NYU

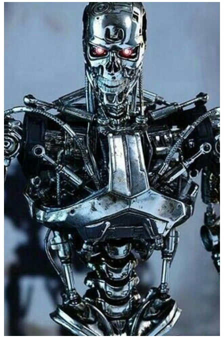
Bayesian Model Selection, the Marginal Likelihood, and Generalization

Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, Andrew Gordon Wilson

ICML 2022

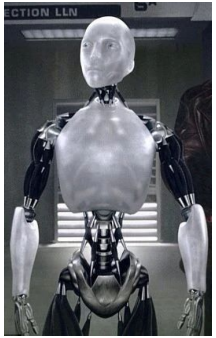
How do we perform model selection?

How do we select between scientific hypotheses or trained models that are entirely consistent with observations?



Model 1

Model 2



Model 3

Model 4

The marginal likelihood or the evidence

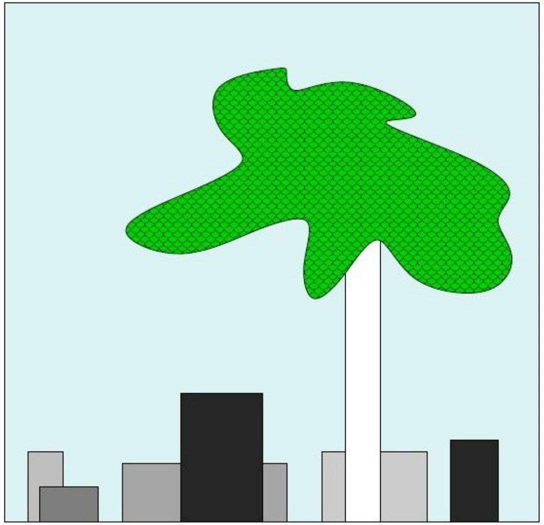
The probability (density) that we would generate a dataset D with a model $\mathcal{M}$ if we randomly sample from a prior over its parameters,

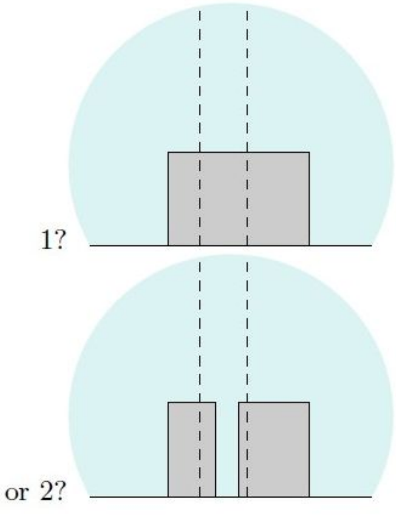
$$p(D\,|\mathcal{M}) = \int p(D|w,\mathcal{M})\; p(w|\mathcal{M})\; dw$$

We usually use the log-marginal likelihood (LML) $\log p(D\,|\mathcal{M})$.

Occam's razor

We should accept the ***simplest*** *explanation that fits the data*

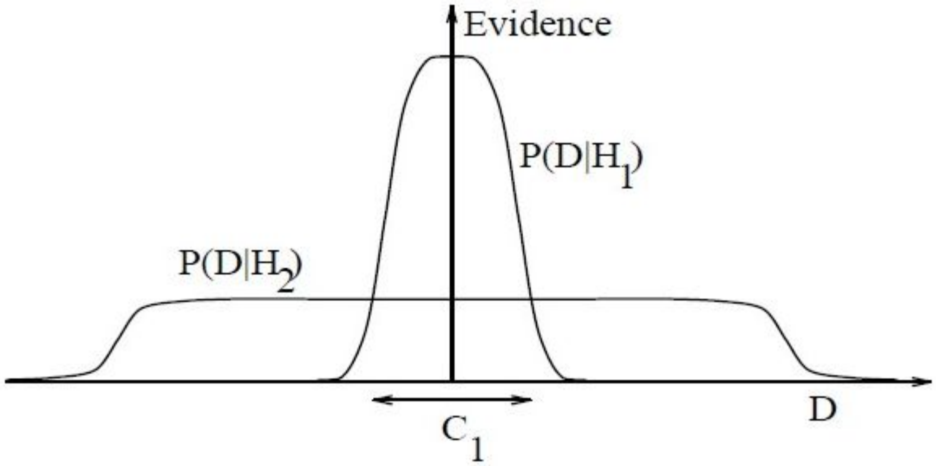




Figures from Information Theory, Inference, and Learning Algorithms, page 343, David J. C. MacKay

The marginal likelihood encodes Occam's razor

The most constrained model which can fit the data wins, encapsulating "Occam's razor".



Figures from Information Theory, Inference, and Learning Algorithms, page 344, David J. C. MacKay

Contributions

- Distinguish between selecting over prior scientific hypotheses, for which the marginal likelihood is reasonable, and predicting the generalization of models after training.

Contributions

- Distinguish between selecting over prior scientific hypotheses, for which the marginal likelihood is reasonable, and predicting the generalization of models after training.

- Describe conceptual and practical issues in using the marginal likelihood for selecting between trained models and hyperparameter learning, including a variety of mechanisms for over- and under-fitting, and approximate inference.

Contributions

- Distinguish between selecting over prior scientific hypotheses, for which the marginal likelihood is reasonable, and predicting the generalization of models after training.

- Describe conceptual and practical issues in using the marginal likelihood for selecting between trained models and hyperparameter learning, including a variety of mechanisms for over- and under-fitting, and approximate inference.

- Demonstrate that the marginal likelihood can be *negatively* correlated with the generalization of trained neural network architectures.

Contributions

- Distinguish between selecting over prior scientific hypotheses, for which the marginal likelihood is reasonable, and predicting the generalization of models after training.

- Describe conceptual and practical issues in using the marginal likelihood for selecting between trained models and hyperparameter learning, including a variety of mechanisms for over- and under-fitting, and approximate inference.

- Demonstrate that the marginal likelihood can be *negatively* correlated with the generalization of trained neural network architectures.

- Demonstrate that a conditional marginal likelihood is more aligned with generalization and more practical for large-scale hyperparameter learning.

Background

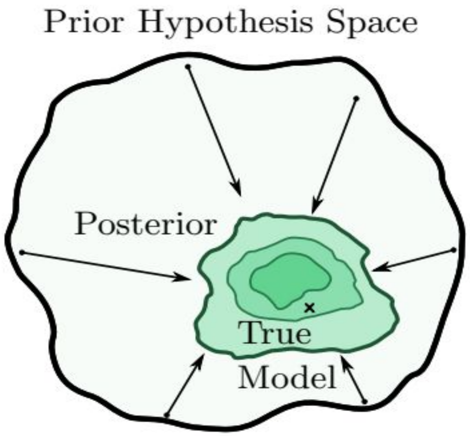
Bayesian learning

- Posterior weighted Bayesian model average (BMA):

$$p_{BMA}(y|x,D) = \int p(y|x,w)\, p(w|D)\, dw$$

Posterior → $p(w|D)$ ∝ Likelihood → $p(D|w)$ × Prior → $p(w)$

$$p(w|D) \propto p(D|w) \times p(w)$$



Bayesian learning

- Posterior weighted Bayesian model average (BMA):

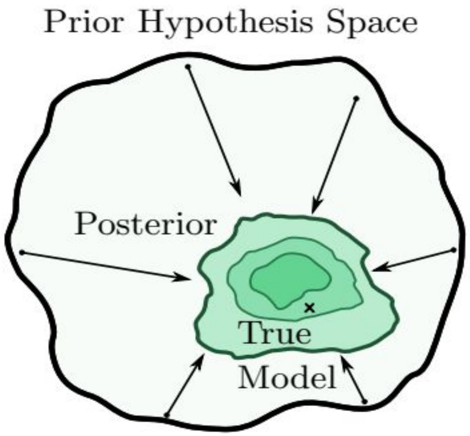
$$p_{BMA}(y|x, D) = \int p(y|x, w)\, p(w|D)\, dw$$

- Standard training:

$$w_{MAP} = argmax_w \log p(w|D)$$
$$= argmax_w [\log p(D|w) + \log p(w)]$$

Posterior Likelihood Prior

$$p(w|D) \propto p(D|w) \times p(w)$$

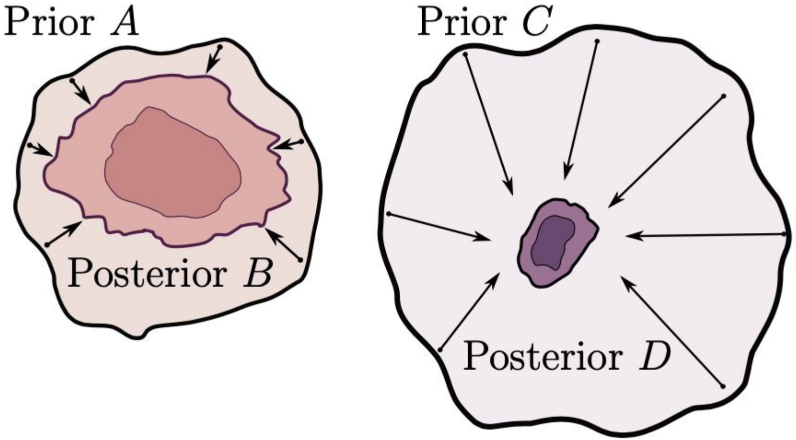


The pitfalls of the marginal likelihood

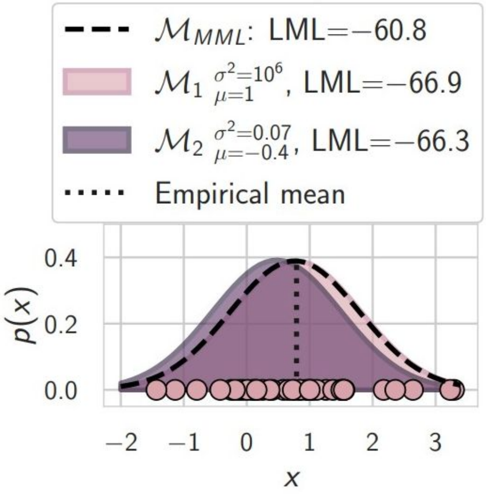
The marginal likelihood penalizes diffuse priors

Density estimation example:

$$x \sim \mathcal{N}(u, 1), u \sim \mathcal{N}(\mu, \sigma^2)$$



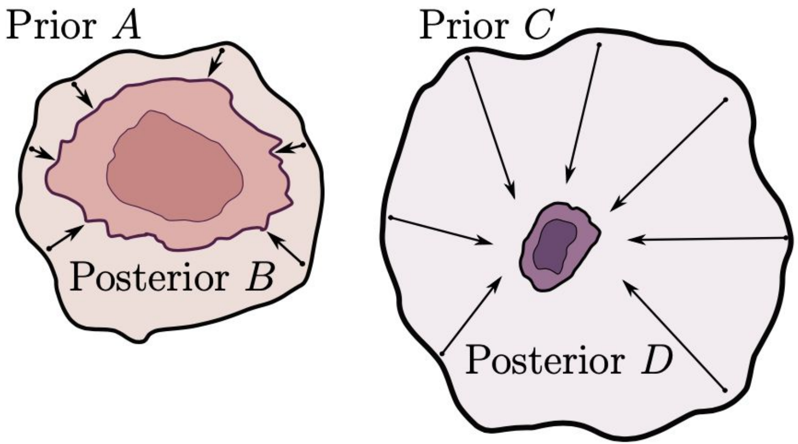
$$p(w|D) \propto p(D|w) \times p(w)$$



The marginal likelihood penalizes diffuse priors

Density estimation example:

$$x \sim \mathcal{N}(u, 1), u \sim \mathcal{N}(\mu, \sigma^2)$$



$$p(w|D) \propto p(D|w) \times p(w)$$

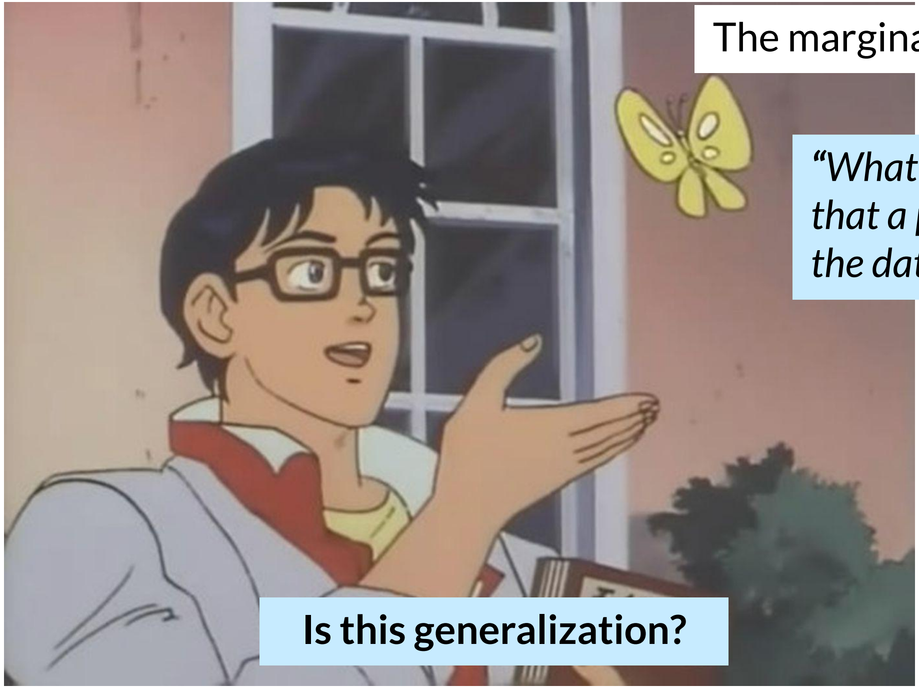
- Prior A → Posterior B
- Prior C → Posterior D

- Prior A > Prior C
- Posterior D > Posterior B

The marginal likelihood is NOT generalization

The generalization question:

“How likely is the posterior, conditioned on the training data, to have generated withheld points drawn from the same distribution?”



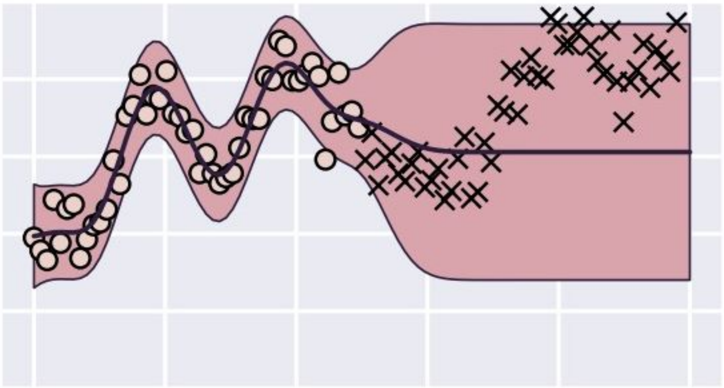
“What is the probability that a prior generated the data?”

The marginal likelihood can overfit - GPs

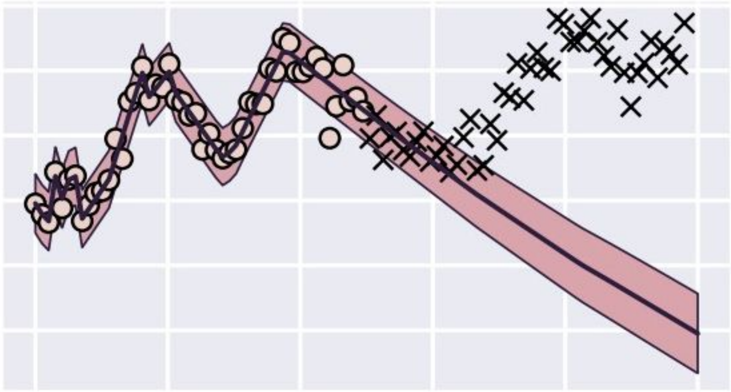
$$f(X) \sim \mathcal{N}\left(m(X), k(X,X)\right);\; k(x,x') = \exp\left(-\frac{1}{2l^2}\left|\left|x - x'\right|\right|^2\right)$$

Given enough flexibility with the prior mean of a Gaussian process, the marginal likelihood **overfits** the data, providing **poor overconfident** predictions outside of the train region.

$m(X) = \mu$

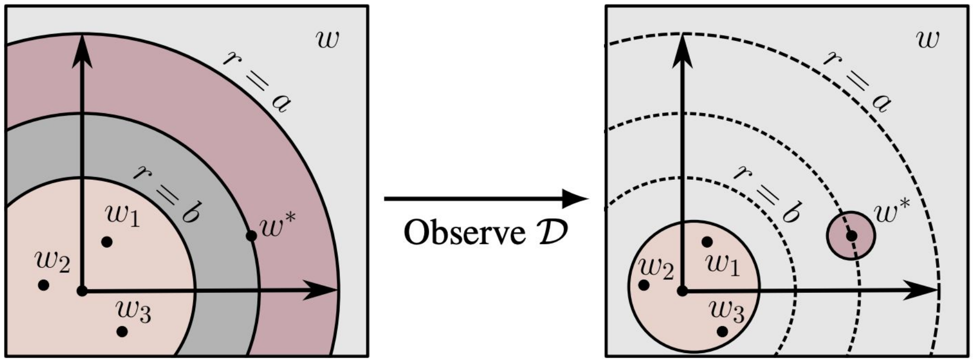


$m(X) = MLP$



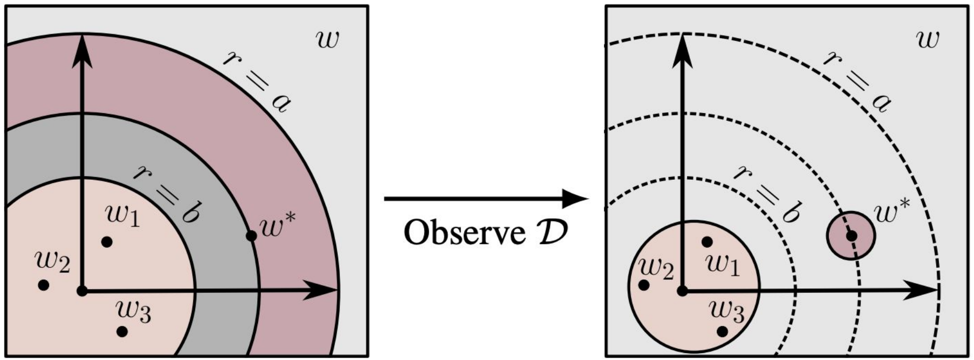
The marginal likelihood can also underfit

The LML will not support optimal solutions if it requires supporting other solutions that do not provide a good fit to the data, leading to underfitting.

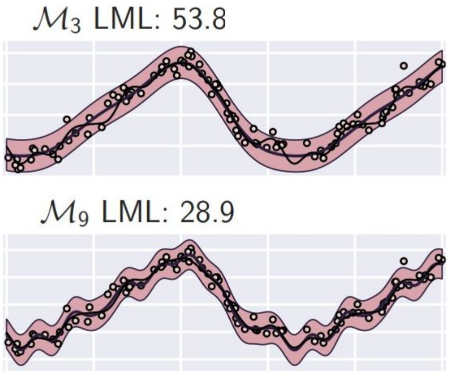


The marginal likelihood can also underfit

The LML will not support optimal solutions if it requires supporting other solutions that do not provide a good fit to the data, leading to underfitting.



An example of the LML favoring an overly simple model:



Decomposition of the marginal likelihood

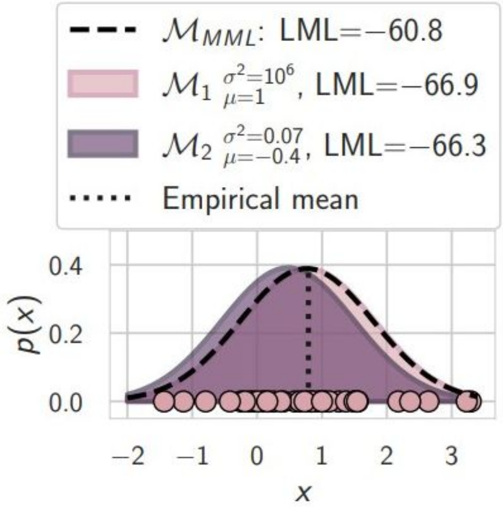
The log marginal likelihood (LML) can be decomposed as follows:

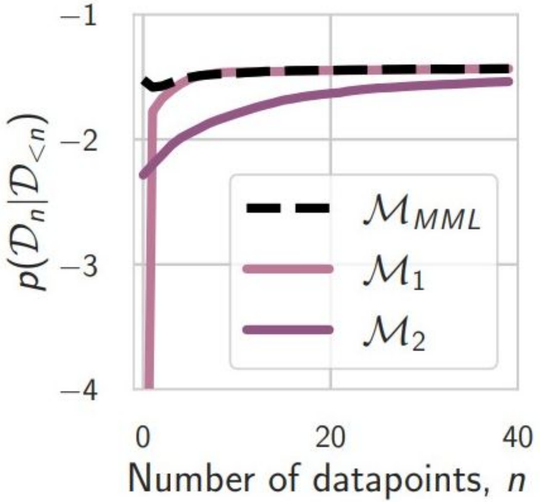
$$\log p(D \mid \mathcal{M}) = \sum_{i=1}^{n} \log p(D_i | D_{<i}, \mathcal{M}),$$

$\log p(D_i | D_{<i}, \mathcal{M})$: the predictive log-likelihood of the data point D_i under the Bayesian model average after observing the data $D_{<i}$ containing all samples before i.

Decomposition of the marginal likelihood

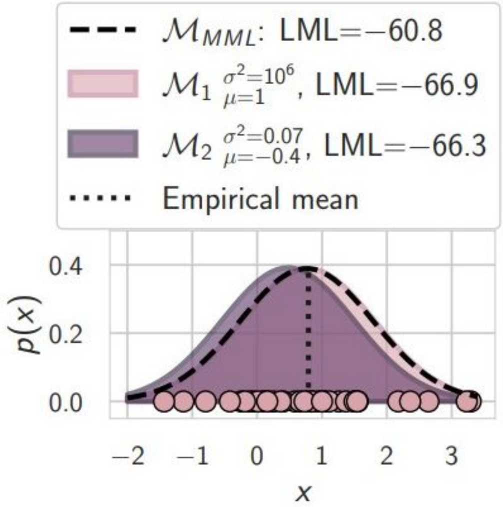
- We can decompose the LML: $\log p(D \mid \mathcal{M}) = \sum_{i=1}^{n} \log p(D_i \mid D_{<i}, \mathcal{M})$.
- Back to the density estimation example: $x \sim \mathcal{N}(u, 1), u \sim \mathcal{N}(\mu, \sigma^2)$

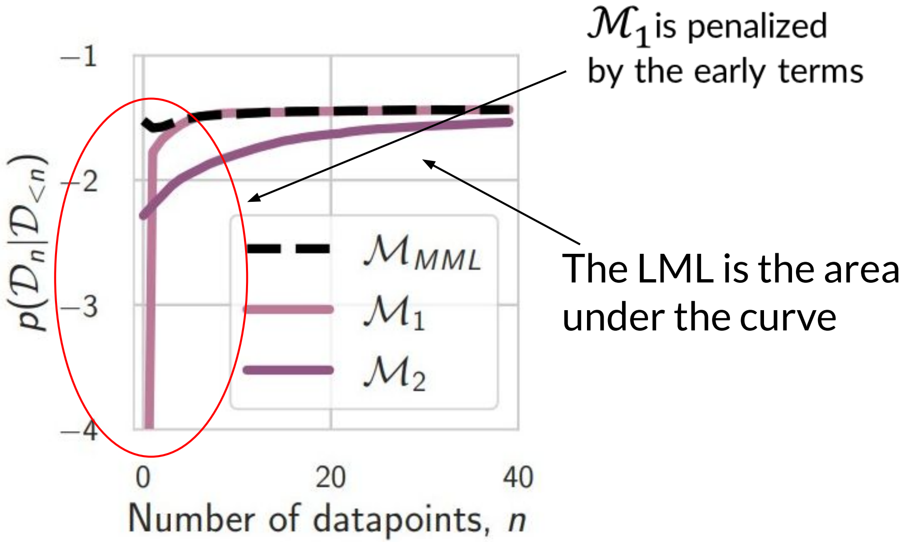




Decomposition of the marginal likelihood

- We can decompose the LML: $\log p(D \mid \mathcal{M}) = \sum_{i=1}^{n} \log p(D_i \mid D_{<i}, \mathcal{M})$.
- Back to the density estimation example: $x \sim \mathcal{N}(u, 1), u \sim \mathcal{N}(\mu, \sigma^2)$

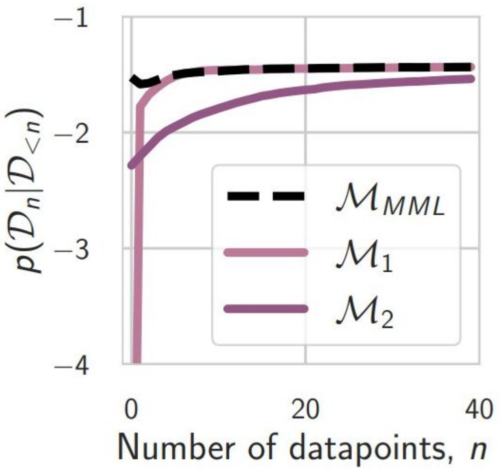




The conditional marginal likelihood

- What if we formed a posterior over a subset of the data and used it as a prior to compute LML for the rest of the data?
- This is equivalent to ignoring the first *m* terms in the LML decomposition.
- We define the conditional log marginal likelihood (CLML):

$$\log p(D_{\geq m}|D_{<m}, \mathcal{M}) = \sum_{i=m}^{n} \log p(D_i|D_{<i}, \mathcal{M})$$



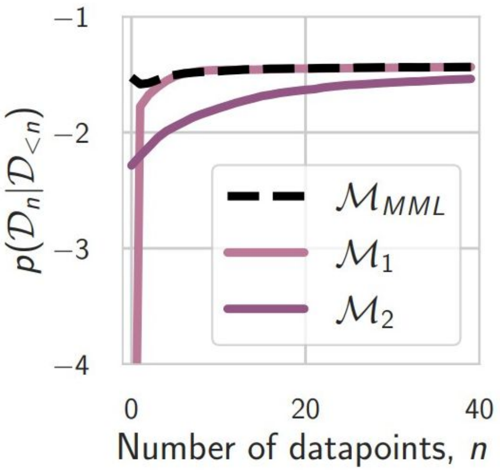
The conditional marginal likelihood

- What if we formed a posterior over a subset of the data and used it as a prior to compute LML for the rest of the data?

- This is equivalent to ignoring the first *m* terms in the LML decomposition.

- We define the conditional log marginal likelihood (CLML):

$$\log p(D_{\geq m}|D_{<m}, \mathcal{M}) = \sum_{i=m}^{n} \log p(D_i|D_{<i}, \mathcal{M})$$



- *The CLML has been considered for reducing prior sensitivity, but not to address underfitting, hyperparameter learning, neural architecture search, or model comparison with approximate inference.*

The conditional marginal likelihood: experimental results!

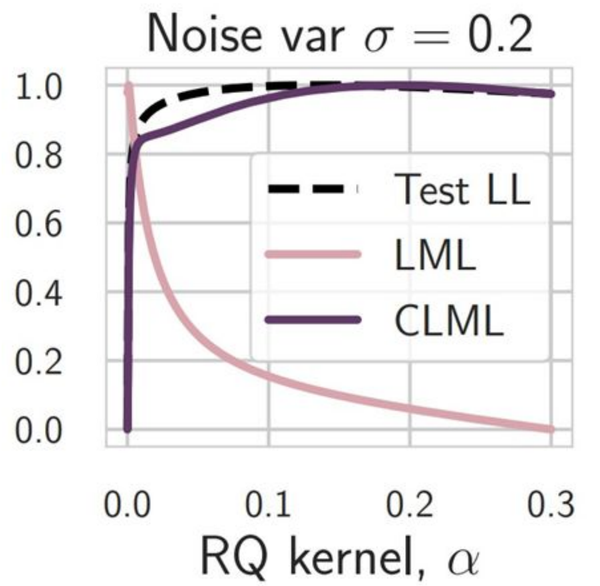
The CLML is more aligned with generalization

- Rational quadratic (RQ) kernel:

$$k_{RQ}(x, x') = a^2 \left(1 + ||x - x'||^2 / (2\,\alpha\, l^2)\right)^{-\alpha}$$

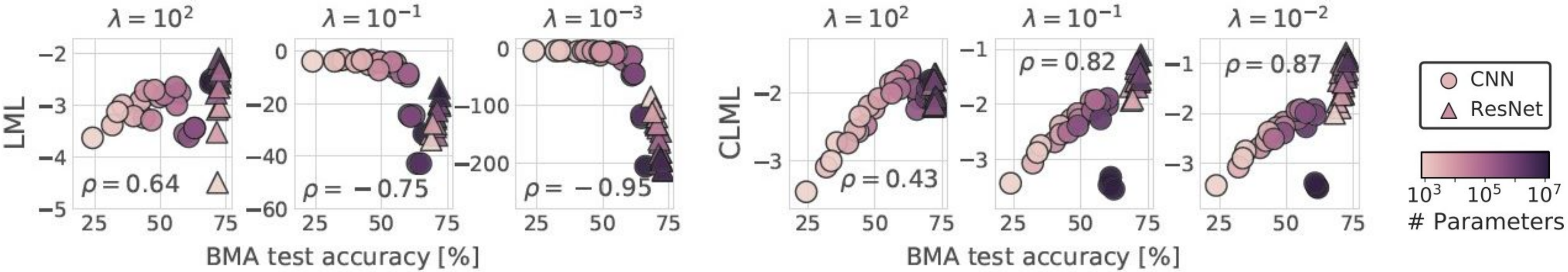
- The LML is misaligned with the shape of the test log-likelihood for large noise observation values.

- The CLML is more robust to model misspecification.



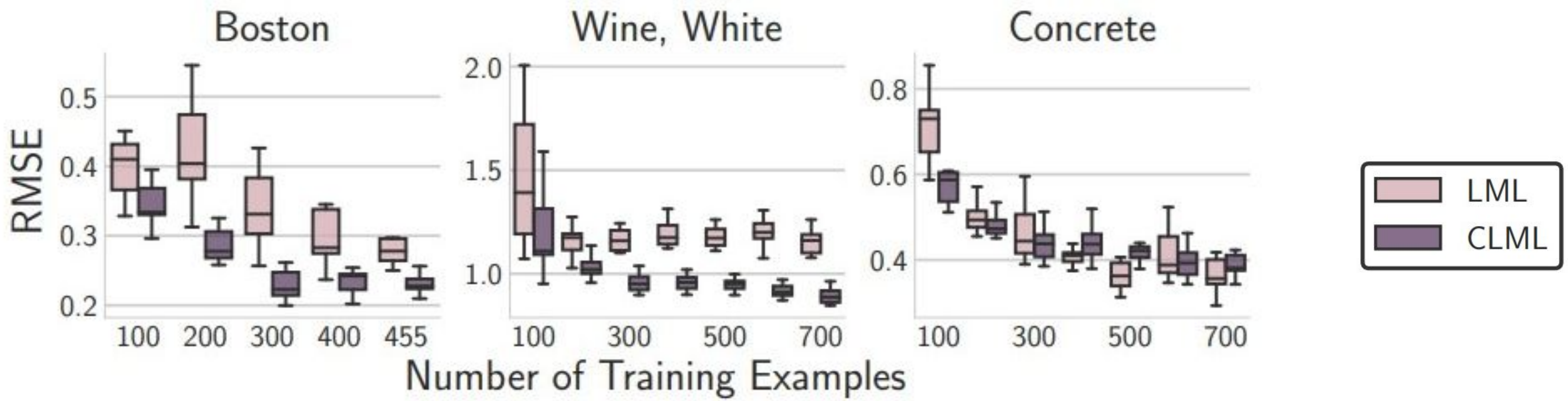
The CLML for neural architecture search, CIFAR-100

- The LML is not always aligned with generalization.
- CLML is aligned with generalization for all prior precisions!



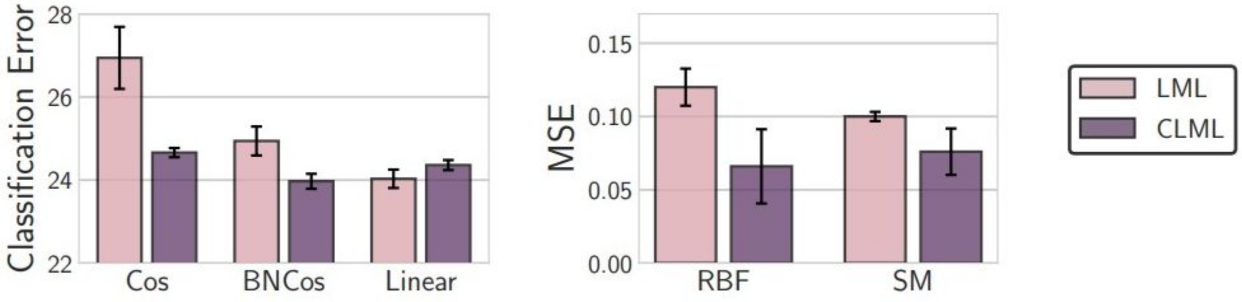
The CLML for deep kernel learning (DKL), regression

- CLML optimization outperforms LML optimization in low data regimes.



The CLML for deep kernel learning (DKL), classification

- CLML optimization outperforms LML optimization for different kernels and transfer learning tasks.

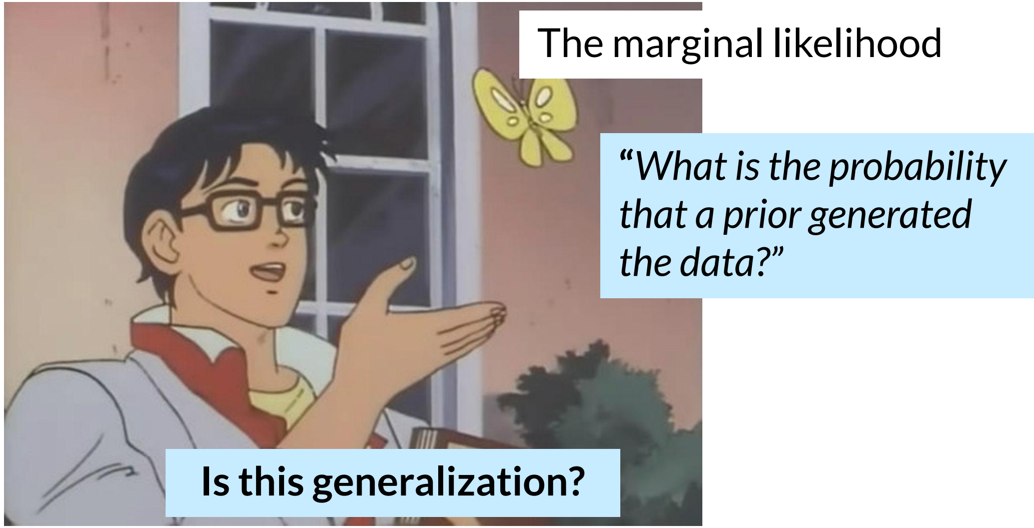


(b) Transfer to Omniglot (c) Transfer to QMUL

Concluding remarks

Conclusion

- The marginal likelihood story is more *nuanced*: “how likely is my prior to have generated the data?” ≠ “how likely is my posterior to make good predictions?”



Paper Code

Conclusion

- The marginal likelihood is reasonable for comparing *fixed prior scientific hypotheses*, but answers the wrong question for *predicting the generalization* of trained models.

- The marginal likelihood can overfit and underfit.

- The CLML provides an alternative to the LML that addresses underfitting.

Find us during the poster session: poster 828, hall E, between 6 and 8 pm!