Bayesian optimization: motivation and potential research directions

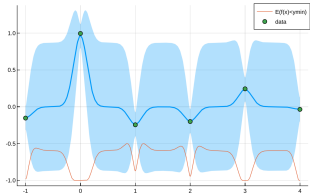
FJFI seminar, 6.3.2024

Václav Šmídl

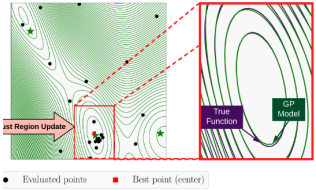
March 7, 2024

Ouline

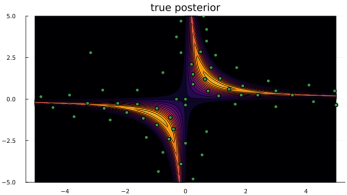
1. Motivation



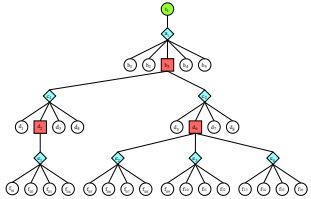
2. High-dimensions



3. Posterior Distribution

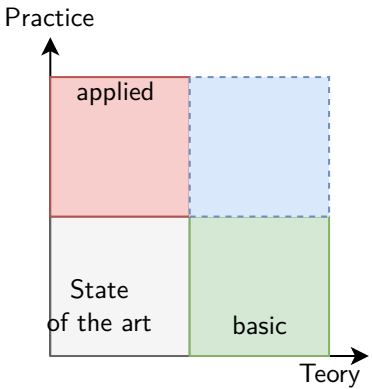


4. Structured Data



Practically motivated theoretical research

Research types

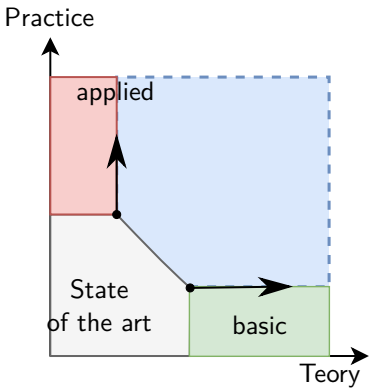


Theoretical research

- Development of abstract well defined concepts

Practically motivated theoretical research

Research types



Theoretical research

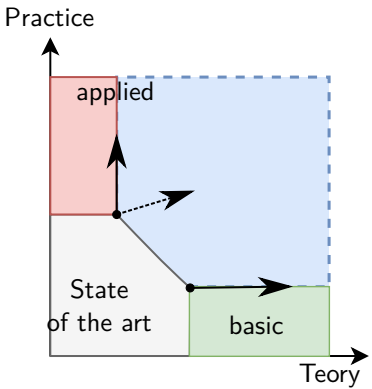
- Development of abstract well defined concepts

Practical research

- Problem definition takes often more time than solution
- Simplicity is essential!

Practically motivated theoretical research

Research types



Theoretical research

- Development of abstract well defined concepts

Practical research

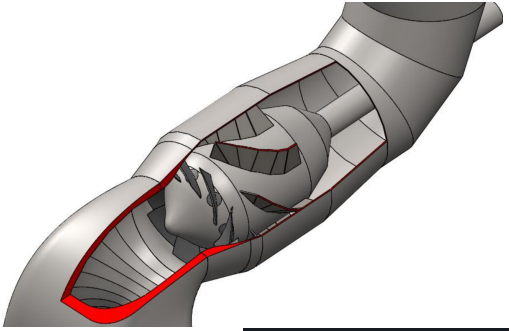
- Problem definition takes often more time than solution
- Simplicity is essential!

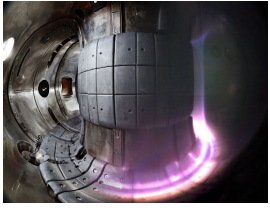
Best of both worlds

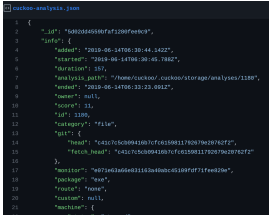
- Practically motivated theoretical problems
- Need for a good partner in the industry

Common problem: expensive data

1. Optimization of machine design
   - one simulation takes a day
2. Design of monitoring network for a tokamak
   - New measurement device is expensive
3. Detection of novel mallware
   - human analysis expensive and slow







Bayesian Optimization Classic

- Močkus, Jonas. **On Bayesian methods for seeking the extremum**. In Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974, pp. 400-404. Springer Berlin Heidelberg, 1975.

Bayesian Optimization Classic

- Močkus, Jonas. **On Bayesian methods for seeking the extremum**. In Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974, pp. 400-404. Springer Berlin Heidelberg, 1975.
- Jones, Donald R., Matthias Schonlau, and William J. Welch. **Efficient global optimization of expensive black-box functions**. Journal of Global optimization 13, no. 4 (1998): 455.
- Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. **Taking the human out of the loop: A review of Bayesian optimization**. Proceedings of the IEEE 104, no. 1 (2015): 148-175.

Bayesian Optimization Classic

- Močkus, Jonas. **On Bayesian methods for seeking the extremum**. In Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974, pp. 400-404. Springer Berlin Heidelberg, 1975.
- Jones, Donald R., Matthias Schonlau, and William J. Welch. **Efficient global optimization of expensive black-box functions**. Journal of Global optimization 13, no. 4 (1998): 455.
- Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. **Taking the human out of the loop: A review of Bayesian optimization**. Proceedings of the IEEE 104, no. 1 (2015): 148-175.

Becomes standard tool:

- part of google tools for tuning of Neural Networks
- winner of many benchmarks
- rules of thumb of using:
  - when the number of data <1000
  - when problem dimension <20

Why it doesn't work in higher-dimensions?

1. Curse of dimensionality (Binois, Wycoff, 2022):

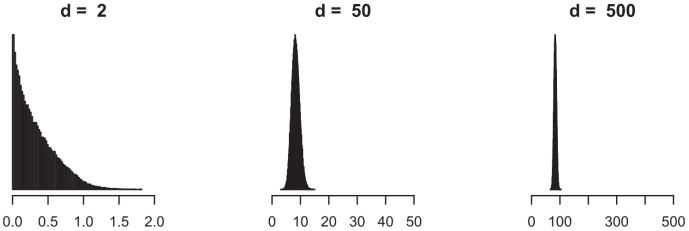


Fig. 2. Distances Concentrate in High Dimension: Randomly sampling 10,000 points according to the uniform measure in $[0,1]^d$ and then calculating squared inter-point distances reveals that these concentrate within the bounds of possible values in high dimension: $[0, d]$.

Why it doesn't work in higher-dimensions?

1. Curse of dimensionality (Binois, Wycoff, 2022):

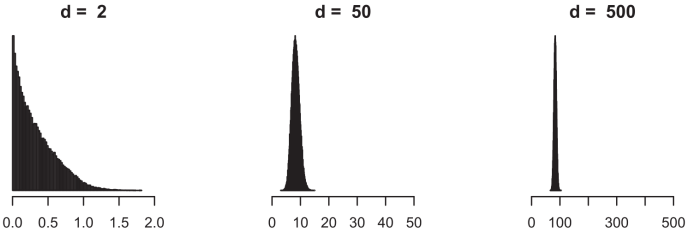
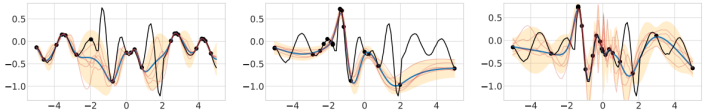


Fig. 2. Distances Concentrate in High Dimension: Randomly sampling 10,000 points according to the uniform measure in $[0,1]^d$ and then calculating squared inter-point distances reveals that these concentrate within the bounds of possible values in high dimension: $[0, d]$.

2. Homogeneity of the model (Li, Rudner, Wilson, 2023)
   - the distance function is same for all points



(a) GP (b) HMC (c) I-BNN

Heterogenous GP models

1. Deep GP: $f(x) \sim \mathcal{GP}()$, $x = g(z) \sim \mathcal{GP}()$
   - Kurt Cutajar, Mark Pullin, Andreas Damianou, Javier González, and Neil Lawrence. 2018. Deep Gaussian processes for multi-fidelity modeling. In Proceedings of the International Conference on Neural Information Processing Systems.
2. Non-stationary kernels:
   - Christopher J. Paciorek and Mark J. Schervish. 2006. Spatial modelling using a new class of nonstationary covariance functions. Environmetrics: The Official Journal of the International Environmetrics Society 17, 5 (2006), 483–506

Methods have no analytical solution and are inaccurate and computationally costly!

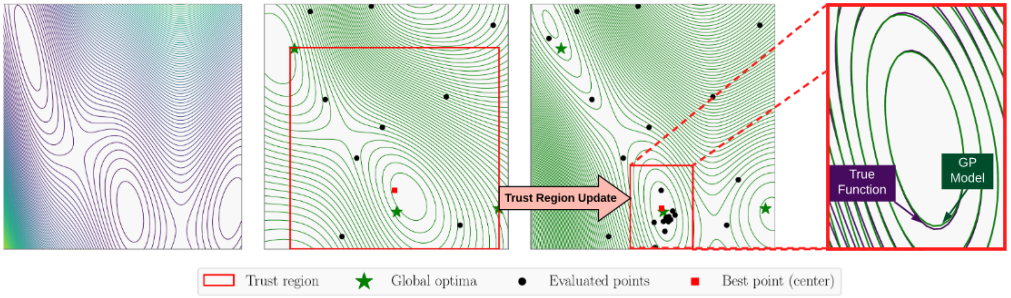
Deep Neural networks

Neal 1995: Gaussian Process is a limit of multilayer perceptron with infinitely many neurons on the hidden layer.

Open problem

Combining BO with Trust region: TuRBO

- Uses the idea of Trust Region to fit GP only locally
- Keeps multiple TR simultaneously and samples from them using the **Multi-armed Bandit** methods
- David Eriksson, Michael Pearce, Jacob Gardner, Ryan D. Turner, and Matthias Poloczek. 2019. Scalable global optimization via local Bayesian optimization. In Proceedings of the 33rd International Conference on Neural Information Processing Systems. 5496–5507



Looking for a set of solutions

Task: Decide if we need to buy a new (expensive) sensor for Tokamak

Reason: Existing sensor network may not see clearly enough what is going on in the plasma.

Evaluation: How will the new measurement reduce uncertainty?

Looking for: quantification of uncertainty:

- a set of admissible solution (Diversity in BO)
- posterior distribution / likelihood function
  - Gutmann, M.U. and Cor, J., 2016. Bayesian optimization for likelihood-free inference of simulator-based statistical models. Journal of Machine Learning Research, 17(125), pp.1-47.

Illustrative example

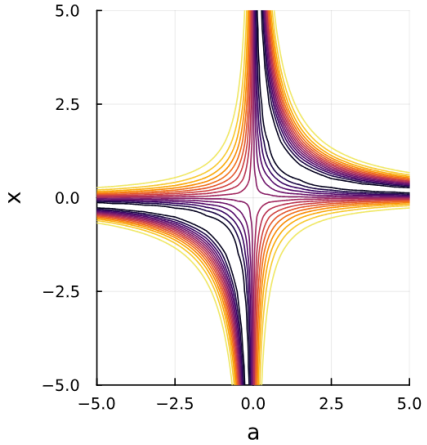
Example:
unknown variables a, x observed via

$$y = ax + e_1$$
$$p(y|a, x) = \mathcal{N}(ax, \sigma_e)$$

every evaluation of $p(y|a, x)$ is expensive!

Visualization of the posterior for
evaluated on a grid of 300x300 points for

$$y = 1, \sigma_e = 1.$$



- imagine every evaluation costs 100$

How it works

- The GP is used to model the model deviation Δ_θ, present in the log-likelihood, e.g.

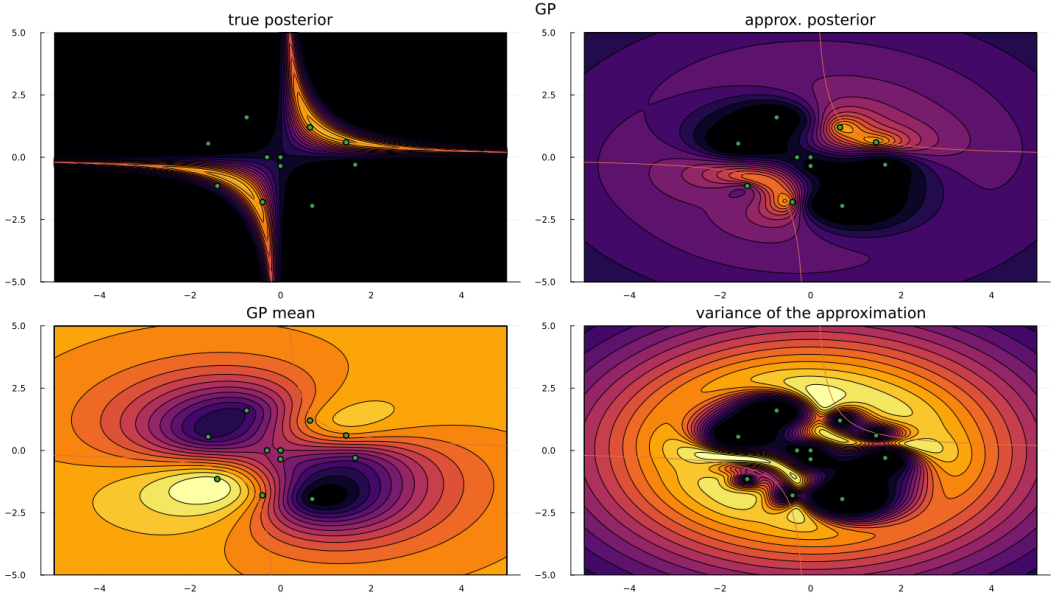
$$\log p(y|a,x) = -\frac{1}{2\sigma}\Delta_\theta^2$$

- The estimate of the density is available in closed form

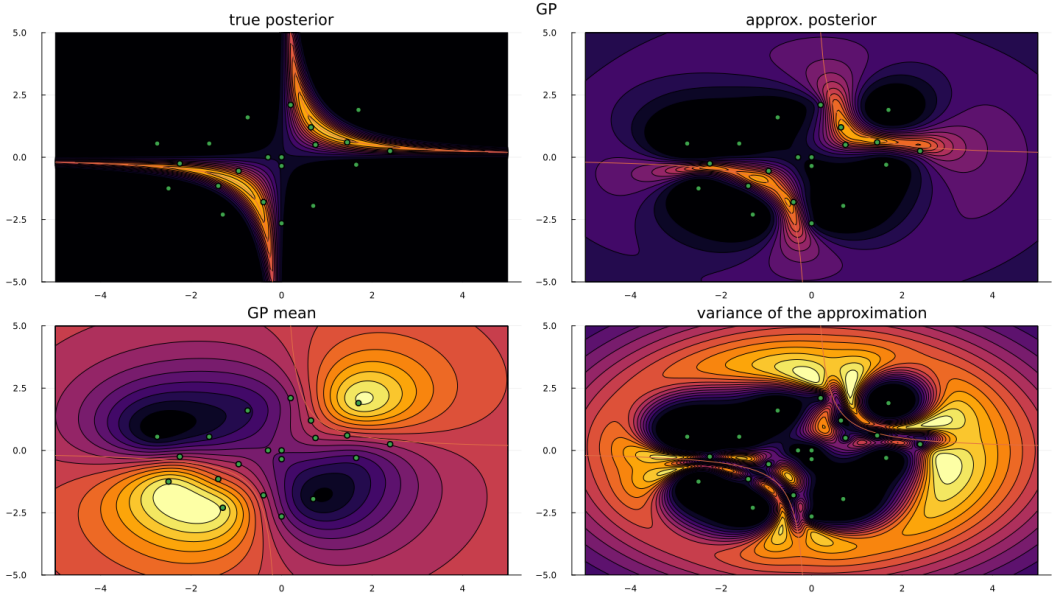
$$\begin{aligned}
\mathbb{E}\left[p(y_o|\theta)p(\theta)\right] &= p(\theta)\mathcal{N}(\mu_d(\theta) \mid 0, \sigma_e^2 + \sigma_d^2(\theta)) \\
\mathbb{V}\left[p(y_o|\theta)p(\theta)\right] &= p(\theta)^2 \frac{1}{2\sqrt{\pi}} \left[\frac{1}{\sigma_e}\mathcal{N}\left(\mu_d(\theta) \mid 0, \frac{1}{2}(\sigma_e^2 + 2\sigma_d^2(\theta))\right)\right. \\
&\quad \left. - \frac{1}{\sqrt{\sigma_e^2 + \sigma_d^2}}\mathcal{N}\left(\mu_d(\theta) \mid 0, \frac{1}{2}(\sigma_e^2 + \sigma_d^2)\right)\right]
\end{aligned}$$

- Acquisition function: minimize the variance of the likelihood function!

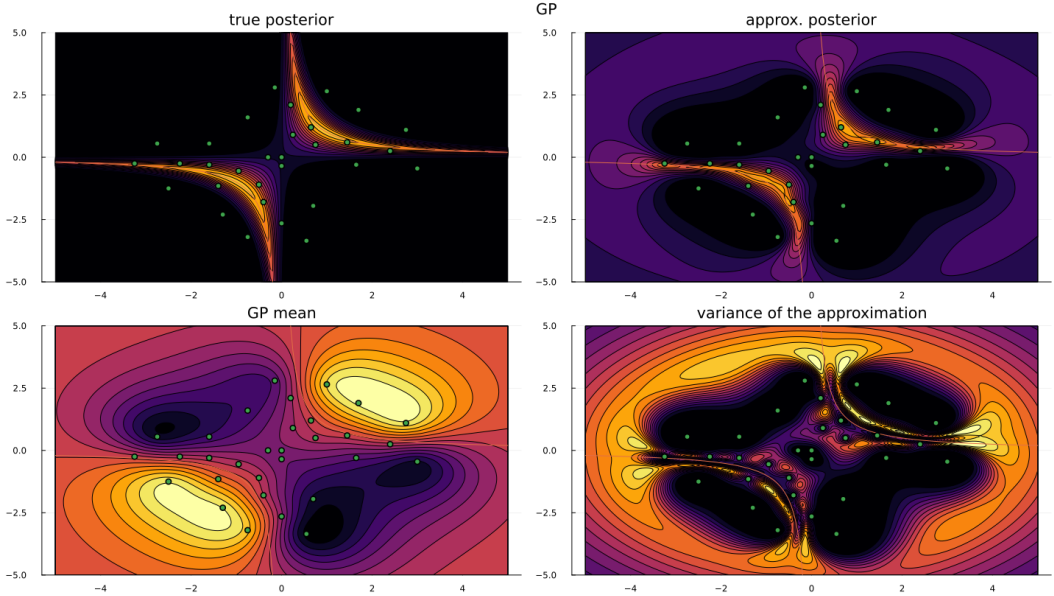
How it works in practice



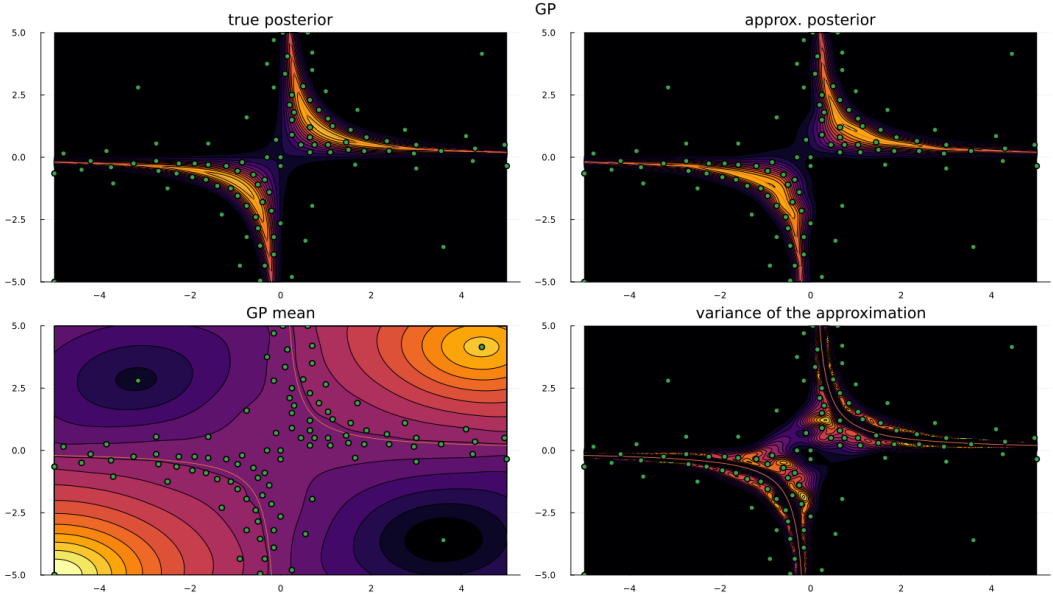
How it works in practice



How it works in practice



How it works in practice



Structured Data

Expensive experiments may have outcomes in different form:

Malware reports are JSON files: 10kb – 500MB

Looking for classifier: $y = f(\text{JSON}), y \in \{0, 1\}$ done by Neural Networks

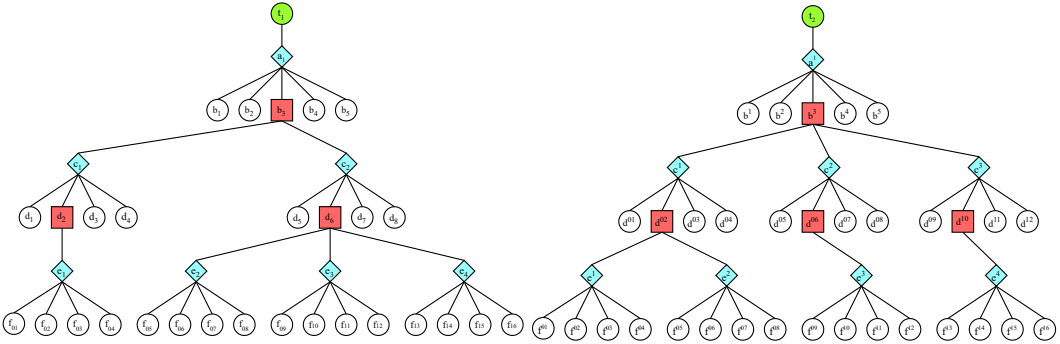
Label problem: supervised training works extremely well for large number of the data pairs.

- Malware evolves:
  every day the system receives thousands of attempts to smuggle a new malware
- Human analysts can no process thousand samples per day
- Selection of the most "informative" for the modeling

How to define a kernel? $K(\text{JSON}, \text{JSON}')$

Recursive definition of a metric on trees

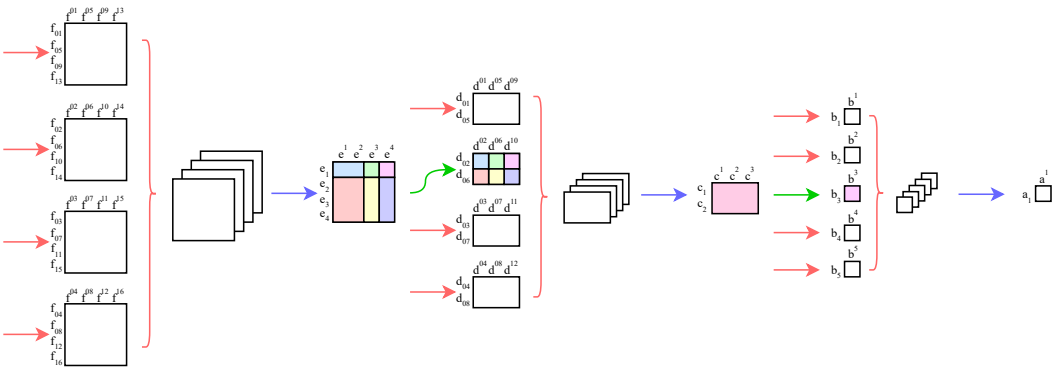
Consider two trees:



Their metric is build by a recursive combination of:

- leaf metrics (discrete, continuous, string)
- (weighted) product metrics
- Set metrics (Wasserstein, Chamfer)

Composite metric with 23 parameters (Mutagenesis dataset)



The goal is to lean the metric parameters

- in supervised setting, standard distance-based classifiers (kNN, SVM)
  - hyper-parameter search by CV
- in unsupervised, we can define kernel methods (GP)

GP on heterogeneous trees

Kernel

$$K(J, J') = \exp\left(-\frac{1}{2}|D_\theta(J, J')|\right)$$

and using Type-II maximum likelihood approach (Zorek et. al., to be published):

method/dataset	Mut.	Hepatitis	Chess	Genes	Webkp	Cora	MUTAG	BZR
HMIL classifier	87.8	92.5	41.5	98.8	82.0	**85.3**	91.0	88.2
kNN-TED	86.5	64.0	36.4	44.2	46.0	27.3	add	add
TMD	—	—	—	—	—	—	92.2	85.5[2]
kNN-HD (ours)	95.5	**94.0**	**52.5**	**100**	**86.1**	85.2	92.8	88.6
SVM-HD (ours)	**96.4**	74.3	42.4	**100**	85.3	80.6	**93.7**	**89.8**
GP-HD (ours)	91.9	84.3	41.2	93.6	add	add	75.7	87.3

Table 2: Classification experiment results. Results are reported using an accuracy score. For datasets MUTAG and BZR homogeneous graphs were unrolled to trees with depth L=4.

Potential improvement: Bayesian estimate?

Open problems

Conventional BO:

- Locality proved to be essential!
  - Why only rectangular boxes? CMA?
- Non-stationary processes $\approx$ Neural Networks

Posterior probability via BO:

- Same problem with non-stationarity
  - interpret TR as components of a mixture?

GPs on heterogenous tree structured data

- Sensitivity to kernel choice
- Poor hyper-parameter estimates