Statistical Geometry of Group Actions: Symmetry, Transport, Estimation

Q. Berthet, P. Bartlett, R. Elie, Y.-H. Wu

arXiv:2611.10111v1 [math.ST] 6 Nov 2026

We study variational inference in the presence of distributional symmetry, focusing on the interaction between two distinct group actions on a sample space: a *symmetry* action that leaves a reference distribution invariant, and a *transport* action that generates a parametric variational family by pushforward. By analyzing the algebraic structure of the total transformation group generated by both actions and the induced adjoint representation of the symmetry group on the transport Lie algebra, we establish that the Riemannian gradient of the Kullback–Leibler (KL) divergence at the true transport parameter is constrained to lie in the trivial representation of the symmetry group. Consequently, symmetry forces the gradient to vanish along all non-trivial representations. Combining this representation-theoretic constraint with parameter identifiability yields an orthogonal decomposition of the transport Lie algebra into three interpretable subspaces: an *identifiability kernel* (along which the objective is locally constant), a *bias subspace* (where the gradient may be nonzero due to model misspecification), and a *symmetry-breaking subspace* (where the gradient is forced to vanish by symmetry alone). Finally, we characterize the residual bias directions in terms of distributional mismatch between the base model and the standardized target, and derive an explicit Stein-type moment-matching condition that eliminates this bias, extending the location-scale results of Margossian and Saul (2025) to general Lie group transformation families.

1. Introduction

In modern statistical practice, variational inference (Blei et al., 2017; Jordan et al., 1999; Wainwright and Jordan, 2008) has become a primary methodology for scalable approximate inference, most notably for approximating intractable posterior distributions in Bayesian models. Rather than sampling directly from a complex target distribution p on a sample space $\mathcal{X}$, variational inference casts inference as an optimization problem: one seeks a distribution $q \in \mathcal{Q}$ within a tractable *variational family* that minimizes a statistical divergence—most commonly the Kullback–Leibler (KL) divergence $\mathrm{KL}(q\|p)$. In practice, the variational family $\mathcal{Q}$ is nearly always misspecified with respect to p: factorized mean-field families sacrifice dependence structure for tractability, while structured transformation families and normalizing flows (Papamakarios et al., 2021; Rezende and Mohamed, 2015) trade computational cost for richer geometry. Because p rarely belongs to $\mathcal{Q}$, a foundational statistical question arises: *even under model misspecification, when does the variational optimum exactly recover meaningful structural parameters of the target distribution?*

Symmetry offers a natural geometric principle to address this question. In classical mathematical statistics, group invariance is foundational: equivariance principles govern the theory of optimal invariant estimators and hypothesis tests (Eaton, 1989; Lehmann and Romano, 2005; Wijsman, 1990), the Hunt–Stein theorem establishes conditions under which minimax procedures can be chosen to be group-invariant (Bondar and Milnes, 1981), and the interplay between sufficiency and invariance is central to statistical decision theory (Hall et al., 1965). In contemporary machine learning, group invariance and equivariance have similarly guided the design of structured neural architectures (Bronstein et al., 2021; Cohen and Welling, 2016; Weiler and Cesa, 2019). By contrast, the role of distributional symmetry in shaping the optimization landscape of variational inference—specifically, how shared invariances between the variational base distribution and the standardized

target distribution constrain stationary points of the variational objective—has only recently begun to receive systematic study.

A striking step in this direction was taken by Margossian and Saul (2025), who demonstrated that variational inference can achieve exact parameter recovery in misspecified location families (generated by translations $x \mapsto x+\mu$) and location-scale families (generated by affine transformations $x \mapsto \mu + Ax$) whenever the base distribution q_0 and the standardized target share a common symmetry. Specifically, they established that the gradient of the KL divergence vanishes at the true location parameter under central reflection symmetry ($x \mapsto -x$), and at the true correlation matrix under elliptical symmetry, while uncovering a persistent scale bias when estimating the full covariance matrix. These findings led the authors to pose the following general question:

> "*Finally, an intriguing conjecture, also deserving of study, is that for each symmetry of p, there is a corresponding statistic that VI under mild conditions can exactly recover.*"

In subsequent work, Margossian et al. (2025b) extended these location-scale recovery guarantees from the standard variational objective $\mathrm{KL}(q\|p)$ to general f-divergences without requiring log-concavity. In the present paper, we address this conjecture in full generality by developing a Lie-theoretic and representation-theoretic framework for transformation-based variational inference. While our first-order symmetry constraint (Theorem 6.2) relies only on diffeomorphism invariance of the divergence and therefore applies broadly to any invariant loss functional, working explicitly with the KL divergence $\mathrm{KL}(q_g\|p)$ allows us to derive exact Stein-type moment-matching conditions that govern the remaining non-stationary directions. Our analysis reveals precisely how the algebraic compatibility between the symmetry group of the distribution and the Lie group of variational transports determines which parameter components are recovered without bias and which remain susceptible to model misspecification.

Concretely, we study the interaction between two group actions on the sample space $\mathcal{X}$: a compact *symmetry group* H acting by isometries, and a *transport Lie group* G acting by diffeomorphisms to generate the variational family. This formulation connects classical statistical transformation models (Eaton, 1989) and Lie group representation theory (Fulton and Harris, 1991; Hall, 2015; Knapp, 2002) with information geometry (Amari, 2016; Murray and Rice, 1993) and Riemannian optimization on manifolds (Absil et al., 2008; Boumal, 2023).

To formalize the estimation problem, we consider variational families obtained by pushing forward a fixed reference distribution $q_0 \in \mathcal{P}(\mathcal{X})$ under the action of G: if $Z \sim q_0$, then $X = g \cdot Z \sim q_g$, yielding the variational optimization problem

$$\min_{g \in G} \mathrm{KL}(q_g\|p)\,.$$

On the data side, we assume that the target distribution p is generated by transporting an unknown standardized distribution $\tilde{p}$ by a true parameter $g^* \in G$ that we wish to estimate; equivalently, $\tilde{p}$ is the *pullback* of p under g^*, meaning that if $X \sim p$, then the standardized variable $\xi = (g^*)^{-1} \cdot X$ follows $\tilde{p}$. Importantly, we do not assume that $\tilde{p}$ coincides with q_0, so the variational family $\mathcal{Q}_G = \{q_g : g \in G\}$ may be misspecified ($p \notin \mathcal{Q}_G$). Instead, we assume a shared *structural symmetry*: both the variational base q_0 and the standardized target $\tilde{p}$ are invariant under the action of the symmetry group H. Geometrically, this means that q_0 and $\tilde{p}$ share the same notion of centering or alignment around the origin, and the statistical question is whether this shared invariance suffices to make g^* a stationary point (and local minimizer) of the variational loss $g \mapsto \mathrm{KL}(q_g\|p)$.

Main contributions. We characterize the first-order geometry of transport variational inference at g^* through the conjugation action of H on G and its linearization, the adjoint representation of H on the Lie algebra $\mathfrak{g} = T_eG$:

1. **A Lie-theoretic framework for symmetric transport VI** (Sections 3–4). We formulate a general statistical framework pairing a compact symmetry group H with an H-normalized transport Lie group G. Analyzing the total transformation group $K = GH$ and the intersection subgroup $N = G \cap H$ allows us to separate parameter identifiability on the coset manifold $\mathcal{M} = G/N$ from first-order estimation bias, unifying location, scale, directional, and permutation models under a single geometric theory.
2. **Gradient constraint and orthogonal Lie algebra decomposition** (Theorem 6.2 and Proposition 6.3). We prove that the Riemannian gradient $\nabla_G \mathcal{L}(g^*)$ is constrained to lie in the trivial representation $\mathfrak{g}^H = \{u \in \mathfrak{g} : \mathrm{Ad}_h u = u \ \forall h \in H\}$ of the adjoint action of H on $\mathfrak{g}$. Combining this invariance with the identifiability structure yields an orthogonal decomposition of the transport Lie algebra,
$$\mathfrak{g} = \mathfrak{n} \ \oplus \ \mathfrak{g}^H_{\text{bias}} \ \oplus \ \mathfrak{g}^\perp,$$
into an *identifiability kernel* $\mathfrak{n}$ (along which the loss is locally constant), a *bias subspace* $\mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H \cap \mathfrak{n}^\perp$ (consisting of symmetry-preserving perturbations where model mismatch can induce a nonzero gradient), and a *symmetry-breaking subspace* $\mathfrak{g}^\perp = (\mathfrak{g}^H)^\perp \cap \mathfrak{n}^\perp$ (along which the gradient is forced to vanish by symmetry alone). Whenever $\mathfrak{g}^H_{\text{bias}} = \{0\}$, g^* is guaranteed to be a first-order stationary point regardless of model misspecification between q_0 and $\tilde{p}$.
3. **Stein-type moment-matching condition and debiasing recipe** (Theorem 6.5). When the bias subspace $\mathfrak{g}^H_{\text{bias}}$ is nontrivial, we derive an exact necessary and sufficient condition for g^* to be a critical point. Using integration by parts along the infinitesimal generators V_u of the group action, we show that the gradient along $u \in \mathfrak{g}^H_{\text{bias}}$ vanishes if and only if the directional scores of the log-potentials of q_0 and $\tilde{p}$ match in expectation under q_0. This characterization provides a constructive calibration procedure to eliminate estimation bias by rescaling or shifting the base distribution q_0 along $\exp(\mathfrak{g}^H_{\text{bias}})$ while preserving its H-invariance.

Organization. The remainder of the paper is organized as follows. Section 2 situates our work within the literature on variational inference, group invariance in statistics, and Riemannian optimization. Section 3 introduces transport variational families, H-conjugate symmetry, and the structural normalization condition, together with six guiding examples that we trace throughout the paper. Section 4 analyzes the total transformation group $K = GH$, the conjugation homomorphism, and the three resulting identifiability regimes on the quotient space $\mathcal{M} = G/N$. Section 5 develops the required differential calculus on G and introduces the adjoint representation Ad_H and its trivial subspace $\mathfrak{g}^H$. Section 6 presents our main theoretical results: the gradient constraint theorem, the orthogonal decomposition of $\mathfrak{g}$, and the moment-matching theorem. Section 7 discusses gauge-fixing vs. base calibration, second-order curvature conditions, and extensions. Technical proofs and detailed example derivations are deferred to the Appendices.

2. Related work

Variational inference and transport-based families. The variational formulation of approximate inference, rooted in statistical physics and convex duality, was brought to prominence in statistics and machine learning by Jordan et al. (1999), Wainwright and Jordan (2008), and Blei et al. (2017). The development of black-box variational inference and reparameterization gradients (Ranganath et al., 2014; Rezende and Mohamed, 2015) subsequently enabled optimization over general continuous families where expectations under q_g are evaluated by pushing forward samples $Z \sim q_0$ through a differentiable transformation $x = g(Z)$. In the normalizing flow literature (Kobyzev et al., 2021; Papamakarios et al., 2021; Rezende and Mohamed, 2015), the transformation g is typically parameterized by a deep composition of general diffeomorphisms. Our framework studies the structured setting in which the transport maps form a finite-dimensional Lie group G acting on $\mathcal{X}$: this encompasses

classical parametric families (such as location, scale, and directional families) while endowing the variational parameter space with an explicit Lie algebra that makes global and local symmetry analysis tractable.

Group invariance in mathematical statistics and machine learning. Group actions have long served as a unifying language in theoretical statistics (Eaton, 1989; Lehmann and Romano, 2005; Wijsman, 1990). In classical statistical decision theory, when a family of distributions is closed under a group action, one seeks estimators that are *equivariant*—such as the Pitman estimator for location or invariant Bayes estimators under right Haar measure—and hypothesis tests that are invariant with respect to nuisance group actions. In those classical works, the statistical model is typically well-specified and the group acts transitively on the parameter space to reduce risk. Our perspective differs in two important respects: first, we study the deterministic optimization landscape of a divergence functional under model misspecification ($p \notin \mathcal{Q}_G$); second, the symmetry group H of the standardized distribution and the transport group G indexing the variational family need not coincide, and their interaction via conjugation governs which parameters are unbiasedly estimated. In a complementary direction, geometric deep learning and equivariant neural networks (Bronstein et al., 2021; Cohen and Welling, 2016; Weiler and Cesa, 2019) build group symmetries directly into function approximators and normalizing flows (Köhler et al., 2020); rather than constraining the variational family to produce strictly H-invariant distributions (which would preclude estimating symmetry-breaking parameters g^*), our work uses H-invariant base distributions pushed forward by G to estimate the unknown transformation g^*.

Information geometry and optimization on Lie groups. Differential-geometric approaches to statistical estimation have been extensively studied in information geometry (Amari, 2016; Murray and Rice, 1993), where the Fisher–Rao metric defines a natural Riemannian structure on parametric families, and in optimization on smooth manifolds and Lie groups (Absil et al., 2008; Boumal, 2023). When a variational family is indexed by a Lie group G, left-trivialization identifies the tangent space T_gG at any parameter g with the Lie algebra $\mathfrak{g} = T_eG$. By equipping $\mathfrak{g}$ with an Ad_H-invariant inner product (guaranteed by the compactness of H), we connect Riemannian first-order stationarity directly to the decomposition of $\mathfrak{g}$ into irreducible representations of H.

Symmetry in variational inference. The most direct predecessors of the present work are Margossian and Saul (2025) and Margossian et al. (2025b), who proved that location-scale variational inference exactly recovers the target mean under even symmetry and the target correlation matrix under elliptical symmetry across general f-divergences, and Zellinger and Vergari (2026), who extended even-symmetry location recovery to forward KL and α-divergences without log-concavity assumptions. Complementing these positive results, Margossian et al. (2025a) established an impossibility theorem for factorized (mean-field) variational approximations of non-factorized targets, proving that a mean-field family can accurately recover at most one of three canonical dispersion summaries (marginal variances, marginal precisions, or generalized variance). This impossibility underscores the necessity of non-factorized group-transport families such as full $\mathrm{GL}(d)$ covariance models. Concurrently with our work, Marks et al. (2026) developed a global theory of symmetry-induced statistic recovery in variational inference, characterizing when global minimizers inherit target symmetries and applying their framework to von Mises–Fisher distributions on the sphere, while Ghafourpour et al. (2026) investigated radial transport families. Our differential-geometric approach is directly complementary: by linearizing the conjugation action of H on G via the adjoint representation, we provide an explicit subspace decomposition of the transport Lie algebra $\mathfrak{g}$ that simultaneously explains both why symmetry-breaking gradient directions vanish and why symmetry-preserving directions ($\mathfrak{g}^H_{\text{bias}}$) exhibit estimation bias unless calibrated by moment matching.

3. Problem description

In parametric variational inference (Blei et al., 2017; Ranganath et al., 2014; Wainwright and Jordan, 2008), one approximates an intractable target distribution $p \in \mathcal{P}(\mathcal{X})$ on a sample space $\mathcal{X}$ by selecting a candidate distribution q_θ from a parameterized family $\mathcal{Q}_\Theta = \{q_\theta \in \mathcal{P}(\mathcal{X}) : \theta \in \Theta\}$ that minimizes a divergence functional over $\theta \in \Theta$. In this paper, we study structured variational families generated by the smooth action of a *transport* Lie group G on $\mathcal{X}$, in the presence of a *symmetry* group H acting isometrically on $\mathcal{X}$. For any $g \in G$ (respectively, $h \in H$) and any point $x \in \mathcal{X}$, we denote the group action on the sample space by $g \cdot x \in \mathcal{X}$ (respectively, $h \cdot x \in \mathcal{X}$).

Viewing both G and H through their actions on the common domain $\mathcal{X}$ serves two fundamental purposes. First, any measurable action on $\mathcal{X}$ canonically induces a pushforward action on probability measures over $\mathcal{X}$: transforming a random variable $Z \sim q_0$ via $X = g \cdot Z$ transports its law to $q_g = g_\sharp q_0$. Second, because both G and H act as transformations of $\mathcal{X}$, they can be composed within the ambient group $\mathrm{Aut}(\mathcal{X})$ of diffeomorphisms of $\mathcal{X}$, even when G and H are defined a priori as distinct algebraic groups. As we show throughout the paper, the non-commutativity of these two actions on $\mathcal{X}$—captured algebraically by the conjugation of transports in G by symmetries in H—is the central mechanism governing parameter identifiability and unbiased estimation in variational inference.

3.1. Group transport variational inference

We begin by formalizing variational families generated by pushing forward a fixed reference measure q_0 under a Lie group action. From a statistical standpoint, this construction generalizes classical location-scale transformation models (Eaton, 1989) to arbitrary smooth manifolds and Lie groups; from a machine learning standpoint, it corresponds to a normalizing flow (Papamakarios et al., 2021; Rezende and Mohamed, 2015) whose transformation maps are structured as a Lie group G. Throughout the paper, we restrict attention to distributions admitting strictly positive, smooth (or piecewise smooth, locally Lipschitz) densities with respect to the canonical reference measure on $\mathcal{X}$ (Lebesgue measure when $\mathcal{X} = \mathbb{R}^d$, or the Riemannian volume measure when $\mathcal{X}$ is a Riemannian manifold), and we identify each probability measure with its density.

Definition 3.1 (Transport variational inference)**.** Let $\mathcal{X}$ be a smooth *data domain* (such as Euclidean space $\mathbb{R}^d$ or a smooth Riemannian manifold) and let G be a Lie group acting smoothly and effectively on $\mathcal{X}$ by diffeomorphisms, written $(g, x) \mapsto g \cdot x \in \mathcal{X}$ (where acting *effectively* simply means that distinct group elements $g \neq g'$ define distinct transformations on $\mathcal{X}$). We call G the *parameter transport group*. Given a base distribution $q_0 \in \mathcal{P}(\mathcal{X})$ and a target data distribution $p \in \mathcal{P}(\mathcal{X})$, both with positive densities on $\mathcal{X}$, we define:

- The *transported distribution* $q_g = g_\sharp q_0$: the law of $X = g \cdot Z$ when $Z \sim q_0$, whose density is given by the change-of-variables formula

$$q_g(x) = q_0(g^{-1} \cdot x)\, |\det J_{g^{-1}}(x)|\,,$$

  where $J_{g^{-1}}(x)$ denotes the Jacobian of the inverse map $x \mapsto g^{-1} \cdot x$ with respect to the reference volume measure on $\mathcal{X}$ (in particular, $|\det J_{g^{-1}}(x)| = 1$ whenever G acts by isometries).
- The *transport variational family* $\mathcal{Q}_G = \{q_g \in \mathcal{P}(\mathcal{X}) \mid g \in G\}$.
- The *variational inference objective* $\mathcal{L} : G \to \mathbb{R}$, defined via the Kullback–Leibler divergence:

$$\mathcal{L}(g) = \mathrm{KL}(q_g \| p) = \mathbf{E}_{x \sim q_g}\left[\log q_g(x) - \log p(x)\right] .$$

Illustrative examples of transport families. To ground Definition 3.1, consider three classical parametric families obtained as group pushforwards of a standard base distribution q_0:

- **Gaussian location family**: On $\mathcal{X} = \mathbb{R}^d$, let $G = (\mathbb{R}^d, +)$ act by translation $g \cdot x = x + \mu_g$. Taking the standard Gaussian base $q_0 = \mathcal{N}(0, I_d)$ yields the homoscedastic location family

$$Q_G^{\text{mean}} = \{\mathcal{N}(\mu_g, I_d) \mid \mu_g \in \mathbb{R}^d\}\,.$$

- **Gaussian scale/covariance family**: On $\mathcal{X} = \mathbb{R}^d$, let the general linear group $G = \mathrm{GL}(d)$ act linearly by $g \cdot x = A_g x$. If $Z \sim q_0 = \mathcal{N}(0, I_d)$, then $A_g Z \sim \mathcal{N}(0, A_g A_g^\intercal)$, generating the centered Gaussian covariance family

$$Q_G^{\text{cov}} = \{\mathcal{N}(0, A_g A_g^\intercal) \mid A_g \in \mathrm{GL}(d)\}\,.$$

- **Directional von Mises–Fisher family on the sphere**: On the unit sphere $\mathcal{X} = \mathcal{S}^2 \subset \mathbb{R}^3$, let $G = \mathrm{SO}(3)$ act by rotations $g \cdot x = R_g x$. Taking a von Mises–Fisher base distribution $q_0(x) = c_\kappa \exp(\kappa e_3^\intercal x)$ ($\kappa > 0$) concentrated around the polar axis $e_3 = (0, 0, 1)^\intercal$ (the North pole) yields the rotational family

$$Q_G^{\text{ax}} = \{c_\kappa \exp(\kappa v_g^\intercal x) \mid v_g = R_g e_3,\ R_g \in \mathrm{SO}(3)\}\,,$$

  indexed by the transported modal direction $v_g = R_g e_3 \in \mathcal{S}^2$.

When the target distribution p is itself generated by transporting a centered distribution $\tilde{p}$ via an unknown true transformation $g^* \in G$, a fundamental statistical question is whether minimizing $\mathcal{L}(g)$ recovers g^*—or, at a first-order differential level, whether g^* is a stationary point of $\mathcal{L}$. Because G is a smooth manifold rather than a flat vector space in general, first-order stationarity is characterized through the *Riemannian gradient* $\nabla_G \mathcal{L}(g^*)$, which is an element of the Lie algebra $\mathfrak{g} = T_e G$ tangent to G at the identity (defined formally in Section 5). By left-trivialization, $\nabla_G \mathcal{L}(g)$ governs the first-order variation of $\mathcal{L}$ along one-parameter subgroup perturbations $t \mapsto g \exp(tu)$ for $u \in \mathfrak{g}$:

$$\mathcal{L}\big(g \cdot \exp(tu)\big) = \mathcal{L}(g) + t\langle \nabla_G \mathcal{L}(g), u\rangle_{\mathfrak{g}} + o(t)\,.$$

Thus, g^* is a first-order critical point of the variational loss if and only if $\nabla_G \mathcal{L}(g^*) = 0$ in $\mathfrak{g}$.

Remarks on pushforwards, overparametrization, and notation.

- *Pushforward versus pullback*: In measure-theoretic terminology, the transported distribution $q_g = g_\sharp q_0$ is the *pushforward* of q_0 by the action of g. Conversely, standardizing a distribution p by applying the inverse transformation $(g^*)^{-1}$ defines its *pullback* $\tilde{p} = (g^*)^{-1}_\sharp p$ (Definition 3.2).
- *Overparametrization and coset identifiability*: In several natural families—such as the Gaussian covariance family ($G = \mathrm{GL}(d)$) and the spherical axis family ($G = \mathrm{SO}(3)$)—the map $g \mapsto q_g$ is not injective: distinct group elements $g \neq g'$ can produce the exact same distribution $q_g = q_{g'}$ (for instance, whenever $A_g A_g^\intercal = A_{g'} A_{g'}^\intercal$, which occurs when $A_{g'} = A_g U$ for $U \in \mathrm{O}(d)$). In group-theoretic terminology, the set of transformations $n \in G$ that leave the base distribution unchanged ($q_n = q_0$) forms a subgroup called the *stabilizer* of q_0, and the equivalence class of all parameters $\{gn : n \text{ in the stabilizer}\}$ that generate the same distribution q_g is called a *left coset*. In such settings, g^* is identifiable only up to this stabilizer subgroup, while the induced distribution q_{g^*} and the coset parameter on the quotient space (Section 4) remain well-defined.
- *Notation*: Where no ambiguity arises, we identify a group element $g \in G$ directly with its concrete transformation parameter, writing $\mu \in \mathbb{R}^d$ for translations, $A \in \mathrm{GL}(d)$ for invertible matrices, or $R \in \mathrm{SO}(d)$ for rotations.

3.2. Distribution symmetries

To understand when variational inference can recover g^* despite functional misspecification between q_0 and the target distribution, we formalize the notion of distributional invariance under a group H of

isometries of $\mathcal{X}$, and define *conjugate symmetry* around a transport parameter $g^* \in G$. Statistically, conjugate symmetry expresses the familiar modeling assumption that a random variable $X \sim p$ becomes symmetric around the origin after being standardized by $(g^*)^{-1}$, i.e., the standardized variable $\xi = (g^*)^{-1} \cdot X \sim \tilde{p}$ has an H-invariant distribution.

Definition 3.2 (H-symmetry and conjugate symmetry)**.** Let H be a non-trivial compact group acting effectively on $\mathcal{X}$ by isometries (written linearly as $h \cdot x = M_h x$ with $M_h \in \mathrm{O}(d)$ when $\mathcal{X} \subseteq \mathbb{R}^d$).

- A distribution $q \in \mathcal{P}(\mathcal{X})$ is *H-symmetric* (or *H-invariant*) if $q(h \cdot z) = q(z)$ for all $h \in H$ and $z \in \mathcal{X}$.
- Given a target distribution $p \in \mathcal{P}(\mathcal{X})$ and a transport element $g^* \in G$, the *pullback distribution* $\tilde{p} = (g^*)_\sharp^{-1} p$ is the law of the standardized variable $\xi = (g^*)^{-1} \cdot X$ when $X \sim p$, with density

$$\tilde{p}(\xi) = p(g^* \cdot \xi)\,|\det J_{g^*}(\xi)|\,.$$

- A distribution $p \in \mathcal{P}(\mathcal{X})$ has *H-conjugate symmetry around $g^* \in G$* if its pullback $\tilde{p}$ under g^* is H-symmetric, meaning that for all $h \in H$ and $\xi \in \mathcal{X}$,

$$\tilde{p}(h \cdot \xi) = \tilde{p}(\xi)\,.$$

By construction, the pullback operation inverts the transport pushforward: if $X \sim p$, then $(g^*)^{-1} \cdot X \sim \tilde{p}$, and conversely if $\xi \sim \tilde{p}$, then $g^* \cdot \xi \sim p$. Thus, p has H-conjugate symmetry around g^* precisely when p is obtained by transporting an H-symmetric reference distribution $\tilde{p}$ by g^*.

Examples of H-symmetry and conjugate symmetry.

- *Central reflection symmetry around a location parameter*: Let $\mathcal{X} = \mathbb{R}^d$, let $H = \mathbb{Z}/2\mathbb{Z}$ act by central reflection $\{I_d, -I_d\}$, and let $G = (\mathbb{R}^d, +)$ act by translation $x \mapsto x + \mu$. A distribution $q \in \mathcal{P}(\mathbb{R}^d)$ is H-symmetric if and only if it is even, $q(-z) = q(z)$. For a target distribution $p \in \mathcal{P}(\mathbb{R}^d)$ and a location parameter $\mu^* \in \mathbb{R}^d$, the pullback is the centered density $\tilde{p}(\xi) = p(\mu^* + \xi)$, corresponding to the centered random variable $\xi = X - \mu^*$. Therefore, p has H-conjugate symmetry around μ^* if and only if $\tilde{p}(\xi) = \tilde{p}(-\xi)$, or equivalently $p(\mu^* + \xi) = p(\mu^* - \xi)$ for all $\xi \in \mathbb{R}^d$. This recovers the classical setting of location estimation for centrally symmetric distributions studied in Margossian and Saul (2025).
- *Canonical isometry groups in statistics*: Beyond central reflection $\mathbb{Z}/2\mathbb{Z}$, several compact isometry groups H arise naturally in statistical modeling: the full orthogonal group $\mathrm{O}(d)$ on $\mathbb{R}^d$ (characterizing spherical and radial distributions, or elliptical distributions after affine standardization); the symmetric group $\mathbb{S}_n$ acting on $\mathbb{R}^n$ by coordinate permutation (characterizing exchangeable random vectors); the dihedral groups D_k on $\mathbb{R}^2$ (characterizing discrete rotational and reflectional symmetries, such as ℓ_m-norm contours); and the planar rotation group $\mathrm{SO}(2) \subset \mathrm{SO}(3)$ fixing a polar axis e_3 on the sphere $\mathcal{S}^2$ (characterizing axially symmetric directional distributions).

Remark (Geometric meaning of conjugate symmetry). Written directly in terms of the target density p without reference to the pullback $\tilde{p}$, H-conjugate symmetry around g^* states that for every $h \in H$ and $x \in \mathcal{X}$,

$$p\big((g^* \circ h \circ (g^*)^{-1}) \cdot x\big)\,\big|\det J_{g^* \circ h \circ (g^*)^{-1}}(x)\big| = p(x)\,.$$

That is, while the standardized distribution $\tilde{p}$ is invariant under the base symmetry group H centered at the origin, transporting $\tilde{p}$ by g^* shifts its symmetries to the *conjugated subgroup* $g^* H (g^*)^{-1}$. Consequently, specifying a symmetric variational inference problem involves an inherently coupled modeling choice between the symmetry group H of the standardized data and the transport group G used to relocate the distribution. We formalize this coupling next.

3.3. Transport variational inference with symmetry

We now combine the transport family of Definition 3.1 with the symmetry structure of Definition 3.2. Suppose that the variational base distribution q_0 is H-symmetric, and that the target distribution p has H-conjugate symmetry around an unknown true parameter $g^* \in G$. Even when the functional form of q_0 differs substantially from that of the pullback $\tilde{p}$ (for instance, when q_0 is Gaussian while $\tilde{p}$ is heavy-tailed or multimodal), both distributions share the same invariance group H. Our central question is whether this shared symmetry suffices for variational inference to identify g^*, and specifically whether $\nabla_G \mathcal{L}(g^*) = 0$.

Margossian and Saul (2025) analyzed two special instances of this problem (corresponding to Examples 1 and 4 below), showing that the variational gradient vanishes at the true location parameter under reflection symmetry, but exhibits a nonzero radial scaling bias when estimating a covariance matrix under elliptical symmetry. Our goal is to establish a general theory covering arbitrary pairs (G, H) that explains both phenomena and characterizes $\nabla_G \mathcal{L}(g^*)$ completely.

Definition 3.3 (Transport variational inference with symmetry)**.** In the setting of Definition 3.1, let H be a non-trivial compact group acting effectively on $\mathcal{X}$ by isometries (the *symmetry group*). Assume that:

(i) the variational base distribution $q_0 \in \mathcal{P}(\mathcal{X})$ is H-symmetric (with stabilizer $\mathrm{Stab}_G(q_0) = G \cap H$ in G);
(ii) the target distribution $p \in \mathcal{P}(\mathcal{X})$ has H-conjugate symmetry around a true parameter $g^* \in G$; and
(iii) the transport group G is *normalized* by H within $\mathrm{Aut}(\mathcal{X})$, meaning that $\Phi_h(g) := h \circ g \circ h^{-1} \in G$ for all $g \in G$ and $h \in H$—in plain terms, expressing any parameter transformation $g \in G$ in a coordinate system rotated or reflected by $h \in H$ yields another valid parameter transformation inside G.

Symmetric transport variational inference then consists in solving the optimization problem

$$\min_{g \in G} \mathcal{L}(g) = \min_{g \in G} \mathrm{KL}(q_g \| p) \,.$$

Remark (Why G must be normalized by H). Condition (iii) in Definition 3.3 requires that conjugating any transport $g \in G$ by any symmetry $h \in H$ produces a transformation $\Phi_h(g) = h \circ g \circ h^{-1}$ that still belongs to the variational transport group G. This structural consistency condition has a clear geometric and statistical motivation.

Consider perturbing the variational parameter away from g^* by a relative transport $g \in G$, evaluating $\mathcal{L}(g^* g) = \mathrm{KL}(q_g \| \tilde{p})$. Because both q_0 and $\tilde{p}$ are H-invariant, rotating the perturbation by any symmetry $h \in H$ produces a symmetrically equivalent perturbation $h \circ g \circ h^{-1}$ that achieves the exact same KL divergence against $\tilde{p}$. However, for this symmetry of the objective to constrain the gradient within our parameter space G, the rotated perturbation $h \circ g \circ h^{-1}$ must actually lie in G. If $h \circ g \circ h^{-1} \notin G$, the variational family truncates the symmetry orbit of the perturbation. For example, if $\tilde{p}$ is rotationally invariant in the plane ($H = \mathrm{SO}(2)$) but the variational family only allows horizontal translations along the x_1-axis, rotating a horizontal translation by 90° produces a vertical translation along the x_2-axis, which lies outside G. Requiring G to be normalized by H ensures that the variational parameter space is closed under the symmetry transformations we seek to exploit, and guarantees that the quotient parameter space $\mathcal{M} = G/(G \cap H)$ (Section 4) and the adjoint representation on the Lie algebra $\mathfrak{g}$ (Section 5) are well-defined.

Guiding examples. Throughout the paper, we trace six concrete examples that span location estimation, directional statistics, covariance estimation, non-Gaussian orientation estimation, and exchangeable models. These examples allow every algebraic object in our theory to be interpreted directly in familiar statistical terms:

1. **Reflection (Univariate symmetric location model)**: On the real line $\mathcal{X} = \mathbb{R}$, let $H = \mathbb{Z}/2\mathbb{Z} = \{1, -1\}$ act by sign reflection $x \mapsto \pm x$, and let $G = (\mathbb{R}, +)$ act by translation $x \mapsto x + \mu$. Choosing the standard normal base $q_0 = \mathcal{N}(0, 1)$ generates the unit-variance Gaussian location family $Q_{\mathbb{R}} = \{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}$. Any target distribution p that is symmetric around a true location $\mu^* \in \mathbb{R}$—such as a symmetric bimodal Gaussian mixture $p = \frac{1}{2}\mathcal{N}(\mu^* + \delta, 1) + \frac{1}{2}\mathcal{N}(\mu^* - \delta, 1)$, a Student-$t$ distribution centered at μ^*, or a Laplace distribution—has H-conjugate symmetry around μ^* even though $p \notin Q_{\mathbb{R}}$.
2. **Euclidean (Multivariate radially symmetric location model)**: On $\mathcal{X} = \mathbb{R}^d$ ($d \geq 2$), let $H = \mathrm{O}(d)$ be the orthogonal group of rotations and reflections around the origin, and let $G = (\mathbb{R}^d, +)$ act by translation $x \mapsto x + \mu$. With $q_0 = \mathcal{N}(0, I_d)$, the variational family is $Q_{\mathbb{R}^d} = \{\mathcal{N}(\mu, I_d) \mid \mu \in \mathbb{R}^d\}$, while the target $p(x) = \tilde{p}(x - \mu^*)$ can be any radially symmetric distribution $\tilde{p}(\xi) = f(\|\xi\|_2)$ shifted to center $\mu^* \in \mathbb{R}^d$.
3. **Sphere (Directional estimation for axially symmetric distributions)**: On the unit sphere $\mathcal{X} = \mathcal{S}^2 \subset \mathbb{R}^3$, let $H = \mathrm{SO}(2)$ be the subgroup of $\mathrm{SO}(3)$ consisting of rotations around the polar axis $e_3 = (0, 0, 1)^\top$ (the North pole), and let $G = \mathrm{SO}(3)$ act by 3D rotations $x \mapsto Rx$. Using a von Mises–Fisher base distribution $q_0(x) = c_\kappa \exp(\kappa e_3^\top x)$ ($\kappa > 0$) produces the variational family $Q_{\mathrm{SO}(3)} = \{q_R(x) = c_\kappa \exp(\kappa (Re_3)^\top x) \mid R \in \mathrm{SO}(3)\}$, which is effectively indexed by the modal direction $v = Re_3 \in \mathcal{S}^2$. The target $p(x) = \tilde{p}((R^*)^\top x)$ is obtained by rotating any rotationally symmetric spherical distribution $\tilde{p}$ around e_3 so that its axis of symmetry aligns with $v^* = R^* e_3$.
4. **Covariance (Elliptical scatter and covariance estimation)**: On $\mathcal{X} = \mathbb{R}^d$ ($d \geq 2$), let $H = \mathrm{O}(d)$ act by orthogonal transformations and let $G = \mathrm{GL}(d)$ act linearly by $x \mapsto Ax$. Taking $q_0 = \mathcal{N}(0, I_d)$ yields the zero-mean Gaussian family $Q^{\mathrm{cov}}_{\mathrm{GL}(d)} = \{\mathcal{N}(0, AA^\top) \mid A \in \mathrm{GL}(d)\}$, which parameterizes all symmetric positive-definite covariance matrices $\Sigma = AA^\top \in \mathrm{SPD}(d)$. If the target $p(x) = |\det A^*|^{-1} \tilde{p}((A^*)^{-1}x)$ is generated by applying an invertible linear transformation A^* to any spherically symmetric distribution $\tilde{p}$ (such as a multivariate t or Kotz-type elliptical distribution), then p is an elliptical distribution with scatter structure $\Sigma^* = A^*(A^*)^\top$ and possesses $\mathrm{O}(d)$-conjugate symmetry around A^*.
5. **Square (Orientation estimation under discrete dihedral symmetry)**: On the plane $\mathcal{X} = \mathbb{R}^2$, let $H = D_4$ be the dihedral symmetry group of the square (generated by 90° rotations and coordinate-axis reflections), and let $G = \mathrm{SO}(2)$ act by continuous planar rotations R_θ. A natural D_4-invariant base distribution is the generalized Gaussian $q_0(x) \propto \exp(-\gamma_m \|x\|_m^m)$ with $m > 2$ (for example, $m = 4$), whose level sets are ℓ_m-balls with fourfold corners. Rotating q_0 by $R_\theta \in \mathrm{SO}(2)$ generates the orientation family $Q^{\mathrm{square}}_{\mathrm{SO}(2)} = \{q_0(R_\theta^\top x) \mid R_\theta \in \mathrm{SO}(2)\}$ modeling distributions whose square-symmetric contours are tilted by an angle θ, against a target $p(x) = \tilde{p}(R_{\theta^*}^\top x)$ tilted by a true angle θ^*.
6. **Permutation (Location estimation for exchangeable coordinates)**: On $\mathcal{X} = \mathbb{R}^n$, let $H = \mathbb{S}_n$ be the symmetric group of coordinate permutations $(x_1, \ldots, x_n) \mapsto (x_{\sigma(1)}, \ldots, x_{\sigma(n)})$, which characterizes exchangeable random variables, and let $G = (\mathbb{R}^n, +)$ act by coordinate-wise translation $x \mapsto x + \mu$. For instance, taking an exchangeable product Laplace base $q_0(x) = c_\kappa \exp(-\kappa \|x\|_1)$ yields the location family $Q^{\mathrm{perm}}_{\mathbb{R}^n} = \{c_\kappa \exp(-\kappa \|x - \mu\|_1) \mid \mu \in \mathbb{R}^n\}$, while the target $p(x) = \tilde{p}(x - \mu^*)$ is obtained by shifting any exchangeable distribution $\tilde{p}$ (potentially with dependent coordinates, such as an equicorrelated normal or exchangeable mixture) by a location vector $\mu^* \in \mathbb{R}^n$.

Table 1 summarizes the sample spaces, transport groups, and symmetry groups for these six guiding

examples.

Table 1 | Summary of sample spaces, transport groups, and symmetry groups for the six guiding examples.

Example	Data Domain ($\mathcal{X}$)	Transport Group (G)	Symmetry Group (H)	Symmetry Action
1. Reflection	$\mathbb{R}$	$(\mathbb{R}, +)$	$\mathbb{Z}/2\mathbb{Z}$	Sign flip ($x \mapsto -x$)
2. Euclidean	$\mathbb{R}^d$	$(\mathbb{R}^d, +)$	$\mathrm{O}(d)$	Rotations / reflections
3. Sphere	$\mathcal{S}^2 \subset \mathbb{R}^3$	$\mathrm{SO}(3)$	$\mathrm{SO}(2)$	Axial rotations around e_3
4. Covariance	$\mathbb{R}^d$	$\mathrm{GL}(d)$	$\mathrm{O}(d)$	Orthogonal basis change
5. Square	$\mathbb{R}^2$	$\mathrm{SO}(2)$	D_4	90° rot. & axis flips
6. Permutation	$\mathbb{R}^n$	$(\mathbb{R}^n, +)$	$\mathbb{S}_n$	Coordinate permutations

4. Group interactions and identifiability

Before analyzing first-order stationarity of the variational loss $\mathcal{L}(g) = \mathrm{KL}(q_g \| p)$, we must address a foundational statistical question: *which parameters are identifiable from the variational family* $\mathcal{Q}_G = \{q_g : g \in G\}$*?* Because the base distribution q_0 is invariant under the symmetry group H, any transport $n \in G$ that also belongs to H leaves q_0 unchanged ($q_n = q_0$), and more generally right-multiplying any $g \in G$ by $n \in G \cap H$ produces the identical distribution $q_{gn} = q_g$. Furthermore, even when $h \in H$ does not belong to G, composing transports in G with symmetries in H on the sample space $\mathcal{X}$ generates a larger ambient group of transformations. In this section, we formalize this ambient group $K = GH$, analyze the conjugation action $\Phi_h(g) = h \circ g \circ h^{-1}$ across our six guiding examples, and classify the resulting statistical identifiability regimes on the quotient manifold $\mathcal{M} = G/(G \cap H)$.

4.1. Total transformation group and semidirect product structure

When G and H both act on the sample space $\mathcal{X}$, their compositions generate a combined group of transformations inside $\mathrm{Aut}(\mathcal{X})$. The algebraic structure of this total group is governed by the normalization condition $\Phi_h(g) = h \circ g \circ h^{-1} \in G$ introduced in Definition 3.3.

Definition 4.1 (Total transformation group)**.** Let $K = GH \subseteq \mathrm{Aut}(\mathcal{X})$ denote the total group of transformations on $\mathcal{X}$ generated by composing actions of G and H. Every element $k \in K$ can be written as $k = g \circ h$ for some $g \in G$ and $h \in H$, acting on $x \in \mathcal{X}$ by applying the symmetry h first and the transport g second:

$$k \cdot x = g \cdot (h \cdot x)\,.$$

Because G is normalized by H, the composition of any two elements $k_2 = g_2 \circ h_2$ and $k_1 = g_1 \circ h_1$ in K remains in GH and obeys the semidirect multiplication law

$$(g_2 \circ h_2) \circ (g_1 \circ h_1) = \big(g_2 \circ \Phi_{h_2}(g_1)\big) \circ (h_2 \circ h_1)\,,$$

where $\Phi_h(g) = h \circ g \circ h^{-1} \in G$ denotes the conjugation of g by h. Depending on the intersection subgroup $N = G \cap H$, the total group K takes one of three algebraic forms:

(i) When $N = G \cap H = \{e\}$ (disjoint actions), K is isomorphic to the internal semidirect product $G \rtimes_\Phi H$: every transformation $k \in K$ decomposes uniquely into a symmetry $h \in H$ followed by a transport $g \in G$, and composing two such transformations twists the first transport parameter g_1 by h_2 via $\Phi_{h_2}(g_1)$ (just as rotating an affine-shifted variable rotates its translation vector). For location families where $G = (\mathbb{R}^d, +)$ acts by translation and H acts linearly (Examples 1, 2, and 6), $K = \mathbb{R}^d \rtimes H$ is the group of affine isometries generated by H and translations (in particular, the full Euclidean group $\mathrm{E}(d) = \mathbb{R}^d \rtimes \mathrm{O}(d)$ in Example 2).

(ii) When $H \subseteq G$, so that $N = H$ (internal symmetries), every symmetry is already a transport transformation, and the total group collapses to $K = G$ (as in Example 3 where $K = \mathrm{SO}(3)$ and Example 4 where $K = \mathrm{GL}(d)$).

(iii) In the general overlapping case where $\{e\} \subsetneq N \subsetneq H$, K is isomorphic to the quotient $(G \rtimes_\Phi H)/N$, where N is embedded anti-diagonally via $n \mapsto (n, n^{-1})$ (as in Example 5, where $G = \mathrm{SO}(2)$ and $H = D_4$ intersect in $N = C_4$ and together generate the full planar orthogonal group $K = \mathrm{O}(2)$).

Remark (Closure under conjugation and well-posedness of K). Definition 4.1 underscores why the normalization requirement $\Phi_h(g) \in G$ from Section 3 is algebraically indispensable. Rewriting the composition $h_2 \circ g_1$ as $(h_2 \circ g_1 \circ h_2^{-1}) \circ h_2 = \Phi_{h_2}(g_1) \circ h_2$ allows any sequence of alternating transports and symmetries to be re-ordered into a single symmetry followed by a single transport in G. If G were not normalized by H, composing transports with symmetries would generate transformations outside GH, and differentiating the conjugation map at the identity (Section 5) would produce tangent vectors pointing outside the Lie algebra $\mathfrak{g}$ of the variational family.

4.2. The conjugation action in detail

The conjugation map $\Phi : H \to \mathrm{Aut}(G)$, $h \mapsto \Phi_h$, where $\Phi_h(g) = h \circ g \circ h^{-1}$, is the fundamental algebraic bridge connecting the symmetry group H to the transport group G. Before passing to the tangent space in Section 5, we establish the basic structural properties of Φ, examine its geometric meaning on $\mathcal{X}$, and compute its action, kernel, and fixed-point subgroup across all six guiding examples.

Proposition 4.2 (Conjugation as a group homomorphism)**.** *The map $\Phi : H \to \mathrm{Aut}(G)$, $h \mapsto \Phi_h$, is a group homomorphism: $\Phi_{h_1 h_2} = \Phi_{h_1} \circ \Phi_{h_2}$ for all $h_1, h_2 \in H$. Moreover, for each $h \in H$, the map $\Phi_h : G \to G$ is a Lie group automorphism, satisfying $\Phi_h(g_1 g_2) = \Phi_h(g_1)\Phi_h(g_2)$ and $\Phi_h(e_G) = e_G$.*

Proof. For any $h_1, h_2 \in H$ and $g \in G$, associativity of composition in $\mathrm{Aut}(\mathcal{X})$ yields

$$\Phi_{h_1 h_2}(g) = (h_1 h_2) \circ g \circ (h_1 h_2)^{-1} = h_1 \circ (h_2 \circ g \circ h_2^{-1}) \circ h_1^{-1} = \Phi_{h_1}(\Phi_{h_2}(g)) \,.$$

Similarly, for any $g_1, g_2 \in G$, inserting $h^{-1} \circ h = \mathrm{id}_{\mathcal{X}}$ gives

$$\Phi_h(g_1 g_2) = h \circ g_1 \circ (h^{-1} \circ h) \circ g_2 \circ h^{-1} = \Phi_h(g_1)\Phi_h(g_2) \,,$$

and $\Phi_h(e_G) = h \circ e_G \circ h^{-1} = e_G$. Smoothness of Φ_h follows from the smooth group structure of K. □

Two subgroups associated with the conjugation action Φ play a recurring role in our analysis:

- The *kernel of conjugation* (or *centralizer* of G in H),

$$\ker(\Phi) = \{h \in H : \Phi_h(g) = g \ \ \forall g \in G\} = C_H(G) \,,$$

  which consists of all symmetries in H that commute with every transport in G (meaning that transforming the coordinate system by h leaves every parameter $g \in G$ unchanged). When $\ker(\Phi) = \{e_H\}$, the conjugation action of H on G is faithful: every non-trivial symmetry induces a non-trivial reparametrization of the transport group.
- The *fixed-point subgroup* (or *H-centralizer* in G),

$$G^H = \{g \in G : \Phi_h(g) = g \ \ \forall h \in H\} \,,$$

  which consists of all transports in G that commute with every symmetry in H. Because each Φ_h is a group automorphism fixing e_G, G^H is a closed Lie subgroup of G. Statistically, G^H consists of those parameter transformations (such as global isotropic scaling) that do not break the H-symmetry of the model: if q_0 is H-symmetric and $g \in G^H$, then q_g remains H-symmetric.

Geometric interpretation. Conjugation describes how a change of reference frame by a symmetry $h \in H$ re-expresses a transport transformation $g \in G$. Suppose we wish to apply a transport g in a coordinate system that has been transformed by h. Moving from the original coordinates to the h-transformed coordinates applies h^{-1}; performing the transport in those coordinates applies g; and returning to the original frame applies h. The resulting transformation on $x \in \mathcal{X}$ is $(h \circ g \circ h^{-1}) \cdot x = \Phi_h(g) \cdot x$. Equivalently, for every $x \in \mathcal{X}$, $g \in G$, and $h \in H$, the following diagram commutes:

$$\begin{array}{ccc} x & \xrightarrow{\quad g \quad} & g \cdot x \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle h} \\ h \cdot x & \xrightarrow{\quad \Phi_h(g) \quad} & h \cdot (g \cdot x) = \Phi_h(g) \cdot (h \cdot x) \end{array}$$

That is, transporting a point by g and then applying the symmetry h is identical to first applying the symmetry h and then transporting by the conjugated parameter $\Phi_h(g) \in G$.

Detailed analysis of the guiding examples. We now compute the total group K, the conjugation map Φ_h, the kernel $\ker(\Phi)$, and the fixed-point subgroup G^H explicitly for each of our six guiding examples:

1. **Reflection** ($G = (\mathbb{R}, +)$, $H = \mathbb{Z}/2\mathbb{Z} = \{e, \sigma\}$ with $\sigma \cdot x = -x$). The total transformation group is the one-dimensional semidirect product $K = \mathbb{R} \rtimes (\mathbb{Z}/2\mathbb{Z})$ of translations and sign reflections. Denoting translation by $T_\mu(x) = x + \mu$, conjugating T_μ by the reflection σ yields
$$(\sigma \circ T_\mu \circ \sigma^{-1})(x) = \sigma(T_\mu(-x)) = -(-x + \mu) = x - \mu = T_{-\mu}(x)\,.$$
Thus $\Phi_\sigma(\mu) = -\mu$: reflecting the coordinate axis reverses the direction of the translation shift. The kernel $\ker(\Phi) = \{e\}$ is trivial, and the fixed-point subgroup is $G^H = \{\mu \in \mathbb{R} : -\mu = \mu\} = \{0\}$, indicating that no nonzero translation preserves reflection symmetry around the origin.
2. **Euclidean** ($G = (\mathbb{R}^d, +)$, $H = \mathrm{O}(d)$ with $h \cdot x = Ux$ for $U \in \mathrm{O}(d)$). The total group is the full Euclidean group $K = \mathrm{E}(d) = \mathbb{R}^d \rtimes \mathrm{O}(d)$. Conjugating a translation $T_\mu(x) = x + \mu$ by an orthogonal transformation $U \in \mathrm{O}(d)$ gives
$$(U \circ T_\mu \circ U^{-1})(x) = U(U^{-1}x + \mu) = x + U\mu = T_{U\mu}(x)\,,$$
so $\Phi_U(\mu) = U\mu$: an orthogonal change of basis rotates or reflects the location parameter μ by the same matrix U. The conjugation orbits in $G = \mathbb{R}^d$ are concentric spheres $\{\mu \in \mathbb{R}^d : \|\mu\|_2 = r\}$. Because only the identity matrix fixes every vector in $\mathbb{R}^d$, the action is faithful ($\ker(\Phi) = \{I_d\}$), and the fixed-point subgroup is trivial, $G^H = \{0\}$.
3. **Sphere** ($G = \mathrm{SO}(3)$, $H = \mathrm{SO}(2)$ embedded as planar rotations around e_3). Because $H \subset G$, the total group collapses to $K = \mathrm{SO}(3)$. For an axial symmetry rotation $h = R_\alpha^{e_3} \in \mathrm{SO}(2)$ by angle α around e_3 and a general 3D rotation $g = R_\beta^\nu \in \mathrm{SO}(3)$ by angle β around a unit axis $\nu \in \mathcal{S}^2$, matrix conjugation yields
$$\Phi_h(g) = R_\alpha^{e_3} R_\beta^\nu (R_\alpha^{e_3})^{-1} = R_\beta^{R_\alpha^{e_3}\nu}\,.$$
Geometrically, conjugating a rotation by $R_\alpha^{e_3}$ preserves the rotation angle β while rotating its axis ν around the North–South pole e_3 by angle α. A rotation R_β^ν commutes with all rotations around e_3 if and only if its axis ν is parallel to e_3; hence the fixed-point subgroup is $G^H = \{R_\gamma^{e_3} : \gamma \in [0, 2\pi)\} = \mathrm{SO}(2) = H$, which coincides exactly with the internal symmetry group $N = G \cap H = H$.
4. **Covariance** ($G = \mathrm{GL}(d)$, $H = \mathrm{O}(d)$ with $g \cdot x = Ax$). Here again $H \subset G$, so the total group is $K = \mathrm{GL}(d)$. For an orthogonal matrix $U \in \mathrm{O}(d)$ and an invertible matrix $A \in \mathrm{GL}(d)$, conjugation acts by orthogonal similarity:
$$\Phi_U(A) = UAU^{-1} = UAU^\top\,.$$

Statistically, conjugating A by U corresponds to rotating the coordinate axes of the covariance matrix $\Sigma = AA^\top \mapsto U\Sigma U^\top$, which rotates the principal axes of the ellipsoid while preserving its eigenvalues (and hence its trace and determinant). Any matrix $A \in \mathrm{GL}(d)$ satisfying $UAU^\top = A$ for all $U \in \mathrm{O}(d)$ must remain invariant under all orthogonal rotations and coordinate reflections, which forces A to have equal scaling in every direction and no off-diagonal couplings (formally, by Schur's lemma applied to the irreducible action of $\mathrm{O}(d)$ on $\mathbb{R}^d$). Thus the fixed-point subgroup is the one-dimensional subgroup of isotropic dilations, $G^H = \{\lambda I_d : \lambda \in \mathbb{R} \setminus \{0\}\}$, whereas the kernel of Φ is $\ker(\Phi) = \{\pm I_d\}$.

5. **Square** ($G = \mathrm{SO}(2)$, $H = D_4$). The planar rotation group $G = \mathrm{SO}(2)$ and the dihedral group $H = D_4$ intersect in the fourfold rotational subgroup $N = C_4 = \{R_{k\pi/2} : k = 0, 1, 2, 3\}$ and together generate the full orthogonal group $K = \mathrm{O}(2)$. The eight elements of D_4 consist of four rotations $r_k = R_{k\pi/2}$ and four reflections s_k. Because $\mathrm{SO}(2)$ is abelian, all rotations $r_k \in C_4$ commute with every $R_\theta \in \mathrm{SO}(2)$, giving $\Phi_{r_k}(R_\theta) = R_\theta$. By contrast, any reflection $s_k \in D_4$ reverses planar orientation and therefore inverts the rotation angle:
$$\Phi_{s_k}(R_\theta) = s_k R_\theta s_k^{-1} = R_{-\theta}\,.$$
Consequently, the conjugation homomorphism $\Phi : D_4 \to \mathrm{Aut}(\mathrm{SO}(2))$ factors through $D_4/C_4 \cong \mathbb{Z}/2\mathbb{Z}$ with kernel $\ker(\Phi) = C_4$, and its fixed-point set in $\mathrm{SO}(2)$ is the discrete two-element subgroup $G^H = \{R_0, R_\pi\} \cong \mathbb{Z}/2\mathbb{Z}$.
6. **Permutation** ($G = (\mathbb{R}^n, +)$, $H = \mathbb{S}_n$ with $\sigma \cdot x = P_\sigma x$). The total group is the semidirect product $K = \mathbb{R}^n \rtimes \mathbb{S}_n$ of translations and coordinate permutations. Conjugating a translation $T_\mu(x) = x + \mu$ by a permutation matrix P_σ gives
$$(P_\sigma \circ T_\mu \circ P_\sigma^{-1})(x) = P_\sigma(P_\sigma^{-1}x + \mu) = x + P_\sigma\mu = T_{P_\sigma\mu}(x)\,,$$
so $\Phi_\sigma(\mu) = P_\sigma\mu$: coordinate permutations act on the location parameter by permuting its entries. The conjugation orbit of $\mu \in \mathbb{R}^n$ consists of all vectors sharing the same empirical multiset of coordinates. The action is faithful ($\ker(\Phi) = \{e\}$), and a location vector μ is fixed by all permutations if and only if all of its coordinates are equal:
$$G^H = \{\mu \in \mathbb{R}^n : P_\sigma\mu = \mu \ \ \forall \sigma \in \mathbb{S}_n\} = \mathrm{span}(\mathbf{1})\,,$$
where $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ denotes the all-ones vector. Statistically, shifting an exchangeable distribution by $c\mathbf{1}$ shifts the common location of all coordinates simultaneously while preserving exchangeability.

Conjugation and distributional invariance of the variational objective. The following proposition explains why the conjugation map Φ_h directly governs the symmetries of the variational loss landscape $\mathcal{L}(g) = \mathrm{KL}(q_g \| p)$ around g^*.

Proposition 4.3 (Conjugation and transport of symmetric distributions)**.** *Assume that the base distribution $q_0 \in \mathcal{P}(\mathcal{X})$ is H-symmetric, and let $g \in G$. For any $h \in H$, let $q_{hg} := h_\sharp q_g$ denote the pushforward of q_g under the symmetry h, i.e., the distribution of $h \cdot (g \cdot Z)$ when $Z \sim q_0$ (note that the notation q_{hg} refers to the pushforward measure $(h \circ g)_\sharp q_0$, even when $h \notin G$). Then:*

(i) Pushforward equivariance*: $q_{\Phi_h(g)} = q_{hg}$, meaning that $q_{\Phi_h(g)}(x) = q_g(h^{-1} \cdot x)$ for all $x \in \mathcal{X}$.*
(ii) Loss under conjugated transport*: $\mathcal{L}(\Phi_h(g)) = \mathrm{KL}(q_{hg} \| p)$ for all $h \in H$ and $g \in G$.*
(iii) Conjugation invariance of the relative loss around g^*: If, in addition, the target p has H-conjugate symmetry around $g^* \in G$, then for every relative transport $g \in G$ and every symmetry $h \in H$,*
$$\mathcal{L}\big(g^*\Phi_h(g)\big) = \mathcal{L}(g^*g)\,.$$

Proof. To prove (i), observe that $q_{\Phi_h(g)}$ is the law of $\Phi_h(g) \cdot Z = (h \circ g \circ h^{-1}) \cdot Z$ when $Z \sim q_0$. Because q_0 is H-symmetric, the random variable $Z' = h^{-1} \cdot Z$ has the same distribution q_0 as Z. Therefore,

$$\Phi_h(g) \cdot Z = h \cdot \big(g \cdot (h^{-1} \cdot Z)\big) \overset{d}{=} h \cdot (g \cdot Z) \sim q_{hg}\,,$$

which establishes $q_{\Phi_h(g)} = q_{hg}$, and part (ii) follows immediately from the definition of $\mathcal{L}$. For part (iii), recall that $p = (g^*)_\sharp \tilde{p}$ where $\tilde{p}$ is H-symmetric ($\tilde{p} = h_\sharp \tilde{p}$). By the invariance of the KL divergence under the diffeomorphism $(g^*)^{-1}$, we have $\mathcal{L}(g^* g') = \mathrm{KL}(q_{g^* g'} \| p) = \mathrm{KL}(q_{g'} \| \tilde{p})$ for any $g' \in G$. Applying this identity with $g' = \Phi_h(g)$ and invoking part (i) together with the H-invariance of $\tilde{p}$ yields

$$\mathcal{L}\big(g^* \Phi_h(g)\big) = \mathrm{KL}\left(q_{\Phi_h(g)} \,\middle\|\, \tilde{p}\right) = \mathrm{KL}\left(h_\sharp q_g \,\middle\|\, \tilde{p}\right) = \mathrm{KL}\left(h_\sharp q_g \,\middle\|\, h_\sharp \tilde{p}\right) = \mathrm{KL}(q_g \| \tilde{p}) = \mathcal{L}(g^* g)\,,$$

where the penultimate equality again uses the invariance of KL divergence under the diffeomorphism h. □

Proposition 4.3 provides the key geometric insight into how symmetry shapes the variational objective. Part (i) shows that because q_0 is H-symmetric, applying a symmetry h to the transported distribution q_g is equivalent to replacing the parameter g by its conjugate $\Phi_h(g) \in G$—at the distribution level, q_g and $q_{\Phi_h(g)}$ have the exact same shape, viewed from an h-rotated angle. Part (iii) then shows that when p is H-conjugate symmetric around g^*, stepping away from g^* by a relative transport g or by its conjugated version $\Phi_h(g)$ yields the exact same KL divergence, so the local optimization landscape around g^* cannot distinguish between conjugate relative transports and $\mathcal{L}$ is constant along every conjugation orbit $\{g^* \Phi_h(g) : h \in H\}$. Notice that this conjugation invariance $\mathcal{L}(g^* \Phi_h(g)) = \mathcal{L}(g^* g)$ is far richer than the right-invariance $\mathcal{L}(gn) = \mathcal{L}(g)$ for $n \in N = G \cap H$: right-invariance under N merely reflects parameter redundancy *within* G (since $q_{gn} = q_g$), whereas conjugation invariance captures how symmetries in H act on G *from the outside* to relate distinct distributions $q_{g^* g}$ and $q_{g^* \Phi_h(g)}$ that are symmetric counterparts of one another around g^*.

4.3. Quotient space and statistical identifiability regimes

We now formalize the parameter redundancy induced by the overlap between the transport group G and the symmetry group H. Because q_0 is H-invariant, any transformation in the total group $K = GH$ that differs on the right by an element of H produces the exact same pushforward distribution on $\mathcal{X}$. Restricting to the transport group G, two parameters $g, g' \in G$ satisfy $g' = gn$ for some $n \in N := G \cap H$ if and only if they belong to the same left coset gN in the quotient space

$$\mathcal{M} := G/N \cong K/H\,, \qquad N = G \cap H\,.$$

In statistical terms, the subgroup $\mathrm{Stab}_G(q_0) := \{g \in G : g_\sharp q_0 = q_0\}$ (the *stabilizer* of q_0 in G) is the set of redundant parameter transformations that leave the base distribution q_0 unchanged, and forming the *quotient space* $\mathcal{M} = G/N$ simply means grouping together all parameters $\{gn : n \in N\}$ that produce the exact same distribution q_g into a single equivalence class $[g] = gN$ (just as a covariance matrix $\Sigma = AA^\intercal$ groups together all matrix square roots $\{AU : U \in \mathrm{O}(d)\}$). Assuming that the stabilizer of q_0 within G is precisely $N = G \cap H$ (i.e., $\mathrm{Stab}_G(q_0) = G \cap H$), the map $[g] = gN \mapsto q_g$ is injective on $\mathcal{M}$, so the smooth coset manifold $\mathcal{M} = G/N$ represents the space of *statistically identifiable parameters* of the variational family $\mathcal{Q}_G$. Depending on the structure of the intersection $N = G \cap H$, our examples fall into three distinct identifiability regimes:

1. Separable symmetry ($N = G \cap H = \{e\}$). When the transport group G and the symmetry group H intersect only at the identity transformation, the quotient manifold coincides with the transport group itself, $\mathcal{M} \cong G$. In this regime, no non-trivial transport leaves q_0 invariant, so the parameter $g \in G$ is **globally identifiable** from q_g:

- **Examples 1, 2, and 6 (Reflection, Euclidean, and Permutation location families)**: In all three translation models, $G = (\mathbb{R}^d, +)$ (or $\mathbb{R}^n$) contains no non-trivial linear isometries around the origin, so $N = G \cap H = \{0\}$ and $\mathcal{M} \cong \mathbb{R}^d$ (or $\mathbb{R}^n$). Every location parameter μ uniquely indexes a distinct distribution $q_\mu \in \mathcal{Q}_G$. Crucially, however, global parameter identifiability on $\mathcal{M}$ does not by itself guarantee that μ^* minimizes $\mathcal{L}(\mu)$ under model misspecification: as we show in Section 6, Examples 1 and 2 are always stationary at μ^*, whereas Example 6 can exhibit an additive location bias along the diagonal span($\mathbf{1}$).

2. Internal symmetry ($H \subset G$, so $N = H$). When the symmetry group H is a subgroup of the transport group G, the total group collapses to $K = G$ and the identifiable parameter space is the homogeneous quotient manifold $\mathcal{M} = G/H$ of left cosets $[g] = gH$. In this regime, the group element $g \in G$ is **overparametrized**, and only the coset $[g] \in G/H$ is statistically identifiable:

- **Example 3 (Sphere)**: With $G = \mathrm{SO}(3)$ and $H = N = \mathrm{SO}(2)$ (rotations around e_3), the quotient manifold is $\mathcal{M} = \mathrm{SO}(3)/\mathrm{SO}(2) \cong \mathcal{S}^2$. Right-multiplying a rotation matrix $R \in \mathrm{SO}(3)$ by any axial rotation $R_\alpha^{e_3} \in \mathrm{SO}(2)$ leaves the image of the North pole $v = Re_3 \in \mathcal{S}^2$ unchanged, so the identifiable statistic is precisely the modal axis $v^* = R^* e_3 \in \mathcal{S}^2$.
- **Example 4 (Covariance)**: With $G = \mathrm{GL}(d)$ and $H = N = \mathrm{O}(d)$, the quotient manifold is $\mathcal{M} = \mathrm{GL}(d)/\mathrm{O}(d) \cong \mathrm{SPD}(d)$, the cone of $d \times d$ symmetric positive-definite matrices. Two invertible matrices $A, A' \in \mathrm{GL}(d)$ belong to the same left coset $A\,\mathrm{O}(d)$ if and only if $A' = AU$ for some $U \in \mathrm{O}(d)$, which is equivalent to $A'(A')^\top = AA^\top = \Sigma$. Thus, the factor A^* is identifiable up to right-multiplication by $\mathrm{O}(d)$ (with continuous gradient optimization operating within a single connected component such as $\mathrm{GL}^+(d)$, where the connected stabilizer fiber is $\mathrm{GL}^+(d) \cap \mathrm{O}(d) = \mathrm{SO}(d)$), and the identifiable parameter on $\mathcal{M}$ is the covariance matrix $\Sigma^* = A^*(A^*)^\top$.

3. Partial overlap ($\{e\} \subsetneq N \subsetneq H$). When G and H share a proper, non-trivial subgroup $N = G \cap H$, the identifiable parameter space is the quotient $\mathcal{M} = G/N$, while the remaining symmetries in $H \setminus N$ act non-trivially on $\mathcal{M}$ from the outside via conjugation:

- **Example 5 (Square)**: With $G = \mathrm{SO}(2)$ and $H = D_4$, the intersection is the discrete fourfold rotation group $N = C_4 = \{R_{k\pi/2} : k = 0, 1, 2, 3\}$. Because rotating a square-symmetric distribution by 90° produces the exact same distribution, the orientation angle $\theta \in [0, 2\pi)$ is identifiable only modulo $\pi/2$, on the quotient manifold $\mathcal{M} = \mathrm{SO}(2)/C_4 \cong \mathcal{S}^1$. Meanwhile, the reflections in $D_4 \setminus C_4$ do not belong to G, yet they act on $\mathcal{M}$ via conjugation ($\theta \mapsto -\theta$), which will prove decisive in forcing the gradient at θ^* to vanish.

Remark (Identifiability versus unbiased stationarity). Because $q_{gn} = q_g$ for all $n \in N$, the variational loss satisfies $\mathcal{L}(gn) = \mathcal{L}(g)$ for every $g \in G$ and $n \in N$. When N is a positive-dimensional Lie subgroup (as in Examples 3 and 4), $\mathcal{L}$ is identically constant along the coset fibers gN, so the directional derivative of $\mathcal{L}$ at g^* automatically vanishes along all tangent directions to N. When N is finite (as in Example 5), $\mathcal{L}$ is periodic across the $|N|$ equivalent representatives $\{g^* n : n \in N\}$. However, this fiber invariance only accounts for directions *inside* N; whether the gradient vanishes along the identifiable directions transverse to N depends on the adjoint action of the full symmetry group H on the Lie algebra of G, which we develop in the next two sections.

5. Differentiation on the transport Lie group

To characterize when the true transport parameter $g^* \in G$ is a first-order stationary point of the variational loss $\mathcal{L}(g) = \mathrm{KL}(q_g \| p)$, we require differential calculus on the Lie group G. In classical

parametric statistics, one typically differentiates a loss function by adding a small coordinate vector $\theta + tu$ in $\mathbb{R}^p$. For transformation parameters such as rotation matrices $R \in \mathrm{SO}(d)$ or scale matrices $A \in \mathrm{GL}(d)$, however, an additive perturbation $A + tU$ either leaves the manifold of valid parameters (breaking orthogonality) or depends on the current scale A. Because G is both a group and a smooth manifold, local perturbations around any parameter $g \in G$ can instead be parameterized canonically through the tangent space at the identity—the Lie algebra $\mathfrak{g} = T_eG$—via left multiplication $t \mapsto g\exp(tu)$ (where $\exp(tu) \approx I + tu$ for matrix groups). Writing local perturbations multiplicatively as $g \cdot \exp(tu)$ rather than additively (known geometrically as *left-trivialization*) measures *relative* (or equivariant) parameter changes in a single fixed vector space $\mathfrak{g}$ at the identity. From a statistical perspective, this multiplicative parameterization is natural for transformation models: perturbing g^* to $g^*\exp(tu)$ corresponds, after pulling back by $(g^*)^{-1}$, to perturbing the standardized base distribution q_0 by $\exp(tu)$ near the identity e_G. We briefly review the necessary tools from Lie theory and Riemannian optimization; see Hall (2015) and Boumal (2023) for comprehensive treatments.

5.1. Lie group calculus and left-trivialized Riemannian gradients

Definition 5.1 (Lie algebra, exponential map, and Riemannian gradient)**.** Let G be a Lie group with identity element e_G. We recall the following differential-geometric objects:

- The *Lie algebra* $\mathfrak{g} := T_{e_G}G$ is the tangent vector space to G at the identity, which we equip with a positive-definite inner product $\langle\cdot,\cdot\rangle_{\mathfrak{g}}$ (defining a left-invariant Riemannian metric on G).
- The *exponential map* $\exp : \mathfrak{g} \to G$ maps each tangent vector $u \in \mathfrak{g}$ to the time-1 point $\gamma_u(1)$ of the unique one-parameter subgroup $\gamma_u(t) = \exp(tu)$ satisfying $\gamma_u(0) = e_G$ and $\gamma_u'(0) = u$. The map exp is a smooth local diffeomorphism from a neighborhood of $0 \in \mathfrak{g}$ onto a neighborhood of $e_G \in G$, so a neighborhood of any $g \in G$ is parameterized via $u \mapsto g \cdot \exp(u)$.
- The (left-trivialized) *Riemannian gradient* $\nabla_G\mathcal{L}(g) \in \mathfrak{g}$ of a smooth objective $\mathcal{L} : G \to \mathbb{R}$ at $g \in G$ is the unique element of $\mathfrak{g}$ representing the first-order Taylor expansion of $u \mapsto \mathcal{L}(g \cdot \exp(u))$ at $u = 0$:

$$\mathcal{L}(g \cdot \exp(u)) = \mathcal{L}(g) + \langle \nabla_G\mathcal{L}(g), u\rangle_{\mathfrak{g}} + O(\|u\|^2) \quad \forall u \in \mathfrak{g}\,. \tag{1}$$

Concrete form of the exponential map and Riemannian gradient. In statistical applications, G is either an additive vector space or a matrix Lie group, where both the exponential map and the left-trivialized gradient take explicit forms:

- **Additive translation group** $G = (\mathbb{R}^d, +)$: The group operation is vector addition, the identity is $e_G = 0$, the Lie algebra is $\mathfrak{g} = \mathbb{R}^d$ equipped with the standard Euclidean inner product, and the exponential map is the identity $\exp(u) = u$. Perturbing $\mu \in G$ via $\mu \cdot \exp(tu)$ is simply $\mu + tu$, and the Riemannian gradient coincides with the ordinary Euclidean gradient:

$$\nabla_G\mathcal{L}(\mu) = \nabla_{\mathrm{Euc}}\mathcal{L}(\mu) \in \mathbb{R}^d\,.$$

- **Matrix Lie groups** $G \subseteq \mathrm{GL}(d)$: When G is a closed subgroup of $\mathrm{GL}(d)$, $\exp(U) = \sum_{k=0}^{\infty} \frac{1}{k!}U^k$ is the standard matrix exponential, and $g \cdot \exp(tU) = ge^{tU}$ is matrix multiplication. If $\mathcal{L}$ is extended smoothly to an open neighborhood in $\mathbb{R}^{d\times d}$ with Euclidean matrix of partial derivatives $\nabla_{\mathrm{Euc}}\mathcal{L}(g) \in \mathbb{R}^{d\times d}$, then for any $U \in \mathfrak{g}$, the chain rule gives

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathcal{L}(ge^{tU}) = \langle \nabla_{\mathrm{Euc}}\mathcal{L}(g),\, gU\rangle_{\mathrm{Frob}} = \langle g^\top \nabla_{\mathrm{Euc}}\mathcal{L}(g),\, U\rangle_{\mathrm{Frob}}\,.$$

  Consequently, under the Frobenius inner product $\langle U, V\rangle_{\mathfrak{g}} = \mathrm{tr}(U^\top V)$, the Riemannian gradient in $\mathfrak{g}$ is obtained by left-multiplying the Euclidean gradient by $g^\top$ (transporting the tangent vector from

T_gG back to $T_IG = \mathfrak{g}$) and projecting orthogonally onto the subspace $\mathfrak{g} \subseteq \mathbb{R}^{d\times d}$:

$$\nabla_G \mathcal{L}(g) = \mathrm{proj}_{\mathfrak{g}}(g^\top \nabla_{\mathrm{Euc}} \mathcal{L}(g)) \ .$$

For the general linear group $G = \mathrm{GL}(d)$, the Lie algebra is the full matrix space $\mathfrak{g} = \mathfrak{gl}(d) = \mathbb{R}^{d\times d}$, so $\mathrm{proj}_{\mathfrak{g}}$ is the identity and $\nabla_G \mathcal{L}(A) = A^\top \nabla_{\mathrm{Euc}} \mathcal{L}(A)$. For the special orthogonal group $G = \mathrm{SO}(d)$, the Lie algebra $\mathfrak{g} = \mathfrak{so}(d) = \{U \in \mathbb{R}^{d\times d} : U^\top = -U\}$ consists of skew-symmetric matrices, and orthogonal projection onto $\mathfrak{so}(d)$ is the skew-symmetrization $\mathrm{proj}_{\mathfrak{so}(d)}(M) = \frac{1}{2}(M - M^\top)$.

5.2. Adjoint representation and the trivial subspace $\mathfrak{g}^H$

In Section 4.2, we saw that the symmetry group H acts on the transport group G by conjugation automorphisms $\Phi_h(g) = h \circ g \circ h^{-1}$. While $\Phi_h : G \to G$ is in general a nonlinear map on the manifold G, it fixes the identity element ($\Phi_h(e_G) = e_G$ for every $h \in H$). Differentiating Φ_h at e_G therefore induces a *linear representation* of H on the tangent space $\mathfrak{g} = T_{e_G}G$: the *adjoint representation*.

Definition 5.2 (Adjoint representation and trivial subspace)**.** The *adjoint representation* of H on $\mathfrak{g}$ is the group homomorphism $\mathrm{Ad} : H \to \mathrm{GL}(\mathfrak{g})$, $h \mapsto \mathrm{Ad}_h$, defined as the differential of the conjugation map Φ_h at the identity e_G. For any $h \in H$ and $u \in \mathfrak{g}$:

$$\mathrm{Ad}_h \cdot u := (d\Phi_h)_{e_G}(u) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \Phi_h(\exp(tu)) \in \mathfrak{g} \, .$$

The *trivial representation* (or *H-fixed subspace*) $\mathfrak{g}^H \subseteq \mathfrak{g}$ is the linear subspace of tangent directions left invariant by the adjoint action of every element of H (called "trivial" in representation theory because every $h \in H$ acts on $\mathfrak{g}^H$ as the identity operator, leaving its vectors unchanged):

$$\mathfrak{g}^H = \{v \in \mathfrak{g} \ : \ \mathrm{Ad}_h \cdot v = v \ \ \forall h \in H\} \, .$$

Compactness of H and Ad_H-invariant inner products. Because H is assumed to be a compact topological group, it admits a unique normalized Haar probability measure. By averaging any initial inner product $\langle \cdot, \cdot \rangle_0$ on $\mathfrak{g}$ over H via $\langle u, v \rangle_{\mathfrak{g}} := \int_H \langle \mathrm{Ad}_h u, \mathrm{Ad}_h v \rangle_0 \, \mathrm{dHaar}(h)$, we are guaranteed an Ad_H-invariant inner product on $\mathfrak{g}$ satisfying

$$\langle \mathrm{Ad}_h u, \mathrm{Ad}_h v \rangle_{\mathfrak{g}} = \langle u, v \rangle_{\mathfrak{g}} \qquad \forall h \in H, \ \ \forall u, v \in \mathfrak{g} \, .$$

With respect to this inner product, each linear operator $\mathrm{Ad}_h : \mathfrak{g} \to \mathfrak{g}$ is orthogonal ($\mathrm{Ad}_h^\top = \mathrm{Ad}_h^{-1} = \mathrm{Ad}_{h^{-1}}$). Note that for all of our guiding examples, the standard Euclidean inner product on $\mathbb{R}^d$ and the Frobenius inner product on $\mathfrak{gl}(d)$ or $\mathfrak{so}(d)$ are already Ad_H-invariant without requiring explicit averaging.

Statistical and geometric interpretation of Ad_h and $\mathfrak{g}^H$. Just as $\Phi_h(g)$ represents the transport g viewed in an h-transformed coordinate frame, $\mathrm{Ad}_h \cdot u$ represents the infinitesimal parameter perturbation $u \in \mathfrak{g}$ rotated by the symmetry h. In our two primary classes of transport groups, Ad_h takes a particularly transparent form:

- *Translation groups* ($G = (\mathbb{R}^d, +)$ with H acting linearly by $h \cdot x = M_h x$): Because $\Phi_h(\mu) = M_h \mu$ is already linear in μ, its derivative at 0 is simply matrix-vector multiplication:

$$\mathrm{Ad}_h \cdot u = M_h u \qquad (u \in \mathbb{R}^d) \, .$$

- *Matrix groups* ($G \subseteq \mathrm{GL}(d)$ with H acting linearly by $M_h \in \mathrm{O}(d)$): Because $\Phi_h(g) = M_h g M_h^{-1}$ is matrix conjugation, differentiating $t \mapsto M_h e^{tU} M_h^{-1}$ at $t = 0$ yields matrix similarity on the Lie algebra:

$$\mathrm{Ad}_h \cdot U = M_h U M_h^{-1} = M_h U M_h^\top \qquad (U \in \mathfrak{g} \subseteq \mathbb{R}^{d\times d}) .$$

The trivial subspace $\mathfrak{g}^H = \{u \in \mathfrak{g} : \mathrm{Ad}_h \cdot u = u \ \forall h \in H\}$ is precisely the Lie algebra of the fixed-point subgroup $G^H = \{g \in G : \Phi_h(g) = g \ \forall h \in H\}$ introduced in Section 4.2. Statistically, a tangent vector u belongs to $\mathfrak{g}^H$ if and only if the one-parameter flow $\exp(tu)$ commutes with every symmetry $h \in H$, which means that pushing an H-symmetric distribution q_0 forward along $\exp(tu)$ *preserves its H-symmetry*. Conversely, any direction u orthogonal to $\mathfrak{g}^H$ breaks the H-symmetry of q_0.

Guiding examples: Adjoint action and trivial subspace $\mathfrak{g}^H$. We now determine $\mathfrak{g}$, the linear operator Ad_h, and the trivial subspace $\mathfrak{g}^H$ for each of the six guiding examples:

1. **Reflection** ($G = (\mathbb{R}, +)$, $H = \mathbb{Z}/2\mathbb{Z} = \{1, -1\}$): The Lie algebra is $\mathfrak{g} = \mathbb{R}$, and the non-trivial reflection $\sigma = -1$ acts on tangent shifts $u \in \mathbb{R}$ by sign inversion, $\mathrm{Ad}_\sigma u = -u$. The only shift direction satisfying $-u = u$ is $u = 0$, so the trivial subspace vanishes: $\mathfrak{g}^H = \{0\}$. Every nonzero translation breaks reflection symmetry around the origin.
2. **Euclidean** ($G = (\mathbb{R}^d, +)$, $H = \mathrm{O}(d)$): The Lie algebra is $\mathfrak{g} = \mathbb{R}^d$, and each orthogonal matrix $U \in \mathrm{O}(d)$ acts on tangent vectors $u \in \mathbb{R}^d$ by $\mathrm{Ad}_U u = Uu$. Because no nonzero vector in $\mathbb{R}^d$ is invariant under all orthogonal transformations, the trivial subspace is again trivial: $\mathfrak{g}^H = \{0\}$.
3. **Sphere** ($G = \mathrm{SO}(3)$, $H = \mathrm{SO}(2)$ fixing e_3): The Lie algebra $\mathfrak{g} = \mathfrak{so}(3)$ of 3×3 skew-symmetric matrices can be identified with $\mathbb{R}^3$ via the hat isomorphism $\omega \in \mathbb{R}^3 \leftrightarrow [\omega]_\times \in \mathfrak{so}(3)$, where $[\omega]_\times x = \omega \times x$. Under this identification, matrix conjugation $\mathrm{Ad}_h([\omega]_\times) = R_\alpha^{e_3} [\omega]_\times (R_\alpha^{e_3})^\top = [R_\alpha^{e_3} \omega]_\times$ acts on the instantaneous rotation axis $\omega \in \mathbb{R}^3$ simply by rotating ω around e_3 by angle α. The vectors invariant under all rotations around e_3 are those parallel to e_3 itself, yielding a one-dimensional trivial subspace:

$$\mathfrak{g}^H = \mathrm{span}([e_3]_\times) \cong \mathrm{span}(e_3) .$$

   Note that this subspace coincides with the Lie algebra $\mathfrak{n} = \mathfrak{so}(2)$ of the stabilizer $N = \mathrm{SO}(2)$ of q_0.
4. **Covariance** ($G = \mathrm{GL}(d)$, $H = \mathrm{O}(d)$): The Lie algebra is the space of all $d \times d$ matrices $\mathfrak{g} = \mathfrak{gl}(d) = \mathbb{R}^{d\times d}$, on which $U \in \mathrm{O}(d)$ acts by orthogonal similarity $\mathrm{Ad}_U(M) = UMU^\top$. A matrix $M \in \mathfrak{gl}(d)$ belongs to $\mathfrak{g}^H$ if and only if $UMU^\top = M$ (i.e., $UM = MU$) for all $U \in \mathrm{O}(d)$. By Schur's lemma, the only matrices commuting with all orthogonal transformations are scalar multiples of the identity matrix:

$$\mathfrak{g}^H = \mathrm{span}(I_d) = \{\lambda I_d : \lambda \in \mathbb{R}\} .$$

   Statistically, $\mathrm{span}(I_d)$ generates the one-parameter subgroup $\exp(tI_d) = e^t I_d$ of isotropic scalings, which dilate or contract the overall variance of an $\mathrm{O}(d)$-symmetric distribution without breaking its spherical symmetry.
5. **Square** ($G = \mathrm{SO}(2)$, $H = D_4$): The Lie algebra of planar rotations is one-dimensional, $\mathfrak{g} = \mathfrak{so}(2) \cong \mathbb{R}$, parameterized by the angular velocity $\dot\theta \in \mathbb{R}$. Differentiating $\Phi_{r_k}(R_\theta) = R_\theta$ and $\Phi_{s_k}(R_\theta) = R_{-\theta}$ at $\theta = 0$ shows that rotations $r_k \in C_4$ act trivially ($\mathrm{Ad}_{r_k}\dot\theta = \dot\theta$), whereas reflections $s_k \in D_4 \setminus C_4$ reverse the angular velocity ($\mathrm{Ad}_{s_k}\dot\theta = -\dot\theta$). Because an invariant direction must be fixed by *all* elements of D_4 including reflections, we obtain $\mathfrak{g}^H = \{0\}$.
6. **Permutation** ($G = (\mathbb{R}^n, +)$, $H = \mathbb{S}_n$): The Lie algebra is $\mathfrak{g} = \mathbb{R}^n$, and each permutation $\sigma \in \mathbb{S}_n$ acts on tangent shift vectors $u \in \mathbb{R}^n$ by permuting their coordinates, $\mathrm{Ad}_\sigma u = P_\sigma u$. Under the standard representation of $\mathbb{S}_n$, $\mathbb{R}^n$ decomposes orthogonally into the one-dimensional trivial representation

of constant vectors and the $(n-1)$-dimensional standard irreducible representation of zero-sum vectors:

$$\mathfrak{g}^H = \{u \in \mathbb{R}^n : P_\sigma u = u \ \ \forall \sigma \in \mathbb{S}_n\} = \mathrm{span}(\mathbf{1}), \qquad (\mathfrak{g}^H)^\perp = \left\{u \in \mathbb{R}^n : \sum_{i=1}^n u_i = 0\right\}.$$

Statistically, perturbations along $\mathrm{span}(\mathbf{1})$ shift all coordinates by a common scalar additive constant, preserving exchangeability, whereas perturbations in $(\mathfrak{g}^H)^\perp$ alter the relative offsets between coordinates and break exchangeability.

As we prove in Section 6, the Riemannian gradient $\nabla_G \mathcal{L}(g^*)$ is forced to lie inside $\mathfrak{g}^H$. Whenever $\mathfrak{g}^H = \{0\}$ (Examples 1, 2, and 5), first-order stationarity $\nabla_G \mathcal{L}(g^*) = 0$ follows immediately from symmetry alone; when $\mathfrak{g}^H \neq \{0\}$ (Examples 3, 4, and 6), the interplay between $\mathfrak{g}^H$ and the identifiability kernel $\mathfrak{n}$ determines whether any nonzero gradient component can survive.

6. Main results: symmetry constraints, Lie algebra decomposition, and moment matching

We now establish our main theoretical results characterizing the first-order optimization landscape of the variational objective $\mathcal{L}(g) = \mathrm{KL}(q_g \| p)$ at the true transport parameter g^*. Geometrically, our results rest on a fundamental principle of invariant analysis: *the gradient of a group-invariant functional at a symmetric point must itself be invariant under the group action*. To picture this concretely, consider a rotationally invariant (radial) function on $\mathbb{R}^d$: computing its gradient at a rotated point on the same circular level set yields an equally rotated gradient vector, and at the center of symmetry itself (the origin), the gradient vector must remain unchanged under every rotation—which forces it to be the zero vector. When translated to the Lie group setting via the pullback loss around g^*, this principle implies that the Riemannian gradient $\nabla_G \mathcal{L}(g^*)$ is fixed by the adjoint action Ad_h for every $h \in H$, and therefore must reside in the trivial representation $\mathfrak{g}^H$.

6.1. Warm-up: Gradient constraint for translation families

To build concrete statistical intuition before treating general Lie groups, we first state and prove the gradient constraint for additive location families $G = (\mathbb{R}^d, +)$, where the Lie algebra is $\mathbb{R}^d$ and the Riemannian gradient reduces to the standard Euclidean gradient. This proposition generalizes the univariate reflection argument of Margossian and Saul (2025) (where $d = 1$ and $H = \{\pm 1\}$) to arbitrary linear isometry groups $H \subseteq \mathrm{O}(d)$ in d dimensions, and illustrates how the expected score of the standardized target interacts with H-invariance.

Proposition 6.1 (Gradient constraint — translation case)**.** *In the setting of Definition 3.3, suppose that $G = (\mathbb{R}^d, +)$ acts on $\mathcal{X} = \mathbb{R}^d$ by translation $g \cdot x = x + \mu_g$, and that H acts linearly by orthogonal transformations $h \cdot x = M_h x$. Then the Euclidean gradient of $\mathcal{L}(\mu) = \mathrm{KL}(q_\mu \| p)$ at the true location parameter μ^* satisfies*

$$\nabla \mathcal{L}(\mu^*) = M_h \nabla \mathcal{L}(\mu^*), \quad \forall h \in H.$$

Equivalently, $\nabla \mathcal{L}(\mu^*)$ lies in the H-fixed subspace $(\mathbb{R}^d)^H = \{v \in \mathbb{R}^d : M_h v = v \ \forall h \in H\}$.

Proof. Applying the translation change of variables $x = \mu + \xi$ (with unit Jacobian determinant), the variational objective at any location $\mu \in \mathbb{R}^d$ can be written as an expectation under the fixed base distribution $\xi \sim q_0$:

$$\mathcal{L}(\mu) = \int_{\mathbb{R}^d} q_0(\xi) \big[\log q_0(\xi) - \log p(\mu + \xi)\big] \, \mathrm{d}\xi.$$

Differentiating with respect to μ under the integral sign and evaluating at $\mu = \mu^*$, where $p(\mu^* + \xi) = \tilde{p}(\xi)$ by definition of the standardized pullback $\tilde{p}$, yields the expected score of $\tilde{p}$ under q_0:

$$\nabla_G \mathcal{L}(\mu^*) = \nabla \mathcal{L}(\mu^*) = -\int_{\mathbb{R}^d} q_0(\xi) \nabla_\mu \big[\log p(\mu^* + \xi)\big]\, \mathrm{d}\xi = -\int_{\mathbb{R}^d} q_0(\xi) \nabla \log \tilde{p}(\xi)\, \mathrm{d}\xi\,.$$

Because $\tilde{p}$ is H-symmetric ($\tilde{p}(M_h\xi') = \tilde{p}(\xi')$), the chain rule implies that its score function is H-equivariant: $M_h^\top \nabla \log \tilde{p}(M_h\xi') = \nabla \log \tilde{p}(\xi')$, or equivalently $\nabla \log \tilde{p}(M_h\xi') = M_h \nabla \log \tilde{p}(\xi')$ since $M_h \in \mathrm{O}(d)$ is orthogonal. Performing the orthogonal change of variables $\xi = M_h\xi'$ (with $|\det M_h| = 1$) and using the H-invariance $q_0(M_h\xi') = q_0(\xi')$, we obtain for every $h \in H$:

$$\nabla \mathcal{L}(\mu^*) = -\int_{\mathbb{R}^d} q_0(M_h\xi')\, \nabla \log \tilde{p}(M_h\xi')\, \mathrm{d}\xi' = -M_h \int_{\mathbb{R}^d} q_0(\xi')\, \nabla \log \tilde{p}(\xi')\, \mathrm{d}\xi' = M_h \nabla \mathcal{L}(\mu^*)\,.$$

Thus $v^* = \nabla \mathcal{L}(\mu^*)$ satisfies $M_h v^* = v^*$ for all $h \in H$, i.e., $\nabla \mathcal{L}(\mu^*) \in (\mathbb{R}^d)^H$. In Examples 1 (Reflection) and 2 (Euclidean), $(\mathbb{R}^d)^H = \{0\}$, which immediately forces $\nabla \mathcal{L}(\mu^*) = 0$; in Example 6 (Permutation), it constrains $\nabla \mathcal{L}(\mu^*) \in \mathrm{span}(\mathbf{1})$. □

6.2. Gradient constraint: the general Lie group case

We now establish our first main theorem for an arbitrary transport Lie group G normalized by a compact symmetry group H. In the translation setting of Proposition 6.1, the matrix M_h acted directly on the parameter gradient because the conjugation of a translation μ by M_h is simply $\Phi_h(\mu) = M_h\mu$. For a general Lie group G, the natural counterpart of M_h on the tangent space $\mathfrak{g} = T_eG$ is the adjoint operator $\mathrm{Ad}_h = (d\Phi_h)_{e_G}$. Replacing Euclidean invariance with the conjugation invariance of the relative loss established in Proposition 4.3(iii) yields the general gradient constraint.

Theorem 6.2 (Gradient constraint from symmetry)**.** *Under the assumptions of Definition 3.3, let $\mathfrak{g}$ be equipped with an Ad_H-invariant inner product $\langle\cdot,\cdot\rangle_\mathfrak{g}$. Then the Riemannian gradient of the variational objective $\mathcal{L}(g) = \mathrm{KL}(q_g \| p)$ at the true transport parameter $g^* \in G$ satisfies*

$$\nabla_G \mathcal{L}(g^*) = Ad_h \cdot \nabla_G \mathcal{L}(g^*)\,, \quad \forall h \in H\,.$$

Equivalently, $\nabla_G \mathcal{L}(g^*)$ is constrained to lie in the trivial representation of H in $\mathfrak{g}$:

$$\nabla_G \mathcal{L}(g^*) \in \mathfrak{g}^H\,.$$

Proof. For any tangent vector $u \in \mathfrak{g}$, consider the perturbed parameter $g^* \exp(u) \in G$. Because the Kullback–Leibler divergence is invariant under simultaneous application of the diffeomorphism $(g^*)^{-1}$ to both measures, we can pull the variational objective back to the standardized reference space around the identity e_G:

$$\mathcal{L}(g^* \exp(u)) = \mathrm{KL}\left(q_{g^*\exp(u)} \,\middle\|\, p\right) = \mathrm{KL}\left(q_{\exp(u)} \,\middle\|\, \tilde{p}\right) =: \mathcal{L}_0(\exp(u))\,.$$

By Definition 5.1, the left-trivialized Riemannian gradient of $\mathcal{L}$ at g^* coincides with the Riemannian gradient of the standardized loss $\mathcal{L}_0$ at e_G, and hence with the Euclidean gradient at $0 \in \mathfrak{g}$ of the pullback objective $J : \mathfrak{g} \to \mathbb{R}$ defined by $J(u) := \mathcal{L}_0(\exp(u)) = \mathcal{L}(g^* \exp(u))$.

We first show that the function $J : \mathfrak{g} \to \mathbb{R}$ is strictly invariant under the adjoint action of H, i.e., $J(\mathrm{Ad}_h \cdot u) = J(u)$ for all $h \in H$ and $u \in \mathfrak{g}$:

$$\begin{aligned} J(\mathrm{Ad}_h \cdot u) &= \mathcal{L}_0\big(\exp(\mathrm{Ad}_h \cdot u)\big) = \mathcal{L}_0\big(\Phi_h(\exp(u))\big) \quad (\dagger) \\ &= \mathrm{KL}\left(q_{\Phi_h(\exp(u))} \,\middle\|\, \tilde{p}\right) = \mathrm{KL}\left(q_{h\exp(u)} \,\middle\|\, \tilde{p}\right) \quad (*) \\ &= \mathrm{KL}\left(q_{h\exp(u)} \,\middle\|\, h_\sharp \tilde{p}\right) \quad (**) \\ &= \mathrm{KL}\left(q_{\exp(u)} \,\middle\|\, \tilde{p}\right) = J(u)\,. \end{aligned}$$

Each step is justified by the algebraic and measure-theoretic structure established earlier:

- **Step (†) (Naturality of the exponential map):** $\exp(\mathrm{Ad}_h \cdot u) = \Phi_h(\exp(u))$. Since $\Phi_h : G \to G$ is a Lie group automorphism (Proposition 4.2), the curve $t \mapsto \Phi_h(\exp(tu))$ is a one-parameter subgroup of G whose initial velocity at $t = 0$ is $(d\Phi_h)_{e_G}(u) = \mathrm{Ad}_h \cdot u$ by Definition 5.2; by uniqueness of one-parameter subgroups, $\Phi_h(\exp(tu)) = \exp(t\mathrm{Ad}_h \cdot u)$.
- **Step (*) (Pushforward equivariance of q_0):** $q_{\Phi_h(\exp(u))} = q_{h\exp(u)} = h_\sharp q_{\exp(u)}$, which follows directly from Proposition 4.3(i) using the H-invariance of q_0.
- **Step (**) (H-invariance of the pullback $\tilde{p}$):** Because p has H-conjugate symmetry around g^*, its pullback $\tilde{p}$ satisfies $h_\sharp \tilde{p} = \tilde{p}$, and the final equality follows by applying the diffeomorphism h^{-1} to both arguments of the KL divergence.

Differentiating the invariance identity $J(\mathrm{Ad}_h \cdot u) = J(u)$ at $u = 0$ along an arbitrary direction $u \in \mathfrak{g}$ via the first-order expansion (1) gives

$$\langle \nabla_G \mathcal{L}(g^*), u\rangle_{\mathfrak{g}} = \langle \nabla J(0), u\rangle_{\mathfrak{g}} = \langle \nabla J(0), \mathrm{Ad}_h \cdot u\rangle_{\mathfrak{g}} = \langle \mathrm{Ad}_h^\intercal \nabla J(0), u\rangle_{\mathfrak{g}} \qquad \forall u \in \mathfrak{g}\,.$$

Because the inner product $\langle\cdot,\cdot\rangle_{\mathfrak{g}}$ is Ad_H-invariant, Ad_h is an orthogonal operator on $\mathfrak{g}$, so $\mathrm{Ad}_h^\intercal = \mathrm{Ad}_h^{-1} = \mathrm{Ad}_{h^{-1}}$ (as is also immediately verified in coordinates: for translation groups $\mathrm{Ad}_h^\intercal = M_h^\intercal = M_h^{-1}$, and for matrix groups under the Frobenius inner product $\langle \mathrm{Ad}_h U, V\rangle = \mathrm{tr}((M_h U M_h^\intercal)^\intercal V) = \mathrm{tr}(U^\intercal M_h^\intercal V M_h) = \langle U, \mathrm{Ad}_{h^{-1}} V\rangle$). Since $u \in \mathfrak{g}$ is arbitrary, we conclude that $\nabla_G \mathcal{L}(g^*) = \mathrm{Ad}_{h^{-1}} \cdot \nabla_G \mathcal{L}(g^*)$ for all $h \in H$. Replacing h by h^{-1} yields $\nabla_G \mathcal{L}(g^*) \in \mathfrak{g}^H$. □

6.3. Gradient localization and orthogonal Lie algebra decomposition

Theorem 6.2 establishes that the Riemannian gradient $\nabla_G \mathcal{L}(g^*)$ is confined to the trivial representation $\mathfrak{g}^H$. At the same time, our identifiability analysis in Section 4 showed that whenever the intersection subgroup $N = G \cap H$ has positive dimension, the variational objective $\mathcal{L}(g^* \exp(tu)) = \mathcal{L}(g^*)$ is identically constant along the Lie algebra $\mathfrak{n} = T_e N$ of N. Combining these two independent structural constraints restricts the gradient to the orthogonal complement of $\mathfrak{n}$ inside $\mathfrak{g}^H$, and induces a canonical three-way orthogonal decomposition of the entire transport Lie algebra $\mathfrak{g}$. From a statistical estimation standpoint, this decomposes all local parameter perturbations around g^* into three intuitive categories: (1) *unidentifiable directions* $\mathfrak{n}$ that do not change the distribution q_{g^*} at all; (2) *symmetry-preserving directions* $\mathfrak{g}^H_{\text{bias}}$ (such as global scale) where mismatch between q_0 and $\tilde{p}$ can induce estimation bias; and (3) *symmetry-breaking directions* $\mathfrak{g}^\perp$ (such as location, orientation, or correlation shape) where symmetry forces the gradient at g^* to vanish.

Proposition 6.3 (Gradient localization and Lie algebra decomposition)**.** *Under the setting of Definition 3.3 with an Ad_H-invariant inner product on $\mathfrak{g}$, the transport Lie algebra $\mathfrak{g}$ admits the orthogonal direct sum decomposition*

$$\mathfrak{g} = \underbrace{\mathfrak{n}}_{\text{identifiability kernel (flat)}} \oplus \underbrace{\mathfrak{g}^H_{bias}}_{\text{bias directions}} \oplus \underbrace{\mathfrak{g}^\perp}_{\text{symmetry-breaking (forced to 0)}},$$

where the three mutually orthogonal subspaces are defined as follows:

- ***Identifiability kernel*** $\mathfrak{n} := \{u \in \mathfrak{g} \mid \exp(tu) \in N = G \cap H \ \ \forall t \in \mathbb{R}\}$*: the Lie algebra of the stabilizer intersection $N = G \cap H$. Along any direction $u \in \mathfrak{n}$, the transported distribution $q_{g^*\exp(tu)} = q_{g^*}$ is unchanged, so $\mathcal{L}(g^* \exp(tu)) = \mathcal{L}(g^*)$ is locally constant and* $\mathrm{proj}_{\mathfrak{n}}\big(\nabla_G \mathcal{L}(g^*)\big) = 0$.
- ***Bias subspace*** $\mathfrak{g}^H_{bias} := \mathfrak{g}^H \cap \mathfrak{n}^\perp$*: the subspace of H-invariant tangent directions orthogonal to the identifiability kernel $\mathfrak{n}$. Perturbations along $\mathfrak{g}^H_{bias}$ deform the distribution q_{g^*} while preserving its H-conjugate symmetry, and represent the only directions along which $\nabla_G \mathcal{L}(g^*)$ can be nonzero.*

- ***Symmetry-breaking subspace*** $\mathfrak{g}^\perp := (\mathfrak{g}^H)^\perp \cap \mathfrak{n}^\perp$*: the subspace of identifiable tangent directions that transform non-trivially under* Ad_H*. By Theorem 6.2, the gradient vanishes identically on this subspace:* $\mathrm{proj}_{\mathfrak{g}^\perp}\big(\nabla_G \mathcal{L}(g^*)\big) = 0$.

Consequently, the Riemannian gradient at the true transport parameter g^* *satisfies*

$$\nabla_G \mathcal{L}(g^*) \in \mathfrak{g}^H_{bias}\,.$$

Proof. By Theorem 6.2, $\nabla_G \mathcal{L}(g^*) \in \mathfrak{g}^H$. Furthermore, for any $u \in \mathfrak{n}$ and any $t \in \mathbb{R}$, we have $\exp(tu) \in N = G \cap H \subseteq H$. Because q_0 is H-invariant, $q_{\exp(tu)} = (\exp(tu))_\sharp q_0 = q_0$, which implies $q_{g^* \exp(tu)} = q_{g^*}$ and hence $\mathcal{L}(g^* \exp(tu)) = \mathcal{L}(g^*)$ for all $t \in \mathbb{R}$. Differentiating at $t = 0$ gives $\langle \nabla_G \mathcal{L}(g^*), u\rangle_{\mathfrak{g}} = 0$ for all $u \in \mathfrak{n}$, so $\nabla_G \mathcal{L}(g^*) \in \mathfrak{n}^\perp$. Combining both inclusions yields $\nabla_G \mathcal{L}(g^*) \in \mathfrak{g}^H \cap \mathfrak{n}^\perp = \mathfrak{g}^H_{\text{bias}}$.

To prove that $\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{g}^H_{\text{bias}} \oplus \mathfrak{g}^\perp$ is an orthogonal direct sum spanning $\mathfrak{g}$, note first that $\mathfrak{g}^H_{\text{bias}} \subseteq \mathfrak{n}^\perp$ and $\mathfrak{g}^\perp \subseteq \mathfrak{n}^\perp$ are both orthogonal to $\mathfrak{n}$ by definition, and $\mathfrak{g}^H_{\text{bias}} \subseteq \mathfrak{g}^H$ is orthogonal to $\mathfrak{g}^\perp \subseteq (\mathfrak{g}^H)^\perp$. Next, observe that $N = G \cap H$ is normalized by H: for any $n \in N$ and $h \in H$, $\Phi_h(n) = hnh^{-1}$ lies in G (since G is normalized by H) and in H (since H is a group), so $\Phi_h(N) = N$. Differentiating at e_G implies that $\mathfrak{n}$ is an Ad_H-invariant subspace of $\mathfrak{g}$ ($\mathrm{Ad}_h(\mathfrak{n}) = \mathfrak{n}$ for all $h \in H$; see Appendix B for details). Because $\mathfrak{n}$ is Ad_H-invariant and the inner product is Ad_H-invariant, the orthogonal projection onto $\mathfrak{g}^H$ maps $\mathfrak{n}$ into itself, so $\mathfrak{n}$ decomposes orthogonally as

$$\mathfrak{n} = (\mathfrak{n} \cap \mathfrak{g}^H) \;\oplus\; (\mathfrak{n} \cap (\mathfrak{g}^H)^\perp)\,.$$

Taking orthogonal complements inside $\mathfrak{g}^H$ and $(\mathfrak{g}^H)^\perp$ respectively gives $\mathfrak{g}^H = (\mathfrak{n} \cap \mathfrak{g}^H) \oplus \mathfrak{g}^H_{\text{bias}}$ and $(\mathfrak{g}^H)^\perp = (\mathfrak{n} \cap (\mathfrak{g}^H)^\perp) \oplus \mathfrak{g}^\perp$. Summing these decompositions over $\mathfrak{g} = \mathfrak{g}^H \oplus (\mathfrak{g}^H)^\perp$ yields $\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{g}^H_{\text{bias}} \oplus \mathfrak{g}^\perp$. □

Guiding examples: Three-way decomposition and statistical interpretation. We now evaluate the orthogonal decomposition $\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{g}^H_{\text{bias}} \oplus \mathfrak{g}^\perp$ for each of our six guiding examples, illustrating how Proposition 6.3 separates overparametrization, estimation bias, and unbiased recovery:

1. **Reflection** ($G = (\mathbb{R}, +)$, $H = \mathbb{Z}/2\mathbb{Z}$): The Lie algebra is $\mathfrak{g} = \mathbb{R}$. Because $N = \{0\}$ is trivial and $\mathrm{Ad}_\sigma u = -u$ leaves only 0 fixed, both the identifiability kernel and the bias subspace vanish:

   $$\mathfrak{n} = \{0\}\,, \qquad \mathfrak{g}^H_{\text{bias}} = \{0\}\,, \qquad \mathfrak{g}^\perp = \mathbb{R}\,.$$

   Every nonzero translation breaks reflection symmetry, so the entire parameter space lies in $\mathfrak{g}^\perp$ and $\nabla_G \mathcal{L}(\mu^*) = 0$ is guaranteed for any symmetric target p.
2. **Euclidean** ($G = (\mathbb{R}^d, +)$, $H = \mathrm{O}(d)$): The Lie algebra is $\mathfrak{g} = \mathbb{R}^d$, with $N = \{0\}$ and $\mathfrak{g}^H = \{0\}$. Thus

   $$\mathfrak{n} = \{0\}\,, \qquad \mathfrak{g}^H_{\text{bias}} = \{0\}\,, \qquad \mathfrak{g}^\perp = \mathbb{R}^d\,.$$

   All d translation coordinates are symmetry-breaking directions, forcing $\nabla_G \mathcal{L}(\mu^*) = 0$.
3. **Sphere** ($G = \mathrm{SO}(3)$, $H = \mathrm{SO}(2)$ fixing e_3): Under the isomorphism $\mathfrak{g} = \mathfrak{so}(3) \cong \mathbb{R}^3$, the stabilizer $N = \mathrm{SO}(2)$ has Lie algebra $\mathfrak{n} = \mathrm{span}(e_3)$, consisting of infinitesimal rotations around the symmetry axis e_3. As computed in Section 5, the trivial representation is also $\mathfrak{g}^H = \mathrm{span}(e_3)$. Because $\mathfrak{g}^H$ is entirely contained within the identifiability kernel $\mathfrak{n}$, the bias subspace vanishes:

   $$\mathfrak{n} = \mathrm{span}(e_3)\,, \qquad \mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H \cap \mathfrak{n}^\perp = \{0\}\,, \qquad \mathfrak{g}^\perp = \mathrm{span}(e_1, e_2)\,.$$

   Geometrically, $\mathfrak{g}^\perp = \mathrm{span}(e_1, e_2)$ consists of infinitesimal tilts of the modal axis e_3 toward the equatorial plane (spanning the tangent space $T_{e_3}\mathcal{S}^2$ of the quotient manifold $\mathcal{M} = \mathrm{SO}(3)/\mathrm{SO}(2) \cong \mathcal{S}^2$). Symmetry forces the gradient along all equatorial tilts to vanish, while the axial spin direction $\mathfrak{n}$ is flat by identifiability, guaranteeing $\nabla_G \mathcal{L}(R^*) = 0$.

4. **Covariance** ($G = \mathrm{GL}(d)$, $H = \mathrm{O}(d)$): The Lie algebra $\mathfrak{g} = \mathfrak{gl}(d) = \mathbb{R}^{d\times d}$ equipped with the Frobenius inner product decomposes into skew-symmetric matrices and symmetric matrices. The stabilizer $N = \mathrm{O}(d)$ has Lie algebra $\mathfrak{n} = \mathfrak{so}(d) = \{U \in \mathbb{R}^{d\times d} : U^\top = -U\}$, while the trivial subspace is the one-dimensional space of isotropic scalings $\mathfrak{g}^H = \mathrm{span}(I_d)$. Since I_d is symmetric and every matrix in $\mathfrak{so}(d)$ is skew-symmetric (and traceless), $\mathfrak{g}^H$ is completely orthogonal to $\mathfrak{n}$ ($\mathrm{tr}(I_d^\top U) = \mathrm{tr}(U) = 0$ for all $U \in \mathfrak{so}(d)$). We therefore obtain the classical Cartan-type matrix decomposition:

$$\mathfrak{n} = \mathfrak{so}(d)\,, \qquad \mathfrak{g}^H_{\text{bias}} = \mathrm{span}(I_d)\,, \qquad \mathfrak{g}^\perp = \mathrm{Sym}_0(d) := \{S \in \mathbb{R}^{d\times d} : S = S^\top,\ \mathrm{tr}(S) = 0\}\,.$$

   This decomposition has a striking statistical interpretation. The identifiability kernel $\mathfrak{n} = \mathfrak{so}(d)$ generates orthogonal right-rotations of A^* that leave the covariance matrix $\Sigma^* = A^*(A^*)^\top$ invariant. The symmetry-breaking subspace $\mathfrak{g}^\perp = \mathrm{Sym}_0(d)$ of traceless symmetric matrices parameterizes *anisotropic shape and correlation deformations* (changing the ratios of eigenvalues and orientation of principal axes while preserving volume to first order); along all $\frac{1}{2}d(d+1) - 1$ such directions, the gradient at A^* is forced to vanish by elliptical symmetry! However, along the single remaining scalar direction $\mathfrak{g}^H_{\text{bias}} = \mathrm{span}(I_d)$, the gradient satisfies $\nabla_G \mathcal{L}(A^*) = \lambda I_d$ for some scalar $\lambda \in \mathbb{R}$, corresponding to a *multiplicative global scale bias* between the dispersion of q_0 and that of $\tilde{p}$.
5. **Square** ($G = \mathrm{SO}(2)$, $H = D_4$): The Lie algebra is $\mathfrak{g} = \mathfrak{so}(2) \cong \mathbb{R}$. Because the overlap subgroup $N = C_4$ is discrete, its Lie algebra is zero-dimensional ($\mathfrak{n} = \{0\}$). At the same time, the reflections in D_4 force $\mathfrak{g}^H = \{0\}$. Consequently,

$$\mathfrak{n} = \{0\}\,, \qquad \mathfrak{g}^H_{\text{bias}} = \{0\}\,, \qquad \mathfrak{g}^\perp = \mathbb{R}\,.$$

   Every infinitesimal planar rotation breaks the reflection symmetry of the square, so $\nabla_G \mathcal{L}(R_{\theta^*}) = 0$ is guaranteed by D_4-invariance.
6. **Permutation** ($G = (\mathbb{R}^n, +)$, $H = \mathbb{S}_n$): The Lie algebra is $\mathfrak{g} = \mathbb{R}^n$ with trivial intersection $N = \{0\}$, so $\mathfrak{n} = \{0\}$. The trivial representation of $\mathbb{S}_n$ is the one-dimensional diagonal subspace $\mathfrak{g}^H = \mathrm{span}(\mathbf{1})$, which therefore coincides entirely with the bias subspace:

$$\mathfrak{n} = \{0\}\,, \qquad \mathfrak{g}^H_{\text{bias}} = \mathrm{span}(\mathbf{1})\,, \qquad \mathfrak{g}^\perp = \left\{u \in \mathbb{R}^n : \sum_{i=1}^n u_i = 0\right\}.$$

   Statistically, the $(n-1)$-dimensional symmetry-breaking subspace $\mathfrak{g}^\perp$ of zero-sum vectors governs all *relative contrasts* $\mu_i - \mu_j$ between coordinates: exchangeability forces the gradient at μ^* to vanish along every contrast vector $u \in \mathfrak{g}^\perp$, ensuring that relative coordinate differences are estimated without bias. By contrast, along the common-location direction $\mathfrak{g}^H_{\text{bias}} = \mathrm{span}(\mathbf{1})$, the gradient takes the form $\nabla_G \mathcal{L}(\mu^*) = \lambda \mathbf{1}$ for some $\lambda \in \mathbb{R}$, representing a uniform *additive location bias* across all coordinates if q_0 and $\tilde{p}$ are not centered at the same scalar mean/median level.

Proposition 6.3 unifies the phenomena observed in Margossian and Saul (2025) and extends them across general Lie groups. Whenever $\mathfrak{g}^H_{\text{bias}} = \{0\}$—either because $\mathfrak{g}^H = \{0\}$ (Reflection, Euclidean, Square) or because $\mathfrak{g}^H = \mathfrak{n}$ is absorbed into the stabilizer of q_0 (Sphere)—symmetry alone guarantees $\nabla_G \mathcal{L}(g^*) = 0$. Conversely, when $\mathfrak{g}^H_{\text{bias}} \neq \{0\}$ (Covariance and Permutation), symmetry eliminates all gradient components in $\mathfrak{g}^\perp$ while isolating any potential misspecification bias to the low-dimensional subspace $\mathfrak{g}^H_{\text{bias}}$. Table 2 summarizes the algebraic invariants and decomposition across all six guiding examples.

6.4. Bias from model mismatch and the Stein moment-matching condition

When the bias subspace $\mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H \cap \mathfrak{n}^\perp$ is non-trivial, Proposition 6.3 shows that $\nabla_G \mathcal{L}(g^*)$ lies in $\mathfrak{g}^H_{\text{bias}}$ but need not vanish. Statistically, a nonzero gradient component along $\mathfrak{g}^H_{\text{bias}}$ arises because

Table 2 | Algebraic invariants and Lie algebra decomposition $\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{g}^H_{\text{bias}} \oplus \mathfrak{g}^\perp$ for the six guiding examples. The final column indicates whether $\nabla_G \mathcal{L}(g^*) = 0$ is guaranteed by symmetry alone ($\mathfrak{g}^H_{\text{bias}} = \{0\}$).

Example	$\mathcal{X}$	G	H	$N = G \cap H$	Regime	$\mathfrak{g}^H$	$\mathfrak{g}^H_{\text{bias}}$	$\nabla = 0$?
1. Reflection	$\mathbb{R}$	$(\mathbb{R}, +)$	$\mathbb{Z}/2\mathbb{Z}$	$\{e\}$	Separable	$\{0\}$	$\{0\}$	Yes
2. Euclidean	$\mathbb{R}^d$	$(\mathbb{R}^d, +)$	$\mathrm{O}(d)$	$\{e\}$	Separable	$\{0\}$	$\{0\}$	Yes
3. Sphere	$\mathcal{S}^2$	$\mathrm{SO}(3)$	$\mathrm{SO}(2)$	$\mathrm{SO}(2)$	Internal	$\mathrm{span}(e_3)$	$\{0\}$	Yes
4. Covariance	$\mathbb{R}^d$	$\mathrm{GL}(d)$	$\mathrm{O}(d)$	$\mathrm{O}(d)$	Internal	$\mathrm{span}(I_d)$	$\mathrm{span}(I_d)$	No
5. Square	$\mathbb{R}^2$	$\mathrm{SO}(2)$	D_4	C_4	Partial	$\{0\}$	$\{0\}$	Yes
6. Permutation	$\mathbb{R}^n$	$(\mathbb{R}^n, +)$	$\mathbb{S}_n$	$\{e\}$	Separable	$\mathrm{span}(\mathbf{1})$	$\mathrm{span}(\mathbf{1})$	No

perturbations $\exp(tu)$ with $u \in \mathfrak{g}^H_{\text{bias}}$ transform the base distribution q_0 into a distinct H-symmetric distribution $q_{\exp(tu)}$. If the standardized target $\tilde{p}$ and the base q_0 differ in their scale or position along this symmetry-preserving subgroup $\exp(\mathfrak{g}^H_{\text{bias}})$, the KL divergence pulls the variational parameter away from g^* along $\mathfrak{g}^H_{\text{bias}}$ to compensate for the mismatch between q_0 and $\tilde{p}$.

Illustrative examples of estimation bias under model mismatch. Before deriving the general analytical criterion for when the bias vanishes, we illustrate how mismatch along $\mathfrak{g}^H_{\text{bias}}$ manifests in three concrete settings:

- **Multiplicative scale bias in covariance estimation (Example 4)**: Suppose the base distribution is $q_0 = \mathcal{N}(0, I_d)$, while the standardized pullback of the target is $\tilde{p} = \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \neq 1$ (so that $p = \mathcal{N}(0, \sigma^2 A^*(A^*)^\top)$). Both q_0 and $\tilde{p}$ are $\mathrm{O}(d)$-symmetric, yet their radial scales differ by σ. Consequently, $\mathcal{L}(A)$ is minimized at $\hat{A} = \sigma A^*$ rather than A^*, and the gradient at A^* has a nonzero component $\nabla_G \mathcal{L}(A^*) = (\sigma^{-2} - 1) I_d \in \mathrm{span}(I_d) = \mathfrak{g}^H_{\text{bias}}$ (so that the negative gradient $-\nabla_G \mathcal{L}(A^*) = (1 - \sigma^{-2}) I_d$ drives A toward σA^*).
- **Common-location additive bias under exchangeability (Example 6)**: Suppose the base distribution is $q_0(x) \propto \exp(-\kappa \|x\|_1)$ (centered at the origin), while the standardized pullback is $\tilde{p}(x) \propto \exp(-\kappa \|x - \beta \mathbf{1}\|_1)$ with $\beta \neq 0$. Because shifting all coordinates by $\beta \mathbf{1}$ commutes with all coordinate permutations in $\mathbb{S}_n$, $\tilde{p}$ remains fully $\mathbb{S}_n$-symmetric! However, q_0 and $\tilde{p}$ have different common coordinate medians (0 versus β), so $\mathcal{L}(\mu)$ is minimized at $\hat{\mu} = \mu^* + \beta \mathbf{1}$ rather than μ^*, producing a nonzero gradient at μ^* along $\mathrm{span}(\mathbf{1}) = \mathfrak{g}^H_{\text{bias}}$.
- **Angular drift under chiral C_4 symmetry (Modified Example 5)**: Suppose we weaken the symmetry of the standardized target $\tilde{p}$ in Example 5 from the full dihedral group D_4 to the pure rotational subgroup $H' = C_4$ (eliminating reflections) by setting $\tilde{p}(x) = q_0(R_\alpha^\top x)$ for an angle $\alpha \notin (\pi/4)\mathbb{Z}$. Because $C_4 \subset \mathrm{SO}(2)$ is abelian, conjugation by $r_k \in C_4$ acts trivially on $\mathrm{SO}(2)$, so $\mathrm{Ad}_{r_k} = \mathrm{id}$ and the trivial subspace expands from $\{0\}$ to the entire Lie algebra: $\mathfrak{g}^{C_4}_{\text{bias}} = \mathfrak{g} = \mathfrak{so}(2) \cong \mathbb{R}$. With reflections removed, symmetry imposes no constraint on the rotational gradient, and the variational minimizer shifts to $\theta^* + \alpha \pmod{\pi/2}$.

Infinitesimal generators and the Stein moment-matching condition. To characterize precisely when $\nabla_G \mathcal{L}(g^*) = 0$ in the presence of a non-trivial bias subspace $\mathfrak{g}^H_{\text{bias}}$, we express the directional derivative of the pullback loss $J(u) = \mathrm{KL}(q_{\exp(u)} \| \tilde{p})$ at $u = 0$ in terms of the log-density potentials of q_0 and $\tilde{p}$. Write

$$q_0(z) \propto \exp(-\psi(z)), \qquad \tilde{p}(z) \propto \exp(-\varphi(z)),$$

where $\psi, \varphi : \mathcal{X} \to \mathbb{R}$ are smooth H-invariant potential functions. Describing how an infinitesimal transport $u \in \mathfrak{g}$ deforms points $z \in \mathcal{X}$ requires the fundamental vector field (or *velocity field*) induced by the group action:

Definition 6.4 (Infinitesimal generator / velocity field)**.** For each tangent direction $u \in \mathfrak{g}$, the *velocity field* (or *infinitesimal generator*) $V_u : \mathcal{X} \to T\mathcal{X}$ is the vector field on $\mathcal{X}$ assigning to each point $x \in \mathcal{X}$ its instantaneous velocity under the flow $t \mapsto \exp(tu) \cdot x$:

$$V_u(x) := \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (\exp(tu) \cdot x) .$$

For our primary transport groups, the velocity field $V_u(x)$ takes an immediate closed form:

- **Additive translations** ($G = (\mathbb{R}^d, +)$, $\exp(tu) \cdot x = x + tu$): $V_u(x) = u$ (a constant vector field on $\mathbb{R}^d$).
- **Linear matrix transformations** ($G = \mathrm{GL}(d)$, $\exp(tU) \cdot x = e^{tU}x$): $V_U(x) = Ux$ (a linear vector field on $\mathbb{R}^d$).
- **Rotations on the sphere** ($G = \mathrm{SO}(3)$, $\mathfrak{g} \cong \mathbb{R}^3$): $V_\omega(x) = \omega \times x$ (the tangential rotational velocity field on $\mathcal{S}^2$).

When a random variable $Z \sim q_0$ is pushed forward along the flow $x(t) = \exp(tu) \cdot Z$, the KL divergence $J(tu) = \mathrm{KL}(q_{\exp(tu)} \| \tilde{p}) = \mathbf{E}_{Z\sim q_0}[\log q_0(Z) - \log \tilde{p}(\exp(tu) \cdot Z) - \log |\det J_{\exp(tu)}(Z)|]$ varies at first order through two distinct mechanisms:

(i) *Cross-entropy / target potential variation*: Differentiating $-\log \tilde{p}(\exp(tu) \cdot Z) = \varphi(\exp(tu) \cdot Z) +$ const at $t = 0$ yields the directional derivative of the target potential along the velocity field, $\mathbf{E}_{q_0}[\langle \nabla\varphi(Z), V_u(Z)\rangle]$.

(ii) *Entropic / Jacobian volume variation*: By Jacobi's formula, differentiating the log-Jacobian determinant $-\log |\det J_{\exp(tu)}(Z)|$ at $t = 0$ yields the negative divergence of the velocity field, $-\mathbf{E}_{q_0}[\mathrm{div}(V_u)(Z)]$. Applying integration by parts (Stein's identity (Stein, 1981); cf. Hyvärinen (2005)) against the density $q_0(z) \propto \exp(-\psi(z))$ re-expresses this volume expansion rate directly in terms of the base potential score:

$$\mathbf{E}_{Z\sim q_0}\big[\mathrm{div}(V_u)(Z)\big] = -\mathbf{E}_{Z\sim q_0}\big[\langle \nabla \log q_0(Z), V_u(Z)\rangle\big] = \mathbf{E}_{Z\sim q_0}\big[\langle \nabla\psi(Z), V_u(Z)\rangle\big] .$$

Combining these two terms yields our second main theorem, proven in detail in Appendix A.

Theorem 6.5 (Stein moment-matching condition for stationarity)**.** *Under the setting of Definition 3.3, assume that $q_0 \propto \exp(-\psi)$ and $\tilde{p} \propto \exp(-\varphi)$ have differentiable (or locally Lipschitz and weakly differentiable a.e.) log-potentials ψ and φ satisfying the integrability condition (locally uniformly for small $|t|$) $\mathbf{E}_{z\sim q_0}\big[\|\nabla\psi(z)\|\|V_u(z)\| + \|\nabla\varphi(z)\|\|V_u(z)\| + |\mathrm{div}(V_u)(z)|\big] < \infty$ and the boundary condition $\lim_{R\to\infty} \int_{\partial B_R} q_0(z)\|V_u(z)\|\,\mathrm{d}S(z) = 0$ (which holds automatically when $\mathcal{X}$ is compact, and on $\mathbb{R}^d$ whenever q_0 has sub-Gaussian or faster tail decay and $\|V_u(z)\|$ grows at most polynomially). Then for any $u \in \mathfrak{g}$, the directional derivative of the pullback loss $J(u) = \mathrm{KL}(q_{\exp(u)} \| \tilde{p})$ at $u = 0$ is given by*

$$\langle \nabla_G \mathcal{L}(g^*), u\rangle_{\mathfrak{g}} = \langle \nabla J(0), u\rangle_{\mathfrak{g}} = \mathbf{E}_{z\sim q_0}\left[\langle \nabla\varphi(z), V_u(z)\rangle\right] - \mathbf{E}_{z\sim q_0}\left[\langle \nabla\psi(z), V_u(z)\rangle\right] .$$

In particular, since $\nabla_G \mathcal{L}(g^*) \in \mathfrak{g}^H_{\mathrm{bias}}$ by Proposition 6.3, the true transport parameter g^* is a first-order critical point of $\mathcal{L}$ ($\nabla_G \mathcal{L}(g^*) = 0$) if and only if the following moment-matching condition holds for every basis direction $u \in \mathfrak{g}^H_{\mathrm{bias}}$:

$$\mathbf{E}_{z\sim q_0}\left[\langle \nabla\psi(z), V_u(z)\rangle\right] = \mathbf{E}_{z\sim q_0}\left[\langle \nabla\varphi(z), V_u(z)\rangle\right] . \tag{2}$$

Statistically, the quantity $\langle \nabla\psi(z), V_u(z)\rangle = -\langle \nabla \log q_0(z), V_u(z)\rangle$ is the *directional score function* of the base model along the transformation generator V_u (for instance, the radial virial score $z^\top \nabla\psi(z)$ measuring overall dispersion when $V_u(z) = z$ is an isotropic scaling, or the summed score $\mathbf{1}^\top \nabla\psi(z)$

measuring common location when $V_u(z) = \mathbf{1}$ is a diagonal shift). Equation (2) therefore has a transparent statistical meaning: the true parameter g^* is a stationary point of the variational objective if and only if the standardized target $\tilde{p}$ and the base distribution q_0 have matching expected directional scores under q_0 along every symmetry-preserving bias direction $u \in \mathfrak{g}^H_{\text{bias}}$.

Remark (Constructive debiasing by base distribution calibration). When the moment-matching identity (2) is violated, $\nabla_G \mathcal{L}(g^*)$ has a nonzero component in $\mathfrak{g}^H_{\text{bias}}$, reflecting a calibration mismatch between the variational base q_0 and the standardized target $\tilde{p}$ along the symmetry-preserving subgroup $\exp(\mathfrak{g}^H_{\text{bias}})$. Crucially, Theorem 6.5 provides a direct constructive procedure to eliminate this bias without altering the symmetry structure of the variational family: one recalibrates the base distribution by pushing q_0 forward along an element $g_\Omega = \exp(u_\Omega)$ with $u_\Omega \in \mathfrak{g}^H_{\text{bias}}$, defining the calibrated base

$$q_\Omega := (g_\Omega)_\sharp q_0 \,.$$

Because every $u_\Omega \in \mathfrak{g}^H_{\text{bias}} \subseteq \mathfrak{g}^H$ generates transformations in the centralizer G^H commuting with H, the recalibrated base q_Ω remains strictly H-symmetric. Specifically, replacing $q_0 \propto \exp(-\psi)$ by $q_\Omega \propto \exp(-\psi_\Omega)$ in (2) and pulling back to $z' \sim q_0$ via $z = g_\Omega \cdot z'$ (using Stein's identity on the left-hand side) yields the calibration equation

$$\mathbf{E}_{z'\sim q_0}\left[\operatorname{div}(V_u)(g_\Omega \cdot z')\right] = \mathbf{E}_{z'\sim q_0}\left[\langle \nabla\varphi(g_\Omega \cdot z'), V_u(g_\Omega \cdot z')\rangle\right], \qquad \forall u \in \mathfrak{g}^H_{\text{bias}},$$

which admits a solution $u_\Omega \in \mathfrak{g}^H_{\text{bias}}$ whenever the left-hand side lies in the range of the right-hand side along $\exp(\mathfrak{g}^H_{\text{bias}})$. Because $\dim(\mathfrak{g}^H_{\text{bias}})$ is typically very small (equal to 1 in both the Covariance and Permutation examples), solving this condition reduces to a simple scalar algebraic equation for u_Ω. When the functional form of the standardized potential φ is known (for example, when approximating a known heavy-tailed elliptical family by a Gaussian family), u_Ω can be computed offline prior to optimization; when φ is unknown, one can alternatively eliminate $\mathfrak{g}^H_{\text{bias}}$ by restricting the transport group G to a subgroup $G_0 \subset G$ (structural gauge fixing), as discussed in Section 7.

Guiding examples: Explicit moment-matching equations and base calibration. We conclude this section by working out the moment-matching condition (2) explicitly in our three examples with non-vanishing directional derivatives (full derivations are provided in Appendix C):

- **Covariance (Example 4)**: Here $\mathfrak{g}^H_{\text{bias}} = \operatorname{span}(I_d)$ is one-dimensional, and the single basis generator $u = I_d$ induces the radial scaling velocity field $V_{I_d}(z) = I_d z = z$. For a recalibrated isotropic Gaussian base $q_\Omega = \mathcal{N}(0, \lambda^2 I_d)$ with potential $\psi_\lambda(z) = \|z\|_2^2/(2\lambda^2)$, the left-hand side of (2) is the virial expectation

$$\mathbf{E}_{z\sim q_\Omega}\left[\langle \nabla\psi_\lambda(z), z\rangle\right] = \frac{1}{\lambda^2}\mathbf{E}_{z\sim q_\Omega}\left[\|z\|_2^2\right] = d \,.$$

  When the standardized target is $\tilde{p} = \mathcal{N}(0, \sigma^2 I_d)$ with $\varphi(z) = \|z\|_2^2/(2\sigma^2)$, the right-hand side is $\sigma^{-2}\mathbf{E}_{q_\Omega}[\|z\|_2^2] = d\lambda^2/\sigma^2$, so (2) holds if and only if $\lambda = \sigma$. More generally, for any non-Gaussian spherically symmetric target $\tilde{p}(z) \propto \exp(-\varphi(z))$ with radial potential $\varphi(z) = \rho(\|z\|_2)$, we have $\langle \nabla\varphi(z), z\rangle = \rho'(\|z\|_2)\|z\|_2$, and the exact scale calibration condition for $\lambda > 0$ reads

$$\frac{1}{\lambda^2}\mathbf{E}_{z\sim\mathcal{N}(0,\lambda^2 I_d)}\left[\|z\|_2^2\right] = d = \mathbf{E}_{z\sim\mathcal{N}(0,\lambda^2 I_d)}\left[\rho'(\|z\|_2)\|z\|_2\right].$$

  Choosing λ to satisfy this scalar equation eliminates the radial scale bias at A^*, making $\hat{A} = A^*$ stationary and recovering the target scatter parameter $\Sigma^* = A^*(A^*)^\top$ (while the fitted Gaussian density itself has covariance $\lambda^2\Sigma^*$).

- **Permutation (Example 6)**: Here $\mathfrak{g}_{\text{bias}}^H = \text{span}(\mathbf{1})$ is one-dimensional, and the generator $u = \mathbf{1}$ induces the constant velocity field $V_{\mathbf{1}}(z) = \mathbf{1}$. The moment-matching condition (2) therefore requires the expected sum of partial derivatives of the potentials to match under q_0:

$$\mathbf{E}_{z\sim q_0}\left[\langle \nabla\psi(z), \mathbf{1}\rangle\right] = \mathbf{E}_{z\sim q_0}\left[\langle \nabla\varphi(z), \mathbf{1}\rangle\right].$$

  For the product Laplace base $q_0(z) \propto \exp(-\kappa\|z\|_1)$, we have $\langle \nabla\psi(z), \mathbf{1}\rangle = \kappa \sum_{i=1}^n \text{sign}(z_i)$, whose expectation under q_0 is 0. If $\tilde{p}(z) \propto \exp(-\kappa\|z - \beta\mathbf{1}\|_1)$, shifting the base distribution along $\mathfrak{g}_{\text{bias}}^H$ by $u_\Omega = \lambda\mathbf{1}$ to $q_\Omega(z) \propto \exp(-\kappa\|z - \lambda\mathbf{1}\|_1)$ achieves $\mathbf{E}_{q_\Omega}[\langle \nabla\varphi(z), \mathbf{1}\rangle] = 0$ precisely when $\lambda = \beta$.
- **Square vs. Chiral C_4 rotation (Example 5)**: For the quartic potential $\psi(z) = \kappa(x^4 + y^4)$ on $\mathbb{R}^2$ (written in polar coordinates $z = (r\cos\phi, r\sin\phi)$ as $\psi(r,\phi) = \kappa r^4(\frac{3}{4} + \frac{1}{4}\cos 4\phi)$) and the tilted target pullback $\varphi(z) = \psi(R_{-\alpha}z)$, the rotational generator $\dot{\theta} \in \mathfrak{so}(2) \cong \mathbb{R}$ induces the tangential velocity field $V_{\dot{\theta}}(x,y) = \dot{\theta}(-y,x)^\top = \dot{\theta}\frac{\partial}{\partial\phi}$. The directional derivatives of ψ and φ along $V_{\dot{\theta}}$ are

$$\langle \nabla\psi(z), V_{\dot{\theta}}(z)\rangle = -\kappa\dot{\theta}r^4\sin(4\phi), \qquad \langle \nabla\varphi(z), V_{\dot{\theta}}(z)\rangle = -\kappa\dot{\theta}r^4\sin\big(4(\phi-\alpha)\big).$$

  Because q_0 is invariant under the reflection $\phi \mapsto -\phi$, the odd trigonometric moment vanishes ($\mathbf{E}_{q_0}[r^4\sin(4\phi)] = 0$), and expanding $\sin(4(\phi-\alpha)) = \sin(4\phi)\cos(4\alpha) - \cos(4\phi)\sin(4\alpha)$ yields

$$\langle \nabla J(0), \dot{\theta}\rangle_{\mathfrak{g}} = -\kappa M_4 \sin(4\alpha)\,\dot{\theta}, \qquad \text{where } M_4 := -\mathbf{E}_{z\sim q_0}\left[r^4\cos(4\phi)\right] > 0.$$

  Consequently, $\nabla J(0) = 0$ if and only if $\sin(4\alpha) = 0$, i.e., $\alpha \equiv 0 \pmod{\pi/4}$, which is precisely the condition that $\tilde{p}$ possesses the reflection symmetries of D_4 aligned with the axes or diagonals of q_0.

7. Discussion and future directions

We have developed a general Lie-theoretic framework characterizing how distributional symmetries of the target and base measures interact with transport-based variational families. By studying the conjugation action of the symmetry group H on the transport group G and its linearization—the adjoint representation Ad_H on the transport Lie algebra $\mathfrak{g}$—Theorem 6.2 and Proposition 6.3 establish that the Riemannian gradient $\nabla_G\mathcal{L}(g^*)$ decomposes across three mutually orthogonal subspaces: the identifiability kernel $\mathfrak{n}$ (where the objective is locally constant), the symmetry-breaking subspace $\mathfrak{g}^\perp$ (where the gradient is forced to vanish by symmetry), and the symmetry-preserving bias subspace $\mathfrak{g}_{\text{bias}}^H = \mathfrak{g}^H \cap \mathfrak{n}^\perp$ (where model misspecification can induce a nonzero gradient). Theorem 6.5 complements this geometric decomposition by providing an exact Stein-type moment-matching condition characterizing when the gradient along $\mathfrak{g}_{\text{bias}}^H$ vanishes.

Addressing the conjecture of Margossian and Saul (2025). Our results provide a precise algebraic answer to the conjecture of Margossian and Saul (2025) that each symmetry of a target distribution p corresponds to a statistic that variational inference can recover under mild conditions. Within the class of Lie group transformation models, given a compact symmetry group H of the standardized target $\tilde{p}$, any H-normalized transport group G defines a smooth manifold of identifiable parameters $\mathcal{M} = G/(G\cap H)$ whose tangent space at the true coset $[g^*]$ is canonically identified with $\mathfrak{n}^\perp = \mathfrak{g}_{\text{bias}}^H \oplus \mathfrak{g}^\perp$. The components of $[g^*]$ along the symmetry-breaking directions $\mathfrak{g}^\perp$ are automatically first-order stationary points of the variational objective $\text{KL}(q_g\|p)$ for *any* H-symmetric base distribution q_0, without requiring parametric specification ($p \in \mathcal{Q}_G$). Whenever $\mathfrak{g}_{\text{bias}}^H = \{0\}$, the entire coset $[g^*] \in G/N$ is recovered at first order by symmetry alone.

Structural gauge fixing versus base calibration. When the bias subspace $\mathfrak{g}_{\text{bias}}^H$ is nontrivial, our framework highlights two distinct remedies for eliminating estimation bias:

(i) *Base calibration via moment matching (Theorem 6.5)*: One retains the full transport group G and calibrates the base distribution $q_0 \mapsto q_\Omega = (\exp(u_\Omega))_\sharp q_0$ along $u_\Omega \in \mathfrak{g}^H_{\text{bias}}$ so that the directional score moments (2) match, recovering the full parameter g^* (modulo N).

(ii) *Structural gauge fixing (subgroup restriction)*: When calibrating q_0 is impractical because the radial profile of $\tilde{p}$ is unknown, one can instead restrict the variational family to a smaller, H-normalized Lie subgroup $G_0 \subset G$ whose Lie algebra $\mathfrak{g}_0 = T_e G_0$ is a Lie subalgebra ($[\mathfrak{g}_0, \mathfrak{g}_0] \subseteq \mathfrak{g}_0$) transverse to $\mathfrak{g}^H_{\text{bias}}$, i.e., $\mathfrak{g}_0 \cap \mathfrak{g}^H_{\text{bias}} = \{0\}$. In the restricted family Q_{G_0}, the bias subspace vanishes identically, so the projection g_0^* of g^* onto $G_0/(G_0 \cap H)$ is guaranteed to be a stationary point for any H-symmetric base q_0 (whenever g^* factors as $g^* = g_0^* g^*_{\text{bias}} n$ with $g_0^* \in G_0$, $g^*_{\text{bias}} \in G^H$, and $n \in N$, so that pulling p back by g_0^* preserves H-invariance), at the cost of no longer estimating the parameter component along $\mathfrak{g}^H_{\text{bias}}$.

Covariance estimation (Example 4, $G = \mathrm{GL}(d)$, $H = \mathrm{O}(d)$) provides an instructive comparison between these strategies. Here the bias subspace is the one-dimensional trace direction $\mathfrak{g}^H_{\text{bias}} = \mathrm{span}(I_d)$. From a Lie-theoretic standpoint, the canonical H-normalized subgroup of $\mathrm{GL}(d)$ that excises $\mathrm{span}(I_d)$ is the special linear group $G_0 = \mathrm{SL}(d) = \{A \in \mathrm{GL}(d) : \det A = 1\}$, whose Lie algebra $\mathfrak{sl}(d) = \{M \in \mathfrak{gl}(d) : \mathrm{tr}(M) = 0\}$ is precisely $\mathfrak{n} \oplus \mathfrak{g}^\perp = \mathfrak{so}(d) \oplus \mathrm{Sym}_0(d)$. Restricting transport to $\mathrm{SL}(d)$ yields the quotient manifold $\mathcal{M}_0 = \mathrm{SL}(d)/\mathrm{SO}(d)$ of unit-determinant shape matrices ($\det \Sigma = 1$), recovering the exact covariance shape while leaving only the single scalar generalized variance $\det \Sigma^*$ unestimated. By comparison, Margossian and Saul (2025) optimize over the full unconstrained group $G = \mathrm{GL}(d)$ (obtaining a stationary point $\hat{\Sigma} = c\Sigma^*$ proportional to Σ^*) and eliminate the scale factor $c > 0$ post-hoc by normalizing $\hat{\Sigma}$ into the *correlation matrix* ($\Sigma_{ii} = 1$ for $i = 1, \ldots, d$). Our decomposition clarifies both post-hoc normalization and prior subgroup restriction: from the unconstrained optimum $\hat{\Sigma} = c\Sigma^*$, one can in fact extract the full unit-determinant shape matrix $(\det \hat{\Sigma})^{-1/d}\hat{\Sigma} \in \mathrm{SL}(d)/\mathrm{SO}(d)$ (preserving the $d - 1$ ratios of marginal variances that are discarded by the correlation matrix), while if one wishes to restrict the optimization domain a priori to a subgroup $G_0 \subset \mathrm{GL}(d)$, restricting to $\mathrm{SL}(d)$ preserves full $\mathrm{O}(d)$ conjugation invariance whereas enforcing unit diagonal entries breaks it. Similarly, in the Permutation model (Example 6, $G = (\mathbb{R}^n, +)$, $H = \mathbb{S}_n$), restricting G to the $(n-1)$-dimensional zero-sum hyperplane $G_0 = \{\mu \in \mathbb{R}^n : \sum_{i=1}^n \mu_i = 0\} = \mathfrak{g}^\perp$ is normalized by $\mathbb{S}_n$ and guarantees unbiased recovery of all centered coordinate offsets $\mu^* - \bar{\mu}^* \mathbf{1}$.

From first-order stationarity to second-order and global optimality. Theorem 6.2 is a first-order result: it guarantees that $\nabla_G \mathcal{L}(g^*) = 0$ whenever $\mathfrak{g}^H_{\text{bias}} = \{0\}$ (or when moment matching holds), which means g^* is a critical point of $\mathcal{L}$. Whether g^* is a strict local minimizer (modulo the flat coset fiber N) or a global minimizer depends on higher-order curvature properties of the target and base distributions. Specifically, g^* defines a strict local minimum on the quotient manifold $\mathcal{M} = G/N$ if the Riemannian Hessian $\mathrm{Hess}_G \mathcal{L}(g^*) = \nabla^2 J(0)$ is strictly positive definite on the horizontal subspace $\mathfrak{n}^\perp = \mathfrak{g}^H_{\text{bias}} \oplus \mathfrak{g}^\perp$. When the target p is strongly multimodal—for instance, in Example 1 when $p = \frac{1}{2}\mathcal{N}(\mu^* + \delta, 1) + \frac{1}{2}\mathcal{N}(\mu^* - \delta, 1)$ has two widely separated modes ($\delta \gg 1$) while $q_0 = \mathcal{N}(0, 1)$ is unimodal with unit variance—the symmetry center μ^* remains a critical point by Theorem 6.2, but bifurcates from a global minimum into a local maximum (or a saddle point in higher dimensions) straddled by two symmetry-broken global minima near $\mu^* \pm \delta$. For translation families ($G = \mathbb{R}^d$), log-concavity of p ensures convexity of $\mathcal{L}(\mu)$ and hence global uniqueness of the stationary point μ^*; extending such global uniqueness guarantees to non-abelian Lie groups via geodesic convexity on Riemannian quotient manifolds $\mathcal{M} = G/N$ (Boumal, 2023) is an important direction for future research.

Further extensions. Several additional directions merit investigation. First, in empirical applications, target distributions often satisfy H-conjugate symmetry only approximately; combining our representation-theoretic decomposition with perturbation bounds on $\nabla J(0)$ would quantify how an ϵ-asymmetry in $\tilde{p}$ induces an $O(\epsilon)$ drift of the variational stationary point along $\mathfrak{g}^{\perp}$. Second, while Theorem 6.2 already holds for any divergence functional that is invariant under diffeomorphisms of $\mathcal{X}$ (including all f-divergences and Rényi α-divergences, since steps (†), (*), and (**) in its proof rely solely on pushforward invariance), extending the Stein moment-matching formula of Theorem 6.5 to general f-divergences and Wasserstein distances will yield analogous base-calibration recipes beyond the KL divergence. Finally, our assumption that H acts on $\mathcal{X}$ by linear isometries can be relaxed to arbitrary isometric actions of compact Lie groups on Riemannian manifolds $(\mathcal{X}, \mathrm{vol}_{\mathcal{X}})$: replacing the Euclidean gradient and divergence in Theorem 6.5 by the Riemannian gradient $\nabla_{\mathcal{X}}$ and Laplace–Beltrami divergence $\mathrm{div}_{\mathcal{X}}$ leaves all our main theorems intact.

References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

S.-i. Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, 2016.

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

J. V. Bondar and P. Milnes. Amenability: A survey for statistical applications of Hunt–Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57 (1):103–128, 1981.

N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

M. L. Eaton. *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics and American Statistical Association, 1989.

W. Fulton and J. Harris. *Representation Theory: A First Course*, volume 129 of *Graduate Texts in Mathematics*. Springer, 1991.

L. Ghafourpour, S. Chewi, A. Figalli, and A.-A. Pooladian. Variational inference via radial transport. In *International Conference on Artificial Intelligence and Statistics*, 2026.

B. C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2015.

W. J. Hall, R. A. Wijsman, and J. K. Ghosh. The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Mathematical Statistics*, 36(2):575–614, 1965.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

A. W. Knapp. *Lie Groups Beyond an Introduction*, volume 140 of *Progress in Mathematics*. Birkhäuser, 2 edition, 2002.

I. Kobyzev, S. J. D. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

J. Köhler, L. Klein, and F. Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, pages 5361–5370. PMLR, 2020.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 3 edition, 2005.

C. C. Margossian and L. K. Saul. Variational inference in location-scale families: Exact recovery of the mean and correlation matrix. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pages 3466–3474. PMLR, 2025.

C. C. Margossian, L. Pillaud-Vivien, and L. K. Saul. Variational inference for uncertainty quantification: An analysis of trade-offs. *Journal of Machine Learning Research*, 26(202):1–41, 2025a.

C. C. Margossian, I. E. Rankin, and L. K. Saul. Generalized guarantees for variational inference in the presence of even and elliptical symmetry. *arXiv preprint arXiv:2511.01064*, 2025b.

D. Marks, D. Paccagnan, and M. van der Wilk. Symmetry guarantees statistic recovery in variational inference. *arXiv preprint arXiv:2604.18310*, 2026.

M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*. Chapman & Hall/CRC, 1993.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, volume 33, pages 814–822. PMLR, 2014.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9 (6):1135–1151, 1981.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

M. Weiler and G. Cesa. General E(2)-equivariant steerable CNNs. *Advances in Neural Information Processing Systems*, 32, 2019.

R. A. Wijsman. *Invariant Measures on Groups and Their Use in Statistics*, volume 14 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 1990.

L. Zellinger and A. Vergari. Even more guarantees for variational inference in the presence of symmetries. *arXiv preprint arXiv:2604.21407*, 2026.

A. Proof of Theorem 6.5

We provide the complete proof of the Stein moment-matching condition (Theorem 6.5). Recall that $q_0(z) = Z_0^{-1}\exp(-\psi(z))$ and $\tilde{p}(z) = Z_{\tilde{p}}^{-1}\exp(-\varphi(z))$ have differentiable potentials $\psi, \varphi : \mathcal{X} \to \mathbb{R}$ satisfying the integrability and boundary conditions stated in Theorem 6.5.

Proof of Theorem 6.5. Recall from the proof of Theorem 6.2 that the pullback objective around g^* is given by $J(u) = \mathcal{L}(g^*\exp(u)) = \mathrm{KL}(q_{\exp(u)}\|\tilde{p})$ for $u \in \mathfrak{g}$. We compute the directional derivative $\langle \nabla_G \mathcal{L}(g^*), u\rangle_{\mathfrak{g}} = \langle \nabla J(0), u\rangle_{\mathfrak{g}} = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} J(tu)$ for an arbitrary tangent vector $u \in \mathfrak{g}$.

Step 1: Pullback to the fixed base measure q_0. Under the diffeomorphism $x = \exp(tu)\cdot z$ with $z \sim q_0$, the pushforward density $q_{\exp(tu)}$ satisfies $q_{\exp(tu)}(\exp(tu)\cdot z)\,|\det J_{\exp(tu)}(z)| = q_0(z)$. Substituting the change of variables $x = \exp(tu)\cdot z$ (with $\mathrm{d}x = |\det J_{\exp(tu)}(z)|\,\mathrm{d}z$) into the definition of the KL divergence gives

$$J(tu) = \int_{\mathcal{X}} q_0(z)\Big[\log q_0(z) - \log \tilde{p}\big(\exp(tu)\cdot z\big) - \log\big|\det J_{\exp(tu)}(z)\big|\Big]\,\mathrm{d}z\,.$$

Note that the first integrand term $\log q_0(z)$ is independent of t and therefore does not contribute to the derivative with respect to t.

Step 2: Differentiation at $t = 0$. Differentiating under the integral sign at $t = 0$ (justified by dominated convergence under a standard integrable local dominating bound on $t \in (-\epsilon, \epsilon)$ together with the integrability assumption of Theorem 6.5), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} J(tu) = -\int_{\mathcal{X}} q_0(z)\left[\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\log\tilde{p}\big(\exp(tu)\cdot z\big) + \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\log\big|\det J_{\exp(tu)}(z)\big|\right]\mathrm{d}z\,.$$

For the first term, since $\log\tilde{p}(x) = -\varphi(x) - \log Z_{\tilde{p}}$ and $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}(\exp(tu)\cdot z) = V_u(z)$ by Definition 6.4, the chain rule yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\log\tilde{p}\big(\exp(tu)\cdot z\big) = \big\langle \nabla\log\tilde{p}(z), V_u(z)\big\rangle = -\big\langle\nabla\varphi(z), V_u(z)\big\rangle\,.$$

For the second term, the flow map $\phi_t(z) = \exp(tu)\cdot z$ satisfies $\phi_0(z) = z$ (with $J_{\phi_0}(z) = I_d$) and $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} J_{\phi_t}(z) = \nabla V_u(z)$. Applying Jacobi's formula for the derivative of a log-determinant gives the divergence of the velocity field V_u:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\log\big|\det J_{\exp(tu)}(z)\big| = \mathrm{tr}\big(\nabla V_u(z)\big) = \mathrm{div}(V_u)(z) = \sum_{i=1}^{d}\frac{\partial (V_u)_i}{\partial z_i}(z)\,.$$

Step 3: Integration by parts (Stein's identity). Using the product rule $\mathrm{div}(q_0 V_u) = q_0\,\mathrm{div}(V_u) + \langle\nabla q_0, V_u\rangle$ and the divergence theorem on $\mathcal{X}$, the boundary integral vanishes by our boundary condition, and since $\nabla q_0(z) = q_0(z)\nabla\log q_0(z) = -q_0(z)\nabla\psi(z)$, we obtain

$$\begin{aligned}\int_{\mathcal{X}} q_0(z)\,\mathrm{div}(V_u)(z)\,\mathrm{d}z &= -\int_{\mathcal{X}}\big\langle\nabla q_0(z), V_u(z)\big\rangle\,\mathrm{d}z\\ &= -\int_{\mathcal{X}} q_0(z)\big\langle\nabla\log q_0(z), V_u(z)\big\rangle\,\mathrm{d}z = \int_{\mathcal{X}} q_0(z)\big\langle\nabla\psi(z), V_u(z)\big\rangle\,\mathrm{d}z\,.\end{aligned}$$

Step 4: Combining both contributions. Substituting the results of Steps 2 and 3 into the derivative of $J(tu)$ yields

$$\langle \nabla_G \mathcal{L}(g^*), u\rangle_{\mathfrak{g}} = \langle \nabla J(0), u\rangle_{\mathfrak{g}} = \mathbf{E}_{z\sim q_0}\Big[\big\langle \nabla\varphi(z), V_u(z)\big\rangle\Big] - \mathbf{E}_{z\sim q_0}\Big[\big\langle \nabla\psi(z), V_u(z)\big\rangle\Big].$$

Finally, by Proposition 6.3, $\nabla_G\mathcal{L}(g^*)$ already lies in $\mathfrak{g}^H_{\text{bias}}$ (its components along $\mathfrak{n}$ and $\mathfrak{g}^\perp$ vanish identically). Thus $\nabla_G\mathcal{L}(g^*) = 0$ if and only if $\langle \nabla_G\mathcal{L}(g^*), u\rangle_{\mathfrak{g}} = 0$ for all $u \in \mathfrak{g}^H_{\text{bias}}$, which is equivalent to (2). □

B. Ad_H-invariance of $\mathfrak{n}$ and subspace decomposition details

The orthogonal decomposition $\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{g}^H_{\text{bias}} \oplus \mathfrak{g}^\perp$ in Proposition 6.3 uses the fact that the identifiability kernel $\mathfrak{n} = T_eN$ is an Ad_H-invariant subspace of $\mathfrak{g}$. We record the detailed verification here and display the explicit decomposition $\mathfrak{n} = (\mathfrak{n}\cap\mathfrak{g}^H) \oplus (\mathfrak{n}\cap(\mathfrak{g}^H)^\perp)$ across all six guiding examples.

B.1. Proof of Ad_H-invariance of $\mathfrak{n}$

Recall that $N = G \cap H$ is the intersection of the transport group G and the symmetry group H inside the total group $K = GH$. For any $n \in N$ and $h \in H$, we have $\Phi_h(n) = h\circ n\circ h^{-1} \in G$ because G is normalized by H (Definition 3.3), and simultaneously $h\circ n\circ h^{-1} \in H$ because $n \in H$ and H is a group. Hence $\Phi_h(n) \in G\cap H = N$ for all $n \in N$ and $h \in H$.

Now let $u \in \mathfrak{n}$. By definition of $\mathfrak{n}$, the one-parameter subgroup $t \mapsto \exp(tu)$ lies entirely in N for all $t \in \mathbb{R}$. Applying Φ_h, the smooth curve $t \mapsto \Phi_h(\exp(tu))$ also lies entirely in N for all $t \in \mathbb{R}$, and passes through e_G at $t = 0$. Its velocity vector at $t = 0$ therefore belongs to the tangent space $T_eN = \mathfrak{n}$:

$$\mathrm{Ad}_h \cdot u = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \Phi_h(\exp(tu)) \in \mathfrak{n}.$$

Thus $\mathrm{Ad}_h(\mathfrak{n}) = \mathfrak{n}$ for every $h \in H$. Because the inner product $\langle\cdot,\cdot\rangle_{\mathfrak{g}}$ is Ad_H-invariant, the orthogonal averaging projection $P_{\mathfrak{g}^H} = \int_H \mathrm{Ad}_h \,\mathrm{dHaar}(h)$ onto the trivial representation $\mathfrak{g}^H$ maps $\mathfrak{n}$ into $\mathfrak{n}\cap\mathfrak{g}^H$, and the complementary projection $I - P_{\mathfrak{g}^H}$ maps $\mathfrak{n}$ into $\mathfrak{n}\cap(\mathfrak{g}^H)^\perp$. Consequently, $\mathfrak{n}$ splits orthogonally as

$$\mathfrak{n} = (\mathfrak{n}\cap\mathfrak{g}^H) \;\oplus\; (\mathfrak{n}\cap(\mathfrak{g}^H)^\perp),$$

which ensures that $\mathfrak{n}$, $\mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H \cap \mathfrak{n}^\perp$, and $\mathfrak{g}^\perp = (\mathfrak{g}^H)^\perp \cap \mathfrak{n}^\perp$ are mutually orthogonal and sum to $\mathfrak{g}$.

B.2. Explicit subspace verification for the six guiding examples

- **Separable regime (Examples 1, 2, and 6)**: Here $N = \{e\}$, so $\mathfrak{n} = \{0\}$ and both $\mathfrak{n}\cap\mathfrak{g}^H = \{0\}$ and $\mathfrak{n}\cap(\mathfrak{g}^H)^\perp = \{0\}$ are trivial. Consequently, $\mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H$ and $\mathfrak{g}^\perp = (\mathfrak{g}^H)^\perp$.
- **Sphere (Example 3)**: Here $N = H = \mathrm{SO}(2)$ is abelian, so the adjoint action of $\mathrm{SO}(2)$ on its own Lie algebra $\mathfrak{n} = \mathrm{span}(e_3)$ is trivial, giving $\mathfrak{n} = \mathfrak{g}^H = \mathrm{span}(e_3)$. Thus $\mathfrak{n}\cap\mathfrak{g}^H = \mathfrak{n}$ and $\mathfrak{n}\cap(\mathfrak{g}^H)^\perp = \{0\}$: the entire identifiability kernel lies inside the trivial subspace $\mathfrak{g}^H$, leaving $\mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H \cap \mathfrak{n}^\perp = \{0\}$.
- **Covariance (Example 4)**: Here $N = H = \mathrm{O}(d)$ has Lie algebra $\mathfrak{n} = \mathfrak{so}(d)$ (skew-symmetric matrices), whereas $\mathfrak{g}^H = \mathrm{span}(I_d)$ (isotropic scalar matrices). Because the only matrix that is both symmetric (a multiple of I_d) and skew-symmetric is the zero matrix, we have $\mathfrak{n}\cap\mathfrak{g}^H = \{0\}$ and $\mathfrak{n}\cap(\mathfrak{g}^H)^\perp = \mathfrak{so}(d)$: for $d \geq 2$, the entire identifiability kernel $\mathfrak{so}(d)$ lies in the non-trivial representation $(\mathfrak{g}^H)^\perp$. Consequently, $\mathfrak{g}^H_{\text{bias}} = \mathfrak{g}^H = \mathrm{span}(I_d)$, and $\mathfrak{g}^\perp = (\mathfrak{g}^H)^\perp \cap \mathfrak{so}(d)^\perp = \mathrm{Sym}_0(d)$ is the space of traceless symmetric matrices.
- **Square (Example 5)**: Here $N = C_4$ is a finite group of order 4, so its Lie algebra is zero-dimensional ($\mathfrak{n} = \{0\}$), and the decomposition of $\mathfrak{n}$ is again trivial.

C. Detailed derivations for the guiding examples

C.1. Covariance example: Explicit scale bias and base calibration

Consider Example 4 ($G = \mathrm{GL}(d)$, $H = \mathrm{O}(d)$) with standard Gaussian base $q_0 = \mathcal{N}(0, I_d)$ and standardized target pullback $\tilde{p} = \mathcal{N}(0, \sigma^2 I_d)$ for some $\sigma > 0$. The Lie algebra is $\mathfrak{g} = \mathfrak{gl}(d)$, and by Proposition 6.3 the bias subspace is $\mathfrak{g}^H_{\mathrm{bias}} = \mathrm{span}(I_d)$.

For the basis generator $u = I_d$, the velocity field is $V_{I_d}(z) = I_d z = z$. The log-density potentials (up to additive constants) are $\psi(z) = \frac{1}{2}\|z\|_2^2$ and $\varphi(z) = \frac{1}{2\sigma^2}\|z\|_2^2$, with gradients $\nabla\psi(z) = z$ and $\nabla\varphi(z) = \sigma^{-2}z$. Substituting these into Theorem 6.5 gives the directional derivative of the pullback loss along I_d:

$$\begin{aligned}\langle \nabla_G \mathcal{L}(A^*), I_d\rangle_{\mathrm{Frob}} &= \mathbf{E}_{z\sim q_0}\left[\langle \nabla\varphi(z), z\rangle\right] - \mathbf{E}_{z\sim q_0}\left[\langle \nabla\psi(z), z\rangle\right] \\ &= \frac{1}{\sigma^2}\mathbf{E}_{z\sim q_0}\left[\|z\|_2^2\right] - \mathbf{E}_{z\sim q_0}\left[\|z\|_2^2\right] = d\left(\frac{1}{\sigma^2} - 1\right).\end{aligned}$$

Because $\|I_d\|^2_{\mathrm{Frob}} = d$ and $\nabla_G\mathcal{L}(A^*) \in \mathrm{span}(I_d)$, we obtain the exact Riemannian gradient at A^*:

$$\nabla_G \mathcal{L}(A^*) = \left(\frac{1}{\sigma^2} - 1\right) I_d\,.$$

This vanishes if and only if $\sigma^2 = 1$. When $\sigma^2 \neq 1$, recalibrating the base distribution along $\exp(\mathfrak{g}^H_{\mathrm{bias}}) = \{\lambda I_d : \lambda > 0\}$ by choosing $g_\Omega = \sigma I_d$ produces the calibrated base $q_\Omega = (g_\Omega)_\sharp q_0 = \mathcal{N}(0, \sigma^2 I_d)$ with potential $\psi_\Omega(z) = \frac{1}{2\sigma^2}\|z\|_2^2$. The moment-matching condition (2) is then satisfied identically, and $\nabla_G\mathcal{L}(A^*) = 0$.

C.2. Square example: Polar coordinate derivation of the angular gradient

Consider Example 5 ($G = \mathrm{SO}(2)$, $H = D_4$) with quartic base distribution $q_0(z) \propto \exp(-\kappa(x^4 + y^4))$ on $\mathbb{R}^2$, where $z = (x, y)^\intercal$ and $\kappa > 0$. Passing to polar coordinates $x = r\cos\phi$, $y = r\sin\phi$ and using the identity $\cos^4\phi + \sin^4\phi = (\cos^2\phi + \sin^2\phi)^2 - \frac{1}{2}\sin^2(2\phi) = \frac{3}{4} + \frac{1}{4}\cos(4\phi)$, the base potential takes the form

$$\psi(r, \phi) = \kappa(r^4\cos^4\phi + r^4\sin^4\phi) = \kappa r^4\left(\frac{3}{4} + \frac{1}{4}\cos(4\phi)\right).$$

For an angular velocity $u = \dot\theta \in \mathfrak{so}(2) \cong \mathbb{R}$, the flow $\exp(t\dot\theta)\cdot z = R_{t\dot\theta}z$ has velocity field $V_{\dot\theta}(x, y) = \dot\theta(-y, x)^\intercal$. Consequently, the directional derivative of ψ along $V_{\dot\theta}$ is

$$T_0(z) := \left\langle \nabla\psi(z), V_{\dot\theta}(z)\right\rangle = 4\kappa\dot\theta\left(-x^3y + xy^3\right) = \dot\theta\frac{\partial\psi}{\partial\phi}(r, \phi) = -\kappa\dot\theta r^4\sin(4\phi)\,.$$

Now suppose the standardized target pullback is rotated by an angle α, so that $\varphi(z) = \psi(R_{-\alpha}z) = \kappa r^4\left(\frac{3}{4} + \frac{1}{4}\cos(4(\phi - \alpha))\right)$. The directional derivative of φ along $V_{\dot\theta}$ is similarly

$$T_\alpha(z) := \left\langle \nabla\varphi(z), V_{\dot\theta}(z)\right\rangle = -\kappa\dot\theta r^4\sin\left(4(\phi - \alpha)\right).$$

By Theorem 6.5, the directional derivative of the pullback loss J at 0 is

$$\langle \nabla J(0), \dot\theta\rangle_{\mathfrak{g}} = \mathbf{E}_{z\sim q_0}\left[T_\alpha(z) - T_0(z)\right] = -\kappa\dot\theta\,\mathbf{E}_{z\sim q_0}\left[r^4\left(\sin(4(\phi - \alpha)) - \sin(4\phi)\right)\right].$$

Because the density $q_0(r, \phi) \propto \exp(-\kappa r^4(\frac{3}{4} + \frac{1}{4}\cos(4\phi)))$ is an even function of ϕ while $\sin(4\phi)$ is odd, we have $\mathbf{E}_{q_0}[r^4\sin(4\phi)] = 0$. Expanding $\sin(4(\phi - \alpha)) = \sin(4\phi)\cos(4\alpha) - \cos(4\phi)\sin(4\alpha)$ therefore simplifies the gradient to

$$\langle \nabla J(0), \dot\theta\rangle_{\mathfrak{g}} = \kappa\dot\theta\sin(4\alpha)\,\mathbf{E}_{z\sim q_0}\left[r^4\cos(4\phi)\right] = -\kappa M_4\sin(4\alpha)\,\dot\theta\,,$$

where $M_4 := -\mathbf{E}_{z\sim q_0}[r^4\cos(4\phi)]$. Note that M_4 is strictly positive for every $\kappa > 0$: for each fixed radius $r > 0$, the conditional angular density $p(\phi \mid r) \propto \exp(-\frac{1}{4}\kappa r^4\cos(4\phi))$ places strictly greater mass on angles where $\cos(4\phi) < 0$ (near the diagonals $\phi = \pm\pi/4, \pm 3\pi/4$) than on angles where $\cos(4\phi) > 0$ (near the coordinate axes), ensuring $\mathbf{E}_{q_0}[r^4\cos(4\phi)] < 0$. Thus $\nabla J(0) = 0$ if and only if $\sin(4\alpha) = 0$, i.e., $\alpha \in (\pi/4)\mathbb{Z}$.