirror descent on the simplex. At each step, given observation $(a_t^{\text{self}}, a_t^{\text{opp}})$ and history h_{t-1},

$$\ell_t(\lambda) = -\log\Bigl(\sum_{z=1}^{k} \lambda_z\, P(a_t^{\text{opp}} \mid h_{t-1}, z)\Bigr), \tag{42}$$

$$\lambda_{t+1} = \text{softmax}\bigl(\log \lambda_t - \eta \nabla \ell_t(\lambda_t)\bigr), \tag{43}$$

where the likelihoods $P(\cdot \mid h, z)$ are evaluated using the *learned* prototype policies (not ground-truth strategies). We grid-search $\eta \in \{0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10\}$ per episode and report the configuration with a minimum mean KL divergence to the transformer beliefs.

E.10. Evaluation Protocol

We train for 67 epochs with AdamW (learning rate 10^{-3}, cosine decay, weight decay 10^{-4}) and evaluate over $N_{\text{seeds}} = 3$ independent runs. For each seed, we tested all four ground-truth prototypes. As previously noted, η is defined by the effective interaction horizon, the definition of which is currently out of the paper's scope.

KL trajectory: $\mathrm{KL}(\lambda_t^{\text{model}} \parallel \lambda_t^{\text{MD}})$ per turn, averaged over the seeds.

Fisher-Rao geometry. Because the probability simplex Δ^{k-1} equipped with the Fisher-Rao metric is a Riemannian manifold, we can compare the transformer belief trajectory to the natural gradient flow predicted by Theorem 2. For two beliefs $p, q \in \Delta^{k-1}$, the Fisher-Rao distance is

$$d_{\mathrm{FR}}(p, q) = 2 \arccos\Big(\sum_{j=1}^{k} \sqrt{p_j q_j}\Big). \tag{44}$$

From this we compute:

- **FR distance to MD baseline**: mean $d_{\mathrm{FR}}(\lambda_t^{\text{model}}, \lambda_t^{\text{MD}})$ over the episode.
- **Path length**: $\sum_{t=0}^{T-1} d_{\mathrm{FR}}(\lambda_{t+1}, \lambda_t)$, the total distance travelled by the belief.
- **Straight-line distance**: $d_{\mathrm{FR}}(\lambda_0, \lambda_T)$, the direct geodesic from prior to final belief.
- **Geodesic ratio**: path length divided by straight-line distance. A ratio of 1 indicates that the trajectory lies exactly on a Fisher-Rao geodesic, and larger values indicate a deviation from the natural-gradient flow.

The geodesic ratio serves as a geometric diagnostic: Theorem 2 predicts that a depth-L transformer approximates the continuous natural-gradient flow with error $O(\bar{\tau}^2/L)$, so its ratio should be closer to 1 than a single-step discrete MD baseline.

Prototype recovery. After training, we matched the learned sigmoid-constrained policies $\sigma(p_z) \in [0, 1]^4$ to the ground-truth prototypes via the Hungarian algorithm on pairwise MSE and reported the minimum-cost MSE and maximum error.

All time-series are plotted as mean $\pm$ standard deviation across seeds; bar plots show mean $\pm$ standard deviation of cumulative reward.

E.11. Empirical Results

Optimisation landscape. Figure 1 shows the mean cumulative loss when training in a pool of diverse opponents. As we did not optimise the hyperparameters and the loss structure, the landscape was quite noisy and prone to overfitting. However, this demonstrates that the PPI training objective is functional.

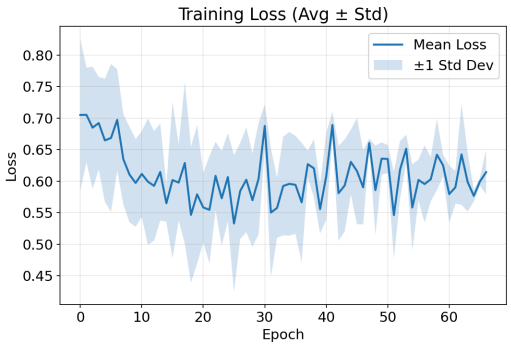

Figure 1: Training loss (mean $\pm$ std, 5 seeds) against each memory-1 prototype over 20 rounds of IPD.

As depicted in Figure 2, during the inference time, the KL divergence between model beliefs and their Mirror Descent approximations remains low and decreases during the inference stage which is in agreement with the theory. However, it is worth noting that there are some dynamical anomalies (for instance, sudden jumps for GRIM and ALWAYSCOOPERATE) that are left for future investigations.

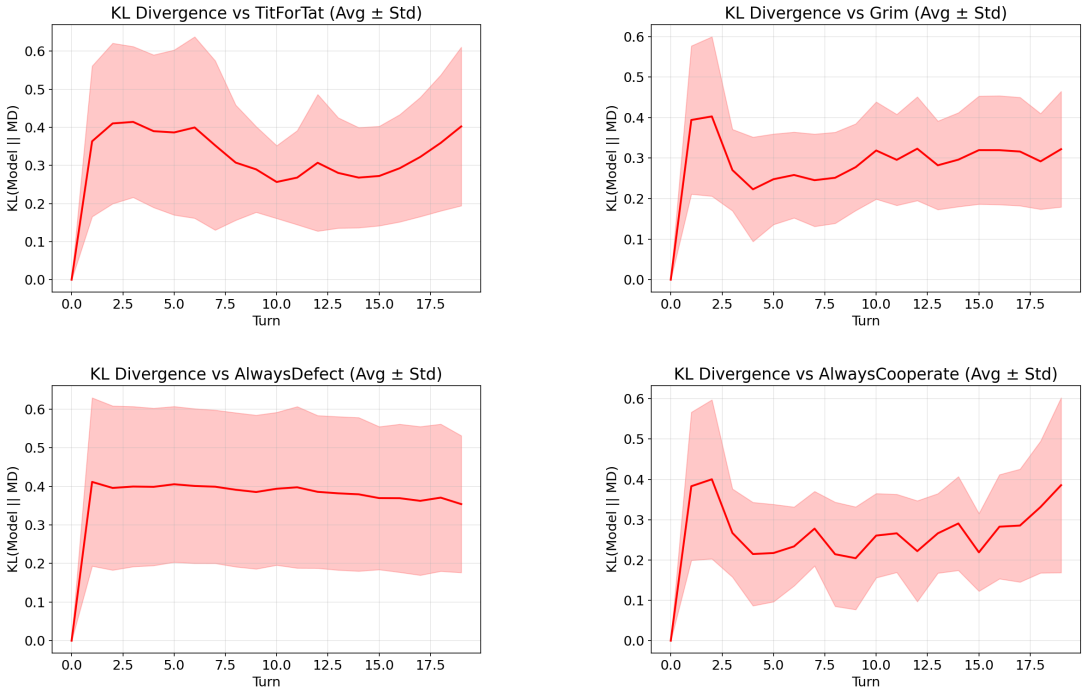

Figure 2: KL Divergence $\mathrm{KL}(\lambda_t^{\text{model}} \,\|\, \lambda_t^{\text{MD}})$ per turn of inference (mean $\pm$ std, 3 seeds) over 4 test prototypes. As it steadily decreases and stays relatively low 0.3 ± 0.15, the model approaches behaviour established by Theorem 2.

Reward structure. Figure 3 shows the mean cumulative reward against each opponent. Against ALWAYSCOOPERATE and TITFORTAT the agent achieves the highest returns (≈ 67 and ≈ 65), reflecting successful adaptation to cooperative or conditionally cooperative strategies. Returns against GRIM (≈ 53) and ALWAYSDEFECT (≈ 45) are lower but positive, indicating that the agent learns to avoid unconditional exploitation. The comparatively large variance against TITFORTAT (error bar $\approx \pm 11$) suggests that the agent occasionally fails to lock in mutual cooperation early in the episode.

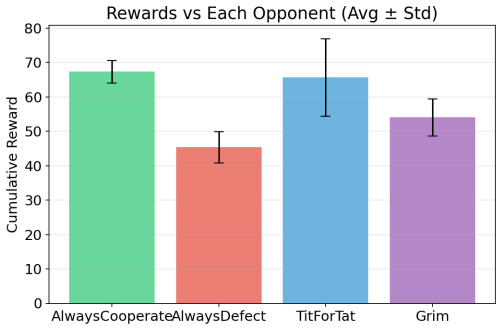

Figure 3: Cumulative reward (mean $\pm$ std, 3 seeds) against each memory-1 prototype after 20 rounds of IPD.

Learned prototype geometry. Figure 4 shows a PCA projection of the learned prototype keys (averaged over the heads in the final layer). The four prototypes form a well-separated structure in embedding space, with pairwise Euclidean distances ranging from 0.12 to 0.37. However, the learned policies do *not* coincide with the ground-truth canonical strategies: the prototype-recovery MSE is ≈ 0.22 (Figure 6), well above the 0.05 threshold, which would indicate exact recovery. This is expected because the model is free to discover any four basis strategies that span the space of opponent behaviours useful for prediction. The fact that the keys are distinct and the agent still achieves high reward confirms that the transformer has learned a *latent* basis of opponent types, not merely memorised the training distribution.

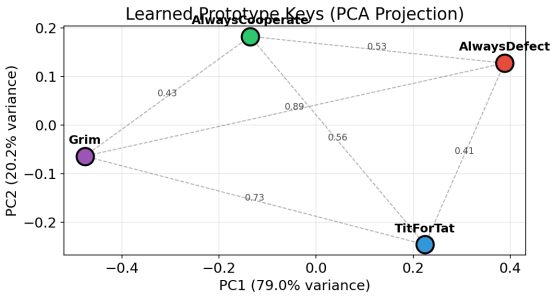

Figure 4: PCA projection of the learned prototype keys (final layer, averaged over heads). Pairwise distances are indicated on the dashed edges.

In-context belief dynamics. Figure 5 plots the belief trajectory when the true opponent is TITFORTAT. Because the learned prototypes do not map one-to-one onto the canonical strategies, the "correct" prototype is not uniquely identifiable; instead, the model distributes mass over the learned basis elements that best explain the observed behaviour. The belief stabilises within the first two turns and remains relatively flat thereafter, indicating that the model rapidly commits to an interpretable mixture and then updates it conservatively, a hallmark of entropy-regularised mirror descent with a small effective learning rate.

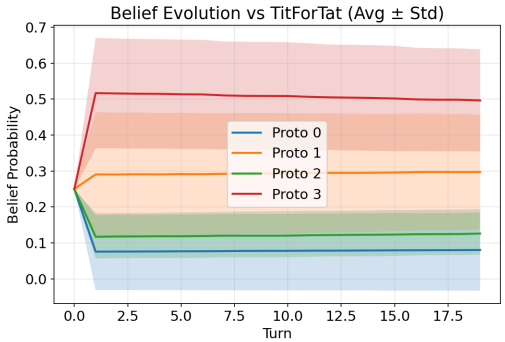

Figure 5: Belief evolution vs. TITFORTAT (mean $\pm$ std, 5 seeds). Because the prototypes are learned, the curves do not align one-to-one with the canonical strategy names; rather, they show how the model allocates mass across its discovered basis.

Fisher-Rao geometry and MD alignment. Figure 6 (left) reports the mean Fisher-Rao distance between the transformer beliefs, and the explicit mirror-descent baseline. The distance is moderate (≈ 0.45–0.60), which is unsurprising because the MD baseline uses the same learned prototypes but updates with a single fixed learning rate, whereas the transformer composes $L = 3$ layers with an implicit effective rate $\bar{\tau} = L\eta$. More striking is the geodesic ratio (middle panel): the transformer trajectory has a ratio ≈ 1.1 for all opponents, i.e. it stays almost exactly on the Fisher-Rao geodesic between the initial and final beliefs. By contrast, the explicit MD baseline deviates substantially from geodesic flow (ratio up to ≈ 2 against ALWAYSCOOPERATE). This provides strong empirical evidence that the depth-L composition of attention layers approximates the *continuous-time* natural-gradient flow on the simplex, as predicted by Theorem 2, better than a discrete single-step MD algorithm with identical prototypes.

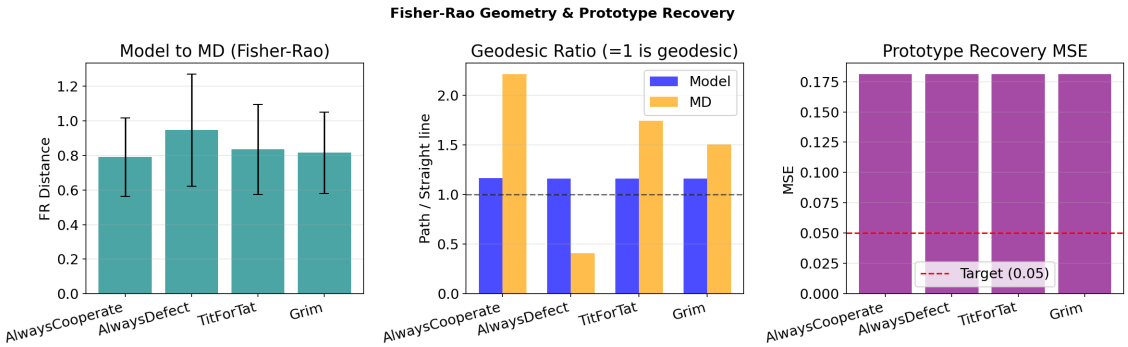

Figure 6: **Left:** Fisher-Rao distance between transformer beliefs and the explicit MD baseline (mean $\pm$ std, 3 seeds). **Middle:** Geodesic ratio (path length / straight-line distance) on the Fisher-Rao manifold; a ratio of 1 indicates perfect geodesic flow. **Right:** Prototype-recovery MSE relative to the ground-truth canonical strategies.

Mechanistic validation. Figure 7 displays the softmax attention weights (averaged over heads) from the final layer for an episode against TITFORTAT. The horizontal band at sequence position ≈ 100 corresponds to the PREDICT token; its weights are concentrated on a single prototype index, confirming that the PREDICT token's attention distribution implements a sharp belief λ_t over the learned opponent basis, exactly as posited in Theorem 2.

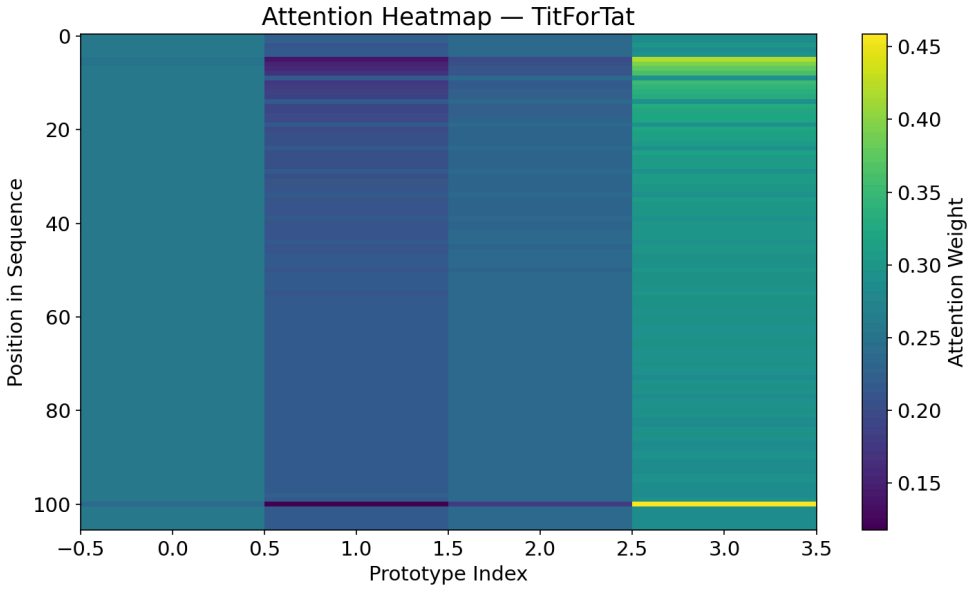

Figure 7: Final-layer attention weights (averaged over heads) for an episode against TITFORTAT. The PREDICT token (horizontal band at position ≈ 100) attends selectively to one learned prototype.

Stuctural scalability. Figure 8 demonstrates that changing number of layers improves geodesic ration as posited in Theorem 2.

Summary. The toy experiment supports three core claims: (i) the disentangled transformer discovers a latent basis of opponent strategies (keys are well-separated, MSE to ground truth is non-zero but the basis is functional); (ii) the attention weights at the PREDICT position evolve as a belief distribution over this learned basis, with trajectories that closely follow Fisher-Rao geodesics; and (iii) this in-context belief updating enables non-trivial best-response behaviour, yielding positive cumulative rewards against all four canonical IPD strategies. Nevertheless, there are many unanswered questions. For example, what would happen if we tested the model against out-of-distribution prototypes, and how robust are the learnable prototypes to different types of non-stationarities.

E.12. Scaling: Depth and Effective Horizon

Theorem 2 predicts that a depth-L transformer composes L mirror-descent steps with an effective horizon $\bar{\tau} = L\eta$, yielding a geodesic-flow approximation with an error $O(\bar{\tau}^2/L)$. To test this, we trained models at depths $L \in \{1, 3, 6\}$ with per-layer step size scaled inversely as $\eta = 0.1/L$, while keeping $\bar{\tau} = 0.1$ constant. Training used 33 epochs and a single attention head per layer, reflecting compute budget constraints. We report the results averaged across all opponents.

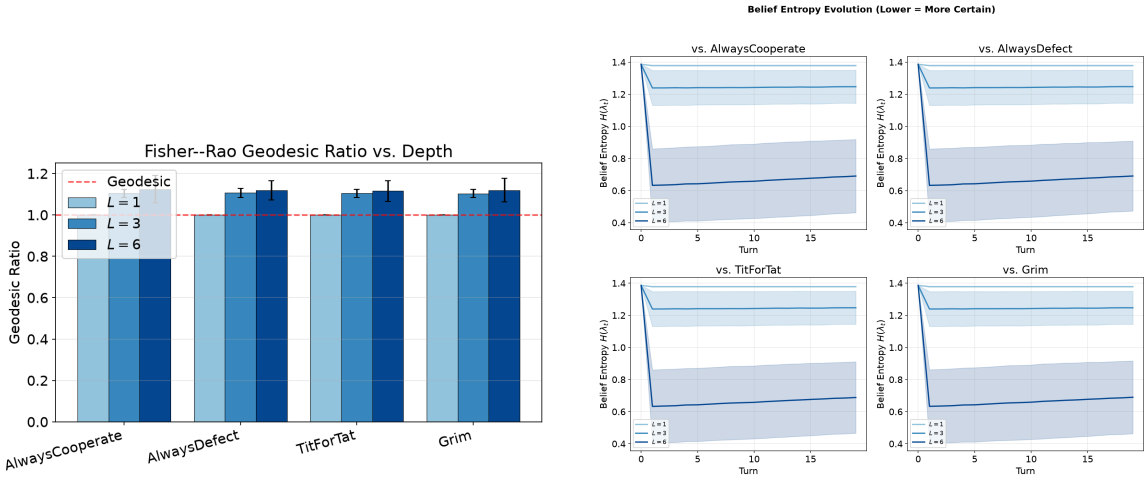

Figure 8: **Left:** Fisher-Rao geodesic ratio for $L \in \{1, 3, 6\}$ with $\eta = 0.1/L$. All depths remained near the geodesic (dashed line at 1.0). **Right:** Belief entropy evolution; deeper models converge to sharper beliefs faster.

Geodesic ratio. Figure 8 (left) reports the Fisher-Rao geodesic ratio. All depths achieve ratios near 1.0 (dashed line), with $L = 1$ closest to exact geodesic flow and $L = 3, 6$ showing mild deviation consistent with the $O(\bar{\tau}^2/L)$ second-order error. Notably, $L = 6$ does not degrade relative to $L = 3$, suggesting that the concatenative dimension growth in our architecture is offset by the smaller per-layer step size.

Belief dynamics. Figure 8 (right) plots belief entropy $H(\lambda_t)$ over time. All models start from maximum entropy (≈ 1.39 nats for $k = 4$) and drop sharply after the first observation, and then decreased. $L = 6$ achieves the lowest terminal entropy (≈ 0.65 for $L = 1$), indicating that deeper architectures extract more confident opponent identification from the same interaction history data.

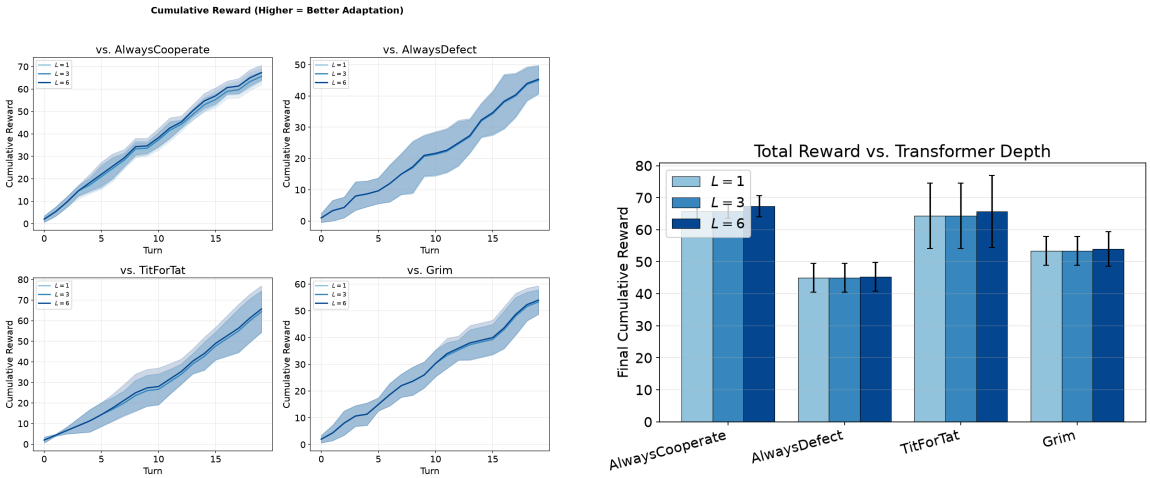

Figure 9: **Left:** Cumulative reward trajectories (mean $\pm$ std, 3 seeds). **Right:** Final cumulative reward by depth. Depth improves belief sharpness but not net reward, suggesting that $L = 1$ already suffices for best-response play in this toy setting.

Task performance. Figure 9 shows the cumulative reward trajectories and final totals. Reward curves are nearly identical across depths, with all models converging to the same final performance ($\approx$ 53–54 per opponent). The variance bands overlap almost completely, confirming that depth affects *belief geometry* without altering *task-level* adaptation.

Summary. With a fixed effective horizon $\bar{\tau}$, deeper transformers maintain near-geodesic belief dynamics while producing sharper beliefs. However, the task (20-round IPD against memory-1 opponents) is too easy for depth to translate into reward gains; even a single attention layer achieves near-optimal return.

Camera-Ready Edits Summary

Reviewer MaXE (Rating: 3, Clear Rejection)

1. *"The connection from in-context mirror descent to convergence toward Predictive Equilibrium is not justified."*
   Fix: Theorem 2 states only the single-agent mirror-descent/attention equivalence (proven in Appendix D). The multi-agent convergence claim is explicitly demoted to Conjecture 4, which is clearly labelled as conjectural.

2. *"The paper should more clearly distinguish between established results, proposed interpretations, and conjectures."*
   Fix: Three categories now typographically distinct: Theorem 2 (established, with proof), Remark 3 (proposed interpretation), Conjecture 4 (conjectural).

3. *"I would recommend presenting Theorem 2 as an informal conjecture or position statement unless a full proof is provided."*
   Fix: Full proof retained in Appendix D (Steps 1-5). Theorem 2 kept as theorem because proof is complete; only multi-agent extension demoted.

4. *"The notation is unnecessarily heavy and sometimes inconsistent."*
   Fix: Entire §2 replaced with concise opponent-as-mixture setup. The full POSG tuple was removed.

5. *"Symbols such as x_t, h_t, s_t, θ, and λ should be used more consistently."*
   Fix: $x_{\leq t}$ = interaction history, h_{t-1} = last joint history, λ = belief, θ = model parameters / dual coordinates. No collisions remain.

6. *"Apparent inconsistency between Δ^{k-1} and Δ^{q-1}."*
   Fix: Standardised to Δ^{k-1} everywhere.

7. *"Equation (5), where the minimisation variable is unclear."*
   Fix: PE definition now explicitly minimises over θ^i with clear subscript.

8. *"A table mapping latent mixture model concepts to opponent modelling concepts would make the argument much easier to follow."*
   Fix: Added Table 1 in §2: 6-row mapping (mixture weights $\rightarrow$ attention weights, prototypes $\rightarrow$ keys, etc.).

9. *"The writing also needs polishing. There are several typos and grammatical issues."*
   Fix: Fixed: "cotribution" $\rightarrow$ "contribution", "distributiosn" $\rightarrow$ "distributions", "independetly" $\rightarrow$ "independently", "equilibriua" $\rightarrow$ "equilibria", "AllwaysDefect/Cooperate" $\rightarrow$ "AlwaysDefect/Cooperate", "AllCoordinate" $\rightarrow$ "AlwaysCooperate", "unfeasible" $\rightarrow$ "infeasible", "nonstationary" $\rightarrow$ "non-stationary", "ansd" $\rightarrow$ "and", "Stuctural" $\rightarrow$ "Structural", "ration" $\rightarrow$ "ratio", "econputed" $\rightarrow$ "recomputed".

10. *"The paper would benefit from a simpler exposition."*
    Fix: §2 cut by $\sim$50% (POSG removed, MUPI/SEE details moved to Appendix C). §3 opens with intuitive geometry motivation before any theorem.

Reviewer LFHC (Rating: 4, Ok but not good enough — Rejection)

1. *"The paper does not discuss how far the interpretation survives under realistic architectural conditions."*
   Fix: Added explicit limitation in §6, bullet 1: finite prototype set, out-of-distribution opponents, non-stationarity.

2. *"The mutual coupled dynamics case is deferred to future work in the limitations section despite being the core motivating scenario."*
   Fix: Conjecture 4 explicitly addresses coupled dynamics with scope: convergence "when the Fisher curvature at the prior is positive definite". §6 bullet 2 elaborates on negative curvature and non-convergence.

3. *"Providing some experimental measurements to bridge the gap between the theoretical position and empirical plausibility would highly improve the paper's credibility."*
   Fix: Added empirical paragraph in §4: 3-layer transformer on IPD, Fisher–Rao geodesic ratio ≈ 1.1 vs. MD baseline ≈ 2.0, PCA distances 0.12–0.37. Full experiment in Appendix E (67 epochs, 3 seeds, 4 prototypes, depth scaling $L \in \{1, 3, 6\}$). Figures 1–9.

4. *"Page 2 typo — 'Main cotribution of the this work is not an algorithm'."*
   Fix: "The main contribution of this work is not an algorithm."

5. *"Page 4 typo — 'For probability distributiosn, Fisher-Rao metric'."*
   Fix: "For probability distributions, the Fisher-Rao metric…"

Additional Improvements (Not Explicitly Requested)

- **Completed abstract sentence.** Was cut off; now reads "…bridging the gap between agent modelling and equilibrium-based multi-agent learning. A toy experiment on the Iterated Prisoner's Dilemma supports these geometric predictions."

- **Added depth-scaling experiment** (Appendix E.12, Figures 8–9). Tests Theorem 2 prediction that error is $O(\bar{\tau}^2/L)$. Shows $L = 1, 3, 6$ maintains a geodesic ratio ≈ 1.0.