Entropic algorithms and wide flat minima

Carlo Lucibello

Artificial Intelligence Lab

Bocconi

in collaboration with:

F. Pittorino, C. Feinauer, E. Malatesta, G. Perugini, C. Baldassi, M. Negri, E. Demyanenko, R. Zecchina





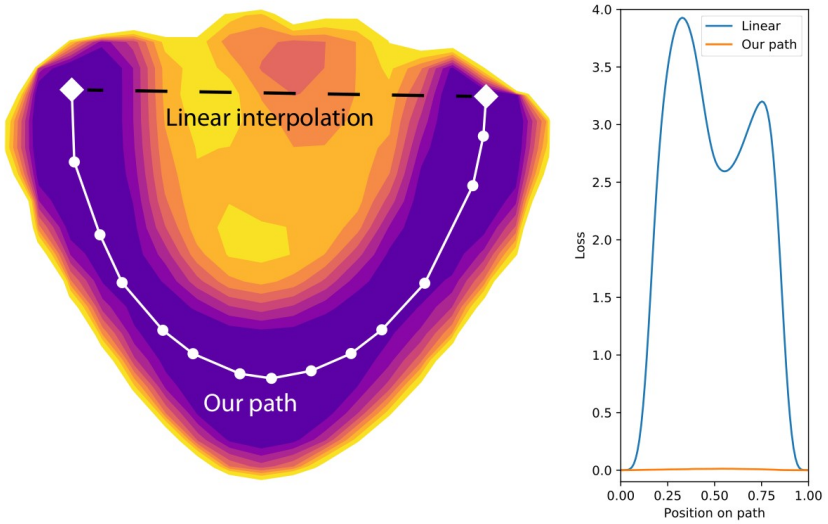




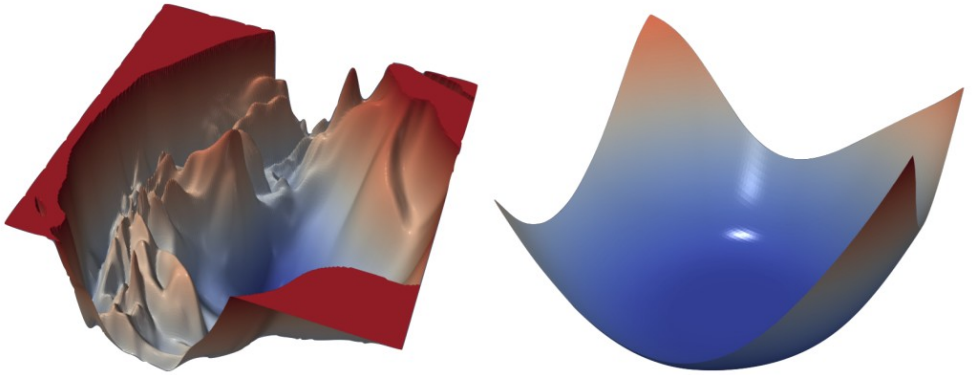
Complex Loss Landscapes for NN

Non-convex landscape, many local and global minima in principle, but...

There seems to be a flat non-convex “bottom” connecting “accessible” minimizers.



From Draxler et al. ‘18, “Essentially no barriers...”



(a) ResNet-110, no skip connections (b) DenseNet, 121 layers

From Li et al. ‘17, “Visualizing...”

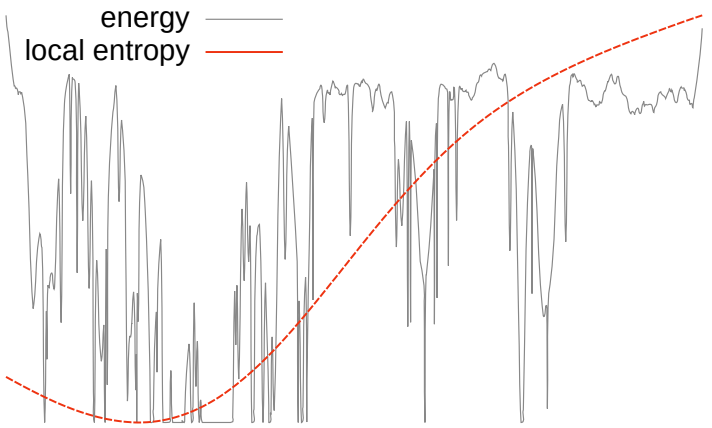
Architectural choices (e.g. loss, activations, batch-norm, skip-connections) influence the roughness and the large-scale structure of the landscape

Training choices (e.g. dropout, SGD batch-size) may lead to minima with different characteristics

Local Entropy loss

- How to tell apart good minima (good at generalizing) from bad minima?
- We (and others) argue **flatness** is a good measure of generalization
- We define the **local entropy loss** as a way to characterize **flatness**, and as a new auxiliary loss

$$\mathcal{L}_{\mathrm{LE}}(W) = -\frac{1}{\beta}\log\int dW'\ e^{-\beta\mathcal{L}(W')-\frac{1}{2}\beta\gamma\|W-W'\|^2}$$



Past analysis of shallow networks

- Using non-rigorous replica theory on shallow neural network models, either with continuous and discrete weights, we consistently observe that miniminizer with **lower local entropy loss** have **better generalization performance** [1,2,3]
- Use of **ReLU activations**, instead of e.g. tanh, induces **wider and flatter minima** [2]
- Same for **cross-entropy** loss compared to error-counting or MSE [3]

[1] Baldassi, Borgs, Chayes, Ingrosso, Lucibello, Saglietti, Zecchina, PNAS ‘16
[2] Baldassi, Malatesta, Zecchina, PRL ‘19
[3] Baldassi, Pittorino, Zecchina, PNAS ‘19

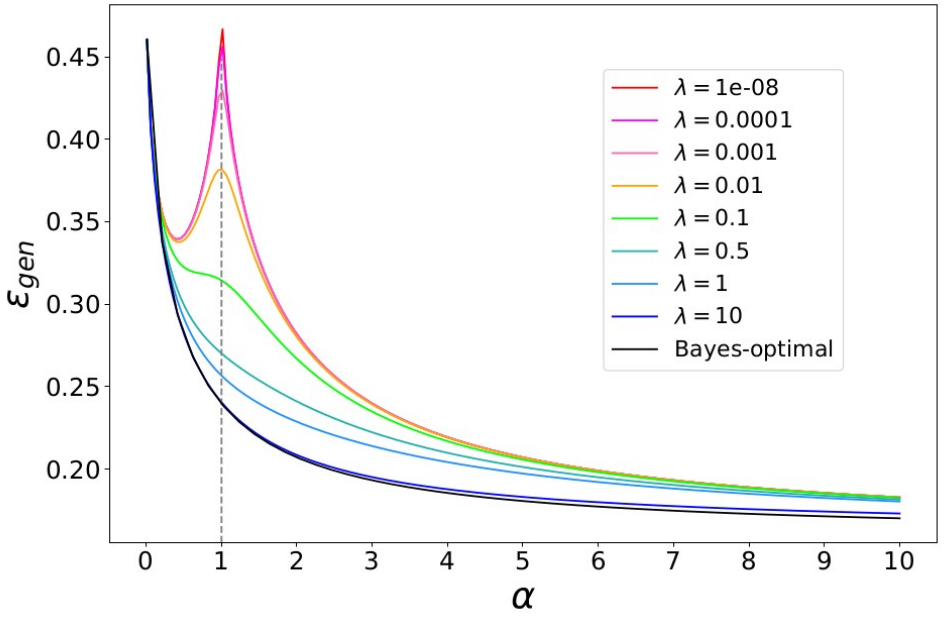
Gaussian Mixtures

Consider a binary mixture of Gaussians, where datapoints for the two classes are generated as a corruption of the true signals $\sigma \mathbf{v}^*, \text{with } \sigma = -1 \text{ and } +1 \text{ respectively}$

$$\mathbf{x}^\mu \sim \mathcal{N}(\sigma^\mu \mathbf{v}^*, \Delta I_N) \qquad \mu = 1, \ldots, \alpha N$$

The student is a perceptron, predicting labels with

$$\hat{\sigma} = sign(\mathbf{w} \cdot \mathbf{x})$$



[1] Mignacco, Krzakala, Lu, Zdeborová 2020

In Ref [1] it was shown that a student minimizing a MSE or a logistic loss, achieves Bayesian optimal error in the limit of infinite L2 regularization.

On the other hand, prediction in this model and in deep nets with ReLU activations are norm independent.

The L2 regularization strategy makes sense only in presence of a differentiable surrogate loss and doesn't seem very general.

Gaussian Mixtures

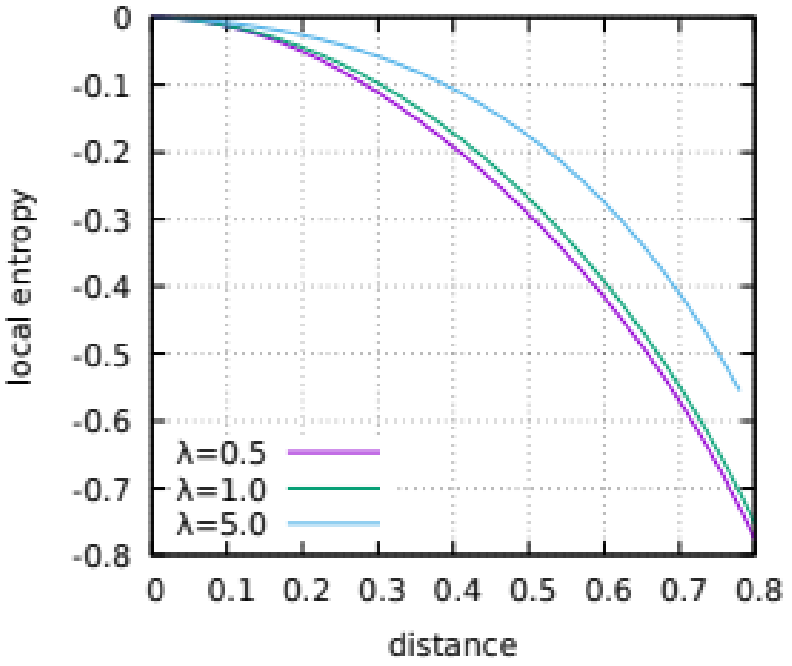
We consider the loss minimizers obtained with different regularization parameter, and rescale them to unit norm.

We consider the normalized local entropy, i.e. the volume fraction around a configuration within distance d and with lower training error

$$\Phi(w,d) = \log \frac{\int_{S_N} dw' \ \theta(\epsilon(w) - \epsilon(w'))\theta(d - \|w - w'\|^2)}{\int_{S_N} dw' \ \theta(d - \|w - w'\|^2)}$$

Once again, we obtain that optimal configurations are characterized by higher and flatter normalized local entropy

We obtained Bayesian optimality by using a more general optimization principle



Algorithm 1: Entropy-SGD

Local Entropy hard to compute

$$\mathcal{L}_{\mathrm{LE}}(W) = -\frac{1}{\beta} \log \int dW' \; e^{-\beta \mathcal{L}(W') - \beta \frac{1}{2} \gamma \|W - W'\|^2}$$

but the gradient

$$\nabla \mathcal{L}_{\mathrm{LE}}(W) = \gamma \langle W - W' \rangle$$

can be approximated by Stochastic Gradient Langevin Dynamics.
The corresponding algorithm is called Entropy-SGD [1]

[1] Chaudhari, Choromanska, Soatto, LeCun, Baldassi, Borgs, Chayes, Sagun, Zecchina, ICLR 2016

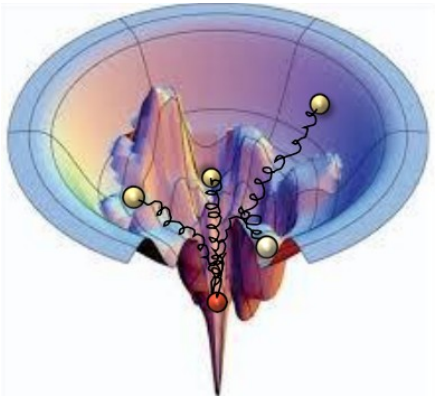
Algorithm 2: Replicated-SGD

Another class of entropic algorithms can be derived starting from

$$p(w) \propto e^{-\beta y \mathcal{L}_{\mathrm{LE}}(w)}$$

For y integer, one can use the local entropy definition
to obtain the statistical measure of a system with y+1 replicas,
then integrate out the original one and obtain

$$p(\{w^a\}_{a=1}^{y}) \propto e^{-\beta \mathcal{L}_{\mathrm{R}}(\{w^a\})}$$



where

$$\mathcal{L}_R(\{w^a\}_a) = \sum_{a=1}^{y} \mathcal{L}(w^a) + \frac{1}{2}\gamma \sum_{a=1}^{y} \|w^a - \bar{w}\|^2,$$

with $\bar{w} = \frac{1}{y}\sum_a w^a$. Now one can perform SGD on the replicated loss.

Deep Learning Experiments

Algorithm 1: Entropy-SGD (eSGD)

Input : w
Hyper-parameters : $L, \eta, \gamma, \eta', \epsilon, \alpha$

for $t = 1, 2, \ldots$ **do**
  $w', \mu \leftarrow w$
  for $l = 1, \ldots, L$ **do**
    $\Xi \leftarrow$ sample minibatch
    $dw' \leftarrow \nabla \mathcal{L}\left(w'; \Xi\right) + \gamma\left(w' - w\right)$
    $w' \leftarrow w' - \eta' dw' + \sqrt{\eta'}\, \epsilon\, \mathcal{N}(0, I)$
    $\mu \leftarrow \alpha\mu + (1 - \alpha)\, w'$
  $w \leftarrow w - \eta\left(w - \mu\right)$

Algorithm 2: Replicated-SGD (rSGD)

Input : $\{w^a\}$
Hyper-parameters : y, η, γ, K

for $t = 1, 2, \ldots$ **do**
  $\bar{w} \leftarrow \frac{1}{y} \sum_{a=1}^{y} w^a$
  for $a = 1, \ldots, y$ **do**
    $\Xi \leftarrow$ sample minibatch
    $dw^a \leftarrow \nabla \mathcal{L}\left(w^a; \Xi\right)$
    if $t = 0 \bmod K$ **then**
      $dw^a \leftarrow dw^a + K\gamma\left(w^a - \bar{w}\right)$
    $w^a \leftarrow w^a - \eta\, dw^a$

Deep Learning Experiments

Dataset	Model	Baseline	rSGD	eSGD	rSGD$\times y$
CIFAR-10	SmallConvNet	16.5 ± 0.2	16 ± 0.1	14.7 ± 0.3	15.5 ± 0.3
	ResNet-18 [28]	13.1 ± 0.3	12.4 ± 0.3	12.1 ± 0.3	11.8 ± 0.1
	ResNet-110 [28]	6.4 ± 0.1	6.2 ± 0.2	6.2 ± 0.1	5.3 ± 0.1
	PyramidNet+ShakeDrop [29, 30]	2.0			1.8
CIFAR-100	PyramidNet+ShakeDrop [29, 30]	13.9			12.7
	EfficientNet-B0 [27]	20.5			19.5
Tiny ImageNet	ResNet-50 [28]	45.2 ± 1.2	41.5 ± 0.3	41.7 ± 1	39.2 ± 0.3
	DenseNet-121 [31]	41.4 ± 0.3	39.8 ± 0.2	38.6 ± 0.4	38.9 ± 0.3

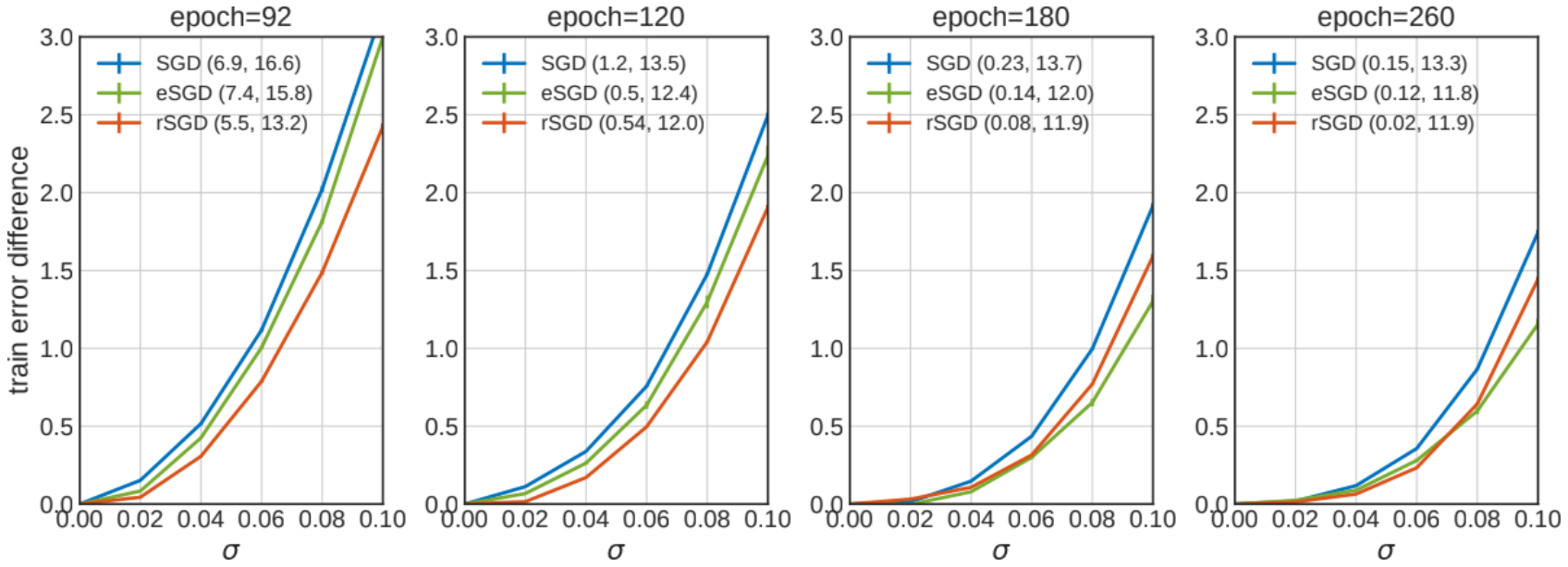
A cheap flatness measure

We want to verify that our entropic algorithm effectively find flatter minima.

Local Entropy is expensive to compute, so we introduce a cheap flatness measure

$$\delta\epsilon(w) = \mathbb{E}_z\, \epsilon(w(1+\sigma z)) - \epsilon(w)$$

Experiment confirm that entropic algorithm find flatter minima (according to this flatness definition) and generalize better



Conclusions and Perspectives

- For Gaussian mixtures classification, we showed that the optimal classificators corresponds to flatter minima in the training error/loss landscape
- For deep networks, we showed entropic algorithms outperform standard ones.
- Work remains to be done to make the algorithms hyper-parameter free
- Application of local entropy framework to devise tighter PAC-Bayesian bounds is under investigation

arXiv:2006.07897 Entropic gradient descent algorithms and wide flat minima

Thank you!