Scalable Multi-Class Gaussian Process Classification via Data Augmentation

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

[1]TU Berlin, Germany, [2]TU Kaiserslautern, Germany

Technische Universität Berlin

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

TL;DR

- We propose a **new scalable multi-class Gaussian process classification** approach building on a novel **modified softmax likelihood function**.
- This likelihood allows for a latent variable augmentation that leads to a **conditionally conjugate model** and enables **efficient variational inference** via **block coordinate ascent updates.**
- **Sparse Gaussian Processes** with independent inducing points and kernel parameters.

GP Multi-Class Problem

- Dataset with inputs $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ and labels $\boldsymbol{y} = (y_1, \dots, y_N)$ where $y_i \in \{1, \dots, C\}$
- Classical softmax : $p(y_i = k|\boldsymbol{f}_i) = \frac{\exp(f_i^k)}{\sum_{c=1}^C \exp(f_i^c)}$
- With C functions $\boldsymbol{f} = (f^1, \dots, f^C)$ with $f_i^c = f^c(\boldsymbol{x}_i)$ independent priors : $f^c \sim \mathrm{GP}(0, k^c)$

Logistic-softmax likelihood

$$\boxed{p(y_i = k|\boldsymbol{f}_i) = \frac{\sigma(f_i^k)}{\sum_{c=1}^C \sigma(f_i^c)}}$$

- $\sigma(x) = (1 + \exp(-x))^{-1}$ is the logistic function
- Softmax with transformation $h(f) = \log(\sigma(f))$
- Generalization of binary classification with $f^2 = -f^1$

Augmentations to conditional conjugacy

Augm. 1 : Gamma

$$\frac{1}{x} = \int_0^\infty \exp(-\lambda x) d\lambda \Longrightarrow p(y_i = k|\boldsymbol{f}_i) = \sigma(f_i^k) \int_0^\infty \exp\left(-\lambda^i \sum_{c=1}^C \sigma(f_i^c)\right) d\lambda^i$$

$$\Longrightarrow p(y_i = k, \lambda^i|\boldsymbol{f}_i) = \sigma(f_i^k) \prod_{c=1}^C \exp\left(-\lambda^i \sigma(f_i^c)\right)$$

Augm. 2 : Poisson

- σ is bounded : $\sigma(x) = 1 - \sigma(-x)$

$$\Longrightarrow p(y_i = k|\lambda^i, \boldsymbol{f}_i) = \sigma(f_i^k) \prod_{c=1}^C \exp\left(-\lambda^i (1 - \sigma(-f_i^c))\right)$$

$$\exp(x) = \sum_{n=1}^\infty x^n / n!$$

$$\Longrightarrow = \sigma(f_i^k) \prod_{c=1}^C \sum_{n_i^c=1}^\infty \exp(-\lambda^i) \left(\lambda^i \sigma(-f_i^c)^{n_c^i} / n_c^i!\right)$$

$$\Longrightarrow p(y_i = k, \{n_c^i\}_{c=1}^C|\lambda^i, \boldsymbol{f}_i) = \sigma(f_i^k) \prod_{c=1}^C \mathrm{Po}(n_c^i|\lambda^i) \sigma(-f_i^c)^{n_c^i}$$

Augm. 3 : Pólya-Gamma

$$\sigma^n(x) = 2^{-n} \int_0^\infty \exp\left(\frac{nx}{2} - \frac{\omega x^2}{2}\right) \mathrm{PG}(\omega|n, 0) d\omega, \quad \underbrace{\omega_1}_{\mathrm{PG}(a,s)} + \underbrace{\omega_2}_{\mathrm{PG}(b,s)} = \underbrace{\omega_3}_{\mathrm{PG}(a+b,s)}$$

$$\Longrightarrow \prod_{c=1}^C \int_0^\infty 2^{-(\delta_{kc} + n_i^c)} \exp\left(\frac{(\delta_{kc} - n_i^c) f_i^c)}{2} - \frac{(f_i^c)^2}{2} \omega_i^c\right) \mathrm{PG}(\omega_i^c|\delta_{kc} + n_i^c, 0) d\omega_i^c$$

Conjugate Likelihood

$$p\left(y_i = k|\{\omega_i^c\}_{c=1}^C, \{n_i^c\}_{c=1}^C, \lambda_i, \boldsymbol{f_i}\right)$$

$$\prod_{c=1}^C 2^{-(\delta_{kc} + n_i^c)} \exp\left(\frac{(\delta_{kc} - n_i^c) f_i^c)}{2} - \frac{(f_i^c)^2}{2} \omega_i^c\right)$$

Inference

Inducing Points (VFE)

$$p(\boldsymbol{f^c}) = \int p(\boldsymbol{f^c}|\boldsymbol{u^c}) p(\boldsymbol{u^c}) d\boldsymbol{u^c} \qquad \boldsymbol{u^c} \sim \mathrm{GP}(0, k^c)$$

Mean Field Approximation

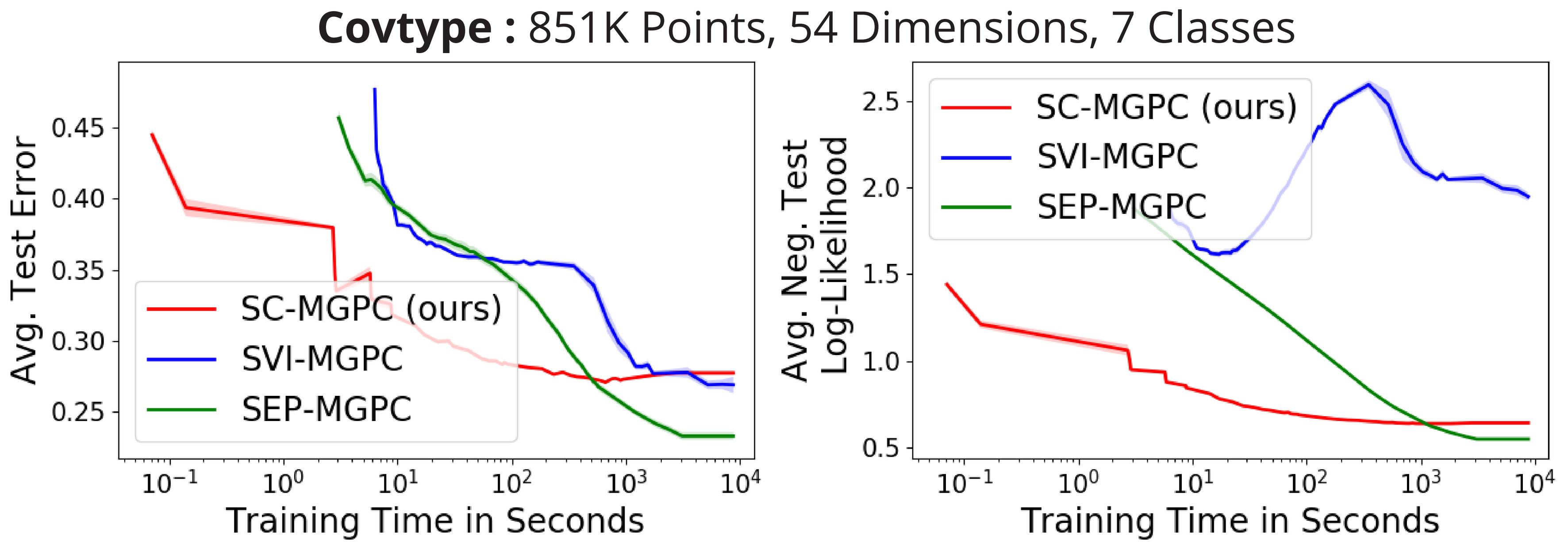
$$p(\boldsymbol{u}, \boldsymbol{\omega}, \boldsymbol{n}, \boldsymbol{\lambda}|\boldsymbol{y}) \approx \prod_{c=1}^C q(\boldsymbol{u^c}) \prod_{i=1}^N q(\lambda_i) q(n_i^c, \omega_i^c)$$

Block CAVI Updates

$$q^*(\theta) \propto \exp\left(\mathbb{E}_{q/\theta}\left[\log p(\theta, \boldsymbol{\Theta}_{/\boldsymbol{\theta}}, \boldsymbol{y})\right]\right)$$

Experiments

Covtype : 851K Points, 54 Dimensions, 7 Classes



DNA : 3386 Points, 180 Dimensions, 3 Classes