Riemannian Laplace approximations for Bayesian neural networks

Federico Bergamin, Pablo Moreno-Muñoz, Søren Hauberg, Georgios Arvanitidis

Section for Cognitive Systems, Technical University of Denmark

1 Approximate inference

Bayesian neural network

Computing $p(\theta|\mathcal{D})$ is the biggest challenge for Bayesian deep learning

$p(\theta)$ prior

$p(\mathcal{D}|\theta)$ likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$ posterior

Laplace approximation

I. Fit a neural network using SGD: $\theta_{\text{MAP}} = \arg\min \overbrace{-\log\left(p(\mathcal{D}|\theta)p(\theta)\right)}^{\mathcal{L}(\theta)}$

II. Approximate the posterior distribution by

$$q(\theta) = \mathcal{N}(\theta; \theta_{\text{MAP}}, H_\theta^{-1})$$ Several choices for approximating the Hessian

III. Compute the predictive distribution as

$$p(y|\mathbf{x}', \mathcal{D}) \approx \int p(y|\mathbf{x}', \mathcal{D}, \theta) q(\theta) \mathrm{d}\theta$$
$$= \frac{1}{S}\sum_{s=1}^{S} p(y|\mathbf{x}', \mathcal{D}, \theta_s), \;\; \theta_s \sim q(\theta)$$

Problem

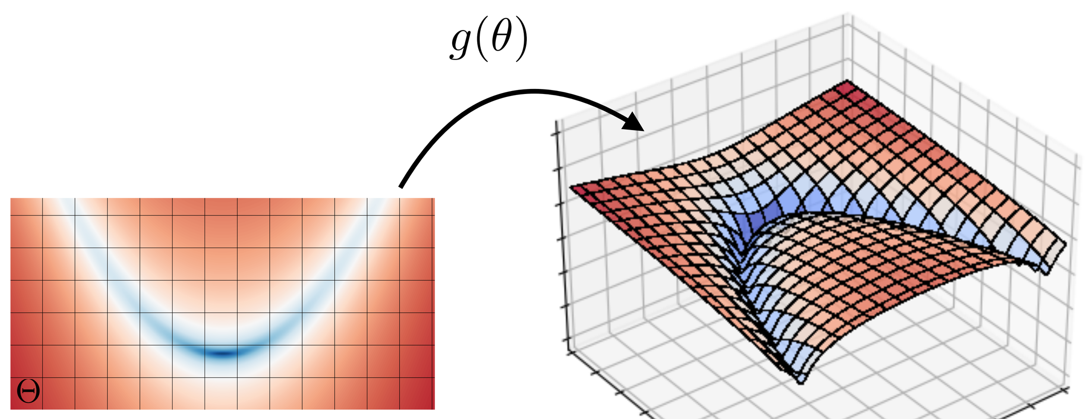
- The Gaussian distribution does not take into account the nonlinear structure of the posterior;
- Probability mass spreads in low posterior regions leading to suboptimal behaviour.

Research question

- Can we define a **flexible approximate posterior** that adapts to the nonlinear structure of the true posterior while sampling remains efficient?

2 How can differential geometry help us?

Assumption: the loss surface changes smoothly wrt to θ



Parameter space is the intrinsic coordinates of our manifold

Parametrization of the loss surface

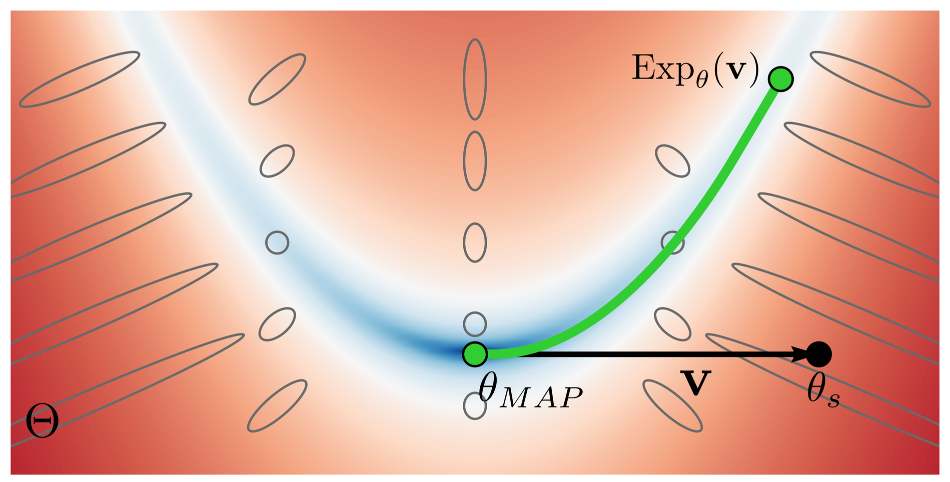
Immersion function $g(\theta) = [\theta, \mathcal{L}(\theta)]$

Riemannian metric $\mathbf{M}(\theta) = \mathbf{J}_g(\theta)^T \mathbf{J}_g(\theta)$

$$\boxed{\mathbf{M}(\theta) = \mathbb{I}_K + \nabla_\theta \mathcal{L}(\theta) \nabla_\theta \mathcal{L}(\theta)^T}$$

This gives us a notion of a local inner product in the intrinsic coordinates of the manifold

Using the Riemannian metric to solve exponential maps



Exponential map tend to stay in region of the parameter space that has low loss

Given a metric, we can compute a curve in the intrinsic space by solving the following ODE system

IVP: result is a geodesic **exponential map**

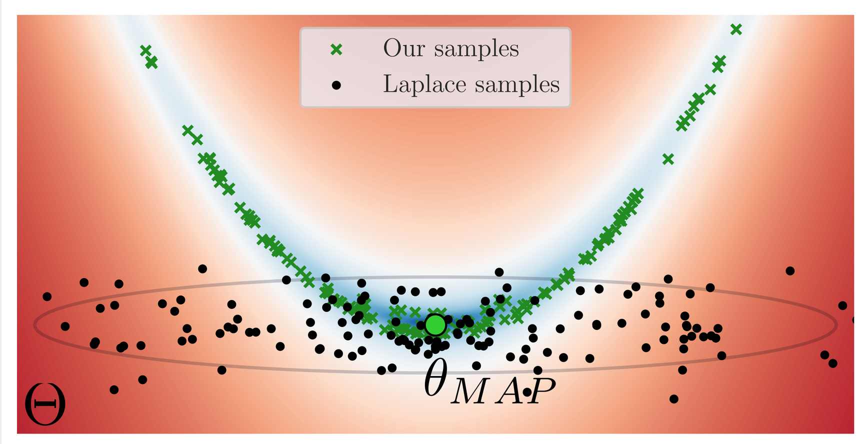
$$\ddot{c}(t) = -\frac{\mathbf{M}^{-1}(c(t))}{2}\left[2\left[\frac{\partial \mathbf{M}(c(t))}{\partial c_1(t)}, \ldots, \frac{\partial \mathbf{M}(c(t))}{\partial c_K(t)}\right] - \frac{\partial \mathrm{vec}[\mathbf{M}(c(t))]^T}{\partial c(t)}\right](\dot{c}(t) \otimes \dot{c}(t))$$

With our metric above, this simplifies to

Can be easily compute via automatic differentiation and using a IVP solver from scipy

$$\ddot{c}(t) = -\frac{\nabla_\theta \mathcal{L}(c(t))}{1 + \langle \nabla_\theta, \nabla_\theta \mathcal{L}(c(t)) \mathcal{L}(c(t)) \rangle}\langle \dot{c}(t), \mathbf{H}_\theta[\mathcal{L}](c(t))\dot{c}(t)\rangle$$

3 Our Riemannian Laplace approximation



Idea

Laplace samples can be used to generate geodesics that stay within the loss generating better samples.

I. Fit classic Laplace approximation $q(\theta) = \mathcal{N}(\theta; \theta_{\text{MAP}}, H_\theta^{-1})$

II. Sample $\theta_s \sim q(\theta)$

and compute initial velocities $\mathbf{v}_s = \theta_{MAP} - \theta_s$

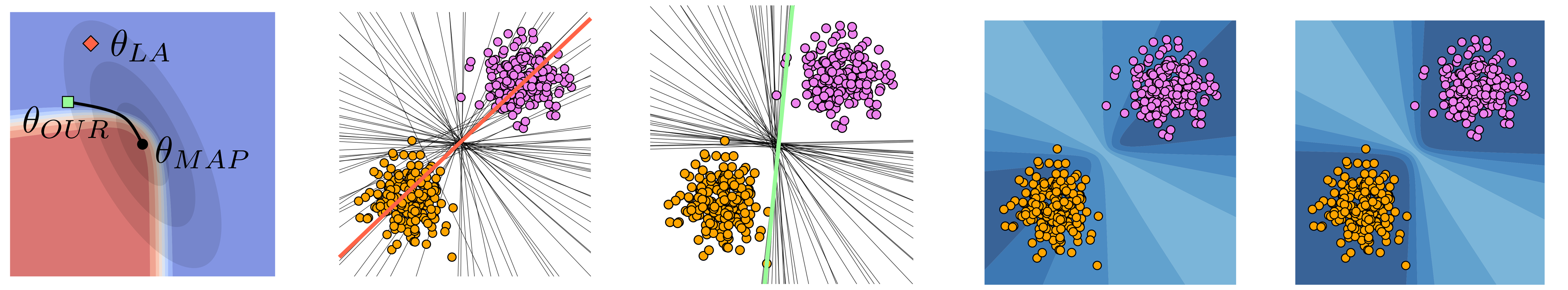
III. Compute predictive distribution as

$$p(y|\mathbf{x}', \mathcal{D}) \approx \int p(y|\mathbf{x}', \mathcal{D}, \theta) q(\theta) \mathrm{d}\theta$$
$$= \frac{1}{S}\sum_{s=1}^{S} p(y|\mathbf{x}', \mathcal{D}, \mathrm{Exp}_{\theta_{\text{MAP}}}(\mathbf{v_s}))$$

4 Experiments

Logistic Regression

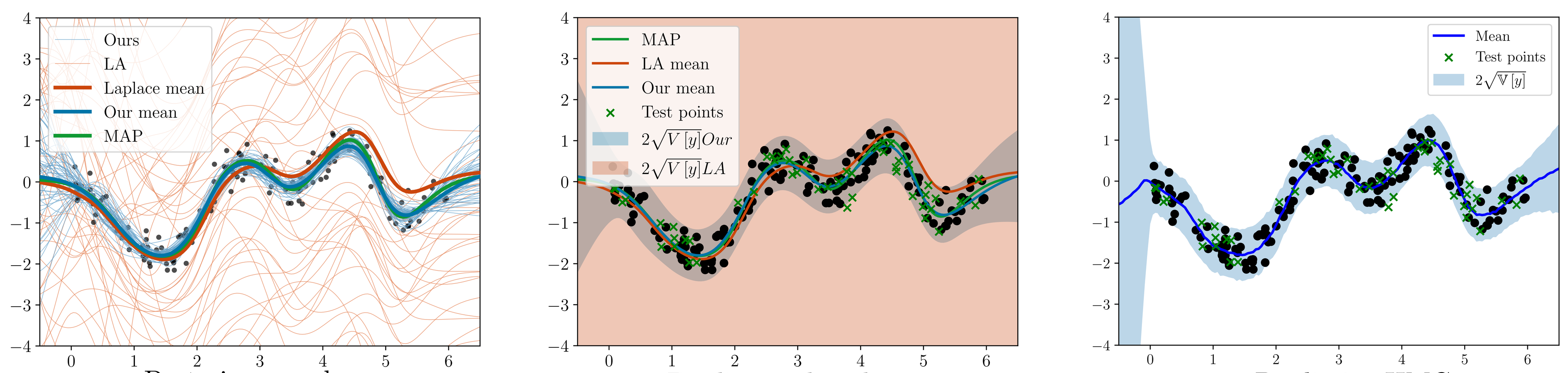
We consider a logistic regressor $\sigma(\mathbf{x}^T\theta + b)$ on a linearly separable dataset and and the posterior with respect to θ fixing the optimal b_*



True posterior | LA sample | Riemannian LA sample | Confidence LA | Confidence Riemannian LA

Regression

We consider a regression problem with DNN, where classic LA is known to perform poorly even if Hessian is not particularly ill-conditioned.



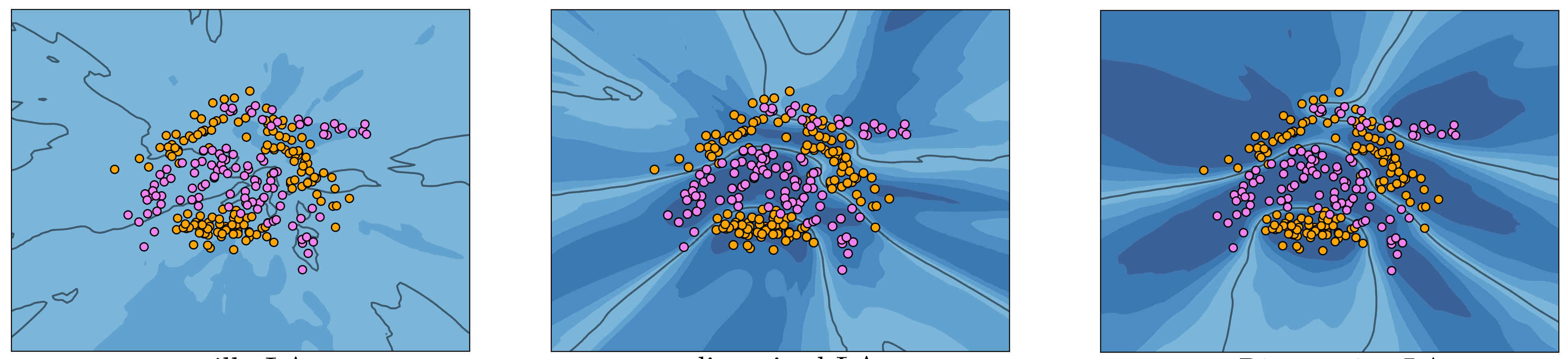
Posterior samples | Predictive distribution | Predictive HMC

1-layer NN with 15 hidden units

Neil Lawrence, "Variational Inference in Probabilistic Models", PhD thesis, 2000
Ritter, Botev, and Barber. A scalable laplace approximation for neural networks. ICLR 2018

Classification

We consider nonlinear classification problems and our method performs better than linearized LA.



vanilla LA | linearized LA | Riemannian LA

2-layer NN with 16 hidden units per layer

MORE EXPERIMENTS AND TABLES IN THE PAPER...

5 Future directions

Directions that open up from this work:

- Other Riemannian metrics with different properties can be considered;
- The proposed metric can be computed efficiently thanks to auto-differentiation techniques. Additional approximations e.g. KFAC or diagonal Hessian, can make it even faster to compute;
- Tailored-made solvers that exploit the structure and behaviour of our ODE system, i.e. geodesics start from low and move towards higher loss, can be beneficial for scaling this method to bigger datasets and networks.

CHECK OUT THE PAPER!

Neural Information Processing Systems (NeurIPS), 2023