On the Importance of Priors in Bayesian Deep Learning

Dr. Vincent Fortuin

RIKEN AIP (remotely)

April 2022

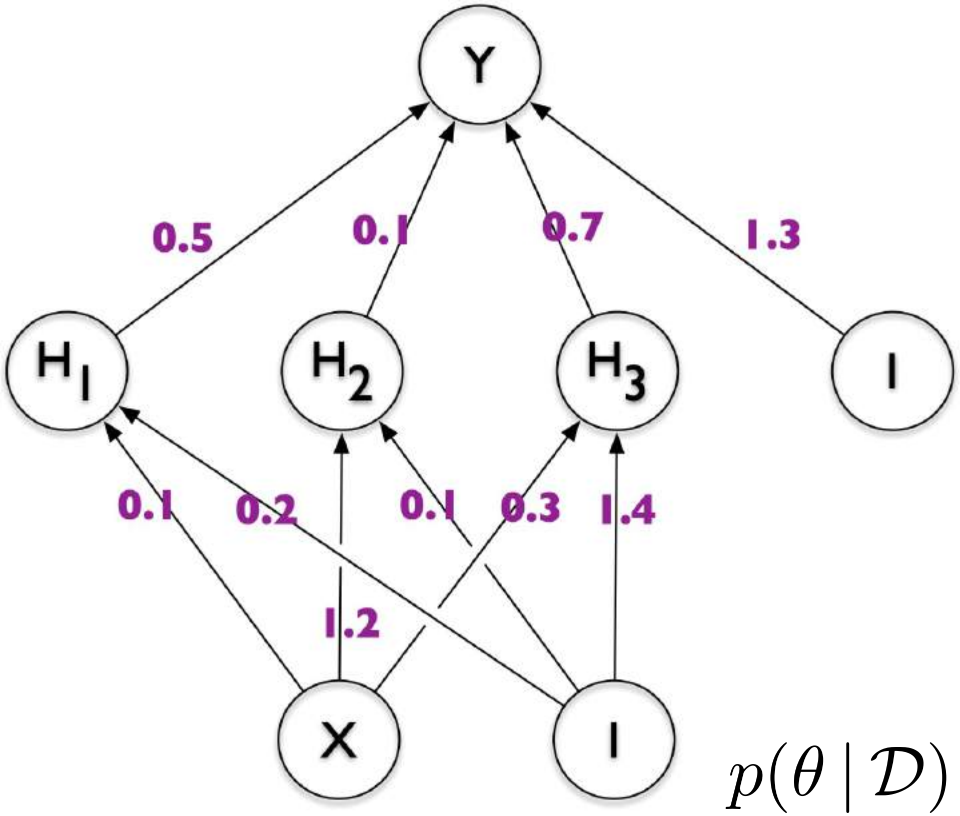
Agenda

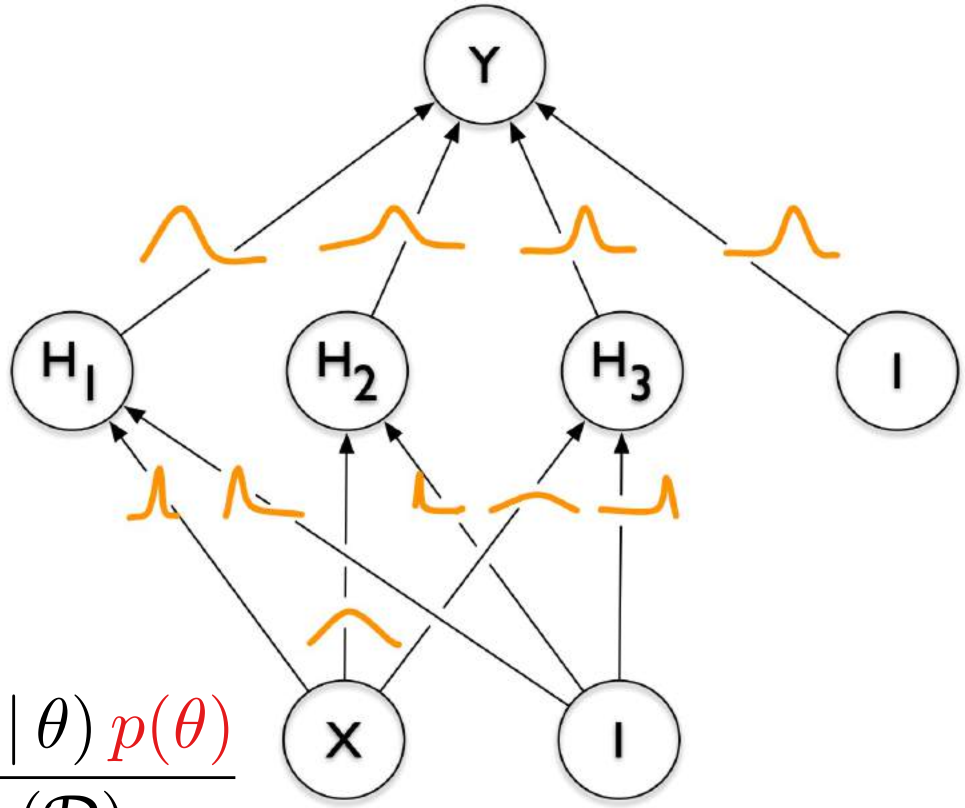
- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation
- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning
- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

Agenda

- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation
- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning
- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

Background: Bayesian Neural Networks

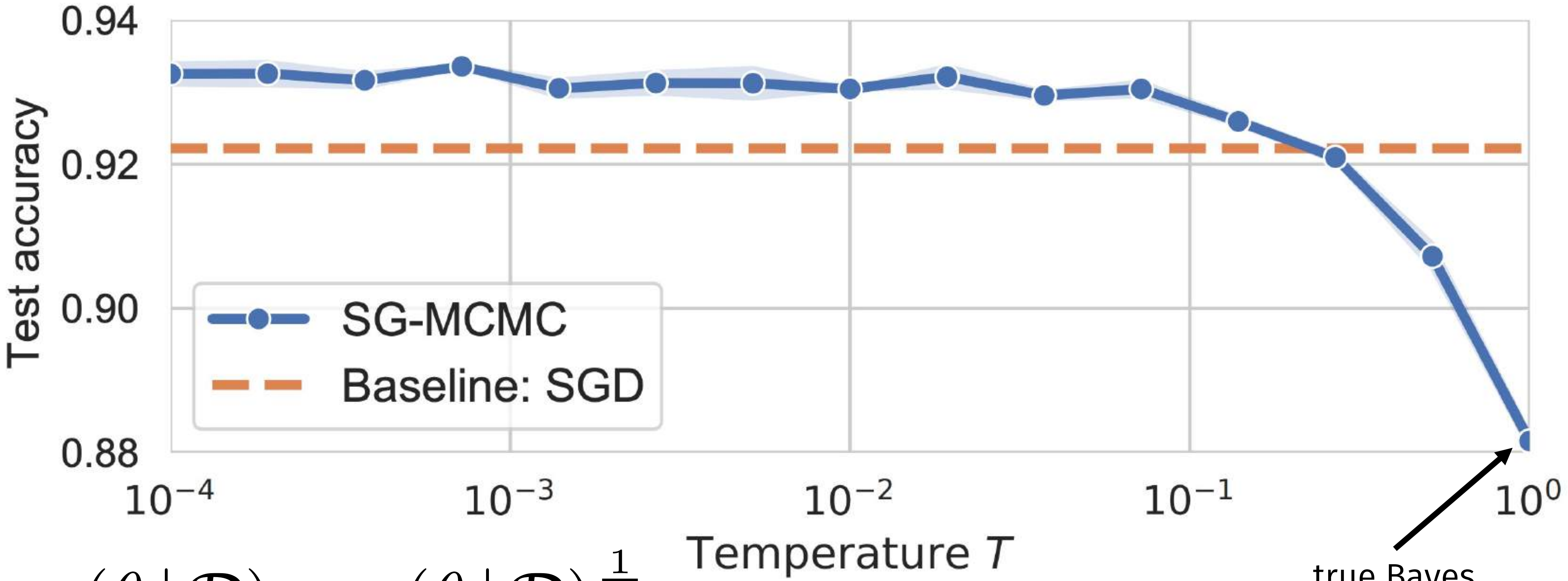




$$p(\theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta)\, {\color{red} p(\theta)}}{p(\mathcal{D})}$$

[Blundell, Cornebise, Kavukcuoglu, Wierstra 2015]

Motivation: Cold-posterior effect



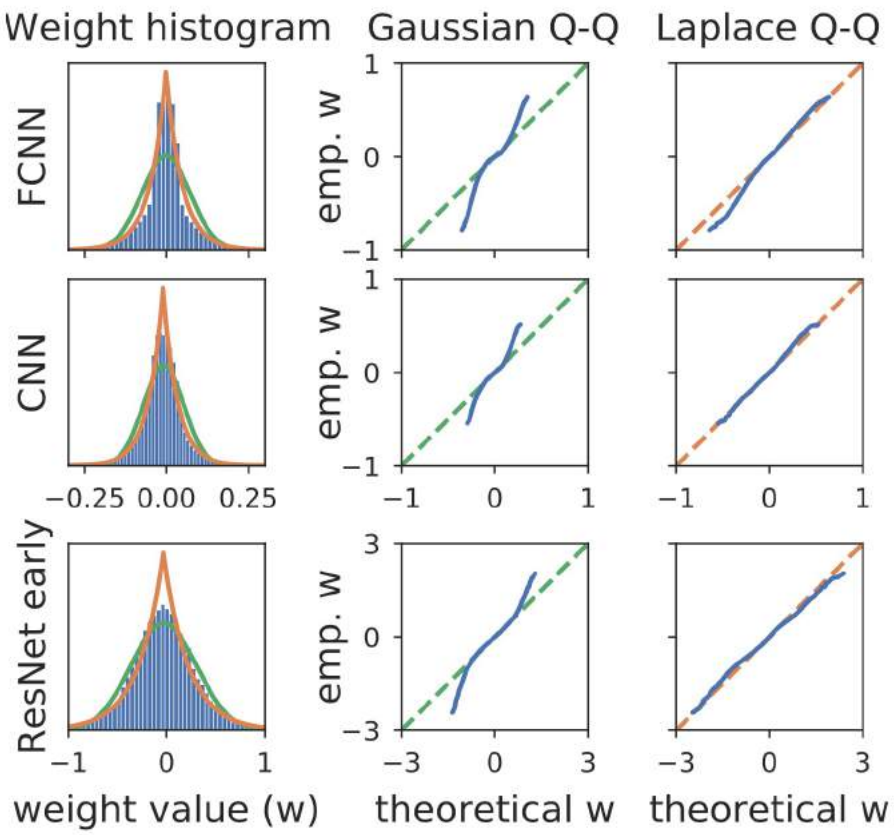
$$p_T(\theta \,|\, \mathcal{D}) \propto p(\theta \,|\, \mathcal{D})^{\frac{1}{T}}$$

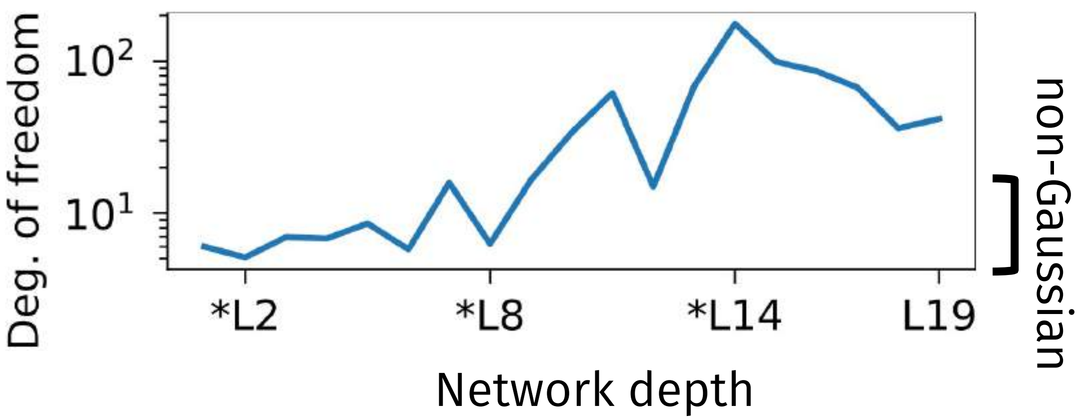
true Bayes posterior

[Wenzel, Roth, Veeling, Swiatkowski, Tran, Mandt, Snoek, Salimans, Jenatton, Nowozin 2020]

Empirical FCNN weights are heavy-tailed

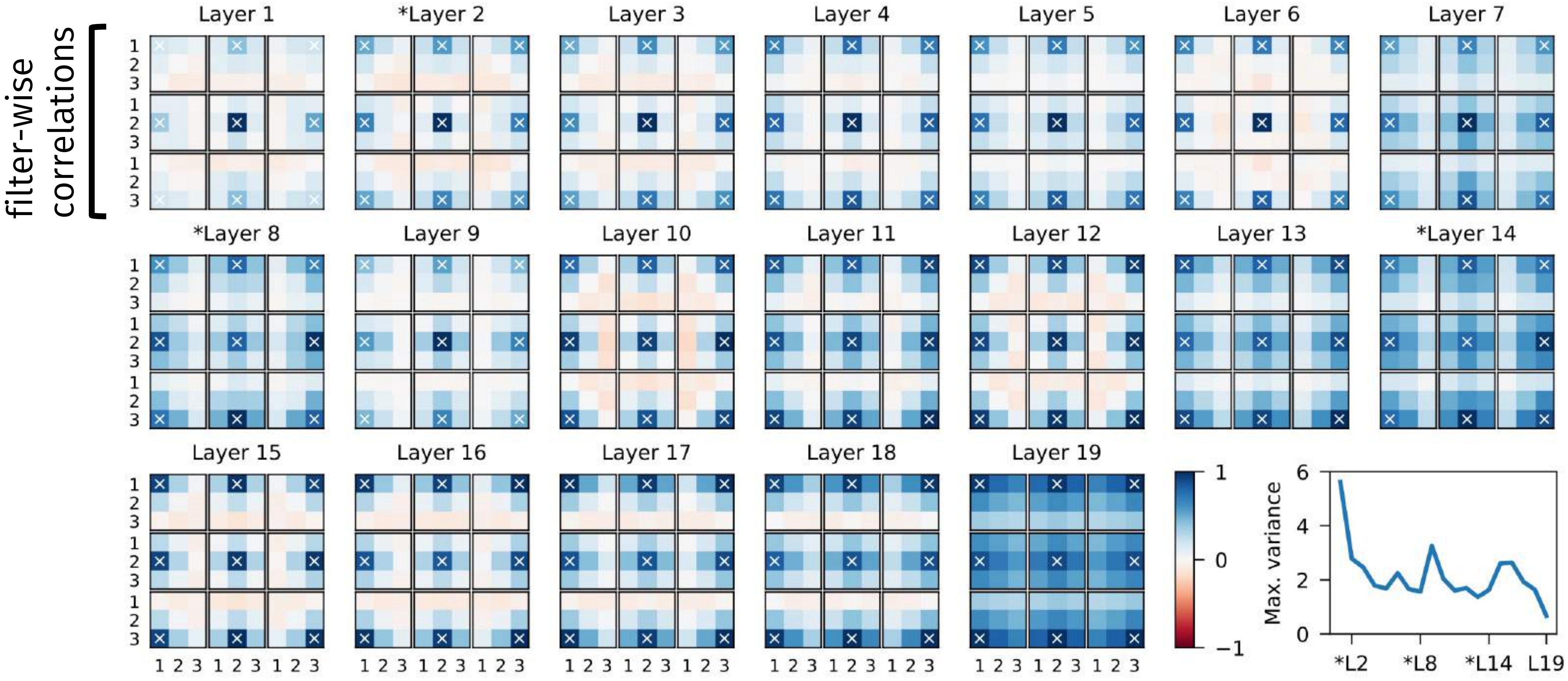
[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]





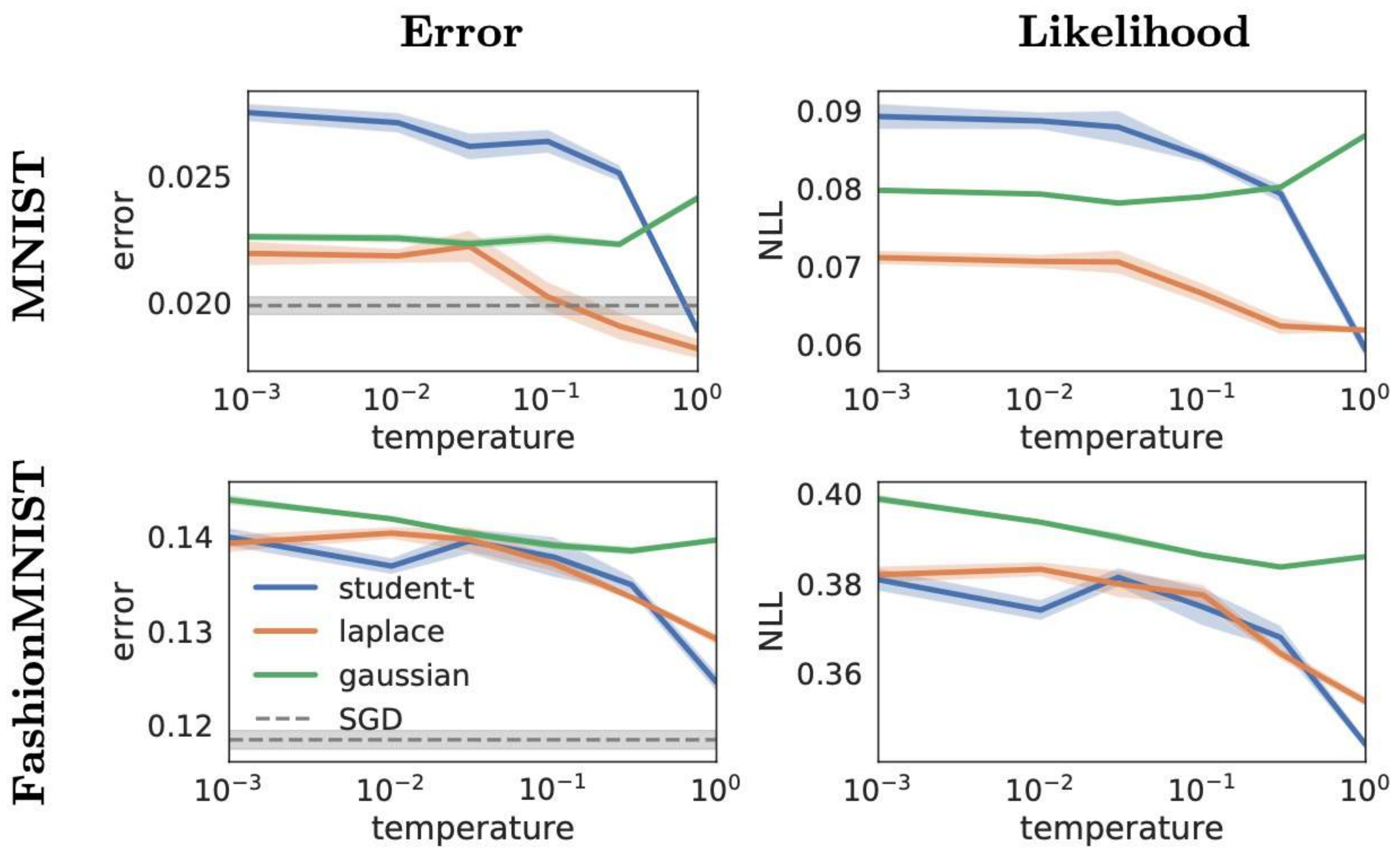
Empirical CNN weights are correlated

[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]



Bayesian FCNNs with different priors

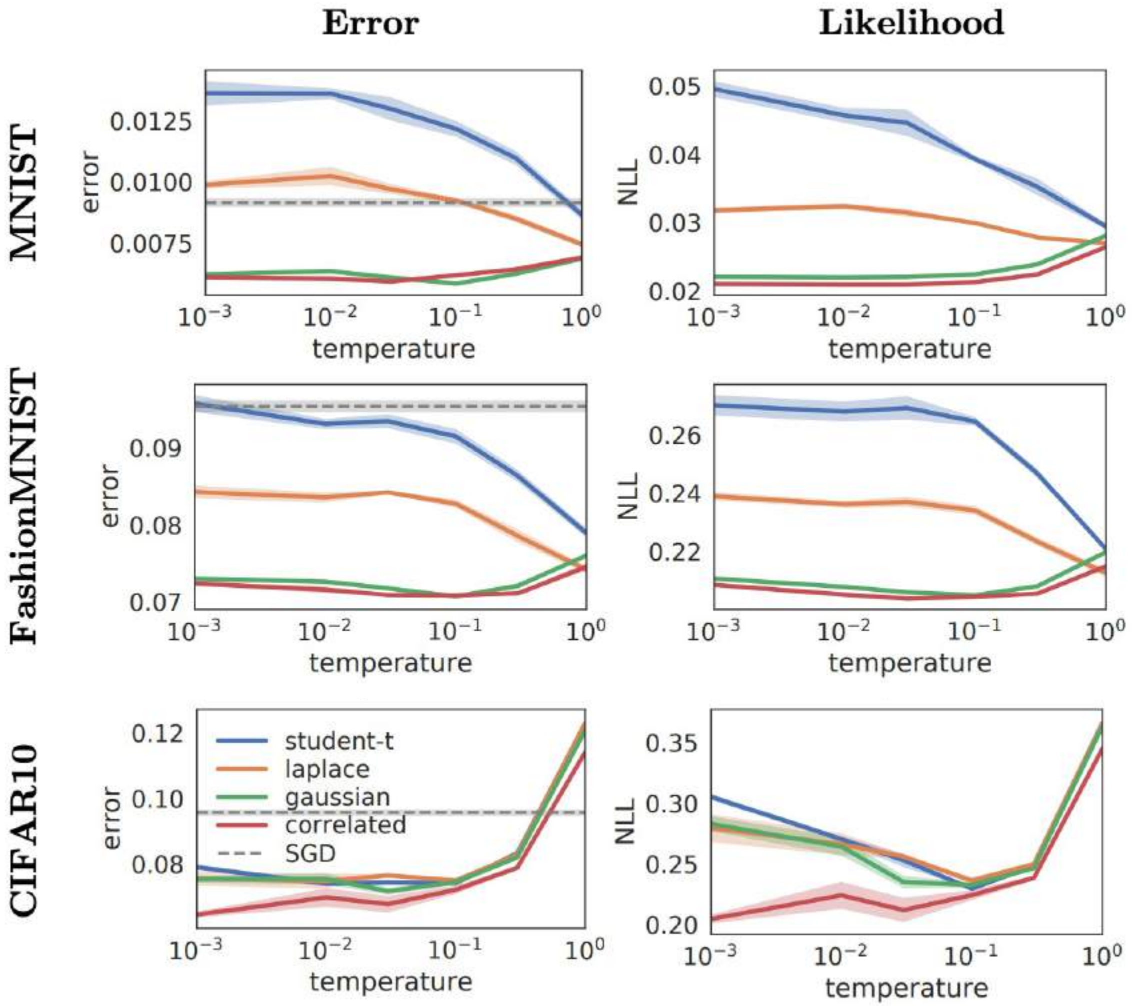
[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]



heavy-tailed priors
perform better
and reduce
cold-posterior effect

Bayesian CNNs with different priors

[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]



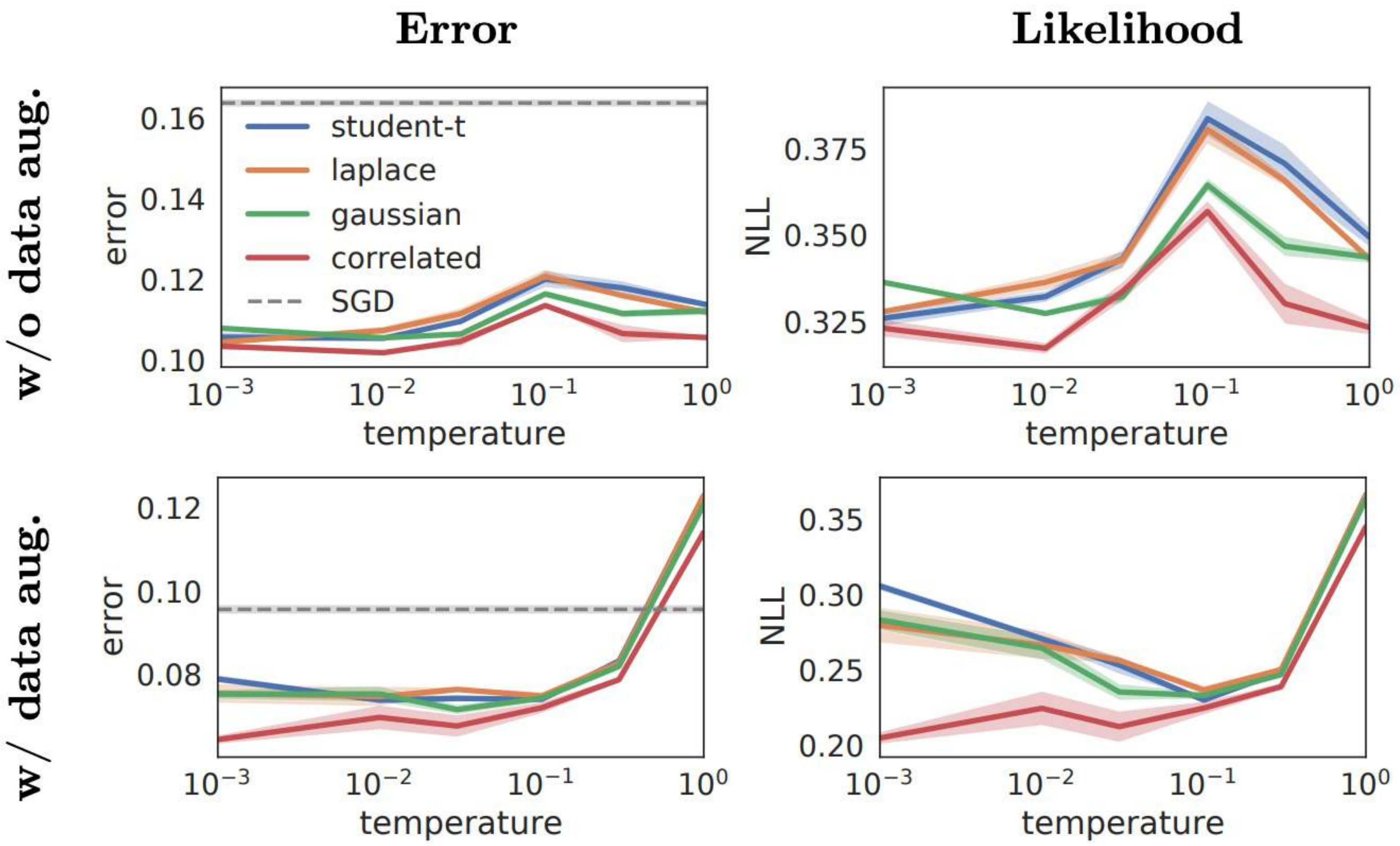
correlated prior performs better
but retains cold-posterior effect

Agenda

- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation
- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning
- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

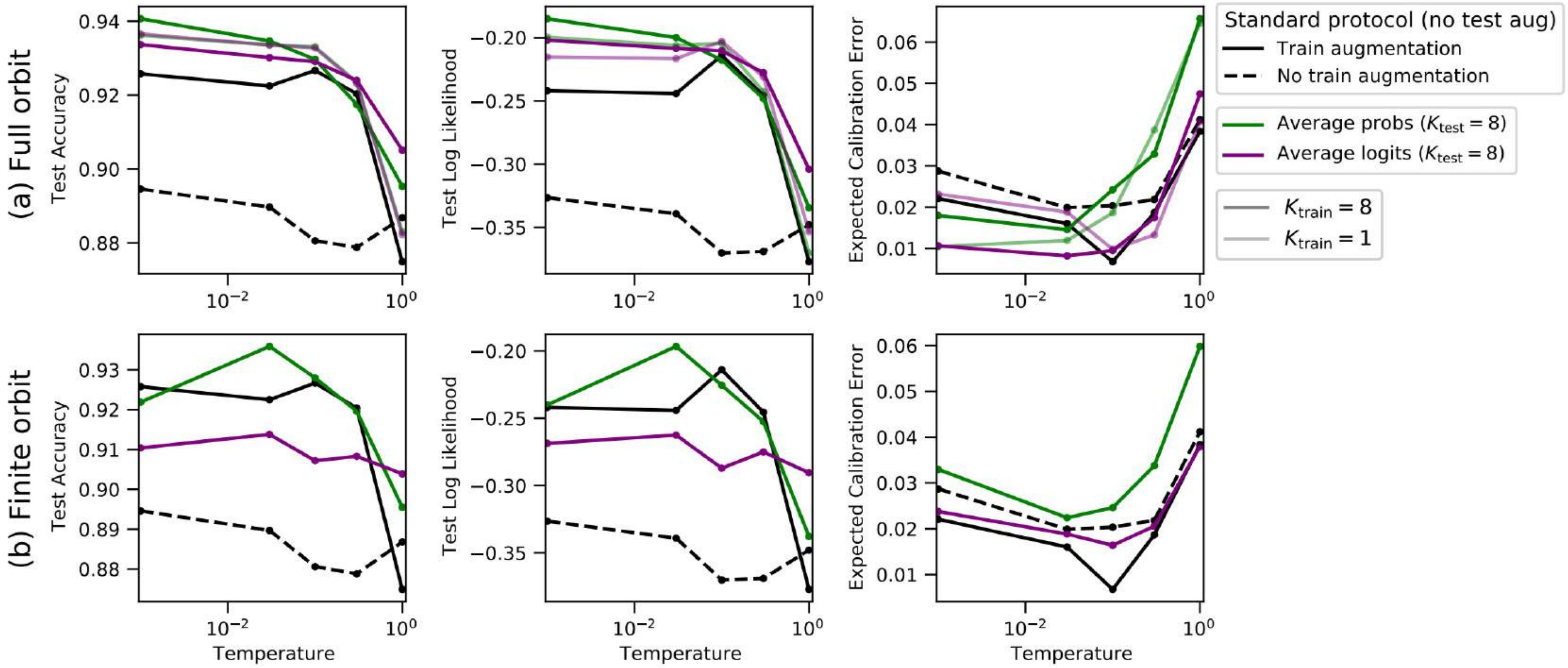
Caveat: Data augmentation plays a role!

[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]



Averaging logits/probs doesn't help

[Nabarro, Ganev, Garriga-Alonso, F, van der Wilk, Aitchison. arXiv 2021]



Agenda

- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation
- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning
- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

Marginal likelihood prior selection

[Immer, Bauer, F, Rätsch, Khan. ICML 2021]

MAP solution

$$\log p(\mathcal{D}|\mathcal{M}) \approx \log q(\mathcal{D}|\mathcal{M}) := \log p(\mathcal{D}, \boldsymbol{\theta}_*|\mathcal{M}) - \frac{1}{2}\log\left|\frac{1}{2\pi}\mathbf{H}_{\boldsymbol{\theta}_*}\right|$$

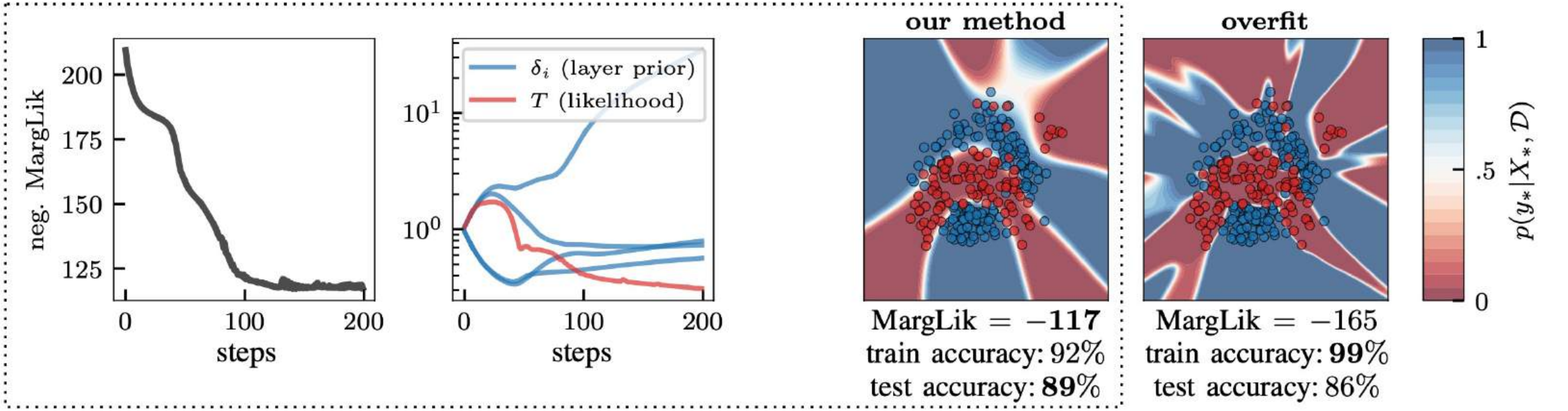
Hessian

Jacobian

$$\mathbf{H}_{\boldsymbol{\theta}} \approx \mathbf{H}_{\boldsymbol{\theta}}^{\mathrm{GGN}} = \mathbf{J}_{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{L}_{\boldsymbol{\theta}}\mathbf{J}_{\boldsymbol{\theta}} + \mathbf{P}_{\boldsymbol{\theta}}$$

$$|\mathbf{H}_{\boldsymbol{\theta}}^{\mathrm{GGN}}| \approx |\mathbf{H}_{\boldsymbol{\theta}}^{\mathrm{KFAC}}| = \prod_l \prod_{ij} \mathbf{q}_i^{(l)}\mathbf{w}_j^{(l)} + p_{\theta}^{(l)}$$

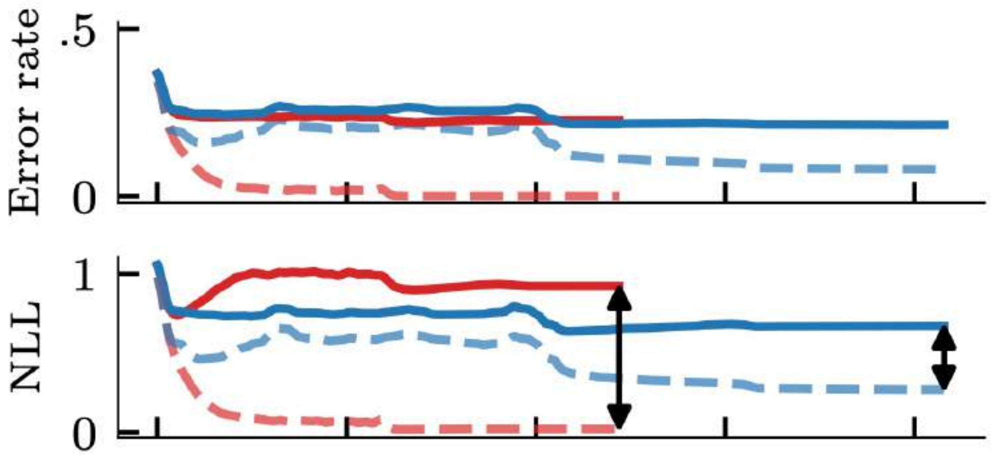
Step 1: Optimize Marginal-Likelihood wrt. hyperparameters **Step 2:** Compare marginal likelihood of models



ML-II prior improves generalization

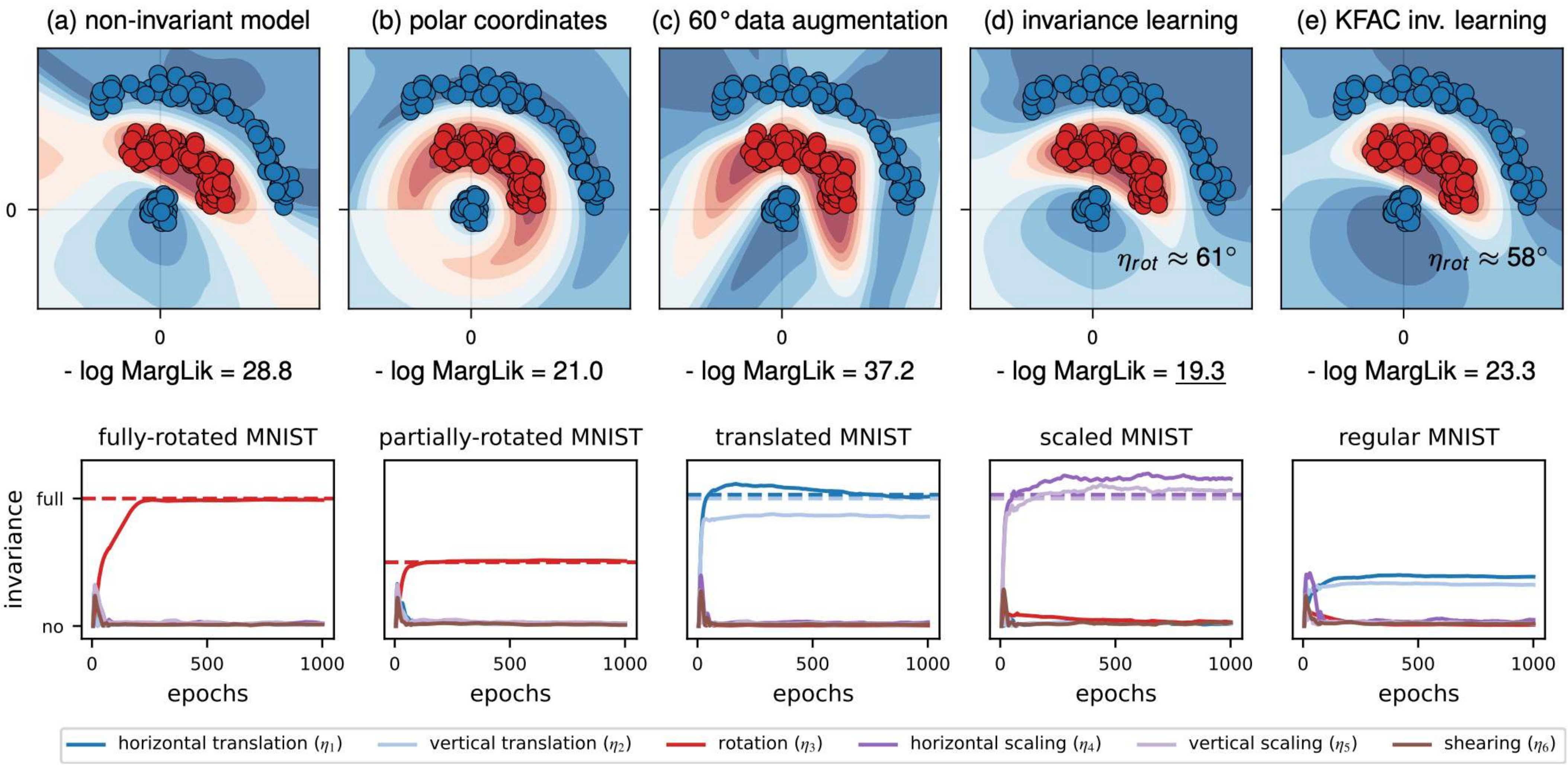
[Immer, Bauer, F, Rätsch, Khan. ICML 2021]

		cross-validation		marginal likelihood optimization					
				KFAC			diagonal EF		
Dataset	Model	accuracy	logLik	accuracy	logLik	MargLik	accuracy	logLik	MargLik
MNIST	**MLP**	98.22	−0.061	98.38	−0.053	−0.158	97.05	−0.095	−0.553
	CNN	99.40	**−0.017**	**99.46**	**−0.016**	**−0.064**	**99.45**	−0.019	−0.134
FMNIST	**MLP**	88.09	−0.347	89.83	−0.305	−0.468	85.72	−0.400	−0.756
	CNN	91.39	−0.258	**92.06**	**−0.233**	**−0.401**	91.69	**−0.233**	−0.570
CIFAR10	**CNN**	77.41	−0.680	80.46	−0.644	−0.967	80.17	−0.600	−1.359
	ResNet	83.73	−1.060	**86.11**	−0.595	**−0.717**	**85.82**	**−0.464**	−0.876



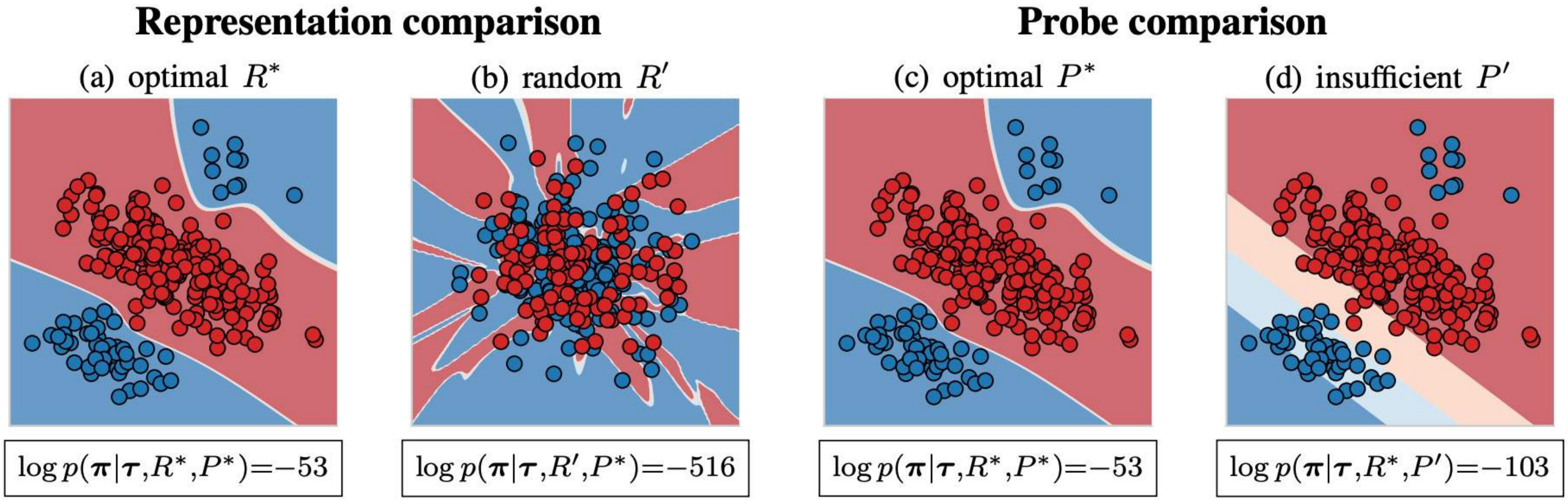
Sidenote: Learning invariances

[Immer, van der Ouderaa, F, Rätsch, van der Wilk. arXiv 2022]



Another sidenote: Linguistic probing

[Immer, Torroba-Hennigen, F, Cotterell. ACL 2022]

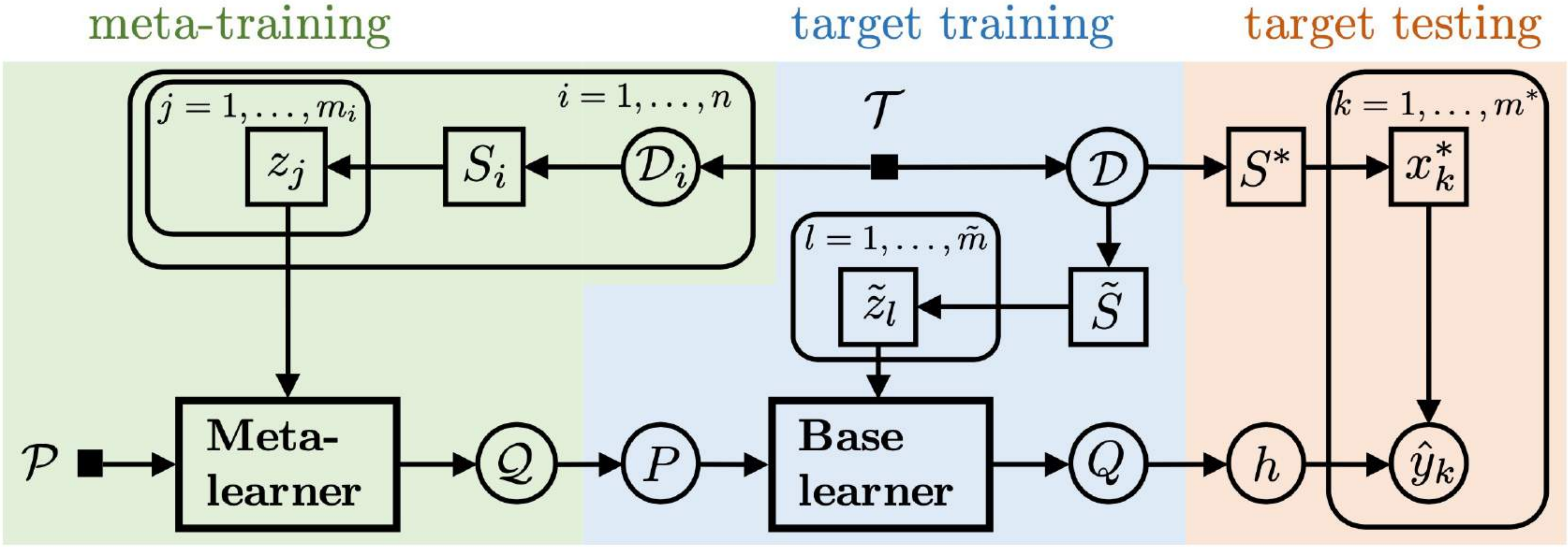


Agenda

- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation
- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning
- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

PAC-Bayesian meta-learning

[Rothfuss, F, Josifoski, Krause. ICML 2021]



PAC-Bayesian meta-learning

[Rothfuss, F, Josifoski, Krause. ICML 2021]

$$\mathcal{Q}^*(P) = \frac{\mathcal{P}(P) \exp\left(\frac{\lambda}{n\beta + \lambda} \sum_{i=1}^{n} \ln Z_\beta(S_i, P) \right)}{Z^{II}(S_1, \dots, S_n, \mathcal{P})}$$

hyperprior

marginal likelihood

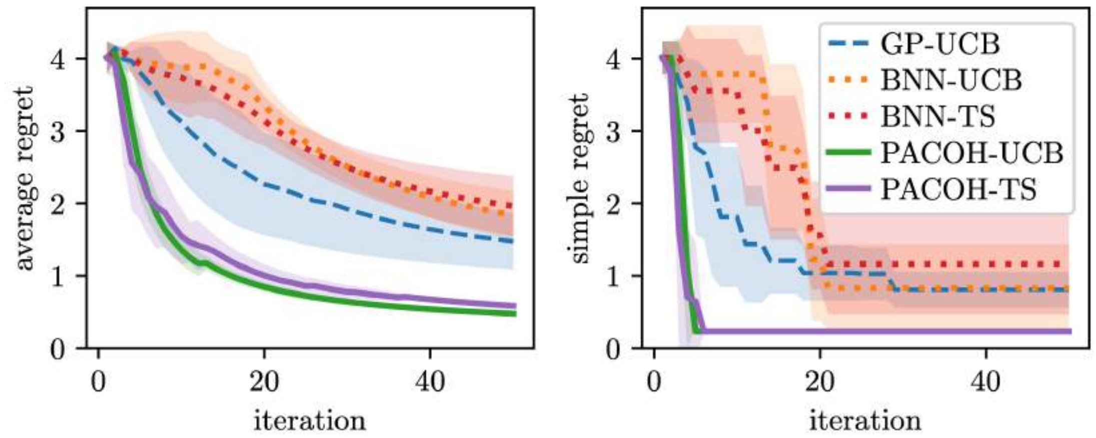
hyperposterior

PAC-Bayesian meta-learning

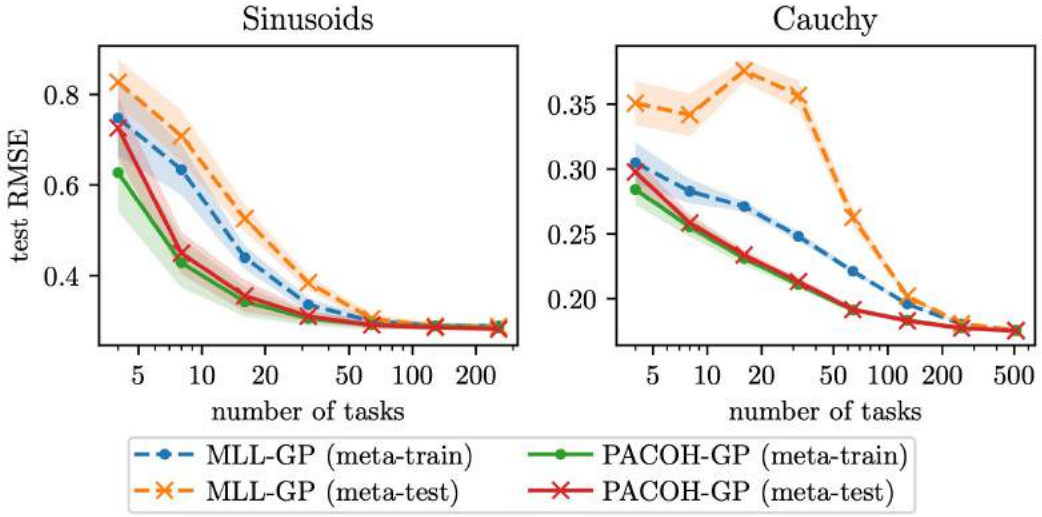
[Rothfuss, F, Josifoski, Krause. ICML 2021]

	Accuracy	Calibration error
Vanilla BNN (Liu & Wang, 2016)	0.795 ± 0.006	0.135 ± 0.009
MLAP (Amit & Meir, 2018)	0.700 ± 0.0135	0.108 ± 0.010
MAML (Finn et al., 2017)	0.693 ± 0.013	0.109 ± 0.011
BMAML (Kim et al., 2018)	0.764 ± 0.025	0.191 ± 0.018
PACOH-NN (ours)	$\mathbf{0.885 \pm 0.090}$	$\mathbf{0.091 \pm 0.010}$

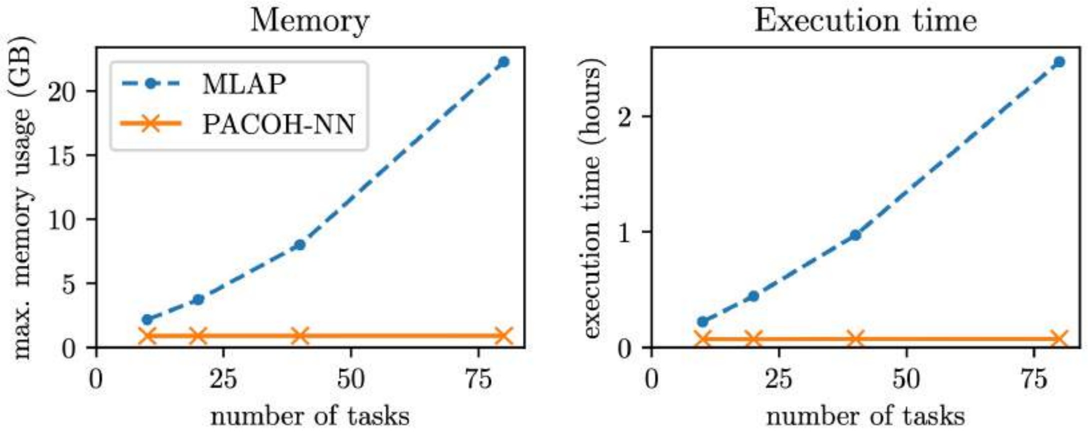
Few-shot learning on Omniglot



Bandit task



Meta-overfitting

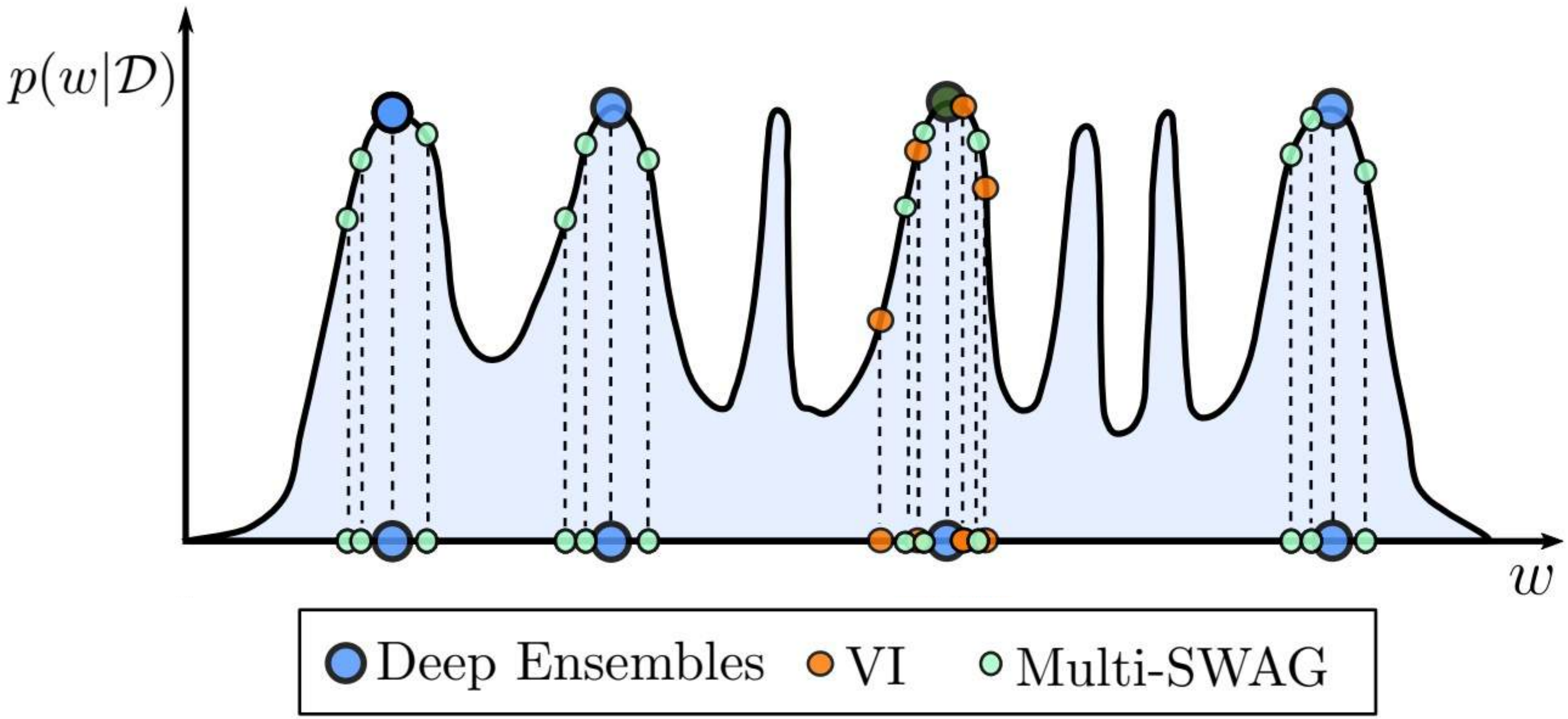


Runtimes

Agenda

- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation
- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning
- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

Background: Posterior coverage



[Wilson, Izmailov 2020]

Repulsive deep ensembles

[D'Angelo, F. NeurIPS 2021]

Standard deep ensembles:

$$\mathbf{w}_i^{t+1} \leftarrow \mathbf{w}_i^t + \epsilon_t \phi(\mathbf{w}_i^t)$$

$$\phi(\mathbf{w}_i^t) = \nabla_{\mathbf{w}_i^t} \log p(\mathbf{w}_i^t | \mathcal{D})$$

Repulsive deep ensembles:

kernel

$$\phi(\mathbf{w}_i^t) = \nabla_{\mathbf{w}_i^t} \log p(\mathbf{w}_i^t | \mathcal{D}) - \mathcal{R}\left(\left\{\nabla_{\mathbf{w}_i^t} k(\mathbf{w}_i^t, \mathbf{w}_j^t)\right\}_{j=1}^{n}\right)$$

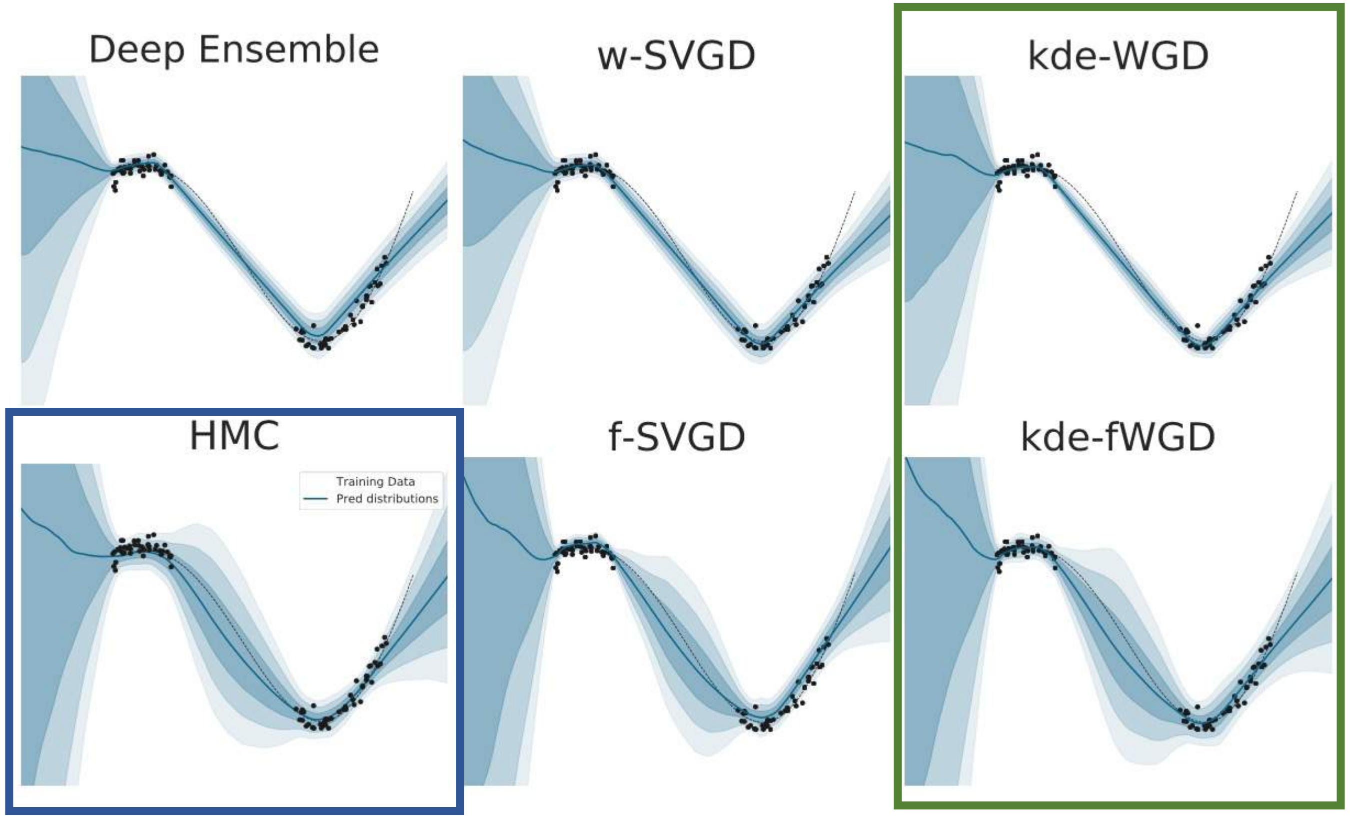
Function-space repulsive deep ensembles:

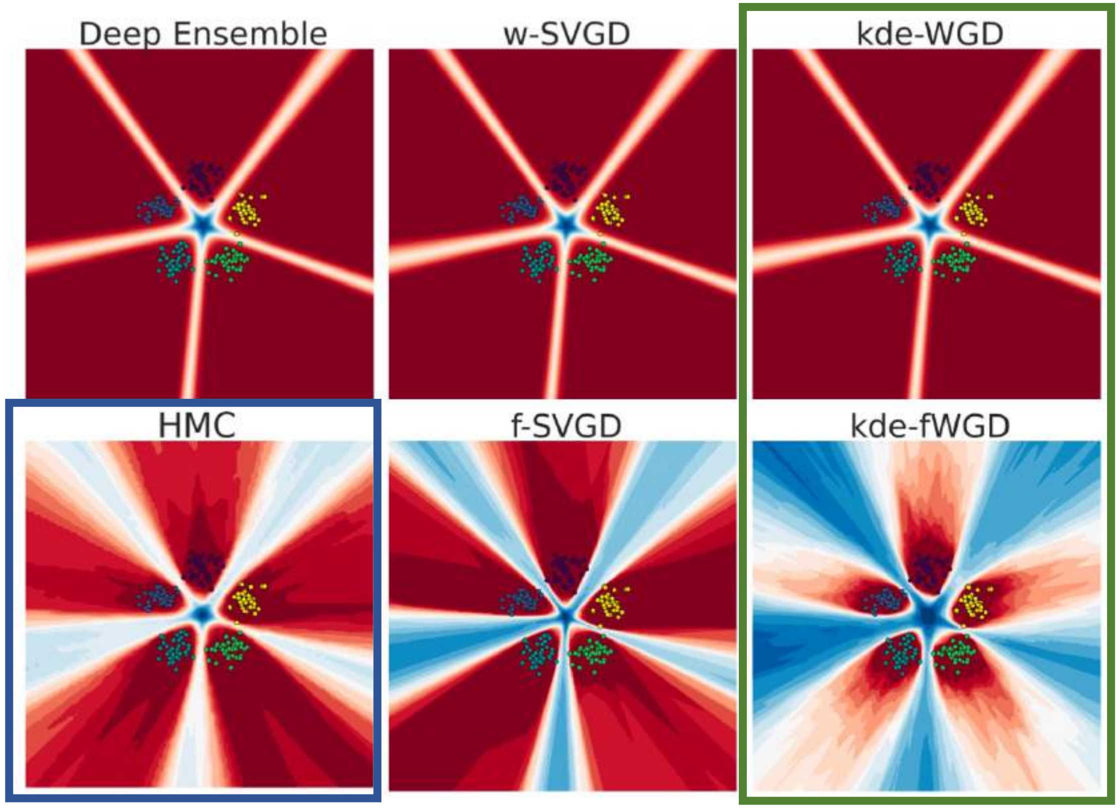
canonical projection

$$\phi(\mathbf{w}_i^t) = \left(\frac{\partial \boldsymbol{f}_i^t}{\partial \mathbf{w}_i^t}\right)^\top \left[\nabla_{\boldsymbol{f}_i^t} \log p(\boldsymbol{f}_i^t | \mathcal{D}) - \mathcal{R}\left(\left\{\nabla_{\boldsymbol{f}_i^t} k(\pi_B(\boldsymbol{f}_i^t), \pi_B(\boldsymbol{f}_j^t))\right\}_{j=1}^{n}\right)\right]$$

Repulsion approximates the posterior

[D‘Angelo, F. NeurIPS 2021]





„gold standard“

our model

„gold standard“

our model

Agenda

- Pathologies of common BNN priors
  - BNN priors and the cold posterior effect
  - The role of data augmentation

- How to find better priors
  - Empirical Bayes using the marginal likelihood
  - (PAC-)Bayesian meta-learning

- How to use function-space priors
  - Repulsive deep ensembles
  - GP priors in the latent space

GP priors in the latent space

[F, Collier, Wenzel, Liu, Allingham, Tran, Lakshminarayanan, Berent, Jenatton, Kokiopoulou. AABI 2022]

SNGP-distributed latent means
(correlated across data points)

SNGP kernel
(RFF approximation)

$$\boldsymbol{f}_c \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_\theta(\boldsymbol{x}, \boldsymbol{x}))$$

$$\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{f}_i, \boldsymbol{\Sigma}(\boldsymbol{x}_i; \varphi))$$

heteroscedastic logits
(correlated across classes)

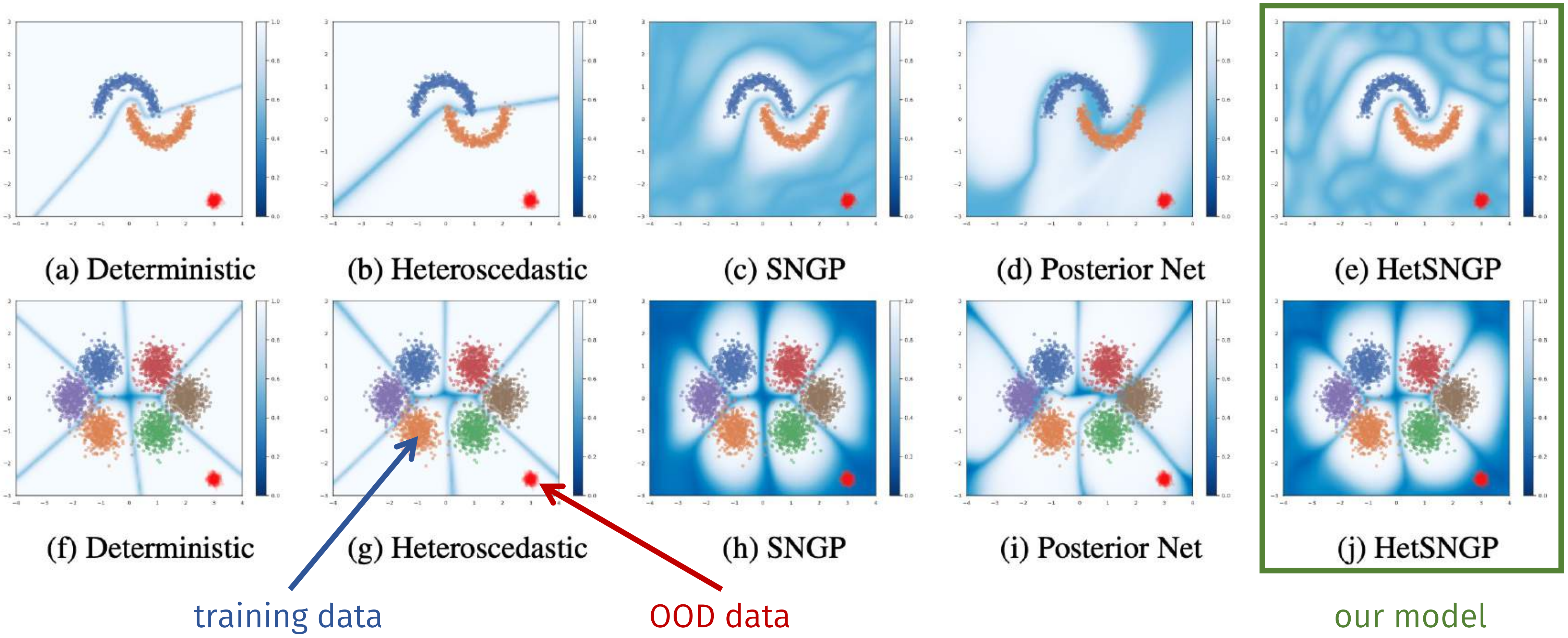
label noise covariance
(possibly low-rank)

$$p(y_i = c \,|\, \boldsymbol{u}_i) = \mathbb{1}\left[c = \arg\max_k \, u_{ik}\right]$$

output probabilities (approximated by softmax)

Distance-aware OOD uncertainties

[F, Collier, Wenzel, Liu, Allingham, Tran, Lakshminarayanan, Berent, Jenatton, Kokiopoulou. AABI 2022]



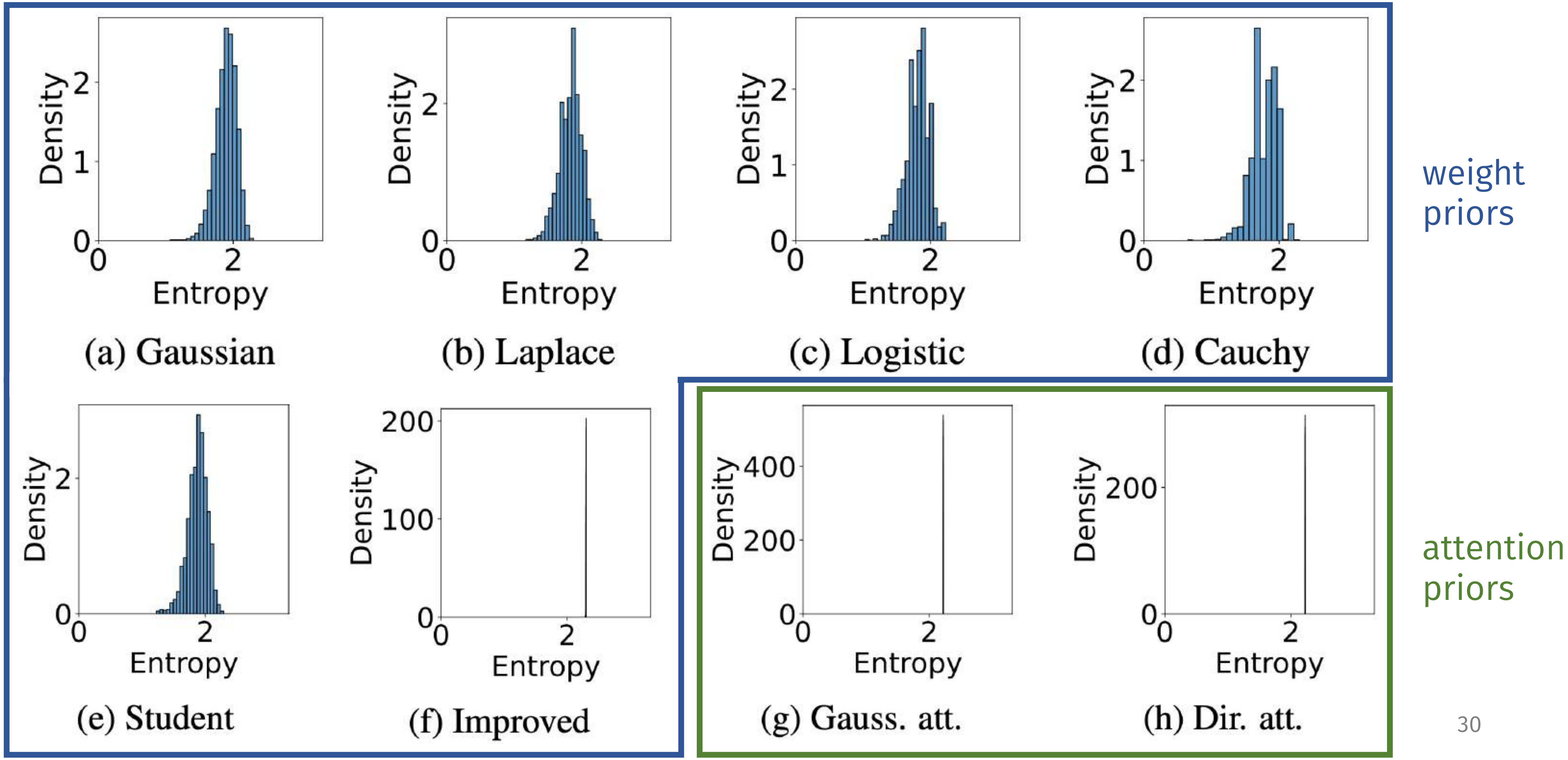
Label noise modeling in real datasets

[F, Collier, Wenzel, Liu, Allingham, Tran, Lakshminarayanan, Berent, Jenatton, Kokiopoulou. AABI 2022]

Method	↑ID prec@1	↑Im Acc	↑ImC Acc	↑ImA Acc	↑ImR Acc	↑ImV2 Acc
Det.	0.471 $\pm$ 0.000	0.800 $\pm$ 0.000	0.603 $\pm$ 0.000	0.149 $\pm$ 0.000	0.311 $\pm$ 0.000	0.694 $\pm$ 0.000
Het.	**0.480** $\pm$ 0.001	0.796 $\pm$ 0.002	0.590 $\pm$ 0.001	0.132 $\pm$ 0.004	0.300 $\pm$ 0.006	0.687 $\pm$ 0.000
SNGP	0.468 $\pm$ 0.001	0.799 $\pm$ 0.001	0.602 $\pm$ 0.000	0.165 $\pm$ 0.003	0.328 $\pm$ 0.005	0.696 $\pm$ 0.003
HetSNGP	0.477 $\pm$ 0.001	**0.806** $\pm$ 0.001	**0.613** $\pm$ 0.003	**0.172** $\pm$ 0.007	**0.336** $\pm$ 0.002	**0.705** $\pm$ 0.001

Sidenote: Attention prior in transformers

[Cinquin, Immer, Horn, F. AABI 2022]



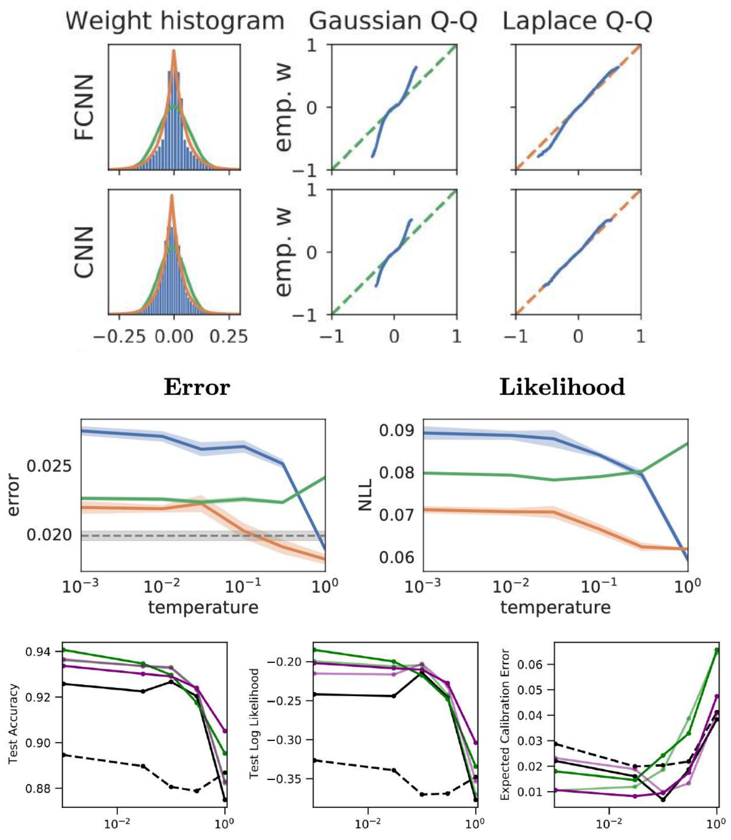
Improving the prior helps in all tasks

[Cinquin, Immer, Horn, F. AABI 2022]

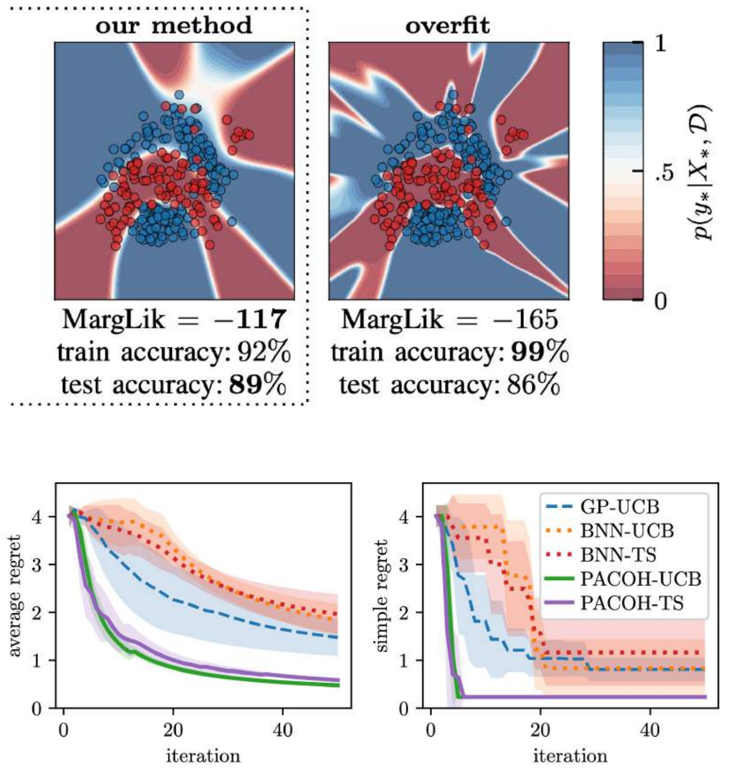
Dataset	Gauss. VI	Laplace VI	Logistic VI	Cauchy VI	Student VI
M1	1.40%	3.80%	4.12%	1.85%	2.79%
M2	2.85%	3.06%	2.76%	4.36%	2.70%
POS	0.12%	2.05%	2.16%	0.87%	-0.32%
MNIST	26.95%	33.31%	31.36%	5.66%	26.94%

Percentage of improvement changing from standard to improved prior

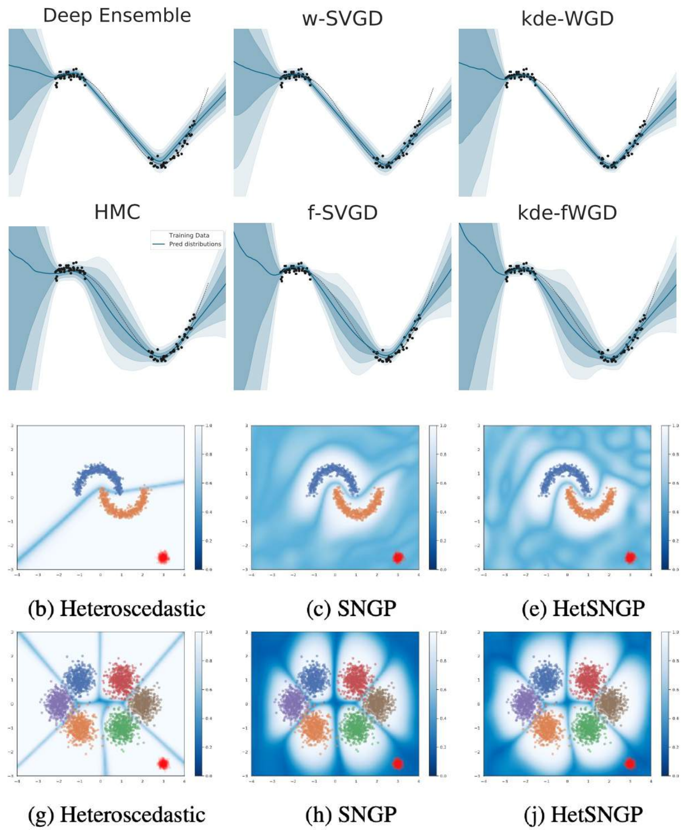
Take-home messages



Common weight-space BNN priors have several pathologies



We can find better priors using ML-II or meta-learning



Function-space priors can work better than weight-space ones

Thank you!

Deepmind

Matthias Bauer

EPF Lausanne

Martin Josifoski

ETH Zürich

Tristan Cinquin
Ryan Cotterell
Francesco D'Angelo
Max Horn
Alexander Immer
Andreas Krause
Gunnar Rätsch
Jonas Rothfuss

Google

Jesse Berent
Mark Collier
Rodolphe Jenatton
Effrosyni Kokiopoulou
Balaji Lakshminarayanan
Jeremiah Liu
Dustin Tran
Florian Wenzel

Imperial College London

Seth Nabarro
Tycho van der Ouderaa
Mark van der Wilk

MIT

Lucas Torroba-Hennigen

RIKEN

Mohammad Emtiyaz Khan

University of Bristol

Laurence Aitchison
Stoil Ganev

University of Cambridge

James Allingham
Adrià Garriga-Alonso
Sebastian Ober
Richard Turner

fortuin.github.io

vbf21@cam.ac.uk

@vincefort