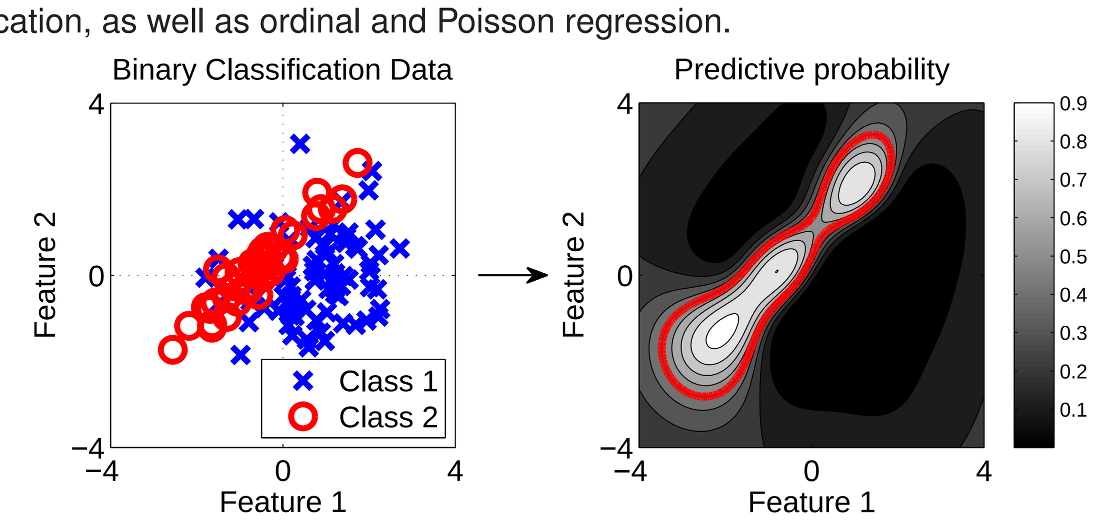
Fast Bayesian Inference for Non-conjugate Gaussian Process Regression

Mohammad Emtiyaz Khan, Shakir Mohamed, and Kevin P. Murphy
Department of Computer Science, University of British Columbia

Introduction

Motivation: Non-parametric regression using Gaussian processes is one of the most popular and widely used models in machine learning, with application to binary and multi-class classification, as well as ordinal and Poisson regression.



Problem: For real-valued outputs, we can combine the GP prior with a Gaussian likelihood and perform exact posterior inference in closed form. For problems, such as classification, the likelihood is no longer conjugate to the GP prior and **exact inference becomes intractable**.

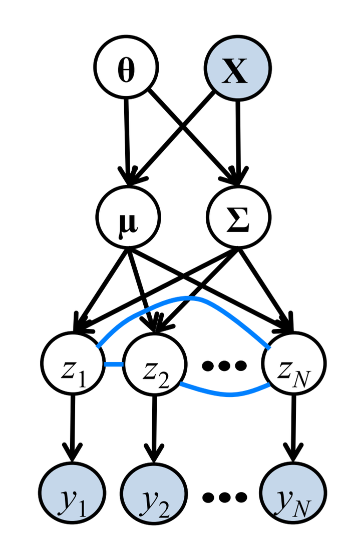
Solution: We make the following contributions for fast and tractable Bayesian inference.

- We derive a **concave lower bound** to the log marginal likelihood.
- We derive a **convergent algorithm** for lower bound maximization.

Advantages:

- Reduction of number of variational parameter from $O(N^2)$ to $O(N)$.
- **Fast convergence** due to concavity.
- Computation cost identical to EP, but **convergent and no numerical problems**.

Gaussian Process Regression



Given observation y_n with features $\mathbf{x}_n$, GPs use a nonlinear latent function $z(\mathbf{x}_n)$ to model y_n.

The GP prior $p(\mathbf{z}|\mathbf{X},\theta) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is characterized by,

- **Mean function:** e.g. zero mean function $\mu(\mathbf{x}) = 0$.
- **Covariance function:**, e.g. squared-exponential, $\Sigma(\mathbf{x}_i,\mathbf{x}_j) = \sigma^2 \exp[-(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)/(2s)]$.
- **Hyperparameters:** $\theta = (s, \sigma)$.

The likelihood is factorial where each $p(y_n|z_n)$ depends on the type of observations (see the table in the next column).

$$p(\mathbf{y}|\mathbf{z}) = \prod_{n=1}^{N} p(y_n|z_n)$$

Concave Lower Bounds

Variational lower bound: Computation of the marginal likelihood is intractable since the likelihood is not conjugate to the Gaussian prior. Using Jensen's inequality, we can obtain a lower bound to the marginal likelihood.

$$\begin{aligned}\mathcal{L}(\theta) &= \log \int p(\mathbf{y}|\mathbf{z})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{z} = \log \int \frac{p(\mathbf{y}|\mathbf{z})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})}\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})d\mathbf{z}\\ &\geq \int \mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V}) \log \frac{\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})} + \int \mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V}) \log p(\mathbf{y}|\mathbf{z})d\mathbf{z}\\ &\geq -KL\left[\mathcal{N}(\mathbf{m},\mathbf{V})||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})\right] + \sum_{n=1}^{N}\int \mathcal{N}(z_n|m_n,V_{nn}) \log p(y_n|z_n) d\mathbf{z}\end{aligned}$$

The second integral is not tractable for many distributions. We make use of tractable **local variational bounds**, such that $\mathbb{E}[\log p(y_n|z_n)] \geq f_b(y_n, m_n, V_{nn})$.

$$\mathcal{L}_J(\theta,\mathbf{m},\mathbf{V}) = \frac{1}{2}\left[\log|\mathbf{V}\boldsymbol{\Sigma}^{-1}| - \mathrm{tr}(\mathbf{V}\boldsymbol{\Sigma}^{-1}) - (\mathbf{m}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}) + N\right] + \sum_{n=1}^{N} f_b(y_n, m_n, V_{nn})$$

Concave Lower Bounds

Concave Bound:

- Our variational bound is strictly concave when f_b is **jointly concave** with respect to $\mathbf{m}, \mathbf{V}$.
- Given $\mathbf{V}$, optimization w.r.t. $\mathbf{m}$ is a **non-linear least-squares** function.
- Given $\mathbf{m}$, optimization w.r.t. $\mathbf{V}$ is a form of **covariance selection** or graphical Lasso.

But still $O(N^2)$ variational parameters!

Concave Local Variational Bound: List of non-conjugate likelihoods with concave local variational bounds (LVBs).

Type	Distribution	$p(y\|z)$	$\mathbb{E}[\log p(y\|z)]$	LVBs
Count	Poisson	$p(y=k\|z) = \frac{e^{-e^{\eta}}e^{kz}}{k!}$	$ym - \exp(m+v/2) - \log y!$	Analytical
Binary	Bernoulli logit	$p(y=1\|z) = \sigma(z)$	$ym - \mathbb{E}[\mathrm{llp}(z)]$	Piecewise Bounds
Categorical	Multinomial logit	$p(y=k\|\mathbf{z}) = e^{z_k - \mathrm{lse}(\mathbf{z})}$	$\mathbf{y}^T\mathbf{m} - \mathbb{E}[\mathrm{lse}(\mathbf{z})]$	Blei, Bouchard, etc.
Ordinal	Cumulative logit	$p(y\leq k\|z) = \sigma(\phi_k - z)$	$m - \mathbb{E}[\mathrm{llp}(-\phi_y + z) + \mathrm{llp}(-\phi_{y-1}+z)]$	Piecewise Bounds

Here, $\sigma(z) = 1/(1+e^{-z})$, $llp(x) = \log(1+\exp(x))$, lse is the log-sum-exp function, and ϕ_k are real numbers such that $\phi_1 < \phi_2 < \ldots < \phi_K$, for K ordered categories.

It is easy to see that $\mathbb{E}[\log p(y|z)]$ is concave for Poisson distribution, since $\exp(m + v/2)$ is convex. Concavity for other distributions can be obtained in a similar way by bounding the red (highlighted) part by a convex function.

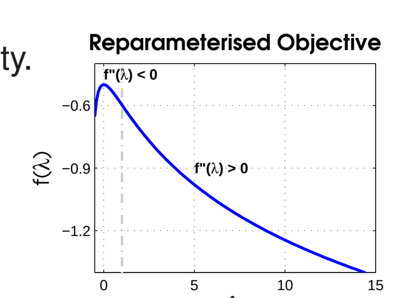
A Fast Convergent Algorithm

We reduce the number of parameters from $O(N^2)$ to $O(N)$ by using the structure of $\mathbf{V}$. Derivative with respect to $\mathbf{V}$ takes the following form,

$$\overset{\mathbf{K} = \mathbf{V}^{-1}}{\begin{bmatrix}\mathbf{K}_{11} & \mathbf{k}_{12}\\ \mathbf{k}_{21} & k_{22}\end{bmatrix}} - \overset{\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}}{\begin{bmatrix}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12}\\ \boldsymbol{\Omega}_{21} & \Omega_{22}\end{bmatrix}} + \overset{\mathrm{diag}(\frac{\partial f_b}{\partial v})}{\begin{bmatrix}\begin{smallmatrix}\ddots & \mathbf{0}\\ \frac{\partial f_b}{\partial \mathbf{v}_{11}} & \\ \mathbf{0} & \ddots\end{smallmatrix} & \mathbf{0}\\ \mathbf{0} & \frac{\partial f_b}{\partial v_{22}}\end{bmatrix}} = 0$$

The naive reparameterization $V = (\Sigma^{-1} + \lambda)^{-1}$ destroys concavity.

$$\begin{aligned} f(V) &= [\log(V\Sigma^{-1}) - V\Sigma^{-1}]/2 + f_b(y,m,V)\\ f(\lambda) &= [-\log(1+\Sigma\lambda) - (1+\Sigma\lambda)^{-1}]/2 + f_b(y,m,V)\\ \nabla_\lambda^2 f(\lambda) &= \frac{1}{2}[\Sigma/(1+\Sigma\lambda)]^2(\Sigma\lambda - 1) + \nabla_\lambda^2 f_b(y,m,V)\end{aligned}$$



We use two facts. First, $K_{ij} = \Omega_{ij}, \forall i \neq j$ and second, the following relation between $\mathbf{K}$ and $\mathbf{V}$,

$$\begin{bmatrix}\mathbf{V}_{11} & \mathbf{v}_{12}\\ \mathbf{v}_{21} & v_{22}\end{bmatrix} = \begin{bmatrix}\mathbf{K}_{11}^{-1} + \frac{\mathbf{K}_{11}^{-1}\mathbf{k}_{12}\mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}}{k_{22} - \mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}\mathbf{k}_{12}} & -\frac{\mathbf{K}_{11}^{-1}\mathbf{K}_{12}}{k_{22} - \mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}\mathbf{k}_{12}}\\ -\frac{\mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}}{k_{22} - \mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}\mathbf{k}_{12}} & \frac{1}{k_{22} - \mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}\mathbf{k}_{12}}\end{bmatrix}$$

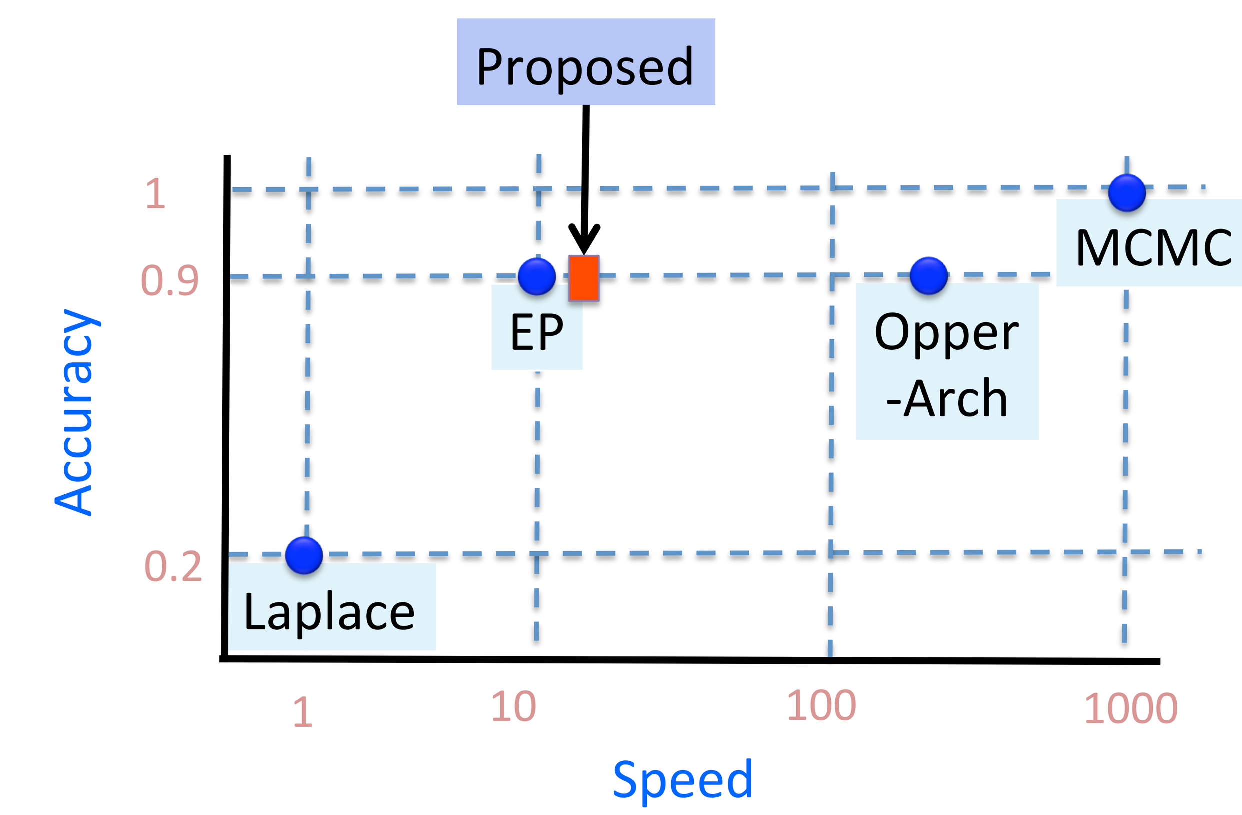
We fix all elements of $\mathbf{K}$ and update k_{22} via v_{22}. Define $\widetilde{k}_{22} = \mathbf{k}_{12}^T\mathbf{K}_{11}^{-1}\mathbf{k}_{12}$.

$$k_{22} = \Omega_{22} + 2\frac{\partial f_b}{\partial v_{22}} \quad \text{and} \quad v_{22} = 1/\left(k_{22} - \widetilde{k}_{22}\right) \quad \Rightarrow v_{22} \leftarrow 1/\left(\Omega_{22} + 2\frac{\partial f_b}{\partial v_{22}} - \widetilde{k}_{22}\right)$$

Proposed Algorithm	Expectation Propagation
Update **V**.	Update **m** and **V**
► Compute $\widetilde{k}_{22} \leftarrow k_{22} - 1/v_{22}$.	► Compute cavity parameters $\tilde{m}_{-2}, \tilde{v}_{-22}$.
► Update marginal variance v_{22}.	► Update site parameters $\tilde{m}_2, \tilde{v}_{22}$.
► Rank 1 update **V**.	► Rank 1 update **V**.
Update **m** with non-linear least-squares.	► Update **m**.

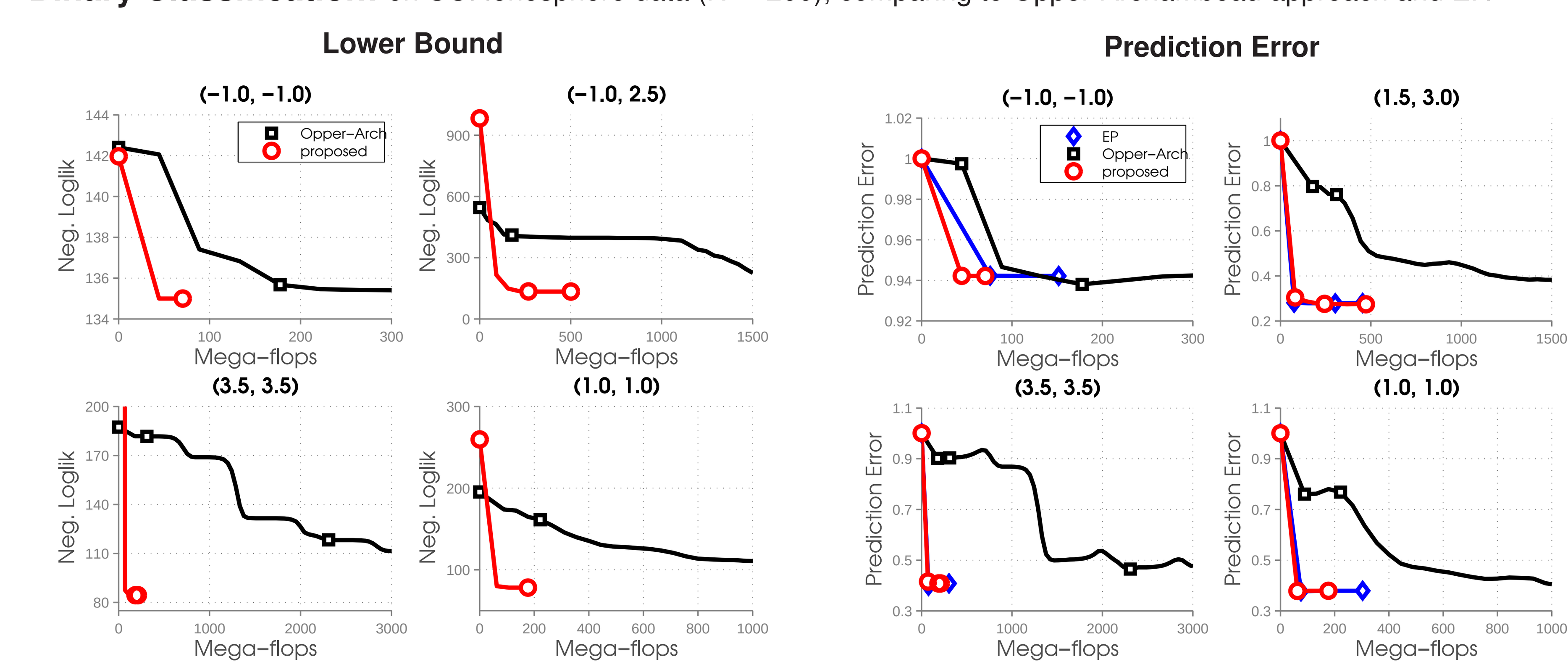
Conclusions

We add our proposed method to the comparison of Rasmussen and Nickisch (2008) for binary GPs. Our approach is as fast as EP and have the same accuracy as well, but unlike EP, our approach is principled, convergent, and free from numerical issue.

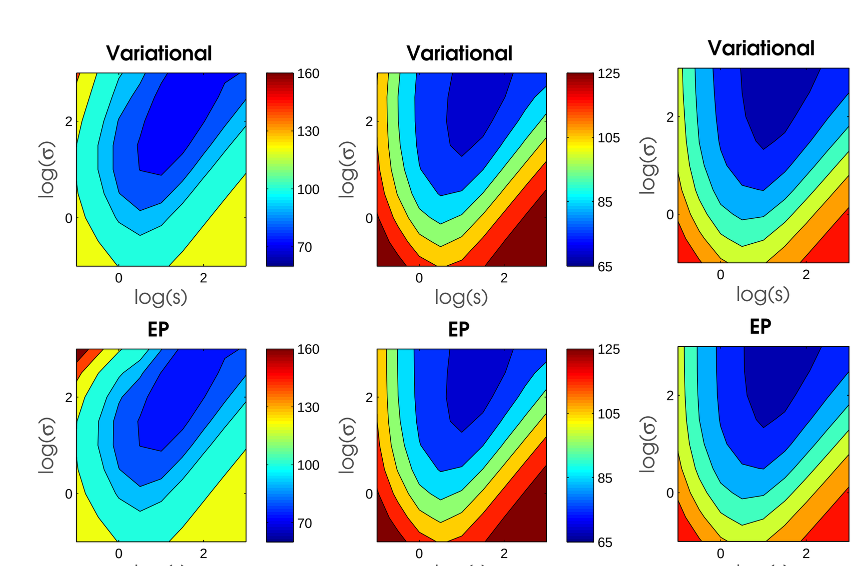


Results

Binary Classification: on UCI Ionosphere data ($N = 200$), comparing to Opper-Archambeau approach and EP.



EP and our algorithm give almost identical results.



KL Bound | EP Approx. | Prediction Error

Multi-class Classification
on UCI forensic glass data ($N = 214, K = 6$).