Talk for Google Research

# Pathwise Conditioning and Non-Euclidean Gaussian Processes

UNIVERSITY OF CAMBRIDGE

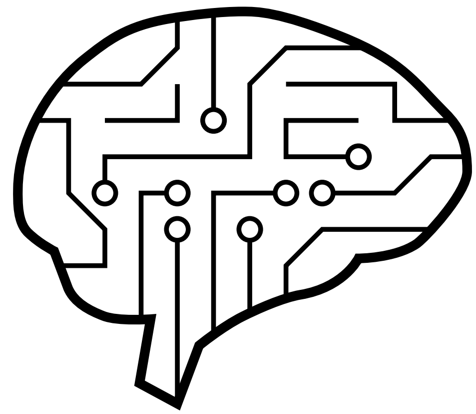
Alexander Terenin

https://avt.im/ · @avt_im

# Huge Interest in Machine Learning

Computer Vision

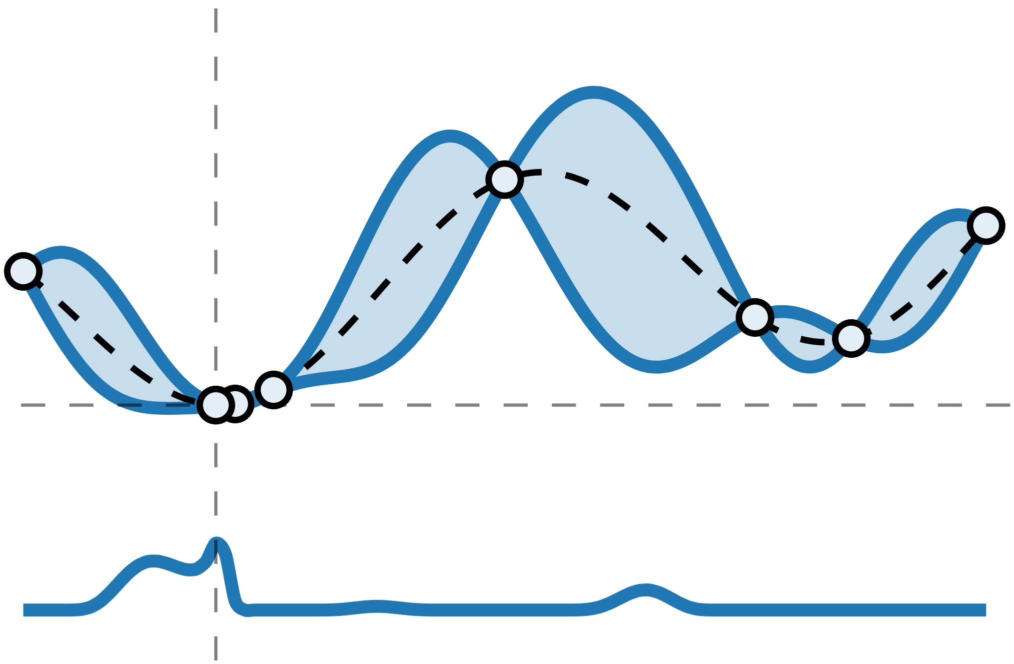

Natural Language Processing



Robotics

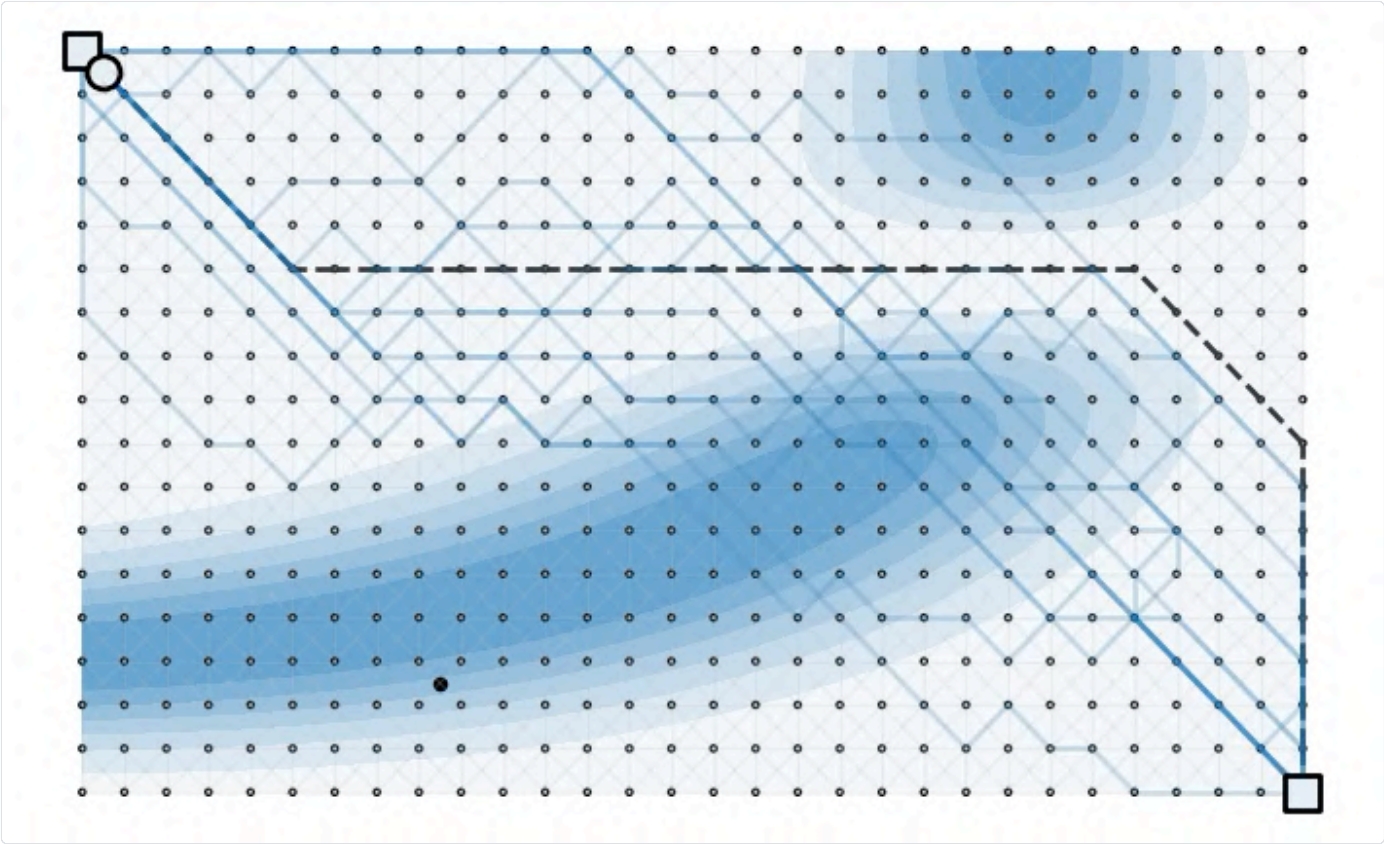
Data: usually viewed as fixed

# Bayesian Optimization



*Automatic explore-exploit tradeoff*

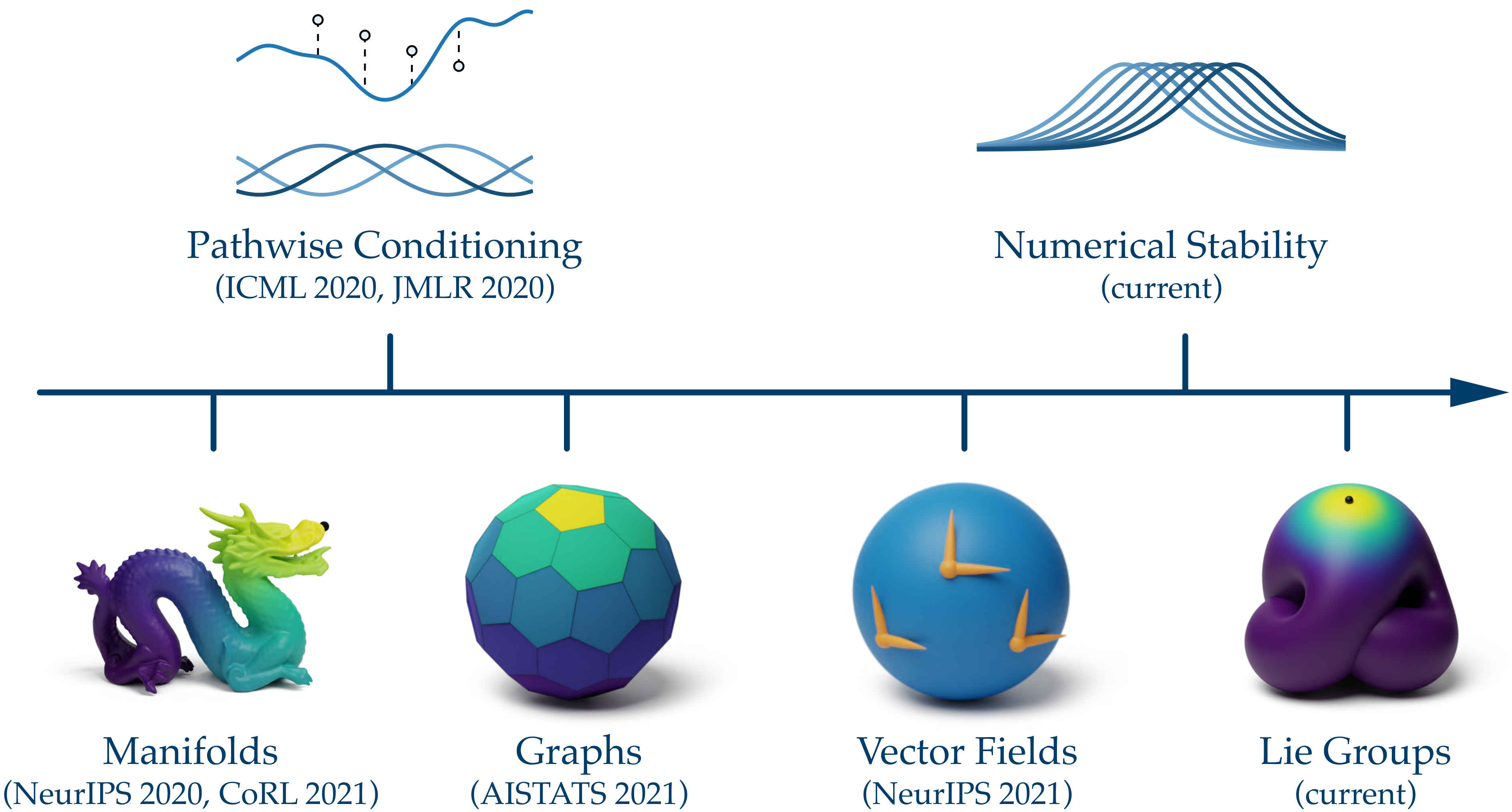
# From Bayesian Optimization to Bayesian Interactive Decision-making



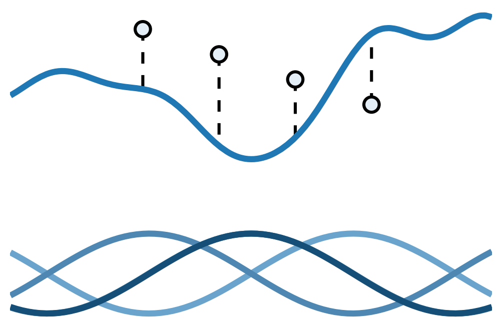
Up next: technical fundamentals
which make this possible

Neiswanger et al. (2020)

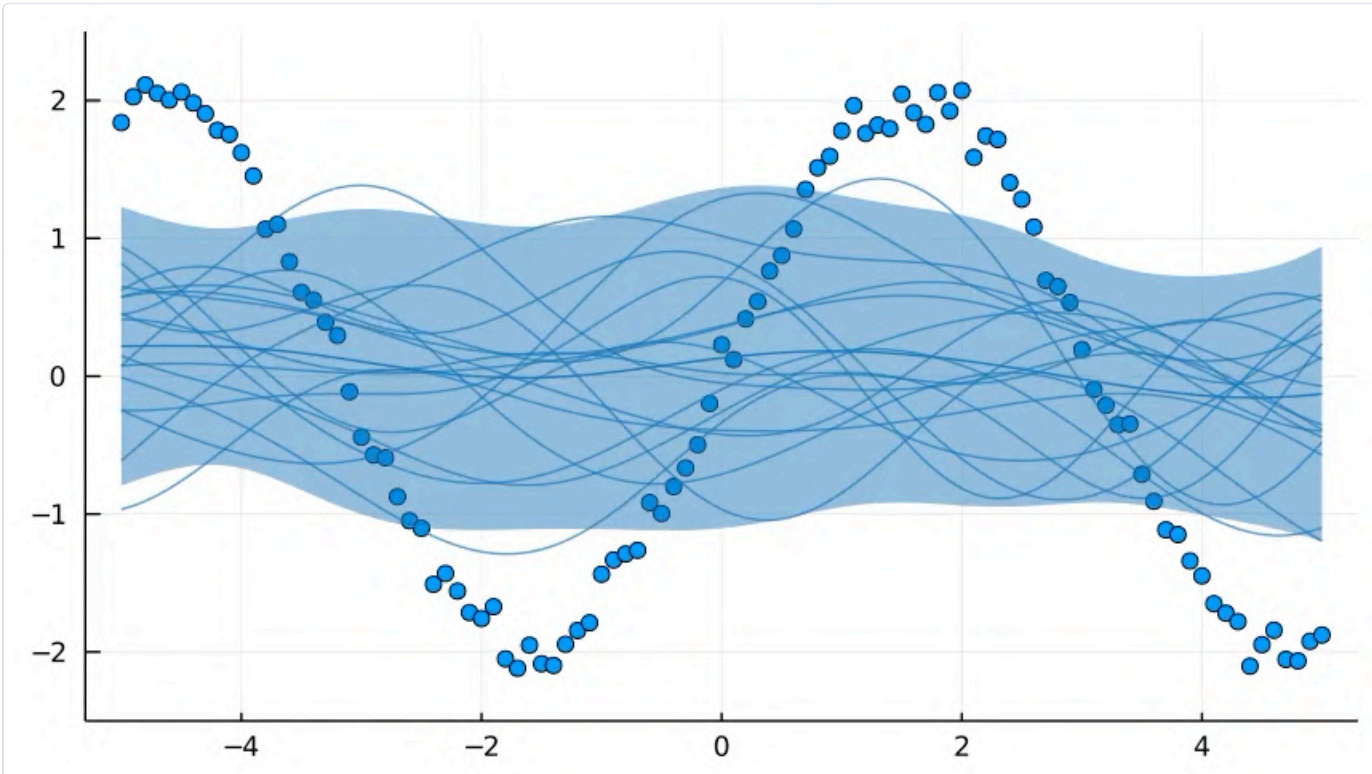
# Research Outline



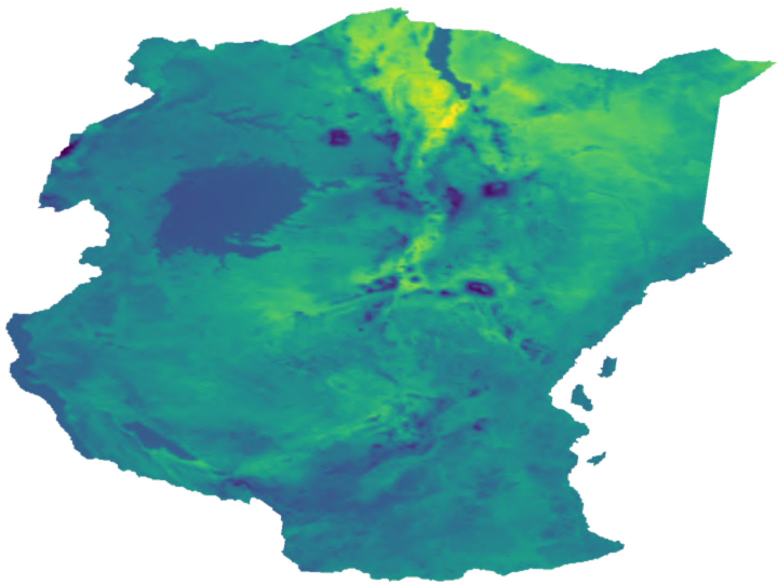
# Pathwise Conditioning of Gaussian Processes



# Gaussian Processes

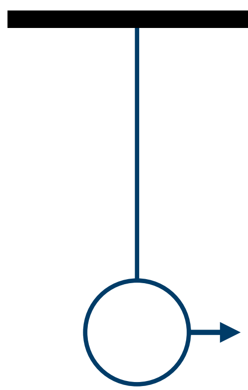


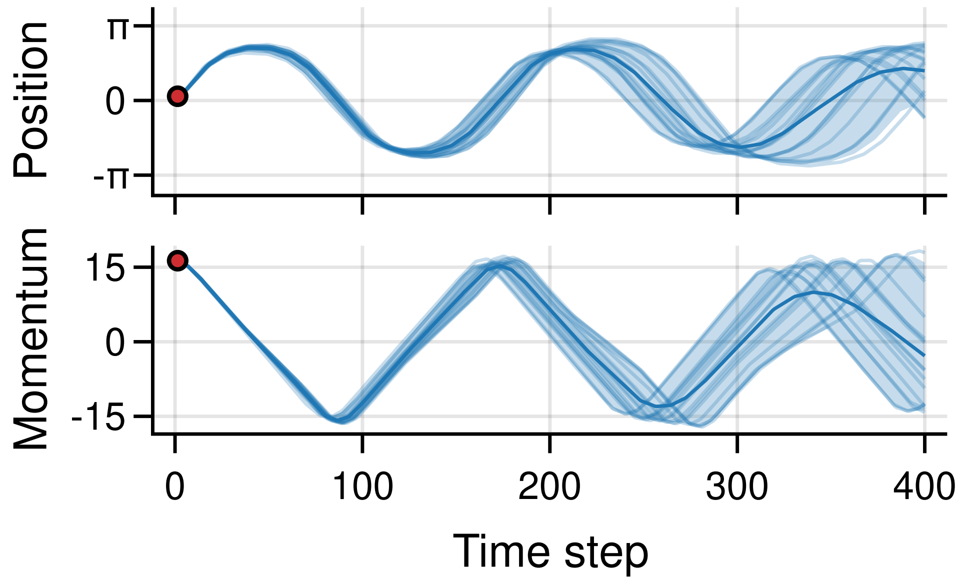
# Geospatial Learning



Large-scale probabilistic interpolation

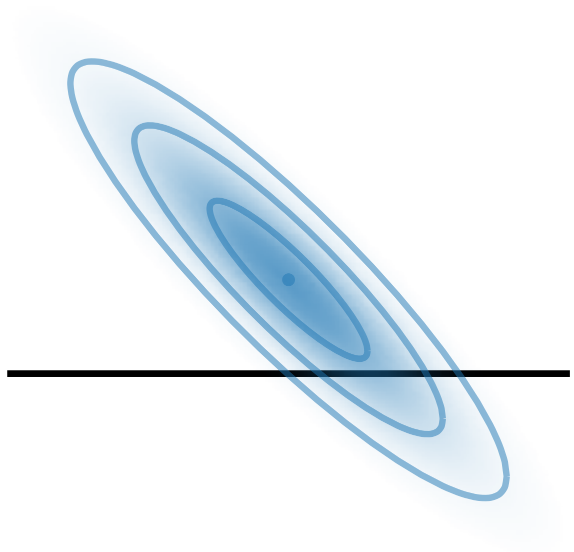
# Modeling Dynamical Systems with Uncertainty



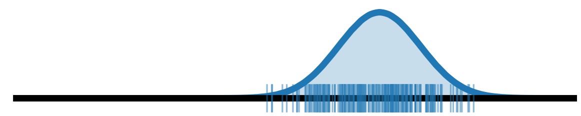



$$x_0 \qquad x_1 = x_0 + f(x_0)\Delta t \qquad x_2 = x_1 + f(x_1)\Delta t \qquad ..$$

# Conditioning of Multivariate Gaussians



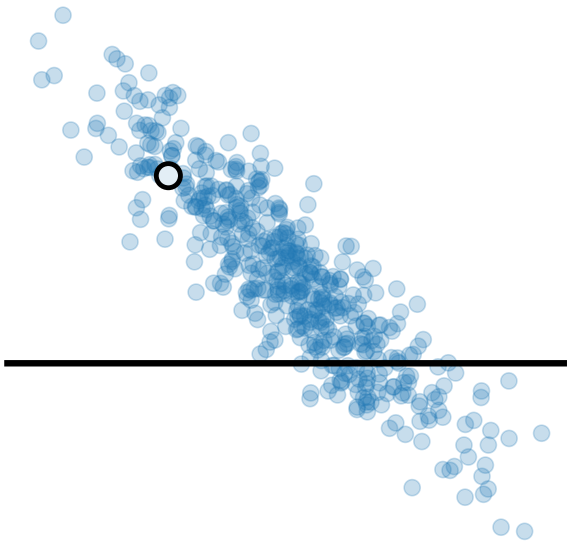
$$(\boldsymbol{\theta}, \boldsymbol{y}) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



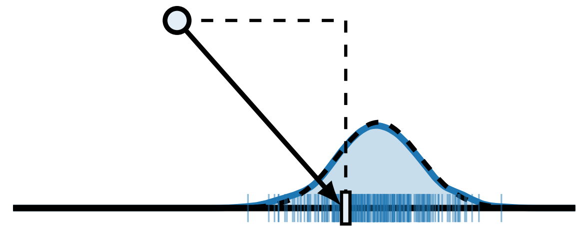
$$(\boldsymbol{\theta} \mid \boldsymbol{y} = \boldsymbol{\gamma}) \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}})$$

Calculate conditional, draw samples

# Conditioning of Multivariate Gaussians



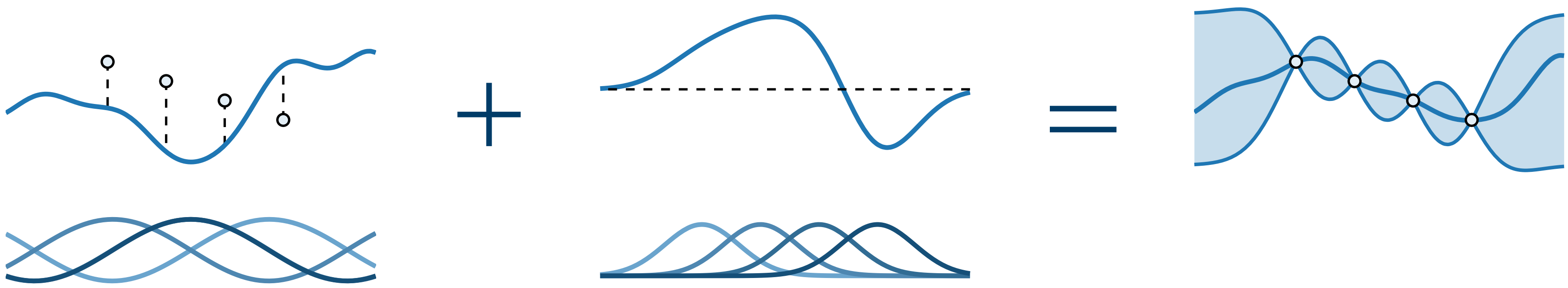
$$(\boldsymbol{\theta}, \boldsymbol{y}) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



$$(\boldsymbol{\theta} \mid \boldsymbol{y} = \boldsymbol{\gamma}) = \boldsymbol{\theta} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{y}} \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{y}}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{y}})$$

Draw samples, transform into conditional

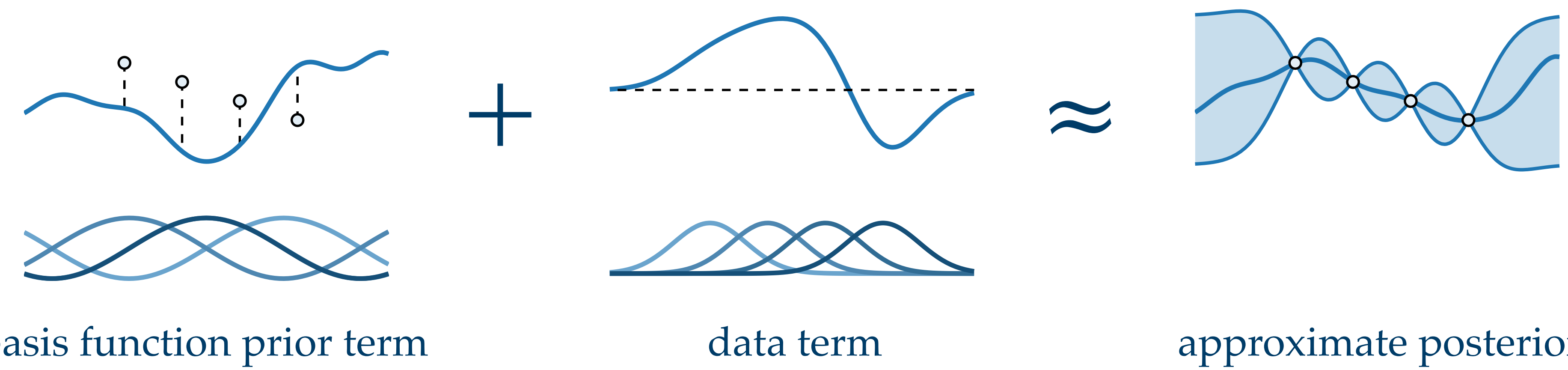
# Pathwise Conditioning



prior term | data term | posterior Gaussian process

$$(f \mid \boldsymbol{y})(\cdot) = \underbrace{f(\cdot)}_{\mathcal{O}(N_*^3)} + \mathbf{K}_{(\cdot)\boldsymbol{x}} \underbrace{\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}}^{-1}}_{\mathcal{O}(N^3)} (\boldsymbol{y} - f(\boldsymbol{x}))$$

# Efficient Sampling




$$(\boldsymbol{f} \mid \boldsymbol{y})(\cdot) \approx \underbrace{\sum_{i=1}^{L} w_i \phi_i(\cdot)}_{\text{basis function prior approx.}} + \sum_{j=1}^{N} v_j \boldsymbol{k}(x_j, \cdot)$$

**$\mathcal{O}(N_*)$ complexity in num. test points**

## Acquisition Functions for Bayesian Optimization

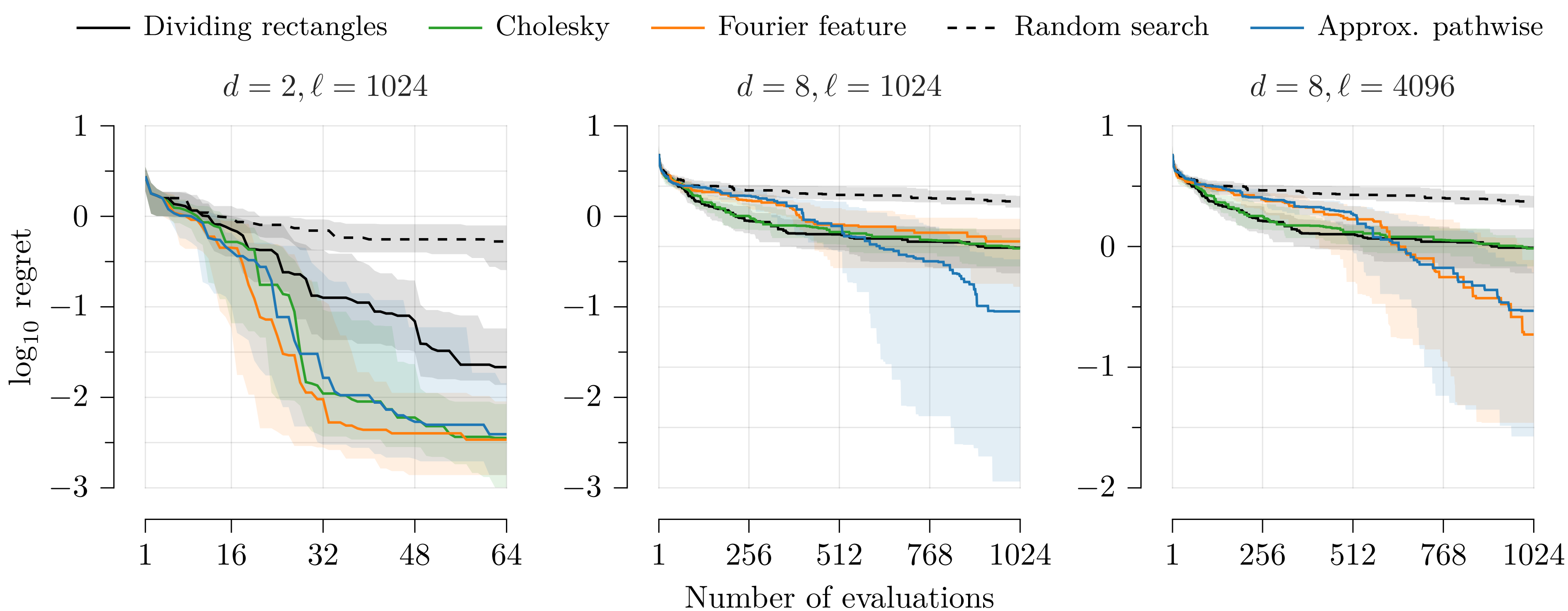
Thompson sampling: choose

$$x_{n+1} = \operatorname*{arg\,min}_{x \in \mathcal{X}} \alpha_n(x) \qquad \alpha_n \sim f \mid \boldsymbol{y}.$$

Pathwise conditioning $\rightsquigarrow$ easy to calculate $\alpha_n$

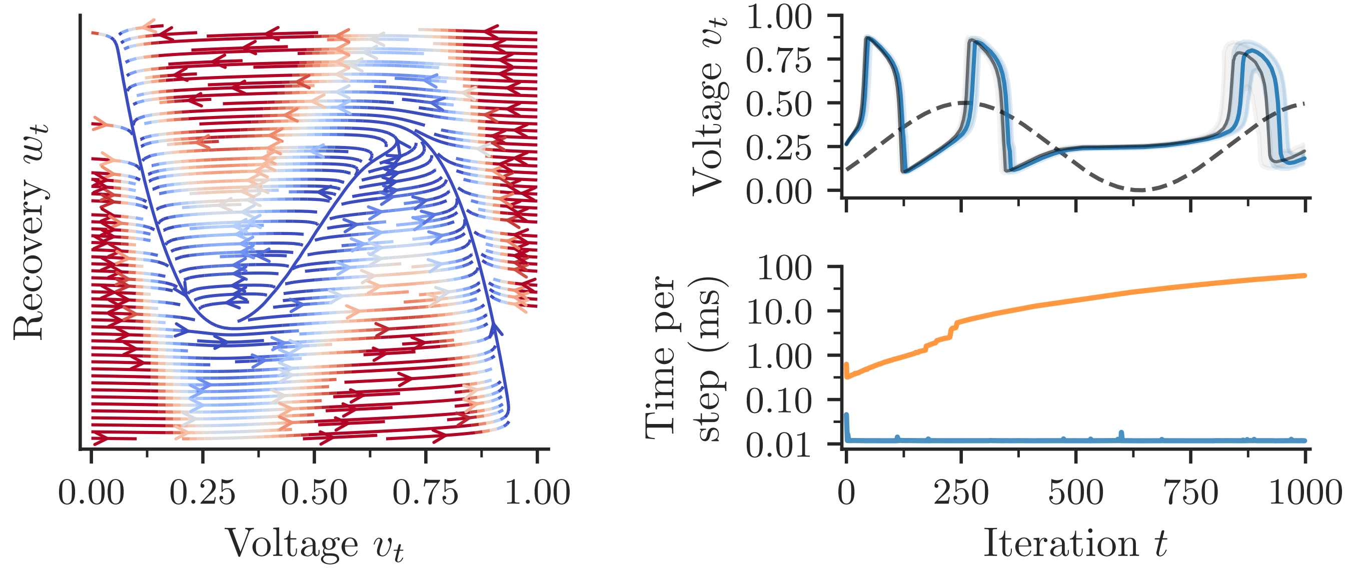
More generally: $\alpha_n(x) = \mathbb{E}_{\psi \sim f \mid \boldsymbol{y}}(A_\psi(x))$

# Parallel Bayesian Optimization: Thompson Sampling



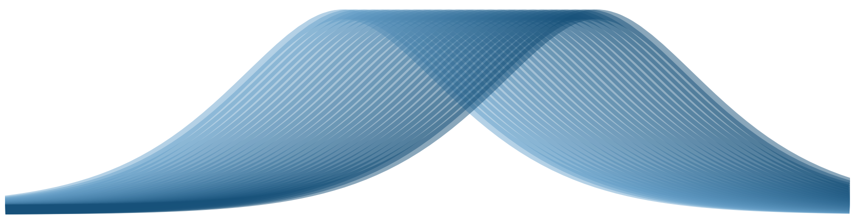

Better regret curves owing to reduced approximation error

# Learning Dynamical Systems




Accurate and efficient $\mathcal{O}(N_*)$ propagation of uncertainty

# Sparse Gaussian Processes



$$(f \mid \boldsymbol{y})(\cdot) = f(\cdot) + \sum_{i=1}^{N} v_i^{(\boldsymbol{y})} k(x_i, \cdot)$$

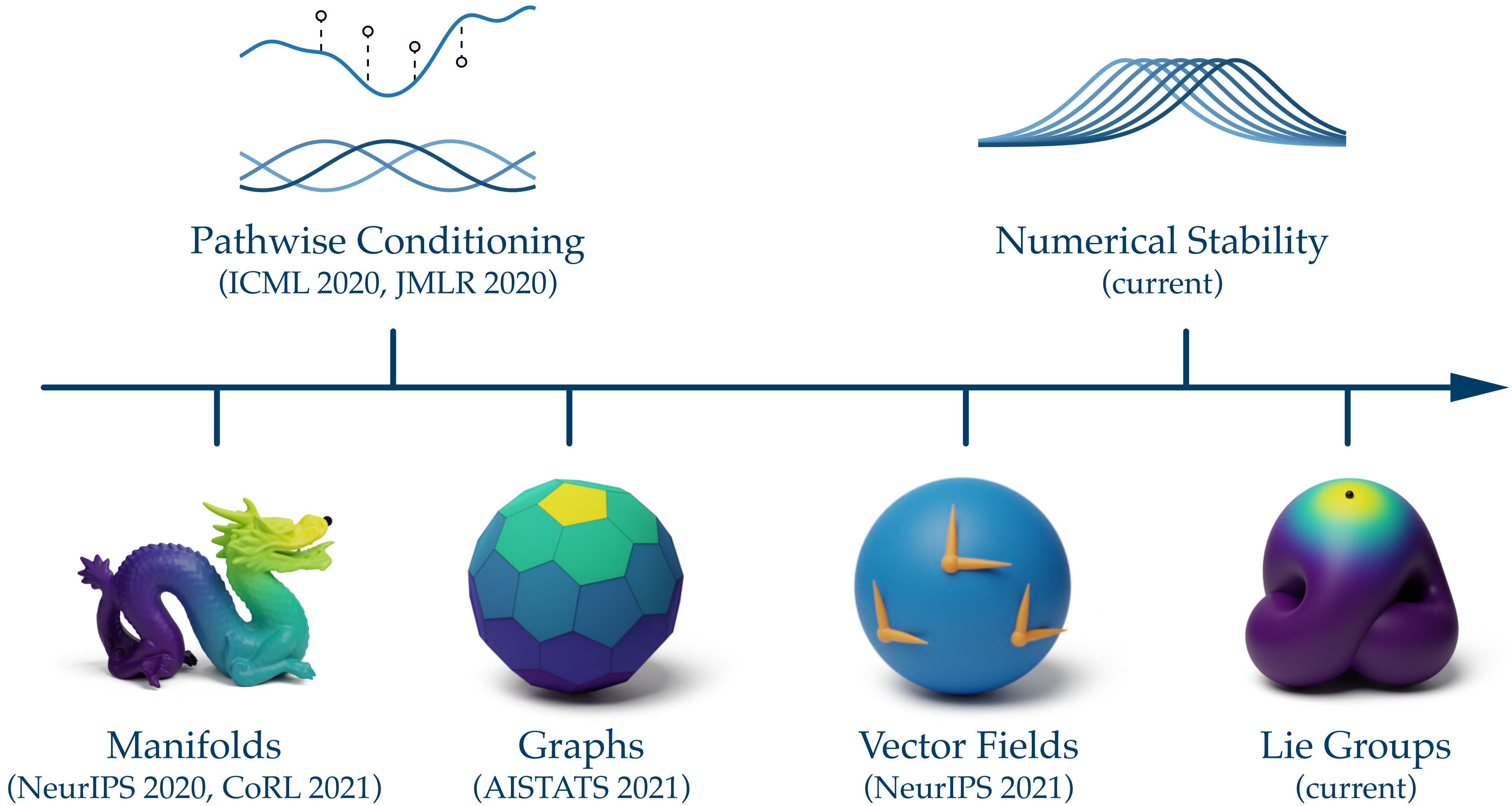
$$(f \mid \boldsymbol{y})(\cdot) \approx f(\cdot) + \sum_{j=1}^{M} v_j^{(\boldsymbol{u})} k(z_j, \cdot)$$

Exact Gaussian process: $\mathcal{O}(N^3)$

Sparse Gaussian process: $\mathcal{O}(NM^2)$

Works well under *data-redundancy*

# Research Outline



# Numerical Stability of Gaussian Processes

# Numerical Stability

Condition number: quantifies difficulty of solving $\mathbf{A}^{-1}\boldsymbol{b}$

$$\operatorname{cond}(\mathbf{A}) = \lim_{\varepsilon \to 0} \sup_{\|\boldsymbol{\delta}\| \leq \varepsilon \|\boldsymbol{b}\|} \frac{\left\|\mathbf{A}^{-1}(\boldsymbol{b}+\boldsymbol{\delta}) - \mathbf{A}^{-1}\boldsymbol{b}\right\|_2}{\varepsilon \left\|\mathbf{A}^{-1}\boldsymbol{b}\right\|_2} = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

$\lambda_{\min}, \lambda_{\max}$: eigenvalues

# Cholesky Factorization

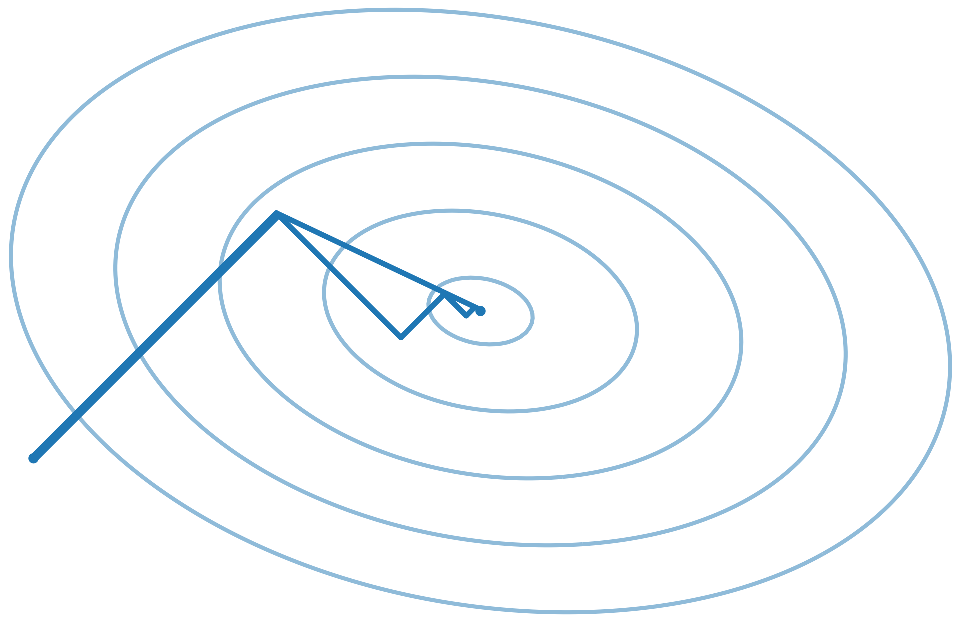
**Result.** Let $\mathbf{A}$ be a symmetric positive definite matrix of size $N \times N$, where $N > 10$. Assume that

$$\text{cond}(\mathbf{A}) \leq \frac{1}{2^{-t} \times 3.9N^{3/2}}$$

where $t$ is the length of the floating point mantissa, and that $3N2^{-t} < 0.1$. Then floating point Cholesky factorization will succeed, producing a matrix $\mathbf{L}$ satisfying

$$\mathbf{L}\mathbf{L}^T = \mathbf{A} + \mathbf{E} \qquad\qquad \|\mathbf{E}\|_2 \leq 2^{-t} \times 1.38N^{3/2} \|\mathbf{A}\|_2 .$$

# Conjugate Gradients



Refinement of gradient descent for solving linear systems $\mathbf{A}^{-1}\boldsymbol{b}$

Convergence rate depends on $\mathbf{cond}(\mathbf{A})$

# Condition Numbers of Kernel Matrices

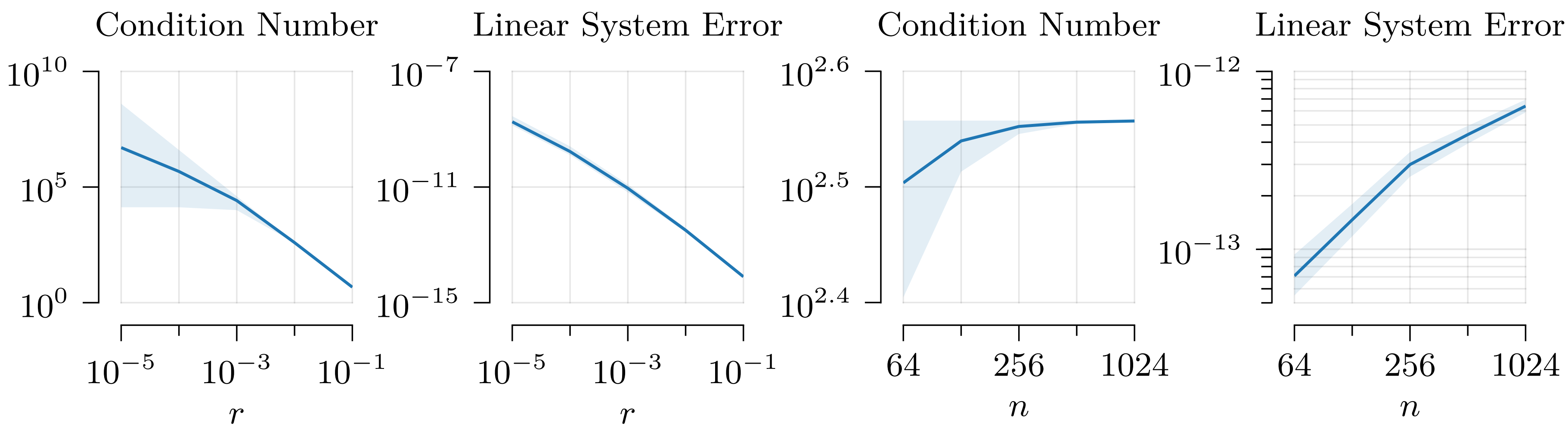
Are kernel matrices always well-conditioned? **No.**

One-dimensional time series on grid: Kac–Murdock–Szegö matrix

$$\mathbf{K}_{\boldsymbol{xx}} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}$$

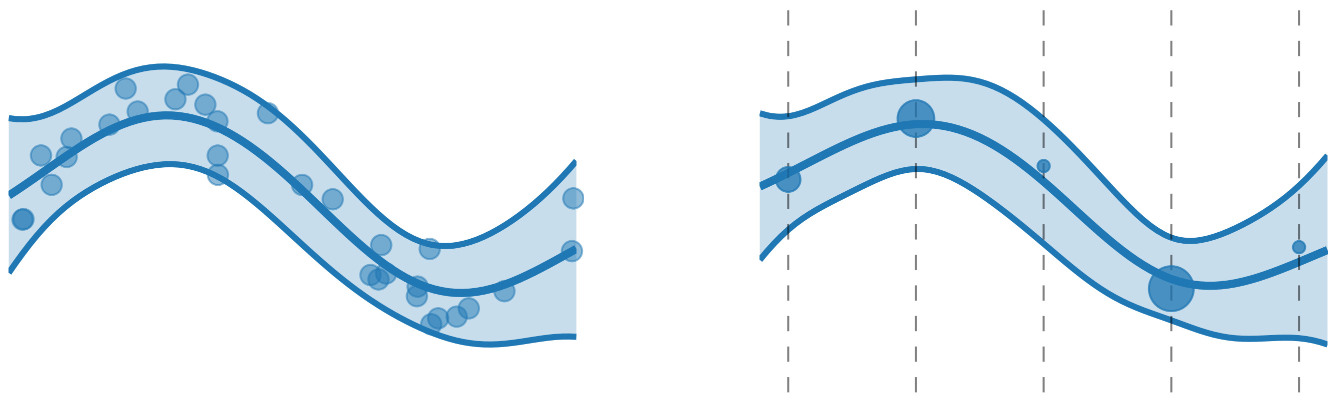
for which $\frac{1+2\rho+2\rho\varepsilon+\rho^2}{1-2\rho-2\rho\varepsilon+\rho^2} \leq \operatorname{cond}(\mathbf{K}_{\boldsymbol{xx}}) \leq \frac{(1+\rho)^2}{(1-\rho)^2}$, where $\varepsilon = \frac{\pi^2}{N+1}$.

# Condition Numbers of Kernel Matrices




Problem: too much correlation ⇝ *points too close by*

# Minimum Separation

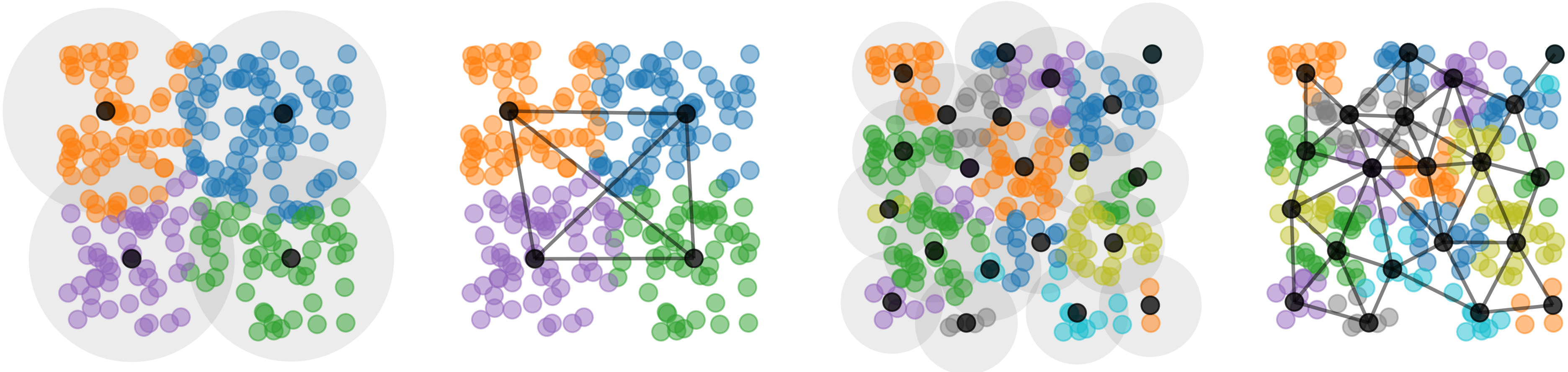


*Separation*: minimum distance between distinct $z_i$ and $z_j$

**Proposition.** Assuming spatial decay and stationarity, separation controls $\mathrm{cond}(\mathbf{K}_{zz})$ uniformly in $M$.

Idea: use this to select numerically stable inducing points

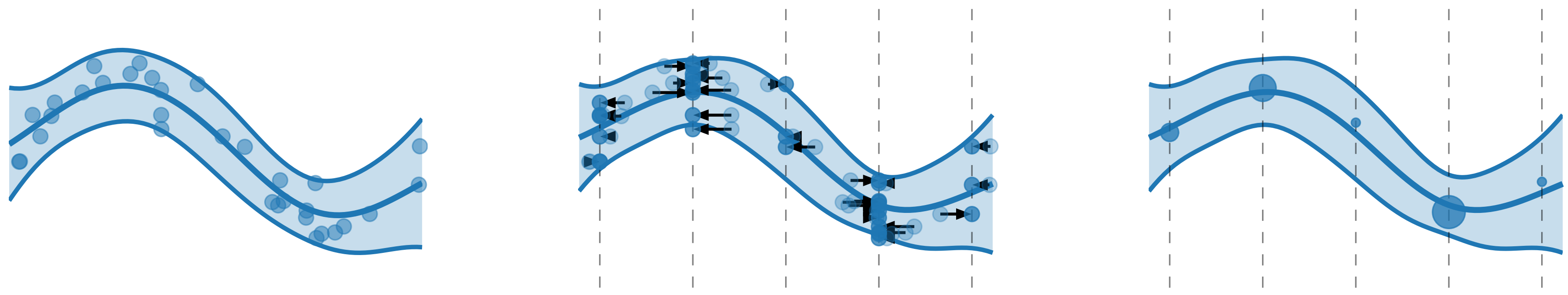
# Inducing Point Selection via Cover Trees



*Spatial resolution*: maximum distance from $x_i$ to nearest $z_j$

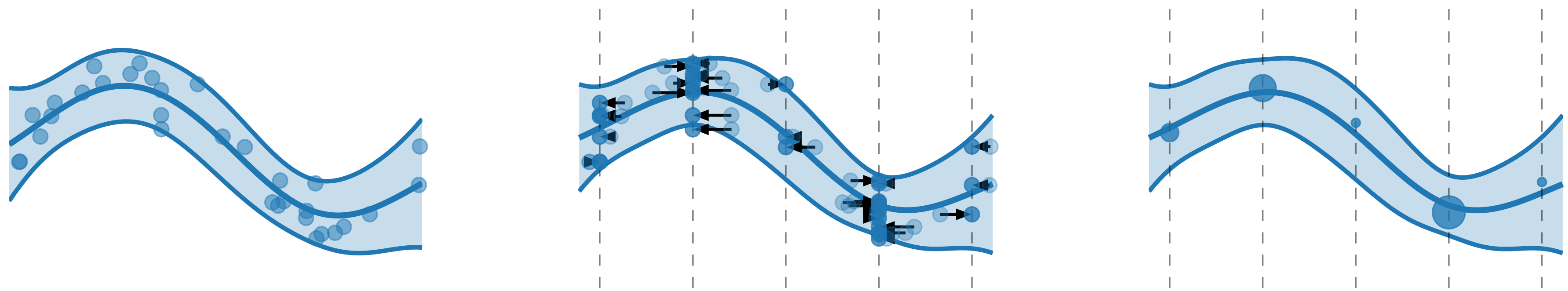
*Cover tree algorithm*: guarantee spatial resolution and separation

# The Clustered-data Approximation



Replace large dataset with re-weighted sparser dataset

# The Clustered-data Approximation

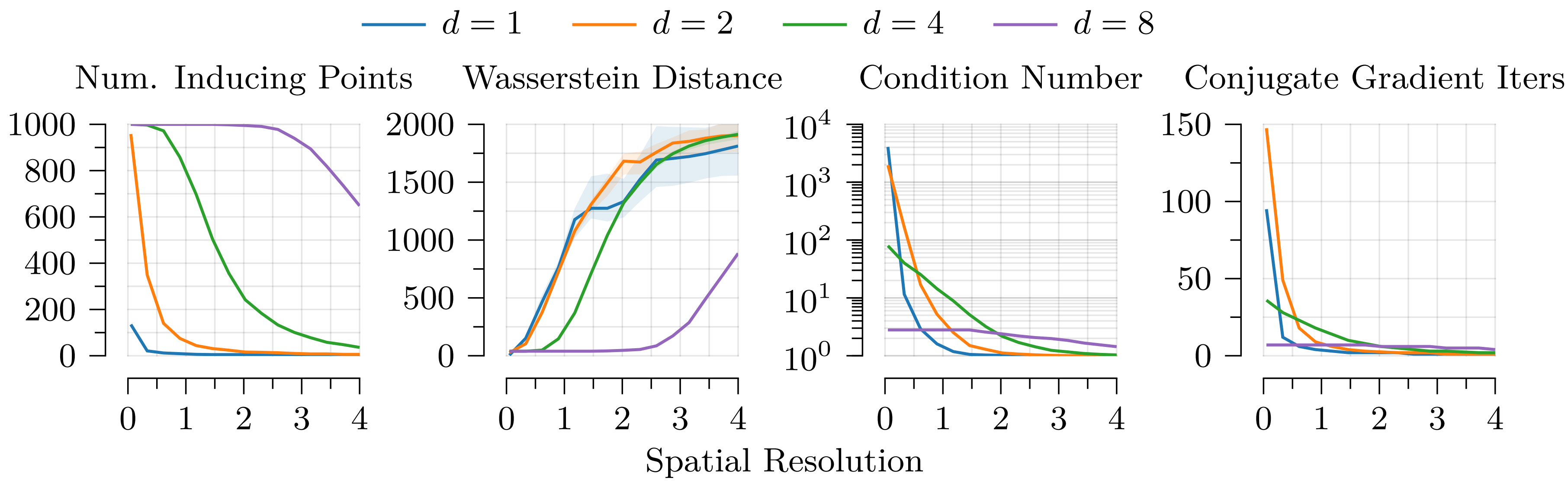


$$(f \mid \boldsymbol{y})(\cdot) = f(\cdot) + \mathbf{K}_{(\cdot)\boldsymbol{z}}(\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} + \boldsymbol{\Lambda})^{-1}(\boldsymbol{u} - f(\boldsymbol{z}) - \boldsymbol{\epsilon}) \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Lambda})$$

Minimum eigenvalue: $\lambda_{\min}(\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} + \boldsymbol{\Lambda}) \geq \min_{i=1,..,M} \Lambda_{ii}$

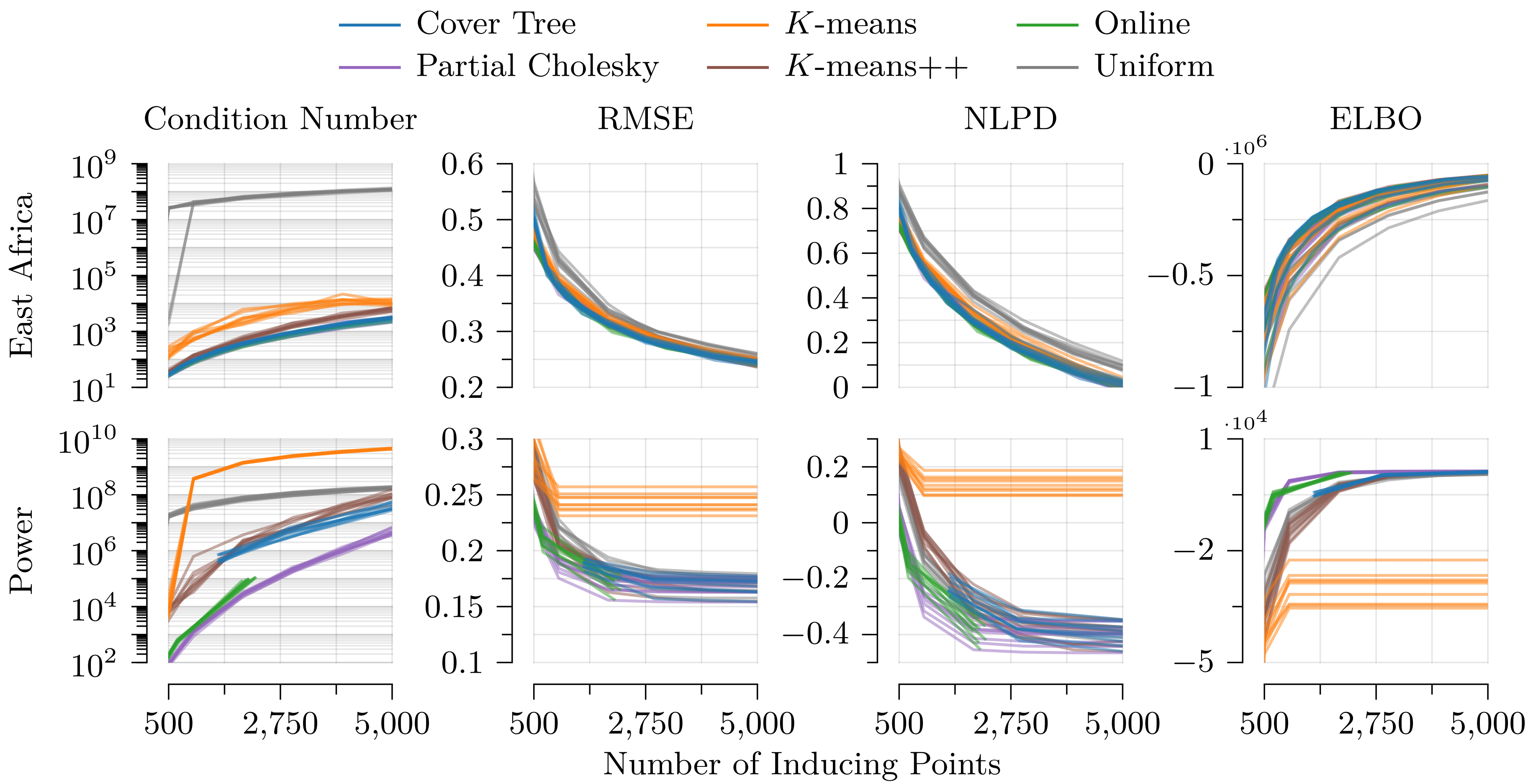
Clustered-data approximation results in *extra stability*

# Empirical Accuracy and Computational Cost



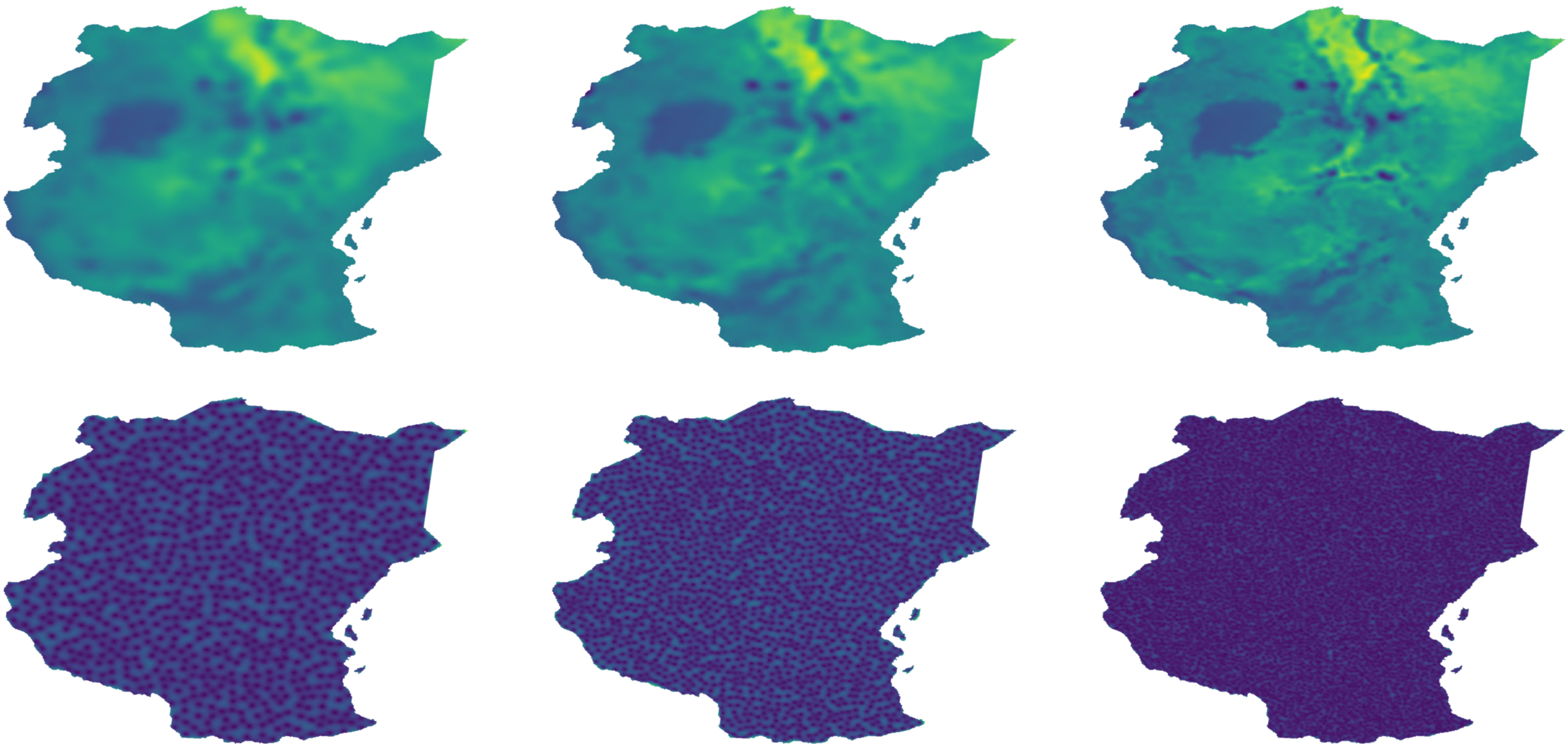

Spatial resolution directly controls error and computational cost

# Inducing Point Selection Comparison



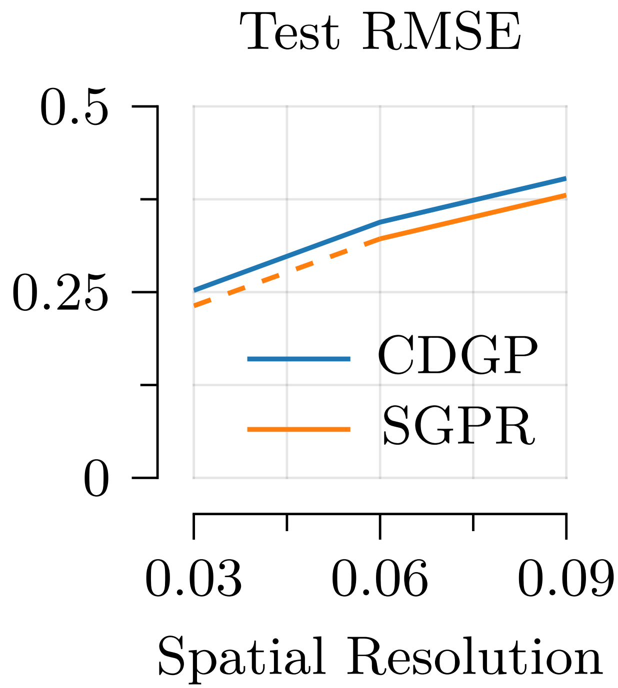

Good performance-stability tradeoff in geospatial settings

# Geospatial Learning



$\varepsilon = 0.09, M = 902$ $\varepsilon = 0.06, M = 1934$ $\varepsilon = 0.03, M = 6851$

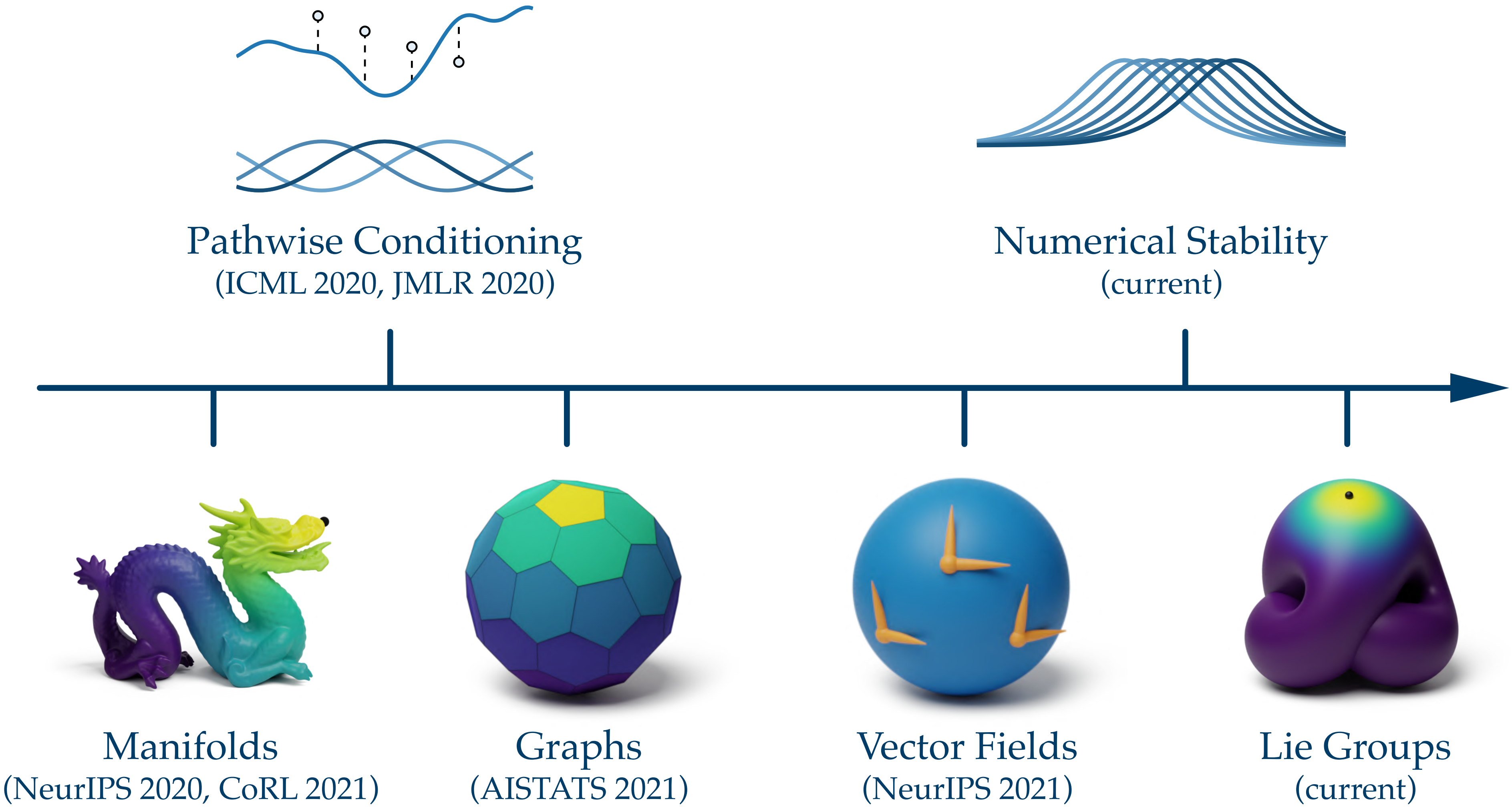
# Approximation Error and Numerical Stability



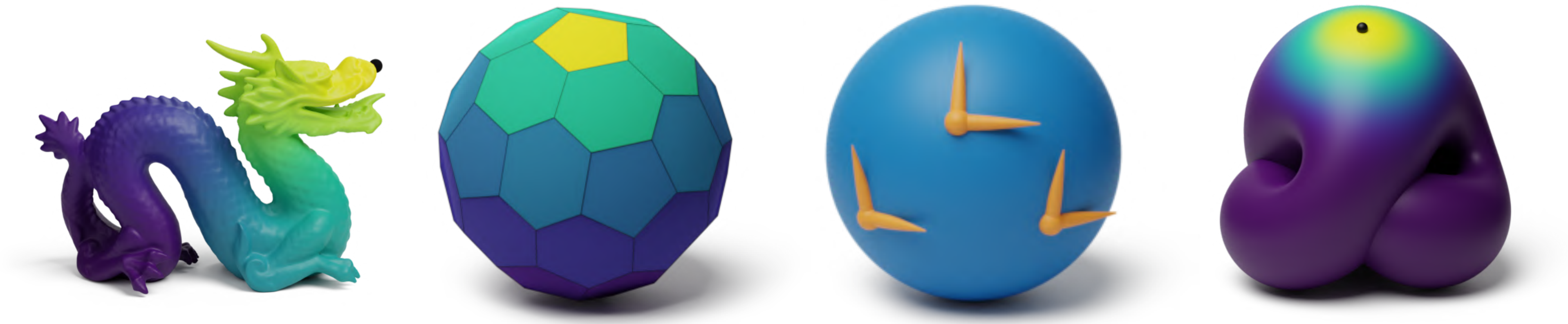

Dashed line: increased jitter due to Cholesky failure in floating point

Slightly lower approximation accuracy, but better stability

# Research Outline

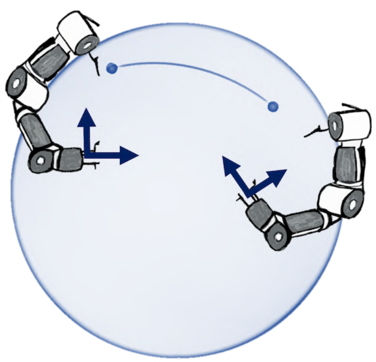


# Non-Euclidean Gaussian Processes

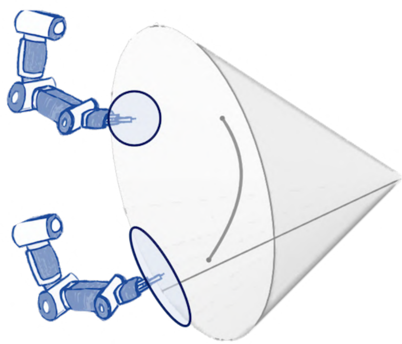


AISTATS 2021 Best Student Paper

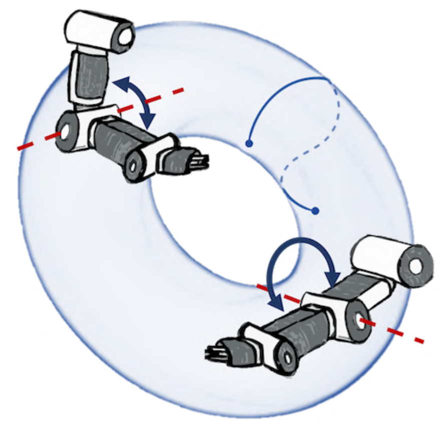
# Bayesian Optimization in Robotics



Orientation: sphere



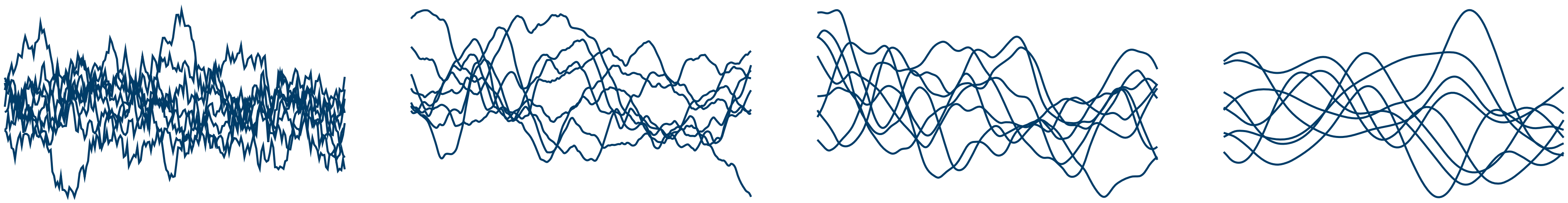
Manipulability:
SPD manifold



Joint postures: torus

Jaquier et al. (2020)

# Matérn Gaussian Processes



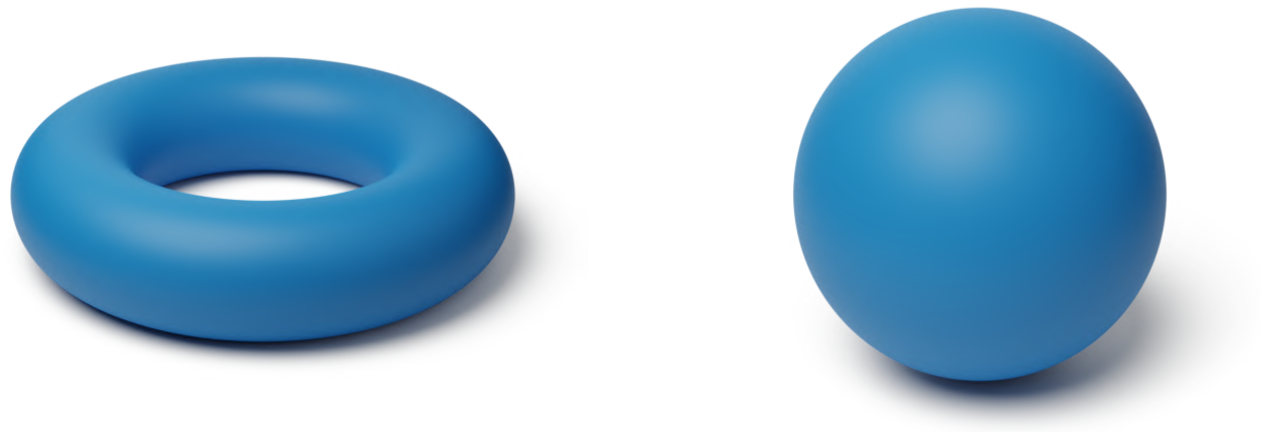
$\nu = 1/2$ $\quad$ $\nu = 3/2$ $\quad$ $\nu = 5/2$ $\quad$ $\nu = \infty$

$$k_\nu(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \sqrt{2\nu} \frac{\|x - x'\|}{\kappa} \right)^\nu K_\nu \left( \sqrt{2\nu} \frac{\|x - x'\|}{\kappa} \right)$$

$\sigma^2$: variance $\quad$ $\kappa$: length scale $\quad$ $\nu$: smoothness

$\nu \to \infty$: recovers squared exponential kernel

# Riemannian Geometry



How should Matérn kernels generalize to this setting?

# Geodesics

$$k_{\infty}^{(d_g)}(x, x') = \sigma^2 \exp\left(-\frac{d_g(x, x')^2}{2\kappa^2}\right)$$

**Theorem.** (Feragen et al.) Let $M$ be a complete Riemannian manifold without boundary. If $k_{\infty}^{(d_g)}$ is positive semi-definite for all $\kappa$, then $M$ is isometric to a Euclidean space.

Need a different candidate generalization

Feragen et al. (2015)

# Stochastic Partial Differential Equations

$$\underbrace{\left(\frac{2\nu}{\kappa^2} - \Delta\right)^{\frac{\nu}{2}+\frac{d}{4}} f = \mathcal{W}}_{\text{Matérn}} \qquad \underbrace{e^{-\frac{\kappa^2}{4}\Delta} f = \mathcal{W}}_{\text{squared exponential}}$$

$\Delta$: Laplacian $\quad \mathcal{W}$: (rescaled) white noise

$e^{-\frac{\kappa^2}{4}\Delta}$: (rescaled) heat semigroup

Generalizes well to the Riemannian setting

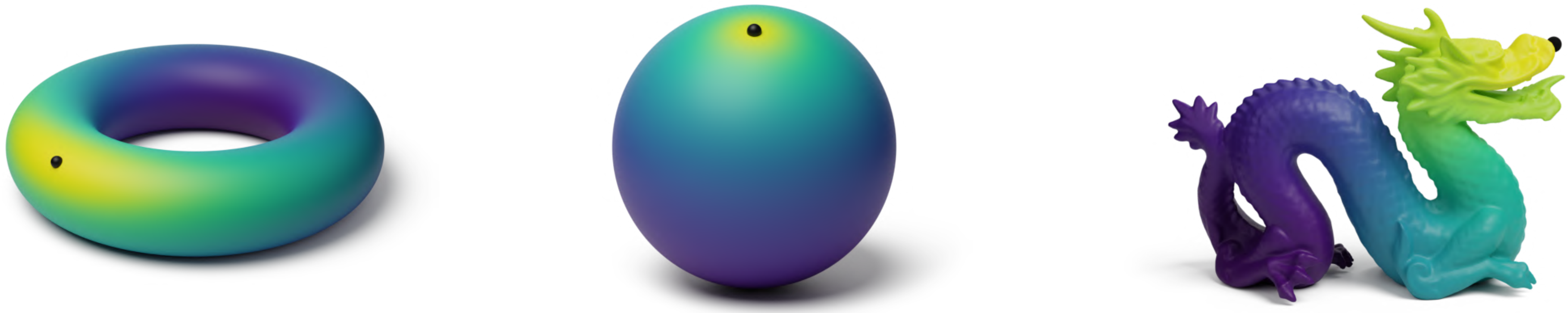
Whittle (1963)
Lindgren et al. (2011)

# Riemannian Matérn Kernels: compact spaces

$$k_\nu(x, x') = \frac{\sigma^2}{C_\nu} \sum_{n=0}^{\infty} \left( \frac{2\nu}{\kappa^2} - \lambda_n \right)^{\nu - \frac{d}{2}} f_n(x) f_n(x')$$

$\lambda_n$, $f_n$: Laplace–Beltrami eigenpairs
Analytic expressions for circle, sphere, ..

# Riemannian Matérn Kernels

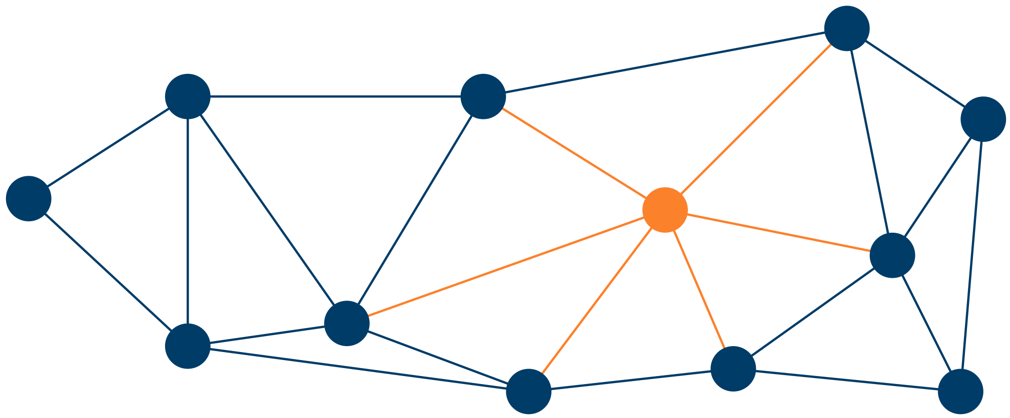


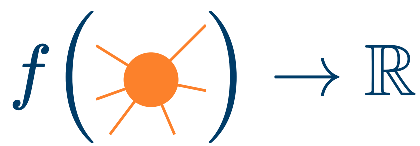
$$k_{1/2}(\bullet, \cdot)$$

# Example: regression on the surface of a dragon

(a) Ground truth (b) Posterior mean (c) Std. deviation (d) Posterior sample

# Weighted Undirected Graphs





# The Graph Laplacian

$$(\boldsymbol{\Delta} \boldsymbol{f})(x) = \sum_{x' \sim x} w_{xx'}(f(x) - f(x'))$$

$$\boldsymbol{\Delta} = \mathbf{D} - \mathbf{W}$$

$\mathbf{D}$: degree matrix    $\mathbf{W}$: (weighted) adjacency matrix

Note: different sign convention, analogous to Euclidean $-\Delta$

# Graph Matérn Gaussian Processes

$$\underbrace{\left(\frac{2\nu}{\kappa^2} + \boldsymbol{\Delta}\right)^{\frac{\nu}{2}} \boldsymbol{f} = \boldsymbol{\mathcal{W}}}_{\text{Matérn}} \qquad \underbrace{e^{\frac{\kappa^2}{4}\boldsymbol{\Delta}} \boldsymbol{f} = \boldsymbol{\mathcal{W}}}_{\text{squared exponential}}$$

$\boldsymbol{\Delta}$: graph Laplacian $\quad \boldsymbol{\mathcal{W}}$: standard Gaussian

# Graph Matérn Gaussian Processes

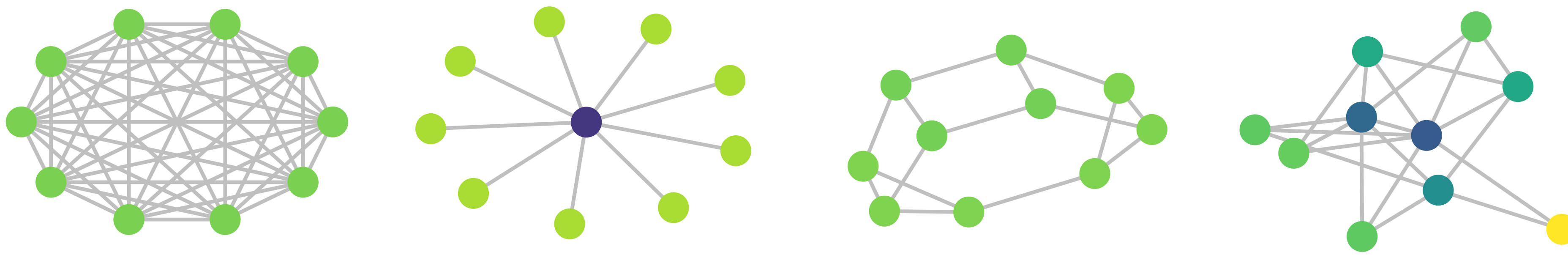
$$\underbrace{\boldsymbol{f} \sim \mathrm{N}\left(\mathbf{0}, \left(\frac{2\nu}{\kappa^2} + \boldsymbol{\Delta}\right)^{-\nu}\right)}_{\text{Matérn}} \qquad \underbrace{\boldsymbol{f} \sim \mathrm{N}\left(\mathbf{0}, e^{-\frac{\kappa^2}{2}\boldsymbol{\Delta}}\right)}_{\text{squared exponential}}$$

# Graph Fourier Features

$$k_\nu(x, x') = \frac{\sigma^2}{C_\nu} \sum_{n=1}^{|G|} \left( \frac{2\nu}{\kappa^2} + \lambda_n \right)^{-\nu} \boldsymbol{f}_n(x) \boldsymbol{f}_n(x')$$

$\lambda_n, \boldsymbol{f}_n$: eigenvalues and eigenvectors of graph Laplacian

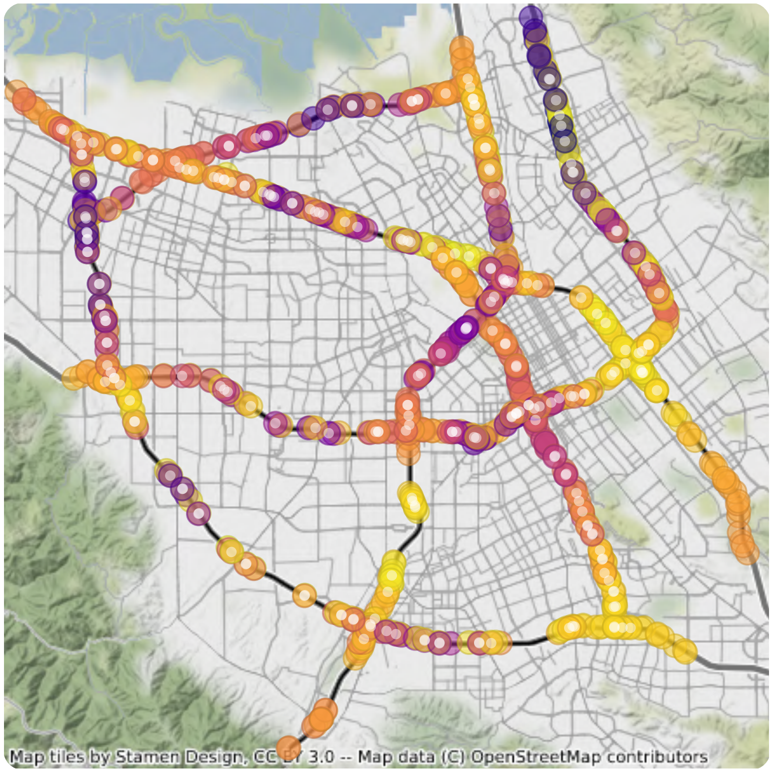
# Prior Variance



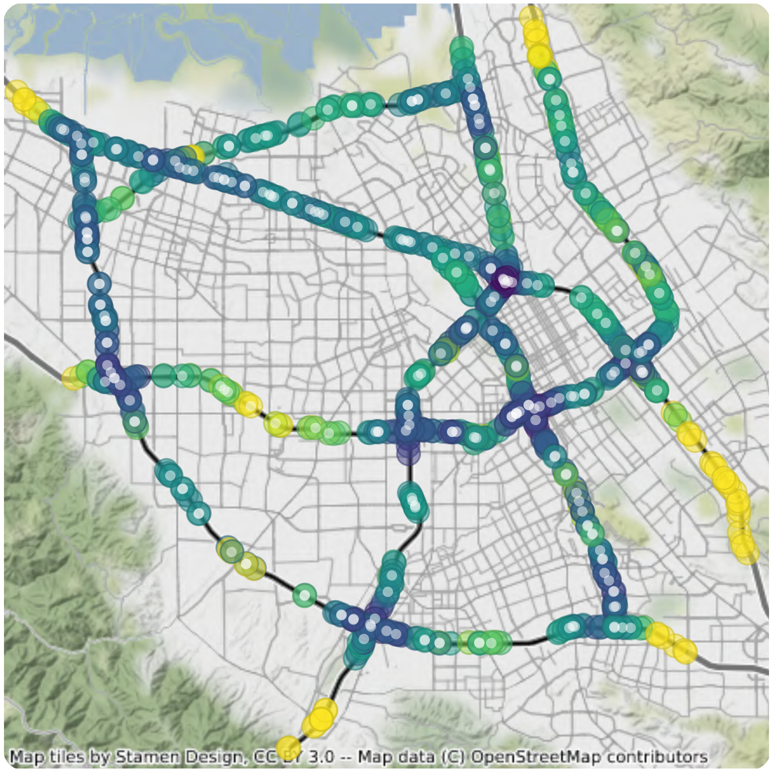
(a) Complete graph (b) Star graph (c) Random graph (d) Random graph

Prior variance depends on geometry

# Example: Graph Interpolation of Traffic

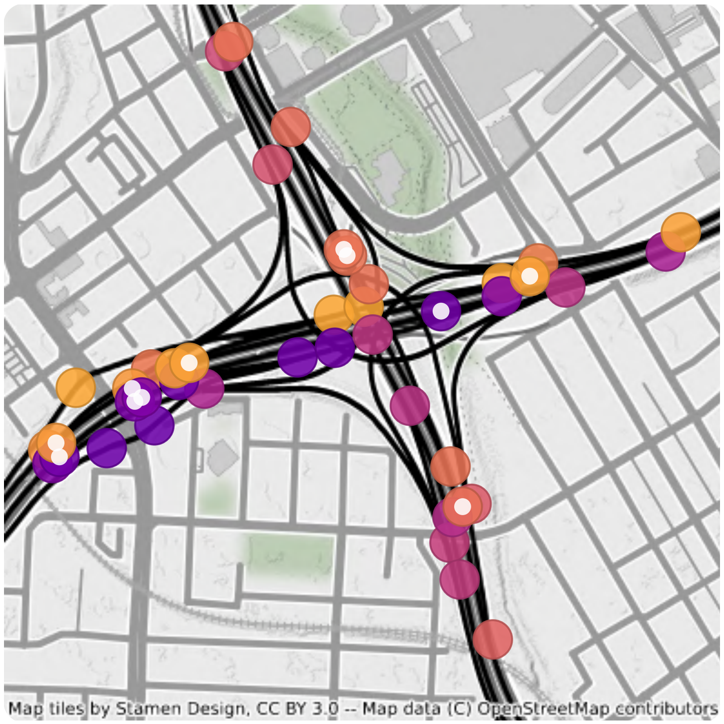



(a) Predictive mean

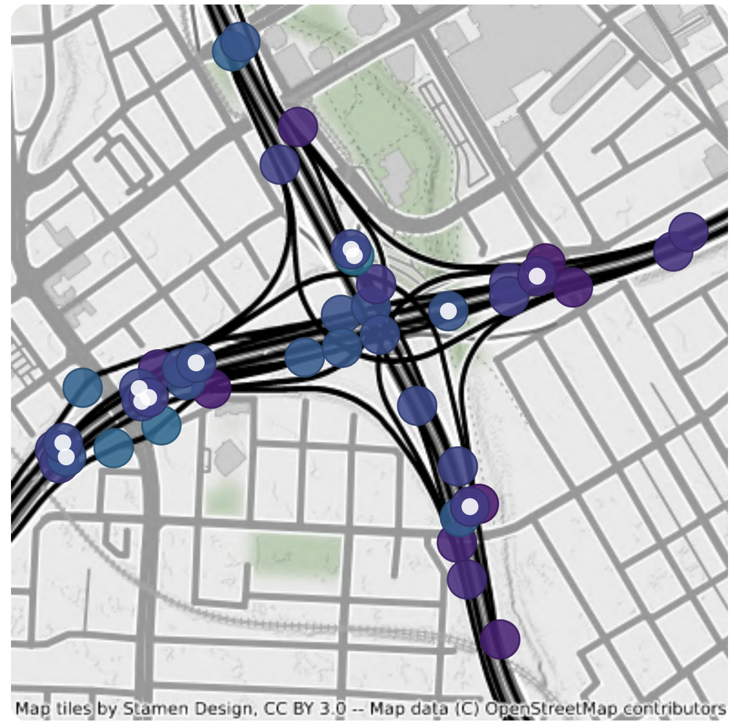



(b) Standard deviation

# Example: Graph Interpolation of Traffic

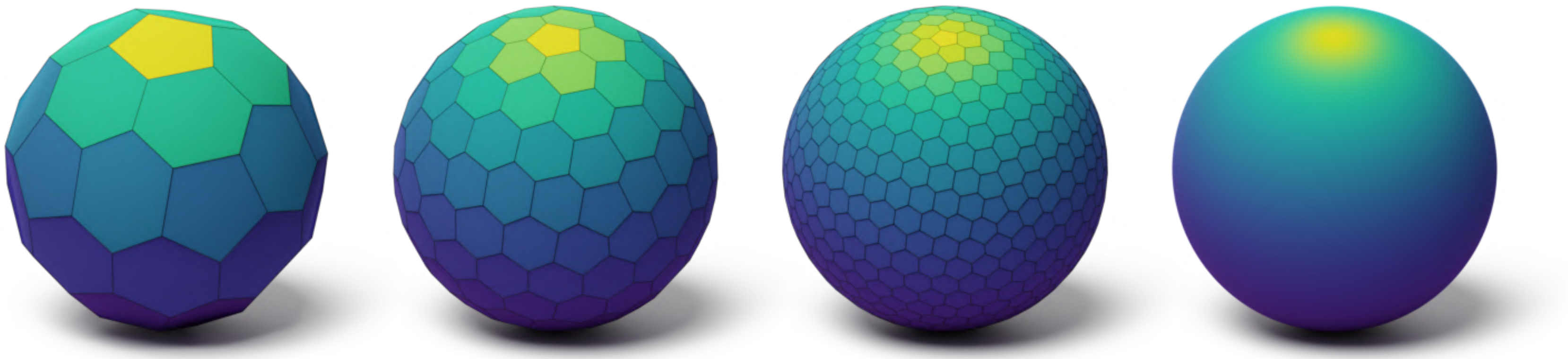



(a) Predictive mean

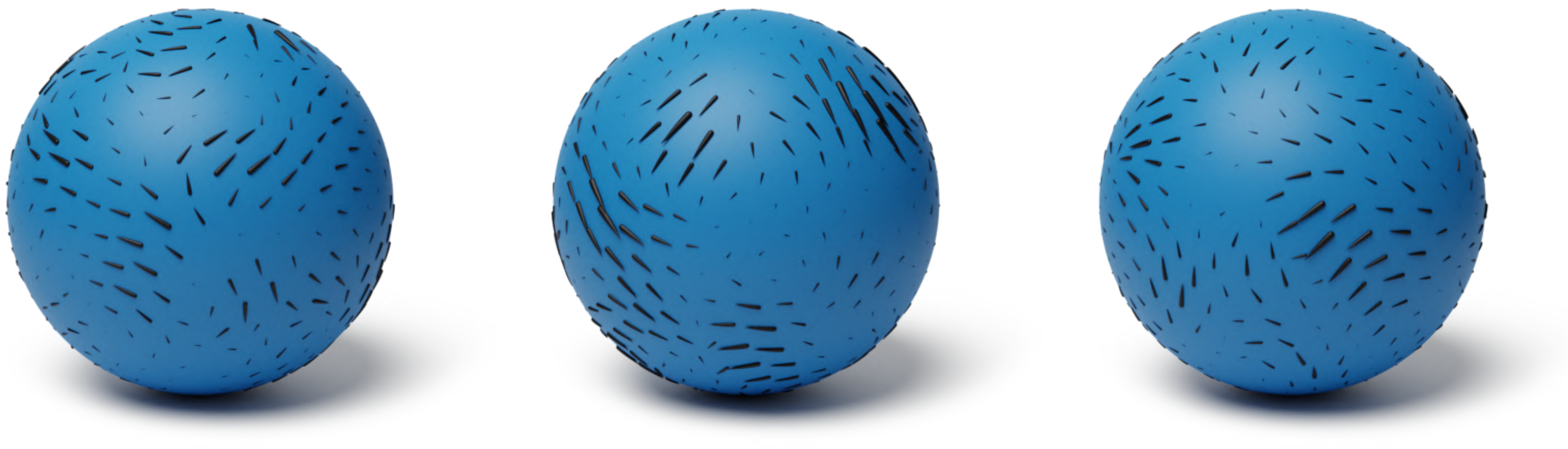



(b) Standard deviation

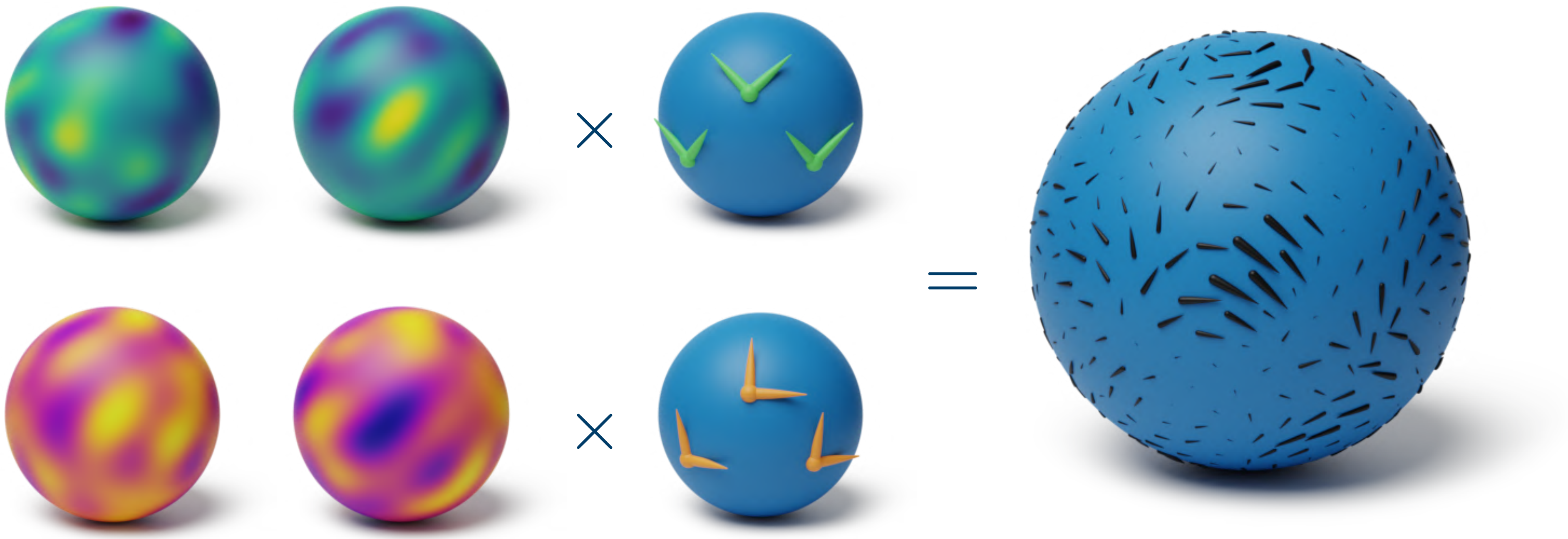
# Riemannian Limits



# Gaussian Vector Fields

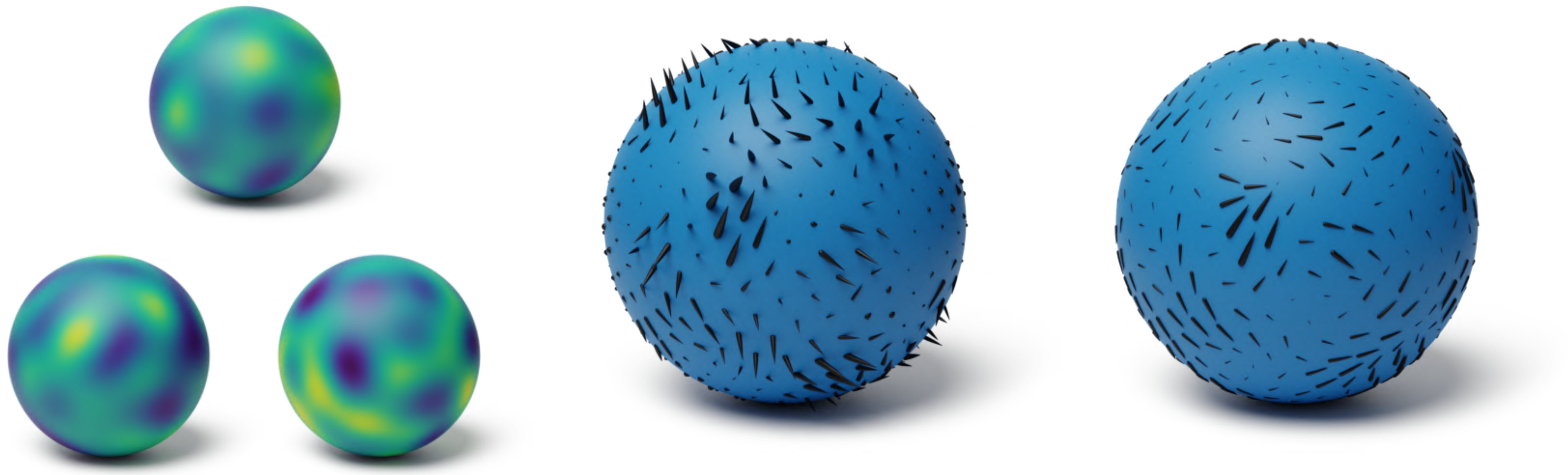


# Frames



Same vector field: represented differently in different frames

# Projected Kernels



Construct vector fields by embedding and flattening scalar fields

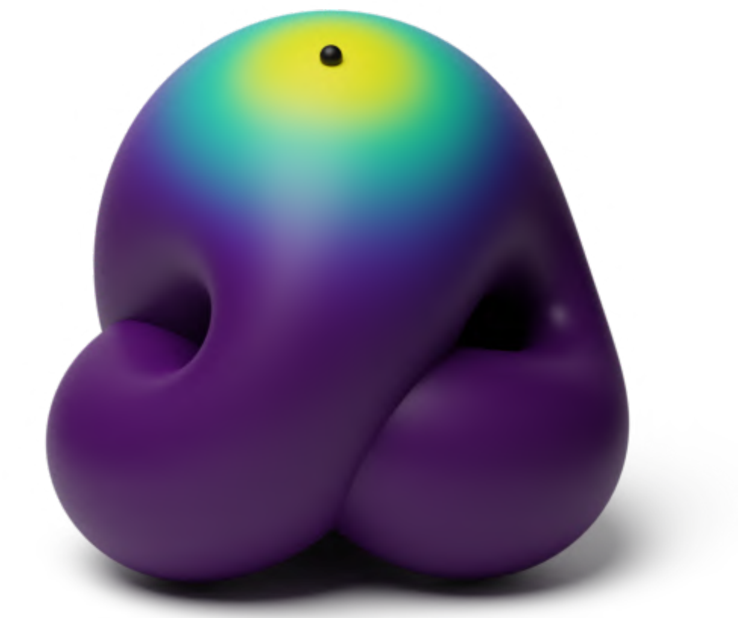
## Stationary kernels on Lie groups

$$k(g_1, g_2) = \sum_{n=1}^{\infty} a(\lambda_n) \boldsymbol{f}_n(g_1) \boldsymbol{f}_n(g_2)$$
$$= \sum_{\lambda \in \Lambda} a^{(\lambda)} \operatorname{Re} \chi^{(\lambda)}(g_2^{-1} \cdot g_1)$$

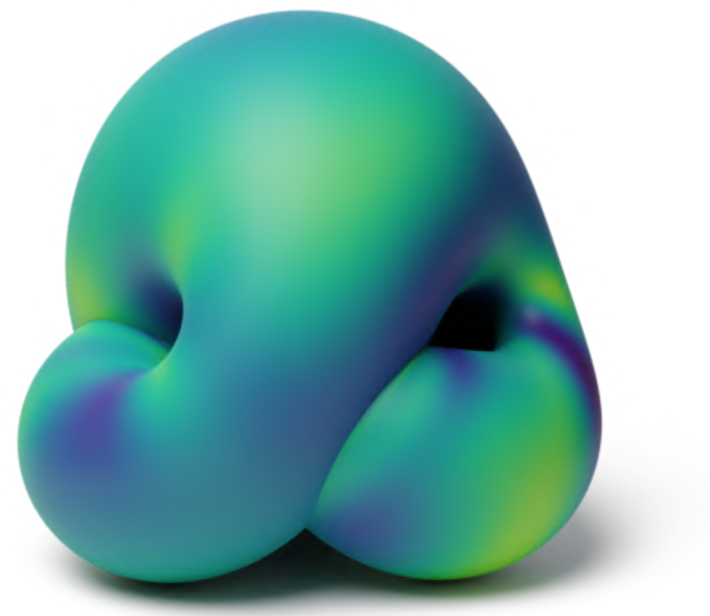
Computable numerically using representation-theoretic quantities

$\underbrace{\text{Sphere:}}_{\text{homogeneous space}}$ spherical harmonics $\rightsquigarrow$ Gegenbauer polynomials

# Example: real projective space

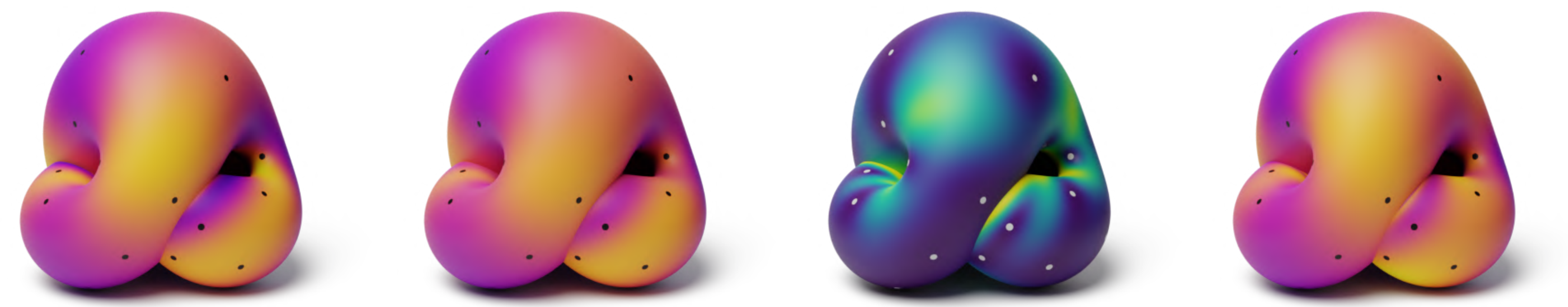


Kernel



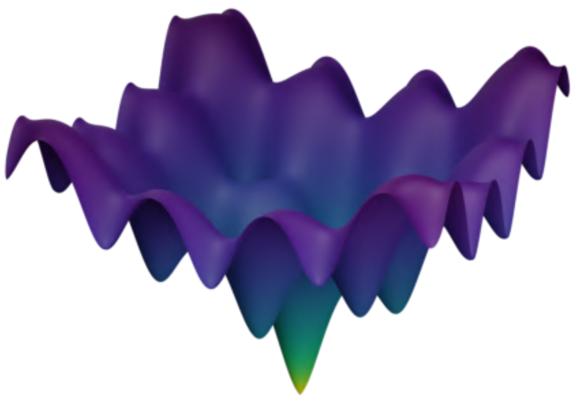
Prior samples

# Example: regression on a real projective space

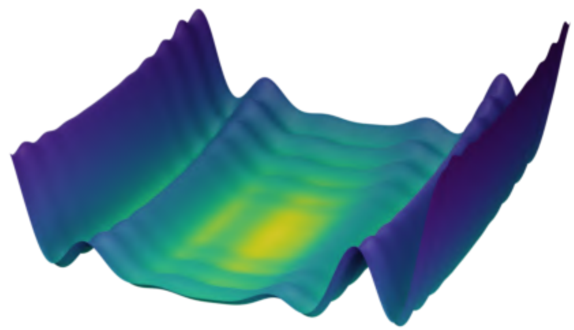


(a) Ground truth (b) Posterior mean (c) Std. deviation (d) Posterior sample

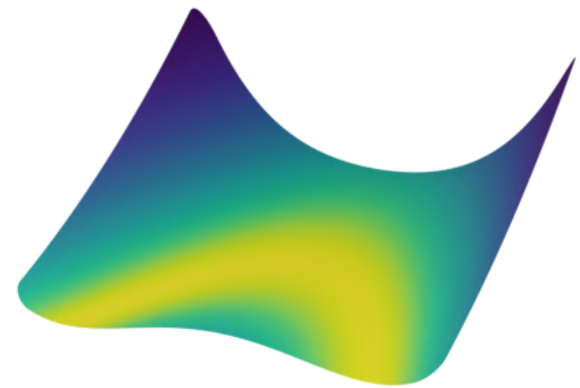
# Geometry-aware Bayesian Optimization



Ackley function

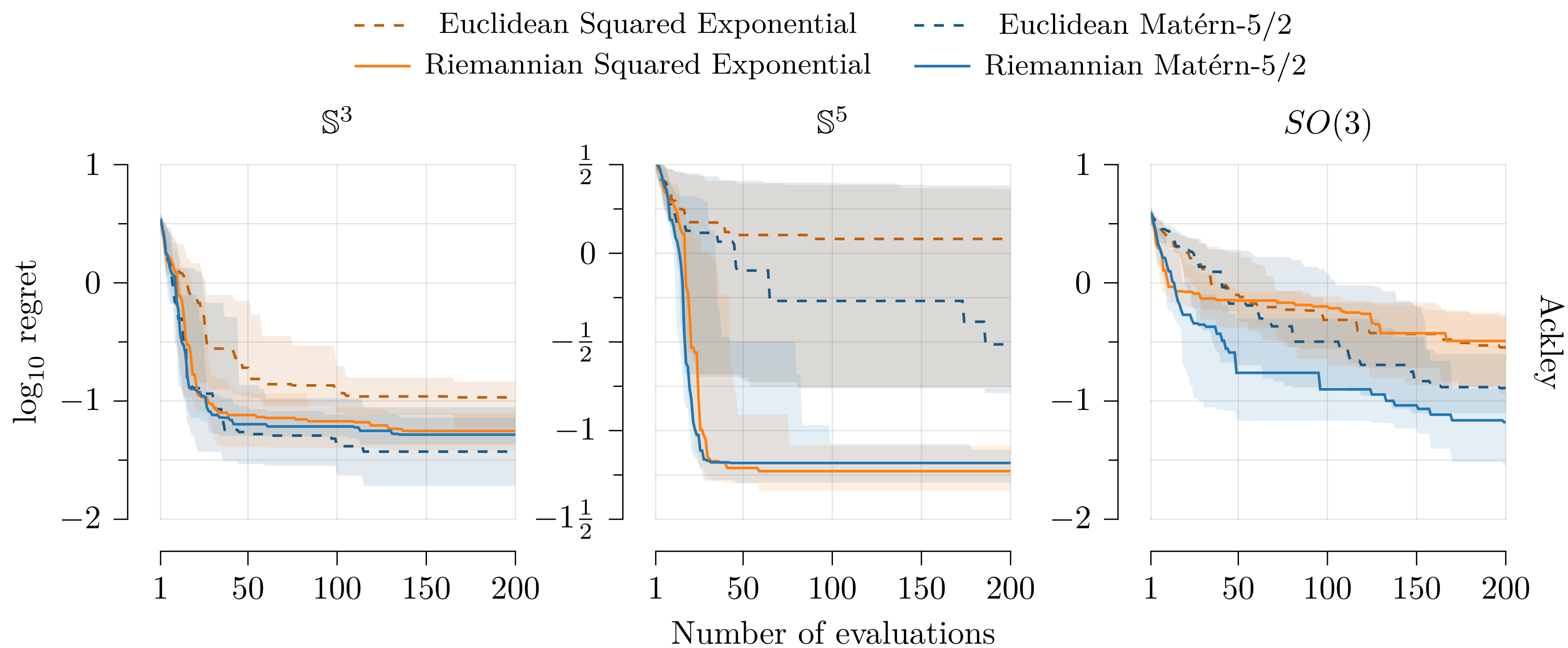


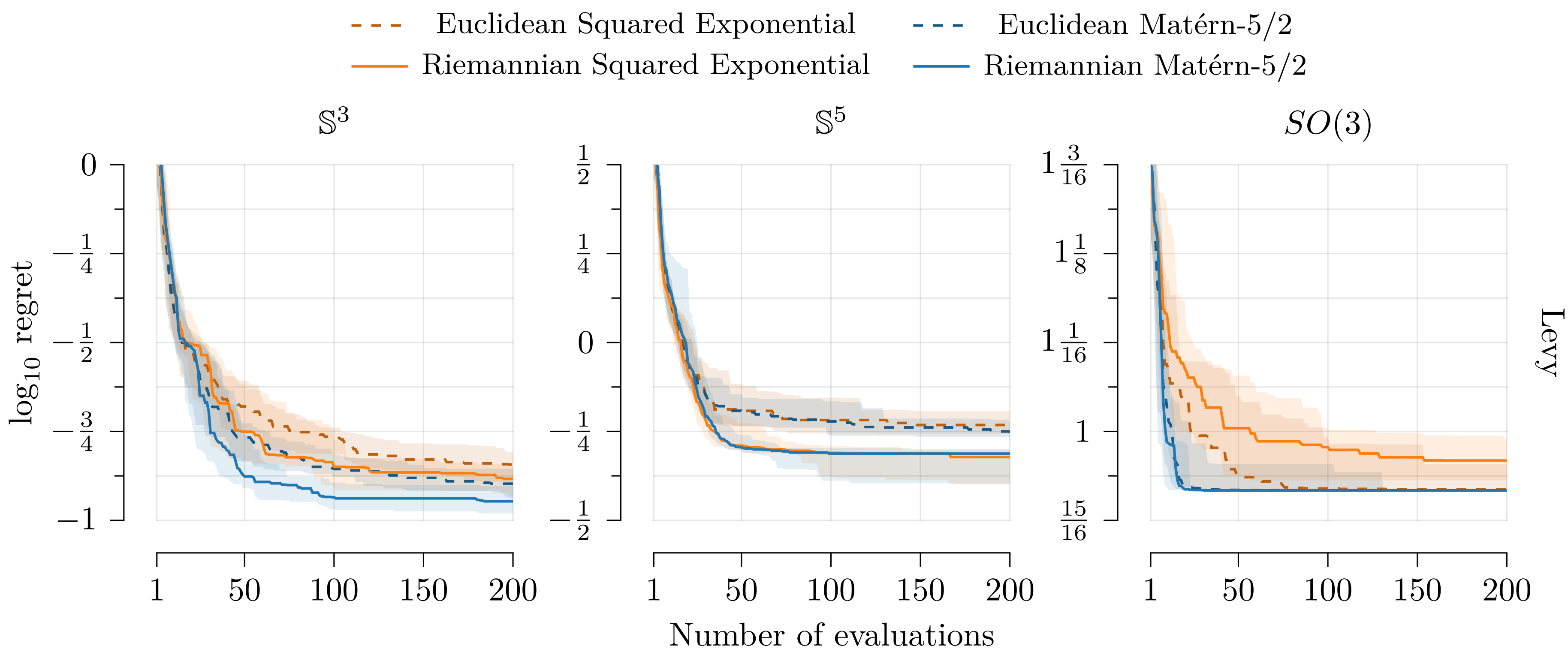
Levy function



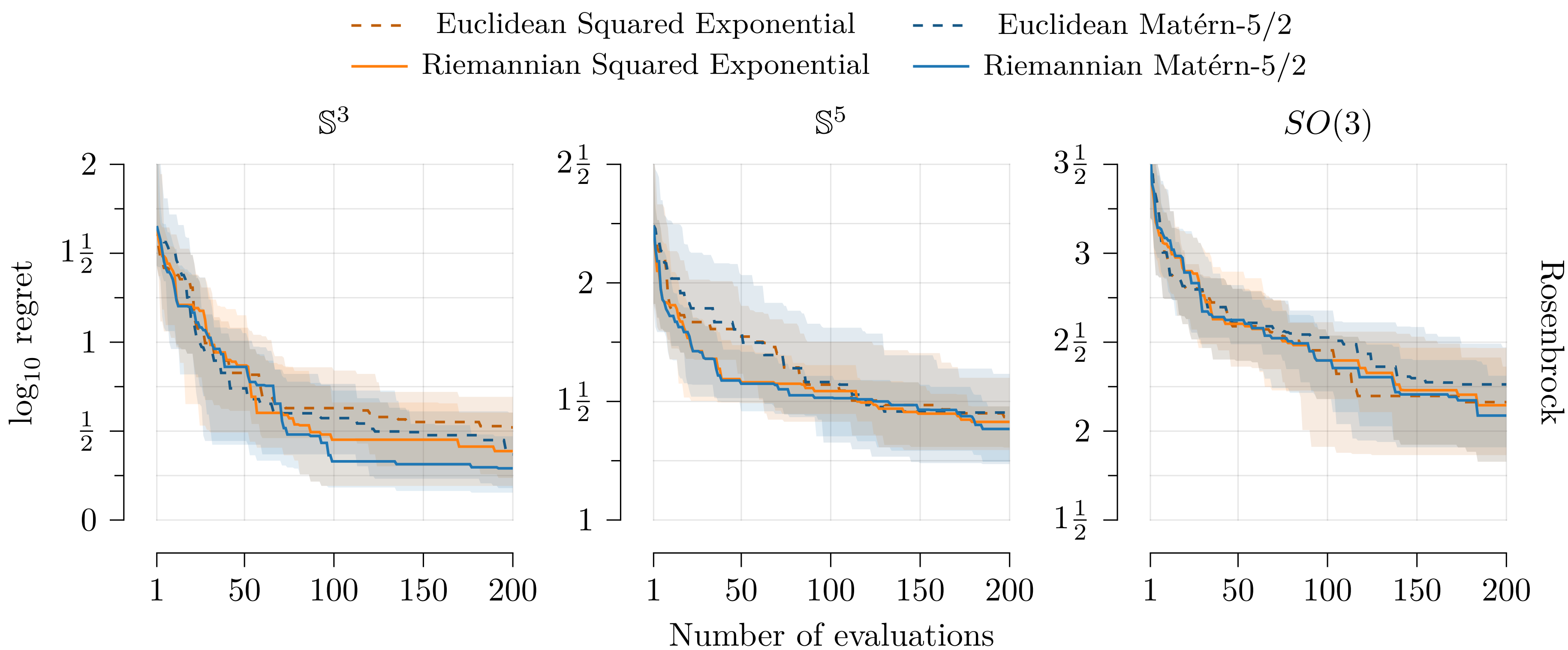
Rosenbrock function

# Geometry-aware Bayesian Optimization

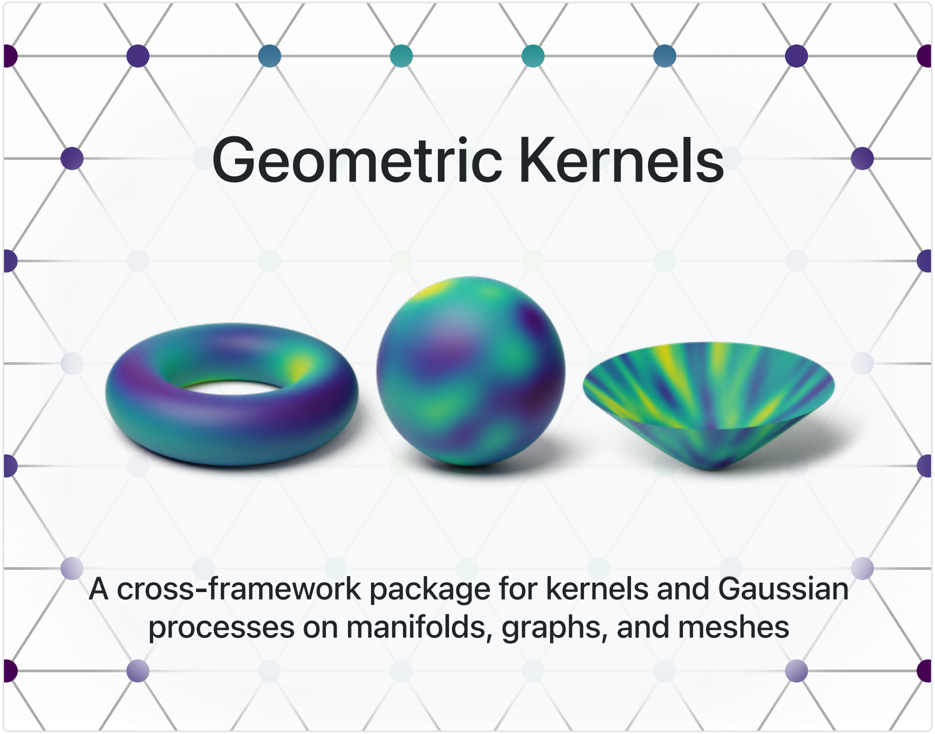




Euclidean Squared Exponential
Euclidean Matérn-5/2
Riemannian Squared Exponential
Riemannian Matérn-5/2
$\mathbb{S}^3$
$\mathbb{S}^5$
$SO(3)$
$\log_{10}$ regret
Levy
Number of evaluations


Euclidean Squared Exponential
Euclidean Matérn-5/2
Riemannian Squared Exponential
Riemannian Matérn-5/2
$\mathbb{S}^3$
$\mathbb{S}^5$
$SO(3)$
$\log_{10}$ regret
Rosenbrock
Number of evaluations
2
$1\frac{1}{2}$
1
$\frac{1}{2}$
0
$2\frac{1}{2}$
2
$1\frac{1}{2}$
1
$3\frac{1}{2}$
3
$2\frac{1}{2}$
2
$1\frac{1}{2}$
1
50
100
150
200

# Geometric Kernels in Python




HTTPS://GEOMETRIC-KERNELS.GITHUB.IO/

# Thank you!

HTTPS://AVT.IM/ · @AVT_IM

J. T. Wilson,* V. Borovitskiy,* P. Mostowsky,* A. Terenin,* M. P. Deisenroth. Efficiently Sampling Functions from Gaussian Process Posteriors. *International Conference on Machine Learning*, 2020. **Honorable Mention for Outstanding Paper Award.**

J. T. Wilson,* V. Borovitskiy,* P. Mostowsky,* A. Terenin,* M. P. Deisenroth. Pathwise Conditioning of Gaussian Process. *Journal of Machine Learning Research*, 2021.

V. Borovitskiy,* P. Mostowsky,* A. Terenin,* M. P. Deisenroth. Matérn Gaussian Processes on Riemannian Manifolds. *Advances in Neural Information Processing Systems*, 2020.

V. Borovitskiy,* I. Azangulov,* P. Mostowsky,* A. Terenin,* M. P. Deisenroth, N. Durrande. Matérn Gaussian Processes on Graphs. *Artificial Intelligence and Statistics*, 2021. **Best Student Paper Award.**

M. J. Hutchinson,* A. Terenin,* V. Borovitskiy,* S. Takao,* Y. W. Teh, M. P. Deisenroth. Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Independent Projected Kernels. *Advances in Neural Information Processing Systems*, 2021.

N. Jaquier,* V. Borovitskiy,* A. Smolensky, A. Terenin, T. Asfour, L. Rozo. Geometry-aware Bayesian Optimization in Robotics using Riemannian Matérn Kernels. *Conference on Robot Learning*, 2021.

I. Azangulov, A. Smolensky, A. Terenin, V. Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces I: the Compact Case. *arXiv: 2208.14960*, 2022.

A. Terenin,* D. R. Burt,* A. Artemev, S. Flaxman, M. van der Wilk, C. E. Rasmussen, H. Ge. Numerically Stable Sparse Gaussian Processes via Minimum Separation using Cover Trees. *arXiv: 2210.07893*, 2022.

*Equal contribution

UNIVERSITY OF CAMBRIDGE