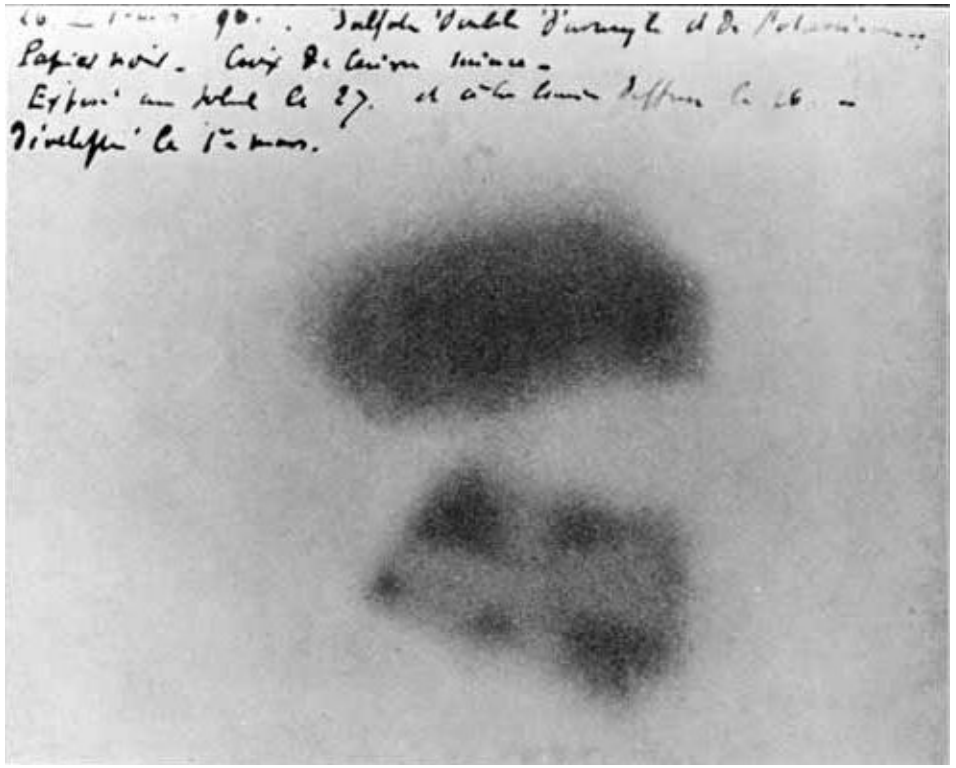
Local Learning Coefficient Estimation

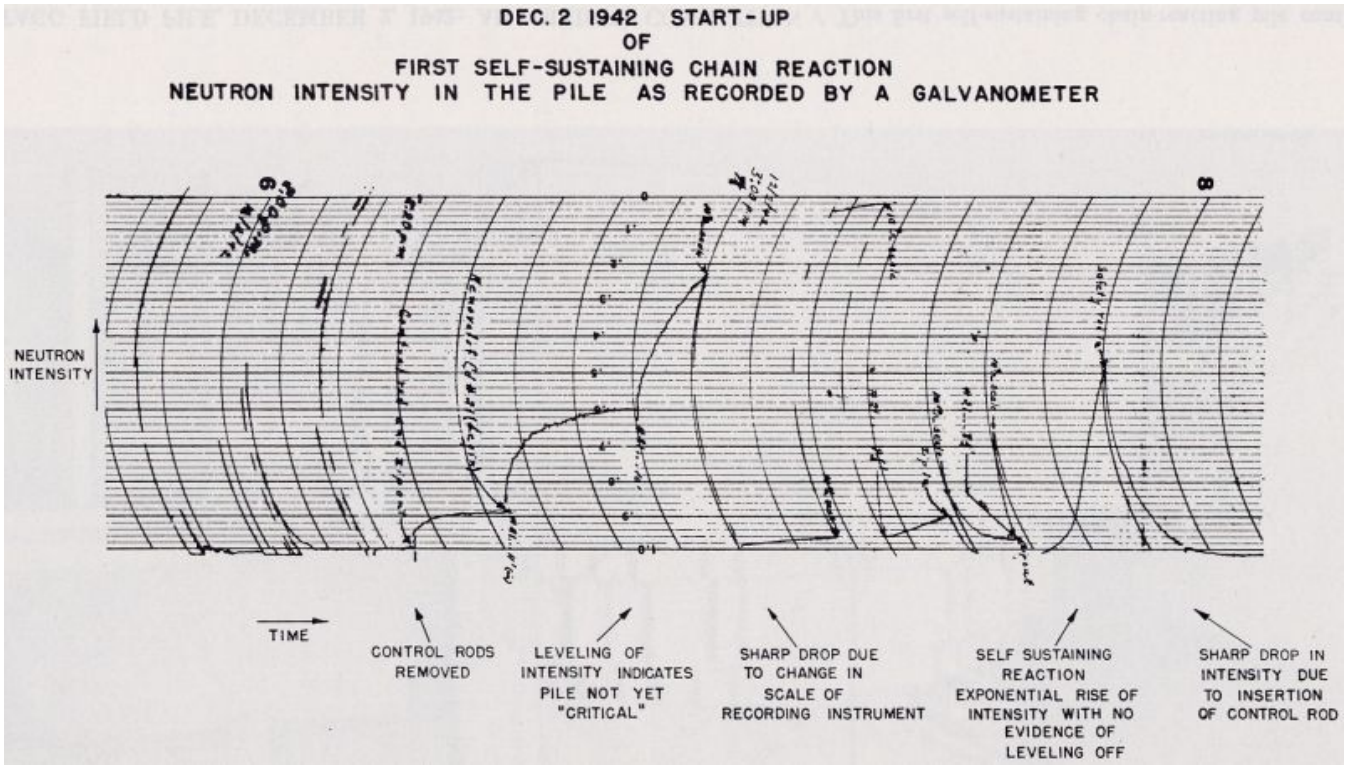
The story so far and open problems

Why does this matter?

… if you care about SLT



SOTA radiation detection
1896



SOTA radiation detection
1942

Background

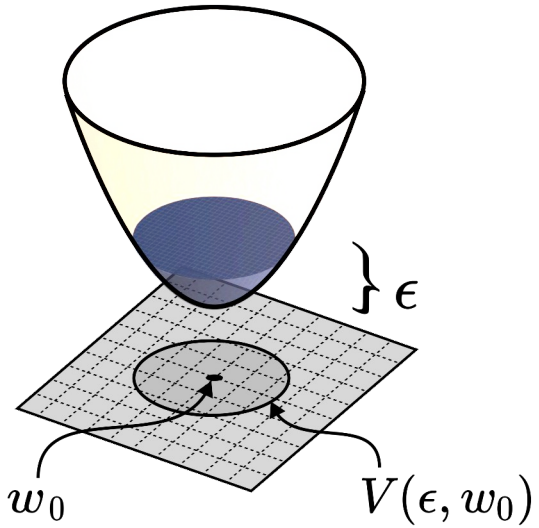
Setup and notation

- We consider a model distribution $p(x|w)$ determined in some way by a neural network, where $w \in \mathbb{R}^d$ are the weights
- Given iid data $X_1, \ldots, X_n$ drawn from $q(x)$, consider:
  - Empirical negative log-likelihood: $L_n(w) = \frac{-1}{n} \sum_{i=1}^{n} \log p(X_i|w)$
  - Expected negative log-likelihood: $L(w) = -\int \log p(x|w) q(x) dx$
- Eg:
  - Regression: $p(x_{\text{in}}, x_{\text{out}}|w) \propto \exp\left(\frac{-1}{2} \|x_{\text{out}} - f(x_{\text{in}}, w)\|^2\right) q(x_{\text{in}})$, MSE loss
  - Autoregressive sequence: $p(x_1, \ldots, x_L|w) = \prod_{l=1}^{L} p(x_l|x_{1:l}, w)$, X-entropy

The local learning coefficient $\lambda(w_0)$

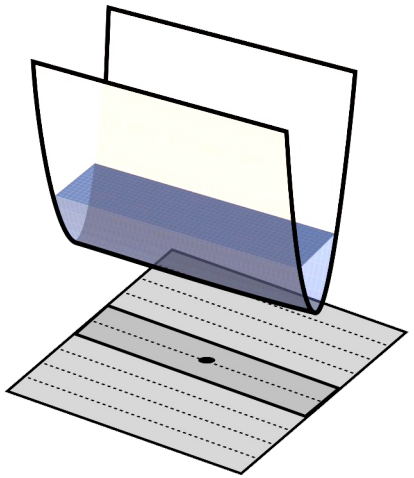
$$\text{volume}\{w \in W \mid L(w) \leq \epsilon\} \approx c\,\epsilon^{\lambda(w_0)}(-\log\epsilon)^{m(w_0)-1} \text{ as } \epsilon \to 0$$

$L(w) = w_1^2 + w_2^2$



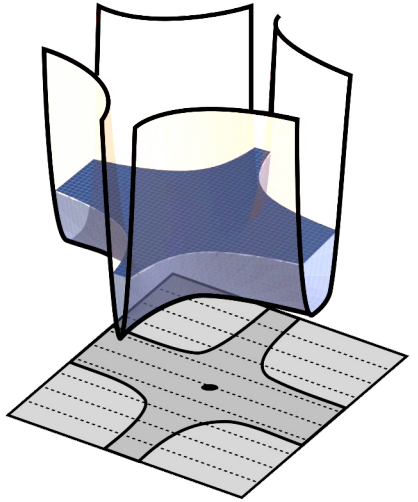
$\lambda(w_0) = 1$

$L(w) = w_1^2$



$\lambda(w_0) = \frac{1}{2}$

$L(w) = w_1^4 w_2^2$



$\lambda(w_0) = \frac{1}{4}$

Figure credit Jesse

The Story So Far

Definition (Lau et al. 2024)**:** Let $w_0 \in W$ be a nbhd and $\varphi(w)$ a prior supported on W. The *WBIC estimator* $\hat{\lambda}_n(w_0)$ of the local learning coefficient $\lambda(w_0)$

$$\hat{\lambda}_n(w_0) = n\beta(\mathbb{E}_w^\beta[L_n(w)] - L(w))$$

where

$$\mathbb{E}_w^\beta[L_n(w)] = \int_W L_n(w) \exp\big(-n\beta L_n(w)\big)\varphi(w)dw$$

Theorem* (Watanabe 2013)**:** Let $\beta = \beta_0/\log n$. For some sufficiently local nbhd W of w_0, $\hat{\lambda}_n(w_0)$ is an asymptotically consistent estimator of $\lambda(w_0)$.

*Some conditions are missing, see Watanabe (2013) or Hitchcock and Hoogland

Estimation in practice

- We can't do the integral $\mathbb{E}_w^\beta[L_n(w)] = \int_w \ldots dw$.
  - Use MCMC to obtain samples $w_1, \ldots, w_T$ such that

$$\mathbb{E}_w^\beta[L_n(w)] \approx \frac{1}{T}\sum_{t=1}^{T} L_n(w_t)$$

- We want to do this in big models; need each step $O(\text{backprop})$
  - Use a stochastic-gradient MCMC algorithm like SGLD

- Wait, our estimator requires a sufficiently local nbhd W of w_0?
  - Use prior $\varphi_\gamma(w) \propto \exp(-\frac{\gamma}{2}(w - w_0))$ where $\gamma > 0$ controls localization

Putting this together

- For SGLD, we generate $w_1, \ldots, w_T$ by $w_{t+1} = w_t + \Delta w_t$ where

$$\Delta w_t = -\frac{\epsilon}{2}\left(\gamma(w_t - w_0) + n\beta\nabla L_{m,t}(w_t)\right) + \sqrt{\epsilon}\eta_t$$

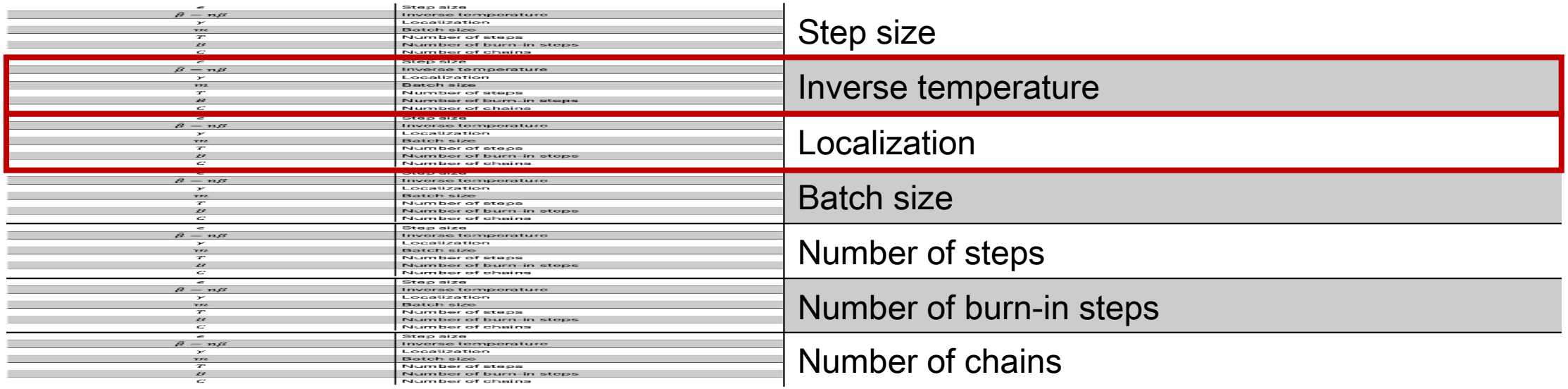
prior posterior noise

where $\epsilon > 0$ is a chosen step size, $\nabla L_{m,t}(w) \approx \nabla L_n(w)$ is a batch-size m estimate of the gradient, $\eta_1, \ldots, \eta_T$ are iid normal.

- LLC is estimated as

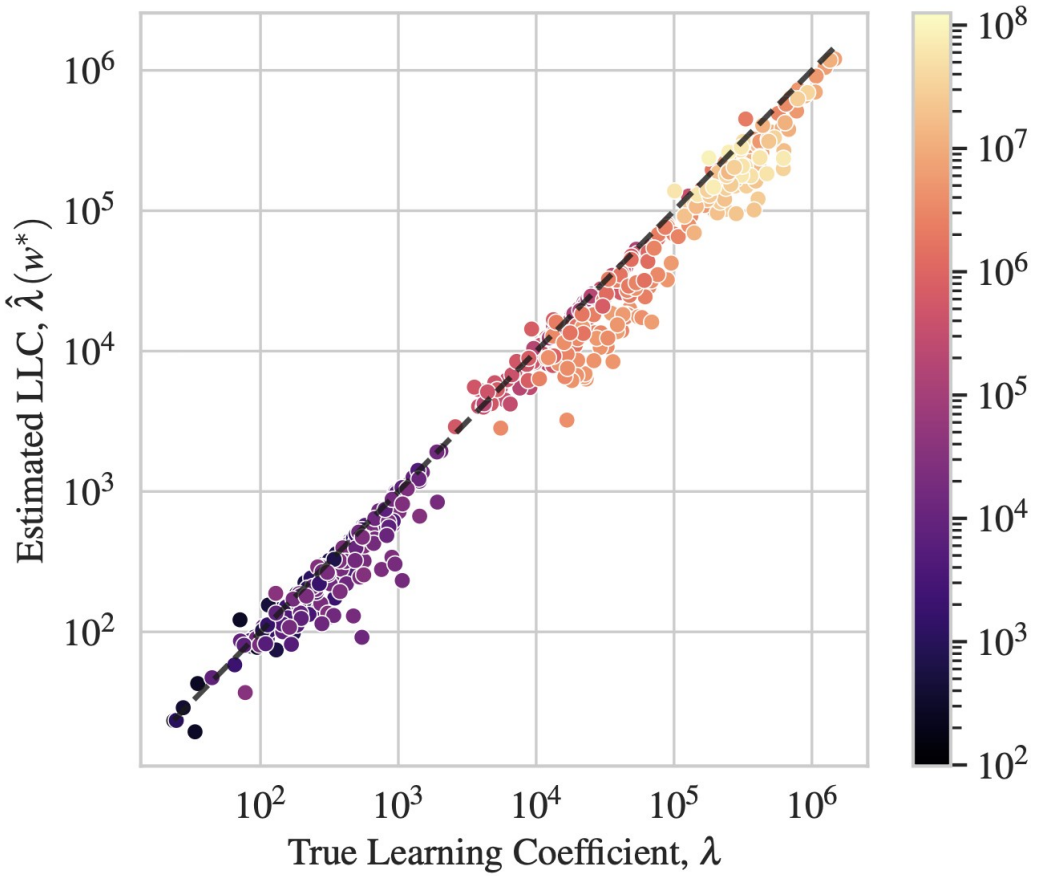
$$\bar{\lambda}(w_0) = n\beta\left(\frac{1}{T-B}\sum_{t=B}^{T} L_{m,t}(w_t) - L_{m,0}(w_0)\right)$$

Hyperparameters in one place

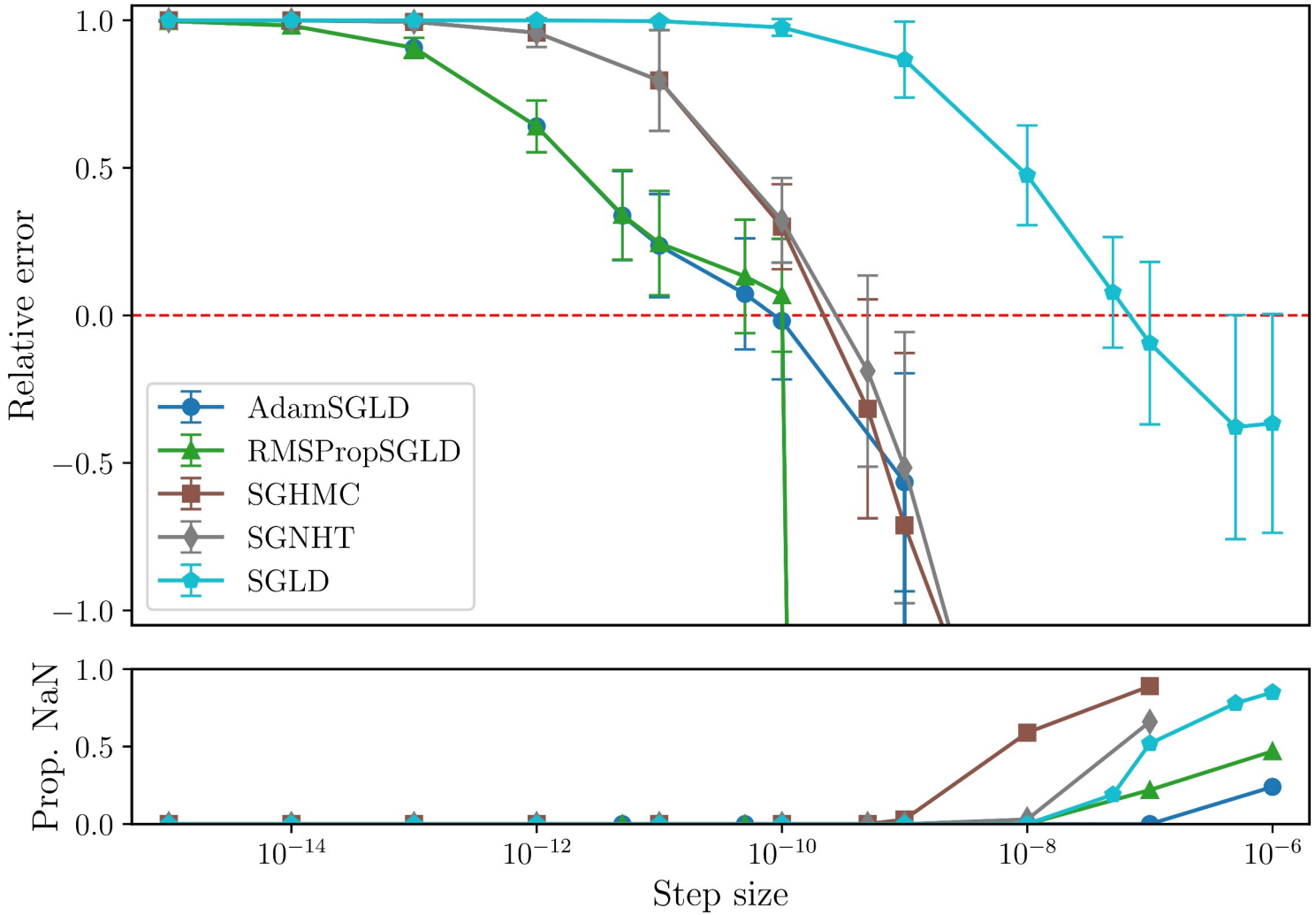


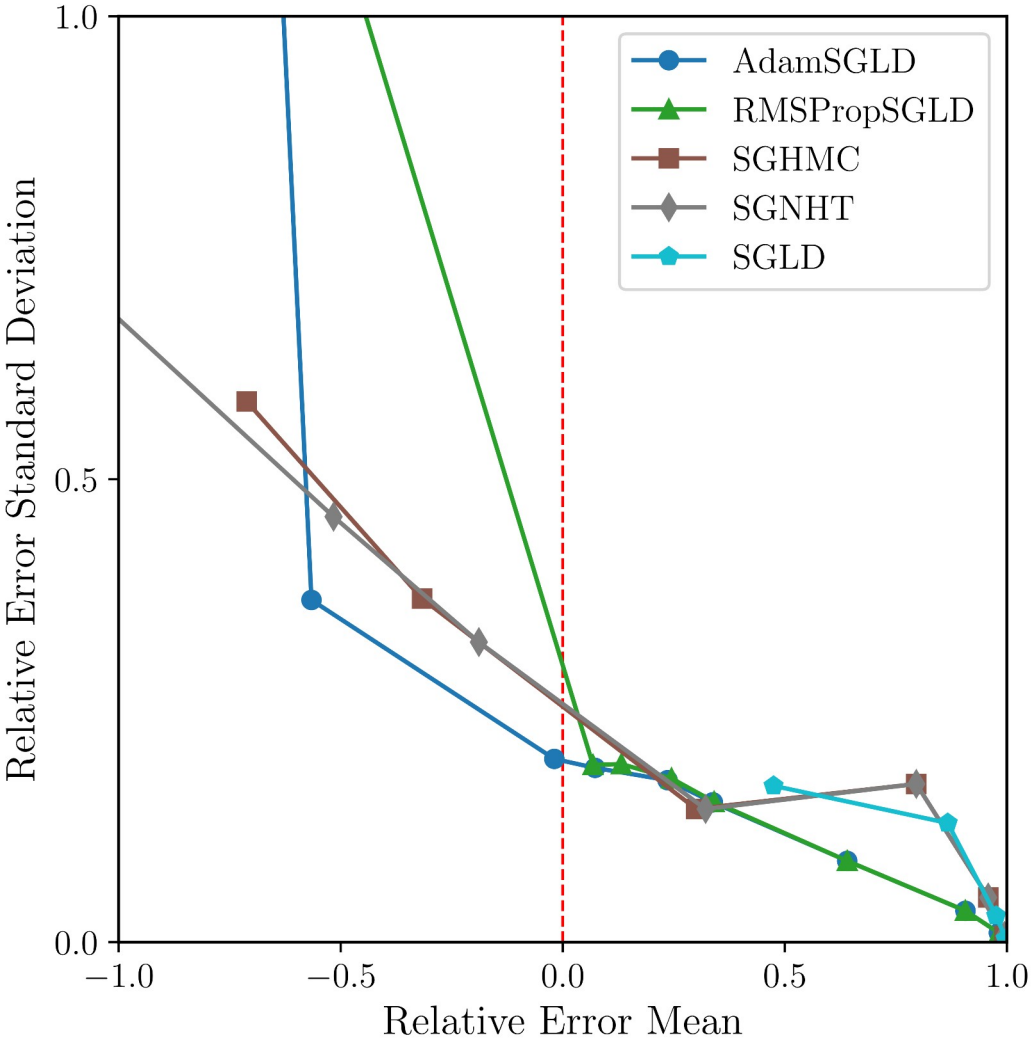
Empirical verification in DLNs

- Ground-truth LC values are known for DLNs (Aoyagi, 2024 and Lehalleur and Rimányi, 2024)
- Lau et al 2024 checks SGLD-based LLC estimation in DLNs



Empirical verification in DLNs

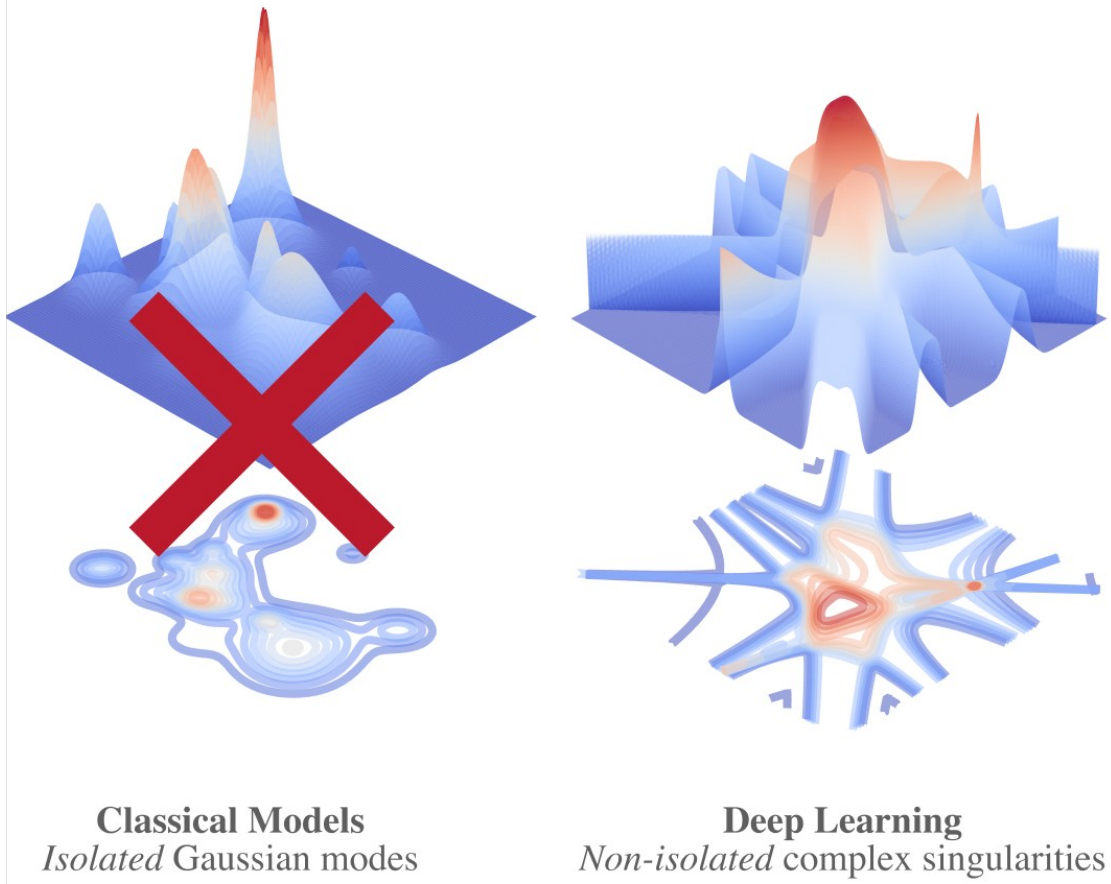




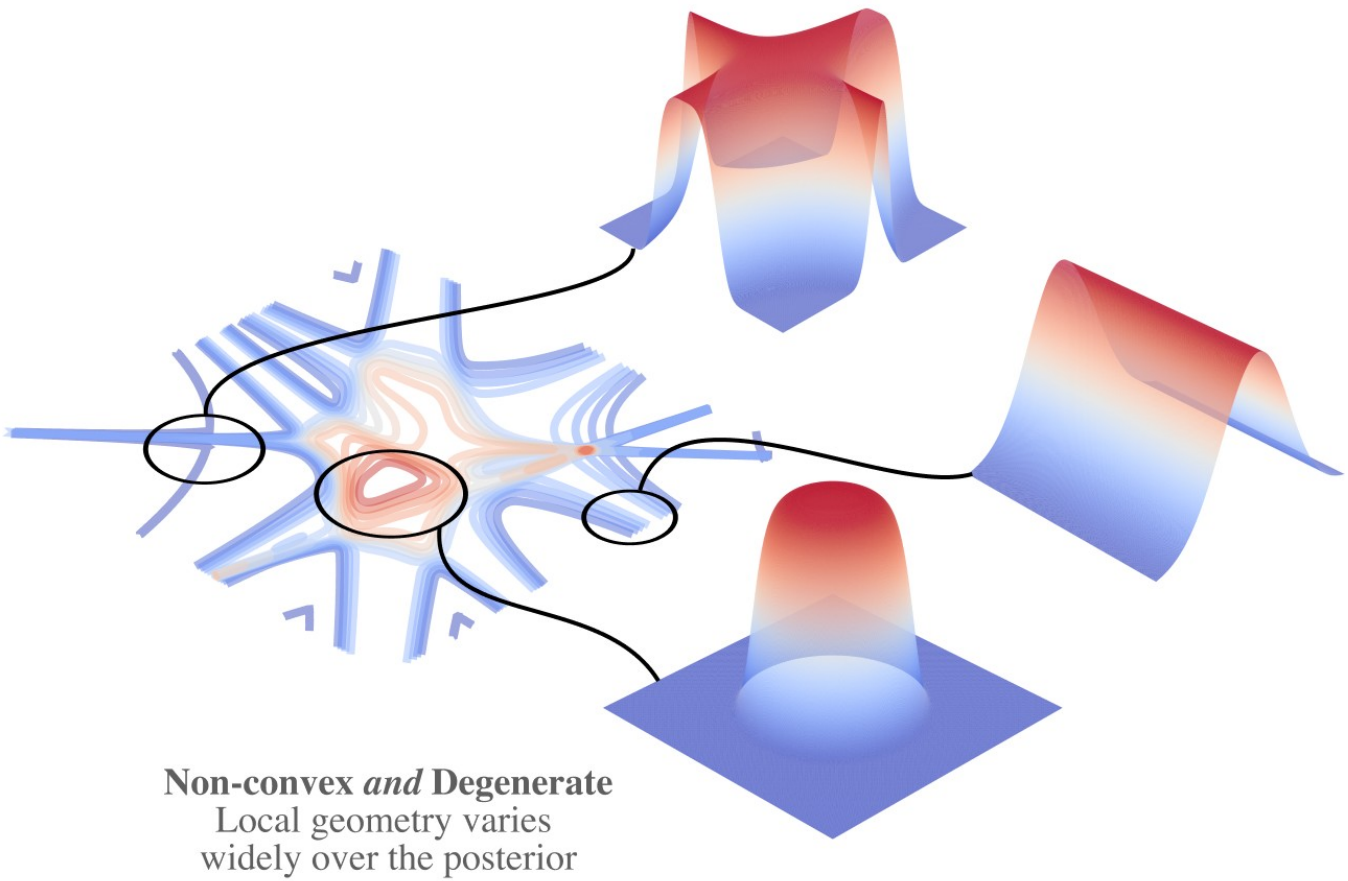
Hitchcock and Hoogland, 2025

Open Problems

How to *globally* sample from *non-convex* posteriors? → **How to *locally* sample from *degenerate* posteriors?**



(a) **False simplification:** Neural-network posteriors are not just multimodal but highly *degenerate*: changes in weights do not imply changes in posterior density.



(b) **Open problem:** Theory and practice still lack guarantees for local exploration, so progress demands a shift from a focus on global mode-hunting to understanding local geometry.

Figure credit Jesse

Open Problem 1: DLNs are ‘mildly singular’

- The degeneracy in deep linear networks is not as complex as it could be (Lehalleur and Rimányi, 2024)
  - Minimally singular except at particular points
- It is possible LLC estimation works well in DLNs, but is more difficult in complex degenerate landscapes
- There are a few other situations where ground-truth LLCs are known, but these seem further from usual neural networks than DLNs

Open Problem 2: Guarantees of convergence

- Proofs of convergence for SGLD (as used for LLC estimation) do not hold e.g. for DLNs
- In fact, some results suggest SGLD should eventually *diverge*
  - What distribution are we estimating from in the meantime?
- Why can we estimate the LLC in so few samples?
- What kind of observables of the loss landscape can we reliably estimate?

Open Problem 3: Locality and temperature

- What neighborhood are we *really* sampling from?
- How do we set the hyperparameters $\mathrm{n}\beta$ and γ?
  - Chen and Murfet (2025) suggest we should have $Mn\beta < \gamma$ in sequence models
  - But finding hyperparameters is difficult
- LLC estimates can be 'unreasonably' low, and/or temperature dependent (while still being ordinally consistent across different hyperparameter choices)

Open problem 4: Better samplers?

- Current vibe is that RMSPropSGLD (SGLD with adaptive preconditioning based on RMSProp) is ‘best’
- Can we do better?
  - ‘Better’ ≈ less sensitive to hyperparameter choices

- Watanabe (2013) “A widely applicable Bayesian information criterion”
- Aoyagi (2024) “Consideration on the learning efficiency of multiple-layered neural networks with linear units”
- Lau et al (2024) “The Local Learning Coefficient: A Singularity-Aware Complexity Measure”
- Lehalleur and Rimányi (2024) “Geometry of fibers of the multiplication map of deep linear neural networks”
- Chen and Murfet (2025) “Modes of Sequence Models and Learning Coefficients”
- Hitchcock and Hoogland (2025) “From Global to Local: A Scalable Benchmark for Local Posterior Sampling”