Information Geometry of Gaussian Processes

Shotaro Akaho (https://shotaroakaho.github.io/misc/)
ISM

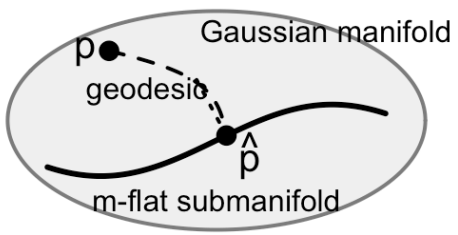
Motivation

> Gaussian processes are infinite-dimensional Gaussian models
> — but their information geometry is surprisingly nontrivial.

- KL divergence depends on input choices
- naive infinite-dimensional KL often diverges
- we introduce averaged KL divergence and IGP geometry (Akaho+2025)

Why Information Geometry for GPs?

Multivariate Gaussian



- Exponential family (dually flat mfd)
- KL divergence
- Geodesics and projections

Gaussian Process

- Infinite-dimensional Gaussian
- Uncertainty over functions
- Can we define its info-geometry?

Key question

Can the information geometry of multivariate Gaussians be extended to Gaussian processes?

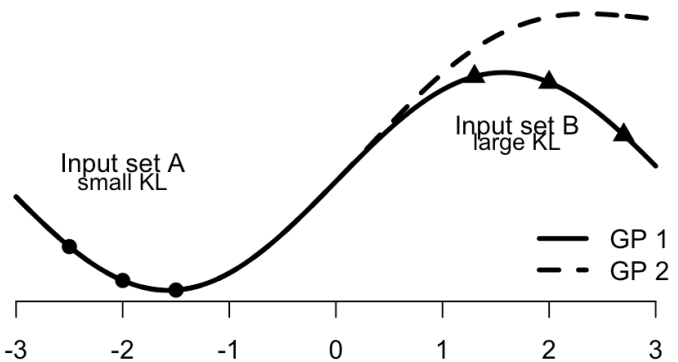
The Difficulty: KL Depends on Inputs

Finite-dimensional Gaussian

For multivariate Gaussians,

$D\left[N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)\right]$

is uniquely defined



$$D[N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)] = \frac{1}{2}\left(\log(|\Sigma_2|/|\Sigma_1|) + \|\mu_1 - \mu_2\|^2_{\Sigma_2^{-1}} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - n\right).$$

Gaussian Process

For GPs, $D[GP_1, GP_2] \quad ?$

The result depends on $X = (X_1, \ldots, X_n)$.

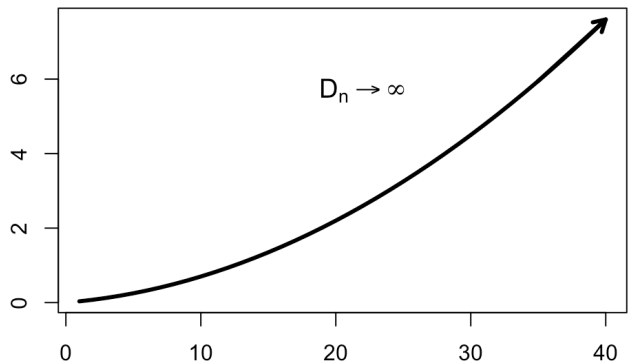
Key Observation
Different choices of input points may produce different KL divergences. Therefore, a KL divergence between Gaussian processes is not uniquely defined unless the input set is specified.

Naive Infinite-Dimensional KL Fails

A natural idea

For finite inputs,

$$D_n \equiv D\left[N_1(X_1, \ldots, X_n), N_2(X_1, \ldots, X_n)\right].$$



Why not increase the number of inputs?

$$D[GP_1, GP_2] \equiv \lim_{n\to\infty} D_n ?$$

- In general, the limit diverges.
- Covariance matrices become nearly singular as inputs get dense.
- Numerical evaluation becomes unstable.

Take-home message Increasing the number of input points does not provide a satisfactory definition of KL divergence for general Gaussian processes. (cf. Ishibashi+2022)

Averaged KL Divergence

Key Idea

Instead of fixing the input locations, $X = (X_1, \ldots, X_k)$, we introduce a probability distribution $q(X)$ over the input space and average the KL divergence.

$$D_{q,k}[GP_1, GP_2] \equiv \int q(X)\, D\left[N_1(X), N_2(X)\right]\, dX$$

- Compare GPs through their finite-dimensional Gaussian marginals.
- Average over possible input configurations.
- Remove dependence on a particular choice of inputs.

Advantages

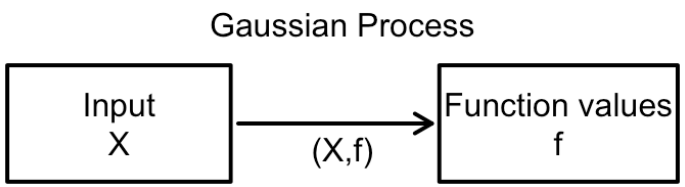
- No arbitrary input set selection
- Covariance information can be incorporated $(k \geq 2)$
- Applicable to general Gaussian processes

A Probabilistic Interpretation

Joint distribution of inputs and outputs

Instead of considering only function values, $f = (f(X_1), \ldots, f(X_k))$,

we consider the joint random variable (X, f).



For each GP, $p(X, f) = q(X)\, p(f \mid X)$.

- $q(X)$: distribution of input locations
- $p(f \mid X)$: Gaussian induced by the GP

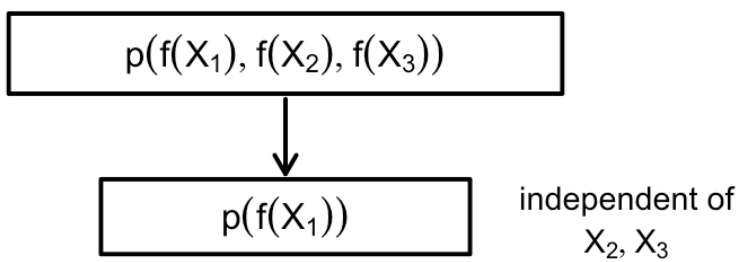
The averaged KL divergence becomes

$$D_{q,k}[GP_1, GP_2] = D[p_1(X, f), p_2(X, f)]\,.$$

Instead of comparing GPs directly, we compare two ordinary probability distributions on the extended space (X, f) (i.e. The averaged KL is an ordinary KL divergence on the extended space (X, f).)

Why Do We Need IGPs?

A Gaussian Process Must Be Consistent



For a valid GP, the marginal distribution of $f(X_1)$ must not depend on X_2, X_3.

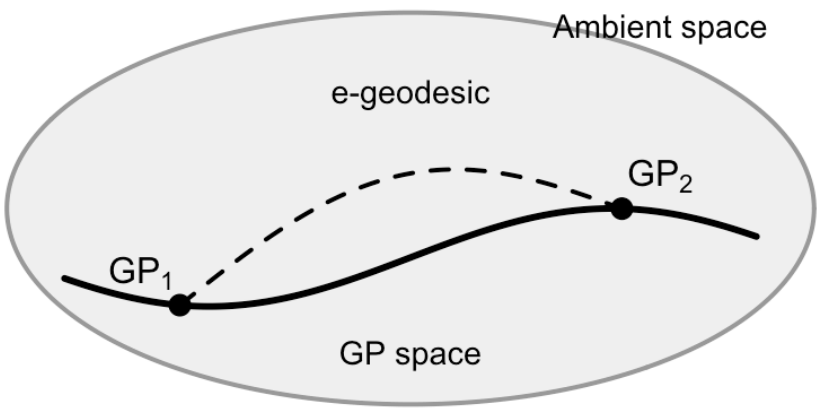
This is the consistency condition.

Every finite-dimensional Gaussian distribution must fit together coherently.

Geometric Question

Can we construct a geometric structure that preserves this consistency?

e-Geodesics Leave the GP Space



Two Gaussian processes are consistent.

Their e-geodesic is generally inconsistent.

$$GP_1,\ GP_2 \in S_{GP}$$

but

$$\text{e-geodesic}(GP_1, GP_2) \not\subset S_{GP}.$$

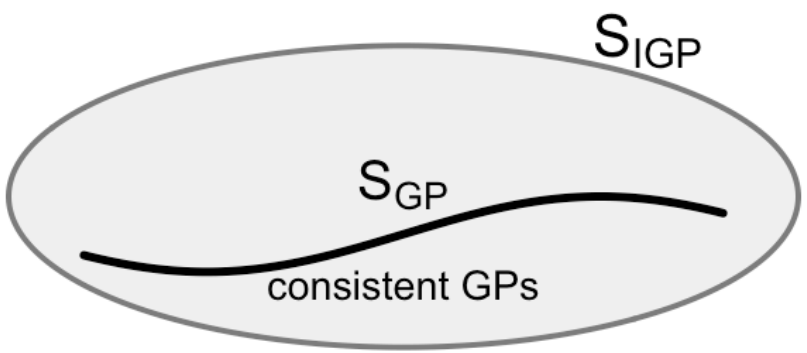
Therefore, S_{GP} is **not e-flat**.

We need a larger ambient space.

Integrated Gaussian Processes (IGPs)

Key Idea

Relax the consistency condition of Gaussian processes.



Gaussian Process

- Consistent marginals
- Valid stochastic process

Integrated GP

- Mean and covariance defined for each input tuple
- Consistency is not required
- $S_{GP} \subset S_{IGP}$

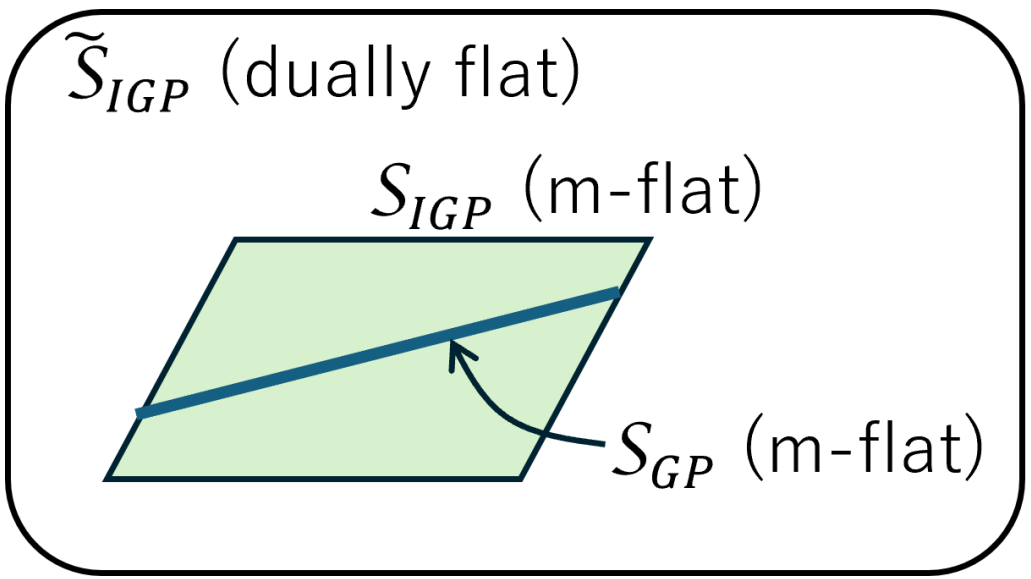
Benefit

The larger space S_{IGP} admits a simple geometric structure, while ordinary GPs appear as a constrained subset.

Geometric Structure

Main Result

The space of Integrated Gaussian Processes admits a dually flat structure.



Properties

- $\widetilde{S}_{IGP}$ is dually flat
- S_{IGP} and S_{GP} are m-flat
- S_{GP} is generally not e-flat

Therefore,

projection and divergence calculations can be performed in the ambient geometric space.

Hierarchy: $S_{GP} \subset S_{IGP} \subset \widetilde{S}_{IGP}$

Key Message

GPs inherit information-geometric structure through their embedding into the larger IGP manifold.

e- and m-Coordinates

Dual Coordinates of Gaussian Families

For a multivariate Gaussian,

$$p(x) = \exp\left(\xi^\top x + \langle \Xi, xx^\top \rangle - \psi(\xi, \Xi)\right),$$

the information-geometric coordinates are

e-coordinates

Natural parameters

$\xi = \Sigma^{-1}\mu, \Xi = -\frac{1}{2}\Sigma^{-1}.$

m-coordinates

Expectation parameters

$\zeta = \mu, Z = \mu\mu^T + \Sigma$

For IGPs, the same coordinates are assigned to every input tuple X, $(\xi(X), \Xi(X))$ and $(\zeta(X), Z(X))$

Why are m-coordinates important?

Convex combinations in m-coordinates preserve GP consistency.

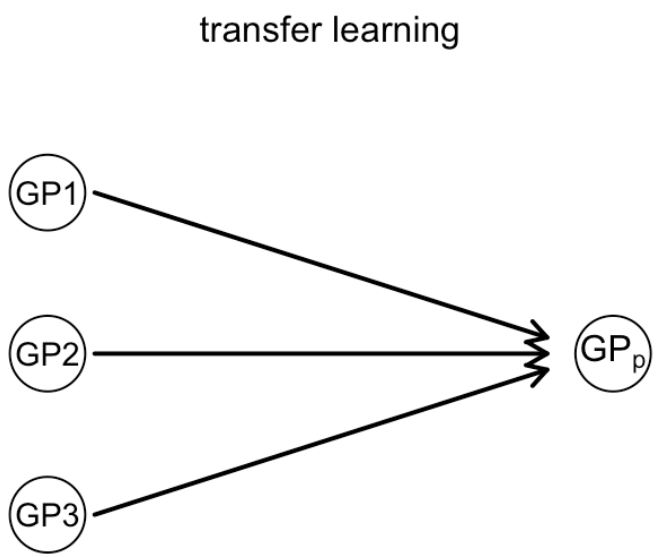
Transfer Learning for Gaussian Processes

Motivation

Suppose that we already have several well-trained source GPs, $GP_1, \ldots, GP_c$, but only a small amount of data is available for a new target task.

Typical situation

- Source tasks: many observations
- Target task: few observations



Goal

Use information from source GPs to improve the target GP.

Question:

How should we combine the source models in a principled way?

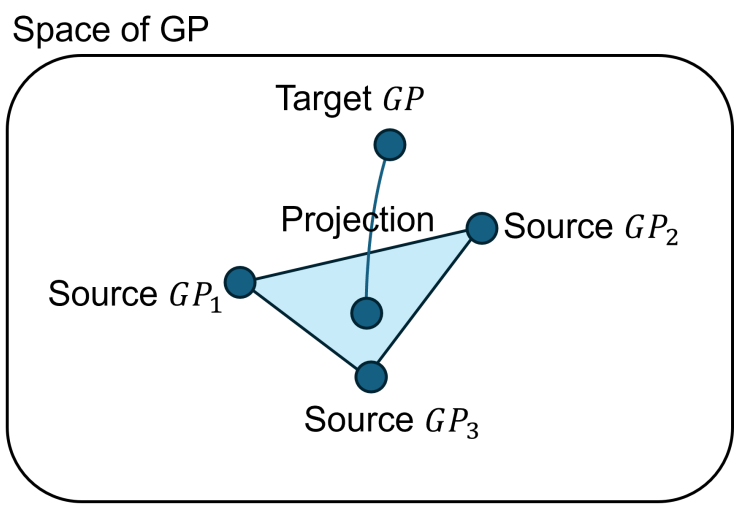
Our Answer

Formulate transfer learning as a geometric projection problem.

Transfer Learning as e-Projection

Geometric Formulation

Let $T = GP_1, \ldots, GP_c$ be source Gaussian processes. Their m-coordinates define a convex set $M(T)$.



Transfer learning of GP by projection

Target GP: GP_p

Projected GP:

$$\widehat{GP} = \arg\min_{GP \in M(T)} D_{q,k}(GP, GP_p)$$

The projected model is the closest point in the convex hull of source GPs.

Interpretation

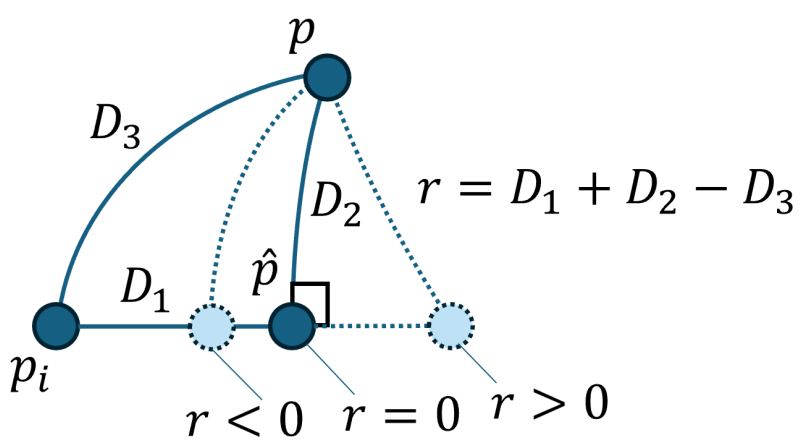
Transfer learning becomes an information-geometric projection problem.

Computing the Projection

Challenge

The parameter space of Gaussian processes is infinite-dimensional(Takano+2016).

Direct optimization of $\widehat{GP} = \arg\min_{GP \in M(T)} D_{q,k}(GP, GP_p)$ is generally intractable.



Use the generalized Pythagorean theorem.

For each source GP_i, let $r_i \equiv D(GP_i, \widehat{GP}_t) + D(\widehat{GP}_t, GP_p) - D(GP_i, GP_p)$ at t.

- $r_i > 0$: increase weight
- $r_i < 0$: decrease weight
- $r_i = 0$: projection condition

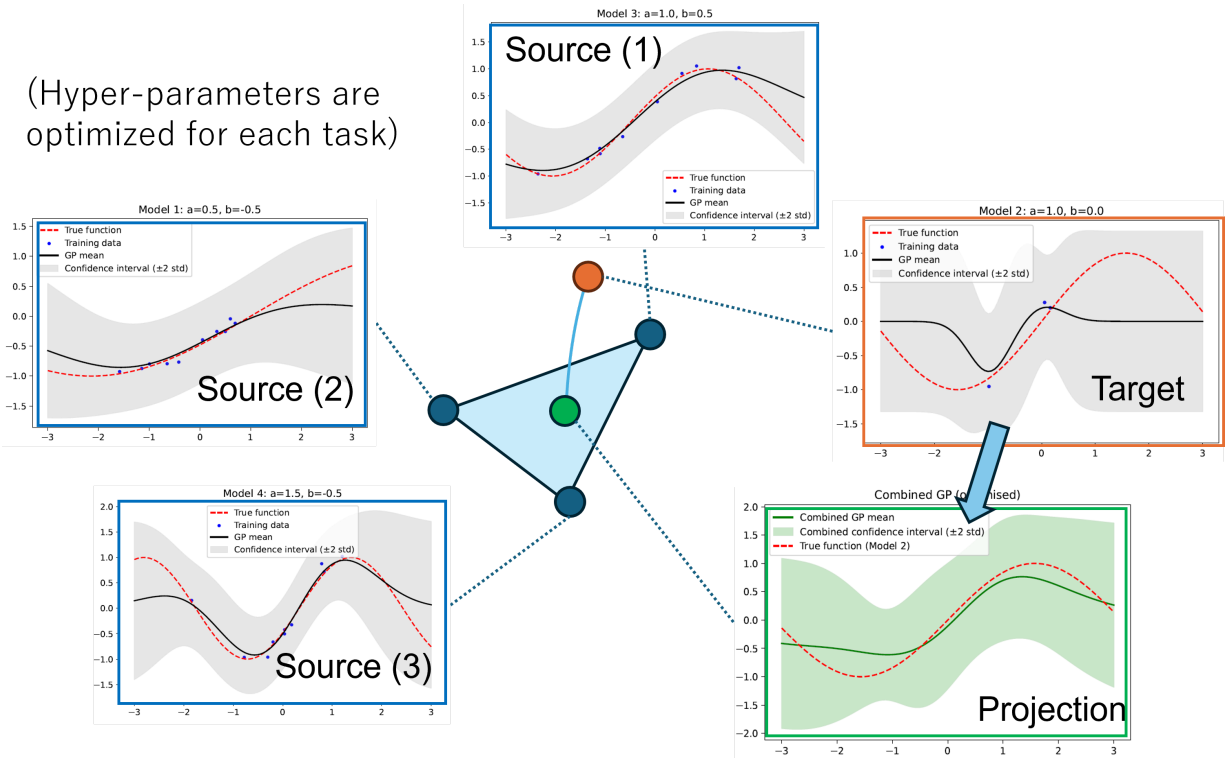
Key Advantage

Only divergence evaluations are required.

No explicit optimization in infinite-dimensional coordinates is needed.

Numerical Example

(Hyper-parameters are optimized for each task)



Observation

- The target GP is uncertain because only a few samples are available.
- Source GPs contain useful information but are not identical to the target.
- The projected GP gives a better estimate of the target function.

The projection is computed only from averaged KL divergences.

Conclusions

Story

Traditional KL for GPs is problematic.

$\Downarrow$

Averaged KL divergence

$\Downarrow$

IGP as a dually flat ambient space

$\Downarrow$

GP space as an m-flat submanifold

$\Downarrow$

Transfer learning as e-projection

Future Directions

- Extrapolation and positive definiteness
- Dimension reduction on GP spaces
- Symmetry breaking between m- and e-geometries
- Extensions to copula models and optimal transport

Remark 1. KL When Two Inputs Become Close

Consider the KL divergence between the Gaussians at $(X_1, X_2) \in \mathbb{R}^2$ induced by two GPs GP_A, GP_B.

$$KL_2(f(X_1), f(X_2)) \equiv D[N(\mu_A, \Sigma_A), N(\mu_B, \Sigma_B) \mid (X_1, X_2)]$$

$$= \frac{1}{2}\left(\log(|\Sigma_B|/|\Sigma_A|) + \|\mu_A - \mu_B\|^2_{\Sigma_B^{-1}} + \mathrm{tr}(\Sigma_B^{-1}\Sigma_A) - 2\right).$$

Question: If $X_2 \to X_1$, does $KL_2(f(X_1), f(X_2))$ converge to the one-dimensional KL divergence $KL_1(f(X_1)) = D[p_A(f(X_1)), p_B(f(X_1))]$?

Note: The covariance matrix becomes nearly singular, so the limiting behavior is non-trivial.

Question: If $X_2 \to X_1$, does $KL_2(f(X_1), f(X_2))$ converge to the one-dimensional KL divergence $KL_1(f(X_1)) = D[p_A(f(X_1)), p_B(f(X_1))]$?

Answer: In general, KL_2 does **not** converge to KL_1.

It converges to the KL divergence of the joint distribution of f and f' at X_1.

Indeed, assuming sufficient smoothness,

$$\lim_{X_2 \to X_1} KL_2(f(X_1), f(X_2)) = \lim_{\epsilon \to 0} KL_2(f(X_1), f(X_1 + \epsilon))$$

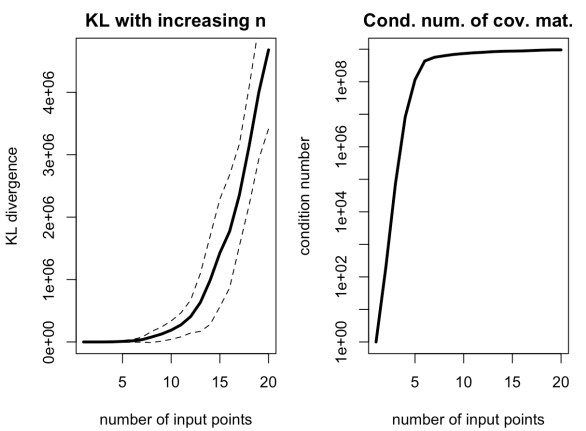
$$= \lim_{\epsilon \to 0} KL_2\left(f(X_1), \frac{f(X_1+\epsilon) - f(X_1)}{\epsilon}\right)$$

$$= D[p_A(f(X_1), f'(X_1)), p_B(f(X_1), f'(X_1))]\,.$$

cf. $\quad KL_1(f(X_1)) = D[p_A(f(X_1)), p_B(f(X_1))]]$

Remark 2. Calculating Averaged KL Divergence

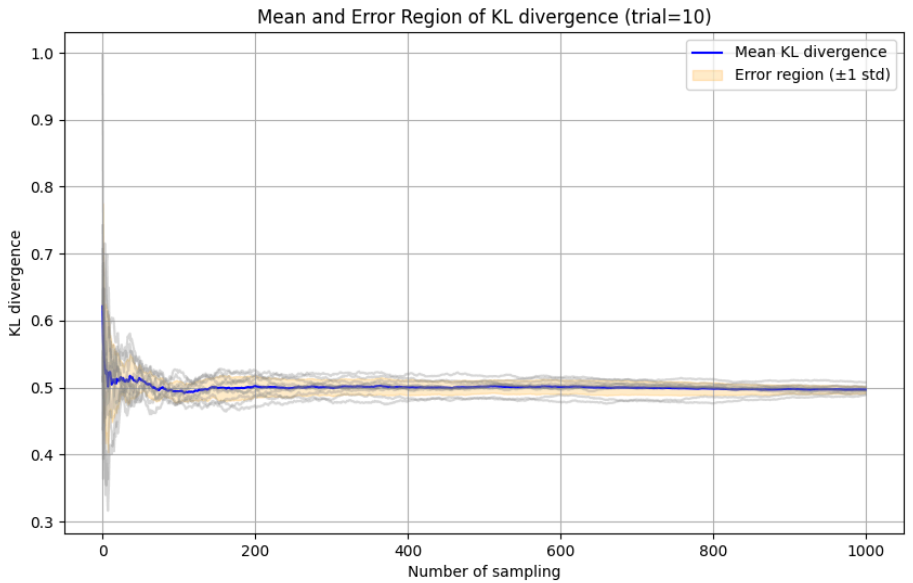
- Naive finite-dim KL, $D_n = D[N_1(X_1, \ldots, X_n), N_2(X_1, \ldots, X_n)]$, typically grows with n.
- Averaged KL depends on the input prior q and the number of points k.
- $k = 1$ only compares marginal mean and variance.
- $k = 2$ already incorporates covariance structure.
- Monte Carlo integration is useful for large k.
- Gauss–Hermite quadrature is effective for small k, especially $k = 2$ or 3 and q is Gaussian.



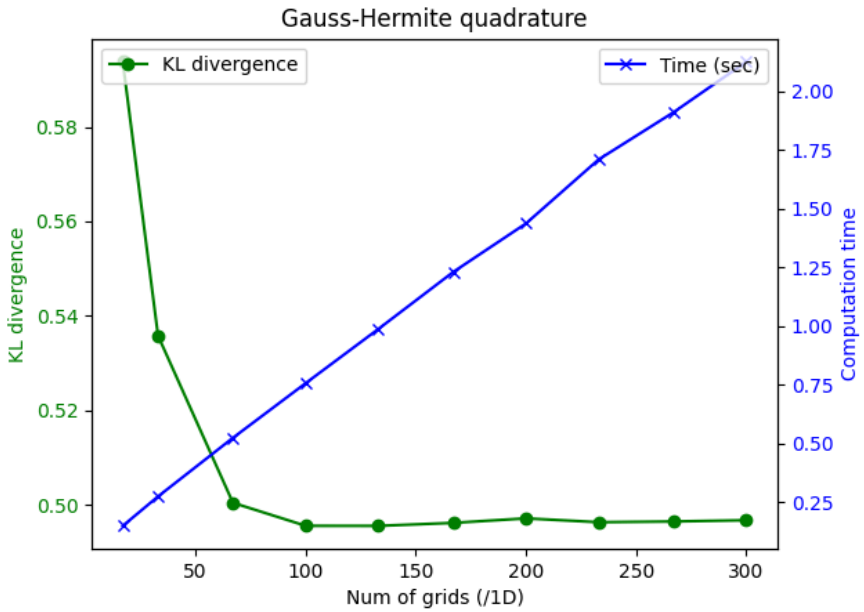
Averaged KL replaces an unstable infinite-dimensional limit with a controlled finite-dimensional integration problem.

Calculation of KL divergence

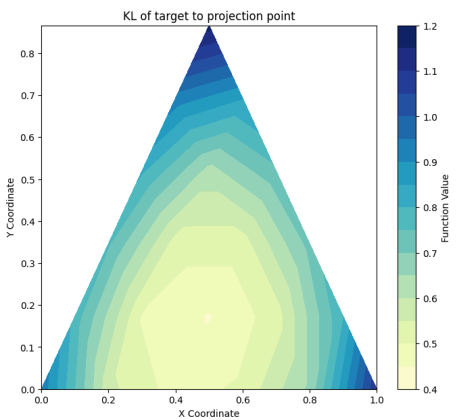
Monte Carlo estimation



Gauss-Hermite quadrature



Contour of KL



Remark 3. Why is the GP Space not e-flat?

Consider two Gaussian processes with covariance functions $v_1(x, x'), v_2(x, x')$.

For a pair of inputs $X = (X_1, X_2)$, their covariance matrices are $V_i(X) = \begin{pmatrix} v_i(X_1, X_1) & v_i(X_1, X_2) \\ v_i(X_2, X_1) & v_i(X_2, X_2) \end{pmatrix}$.

e-geodesic

The e-coordinate of a Gaussian distribution is $\Xi_i(X) = -\frac{1}{2} V_i(X)^{-1}$.

Thus $\Xi_t(X) = (1-t)\Xi_1(X) + t\Xi_2(X)$, i.e., $V_t(X) = \left((1-t)V_1(X)^{-1} + tV_2(X)^{-1}\right)^{-1}$.

Consistency condition

For a valid GP, $V_t(X_1, X_1) = v_t(X_1, X_1)$ must be independent of X_2.

However, $V_t(X) = \left((1-t)V_1(X)^{-1} + tV_2(X)^{-1}\right)^{-1}$ depends on X_2.

Therefore the marginal variance generally depends on the other input point, i.e., The e-geodesic generally violates consistency: $\text{e-geodesic}(GP_1, GP_2) \notin S_{GP}$. Hence S_{GP} is not e-flat.

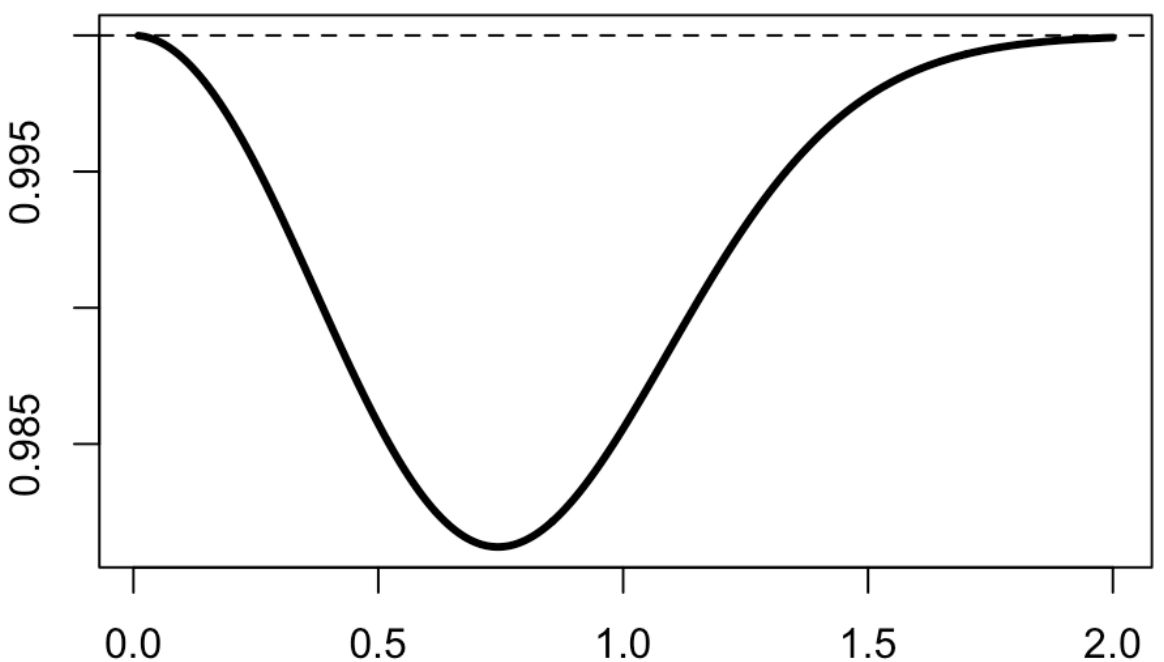
Numerical Illustration

Consider $v_1(x, x') = \exp\left(-(x - x')^2\right)$, and $v_2(x, x') = \exp\left(-2(x - x')^2\right)$.

Let $X = (0, h)$ and $t = 1/2$ (midpoint).

For a genuine GP, $v(0, 0) = 1$ for all h.

The following figure shows $[V_t(0, h)]_{11}$, where $V_t = \left(\frac{1}{2}V_1^{-1} + \frac{1}{2}V_2^{-1}\right)^{-1}$.



Remark 4. The Ambient Exponential Family

Recall the joint distribution $p(X, f) = q(X)p(f|X)$.

For every input tuple (X), the conditional distribution is Gaussian: $p(f|X) = N(\mu(X), \Sigma(X))$.

Using natural parameters $\xi(X) = \Sigma(X)^{-1}\mu(X), \Xi(X) = -\frac{1}{2}\Sigma(X)^{-1}$, we obtain

$$p(X, f) = \exp\left(\langle \xi(X), f\rangle + \langle \Xi(X), ff^\top\rangle - \psi(\xi(X), \Xi(X))\right) q(X).$$

Ambient Space

This is a subset of the ambient space $\widetilde{S}_{IGP}$ of the exponential family

$$p(X, f) = \exp\left(\langle \xi(X), f\rangle + \langle \Xi(X), ff^\top\rangle + \alpha(X)\right).$$

Let $\nu(X)$ be the dual coordinate corresponding to $\alpha(X)$, S_{IGP} is the space restricted $\nu(X) = q(X)$, which means S_{IGP} is an m-flat submanifold of $\widetilde{S}_{IGP}$

References

1. S. Akaho and H. Ishibashi.
   Information geometry of Gaussian processes and its applications to transfer learning.
   Information Geometry, 2025.
   DOI: 10.1007/s41884-025-00180-5.

2. H. Ishibashi and S. Akaho.
   Principal component analysis for Gaussian process posteriors.
   Neural Computation, 34(5):1189–1219, 2022.

3. K. Takano, H. Hino, S. Akaho, and N. Murata.
   Nonparametric e-mixture estimation.
   Neural Computation, 28(12):2687–2725, 2016.