

Scalable Gaussian Processes with Billions of Inducing Inputs via Tensor Train Decomposition

Pavel Izmailov^{1,4} Alexander Novikov^{2,3} Dmitry Kropotov⁴

¹ Cornell University ² National Research University Higher School of Economics

³ Institute of Numerical Mathematics RAS ⁴ Lomonosov Moscow State University

Summary

- Gaussian processes are flexible and elegant models, but exact inference requires $\mathcal{O}(n^3)$ computations, where n is the number of training data
- We propose the Tensor Train GP (TT-GP) framework with linear complexity $\mathcal{O}(n)$
- TT-GP allows to build flexible posterior approximations and train expressive deep kernels by using billions of inducing points
- TT-GP achieves state-of-the-art results on several important benchmarks both with RBF and deep kernels

Inducing Inputs and Structured Kernel Interpolation

- Inducing inputs are imaginary data points that allow to speed up GP inference
- SKI (Wilson and Nickisch, 2015): set inducing points on a multi-dimensional grid

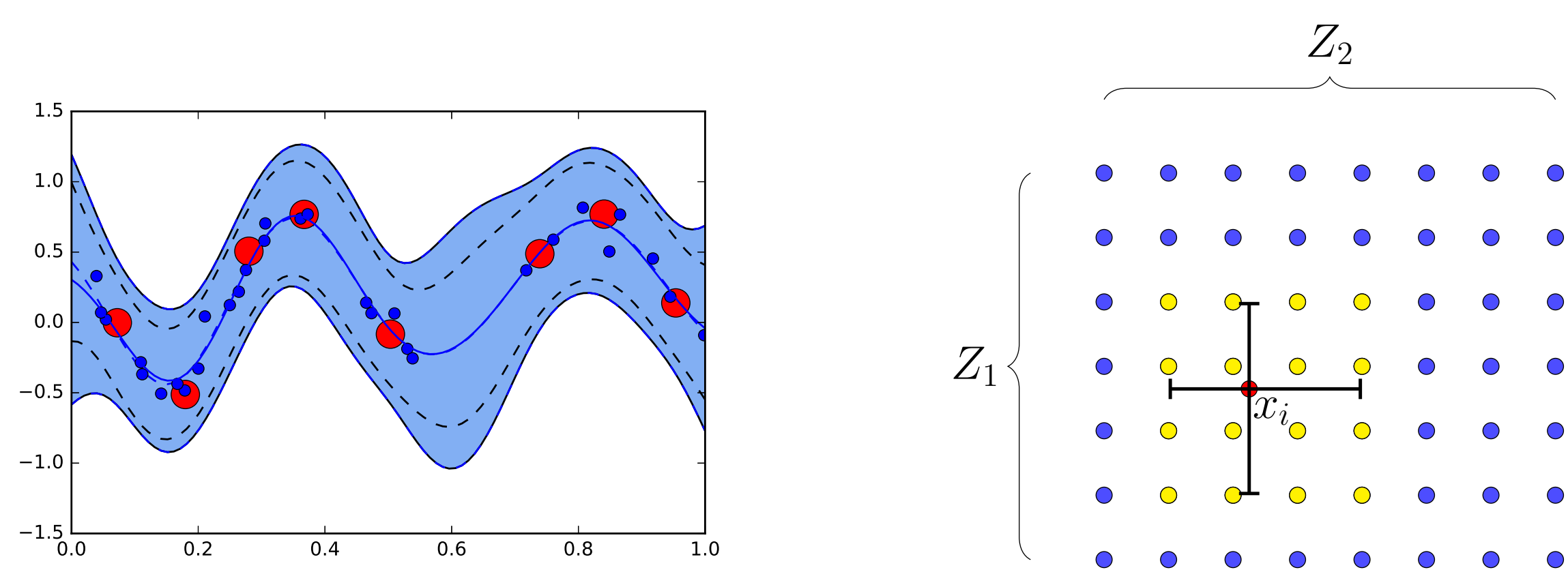
$$Z = Z^1 \times Z^2 \times \dots \times Z^D.$$

- Assume the kernel decomposes as

$$k(x, x') = k^1(x^1, x'^1) \cdot k^2(x^2, x'^2) \cdot \dots \cdot k^D(x^D, x'^D)$$

Covariance matrix $K_{mm} \in \mathbb{R}^{m \times m}$ computed at the inducing points takes form

$$K_{mm} = K_{m_1 m_1}^1 \otimes K_{m_2 m_2}^2 \otimes \dots \otimes K_{m_D m_D}^D$$



- $|K_{mm}|$ and K_{mm}^{-1} can be computed efficiently
- Inducing points can be considered as interpolation points for the kernel

$$k_i \approx K_{mm} w_i,$$

where $k_i \in \mathbb{R}^m$ is the vector of covariances between the i -th training object and the inducing points, $w_i \in \mathbb{R}^m$ is the vector of interpolation coefficients

- KISS-GP uses cubic convolutional interpolation for which

$$w_i = w_i^1 \otimes w_i^2 \otimes \dots \otimes w_i^D$$

Tensor Train Format

Tensor \mathcal{A} is said to be represented in Tensor Train (Oseledets, 2011) format if:

$$\mathcal{A}_{i_1 \dots i_d} = \underbrace{G_1[i_1]}_{1 \times r} \underbrace{G_2[i_2]}_{r \times r} \dots \underbrace{G_d[i_d]}_{r \times 1}$$

An example of computing one element of a 4-dimensional tensor:

$$\mathcal{A}_{2423} = \underbrace{G_1}_{i_1=2} \times \underbrace{G_2}_{i_2=4} \times \underbrace{G_3}_{i_3=2} \times \underbrace{G_4}_{i_4=3}$$

- TT-format uses $\mathcal{O}(dnr^2)$ memory to approximate a tensor with n^d elements
- Allows for efficient implementation of linear algebra operations

Gaussian Process ELBO

Evidence Lower Bound (Hensman et al., 2013) with KISS-GP approximation of k_i :

$$\log p(y) \geq \sum_{i=1}^n \left(\log \mathcal{N}(y_i | w_i^T \mu, \sigma^2) - \frac{1}{2\sigma^2} (\text{var}_i + w_i^T K_{mm} w_i) - \frac{1}{2\sigma^2} \text{tr}(w_i^T \Sigma w_i) \right) - \frac{1}{2} \left(\log \frac{|K_{mm}|}{|\Sigma|} - m + \text{tr}(K_{mm}^{-1} \Sigma) + \mu^T K_{mm}^{-1} \mu \right)$$

where

- σ^2 is the noise variance
- $\mu \in \mathbb{R}^m$, $\Sigma \in \mathbb{R}^{m \times m}$ variational parameters
- $\tilde{K}_{ii} = \text{var} - k_i^T K_{mm}^{-1} k_i$, where var is the prior variance of the process at any point

TT-GP

- Set inducing points Z on a grid in the feature space
- Restrict Σ to be in a Kronecker product format

$$\Sigma = \Sigma^1 \otimes \Sigma^2 \otimes \dots \otimes \Sigma^D$$

- Represent μ as a d -dimensional tensor, restrict to be in TT format with TT-ranks r
- Maximize ELBO with respect to TT-cores of μ , Kronecker factors of Σ using SGD

Properties

- Linear computational complexity in the size of the data $\mathcal{O}(nDm^{1/D}r^2 + Dm^{1/D}r^3 + Dm^{3/D})$. TT-ranks are in general on the scale of $r \approx 10$. Here $m = m_0^D$
- TT-GP can be applied for very large n and m

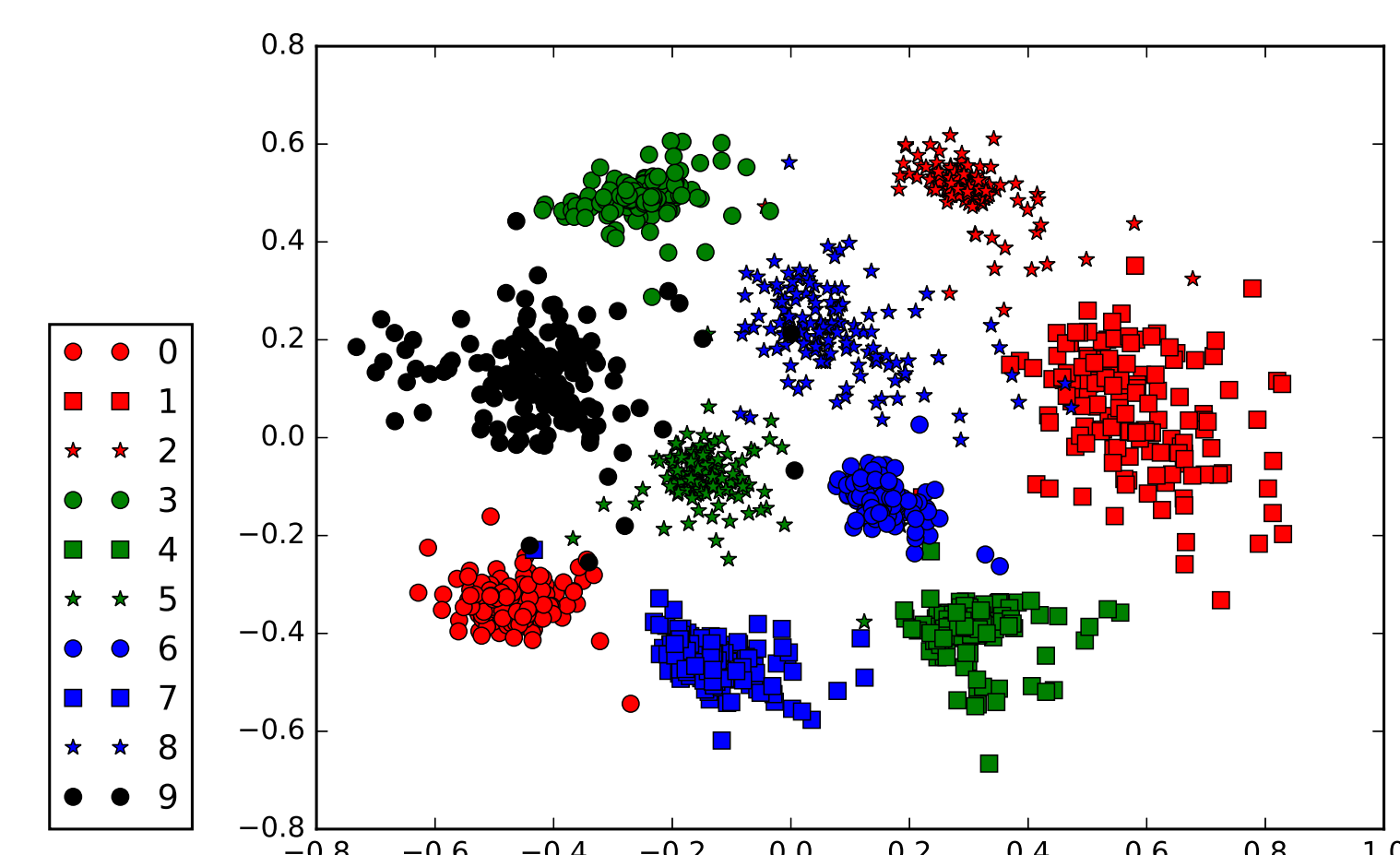
RBF Kernel Experiments

Comparison with SVI-GP (Hensman et al., 2013) on regression and classification tasks:

Dataset			SVI-GP / KLSP-GP			TT-GP			
Name	n	D	acc.	m	t (s)	acc.	m	d	t (s)
Powerplant	7654	4	0.94	200	10	0.95	35^4	-	5
Protein	36584	9	0.50	200	45	0.56	30^9	-	40
YearPred	463K	90	0.30	1000	597	0.32	10^6	6	105
Airline	$6M$	8	0.665*	-	-	0.694	20^8	-	5200
svmguidel	3089	4	0.967	200	4	0.969	20^4	-	1
EEG	11984	14	0.915	1000	18	0.908	12^{10}	10	10
covtype bin	465K	54	0.817	1000	320	0.852	10^6	6	172

Deep Kernel Experiments

Embedding learned by TT-GP with a deep kernel on *digits* dataset:



Comparison with SV-DKL (Wilson et al., 2016) and stand-alone DNN:

Dataset		SV-DKL	DNN		TT-GP		
Name	n	acc.	acc.	t (s)	acc.	d	t (s)
Airline	$6M$	0.781	0.780	1055	0.788 ± 0.002	2	1375
CIFAR-10	$50K$	—	0.915	166	0.908 ± 0.003	9	220
MNIST	$60K$	—	0.993	23	0.9936 ± 0.0004	10	64

References

- J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 282–290. AUAI Press, 2013.
- I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.
- A. G. Wilson and H. Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784, 2015.
- A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems*, pages 2586–2594, 2016.