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

Pavel Izmailov[1,4] **Alexander Novikov**[2,3] **Dmitry Kropotov**[4]

[1] Cornell University [2] National Research University Higher School of Economics
[3] Institute of Numerical Mathematics RAS [4] Lomonosov Moscow State University

Motivation

Motivation

Structured Kernel Interpolation (SKI)

Wilson and Nickisch (2015)

- Set inducing points on a multi-dimensional grid

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

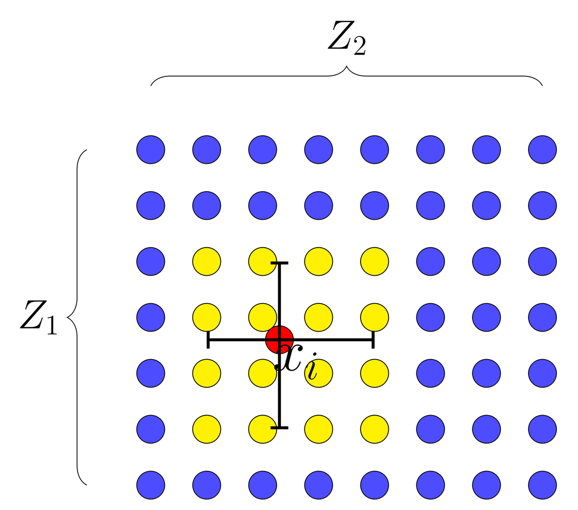
- Assume the kernel decomposes as

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

Then covariance matrix K_{mm} takes form

$$K_{mm} = K^1_{m_1 m_1} \otimes K^2_{m_2 m_2} \otimes \ldots \otimes K^D_{m_D m_D}.$$

- $|K_{mm}|$ and K_{mm}^{-1} can be computed efficiently.



- Inducing points can be considered as inteprolation points for the kernel.

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

where $w_i \in \mathbb{R}^m$ is the vector of interpolation coefficients.

- KISS-GP uses cubic convolutional interpolation for which

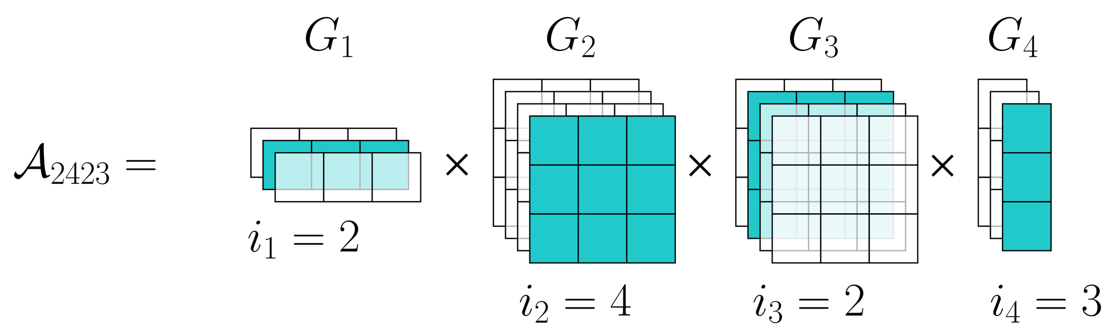
$$w_i = w_i^1 \otimes w_i^2 \otimes \ldots \otimes w_i^D.$$

Tensor Train Format

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

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

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



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

Gaussian Process ELBO

Using KISS-GP approximation of $k_i \approx K_{mm} w_i$, we can rewrite the ELBO as

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

where

- $K_{mm} \in \mathbb{R}^{m \times m}$ is the covariance matrix computed at the inducing points
- $k_i \in \mathbb{R}^m$ is the vector of covariances between the i-th training object and the inducing points
- 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

$$\Sigma = \Sigma^1 \otimes \Sigma^2 \otimes \ldots \otimes \Sigma^D.$$

- Restrict μ to be in TT format with TT-ranks r.

Properties

- The computational complexity is linear 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 $m = m_0^D$.
- To fit the process we can use simple SGD
- The method can be applied for large n, moderately large D, and huge m

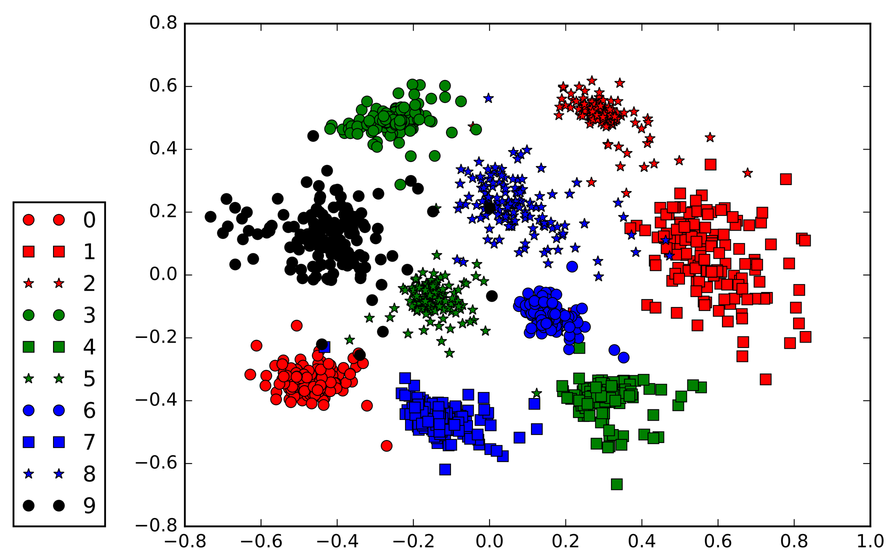
RBF Kernel Experiments

Hensman et al. (2013)

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
svmguide1	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

Wilson et al. (2016)



Dataset	SV-DKL	DNN		TT-GP		
Name	acc.	acc.	t (s)	acc.	d	t (s)
Airline	0.781	0.780	1055	0.788 ± 0.002	2	1375
CIFAR-10	–	0.915	166	0.908 ± 0.003	9	220
MNIST	–	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.