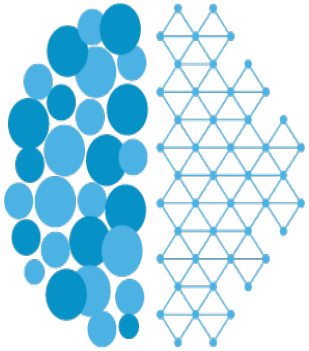
On stochastic gradient descent, flatness and generalization

Yoshua Bengio

July 14, 2018

ICML'2018 Workshop on nonconvex optimization



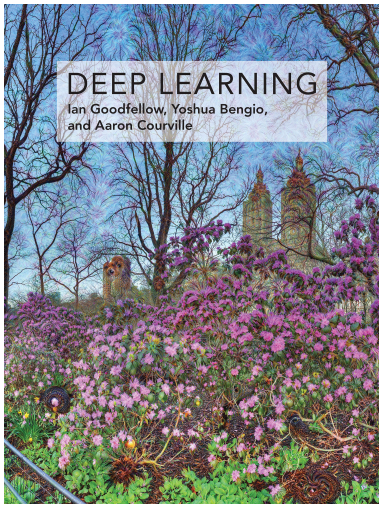
MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

Université de Montréal

IVADO

PLUG: **Deep Learning**, MIT Press, chapters for free online



Disentangling optimization and generalization

- The traditional ML picture is that optimization and generalization are neatly separated aspects
- That makes theory easier to handle, separately
- Unfortunately not the case
- SGD variants influence optimization AND generalization

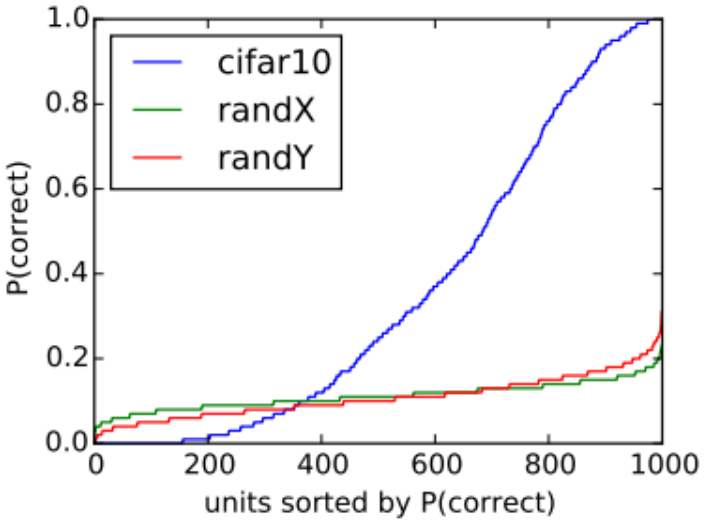
Memorization in Deep Networks

Mostly from preprint arXiv:1706.05394
Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, Simon Lacoste-Julien

Memorization in Deep Networks

- Deep networks trained with SGD generalize well due to its implicit regularization effect (Zhang et al 2016)
- Deep networks achieve ~100% train accuracy on random data (Zhang et al 2016)
- Do deep networks also memorize real data?

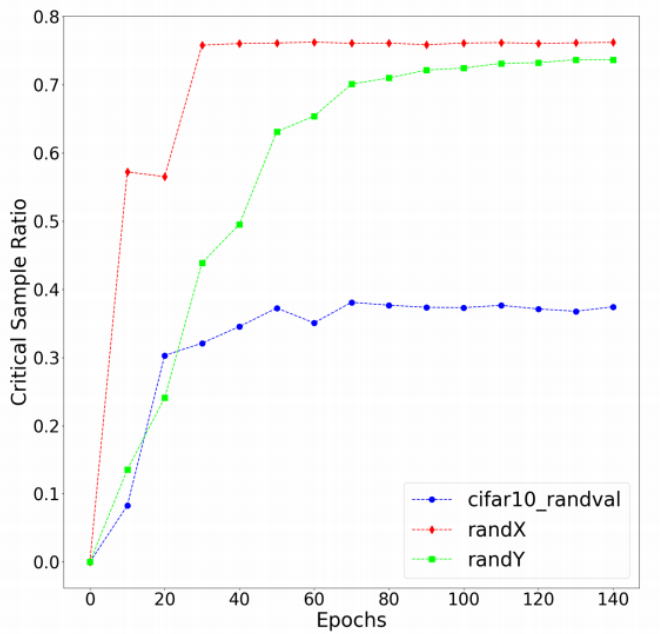
Real data has Dominant Patterns



Fraction of times each of 1000 samples is classified correctly after 1 epoch across 100 runs

- Real data: some samples are learned first.
- Random data: samples are learned in arbitrary order.

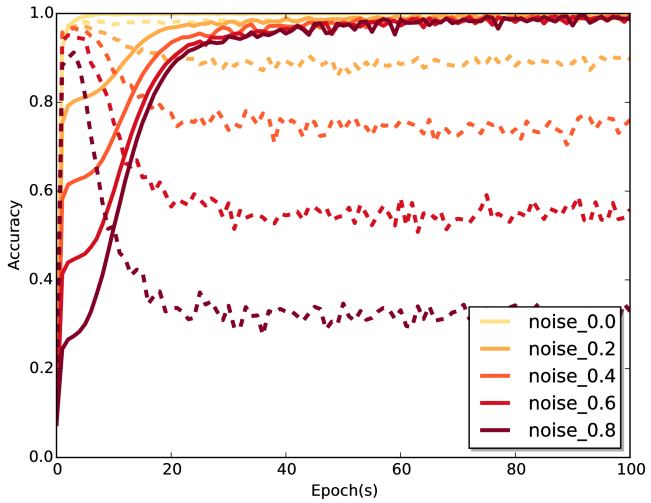
Larger Margin on Real data



Critical sample ratio = fraction of samples which have adversarial examples in their vicinity

- Real data: distance from decision boundary is large
- Random data: distance from decision boundary is small

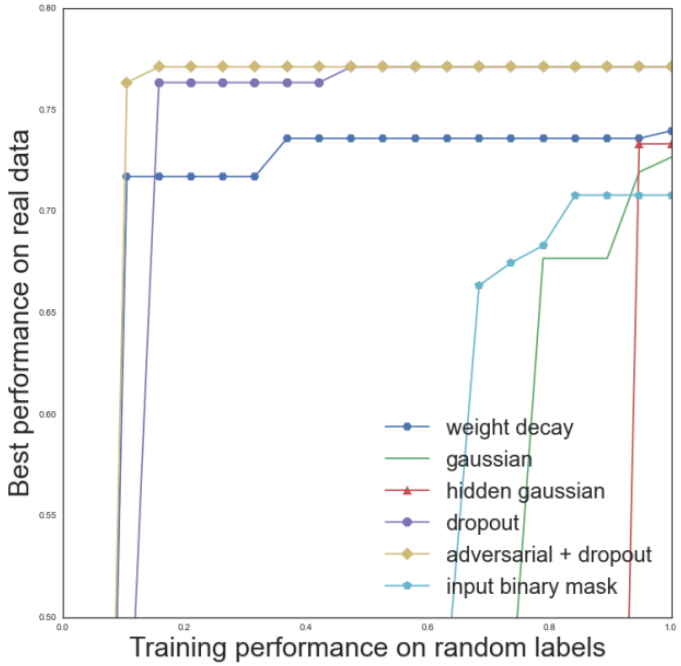
Patterns come First



Train (full) and validation (dotted) accuracy on MNIST during training with noisy labels

- Validation accuracy peaks before falling
- Patterns in real data learned before overfitting noise

Regularization Hinders Memorization



Best validation performance (picked across hyper parameter grid) on real data vs. training performance on noise labels for the same model, for different regularizers.

- Dropout is best at hindering memorization
- Maintains performance on real data for reduced memorization on random data.

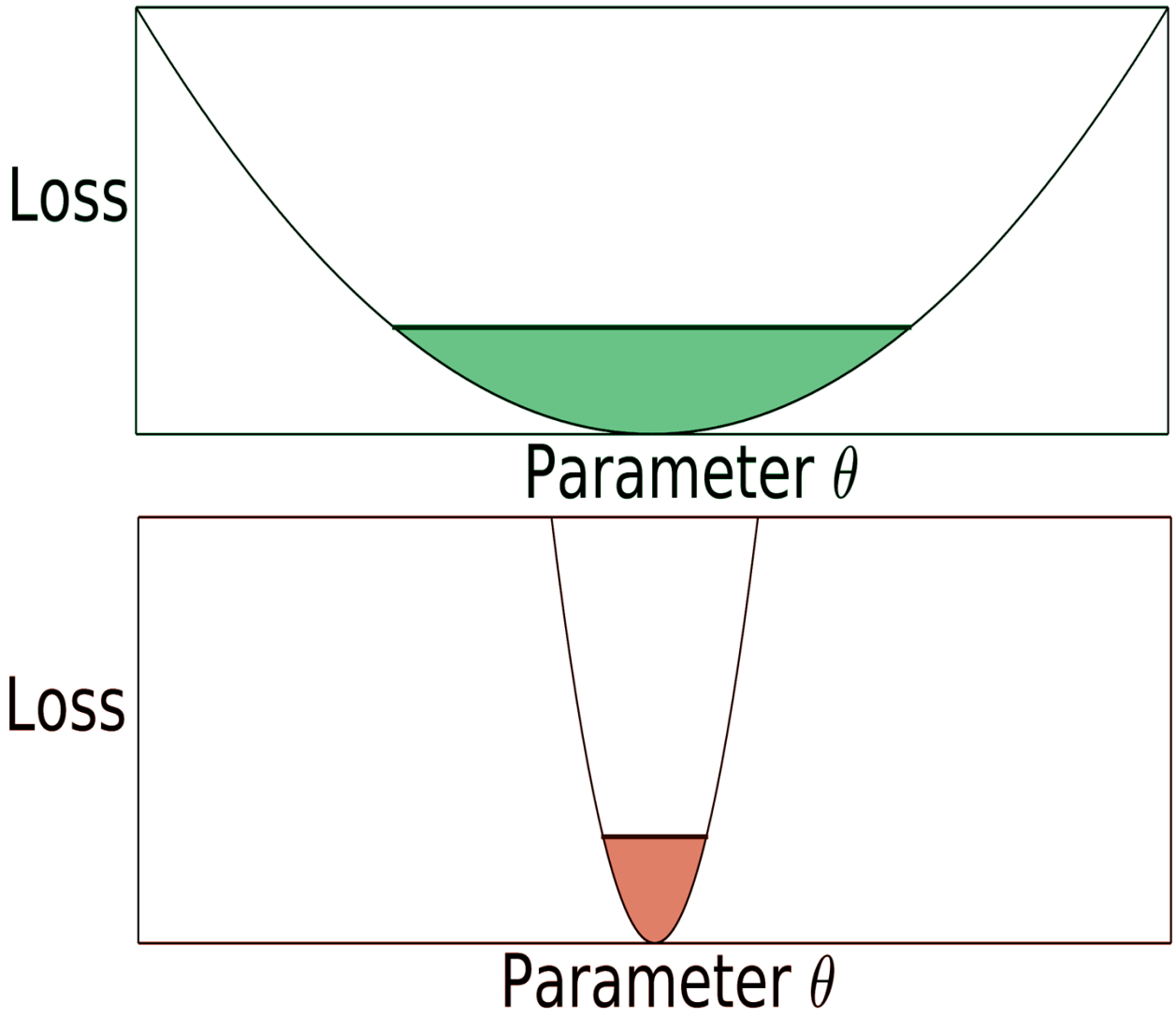
Take Home Message

- **DNNs learn patterns before memorizing noise**
  - **Why?**
  - **Does it have to do with SGD?**

On the relevance of loss function geometry for generalization

Laurent Dinh, Razvan Pascanu,
Samy Bengio, Yoshua Bengio

Flatness



Reparametrization

$$\eta = g^{-1}(\theta) \qquad L_\eta(\eta) = L\left(g(\eta)\right)$$

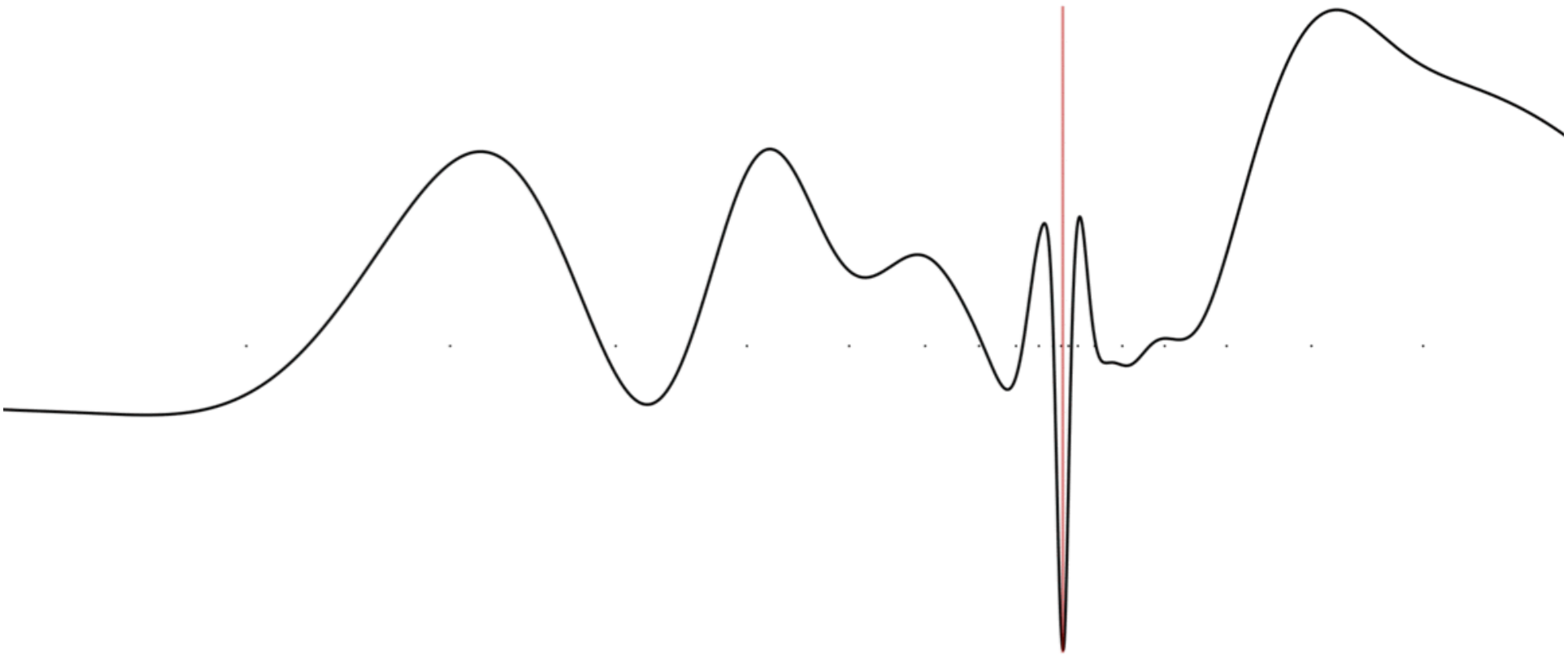
-
- Differentiation at critical point

$$(\nabla^2 L_\eta)(\eta) = (\nabla g)(\eta)^T (\nabla^2 L)\left(g(\eta)\right)(\nabla g)(\eta)$$

- Flat minima $\xrightarrow{g}$ Sharp minima

  Sharp minima $\xrightarrow{g}$ Flat minima

Reparametrization



Sharp minima can generalize
Flat minima can poorly generalize

Eppur, si muove!

And yet, it generalizes!

Factors influencing Minima in SGD

Mostly from preprint arXiv:1711.04623

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, Amos Storkey

Behavior of SGD

- Small mini-batch finds wider minima (Keskar et al 2016)
- What dynamics/factors govern the quality of minima found by SGD?

SGD as Stochastic Differential Equation

- Mini-batch gradient $\mathbf{g}^{(S)}(\boldsymbol{\theta})$ (due to CLT), batch size S:

- SGD with learning rate η is described by:

$$\mathbf{g}^{(S)}(\boldsymbol{\theta}) = \underbrace{\mathbf{g}(\boldsymbol{\theta})}_{\frac{\partial L(\theta)}{\partial \theta}} + \frac{1}{\sqrt{S}}\Delta\mathbf{g}(\boldsymbol{\theta}), \text{ where } \Delta\mathbf{g}(\boldsymbol{\theta}) \sim N(0, \mathbf{C}(\boldsymbol{\theta}))$$

- Continuous stochastic differential equation (SDE) form: (Li et al 2017)

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \boldsymbol{g}^{(S)}(\boldsymbol{\theta})$$

$$\frac{d\boldsymbol{\theta}}{dt} = -\eta \boldsymbol{g}(\boldsymbol{\theta}) + \frac{\eta}{\sqrt{S}}\mathbf{B}(\boldsymbol{\theta})\mathbf{f}(t)$$

If small enough learning rate, ie. small steps

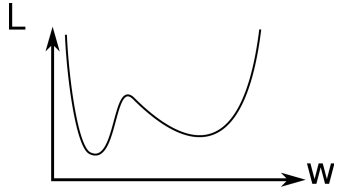
Note: $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{B}(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{B}(\boldsymbol{\theta})$

Equilibrium Distribution of SGD

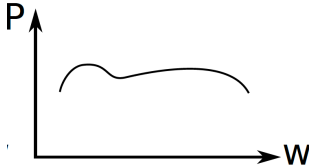
- The equilibrium distribution of this SDE is given by:
- ~Inverse relation between loss and density

$$P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2L(\boldsymbol{\theta})}{n\sigma^2}\right)$$

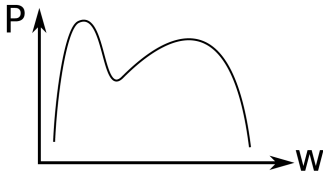
- Noise n controls the granularity of the equilibrium distribution



$n = \eta/S$



n=10

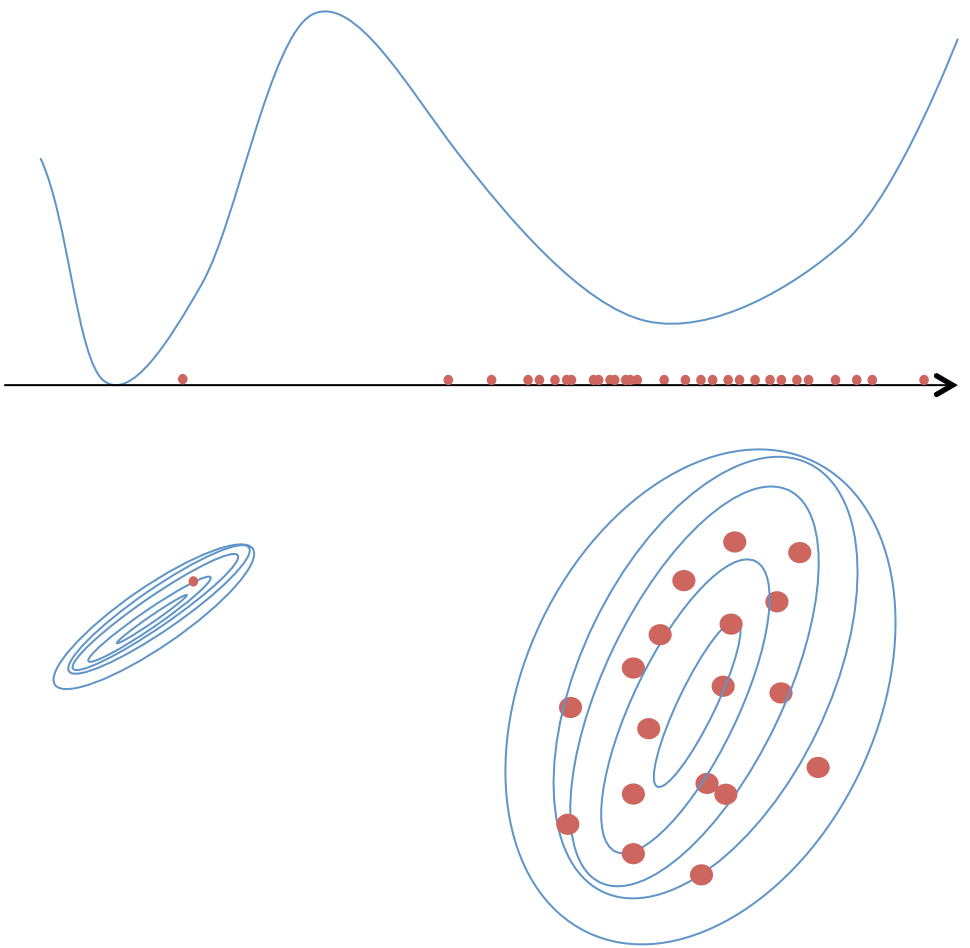


n=1

Note: η = learning rate, S = batch size, σ^2 = fixed isotropic gradient variance

SGD Moves a Cloud of Points

- Consider the last *k* values of θ
- Form a cloud of points
- The cloud gradually moves with SGD updates
- The width of the cloud grows with the noise level (l.rate/BS)
- It cannot go in valleys sharper than that width



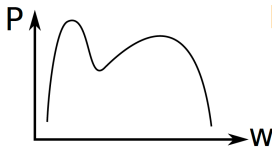
Implications of the Theory

- Probability of ending in a minima A described by Hessian $\mathbf{H}_A$:

$$p_A \propto \frac{1}{\sqrt{\det \mathbf{H}_A}} \exp\left(-\frac{2}{n\sigma^2} L_A\right)$$

- In general, minima with larger volume is favored more (simply because it has higher probability mass)
- Higher noise n prioritizes width (volume) over depth
- Final equilibrium distribution is unchanged when learning rate and batch size are scaled proportionally $\eta \to \beta\eta$, $S \to \beta S$

$$P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2L(\boldsymbol{\theta})}{n\sigma^2}\right)$$



Note: n = η/S, η = learning rate, S = batch size, σ^2 = fixed isotropic gradient variance

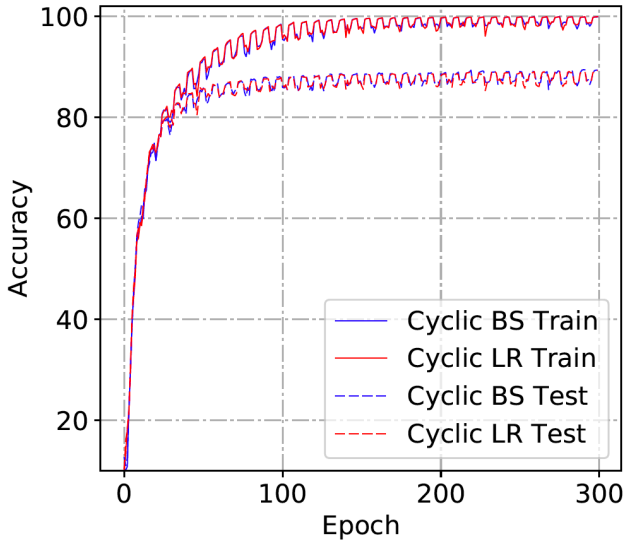
Experimental Results

Smaller Noise –
Sharper Bowl

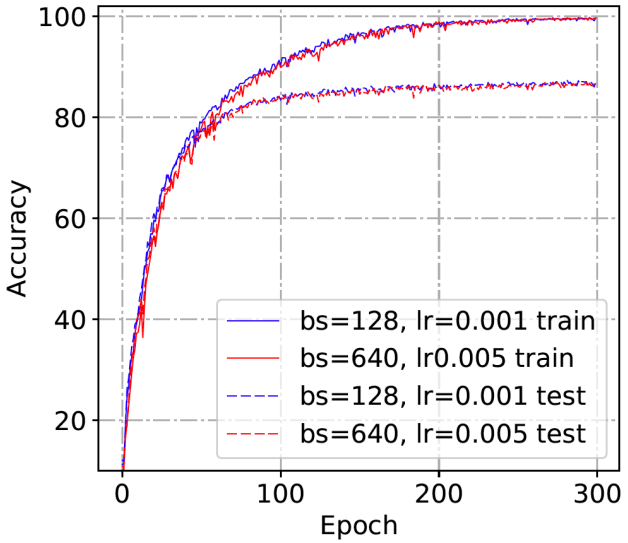
Equal noise –
Equal Width

Same Noise - Same Learning Dynamics

- Theory talks about final equilibrium distribution but seems to apply along trajectory as well
- But even learning dynamics is similar when learning rate and batch size are scaled proportionally $\eta \to \beta\eta$, $S \to \beta S$



Cyclic Learning Rate and Cyclic Batch-size



Constant Learning Rate and Constant Batch-size

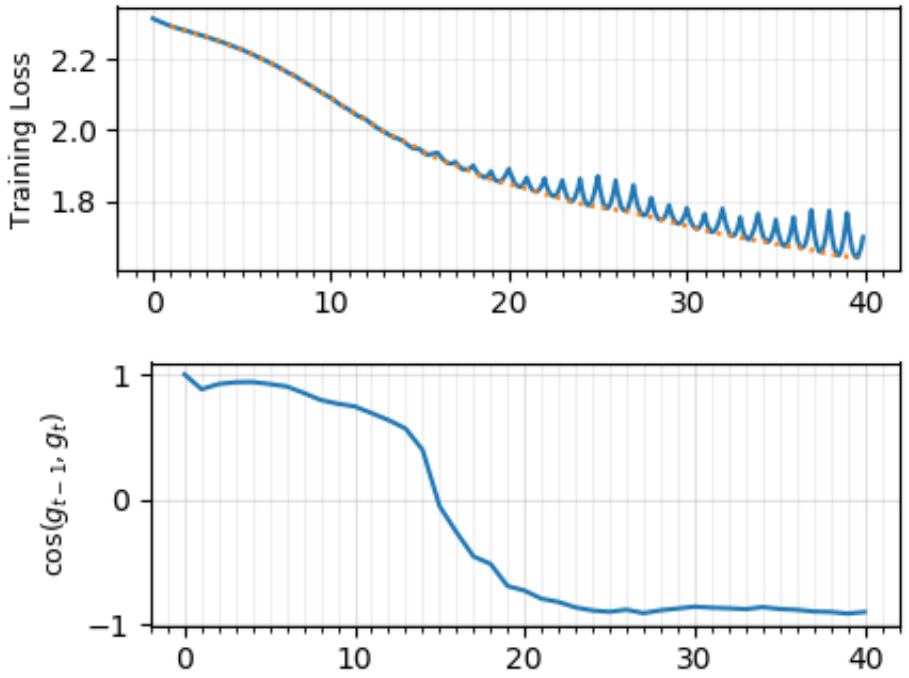
Take Home Messages

- DNNs learn patterns before memorizing noise
- Regularization hinders memorization
- The quality of final minima and learning dynamics is similar when learning rate and batch size are scaled proportionally
- Larger noise favors large volume minima over deep ones
- Larger noise (e.g. due to BS or l.rate) hinders memorization

A Walk with SGD

Xing, Arpit, Tsirigotis & Bengio ArXiv:1802.08770

- Interpolate in parameter space between minibatch SGD updates and see convex shape
- After initial phase, updates bounce off valley floor, which monotonically improves, traversing larger distances with smaller batch sizes (BS)
- Learning rate: height from floor
- BS: exploration noise
- Pure GD gets stuck on floor, while SGD finds flatter regions, which generalize better



Sharpest Directions Along the SGD Trajectory
(Jastrzębski, Kenton, Ballas, Fischer, Bengio, Storkey)

- Even at the beginning of training, a high learning rate or small batch size influences SGD to visit flatter loss regions.
- the largest eigenvalues appears to always follow a similar pattern, with a fast increase in the early phase and a decrease thereafter, where the peak value is determined by the learning rate and batch size.
- altering the learning rate just in the direction of the eigenvectors associated with the largest eigenvalues, SGD can be steered towards regions which are an order of magnitude sharper but correspond to models with similar generalization, **confirming that curvature of the endpoint found by SGD is not predictive of its generalization properties**.

Montreal Institute for Learning Algorithms

MILA

Université de Montréal