

A Gauge Theory Foundation of Attention

RC Dennis

Abstract

We present a unified gauge-theoretic formulation of attention and message passing in Transformer-like architectures and connect this as a bridge to Friston's free energy principle in neuroscience and cognition.

In our framework, each token (agent) is modeled as a local section of an associated bundle over a base manifold \mathcal{C} , carrying an exponential family belief field (specifically we consider multi-variate Gaussians $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ whose mean and covariance define local epistemic states and metrics).

Inter-token communication is mediated by a non-abelian gauge connection A with parallel-transport operators Ω_{ij} (which we consider in $SO(3)$, representing orientation changes between local frames).

Attention weights arise naturally as normalized exponentials of information-geometric divergences, $\beta_{ij} \propto \text{softmax}_j(-\text{KL}(q_i \parallel \Omega_{ij} q_j))$, so that belief alignment, rather than ad hoc dot products, governs interaction.

In the isotropic, flat-bundle, delta-function limit (where all covariances are scalar and local frames coincide) this expression reduces exactly to the canonical Transformer rule $\beta_{ij} \propto \text{softmax}(Q_i K_j^\top)$, with $Q_i = \mu_i^T e^{\phi_i}$ and $K_j = \mu_j^T e^{-\phi_j}$ (where ϕ_i are the gauge frames).

Thus standard attention emerges as the trivial-connection, constant-metric limit of a more general gauge-covariant update law.

Curvature of the global connection $F = dA + A \wedge A$ measures the path dependence of multi-hop attentions, providing a natural notion of holonomy and emergent higher-order structure.

Finally, we show that statistical observations correspond to a symmetry breaking term allowing agents/tokens to flow towards unique embedding vectors. In this view an "attention" term, $\beta_{ij} \text{KL}(q_i \parallel \Omega_{ij} q_j)$ is postulated as a variational functional-energy term to be minimized under gradient descent. If this can be made rigorous then this model will realize the unification of the free energy principle and methods in machine learning. Interestingly, much as Friston's FEP is analogous to the Helmholtz thermodynamic Potential our frame work is analogous to the Grand Potential.

Our construction fuses information geometry, non-abelian gauge theory, and Transformer methods into a single mathematical formalism, opening a path toward curvature-aware, uncertainty-sensitive neural architectures grounded in geometric and variational principles.

1. Bridging the Free Energy Principle and Machine Learning

We show that attention is a limiting case of a more general geometric framework. Due to this we speculate that a bridge exists between Friston's free energy principle and the methods of deep learning and machine learning architectures. If such a bridge exists then machine learning may be able to leverage gauge theoretic methods to advance deep learning architectures and similarly new insights into cognition and neuroscience could possibly be advanced.

Here we propose a geometric framework which seeks to bridge this gap.

1.1. The Mathematical Model

We introduce a general formal model underlying the transformer architecture. Our model is geometrically and mathematically rich allowing hierarchical emergence of meta-agents and evolution.

We model each agent as a local section of an associated bundle to a principal fiber bundle whose base space \mathcal{C} .

Beliefs and models are encoded in the associated bundle constructed from a structure group G acting on the fiber (\mathcal{B}_q) . These structures allow us to formalize both intra-agent inference and inter-agent communication using gauge-theoretic transport.

specifically, let $\pi : \mathcal{N} \rightarrow \mathcal{C}$ be a smooth principal G -bundle where \mathcal{C} is a smooth manifold, G is a Lie group acting freely and transitively on the right on \mathcal{N} .

The projection satisfies $\pi(n \cdot g) = \pi(n)$ for all $g \in G, n \in \mathcal{N}$.

Let $\rho : G \rightarrow \text{Aut}(\mathcal{B}_q)$ be a representation of the Lie group G on a smooth manifold \mathcal{B}_q . Depending on context, \mathcal{B}_q may be modeled as a K -dimensional probability simplex Δ^K (e.g., for categorical distributions) or as a statistical manifold equipped with a suitable information geometry.

Importantly, the relevant geometric structures — such as divergence measures, metrics, and connections — remain well-defined on these fibers even when they lack linear structure (see e.g. Amari's information geometry for examples of dually flat but non-linear statistical manifolds).

Next, the associated bundle is defined as:

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

where

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

This gives a fiber bundle $\pi_{\mathcal{E}_u} : \mathcal{E}_u \rightarrow \mathcal{C}$ with fiber \mathcal{B}_q . Generally speaking the

Definition:

An agent is a local section of the associated bundle over \mathcal{C}

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

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

where \mathcal{U}_i are open local subsets of \mathcal{C} .

Definition:

A multi-agent (\mathcal{M}) over \mathcal{C} is a tuple of agents (where \mathcal{I} is an index set)

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

Definition:

A meta-agent is a multi-agent whose agents share identical generative models. Furthermore, we say that a set of agents is "epistemically dead" if they identically share both beliefs and models. Note, importantly that even if the agents composing a meta-agent are epistemically dead this does NOT imply the meta-agent itself is epistemically dead. In fact, this is a feature of our framework - epistemically dead agents can be integrated out!

Next, in the standard way, via horizontal lifting from \mathcal{N} to \mathcal{E}_q we generate a variety of morphisms and induced connections across scales (i, j) . In particular parallel transport and cross-scale transport operators:

1. $\Omega^s : \Gamma^s(\mathcal{B}_q) \rightarrow \Gamma^s(\mathcal{B}_q)$
2. $\Lambda_{s'}^s : \Gamma^s(\mathcal{B}_q) \rightarrow \Gamma^{s'}(\mathcal{B}_q)$

where $\Gamma(\mathcal{B}_q)$ denotes the space of smooth sections of \mathcal{B}_i over the relevant open subset of \mathcal{C} . Additionally Λ_j^i represents the parallel transport operator between agent at scale s to s' .

1.2. Types of Parallel Transport in Epistemic Geometry

In our framework, parallel transport arises in several distinct but interconnected settings, depending on whether the transport occurs along the base manifold \mathcal{C} or across agent frames. We distinguish the following cases:

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

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

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

This operator maps fiber elements between different base points:

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

This is the canonical notion of parallel transport along the base, lifted via the connection to fibers.

1.2.2. 2. Vertical Transport Within a Single Fiber \mathcal{B}_c

At a fixed point $c \in \mathcal{C}$, the fiber \mathcal{B}_c is a manifold in its own right. In our setting, this fiber typically represents a statistical manifold (e.g., probability simplex Δ_K , or an exponential family), equipped with an intrinsic Riemannian or dual-affine geometry.

Parallel transport within \mathcal{B}_c may occur along a curve $\eta(\tau) \subset \mathcal{B}_c$ with tangent vector $\dot{\eta}(\tau)$ governed by a connection ∇ intrinsic to \mathcal{B} :

$$\nabla_{\dot{\eta}} V = 0, \quad \text{for parallel vector field } V(\tau).$$

In the information geometry setting (Amari), \mathcal{B}_c may carry a pair of dual connections $(\nabla^{(e)}, \nabla^{(m)})$ associated with exponential and mixture families. Curvature in this fiber space is defined via the Riemann tensor $R^{\mathcal{B}}$:

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

Thus, even at a single point $c \in \mathcal{C}$, the fiber \mathcal{B}_c may be a curved internal space with nontrivial geodesics and holonomies. These paths represent purely epistemic transformations of beliefs or models that do not involve movement along \mathcal{C} .

1.2.3. 3. Intra-Agent Spatial Transport (Within Fiber Bundle)

To compare beliefs $q(c_1)$ and $q(c_2)$ held by the same agent over two nearby points $c_1, c_2 \in \mathcal{U}_i \subset \mathcal{C}$, we apply the agent's gauge frame $\phi_i(c)$ to define a connection $A_\mu^{(i)} = \partial_\mu \phi_i(c)$ and use:

$$T_{c_1 \rightarrow c_2}^{(i)} := \mathcal{P} \exp \left(- \int_{c_1}^{c_2} A_\mu^{(i)} dc^\mu \right). \quad (1)$$

This defines gauge-covariant parallel transport between fibers over space, for a fixed agent.

1.2.4. 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 its own gauge frame $\phi_i(c), \phi_j(c) \in \mathfrak{g}$. The inter-agent gauge transformation is given by:

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

and transports beliefs or models from agent j 's frame to agent i 's:

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

This is essential for defining KL alignment between agents:

$$D_{\text{KL}} [q_i(c) \parallel \Omega_{ij}(c) \cdot q_j(c)].$$

1.2.5. 5. Composite Transport and Holonomy

More generally, a belief may be transported along a path $\gamma = [(i_1, c_1) \rightarrow \dots \rightarrow (i_n, c_n)]$ consisting of both spatial transport within agents and inter-agent frame shifts. The total parallel transport operator is:

$$P_\gamma = T_{c_1 \leftarrow c_n}^{(i_n)} \cdot \Omega_{i_n i_{n-1}} \cdots \Omega_{i_2 i_1} \cdot T_{c_2 \leftarrow c_1}^{(i_1)},$$

This defines the holonomy of belief transport:

$$\text{Hol}_\gamma(q) := P_\gamma \cdot q.$$

If $\text{Hol}_\gamma(q) \neq \text{Id}$, the loop encloses nontrivial epistemic curvature.

1.2.6. Summary

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

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

2. Transformer Attention as a Special Case of Gauge-Covariant Probabilistic Message Passing

In this section we show that standard Transformer self-attention is recovered as a δ -function limiting case of our gauge-covariant, uncertainty-aware agent-agent coupling in our generalized variational energy framework where, in this view, tokens are generally "fuzzy" embeddings (vector μ and Σ ellipse under gauge frame ϕ). In the δ -function limit this becomes the standard token embedding.

In particular, we apply the following to our framework:

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

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

and

(iii) isotropic uncertainty (identical spherical covariance for each agent).

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.

2.1. Setup: agents as Gaussian beliefs in local frames

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

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

where $\mu_i \in \mathbb{R}^d$ is the agent's mean representation in its local frame, and $\Sigma_i \in \mathbb{R}^{d \times d}$ 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.

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

That is, in our variational energy function we have the term

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

where β_{ij} is given by

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

where

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

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

2.2. Assumptions for the Transformer limit

We now impose the three simplifying assumptions.

2.2.1. 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.

2.2.2. 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 ."

2.2.3. 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

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

To make contact with the simplest Transformer form, we additionally absorb $\Omega \Omega^\top$ into a learned rescaling of Ω and into 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

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

Note: we consider $SO(3)$ 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 cant 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!

2.2.4. The compatibility score becomes a dot product

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

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

Note: the trace (canceled by the dimension term) and log det terms vanish.

Next, we then have

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

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 the above expression, the term $-\frac{1}{2\sigma^2} \|\mu_i\|^2$ and the additive constant do not depend on j and therefore cancel under softmax.

Then, 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) :

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

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 $\Omega_{ij} = e^{\phi_i} e^{-\phi_j} = AB^T$

Next, we define the query and key vectors as

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

Then

$$Q_i K_j^\top = \mu_i^\top A B^\top \mu_j = \frac{1}{\sigma^2} \mu_i^\top \Omega \mu_j.$$

Therefore 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 $Q_i K_j^\top$.

The remaining term $-\frac{1}{2\sigma^2} \|\Omega \mu_j\|^2$ depends only on j .

This acts as a key-dependent bias. Such additive, key-specific biases do not change the functional form of attention: they simply shift each column of the attention score matrix before the row-wise softmax. Standard Transformers typically omit this bias, but including or omitting it does not alter the bilinear dot-product structure that defines attention.

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

$$\tilde{s}_{ij} \approx Q_i K_j^\top.$$

Consequently,

$$\beta_{ij} = \text{softmax}_j(\tilde{s}_{ij}) \approx \text{softmax}_j(Q_i K_j^\top),$$

which is the standard Transformer attention weighting rule.

2.3. The aggregation becomes the V -projection sum

Recall our gauge-covariant aggregation rule

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

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

Therefore, define a value projection

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

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 $\alpha_{ij} = \text{softmax}_j(Q_i K_j^\top)$ 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.

2.4. Conclusion

We have shown the following:

1. Under our assumptions, the score $s_{ij} = -\text{KL}(q_i \parallel \Omega_{ij} q_j)$ 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 $Q K^\top$.
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 $\sigma^2 \mathbb{I}$, our gauge-covariant, uncertainty-aware message passing law (or $q_i = \delta$ if you like)

$$\beta_{ij} \propto \text{softmax}_j \left(-\text{KL}(q_i \parallel \Omega_{ij} q_j) \right),$$

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

becomes exactly the canonical Transformer attention

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

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

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.

2.5. Generalized Variational Energy

Following Friston[^{^1}] we define a variational free energy as

$$\mathcal{F}[q] = D_{KL}[q(c) \mid p(c)] - \mathbb{E}_{q(c)}[\log p(o|c)].$$

where $q(c)$ represents an agents beliefs over c , $p(c)$ is the agent's prior, and $p(o|c)$ consitutes the agent's likelihood mapping.

For simplicity, in this study, we will not consider the "active" inference term - we shall concern ourselves with how agents passively condense into meta-agents and quasi-meta agents and its connection to machine learning architectures.

Given our generalized gauge theory model we extend this principle to agents who may share information/beliefs/models via parallel transport. We define a generalized variational energy for an agent A as:

$$\mathcal{V}_A = D_{KL}[q_A(c) \parallel \Phi p_A(c)] - \mathbb{E}_{q_A}[\log p_A(o|c)] + \sum_i \mathcal{V}_{\Lambda_i},$$

where \mathcal{V}_{Λ_i} represents possible interactions between other agents and (quasi)-meta-agents mediated by the various gauge connections above.

In our current considerations we will only focus on the induced connection Ω such that agents can compare beliefs between different sections of \mathcal{B}_q where, again,

$$\Omega : \Gamma(\mathcal{B}_q) \rightarrow \Gamma(\mathcal{B}_q).$$

The induced connection provides a parallel transport of beliefs which in turn defines the coordination structure used in our generalized variational energy term. This construction is well defined even when the fibers are not vector spaces so long as they carry a smooth G -action[³⁴].

Therefore, we have as a generalization

$$\begin{aligned} \mathcal{V}[q_i(c)] &= \alpha D_{KL}[(q_i(c)|p_i(c))] \\ &+ \sum_j \beta_{ij} D_{KL}[q_i(c)|\Omega_{ij}q_j(c)] \\ &- \sum_i \mathbb{E}_{q_i(c)}[\log p_i(o_i|c)]. \end{aligned}$$

where α, β represents general couplings (we generally take $\alpha = 1$. We note in passing that this expression bears remarkable similarities to the Grand potential $\Upsilon = U - TS + \mu N$ in standard thermodynamics (where μN is analogous to the Ω term)!

The β_{ij} coupling warrants more consideration. We tentatively assume that agents will favor alignment with other agents who share their beliefs. A natural candidate for β_{ij} is then a Boltzmann distribution given by

$$\beta_{ij}(c) = \frac{\exp\left[-\frac{1}{\kappa} \text{KL}(q_i(c) \parallel \Omega_{ij}[q_j](c))\right] m_{ij}(c)}{\sum_k \exp\left[-\frac{1}{\kappa} \text{KL}(q_i(c) \parallel \Omega_{ik}[q_k](c))\right] m_{ik}(c)}$$

where $m_{ij}(x)$ is the agents support (or mask).

Note: for full generality we could construct analogous terms for agent-agent models.

2.6. 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 $\mathfrak{g} = \text{Lie}(G)$, representing the gauge frame of agent i . Then, for each pair of agents (i, j) , we define the induced transport operator as:

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

where the exponential map $\exp : \mathfrak{g} \rightarrow G$ 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:

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

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, $\Omega_{ii} = \text{Id}$, 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.

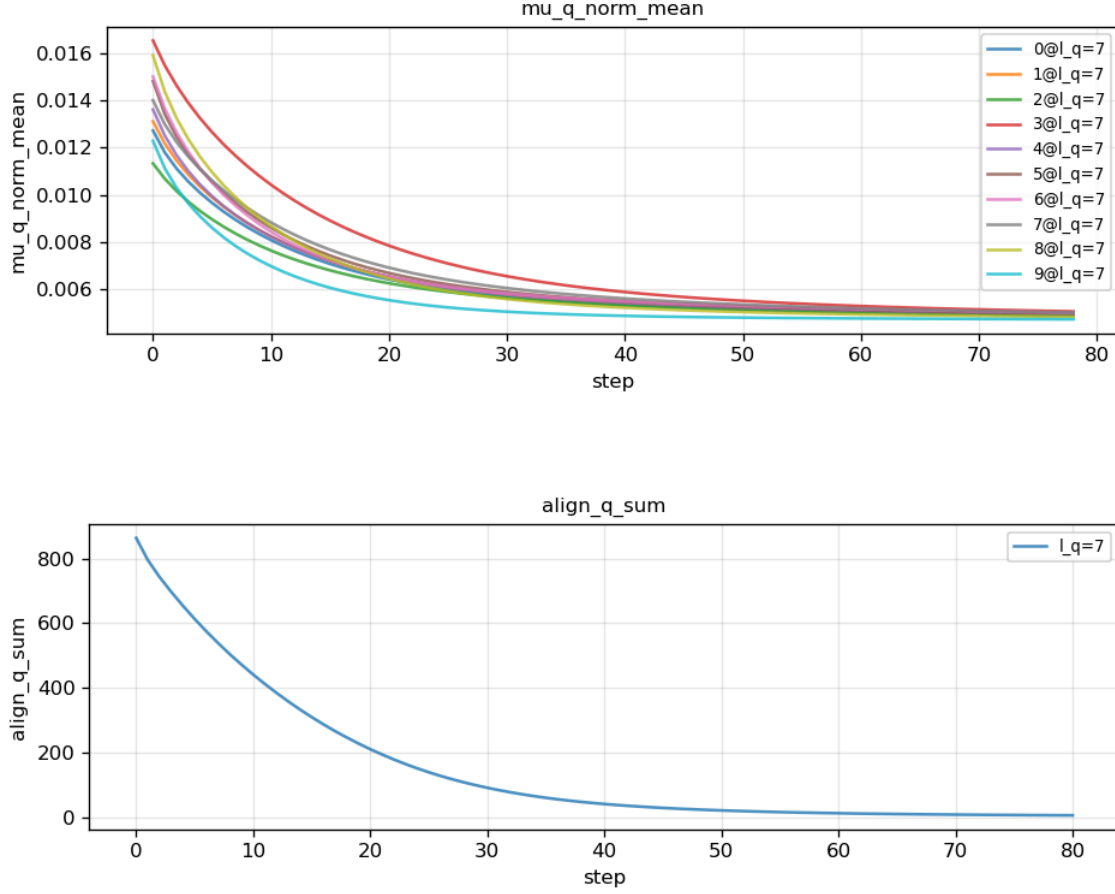
In our framework, each agent maintains a fixed gauge frame $\phi_i(c)$, which defines their local epistemic perspective. This frame serves as the reference against which other agents' beliefs are compared. The term $D_{\text{KL}}[q_i(c) \parallel \Omega_{ij}(c) \cdot q_j(c)]$ then quantifies the epistemic 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 consciousness, organizing all incoming beliefs and observations relative to a quasi-static internal model.

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

$$\begin{aligned} \mathcal{V}[q(c)] &= \alpha \sum_i D_{\text{KL}}[(q_i(c) | p_i(c))] \\ &\quad + \sum_{ij} \beta_{ij} D_{\text{KL}}[q_i(c) | \Omega_{ij} q_j(c)] \end{aligned}$$

Detailed simulations and analytic derivations produce agent belief and frame flow towards a shared constant vacuum state $\mu_i(c), \phi_i(c) \rightarrow \mu^*, \phi^*$ 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 $SO(3)$).



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 $-\mathbb{E}_{q_i(c)}[\log p(o_i|c)] \rightarrow -\log p(o_i|c)$ the standard negative log-likelihood term
3. but if we take observations to be Gaussian then this simplifies to

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

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

which is the standard machine learning loss function!

With only attention (which we have shown is due to the $\sum_j \beta_{ij} D_{\text{KL}}[q_i(c)|\Omega_{ij}q_j(c)]$ 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.

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

3. 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.

4. Appendix

4.0.1. A Tantalizing Aside

This interpretation, pregnant with possibilities, opens the door to a rich epistemic ontology: communication becomes a gauge-theoretic alignment of perspectives, and disagreement (or confusion) corresponds to curvature in some "semantic" field. Agents with similar gauge frames can align beliefs easily, while those with strongly divergent frames experience epistemic dissonance.

Additionally, for future exploration, human language itself may possibly be considered as a gauge theory. Each speaker maintains a local linguistic frame — a grammar, lexicon,

and semantic structure — and successful communication requires parallel transport of meaning across these frames. Misunderstandings arise when the linguistic gauge transformations between individuals are ill-defined or insufficient to fully align perspectives. In this view, syntax encodes local structural constraints, semantics provides the fiber content, and discourse becomes a path-dependent operation on shared meaning spaces. The geometric framework developed here may thus offer a natural foundation for modeling language as a structured system of epistemic transport.

Interestingly and tantalizingly this framework might allow one to study the pullbacks of beliefs/models from the fiber to the base manifold. The fiber is manifestly informational. This suggests a wild interpretation that agent qualia might be modeled as the pull back of informational geometries to the base manifold. E.g. agents at a point c in \mathcal{C} pullback their beliefs/models via their personal sections - they "experience" different phenomena over the same underlying base space. This, however, is fantastically speculative but it may allow new pursuits in Wheeler's "It from Bit" ontology!

As another example, constructivist theories of physics might have a firm mathematical model with which to explore. For example, if we consider this base manifold as a Kantian noumenon then we might speculate that physics itself is a generative model of cognition that humans have evolutionarily flowed towards sharing. In this view gauge invariance in physics is necessitated by the gauge frame alignment of human agents! There is much more to say.

Here we write the most general Vacuum Variational Action Functional our framework can offer. Notice that in this formulation we build TWO associated bundles and write bundle morphisms $\Phi, \tilde{\Phi}$ between them. In this most general formulation new cross-fiber terms can be written (γ_{ij}) .

$$\begin{aligned}
S = & \underbrace{\sum_i \int_{\mathcal{C}} \alpha_i \text{KL}(q_i(x) \parallel \tilde{\Phi}_i[p_i](x)) \, dx}_{(1) \text{ Self term: belief} \rightarrow \text{model alignment}} \\
& + \underbrace{\sum_i \int_{\mathcal{C}} \tilde{\alpha}_i \text{KL}(p_i(x) \parallel \Phi_i[q_i](x)) \, dx}_{(2) \text{ Feedback term: model} \rightarrow \text{belief (currently not studied)}} \\
& + \underbrace{\sum_i \sum_j \int_{\mathcal{C}} \beta_{ij}(x) \text{KL}(q_i(x) \parallel \Omega_{ij}^{(q)}[q_j](x)) \, dx}_{(3) \text{ Alignment across belief fibers}} \\
& + \underbrace{\sum_i \sum_j \int_{\mathcal{C}} \tilde{\beta}_{ij}(x) \text{KL}(p_i(x) \parallel \tilde{\Omega}_{ij}^{(p)}[p_j](x)) \, dx}_{(4) \text{ Alignment across model fibers}} \\
& + \underbrace{\sum_i \sum_j \int_{\mathcal{C}} \tilde{\gamma}_{ij}(x) \text{KL}(q_j(x) \parallel \Omega_{ji}^{(q)}[\tilde{\Phi}_i[p_i]](x)) \, dx}_{(5) \text{ Cross-fiber coupling: } p_i \text{ pushed to } q_j \text{ frame (Case A)}} \\
& + \underbrace{\sum_i \sum_j \int_{\mathcal{C}} \gamma_{ij}(x) \text{KL}(p_i(x) \parallel \Phi_i[\Omega_{ij}^{(q)}[q_j]](x)) \, dx}_{(6) \text{ Cross-fiber coupling: } q_j \text{ pushed to } p_i \text{ frame (Case B)}} \\
& + \underbrace{\lambda_A \int_{\mathcal{C}} \frac{1}{2} \sum_{a < b} \|F_{ab}(x)\|^2 \, dx}_{(7) \text{ Gauge curvature energy}} \\
& + \underbrace{\lambda_\phi \int_{\mathcal{C}} \frac{1}{2} |D\phi(x)|^2 \, dx}_{(8) \text{ Optional gauge-fixing or smoothness penalty } (\phi\text{-mass term)}} \\
& + (\text{other regularizers: Fisher metric, mass, etc.})
\end{aligned}$$