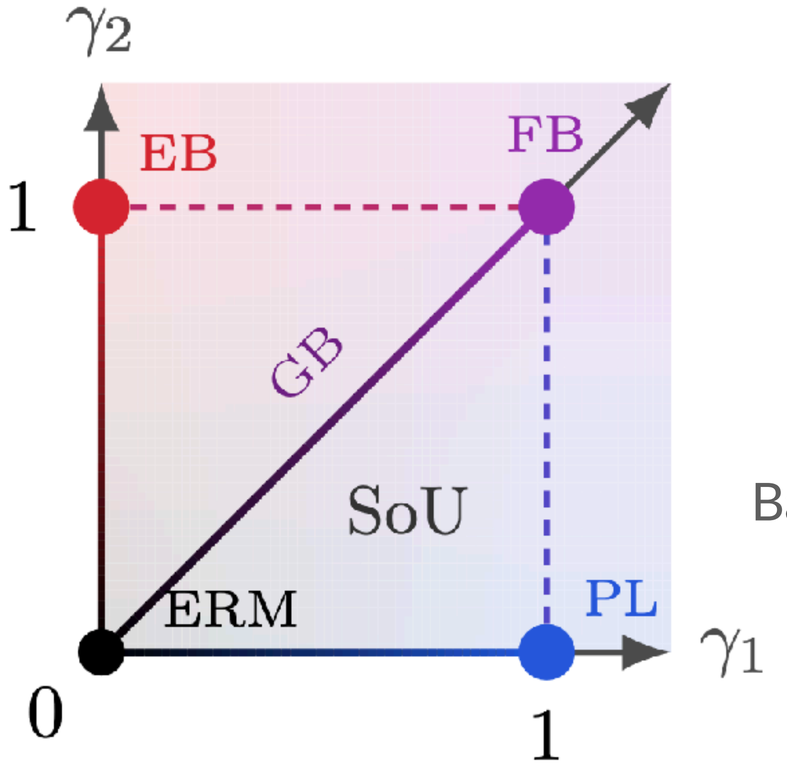
Empowering Bayes with sources of uncertainty
from confidence modelling to sparse estimation



Julyan Arbel, Inria Grenoble

Joint work with
Rafael Mouallem Rosa, Inria Grenoble
Hien Duy Nguyen, La Trobe and Kyushu

BayesAI meeting, Edinburgh, Oct 29, 2026





Bayesian inference with sources of uncertainty (SoU)

Uncertainty

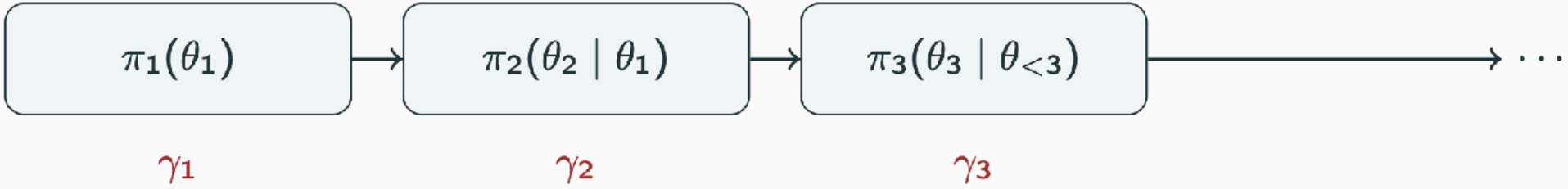
Under the Bayesian paradigm, beliefs are expressed through probability distributions.

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \qquad \longmapsto \qquad \pi(\boldsymbol{\theta})$$

Sources of uncertainty

A factorisation of the prior separates this uncertainty into components:

$$\pi(\boldsymbol{\theta}) = \pi_1(\theta_1) \prod_{k=2}^{K} \pi_k(\theta_k \mid \theta_{<k}).$$



SoU: attach a confidence level γ_k to each prior component.

From Bayesian inference to SoU inference

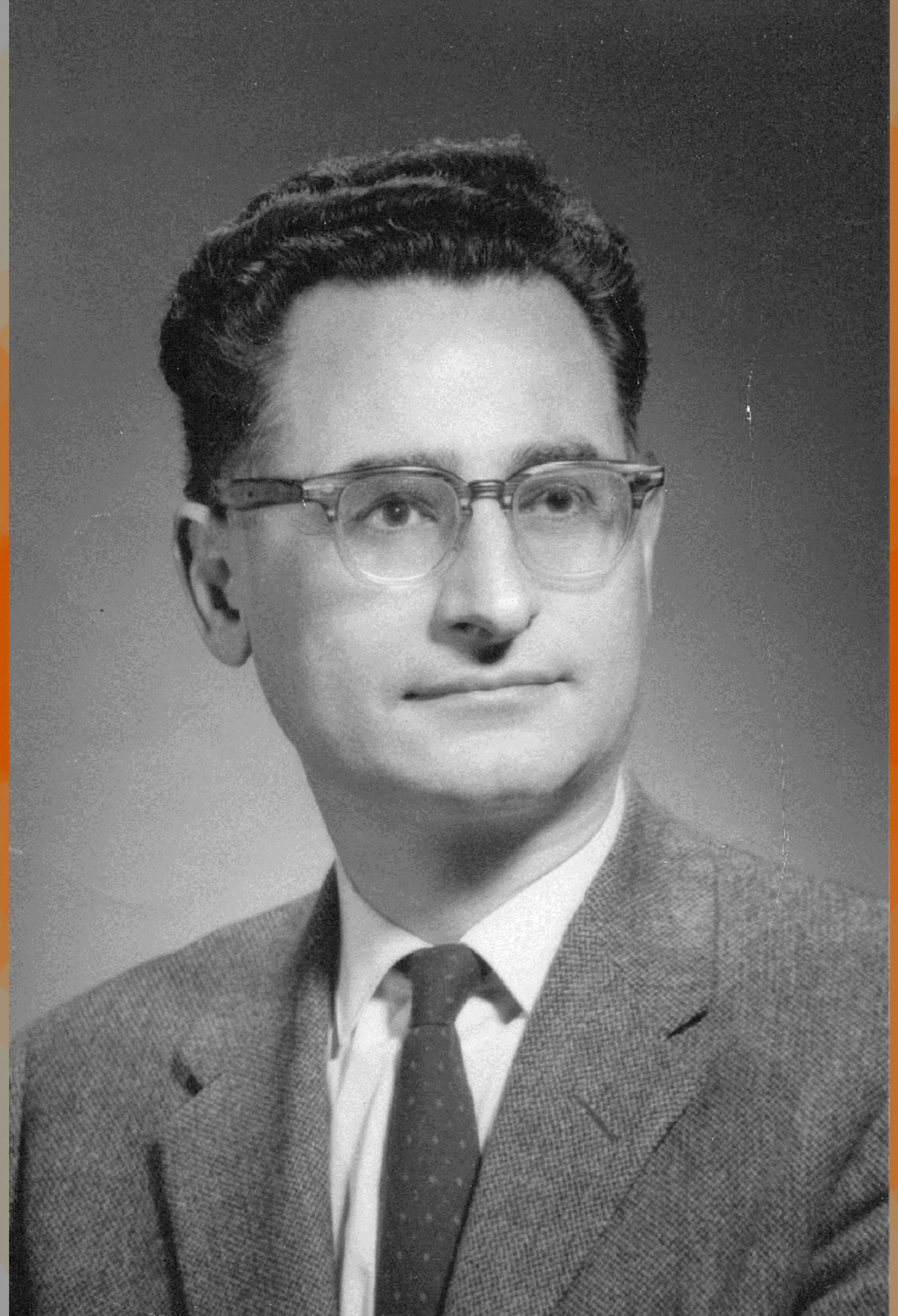
Bayesian inference

$$\underbrace{p(\boldsymbol{X} \mid \boldsymbol{\theta})}_{\text{likelihood}}, \qquad \underbrace{\pi_2(\theta_2 \mid \theta_1),\, \pi_1(\theta_1)}_{\text{prior}} \xrightarrow{\text{update rule}} \underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{X})}_{\text{posterior}}$$

SoU inference

$$\underbrace{p(\boldsymbol{X} \mid \boldsymbol{\theta})}_{\text{likelihood}}, \qquad \underbrace{(\pi_2(\theta_2 \mid \theta_1), \gamma_2),\, (\pi_1(\theta_1), \gamma_1)}_{\text{prior and confidence}} \xrightarrow{\text{update rule}} \underbrace{p_\gamma(\boldsymbol{\theta} \mid \boldsymbol{X})}_{\text{posterior}}$$

The new modelling input is confidence in each prior component.



Bayesian inference as an optimisation problem

Let $L_n(\theta) = \sum_{i=1}^n \ell(\theta, X_i)$. We define

$$q^* \in \arg\min_q \left\{ \underbrace{\frac{1}{n}\mathbb{E}_q[L_n(\theta)]}_{\text{fit the data}} + \underbrace{\frac{\gamma_1}{n}\mathrm{KL}(q(\theta\ )\|\pi(\theta\ ))}_{\text{stay close to the prior, source by source}} \right\}$$

SoU inference as an optimisation problem

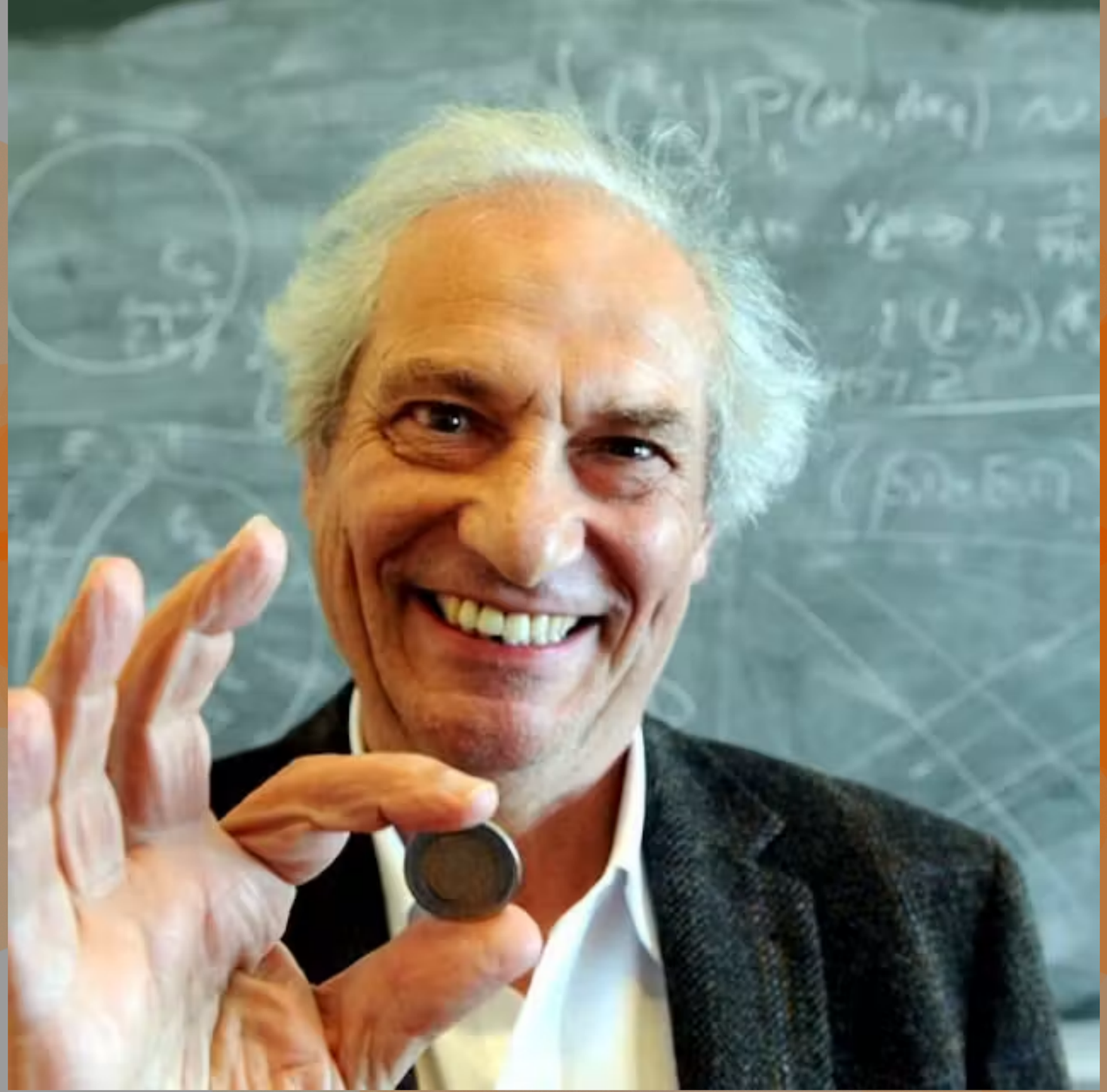
Let $L_n(\theta) = \sum_{i=1}^n \ell(\theta, X_i)$. We define

$$q^* \in \arg\min_q \left\{ \underbrace{\frac{1}{n}\mathbb{E}_q[L_n(\theta)]}_{\text{fit the data}} + \underbrace{\frac{\gamma_1}{n}\mathrm{KL}(q(\theta_1)\|\pi(\theta_1)) + \sum_{k=2}^{K} \frac{\gamma_k}{n}\mathbb{E}_{q(\theta_{<k})}[\mathrm{KL}(q(\theta_k \mid \theta_{<k})\|\pi(\theta_k \mid \theta_{<k}))]}_{\text{stay close to the prior, source by source}} \right\}$$

(Black-Box Generalised Variational Inference (BBGVI) algorithm of Knoblauch et al., 2022)

Framework	Loss	Confidence structure
Bayesian inference	negative log-likelihood	$\gamma_k = 1$ for all k
Generalised Bayes	general loss	$\gamma_k = \gamma$ for all k
SoU inference	general loss	γ_k source-specific

The data–prior trade-off is now controlled separately for each source of uncertainty.



Example: exponential family

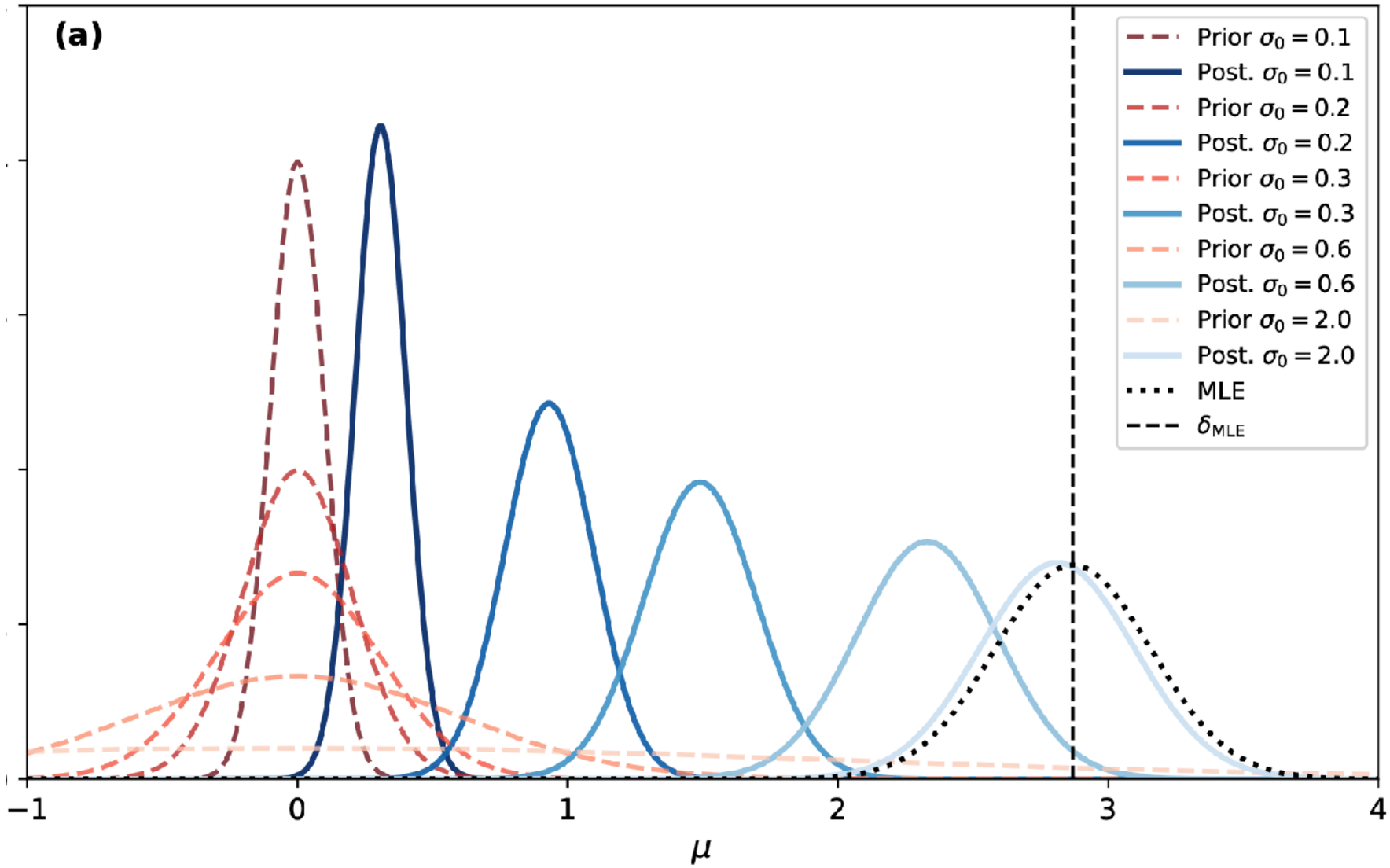
$$p(\boldsymbol{X}_{1:n} \mid \boldsymbol{\theta}) = \left[\prod_{i=1}^{n} h(X_i)\right] c(\boldsymbol{\theta})^n \exp\left\{\boldsymbol{\theta}^\top \sum_{i=1}^{n} T(X_i)\right\}$$

$$\pi(\boldsymbol{\theta} \mid N_0, T_0) = \kappa(N_0, T_0)\, c(\boldsymbol{\theta})^{N_0} \exp\left\{N_0 \boldsymbol{\theta}^\top T_0\right\}$$ (Diaconis & Ylvisaker, 1979)

Bayes

$$N_n := N_0 + n, \quad T_n := \alpha_n T_0 + (1-\alpha_n)\bar{T}_n, \quad \alpha_n := \frac{N_0}{N_0+n}, \quad \bar{T}_n := \frac{1}{n}\sum_{i=1}^{n} T(X_i)$$

Bayes N(μ, σ^2)-N($0,\sigma_0^2$), varying σ_0



Example: exponential family

$$p(\boldsymbol{X}_{1:n} \mid \boldsymbol{\theta}) = \left[\prod_{i=1}^{n} h(X_i)\right] c(\boldsymbol{\theta})^n \exp\left\{\boldsymbol{\theta}^\top \sum_{i=1}^{n} T(X_i)\right\}$$

$$\pi(\boldsymbol{\theta} \mid N_0, T_0) = \kappa(N_0, T_0)\, c(\boldsymbol{\theta})^{N_0} \exp\left\{N_0 \boldsymbol{\theta}^\top T_0\right\}$$ (Diaconis & Ylvisaker, 1979)

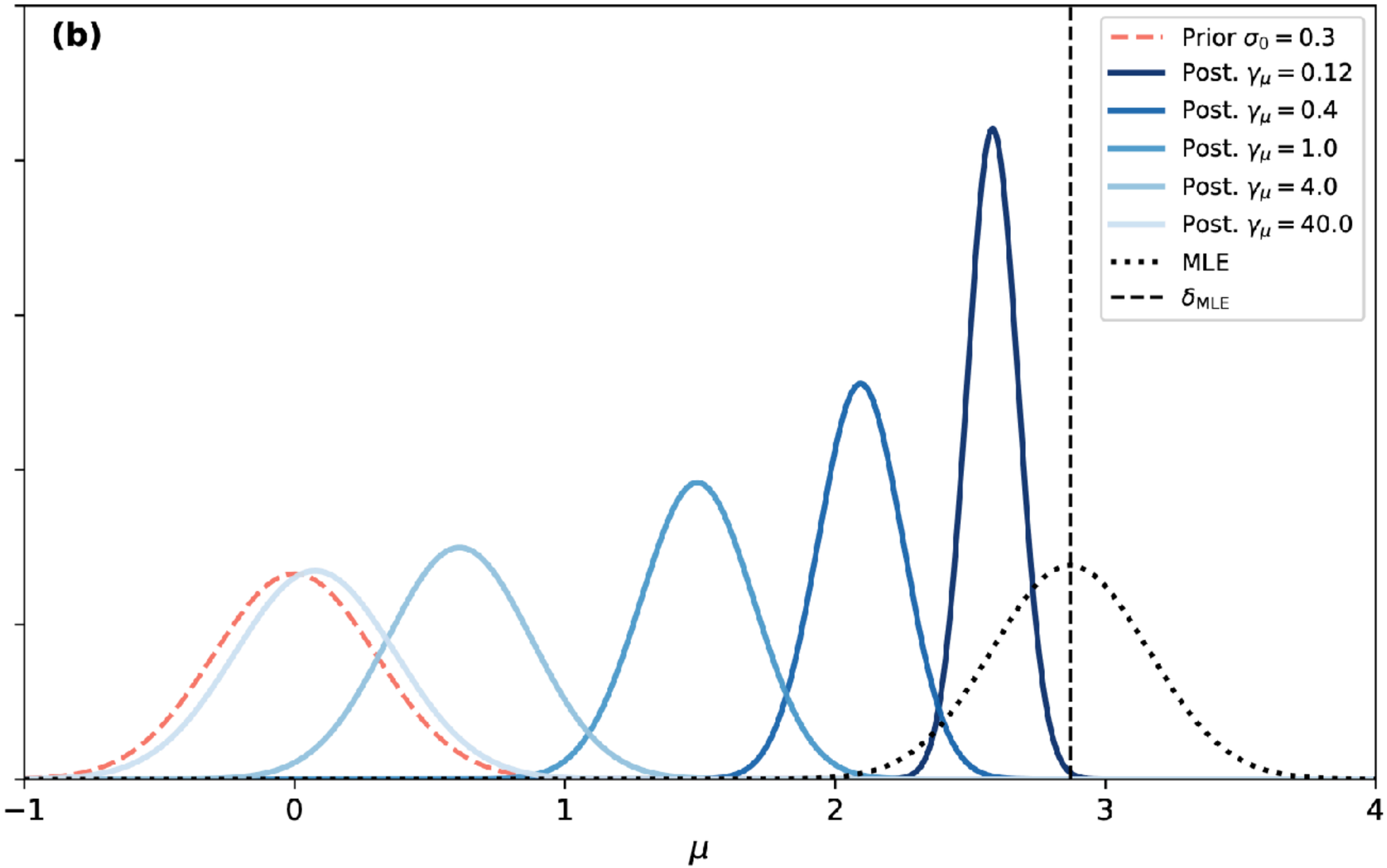
Bayes

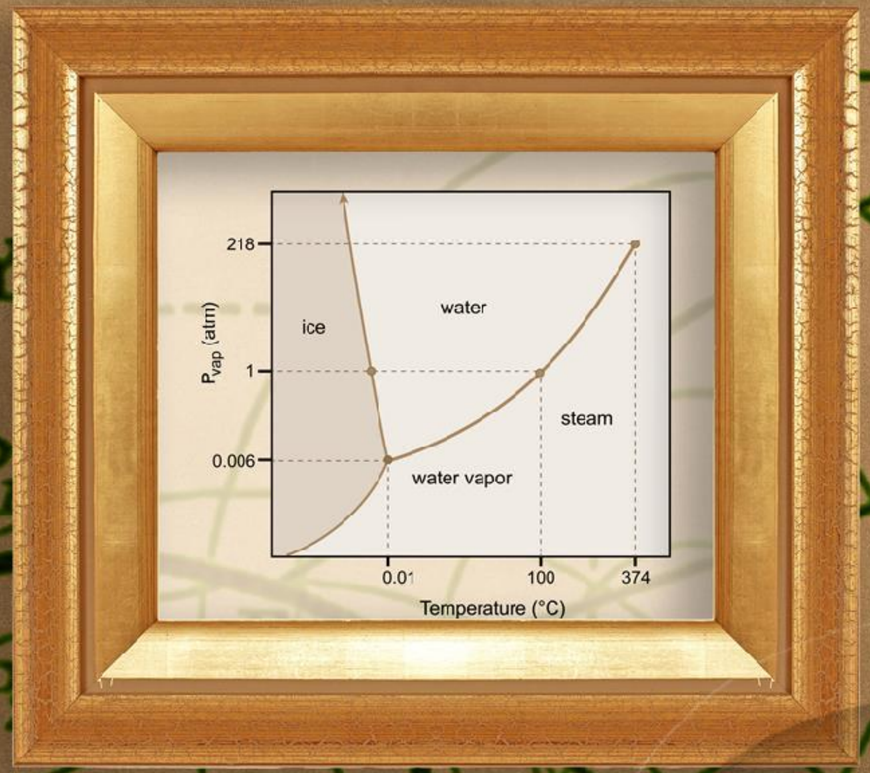
$$N_n := N_0 + n, \quad T_n := \alpha_n T_0 + (1-\alpha_n)\bar{T}_n, \quad \alpha_n := \frac{N_0}{N_0+n}, \quad \bar{T}_n := \frac{1}{n}\sum_{i=1}^{n} T(X_i)$$

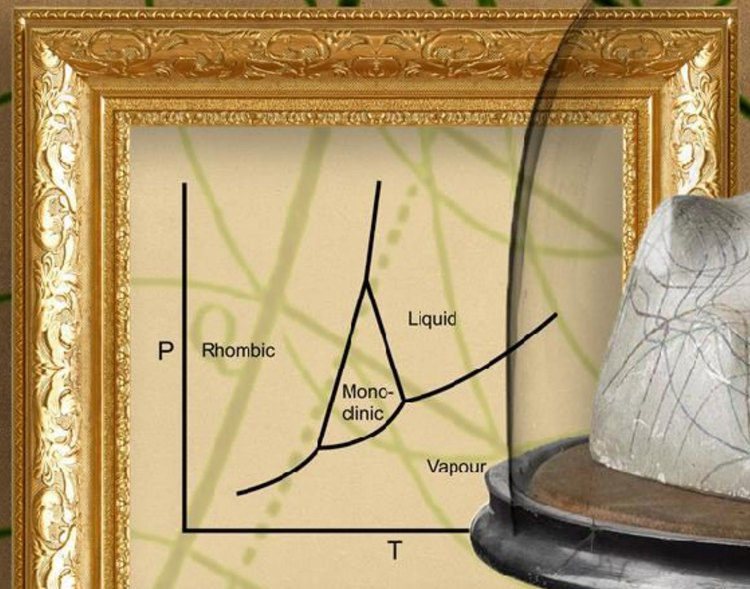
SoU

$$N_n = N_0 + \frac{n}{\gamma}, \quad \alpha_n = \frac{\gamma N_0}{\gamma N_0 + n}$$

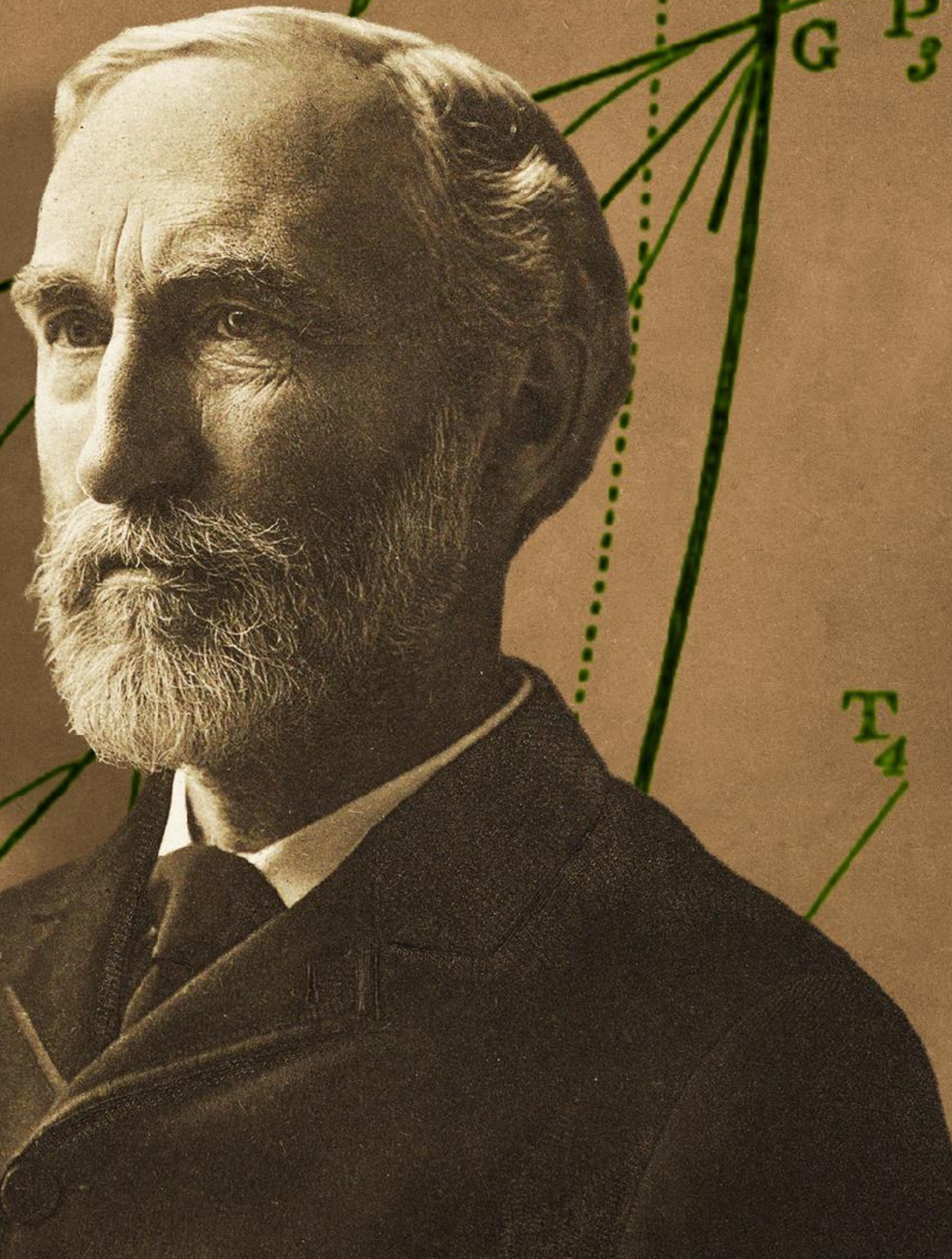
SoU N(μ, σ^2)-N(0,σ_0^2), varying γ_μ




218
1
0.006
P_{vap} (atm)
ice
water
steam
water vapor
0.01
100
374
Temperature (°C)


P
Rhombic
Liquid
Monc-
clinic
Vapour
T





Theoretical results

1. Closed-form characterization

The unrestricted SoU posterior can be constructed by a backward recursion over sources. Each conditional block is a Gibbs-type update with its own confidence parameter γ_k.

2. Objective collapse

$$\|\gamma\|_\infty \to 0 \quad \Rightarrow \quad q^*_\gamma \Rightarrow \delta_{\widehat{\theta}_n}$$

For fixed data, low confidence removes the prior regularisation and recovers empirical risk minimization.

3. Consistency

$$q^*_n \Rightarrow \delta_{\theta_0}$$

As $n \to \infty$, the SoU posterior concentrates around the population risk minimizer, under standard regularity conditions and a controlled scaling of γ_n.

Two-block case

$$\mathcal{L}_{\gamma_2}(\boldsymbol{\theta}_1) = \left[\int p(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{1/\gamma_2} \pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1) \, d\boldsymbol{\theta}_2 \right]^{\gamma_2}$$

$$q_{\gamma_2}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{X}) \propto p(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{1/\gamma_2} \pi_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$$

$$q_{\gamma_1,\gamma_2}(\boldsymbol{\theta}_1 \mid \boldsymbol{X}) \propto \pi_1(\boldsymbol{\theta}_1) \mathcal{L}_{\gamma_2}(\boldsymbol{\theta}_1)^{1/\gamma_1}.$$

$$q_{\gamma_1,\gamma_2}(\boldsymbol{\theta}_2 \mid \boldsymbol{X}) = \int q_{\gamma_2}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{X}) q_{\gamma_1,\gamma_2}(\boldsymbol{\theta}_1 \mid \boldsymbol{X}) \, d\boldsymbol{\theta}_1$$

Selective limits I

θ_1 hyperparameter block

θ_2 target parameter block

Set $\gamma_2 = 1$: standard Bayesian treatment

$$\mathcal{L}_1(\boldsymbol{\theta}_1) = \int p(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1) \, d\boldsymbol{\theta}_2 = m(\boldsymbol{X} \mid \boldsymbol{\theta}_1)$$

$$q_{\gamma_1,1}(\boldsymbol{\theta}_1 \mid \boldsymbol{X}) \propto \pi(\boldsymbol{\theta}_1) m(\boldsymbol{X} \mid \boldsymbol{\theta}_1)^{1/\gamma_1}$$

Take $\gamma_1 \to 0$: $\quad q_{\gamma_1,1}(\boldsymbol{\theta}_1 \mid \boldsymbol{X}) \Longrightarrow \delta_{\widehat{\boldsymbol{\theta}}_1}, \qquad \widehat{\boldsymbol{\theta}}_1 = \arg\max_{\boldsymbol{\theta}_1} m(\boldsymbol{X} \mid \boldsymbol{\theta}_1)$

$$q_{\gamma_1,1}(\boldsymbol{\theta}_2 \mid \boldsymbol{X}) \longrightarrow q(\boldsymbol{\theta}_2 \mid \widehat{\boldsymbol{\theta}}_1, \boldsymbol{X})$$

That's *empirical Bayes*
(Carlin, Louis, 2000)

Selective limits II

θ_1 target parameter block

θ_2 nuisance parameter block

Set $\gamma_1 = 1$: standard Bayesian treatment

$$q_{\gamma_2}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{X}) \propto p(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{1/\gamma_2} \pi_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$$

Take $\gamma_2 \to 0$: $q_{\gamma_2}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1, \boldsymbol{X}) \Longrightarrow \delta_{\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)}$, $\quad \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1) = \arg\max_{\boldsymbol{\theta}_2} p(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

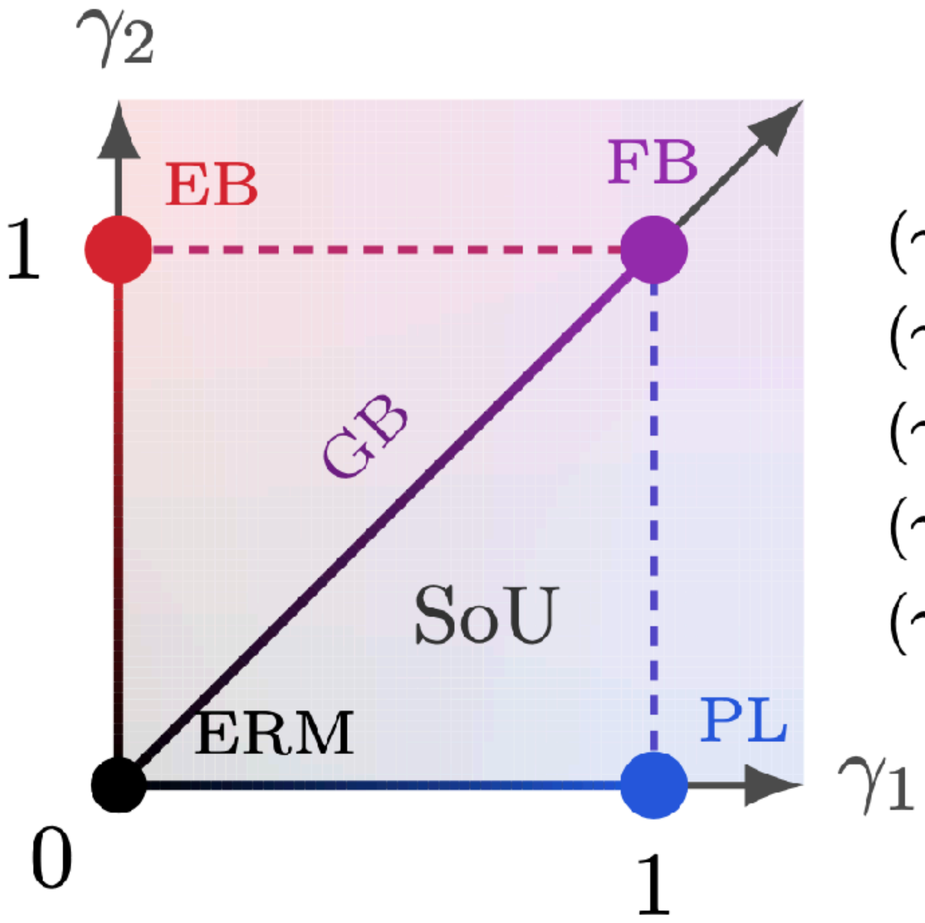
$$\mathcal{L}_{\gamma_2}(\boldsymbol{\theta}_1) \longrightarrow \sup_{\boldsymbol{\theta}_2} p(\boldsymbol{X} \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) =: L_{\text{prof}}(\boldsymbol{\theta}_1)$$

$$q_{1,\gamma_2}(\boldsymbol{\theta}_1 \mid \boldsymbol{X}) \longrightarrow q_{\text{prof}}(\boldsymbol{\theta}_1 \mid \boldsymbol{X}) \propto \pi(\boldsymbol{\theta}_1) L_{\text{prof}}(\boldsymbol{\theta}_1)$$

That's *a profiled likelihood*
(Murphy and van der Vaart, 2000)

Selective limits

SoU relaxes homogeneity, making selective collapse possible



$(\gamma_1, \gamma_2) = (1, 1)$ full Bayes (FB),
$(\gamma_1, \gamma_2) = (\gamma, \gamma)$ generalised Bayes (GB),
$(\gamma_1, \gamma_2) \to (0, 1)$ empirical Bayes (EB),
$(\gamma_1, \gamma_2) \to (1, 0)$ profile likelihood (PL),
$(\gamma_1, \gamma_2) \to (0, 0)$ empirical risk minimisation (ERM).



Choosing confidence I

Predictive calibration.

PPD $$p_{\boldsymbol{\gamma}}(\widetilde{X} \mid \boldsymbol{X}) = \int p(\widetilde{X} \mid \boldsymbol{\theta}) q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{X})\, d\boldsymbol{\theta}$$

Estimated predictive score:
held-out data or cross-validation $$\widehat{\mathcal{S}}_{\mathrm{pred}}(p_{\boldsymbol{\gamma}})$$

$$\widehat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma} \in \Gamma} \widehat{\mathcal{S}}_{\mathrm{pred}}(p_{\boldsymbol{\gamma}})$$

Choosing confidence II

Uncertainty calibration.

θ_t target block, $\gamma_t = 1$

θ_s source to be calibrated, $\gamma_s = \gamma$

Focus is SoU posterior on θ_t: $q_\gamma(\boldsymbol{\theta}_t \mid \boldsymbol{X})$

Posterior credible region: $q_\gamma\left(\boldsymbol{\theta}_t \in C_{\gamma,\alpha}(\boldsymbol{X}) \mid \boldsymbol{X}\right) = 1 - \alpha$

Choose γ such that it induces correct freq. coverage:
(in practice eg by bootstrap)

$$P_0\left\{\boldsymbol{\theta}_{t,0} \in C_{\gamma,\alpha}(\boldsymbol{X})\right\} = 1 - \alpha$$

Choosing confidence III

Information matching.

Matching asymptotic information of SoU posterior to a reference distribution.
Homogeneous-confidence case (Lyddon et al. 2019).

Expected loss minimiser: $$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \mathbb{E}_{P_0} \left[\ell(\boldsymbol{\theta}, X) \right]$$

$$I(\boldsymbol{\theta}_0) = \mathbb{E}_{P_0} \left[\nabla \ell(\boldsymbol{\theta}_0, X) \nabla \ell(\boldsymbol{\theta}_0, X)^\top \right], \qquad J(\boldsymbol{\theta}_0) = \mathbb{E}_{P_0} \left[\nabla^2 \ell(\boldsymbol{\theta}_0, X) \right]$$

Generalised posterior & loss-likelihood bootstrap are asymptotically Gaussian:

$$\Sigma_{\mathrm{GB}}(\gamma) = \gamma J(\boldsymbol{\theta}_0)^{-1}, \qquad \Sigma_{\mathrm{LL}} = J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}$$

Matching Fisher information numbers: $$\gamma^* = \frac{\operatorname{tr}\{J(\boldsymbol{\theta}_0)\}}{\operatorname{tr}\{J(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0)^{-1} J(\boldsymbol{\theta}_0)^\top\}}$$



Global-local shrinkage priors

Normal means model

$$y_{ij} \mid \beta_i, \sigma^2 \sim N(\beta_i, \sigma^2), \quad \beta_i \mid \tau_i, \nu \sim N(0, \tau_i^2 \nu^2).$$

ν: global shrinkage; τ_i: local scale

↓

Conditionally on the scales

$$\mathbb{E}[\beta_i \mid y_i, \tau_i, \nu] = (1 - \kappa_i)\bar{y}_i + \kappa_i 0, \qquad \kappa_i = \frac{1}{1 + n\tau_i^2 \nu^2 / \sigma^2}.$$

Standard global-local

$\pi_\tau(\tau_i) \quad \rightsquigarrow \quad$ shrinkage profile of κ_i

The choice of π_τ is central.

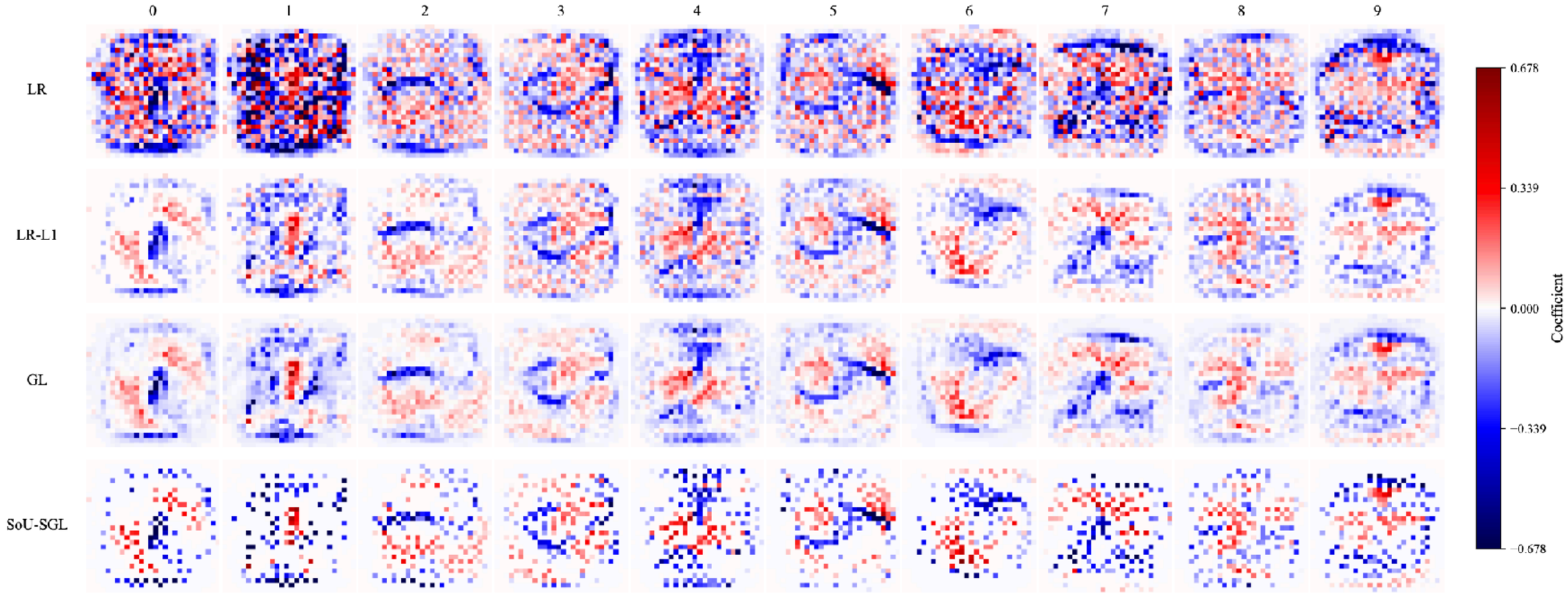
SoU-SGL

$\gamma_\tau \approx 0, \qquad \pi_\tau(\tau_i)^{\gamma_\tau} \to 1$

The choice of π_τ looses influence.

Logistic regression: sparse coefficient maps

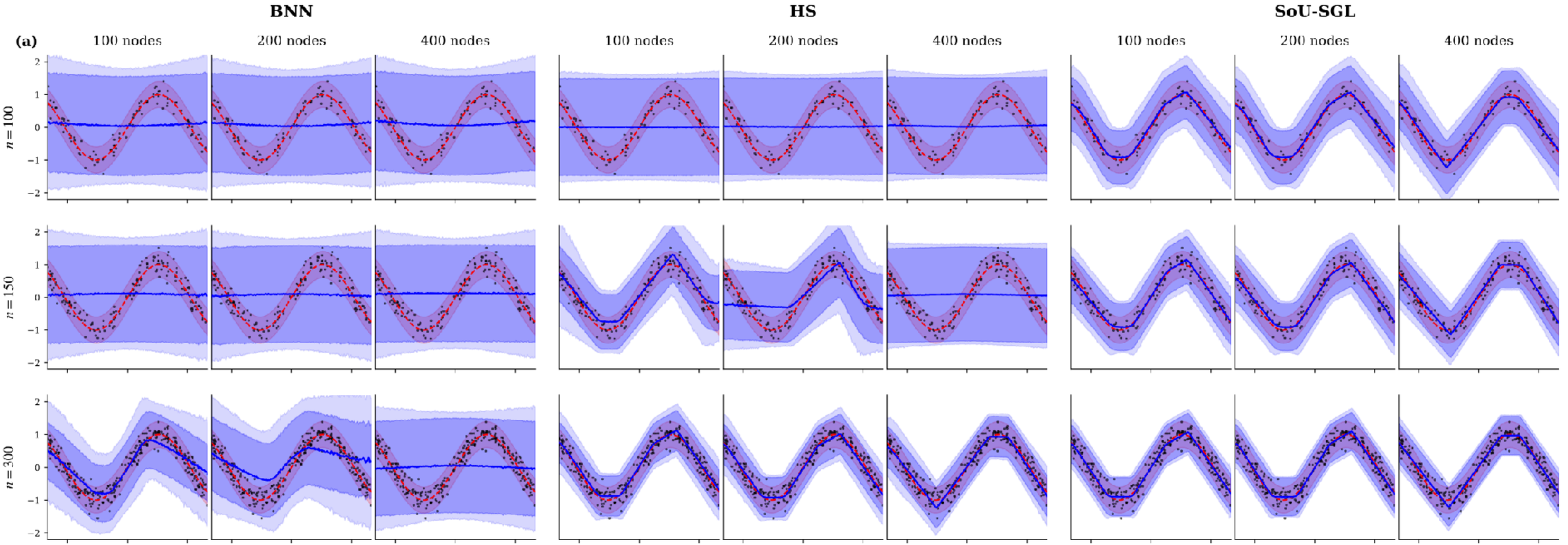
MNIST classification: one-vs-rest logistic regression for digits 0, . . . , 9.





Toy BNN: uncertainty and pruning

Toy regression data: $y = sin(x) + \varepsilon$, $\varepsilon \sim N(0, 0.22)$.

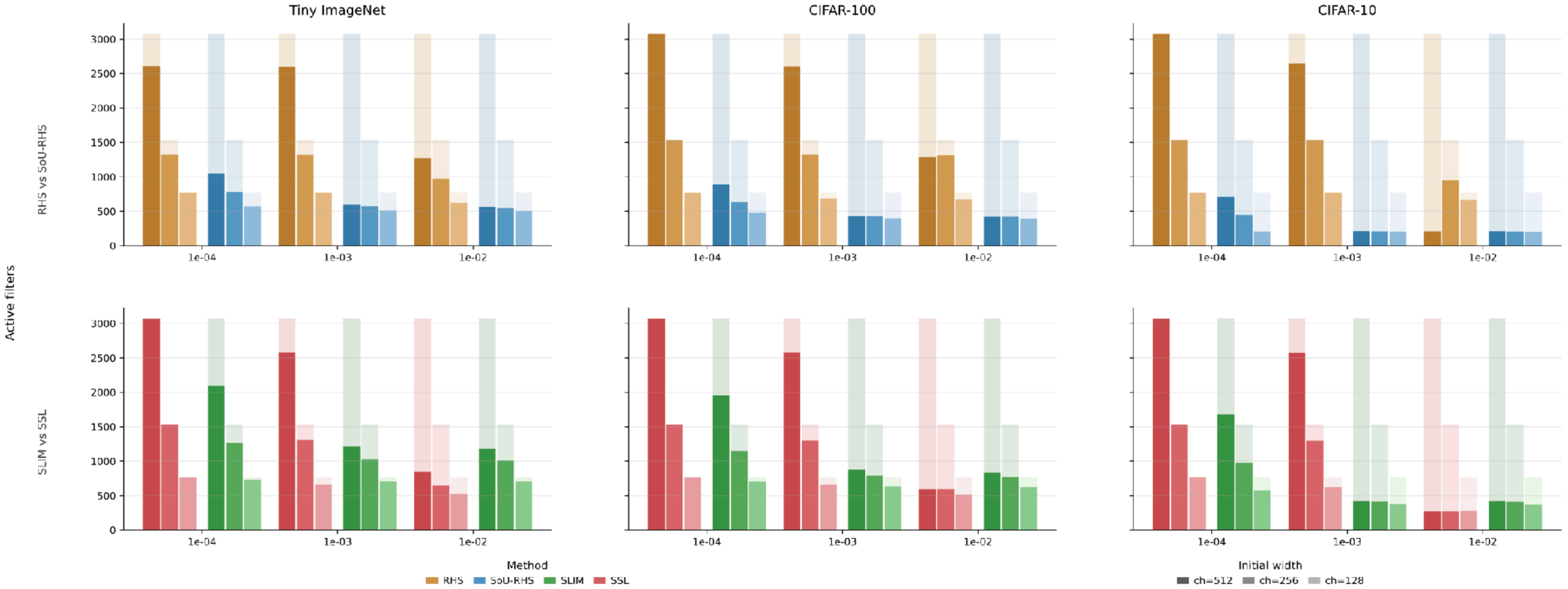


MNIST BNN: compact architectures

2-layer neural networks trained with varying initial widths

Initial width	50	100	200	400	800	1200	1600
Active units, layer 1	26	26	32	30	36	33	31
Active units, layer 2	29	32	24	20	19	15	15
Accuracy	.978	.977	.977	.977	.979	.976	.977

CNN with SoU regularized horseshoe



CNN with SoU regularized horseshoe

Dataset	Width	Stat	NN	BNN	SoU-RHS (ours)	SLIM 10^{-3}	SLIM 10^{-4}	SSL 10^{-3}	SSL 10^{-4}
Tiny ImageNet	128	Acc (↑)	0.356	0.321	**0.465**	0.332	0.361	0.346	0.355
		NLL (↓)	7.932	2.900	**2.237**	4.703	5.924	4.954	7.213
		ECE (↓)	0.510	**0.078**	0.101	0.380	0.448	0.400	0.487
		#AF(768) (↓)	–	–	**509**	702	751	522	694
	256	Acc (↑)	0.417	0.294	**0.432**	0.308	0.411	0.372	0.396
		NLL (↓)	6.871	3.073	**2.387**	4.377	3.780	4.571	5.723
		ECE (↓)	0.453	0.084	**0.082**	0.331	0.319	0.372	0.428
		#AF(1536) (↓)	–	–	**576**	1010	1370	645	1148
	512	Acc (↑)	**0.496**	0.243	0.410	0.061	0.431	0.393	0.423
		NLL (↓)	5.527	3.344	**2.472**	7.673	3.280	4.184	5.270
		ECE (↓)	0.384	**0.058**	0.059	0.397	0.273	0.346	0.401
		#AF(3072) (↓)	–	–	**594**	1177	2018	849	1433

Takeaways

1. **Confidence as modelling input.**
   Each source of uncertainty can have its own confidence parameter.

2. **Confidence controls posterior concentration.**
   It is distinct from prior informativeness.

3. **Low local confidence turns shrinkage into sparsity.**
   With global-local priors, $\gamma_\tau \approx 0$ yields SoU-SGL.

4. **In BNNs, this becomes self-pruning.**
   The network adapts its effective architecture during training.