# Supplementary Material – Scalable Gaussian Process Separation for Kernels with a Non-Stationary Phase

**Jan Graßhoff**[1] **Alexandra Jankowski**[1] **Philipp Rostalski**[1]

### Overview of Kronecker and Toeplitz Methods

Table 1: Structure exploiting inference and learning methods. All kernels are assumed to be stationary and $\hat{K} = K + \sigma^2 I$.

| Kernel and Inputs | Matrix | Linear Solve | Log Determinant |
|---|---|---|---|
| Kernel is separable: $k(\mathbf{x}, \mathbf{z}') = \prod_{d=1}^{D} k_d(\mathbf{x}^{(d)}, \mathbf{z}^{(d)})$ | $K = \bigotimes_{d=1}^{D} K_d$ and | noise-free: $K^{-1}\mathbf{y} = \bigotimes_{d=1}^{D} K_d^{-1}\mathbf{y}$ | noise-free: $\log\lvert K\rvert = \sum_i V_{i,i}$ |
| Inputs on a rectilinear grid: $X = \mathcal{X}_1 \times \cdots \times \mathcal{X}_D$ | $K = QVQ^\top$ | noisy: $\hat{K}^{-1}\mathbf{y} = Q(V + \sigma^2 I)^{-1}Q^\top\mathbf{y}$ | noisy: $\log\lvert\hat{K}\rvert = \sum_i \left(V_{i,i} + \sigma^2\right)$ |
| Kernel: $k(x, x')$ Inputs: $x \in \mathbb{R}$ and equispaced | $K$ is Toeplitz | LCG with fast MVMs | (1) circulant approx. (Wilson et al., 2015) (2) stoch. trace estim. (Dong et al., 2017) |
| Kernel is separable Inputs unstructured | $K \approx W K_{UU} W^\top$ (Wilson & Nickisch, 2015) | LCG with fast MVMs | (1) scaled eigenvalues (Wilson et al., 2015) (2) stoch. trace estim. (Dong et al., 2017) |

### Structure Exploitation for Kernels with a Non-Stationary Phase

Table 2: Comparison of the standard SKI (Wilson & Nickisch, 2015) approach using equidistant inducing points $U$ and warpSKI. Inputs may be unstructured (or have partial grid structure). The kernels $k$ (and $k_i$) are assumed to be stationary and separable. The functions $\phi_i : \mathcal{D} \to \mathcal{D}_i$ and $\phi : \mathcal{D} \to \mathcal{D}_1$ with $\mathcal{D}_{\text{in}} \subseteq \mathbb{R}^D$, $\mathcal{D}_i \subseteq \mathbb{R}^D$ are invertible functions. The linear solve $\hat{K}^{-1}\mathbf{y}$ is done by conjugate gradients and the log determinant is approximated using stochastic trace estimation.

| Kernel | equidistant $U$ recovers... | warpSKI recovers... |
|---|---|---|
| $k(\phi(x), \phi(x'))$ with $x \in \mathbb{R}$ | – | Toeplitz structure |
| $\sum_i k_i(\phi_i(x), \phi_i(x'))$ with $x \in \mathbb{R}$ | – | sum over Toeplitz structures |
| $k(\phi(\mathbf{x}), \phi(\mathbf{x}'))$, $\mathbf{x} \in \mathbb{R}^D$ and $\phi$ is not an elementwise fnc. | – | Kronecker and Toeplitz structure |
| $k(\phi(\mathbf{x}), \phi(\mathbf{x}'))$ $\mathbf{x} \in \mathbb{R}^D$ and $\phi$ is an elementwise fnc. | Kronecker structure | Kronecker and Toeplitz structure |
| $\sum_i k_i(\phi_i(\mathbf{x}), \phi_i(\mathbf{x}'))$ $\mathbf{x} \in \mathbb{R}^D$ and $\phi_i$ are elementwise fnc. | sum over Kronecker structures | sum over Kronecker and Toeplitz structures |

[1]Institute for Electrical Engineering in Medicine, Universität zu Lübeck, Germany. Correspondence to: Jan Graßhoff <j.grasshoff@uni-luebeck.de>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

**Experimental Data - Numerical Results**

Table 3: Inference runtime (s) for the numerical experiment with $n$ data points and $m$ inducing points. The results are averages over five samples ± one standard deviation.

| Points | Inducing Points | | |
|---|---|---|---|
| | $m = 9933$ | $m = 48824$ | $m = 74836$ |
| $n = 10^{2.75}$ | 0.11±0.02 | 0.28±0.05 | 0.47±0.07 |
| $n = 10^3$ | 0.14±0.03 | 0.43±0.05 | 0.64±0.08 |
| $n = 10^3.5$ | 0.34±0.06 | 1.10±0.10 | 1.61±0.17 |
| $n = 10^4$ | 1.95±0.48 | 3.80±0.68 | 4.76±0.89 |
| $n = 10^4.5$ | 19.9±4.4 | 25.8±3.7 | 26.3±4.7 |
| $n = 10^5$ | 214±44 | 228±37 | 241±49 |

Table 4: Likelihood evaluation time (s) for the numerical experiment with $n$ data points and $m$ inducing points. The results are averages over five samples ± one standard deviation.

| Points | Inducing Points | | |
|---|---|---|---|
| | $m = 9933$ | $m = 48824$ | $m = 74836$ |
| $n = 10^{2.75}$ | 1.81±0.37 | 4.41±1.34 | 8.17±2.10 |
| $n = 10^3$ | 1.81±0.37 | 5.74±1.06 | 10.0±1.7 |
| $n = 10^3.5$ | 1.92±0.39 | 6.15±1.16 | 10.9±1.9 |
| $n = 10^4$ | 2.35±0.48 | 6.67±0.68 | 12.0±2.1 |
| $n = 10^4.5$ | 6.19±1.45 | 10.1±1.7 | 14.4±2.8 |
| $n = 10^5$ | 17.0±3.7 | 19.8±3.6 | 26.0±5.2 |

Table 5: Learning runtime (s) for the numerical experiment with $n$ data points and $m$ inducing points. The results are averages over five samples ± one standard deviation.

| Points | Inducing Points | | |
|---|---|---|---|
| | $m = 9933$ | $m = 48824$ | $m = 74836$ |
| $n = 10^{2.75}$ | 28.3±9.3 | 84.6±40.5 | 166±53 |
| $n = 10^3$ | 29.0±5.3 | 94.3±34.2 | 159±40 |
| $n = 10^3.5$ | 26.6±16.0 | 92.0±20.8 | 177±37 |
| $n = 10^4$ | 55.1±7.8 | 117±35 | 214±33 |
| $n = 10^4.5$ | 130±49 | 163±23 | 252±88 |
| $n = 10^5$ | 309±110 | 458±79 | 398±153 |

Table 6: RMSE for the numerical experiment with $n$ data points and $m$ inducing points. The results are averages over five samples ± one standard deviation.

| Points | Inducing Points | | |
|---|---|---|---|
| | $m = 9933$ | $m = 48824$ | $m = 74836$ |
| $n = 10^{2.75}$ | 0.29±0.02 | 0.30±0.03 | 0.29±0.03 |
| $n = 10^3$ | 0.28±0.02 | 0.25±0.03 | 0.26±0.01 |
| $n = 10^3.5$ | 0.20±0.02 | 0.19±0.01 | 0.19±0.01 |
| $n = 10^4$ | 0.17±0.01 | 0.16±0.02 | 0.16±0.02 |
| $n = 10^4.5$ | 0.16±0.03 | 0.15±0.02 | 0.16±0.02 |
| $n = 10^5$ | 0.15±0.01 | 0.15±0.02 | 0.15±0.02 |

## References

Dong, K., Eriksson, D., Nickisch, H., Bindel, D., and Wilson, A. Scalable log determinants for gaussian process kernel learning. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6327–6337, 2017.

Wilson, A. G. and Nickisch, H. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning (ICML)*, pp. 1775–1784, 2015.

Wilson, A. G., Dann, C., and Nickisch, H. Thoughts on massively scalable gaussian processes. *CoRR*, abs/1511.01870, 2015. URL http://arxiv.org/abs/1511.01870.