Department of Information Engineering

Understanding the loss landscape of DNNs

Student: Luca Zancato
Supervisor: Prof. Alessandro Chiuso

Control Days 2019
9 May 2019, Padova

Università degli Studi di Padova

Overview

Outline of the presentation:

Overview

Outline of the presentation:

1. Machine Learning Framework
2. Optimization for Deep Neural Networks

Outline of the presentation:

1. Machine Learning Framework
2. Optimization for Deep Neural Networks
3. Simplified model design

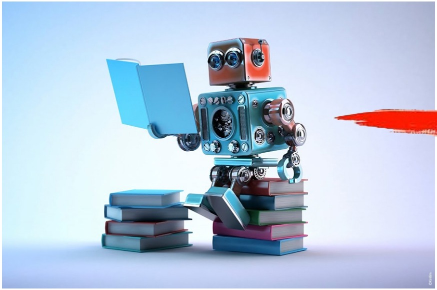
Overview

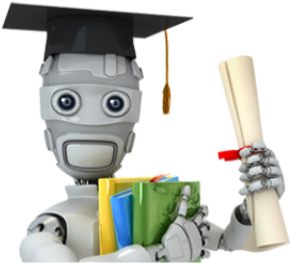
Outline of the presentation:

Machine Learning Framework

Experience

Expertise





Machine Learning Framework

Ingredients of Machine Learning:

- **Training data**: $(y_i, \mathbf{x}_i) \quad 1, ..., m$
- **Model:** $h_\theta \in \mathcal{H}$

Ingredients of Machine Learning:

- **Training data**: $(y_i, \mathbf{x}_i) \quad 1, ..., m$
- **Model:** $h_\theta \in \mathcal{H}$

Empirical Loss function: $L(\theta) := \sum_{i=1}^{m} l(y_i, h_\theta(\mathbf{x}_i))$

Ingredients of Machine Learning:

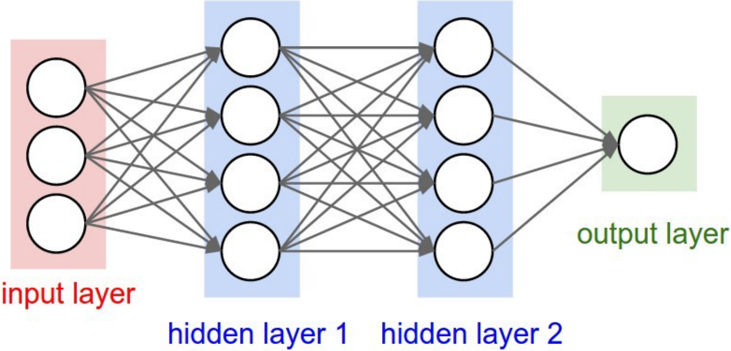
- **Training data**: $(y_i, \mathbf{x}_i) \quad 1, ..., m$
- **Model:** $h_\theta \in \mathcal{H}$

Empirical Loss function: $L(\theta) := \sum_{i=1}^{m} l(y_i, h_\theta(\mathbf{x}_i))$

$$\downarrow$$

$$\theta_{opt} := \arg\min_{\theta \in \mathbb{R}^d} L(\theta)$$

1 What is a DNN? $\Longrightarrow$ Weighted directed acyclic graph



Composed by **many layers** performing non-linear transformations

1. What is a DNN? $\Longrightarrow$ Weighted directed acyclic graph
2. How to choose the parameters? $\Longrightarrow$ Gradient Descent

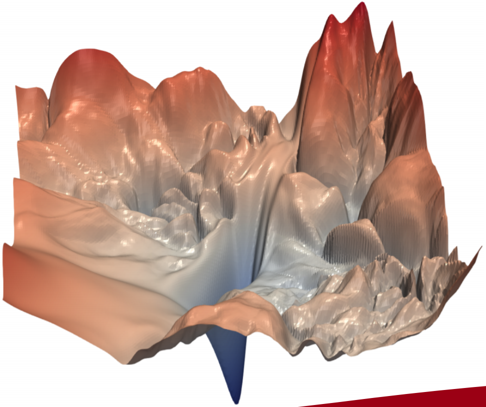
Deep Neural Networks (DNNs)

Università degli Studi di Padova

1. What is a DNN? $\Longrightarrow$ Weighted directed acyclic graph
2. How to choose the parameters? $\Longrightarrow$ Gradient Descent

Optimization problems:

- High dimensional **non convex** empirical loss function



1. What is a DNN? $\implies$ Weighted directed acyclic graph
2. How to choose the parameters? $\implies$ Gradient Descent

OPTIMIZATION PROBLEMS:

- High dimensional **non convex** empirical loss function
- **Computations**

Deep Neural Networks (DNNs)

Università degli Studi di Padova

1. What is a DNN? $\Longrightarrow$ Weighted directed acyclic graph
2. How to choose the parameters? $\Longrightarrow$ Gradient Descent

OPTIMIZATION PROBLEMS:

- High dimensional **non convex** empirical loss function
- **Computations**
- **Generalization** properties of θ_{opt}?

Optimization vs ML

Literature focuses on:

1. **Why** and **when** is deep learning **effective**?
2. Good **optimization** $\implies$ Good **generalization**?

Optimization vs ML

Literature focuses on:

1. **Why** and **when** is deep learning **effective**?
2. Good **optimization** $\implies$ Good **generalization**?

By means of:

- **Stochastic optimization algorithms**
  - SGD
  - RMSProp
  - Adam

Optimization vs ML

Literature focuses on:

1. **Why** and **when** is deep learning **effective**?
2. Good **optimization** $\implies$ Good **generalization**?

By means of:

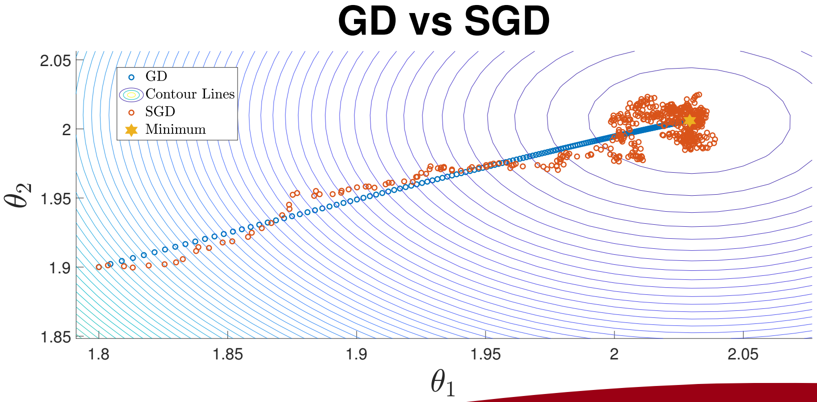
- **Stochastic optimization algorithms**
  - SGD
  - RMSProp
  - Adam
- **Geometry of the empirical loss landscape**

Gradient Descent **SGD**

$$\theta_{k+1} = \theta_k - \eta \sum_{i=1}^{m} \underbrace{\nabla l(\theta_k, \mathbf{x}_i)}_{\text{gradient loss}} \implies \theta_{k+1} = \theta_k - \eta \sum_{i=1}^{B} \nabla l(\theta_k, \mathbf{x}_i)$$

$$m \gg B$$

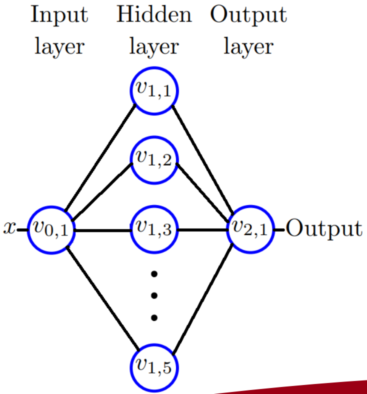
$$\text{SGD} \implies \theta_{k+1} = \underbrace{\theta_k - \eta \sum_{i=1}^{m} \nabla l(\theta_k, \mathbf{x}_i)}_{\text{GD dynamics}} + \underbrace{v_k}_{\text{noise term}}$$



Simplified model and Local minima

Consider a regression task from $\mathbb{R}$ to $\mathbb{R}$:

$$h_{\alpha,\mathbf{w},\mathbf{b}}(\mathbf{x}) = \sum_{i=1}^{d} \alpha_i \tanh(\mathbf{w} \cdot (\mathbf{x} - \mathbf{b}_i))$$



Consider a regression task from $\mathbb{R}$ to $\mathbb{R}$:

$$h_{\alpha,\mathbf{w},\mathbf{b}}(\mathbf{x}) = \sum_{i=1}^{d} \alpha_i \tanh(\mathbf{w} \cdot (\mathbf{x} - \mathbf{b}_i))$$

$$L(\alpha, \mathbf{w}, \mathbf{b}, \mathbf{X}) = \sum_{i=1}^{m} (y_i - h_{\alpha,\mathbf{w},\mathbf{b}}(\mathbf{x}_i))^2 = \|\mathbf{Y} - \Phi(\mathbf{w}, \mathbf{b}, \mathbf{X})\alpha\|^2$$

Consider a **regression** task from $\mathbb{R}$ to $\mathbb{R}$:

$$h_{\alpha,\mathbf{w},\mathbf{b}}(\mathbf{x}) = \sum_{i=1}^{d} \alpha_i \tanh(\mathbf{w} \cdot (\mathbf{x} - \mathbf{b}_i))$$

Fix the centers $\mathbf{b}_i$:

$$L(\alpha, \mathbf{w}, \mathbf{X}) = \|\mathbf{Y} - \Phi(\mathbf{w}, \mathbf{X})\alpha\|^2$$

And α?

$$\arg\min_{\alpha,\mathbf{w}} L(\alpha, \mathbf{w}, \mathbf{X}) = \arg\min_{\mathbf{w}} \left\|\mathbf{Y} - \Phi(\mathbf{w}, \mathbf{X})\Phi^{\dagger}(\mathbf{w}, \mathbf{X})\mathbf{Y}\right\|^2$$

Simplified model and local minima

Design of a simplified model:

$$L(\underbrace{\theta}_{\mathbb{R}^d}) \Longrightarrow L(\underbrace{w}_{\mathbb{R}^p}) \quad d \gg p$$

Simplified model and local minima

Design of a simplified model:

$$L(\underbrace{\theta}_{\mathbb{R}^d}) \Longrightarrow L(\underbrace{w}_{\mathbb{R}^p}) \quad d \gg p$$

Is the geometry of the simplified model rich enough?

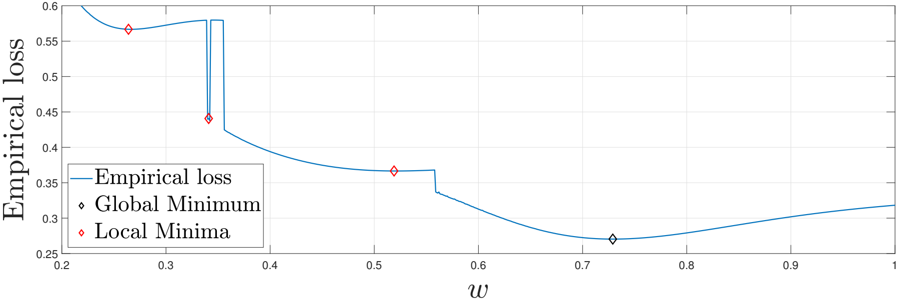
Simplified model and local minima

Università degli Studi di Padova

Design of a simplified model:

$$L(\underbrace{\theta}_{\mathbb{R}^d}) \implies L(\underbrace{w}_{\mathbb{R}^p}) \quad d \gg p$$

Is the geometry of the simplified model rich enough?



Continuous time approximation of SGD

Università degli Studi di Padova

How to **approximate** the following stochastic difference equation?

$$\theta_{k+1} = \underbrace{\theta_k - \eta \nabla L(\theta_k)}_{\text{GD dynamics}} + \underbrace{v_k(\theta_k)}_{\text{colored noise}}$$

$$\Big\downarrow$$

$$d\theta(t) = - \underbrace{\nabla L(\theta)}_{\text{drift term}} \, dt + \underbrace{\sqrt{\beta D(\theta)}}_{\text{diffusion matrix}} \, dW(t)$$

With $\beta = \frac{\eta}{B}$ **Temperature**

Asymptotic Distribution

Let $\rho(\theta, t)$ be the probability distribution over the parameter space. It is well known that this is ruled by the **Fokker-Planck** equation:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \Big(\nabla L(\theta)\rho + \beta \nabla \cdot (D(\theta)\rho) \Big)$$

Asymptotic Distribution

Let $\rho(\theta, t)$ be the probability distribution over the parameter space. It is well known that this is ruled by the **Fokker-Planck** equation:

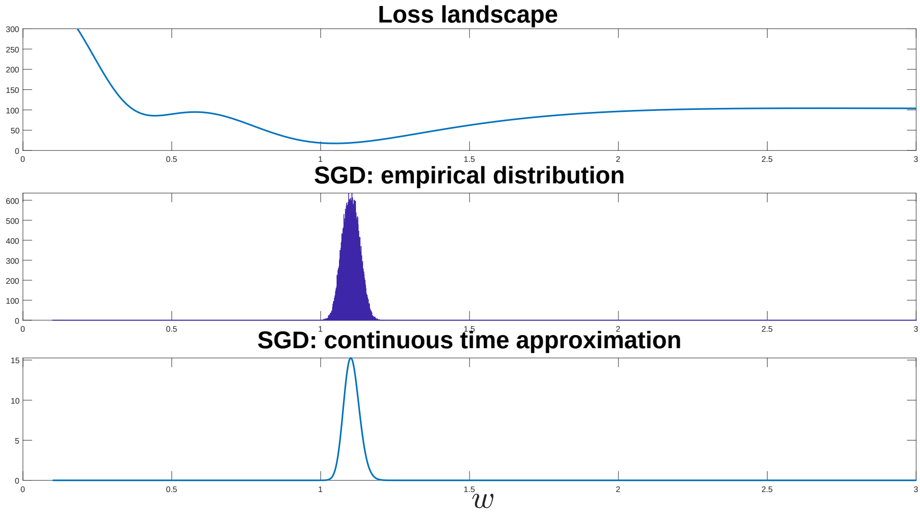
$$\frac{\partial \rho}{\partial t} = \nabla \cdot \Big(\nabla L(\theta)\rho + \beta \nabla \cdot (D(\theta)\rho) \Big)$$

Let the **Stationary distribution** $\rho^{SS}(\theta) : \mathbb{R}^d \to \mathbb{R}$

$$\rho^{SS}(\theta) \propto e^{-\beta^{-1}\Phi(\theta)} = e^{-\frac{B}{\eta}\Phi(\theta)}$$

where $\Phi(\theta)$ s.t. $\frac{\partial \rho^{SS}}{\partial t} = 0$

Asymptotic distribution approximation



Deterministic gradient vs Stochastic gradient

$$\nabla L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \nabla l_i(\theta) \implies \nabla G(\theta) = \frac{1}{B}\sum_{i\in\mathcal{B}} \nabla l_i(\theta)$$

with $i_1, ..., i_B \in \mathcal{B}$ i.i.d. r.v. in the set $1, 2, ..., m$

Covariance matrix decomposition

Deterministic gradient vs Stochastic gradient

$$\nabla L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla l_i(\theta) \implies \nabla G(\theta) = \frac{1}{B} \sum_{i \in \mathcal{B}} \nabla l_i(\theta)$$

with $i_1, \ldots, i_B \in \mathcal{B}$ i.i.d. r.v. in the set $1, 2, \ldots, m$

$$\mathbb{E}_{i_1, i_2, \ldots, i_B} \left[\nabla G(\theta) \right] = \nabla L(\theta)$$

Covariance matrix decomposition

Deterministic gradient vs Stochastic gradient

$$\nabla L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \nabla l_i(\theta) \implies \nabla G(\theta) = \frac{1}{B}\sum_{i\in\mathcal{B}} \nabla l_i(\theta)$$

with $i_1, ..., i_B \in \mathcal{B}$ i.i.d. r.v. in the set $1, 2, ..., m$

$$\mathbb{E}_{i_1,i_2,\dots,i_B}\left[\nabla G(\theta)\right] = \nabla L(\theta)$$

$$Var_{i_1,i_2,\dots,i_B}\left[\nabla G(\theta)\right] = \frac{1}{B}\left(\frac{\sum_{j=1}^{m} \nabla l_j(\theta)\nabla l_j(\theta)^T}{m} - \nabla G(\theta)\nabla G(\theta)^T\right)$$

Covariance matrix decomposition

Assumptions:

1. Quadratic loss:
$$l_i(\theta) = (y_i - h_\theta(\mathbf{x}_i))^2 \implies \nabla l_i(\theta) = -2\underbrace{(y_i - h_\theta(\mathbf{x}_i))}_{:= e_\theta(\mathbf{x}_i)} \nabla h_\theta(\mathbf{x}_i)$$

Covariance matrix decomposition

Università degli Studi di Padova

Assumptions:

1. Quadratic loss:
   $l_i(\theta) = (y_i - h_\theta(\mathbf{x}_i))^2 \implies \nabla l_i(\theta) = -2\underbrace{(y_i - h_\theta(\mathbf{x}_i))}_{:=e_\theta(\mathbf{x}_i)}\nabla h_\theta(\mathbf{x}_i)$

2. Noisy data generative process:

$$y_i = h_{\theta^*}(\mathbf{x}_i) + \epsilon_i \quad \forall i$$

   with ϵ_i i.i.d. with $\mathbb{E}[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$

Assumptions:

1. Quadratic loss:
$l_i(\theta) = (y_i - h_\theta(\mathbf{x}_i))^2 \implies \nabla l_i(\theta) = -2\underbrace{(y_i - h_\theta(\mathbf{x}_i))}_{:=e_\theta(\mathbf{x}_i)} \nabla h_\theta(\mathbf{x}_i)$

2. Noisy data generative process:

$$y_i = h_{\theta^*}(\mathbf{x}_i) + \epsilon_i \quad \forall i$$

with ϵ_i i.i.d. with $\mathbb{E}[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$

3. θ_{loc} is a local minimum i.e. $\nabla L(\theta_{loc}) = 0 \not\Longrightarrow \nabla G(\theta_{loc}) = 0$

Then:

$$e_\theta(x_i) = y_i - h_\theta(x_i) \underbrace{=}_{hyp1} h_{\theta^*}(x_i) + \epsilon_i - h_\theta(x_i) = \underbrace{h_{\theta^*}(x_i) - h_\theta(x_i)}_{:=\Delta_\theta(x_i)} + \epsilon_i$$

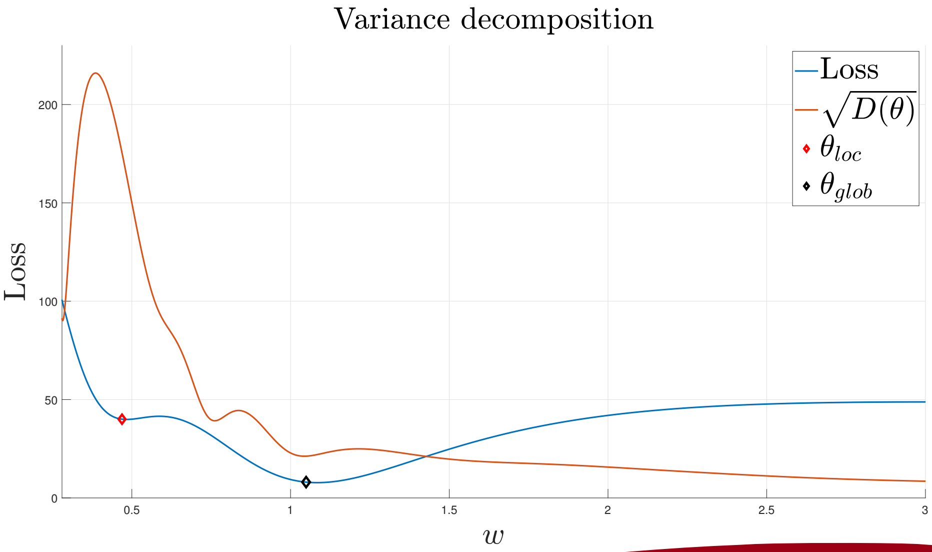
Covariance matrix decomposition

Then:

$$e_\theta(x_i) = y_i - h_\theta(x_i) \underbrace{=}_{hyp1} h_{\theta^*}(x_i) + \epsilon_i - h_\theta(x_i) = \underbrace{h_{\theta^*}(x_i) - h_\theta(x_i)}_{:=\Delta_\theta(x_i)} + \epsilon_i$$

$$\mathbb{E}_e[Var_w[G(\theta)]] \propto$$

$$\Bigg[\underbrace{\sigma^2 \sum_{i=1}^{m} \nabla h_\theta(x_i) \nabla h_\theta(x_i)^T}_{\text{intrinsic noise term}} + \underbrace{\sum_{i=1}^{m} \Delta_\theta^2(x_i) \nabla h_\theta(x_i) \nabla h_\theta(x_i)^T}_{\text{exploration noise term}} \Bigg]$$


Variance decomposition
Loss
$\sqrt{D(\theta)}$
θ_{loc}
θ_{glob}
Loss
0
50
100
150
200
0.5
1
1.5
2
2.5
3
w

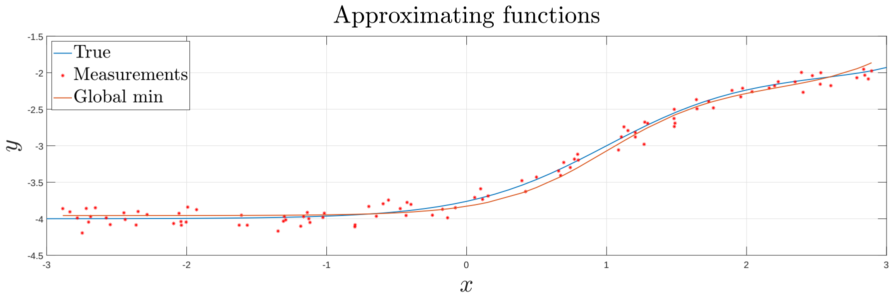
From asymptotic to transient dynamic

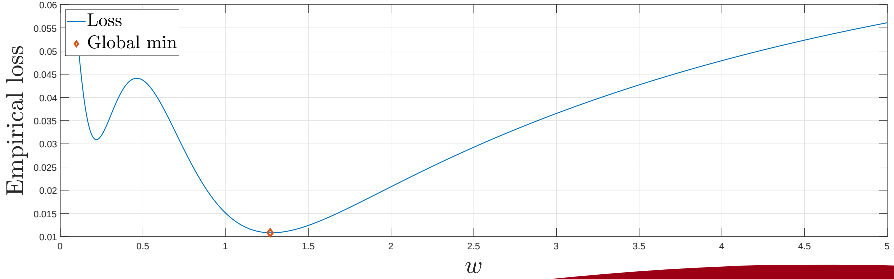
Up to now we described **asymptotic** behaviours!

In practice we are also interested in reaching as quickly as possible the stationary distribution.

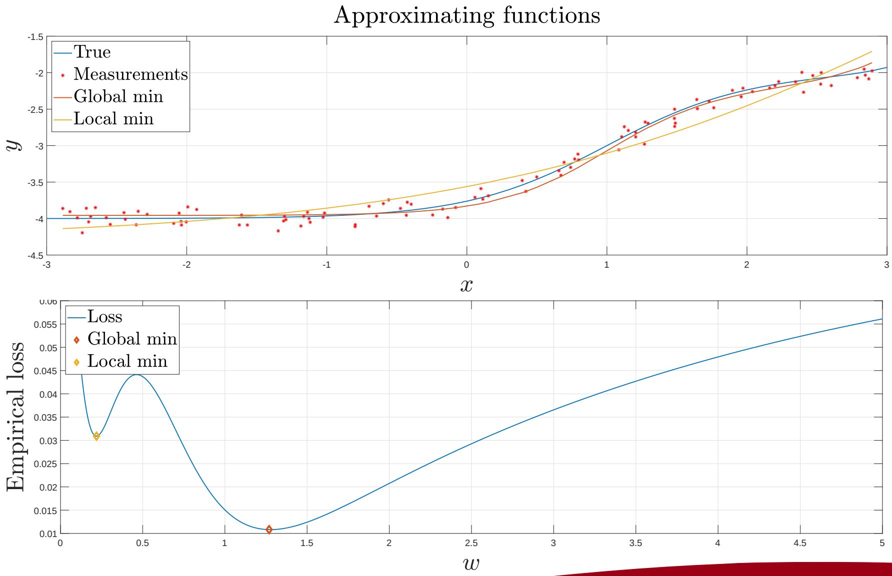
How to study the **transient dynamic**?

Genesis of local minima

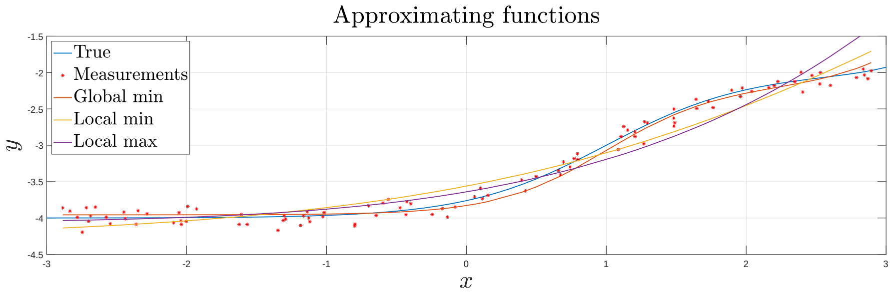


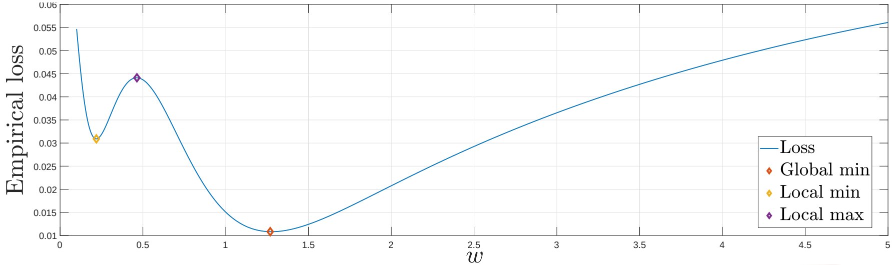


Genesis of local minima



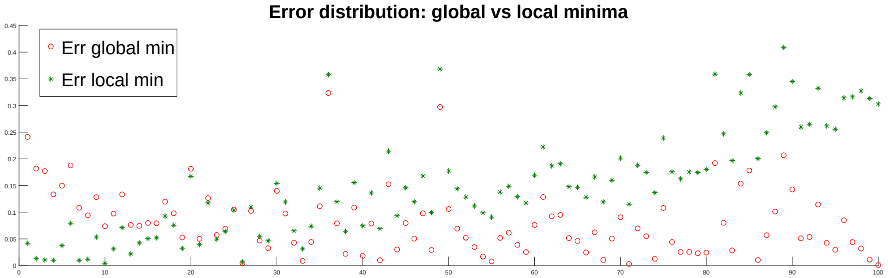
Genesis of local minima

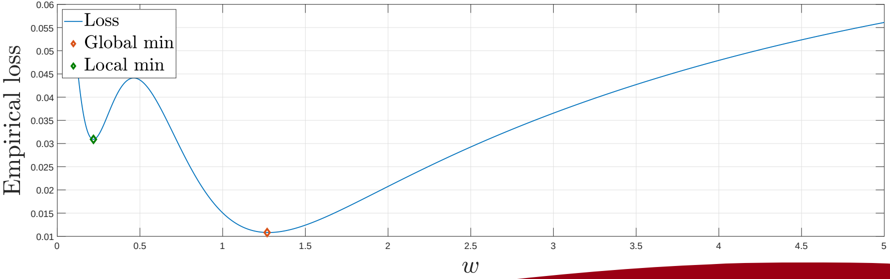




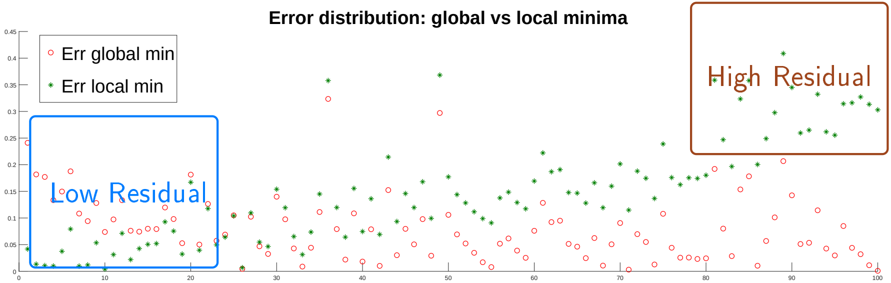
A Local Analysis on the errors distribution

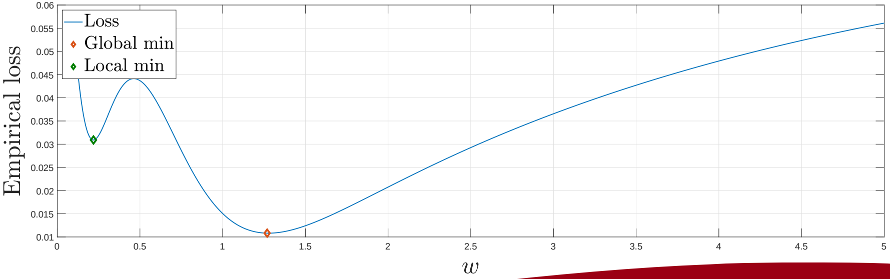
UNIVERSITÀ DEGLI STUDI DI PADOVA



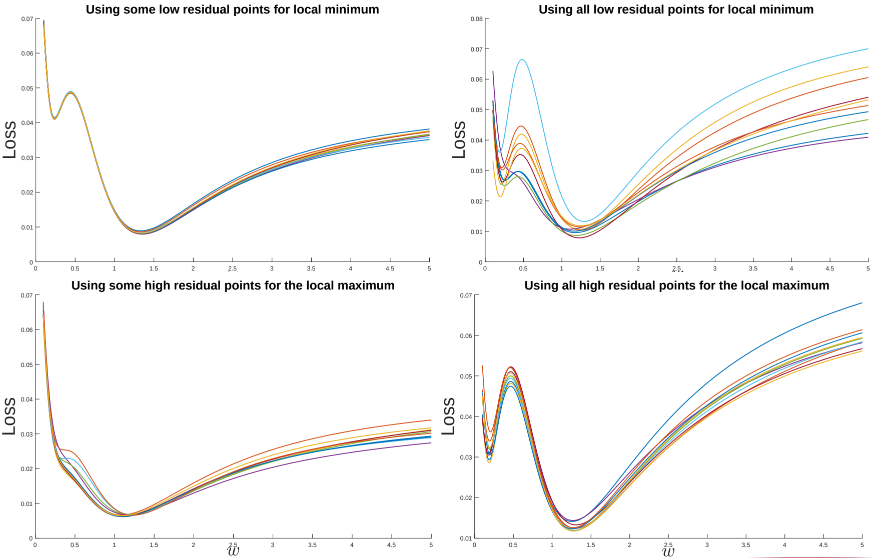


A Local Analysis on the errors distribution





A Local Analysis on the errors distribution



A new perspective: importance sampling

Local maxima are slowing the optimization!

If caused by high residual $\implies$ we can **easily identify** them

A new perspective: importance sampling

Local maxima are slowing the optimization!

If caused by high residual $\implies$ we can **easily identify** them

Dynamic outliers: data points that are not well fitted by the model under the current parameter θ_k

Local maxima are slowing the optimization!

If caused by high residual $\implies$ we can **easily identify** them

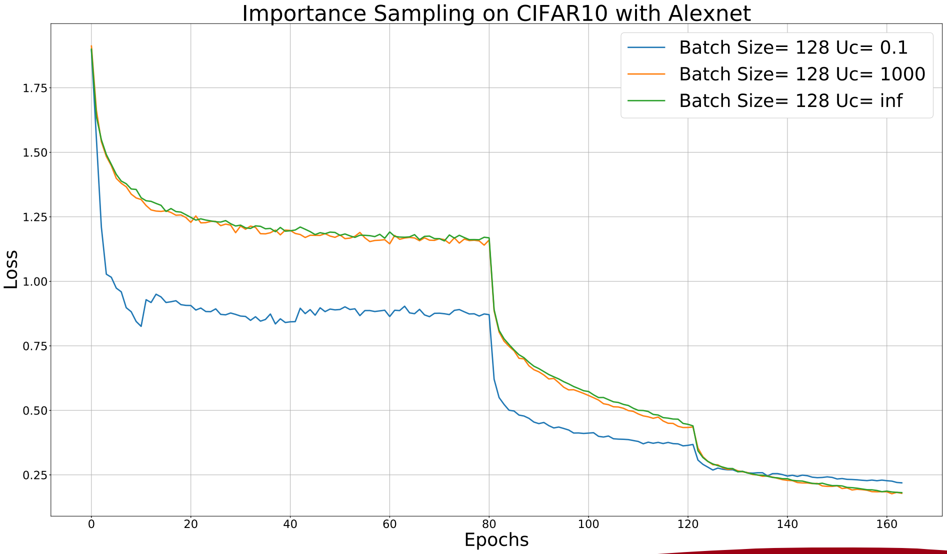
Dynamic outliers: data points that are not well fitted by the model under the current parameter θ_k

Importance sampling strategy:

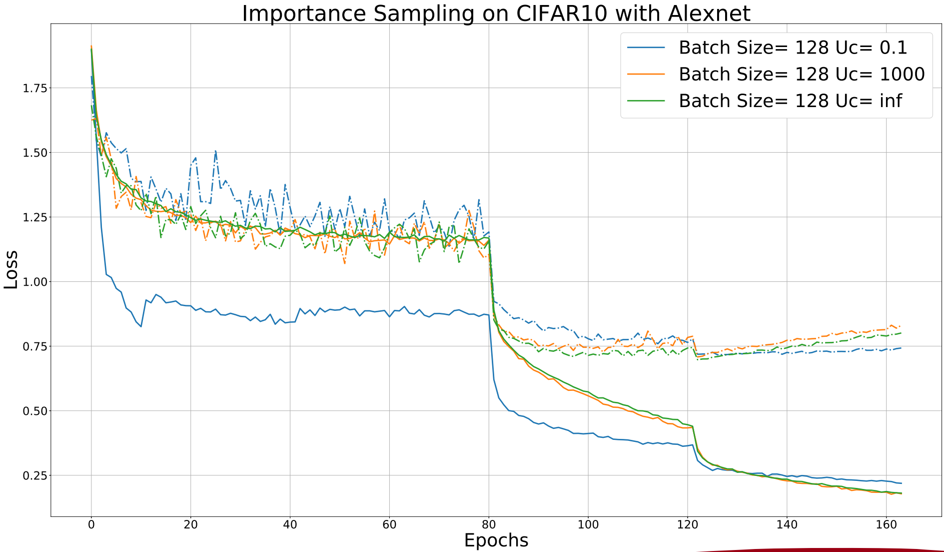
$$p(\mathbf{x}_i) := e^{-\frac{l_i(\theta_k)}{Uc}}$$

Note if $Uc \uparrow \implies$ reducing importance sampling

A new perspective: importance sampling



A new perspective: importance sampling



A new perspective: importance sampling

Conclusions:

1. Stochasticity $\implies$ escaping local minima

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD
3. Instability of high potential local minima

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD
3. Instability of high potential local minima
4. Genesis of local minima

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD
3. Instability of high potential local minima
4. Genesis of local minima
5. Importance sampling $\implies$ smoothing

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD
3. Instability of high potential local minima
4. Genesis of local minima
5. Importance sampling $\implies$ smoothing

Future work:

1. Local minima vs structure of DNNs?

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD
3. Instability of high potential local minima
4. Genesis of local minima
5. Importance sampling $\implies$ smoothing

Future work:

1. Local minima vs structure of DNNs?
2. Avoiding/Escaping local minima in higher dimension?

Conclusions:

1. Stochasticity $\implies$ escaping local minima
2. SDE can be used to describe SGD
3. Instability of high potential local minima
4. Genesis of local minima
5. Importance sampling $\implies$ smoothing

Future work:

1. Local minima vs structure of DNNs?
2. Avoiding/Escaping local minima in higher dimension?
3. Different importance sampling schemes

Thank you for your attention!