Gauge Equivariant Extension of FEP and Attention - Supplementary Information

Robert C. Dennis

1 Appendix

1.1 General Mathematical Framework

1.1.1 Principal Bundle and Associated Bundles

The following geometric and probabilistic constructions comprise standard methods in differential geometry and gauge theory. See reference for details, proofs, and examples [4][3][2][1]

Let $\pi : \mathcal{N} \to \mathcal{C}$ be a smooth principal G-bundle where $\mathcal{C}$ is a smooth manifold (the base space) and G is a Lie group (the structure 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 $\rho_q : G \to \mathrm{Aut}(\mathcal{B}_q)$ and $\rho_p : G \to \mathrm{Aut}(\mathcal{B}_p)$ be representations of G on smooth statistical manifolds $\mathcal{B}_q$ (belief/recognition fiber) and $\mathcal{B}_p$ (model/prior fiber). These fibers are typically:

- K-dimensional probability simplices Δ^K (for categorical distributions), or
- Statistical manifolds with information-geometric structure (e.g., Gaussian manifolds, exponential families)

The associated bundles are:

$$\mathcal{E}_q := \mathcal{N} \times_{\rho_q} \mathcal{B}_q = (\mathcal{N} \times \mathcal{B}_q) \,/\, \sim_q, \tag{1}$$
$$\mathcal{E}_p := \mathcal{N} \times_{\rho_p} \mathcal{B}_p = (\mathcal{N} \times \mathcal{B}_p) \,/\, \sim_p, \tag{2}$$

where $(n \cdot g, b) \sim_u (n, \rho_u(g)b)$ for $u \in \{q, p\}$.

These give fiber bundles $\pi_{\mathcal{E}_q} : \mathcal{E}_q \to \mathcal{C}$ and $\pi_{\mathcal{E}_p} : \mathcal{E}_p \to \mathcal{C}$ with fibers $\mathcal{B}_q(c) \cong \mathcal{B}_q$ and $\mathcal{B}_p(c) \cong \mathcal{B}_p$ at each $c \in \mathcal{C}$.

1.2 Agents and Multi-Agent Systems

Agent
An agent $\mathcal{A}^i$ is a pair of local sections over a domain $\mathcal{U}_i \subset \mathcal{C}$:

$$\mathcal{A}^i = \left(\sigma_q^i, \sigma_p^i\right), \tag{3}$$

where $\sigma_q^i : \mathcal{U}_i \to \mathcal{E}_q$ and $\sigma_p^i : \mathcal{U}_i \to \mathcal{E}_p$.

We write $q_i(c) := \sigma_q^i(c) \in \mathcal{B}_q(c)$ and $p_i(c) := \sigma_p^i(c) \in \mathcal{B}_p(c)$ for the belief and model at base point c.

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

$$\mathcal{M} = \left\{\mathcal{A}^i = \left(\sigma_q^i, \sigma_p^i\right)\right\}_{i \in \mathcal{I}}. \tag{4}$$

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

Meta-Agent and Epistemic Death
A meta-agent is a multi-agent system whose component agents share identical section values on their overlap:

$$q_i(c) = q_j(c), \quad p_i(c) = p_j(c) \quad \text{for } c \in \mathcal{U}_i \cap \mathcal{U}_j. \tag{5}$$

A set of agents is \emph{epistemically dead} if they identically share both beliefs and models. While 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 toward emergence.

1.2.1 Bundle Morphisms and Transport Operators

Via standard horizontal lifting from the principal bundle $\mathcal{N}$ to the associated bundles, we obtain a hierarchy of morphisms and induced connections:

Intra-bundle transport:

- $\Omega_{ij}^{(q)} : \Gamma(\mathcal{B}_q) \to \Gamma(\mathcal{B}_q)$ (belief-to-belief)
- $\Omega_{ij}^{(p)} : \Gamma(\mathcal{B}_p) \to \Gamma(\mathcal{B}_p)$ (model-to-model)

Cross-scale transport:

- $\Lambda_{s'}^{s} : \Gamma^s(\mathcal{B}_q) \to \Gamma^{s'}(\mathcal{B}_q)$ (beliefs across scales)
- $\tilde{\Lambda}_{s'}^{s} : \Gamma^s(\mathcal{B}_p) \to \Gamma^{s'}(\mathcal{B}_p)$ (models across scales)

Inter-bundle morphisms:

- $\Theta^i_j : \Gamma(\mathcal{B}_q) \to \Gamma(\mathcal{B}_p)$ (belief to model)
- $\tilde{\Theta}^i_j : \Gamma(\mathcal{B}_p) \to \Gamma(\mathcal{B}_q)$ (model to belief)

Global bundle morphisms:

- $\Phi : \mathcal{E}_p \to \mathcal{E}_q$ (model bundle to belief bundle)
- $\tilde{\Phi} : \mathcal{E}_q \to \mathcal{E}_p$ (belief bundle to model bundle)

Here $\Gamma(\mathcal{B}_u)$ denotes the space of smooth sections over the base $\mathcal{C}$.

1.2.2 Gauge Frames and Connections

Each agent i possesses a local gauge frame field:

$$\phi_i : \mathcal{U}_i \to \mathfrak{g} = \mathrm{Lie}(G), \tag{6}$$

which induces a local connection one-form:

$$A^{(i)}_\mu(c) = U_i^{-1}(c)\partial_\mu U_i(c), \tag{7}$$

where $U_i(c) = \exp[\phi_i(c)] \in G$. The associated field strength (gauge curvature) is:

$$F^{(i)}_{\mu\nu}(c) = \partial_\mu A^{(i)}_\nu - \partial_\nu A^{(i)}_\mu + [A^{(i)}_\mu, A^{(i)}_\nu] \in \mathfrak{g}. \tag{8}$$

When two agents overlap at $c \in \mathcal{U}_i \cap \mathcal{U}_j$, the inter-agent gauge transformation:

$$\Omega_{ij}(c) = \exp[\phi_i(c)] \exp[-\phi_j(c)] \in G \tag{9}$$

transports agent j's representations into agent i's frame:

$$q_j(c) \mapsto \Omega_{ij}(c) \cdot q_j(c) := \rho(\Omega_{ij}(c)) q_j(c). \tag{10}$$

On overlaps, local connections are related by:

$$A^{(i)}_\mu = \Omega_{ij} A^{(j)}_\mu \Omega^{-1}_{ij} + \Omega_{ij} \partial_\mu \Omega^{-1}_{ij}. \tag{11}$$

1.2.3 Curvature Structure

The full framework incorporates four distinct types of curvature:

Statistical Manifold Curvature

The fiber $\mathcal{B}$ has intrinsic curvature tensor:

$$R^{\mathcal{B}}(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{12}$$

This measures the geometry of the space of probability distributions (e.g., Gaussian manifolds have constant negative curvature).

Gauge Group Curvature.

The structure group G itself is a curved manifold. For $G = SO(3)$:

- Topology: $SO(3) \cong \mathbb{RP}^3$
- Constant positive sectional curvature
- Non-commutativity: $\Omega_{ik} = \Omega_{ij}\Omega_{jk}$ does not commute with $\Omega_{jk}\Omega_{ik}$

Gauge Field Curvature.

The connection A_μ has field strength $F_{\mu\nu}$ measuring path-dependence of parallel transport through the base space $\mathcal{C}$.

Base Manifold Curvature.

If $\mathcal{C}$ carries a Riemannian metric $g_{\mathcal{C}}$, its curvature tensor:

$$R^{\mathcal{C}}(X,Y)Z = \nabla^{\mathcal{C}}_X \nabla^{\mathcal{C}}_Y Z - \nabla^{\mathcal{C}}_Y \nabla^{\mathcal{C}}_X Z - \nabla^{\mathcal{C}}_{[X,Y]} Z \tag{13}$$

affects the geometry of the latent space agents inhabit.

2 Covariance Dynamics and Equilibrium Analysis

2.1 Covariance Gradient of the Generalized Free Energy

Here we derive the gradient of the single-agent free energy $\mathcal{F}_i$ with respect to the covariance Σ_i for agent i; a well known result in information geometry.

The free energy decomposes as

$$\mathcal{F}_i = D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_{j \neq i} \beta_{ij}\, D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j) - \mathbb{E}_{q_i}[\log p(o_i \mid k_i)], \tag{14}$$

where $q_i = \mathcal{N}(\mu_i, \Sigma_i)$, $p_i = \mathcal{N}(\mu_{p,i}, \Sigma_{p,i})$, and $\sum_j \beta_{ij} = 1$ by construction.

2.1.1 Gaussian KL divergence and its derivative

For two Multivariate Gaussians, we have

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1,\Sigma_1)\,\|\,\mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{2}\Big[\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{tr}\big(\Sigma_2^{-1}\Sigma_1\big) \tag{15}$$

$$+ (\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1) - d\Big]. \tag{16}$$

and differentiating w.r.t. Σ_1 (holding μ_1,μ_2,Σ_2 fixed) gives

$$\frac{\partial D_{\mathrm{KL}}}{\partial\Sigma_1} = \frac{1}{2}\left[-\Sigma_1^{-1}+\Sigma_2^{-1}\right]. \tag{17}$$

Applying this to each term in $\mathcal{F}_i$:

$$\frac{\partial}{\partial\Sigma_i}D_{\mathrm{KL}}(q_i\,\|\,p_i) = \frac{1}{2}\left[-\Sigma_i^{-1}+\Sigma_{p,i}^{-1}\right], \tag{18}$$

$$\frac{\partial}{\partial\Sigma_i}D_{\mathrm{KL}}(q_i\,\|\,\Omega_{ij}q_j) = \frac{1}{2}\left[-\Sigma_i^{-1}+(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\right]. \tag{19}$$

The observation term contributes an $O(\Sigma_i^{-1})$ correction that we neglect in the high-precision / strong-alignment regime. Thus

$$\boxed{\frac{\partial\mathcal{F}_i}{\partial\Sigma_i} = \frac{1}{2}\left[-2\Sigma_i^{-1}+\Sigma_{p,i}^{-1}+\sum_j\beta_{ij}(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\right].} \tag{20}$$

Because $\sum_j\beta_{ij}=1$, there is no $(1+\sum_j\beta_{ij})$ prefactor in front of Σ_i^{-1}.

2.2 Fixed-Point Equation and Symmetric Solution

At equilibrium, we set $\partial\mathcal{F}_i/\partial\Sigma_i = 0$, giving

$$\Sigma_i^{-1} = \frac{1}{2}\left[\Sigma_{p,i}^{-1}+\sum_j\beta_{ij}(\Omega_{ij}\Sigma_j\Omega_{ij}^\top)^{-1}\right]. \tag{21}$$

This is a matrix-valued fixed-point equation coupling all agents: each agent's precision is the β-weighted combination of its own prior precision and the transported neighbor precisions.

2.2.1 Homogeneous limit

In the homogenous limit we assume
(i) all agents are identical, so $\Sigma_i = \Sigma_\infty$ for all i,
(ii) $\Omega_{ij} \approx I$ (weak misalignment),
(iii) $\Sigma_{p,i} = \Sigma_0$ (shared prior).
Then (21) becomes

$$\Sigma_\infty^{-1} = \frac{1}{2}\left[\Sigma_0^{-1} + \Sigma_\infty^{-1}\right] \quad \Longrightarrow \quad \Sigma_\infty = \Sigma_0. \tag{22}$$

Hence, in a perfectly symmetric population, the equilibrium covariance reproduces the shared prior.

2.2.2 Alignment-dominated regime

Although $\sum_j \beta_{ij} = 1$, the effective strength of alignment is controlled by the parameter τ in

$$\beta_{ij} = \frac{\exp\left[-\frac{1}{\tau} D_{\mathrm{KL}}\big(q_i \,\|\, \Omega_{ij} q_j\big)\right]}{\sum_k \exp\left[-\frac{1}{\tau} D_{\mathrm{KL}}\big(q_i \,\|\, \Omega_{ik} q_k\big)\right]}. \tag{23}$$

As $\tau \to 0$, β_{ij} becomes sharply peaked on whichever neighbor j best agrees (after transport). In that low-τ limit, the prior precision $\Sigma_{p,i}^{-1}$ becomes negligible relative to the socially enforced alignment term, and

$$\boxed{\Sigma_i^{-1} \;\approx\; \sum_j \beta_{ij} (\Omega_{ij} \Sigma_j \Omega_{ij}^\top)^{-1} \;=\; \left\langle (\Omega_{ij} \Sigma_j \Omega_{ij}^\top)^{-1} \right\rangle_\beta.} \tag{24}$$

Here $\langle \cdot \rangle_\beta$ denotes a β-weighted expectation over neighbors.

Therefore, in the strong-alignment (i.e. small τ) regime, agent i's precision matrix becomes the β-weighted average of its neighbors' transported precisions.

When, additionally, all agents already have approximately equal covariances and the transports are near-identity, i.e. $\Sigma_i \approx \Sigma_j$ and $\Omega_{ij} \approx I$, then (24) implies

$$\Sigma_i^{-1} \approx \Sigma_j^{-1} \quad \Longrightarrow \quad \Sigma_i \approx \Omega_{ij} \Sigma_j \Omega_{ij}^\top, \tag{25}$$

justifying the alignment assumption used in the main text.

2.2.3 Gradient flow dynamics

Finally, consider the gradient flow

$$\frac{d\Sigma_i}{dt} = -\eta_\Sigma \frac{\partial \mathcal{F}_i}{\partial \Sigma_i}, \qquad \eta_\Sigma > 0. \tag{26}$$

Local stability of the equilibrium (21) follows from the positive-definiteness of the Hessian. For the Gaussian KL terms,

$$\frac{\partial^2 D_{\mathrm{KL}}}{\partial \Sigma_1 \, \partial \Sigma_1} \;\sim\; \Sigma_1^{-1} \otimes \Sigma_1^{-1} + \Sigma_2^{-1} \otimes \Sigma_2^{-1}, \tag{27}$$

which is manifestly positive definite for Σ_1, Σ_2.

Hence the covariance alignment fixed-point is an attractor of the variational dynamics.

Therefore, we find that $\Sigma_i \approx \Omega_{ij} \Sigma_j \Omega_{ij}^\top$ emerges from the dynamics itself rather than as an imposed constraint.

3 Relating The Quadratic Forms to Transported KL Divergences

We now show that the pairwise quadratic expectations in (**??**) can be expressed in terms of KL divergences between transported distributions (see appendix for general requirement of the forward KL).

3.1 Exact expansion for Gaussian beliefs

For independent Gaussians $q_i = \mathcal{N}(\mu_{q,i}, \Sigma_{q,i})$ and $q_j = \mathcal{N}(\mu_{q,j}, \Sigma_{q,j})$, the expectation of a quadratic form is

$$\mathbb{E}_{q_i q_j}[\delta^\top A \delta] = \mathrm{tr}\big(A \, \mathrm{Cov}(\delta)\big) + \bar{\delta}^\top A \bar{\delta}, \tag{28}$$

where $\delta = k_i - \Omega_{ij} k_j$, $\bar{\delta} = \mathbb{E}[\delta] = \mu_{q,i} - \Omega_{ij} \mu_{q,j}$, and

$$\mathrm{Cov}(\delta) = \Sigma_{q,i} + \Omega_{ij} \Sigma_{q,j} \Omega_{ij}^\top. \tag{29}$$

Applying (28) with $A = \Lambda_{ij}$:

$$\begin{aligned}\mathbb{E}_{q_i q_j}[(k_i - \Omega_{ij} k_j)^\top \Lambda_{ij} (k_i - \Omega_{ij} k_j)] = \mathrm{tr}\big[\Lambda_{ij} (\Sigma_{q,i} + \Omega_{ij} \Sigma_{q,j} \Omega_{ij}^\top)\big] \\ + (\mu_{q,i} - \Omega_{ij} \mu_{q,j})^\top \Lambda_{ij} (\mu_{q,i} - \Omega_{ij} \mu_{q,j}).\end{aligned} \tag{30}$$

An identical expansion holds for the model channel with Γ_{ij} and $\tilde{\Omega}_{ij}$.

We observe that we can choose the coupling precisions Λ_{ij} and Γ_{ij} to make (30) proportional to a KL divergence.

Specifically, we define

$$\Lambda_{ij} := \tau_{ij}^{(q)} \, (\Omega_{ij}\Sigma_{q,j}\Omega_{ij}^{\top})^{-1}, \qquad \Gamma_{ij} := \tau_{ij}^{(p)} \, (\tilde{\Omega}_{ij}\Sigma_{p,j}\tilde{\Omega}_{ij}^{\top})^{-1}, \tag{31}$$

where $\tau_{ij}^{(q)}, \tau_{ij}^{(p)} > 0$ are dimensionless coupling strengths.

Recall that the KL divergence between two Gaussians $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ and $\Omega_{ij}q_j = \mathcal{N}(\Omega_{ij}\mu_j, \Omega_{ij}\Sigma_j\Omega_{ij}^{\top})$ is

$$D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij}q_j) = \frac{1}{2}\Big[\mathrm{tr}\big((\Omega_{ij}\Sigma_j\Omega_{ij}^{\top})^{-1}\Sigma_i\big) - d_q$$

$$+ \log \frac{\det(\Omega_{ij}\Sigma_j\Omega_{ij}^{\top})}{\det \Sigma_i} + (\mu_i - \Omega_{ij}\mu_j)^{\top}(\Omega_{ij}\Sigma_j\Omega_{ij}^{\top})^{-1}(\mu_i - \Omega_{ij}\mu_j)\Big].$$

When beliefs are approximately aligned (the regime enforced by the coupling itself), the covariances satisfy $\Sigma_i \approx \Omega_{ij}\Sigma_j\Omega_{ij}^{\top}$, and the trace and log-determinant terms approximately cancel. In this alignment regime, as shown in the appendix, the quadratic expectation becomes

$$\frac{1}{4}\mathbb{E}_{q_iq_j}[(k_i - \Omega_{ij}k_j)^{\top}\Lambda_{ij}(k_i - \Omega_{ij}k_j)] \approx \frac{\tau_{ij}^{(q)}}{2} D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij}q_j) + \mathrm{const}, \tag{32}$$

where the constant absorbs dimension-dependent terms and $O(\|\Delta\|^2)$ covariance mismatch corrections.

Similarly for the model channel:

$$\frac{1}{4}\mathbb{E}_{s_is_j}[(m_i - \tilde{\Omega}_{ij}m_j)^{\top}\Gamma_{ij}(m_i - \tilde{\Omega}_{ij}m_j)] \approx \frac{\tau_{ij}^{(p)}}{2} D_{\mathrm{KL}}(s_i \,\|\, \tilde{\Omega}_{ij}s_j) + \mathrm{const}. \tag{33}$$

3.2 Defining normalized alignment weights

To obtain the standard form, we define normalized alignment weights

$$\beta_{ij} := \frac{\tau_{ij}^{(q)}}{2}, \qquad \gamma_{ij} := \frac{\tau_{ij}^{(p)}}{2}. \tag{34}$$

These weights have a clear interpretation: β_{ij} measures the strength of belief alignment between agents i and j, while γ_{ij} measures model alignment strength. In the limit $\beta_{ij} \to \infty$ with fixed γ_{ij}, agents' beliefs are forced to perfect agreement after transport, while their models may still differ. Conversely, $\gamma_{ij} = 0$ decouples model alignment entirely. In all subsequence equations we taked τ to be independent of each agent - a constant global value which we set to 1.

4 Conditional Uniqueness of the Forward KL Divergence via Variational Duality

We now show that, within a broad but well-defined class of variational games, the forward KL divergence $D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$ is the only divergence that yields a closed-form Gibbs-type solution for the belief update and a consistent dual interpretation for the attention weights. Agents locally minimize their agent-specific variational free energy and the system of agents minimize their collective global free energies. This local-global coordination gives rise to the expected forward KL attention term.

Here we are restricting the the matched-fiber case. The full generalization follows by applying the appropriate bundle morphisms/intertwiners.

The uniqueness of this term is conditional and follows from three assumptions:

1. $\mathcal{D}$ is local in c, of f-divergence form $\int q_i(c)\, f\!\left(\dfrac{q_i(c)}{\Omega_{ij}(c) q_j(c)}\right) dc$,
2. the coupling is linear
3. the minimizing belief, q_i^*, remains in the exponential-family (log-linear) class.

4.0.1 The Coupled Variational Problem

Each agent i minimizes a local free-energy functional:

$$F_i[\beta_i] = \min_{q_i} \Big\{ D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_{j \neq i} \beta_{ij}\, \mathcal{D}(q_i, q_j) \Big\}, \tag{35}$$

For notational convenience, we can decompose the KL divergence as:

$$D_{\mathrm{KL}}(q_i \,\|\, p_i) = \langle H_i \rangle_{q_i} + S(q_i) + \text{const}, \tag{36}$$

where

$$H_i(c) := -\log p_i(c) \quad \text{(local "energy")}, \tag{37}$$

$$S(q_i) := -\int q_i(c) \log q_i(c)\, dc \quad \text{(Shannon entropy)}, \tag{38}$$

$$\langle H_i \rangle_{q_i} := \int q_i(c) H_i(c)\, dc \quad \text{(expected energy)}. \tag{39}$$

This gives $D_{\mathrm{KL}}(q_i \,\|\, p_i) = \int q_i(c)[\log q_i(c) - \log p_i(c)]\, dc$.

The attention weights $\beta_{ij} \geq 0$ (with $\sum_j \beta_{ij} = 1$) are subsequently chosen by optimizing

$$\mathcal{J}_i(\beta_i) = \sum_{j \neq i} \beta_{ij} C_{ij} + \tau \sum_{j \neq i} \beta_{ij} \log \beta_{ij}, \tag{40}$$

where C_{ij} denotes the marginal cost of attending to agent j as we've described above.

We shall later identify

$$C_{ij} := \frac{\partial S_i}{\partial \beta_{ij}}, \tag{41}$$

so that attention weights allocate resources in proportion to marginal coordination penalties.

4.0.2 Forward KL and the Geometric-Mean Solution

Let $\mathcal{D}(q_i, q_j) = D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$, where

$$\frac{\delta D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)}{\delta q_i(c)} = \log \frac{q_i(c)}{\Omega_{ij}(c) q_j(c)} + 1.$$

The stationary condition for (35) is

$$H_i(c) + \log q_i(c) + \sum_j \beta_{ij} \left[\log \frac{q_i(c)}{\Omega_{ij}(c) q_j(c)} + 1 \right] = \lambda_i, \tag{42}$$

with λ_i enforcing normalization.

Rearranging and solving for $q_i(c)$ gives a Boltzmann distribution whose mean field is a geometric average of transported neighbor beliefs.

$$q_i^*(c) = \frac{1}{Z_i} e^{-H_i(c)/2} \prod_j [\Omega_{ij}(c) q_j(c)]^{\beta_{ij}/2}, \tag{43}$$

This structure is preserved only for the forward KL divergence; alternative divergences destroy the log-linearity of the exponent.

4.0.3 Dual Relation via the Envelope Theorem

At the stationary value q_i^*, the envelope theorem implies

$$\frac{\partial F_i}{\partial \beta_{ij}} = \mathcal{D}(q_i^*, q_j) = D_{\mathrm{KL}}(q_i^* \,\|\, \Omega_{ij} q_j), \tag{44}$$

such that the marginal cost of increasing attention to j equals the forward KL divergence between the updated belief q_i^* and the transported neighbor $\Omega_{ij}q_j$ into i's frame.

This identifies the attention cost with the KL.

4.0.4 Reverse and Symmetric KL Forms

If instead $\mathcal{D}(q_i, q_j) = D_{\mathrm{KL}}(\Omega_{ij}q_j \,\|\, q_i)$, the stationary condition becomes

$$H_i(c) + \log q_i(c) - \sum_j \beta_{ij} \frac{\Omega_{ij}(c)q_j(c)}{q_i(c)} = \text{const}, \tag{45}$$

introducing terms as $1/q_i$ and leading to a transcendental stationary equation without a closed-form solution destroying the exponential family requirement above.

Likewise, the symmetrized divergence

$$\mathcal{D}_{\mathrm{sym}}(q_i, q_j) = \tfrac{1}{2}[D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij}q_j) + D_{\mathrm{KL}}(\Omega_{ij}q_j \,\|\, q_i)]$$

mixes $\log q_i$ and $1/q_i$ terms, again breaking log-linearity.
Hence, among local f-divergences, only the forward KL preserves exponential-family closure.

4.0.5 Conditional Uniqueness Theorem

Let $\mathcal{D}(q_i, q_j)$ be any local f-divergence

$$\mathcal{D}(q_i, q_j) = \int q_i(c)\, f\left(\frac{q_i(c)}{\Omega_{ij}(c)q_j(c)}\right) dc,$$

that enters linearly in (35), and further suppose that the stationary distribution q_i^* is log-linear in $\{H_i, \Omega_{ij}q_j\}$.

Then the following are equivalent:

1. q_i^* has the geometric-mean Boltzmann form (43);
2. $\mathcal{D}(q_i, q_j) = D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij}q_j)$;
3. $C_{ij} = \dfrac{\partial F_i}{\partial \beta_{ij}} = D_{\mathrm{KL}}(q_i^* \,\|\, \Omega_{ij}q_j)$.

Proof sketch
(2) $\Rightarrow$ (1) follows from direct solution of the stationary condition.

(1) $\Rightarrow$ (2):

Assume the solution is log-linear (43).

Substituting this into the stationarity condition gives

$$\frac{\delta \mathcal{D}}{\delta q_i(c)}\Big|_{q_i^*} = \log \frac{q_i^*(c)}{\Omega_{ij}(c) q_j(c)} + 1.$$

Next, integrating, we find

$$\mathcal{D}(q_i^*, q_j) = \int q_i^*(c) \log \frac{q_i^*(c)}{\Omega_{ij}(c) q_j(c)}\, dc + \text{const.}$$

We require that $\mathcal{D}(q, q) = 0$ thereby fixing the constant to be zero thus producing the forward KL form.

(3) $\Rightarrow$ (2):

By the envelope theorem (44), the only divergence consistent with a linear β_{ij}-coupling and this derivative structure is the forward KL.

4.0.6 Interpretations

1. **Gauge invariance**
   The comparison is made between q_i and the transported $\Omega_{ij} q_j$ in the same frame (and similarly for $j \to i$). Gauge covariance fixes what is compared, while variational duality fixes how it is compared.

2. **Variational duality**

   The forward KL is the only divergence that simultaneously yields:

   - a closed-form Boltzmann solution for q_i, and
   - a consistent dual cost $C_{ij} = \partial F_i / \partial \beta_{ij}$.

3. **Information geometry**

   The forward KL is the Bregman divergence generated by the negative entropy potential, whose Hessian induces the Fisher-Rao metric and yields the m/e-projection (mixed/exponential family) Pythagorean theorem. These global properties are not shared by generic f-divergences.

4.0.7 Summary

Under the natural assumptions of locality, linear coupling, and exponential-family closure, the forward KL divergence

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

is not merely a modeling choice but a necessary consequence of the variational and geometric structure underlying agent coordination. It is furthermore rotationally invariant under $SO(N)$ and $SU(N)$.

4.0.8 Connection to the Gauge-Covariant Free Energy

The conditional uniqueness result above justifies the specific form of the alignment terms appearing in the generalized variational free energy:

$$\mathcal{S} = \sum_i D_{\mathrm{KL}}(q_i \,\|\, p_i) + \sum_i D_{\mathrm{KL}}(s_i \,\|\, r_i) + \sum_{i,j} \beta_{ij}\, D_{\mathrm{KL}}(q_i \,\|\, \Omega_{ij} q_j)$$

$$+ \sum_{i,j} \gamma_{ij}\, D_{\mathrm{KL}}(s_i \,\|\, \tilde{\Omega}_{ij} s_j) - \mathbb{E}_q[\log p(o|\{k_i, m_i\})].$$

Each coupling term, such as

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

is therefore not arbitrary: it arises uniquely from the requirement that

1. belief updates q_i admit an exponential-family form consistent with local free energy minimization,
2. attention weights β_{ij} act as variational dual variables conjugate to those divergences, and
3. comparisons between agents are made in gauge-aligned coordinates through the transport operators Ω_{ij}.

Therefore, the forward KL plays the role of the canonical gauge-covariant coupling between agents' beliefs/models, unifying variational and geometric principles.

Reverse or symmetric divergences would violate at least one of these constraints: they either destroy exponential-family closure, break dual consistency $\big(C_{ij} = \partial F_i / \partial \beta_{ij}\big)$, or fail to respect the gauge-covariant comparison structure. Thus, the gauge-covariant free energy (**??**) naturally inherits the unique forward-KL alignment form as a direct consequence of its underlying variational geometry.

4.1 Softmax Attention via Maximum Entropy Principle

Given alignment cost $C_{ij} = D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)$ above, we seek attention weights β_{ij} that:

1. Minimize expected disagreement: $\sum_j \beta_{ij} C_{ij}$
2. Maximize uncertainty (entropy): $-\sum_j \beta_{ij} \log \beta_{ij}$
3. Satisfy normalization: $\sum_j \beta_{ij} = 1$

Following Jaynes' maximum entropy principle, we maximize entropy subject to the constraint that expected cost equals some target $\langle C \rangle$:

$$\max_{\{\beta_{ij}\}} \left\{ -\sum_j \beta_{ij} \log \beta_{ij} \right\} \tag{46}$$

subject to:

$$\sum_j \beta_{ij} = 1, \tag{47}$$

$$\sum_j \beta_{ij} C_{ij} = \langle C \rangle. \tag{48}$$

Lagrangian:

$$\mathcal{J} = -\sum_j \beta_{ij} \log \beta_{ij} + \lambda \left(\sum_j \beta_{ij} - 1 \right) + \mu \left(\sum_j \beta_{ij} C_{ij} - \langle C \rangle \right). \tag{49}$$

First-order condition:

$$\frac{\partial \mathcal{J}}{\partial \beta_{ij}} = -\log \beta_{ij} - 1 + \lambda + \mu C_{ij} = 0. \tag{50}$$

Solution:

$$\log \beta_{ij} = \lambda - 1 + \mu C_{ij} \quad \Rightarrow \quad \beta_{ij} = K \exp(\mu C_{ij}), \tag{51}$$

where $K = \exp(\lambda - 1)$.
Normalization $\sum_j \beta_{ij} = 1$ gives:

$$\beta_{ij} = \frac{\exp(\mu C_{ij})}{\sum_k \exp(\mu C_{ik})}. \tag{52}$$

Since we want to \emph{minimize} expected cost (not maximize), we choose $\mu = -1/\tau < 0$:

$$\boxed{\beta_{ij} = \frac{\exp(-C_{ij}/\tau)}{\sum_k \exp(-C_{ik}/\tau)} = \mathrm{softmax}_j\left(-\frac{C_{ij}}{\tau}\right)}. \tag{53}$$

Interpretation:

- $\tau \to 0$: Hard attention (argmin over j)
- $\tau \to \infty$: Uniform attention (maximum uncertainty)
- Intermediate τ: Soft attention balancing cost minimization and entropy maximization

Alternative Derivation (Unconstrained):
Equivalently, minimize the free energy functional:

$$F[\beta_i] = \sum_j \beta_{ij} C_{ij} + \tau \sum_j \beta_{ij} \log \beta_{ij}, \tag{54}$$

subject only to $\sum_j \beta_{ij} = 1$.
Lagrangian:

$$\mathcal{J} = \sum_j \beta_{ij} C_{ij} + \tau \sum_j \beta_{ij} \log \beta_{ij} + \lambda \left(\sum_j \beta_{ij} - 1 \right). \tag{55}$$

First-order condition:

$$C_{ij} + \tau(\log \beta_{ij} + 1) + \lambda = 0 \quad \Rightarrow \quad \tau \log \beta_{ij} = -C_{ij} - \tau - \lambda. \tag{56}$$

This immediately gives:

$$\beta_{ij} = \frac{\exp(-C_{ij}/\tau)}{\sum_k \exp(-C_{ik}/\tau)}. \tag{57}$$

This is the unique maximum-entropy distribution consistent with the constraint $\langle C \rangle = \sum_j \beta_{ij} C_{ij}$.

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

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

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

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

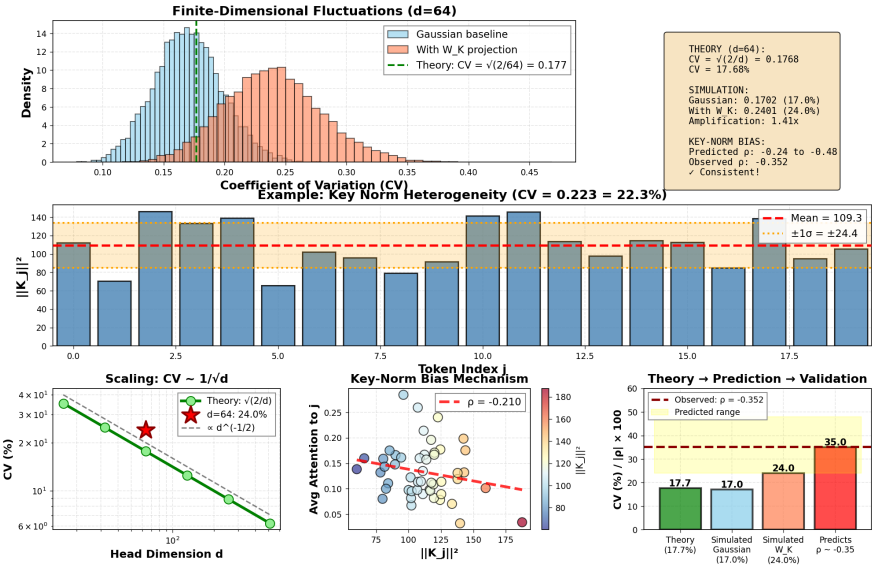

Figure 1: **Finite-dimensional fluctuations in transformer key norms.** (**Top-left**) Distribution of key-norm coefficient of variation (CV) under Gaussian baseline and after W_K projection, compared with theoretical prediction (CV $\approx 17.7\%$). (**Top-right**) Example of individual key norms across tokens illustrating sample CV. (**Bottom-left**) Finite-dimensional scaling of CV following the theoretical law CV $\sim 1/\sqrt{d}$, with BERT-base ($d = 64$) highlighted. (**Bottom-right**) Average attention bias as a function of key-norm magnitude, demonstrating minimal systematic bias even with CV 26%.

5 Variational Gradient Descent: Implementation and Numerical Methods

The variational free energy minimization is implemented through a sophisticated gradient descent scheme that respects the intricate geometric structure of the multi-agent system. At each simulation step, the algorithm orchestrates a sequence of coordinated field updates across all dynamic variables—the statistical parameters (μ_i, Σ_i) for both belief and model fibers, the gauge frame fields $(\phi_i, \tilde{\phi}_i)$, and optionally the global connection field A_μ—while maintaining numerical stability on the constrained manifolds where these quantities naturally live.

5.1 Gradient Accumulation and Energy Terms

The core computational pipeline begins with gradient accumulation, where each agent's variational gradient is constructed by summing contributions from multiple energy terms. For the statistical parameters, we compute gradients arising from the self-consistency term $D_{\mathrm{KL}}(q_i \| p_i)$, the belief alignment coupling $\sum_j \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)$, the model alignment coupling $\sum_j \gamma_{ij} D_{\mathrm{KL}}(s_i \| \tilde{\Omega}_{ij} s_j)$, and the observation likelihood $-\mathbb{E}_{q_i}[\log p(o_i|k_i)]$. Each of these terms contributes a distinct gradient component that must be carefully transported, aggregated, and symmetrized before application.

The alignment terms are particularly intricate, requiring neighbor iteration over all overlapping agent pairs, computation of gauge-transported statistics via the parallel transport operators Ω_{ij}, and evaluation of the KL divergence gradients in the appropriate local frames. These gradients are accumulated into per-agent "inboxes" during a first pass, then drained and combined with self-energy gradients in a second pass to yield the total gradient for each agent.

5.2 Natural Gradient Descent on the Gaussian Manifold

A critical aspect of our implementation is the use of natural gradient descent for the statistical parameters, which exploits the information-geometric structure of the Gaussian manifold. Rather than treating the covariance matrices Σ_i as Euclidean variables, we recognize them as points on the manifold of symmetric positive-definite matrices, equipped with the Fisher-Rao metric. The natural gradient at a point Σ on this manifold is obtained by applying the inverse Fisher metric to the Euclidean gradient, yielding an intrinsic tangent vector that respects the manifold's curved geometry.

In our vectorized implementation, this transformation is performed via `apply_natural_gradient_batch`, which computes the whitened gradient

$$R = \Sigma^{-1/2}(\eta \cdot \nabla)\Sigma^{-1/2} \tag{58}$$

where ∇ is the raw Euclidean gradient and η is the learning rate. This whitening procedure ensures that the gradient step is affine-invariant—the same step size produces comparable effects regardless of the current scale of the covariance matrix. Following the computation of the tangent vector R, we apply a trust-region constraint by clipping its Frobenius norm to a maximum relative step size ρ, typically set to values between 0.1 and 0.5. This prevents excessively large updates that could destabilize the manifold structure or lead to ill-conditioned covariances.

5.3 Manifold Retraction for Covariance Matrices

The actual manifold retraction—the process of mapping the tangent vector back to a valid point on the manifold—is performed using the affine-invariant exponential map by default. The exponential map is computed as

$$\Sigma_{k+1} = \Sigma_k^{1/2} \exp(R) \Sigma_k^{1/2}, \tag{59}$$

where the matrix exponential $\exp(R)$ is evaluated via eigendecomposition: $R = U\Lambda U^\top$ yields $\exp(R) = U \exp(\Lambda) U^\top$ with component-wise exponentials applied to the eigenvalues. This procedure is provably SPD-preserving provided Σ_k is SPD and R is symmetric, which our implementation guarantees through explicit symmetrization at multiple stages.

Additional safeguards include post-retraction sanitization via `sanitize_sigma`, which symmetrizes the result, raises on any floating-point anomalies (NaNs), and applies a spectral floor to the eigenvalues to ensure they remain above a minimum threshold $\epsilon_{\text{SPD}} \sim 10^{-8}$. This spectral regularization is essential for preventing numerical collapse of the covariance matrices during prolonged gradient descent, particularly in regions where the free-energy landscape becomes very flat.

5.4 Gauge Frame Dynamics on $\mathrm{SO}(3)$

For the gauge frame fields ϕ_i and $\tilde{\phi}_i$, which live in the Lie algebra $\mathfrak{so}(3)$, a different geometric structure must be respected. These fields parameterize the local gauge frames as $U_i = \exp(\phi_i)$, and their dynamics are governed by gradients computed via the induced Fisher metric on the group manifold.

The gradients are obtained by accumulating contributions from the self KL divergences, the alignment terms, and any optional (disabled) cross-fiber couplings, each of which produces a co-vector in the dual of the Lie algebra.

These co-vectors are computed using the differential of the matrix exponential map, which relates infinitesimal variations $\delta\phi$ to variations in the group element via

$$\delta(\exp\phi) = \exp(\phi) \cdot d\exp_\phi(\delta\phi), \tag{60}$$

where $d\exp_\phi$ is the derivative of the exponential map at ϕ. Our implementation caches the matrix-valued operator $d\exp^{-1}$ (implemented via the Baker-Campbell-Hausdorff formula truncated at fourth order) and uses it to project co-vectors back to tangent vectors in the algebra. The Fisher metric inverse, computed via `inverse_fisher_metric_field`, provides a natural Riemannian structure that preconditions these gradients, accounting for the non-flat geometry of SO(3).

Once the gradient direction is determined, we apply a step with learning rate η_ϕ (typically 10^{-1} to 10^{-2}), forming a candidate update $\phi_{\text{cand}} = \phi + \delta\phi$. However, the Lie algebra $\mathfrak{so}(3)$ has a natural periodic boundary—angles wrap at $\pm\pi$—and numerical drift can cause ϕ to wander outside the principal domain. To prevent this, we apply a retraction via `retract_phi_principal`, which reflects vectors that exceed the boundary back into the fundamental domain, then slightly nudges them away from the boundary via a small margin (typically 10^{-2}) to avoid numerical instability near the critical points where the exponential map degenerates.

5.5 Learning Rate Configuration

Learning rates are carefully tuned to balance convergence speed against numerical stability, with separate tunable rates for each class of variable reflecting their characteristic time scales. Belief means μ_q and covariances Σ_q use rates $\eta_{\mu,q} \sim 10^{-1}$ and $\eta_{\Sigma,q} \sim 10^{-1}$, while model parameters use slightly smaller values to enforce the separation between fast belief dynamics and slow model learning. Gauge frame fields utilize $\eta_\phi \sim 10^{-1}$ to allow relatively rapid frame adjustment. All rates can be dynamically scaled via a global multiplier, and in regions of steep gradients an adaptive reduction mechanism (not currently enabled by default) can further decrease step sizes to prevent overshooting.

During each update, gradients are accumulated in double precision (float64) to minimize rounding errors, particularly for the gauge frame fields where small numerical discrepancies can accumulate over many steps due to

the group structure. After the gradient step is computed, results are cast back to single precision (float32) for storage and subsequent energy evaluations, balancing numerical accuracy with memory efficiency.

5.6 Convergence Criteria and Diagnostics

The simulation monitors convergence through multiple criteria, primarily tracking the total free energy $\mathcal{F}$ and its rate of change $\Delta\mathcal{F}$. A run is considered converged when the free energy stabilizes, defined as $|\Delta\mathcal{F}| < 10^{-5}$ for at least 200 consecutive steps, indicating the system has reached a stationary point (local minimum) of the variational functional. Typical simulations require between 300 and 8000 steps to reach this criterion, depending on initialization and coupling strengths.

In addition to the global energy, we compute per-agent diagnostics including the norms of belief and model means $\|\mu_{q,i}\|$, $\|\mu_{p,i}\|$, trace-normalized covariances, alignment metrics measuring the KL divergence between neighboring agents, and the distribution of attention weights β_{ij}. These quantities are logged at each step and visualized in-situ to verify that the dynamics exhibit expected behavior—for instance, in vacuum (observation-free) runs, all agent mean norms should converge to a common value, reflecting the rotationally symmetric ground state. When observations are present, these norms diverge as agents specialize, providing a clear signature of spontaneous symmetry breaking.

5.7 Numerical Stability Safeguards

The update procedure includes extensive numerical safeguards to ensure stability over long simulation runs. The `sanitize_sigma` function is invoked after every covariance update to detect and repair ill-conditioned matrices, applying a spectral floor to eigenvalues, capping condition numbers if they exceed configurable thresholds (typically 10^8 to 10^{10}), and optionally renormalizing the trace to prevent runaway growth.

For gauge transport operators Ω_{ij}, the function `safe_omega_inv` verifies that these matrices remain approximately orthogonal by checking the deviation $\|\Omega^\top\Omega - I\|_F$ against a tolerance scaled by machine epsilon; if the deviation is excessive, the matrix is re-orthogonalized via QR decomposition with determinant correction to ensure it remains in $\mathrm{SO}(3)$.

Energy budget tracking, implemented via the `EnergyBudget` class, monitors the decomposition of the total free energy into its constituent terms at each step, verifying that energy conservation is approximately satisfied (within

numerical tolerance) when no external observations are added. Similarly, a `StabilityMonitor` tracks condition numbers of all covariance matrices, checks for the emergence of NaN or infinite values, detects gradient explosions (gradients exceeding preset thresholds), and validates that the SPD property is maintained throughout the evolution. If critical stability issues are detected—such as covariance matrices losing positive-definiteness or gradients diverging—the monitor logs detailed diagnostics and optionally halts the simulation to prevent catastrophic numerical failure.

5.8 Parallelization Strategy

Parallelization is employed to accelerate gradient computation across multiple agents, leveraging the `joblib` library with a `loky` backend for process-based parallelism. The computation is structured so that each worker receives a subset of agents along with read-only access to the shared runtime context and a memory-mapped disk cache containing precomputed expensive quantities such as gauge transport operators and their derivatives. Gradients are computed independently for each agent in parallel, then aggregated in the master process to form the global update. This design scales efficiently up to the number of physical cores available.

The parallel implementation carefully manages memory to avoid duplication of large arrays, using shared memory views where possible and writing intermediate results to disk-backed caches that are accessible across processes. Thread-level parallelism within each worker is deliberately suppressed (via environment variables like `OMP_NUM_THREADS=1`) to prevent oversubscription and maintain cache coherence, as nested parallelism typically degrades performance for these workloads.

5.9 Integration and Orchestration

The overall gradient descent loop integrates these components into a cohesive pipeline. At each step, we first zero all gradient accumulators, then invoke `compute_all_gradients` for each agent, which populates the gradient fields `grad_mu_q`, `grad_Sigma_q`, `grad_mu_p`, `grad_Sigma_p`, `grad_phi`, and `grad_phi_tilde`. These raw gradients are then post-processed—symmetrized for covariance gradients, preconditioned via the Fisher metric, and clipped to trust regions—before being applied to update the agent fields.

Gauge frames are retracted onto the principal domain, covariances are sanitized, and dirty flags are set to invalidate any cached derived quantities that depend on the updated fields. Finally, we compute global metrics and

the total action to assess convergence and prepare for visualization. This entire sequence is orchestrated by the `synchronous_step` function, which ensures that all agents are updated in a coordinated, lock-step fashion at each simulation tick, maintaining consistency of the multi-agent state across the entire base manifold.

The resulting dynamics exhibit stable convergence to stationary points of the free energy functional, with numerical precision sufficient to reliably distinguish between symmetric vacuum states and observation-induced symmetry-broken configurations—precisely the behavior required to validate the gauge-theoretic framework proposed in this work.

5.10 Computational Requirements

The simulations were performed on a workstation with an AMD Ryzen 9 5950X processor (16 cores, 32 threads) and 64 GB RAM. Typical simulation runs with $N = 8$ agents over a 20×20 spatial grid required approximately 60 minutes for convergence (500-800 steps). Memory usage scales as $\mathcal{O}(N \cdot H \cdot W \cdot K^2)$ where $H \times W$ is the spatial resolution and K is the fiber dimension.

5.11 Hyperparameter Configuration

All hyperparameters used in the experiments are documented in `config.py` within the repository. Key parameters include:

Parameter	Value	Description
$\eta_{\mu,q}$	1×10^{-1}	Belief mean learning rate
$\eta_{\Sigma,q}$	1×10^{-1}	Belief covariance learning rate
η_ϕ	1×10^{-1}	Gauge frame learning rate
τ	1.0	Attention temperature
α	1.0	Self-consistency weight
ϵ_{SPD}	1×10^{-8}	Covariance regularization floor
ρ	0.3	Trust region radius

Table 1: Standard hyperparameter configuration for all experiments.

5.12 Random Seed Reproducibility

All experiments use seeded random number generators (NumPy `RandomState`). The seeds used for each figure are documented in the repository's

`experiments/` directory. To reproduce Figure˜**??**, run:

python generalized_simulation.py –config configs/vacuum_exp.py –seed 241

References

[1] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

[2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[3] Theodore Frankel. *The Geometry of Physics: An Introduction*. Cambridge University Press, 3rd edition, 2011.

[4] Mikio Nakahara. *Geometry, Topology and Physics*. CRC Press, 2nd edition, 2003.