Numerically Stable Sparse Gaussian Processes via Minimum Separation using Cover Trees

Alexander Terenin,[*1,2,3] David R. Burt,[*2,4] Artem Artemev,[*3,5,6] Seth Flaxman,[6] Mark van der Wilk,[3,6] Carl Edward Rasmussen,[2,5] Hong Ge[6]
[*]Equal contribution [1]Cornell [2]University of Cambridge [3]Imperial College London [4]MIT [5]Secondmind [6]University of Oxford

Abstract

Gaussian processes are frequently deployed as part of larger machine learning and decision-making systems, for instance in geospatial modeling, Bayesian optimization, or in latent Gaussian models. Within a system, the Gaussian process model needs to perform in a stable and reliable manner to ensure it interacts correctly with other parts of the system. In this work, we study the numerical stability of scalable sparse approximations based on inducing points. To do so, we first review numerical stability, and illustrate typical situations in which Gaussian process models can be unstable. Building on stability theory originally developed in the interpolation literature, we derive sufficient and in certain cases necessary conditions on the inducing points for the computations performed to be numerically stable. For low-dimensional tasks such as geospatial modeling, we propose an automated method for computing inducing points satisfying these conditions. This is done via a modification of the cover tree data structure, which is of independent interest. We additionally propose an alternative sparse approximation for regression with a Gaussian likelihood which trades off a small amount of performance to further improve stability. We provide illustrative examples showing the relationship between stability of calculations and predictive performance of inducing point methods on spatial tasks.

Numerical Stability

When do linear algebra computations work in practice?
This work: gather and synthesize results from various literatures

Key quantity: condition number

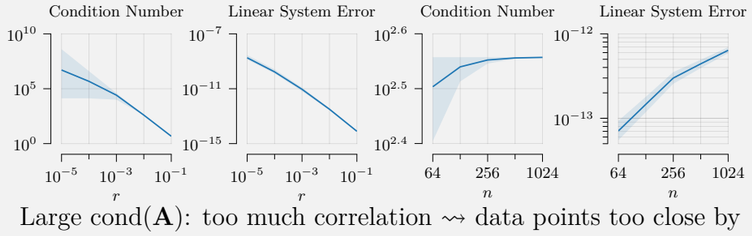
$$\operatorname{cond}(\mathbf{A}) = \lim_{\varepsilon>0} \sup_{\|\boldsymbol{\delta}\| \leq \varepsilon \|\boldsymbol{b}\|} \frac{\|\mathbf{A}^{-1}(\boldsymbol{b}+\boldsymbol{\delta}) - \mathbf{A}^{-1}\boldsymbol{b}\|_2}{\varepsilon \|\mathbf{A}^{-1}\boldsymbol{b}\|_2} = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

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

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

where t *is the floating point mantissa length, and that* $3N2^{-t} < 0.1$. *Then floating point Cholesky factorization will succeed, producing a matrix* $\mathbf{L}$ *satisfying* $\mathbf{L}\mathbf{L}^T = \mathbf{A} + \mathbf{E}$ *and* $\|\mathbf{E}\|_2 \leq 2^{-t} \times 1.38 N^{3/2} \|\mathbf{A}\|_2$.

Kernel matrix condition numbers: depend on properties of $\boldsymbol{x}$



Large $\operatorname{cond}(\mathbf{A})$: too much correlation $\rightsquigarrow$ data points too close by

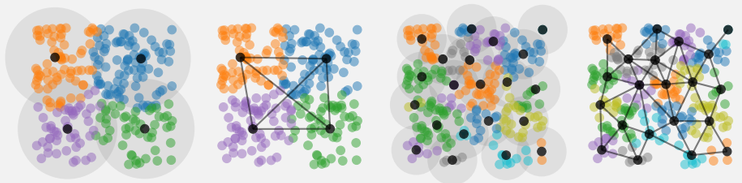
When are Gaussian process computations numerically stable?

Separation radius: $\operatorname{sep}(\boldsymbol{z}) = \min_{i \neq j} \|z_i - z_j\|$

Proposition. *Let* $X \subseteq \mathbb{R}^d$, *and let* k *be stationary, continuous, and satisfy spatial decay. Then there is a* $C^{k,\delta}_{\text{cond}}$ *such that for any* M *and any* $\boldsymbol{z}$ *of size* M *with* $\operatorname{sep}(\boldsymbol{z}) \geq \delta > 0$, *we have* $\operatorname{cond}(\mathbf{K}_{\boldsymbol{zz}}) \leq C^{k,\delta}_{\text{cond}}$.

Follows from results in kernel literature on minimum eigenvalue bounds, and maximum eigenvalue bounds proven under spatial decay via similar techniques based on packing arguments

Inducing Points via Cover Trees

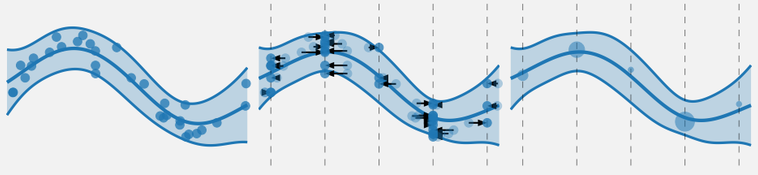


Algorithm to compute minimally separated inducing points:

1. Compute covering of data by picking points one-by-one
2. Compute neighbors within a given radius
3. Repeat recursively

Returns covering of data with minimally separated points
Efficient in geospatial settings: improves numerical stability

The Clustered-data Approximation



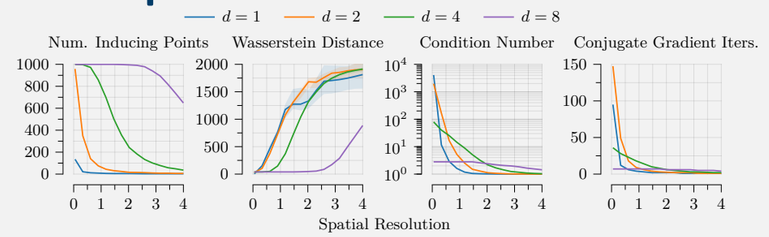
Inducing points: summarize effect of group of nearby points
Simpler way of doing this: merge nearby x-values

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

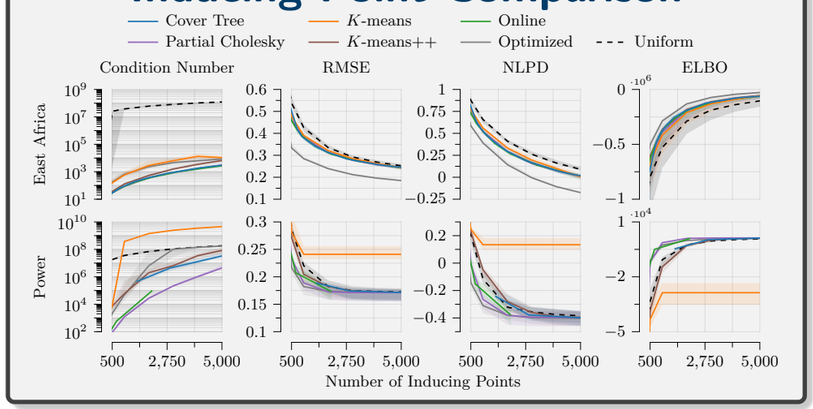
$\boldsymbol{z}$: inducing points $\quad$ $\boldsymbol{\Lambda}$: diagonal noise matrix

Improved stability since $\lambda_{\min}(\mathbf{K}_{\boldsymbol{zz}} + \boldsymbol{\Lambda}) \geq \lambda_{\min}(\mathbf{K}_{\boldsymbol{zz}})$

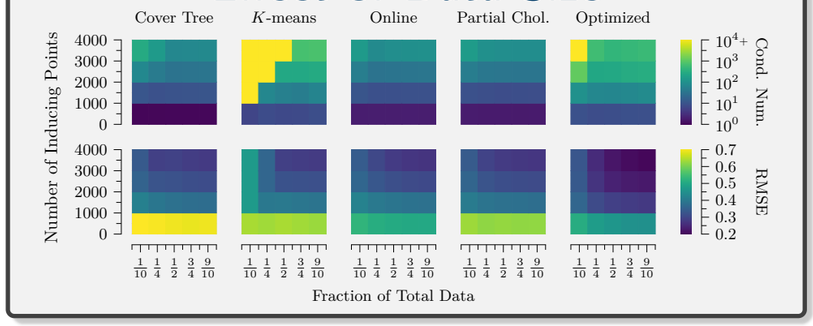
Empirical Cover Tree Behavior



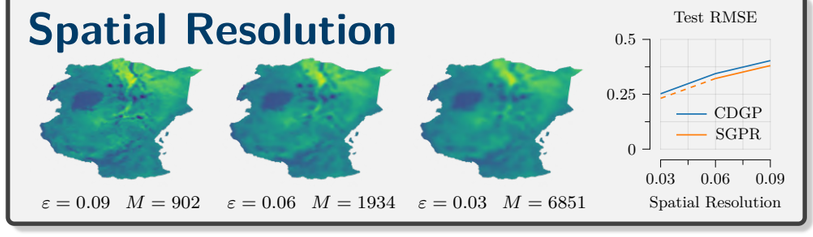
Inducing Point Comparison



Effect of Data Size



Spatial Resolution



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