

BAYESIAN MODEL SELECTION: MARGINAL LIKELIHOOD, CV, CLML, AND BEYOND

Andreas Kirsch
University of Oxford²⁰²³

Occam's Razor

Occam's Razor is the principle that, all else being equal, the simplest explanation tends to be the right one.

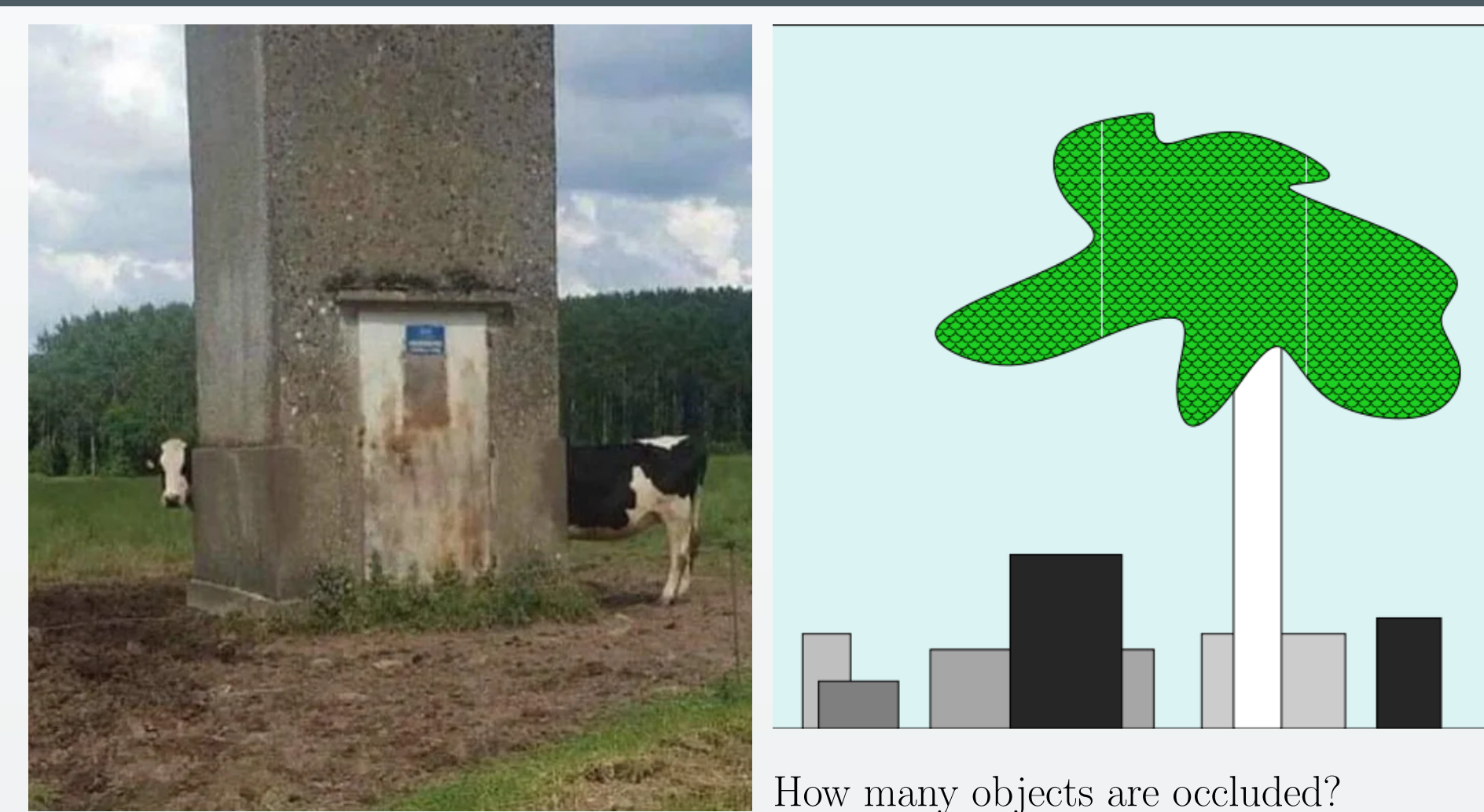
TL;DR

Occam's Razor & Bayesian Model Selection:

Minimize information, maximize likelihood

BUT ambiguous about multiple data points & objectives

Examples (left: Reddit, right: MacKay)



Shannon's Information Content

The *information content* (or *surprisal*) of an event x with probability $p(x)$ is:

$$H[x] = -\log p(x) \quad (1)$$

Likewise, the *entropy* of a random variable \mathbf{X} is:

$$H[\mathbf{X}] = \mathbb{E}_{p(\mathbf{X})} [H[\mathbf{X}]] = -\mathbb{E}_{p(\mathbf{X})} [\log p(\mathbf{X})] \quad (2)$$

Minimum Description Length (MDL)/MLE/MAP

The **Minimum Description Length (MDL)** formalizes Occam's razor by selecting the model that minimizes the sum of the model's description length (complexity) and the data's description length given the model (fit). For a model ϕ and data $(\mathbf{x}_n)_{n=1}^N$, the MDL criterion is:

$$H[\phi] + H[(\mathbf{x}_n)_{n=1}^N | \phi] \quad (3)$$

where $H[\phi]$ is the model's description length and $H[(\mathbf{x}_n)_{n=1}^N | \phi]$ is the data's description length given the model. The model with the lowest MDL score is selected.

Multiple vs Individual Points

For multiple samples (given datasets), information-theoretic quantities for model selection can be computed in different ways, **leading to ambiguity**:

- **Joint Quantities** (e.g., joint marginal cross-entropy, conditional joint marginal information) substitute the dataset directly, e.g. $H[\mathcal{D} | \phi]$, and entail **in-context learning**.
- **Individual Quantities** (e.g., marginal cross-entropy, conditional marginal cross-entropy) focus on model performance over individual points, e.g., $H_{\hat{p}_{\text{data}} || p(\cdot | \phi)}[\mathbf{X}]$, similar to validation/test performance.

Different Information Quantities for Model Selection

Information Quantities

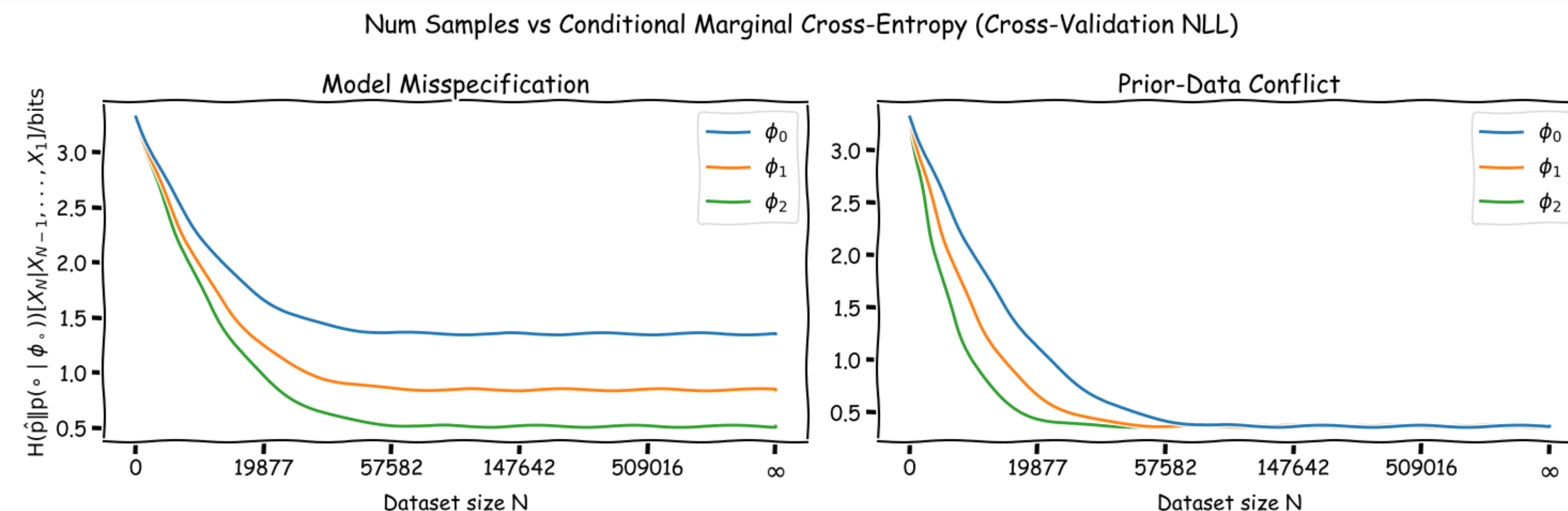
For a dataset $(\mathbf{x}_n)_{n=1}^N = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, we consider the following information-theoretic quantities for model selection:

- **Joint Marginal Cross-Entropy**: $H_{\hat{p}_{\text{data}} || p(\cdot | \phi)}[\{\mathbf{X}_n\}_{n=1}^N]$ The expected joint information content of a dataset $(\mathbf{X}_1, \dots, \mathbf{X}_N)$ under the model's joint prior predictive distribution, averaged over the true data distribution. Equivalent to the **log marginal likelihood (LML)**.
- **Conditional Marginal Cross-Entropy**: $H_{\hat{p}_{\text{data}} || p(\cdot | \phi)}[\mathbf{X}_n | \mathbf{X}_{n-1}, \dots, \mathbf{X}_1]$ The expected information content of a single data point \mathbf{X}_n conditioned on the previous data points $(\mathbf{X}_{n-1}, \dots, \mathbf{X}_1)$ under the model's predictive distribution, averaged over the true data distribution. Equivalent to **leave-one-out cross-validation**.
- **Conditional Joint Marginal Information**: $H[\{\mathbf{x}_n\}_{n=N-k+1}^N | \{\mathbf{x}_n\}_{n=1}^{N-k}, \phi]$ The joint information content of a dataset $\{\mathbf{x}_n\}_{n=N-k+1}^N$ conditioned on a previous dataset $\{\mathbf{x}_n\}_{n=1}^{N-k}$ under the model's joint predictive distribution. This is data-order dependent. Also known as the (negative) **conditional log marginal likelihood (CLML)** (Lotfi et al., 2022, main paper).
- **Conditional Joint Marginal Cross-Entropy**: $H_{\hat{p}_{\text{data}} || p(\cdot | \phi)}[\{\mathbf{X}_n\}_{n=N-k+1}^N | \{\mathbf{X}_n\}_{n=1}^{N-k}]$ The expected joint information content of a dataset $\{\mathbf{X}_n\}_{n=N-k+1}^N$ conditioned on a previous dataset $\{\mathbf{X}_n\}_{n=1}^{N-k}$ under the model's joint predictive distribution, averaged over the true data distribution. Measures the model's online learning (or in-context learning) performance. Also known as the (negative) **conditional log marginal likelihood (CLML)** (Lotfi et al., 2022, appendix).

Different Data Regimes, Model Misspecification and Prior-Data Conflict

TL;DR

1. In the **large-data regime** (or infinite data limit), the (rate of the) joint quantities and individual quantities converge to the same values. Different models perform differently due to different levels of **model misspecification**.
2. In the **low-data regime** (and *low* can still be a lot), these quantities will not have converged, and different models can perform differently due to **model misspecification** and **prior data conflict**, which can even be *anti-correlated*.



Model misspecification occurs when the assumed model class does not contain the true data-generating process. In this case:

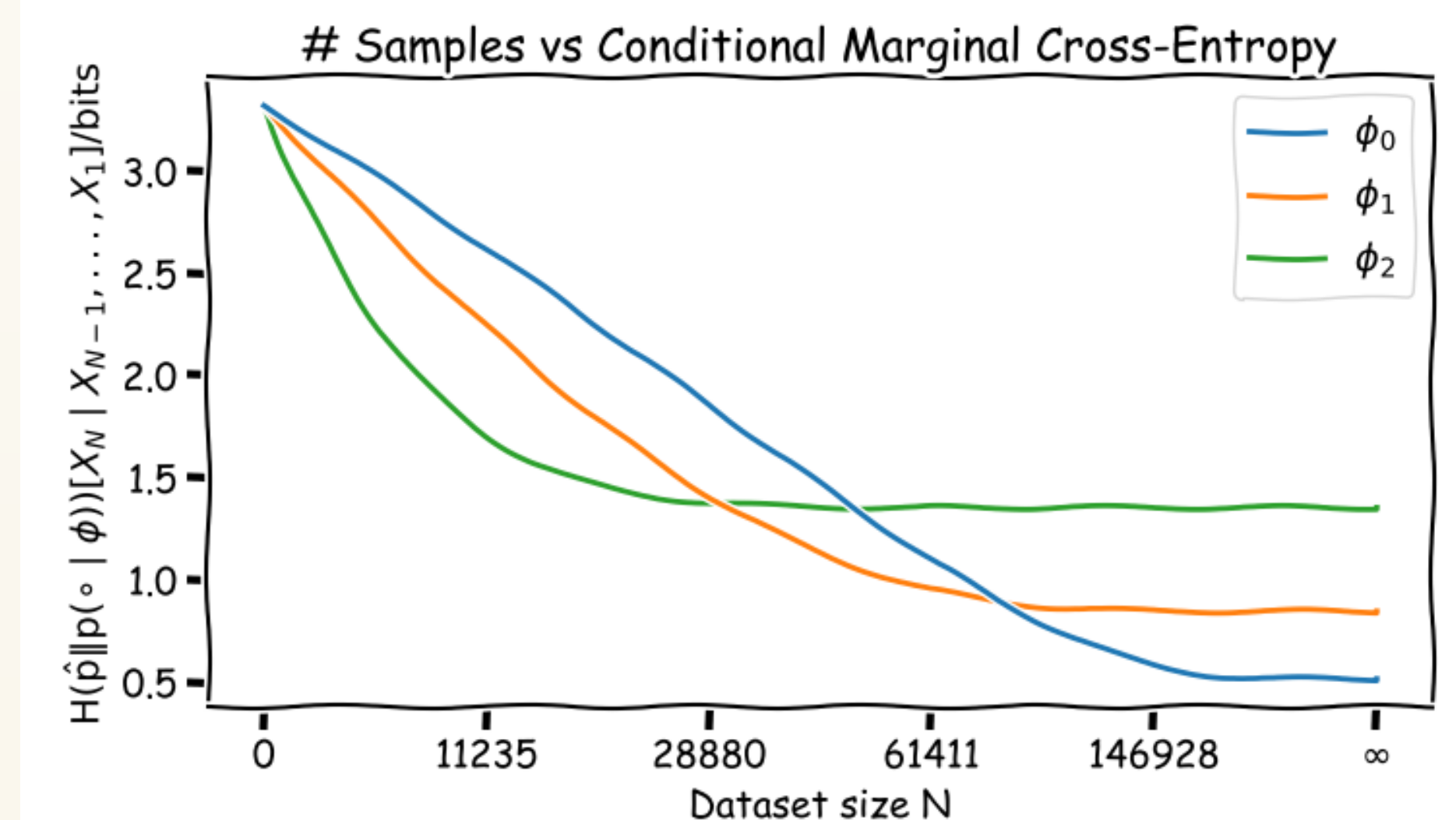
- Different models can have different levels of misspecification.
- In the infinite data limit, the model with the lowest misspecification (i.e., the closest to the true data-generating process) will perform best.

Prior-data conflict arises when the assumed prior distribution is inconsistent with the observed data. In this scenario:

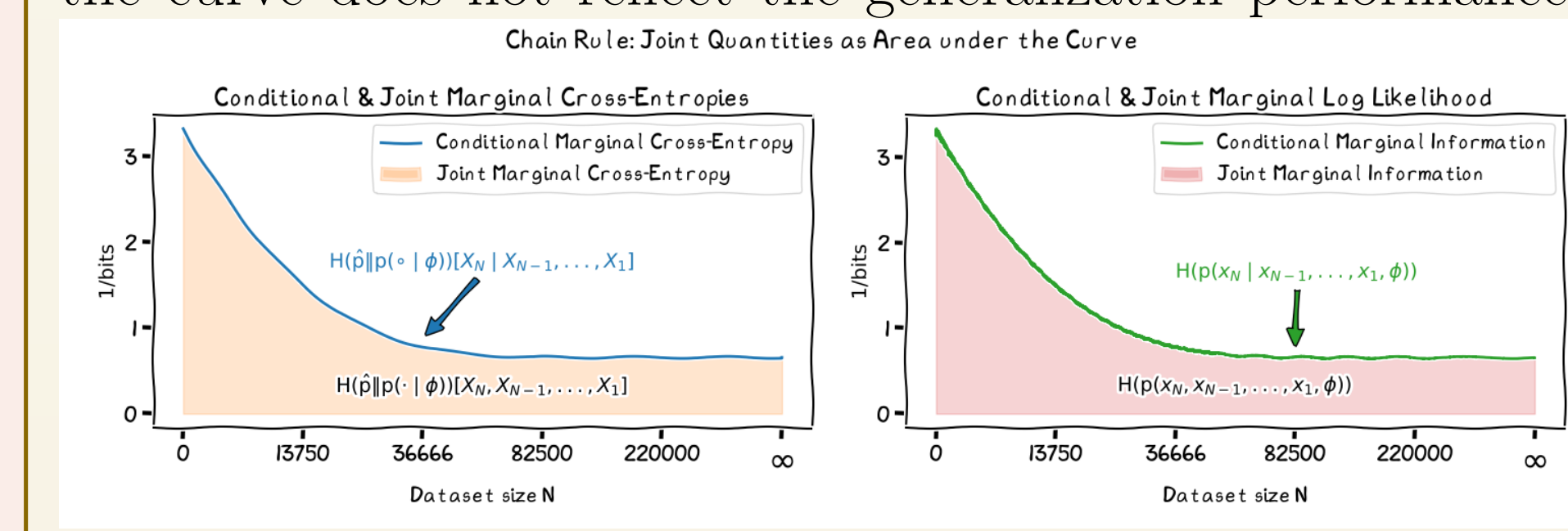
- Models with priors that are more consistent with the observed data will perform better initially.
- The effect of prior-data conflict diminishes as the dataset size increases and the likelihood term dominates the prior term.

Failures & Prior Art

Anti-Correlated Prior-Data Conflict & Model Misspecification



For anti-correlated prior-data conflict and model misspecification, existing methods fail: Training speed methods (TSE, TSE-E, TSE-EMA) (Lyle et al., 2020; Ru et al., 2021) and the conditional log marginal likelihood (CLML) (Fong and Holmes, 2020; Lotfi et al., 2022) *essentially* which approximate the validation loss by averaging under the loss curve might prefer more misspecified models in the low-data regime when the (partial) area under the curve does not reflect the generalization performance.



References

- Fong, E. and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496.
- Lotfi, S., Izmailov, P., Benton, G., Goldblum, M., and Wilson, A. G. (2022). Bayesian model selection, the marginal likelihood, and generalization. In *International Conference on Machine Learning*, pages 14223–14247. PMLR.
- Lyle, C., Schut, L., Ru, R., Gal, Y., and van der Wilk, M. (2020). A bayesian perspective on training speed and model selection. *Advances in neural information processing systems*, 33:10396–10408.
- Ru, B., Lyle, C., Schut, L., Fil, M., van der Wilk, M., and Gal, Y. (2021). Speedy performance estimation for neural architecture search. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

