Entropy-MCMC: Sampling from Flat Basins with Ease

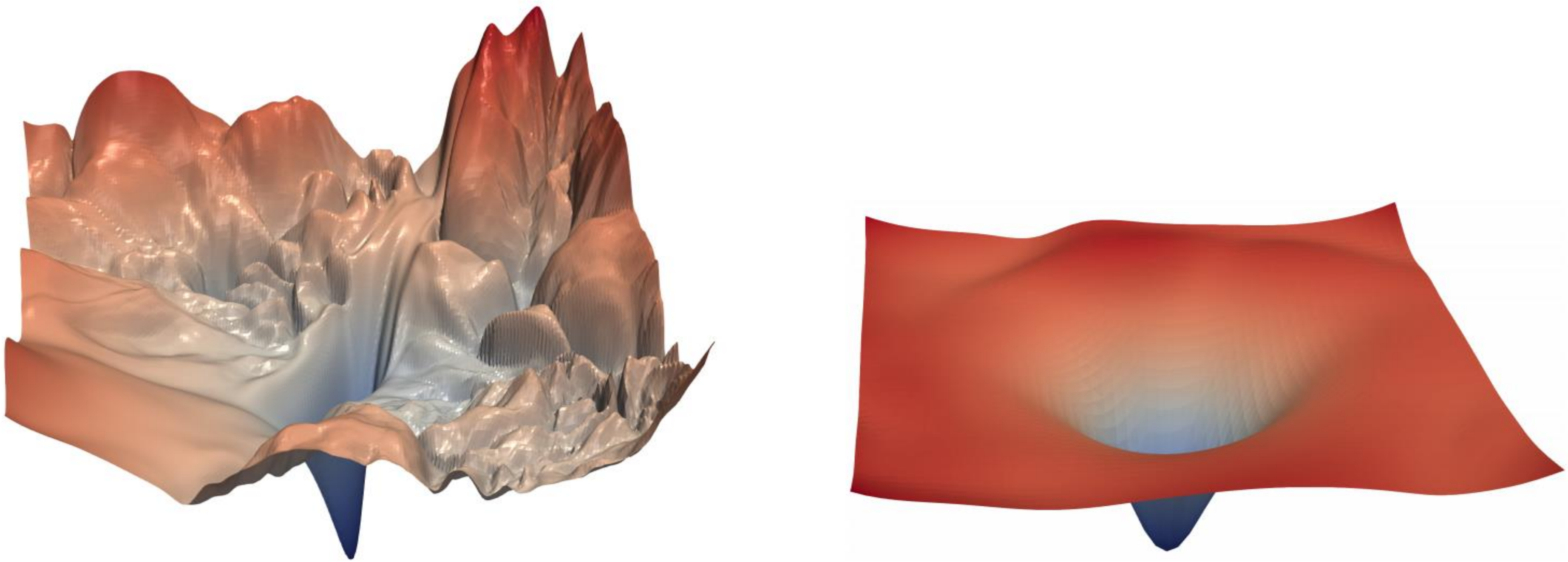
Bolian Li, Ruqi Zhang

Department of Computer Science
Purdue University

Scan for Paper

Introduction: Loss/Energy Landscape

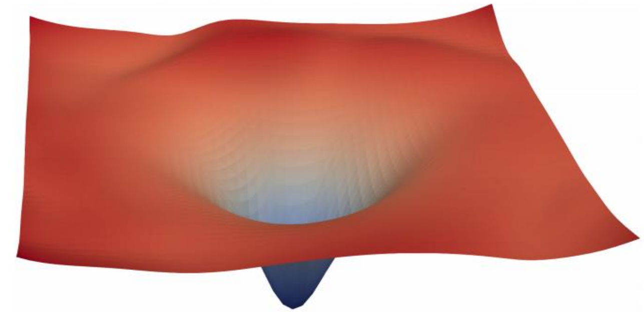
- Empirical observation: Flat minima generalize better.[1]



(a) without skip connections　　(b) with skip connections

[1] Keskar et al. On large-batch training for deep learning: Generalization gap and sharp minima. ICLR 2017.
[2] Li et al. Visualizing the Loss Landscape of Neural Nets. NeurIPS 2018.

Introduction: Motivation



- Energy landscape of DNNs is highly multi-modal.
- Not practical to sample from all modes.
- Flat modes generalize better.
- No MCMC methods consider flat minima before.

[2] Li et al. Visualizing the Loss Landscape of Neural Nets. NeurIPS 2018.

Preliminaries

Local entropy[3]:

$$\mathcal{F}(\boldsymbol{\theta};\eta) = \log \int_{\boldsymbol{\Theta}} \exp\left\{-f(\boldsymbol{\theta}') - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2\right\} d\boldsymbol{\theta}'$$

- Averaged energy within a local region.
- High local entropy indicates flat regions with low energy values.
- The main objective of Entropy-MCMC.

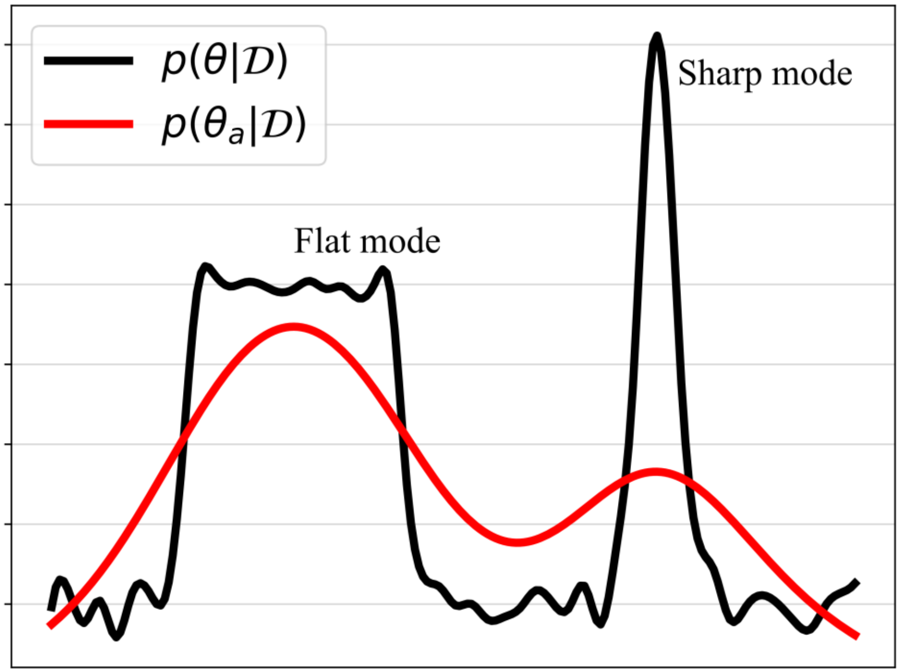
Stochastic gradient Langevin dynamics (SGLD)[4]:

- A standard MCMC algorithm.
- The backbone of Entropy-MCMC implementation.

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} U_{\boldsymbol{\Xi}}(\boldsymbol{\theta}) + \sqrt{2\alpha} \cdot \boldsymbol{\epsilon}$$

[3] Baldassi et al. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. Physical review letters, 2015.
[4] Welling et al. Bayesian learning via stochastic gradient Langevin dynamics. ICML 2011.

Method: Flat Posterior



- Original posterior: multi-modal, hard to sample from

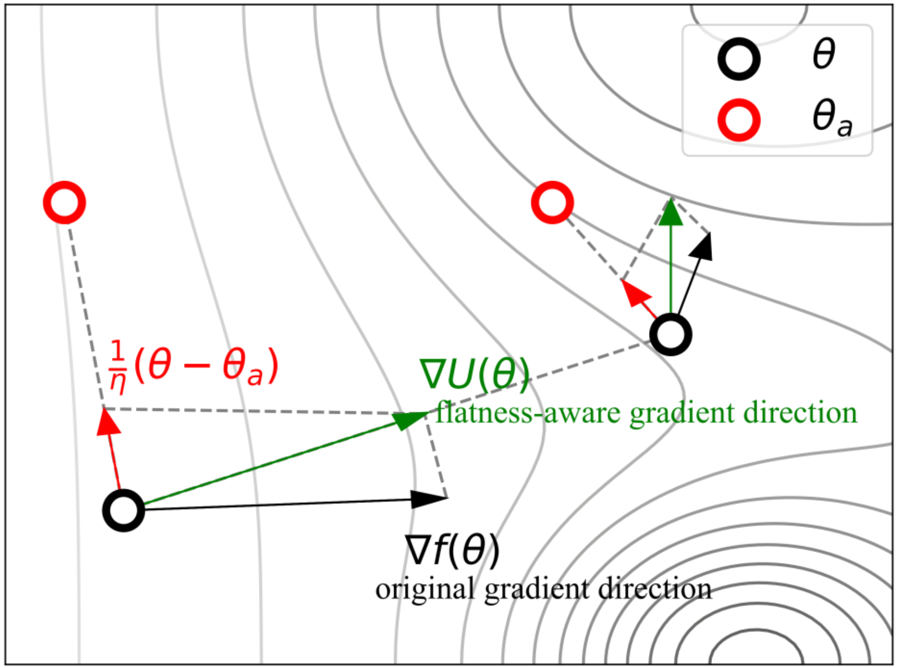
$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-f(\boldsymbol{\theta}))$$

- Flat posterior: fewer modes, smooth, easy to sample from

$$p(\boldsymbol{\theta}_a|\mathcal{D}) \propto \exp \mathcal{F}(\boldsymbol{\theta}_a; \eta) = \int_{\boldsymbol{\Theta}} \exp\left\{-f(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2\right\} d\boldsymbol{\theta}$$

- Flat posterior is computed by the local entropy.

Method: Sampling



- An auxiliary variable **θ_a** to eliminate the integral computation

$$p(\boldsymbol{\theta}_a|\mathcal{D}) \propto \exp \mathcal{F}(\boldsymbol{\theta}_a;\eta) = \boxed{\int_{\boldsymbol{\Theta}}} \exp\left\{-f(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2\right\} d\boldsymbol{\theta}$$

$$p(\widetilde{\boldsymbol{\theta}}|\mathcal{D}) = p(\boldsymbol{\theta}, \boxed{\boldsymbol{\theta}_a}|\mathcal{D}) \propto \exp\left\{-f(\boldsymbol{\theta}) - \frac{1}{2\eta}\|\boldsymbol{\theta}-\boldsymbol{\theta}_a\|^2\right\}$$

- For **θ**, its gradient direction is modified towards flat modes

$$\nabla_{\widetilde{\boldsymbol{\theta}}} U(\widetilde{\boldsymbol{\theta}}) = \begin{bmatrix} \nabla_{\boldsymbol{\theta}} U(\widetilde{\boldsymbol{\theta}}) \\ \nabla_{\boldsymbol{\theta}_a} U(\widetilde{\boldsymbol{\theta}}) \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \boxed{\frac{1}{\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}_a)} \\ \frac{1}{\eta}(\boldsymbol{\theta}_a - \boldsymbol{\theta}) \end{bmatrix}$$

Method: Sampling

Algorithm 1: Entropy-MCMC

Inputs: The model parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, guiding variable $\boldsymbol{\theta}_a \in \boldsymbol{\Theta}$, and dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$;
Results: Collected samples $\mathcal{S} \subset \boldsymbol{\Theta}$;
$\boldsymbol{\theta}_a \leftarrow \boldsymbol{\theta}, \mathcal{S} \leftarrow \emptyset$; `/* Initialize */`
for *each iteration* **do**
  $\boldsymbol{\Xi} \leftarrow$ A mini-batch sampled from $\mathcal{D}$;
  $U_{\boldsymbol{\Xi}} \leftarrow -\log p(\boldsymbol{\Xi}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_a\|^2$;
  $\boxed{\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha\nabla_{\boldsymbol{\theta}} U_{\boldsymbol{\Xi}} + \sqrt{2\alpha} \cdot \boldsymbol{\epsilon}_1 \; ; \quad \boldsymbol{\theta}_a \leftarrow \boldsymbol{\theta}_a - \alpha\nabla_{\boldsymbol{\theta}_a} U_{\boldsymbol{\Xi}} + \sqrt{2\alpha} \cdot \boldsymbol{\epsilon}_2;}$ `/*` $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ `*/`

  if *after burn-in* **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{\theta}, \boldsymbol{\theta}_a\}$; `/* Collect samples */`
  end
end

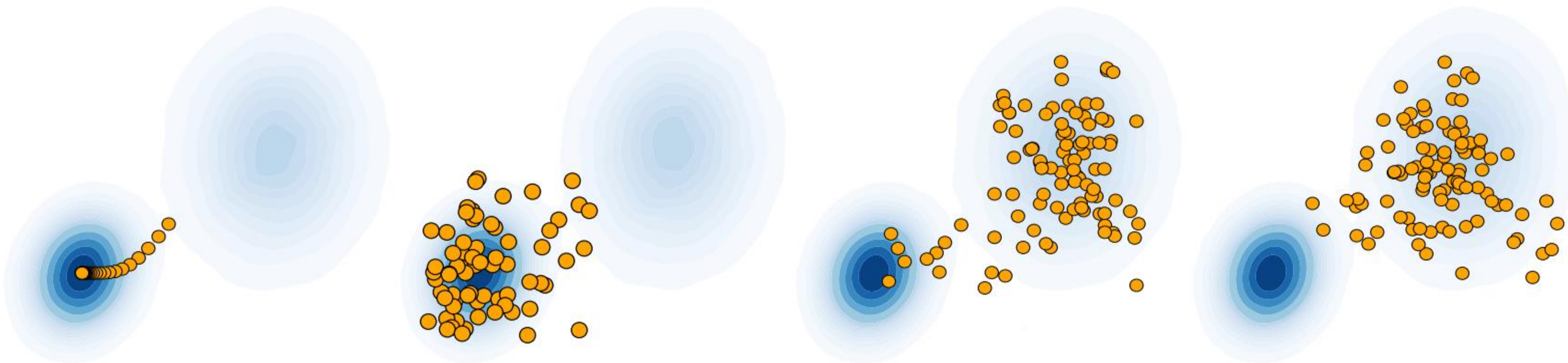
- Inference: Bayesian model averaging

Theoretical Analysis: Convergence Bound

Theorem (informal): Entropy-MCMC converges *faster* than Entropy-SGD and Entropy-SGLD in terms of 2-Wasserstein distance, due to the removal of nested Markov chains.

Experiments: Synthetic Examples

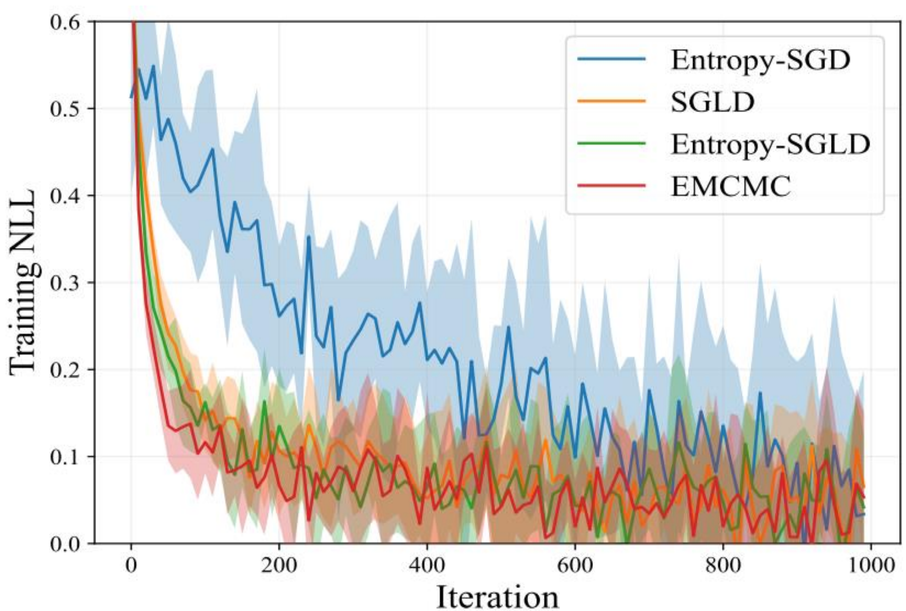
- One sharp mode and one flat mode

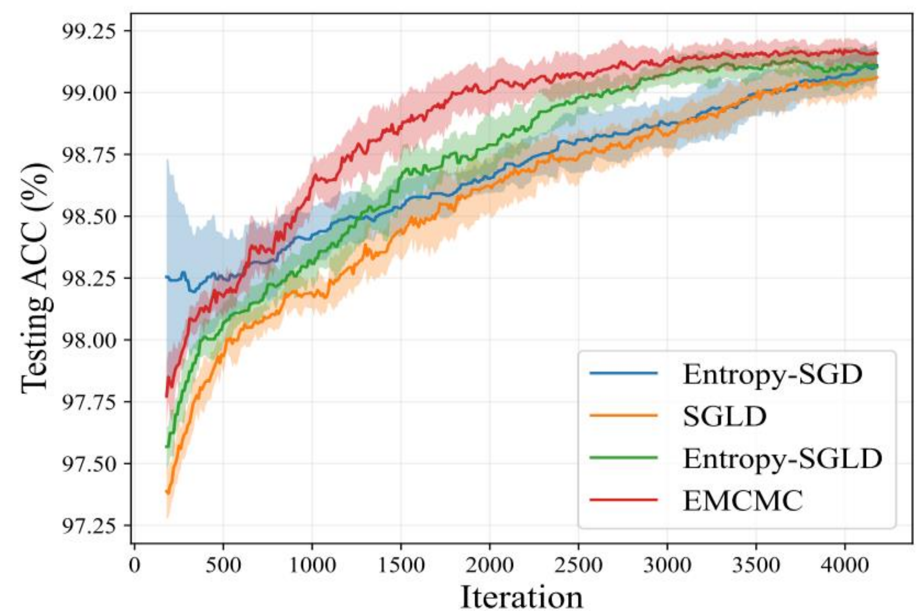


(a) SGD (b) SGLD (c) EMCMC ($\boldsymbol{\theta}$) (d) EMCMC ($\boldsymbol{\theta}_a$)

Experiments: Logistic Regression

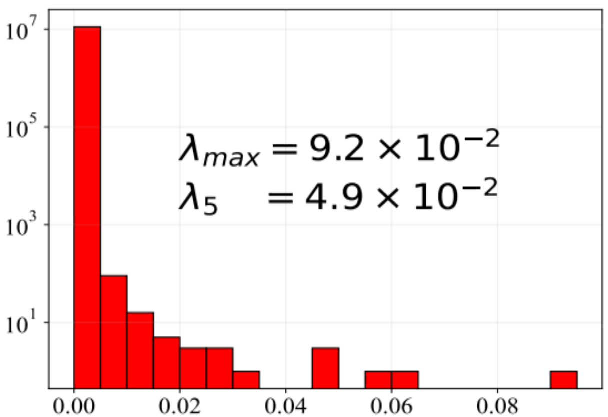
- Entropy-MCMC converges the fastest



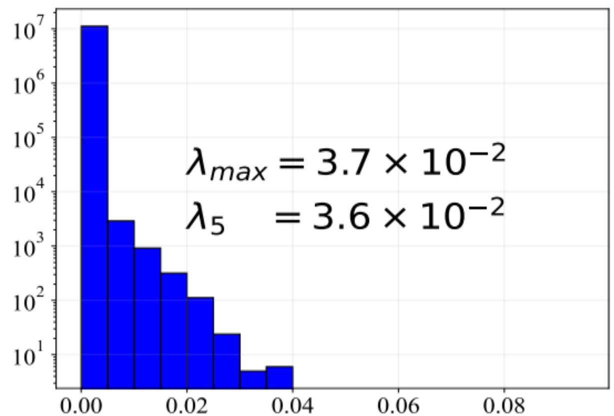


Experiments: Hessian Eigenspectrum

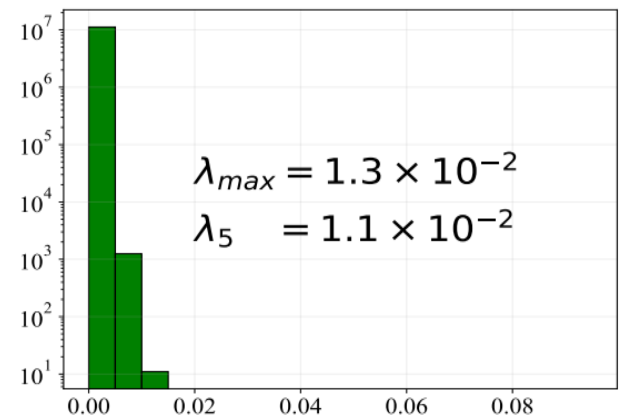
- Lower Hessian eigenvalues indicate more flatness



(a) SGD



(b) SGLD

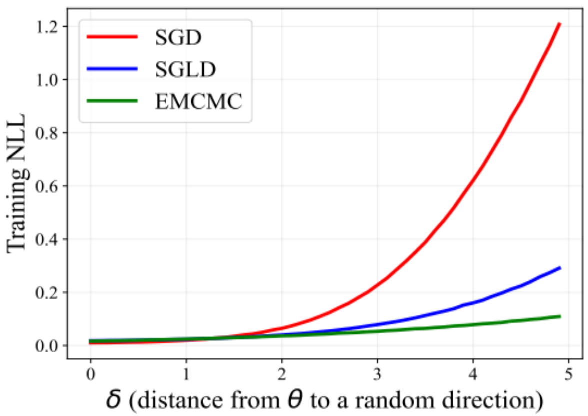


(c) EMCMC

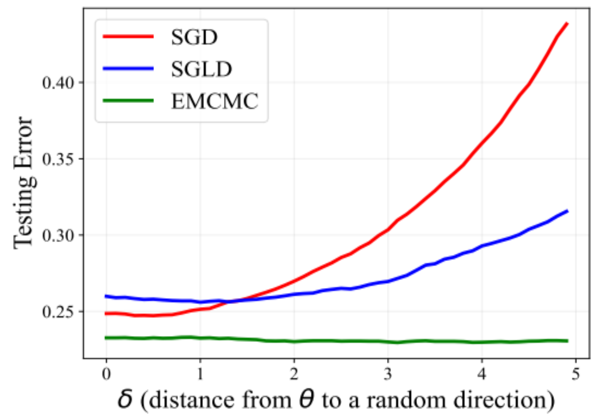
X-axis : eigenvalues
Y-axis : frequency

Experiments: Interpolation

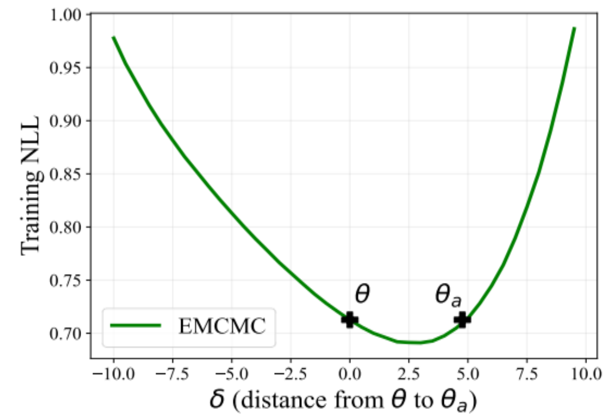
- The mode discovered by Entropy-MCMC is flatter than others



(a) $\boldsymbol{\theta} \to$ Random



(b) $\boldsymbol{\theta} \to$ Random



(c) $\boldsymbol{\theta} \to \boldsymbol{\theta}_a$

Experiments: Image Classification

(a) CIFAR10 and CIFAR100

Method	CIFAR10		CIFAR100	
	ACC (%) $\uparrow$	NLL $\downarrow$	ACC (%) $\uparrow$	NLL $\downarrow$
SGD	94.87 ± 0.04	0.205 ± 0.015	76.49 ± 0.27	0.935 ± 0.021
Entropy-SGD	95.11 ± 0.09	0.184 ± 0.020	77.45 ± 0.03	0.895 ± 0.009
SAM	95.25 ± 0.12	0.166 ± 0.005	78.41 ± 0.22	0.876 ± 0.007
bSAM	95.53 ± 0.09	0.165 ± 0.002	78.92 ± 0.25	0.870 ± 0.005
SGLD	95.47 ± 0.11	0.167 ± 0.011	78.79 ± 0.35	0.854 ± 0.031
Entropy-SGLD	94.46 ± 0.24	0.194 ± 0.020	77.98 ± 0.39	0.897 ± 0.027
EMCMC	$\mathbf{95.69 \pm 0.06}$	$\mathbf{0.162 \pm 0.002}$	$\mathbf{79.16 \pm 0.07}$	$\mathbf{0.840 \pm 0.004}$

(b) Corrupted CIFAR (ACC (%) $\uparrow$)

Severity	1	2	3	4	5
SGD	88.43	82.43	76.20	67.93	55.81
SGLD	88.61	82.46	76.49	69.19	56.98
EMCMC	**88.87**	**83.27**	**77.44**	**70.31**	**58.17**

(c) ImageNet

Metric	NLL $\downarrow$	Top-1 (%) $\uparrow$	Top-5 (%) $\uparrow$
SGD	0.960	76.046	92.776
SGLD	0.921	76.676	93.174
EMCMC	**0.895**	**77.096**	**93.424**

Experiments: OOD detection

- Predictive uncertainty
- Good characterization of posterior leads to good OOD detection

Method	CIFAR10-SVHN		CIFAR100-SVHN	
	AUROC (%) ↑	AUPR (%) ↑	AUROC (%) ↑	AUPR (%) ↑
SGD	**98.30**	**99.24**	71.96	84.08
Entropy-SGD	**98.71**	**99.37**	79.15	86.92
SAM	94.23	95.67	74.56	84.61
SGLD	97.66	98.64	72.51	83.35
Entropy-SGLD	90.07	91.80	71.83	82.89
EMCMC	**98.15**	**99.04**	**81.14**	**87.18**

Conclusion

1. Sampling from the flat basins can improve the generalization of MCMC samples.

2. The proposed joint posterior distribution can eliminate the need for integral computation.

3. Entropy-MCMC can effectively find flat modes and achieve promising empirical results.

Thank You!