清華大學
Tsinghua University

Asymmetric Valleys: Beyond Sharp and Flat Local Minima

Haowei He | Gao Huang | Yang Yuan

IIIS | Department of Automation, Tsinghua University

October 26, 2019

Background: Local Landscape

A good understanding of local landscape is important for
1. Designing better optimization method
2. Explaining when and how a deep network achieves good generalization performance

Previous Work & Our Proposal

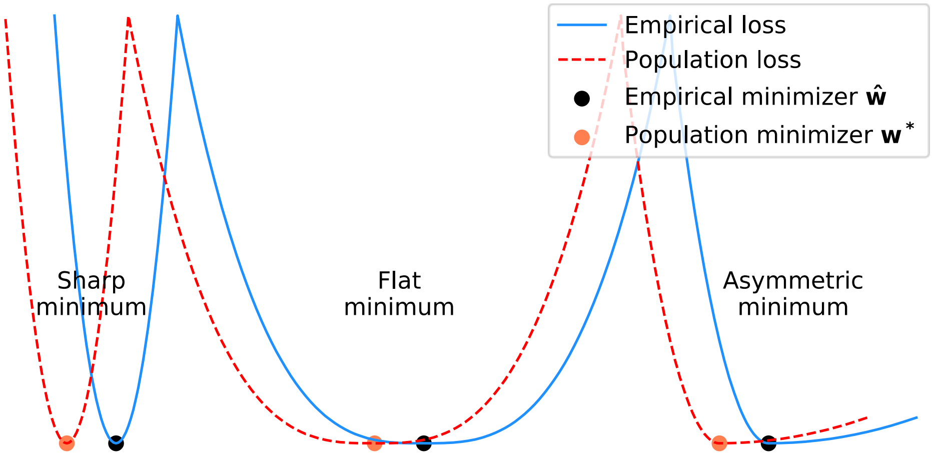
So what is a good local landscape for generalization?

1. Sharp or Flat

Flat minimum generalize better.*

2. Re-parameterization

3. Asymmetric Valley



*On large-batch training for deep learning: Generalization gap and sharp minima. ICLR, 2017.

Previous Work & Our Proposal

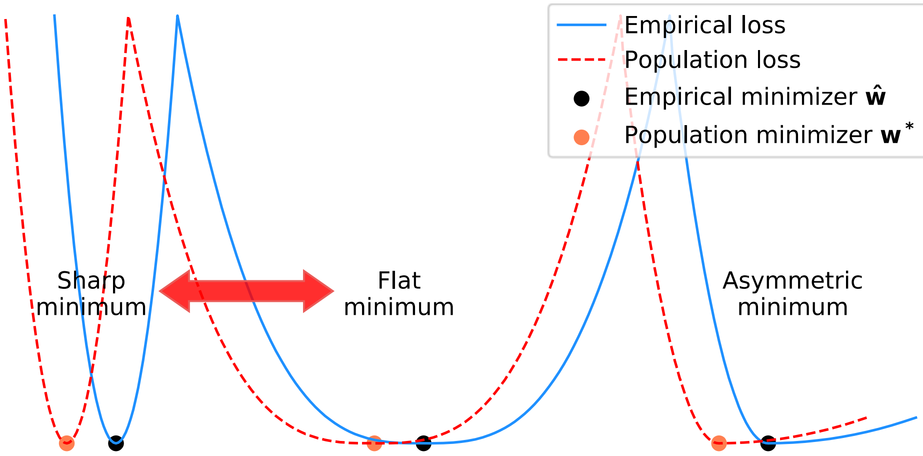
So what is a good local landscape for generalization?

1. Sharp or Flat

2. Re-parameterization

Flat and sharp minimum can convert to each other.*

3. Asymmetric Valley



*Sharp minima can generalize for deep nets. ICML, 2017.

Previous Work & Our Proposal

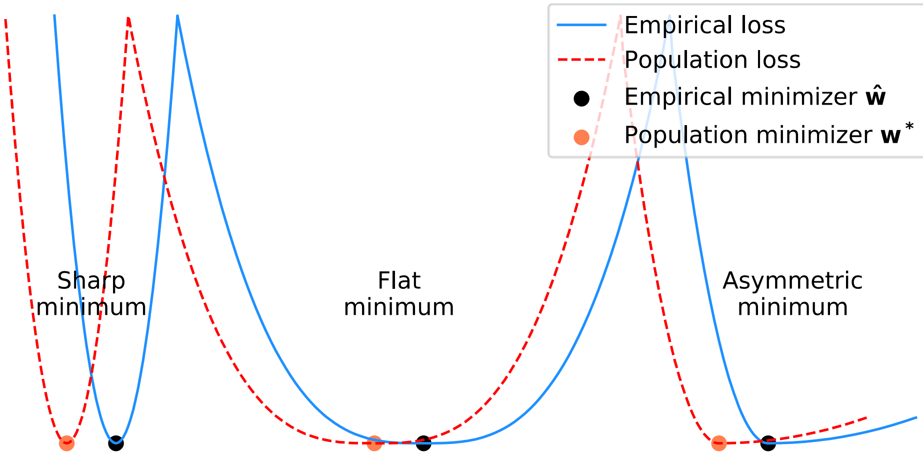
So what is a good local landscape for generalization?

1. Sharp or Flat
2. Re-parameterization
3. Asymmetric Valley

Our work.*

Minimum with asymmetric direction.

Biased solution generalize better.

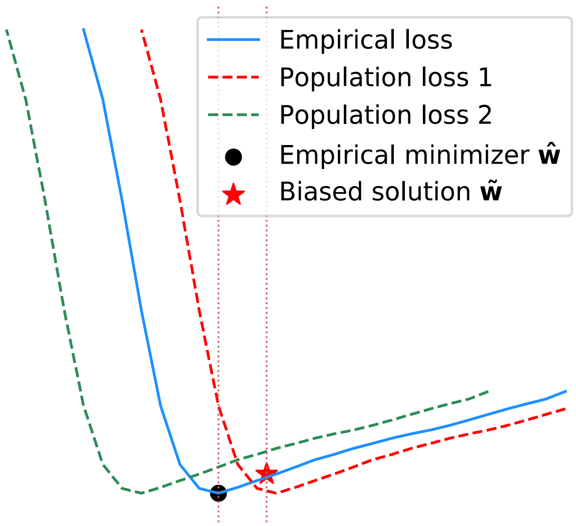


* Asymmetric Valleys: Beyond sharp and flat local minimum. NeurIPS, 2019.

What is asymmetric valley?

Intuitively illustration:

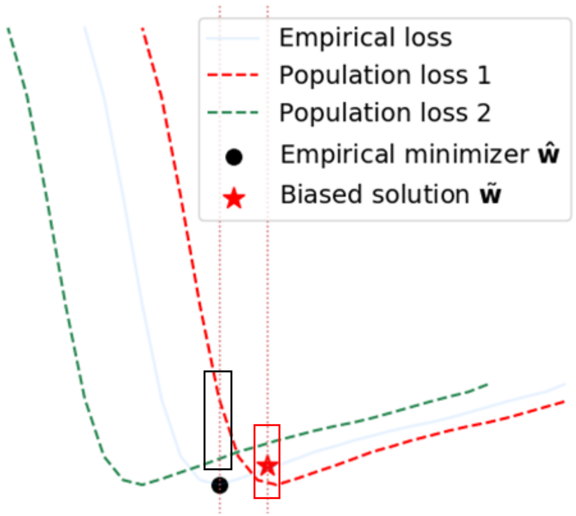
1. Loss grows fast on one side and slowly on the other side.
2. With a random shift between the empirical and population loss, the red star solution has lower population loss in expectation.



What is asymmetric valley?

Intuitively illustration:

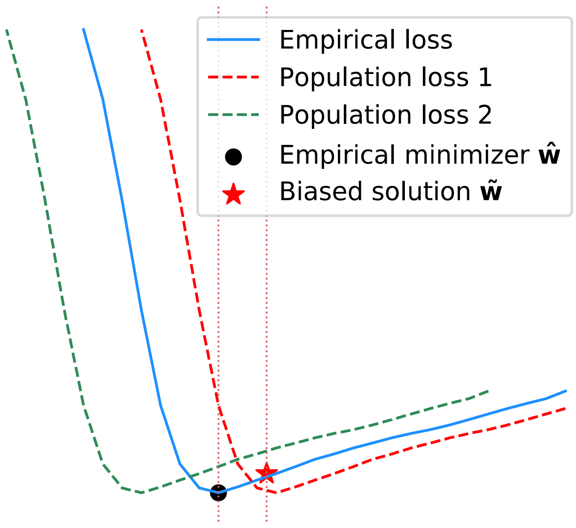
1. Loss grows fast on one side and slowly on the other side.
2. With a random shift between the empirical and population loss, the red star solution has lower population loss in expectation.



What is asymmetric valley?

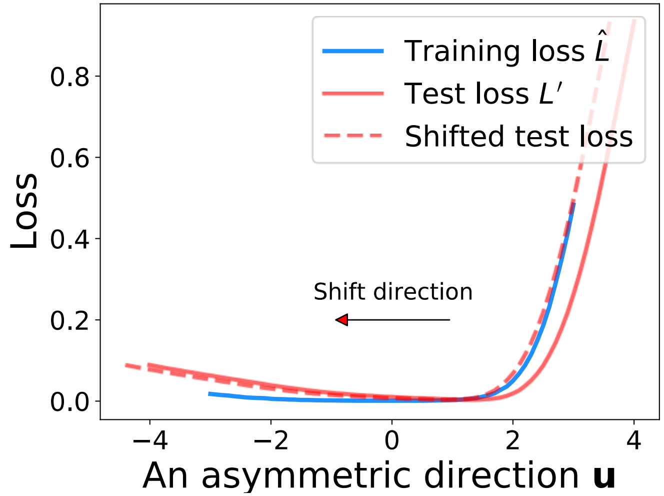
Two interesting implications:

1. converging to *which* local minimum may not be critical. However, it matters a lot *where* the solution locates.
2. the solution with lowest *a priori* generalization error is not necessarily the minimizer of the training loss.

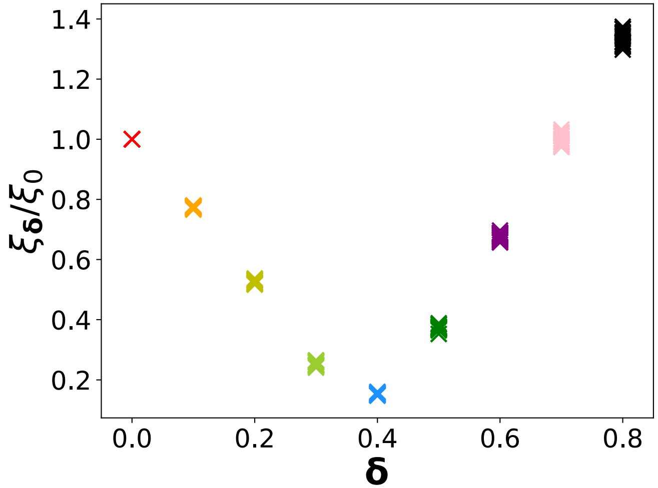


Assumptions and Verification

Random shift assumption: assume a random symmetric horizontal shift between training loss and test loss with 0 expectation.



red dashed line has the best match



δ : relative shift

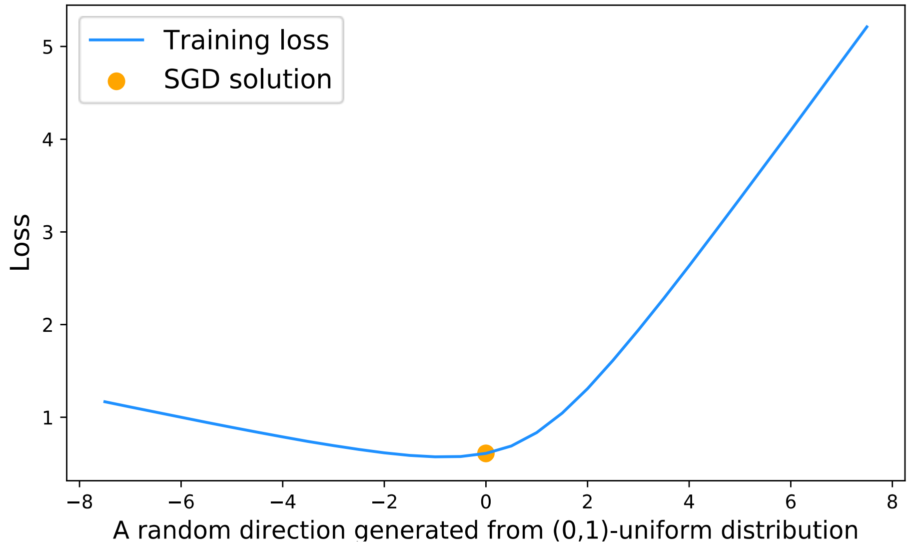
ξ_δ: loss difference with δ shift

Assumptions and Verification

Locally asymmetric assumption: assume locally asymmetric property

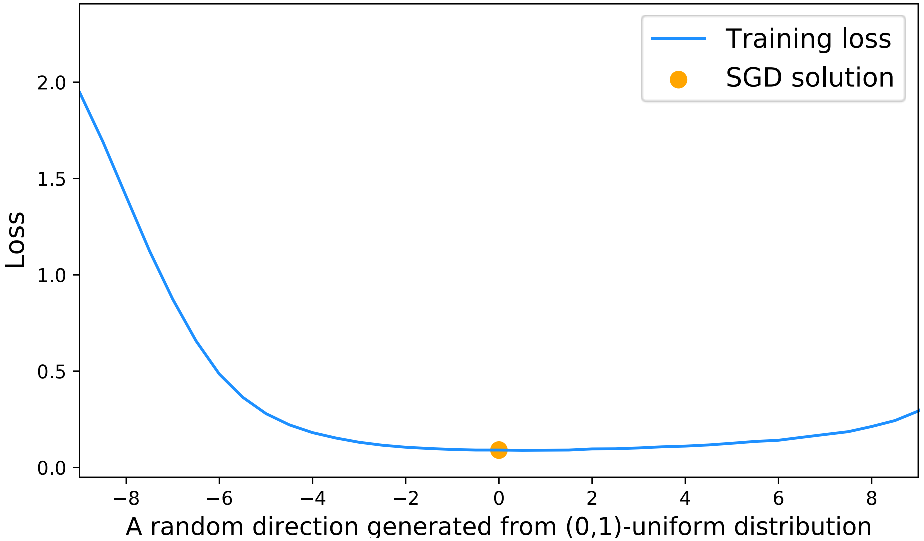
Assumptions and Verification

Locally asymmetric assumption: asymmetric valley in a 2D case
(1 single neuron with its weight, bias and sigmoid activation)



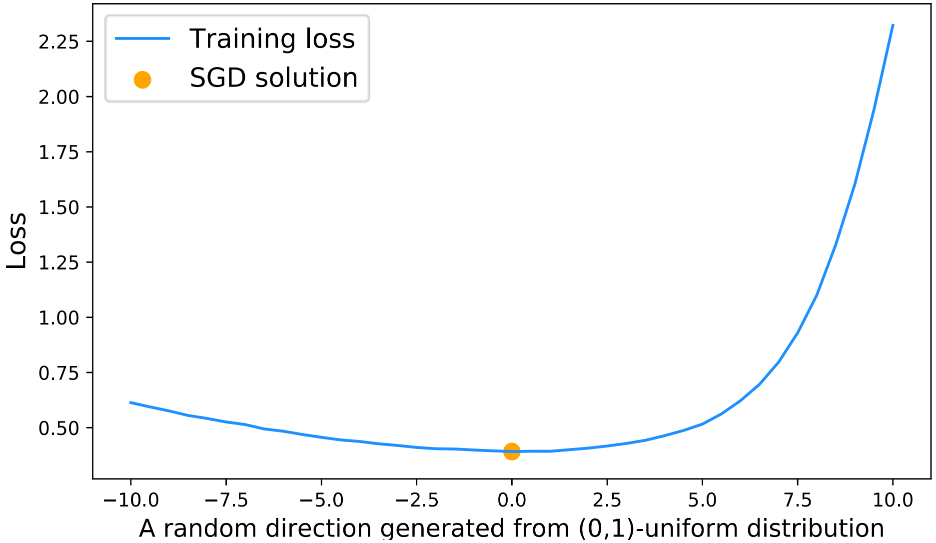
Assumptions and Verification

Locally asymmetric assumption: asymmetric valley in Deep neural networks (DenseNet-100 on CIFAR10)



Assumptions and Verification

Locally asymmetric assumption: asymmetric valley in Deep neural networks (ResNet-164 on CIFAR100)

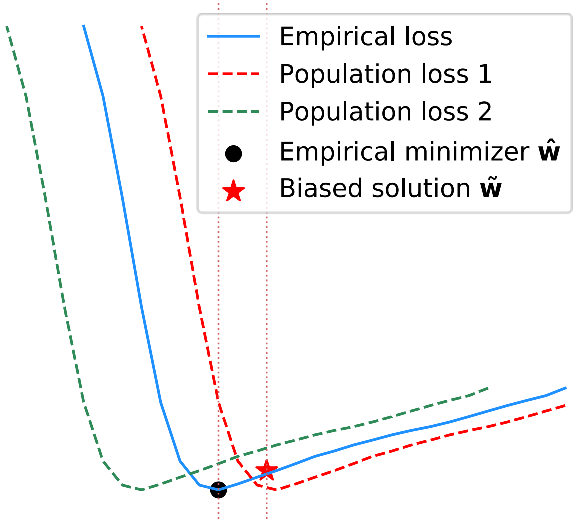


Theorem (informal)

Bias leads to better generalization

$$E_{\delta}L(\hat{w}^*) - E_{\delta}L(\hat{w}^* + c_0) > 0$$

where c_0 is a bias towards the flat side,
$\hat{w}^*$ is an empirical solution

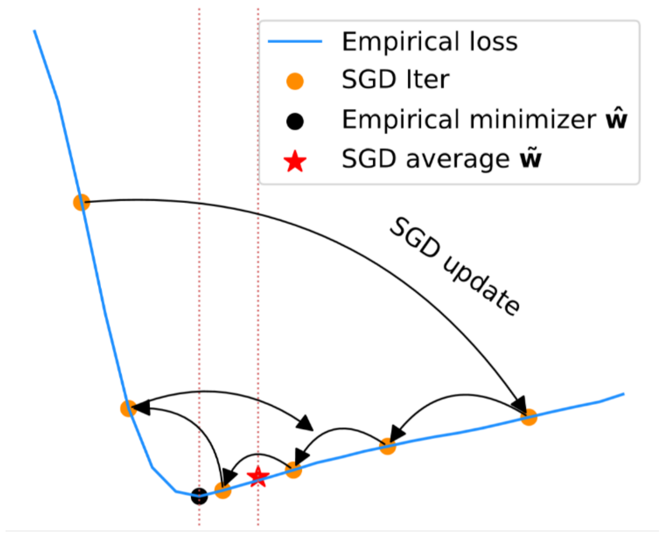


Theorem (informal)

SGD averaging generates a bias

$$E[\overline{w}] > c_0 > 0$$

where c_0 is a bias towards the flat side,
$\overline{w}$ is SGD average



Tsinghua University

Thanks

Asymmetric Valleys: Beyond Sharp and Flat Local Minima

Haowei He | Gao Huang | Yang Yuan

October 26, 2019