ETH zürich

UNIVERSITY OF CAMBRIDGE

On the Importance of Priors in Bayesian Deep Learning

Dr. Vincent Fortuin

Bayesian World Meeting

Montreal, June 2022

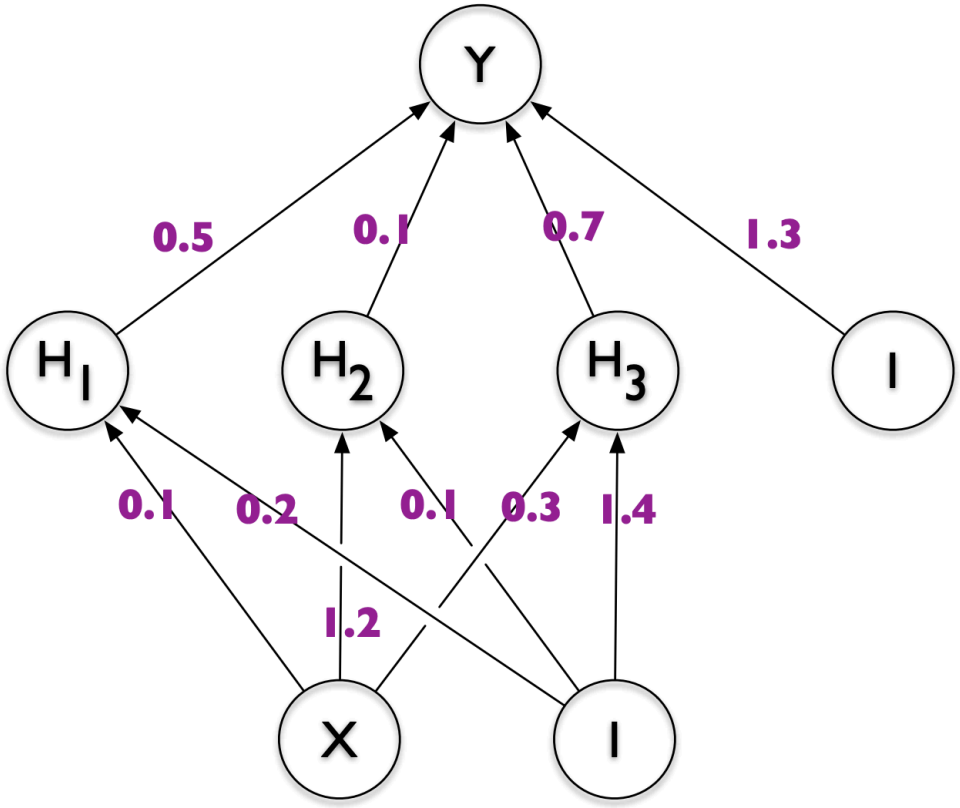
Agenda

- Pathologies of common BNN priors
- How to find better priors using the marginal likelihood
- PAC-Bayesian meta-learning for priors

Agenda

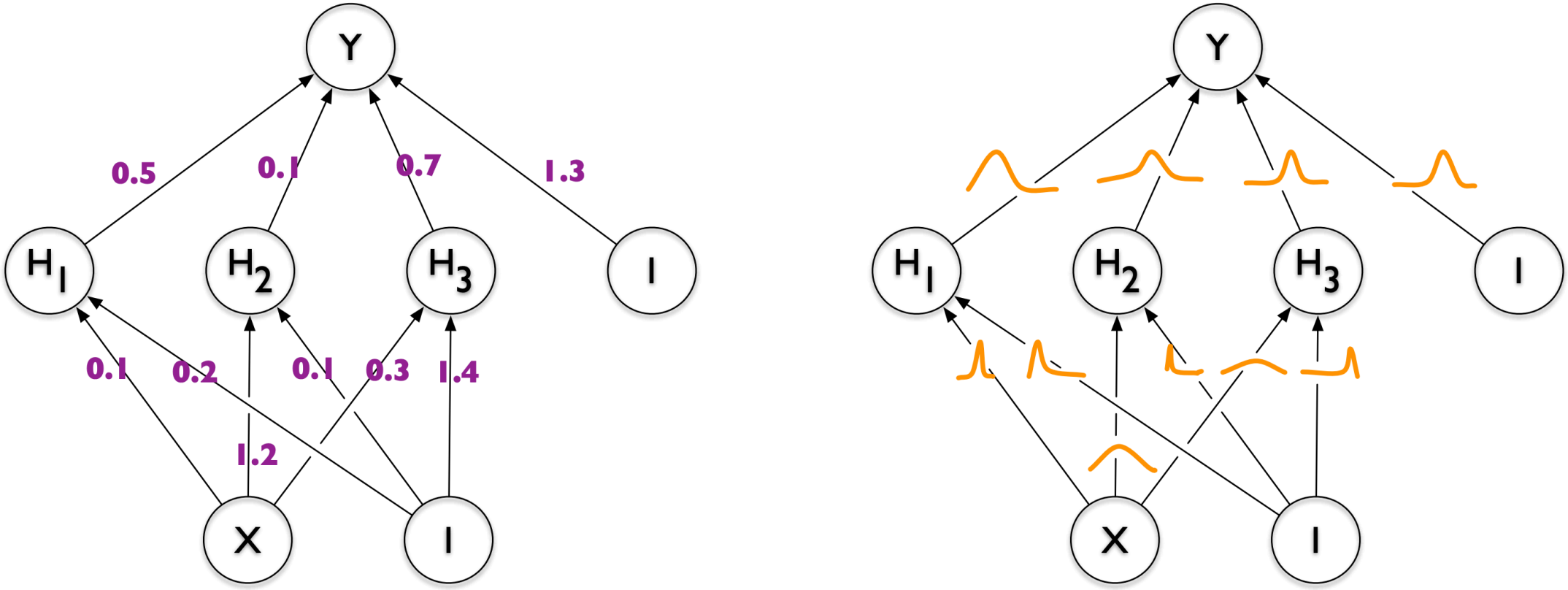
- Pathologies of common BNN priors
- How to find better priors using the marginal likelihood
- PAC-Bayesian meta-learning for priors

Background: Bayesian Neural Networks



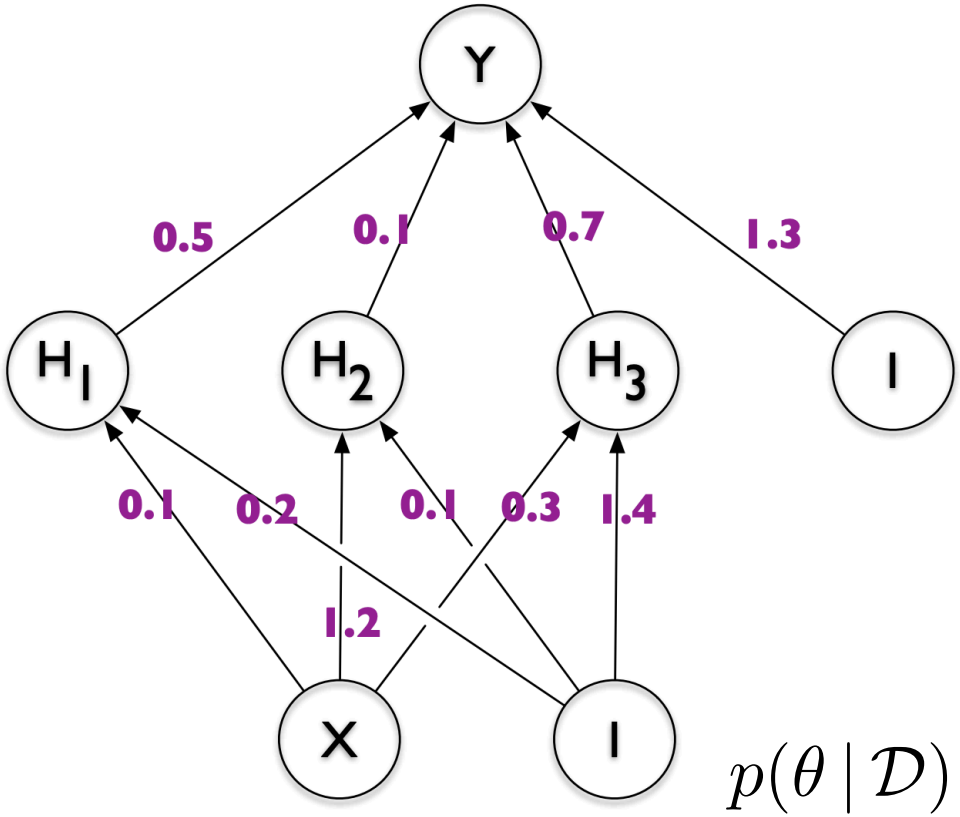
[Blundell, Cornebise, Kavukcuoglu, Wierstra 2015]

Background: Bayesian Neural Networks

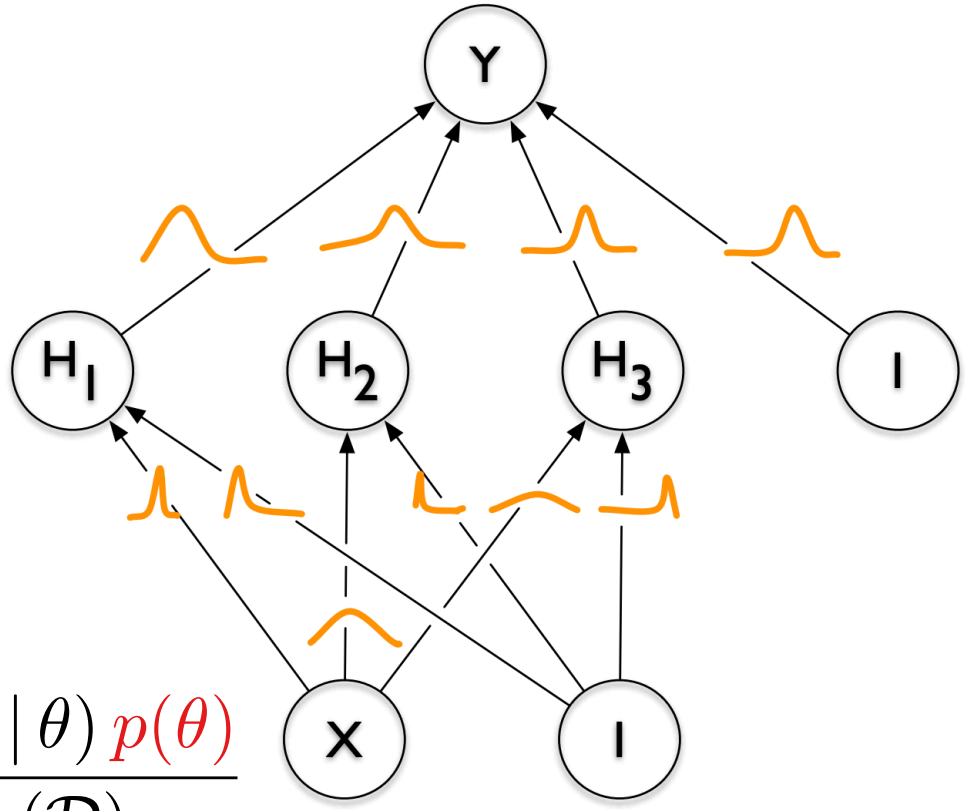


[Blundell, Cornebise, Kavukcuoglu, Wierstra 2015]

Background: Bayesian Neural Networks

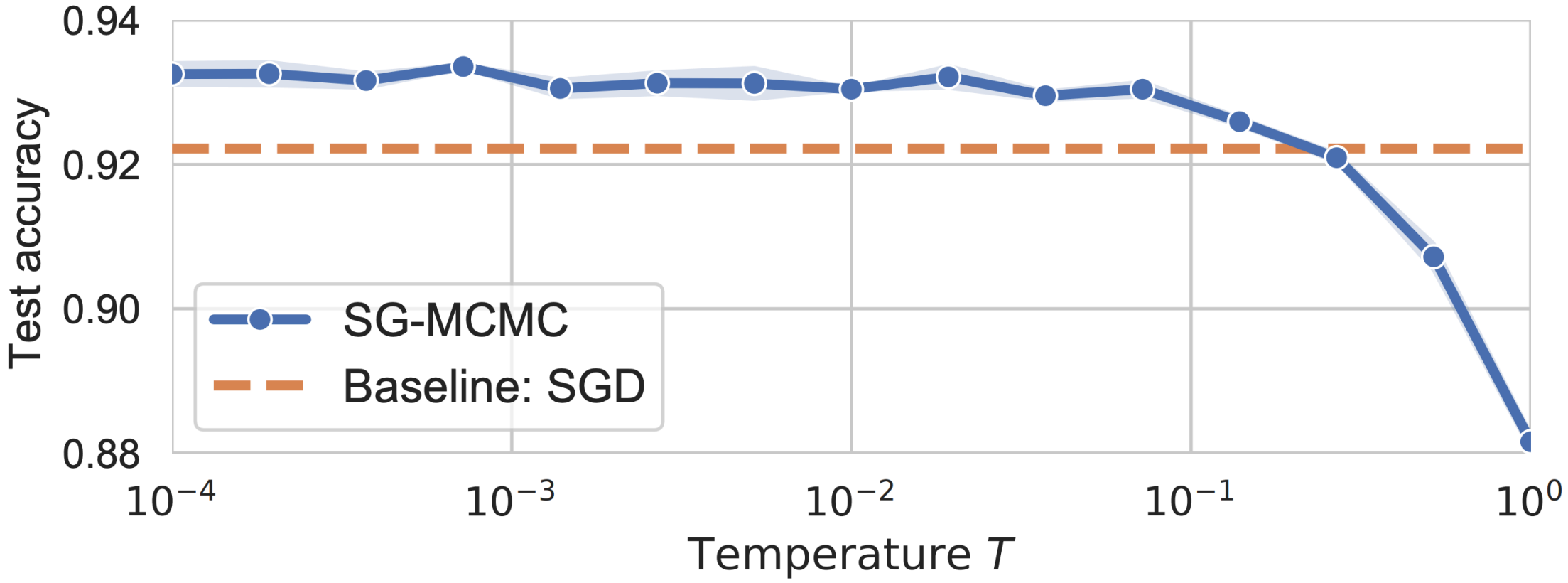


$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}$$



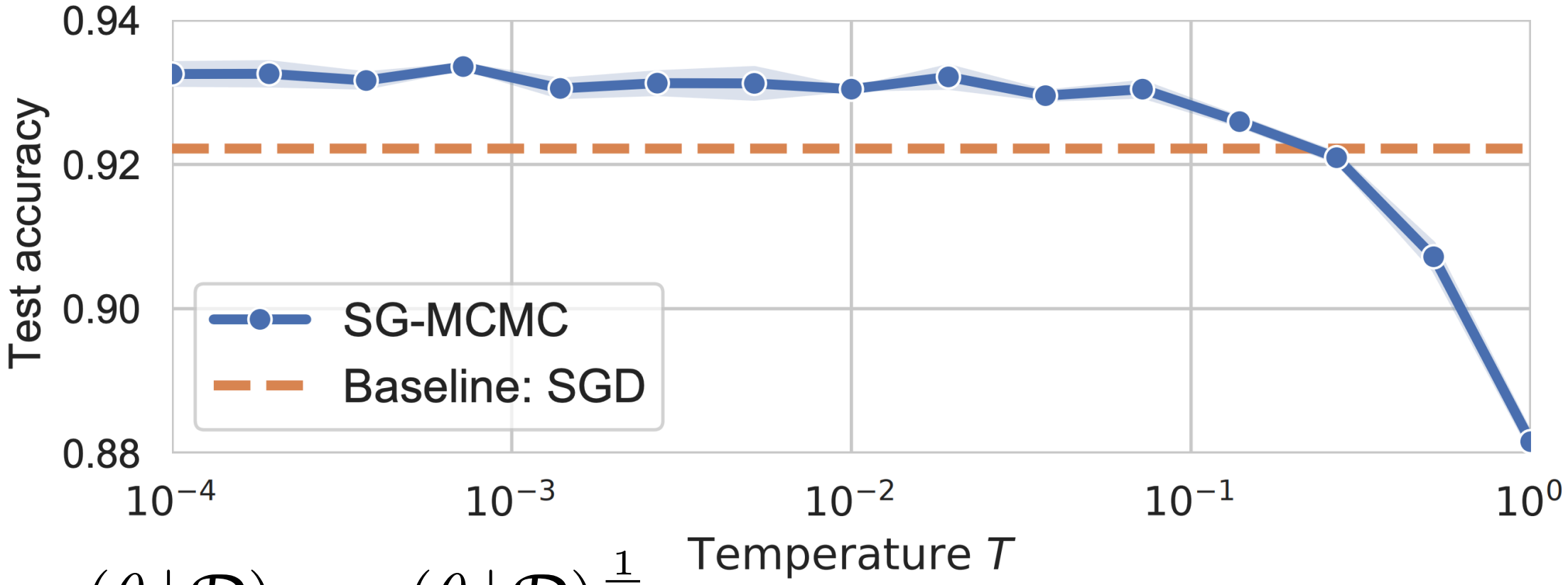
[Blundell, Cornebise, Kavukcuoglu, Wierstra 2015]

Motivation: Cold-posterior effect



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

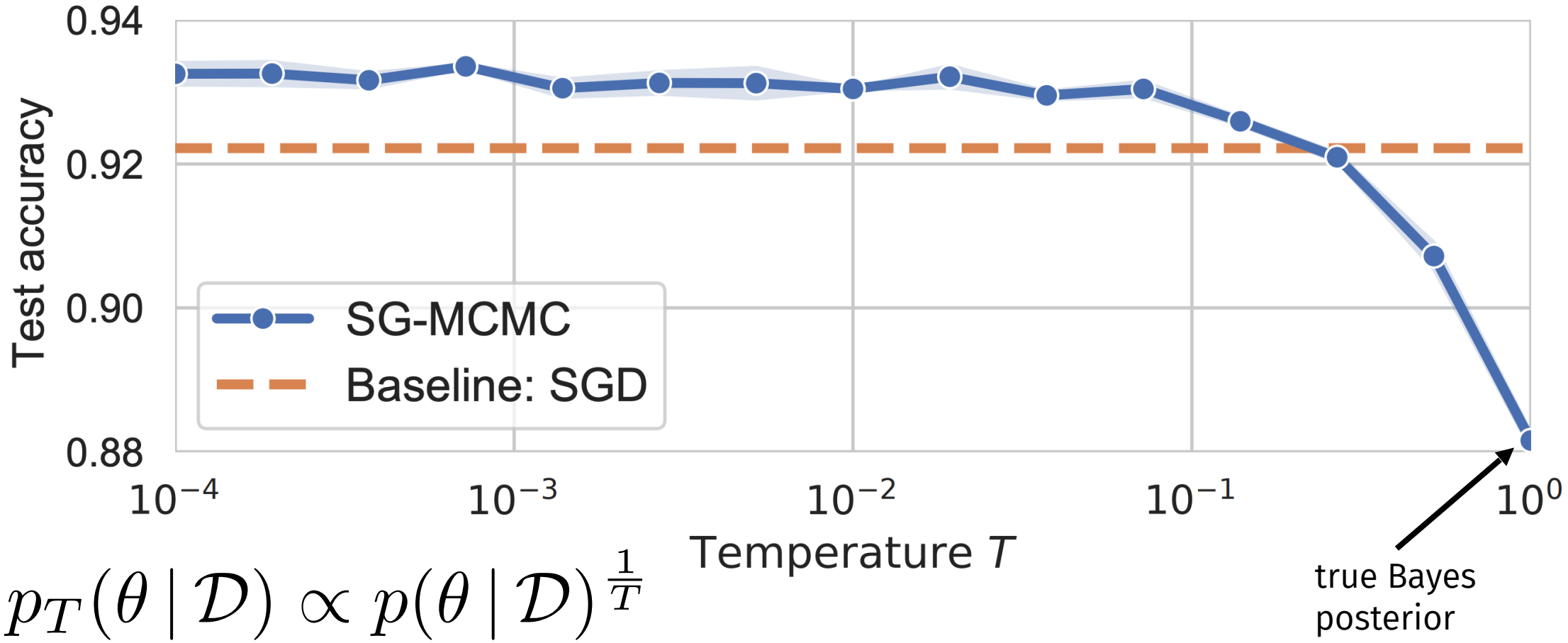
Motivation: Cold-posterior effect



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

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

Motivation: Cold-posterior effect

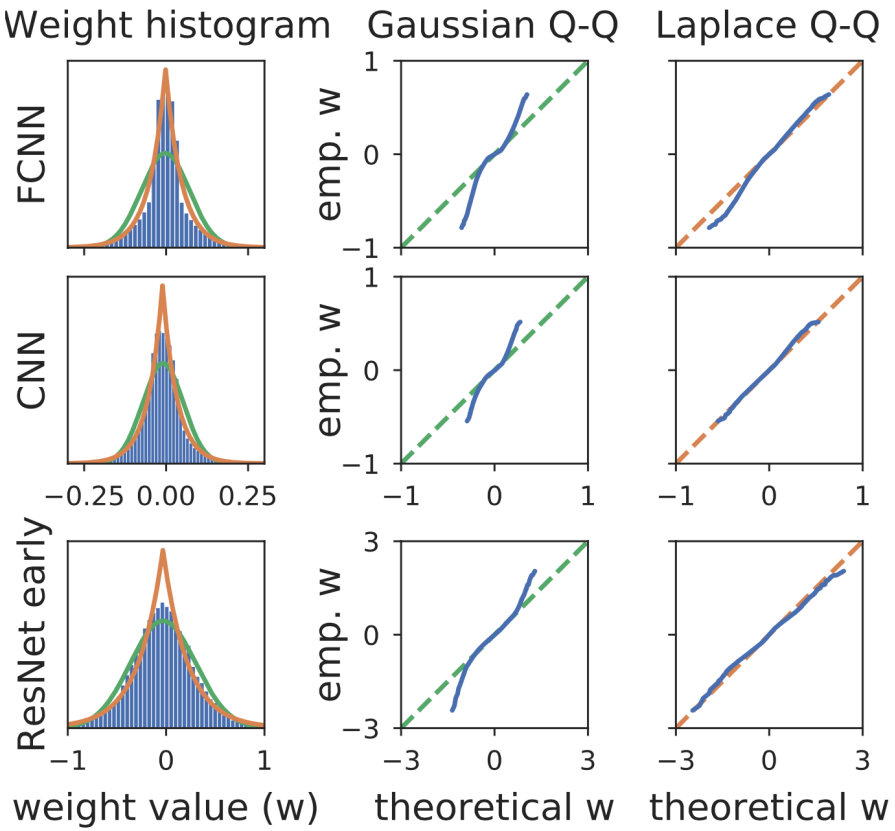


$$p_T(\theta \mid \mathcal{D}) \propto p(\theta \mid \mathcal{D})^{\frac{1}{T}}$$

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

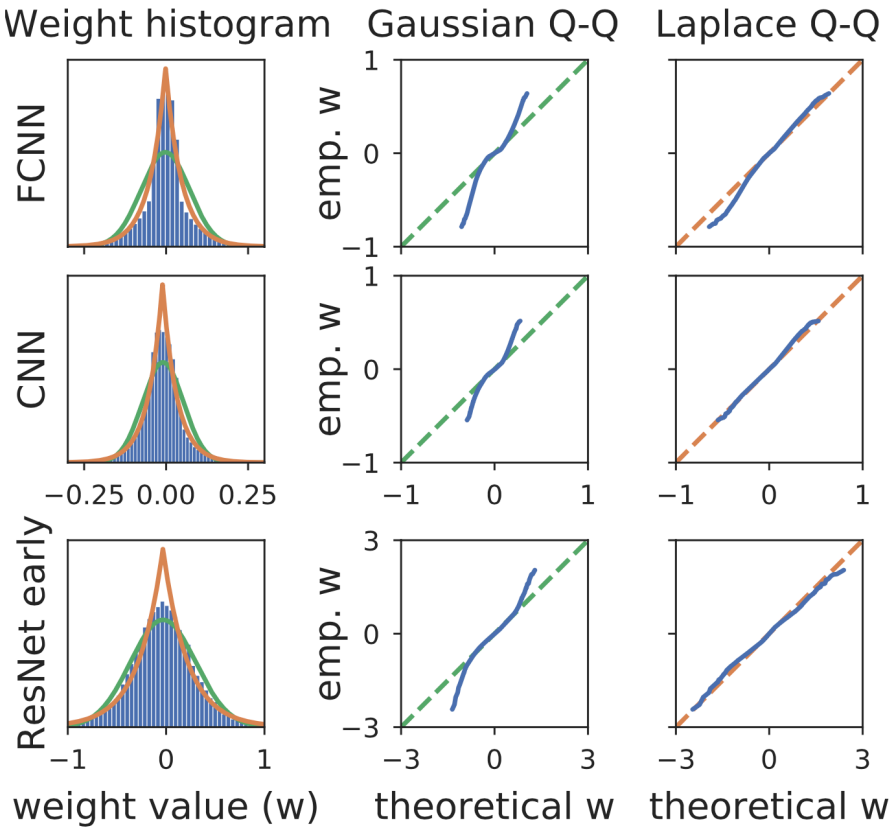
Empirical FCNN weights are heavy-tailed

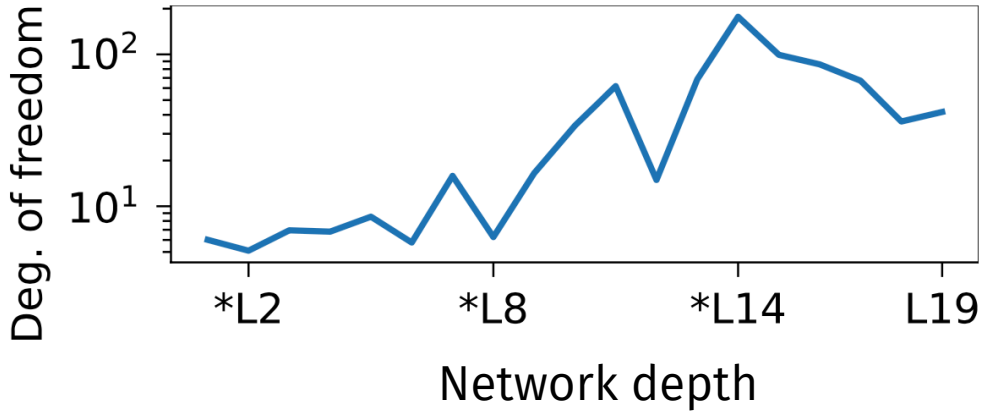
[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]



Empirical FCNN weights are heavy-tailed

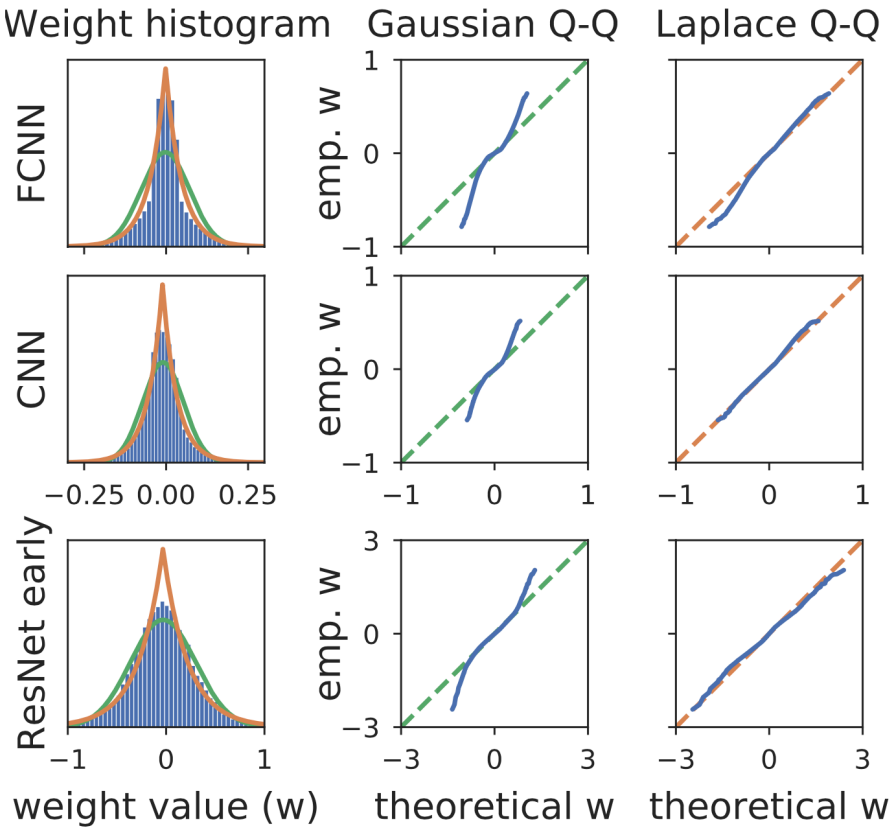
[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]

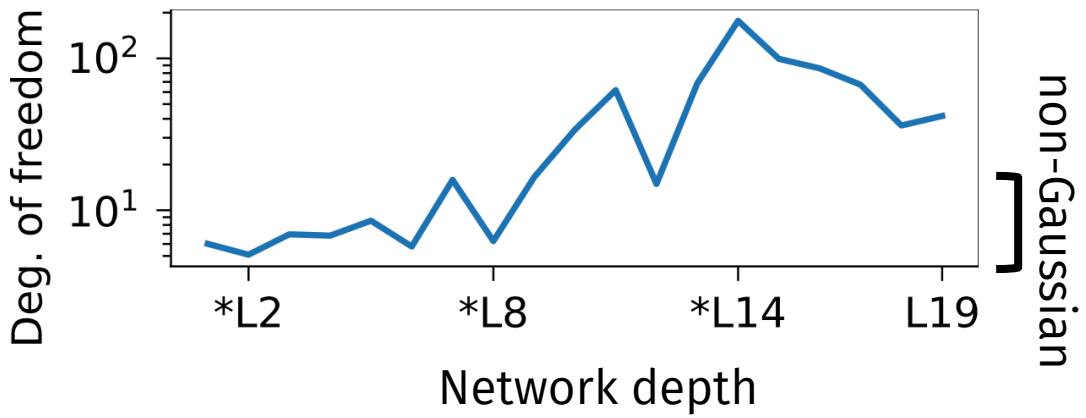




Empirical FCNN weights are heavy-tailed

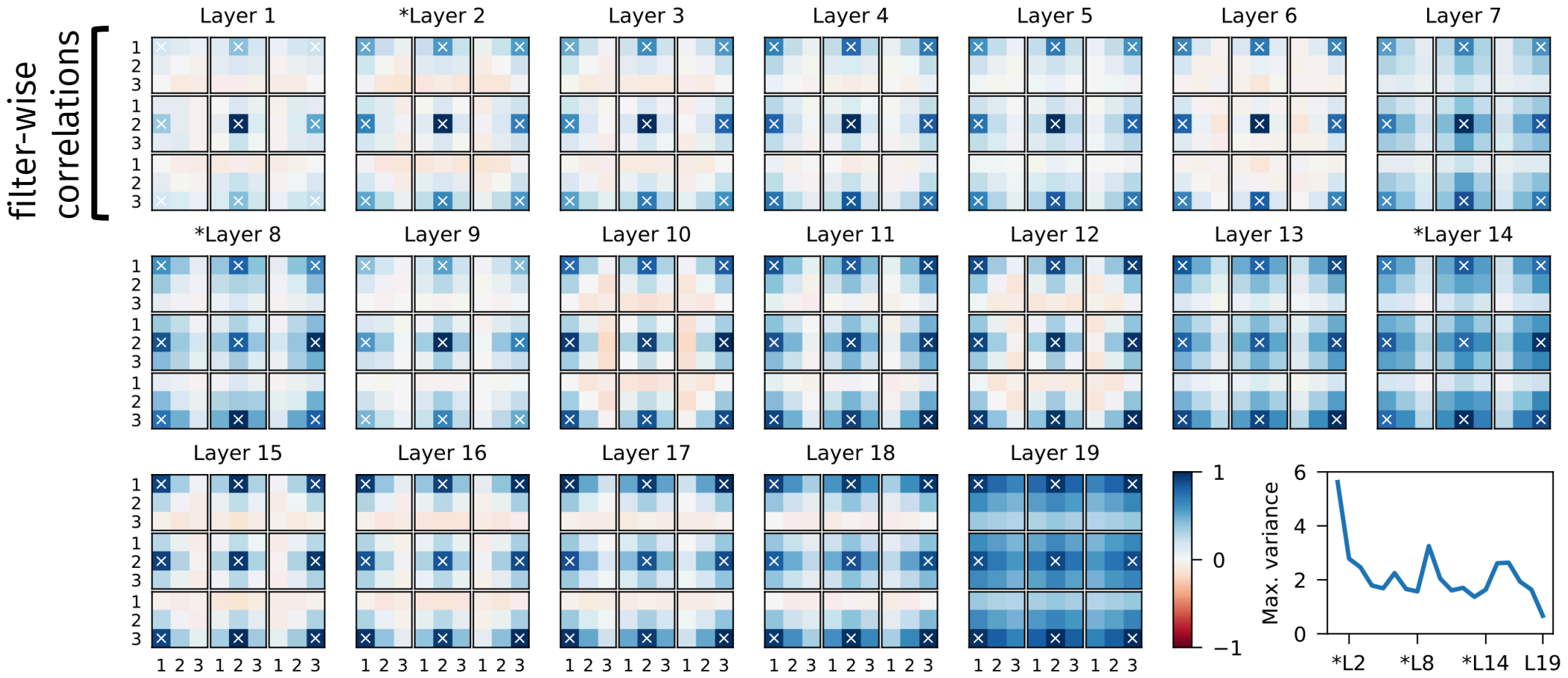
[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]





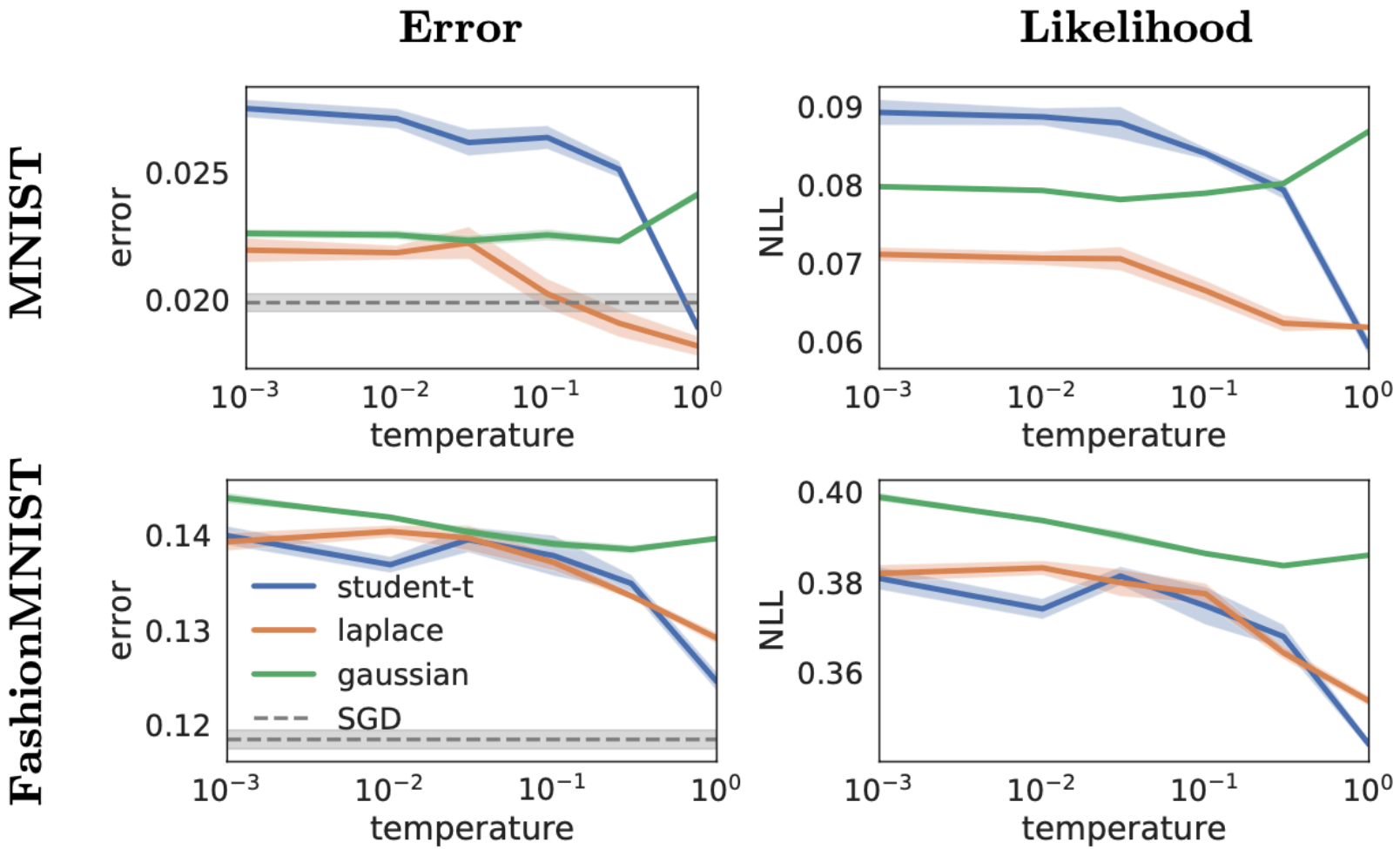
Empirical CNN weights are correlated

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



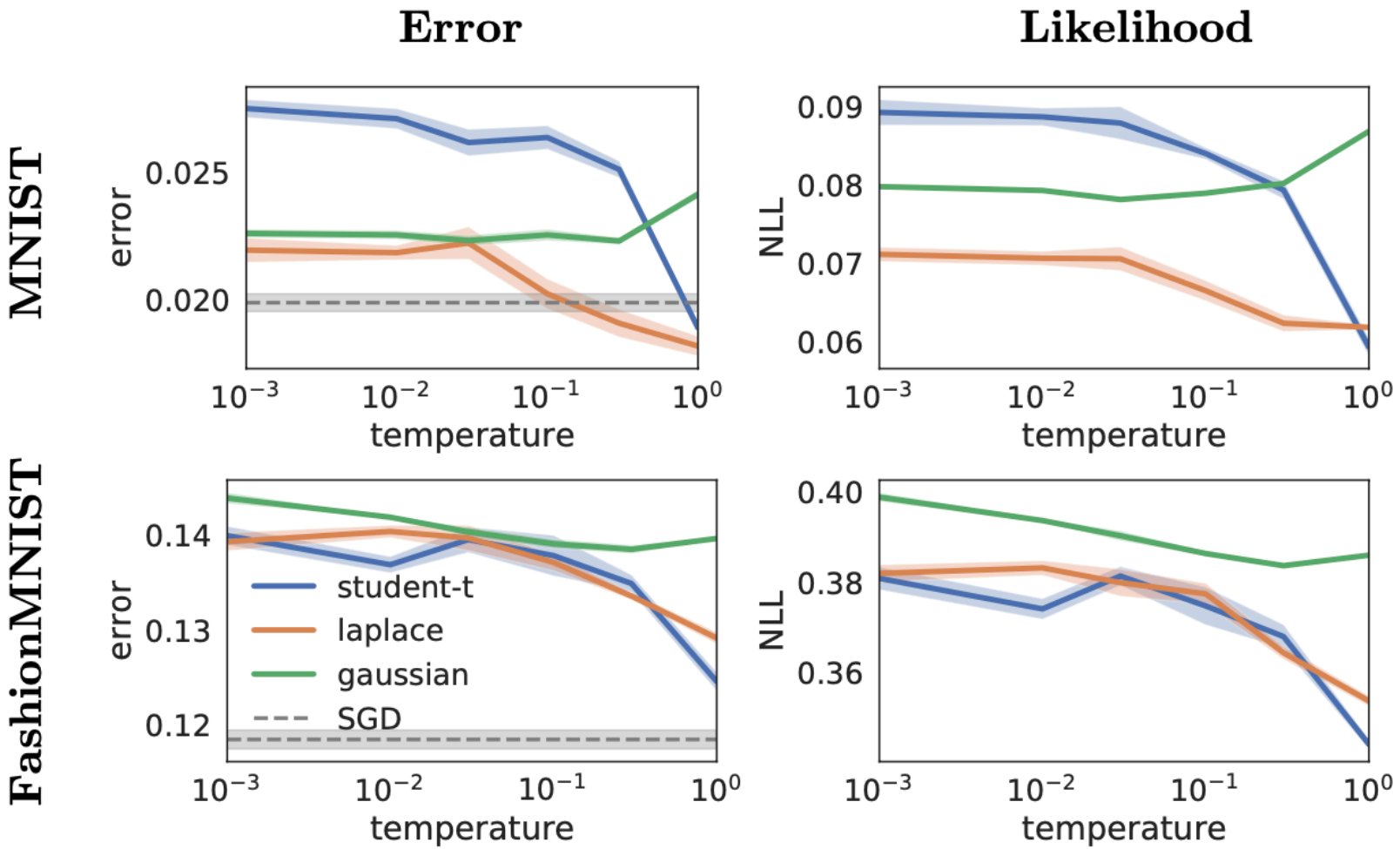
Bayesian FCNNs with different priors

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



Bayesian FCNNs with different priors

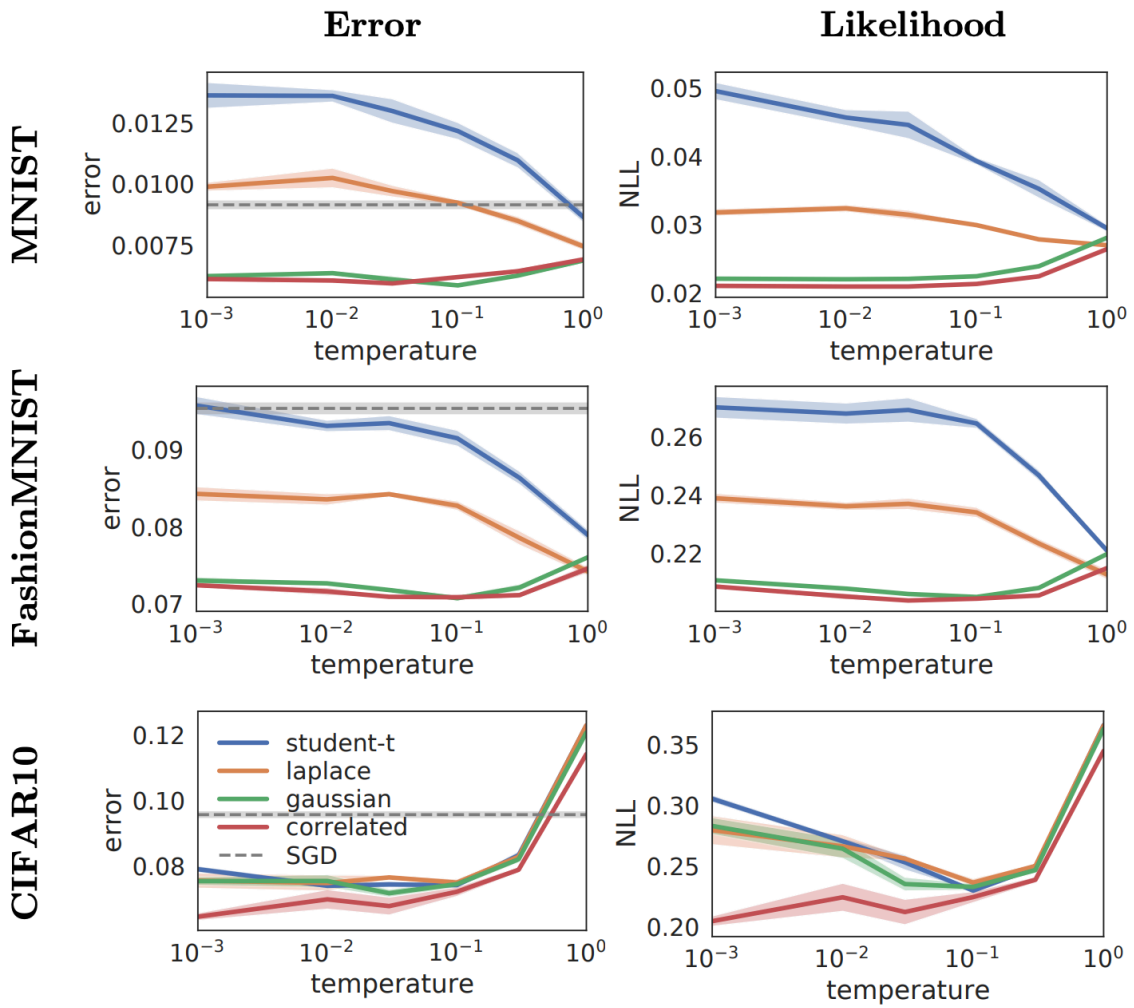
[F, Garriga-Alonso, Ober, Wenzel, Rätsch, Turner, van der Wilk, Aitchison. ICLR 2022]



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

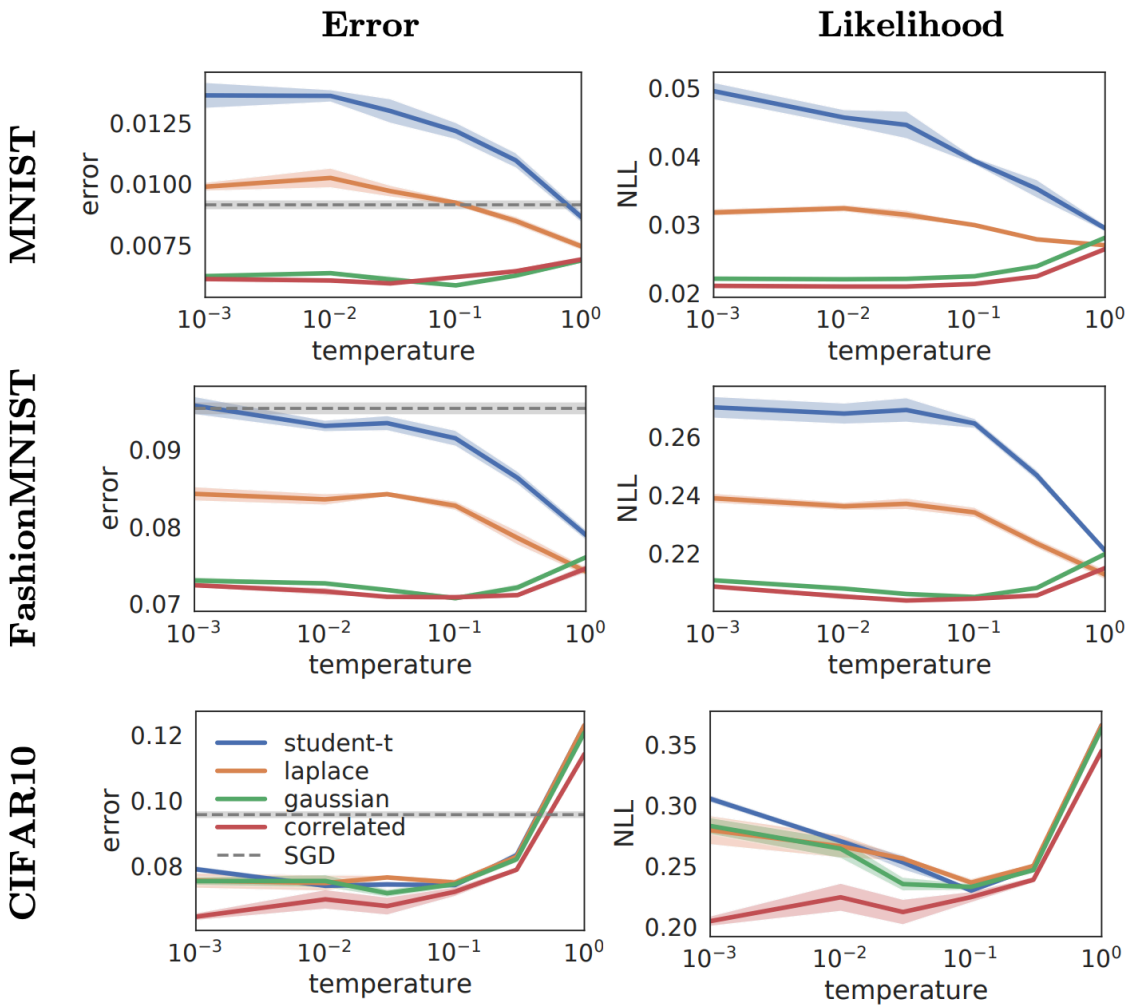
Bayesian CNNs with different priors

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



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

Caveats

Caveats

- Maybe the cold posterior effect is not that bad (temperature might encode a different kind of prior knowledge)

Caveats

- Maybe the cold posterior effect is not that bad (temperature might encode a different kind of prior knowledge)

- FCNNs induce ill-specified function-space priors, but in CNNs and Resnets on images, function-space priors can be better
  - Cold posterior effect is mostly due to data augmentation there

Agenda

- Pathologies of common BNN priors
- How to find better priors using the marginal likelihood
- PAC-Bayesian meta-learning for priors

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

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

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

$$\log p(\mathcal{D}|\mathcal{M}) \approx \log q(\mathcal{D}|\mathcal{M})$$

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

MAP solution

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

$$\log p(\mathcal{D}|\mathcal{M}) \approx \log q(\mathcal{D}|\mathcal{M})$$

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

MAP solution

$$\mathbf{H}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})$$

Hessian

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

$$\log p(\mathcal{D}|\mathcal{M}) \approx \log q(\mathcal{D}|\mathcal{M})$$

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

MAP solution

$$\mathbf{H}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})$$

Hessian

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

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

$$\log p(\mathcal{D}|\mathcal{M}) \approx \log q(\mathcal{D}|\mathcal{M})$$

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

MAP solution

$$\mathbf{H}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})$$

Hessian

Jacobian

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

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

MAP solution

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

$$\mathbf{H}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})$$

Hessian

Jacobian

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

$$\mathbf{P}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{M})$$

Marginal likelihood prior selection

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

$$p(\mathcal{D}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}(\mathbf{x}_n, \boldsymbol{\theta}), \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta}$$

MAP solution

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

$$\mathbf{H}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M})$$

Hessian

Jacobian

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

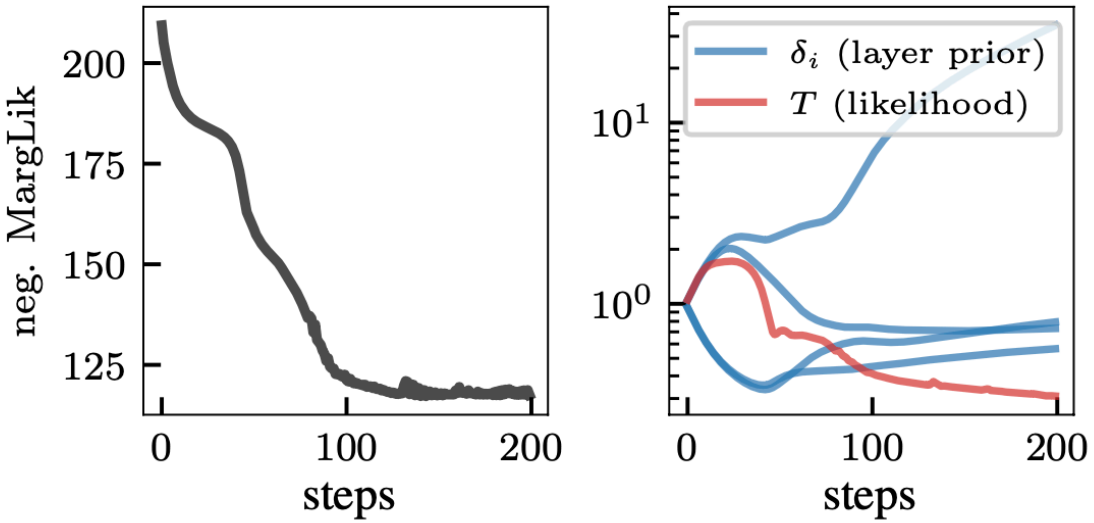
$$\mathbf{P}_{\boldsymbol{\theta}} := -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{M})$$

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

Marginal likelihood prior selection

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

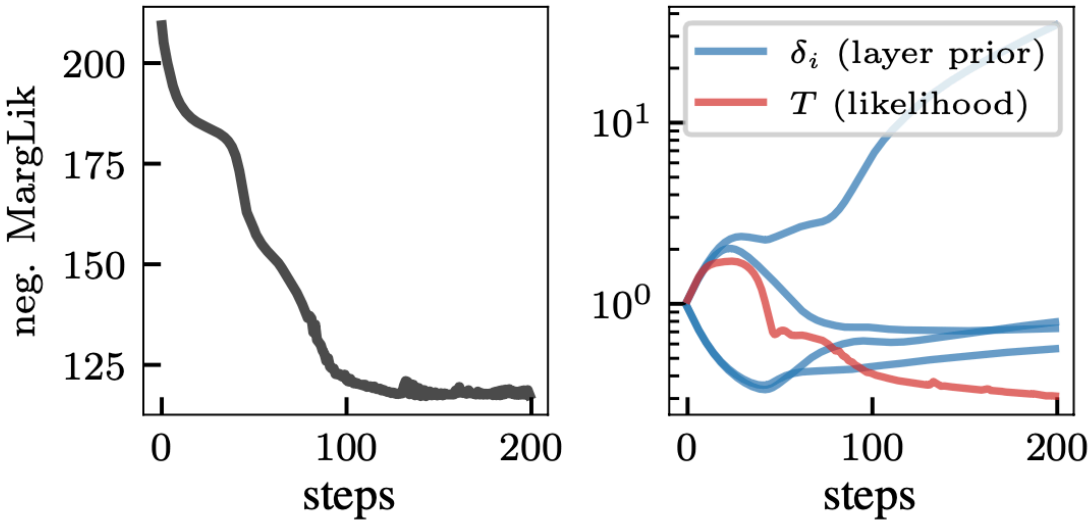
Step 1: Optimize Marginal-Likelihood wrt. hyperparameters



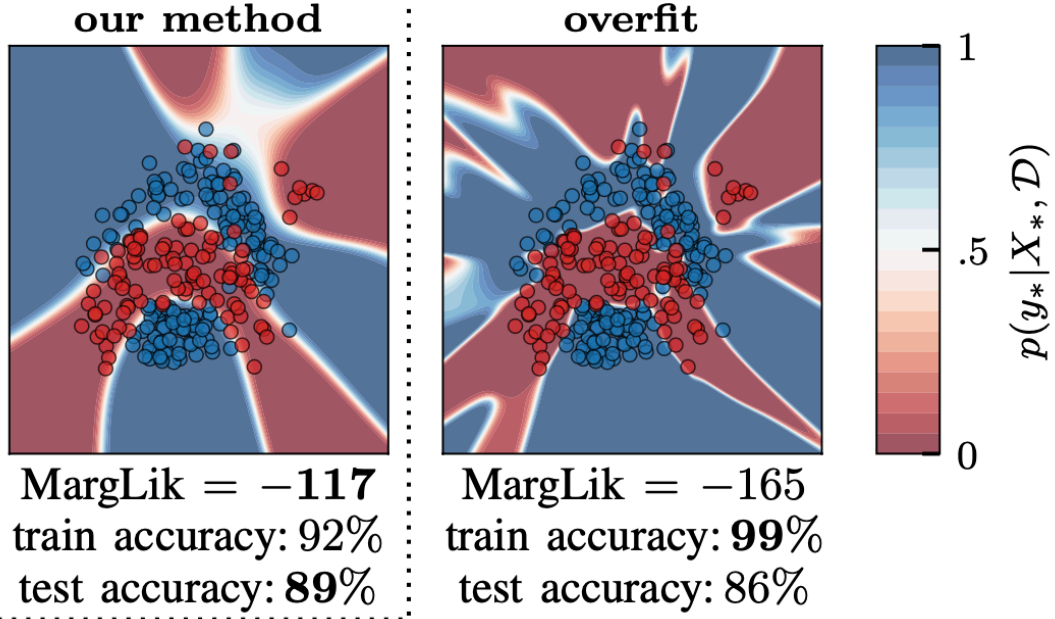
Marginal likelihood prior selection

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

Step 1: Optimize Marginal-Likelihood wrt. hyperparameters



Step 2: Compare marginal likelihood of models



$\text{MargLik} = \mathbf{-117}$
train accuracy: 92%
test accuracy: **89%**

$\text{MargLik} = -165$
train accuracy: **99%**
test accuracy: 86%

ML-II prior improves generalization

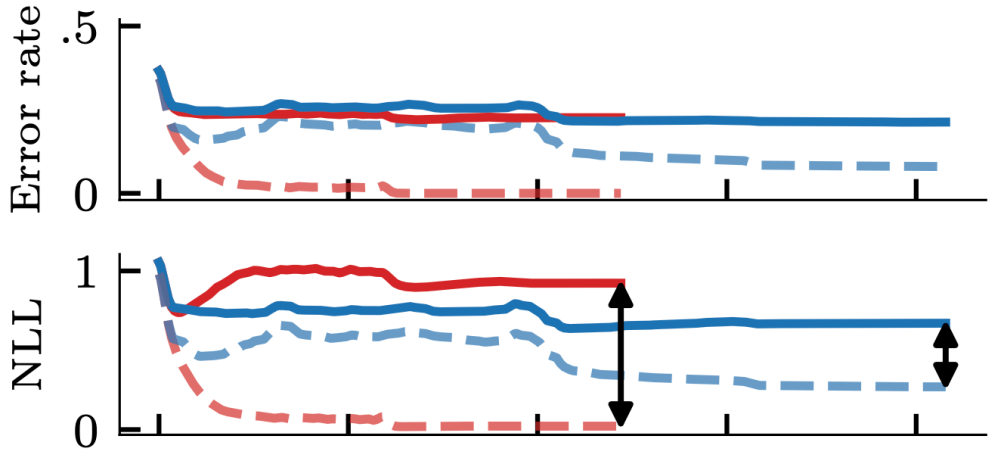
[Immer, Bauer, F, Rätsch, Khan. ICML 2021]

		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

ML-II prior improves generalization

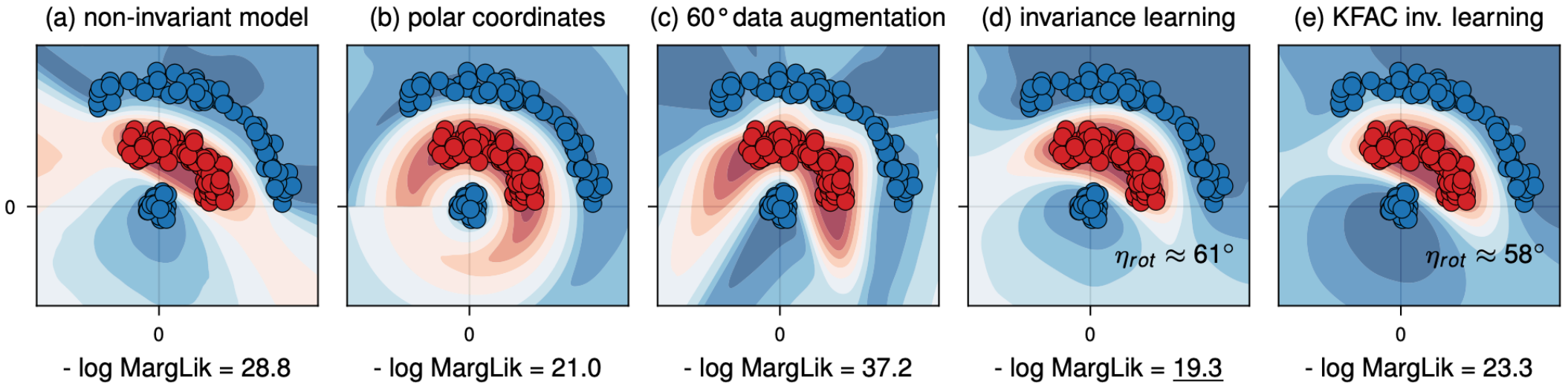
[Immer, Bauer, F, Rätsch, Khan. ICML 2021]

		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



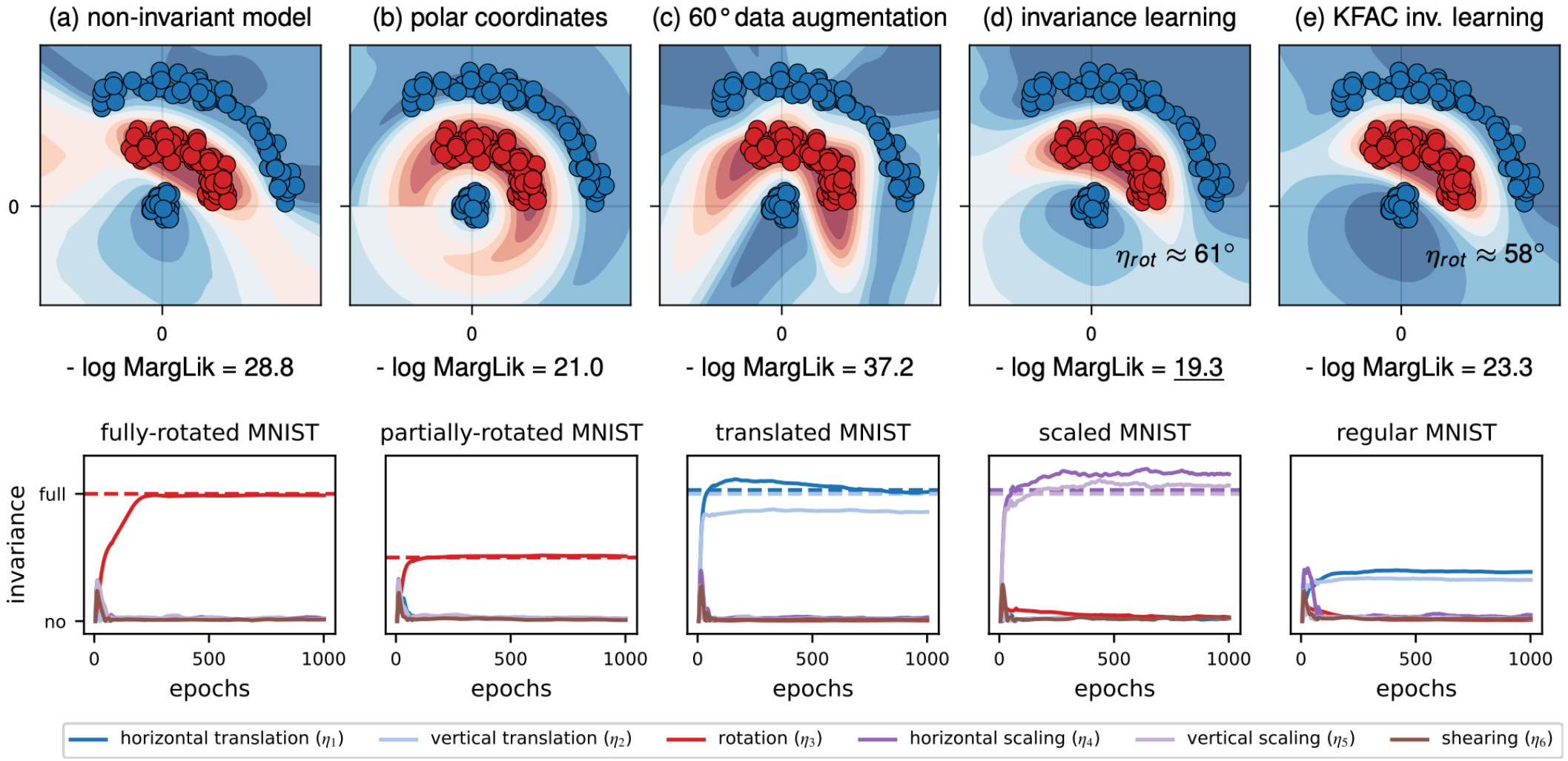
Application: Learning invariances

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



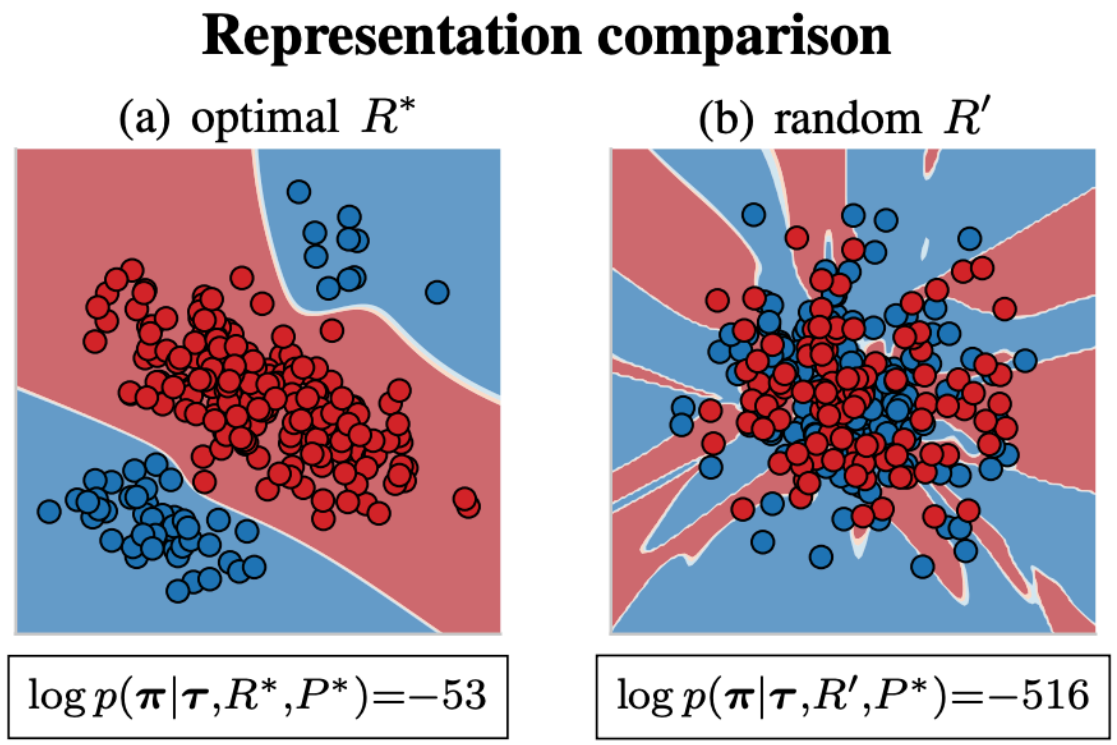
Application: Learning invariances

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



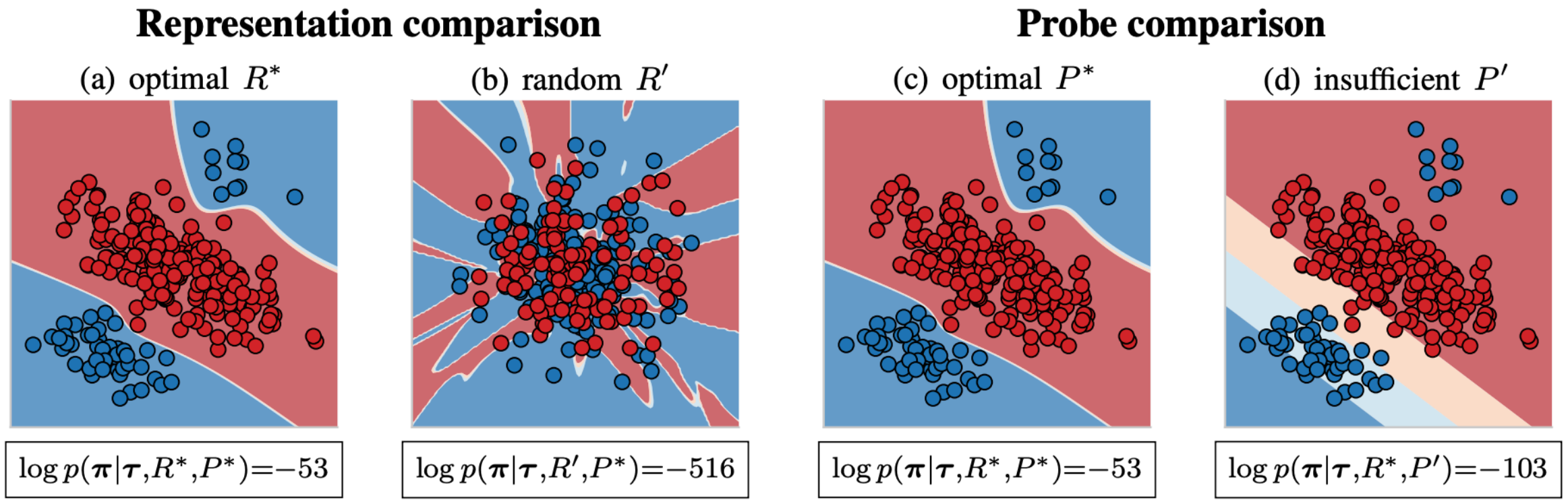
Another application: Linguistic probing

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



Another application: Linguistic probing

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



Caveats

Caveats

- Marginal likelihood measures how well prior explains the data, not how well the posterior generalizes

Caveats

- Marginal likelihood measures how well prior explains the data, not how well the posterior generalizes

- If the prior has a lot of parameters, the marginal likelihood can still overfit

Caveats

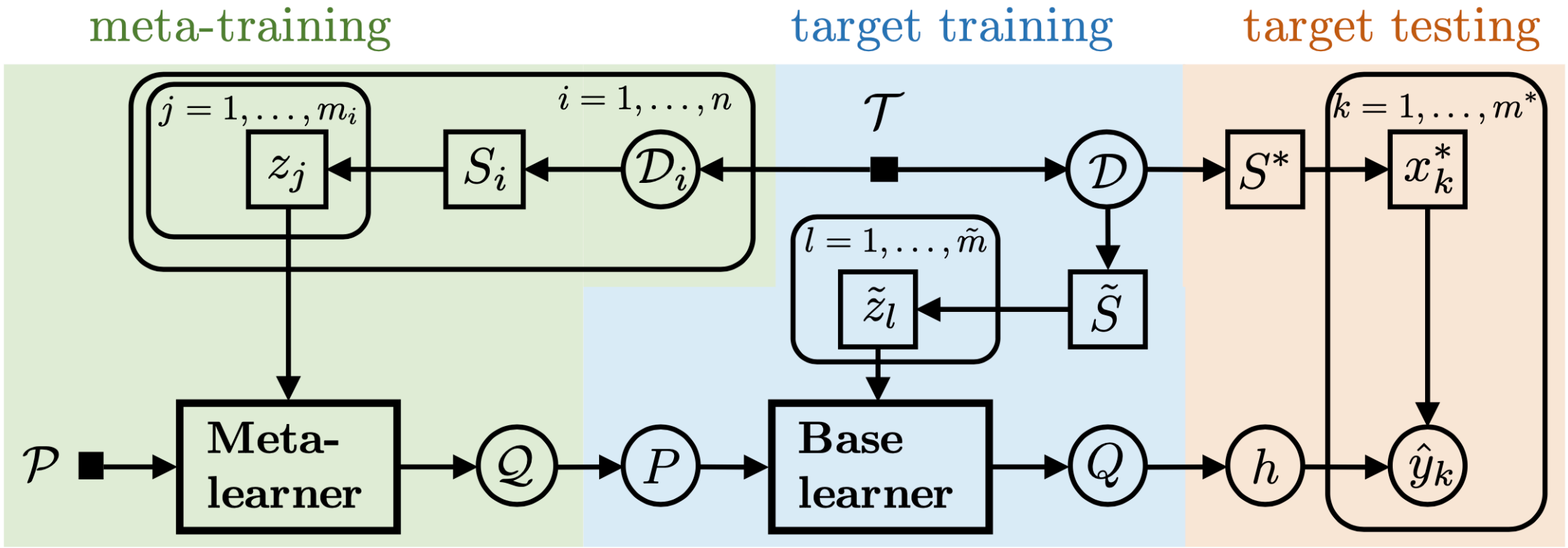
- Marginal likelihood measures how well prior explains the data, not how well the posterior generalizes
- If the prior has a lot of parameters, the marginal likelihood can still overfit
- Laplace-GGN approximation might be more akin to a ”flat minima”-style PAC-Bayes bound than true marginal likelihood

Agenda

- Pathologies of common BNN priors
- How to find better priors using the marginal likelihood
- PAC-Bayesian meta-learning for priors

PAC-Bayesian meta-learning

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



Our meta-learning PAC-bound

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

Theorem 2. *Let $Q : \mathcal{Z}^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$ be a base learner, $\mathcal{P} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ some fixed hyper-prior $\mathcal{P}$ and $\lambda, \beta > 0$. For any confidence level $\delta \in (0, 1]$ the inequality*

$$\mathcal{L}(\mathcal{Q}, \mathcal{T}) \leq \hat{\mathcal{L}}(\mathcal{Q}, S_1, ..., S_n) + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}||\mathcal{P}) + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\beta} \mathbb{E}_{P\sim\mathcal{Q}} \left[D_{KL}(Q(S_i, P)||P)\right] + C(\delta, \lambda, \beta) \tag{4}$$

holds uniformly over all hyper-posteriors $\mathcal{Q} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ with probability $1 - \delta$.

Our meta-learning PAC-bound

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

Theorem 2. *Let $Q : \mathcal{Z}^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$ be a base learner, $\mathcal{P} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ some fixed hyper-prior $\mathcal{P}$ and $\lambda, \beta > 0$. For any confidence level $\delta \in (0, 1]$ the inequality*

$$\mathcal{L}(\mathcal{Q}, \mathcal{T}) \leq \hat{\mathcal{L}}(\mathcal{Q}, S_1, ..., S_n) + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}||\mathcal{P})$$

empirical losses

$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\beta} \mathbb{E}_{P\sim\mathcal{Q}} \left[D_{KL}(Q(S_i, P)||P)\right] \qquad (4)$$

$$+ C(\delta, \lambda, \beta)$$

holds uniformly over all hyper-posteriors $\mathcal{Q} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ with probability $1 - \delta$.

Our meta-learning PAC-bound

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

Theorem 2. *Let $Q : \mathcal{Z}^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$ be a base learner, $\mathcal{P} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ some fixed hyper-prior $\mathcal{P}$ and $\lambda, \beta > 0$. For any confidence level $\delta \in (0, 1]$ the inequality*

$$\mathcal{L}(\mathcal{Q}, \mathcal{T}) \leq \hat{\mathcal{L}}(\mathcal{Q}, S_1, ..., S_n) + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}||\mathcal{P})$$

empirical losses

meta complexity

$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\beta} \mathbb{E}_{P\sim\mathcal{Q}} \left[D_{KL}(Q(S_i, P)||P)\right] \qquad (4)$$

$$+ C(\delta, \lambda, \beta)$$

holds uniformly over all hyper-posteriors $\mathcal{Q} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ with probability $1 - \delta$.

Our meta-learning PAC-bound

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

Theorem 2. *Let $Q : \mathcal{Z}^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$ be a base learner, $\mathcal{P} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ some fixed hyper-prior $\mathcal{P}$ and $\lambda, \beta > 0$. For any confidence level $\delta \in (0, 1]$ the inequality*

$$\mathcal{L}(\mathcal{Q}, \mathcal{T}) \leq \hat{\mathcal{L}}(\mathcal{Q}, S_1, ..., S_n) + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}||\mathcal{P})$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\beta} \mathbb{E}_{P\sim\mathcal{Q}} \left[D_{KL}(Q(S_i, P)||P)\right] \qquad (4)$$

$$+ C(\delta, \lambda, \beta)$$

empirical losses

meta complexity

per-task complexities

holds uniformly over all hyper-posteriors $\mathcal{Q} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$ with probability $1 - \delta$.

PAC-bound for Bayesian base learners

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

Corollary 1. *When choosing a Gibbs posterior* $Q^*(S_i, P) := P(h)\exp(-\beta\hat{\mathcal{L}}(S_i, h))/Z_\beta(S_i, P)$ *as a base learner, under the same assumptions as in Theorem 2, we have*

$$\mathcal{L}(\mathcal{Q}, \mathcal{T}) \leq -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\beta}\mathbb{E}_{P\sim\mathcal{Q}}\left[\ln Z_\beta(S_i, P)\right] + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}||\mathcal{P}) + C(\delta, \lambda, \beta)\ . \tag{5}$$

with probability at least $1 - \delta$.

PAC-bound for Bayesian base learners

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

Corollary 1. *When choosing a Gibbs posterior* $Q^*(S_i, P) := P(h)\exp(-\beta\hat{\mathcal{L}}(S_i, h))/Z_\beta(S_i, P)$ *as a base learner, under the same assumptions as in Theorem 2, we have*

$$\mathcal{L}(\mathcal{Q}, \mathcal{T}) \leq -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\beta}\mathbb{E}_{P\sim\mathcal{Q}}\left[\ln Z_\beta(S_i, P)\right] + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}||\mathcal{P}) + C(\delta, \lambda, \beta)\ . \tag{5}$$

marginal likelihoods

with probability at least $1 - \delta$.

PAC-Bayesian meta-learning

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

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

PAC-Bayesian meta-learning

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

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

hyperposterior

PAC-Bayesian meta-learning

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

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

hyperposterior

hyperprior

PAC-Bayesian meta-learning

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

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

hyperposterior

hyperprior

marginal likelihood

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)	$\mathbf{0.885 \pm 0.090}$	$\mathbf{0.091 \pm 0.010}$

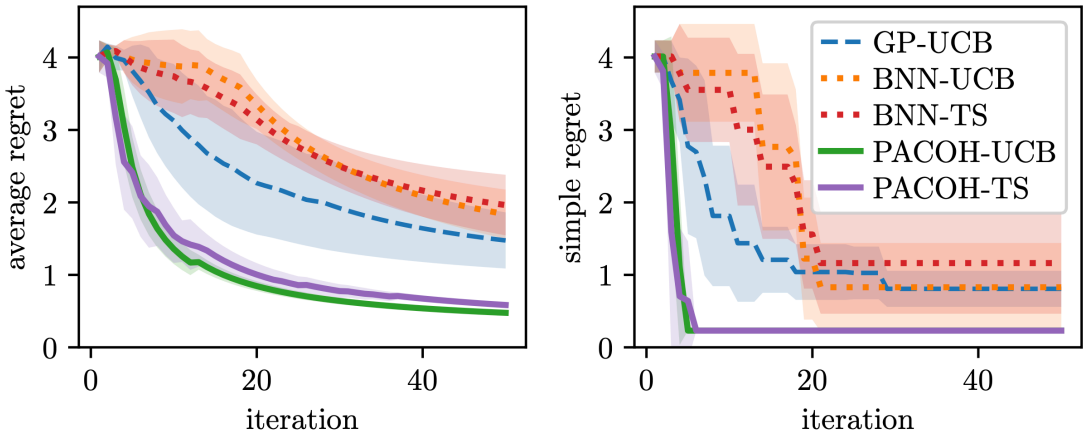
Few-shot learning on Omniglot

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



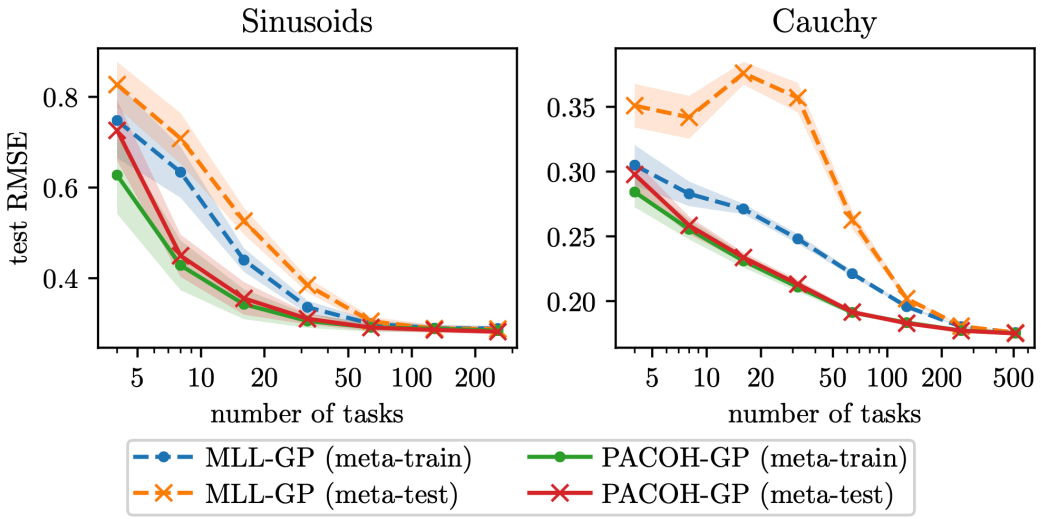
Bandit task

PAC-Bayesian meta-learning

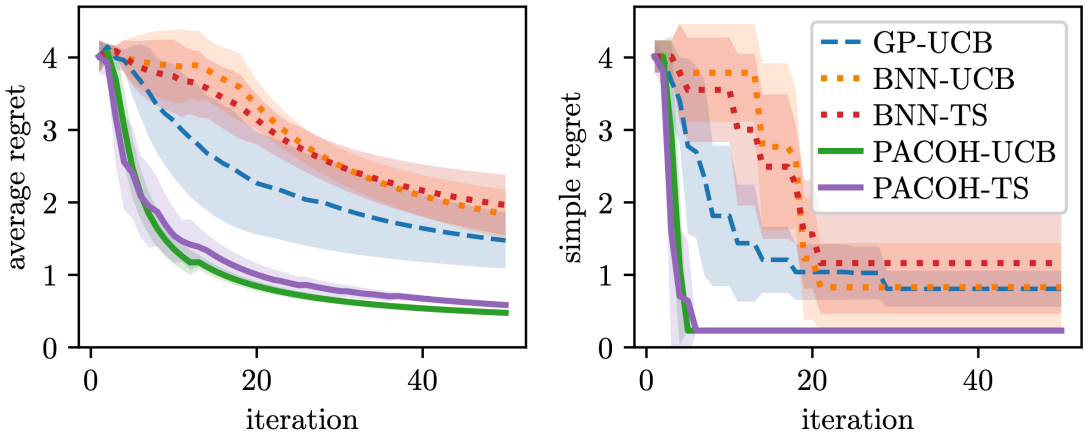
[Rothfuss, F, Josifoski, Krause. ICML 2021]

	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



Meta-overfitting



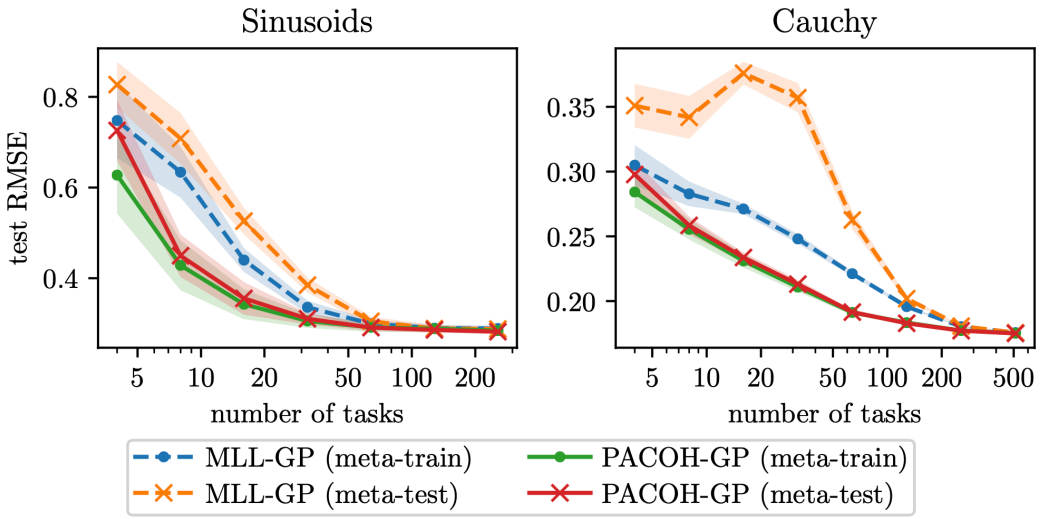
Bandit task

PAC-Bayesian meta-learning

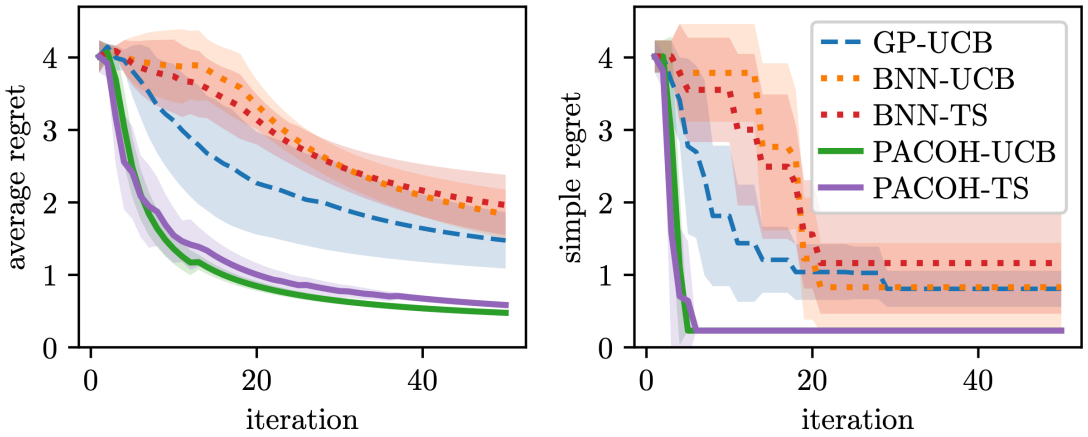
[Rothfuss, F, Josifoski, Krause. ICML 2021]

	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)	$\mathbf{0.885 \pm 0.090}$	$\mathbf{0.091 \pm 0.010}$

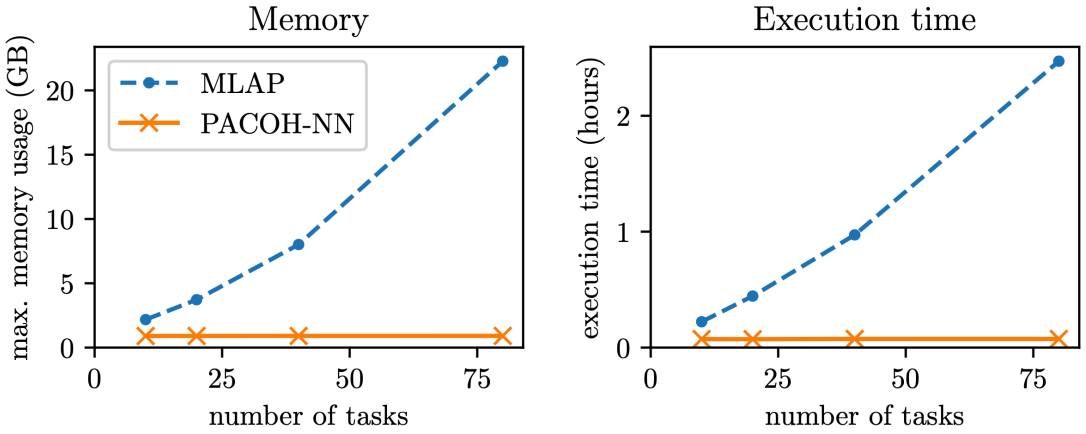
Few-shot learning on Omniglot



Meta-overfitting



Bandit task



Runtimes

Caveats

Caveats

- The bounds themselves are still very loose (and possibly vacuous), so optimizing them is not guaranteed to help

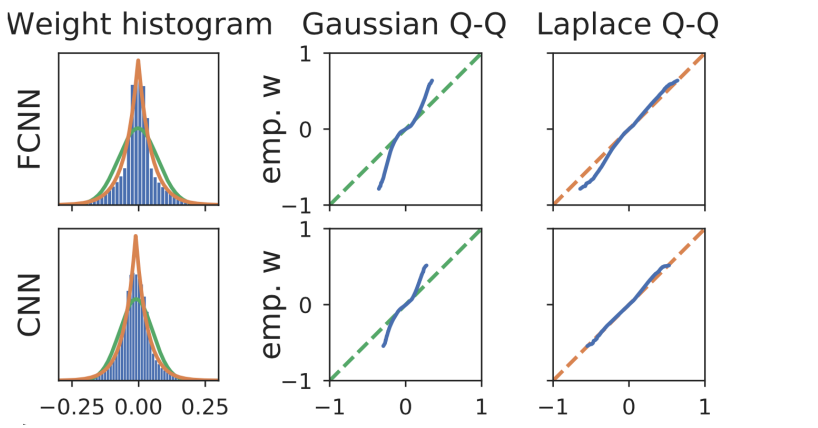
Caveats

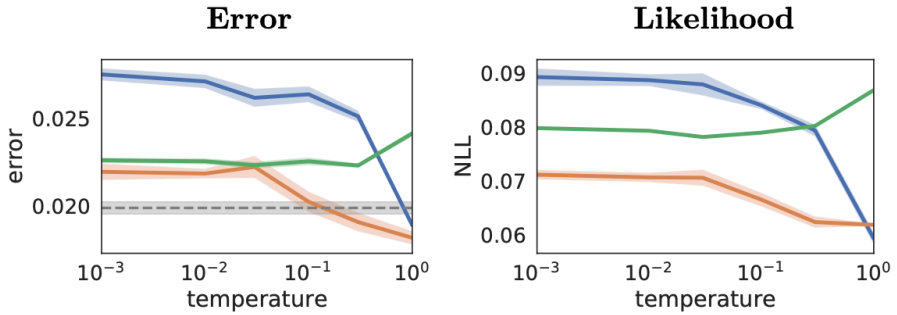
- The bounds themselves are still very loose (and possibly vacuous), so optimizing them is not guaranteed to help

- Choosing λ and β trades off asymptotic consistency with tightness on low data

Take-home messages

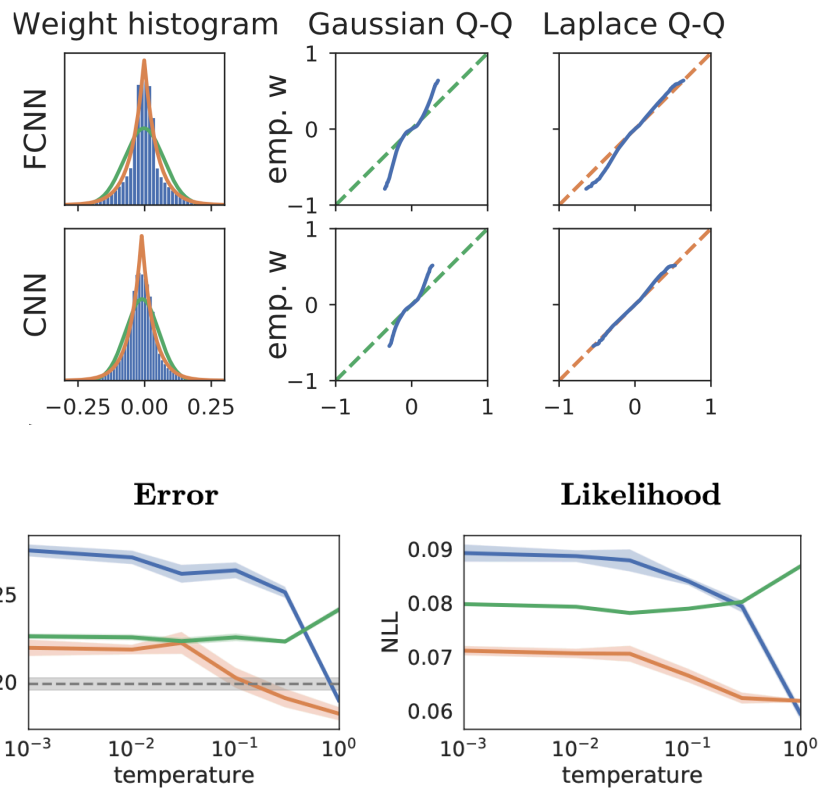
Take-home messages



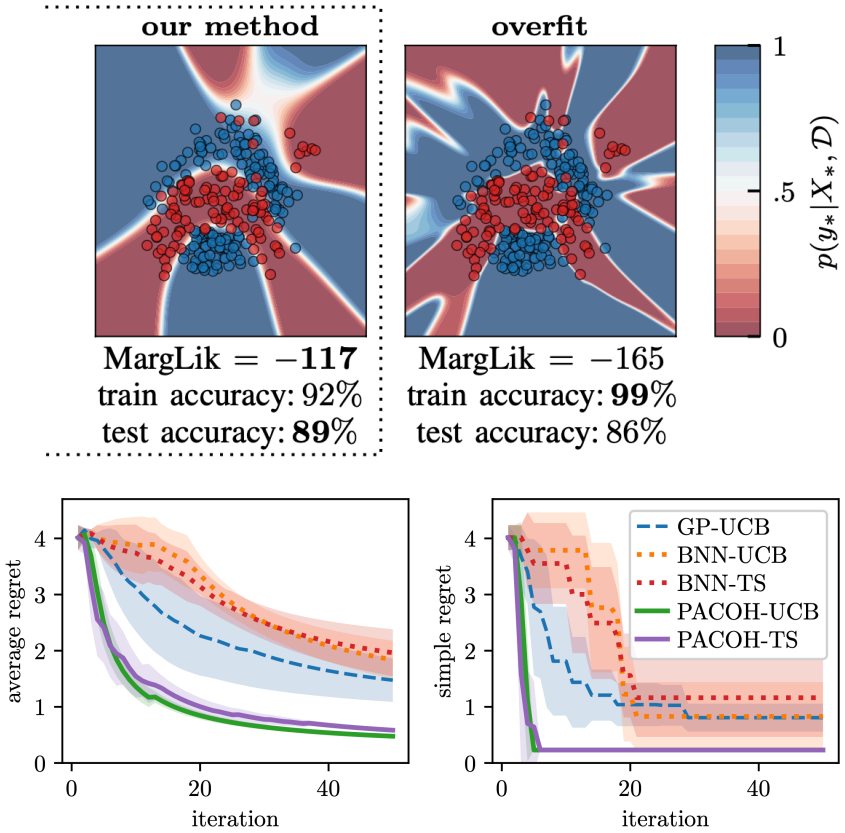


Common Bayesian neural network priors yield sub-optimal performance

Take-home messages



Common Bayesian neural network priors yield sub-optimal performance



We can find better priors using ML-II or (PAC-Bayesian) meta-learning

Further reading

International Statistical Review / Early View

Original Article | Open Access

Priors in Bayesian Deep Learning: A Review

Vincent Fortuin

First published: 11 May 2022

https://doi.org/10.1111/insr.12502

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

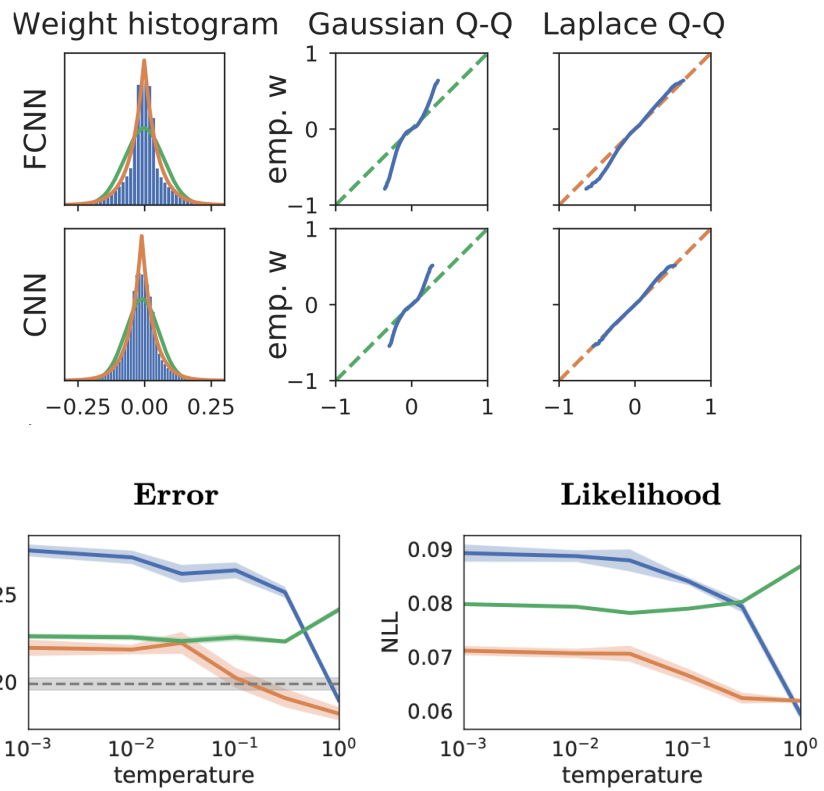
James Allingham
Adrià Garriga-Alonso
Sebastian Ober
Richard Turner

fortuin.github.io

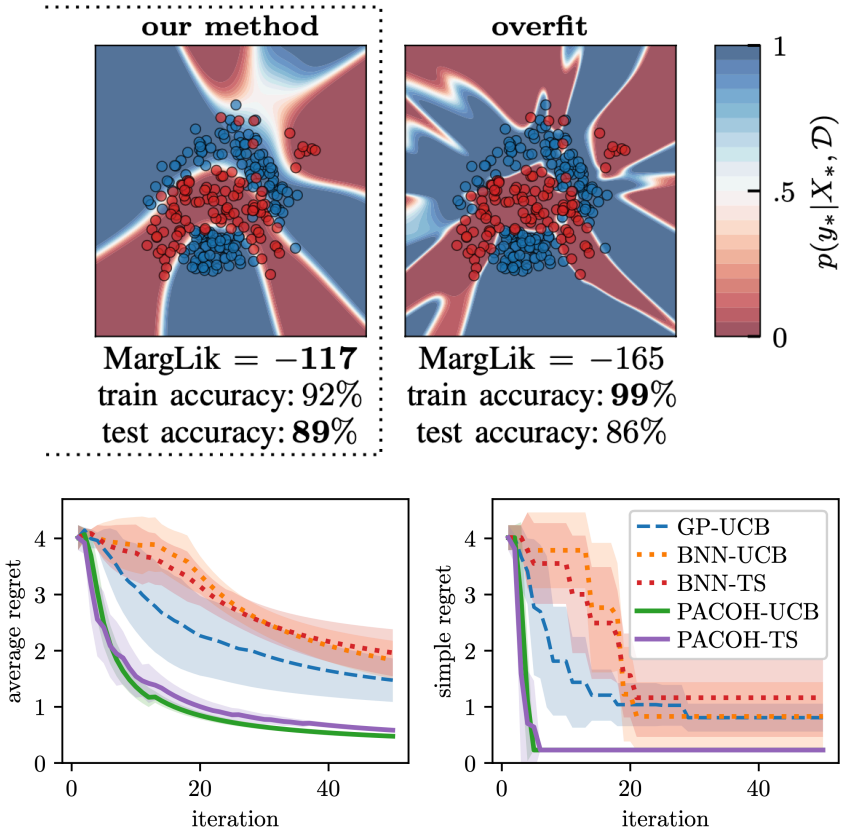
vbf21@cam.ac.uk

@vincefort

Take-home messages



Common Bayesian neural network priors yield sub-optimal performance



We can find better priors using ML-II or (PAC-Bayesian) meta-learning