Scalable Marginal Likelihood Estimation for Model Selection in Deep learning

ICML 2021

Alexander Immer
ETH Zurich

Matthias Bauer
Deepmind

Vincent Fortuin
ETH Zurich

Gunnar Rätsch
ETH Zurich

Emtiyaz Khan
RIKEN AIP

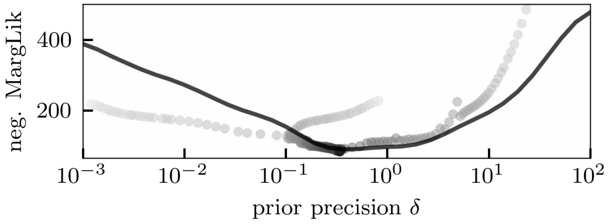
Model Selection in Deep Learning

(1) hyperparameters (regularization) and **(2)** model architecture (ResNet vs CNN).

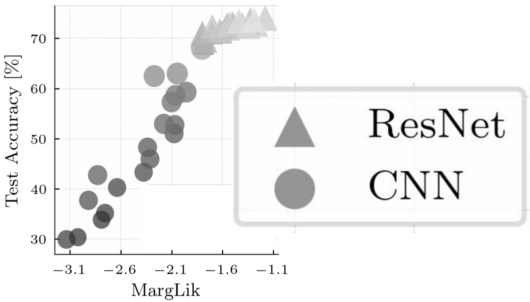
But **validation data** might be **unavailable** (e.g. in continual learning).

We show the training marginal likelihood is viable for model selection in DL!

(1) Differentiable hyperparameters *during training*



(2) Architecture Selection *after training*



Marginal Likelihood Estimation for Deep Learning

1 **Laplace approximation** [1] to the log marginal likelihood

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

$$\log p(\mathcal{D}|\mathcal{M}) \approx \underbrace{\log p(\mathcal{D}|\theta_*, \mathcal{M})}_{\text{Training data fit}} + \underbrace{\log p(\theta_*|\mathcal{M}) - \frac{1}{2}\log\left|\frac{1}{2\pi}\mathbf{H}_{\theta_*}\right|}_{\text{Complexity penalty}}$$

2 **Scalable approximations** to the Hessian

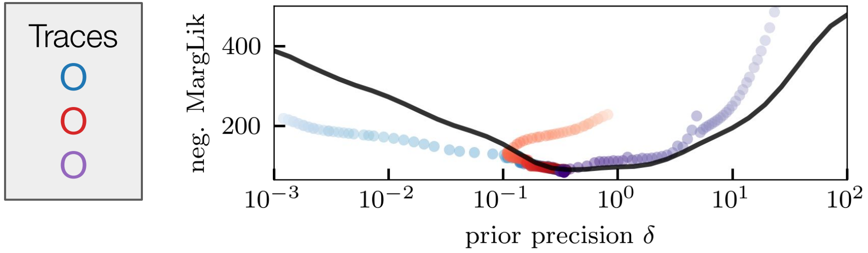
$$\mathbf{H}_{\theta_*} \approx$$

Approx. types		Correlation captured
Gauss-Newton	$\times$	Full
Fisher Information		KFAC (block-diagonal) [3, 4]
Empirical Fisher		Diagonal

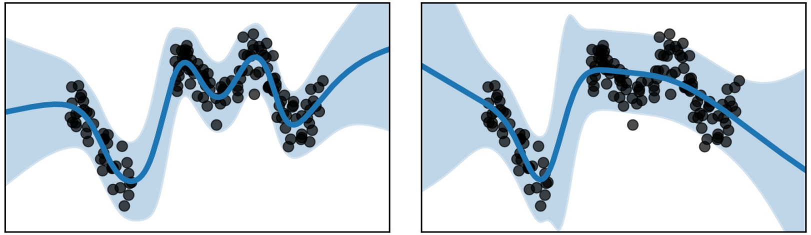
[1] MacKay. "A practical Bayesian framework for backpropagation networks." Neural computation (1992).

Maximizing the Marginal Likelihood during Training

Optimize Hyperparameters during Training
(e.g. regularization)



Compare Architectures
(e.g. #layers)



3 layers, 5221 params
MargLik = -88

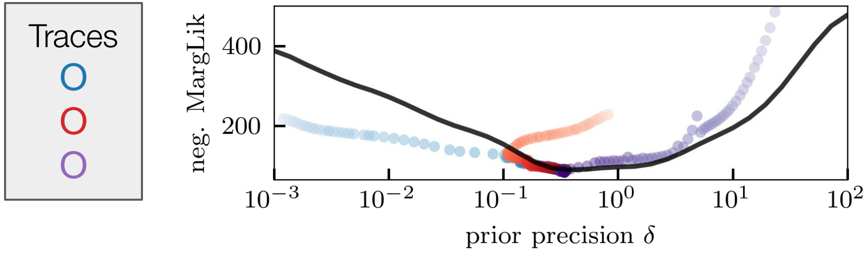
1 layer, 151 params
MargLik = -110

Every epoch:

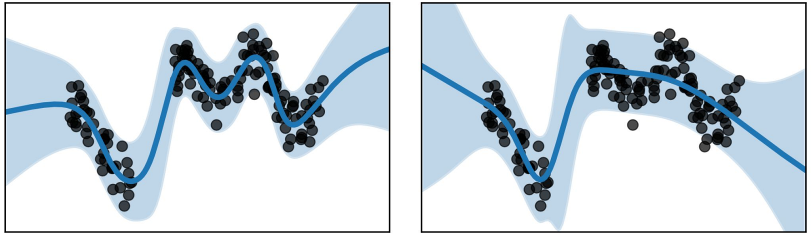
Update network parameters (e.g. Adam)
Differentiate MargLik wrt. hyperparameters
Update differentiable hyperparameters

Maximizing the Marginal Likelihood during Training

Optimize Hyperparameters during Training (e.g. regularization)



Compare Architectures (e.g. #layers)



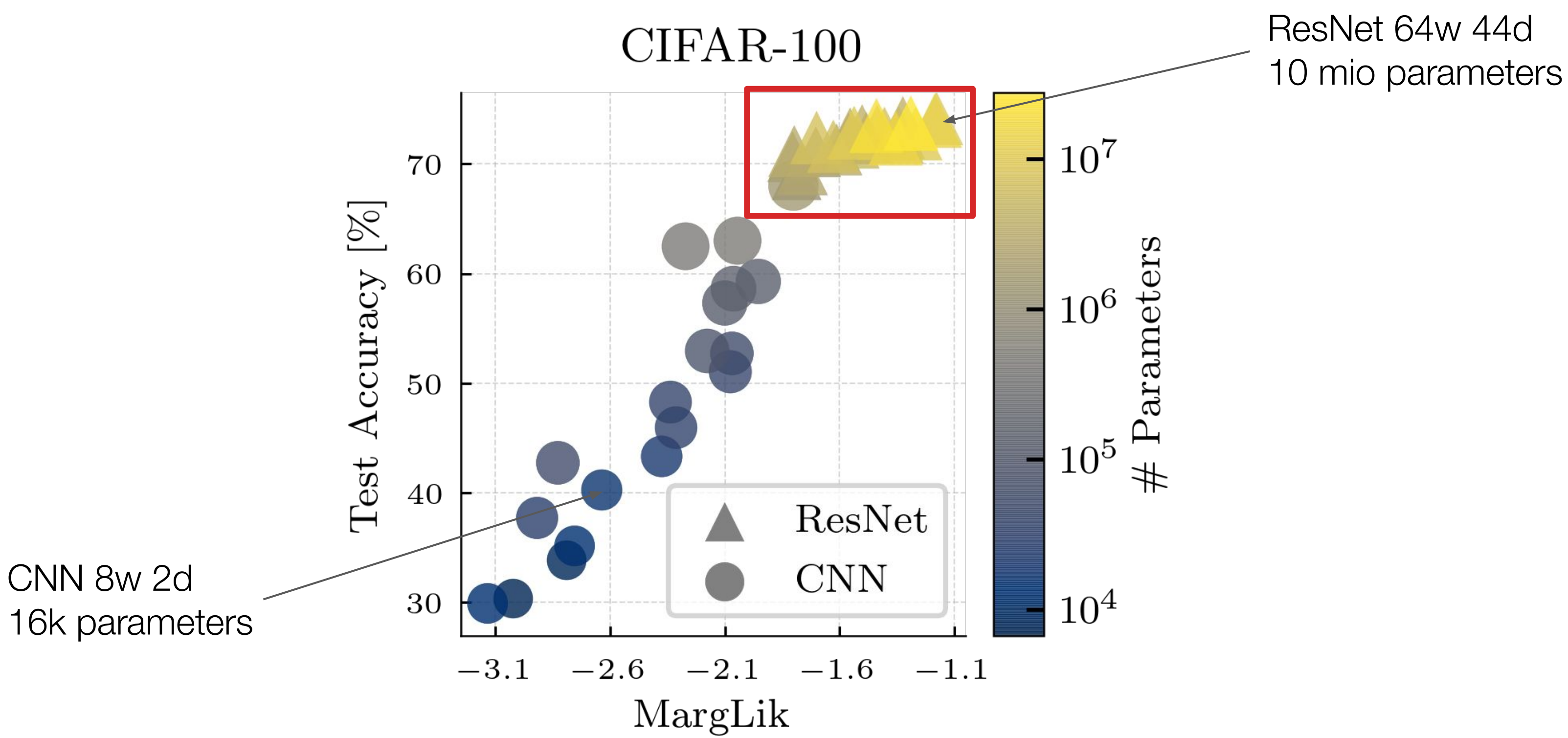
3 layers, 5221 params
MargLik = -88

1 layer, 151 params
MargLik = -110

- On par or better than cross-validation
  - UCI regression/classification, image classification
- Several hundred hyperparameters at once
  - No overhead for some approximations

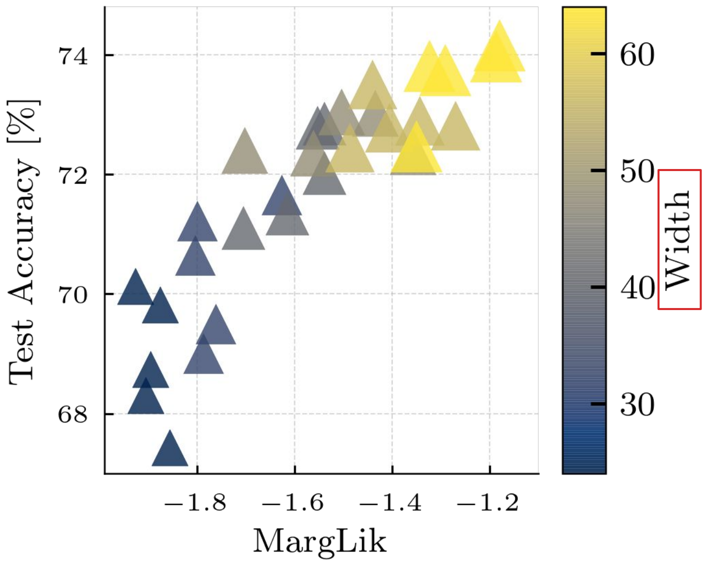
Marginal Likelihood for Architecture Comparison

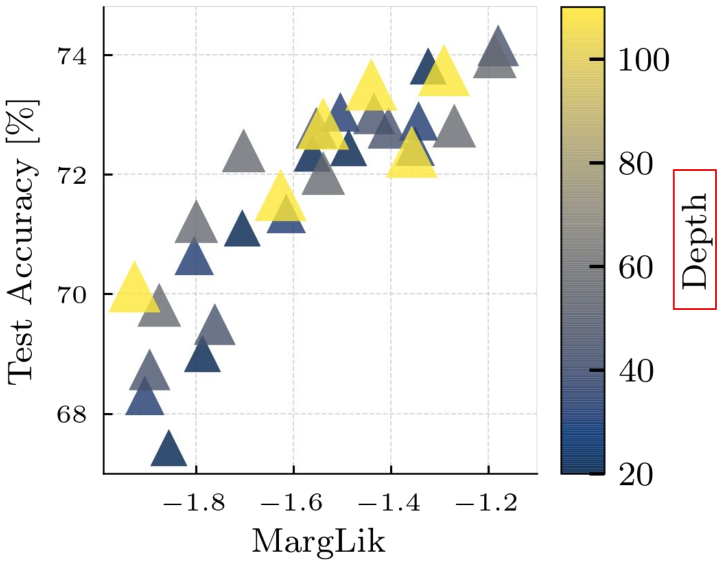
Two architectures (**CNN, ResNet**) + varying **width** (≤ 64) and **depth** (≤ 110)



Marginal Likelihood for Architecture Comparison

ResNets of varying **width** (≤ 64) and **depth** (≤ 110)





→ In line with proposed Wide ResNet architecture [5]

Summary

- Marginal likelihood viable for model selection in DL **without validation data**
- Optimize margLik: **hundreds of hyperparameters during training**
- **Model comparison across architectures** seems possible

References (abbreviated)

[1] MacKay: *"A practical Bayesian framework for backpropagation networks"*

[2] Rasmussen, et al.: "*Occam's razor*"

[3] Martens, et al.: "*Optimizing neural networks with KFAC*"

[4] Ritter, et al.: "A Scalable Laplace Approximation for Neural Networks"

[5] Zagoruyko, et al.: "Wide Residual Networks"