Supplementary Materials

for

Are Variants of the CRPS Sufficient for Generative Modelling?

Jingxiao Xu

Department of Engineering
University of Cambridge

August 14, 2025

Note: This document contains supplementary materials that are not part of the submitted thesis. References to sections in this document can be found in the main thesis.

Contents

S1 Mathematical Proofs and Discussions

S1.1 Multivariate Projection Bound on Energy Distance

Proof of Corollary 3.6. We establish this bound by expressing the energy distance as an integral representation over D-dimensional projections and applying a supremum argument. From Theorem 3.4, we have the integral representation:

$$E^2(X,Y) = \frac{c_d}{2c_1}\int_{\mathbb{S}^{d-1}} E^2(X_\theta, Y_\theta)\mathrm{d}\sigma(\theta). \tag{S1.1}$$

We apply the following Blaschke-Petkantschin decomposition (see [5] Chapter 6.5 and [4] for the classical formulation): for any integrable function f on $\mathbb{S}^{d-1}$:

$$\int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\sigma(\theta) = \int_{G_{d,D}}\int_{\mathbb{S}^{D-1}(P)} f(\omega)J_{d,D}(P)\mathrm{d}\sigma_P(\omega)\mathrm{d}\gamma_D(P), \tag{S1.2}$$

where $G_{d,D}$ denotes the Grassmannian of D-dimensional subspaces of $\mathbb{R}^d$, $\mathbb{S}^{D-1}(P)$ is the unit sphere within subspace P, $\mathrm{d}\sigma_P$ is the surface measure on this sphere, $\mathrm{d}\gamma_D(P)$ is the unique invariant measure on $G_{d,D}$ normalised so that $\gamma_D(G_{d,D}) = 1$, and $J_{d,D}(P)$ is the Jacobian factor. For our specific case with the invariant probability measure $\mathrm{d}\nu(P) = J_{d,D}(P)\mathrm{d}\gamma_D(P)$, we have:

$$\int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\sigma(\theta) = \int_{G_{d,D}}\int_{\mathbb{S}^{D-1}(P)} f(\omega)\mathrm{d}\sigma_P(\omega)\mathrm{d}\nu(P). \tag{S1.3}$$

The normalisation is such that:

$$\int_{G_{d,D}} \mathrm{Vol}(\mathbb{S}^{D-1}(P))\mathrm{d}\nu(P) = \frac{A_{D-1}}{A_{d-1}}, \tag{S1.4}$$

where $\mathrm{Vol}(\mathbb{S}^{D-1}(P)) = A_{D-1}$ for all $P \in G_{d,D}$. Applying this decomposition to our energy distance integral yields:

$$E^2(X,Y) = \frac{c_d}{2c_1}\int_{G_{d,D}}\int_{\mathbb{S}^{D-1}(P)} E^2(X_\omega, Y_\omega)\mathrm{d}\sigma_P(\omega)\mathrm{d}\nu(P). \tag{S1.5}$$

For a fixed D-dimensional subspace P, let $A \in V_D(\mathbb{R}^d)$ be an orthonormal matrix whose rows span P. For any $\omega \in \mathbb{S}^{D-1}(P)$, we can write $\omega = A^T\alpha$ for some unit vector $\alpha \in \mathbb{S}^{D-1}$. The projection satisfies $X_\omega = \omega^T X = \alpha^T AX = \alpha^T X_A$, where $X_A = AX$ is the projection of X onto subspace P. Applying the same argument as in Theorem 1.1 to the D-dimensional space spanned by P, we obtain:

$$\int_{\mathbb{S}^{D-1}(P)} E^2(X_\omega, Y_\omega)\mathrm{d}\sigma_P(\omega) = \frac{2c_1A_{D-1}}{c_D}E^2(X_A, Y_A). \tag{S1.6}$$

Substituting this identity back into our expression:

$$E^2(X,Y) = \frac{c_d}{2c_1}\int_{G_{d,D}} \frac{2c_1A_{D-1}}{c_D}E^2(X_A, Y_A)\mathrm{d}\nu(P) \tag{S1.7}$$

$$= \frac{c_dA_{D-1}}{c_D}\int_{G_{d,D}} E^2(X_A, Y_A)\mathrm{d}\nu(P). \tag{S1.8}$$

To transition from the Grassmannian to the Stiefel manifold, we use the fact that each D-dimensional subspace $P \in G_{d,D}$ corresponds to an equivalence class of orthonormal matrices in $V_D(\mathbb{R}^d) = V_D(\mathbb{R}^d)$ under the action of the orthogonal group $O(D)$. Specifically, if $\pi : V_D(\mathbb{R}^d) \to G_{d,D}$ is the natural projection mapping each orthonormal matrix to the subspace it spans, then for any invariant measure μ on $V_D(\mathbb{R}^d)$ normalised so that $\mu(V_D(\mathbb{R}^d)) = 1$, we have:

$$\int_{G_{d,D}} g(P)\mathrm{d}\nu(P) = \int_{V_D(\mathbb{R}^d)} g(\pi(A))\mathrm{d}\mu(A), \tag{S1.9}$$

for any integrable function g on $G_{d,D}$. Since $E^2(X_A, Y_A)$ depends only on the subspace spanned by the rows of A, we obtain:

$$E^2(X,Y) = \frac{c_d A_{D-1}}{c_D} \int_{V_D(\mathbb{R}^d)} E^2(X_A, Y_A)\mathrm{d}\mu(A). \tag{S1.10}$$

However, the correct normalisation factor requires accounting for the volume ratio between $G_{d,D}$ and $V_D(\mathbb{R}^d)$. This gives:

$$E^2(X,Y) = \frac{c_d A_{d-1}}{c_D A_{D-1}} \int_{V_D(\mathbb{R}^d)} E^2(X_A, Y_A)\mathrm{d}\mu(A), \tag{S1.11}$$

where we have incorporated the factor A_{d-1}/A_{D-1} from the Blaschke-Petkantschin decomposition. Since μ is a probability measure on $V_D(\mathbb{R}^d)$, the integral is bounded by the supremum. Substituting the explicit expressions for the surface areas in terms of gamma functions yields the desired result. □

Weighted Multi-Dimensional Projection Bound

Proof of Corollary **??**. From Corollary 3.6, for each dimension $D_i \in \mathscr{D}$ where $1 \le i \le k$, we have the individual bound:

$$E(X,Y) \le C_{d,D_i} \sup_{A \in V_{D_i}(\mathbb{R}^d)} E(X_A, Y_A), \tag{S1.12}$$

where $V_{D_i}(\mathbb{R}^d)$ denotes the Stiefel manifold of $D_i \times d$ matrices with orthonormal rows, and C_{d,D_i} is the constant from Corollary 3.6. For all $i \in \{1,2,\ldots,k\}$, we can rearrange to obtain a lower bound on each supremum:

$$\sup_{A \in V_{D_i}(\mathbb{R}^d)} E(X_A, Y_A) \ge \frac{E(X,Y)}{C_{d,D_i}} \quad \text{for all } i \in \{1,2,\ldots,k\}. \tag{S1.13}$$

Now consider the weighted sum $\sum_{i=1}^k w_i \sup_{A \in V_{D_i}(\mathbb{R}^d)} E(X_A, Y_A)$. Since each weight w_i is positive, we can apply the lower bounds from (S1.13) to obtain:

$$\sum_{i=1}^k w_i \sup_{A \in V_{D_i}(\mathbb{R}^d)} E(X_A, Y_A) \ge \sum_{i=1}^k w_i \frac{E(X,Y)}{C_{d,D_i}} \tag{S1.14}$$

$$= E(X,Y) \sum_{i=1}^k \frac{w_i}{C_{d,D_i}}, \tag{S1.15}$$

All constants are positive, therefore dividing both sides of inequality (S1.15) by this quantity to obtain:

$$E(X,Y) \leq \left(\sum_{i=1}^{k} \frac{w_i}{C_{d,D_i}}\right)^{-1} \sum_{i=1}^{k} w_i \sup_{A \in V_{D_i}(\mathbb{R}^d)} E(X_A, Y_A). \tag{S1.16}$$

Defining the constant $C_{\mathscr{D},w} \equiv \left(\sum_{i=1}^{k} \frac{w_i}{C_{d,D_i}}\right)^{-1}$ and rearranging the previous inequality, we obtain the desired result. □

S1.2 Optimal Projection for Large Mean Separation

We derive the optimal projection direction in the regime where the mean separation is large.

Lemma S1.1 (Asymptotic expansion of $\mathbb{E}|\mathscr{N}(\delta, s^2)|$ for large $\delta/\sqrt{s^2}$)**.** *Let $a = \delta/\sqrt{s^2}$. As $a \to \infty$ the following expansion holds*

$$\mathbb{E}|\mathscr{N}(\delta, s^2)| = \delta + \frac{2s^3}{\delta^2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\delta^2}{2s^2}\right) + o\left(\frac{s^3}{\delta^2} \exp\left(-\frac{\delta^2}{2s^2}\right)\right) \tag{S1.17}$$

Proof. Using the exact representation

$$\mathbb{E}|\mathscr{N}(\delta, s^2)| = \delta\left(1 - 2\Phi\left(-\frac{\delta}{\sqrt{s^2}}\right)\right) + \sqrt{\frac{2}{\pi}}\sqrt{s^2} \exp\left(-\frac{\delta^2}{2s^2}\right) \tag{S1.18}$$

and the Gaussian tail expansion

$$\Phi(-a) = \frac{\varphi(a)}{a}\left(1 - \frac{1}{a^2} + O\left(\frac{1}{a^4}\right)\right) \tag{S1.19}$$

with $\varphi(a) = \frac{1}{\sqrt{2\pi}} \exp(-a^2/2)$ we substitute $a = \delta/\sqrt{s^2}$ to obtain

$$\delta\left(1 - 2\frac{\varphi(a)}{a}\left(1 - \frac{1}{a^2} + \cdots\right)\right) + \sqrt{\frac{2}{\pi}}\sqrt{s^2} \exp\left(-\frac{\delta^2}{2s^2}\right) \tag{S1.20}$$

Collecting the exponentially small terms reveals cancellation of the leading $\delta \cdot 2\varphi(a)/a$ against part of the second term, and the first surviving correction is of order $\frac{s^3}{\delta^2} \exp(-\delta^2/(2s^2))$ yielding the claimed expansion □

We now substitute this expansion into the definition of the projected energy distance writing

$$D^2_{\text{Gauss}}(\theta) = 2\mathbb{E}|\mathscr{N}(|\theta^\top \Delta|, s^2)| - \mathbb{E}|\mathscr{N}(0, 2\theta^\top \Sigma_X \theta)| - \mathbb{E}|\mathscr{N}(0, 2\theta^\top \Sigma_Y \theta)| \tag{S1.21}$$

and using $\mathbb{E}|\mathscr{N}(0, \tau^2)| = \tau\sqrt{2/\pi}$ to obtain the leading-order approximation in the regime $|\theta^\top \Delta| \gg \sqrt{s^2}$

$$D^2_{\text{Gauss}}(\theta) = 2|\theta^\top \Delta| - \frac{2}{\sqrt{\pi}}\left(\sqrt{\theta^\top \Sigma_X \theta} + \sqrt{\theta^\top \Sigma_Y \theta}\right) + R(\theta) \tag{S1.22}$$

where the remainder $R(\theta)$ is exponentially small in $(\theta^\top \Delta)^2/s^2$ and satisfies

$$R(\theta) = \frac{4s^3}{|\theta^\top \Delta|^2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\theta^\top \Delta)^2}{2s^2}\right) + o\left(\frac{s^3}{|\theta^\top \Delta|^2} \exp\left(-\frac{(\theta^\top \Delta)^2}{2s^2}\right)\right) \tag{S1.23}$$

To find the optimal direction, we maximise the functional $J(\theta)$ from Proposition 3.7. The proof of the proposition follows.

Proof of Proposition 3.7. To find the optimal direction θ, we maximise $J(\theta)$ subject to the constraint $\|\theta\| = 1$. We assume $\theta^\top \Delta \geq 0$ to remove the absolute value. By considering the Lagrangian, we have stationary condition

$$\Delta - \frac{1}{\sqrt{\pi}} \left(\frac{\Sigma_X \theta}{\sqrt{\theta^\top \Sigma_X \theta}} + \frac{\Sigma_Y \theta}{\sqrt{\theta^\top \Sigma_Y \theta}} \right) = \lambda \theta \tag{S1.24}$$

In the large mean separation regime, the optimal direction θ^* is a small perturbation from the mean difference direction. We decompose θ as $\theta = \theta^{(0)} + u$, where $\theta^{(0)} = \frac{\Delta}{\|\Delta\|}$ and $u \perp \Delta$. To isolate the correction u, we project the stationary condition onto the subspace orthogonal to Δ using the projector $P_\perp = I - \theta^{(0)} \theta^{(0)\top}$.

$$-\frac{1}{\sqrt{\pi}} P_\perp \left(\frac{\Sigma_X \theta}{\sqrt{\theta^\top \Sigma_X \theta}} + \frac{\Sigma_Y \theta}{\sqrt{\theta^\top \Sigma_Y \theta}} \right) = \lambda u \tag{S1.25}$$

To make progress, we approximate by linearising the equation. In the large mean separation limit, we approximate the denominators using $\theta \approx \theta^{(0)}$, which gives $\sqrt{\theta^\top \Sigma_X \theta} \approx w_X$ and $\sqrt{\theta^\top \Sigma_Y \theta} \approx w_Y$. Substituting these and $\theta = \theta^{(0)} + u$ into the projected equation gives

$$-\frac{1}{\sqrt{\pi}} P_\perp \left(\frac{\Sigma_X (\theta^{(0)} + u)}{w_X} + \frac{\Sigma_Y (\theta^{(0)} + u)}{w_Y} \right) \approx \lambda u$$

Rearranging to solve for u:

$$-\frac{1}{\sqrt{\pi}} P_\perp \left(\frac{\Sigma_X}{w_X} + \frac{\Sigma_Y}{w_Y} \right) \theta^{(0)} \approx \left(\lambda I + \frac{1}{\sqrt{\pi}} P_\perp \left(\frac{\Sigma_X}{w_X} + \frac{\Sigma_Y}{w_Y} \right) \right) u$$

In the large mean separation regime, the Lagrange multiplier $\lambda \approx \|\Delta\|$ (since $\theta^{(0)}$ is the direction of maximum mean separation) dominates the term in the parenthesis on the right-hand side. We can therefore approximate the equation as

$$u \approx -\frac{1}{\lambda \sqrt{\pi}} P_\perp \left(\frac{\Sigma_X}{w_X} + \frac{\Sigma_Y}{w_Y} \right) \theta^{(0)}$$

Substituting $\lambda \approx \|\Delta\|$ and $\theta^{(0)} = \Delta / \|\Delta\|$ gives an approximate first-order correction. The sign of the correction is chosen to reduce the variance penalty, leading to the stated expression for θ^*. □

S1.3 Gaussian Energy Distance Taylor Expansion

We derive the small mean-difference expansion for the energy distance between Gaussian distributions. Write $\Delta \equiv \mu_X - \mu_Y$, and define the projected variances

$$v_X(\theta) = \theta^\top \Sigma_X \theta, \qquad v_Y(\theta) = \theta^\top \Sigma_Y \theta, \qquad s^2(\theta) = v_X(\theta) + v_Y(\theta) = \theta^\top (\Sigma_X + \Sigma_Y) \theta \tag{S1.26}$$

For the one-dimensional projection along $\theta \in \mathbb{S}^{D-1}$ the energy distance between the Gaussian marginals is

$$D^2_{\text{Gauss}}(\theta) \equiv E^2(\theta^T X, \theta^T Y) = 2\,\mathbb{E}|\mathcal{N}(|\theta^\top \Delta|, s^2)| - \mathbb{E}|\mathcal{N}(0, 2v_X)| - \mathbb{E}|\mathcal{N}(0, 2v_Y)| \tag{S1.27}$$

where $\mathbb{E}|\mathcal{N}(\delta, s^2)|$ is the first absolute moment of a normal distribution, given by

$$\mathbb{E}|\mathcal{N}(\delta, \sigma^2)| = \delta\,(2\Phi(\delta/\sigma) - 1) + \sigma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\delta^2}{2\sigma^2}\right) \tag{S1.28}$$

To analyse the case where $\|\Delta\|$ is small, we expand $\mathbb{E}|\mathcal{N}(|\theta^\top\Delta|, s^2(\theta))|$ in a Taylor series around $|\theta^\top\Delta| = 0$. The first two derivatives of $\mathbb{E}|\mathcal{N}(\delta, s^2)|$ with respect to δ are:

$$\frac{\partial}{\partial\delta}\mathbb{E}|\mathcal{N}(\delta, s^2)| = 2\Phi\left(\frac{\delta}{\sigma}\right) - 1 \tag{S1.29}$$

$$\frac{\partial^2}{\partial\delta^2}\mathbb{E}|\mathcal{N}(\delta, s^2)| = \frac{2}{\sigma}\phi\left(\frac{\delta}{\sigma}\right) = \frac{2}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\delta^2}{2\sigma^2}\right) \tag{S1.30}$$

Evaluating at $\delta = 0$, we find that $\partial_\delta\mathbb{E}|\mathcal{N}(0, s^2)| = 0$ and

$$\frac{\partial^2}{\partial\delta^2}\mathbb{E}|\mathcal{N}(\delta, s^2)|\bigg|_{\delta=0} = \frac{\sqrt{2}}{\sqrt{\pi s^2}} \tag{S1.31}$$

The Taylor expansion is therefore $\mathbb{E}|\mathcal{N}(\delta, s^2)| = \mathbb{E}|\mathcal{N}(0, s^2)| + \frac{1}{2}\frac{\sqrt{2}}{\sqrt{\pi s^2}}\delta^2 + o(\delta^2)$. Substituting this into $D^2_{\text{Gauss}}(\theta)$ gives

$$D^2_{\text{Gauss}}(\theta) = \underbrace{2\mathbb{E}|\mathcal{N}(0, s^2)| - \mathbb{E}|\mathcal{N}(0, 2v_X)| - \mathbb{E}|\mathcal{N}(0, 2v_Y)|}_{\text{call this } B(\theta)} + \frac{\sqrt{2}}{\sqrt{\pi s^2}}(\theta^\top\Delta)^2 + o\Big((\theta^\top\Delta)^2\Big) \tag{S1.32}$$

The leading term $B(\theta)$ depends only on the covariance structure. Hence, when Δ is negligible, the optimal slicing direction θ^* is (to first approximation) the maximiser of $B(\theta)$:

$$\theta^* \approx \arg\max_{\|\theta\|=1}\left[2\sqrt{\frac{2\,\theta^\top(\Sigma_X + \Sigma_Y)\theta}{\pi}} - \sqrt{\frac{4\,\theta^\top\Sigma_X\theta}{\pi}} - \sqrt{\frac{4\,\theta^\top\Sigma_Y\theta}{\pi}}\right] \tag{S1.33}$$

Observe that the bracketed term is proportional to the difference between the quadratic and arithmetic means of the projected variances.

QM-AM Analysis for Covariance Perturbations. We derive the final optimisation form by analysing the quadratic-arithmetic mean structure. Let $\sigma_X(\theta) = \sqrt{v_X(\theta)}$ and $\sigma_Y(\theta) = \sqrt{v_Y(\theta)}$. The objective $B(\theta)$ from the previous appendix can be written as:

$$B(\theta) \propto \sqrt{\frac{\sigma_X(\theta)^2 + \sigma_Y(\theta)^2}{2}} - \left(\frac{\sigma_X(\theta) + \sigma_Y(\theta)}{2}\right) \tag{S1.34}$$

which is the difference between the quadratic and arithmetic means of $\sigma_X(\theta)$ and $\sigma_Y(\theta)$. We assume that the norm of $\Delta\Sigma$ is small compared to that of Σ. The projected variances become:

$$v_X(\theta) = v(\theta) - \delta v(\theta), \qquad v_Y(\theta) = v(\theta) + \delta v(\theta) \tag{S1.35}$$

where $v(\theta) = \theta^\top\Sigma\theta$ and $\delta v(\theta) = \theta^\top\Delta\Sigma\theta$.

For small δv, the quadratic mean term simplifies to:

$$\sqrt{\frac{\sigma_X(\theta)^2 + \sigma_Y(\theta)^2}{2}} = \sqrt{\frac{(v(\theta) - \delta v(\theta)) + (v(\theta) + \delta v(\theta))}{2}} = \sqrt{v(\theta)} \tag{S1.36}$$

For the arithmetic mean term, we expand σ_X and σ_Y to second order in the small parameter $\delta v/v$:

$$\sigma_X(\theta) = \sqrt{v}\left(1-\frac{\delta v}{v}\right)^{1/2} \approx \sqrt{v}\left(1-\frac{1}{2}\frac{\delta v}{v}-\frac{1}{8}\left(\frac{\delta v}{v}\right)^2\right) \tag{S1.37}$$

$$\sigma_Y(\theta) = \sqrt{v}\left(1+\frac{\delta v}{v}\right)^{1/2} \approx \sqrt{v}\left(1+\frac{1}{2}\frac{\delta v}{v}-\frac{1}{8}\left(\frac{\delta v}{v}\right)^2\right) \tag{S1.38}$$

Substituting these expansions back into the expression for $B(\theta)$ yields:

$$B(\theta) \propto \sqrt{v}-\sqrt{v}\left(1-\frac{1}{8}\left(\frac{\delta v}{v}\right)^2\right) = \frac{\sqrt{v}}{8}\left(\frac{\delta v}{v}\right)^2 = \frac{1}{8}\frac{(\delta v)^2}{v^{3/2}} \tag{S1.39}$$

Thus, to leading order, maximising $B(\theta)$ is equivalent to solving the optimisation problem in the proposition 3.9.

S1.4 Spectral Analysis for Isotropic Kernels

We can simplify the problem in the case of isotropic kernels. Mathematically, a covariance matrix Σ arises from an isotropic kernel if there exists a function kernel function $k : \mathbb{R} \to \mathbb{R}$ such that $\Sigma_{ij} = k(\|\mathbf{x}_i - \mathbf{x}_j\|)$ for spatial locations $\mathbf{x}_i, \mathbf{x}_j$. Such kernels are translation-invariant, meaning $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$, and isotropic, meaning they depend only on the Euclidean distance $\|\mathbf{x} - \mathbf{y}\|$ rather than the direction. This assumption is particularly relevant for spatial data where physical processes exhibit homogeneous correlation structures across space, such as in atmospheric or oceanic fields where meteorological variables often display similar correlation patterns regardless of location or orientation. This structural constraint ensures that the resulting covariance operators have spectral profiles that are functions of frequency magnitude alone, allowing us to find the optimal projection direction in the frequency domain.

To make the optimisation tractable and expose its spectral structure, we restrict attention to the case where both covariance operators Σ_X and Σ_Y arise from *symmetric positive definite, translation-invariant, isotropic kernels* on the underlying spatial domain (here a 2D $N \times N$ periodic grid).

Proposition S1.1 (Simplification for Isotropic Kernels)**.** *Let Σ_X and Σ_Y be covariance operators arising from symmetric positive definite, translation-invariant, isotropic kernels on a 2D $N \times N$ periodic grid. Then both operators are jointly diagonalised by the 2D discrete Fourier basis, and the optimisation problem* (3.36) *reduces to:*

$$\mathbf{k}^* = \arg\max_{\mathbf{k}} \frac{(\lambda_Y(\mathbf{k}) - \lambda_X(\mathbf{k}))^2}{(\lambda_X(\mathbf{k}) + \lambda_Y(\mathbf{k}))^{3/2}} \tag{S1.40}$$

where $\lambda_X(\mathbf{k})$ and $\lambda_Y(\mathbf{k})$ are the respective spectral profiles.

Proof. Under the isotropy and translation-invariance assumptions, Σ_X and Σ_Y commute and are jointly diagonalisable by the 2D discrete Fourier basis as eigenvectors. Denoting a Fourier mode by

$$v_{\mathbf{n}}^{(\mathbf{k})} = \exp\left(2\pi \mathrm{i}\,\frac{\mathbf{k}\cdot\mathbf{n}}{N}\right) \tag{S1.41}$$

with $\mathbf{k} \in \{0,\dots,N-1\}^2$ and $\mathbf{n}$ the spatial index. By definition of eigenvectors, have

$$\Sigma_X v^{(\mathbf{k})} = \lambda_X(\mathbf{k})\, v^{(\mathbf{k})}, \qquad \Sigma_Y v^{(\mathbf{k})} = \lambda_Y(\mathbf{k})\, v^{(\mathbf{k})} \tag{S1.42}$$

where $\lambda_X, \lambda_Y > 0$ are the respective spectral profiles. We assume these are monotonically decreasing in $\|\mathbf{k}\|$ with regular decay. Consequently, $\Delta\Sigma$ and Σ are likewise diagonal in the Fourier basis with eigenvalues:

$$\lambda_\Sigma(\mathbf{k}) = \tfrac{1}{2}\big(\lambda_X(\mathbf{k}) + \lambda_Y(\mathbf{k})\big), \qquad \lambda_{\Delta\Sigma}(\mathbf{k}) = \tfrac{1}{2}\big(\lambda_Y(\mathbf{k}) - \lambda_X(\mathbf{k})\big) \tag{S1.43}$$

Monotone decay in spectral values is common amongst kernels. For example, this condition is satisfied by the following common kernels:

- **Gaussian kernel:** $K(\mathbf{r}) = \sigma^2 \exp(-\|\mathbf{r}\|^2/(2\ell^2))$ with Fourier spectrum:

$$\lambda(\mathbf{k}) \propto \sigma^2 \exp(-2\pi^2\ell^2\|\mathbf{k}\|^2) \tag{S1.44}$$

- **Matérn kernel:** With smoothness parameter ν and lengthscale ℓ:

$$\lambda(\mathbf{k}) \propto \left(\tfrac{2\nu}{\ell^2} + 4\pi^2\|\mathbf{k}\|^2\right)^{-(\nu+d/2)} \tag{S1.45}$$

Under these spectral assumptions, the optimisation reduces to a discrete search over Fourier modes. Any optimal θ^* can be taken as one of the basis modes $v^{(\mathbf{k})}$. For a given $\mathbf{k}$:

$$\frac{(\theta^\top \Delta\Sigma\, \theta)^2}{(\theta^\top \Sigma\, \theta)^{3/2}} \quad \longrightarrow \quad \frac{\lambda_{\Delta\Sigma}(\mathbf{k})^2}{\lambda_\Sigma(\mathbf{k})^{3/2}} = \frac{\big(\tfrac{1}{2}(\lambda_Y(\mathbf{k}) - \lambda_X(\mathbf{k}))\big)^2}{\big(\tfrac{1}{2}(\lambda_X(\mathbf{k}) + \lambda_Y(\mathbf{k}))\big)^{3/2}} \tag{S1.46}$$

The optimal frequency trades off the squared spectral discrepancy against the average spectrum raised to the $3/2$ power. □

In the case of Gaussian kernels (also know as Radial Basis Function kernels, squared exponential kernels), we can calculate the optimal projection direction analytically.

Corollary S1.1 (Optimal projection for closely matched Gaussian kernels)**.** *Let Σ_X and Σ_Y be Gaussian covariance operators on the periodic 2D $N \times N$ grid with equal marginal variance σ^2 and lengthscales ℓ and $\ell + \delta\ell$, respectively. Write $\varepsilon = \delta\ell/\ell$ with $|\varepsilon| \ll 1$. Then, to leading order in ε, the Fourier mode that maximises the score*

$$S(\mathbf{k}) = \frac{\big(\lambda_Y(\mathbf{k}) - \lambda_X(\mathbf{k})\big)^2}{\big(\lambda_X(\mathbf{k}) + \lambda_Y(\mathbf{k})\big)^{3/2}} \tag{S1.47}$$

with $\lambda_X(\mathbf{k}) = \sigma^2 \exp(-2\pi^2\ell^2\|\mathbf{k}\|^2)$ and $\lambda_Y(\mathbf{k}) = \sigma^2 \exp(-2\pi^2(\ell+\delta\ell)^2\|\mathbf{k}\|^2)$, has frequency magnitude

$$\|\mathbf{k}^*\| = \frac{\sqrt{2}}{\pi\ell} \tag{S1.48}$$

Proof. Let Σ_X and Σ_Y be Gaussian covariance operators on the periodic 2D $N \times N$ grid with equal variance σ^2 and lengthscales l and $l + \delta l$, respectively. Define $\varepsilon = \frac{\delta l}{l}$ with $|\varepsilon| \ll 1$. Recall the modewise score $S(k)$ which, by isotropy, depends only on $r = ||\mathbf{k}||$:

$$S(r) = \frac{(\lambda_Y(r) - \lambda_X(r))^2}{(\lambda_X(r) + \lambda_Y(r))^{3/2}} \tag{S1.49}$$

where $\lambda_X(r) = \sigma^2 \exp(-2\pi^2 l^2 r^2)$ and $\lambda_Y(r) = \sigma^2 \exp(-2\pi^2 (l+\delta l)^2 r^2)$. To find the maximiser, we perform a leading-order Taylor expansion for small ε. The exponent in $\lambda_Y(r)$ is

$$-2\pi^2 l^2 (1+\varepsilon)^2 r^2 \approx -2\pi^2 l^2 (1+2\varepsilon) r^2 \tag{S1.50}$$

Giving $\lambda_Y(r) \approx \lambda_X(r) \exp(-4\pi^2 l^2 \varepsilon r^2)$. Using the approximation $e^x \approx 1+x$ for small x, the squared difference in the numerator becomes

$$(\lambda_Y(r) - \lambda_X(r))^2 \approx (\lambda_X(r)(1-4\pi^2 l^2 \varepsilon r^2) - \lambda_X(r))^2 = (-4\pi^2 l^2 \varepsilon r^2 \lambda_X(r))^2 \tag{S1.51}$$

Taking only leading order, $\lambda_Y(r) \approx \lambda_X(r)$, so the denominator is approximately $(2\lambda_X(r))^{3/2}$. The score is therefore approximately proportional to

$$S(r) \propto \frac{(-4\pi^2 l^2 \varepsilon r^2 \lambda_X(r))^2}{(2\lambda_X(r))^{3/2}} \propto r^4 \lambda_X(r)^{1/2} \tag{S1.52}$$

Ignoring constant factors, we seek to maximise $\tilde{S}(r) = r^4 \exp(-\pi^2 l^2 r^2)$. To do this, we consider its log-derivative

$$\frac{d}{dr} \log \tilde{S}(r) = \frac{d}{dr}(4 \log r - \pi^2 l^2 r^2) = \frac{4}{r} - 2\pi^2 l^2 r \tag{S1.53}$$

Setting the derivative to zero yields $r^2 = \frac{4}{2\pi^2 l^2} = \frac{2}{\pi^2 l^2}$. Thus, the maximiser is achieved at any Fourier mode with magnitude:

$$\|\mathbf{k}^*\| = r^* = \frac{\sqrt{2}}{\pi l} \tag{S1.54}$$

□

Remark S1.1. *The common variance σ^2 cancels in the corollary's score, so $\mathbf{k}^*$ depends only on $\hat{\ell}$. Hence in the regime where means are well fitted, this motivates the use of projection directions that align with the fourier modes for the spatial data. In the case of multi-dimensional projections, we may take different sets of frequency modes for each dimension, as long as the magnitude satisfies the optimality condition given in corollary S1.1.*

Practical Implementation (optimal 1D projections on an $N \times N$ grid). The result for optimal Gaussian kernels (corollary S1.1) enables us to incorporate our prior knowledge of the lengthscale of interactions within the data, ℓ, directly into our projection design. Under this regime, we can construct projections optimised for practical training purposes. Let each sample $X \in \mathbb{R}^{N \times N}$ (so $d = N^2$ after vectorisation) live on a regular grid $x_p = p\Delta$, $y_q = q\Delta$ for $p, q \in \{0, \ldots, N-1\}$ with spacing $\Delta > 0$ and total side-length $L = N\Delta$. Given an estimate $\hat{\ell}$ of the common Gaussian-kernel lengthscale and a desired number of projections $M \in \mathbb{N}$, define the target frequency radius

$$k_\star \equiv \|\mathbf{k}^*\| = \frac{\sqrt{2}}{\pi \hat{\ell}} \tag{S1.55}$$

Choose M angles $\alpha_m \in [0, \pi)$, equally spaced, e.g. $\alpha_m = \frac{m\pi}{M}$ for $m = 0, \ldots, M-1$. For each projection angle α_m:

1. **Find Discrete Frequencies.** Form the ideal continuous frequency vector $\mathbf{k}^{(m)} = k_\star(\cos \alpha_m, \sin \alpha_m)$ and map it to the nearest discrete DFT bin by

$$(m_x^{(m)}, m_y^{(m)}) = \text{round}\big(L\mathbf{k}^{(m)}\big) \qquad (f_x^{(m)}, f_y^{(m)}) = \Big(\frac{m_x^{(m)}}{L}, \frac{m_y^{(m)}}{L}\Big) \tag{S1.56}$$

   where round is applied componentwise to the nearest integer. This effectively finds the closest integer frequencies with approximately the desired magnitude to project onto.

2. **Obtain Projection Weights.** Define the (unnormalised) cosine weights on the grid by

$$\tilde{\theta}_{p,q}^{(m)} \;=\; \cos\Big(2\pi\big(f_x^{(m)}x_p + f_y^{(m)}y_q\big)\Big), \qquad p,q = 0,\dots,N-1 \tag{S1.57}$$

and normalise to unit Euclidean norm:

$$\theta_{p,q}^{(m)} \;=\; \frac{\tilde{\theta}_{p,q}^{(m)}}{\Big(\sum_{p=0}^{N-1}\sum_{q=0}^{N-1}\big(\tilde{\theta}_{p,q}^{(m)}\big)^2\Big)^{1/2}} \tag{S1.58}$$

Let $\theta^{(m)} \in \mathbb{R}^{N^2}$ denote the vectorisation of $(\theta_{p,q}^{(m)})_{p,q}$, and thus the m-th projection vector.

For training applications with a fixed number of projections (such as $M = d$ in the CRPS loss), the projection vectors $\theta^{(m)}$ need only be calculated once and then reused across all training steps. Similarly, the lengthscale estimate $\hat{\ell}$ should be determined from the observational data y_i prior to training.

S1.5 Projected Gradient Descent for Optimal Slicing

We provide the proof of the Lipschitz smoothness bound used to justify projected gradient descent for selecting optimal slicing directions in the non-isotropic kernel setting.

Proof of Proposition 3.11. Write $a(\theta) = \theta^\top \Delta\Sigma\,\theta$ and $b(\theta) = \theta^\top \Sigma\,\theta$. On $\mathbb{S}^{d-1}$ we have the identities and bounds

$$\nabla a(\theta) = 2\,\Delta\Sigma\,\theta, \quad \nabla b(\theta) = 2\,\Sigma\,\theta, \quad \nabla^2 a(\theta) = 2\,\Delta\Sigma, \quad \nabla^2 b(\theta) = 2\,\Sigma \tag{S1.59}$$

$$|a(\theta)| \le \delta, \quad \mu \le b(\theta) \le M, \quad \|\nabla a(\theta)\| \le 2\delta, \quad \|\nabla b(\theta)\| \le 2M \tag{S1.60}$$

Using $f = a^2 b^{-3/2}$ and the product/chain rules,

$$\nabla f(\theta) = 2a\,b^{-3/2}\nabla a - \frac{3}{2}a^2\,b^{-5/2}\nabla b \tag{S1.61}$$

Differentiate once more and decompose each term as $D(gv) = (\nabla g)\,v^\top + g\,Dv$. For the first term let $g_1 = 4a\,b^{-3/2}$ and $v_1 = \Delta\Sigma\,\theta$, so that

$$\|\nabla g_1\| \le 8\,\delta\,\mu^{-3/2} + 12\,\delta\,M\,\mu^{-5/2}, \qquad |g_1| \le 4\,\delta\,\mu^{-3/2}, \qquad \|v_1\| \le \delta,\ \|Dv_1\| \le \delta \tag{S1.62}$$

Hence

$$\|D(g_1 v_1)\| \le \|\nabla g_1\|\,\|v_1\| + |g_1|\,\|Dv_1\| \le 12\,\delta^2\,\mu^{-3/2} + 12\,\delta^2\,M\,\mu^{-5/2} \tag{S1.63}$$

For the second term let $g_2 = -3a^2 b^{-5/2}$ and $v_2 = \Sigma\,\theta$, so that

$$\|\nabla g_2\| \le 12\,\delta^2\,\mu^{-5/2} + 15\,\delta^2\,M\,\mu^{-7/2}, \qquad |g_2| \le 3\,\delta^2\,\mu^{-5/2}, \qquad \|v_2\| \le M,\ \|Dv_2\| \le M \tag{S1.64}$$

Thus

$$\|D(g_2 v_2)\| \le \|\nabla g_2\|\,\|v_2\| + |g_2|\,\|Dv_2\| \le 15\,\delta^2\,M\,\mu^{-5/2} + 15\,\delta^2\,M^2\,\mu^{-7/2} \tag{S1.65}$$

Adding the two bounds gives

$$\|\nabla^2 f(\theta)\| \le 12\,\delta^2\,\mu^{-3/2} + 27\,\delta^2\,M\,\mu^{-5/2} + 15\,\delta^2\,M^2\,\mu^{-7/2} \quad \text{for all } \|\theta\|_2 = 1 \tag{S1.66}$$

which implies that ∇f is L-Lipschitz on $\mathbb{S}^{d-1}$ with the stated constant L.

□

Proof of proposition 3.12. Define the normalisation projection function as $R_\theta(x) = \frac{\theta+x}{\|\theta+x\|}$. Writing the update as a retraction step gives

$$\theta_{k+1} = R_{\theta_k}\big(\alpha_k \operatorname{grad} f(\theta_k)\big) \tag{S1.67}$$

with the exact scalar

$$\alpha_k = \frac{\eta}{1+\eta\,\langle \theta_k, \nabla f(\theta_k)\rangle} \tag{S1.68}$$

Note from the proof of proposition 3.11 in equation (S1.61) that $\|\nabla f(\theta_k)\| \le G$. Where

$$\|\nabla f(\theta)\| \le G \equiv \delta^2\big(4\,\mu^{-3/2} + 3M\,\mu^{-5/2}\big) \quad \text{for all } \theta \in \mathbb{S}^{d-1} \tag{S1.69}$$

Using this we obtain

$$1 - \eta G \le 1 + \eta\,\langle \theta_k, \nabla f(\theta_k)\rangle \le 1 + \eta G \tag{S1.70}$$

and under $\eta \le 1/(2G)$ this implies

$$\frac{2}{3}\,\eta \le \alpha_k \le 2\eta \tag{S1.71}$$

By the geodesic L-Lipschitzness of the gradient, the Riemannian descent lemma (see A 4.2 and proposition 4.7 of Boumal [2]) for retraction steps yields, for any $\alpha_k \le 1/L$,

$$f\big(R_{\theta_k}(\alpha_k \operatorname{grad} f(\theta_k))\big) \ge f(\theta_k) + \frac{\alpha_k}{2}\,\|\operatorname{grad} f(\theta_k)\|^2 \tag{S1.72}$$

The stepsize condition $\eta \le 1/(2L)$ ensures $\alpha_k \le 2\eta \le 1/L$. Combining with $\alpha_k \ge \frac{2}{3}\eta$ gives

$$f(\theta_{k+1}) \ge f(\theta_k) + \frac{\eta}{3}\,\|\operatorname{grad} f(\theta_k)\|^2 \tag{S1.73}$$

Summing over using telescoping sums $k = 0, \ldots, K-1$ and using $f(\theta_K) \le f_{\max}$ gives

$$\sum_{k=0}^{K-1} \|\operatorname{grad} f(\theta_k)\|^2 \le \frac{3\,(f_{\max} - f(\theta_0))}{\eta} \tag{S1.74}$$

Taking the minimum over all the steps and square rooting gives the result. □

S1.6 Sliced and Energy Distance Bounds for Independent Partitions

Proof of Lemma 3.16. Let $X_\theta = \theta^T X$ and $Y_\theta = \theta^T Y$. By the triangle inequality for the energy distance, we introduce an intermediate random variable $Z_\theta = Z_A + Z_B$, where $Z_A \sim \theta_A^T X_A$ and $Z_B \sim \theta_B^T X_B$ are independent:

$$E(X_\theta, Y_\theta) \le E(X_\theta, Z_\theta) + E(Z_\theta, Y_\theta) \tag{S1.75}$$

The first term, $E(X_\theta, Z_\theta)$, is the projected dependence $\mathrm{Cov}_E(\theta; X_A, X_B)$ by definition. For the second term, $E(Z_\theta, Y_\theta)$, we add and subtract the term $\phi_{Y_{\theta_A}}(t)\phi_{Y_{\theta_B}}(t)$, which represents the characteristic function of a vector with independent components distributed as $\theta_A^T Y_A$ and $\theta_B^T Y_B$.

Applying the Minkowski integral inequality yields (we drop some notation for brevity):

$$E(Z_\theta, Y_\theta) = \left(2c_1 \int_0^\infty \frac{|\phi_{Z_\theta}(t) - \phi_{Y_\theta}(t)|^2}{t^2} \mathrm{d}t\right)^{1/2} \tag{S1.76}$$

$$= \left(2c_1 \int_0^\infty \frac{|\phi_{X_{\theta_A}}(t)\phi_{X_{\theta_B}}(t) - \phi_{Y_\theta}(t)|^2}{t^2} \mathrm{d}t\right)^{1/2} \tag{S1.77}$$

$$= \left(2c_1 \int_0^\infty \frac{|\phi_{X_{\theta_A}}\phi_{X_{\theta_B}} - \phi_{Y_{\theta_A}}\phi_{Y_{\theta_B}} + \phi_{Y_{\theta_A}}\phi_{Y_{\theta_B}} - \phi_{Y_\theta}|^2}{t^2} \mathrm{d}t\right)^{1/2} \tag{S1.78}$$

$$\leq \left(2c_1 \int_0^\infty \frac{|\phi_{X_{\theta_A}}\phi_{X_{\theta_B}} - \phi_{Y_{\theta_A}}\phi_{Y_{\theta_B}}|^2}{t^2} \mathrm{d}t\right)^{1/2} + \left(2c_1 \int_0^\infty \frac{|\phi_{Y_{\theta_A}}\phi_{Y_{\theta_B}} - \phi_{Y_\theta}|^2}{t^2} \mathrm{d}t\right)^{1/2} \tag{S1.79}$$

The second term on the right hand side is precisely the definition of the projected dependence for Y, $\mathrm{Cov}_E(\theta; Y_A, Y_B)$. Consider the first term, noting that $|ab - cd| \leq |a-c||d| + |a||b-d|$ and $\phi(t) \leq 1$ for all t,

$$\left(2c_1 \int_0^\infty \frac{|\phi_{X_{\theta_A}}\phi_{X_{\theta_B}} - \phi_{Y_{\theta_A}}\phi_{Y_{\theta_B}}|^2}{t^2} \mathrm{d}t\right)^{1/2} \leq \left(2c_1 \int_0^\infty \frac{(|\phi_{X_{\theta_A}} - \phi_{Y_{\theta_A}}| + |\phi_{X_{\theta_B}} - \phi_{Y_{\theta_B}}|)^2}{t^2} \mathrm{d}t\right)^{1/2} \tag{S1.80}$$

$$\leq E(X_{\theta_A}, Y_{\theta_A}) + E(X_{\theta_B}, Y_{\theta_B}) \tag{S1.81}$$

Using the scaling property of the energy distance, $E(aZ, aW) = |a|E(Z, W)$, we get

$$E(\theta_A^T X_A, \theta_A^T Y_A) = \|\theta_A\| E(\mathbf{u}^T X_A, \mathbf{u}^T Y_A) \tag{S1.82}$$

Combining the terms yields the stated bound. □

Proof of Theorem 3.17. From Lemma 3.16, letting $\alpha = \|\theta_A\|$ and $\beta = \|\theta_B\|$, and by the definition of the suprema S_A and S_B, we have

$$E(X_\theta, Y_\theta) \leq \alpha S_A + \beta S_B + \mathrm{Cov}_E(\theta; X_A, X_B) + \mathrm{Cov}_E(\theta; Y_A, Y_B). \tag{S1.83}$$

Taking the supremum over all $\theta \in S^{d-1}$ and applying the property that the supremum of a sum is less than or equal to the sum of the suprema:

$$\sup_{\theta \in S^{d-1}} E(X_\theta, Y_\theta) \leq \sup_{\theta \in S^{d-1}} (\alpha S_A + \beta S_B + \mathrm{Cov}_E(\theta; X_A, X_B) + \mathrm{Cov}_E(\theta; Y_A, Y_B)) \tag{S1.84}$$

$$\leq \sup_{\substack{\alpha^2+\beta^2=1 \\ \alpha,\beta \geq 0}} (\alpha S_A + \beta S_B) + \sup_{\theta \in S^{d-1}} \mathrm{Cov}_E(\theta; X_A, X_B) + \sup_{\theta \in S^{d-1}} \mathrm{Cov}_E(\theta; Y_A, Y_B). \tag{S1.85}$$

The supremum of $\alpha S_A + \beta S_B$ subject to $\alpha^2 + \beta^2 = 1$ is $\sqrt{S_A^2 + S_B^2}$ by the Cauchy-Schwarz inequality. Combining the terms gives the desired bound. □

Proof of Theorem 3.18. We start from the inequality in the Generalised Pointwise Slice Bound (Lemma 3.16). Grouping the terms and squaring both sides, using the inequality $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$ we have

$$E^2(X_\theta, Y_\theta) \leq ((\alpha S_A + \beta S_B) + (\mathrm{Cov}_E(\theta; X_A, X_B) + \mathrm{Cov}_E(\theta; Y_A, Y_B)))^2 \tag{S1.86}$$

$$\leq 3(\alpha S_A + \beta S_B)^2 + 3\mathrm{Cov}_E^2(\theta; X_A, X_B) + 3\mathrm{Cov}_E^2(\theta; Y_A, Y_B). \tag{S1.87}$$

By the Cauchy-Schwarz inequality, $(\alpha S_A + \beta S_B)^2 \leq (\alpha^2+\beta^2)(S_A^2+S_B^2) = S_A^2 + S_B^2$.

$$E^2(X_\theta, Y_\theta) \leq 3(S_A^2+S_B^2) + 3\mathrm{Cov}_E^2(\theta; X_A, X_B) + 3\mathrm{Cov}_E^2(\theta; Y_A, Y_B). \tag{S1.88}$$

Integrating this inequality over the sphere S^{d-1} and using the integral representation of $E^2(X,Y)$ along with the definitions of the projected dependences yields

$$\frac{2c_1}{c_d} E^2(X,Y) \leq 3A_{d-1}\left[(S_A^2+S_B^2) + \mathrm{Cov}_E^2(X_A, X_B) + \mathrm{Cov}_E^2(Y_A, Y_B)\right]. \tag{S1.89}$$

Since $C_d^2 = \frac{c_d A_{d-1}}{2c_1}$, substituting and taking the square root yields the desired result. □

S1.7 Multivariate Projection in Independent Structures

We extend the results of section 3.4.1 to projections onto general D-dimensional subspaces, where $1 \leq D < d$. While the concepts generalise naturally, the proofs require careful adaptation to accommodate the matrix structure of multi-dimensional projections.

Definition S1.1 (Multivariate Projected Dependence)**.** *Let $X = (X_A, X_B)$ be a random vector in $\mathbb{R}^d$ and let $A \in V_D(\mathbb{R}^d)$ be a projection matrix onto a D-dimensional subspace, partitioned conformably as $A = [A_A, A_B]$, where A_A is $D \times d_A$ and A_B is $D \times d_B$. The multivariate projected dependence is defined as*

$$Cov_E(A; X_A, X_B) \equiv E(AX, Z_A + Z_B) \tag{S1.90}$$

where $Z_A \sim A_A X_A$ and $Z_B \sim A_B X_B$ are independent D-dimensional random vectors. Its squared value has the integral representation

$$Cov_E^2(A; X_A, X_B) = c_D \int_{\mathbb{R}^D} \frac{|\phi_{AX}(t) - \phi_{A_A X_A}(t)\phi_{A_B X_B}(t)|^2}{||t||^{D+1}} dt \tag{S1.91}$$

The average multivariate projected dependence over all D-dimensional projections is:

$$Cov_{E,D}^2(X_A, X_B) \equiv \int_{V_D(\mathbb{R}^d)} Cov_E^2(A; X_A, X_B) d\mu(A) \tag{S1.92}$$

where μ is the normalised invariant measure on the Stiefel manifold $V_D(\mathbb{R}^d)$.

Lemma S1.2 (Multivariate Pointwise Projection Bound)**.** *Let $X = (X_A, X_B)$ and $Y = (Y_A, Y_B)$ be random vectors in $\mathbb{R}^d$ with finite second moments. For any projection matrix $A = [A_A, A_B] \in V_D(\mathbb{R}^d)$, the projected energy distance satisfies:*

$$E(AX, AY) \leq E(A_A X_A, A_A Y_A) + E(A_B X_B, A_B Y_B) + Cov_E(A; X_A, X_B) + Cov_E(A; Y_A, Y_B) \tag{S1.93}$$

Theorem S1.1 (Multivariate Energy Distance Bound)**.** *Let $X = (X_A, X_B)$ and $Y = (Y_A, Y_B)$ be random vectors in $\mathbb{R}^d$ with finite second moments. The total energy distance is bounded by:*

$$E(X,Y) \leq 2C_{d,D}\sqrt{\mathscr{E}_D^2(X_A,Y_A) + \mathscr{E}_D^2(X_B,Y_B) + Cov_{E,D}^2(X_A,X_B) + Cov_{E,D}^2(Y_A,Y_B)} \tag{S1.94}$$

where $C_{d,D}$ is the constant from Corollary 3.6 and

$$\mathscr{E}_D^2(X_A,Y_A) \equiv \int_{V_D(\mathbb{R}^d)} E^2(A_A X_A, A_A Y_A) d\mu(A) \tag{S1.95}$$

Proof of Lemma S1.2. Introduce the intermediate random variable $Z = Z_A + Z_B$ where $Z_A \sim A_A X_A$ and $Z_B \sim A_B X_B$ are independent. By the triangle inequality:

$$E(AX,AY) \leq E(AX,Z) + E(Z,AY) \tag{S1.96}$$

The first term equals $Cov_E(A; X_A, X_B)$ by definition. For the second term, we apply the triangle inequality and Minkowski's integral inequality on the characteristic functions. Setting $W = W_A + W_B$ where $W_A \sim A_A Y_A$ and $W_B \sim A_B Y_B$ are independent:

$$\begin{aligned} E(Z,AY) &= \left(2c_D \int_0^\infty \frac{|\phi_Z(t) - \phi_{AY}(t)|^2}{|t|^{D+1}} dt\right)^{1/2} \\ &\leq \left(2c_D \int_0^\infty \frac{|\phi_Z(t) - \phi_W(t)|^2}{|t|^{D+1}} dt\right)^{1/2} + \left(2c_D \int_0^\infty \frac{|\phi_W(t) - \phi_{AY}(t)|^2}{|t|^{D+1}} dt\right)^{1/2} \end{aligned}$$

The second term is $Cov_E(A; Y_A, Y_B)$. For the first term, since Z and W have independent components:

$$\begin{aligned} E(Z,W) &= E(Z_A + Z_B, W_A + W_B) \\ &\leq E(Z_A, W_A) + E(Z_B, W_B) \\ &= E(A_A X_A, A_A Y_A) + E(A_B X_B, A_B Y_B) \end{aligned}$$

where the inequality follows from Minkowski's inequality applied to the integral representation. Combining yields the result. □

Proof of Theorem S1.1. Square both sides of the inequality in Lemma S1.2. Using $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$:

$$E^2(AX,AY) \leq 4\left[E^2(A_A X_A, A_A Y_A) + E^2(A_B X_B, A_B Y_B) + Cov_E^2(A; X_A, X_B) + Cov_E^2(A; Y_A, Y_B)\right] \tag{S1.97}$$

Integrate over $V_D(\mathbb{R}^d)$ with respect to $d\mu(A)$:

$$\int_{V_D(\mathbb{R}^d)} E^2(AX,AY) d\mu(A) \leq 4 \int_{V_D(\mathbb{R}^d)} \left[E^2(A_A X_A, A_A Y_A) + \cdots\right] d\mu(A) \tag{S1.98}$$

By Corollary 2.1, the left side equals $\frac{1}{C_{d,D}^2} E^2(X,Y)$. The right side integrals are precisely the average projected quantities by definition. Thus:

$$\frac{1}{C_{d,D}^2} E^2(X,Y) \leq 4\left[\mathscr{E}_D^2(X_A,Y_A) + \mathscr{E}_D^2(X_B,Y_B) + Cov_{E,D}^2(X_A,X_B) + Cov_{E,D}^2(Y_A,Y_B)\right] \tag{S1.99}$$

Taking square roots and rearranging yields the result. □

Table S2.1: Results of the GM, MLP, and FGN models on the low-dimensional Gaussian covariance detection task. Lower is better.

Model	Metric	Training regime		Ground Truth Forecast
		CRPS Training ↓	Energy Training ↓	
GM	Energy Score	0.627 ± 0.012	**0.606 ± 0.011**	0.604 ± 0.011
	CRPS	0.197 ± 0.004	0.196 ± 0.004	0.196 ± 0.004
	MSE	0.128 ± 0.005	0.127 ± 0.005	0.127 ± 0.005
MLP	Energy Score	0.650 ± 0.013	**0.613 ± 0.012**	
	CRPS	0.198 ± 0.004	0.198 ± 0.004	
	MSE	0.130 ± 0.005	0.129 ± 0.005	
FGN	Energy Score	0.633 ± 0.012	**0.610 ± 0.011**	
	CRPS	0.199 ± 0.004	0.197 ± 0.004	
	MSE	0.130 ± 0.005	0.128 ± 0.005	

Values are mean ± standard deviation estimated over 50 runs.

S2 Experimental Details

S2.1 Low dimensional Gaussian Covariance Detection

In low dimensions, the energy score effectively recognises covariance structure, whilst the multivariate CRPS cannot due to its marginal-only focus. We therefore expect CRPS to perform poorly compared to the energy score in this regime.

To demonstrate this, we generate $20,000$ samples from $X \sim \mathcal{N}(0, \Sigma)$ where Σ is a randomly generated 2×2 covariance matrix. We then train a Gaussian model (GM), a two layer feed-forward multi-layer perceptron (MLP), and a stripped-down Functional Generative Network (FGN) models using $n = 100$ samples to learn the unconditional distribution of the generated data (For more detailed description of the models, see Appendix S2.2.1). For baseline comparison, we also simulate the respective scores under data generating distribution forecasts to gauge the performance gap between the models and the ground truth distribution. The quantitative results are presented in Table S2.1. Since we are dealing with two-dimensional data, the model performance can also be verified visually through the training progressions shown in Figures S2.1 and S2.2.

Discussion. The energy score clearly outperforms CRPS in low dimensions, as shown in Table S2.1 and Figures S2.1 and S2.2. The energy score learns correct covariance structure even from poor initialisations, whilst CRPS merely fits marginals.

S2.2 Models

S2.2.1 Synthetic Data

Since our work does not concern too much on inductive biases of the model architecture, throughout our experiments we pick three straightforward deep neural network models to train. The models we use are

1. A Gaussian model (GM) with a trainable mean vector $\mu \in \mathbb{R}^d$ and a trainable $d \times H$

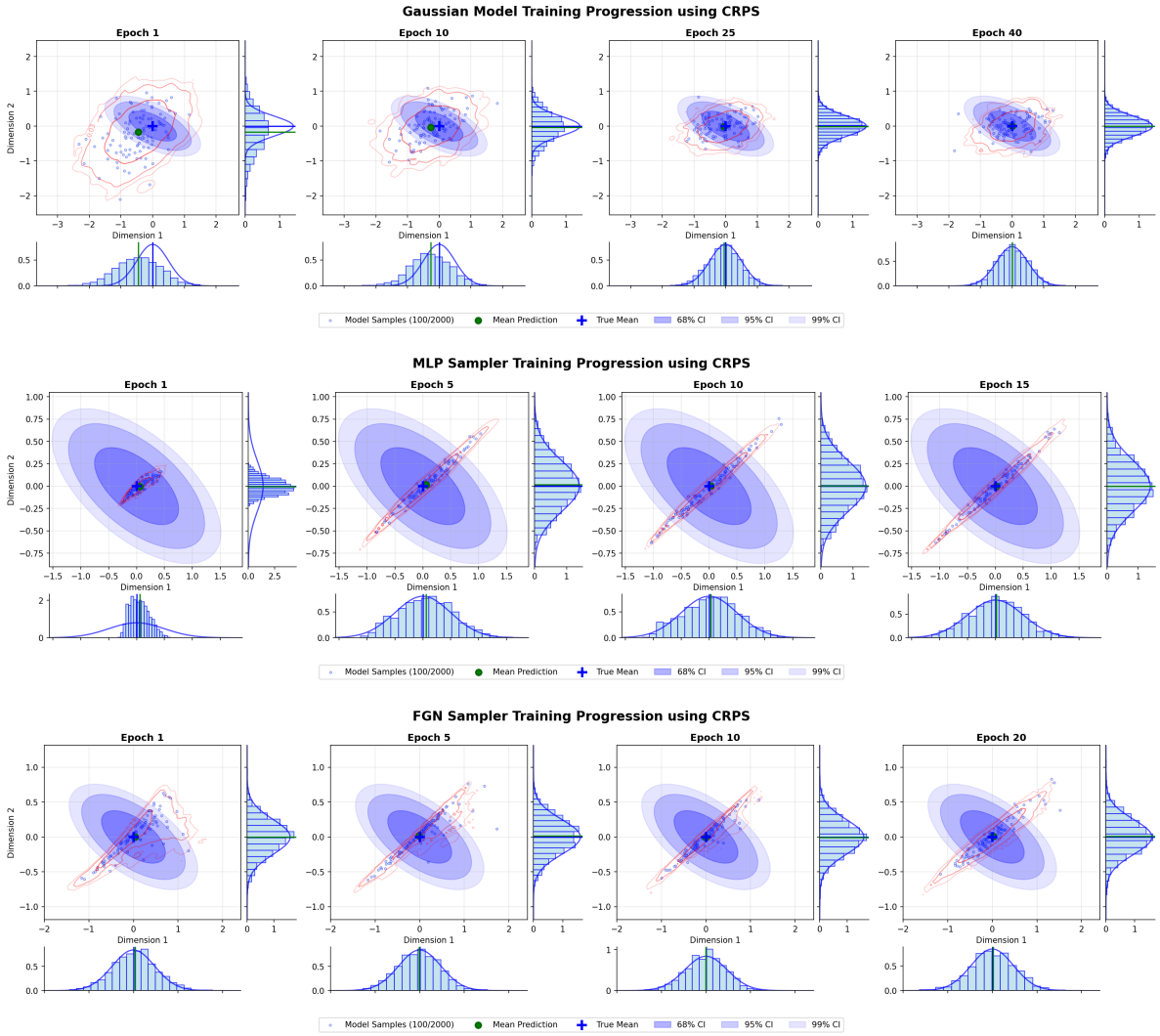

Figure S2.1: Training progression of the CRPS loss for the GM (top), MLP (middle) and FGN (bottom) models. All models fail to learn the covariance structure, but fit onto the marginals well.

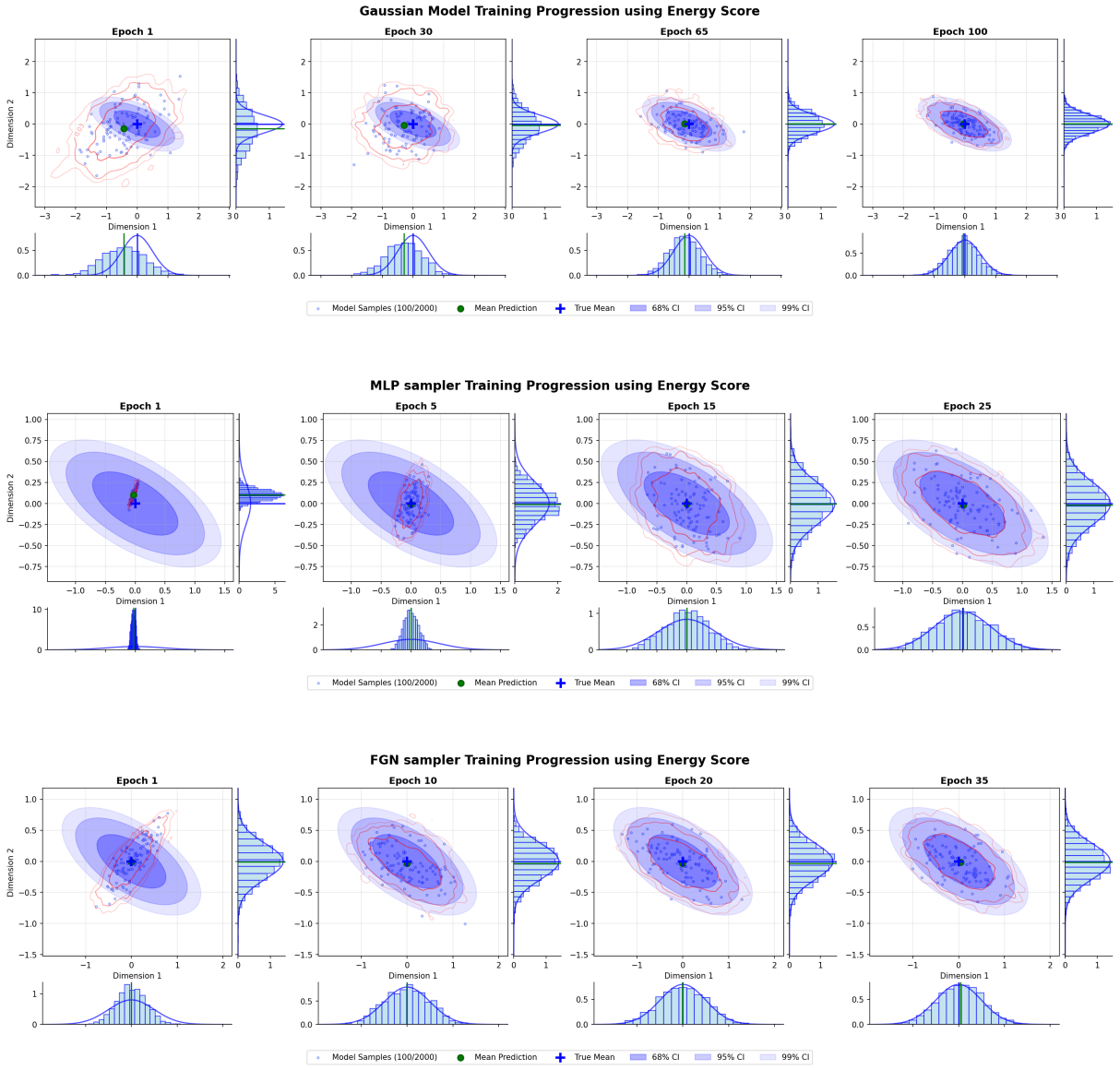

Figure S2.2: Training progression of the energy score for the GM (top), MLP (middle) and FGN (bottom) models. All models learn the covariance structure well, even when starting initialisations are of the wrong correlation.

dimensional matrix A, where samples are generated as:

$$Z_i \sim \mathcal{N}(0, I_H) \quad \text{for } i = 1, \ldots, n \tag{S2.100}$$

$$X_i = AZ_i + \mu \quad \text{for } i = 1, \ldots, n \tag{S2.101}$$

resulting in $X_i \sim \mathcal{N}(\mu, AA^\top)$. Here H is the latent dimension and d is the data dimension. This model does not support conditional generation.

2. A feedforward multi-layer perceptron (MLP) with one hidden layer and ReLU activation. To generate n samples from the conditional distribution given input x, we sample n independent H-dimensional standard Gaussian noise vectors $z_i \sim \mathcal{N}(0, I_H)$ and concatenate each with the conditioning input x to form $[x; z_i]$:

$$h_i = \text{ReLU}(W_1[x; z_i] + b_1) \tag{S2.102}$$

$$y_i = W_2 h_i + b_2 \tag{S2.103}$$

For unconditional generation, x is omitted.

3. A stripped-down Functional Generative Network (FGN) encoder module as proposed by Alet et al. [1]. To generate n samples, we sample noise vectors $z_i \sim \mathcal{N}(0, I_H)$ and encode them as $c_i = W_{\text{enc}} z_i + b_{\text{enc}}$. Each layer applies Conditional Layer Normalisation followed by noise-dependent affine transformation:

$$\hat{h}_i^{(\ell-1)} = \frac{h_i^{(\ell-1)} - \mu_{\ell-1}}{\sqrt{\sigma_{\ell-1}^2 + \varepsilon}} \tag{S2.104}$$

$$\tilde{h}_i^{(\ell-1)} = (\gamma_\ell + \gamma_{\text{proj}}^{(\ell)}(c_i)) \odot \hat{h}_i^{(\ell-1)} + (\beta_\ell + \beta_{\text{proj}}^{(\ell)}(c_i)) \tag{S2.105}$$

$$h_i^{(\ell)} = \text{ReLU}(W_\ell \tilde{h}_i^{(\ell-1)} + b_\ell) \tag{S2.106}$$

where $\odot$ denotes element-wise multiplication.

S2.2.2 2D CFD Data

For 2D CFD data experiments, the ConvCNP [3] is particularly well-suited for this task as it combines the flexibility of Neural Processes with the inductive biases of convolutional networks for spatial data.

Let the context observations be $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^2$ are spatial coordinates and $\mathbf{y}_i \in \mathbb{R}^d$ are the corresponding field values. We use the same model architecture as the one used in Gordon et al. [3], but with the following modifications:

- **Set Convolution Encoder.** The first stage maps irregular spatial observations to a uniform grid $\mathcal{G} = \{\mathbf{g}_j\}_{j=1}^{G^2}$ using set convolutions with RBF kernels. For each grid point $\mathbf{g}_j$, the aggregated representation is computed as:

$$\mathbf{r}_j = \frac{\sum_{i=1}^N k(\mathbf{g}_j, \mathbf{x}_i)\mathbf{y}_i}{\sum_{i=1}^N k(\mathbf{g}_j, \mathbf{x}_i) + \varepsilon} \quad \text{and} \quad \rho_j = \sum_{i=1}^N k(\mathbf{g}_j, \mathbf{x}_i) \tag{S2.107}$$

where $k(\mathbf{g}, \mathbf{x}) = \exp(-\|\mathbf{g} - \mathbf{x}\|^2 / 2\sigma^2)$ is an RBF kernel with learnable length scale σ, and ε prevents division by zero. The density channel ρ_j tracks the amount of information at each grid location, yielding the gridded representation $\mathbf{h}_j = [\mathbf{r}_j; \rho_j] \in \mathbb{R}^{d+1}$.

- **CNN Processor.** The gridded representation is reshaped into a spatial tensor $\mathbf{H} \in \mathbb{R}^{(d+1)\times G\times G}$ and processed through a residual CNN:

$$\mathbf{F} = f_{\text{CNN}}(\mathbf{H}; \theta_{\text{CNN}}) \tag{S2.108}$$

  where f_{CNN} consists of multiple residual blocks with batch normalisation, providing translation equivariance and efficient spatial feature extraction.

- **Grid Interpolator.** The CNN features are mapped back to arbitrary target locations $\{\mathbf{x}_k^*\}_{k=1}^M$ using another RBF-based interpolation:

$$\mathbf{z}_k = \frac{\sum_{j=1}^{G^2} k(\mathbf{x}_k^*, \mathbf{g}_j)\mathbf{f}_j}{\sum_{j=1}^{G^2} k(\mathbf{x}_k^*, \mathbf{g}_j) + \varepsilon} \tag{S2.109}$$

  where $\mathbf{f}_j$ are the flattened CNN features corresponding to grid location $\mathbf{g}_j$.

- **Probabilistic Decoder.** Finally, a probabilistic MLP decoder generates multiple samples by incorporating stochastic latent variables:

$$\mathbf{y}_k^{(s)} = g_{\text{MLP}}([\mathbf{z}_k; \varepsilon^{(s)}]; \theta_{\text{MLP}}) \tag{S2.110}$$

  where $\varepsilon^{(s)} \sim \mathcal{N}(0, \mathbf{I})$ are independent Gaussian latent variables for sample s, and $[\cdot;\cdot]$ denotes concatenation. This produces S samples $\{\mathbf{y}_k^{(s)}\}_{s=1}^S$ at each target location, enabling uncertainty quantification.

References

[1] Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals, 2025. URL https://arxiv.org/abs/2506.10772.

[2] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book.

[3] Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Skey4eBYPS.

[4] Luis Antonio Santalo. *Integral geometry and geometric probability*. Cambridge university press, 2004.

[5] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Springer, 2008.