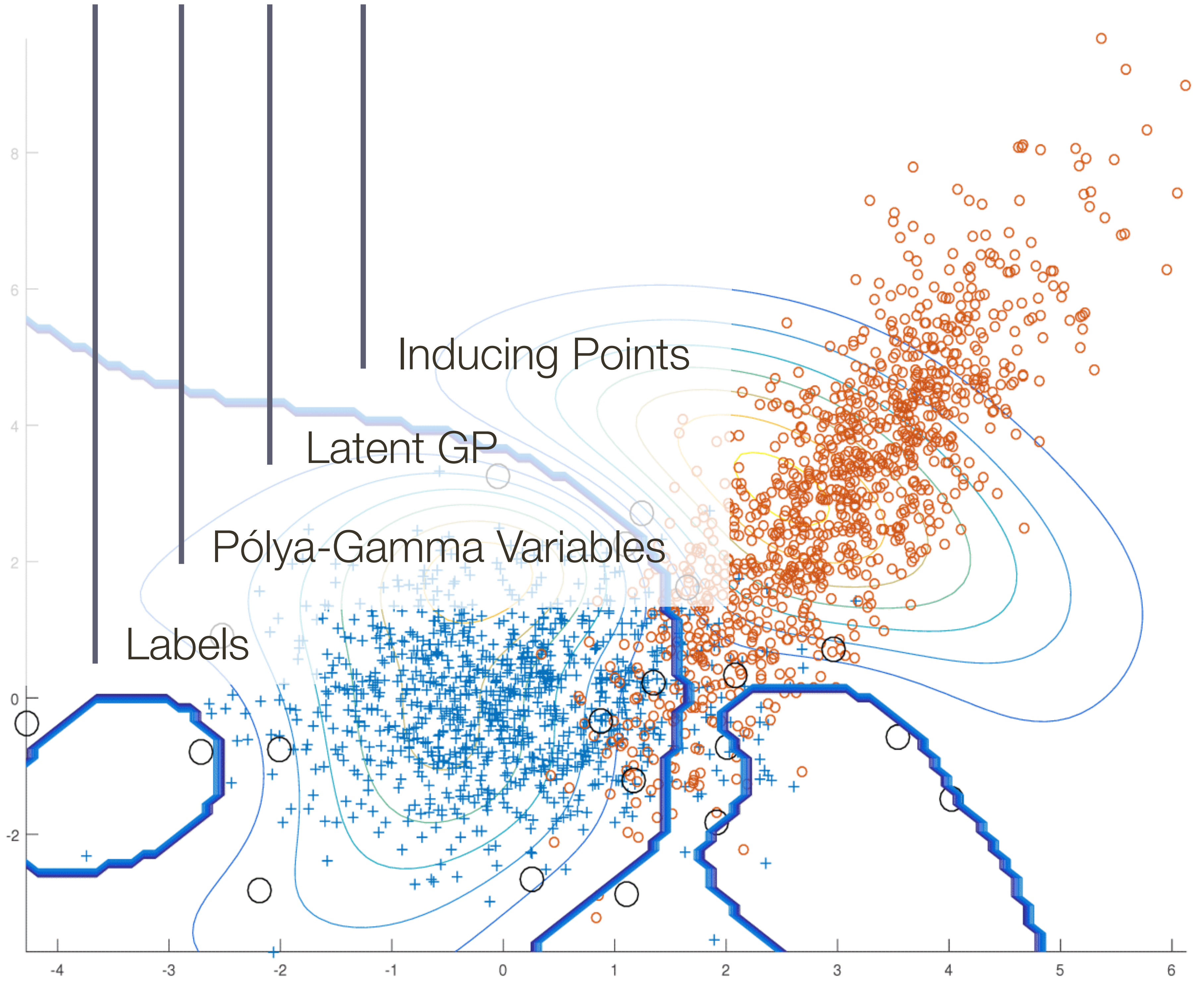
Scalable Logit Gaussian Process Classification

Florian Wenzel
TU Kaiserslautern & HU Berlin / Germany

Joint work with:
Théo Galy-Fajou, Christian Donner,
Marius Kloft and Manfred Opper

$$p(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{f})p(\boldsymbol{\omega})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$$



GP Classification

Training Data

$$X = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$$
$$y = (y_1, \ldots, y_n) \in \{-1, 1\}^n$$

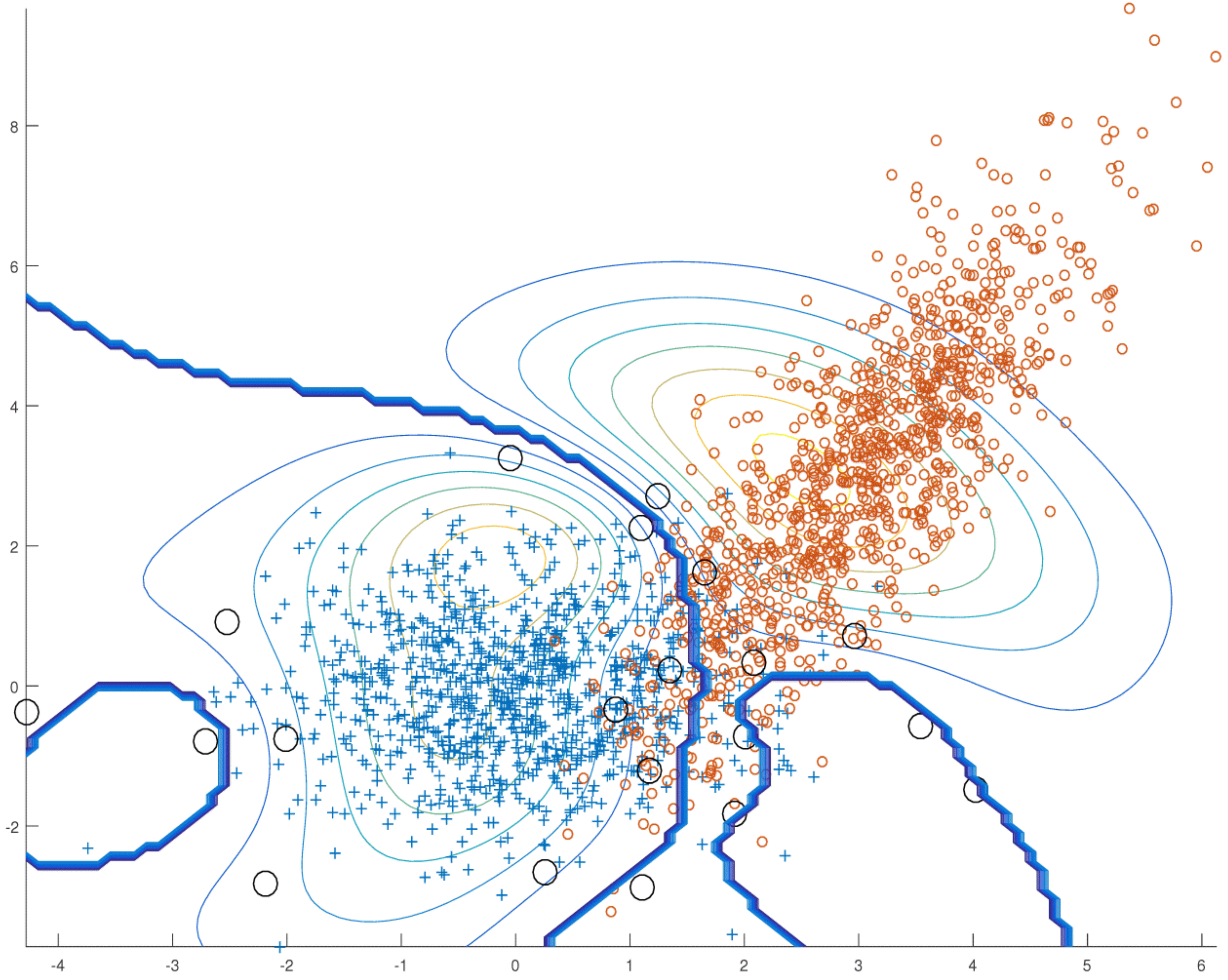
Model

$$p(\boldsymbol{y}|\boldsymbol{f}, X) = \prod_{i=1}^{n} \sigma(y_i f(\boldsymbol{x}_i))$$

$$p(\boldsymbol{f}|X) = \mathcal{N}(\boldsymbol{f}|\mathbf{0}, K_{nn})$$

Using Logit Link

$$\sigma(z) = (1+\exp(-z))^{-1}$$



Florian Wenzel wenzelfl@hu-berlin.de

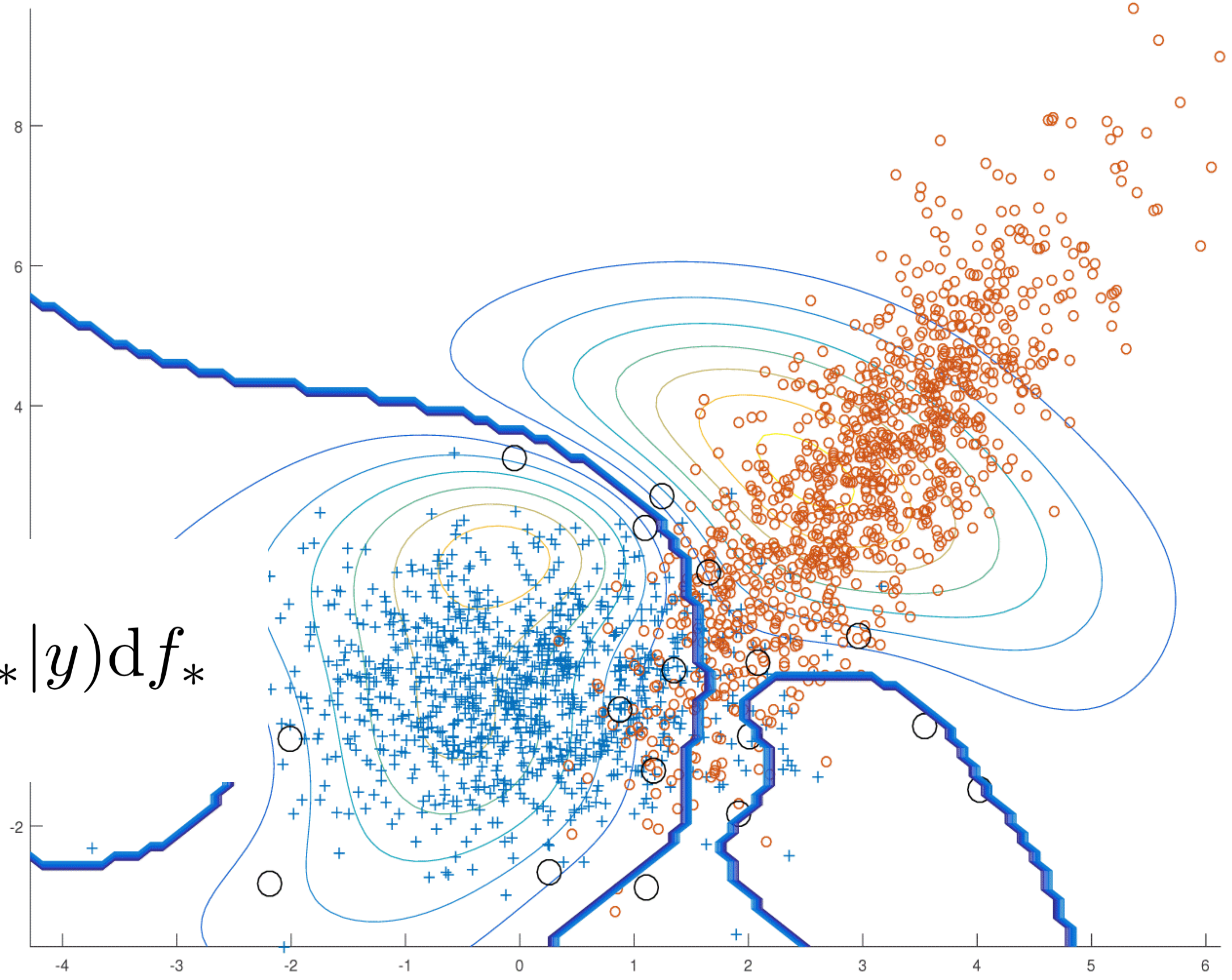
GP Classification

Goal: compute posterior

$$p(f|y,X)$$

Prediction

$$p(y_* = 1|y) = \int \sigma(f_*)p(f_*|y)\mathrm{d}f_*$$

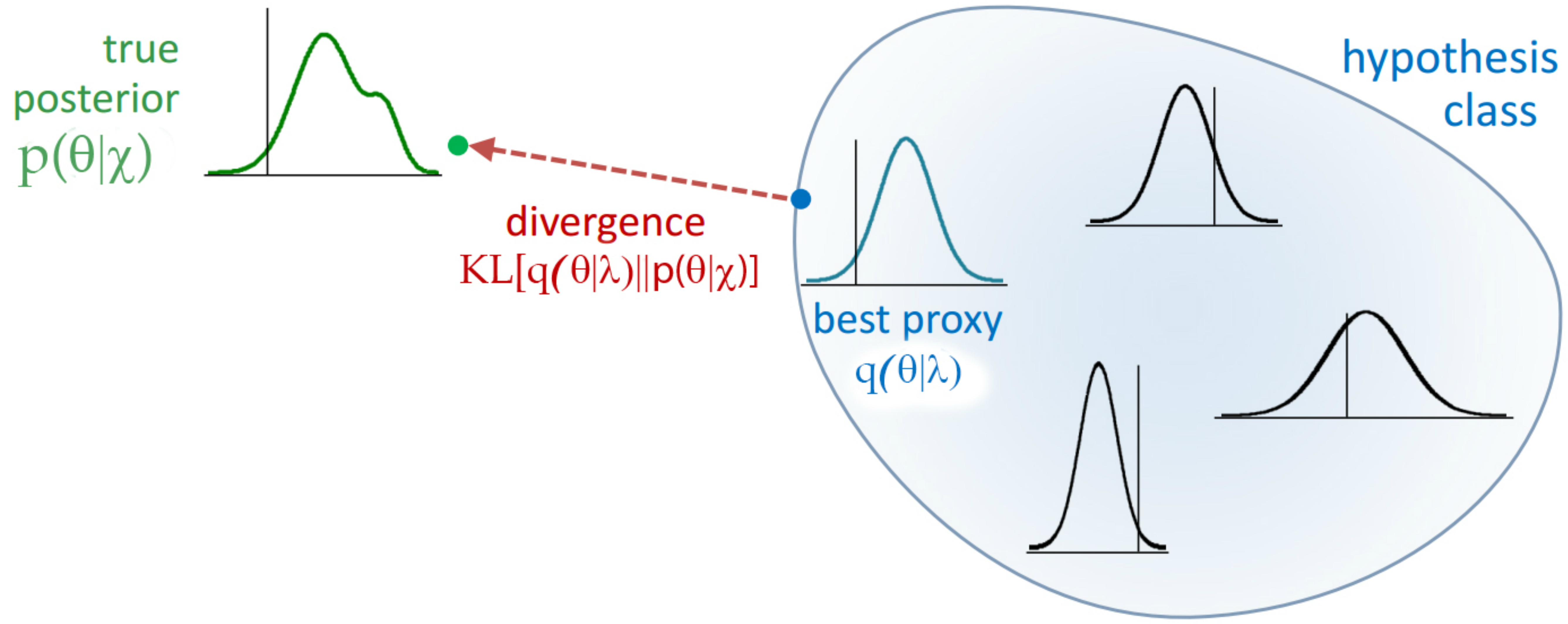


Florian Wenzel wenzelfl@hu-berlin.de

Posterior is intractable

Approximate posterior using variational inference

$$p(f|y, X) \approx q(f)$$



Goals:

- **Faster Algorithm**: efficient closed-form updates
- **Scalability** (millions of data points)

Efficient Updates?

Pólya-Gamma Data Augmentation

How to deal with the non-conjugate logistic likelihood function?

$$p(\boldsymbol{y}|\boldsymbol{f}, X) = \prod_{i=1}^{n} \sigma(y_i f(\boldsymbol{x}_i))$$

Idea:

$$\begin{aligned}
\sigma(z_i) &= (1 + \exp(-z_i))^{-1} \\
&= \frac{\exp(\frac{1}{2} z_i)}{2\cosh(\frac{z_i}{2})} \\
&= \frac{1}{2} \int \exp\left(\frac{z_i}{2} - \frac{z_i^2}{2}\omega_i\right) p(\omega_i) \mathrm{d}\omega_i
\end{aligned}$$

Pólya-Gamma Distribution

$$p(\omega_i) = \mathrm{PG}(\omega_i | 1, 0)$$

Defined by moment generating function

$$\mathbb{E}_{\mathrm{PG}(\omega|\,b,0)}[\exp(-\omega t)] = (\cosh^b(\sqrt{t/2}))^{-1}$$

[Polson & Scott (2013)]

Florian Wenzel wenzelfl@hu-berlin.de

Pólya-Gamma Data Augmentation

$$p(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\omega})p(\boldsymbol{f})p(\boldsymbol{\omega})$$
$$\propto \exp\left(\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{f} - \frac{1}{2}\boldsymbol{f}^\top \Omega \boldsymbol{f}\right) p(\boldsymbol{f})p(\boldsymbol{\omega})$$

In the augmented model the **full conditional distributions** are given in closed-form

p(f | …) is essentially GP Regression

Allows for **efficient closed-form updates** (later more)

Scalability?

Sparse Gaussian Processes

Inference in GPs typically scales $\mathcal{O}(n^3)$

Idea: Introduce m **inducing points** $\boldsymbol{u}$ to represent GP $\boldsymbol{f}$:

$$p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}\left(\boldsymbol{f}|K_{nm}K_{mm}^{-1}\boldsymbol{u}, \tilde{K}\right), \quad p(\boldsymbol{u}) = \mathcal{N}\left(\boldsymbol{u}|0, K_{mm}\right)$$

$$\tilde{K} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$$

Reduces complexity to $\mathcal{O}(m^3)$

[Snelson & Ghahramani 2006; Hensman+ 2013]

Final Model

Scalable Logit GP Classification Model

$$p(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{f})p(\boldsymbol{\omega})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$$

Inducing Points

Latent GP

Pólya-Gamma Variables

Labels

Inference

Apply Variational Inference to marginal joint

$$p(y, \omega, u) \; = \; p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{u})p(\boldsymbol{\omega})p(\boldsymbol{u})$$

Inference

Variational Family

$$q(\boldsymbol{u}, \boldsymbol{\omega}) = q(\boldsymbol{u}) \prod_i q(\omega_i)$$

$$q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{\mu}, \Sigma)$$

$$q(\omega_i) = \mathrm{PG}(\omega_i|1, c_i)$$

Variational Bound (given in closed-form)

$$\begin{aligned}\log p(\boldsymbol{y}) &\geq \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})q(\boldsymbol{\omega})}[\log p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{f})] - \mathrm{KL}\left(q(\boldsymbol{u}, \boldsymbol{\omega})||p(\boldsymbol{u}, \boldsymbol{\omega})\right) \\ &= \sum_i \mathbb{E}_{p(f_i|u)q(\boldsymbol{u})q(\boldsymbol{\omega})}[\log p(y_i|\omega_i, f_i)] - \mathrm{KL}\left(q(\boldsymbol{u}, \boldsymbol{\omega})||p(\boldsymbol{u}, \boldsymbol{\omega})\right)\end{aligned}$$

Inference

Stochastic Variational Inference

Leads to SVI scheme based on **natural gradient updates**

Updates are given in **closed-form** (no sampling / numerical quadrature)

Efficient second-order optimization scheme

Experiments

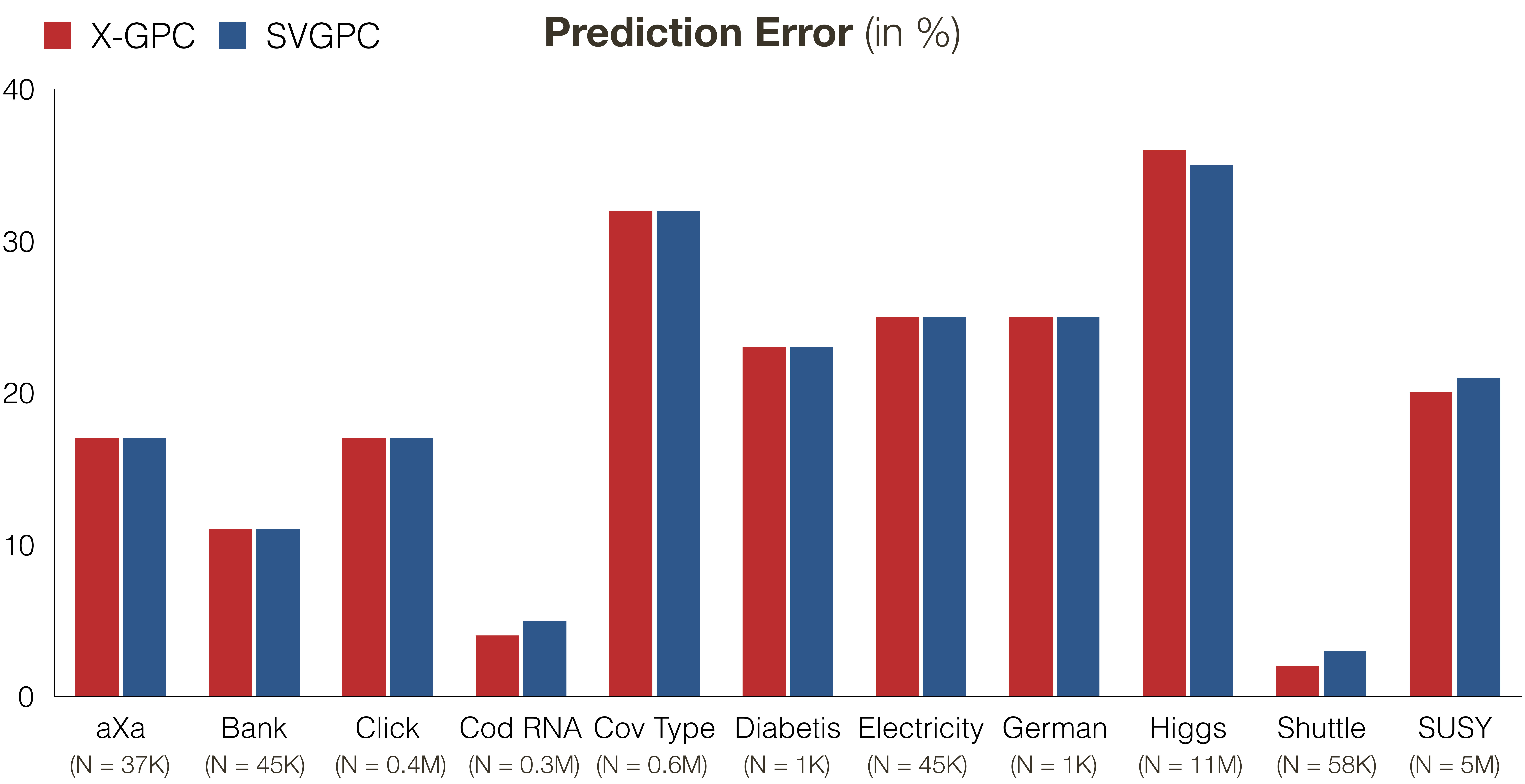
Competitors

X-GPC (our method)
Code: Julia

SVGPC
Code: GPflow (based on Tensorflow)

Scalable Variational Gaussian Process Classification
[Hensman+, AISTATS 2015], Code: github.com/GPflow

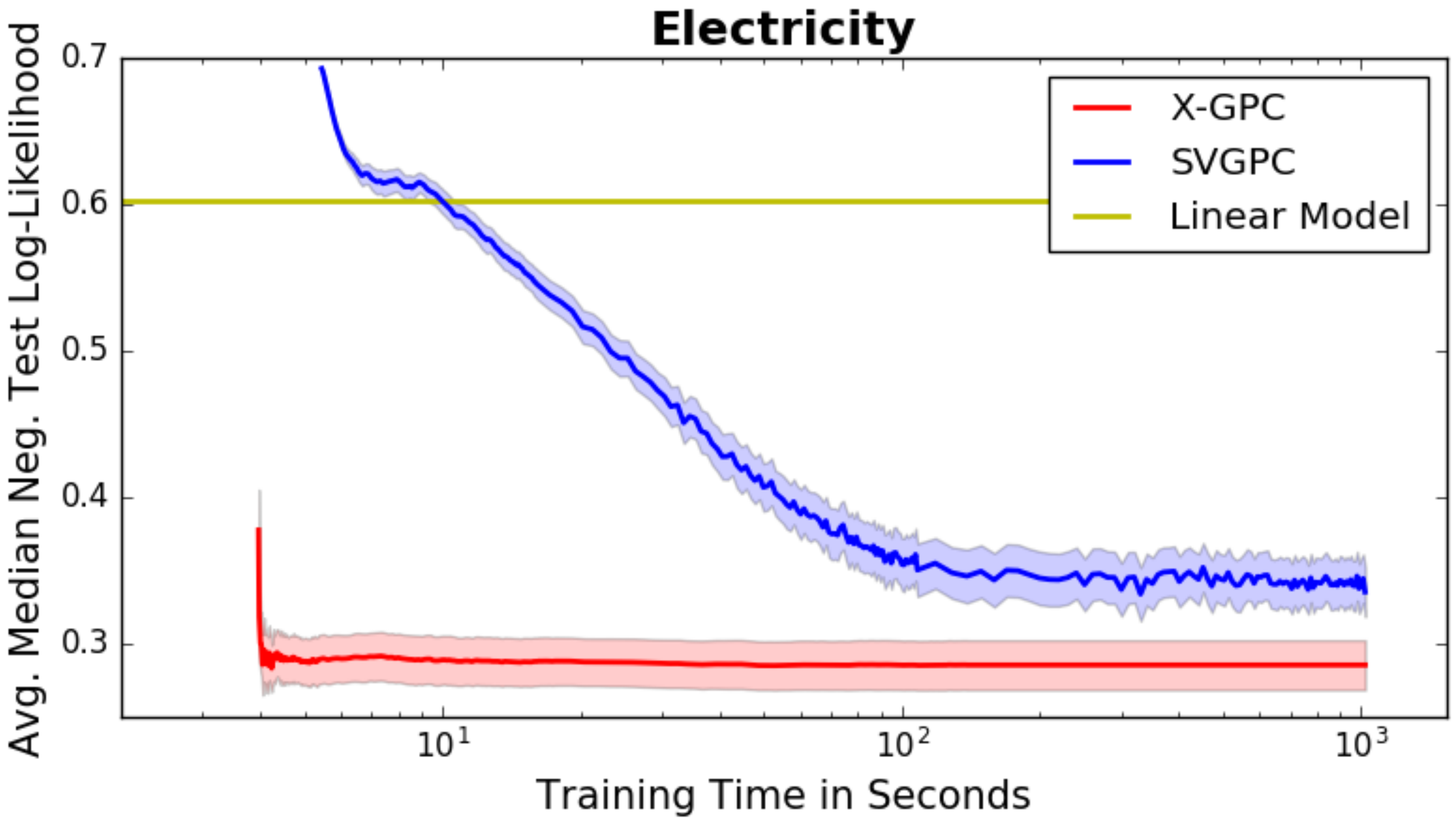
Prediction Error (in %)



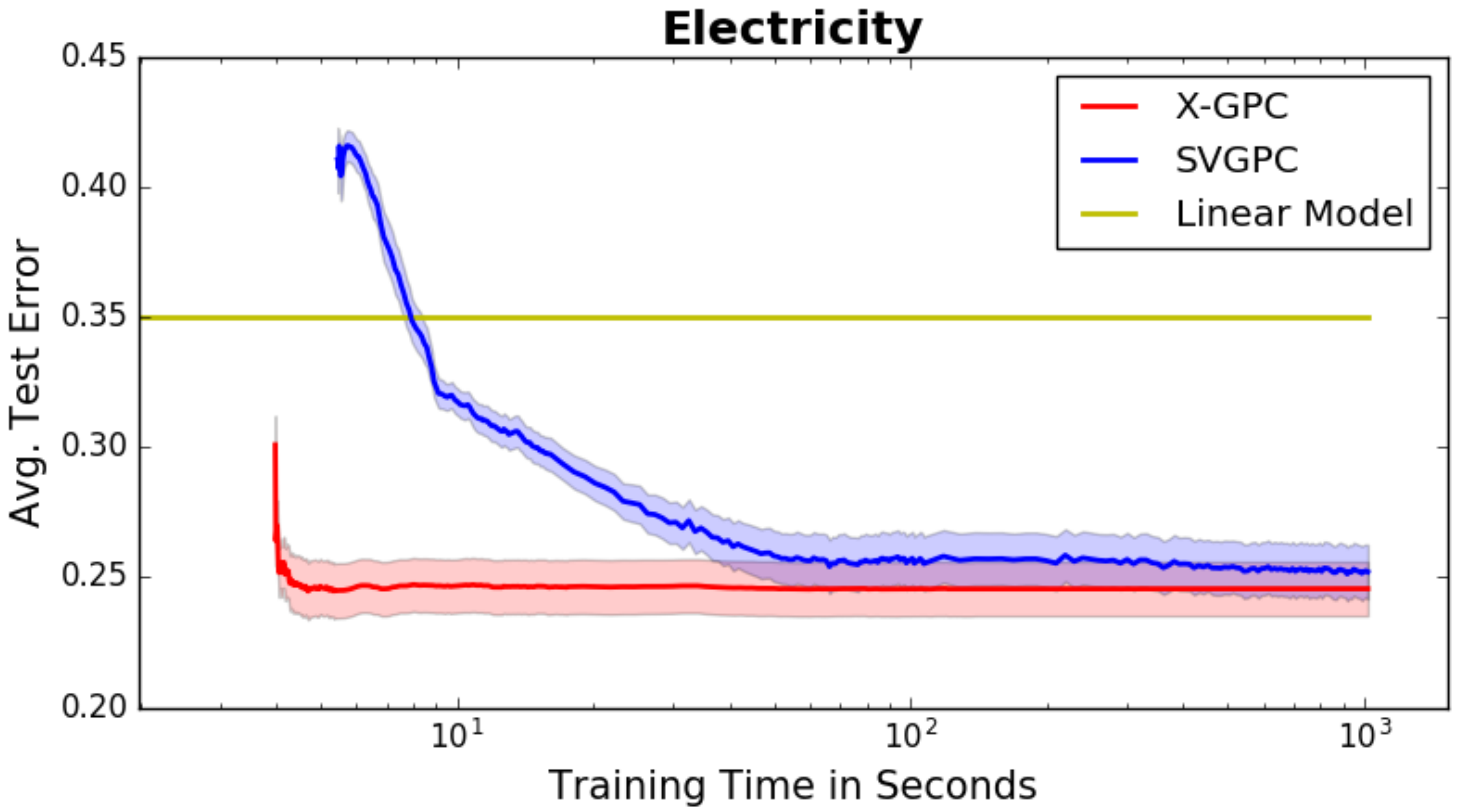
Florian Wenzel wenzelfl@hu-berlin.de


X-GPC
SVGPC
Run Time in sec (log scale!)
10^4
10^3
10^2
10^1
1
10^{-1}
aXa (N = 37K)
Bank (N = 45K)
Click (N = 0.4M)
Cod RNA (N = 0.3M)
Cov Type (N = 0.6M)
Diabetis (N = 1K)
Electricity (N = 45K)
German (N = 1K)
Higgs (N = 11M)
Shuttle (N = 58K)
SUSY (N = 5M)

Predictive Log-Likelihood on Test Set



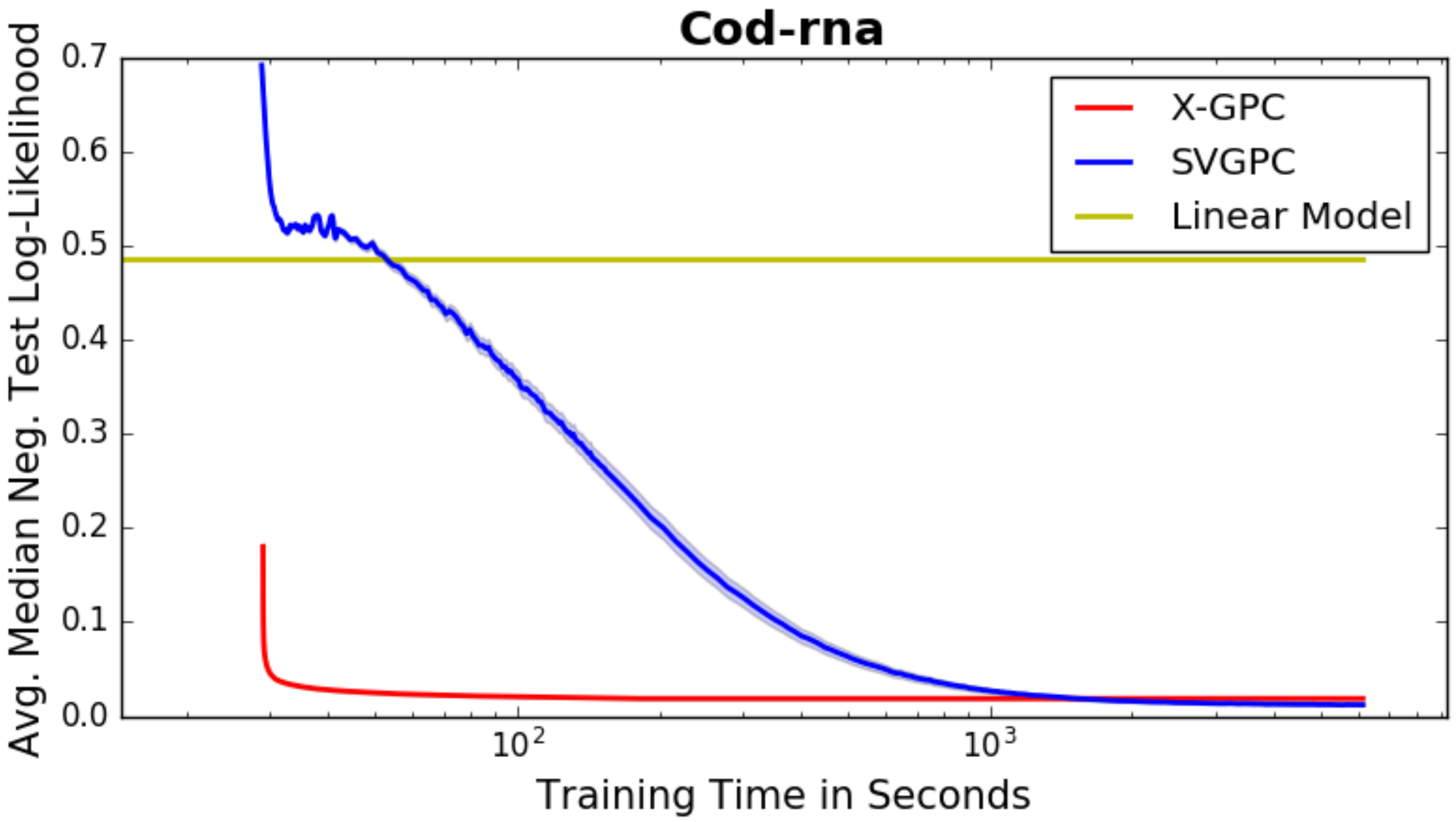
Prediction Error on Test Set



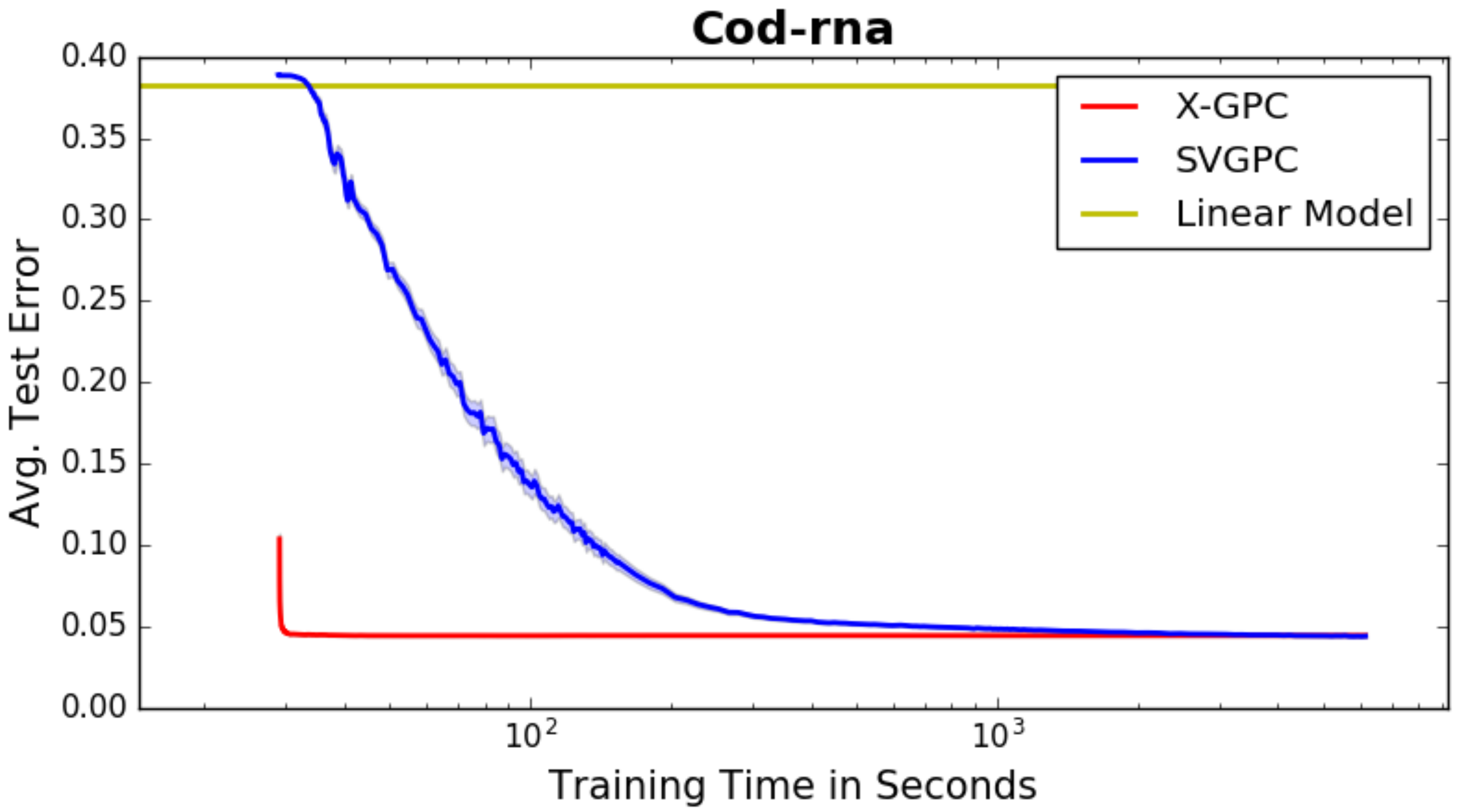
(45K points)

Florian Wenzel wenzelfl@hu-berlin.de

Predictive Log-Likelihood on Test Set



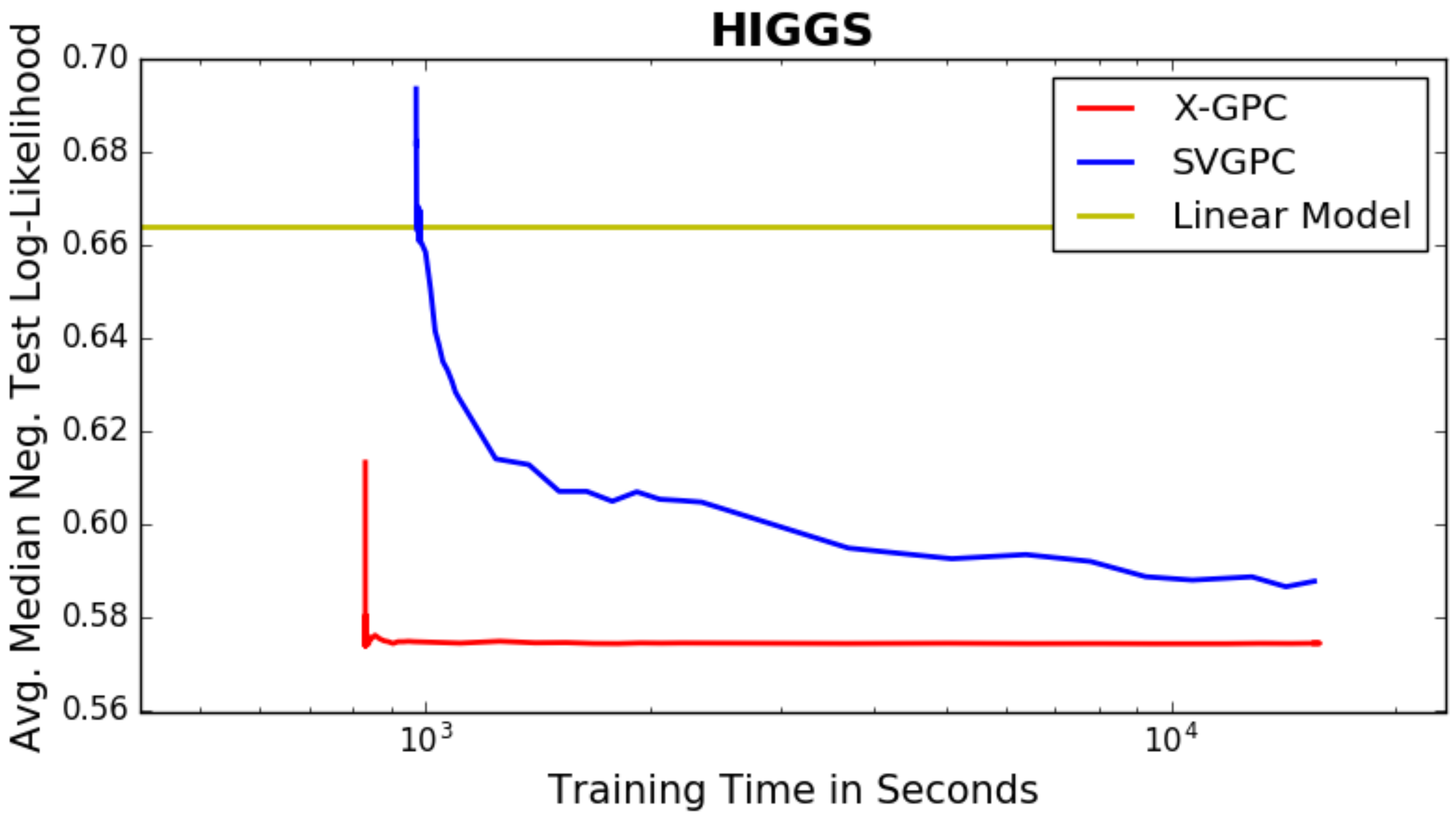
Prediction Error on Test Set



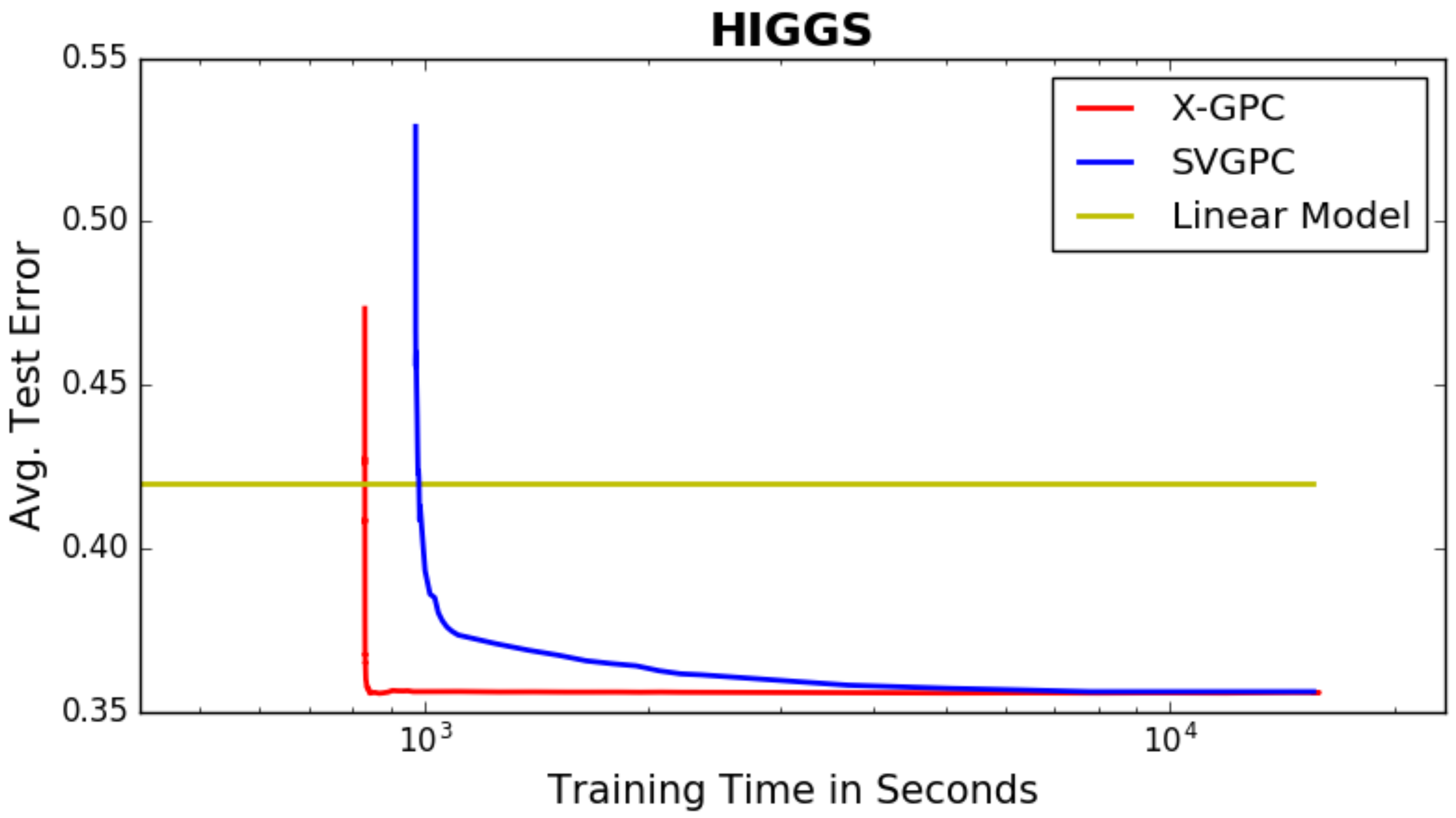
(343K points)

Florian Wenzel wenzelfl@hu-berlin.de

Predictive Log-Likelihood on Test Set



Prediction Error on Test Set



(11M points)

Florian Wenzel wenzelfl@hu-berlin.de

Conclusion

- We propose a fast **Gaussian process classification** method building on **Pólya-Gamma data augmentation and inducing points**.
- **Speedups of up to two orders of magnitude** while being competitive in terms of prediction performance.
- Scales to **millions of data points.**

Future Work

- Scalable Multi-Class GP Classification

Scalable Logit Gaussian Process Classification

Florian Wenzel[1,3], Théo Galy-Fajou[2], Christian Donner[2], Marius Kloft[3] and Manfred Opper[2]

[1] HU Berlin, [2] TU Berlin, [3] TU Kaiserslautern

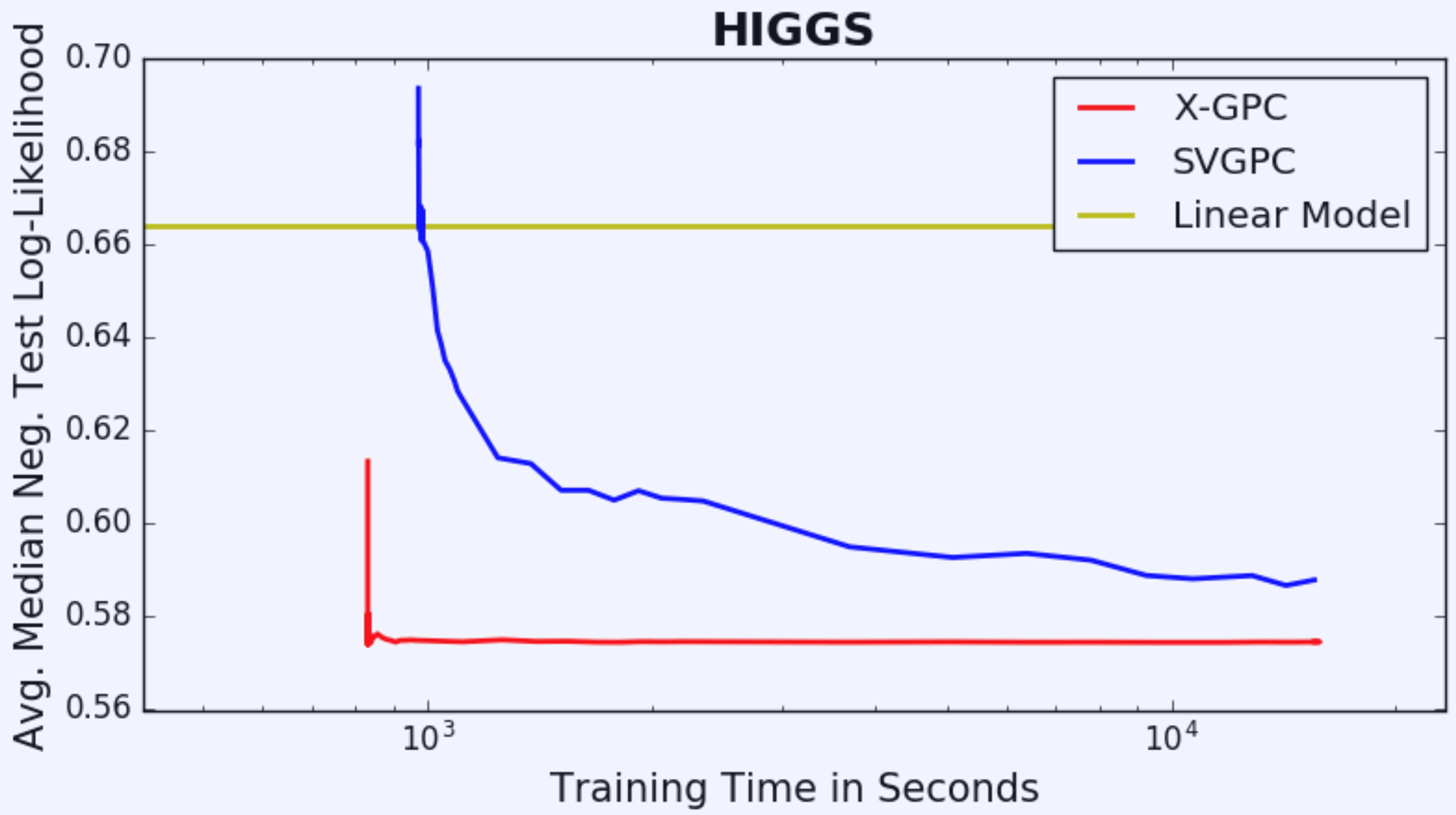
$$p(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{f})p(\boldsymbol{\omega})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$$

Inducing Points

Latent GP

Pólya-Gamma Variables

Labels



Contact:

wenzelfl@hu-berlin.de

www.florian-wenzel.de