4th Symposium on Advances in Approximate Bayesian Inference (AABI) 2022

Linearised Laplace Inference in Networks with Normalisation Layers and the Neural g-Prior

Javier Antorán, James Allingham, Dave Janz, Erik Daxberger, Eric Nalisnick, José Miguel Hernández-Lobato

UNIVERSITY OF CAMBRIDGE

About us

Javier Antorán

James Allingham

David Janz

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF ALBERTA

Max Planck Institute for Intelligent Systems

UNIVERSITY OF AMSTERDAM

Erik Daxberger

Eric Nalisnick

José Miguel Hernández-Lobato

Summary

- We identify a pitfall of the linearised Laplace model evidence for NNs with normalisation layers (batch norm, layer norm, etc) and provide a simple solution

- We propose a new prior for inference in tangent linear models (linearised NNs)

Preliminaries: linearised Laplace in 4 steps

1. Optimise a NN $f(x, w)$ to an optima $w^\star$ of some energy function $L_f(w) = G_f(w) + R(w)$ for

$$G_f(w) = \sum_i \log p(y_i \,|\, f(x_i, w)) \qquad\qquad R(w) = \log p(w)$$

2. Taylor expand f about $w^\star$: $\quad h(x, v) = f(x, w^\star) + \partial_w f(x, w^\star) \cdot (v - w^\star)$

$$L_h(v) = G_h(v) + R(v) \qquad G_h(v) = \sum_i \log p(y_i \,|\, h(x_i, v)) \qquad R(v) = \log p(v)$$

3. Approximate the posterior of the tangent model with a second order expansion about $w^\star$:

$$L(w^\star) + \underbrace{\cancel{\partial_v L_h(w^\star)}}_{0} \cdot (v - w^\star) + 0.5(v - w^\star)^T \partial_v^2 L_h(w^\star)(v - w^\star) \Big\} \text{Gaussian posterior}$$

$w^\star$ is a stationary point

4. For a Gaussian prior $\mathcal{N}(0, \Lambda^{-1})$, estimate model evidence or marginal log-likelihood (MLL)

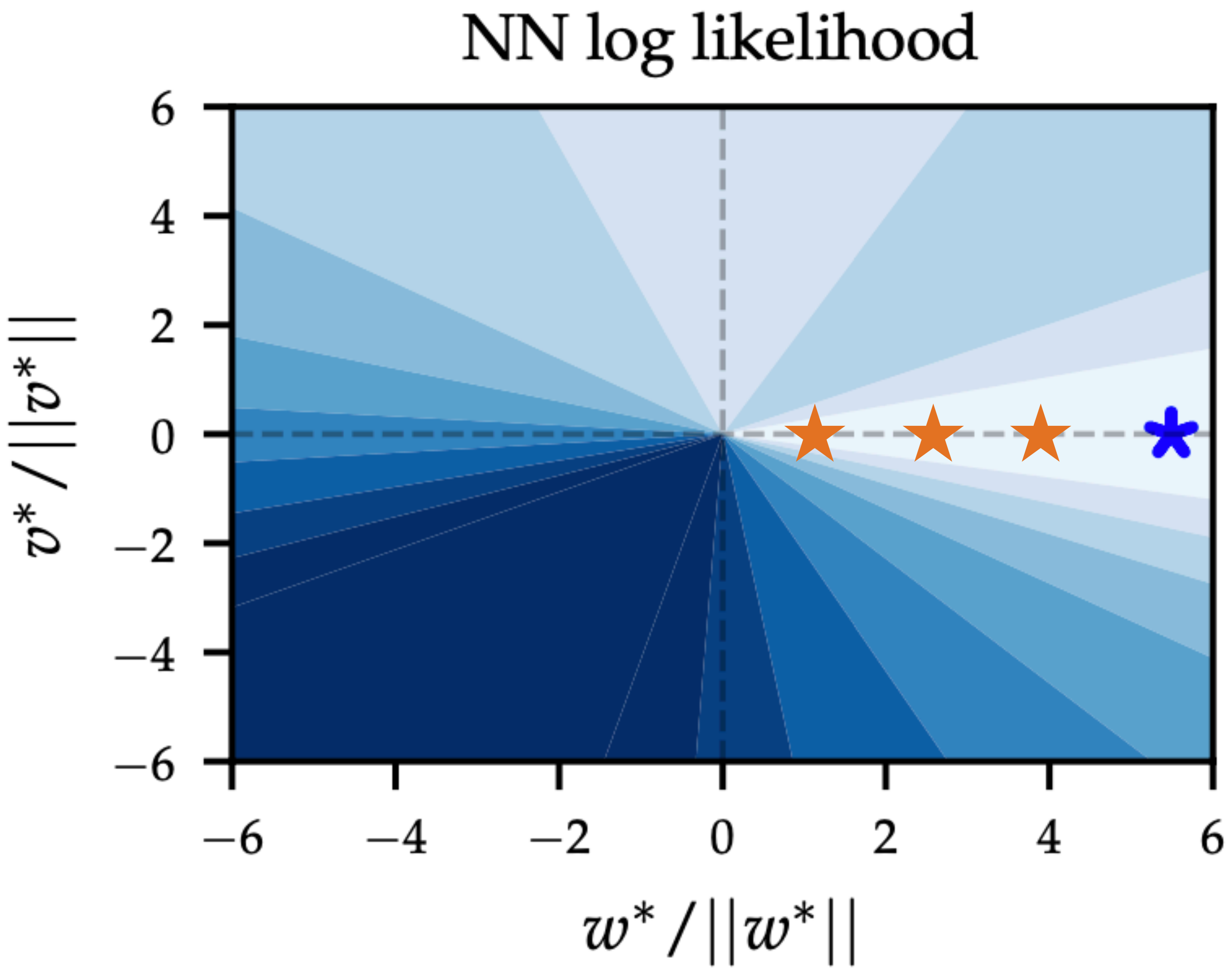
$$\mathcal{M}(\Lambda) = -0.5 \left[||w^\star||_\Lambda^2 + \log \det \frac{H + \Lambda}{\Lambda} \right] + C \qquad \text{with} \qquad \partial_v^2 L_h(w^\star) = H + \Lambda$$

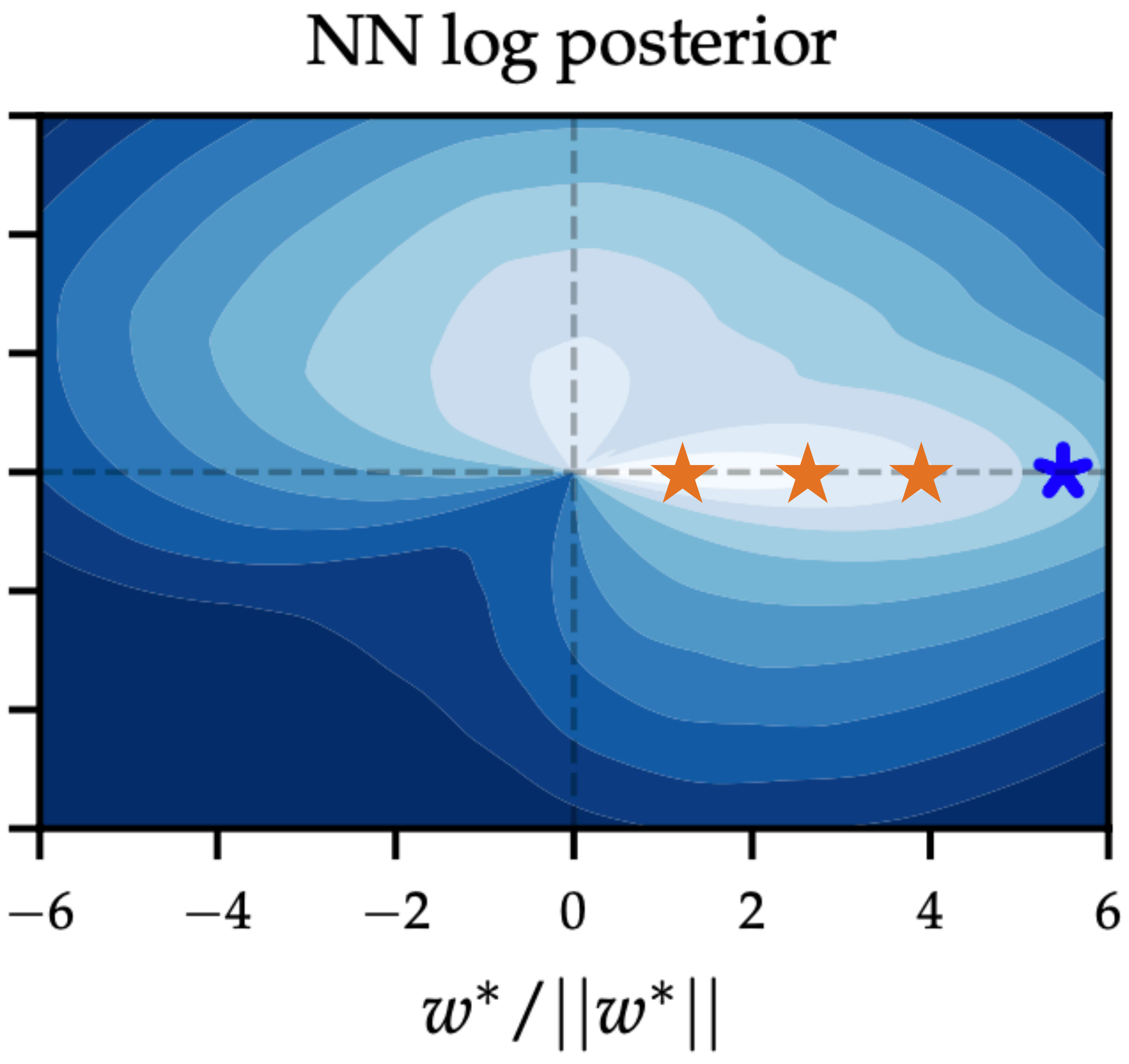
Pathologies introduced by normalisation layers

- Normalisation layers are ubiquitous and introduce scale invariance

$$f(x, w) = f(x, k \cdot w)$$

$$G_f(w) = G_f(k \cdot w)$$



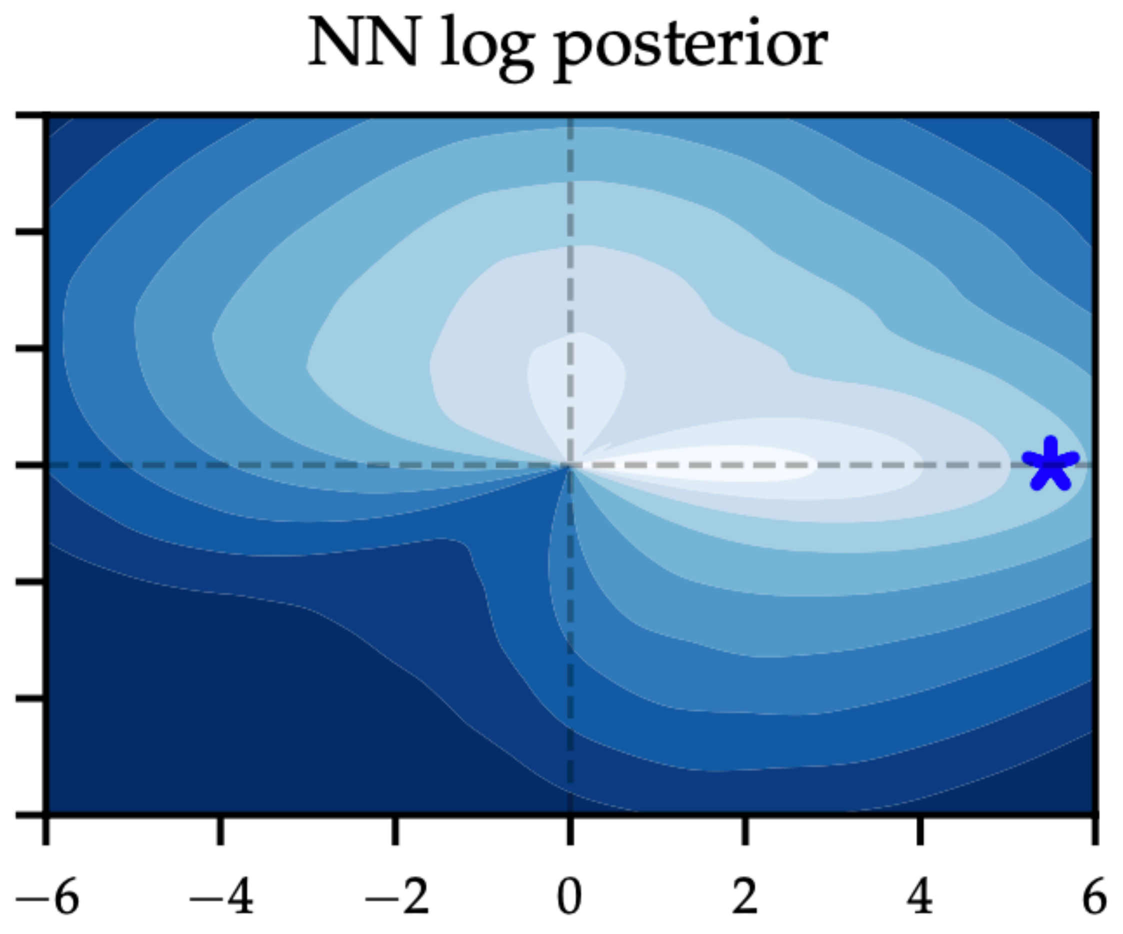


- This invariance is not present in the prior
- We can always obtain a larger prior density as $p(0.5 \cdot w)$
- Thus there exists no posterior mode (MAP)

★ **Linearisation point found with SGD** $w^\star$ **— not a mode of the posterior**

Pathologies introduced by normalisation layers cont.

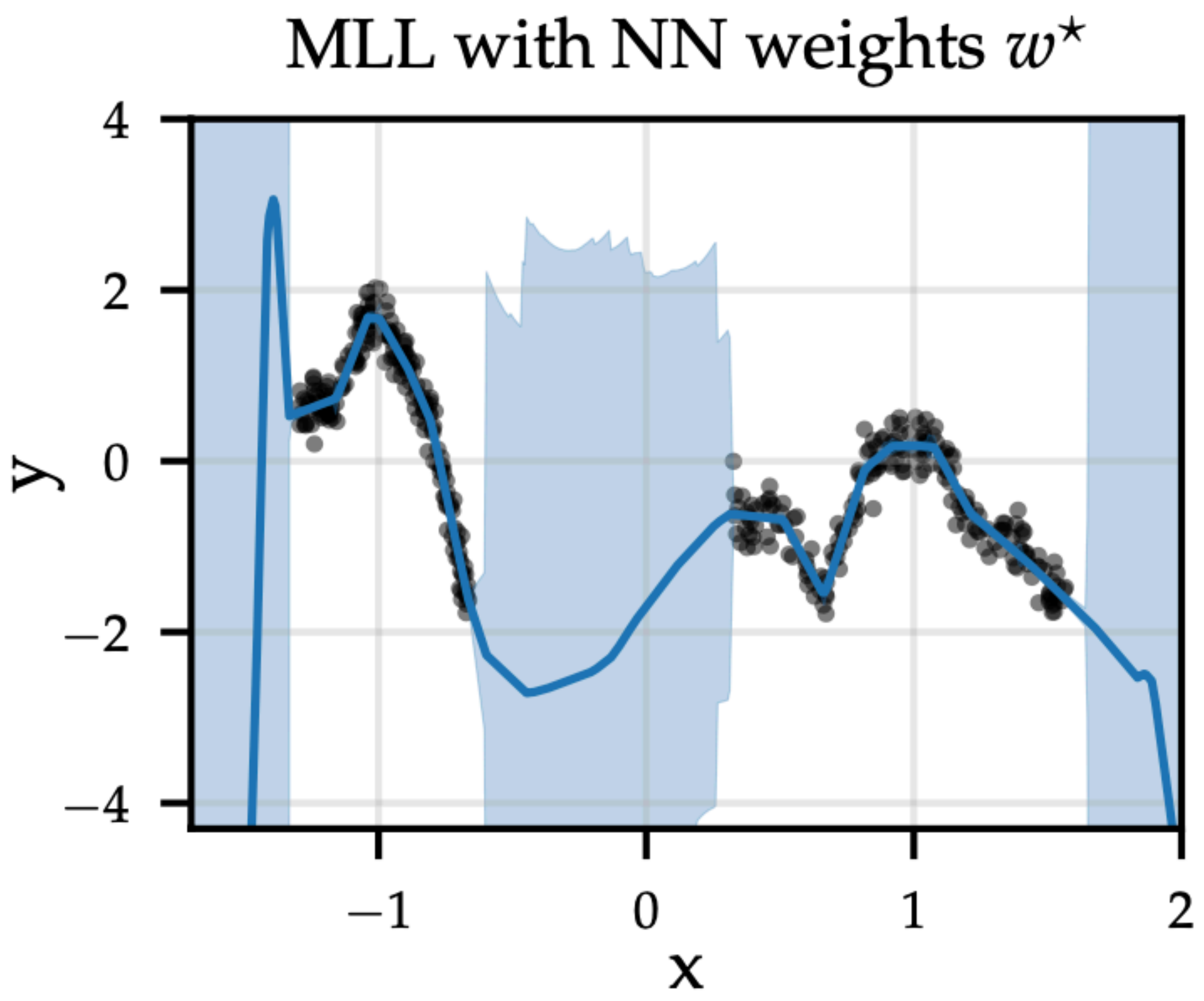
- Normalisation layers preclude the existence of an optima of the NN posterior
- Normalisation layers preclude $w^\star$ from being the MAP of the tangent linear model.



This biases our model evidence estimate

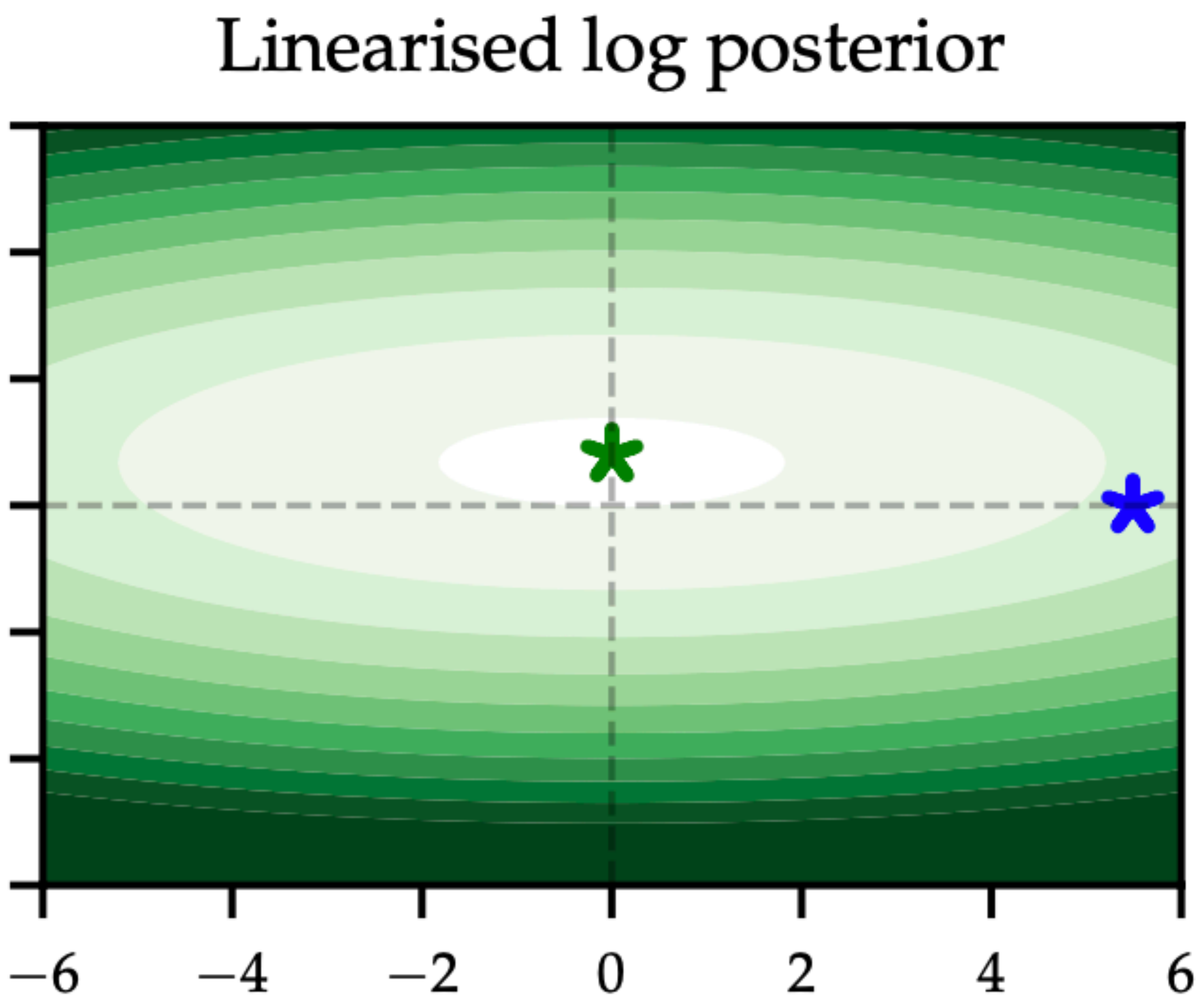
$$\mathscr{M}(\Lambda) = -0.5\left[\underline{||w^\star||^2_\Lambda} + \log\det\frac{H+\Lambda}{\Lambda}\right] + C$$

Leading to a bad $\Lambda^\star$ estimate



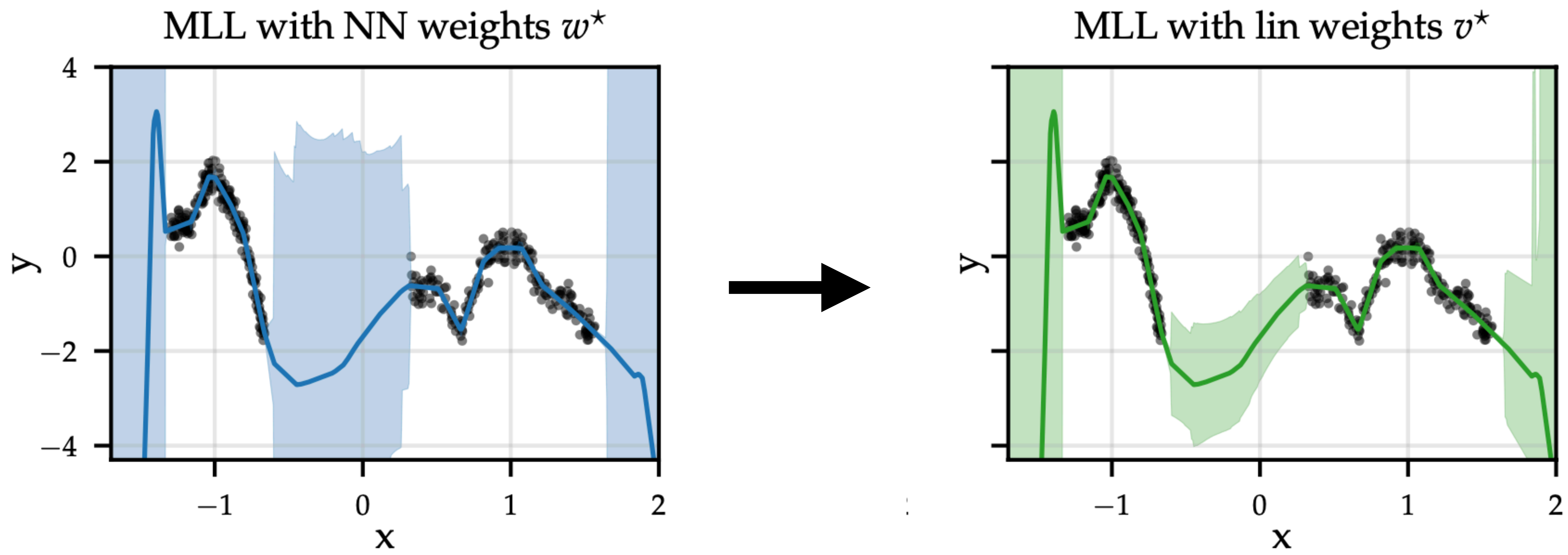
Finding a MAP to the lost linearised evidence

- Fortunately, for any linearisation point $w^\star$ ★ there exists a tangent linear model with a well defined posterior mode $v^\star$ ★



We use the model evidence of this tangent linear model

$$\mathcal{M}(\Lambda) = -0.5\left[\,||v^\star||^2_\Lambda + \log\det\frac{H+\Lambda}{\Lambda}\right] + C$$

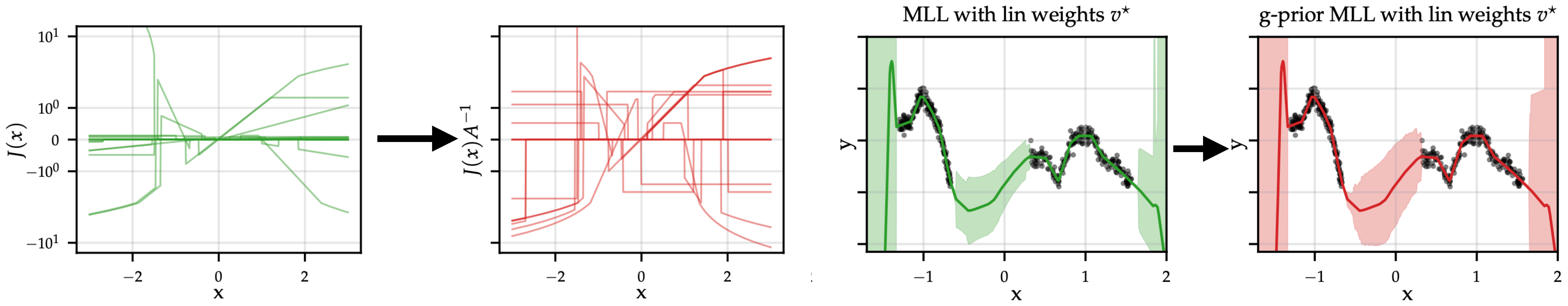


Heterogeneity in the Jacobian basis

- The tangent linear model can be seen as a basis function linear model where the Jacobian of $J = \partial_w f(x, w^\star)$ acts as a basis expansion

- However, different columns of the Jacobian have very different scales

- We extend Zellner's (1996) g-prior to NNs. Same posterior as normalising the second moment of J



Wrapping up

Problem: Direct application of linearised Laplace to neural networks with normalisation layers (batch norm, layer norm, etc) yields spurious model evidence estimates.

Solution: We propose to use the model evidence of the tangent linear model, which does not suffer from normalisation-related pathologies.

- Our results also apply to some recent *normalisation-free* methods. Roughly, these still divide layer outputs by the empirical standard deviation of the weights.

Problem: Different elements of the Jacobian basis expansion have very different scales, making a single choice of regulariser ineffective.

Solution: We extend the scale-invariant g-prior to the neural network setting.