Gauge Equivariant Extension of FEP and Attention

Robert C. Dennis

Abstract

We present a unified gauge-theoretic formulation of attention and message communication in multi-agent Bayesian systems performing variational inference. Each agent is modeled as a smooth local section of an associated bundle with statistical manifold fiber over a base manifold $\mathcal{C}$.

Inter-agent communication is mediated by parallel-transport operators representing frame rotations between the agents' local gauge frames.

Transformer attention weights arise naturally as normalized exponentials of information-geometric divergences such that agent alignment, rather than dot products, govern agent-agent interaction.

We demonstrate that in the isotropic, flat-bundle, delta-function limit (where all local frames coincide globally) the generalized attention weights reduces to the canonical transformer rule $\beta_{ij} \propto \mathbf{softmax}(Q_i K_j^\top)$ with an additional key-dependent bias.

Starting from a generative model of agent-to-agent message exchange, we derive a generalized variational free-energy functional whose stationary points govern agent evolution. In the absence of observations, this functional defines a vacuum theory in which all agents converge to identical beliefs; introducing observations breaks this symmetry and reproduces the familiar machine-learning loss term.

Our construction therefore fuses information geometry, non-abelian gauge theory, and transformer architecture into a single formalism, allowing curvature-aware, uncertainty-sensitive neural architectures to be grounded in standard geometric and variational principles.

1. Introduction

Recent advances in neuroscience and intelligent systems have independently converged on the idea that intelligent systems must integrate perception, inference, and communication under constraints of uncertainty. Friston's Free Energy Principle (FEP) provides a general variational formulation of inference in cognitive systems, whereas the attention mechanism in modern machine learning architectures defines a powerful (although empirically derived) rule for token prediction. Despite their shared reliance on probabilistic inference and pairwise interaction, these two frameworks remain stubbornly separated. In particular, transformer attention lacks an underlying geometric or mathematical foundation and details on how and why modern machine learning architectures work remain obscured in a cloud of mystery. Perhaps the most important feature of these architectures is scaling.

In this report we propose a unified, gauge-equivariant framework that connects these disparate yet similar domains. Our framework is based on geometry whereby each agent

is modeled as a smooth local section of an associated bundle with statistical manifold fibers over a base manifold. Inter-agent communication arises naturally through a non-abelian gauge connection that defines parallel-transport operators between agents' local gauge frames. Within this geometry, attention emerges as a gauge-aligned Kullback–Leibler (KL) term derived directly from the variational free energy of a coupled multi-agent generative model.

We further show that in the flat-bundle, isotropic, delta-function limit attention reduces to the standard transformer dot-product attention QK^T, thereby identifying attention as the degenerate case of a broader geometric law of communication predicated upon the FEP. We further show that hard, one-hot attention encoding is the zero-temperature limit of the FEP agent-agent coupling term and the large temperature limit leads to uniform encoding. We demonstrate this by simulating a toy model of variational gradient descent under generalized free energy of multi-variate Gaussian agents and by applying our alignment expression to a frozen transformer.

Our model suggests that curvature, holonomy, and uncertainty in the latent manifold, base manifold, and gauge group jointly shape to patterns of communication organization of agents. Thus, this gauge-equivariant formulation of the FEP may offer a novel path toward curvature-aware neural architectures and points to a broader unification of cognition, geometry, and machine learning within a single geometric framework.

2. Bridging the Free Energy Principle and Machine Learning

We begin by defining the mathematical model, richness, and its consequences.

We begin with a general geometric construction which, a priori, bears no resemblance to standard machine learning models. The formulation is mathematically rich, naturally supporting hierarchical emergence of meta-agents, scale interactions, and non-trivial holonomy of transport.

Each agent is modeled as a local section of an associated bundle to a principal fiber bundle whose base space is the manifold $\mathcal{C}$, as defined below. Beliefs and models are encoded in the associated bundle constructed from a structure group G acting on a statistical-manifold fiber $\mathcal{B}_q$. This framework enables a unified description of both intra-agent inference and inter-agent communication in terms of gauge-covariant transport.

2.1. Defining the Model

Let $\pi : \mathcal{N} \to \mathcal{C}$ be a smooth principal G-bundle, where $\mathcal{C}$ is a smooth manifold and G a Lie group acting freely and transitively on the right of $\mathcal{N}$. The projection satisfies $\pi(n \cdot g) = \pi(n)$ for all $n \in \mathcal{N}$ and $g \in G$.

Let $\rho : G \to \mathrm{Aut}(\mathcal{B}_q)$ be a representation of G on a smooth statistical manifold $\mathcal{B}_q$. Depending on context, $\mathcal{B}_q$ may be a K-dimensional probability simplex Δ^K (for categorical distributions) or a statistical manifold endowed with information geometry. In this work we take $\mathcal{B}_q$ to be a Gaussian manifold with gauge group $SO(3)$ for clarity and analytic tractability.

Although $\mathcal{B}_q$ need not be a vector space, key geometric structures—divergences, metrics, and connections—remain well-defined on such statistical manifolds (Amari2016). Given the principal bundle $\mathcal{N}$, we define the associated bundle

$$\mathcal{E}_q := \mathcal{N} \times_\rho \mathcal{B}_q = (\mathcal{N} \times \mathcal{B}_q)/\sim, \qquad (1)$$

where $(n \cdot g, b) \sim (n, \rho(g)b)$.

This yields a fiber bundle $\pi_{\mathcal{E}_q} : \mathcal{E}_q \to \mathcal{C}$ with fiber $\mathcal{B}_q$, over which smooth sections can be defined.

2.2. Definition (Agent).

An agent is an open local section of $\mathcal{E}_q$ over $\mathcal{C}$:

$$\mathcal{A}^i = \sigma_q^i(c) = q_i(c)$$

$$\sigma_q^i : \mathcal{U}_i \subset \mathcal{C} \to \mathcal{B}_q,$$

where $\mathcal{U}_i$ is an open subset of $\mathcal{C}$ and i indexes the agent.
When $\mathcal{B}_q$ is Gaussian, an agent is represented by a field of sufficient statistics $(\mu_i(c), \Sigma_i(c))$ defined over $\mathcal{U}_i$.

2.3. Definition (Multi-agent System).

A multi-agent system $\mathcal{M}$ over $\mathcal{C}$ is a collection of agents indexed by $\mathcal{I}$:

$$\mathcal{M} = \{\mathcal{A}^i = \sigma_q^i(c)\}_{i \in \mathcal{I}}. \qquad (2)$$

Agents generally overlap on the intersections $\mathcal{U}_i \cap \mathcal{U}_j$.

2.4. Definition (Meta-agent and Epistemic Death).

A meta-agent is a multi-agent system whose component agents share identical section values on their overlap, i.e.

$$\mu_i(c) = \mu_j(c), \quad \Sigma_i(c) = \Sigma_j(c),$$

or equivalently $q_i(c) = q_j(c)$ for $c \in \mathcal{U}_i \cap \mathcal{U}_j$.

A set of agents is said to be epistemically dead if they identically share both beliefs and models. Notably, while the constituent agents of a meta-agent may be epistemically dead, the meta-agent itself need not be; such agents can be integrated out, yielding coarse-grained higher-order entities and a route towards emergence.

From the horizontal lift of the connection from $\mathcal{N}$ to $\mathcal{E}_q$, we obtain a hierarchy of morphisms and induced connections across scales (i, j). In particular, we define:

1. $\Omega^s : \Gamma^s(\mathcal{B}_q) \to \Gamma^s(\mathcal{B}_q), \qquad (3)$
2. $\Lambda^s_{s'} : \Gamma^s(\mathcal{B}_q) \to \Gamma^{s'}(\mathcal{B}_q), \qquad (4)$

where $\Gamma(\mathcal{B}_q)$ denotes the space of smooth sections of $\mathcal{B}_q$ over $\mathcal{C}$. $\Lambda^s_{s'}$ represents parallel transport between scales s and s'. For the present work we restrict attention to intra-

scale (agent-level) connections and pursue meta-agent emergence and cross-scale communication in future work.

2.5. Types of Parallel Transport in Epistemic Geometry

Parallel transport arises in several distinct but related contexts, depending on whether transport occurs along the base manifold $\mathcal{C}$, within a fiber, or between agent frames.

2.5.1. Horizontal Transport Along the Base Manifold $\mathcal{C}$

Let $\pi : \mathcal{N} \to \mathcal{C}$ be a principal G-bundle, and $\mathcal{E} = \mathcal{N} \times_\rho \mathcal{B}$ the associated bundle. For a path $\gamma : [0,1] \to \mathcal{C}$, a connection one-form $A \in \Omega^1(\mathcal{C}, \mathfrak{g})$ defines parallel transport along γ via the path-ordered exponential:

$$T_\gamma = \mathcal{P} \exp\left(-\int_\gamma A_\mu(c)\, dc^\mu \right) \in G. \quad (5)$$

This maps fiber elements between base points:

$$b(c_0) \in \mathcal{B}_{c_0} \quad \mapsto \quad b(c_1) = \rho(T_\gamma)\, b(c_0) \in \mathcal{B}_{c_1}.$$

This is the canonical notion of base-space parallel transport, needed when tokens are treated as delta-limiting agents.

2.5.2. Vertical Transport Within a Single Fiber $\mathcal{B}_c$

At a fixed base point $c \in \mathcal{C}$, the fiber $\mathcal{B}_q(c)$ is itself a curved statistical manifold. Parallel transport within $\mathcal{B}_q(c)$ along a curve $\eta(\tau)$ with tangent vector $\dot{\eta}(\tau)$ is governed by a connection ∇ intrinsic to $\mathcal{B}$:

$$\nabla_{\dot{\eta}} V = 0. \quad (6)$$

In information geometry, $\mathcal{B}_q(c)$ carries dual connections $(\nabla^{(e)}, \nabla^{(m)})$ associated with exponential and mixture families. Curvature is characterized by the Riemann tensor

$$R^{\mathcal{B}}(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z.$$

Thus, even at a single base point, the fiber admits nontrivial geodesics and holonomies representing purely epistemic transformations of beliefs. These transports represent the agent's dynamic beliefs and models.

2.5.3. Intra-Agent Spatial Transport (Within Fiber Bundle)

To compare beliefs $q_i(c_1)$ and $q_i(c_2)$ held by the same agent over nearby points $c_1, c_2 \in \mathcal{U}_i \subset \mathcal{C}$, we introduce the agent's local gauge frame $U_i(c) = \exp[\phi_i(c)] \in G$ and define the corresponding (pure-gauge) connection

$$A_\mu^{(i)}(c) = U_i^{-1}(c)\, \partial_\mu U_i(c).$$

This ensures the correct transformation law $A_\mu^{(i)} \mapsto g_i A_\mu^{(i)} g_i^{-1} + g_i \partial_\mu g_i^{-1}$ under local gauge transformations.

The associated parallel transport operator is

$$T^{(i)}_{c_1 \to c_2} = \mathcal{P} \exp\left(-\int_{c_1}^{c_2} A^{(i)}_\mu \, dc^\mu \right), \quad (7)$$

which is gauge-covariant in the sense that $T^{(i)}_{c_1 \to c_2} \mapsto g_i(c_2)\, T^{(i)}_{c_1 \to c_2}\, g_i^{-1}(c_1)$.

which realizes gauge-covariant transport between fibers for a fixed agent.

2.5.4. Inter-Agent Frame Transport (At a Shared Point)

When two agents $\mathcal{A}_i$ and $\mathcal{A}_j$ overlap at a point $c \in \mathcal{U}_i \cap \mathcal{U}_j$, each has a local gauge frame $\phi_i(c), \phi_j(c) \in \mathfrak{g}$. The inter-agent gauge transformation

$$\Omega_{ij}(c) = e^{\phi_i(c)} e^{-\phi_j(c)} \in G \quad (8)$$

transports beliefs from j's frame into i's:

$$q_j(c) \mapsto q_i^{(j)}(c) = \rho(\Omega_{ij}(c))\, q_j(c),$$

which defines the gauge-aligned KL divergence

$$D_{\mathrm{KL}}[q_i(c) \,\|\, \Omega_{ij}(c)\, q_j(c)].$$

2.5.5. Composite Transport and Holonomy

More generally, a belief may be transported along a composite path

$$\gamma = [(i_1, c_1) \to \cdots \to (i_n, c_n)],$$

involving both spatial transport within agents and inter-agent frame shifts. The total transport operator is

$$P_\gamma = T^{(i_n)}_{c_1 \leftarrow c_n}\, \Omega_{i_n i_{n-1}} \cdots \Omega_{i_2 i_1}\, T^{(i_1)}_{c_2 \leftarrow c_1}. \quad (9)$$

This defines the holonomy of belief transport,

$$\mathrm{Hol}_\gamma(q) = P_\gamma q,$$

which is nontrivial ($\mathrm{Hol}_\gamma \neq \mathrm{Id}$) when the loop encloses curvature.

Finally, the curvature (or field strength) associated with the connection A_μ is given by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu], \quad (10)$$

which measures the noncommutativity of covariant derivatives and encodes the local holonomy of the gauge field.

2.5.6. Summary

Transport Type	Domain	Operator	Purpose
Horizontal (base)	$c_1 \to c_2 \in \mathcal{C}$	$T_\gamma = \mathcal{P}\exp(-\int A)$	Transport beliefs across space
Vertical (fiber)	$\eta(\tau) \subset \mathcal{B}_c$	∇ or $\rho(\exp\phi)$	Update beliefs
Intra-agent	$q_i(c_1) \to q_i(c_2)$	$T^{(i)}$ from $A_\mu^{(i)}$	Compare distant beliefs
Inter-agent	$q_j(c) \mapsto q_i^{(j)}(c)$	$\Omega_{ij}(c)$	Frame alignment across agents
Composite	$(i_1, c_1) \to \cdots$	P_γ	Transport and holonomy

Table 1: Types of parallel transport in epistemic gauge geometry.

2.6. Derivation of Variational Energy From FEP

2.7. Obtaining $D_{KL}(q_i|\Omega_{ij}q_j)$ from FEP

The generalized $D_{KL}(q_i|\Omega_{ij}q_j)$ term isn't necessarily a new feature in the free energy principle. What is new is promoting probabilities into an associated bundle. We can obtain the $D_{KL}(q_i|\Omega_{ij}q_j)$ term directly from the FEP as follows:

For a single latent variable $k \in \mathcal{B}_q(c)$ over a base manifold point c along with observation o we have the standard variational free energy as

$$\mathcal{F}[q] = D_{KL}(q(k)|p(k)) - \mathbb{E}_q[\log p(o|k)]$$

In a multi-variable ($\dim \mathcal{B}_q = K$) case over a base manifold point c with latents $k_1, \ldots, k_K \in \mathcal{B}_q$ the exact free energy under an approximate posterior $q(k_1, \ldots, k_K)$ is given by

$$\mathcal{F}[q] = \mathbb{E}_q[\log q(k_1, \ldots, k_K)] - \mathbb{E}_q[\log p(k_1, \ldots, k_K; o_1, \ldots, o_K)]$$

Now, if we invoke the mean field assumption on agent's beliefs we have

$$q(k_1, \ldots, k_K) = \prod_i q_i(k_i)$$

Next, we assume the world does not treat agents as independent - i.e. the generative model encodes relationships between agents' latent states.

Therefore we write

$$p(k_1, \ldots, k_K) \propto \prod_i p_i(k_i) \prod_{ij} \psi_{ij}(k_i, k_j)$$

where $p_i(k_i)$ is the local prior for agent i's latent cause. $\psi_{ij}(k_i, k_j)$, meanwhile, is the pairwise interaction describing agent-agent coupling. Here we assume entangled three

(or more) agent interactions to be suppressed focusing primarily on two-agent message passing.

Next, we choose

$$\psi_{ij}(k_i, k_j) \propto e^{-\lambda_{ij} d(k_i, \Omega_{ij} k_j)}$$

where $d(k_i, \Omega_{ij} k_j)$ is a distance-like function between agent i and agent i's interpretation of k_j through the transport operator Ω_{ij}. λ_{ij} is a coupling strength (precision) to be determined later.

Next, we have

$$p(k_1, \dots, k_K; o_1, \dots, o_K) = p(k_1, \dots, k_K) \prod_i p_i(o_i | k_i)$$

such that each agent has its own local observation o_i and additionally, that, given k_i this observation o_i is independent of the other agents.

Therefore, combining terms we have

$$p(k, o) \propto \prod p_i(k_i) \prod_{ij} \psi_{ij}(k_i, k_j) \prod_i p_i(o_i | k_i)$$

Note: all constants will ultimately be absorbed into the partition function Z.

We now write the VFE as

$$\mathcal{F}[q] = \mathbb{E}_q[\log q(k)] - \mathbb{E}_q[\log p(k, o)]$$

and expand the terms.

First, since $q(k) = \prod_i q_i(k_i)$

$$\mathbb{E}_q[\log q(k)] = \sum_i \mathbb{E}_{q_i}[\log q_i(k_i)]$$

Next, from above we expand

$$\log p(k, o) = \sum_i \log p_i(k_i) + \sum_{i,j} \log \psi_{ij}(k_i, k_j) + \sum_i \log p_i(o_i | k_i) - \log Z$$

Then we take the expectation over $q(k) = \prod_m q_k(k_m)$:

$$\mathbb{E}_{q(k)}[\log p(k, o)] = \sum_i \mathbb{E}_{q_i(k_i)} \log p_i(k_i) + \sum_{i,j} \mathbb{E}_{q_i(k_i), q_j(k_j)} \log \psi_{ij}(k_i, k_j) + \sum_i \mathbb{E}_{q_i(k_i)} \log p$$

Plugging these into our VFE we find

$$\mathcal{F}[q] = \sum_i D_{KL}(q_i(k_i) | p_i(k_i)) - \sum_i \mathbb{E}_{q_i}[\log p_i(o_i | k_i)] - \sum_{i,j} \mathbb{E}_{q_i q_j} \log \psi_{ij}(k_i, k_j)$$

We shall focus our attention on the cross-term

$$\sum_{i,j} \mathbb{E}_{q_i q_j} \log \psi_{ij}(k_i, k_j)$$

We previously defined

$$\psi_{ij}(k_i, k_j) \propto e^{-\lambda_{ij} d(k_i, \Omega_{ij} k_j)}$$

Therefore our term becomes

$$\sum_{i,j} \lambda_{ij} \mathbb{E}_{k_i k_j} d(k_i, \Omega_{ij} k_j) + \text{const}$$

Again, the constant can be absorbed by the partition function Z so we shall not consider it further.

Now, in general, any function $d(k_i, k_j)$ satisfying

$$d \geq 0$$

$$d(k, k) = 0$$

$$d \leq \infty \quad \text{and integrable}$$

will suit our compatibility. Generally we can choose any f-divergence here. f-divergences (and especially the KL-divergence) are uniquely suited to produce exponential family compatibilities.

A natural choice for multi-variate Gaussians is the Mahalanobis distance

$$d(k_i, \Omega k_j) = \sqrt{(\Omega_{ij} k_j - k_i)^T \Sigma_i^{-1} (\Omega_{ij} k_j - k_i)}$$

The expectation over q_i, q_j then leads to $D_{KL}(q_i(k_i) | \Omega_{ij} q_j(k_j))$ as the second order expansion of the log-density ratio - namely the KL term. Therefore, we find

$$\mathcal{F}[q] = \sum_i D_{KL}(q_i(k_i) | p_i(k_i)) - \sum_i \mathbb{E}_{q_i}[\log p_i(o_i | k_i)] + \sum_{i,j} \lambda_{ij} D_{KL}(q_i(k_i) | \Omega_{ij} q_j(k_j))$$

Since this was considered at a single general base manifold point it applies to any base manifold point. Note: we are not doing spatial/base transport - this is "in-fiber" transport. Therefore, considering the full base manifold our generalized energy becomes

$$S = \sum_i \int_{\mathcal{C}} \chi_i(c) \Big[\alpha_i \, \mathrm{KL}\big(q_i(c) \,\|\, p_i(c)\big) - \gamma_i \, \mathbb{E}_{q_i(c)}\big[\log p_i(o_i \,|\, c)\big] \Big] \mathrm{d}c \; + \; \sum_{i,j} \int_{\mathcal{C}} \chi_{ij}(c) \, \beta_{ij}(c) \, \mathrm{K}$$

where χ_i and χ_{ij} represent indicator functions of the agent's support and interaction/overlap region respectively. In the most general case we could include gauge-field energies, curvatures, frame-gauge couplings, and other regularizers (see appendix).

We have found that the attention-coupling term in our generalized energy functional is just the mean-field variational free energy of a multi-agent generative prior with

gauge-aligned pairwise factors.

Therefore, multi-agent communication within a gauge covariant formulation allows the FEP to be satisfied as well as allows us to connect attention, transformers, and machine learning to variational inference.

2.8. Obtaining β_{ij}

We shall treat β_{ij} as a set of weights agent i assigns to all other agents j. For fixed i β_{ij} should form a probability distribution. Namely,

$$\sum_j \beta_{ij} = 1$$

We derive the form of β_{ij} by letting agent i choose β_{ij} to minimize its own expected coupling cost to other agents.

First, we define the mismatch cost between agents as

$$C_{ij} = D_{KL}(q_i|\Omega_{ij}q_j)$$

Next, we define the "expected disagreement" as

$$\mathcal{L}_i(\beta_i) = \sum_j \beta_{ij} C_{ij}$$

Upon minimization this, being linear in β_{ij}, would put all weight in the best neighbor j as $j^* = \arg\min_j C_{ij}$. This is a hard attention corresponding to "one-hot" encoding. To obtain a soft attention we introduce an "uncertainty cost" on β_i as

$$\mathcal{L}_i^{ent}(\beta_i) = \kappa \sum_j \beta_{ij} \log \beta_{ij}$$

Incorporating this term penalizes low entropy distributions. Our total objective then becomes

$$\mathcal{J}_i(\beta_i) = \sum_j \beta_{ij} C_{ij} + \kappa \sum_j \beta_{ij} \log \beta_{ij}$$

This now prevents agents from collapsing β_{ij} to a single term: accuracy versus complexity.

Next we optimize β_{ij} via Lagrange multipliers in the standard manner. Define the Lagrangian

$$\mathscr{L}_i(\beta_i, \xi_i) = \sum_j \beta_{ij} C_{ij} + \kappa \sum_j \beta_{ij} \log \beta_{ij} + \xi_i(\sum_j \beta_{ij} - 1)$$

then solve

$$\frac{\partial \mathscr{L}}{\partial \beta_{ij}} = 0$$

and

$$\frac{\partial \mathscr{L}}{\partial \xi_i} = 0$$

We find

$$\frac{\partial \mathscr{L}}{\partial \beta_{ij}} = C_{ij} + \kappa(\log \beta_{ij} + 1) + \xi_i = 0$$

Solving for β_{ij} we find

$$\beta_{ij} = e^{-1} e^{-\frac{\xi_i}{\kappa}} e^{-\frac{C_{ij}}{\kappa}}$$

The first two terms do not depend on j so we have

$$\beta_{ij} = K_i e^{-\frac{C_{ij}}{\kappa}}$$

Now we impose normalization $\sum_j \beta_{ij} = 1$ to find

$$K_i = \frac{1}{\sum_k e^{-\frac{C_{ik}}{\kappa}}}$$

and thus we arrive at our final form

$$\beta_{ij} = \frac{e^{-\frac{C_{ij}}{\kappa}}}{\sum_k e^{-\frac{C_{ik}}{\kappa}}}$$

$$\beta_{ij} = \frac{e^{-\frac{D_{KL}(q_i | \Omega_{ij} q_j)}{\kappa}}}{\sum_k e^{-\frac{D_{KL}(q_i | \Omega_{ik} q_k)}{\kappa}}}$$

Hence, each agent assigns weights β_{ij} according to their relative consistency where κ controls the sharpness of selection. In the limit $\kappa \to 0$ the β_{ij} weights collapse to hard-attention whereas for large κ we approach uniform weighting.

Therefore, given agents as local open sections over the base space our complete attention weights are given as

$$\beta_{ij}(c) = \frac{\exp\left[-\frac{1}{\kappa} \mathrm{KL}\big(q_i(c) \,\big\|\, \Omega_{ij} q_j(c)\big)\right] m_{ij}(c)}{\displaystyle\sum_k \exp\left[-\frac{1}{\kappa} \mathrm{KL}\big(q_i(c) \,\big\|\, \Omega_{ik} q_k(c)\big)\right] m_{ik}(c)}$$

where $m_{ij}(x)$ is the overlap of agent i and agent j support (or mask). Or said simply; their interaction volume.

2.9. Reduction to Transformer Attention

In this section we show that standard transformer self-attention is recovered as a δ-function limiting case of our gauge-covariant, uncertainty-aware framework where, in this view, tokens are generally "fuzzy" embeddings (that is, a field vector μ and Σ ellipse under gauge frame ϕ). In the δ-function limit this becomes the standard token embedding in a globally fixed and flat gauge frame where all variance collapses and we recover the non-probabilistic vector.

In particular, we apply the following to our framework:

(i) a discrete base space (discrete tokens instead of extended open sections - δ-function localization),

(ii) a flat bundle / trivial connection (shared globally defined frame, no curvature),

and

(iii) isotropic uncertainty (identical spherical covariance for each agent which we keep for the derivation and then take the limit as tokens become delta-function distributions in the fiber).

Under these assumptions, our KL-based attention law, β_{ij} reduces exactly to the standard $(QK^\top)V$ form of transformer attention thereby illuminating the geometric source of the ad-hoc dot product similarity. We show this below.

2.10. Setup: Agents as Gaussian Beliefs in Local Frames

In our general formulation, each agent i (token) carries a local state modeled as a Gaussian

$$q_i = \mathcal{N}(\mu_i, \Sigma_i),$$

where $\mu_i \in \mathbb{R}^K$ is the agent's mean representation in its local frame, and $\Sigma_i \in \mathbb{R}^{K\times K}$ is its belief covariance (symmetric positive definite).

Communication between agents i and j is mediated by a gauge-covariant parallel transport operator

$$\Omega_{ij} \in G \subset GL(d),$$

which maps representations expressed in agent j's local frame into agent i's local frame (K-dimensional irreps).

In our general framework agents share states via KL coupling defined by the (negative) KL divergence between i's belief and j's belief transported into i's frame:

That is, in our variational energy function (to be defined below) we have the term

$$\beta_{ij} KL(q_i || \Omega_{ij} q_j)$$

where β_{ij} is given by

$$\beta_{ij} \;=\; \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})} \;=\; \mathrm{softmax}_j\big(s_{ij}\big),$$

such that

$$s_{ij} \;\equiv\; -\mathrm{KL}(q_i \,\|\, \Omega_{ij} q_j),$$

Here $\Omega_{ij} q_j$ is the parallel transported Gaussian with transported mean $\mu_{j\to i} = \Omega_{ij}\mu_j$ and transported covariance $\Sigma_{j\to i} = \Omega_{ij}\Sigma_j\Omega_{ij}^\top$ under group action.

The message (or update) received by agent i is then

$$m_i = \sum_j \beta_{ij}\Omega_{ij}\mu_j$$

The message update is the analog, in our geometry, of the usual attention aggregation $\sum_j \alpha_{ij} V_j$ in a transformer block.

Our goal is to show that, under the above simplifying limits, our expressions reduce to the standard Transformer formulas.

2.11. The Transformer Limit

We now impose the three simplifying assumptions.

2.11.1. Discrete base / tokens as agents.

We consider a base space $\mathcal{C}$ to be a finite index set of positions $\{1, \dots, N\}$ (e.g. token positions in a sequence).

Each agent i is now just "the token at position i". There is no spatial overlap integral; all quantities are evaluated at a single site thereby considerably simplify our variational energies.

2.11.2. Flat Bundle / Trivial Connection.

We next assume there is a single global frame shared by all agents, i.e. no curvature and no position-dependent frame misalignment.

Concretely, we take

$$\Omega_{ij} = W\mathbb{1}_{d\times d} = \Omega \quad \text{for all } i, j,$$

where $\Omega \in \mathbb{R}^{d\times d}$ is a fixed linear map.

Intuitively, this corresponds to a trivial principal bundle with a flat connection: parallel transport between any two sites is global and path-independent. In particular, expressions like "rotate j's state into i's frame" reduce to "apply the same learned linear map Ω."

2.11.3. Isotropic and identical uncertainty.

We assume all agents have the same spherical covariance:

$$\Sigma_i = \sigma^2 \mathbb{1}_{d\times d} \quad \text{for all } i,$$

with $\sigma^2 > 0$. This implies that, after transport by Ω, the comparison metric between two beliefs reduces to a scaled Euclidean/Mahalanobis distance with shared precision $1/\sigma^2$. Equivalently, every agent is equally confident in all directions, and confidence level is the same across agents.

Under these assumptions, the transported covariance becomes

$$\Sigma_{j\to i} = \Omega \Sigma_j \Omega^\top = \sigma^2(\Omega\Omega^\top)$$

To make contact with the simplest transformer form, we can absorb $\Omega\Omega^\top$ into a learned rescaling of Ω , into the variance σ^2, or under an $SO(3)$ gauge group this term becomes the identity. This is equivalent to the standard freedom in attention to choose arbitrary learned projection matrices and overall temperature scaling.

Therefore,

$$\Sigma_{j\to i} \approx \sigma^2 \mathbb{1} \quad \text{and} \quad \Sigma_{j\to i}^{-1} \approx \tfrac{1}{\sigma^2}\mathbb{1}.$$

Note: we are considering $SO(3)$ for simplicity due to multi-variate Gaussians and KL-divergence being invariant under rotation (this is straightforward to show). Interestingly, in the general $SO(3)$ case global token attention cannnot exist due to $SO(3)$ presenting an obstruction, $\pi_1(SO(3)) = \mathbb{Z}_2$. To define a global attention in this case requires lifting to $SU(2)$ - otherwise we can only assign localized agent attentions!

2.11.4. Emergence of the Dot Product Attention

For two Gaussians with identical isotropic covariance $\sigma^2\mathbb{1}$, the Kullback--Leibler divergence reduces to a scaled squared distance between their means (for clarity we no longer write $\mathbb{1}$):

$$\mathrm{KL}\big(\mathcal{N}(\mu_i,\sigma^2) \,\big\|\, \mathcal{N}(\Omega\mu_j,\sigma^2)\big) = \frac{1}{2\sigma^2}\big\|\Omega\mu_j - \mu_i\big\|^2$$

Note: the trace is canceled by the dimension term of the standard KL form for Gaussians and the $\log\det$ terms vanish.

Next, we have

$$s_{ij} = -\mathrm{KL}(q_i \| \Omega q_j) = -\frac{1}{2\sigma^2}\|\Omega\mu_j - \mu_i\|^2$$

Expanding the squared norm,

$$\|\Omega\mu_j - \mu_i\|^2 = \|\Omega\mu_j\|^2 + \|\mu_i\|^2 - 2\,\mu_i^\top(\Omega\mu_j).$$

Therefore

$$s_{ij} = \frac{1}{\sigma^2}\mu_i^\top(\Omega\mu_j) \;-\; \frac{1}{2\sigma^2}\|\Omega\mu_j\|^2 \;-\; \frac{1}{2\sigma^2}\|\mu_i\|^2$$

Next, we fix i and consider the softmax over j.

Any term in s_{ij} that is independent of j (for that fixed i) will pull out and cancel between numerator and denominator of the softmax. In our above expression, the term $-\frac{1}{2\sigma^2}\|\mu_i\|^2$ does not depend on j and therefore falls out under softmax.

Therefore, up to a softmax-equivalent shift, the effective logit for attention is

$$\tilde{s}_{ij} \;\equiv\; \frac{1}{\sigma^2}\mu_i^\top(\Omega\mu_j) \;-\; \frac{1}{2\sigma^2}\|\Omega\mu_j\|^2.$$

The first term in is bilinear in (μ_i, μ_j):

$$\frac{1}{\sigma^2}\mu_i^\top\Omega\mu_j.$$

We therefore define learned projection matrices $A, B \in \mathbb{R}^{d\times d_k}$ such that

$$AB^\top \;=\; \frac{1}{\sigma^2}\Omega.$$

For example, take any matrix factorization of $\frac{1}{\sigma^2}\Omega$, such as an SVD or a learned low-rank factorization; this is generally always possible. Notice that in particular, under an $SO(3)$ gauge group our term would be $\Omega_{ij} = e^{\phi_i}e^{-\phi_j} = AB^T$

Thus we see that we may define the query (Q) and key (K) vectors as

$$Q_i \;\equiv\; \mu_i^\top A \;\in\; \mathbb{R}^{1\times d_k}, \qquad K_j \;\equiv\; \mu_j^\top B \;\in\; \mathbb{R}^{1\times d_k}.$$

Then

$$Q_iK_j^\top \;=\; \mu_i^\top AB^\top\mu_j \;=\; \frac{1}{\sigma^2}\mu_i^\top\Omega\mu_j.$$

Thus the leading compatibility term in $\tilde{s}_{ij}$,
namely $\frac{1}{\sigma^2}\mu_i^\top\Omega\mu_j$, matches exactly the standard Transformer dot product $Q_iK_j^\top$.

The remaining term $-\frac{1}{2\sigma^2}\|\Omega\mu_j\|^2$ depends only on j. This acts as a key-dependent bias. This additional bias is gauge-geometric. Standard Transformers typically omit key dependent biases, however some architectures input such bias by hand. Here, in a geometric manner, each key holds a global self-salience which influences how all other queries attend to them.

Hence, modulo a (learnable) per-key bias, we have shown:

$$\tilde{s}_{ij} \;\approx\; Q_iK_j^\top.$$

Consequently,

$$\beta_{ij} = \mathrm{softmax}_j(\tilde{s}_{ij}) \;\approx\; \mathrm{softmax}_j(Q_iK_j^\top),$$

which is the standard Transformer attention weighting rule.

Next, recall our gauge-covariant aggregation rule - the message communication

$$m_i = \sum_j \beta_{ij}\, \Omega_{ij} \mu_j.$$

Under our present assumptions, $\Omega_{ij} = \Omega$ is the same learned linear map for all pairs (i, j).

Therefore, define a value projection

$$V_j \;\equiv\; \mu_j^\top C, \qquad C \in \mathbb{R}^{d \times d_v}$$

for some learned matrix C.

Since Ω is fixed across (i, j) and linear in μ_j we can absorb Ω into C,

Thus $\Omega_{ij}\mu_j = \Omega\mu_j$ can be parameterized as V_j for a suitable C.

Then

$$m_i = \sum_j \beta_{ij} V_j.$$

Using the correspondence $\beta_{ij} \approx \alpha_{ij}$ with
$\alpha_{ij} = \mathbf{softmax}_j(Q_i K_j^\top)$ is identical to the standard Transformer attention update.

$$z_i = \sum_j \alpha_{ij} V_j.$$

Thus the message m_i in our model becomes z_i in a Transformer layer.

2.11.5. Machine Learning Loss

In our gauge theory framework observations by agents act as a source term in a vacuum variational energy. The vacuum theory is then

$$S[q(c)] = \alpha \sum_i D_{\mathrm{KL}}[(q_i(c)|p_i(c))]$$

$$+ \sum_{ij} \beta_{ij} D_{\mathrm{KL}}[q_i(c)|\Omega_{ij} q_j(c)]$$

Introducing a per-agent source term of observations $\mathbb{E}_{q_i(c)}[\log p(o_i|c)]$ breaks this symmetry allowing for agents to flow towards unique beliefs and gauge frames.

On the machine learning side these observation sources ARE the training data and we minimize the expected loss term. Notice the following:

1. Assume a delta-posterior. This is what is typically considered in machine learning training data
2. then $-\mathbb{E}_{q_i(c)}[\log p(o_i|c)] \to -\log p(o_i|c)$ the standard negative log-likelihood term

3. but if we take observations to be Gaussian then this simplifies to

$$p(o \mid c) = \mathcal{N}(o \mid c, \Sigma) \quad \Rightarrow \quad -\log p(o \mid c) = \tfrac{1}{2}(o-c)^\top \Lambda (o-c) + \text{const.}$$

$$\boxed{\mathcal{L}_{\text{obs}} = \tfrac{1}{2}\|o-c\|_\Sigma^2} \qquad \text{(Mean–squared error loss)}$$

which is the standard machine learning loss function!

We have shown the following:

1. Under our assumptions , the score $s_{ij} = -\mathrm{KL}(q_i \| \Omega_{ij} q_j)$ reduces (up to key-dependent bias terms that are harmless for attention) to a bilinear form $Q_i K_j^\top$ where $Q_i = \mu_i^\top A$ and $K_j = \mu_j^\top B$ are learned linear projections of the agent means. This reproduces the standard Transformer attention logits $QK^\top$.
2. The aggregation rule $m_i = \sum_j \beta_{ij} \Omega_{ij} \mu_j$ reduces to $m_i = \sum_j \alpha_{ij} V_j$ with $V_j = \mu_j^\top C$, matching the Transformer value projection and weighted sum.
3. Therefore, in the limit of (i) discrete sites interpreted as tokens, (ii) a flat global frame with trivial parallel transport, and (iii) isotropic identical uncertainty so that all covariances collapse to $\sigma^2 \mathbb{1}$, our gauge-covariant, uncertainty-aware message passing law (or $q_i = \delta$ if you like)

$$\beta_{ij} \;\propto\; \mathrm{softmax}_j \Big(-\mathrm{KL}(q_i \,\|\, \Omega_{ij} q_j) \Big),$$

$$m_i = \sum_j \beta_{ij}\, \Omega_{ij} \mu_j$$

becomes exactly the canonical Transformer attention

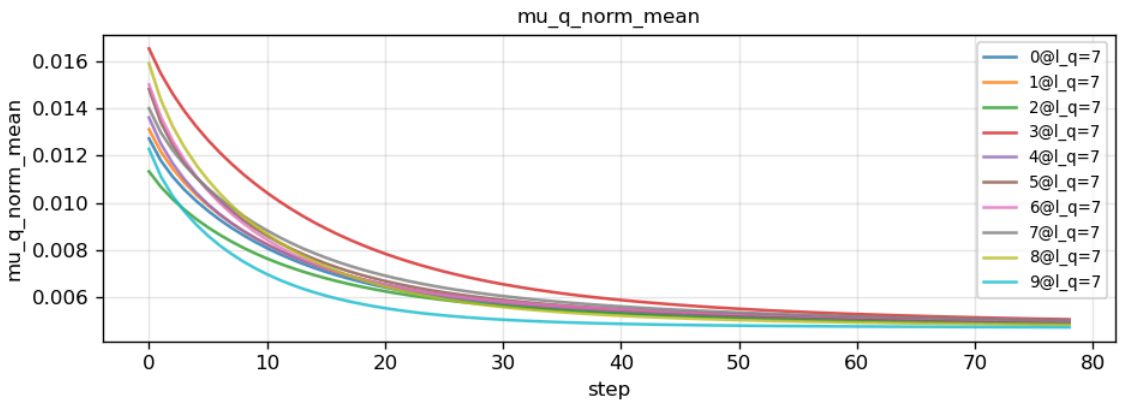
$$\alpha_{ij} = \mathrm{softmax}_j(Q_i K_j^\top),$$

$$z_i = \sum_j \alpha_{ij} V_j.$$

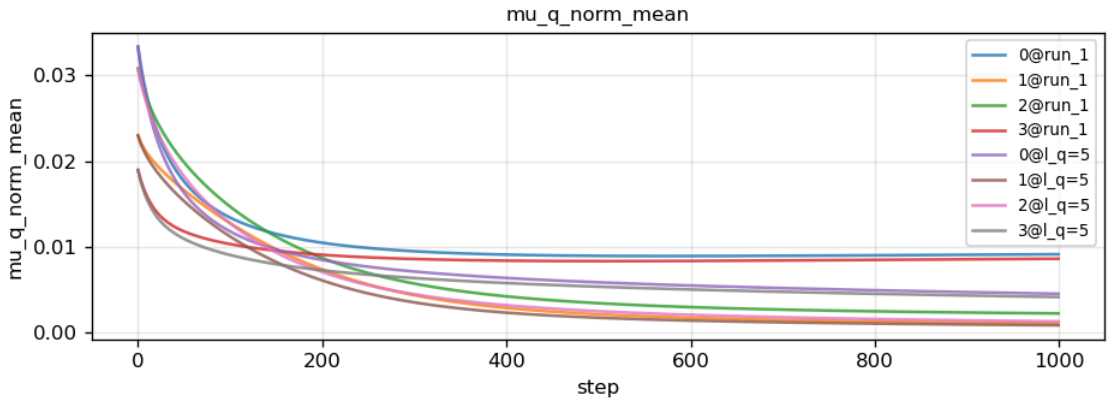
In other words, the standard machine learning dot-product self-attention is the limit of a generalized statistical gauge theory under trivial-connection, isotropic-covariance, discrete-base space ($1D$ lattice) of our gauge-covariant KL-attention rule.

3. Simulations and Empirical Validation

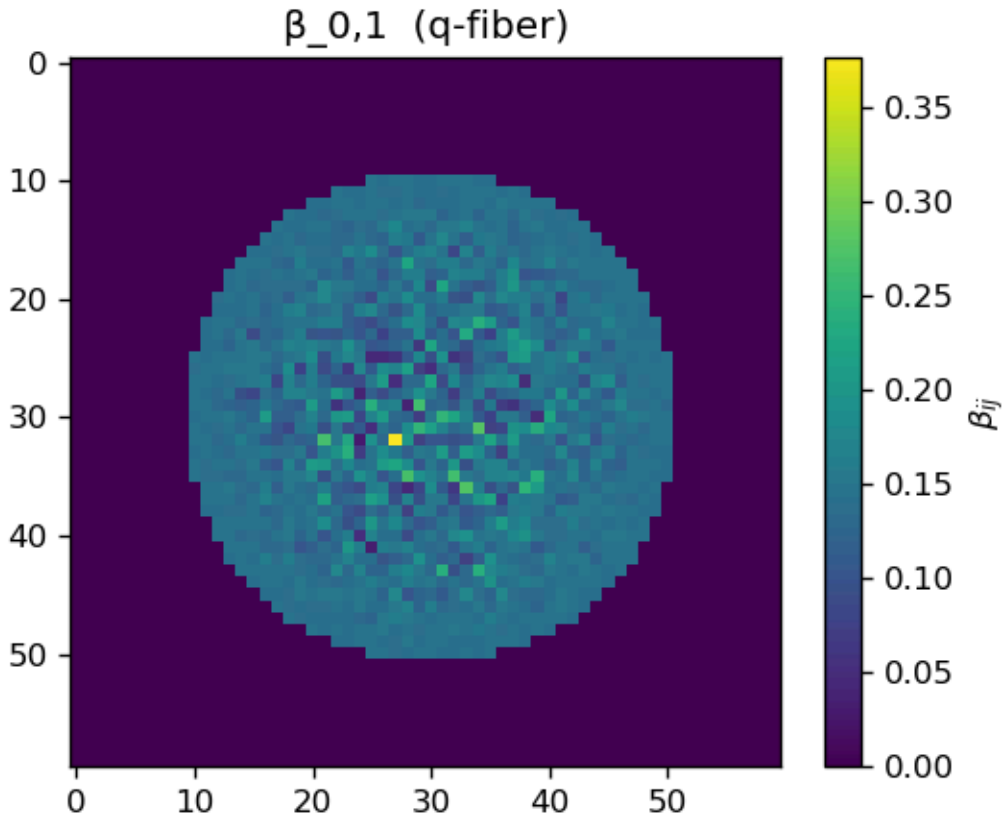
Detailed simulations and analytic derivations produce agent belief and frame flow towards a shared constant vacuum state $\mu_i(c), \phi_i(c) \longrightarrow \mu^*, \phi^*$ under variational gradient descent (all agents share the same models) in the quasi-static regime. All agents share beliefs and gauge frames becoming epistemically dead (see below: $l_q = 7$ is the spin-7 representation of $SO(3)$).



In the case below observations induce specialization.



A typical β_{ij} heat map over the agents' support is given below. The image presents β_{01}. Agents [0] and [1] identically overlap in the two dimensional base space (radius = 20 units, 10 total coincident agents).



4. Discussion

4.1. Gauge Transport via Lie Algebra Frames

The inter-agent transport operator Ω_{ij} serve to compare beliefs (and models if desired) between agents i and j. These operators are defined pointwise over the agent's support in $\mathcal{C}$, and are constructed from local gauge frames associated with each agent.

Let $\phi_i(c) \in \mathfrak{g}$ be a smooth field over $\mathcal{U}_i \subset \mathcal{C}$ valued in the Lie algebra $\mathfrak{g} = \mathrm{Lie}(G)$, representing the gauge frame of agent i. Then, for each pair of agents (i, j), we define the induced transport operator as:

$$\Omega_{ij}(c) := \exp\left(\phi_i(c)\right) \cdot \exp\left(-\phi_j(c)\right),$$

where the exponential map $\exp : \mathfrak{g} \to G$ maps Lie algebra elements to the Lie group G, and the product is taken in the group.

This construction defines a group-valued map $\Omega_{ij}(c) \in G$, which acts on the belief fiber $\mathcal{B}_q$ via the representation ρ. That is, for any section $q_j(c) \in \Gamma(\mathcal{B}_q)$, the transported belief is:

$$\Omega_{ij}(c) \cdot q_j(c) := \rho\left(\Omega_{ij}(c)\right) \cdot q_j(c).$$

These induced transport operators implement a form of gauge-covariant parallel transport between agents. The field $\phi_i(c)$ serves as the agent's gauge frame, and differences in these frames determine the epistemic misalignment between agents.

By construction, $\Omega_{ii} = \mathrm{Id}$, and in general $\Omega_{ij} \neq \Omega_{ji}^{-1}$ unless the gauge fields are flat. This asymmetry encodes epistemic curvature, which we will explore in subsequent sections.

In our framework, each agent maintains a fixed gauge frame $\phi_i(c)$, which defines their local epistemic perspective. This frame serves as the reference against which other agents' beliefs are compared. The term $D_{\mathrm{KL}}[\, q_i(c) \,\|\, \Omega_{ij}(c) \cdot q_j(c) \,]$ then quantifies the epistemic misalignment between agent i and agent j, measured in agent i's own frame. Operationally, this can be interpreted as a communicative act: agent i attempts to transform agent j's beliefs into its own representational basis and evaluates their coherence.

Intuitively, the gauge frame $\phi_i(c)$ can be interpreted as a geometric encoding of the agent's cognitive identity — its internally consistent perspective on the latent manifold $\mathcal{C}$. In this sense, the gauge frame acts as a formal proxy for an agent's "selfhood" or how they interpret their world, organizing all incoming beliefs and observations relative to a quasi-static internal model which gets compared with other agents.

With only attention (which we have shown is due to the $\sum_j \beta_{ij} D_{\mathrm{KL}}[q_i(c)|\Omega_{ij} q_j(c)]$ term and no learning from data all token embeddings will flow to the same embedding - the average of the tokens) and with structured observations will flow to unique embeddings. The mapping (in the case of Gaussians and $SO(3)$ is then complete. This can also be shown easily for categorical distributions).

In this view we hypothesize that our general framework IS a bridge between machine learning and cognition/neuroscience. Attention is possibly all you need to map from machine learning to Friston's free energy principle and a statistical gauge theory (with statistical fibers) is (possibly) all you need to get there.

One major question remains: What generative model, in the Friston FEP view, will produce this attention term - $\sum_j \beta_{ij} D_{\mathrm{KL}}[q_i(c)|\Omega_{ij} q_j(c)]$? If we can find such a generative model then we will have fully mapped the FEP to machine learning.

5. Conclusion

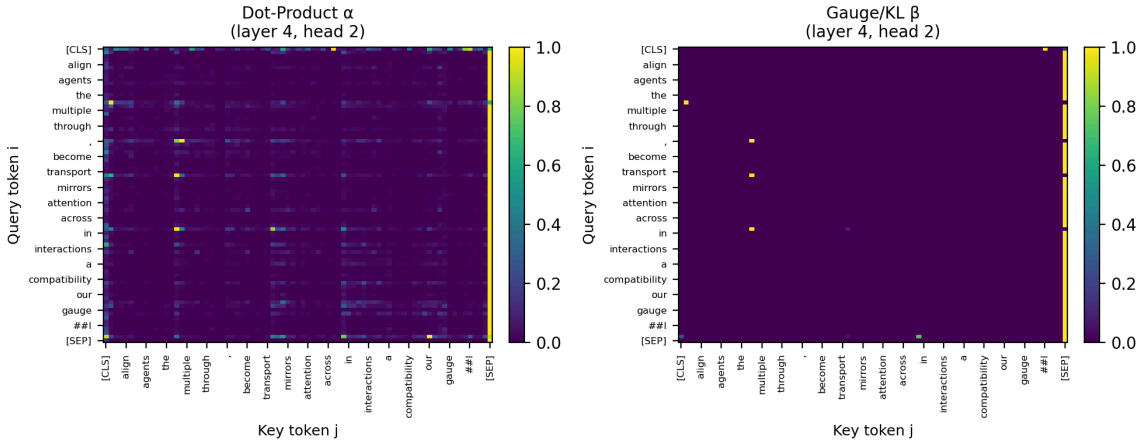
We have shown that attention and transformers are a limiting case of a more general statistical gauge theory where tokens are modeled as agents with certainty of their beliefs (delta-function limit). The attention dot-product is due to an agent-agent "communication" term in a generalized functional variational energy/action. The full framework possesses a vacuum state where all agents flow towards an average belief and gauge frame (embedding) mirroring machine learning without training. We postulate then that agent observations break this symmetry by flowing to unique vector (μ) embeddings (ϕ) and acts as a machine learning loss function.

Furthermore, our framework naturally enables the emergence of higher-scale meta-agents and abstract organizations of agents. In separate studies we have simulated

randomly initialized agents on a two-dimensional grid and have shown that under variational gradient descent meta-agents emerge with cross-scale couplings. Time-scale separation occurs with meta-agents fluctuating on time-scales around $10^4 - 10^6$ times slower than the lower scale agents (where we define time in terms of agent belief updating - i.e. the smallest time scale corresponds to 1 bit of belief updating). .

Our framework then suggests an enticing path towards unifying the variational free energy principle with machine learning architectures by extending the free energy principle to include an agent-agent communication term - attention is all you need.

6. Appendix



6.1. Expectation of $d(k_i, \Omega_{ij} k_j)$

We begin with the definition

$$\mathbb{E}_{q_i q_j}[d(k_i, \Omega_{ij} k_j)] = \tfrac{1}{2}\, \mathbb{E}_{q_i q_j}\left[(\Omega_{ij} k_j - k_i)^\top \Sigma_i^{-1} (\Omega_{ij} k_j - k_i)\right],$$

where $q_i(k_i) = \mathcal{N}(\mu_i, \Sigma_i)$ and $q_j(k_j) = \mathcal{N}(\mu_j, \Sigma_j)$, and the gauge transport $\Omega_{ij} \in SO(3)$ acts linearly on the latent vectors.

For clarity we define

$$\delta := \Omega_{ij} k_j - k_i.$$

Then the product expectation becomes

$$\mathbb{E}_{q_i q_j}[d(k_i, \Omega_{ij} k_j)] = \tfrac{1}{2}\, \mathbb{E}[\delta^\top \Sigma_i^{-1} \delta].$$

For any random vector δ with mean $\bar{\delta} = \mathbb{E}[\delta]$ and covariance $\mathrm{Cov}(\delta)$,

$$\mathbb{E}[\delta^\top A \delta] = \mathrm{tr}(A\, \mathrm{Cov}(\delta)) + \bar{\delta}^\top A \bar{\delta}.$$

Setting $A = \Sigma_i^{-1}$ yields

$$\mathbb{E}_{q_i q_j}[\delta^\top \Sigma_i^{-1} \delta] = \mathrm{tr}(\Sigma_i^{-1}\, \mathrm{Cov}(\delta)) + \bar{\delta}^\top \Sigma_i^{-1} \bar{\delta}.$$

Since k_i and k_j are independent, we have

$$\mathbb{E}[k_i] = \mu_i, \qquad \mathbb{E}[k_j] = \mu_j, \qquad \mathrm{Cov}(k_i) = \Sigma_i, \qquad \mathrm{Cov}(k_j) = \Sigma_j, \qquad \mathrm{Cov}(k_i, k_j) = 0$$

Hence

$$\bar{\delta} = \mathbb{E}[\Omega_{ij} k_j - k_i] = \Omega_{ij}\mu_j - \mu_i,$$

and

$$\mathrm{Cov}(\delta) = \Omega_{ij}\Sigma_j\Omega_{ij}^\top + \Sigma_i.$$

Substituting, we have

$$\begin{aligned}\mathbb{E}_{q_i q_j}[\delta^\top \Sigma_i^{-1} \delta] &= \mathrm{tr}\big(\Sigma_i^{-1}(\Omega_{ij}\Sigma_j\Omega_{ij}^\top + \Sigma_i)\big) + (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i) \\ &= \mathrm{tr}\big(\Sigma_i^{-1}\Omega_{ij}\Sigma_j\Omega_{ij}^\top\big) + \mathrm{tr}(I) + (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i).\end{aligned}$$

Since $\mathrm{tr}(I) = \dim \mathcal{B}_q = K$ we obtain

$$\boxed{\mathbb{E}_{q_i q_j}[d(k_i, \Omega_{ij} k_j)] = \tfrac{1}{2}\big[\mathrm{tr}\big(\Sigma_i^{-1}\Omega_{ij}\Sigma_j\Omega_{ij}^\top\big) + K + (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i)\big].}$$

For comparison, the well known Gaussian KL divergence between $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ and $\Omega_{ij} q_j = \mathcal{N}(\Omega_{ij}\mu_j, \Omega_{ij}\Sigma_j\Omega_{ij}^\top)$ is

$$\begin{aligned}D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) = \tfrac{1}{2}\Big[&\mathrm{tr}\big((\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\Sigma_i\big) - K + \ln\frac{|\Omega_{ij}\Sigma_j\Omega_{ij}^\top|}{|\Sigma_i|} \\ &+ (\Omega_{ij}\mu_j - \mu_i)^\top (\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}(\Omega_{ij}\mu_j - \mu_i)\Big].\end{aligned}$$

When the covariances are nearly aligned, $\Sigma_i \approx \Omega_{ij}\Sigma_j\Omega_{ij}^\top$, then the log-determinant terms are small (higher order) and the two expressions coincide modulo constants.

Thus, to second order,

$$\boxed{\mathbb{E}_{q_i q_j}[d(k_i, \Omega_{ij} k_j)] \;\simeq\; D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) \;+\; \mathrm{const.}}$$

Alternatively, this is equivalent to the KL divergence between nearby distributions. For example, q_θ and $q_{\theta+\delta\theta}$ can be expanded as

$$D_{\mathrm{KL}}(q_\theta \,\|\, q_{\theta+\delta\theta}) = \tfrac{1}{2}\,\delta\theta^\top I_F(\theta)\,\delta\theta + \mathcal{O}(\|\delta\theta\|^3),$$

where $I_F(\theta)$ is the Fisher information matrix.

For Gaussian families, the Fisher metric induces the Mahalanobis form

$$\delta s^2 = (\delta\mu)^\top \Sigma^{-1}\delta\mu + \tfrac{1}{2}\,\mathrm{tr}\big[(\Sigma^{-1}\delta\Sigma)^2\big].$$

Hence the Mahalanobis distance under expectation reproduces the local quadratic (Fisher-metric) form of the KL divergence between Gaussian fiber distributions.

6.2. Generalized Variational Energy

Our fully general variational energy can be written as

$$
\begin{aligned}
S = & \underbrace{\sum_i \int_{\mathcal{C}} \chi_i(c) \left[\alpha_i \, \mathrm{KL}\big(q_i(c) \,\|\, p_i(c)\big) - \gamma_i \, \mathbb{E}_{q_i(c)}\big[\log p_i(o_i \,|\, c)\big] \right] \mathrm{d}c}_{\text{(1) Self / local variational free energy term}} \\
& + \underbrace{\sum_{i,j} \int_{\mathcal{C}} \chi_i(c) \chi_j(c) \, \beta_{ij}(c) \, \mathrm{KL}\big(q_i(c) \,\|\, \Omega_{ij}^{(q)} q_j(c)\big) \, \mathrm{d}c}_{\text{(2) Cross-agent alignment / attention coupling}} \\
& + \underbrace{\lambda_A \int_{\mathcal{C}} \tfrac{1}{2} \sum_{a<b} \big\| F_{ab}(c) \big\|^2 \, \mathrm{d}c}_{\text{(3) Gauge curvature energy (global field consistency)}} \\
& + \underbrace{\lambda_\phi \sum_i \int_{\mathcal{C}} \chi_i(c) \, \tfrac{1}{2} \, \big\| D\phi_i(c) \big\|^2 \, \mathrm{d}c}_{\text{(4) Gauge-fixing / } \phi\text{-smoothness regularization)}} \\
& + (\text{higher-order regularizers: Fisher, mass, or hierarchical couplings}).
\end{aligned}
$$

where

$$
\begin{aligned}
F_{ab} &= \partial_a A_b - \partial_b A_a + [A_a, A_b], && (\text{curvature of the global gauge connection}), \\
D\phi_i &= \nabla \phi_i - A \cdot \phi_i, && (\text{covariant derivative of local frame } \phi_i).
\end{aligned}
$$

In principle, however, cross-scale couplings may emerge under mate-agent renormalization flow and agent-meta-agent coalescence. We relegate these exciting topics to future study.

6.3. Deriving the Pairwise Potential ψ_{ij} from a Normalized Generative Prior

We construct an explicit generative model whose marginal prior over agent latents produces the desired pairwise compatibility factor

$$
\psi_{ij}(k_i, k_j) \;\propto\; \exp\Big[-\lambda_{ij} \, d\big(k_i, \Omega_{ij} k_j\big) \Big],
$$

where $d(\cdot,\cdot)$ will become a Mahalanobis / information-geometric distance measured in agent i's frame.

Let each agent i have a latent state $k_i \in \mathbb{R}^d$ (this lives in the fiber B_q at some base point $c \in C$).
We assume a base (local) prior $p_i(k_i)$ for each agent. For every ordered pair (i,j) we introduce an auxiliary \emph{agreement variable} $z_{ij} \in \mathbb{R}^d$, interpreted as ``what agent i believes agent j looks like, expressed in agent i's gauge frame.'' The gauge transport $\Omega_{ij} \in G \subset \mathrm{GL}(d)$ maps agent j's latent k_j into the frame of agent i.

We define the following joint generative model over $\{k_i\}$ and $\{z_{ij}\}$:

$$p(\{k_i\}, \{z_{ij}\}) = \left[\prod_i p_i(k_i)\right] \left[\prod_{i,j} \mathcal{N}\big(z_{ij}\,;\, k_i, \Lambda_{ij}^{-1}\big)\, \mathcal{N}\big(z_{ij}\,;\, \Omega_{ij} k_j, \Lambda_{ij}^{-1}\big)\right]. \quad (11)$$

Here Λ_{ij} is a (symmetric positive definite) precision matrix that controls how strongly i and j are required to ``agree.''

Intuitively:

- z_{ij} is required to be close to k_i
  (``what I, agent i, think you should look like if you match me''), and
- z_{ij} is simultaneously required to be close to
  $\Omega_{ij} k_j$
  (``what you, agent j, actually look like after I parallel-transport
  you into my frame via Ω_{ij}'').

Because the right-hand side is a product of Gaussians, $p(\{k_i\}, \{z_{ij}\})$ is normalizable.

6.3.1. Marginalizing the agreement variables.

We now integrate out the $\{z_{ij}\}$ to obtain the induced prior over $\{k_i\}$ alone:

$$p(\{k_i\}) = \int \left[\prod_{i,j} dz_{ij}\right] p(\{k_i\}, \{z_{ij}\}). \quad (12)$$

Because the integrand factorizes over (i, j), it suffices to evaluate

$$I_{ij}(k_i, k_j) := \int dz_{ij}\, \mathcal{N}\big(z_{ij}\,;\, k_i, \Lambda_{ij}^{-1}\big)\, \mathcal{N}\big(z_{ij}\,;\, \Omega_{ij} k_j, \Lambda_{ij}^{-1}\big). \quad (13)$$

We use the standard ``product of Gaussians'' identity for equal precision.

Let $m_1, m_2 \in \mathbb{R}^d$ and $\Lambda \succ 0$.

Then

$$\mathcal{N}(z; m_1, \Lambda^{-1})\, \mathcal{N}(z; m_2, \Lambda^{-1}) = \mathcal{N}\Big(m_1; m_2, (2\Lambda)^{-1}\Big)\, \mathcal{N}\Big(z; \tfrac{1}{2}(m_1 + m_2), (\tfrac{1}{2}\Lambda)^{-1}\Big).$$

Integrating over z removes the second Gaussian and leaves

$$\int dz\, \mathcal{N}(z; m_1, \Lambda^{-1})\, \mathcal{N}(z; m_2, \Lambda^{-1}) = \mathcal{N}\Big(m_1; m_2, (2\Lambda)^{-1}\Big). \quad (15)$$

We apply (15) with
$m_1 = k_i$, $m_2 = \Omega_{ij} k_j$, and $\Lambda = \Lambda_{ij}$.

Thus (13) becomes

$$I_{ij}(k_i, k_j) = \mathcal{N}\Big(k_i\,;\, \Omega_{ij} k_j, (2\Lambda_{ij})^{-1}\Big). \quad (16)$$

Writing out the Gaussian explicitly,

$$\mathcal{N}\Big(k_i\,;\,\Omega_{ij}k_j,(2\Lambda_{ij})^{-1}\Big) \propto \exp\Big[-\tfrac{1}{2}\big(k_i-\Omega_{ij}k_j\big)^\top(2\Lambda_{ij})\big(k_i-\Omega_{ij}k_j\big)\Big] \quad (17)$$

$$= \exp\Big[-\big(k_i-\Omega_{ij}k_j\big)^\top\Lambda_{ij}\big(k_i-\Omega_{ij}k_j\big)\Big]. \quad (18)$$

Because Λ_{ij} is symmetric, the quadratic form is unchanged if we flip the sign inside the parentheses:

$$\big(k_i-\Omega_{ij}k_j\big)^\top\Lambda_{ij}\big(k_i-\Omega_{ij}k_j\big) = \big(\Omega_{ij}k_j-k_i\big)^\top\Lambda_{ij}\big(\Omega_{ij}k_j-k_i\big).$$

For interpretability we will write it as $(\Omega_{ij}k_j-k_i)$, i.e.\ ``you, in my frame, minus me.''

Define the pairwise distance-like function

$$d\big(k_i,\Omega_{ij}k_j\big) := \big(\Omega_{ij}k_j-k_i\big)^\top\Lambda_{ij}\big(\Omega_{ij}k_j-k_i\big). \quad (19)$$

Then (16)--(18) yield

$$I_{ij}(k_i,k_j) \propto \exp\Big[-d\big(k_i,\Omega_{ij}k_j\big)\Big]. \quad (20)$$

Substituting (20) into (12), we obtain the marginalized prior over $\{k_i\}$:

$$p(\{k_i\}) \propto \Big[\prod_i p_i(k_i)\Big]\ \prod_{i,j}\exp\Big[-d\big(k_i,\Omega_{ij}k_j\big)\Big]. \quad (21)$$

We now \emph{define} the pairwise potential

$$\psi_{ij}(k_i,k_j) := I_{ij}(k_i,k_j) = \int dz_{ij}\,\mathcal{N}\big(z_{ij};k_i,\Lambda_{ij}^{-1}\big)\,\mathcal{N}\big(z_{ij};\Omega_{ij}k_j,\Lambda_{ij}^{-1}\big) \quad (22)$$

$$\propto \exp\Big[-d\big(k_i,\Omega_{ij}k_j\big)\Big], \quad (23)$$

which is exactly of the desired exponential form with

$$d\big(k_i,\Omega_{ij}k_j\big) = (\Omega_{ij}k_j-k_i)^\top\Lambda_{ij}(\Omega_{ij}k_j-k_i).$$

Thus the normalized generative prior (11)--(21) induces---after marginalizing the agreement variables $\{z_{ij}\}$---a Markov random field over $\{k_i\}$ with pairwise potentials ψ_{ij} that penalize epistemic disagreement between agent i's latent k_i and agent j's latent k_j transported into i's gauge frame via Ω_{ij}.

6.3.2. Connection to KL divergence.

For Gaussian approximate posteriors $q_i(k_i)=\mathcal{N}(k_i;\mu_i,\Sigma_i)$ and $q_j(k_j)=\mathcal{N}(k_j;\mu_j,\Sigma_j)$, the variational free energy contains the term

$$\mathbb{E}_{q_iq_j}\big[d(k_i,\Omega_{ij}k_j)\big] = \mathbb{E}_{q_iq_j}\Big[(\Omega_{ij}k_j-k_i)^\top\Lambda_{ij}(\Omega_{ij}k_j-k_i)\Big].$$

As shown in Appendix~6.1, for aligned covariances $\Sigma_i \approx \Omega_{ij}\Sigma_j\Omega_{ij}^\top$ this expectation matches (up to additive constants and higher-order terms) the KL divergence $\mathrm{KL}\big(q_i \,\|\, \Omega_{ij}q_j\big)$, where $\Omega_{ij}q_j$ denotes the Gaussian obtained by transporting q_j into

agent i's frame via Ω_{ij}. This links the pairwise potential ψ_{ij} to the KL alignment penalty used in the main energy functional.

7. D_{KL}

7.1. Local Equivalence Between Expected Mahalanobis Disagreement and KL Divergence

In this section we prove (up to higher-order terms) that the expected quadratic disagreement penalty used in our pairwise coupling matches the KL divergence between agent beliefs after gauge transport. This validates the replacement of our energy term $\mathbb{E}[d(\cdot,\cdot)]$ by a KL alignment term in the variational free energy.

7.2. Setup.

Each agent i maintains a Gaussian belief over its local latent state $k_i \in \mathbb{R}^d$,

$$q_i(k_i) = \mathcal{N}(k_i\,;\,\mu_i, \Sigma_i), \tag{24}$$

and similarly for agent j,

$$q_j(k_j) = \mathcal{N}(k_j\,;\,\mu_j, \Sigma_j). \tag{25}$$

Gauge transport from agent j's frame to agent i's frame is given by a (group-)linear map $\Omega_{ij} \in G \subset \mathrm{GL}(d)$, which induces the transported belief

$$\Omega_{ij} q_j := \mathcal{N}\big(k\,;\,\Omega_{ij}\mu_j, \Omega_{ij}\Sigma_j\Omega_{ij}^{\top}\big). \tag{26}$$

We denote the transported covariance by

$$\tilde{\Sigma}_j := \Omega_{ij}\Sigma_j\Omega_{ij}^{\top}. \tag{27}$$

7.3. Quadratic disagreement energy.

From the generative construction in Sec.~(3.1)--(3.2), integrating out latent ``agreement'' variables z_{ij} yields an effective pairwise potential of the form

$$\psi_{ij}(k_i, k_j) \;\propto\; \exp\Big[-\,d\big(k_i, \Omega_{ij}k_j\big)\Big], \tag{28}$$

where the disagreement function $d(\cdot,\cdot)$ is a Mahalanobis distance measured in agent i's precision metric:

$$d\big(k_i, \Omega_{ij}k_j\big) := \big(\Omega_{ij}k_j - k_i\big)^{\top}\Lambda_{ij}\big(\Omega_{ij}k_j - k_i\big), \tag{29}$$

with $\Lambda_{ij} \succ 0$ a precision matrix. In the natural choice relevant for information geometry we set $\Lambda_{ij} = \Sigma_i^{-1}$, i.e.\ agent i evaluates disagreement in its own uncertainty metric.

When we form the variational free energy, this pairwise potential contributes the expectation

$$\mathbb{E}_{q_i q_j}\big[d(k_i, \Omega_{ij}k_j)\big] = \mathbb{E}_{k_i\sim q_i,\; k_j\sim q_j}\left[(\Omega_{ij}k_j - k_i)^{\top}\Sigma_i^{-1}(\Omega_{ij}k_j - k_i)\right]. \tag{30}$$

7.4. Exact expansion of the expectation.

Define the random vector

$$\delta := \Omega_{ij} k_j - k_i. \tag{31}$$

Then (30) becomes $\mathbb{E}[\delta^\top \Sigma_i^{-1} \delta]$. For any symmetric matrix A and random vector δ,

$$\mathbb{E}[\delta^\top A \delta] = \operatorname{tr}\big(A \operatorname{Cov}(\delta)\big) + \bar{\delta}^\top A \, \bar{\delta}, \tag{32}$$

where $\bar{\delta} = \mathbb{E}[\delta]$.

Since k_i and k_j are independent under $q_i q_j$, we have

$$\bar{\delta} = \mathbb{E}[\Omega_{ij} k_j - k_i] = \Omega_{ij} \mu_j - \mu_i, \tag{33}$$

and

$$\operatorname{Cov}(\delta) = \Omega_{ij} \Sigma_j \Omega_{ij}^\top + \Sigma_i = \tilde{\Sigma}_j + \Sigma_i. \tag{34}$$

Substituting these into (30) with $A = \Sigma_i^{-1}$ gives

$$\begin{aligned}
\mathbb{E}_{q_i q_j}[d] &= \operatorname{tr}\big(\Sigma_i^{-1}(\tilde{\Sigma}_j + \Sigma_i)\big) + (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1} (\Omega_{ij}\mu_j - \mu_i) && (35)\\
&= \operatorname{tr}\big(\Sigma_i^{-1}\tilde{\Sigma}_j\big) + \operatorname{tr}(\Sigma_i^{-1}\Sigma_i) + (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1} (\Omega_{ij}\mu_j - \mu_i) && (36)\\
&= \operatorname{tr}\big(\Sigma_i^{-1}\tilde{\Sigma}_j\big) + d + (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1} (\Omega_{ij}\mu_j - \mu_i), && (37)
\end{aligned}$$

where $d = \dim(k_i)$. Equation~(37) is \emph{exact}.

7.5. KL divergence between transported beliefs.

The KL divergence between two Gaussians

$q_i = \mathcal{N}(\mu_i, \Sigma_i)$ and $\Omega_{ij} q_j = \mathcal{N}(\Omega_{ij}\mu_j, \tilde{\Sigma}_j)$ is

$$\mathrm{KL}\big(q_i \,\|\, \Omega_{ij} q_j\big) = \frac{1}{2}\Big[\operatorname{tr}\big(\tilde{\Sigma}_j^{-1}\Sigma_i\big) - d + \log \frac{\det \tilde{\Sigma}_j}{\det \Sigma_i} + (\Omega_{ij}\mu_j - \mu_i)^\top \tilde{\Sigma}_j^{-1} (\Omega_{ij}\mu_j - \mu$$

Comparing (37) and (38), we immediately see two structural differences:

(i) Σ_i^{-1} vs.\ $\tilde{\Sigma}_j^{-1}$, and
(ii) the additional log-determinant term
$\log(\det \tilde{\Sigma}_j / \det \Sigma_i)$ in the KL.

We now show that these differences vanish (or become higher order) in the natural alignment regime where the cross-agent coupling drives the two beliefs to agree after gauge transport.

7.6. Local alignment regime.

Define the transported covariance mismatch

$$\Delta := \tilde{\Sigma}_j - \Sigma_i. \tag{39}$$

We assume (and later enforce dynamically) that agents align, i.e.
$\|\Sigma_i^{-1/2}\Delta\Sigma_i^{-1/2}\| \ll 1$.

Equivalently,

$$\tilde{\Sigma}_j = \Sigma_i + \Delta, \qquad \|\Delta\| \text{ is small in the SPD metric.} \tag{40}$$

This assumption is natural in our model: the whole purpose of the coupling term is to drive agents toward $\mu_i \approx \Omega_{ij}\mu_j$ and $\Sigma_i \approx \tilde{\Sigma}_j$.

In this small-Δ regime, we use matrix Taylor expansions.

7.6.1. Inverse expansion.

By the Neumann series,

$$\tilde{\Sigma}_j^{-1} = (\Sigma_i + \Delta)^{-1} = \Sigma_i^{-1} - \Sigma_i^{-1}\Delta\Sigma_i^{-1} + O(\|\Delta\|^2). \tag{41}$$

7.6.2. Log-determinant expansion.

Using

$$\log\det(\Sigma_i + \Delta) = \log\det\Sigma_i + \operatorname{tr}(\Sigma_i^{-1}\Delta) - \tfrac{1}{2}\operatorname{tr}\big((\Sigma_i^{-1}\Delta)^2\big) + O(\|\Delta\|^3)$$

,

we obtain

$$\log\frac{\det\tilde{\Sigma}_j}{\det\Sigma_i} = \operatorname{tr}(\Sigma_i^{-1}\Delta) - \frac{1}{2}\operatorname{tr}\big((\Sigma_i^{-1}\Delta)^2\big) + O(\|\Delta\|^3). \tag{42}$$

7.6.3. KL to second order.

Substitute (41) and (42)
into (38) and keep terms up to second order in Δ.

First, the quadratic mean-mismatch term becomes

$$(\Omega_{ij}\mu_j - \mu_i)^\top\tilde{\Sigma}_j^{-1}(\Omega_{ij}\mu_j - \mu_i) = (\Omega_{ij}\mu_j - \mu_i)^\top\Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i)$$
$$-(\Omega_{ij}\mu_j - \mu_i)^\top\Sigma_i^{-1}\Delta\Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i) + O(\|\Delta$$

Next, for the trace term,

$$\operatorname{tr}(\tilde{\Sigma}_j^{-1}\Sigma_i) = \operatorname{tr}\Big[(\Sigma_i^{-1} - \Sigma_i^{-1}\Delta\Sigma_i^{-1} + O(\|\Delta\|^2))\Sigma_i\Big] \tag{45}$$
$$= \operatorname{tr}\big(I - \Sigma_i^{-1}\Delta + O(\|\Delta\|^2)\big) = d - \operatorname{tr}(\Sigma_i^{-1}\Delta) + O(\|\Delta\|^2). \tag{46}$$

Now substitute (44), (46), and (42) into (38):

$$\mathrm{KL}(q_i \,\|\, \Omega_{ij} q_j) = \frac{1}{2}\Big[\Big(d - \mathrm{tr}(\Sigma_i^{-1}\Delta) + O(\|\Delta\|^2)\Big) - d + \Big(\mathrm{tr}(\Sigma_i^{-1}\Delta) - \tfrac{1}{2}\mathrm{tr}((\Sigma_i^{-1}$$
$$+(\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i)$$
$$-(\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}\Delta\Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i) + O(\|\Delta\|^2)\Big].$$

Observe that the linear $\mathrm{tr}(\Sigma_i^{-1}\Delta)$ terms cancel.

Dropping terms $O(\|\Delta\|^2)$ and higher, we obtain the leading-order approximation

$$\mathrm{KL}(q_i \,\|\, \Omega_{ij} q_j) = \frac{1}{2}(\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i) + O(\|\Delta\|^2). \quad (50)$$

In words: to second order in the covariance mismatch Δ, the KL divergence is dominated by the Mahalanobis norm of the mean difference, measured in agent i's precision Σ_i^{-1}.

7.7. Compare with the expected disagreement energy.

Rewrite (37) by expanding

$$\mathrm{tr}(\Sigma_i^{-1}\tilde{\Sigma}_j) = \mathrm{tr}(\Sigma_i^{-1}(\Sigma_i + \Delta)) = d + \mathrm{tr}(\Sigma_i^{-1}\Delta)$$

$$\mathbb{E}_{q_i q_j}[d] = (\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1}(\Omega_{ij}\mu_j - \mu_i) + \mathrm{tr}(\Sigma_i^{-1}\Delta) + \mathrm{const}, \quad (51)$$

where the constant absorbs the two d terms.

Comparing (50) and (51) shows:

$$\mathbb{E}_{q_i q_j}[d] = 2\,\mathrm{KL}(q_i \,\|\, \Omega_{ij} q_j) + \mathrm{tr}(\Sigma_i^{-1}\Delta) + O(\|\Delta\|^2) + \mathrm{const}. \quad (52)$$

The first term (twice the KL) captures the dominant mean-alignment cost. The second term $\mathrm{tr}(\Sigma_i^{-1}\Delta)$ comes from the mismatch in covariances and is precisely the same ``curvature'' content that, in the KL, is paired with the $\log\det$ term and thus cancels at first order.

The remainder is higher order in Δ plus an additive constant (which does not affect gradients).

7.8. Interpretation.

Equation~(52) justifies our use of $\mathbb{E}[d(k_i, \Omega_{ij}k_j)]$ as an effective KL-alignment penalty in the energy functional: in the alignment regime $\Delta \to 0$, driven by the same coupling, $\mathbb{E}[d]$ and $2\,\mathrm{KL}$ differ only by an additive constant and terms that are $O(\|\Delta\|^2)$ in the local SPD metric on covariances.

Geometrically, this is the standard Fisher-information result from information geometry: for two nearby distributions on a statistical manifold, the KL divergence behaves as a squared geodesic distance to second order.

Here, the gauge transport Ω_{ij} first brings agent j's belief into agent i's local frame, and the quadratic form
$(\Omega_{ij}\mu_j - \mu_i)^\top \Sigma_i^{-1} (\Omega_{ij}\mu_j - \mu_i)$ is exactly that local squared distance in i's precision metric, up to higher-order curvature corrections encoded in Δ.

7.8.1. Conclusion.

We have shown that, after gauge transport, the expected Mahalanobis disagreement $\mathbb{E}[d]$ generated by our \emph{normalized} pairwise prior differs from $2\,\mathrm{KL}(q_i \| \Omega_{ij} q_j)$ only by (i) an additive constant and (ii) $O(\|\Delta\|^2)$ curvature terms that vanish in the intended alignment regime. This provides a principled bridge from the generative graphical model to the KL-based coupling term used in our variational free energy.

References

[1] Friston, K. (2010). *The free-energy principle: a unified brain theory?* **Nature Reviews Neuroscience**, 11(2), 127–138.

[2] Friston, K., Parr, T., & de Vries, B. (2017). *The graphical brain: belief propagation and active inference.* **Network Neuroscience**, 1(4), 381–414.

[3] Parr, T., Pezzulo, G., & Friston, K. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior.* MIT Press.

[4] Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. (2019). A *tale of two densities: active inference is enactive inference.* **Adaptive Behavior**, 27(6), 369–385.

[5] Friston, K., Sajid, N., et al. (2021). *Deep temporal models and active inference.* **Neuroscience & Biobehavioral Reviews**, 128, 279–295.

[6] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is All You Need.* In *Advances in Neural Information Processing Systems (NeurIPS).*

[7] Bahdanau, D., Cho, K., & Bengio, Y. (2015). *Neural Machine Translation by Jointly Learning to Align and Translate.* In *International Conference on Learning Representations (ICLR).*

[8] Dosovitskiy, A., et al. (2021). *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.* In *International Conference on Learning Representations (ICLR).*

[9] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What Does BERT Look at? An Analysis of BERT's Attention.* In *BlackBoxNLP Workshop.*

[10] Amari, S. (2016). *Information Geometry and Its Applications.* Springer.

[11] Amari, S. (1985). *Differential-Geometrical Methods in Statistics.* Springer.

[12] Nakahara, M. (2003). *Geometry, Topology and Physics* (2nd ed.). CRC Press.

[13] Frankel, T. (2012). *The Geometry of Physics: An Introduction* (3rd ed.). Cambridge University Press.

[14] Yang, C. N., & Mills, R. L. (1954). *Conservation of Isotopic Spin and Isotopic Gauge Invariance*. **Physical Review**, 96(1), 191–195.

[15] Clark, A. (2013). *Whatever next? Predictive brains, situated agents, and the future of cognitive science*. **Behavioral and Brain Sciences**, 36(3), 181–204.

[16] Fuchs, C. A., & Schack, R. (2013). *Quantum-Bayesian coherence*. **Reviews of Modern Physics**, 85(4), 1693–1715.

[17] Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). *Geometric deep learning: Grids, groups, graphs, geodesics, and gauges*. **Nature**, 590, 197–205.