Three Factors Influencing Minima in SGD

Stanisław Jastrzębski*[1], Zachary Kenton*[2],
Devansh Arpit[2], Nicolas Ballas[3], Asja Fischer[4],
Yoshua Bengio[2], Amos Storkey[5]

*First two authors contributed equally

[1]Jagiellonian University

[2]MILA, Université de Montréal

[3]Facebook AI Research

[4]University of Bonn

[5]University of Edinburgh

17/11/17

Outline

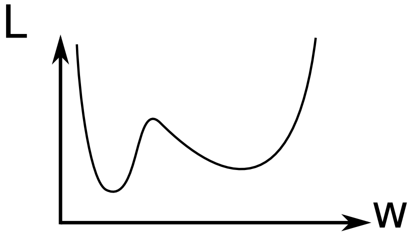
Outline

Motivation

- Why does SGD work so well?
- Originally stochastic part was for computational reasons
- Now view stochastic part as implicit regularization
- Stochastic part essential for good generalization
- Parallelization
- Generalization?

Introduction

- We study the properties of the endpoint of SGD
- Approximate SGD as a stochastic differential equation
- 3 factors control the trade-off between the depth and width of endpoint regions targeted by SGD:
  - Learning rate, η
  - Batch size, S
  - Variance of the loss gradients



Introduction

- Only ratio η/S appears:
  - Higher η/S leads to wider regions
  - Endpoint invariant under rescaling of η and S by same amount
- Experiments are consistent with this

Outline

Theory - Approximate SGD with SDE

- Stochastic gradient, assume by CLT

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

Theory - Approximate SGD with SDE

- Stochastic gradient, assume by CLT

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

Note symmetric positive-semidefinite $\mathsf{C}(\boldsymbol{\theta}) = \mathsf{B}(\boldsymbol{\theta})\mathsf{B}^\top(\boldsymbol{\theta})$.

Theory - Approximate SGD with SDE

- Stochastic gradient, assume by CLT
$$\mathbf{g}^{(S)}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) + \frac{1}{\sqrt{S}}\Delta\mathbf{g}(\boldsymbol{\theta}), \text{ where } \Delta\mathbf{g}(\boldsymbol{\theta}) \sim \mathcal{N}(0, \mathsf{C}(\boldsymbol{\theta}))$$
Note symmetric positive-semidefinite $\mathsf{C}(\boldsymbol{\theta}) = \mathsf{B}(\boldsymbol{\theta})\mathsf{B}^{\top}(\boldsymbol{\theta})$.
- SGD update:
$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta\boldsymbol{g}^{(S)}(\boldsymbol{\theta})$$

Theory - Approximate SGD with SDE

- Stochastic gradient, assume by CLT

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

  Note symmetric positive-semidefinite $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{B}(\boldsymbol{\theta})\mathbf{B}^\top(\boldsymbol{\theta})$.

- SGD update:

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

- Approximate SGD by Stochastic Differential Equation (SDE):

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

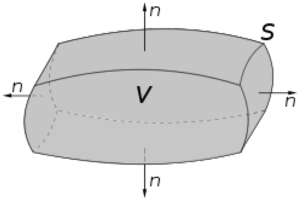
  where $\mathbf{f}(t)$ is a normalized Gaussian time-dependent stochastic term.

Fokker-Planck Equation

- Trade SDE for a deterministic PDE for the distribution $P(\boldsymbol{\theta}, t)$, called Fokker-Planck equation

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{J}$$

$$\text{where } \boldsymbol{J} \equiv \eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot (\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t))$$



Fokker-Planck Equation

- Trade SDE for a deterministic PDE for the distribution $P(\boldsymbol{\theta}, t)$, called Fokker-Planck equation

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{J}$$

$$\text{where } \boldsymbol{J} \equiv \eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot (\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t))$$

- Fokker-Planck common in physics, e.g. diffusion under external force

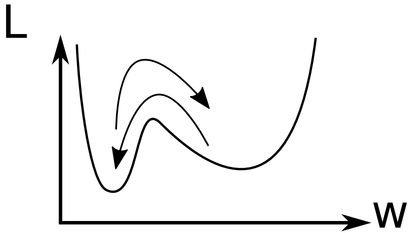
Fokker-Planck Equation

- Trade SDE for a deterministic PDE for the distribution $P(\boldsymbol{\theta}, t)$, called Fokker-Planck equation

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{J}$$

$$\text{where } \boldsymbol{J} \equiv \eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot (\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t))$$

- Fokker-Planck common in physics, e.g. diffusion under external force
- Hard to solve in general, make some simplifying assumptions:
  - Endpoint of SGD: Equilibrium/Detailed Balance: $\boldsymbol{J} = 0$.



Fokker-Planck Equation

- Trade SDE for a deterministic PDE for the distribution $P(\boldsymbol{\theta}, t)$, called Fokker-Planck equation

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{J}$$

$$\text{where } \boldsymbol{J} \equiv \eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot (\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t))$$

- Fokker-Planck common in physics, e.g. diffusion under external force
- Hard to solve in general, make some simplifying assumptions:
  - Endpoint of SGD: Equilibrium/Detailed Balance: $\boldsymbol{J} = 0$.
  - Isotropic variance: $\mathbf{C}(\boldsymbol{\theta}) = \sigma^2 \mathbf{I}$

Fokker-Planck Equation

- Trade SDE for a deterministic PDE for the distribution $P(\boldsymbol{\theta}, t)$, called Fokker-Planck equation

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{J}$$

$$\text{where } \boldsymbol{J} \equiv \eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot (\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t))$$

- Fokker-Planck common in physics, e.g. diffusion under external force
- Hard to solve in general, make some simplifying assumptions:
  - Endpoint of SGD: Equilibrium/Detailed Balance: $\boldsymbol{J} = 0$.
  - Isotropic variance: $\mathbf{C}(\boldsymbol{\theta}) = \sigma^2 \mathbf{I}$
- Then solution is a Boltzmann-Gibbs distribution:

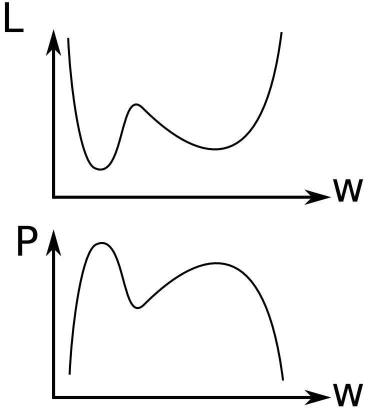
$$P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2}{\sigma^2} \frac{1}{n} L(\boldsymbol{\theta})\right)$$

where $L(\boldsymbol{\theta})$ is the loss and the noise $n \equiv \eta / S$.

Boltzmann-Gibbs distribution

$$P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2}{\sigma^2}\frac{1}{n}L(\boldsymbol{\theta})\right)$$

$$n \equiv \eta/S$$

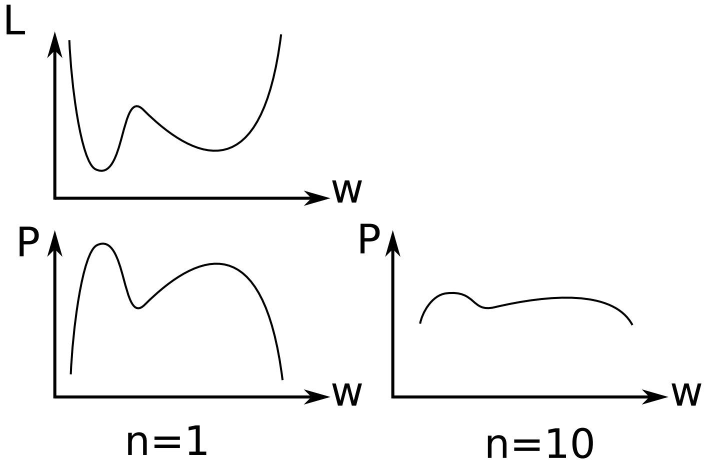


n=1

Boltzmann-Gibbs distribution

$$P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2}{\sigma^2}\frac{1}{n}L(\boldsymbol{\theta})\right)$$

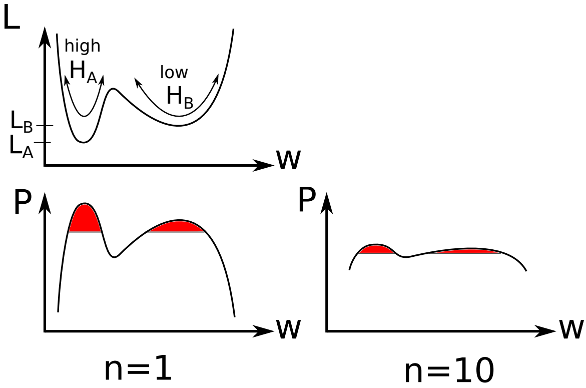
$$n \equiv \eta/S$$



Probability ending in a region near $\boldsymbol{\theta}_A$

- Loss has Hessian $\mathbf{H}_A$ and loss L_A at a minimum $\boldsymbol{\theta}_A$. Probability of ending in region near $\boldsymbol{\theta}_A$

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



Probability ending in a region near $\boldsymbol{\theta}_A$

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

$$n \equiv \eta / S$$

Probability ending in a region near $\boldsymbol{\theta}_A$

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

- 3 factors control trade-off between depth and width:
  - learning rate η
  - batch-size S
  - covariance of the gradients σ^2

Probability ending in a region near $\boldsymbol{\theta}_A$

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

- 3 factors control trade-off between depth and width:
  - learning rate η
  - batch-size S
  - covariance of the gradients σ^2
- The two factors that we directly control only appear in the ratio given by the *noise*, $n = \eta/S$.
  - higher n gives more priority to width and less priority to depth
  - Invariance under simultaneous rescaling $\eta \mapsto c\eta$ and $S \mapsto cS$

Take-Away Theory

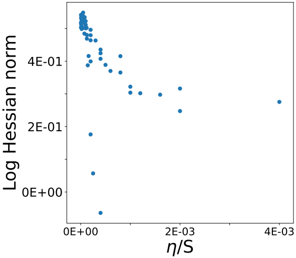
- noise $n = \eta/S$
- Higher n gives more priority to width and less priority to depth
- Invariance under simultaneous rescaling $\eta \mapsto c\eta$ and $S \mapsto cS$

Outline

Outline

Width and n

4-layer Batch Normalized ReLU MLP trained on Fashion-MNIST.
Noise $n = \eta/S$

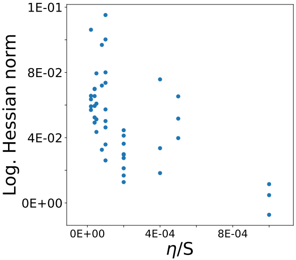


(a) Correlation of $\frac{\eta}{S}$ with logarithm of norm of Hessian.

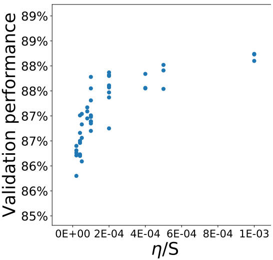
(b) Correlation of $\frac{\eta}{S}$ with validation accuracy.

Width and n

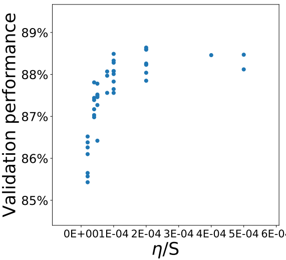
20 layer ReLU MLP without Batch Normalization on FashionMNIST



(a) Good initialization.



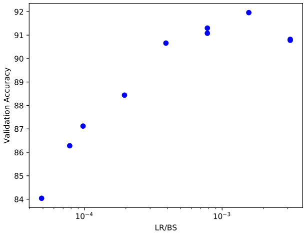
(b) Good initialization.



(c) Bad initialization.

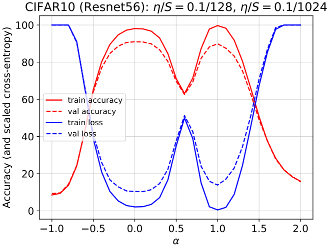
WARNING

- Resnet56 networks on CIFAR10 with different n.
- See there is a peak n for best validation accuracy.
- Higher learning rate to batch-size ratio doesn't necessarily lead to higher validation accuracy

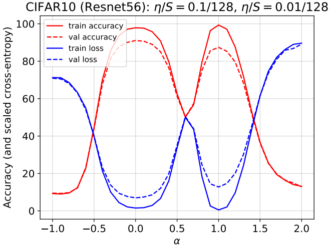


Width and n

- Interpolation of Resnet56 on CIFAR10 for different $n = \frac{\eta}{S}$.
- x-axis, α, corresponds to the interpolation coefficient.
- Consistent with our theory, lower $\frac{\eta}{S}$ ratio leads to sharper minima.



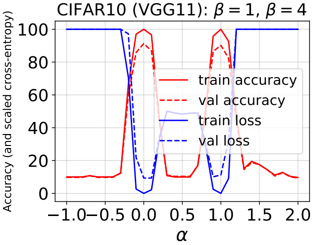
(a) left $\frac{\eta=\mathbf{0.1}}{S=\mathbf{128}}$, right $\frac{\eta=\mathbf{0.1}}{S=\mathbf{1024}}$.



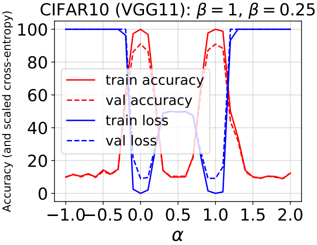
(b) left $\frac{\eta=\mathbf{0.1}}{S=\mathbf{128}}$, right $\frac{\eta=\mathbf{0.01}}{S=\mathbf{128}}$.

Width and n

- Interpolation of VGG-11 on CIFAR10 for same $n = \frac{\eta}{S}$.
- $\frac{\eta = 0.1\times\beta}{S = 50\times\beta}$
- Consistent with theory: approx same n, see approx same width



(a) left: $\beta = 1$, right: $\beta = 4$

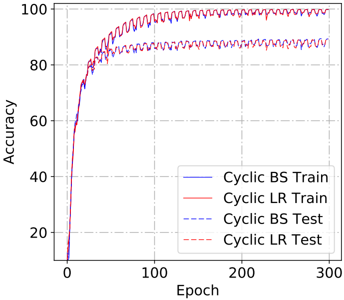


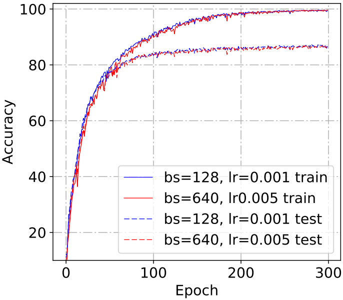
(b) left: $\beta = 1$, right: $\beta = 0.25$

Outline

Simultaneous rescaling $\eta \mapsto c\eta$ and $S \mapsto cS$

- VGG-11 architecture on CIFAR10
- Left: Cyclic learning rate exchanged for cyclic batch size.
- Right: Constant learning rate, batch size.
- Not just endpoint, but also dynamics are approx invariant



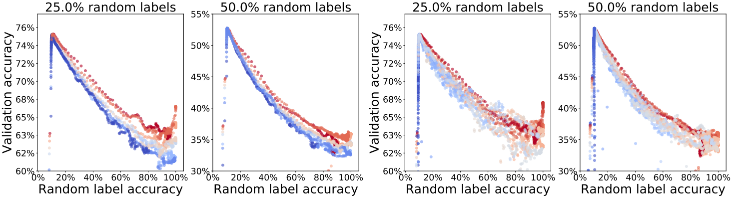


Outline

Memorization and n

Memorization: add random labels, see effect on validation accuracy

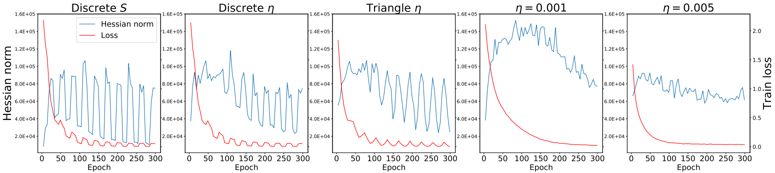
- MLP, 2-layer, each 256 hidden units, ReLU
- Higher value of $n = \frac{\eta}{S}$ is redder
- Left two: momentum with parameter 0.9
- Right two: no momentum
- Specific level of memorization, high n better generalization



Outline

Cyclic Schedules

- VGG-11 on CIFAR10
- Oscillate between sharp and wide regions
- Cyclic find wider minima than baseline

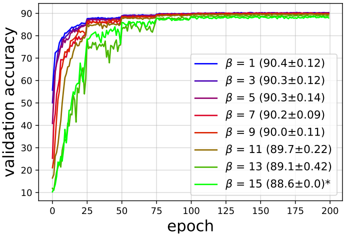


	Test acc	Valid acc	Loss	H. norm.
Discrete η	90.04% $\pm$ 0.18%	90.30% $\pm$ 0.07%	0.048 $\pm$ 0.001	36470
Discrete S	90.07% $\pm$ 0.32%	90.25% $\pm$ 0.06%	0.050 $\pm$ 0.002	13918
Triangle η	90.03% $\pm$ 0.10%	90.04% $\pm$ 0.23%	0.068 $\pm$ 0.002	35310
Baseline	87.70% $\pm$ 0.56%	88.36% $\pm$ 0.13%	0.033 $\pm$ 0.001	57838

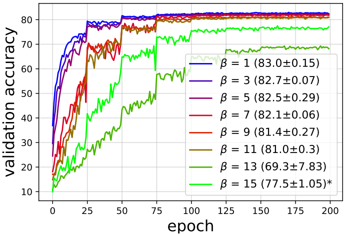
Outline

Breaking down of theory

- VGG-11 on CIFAR10
- Ratio $\frac{\eta=\beta\times 0.1}{S=\beta\times 50}$ fixed
- Break down for large β. Earlier for smaller train size.



(a) Train dataset size 45000



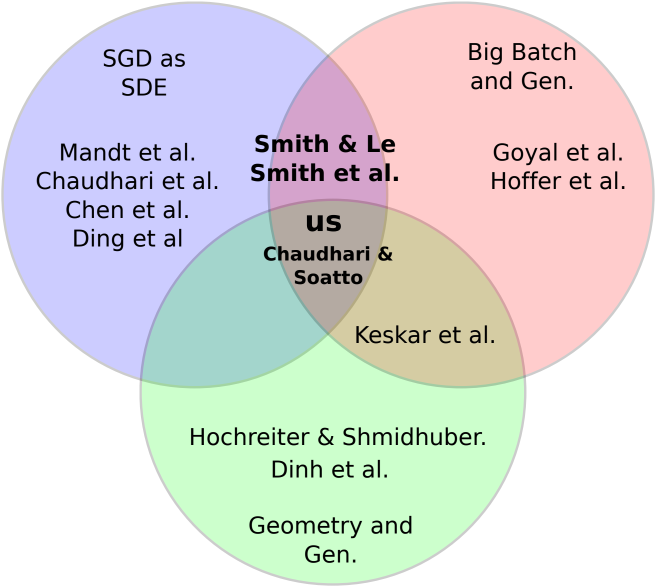
(b) Train dataset size 12000

Outline

Conclusion

- SGD endpoint (from theory, experiments) and dynamics (from experiment) depend on learning rate and batch size through noise $n = \eta/S$
- Higher n gives more priority to width and less priority to depth
- Invariance under simultaneous rescaling $\eta \mapsto c\eta$ and $S \mapsto cS$

Related Work



Discussion

- Isotropic variance
- Dynamics
- Causal links
- Cyclic schedules
- Superconvergence
- Optimality
- CLT assumption is for i.i.d.

Conclusion

- SGD endpoint (from theory, experiments) and dynamics (from experiment) depend on learning rate and batch size through noise $n = \eta/S$
- Higher n gives more priority to width and less priority to depth
- Invariance under simultaneous rescaling $\eta \mapsto c\eta$ and $S \mapsto cS$

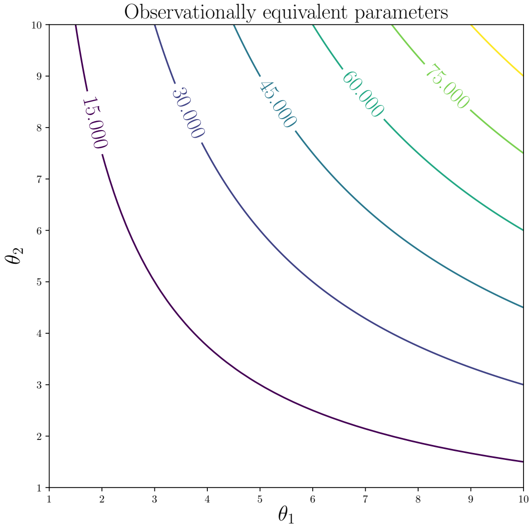
Ratio of Probabilities

$$\frac{p_A}{p_B} = \sqrt{\frac{\det \mathbf{H}_B}{\det \mathbf{H}_A}} \exp\left(\frac{2}{n\sigma^2}\left(L_B - L_A\right)\right)$$

This ratio is invariant to reparametrization

Gradient Descent and Reparametrization

Consider reparametrization of Dinh et al. $(\theta_1, \theta_2) \mapsto (\alpha\theta_1, \alpha^{-1}\theta_2)$



Gradient Descent and Reparametrization

- Gradient Descent only samples one of the equivalent reparametrizations
- Along these GD paths the flatness from Hessians is meaningful