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

\[ p(y, \omega, f, u) = p(y | \omega, f)p(\omega)p(f | u)p(u) \]
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

\[ p(y|f, X) = \prod_{i=1}^{n} \sigma(y_i f(x_i)) \]

\[ p(f|X) = \mathcal{N}(f|0, K_{nn}) \]

Using Logit Link

\[ \sigma(z) = \frac{1}{1 + \exp(-z)} \]
GP Classification

Goal: compute posterior

\[ p(f|y, X) \]

Prediction

\[ p(y_* = 1|y) = \int \sigma(f_*) p(f_*|y) df_* \]
Posterior is intractable

**Approximate posterior** using variational inference

\[ p(f \mid y, X) \approx q(f) \]
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?

\[ p(y|f, X) = \prod_{i=1}^{n} \sigma(y_i f(x_i)) \]

Idea:

\[ \sigma(z_i) = (1 + \exp(-z_i))^{-1} \]

\[ = \frac{\exp\left(\frac{1}{2} z_i\right)}{2 \cosh\left(\frac{z_i}{2}\right)} \]

\[ = \frac{1}{2} \int \exp\left(\frac{z_i}{2} - \frac{z_i^2}{2} \omega_i\right) p(\omega_i) d\omega_i \]

[Polson & Scott (2013)]

Pólya-Gamma Distribution

\[ p(\omega_i) = \text{PG}(\omega_i|1, 0) \]

Defined by moment generating function

\[ \mathbb{E}_{\text{PG}(\omega|b,0)}[\exp(-\omega t)] = (\cosh^b(t/2))^{-1} \]
Pólya-Gamma Data Augmentation

\[
p(y, \omega, f) = p(y|f, \omega)p(f)p(\omega)
\]

\[
\propto \exp\left(\frac{1}{2} y^\top f - \frac{1}{2} f^\top \Omega f\right) p(f)p(\omega)
\]

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

\(p(f|\ldots)\) is essentially GP Regression

Allows for **efficient closed-form updates** (later more)
Scalability?
Sparse Gaussian Processes

Inference in GPs typically scales $O(n^3)$

Idea: Introduce m **inducing points** $u$ to represent GP $f$:

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

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

Reduces complexity to $O(m^3)$

[Snelson & Ghahramani 2006; Hensman+ 2013]
Final Model
Scalable Logit GP Classification Model

\[ p(y, \omega, f, u) = p(y | \omega, f)p(\omega)p(f | u)p(u) \]

- Inducing Points
- Latent GP
- Pólya-Gamma Variables
- Labels
Inference
Apply Variational Inference to marginal joint

\[ p(y, \omega, u) = p(y|\omega, u)p(\omega)p(u) \]
Inference

Variational Family

\[ q(u, \omega) = q(u) \prod_i q(\omega_i) \]

\[ q(u) = \mathcal{N}(u \mid \mu, \Sigma) \]

\[ q(\omega_i) = \text{PG}(\omega_i \mid 1, c_i) \]

Variational Bound (given in closed-form)

\[ \log p(y) \geq \mathbb{E}_{p(f \mid u) q(u) q(\omega)}[\log p(y \mid \omega, f)] - \text{KL} \left( q(u, \omega) \parallel p(u, \omega) \right) \]

\[ = \sum_i \mathbb{E}_{p(f_i \mid u) q(u) q(\omega)}[\log p(y_i \mid \omega_i, f_i)] - \text{KL} \left( q(u, \omega) \parallel p(u, \omega) \right) \]
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](https://github.com/GPflow)
Prediction Error (in %)

- 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

Electricity

<table>
<thead>
<tr>
<th></th>
<th>X-GPC</th>
<th>SVGPC</th>
<th>Linear Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Median Neg. Test Log-Likelihood</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prediction Error on Test Set

Electricity

<table>
<thead>
<tr>
<th></th>
<th>X-GPC</th>
<th>SVGPC</th>
<th>Linear Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Test Error</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(45K points)
(343K points)
Predictive Log-Likelihood on Test Set

Prediction Error on Test Set

HIGGS

X-GPC
SVGPC
Linear Model

Avg. Median Neg. Test Log-Likelihood

Training Time in Seconds

HIGGS

X-GPC
SVGPC
Linear Model

Avg. Test Error

Training Time in Seconds

(11M points)
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

\[ p(y, \omega, f, u) = p(y | \omega, f)p(\omega)p(f | u)p(u) \]

Contact:
wenzelfl@hu-berlin.de
www.florian-wenzel.de