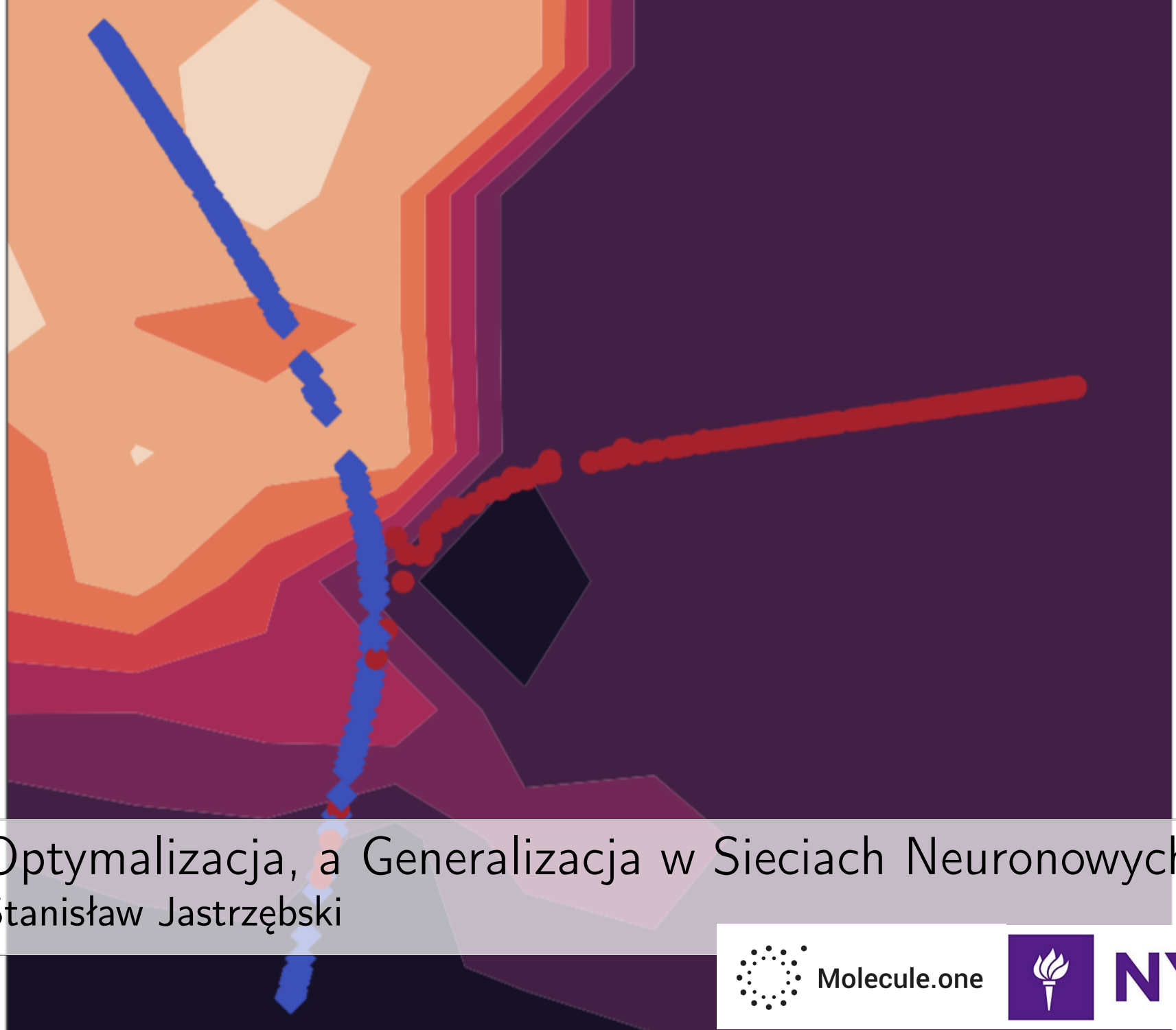


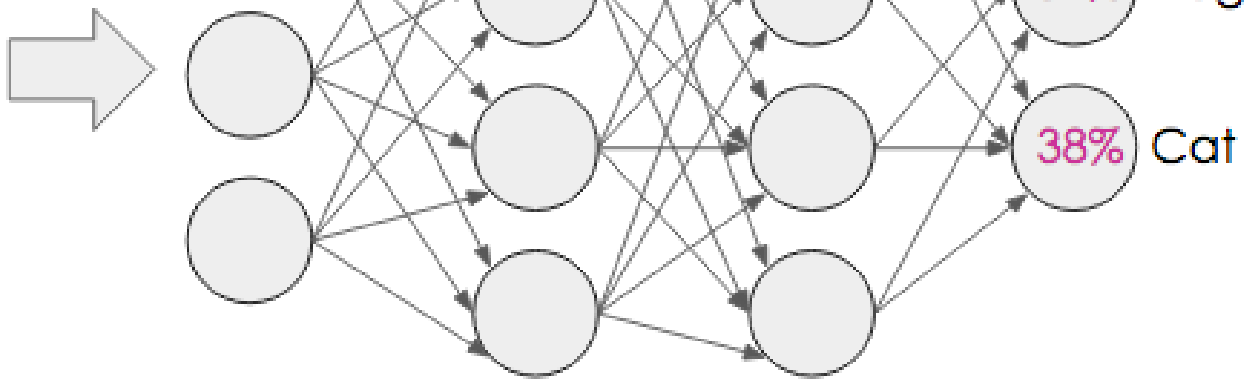
Optymalizacja, a Generalizacja w Sieciach Neuronowych

Stanisław Jastrzębski

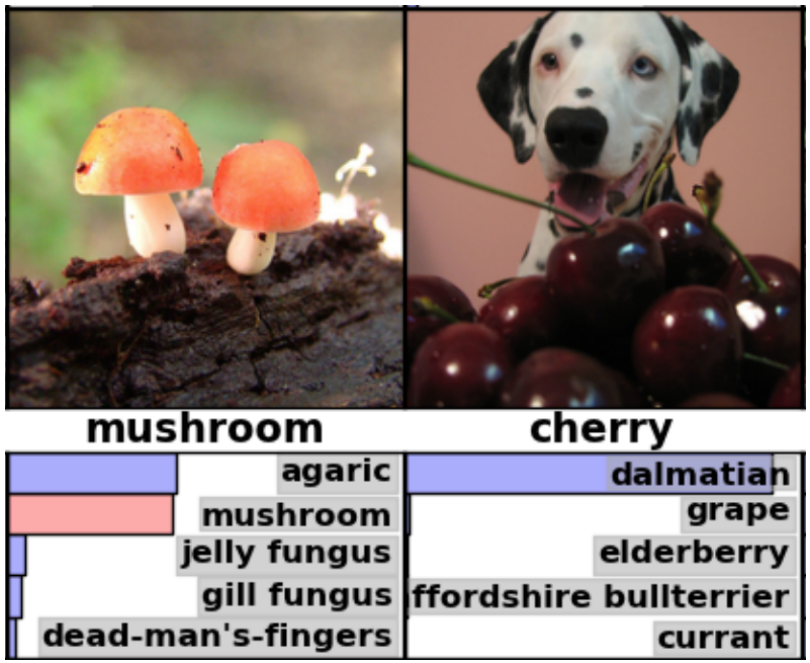
Molecule.one

NYU

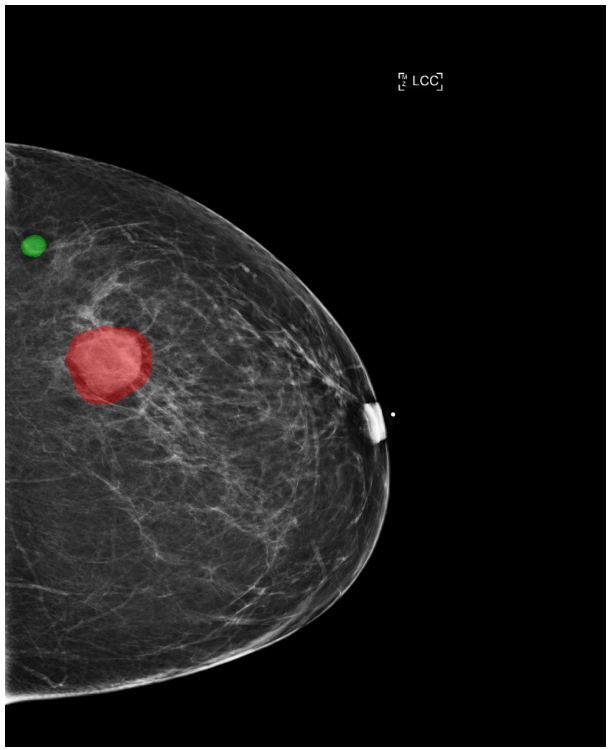
Sieci neuronowe



Sieci neuronowe to niesamowite narzedzie

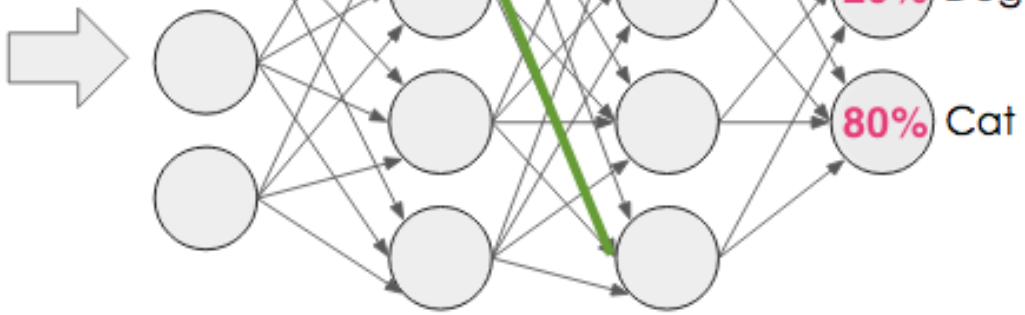


(a) Przewidywania sieci neuronowej na przykładach testowych z ImageNet.

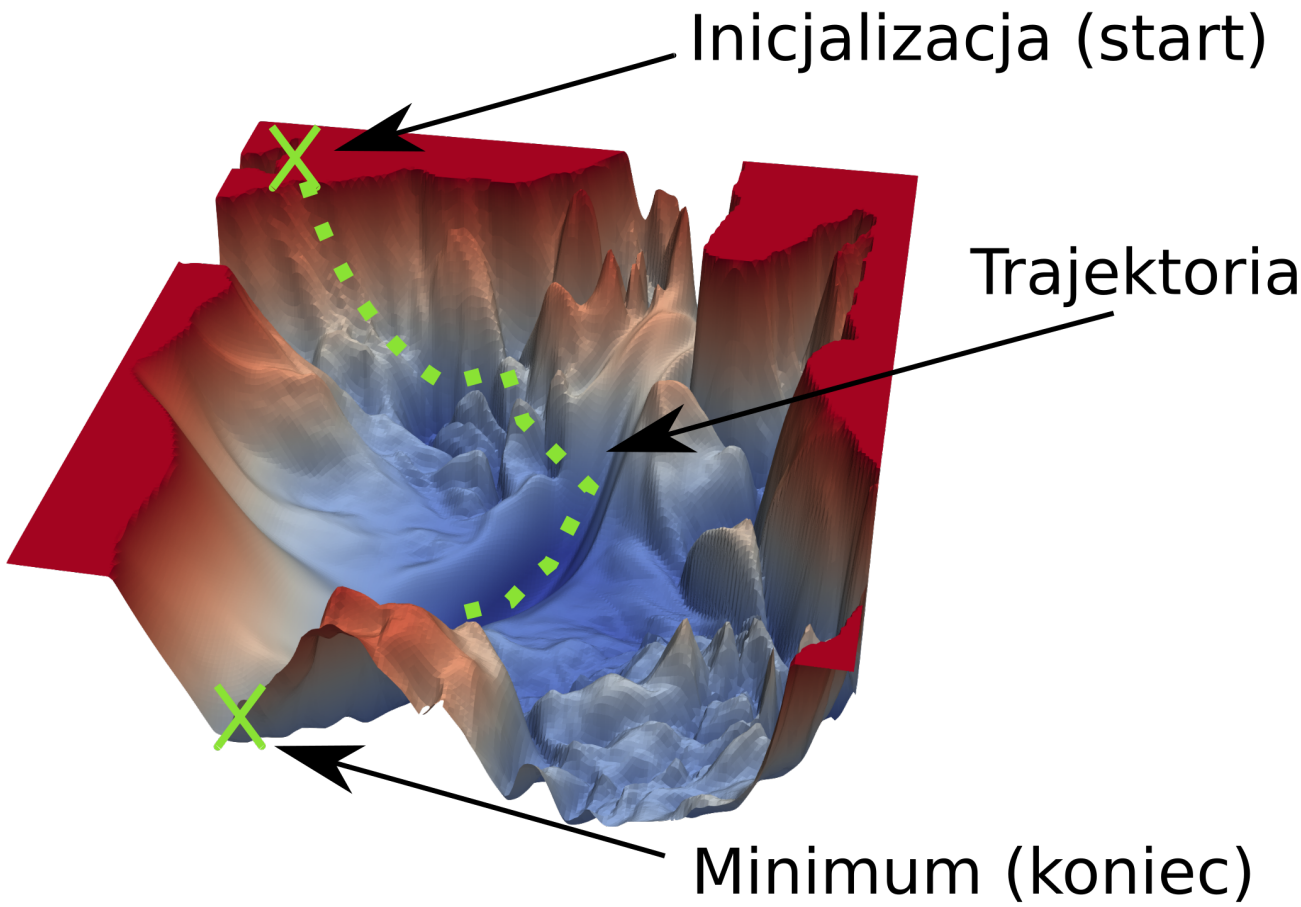


(b) Wykrywanie raka piersi za pomocą sieci neuronowej.

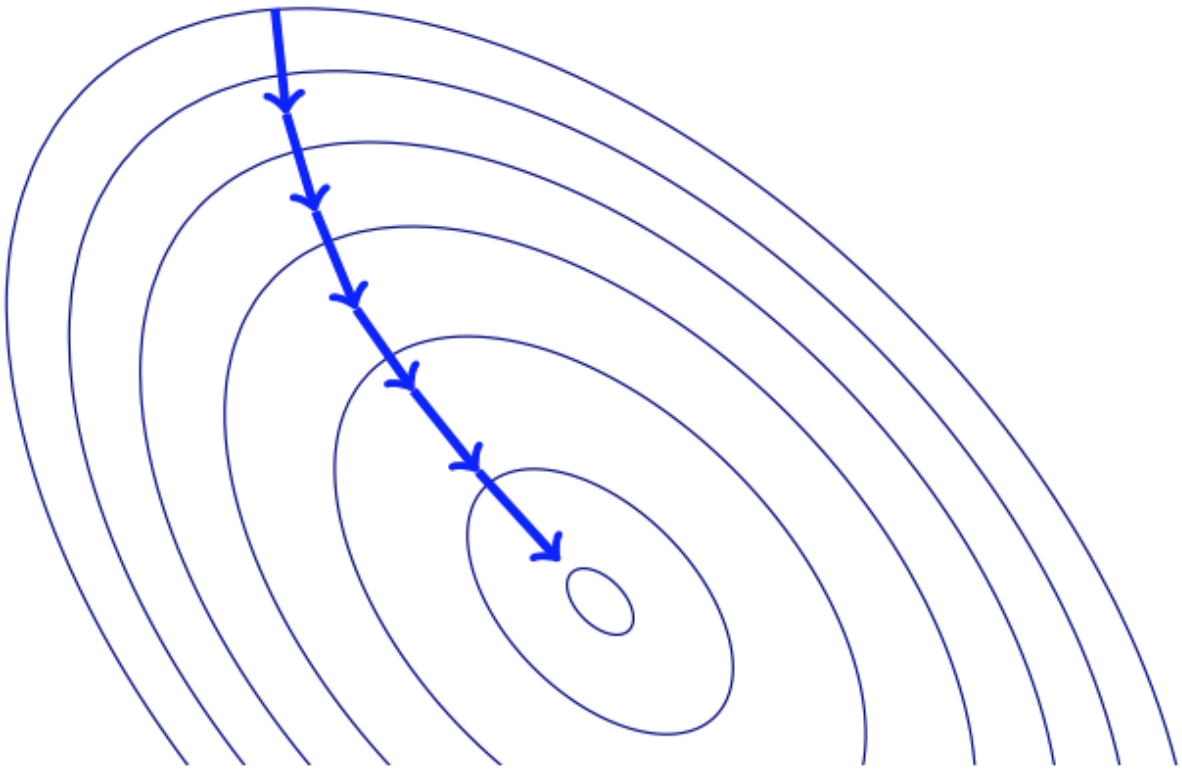
Trenowanie sieci neuronowych



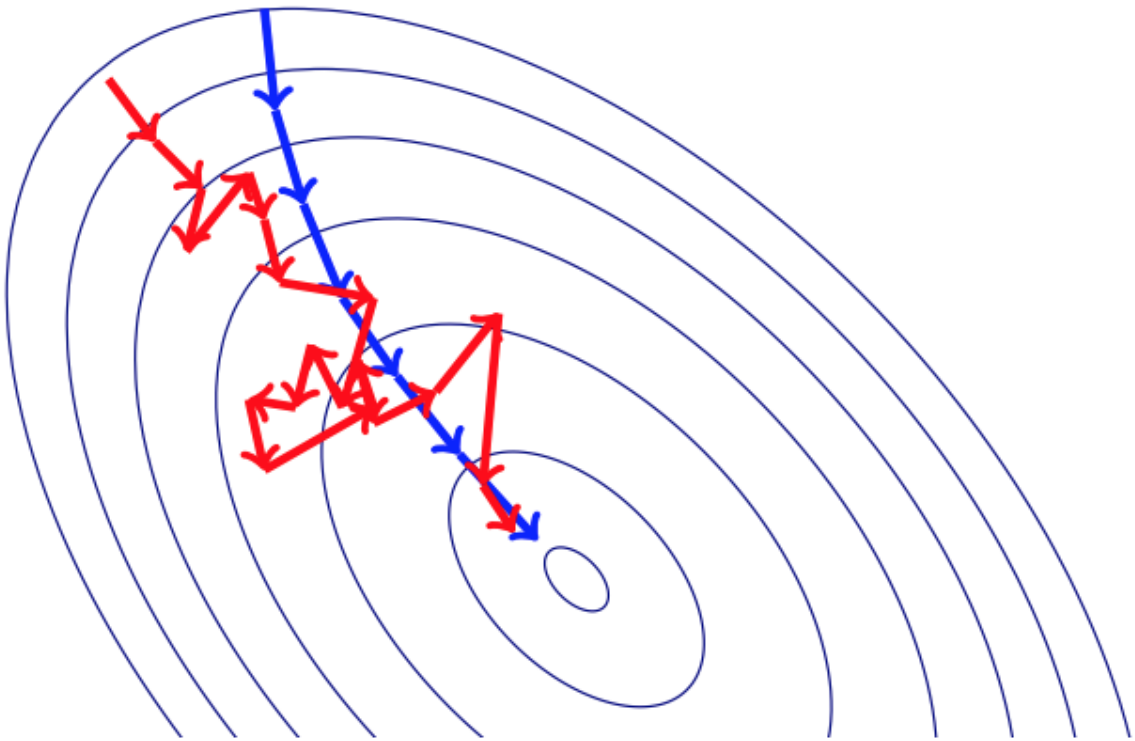
Funkcja kosztu



Metoda najszybszego spadku (GD)



Metoda stochastycznego najszybszego spadku (SGD)



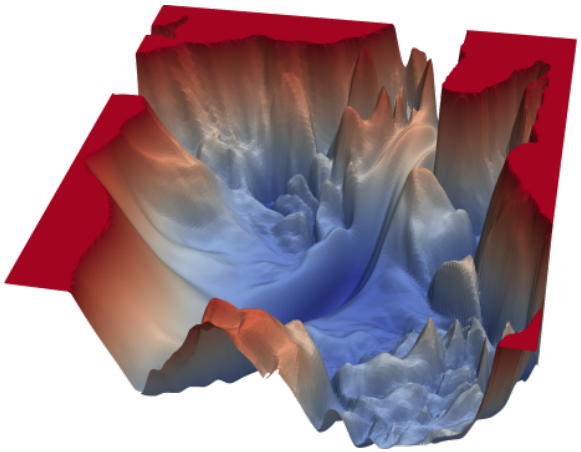
Stochastic Gradient Descent

$$\theta(t+1) = \theta(t) - \eta\hat{g}$$

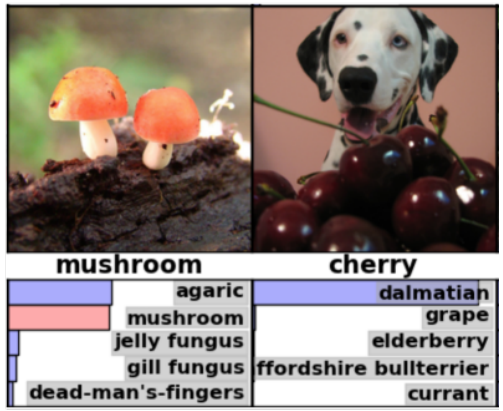
Stochastic Gradient Descent

$$\theta(t+1) = \theta(t) - \eta \hat{g}$$
$$\hat{g} = \frac{1}{S} \sum_{i=1}^{S} \nabla_\theta L(f(x^i), y^i)$$

Problem z sieciami neuronowymi



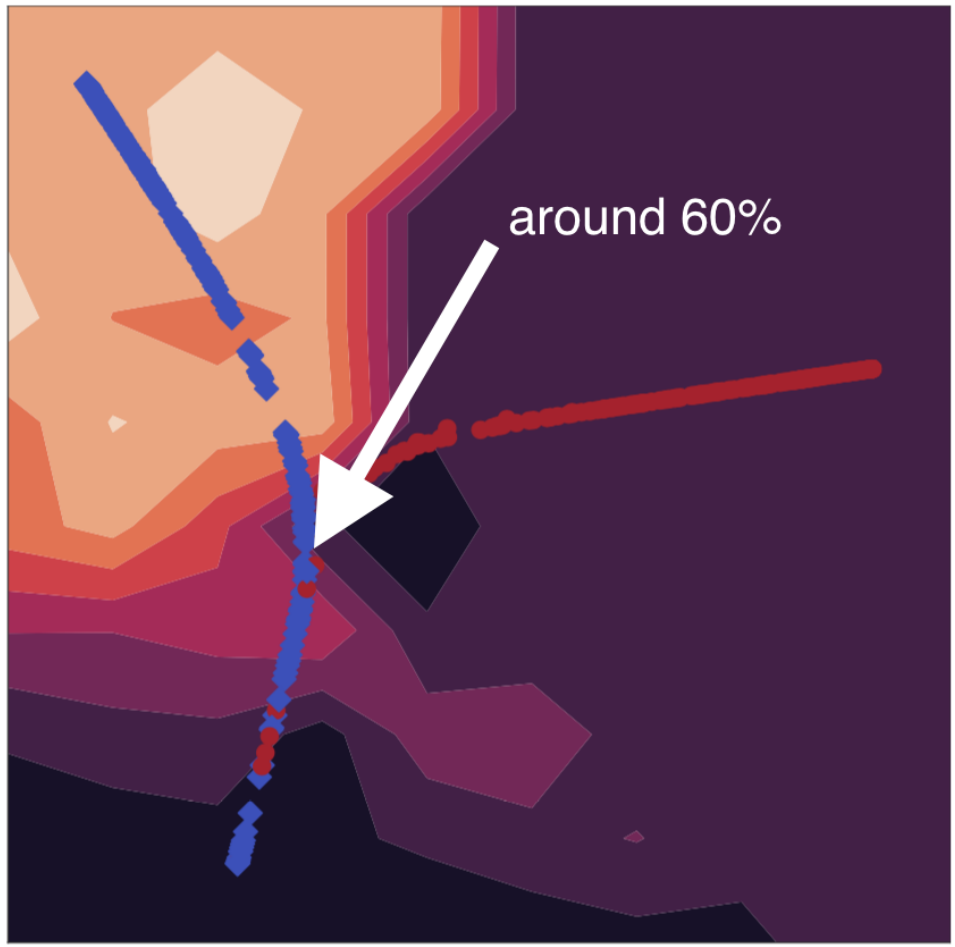
?



Problem badawczy

Jaki jest związek trajektorii optymalizacji z generalizacją sieci neuronowych?

Nowa perspektywa na optymalizację sieci neuronowych



Publikacje

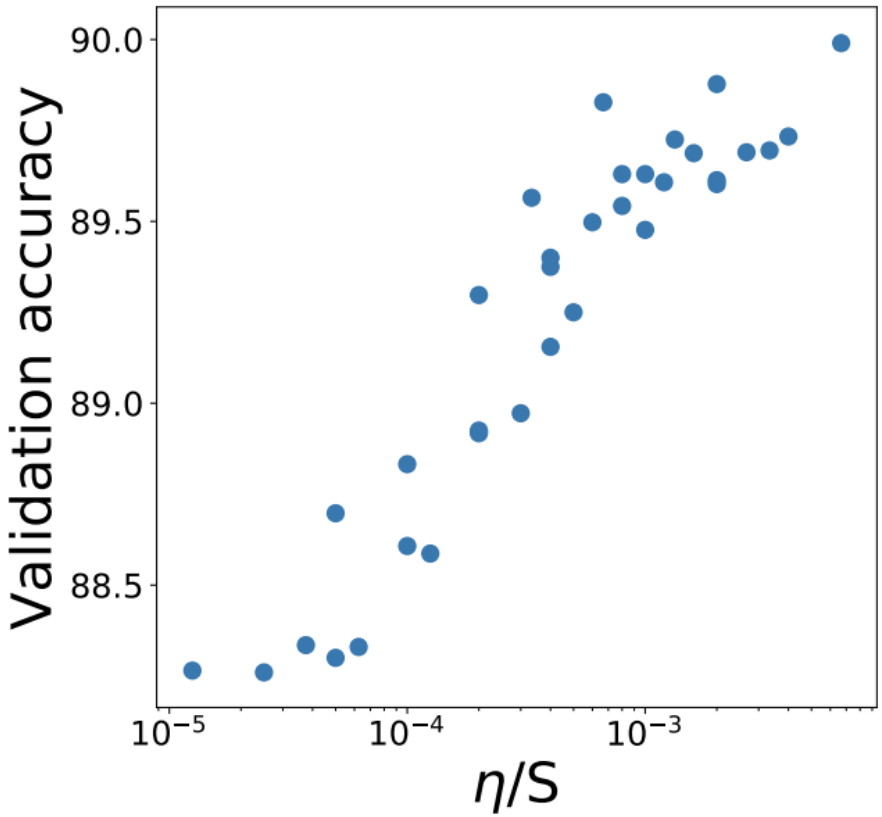
- *A Closer Look at Memorization in Deep Networks*
  Arpit*, **Jastrzebski***, Ballas*, Krueger* et al., ICML 2017
- *Three Factors Influencing Minima in SGD*
  Jastrzebski*, Kenton Z.* et al., ICANN 2018
- *On the Relation Between the Sharpest Directions of DNN Loss and the SGD Step Length*
  Jastrzebski, et al., ICLR 2019
- *Large Scale Structure of Neural Networks Land Scape*
  Fort, **Jastrzebski**, et al., NeurIPS 2019
- *Stiffness: A New Perspective on Generalization of Deep Neural Networks*
  Fort, Novak, **Jastrzebski** et al, in review
- *The Break-Even Point on the Optimization Trajectories of Deep Neural Networks*
  Jastrzebski, et al., in review

1. Czemu krok uczenia zmienia generalizację?

"The learning rate is perhaps the most important hyperparameter. If you have time to tune only one hyperparameter, tune the learning rate"

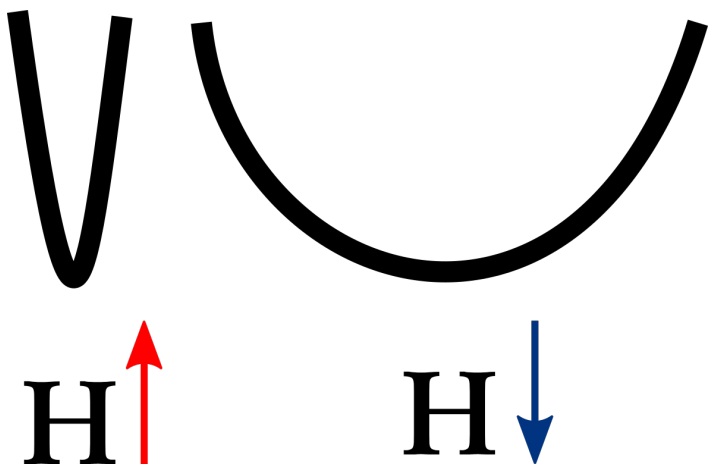
Deep Learning Book, Goodfellow et al

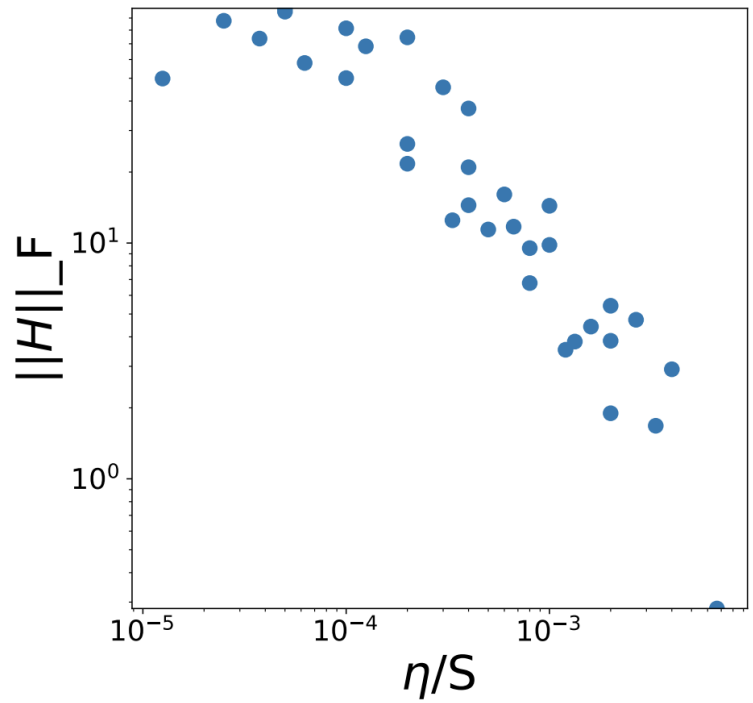
1. Czemu krok uczenia zmienia generalizację?



Three Factors Influencing Minima in SGD

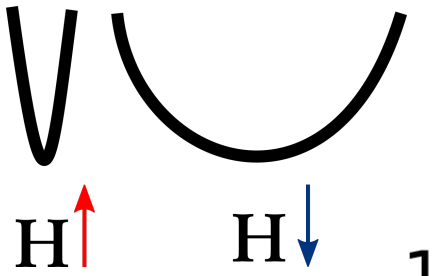
1. Czemu krok uczenia zmienia generalizację?





(a) Norma spektralna Hesjanu w minimum

2. Krok η zmienia trajektorię wcześnie



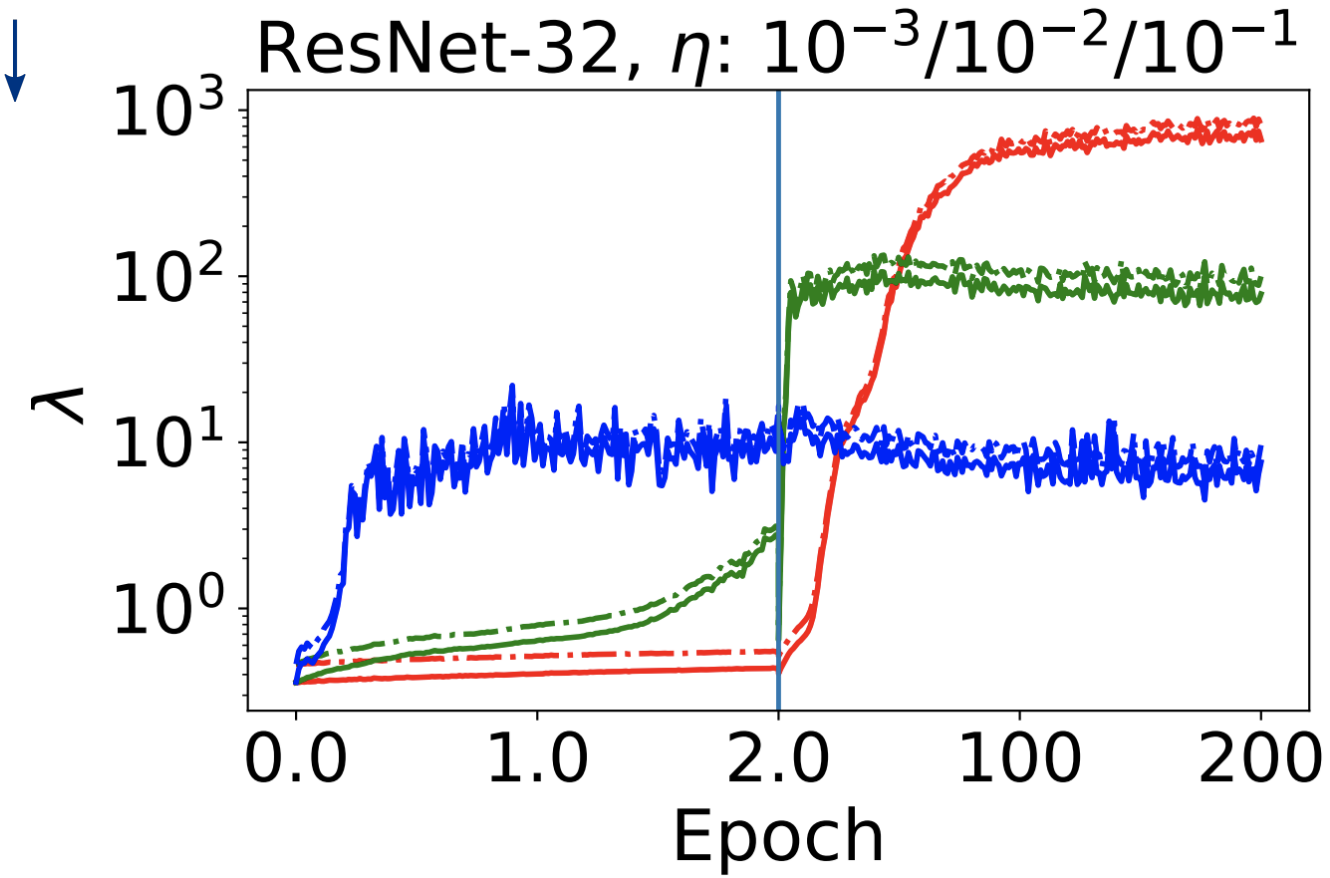
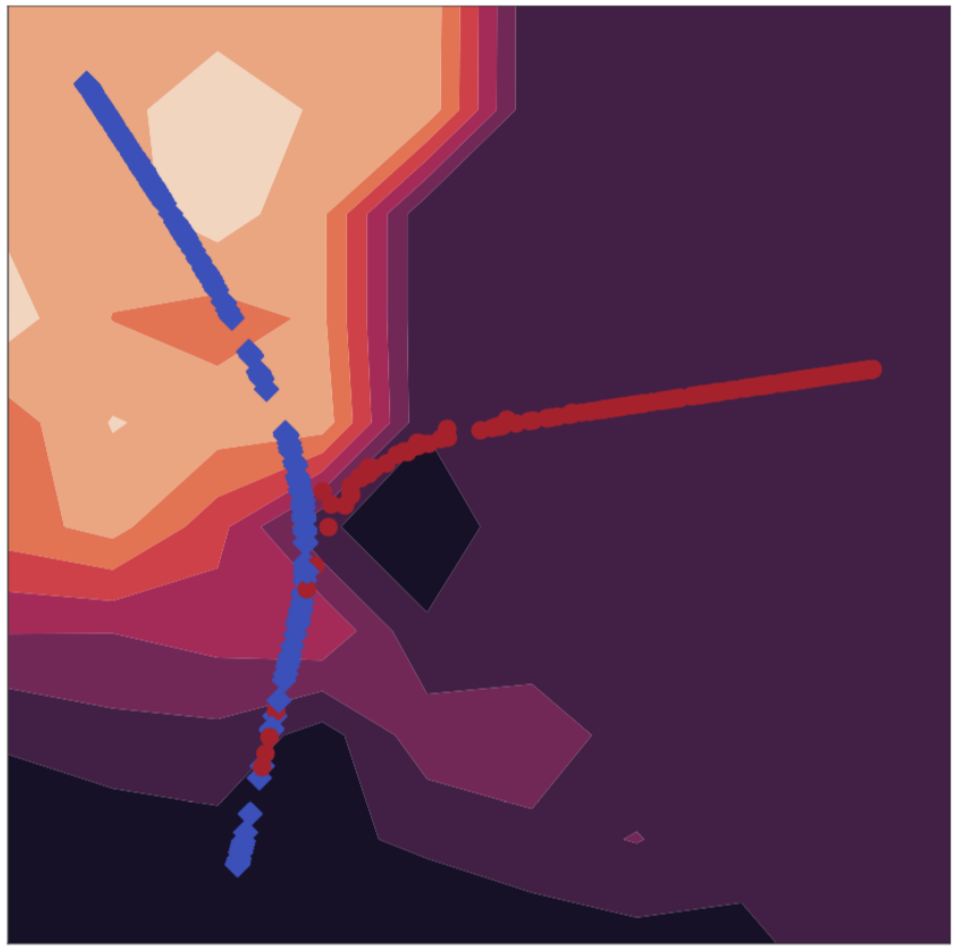


Figure: Norma spektralna Hesjanu w czasie trenowania

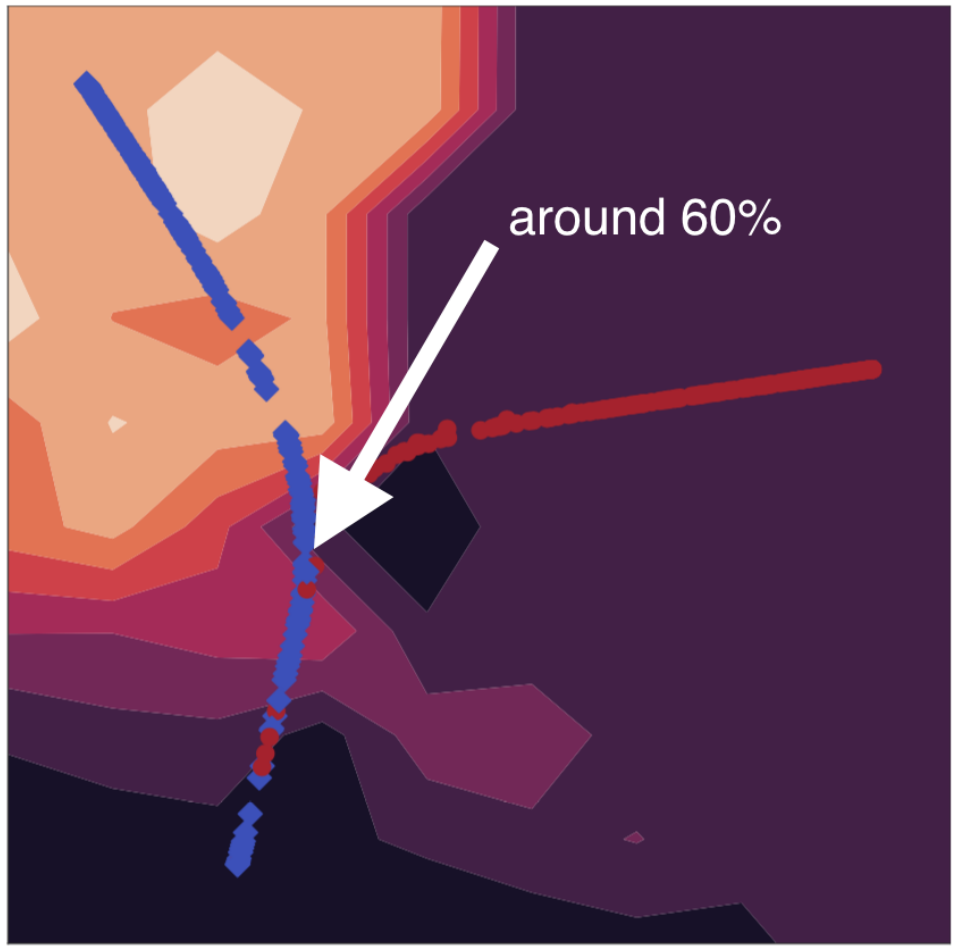
2. Krok η zmienia trajektorię wcześnie



The Break-Even Point on the Optimization Trajectory of Deep Neural Networks

Stiffness: A New Perspective on Generalization

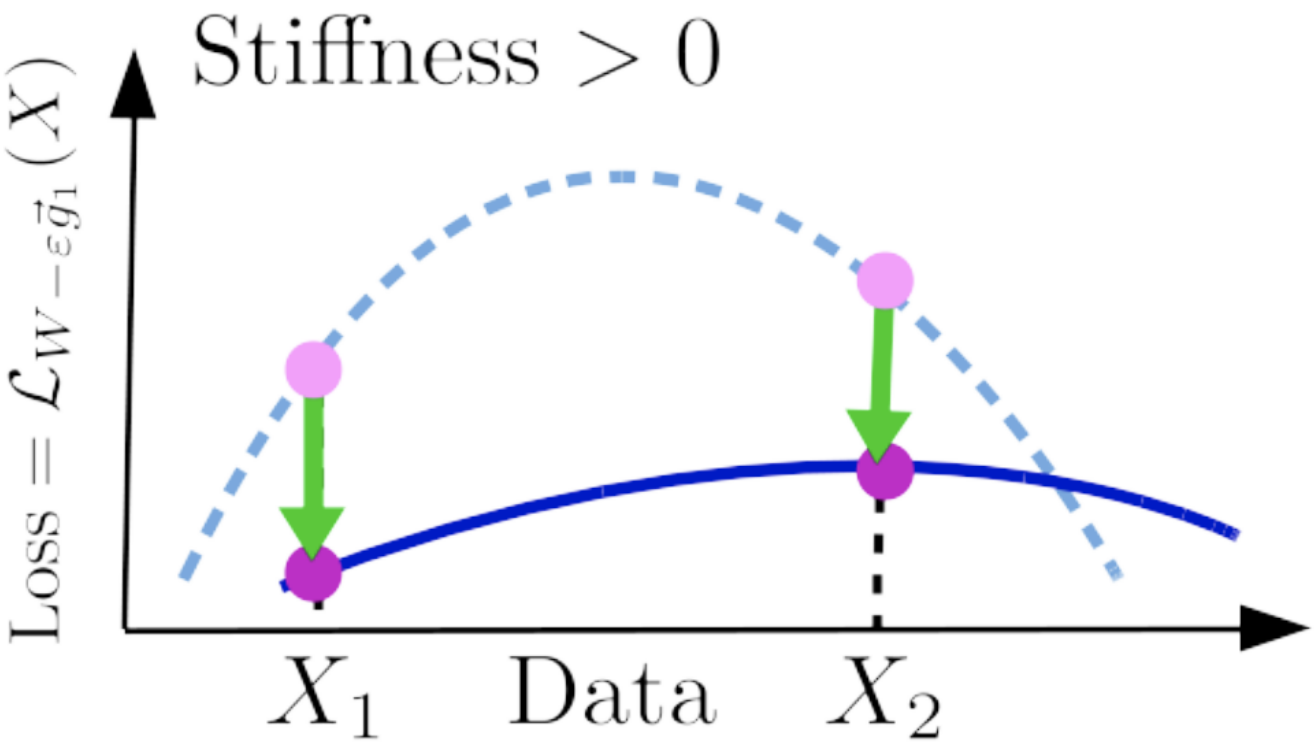
2. Krok η zmienia trajektorię wcześnie

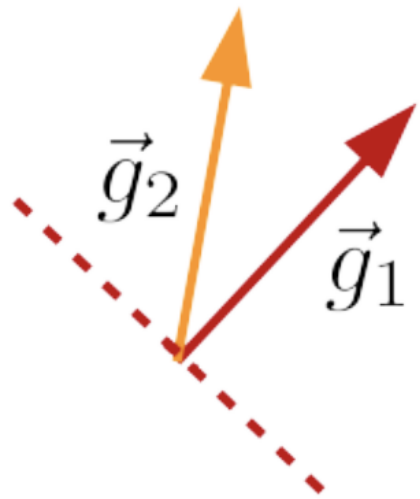


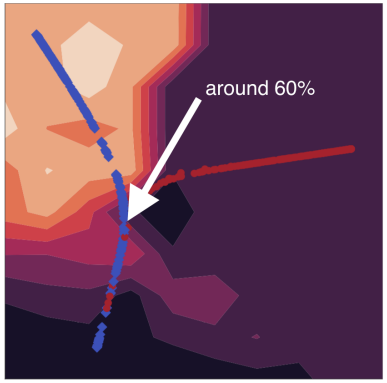
The Break-Even Point on the Optimization Trajectory of Deep Neural Networks

Stiffness: A New Perspective on Generalization

3. Czemu to dobrze?







Podsumowanie

- W sieciach neuronowych optymalizacja i generalizacja są powiązane (np. η)
- Nowa perspektywa: duży krok η w SGD steruje optymalizację od początku w "lepsze" rejony funkcji kosztu