Scalable Logit Gaussian Process Classification

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

[1]Humboldt-Universität zu Berlin, [2]TU Berlin, [3]TU Kaiserslautern

Summary

- We present a Gaussian process classification method building on **Pólya-Gamma data augmentation** and **inducing points**.
- We develop a fast **stochastic variational inference** algorithm based on efficient **natural gradient updates** which are given in closed-form.
- **Speedups of up to two orders of magnitude** while being competitive in terms of prediction performance.

Gaussian Process Classification

- Data: $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \mathbb{R}^{d \times n}$ with labels $\boldsymbol{y} = (y_1, \dots, y_n) \in \{-1, 1\}^n$.
- Logit GP Classification Model:

$$p(y_i|\boldsymbol{f}, \boldsymbol{x}_i) = \sigma(y_i f(\boldsymbol{x}_i)) = (1 + \exp(-y_i f(\boldsymbol{x}_i)))^{-1}$$
$$\boldsymbol{f} \sim \mathrm{GP}(0, K)$$

Pólya-Gamma Data Augmentation

Pólya-Gamma Distribution

- $\omega \sim \mathrm{PG}(b, 0)$, $b > 0$ is defined by the moment generating function $\mathbb{E}_{\mathrm{PG}(\omega|\, b,0)}[\exp(-\omega t)] = (\cosh^b(\sqrt{t/2}))^{-1}$.
- Idea: write **logistic function** in terms of Pólya-Gamma variables

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

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

Pólya-Gamma Augmented Model:

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

Sparse Gaussian Process (Inducing Points)

- Inference in GPs is typically $\mathcal{O}(n^3)$.
- Scalable approximation by using a sparse GP representation with m inducing points $(Z_1, u_1), \dots, (Z_m, u_m)$ (reduced complexity $\mathcal{O}(m^3)$)
- $\boldsymbol{f}$ and the inducing variables $\boldsymbol{u} = (u_1, \dots, u_m)$ are connected via

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

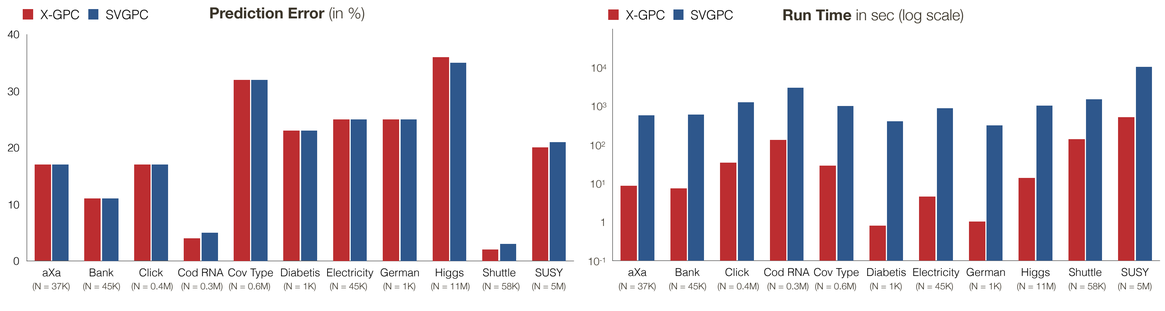
- Where $\tilde{K} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$.

Final Augmented Model

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

- $\boldsymbol{y}$ - labels
- $\boldsymbol{f}$ - latent decision function (modeled as GP)
- $\boldsymbol{u}$ - inducing points
- $\boldsymbol{\omega}$ - Pólya-Gamma variables

Performance and run time on several datasets



Stochastic Variational Inference

- We propose a fast and scalable inference algorithm.
- The lower bound is given in closed-form which enables efficient optimization.
- Efficient second-order optimization based on natural gradient updates.

Variational Approximation

- Apply VI to marginal joint distribution $p(y, \omega, u) = p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{u}) p(\boldsymbol{\omega}) p(\boldsymbol{u})$.
- Variational distribution: $q(\boldsymbol{u}, \boldsymbol{\omega}) = q(\boldsymbol{u}) q(\boldsymbol{\omega})$.
- With $q(\omega_i) = \mathrm{PG}(\omega_i|1, c_i)$ and $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{\mu}, \Sigma)$.

Variational Lower Bound

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

- Is given in closed-form (no sampling needed).

Parameter Updates

(based on mini-batch $\mathcal{S}$ of size s)

- **Pólya-Gamma parameters** (local):

$$c_i = \sqrt{\tilde{K}_{ii} + \boldsymbol{\kappa}_i \Sigma \boldsymbol{\kappa}_i^\top + \boldsymbol{\mu}^\top \boldsymbol{\kappa}_i^\top \boldsymbol{\kappa}_i \boldsymbol{\mu}}$$

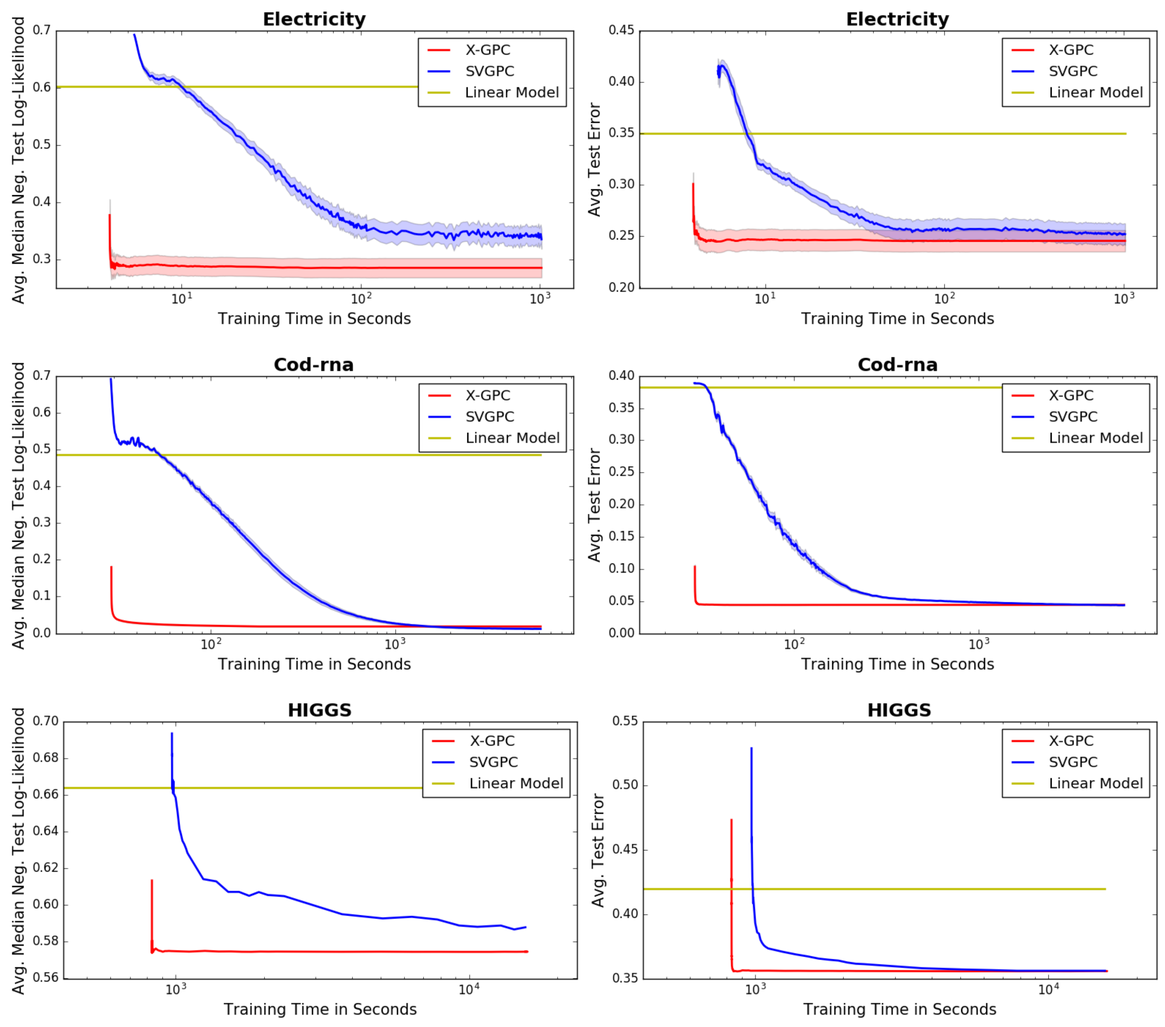
- **GP parameters** in natural parameterization (global):

$$\tilde{\nabla}_{\boldsymbol{\eta}_1} \mathcal{L}_{\mathcal{S}} = \frac{n}{2s} \boldsymbol{\kappa}_{\mathcal{S}}^\top \boldsymbol{y}_{\mathcal{S}} - \boldsymbol{\eta}_1$$
$$\tilde{\nabla}_{\eta_2} \mathcal{L}_{\mathcal{S}} = -\frac{1}{2}\left(K_{mm}^{-1} + \frac{n}{s} \boldsymbol{\kappa}_{\mathcal{S}}^\top \Theta_{\mathcal{S}} \boldsymbol{\kappa}_{\mathcal{S}}\right) - \eta_2$$

- Where $\boldsymbol{\kappa}_i = K_{im}K_{mm}^{-1}$, $\Theta = \mathrm{diag}(\boldsymbol{\theta})$ and $\theta_i = \frac{1}{4c_i} \tanh\left(\frac{c_i}{2}\right)$.

Experiments

Prediction performance as function of time



Contact:

- Mail: wenzelfl@hu-berlin.de
- Web: www.florian-wenzel.de