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

author names withheld

Under Review for NExT-Game 2026

Abstract

Multi-agent learning of agentic models faces a fundamental tension: agents must learn to efficiently adapt to their opponents at test time. Recent work has shown that sequence models can learn to infer opponent strategies in-context—from the interaction history alone—however, mechanism behind this behaviour stays poorly understood despite its 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 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 the updated mixture weights. The fixed points of this dynamics correspond to self-consistent embedded Predictive Equilibria [13, 22]. We hope that the 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 acting.

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

1. Introduction

Foundational models become overly dominant in deep learning applications and established as a cornerstones of language modelling [5] applications and agentic intelligence [17]. 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 interactions of multiple entities, the latter being other models or human entities. These interactions involve completion common goals [8], ensuring that self-interested agents robustly cooperate in general-sum (mixed-motive, [6]) settings remains an important open challenge [23], even as individual agent capabilities have grown significantly.

On the other hand, mechanical interpretability has emerged [15, 21] as a field of algorithmic analysis of the transformer networks, seeing them as a superposition of programmes of different subnetworks that implement specific algorithmic functions — circuits. Recently, D'Angelo and Flammarion [6] showed that its architectural variation can implement a one-step Bayesian predic-

tor through mirror descent which opens a connection between in-context learning and Bayesian predictive intelligence [8, 16].

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

2. Challenges of multi-agent modelling

We formalise the multi-agent interaction as a partially observable stochastic game (POSG, [20]). The game can be written as a tuple $\mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{P}_i, \mathcal{P}_t, \mathcal{P}_r, \mathcal{O}, \mathcal{P}_o, \gamma$, where $\mathcal{I}$ is a finite set of agents, $\mathcal{S}$ — a set of game states, $\mathcal{A}$ — a set of the agents' joint actions $\mathcal{A} = \times_{i\in\mathcal{I}}\mathcal{A}_i$. The game transitions to the next state by $P_t(S_{t+1}|S_t, A_t)$ from an initial state $S_0 \sim P_i(s_0)$. After committing an action, each agents receives a reward r_t^i from the joint factorised distribution $P_r = \times_{i\in\mathcal{I}}P_r^i(r_t^i|s_t, a_t)$ and an observation o_t^i from the observation space $\mathcal{O} = \times_{i\in\mathcal{I}}\mathcal{O}_i$ governed by the function $P_o = P_o(o_t^i|s_t, a_{t-1})$. Action of each agent i can be indexed, such $a \in \mathcal{A} = (a^i, a^{-i})$. Policy of each agent π^i is conditioned on the interaction history $x_{\leq t} = \{o_t^i, a_t^i, r_t^i\}_{t=1}^T$.

A big challenge when analysing multi-agent systems is reliance on the stationarity [11] of the dynamics from the point of each agent and separation the agents from the world it acts upon [13]. There are two major issues with applying existing multi-agent concepts to the agentic systems. First is 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. Those are hefty constraints that require fixed opponent-strategies or a very 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 only for a very narrow choice of scenarios. Besides that, all the above methods treat the agents as a decoupled entity that is not in any way aligned with a task it executes. 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 infeasible. Thus, one proposed solution of the issues is the framework of Embedded predictive intelligence (MUPI, [13]) that makes the agent embedded into the environment: each agent predicts its own future perceptions and actions as a part of the world they inhabit, forming prospective predictive model of themselves as a part of the shared execution setting.

MUPI agents aim to maximise 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 and a_t are self-actions 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:

$$\pi_i^* \in \arg\max_{\pi_i} \mathbb{E}_{\tau\sim p^*_{\theta_i}} \left[\sum_t \gamma^t r_t^i \,\middle|\, \pi_i, \hat{\pi}^*_{-i} \right]. \tag{3}$$

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

$$p_{\theta_i^*}(x_{t+1}, a^{-i}_{t+1} \mid x_{\leq t}, a_{\leq t}) = \mathbb{P}(x_{t+1}, a^{-i}_{t+1} \mid x_{\leq t}, a_{\leq t}; \hat{\pi}^*_{-i}). \tag{4}$$

This objective can be viewed as a close-loop dependency between the model and the data: the agent implements the policy improvement operator that creates empirical distribution over joint interaction histories which, in turn, is used to update the agents' models. It is noteworthy to observe 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 a concept of Predictive equilibrium that approximates SEE for practical implementation. A Predictive equilibrium is a configuration of model parameters θ^* such that:

$$\theta^* = \arg\min_{\theta^i} \mathrm{KL}\Big(\mathbb{P}(x_t^i; \theta^*) \,\Big\|\, p_{\theta_i^*}(x_t^i)\Big) \quad \forall i \in \mathcal{I}, \tag{5}$$

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).

The second issue is, due to enormous expenses [8, 10] of foundational training, most of the interaction happens in-context. Theoretical explanations of in-context learning exist primarily for transformer architectures [23]. Empirical success of LLMs has outpaced theoretical foundations that explain and describe their behaviour. Notably, none of the above mentioned multi-agent method tackle how a fixed-weight models updates their beliefs about other agents from streaming history, not 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. Simply put, the method can be informally described as follows.

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

Theorem 1 (D'Angelo and Flammarion [6] predictor (informal).) *Let $\pi \in \Delta^{q\times q}$ 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 $\mathcal{T}$ [7] such that, given any context sequence $y_{1:t}$ generated from the Mixture of Transition Distributions (MTD, [18]) model, the transformer's output at the final token exactly computes the updated belief*

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

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 the 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 infamous Iterated Prisoner's Dilemma (IPD).The game consists of T independetly played rounds. During each round, 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 4 interaction-induced observations based on committed actions: (C, C), (C, D), (D, C), (D, D). The state s_t is then comprised of all past observations $o_{\leq t}$. The one round reward interaction matrix is demonstrated in Table 1.

A_1/A_2	C	D
C	(1, 1)	(-1, 2)
D	(2, -1)	(0, 0)

Table 1: One-round IPD payoff matrix

In POSG (especially repeated games like IPD [3]), each player faces uncertainty about the opponent's strategy. We can really straightforwardly view the opponent's strategy as it is being drawn from a mixture of policy prototypes $\{\pi_j\}_{j=1}^{k}$: $\pi^i \sim \sum_k \lambda_k \cdot \pi_k$. This is exactly analogous to the MTD setup, where the opponent prototypes $\pi_j,\ j = \{1, ..., k\}$, are different possible types of opponents (AllwaysDefect, AllwaysDefect, TFT, Pavlov, extortionate strategies, for more see [3, 4]) and λ is the player's belief over which type each opponent is.

Therefore, inferring λ in-context in MTD is equivalent to in-context opponent modeling in games. This actually makes the disentangled transformer a powerful tool that can give a mechanistic explanation for how sequence models implement the belief updating required for multi-agent learning and equilibrium selection.

3. Position: Geometry as a missing piece

MUPI framework and its practical implementation, PPI [22], have an issue. They currently formulated, that they are rather descriptive than explanatory: they tell how close the agent to a local or global predictive equilibria. However, there is a lack of information of 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 really stays 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. This is where

interpretability through information geometry or optimal transport becomes essential. For probability distributiosn, Fisher-Rao metric endows the strategy manifold with the following properties [2, 9, 14]: (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; (c) dynamics of the manifold is described by geodesic flows, 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 mirror descent is natural gradient descent on the Fisher manifold [14], its fixed points correspond to self-consistent predictions. Therefore, the Fisher-Rao metric serves as an interpretable connection between two perspectives.

Theorem 2 (Predictive nature of the Transformer (informal)) *Let $\mathcal{M} = \Delta^{k-1}$ be the probability simplex equipped with the Fisher-Rao metric. Let $\{\pi_z\}_{z=1}^{k}$ be memory-1[2] opponent prototypes. Let $\lambda^i \in \mathcal{M}$ denote the agent's belief.*

Given history $x_{\leq t}$, define the negative log-likelihood:

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

There exists a depth-L disentangled transformer $\mathcal{T}_L$ [7] 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:

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

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

We conjecture that, in multi-agent settings, mutual forward passes induce coupled belief dynamics that converge to Predictive Equilibrium (5) when the Fisher curvature at the prior is positive definite.

Because we use Mirror Descent with the negative entropy mirror map $\psi(\lambda) = \sum_j \lambda_j \log \lambda_j$, the update in dual coordinates $\theta = \nabla\psi(\lambda)$ is: $\theta_{t+1} = \theta_t - \eta\nabla\mathcal{L}(\lambda_t)$. This corresponds to natural gradient descent on the Fisher-Rao manifold to first order: $\lambda_{t+1} = \lambda_t - \eta\, g^{-1}(\lambda_t)\nabla\mathcal{L}(\lambda_t) + O(\eta^2)$, where $g(\lambda) = \nabla^2\psi(\lambda) = \mathrm{diag}(1/\lambda_j)$ is the Fisher metric and $\widetilde{\nabla}\mathcal{L} = g^{-1}\nabla\mathcal{L}$ is the natural gradient.

This position could be summarised in a statement:

Attention implements mirror descent, which is first-order natural gradient descent, approximating flow on the Fisher manifold that has fixed points at mixed Predictive equilibriua.

4. Thought experiment

To illustrate the paper's findings, let us do a small thought experiment based on the Axelrod's IPD tournament analysis [3]. Consider two PPI agents in IPD (Table 1) where one agent (Alice) must in-

2. Memory-1 means the agents's decision depends only on previous immediately preceding round.

fer whether her opponent is TFT[3] or Grim[4]. After several rounds of mutual cooperation, 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 AllCoordinate until defection occurs, so early observations are uninformative) which delays its recognition, and Alice remains out of equilibrium. Notably, if Alice was trained on a defection-heavy prior, it would reverse the geometry: Grim becomes easier to recognise. This example illustrates that equilibrium selection is not determined by payoffs alone, but by what geometry of the training distribution looks like.

5. Discussion

In this work, we have proposed a perspective, unifying test-time adaptation and mechanistic interpretability. We support this claim by the following 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 work supports this interpretation. ShapeLLM [19] demonstrates that LLM agents can mutually shape each other through in-context interaction, consistent with coupled mirror descent dynamics.

6. Limitations and Future work

However, it is important to notice that the proposed perspective has substantial limitations:

- The bayesian transformer predictor we base the framework assumes a finite, known set of opponent prototypes, which limits its practical applications. Real opponents may lie outside this set. Moreover, they can be non-stationary. Characterising 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 Meulemans et al. [12]. The joint dynamics may not converge to a unique PE. Curvature of the manifold may become negative caused by chaotic and adversarial behaviour. See, for example, Meulemans et al. [12] for more comprehensive analysis.
- Our analysis assumes the agent observes the opponent's actions and does not handle partial observability which may notably impact the training distribution and variance of the in-context "gradient'.

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

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

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

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

[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.