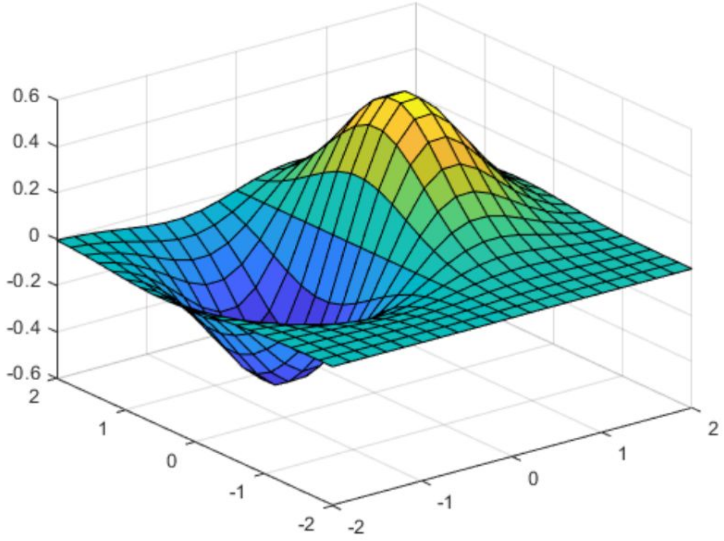
Jak zefektivnit učení NN

Inicializace vah



Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem

Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem
- **Jak dobrá je tato inicializace?**

Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem (odchylkou)
- **Jak dobrá je tato inicializace?**
    - Představme si naší síť, která má 28x28=784 neuronů ve vstupní vrstvě

Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem (odchylkou)
- **Jak dobrá je tato inicializace?**
    - Představme si naší síť, která má 28x28=784 neuronů ve vstupní vrstvě
    - Vezměme příklad, kde je polovina vstupních neuronů je 0 a druhá polovina 1

Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem (odchylkou)
- **Jak dobrá je tato inicializace?**
  - Představme si naší síť, která má 28x28=784 neuronů ve vstupní vrstvě
  - Vezměme příklad, kde je polovina vstupních neuronů je 0 a druhá polovina 1
  - Podíváme se na první neuron ve skryté vrstvě

Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem (odchylkou)
- **Jak dobrá je tato inicializace?**
  - Představme si naší síť, která má 28x28=784 neuronů ve vstupní vrstvě
  - Vezměme příklad, kde je polovina vstupních neuronů je 0 a druhá polovina 1
  - Podíváme se na první neuron ve skryté vrstvě

Inicializace vah

- Používáme np.random.randn(), tj. náhodná čísla ze standardního normálního rozdělení s nulovou střední hodnotou a jednotkových roztpylem
- **Jak dobrá je tato inicializace?**
    - Představme si naší síť, která má 28x28=784 neuronů ve vstupní vrstvě
    - Vezměme příklad, kde je polovina vstupních neuronů je 0 a druhá polovina 1
    - Podíváme se na první neuron ve skryté vrstvě
    - Jeho aktivace je dána vztahem $\sigma(z),\ z = \sum_j w_j x_j + b$

Inicializace vah

- Jeho aktivace je dána vztahem $\sigma(z),\ z = \sum_j w_j x_j + b$
- Polovina výrazů v sumě je nulová (***Xj***=0)

Inicializace vah

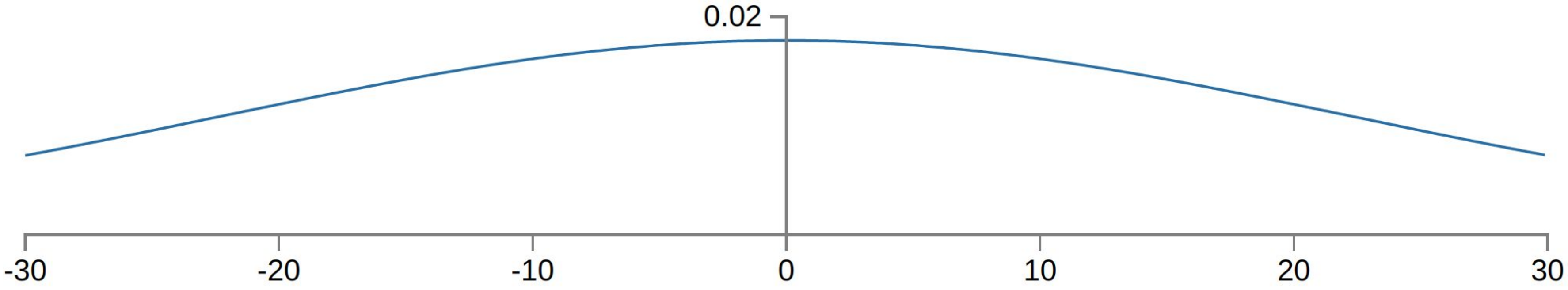
- Jeho aktivace je dána vztahem $\sigma(z),\ z = \sum_j w_j x_j + b$
- Polovina výrazů v sumě je nulová (***Xj***=0)
- ***Z*** je tedy součet 784/2 + 1 = 393 veličin ze standardního normálního rozdělení (392 vah + 1 bias)

Inicializace vah

- Jeho aktivace je dána vztahem $\sigma(z),\ z = \sum_j w_j x_j + b$
- Polovina výrazů v sumě je nulová (***Xj***=0)
- ***Z*** je tedy součet 784/2 + 1 = 393 veličin ze standardního normálního rozdělení (392 vah + 1 bias)
- ***Z*** je tedy také z normálního rozdělení s nulovou střední hodnotou, ale rozptyl = 393

Inicializace vah

- Jeho aktivace je dána vztahem $\sigma(z),\ z = \sum_j w_j x_j + b$
- Polovina výrazů v sumě je nulová (***Xj***=0)
- ***Z*** je tedy součet 784/2 + 1 = 393 veličin ze standardního normálního rozdělení (392 vah + 1 bias)
- ***Z*** je tedy také z normálního rozdělení s nulovou střední hodnotou, ale rozptyl = 393



Inicializace vah

- Z grafu vidíme, že $|z|$ je s velkou pravděpodobností velké, tj. $|z| \gg 1 \implies \sigma(z) \approx 0 \text{ nebo } \sigma(z) \approx 1$

Inicializace vah

- Z grafu vidíme, že $|z|$ je s velkou pravděpodobností velké, tj. $|z| \gg 1 \implies \sigma(z) \approx 0 \text{ nebo } \sigma(z) \approx 1$
- V takovém případě je neuron saturovaný a učení bude pomalé

Inicializace vah

- Z grafu vidíme, že $|z|$ je s velkou pravděpodobností velké, tj. $|z| \gg 1 \implies \sigma(z) \approx 0 \text{ nebo } \sigma(z) \approx 1$
- V takovém případě je neuron saturovaný a učení bude pomalé
- Malé změny ve vahách budou mít miniaturní vliv na aktivaci neuronu, což bude mít ještě menší dopad na neurony v dalších vrstvách a tudíž i na snížení hodnoty účelové funkce

Inicializace vah

- Z grafu vidíme, že $|z|$ je s velkou pravdepodobností velké, tj. $|z| \gg 1 \implies \sigma(z) \approx 0 \text{ nebo } \sigma(z) \approx 1$
- V takovém případě je neuron saturovaný a učení bude pomalé
- Malé změny ve vahách budou mít miniaturní vliv na aktivaci neuronu, což bude mít ještě menší dopad na neurony v dalších vrstvách a tudíž i na snížení hodnoty účelové funkce => to platí pro všechny neurony v síti

Inicializace vah

- Z grafu vidíme, že $|z|$ je s velkou pravdepodobností velké, tj. $|z| \gg 1 \implies \sigma(z) \approx 0 \text{ nebo } \sigma(z) \approx 1$
- V takovém případě je neuron saturovaný a učení bude pomalé
- Malé změny ve vahách budou mít miniaturní vliv na aktivaci neuronu, což bude mít ještě menší dopad na neurony v dalších vrstvách a tudíž i na snížení hodnoty účelové funkce => to platí pro všechny neurony v síti
- (cross entropie nám pomohla tento problém řešit pouze ve výstupní vrstvě)

Inicializace vah

Jak tedy váhy inicializovat?

Inicializace vah

Jak tedy váhy inicializovat?

- Nepoužijeme standardní normální rozdělení, ale normální rozdělení s nulovou střední hodnotou a rozptylem $\frac{1}{n_{l-1}}$, přičemž n_{l-1} je počet neuronů v předchozí vrstvě

Inicializace vah

Jak tedy váhy inicializovat?

- Nepoužijeme standardní normální rozdělení, ale normální rozdělení s nulovou střední hodnotou a rozptylem $\frac{1}{n_{l-1}}$, přičemž n_{l-1} je počet neuronů v předchozí vrstvě
- Bias je jen jeden, takže na výsledek nemá velký vliv => necháme jeho inicializaci tak, jak je

Inicializace vah

Jak tedy váhy inicializovat?

- Nepoužijeme standardní normální rozdělení, ale normální rozdělení s nulovou střední hodnotou a rozptylem $\frac{1}{n_{l-1}}$, přičemž n_{l-1} je počet neuronů v předchozí vrstvě
- Bias je jen jeden, takže na výsledek nemá velký vliv => necháme jeho inicializaci tak, jak je
- Takto získáme $z \sim N\left(0, \sqrt{\frac{3}{2}}\right)$, protože

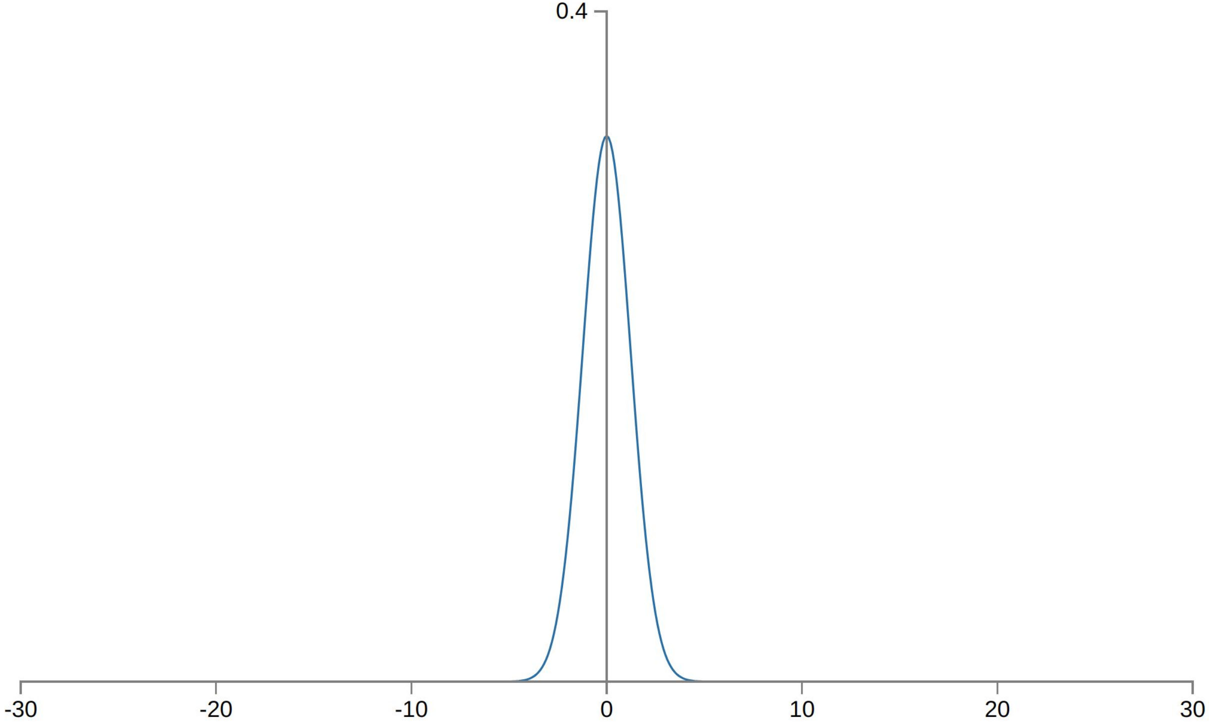
$$var(z) = var\left(\sum_j w_j x_j + b\right) = \sum_{j,\, x_j=1} var(w_j) + var(b_j) =$$

$$= \sum_{j,\, x_j=1} \frac{1}{n_{l-1}} + 1 = \frac{n_{l-1}}{2} \cdot \frac{1}{n_{l-1}} + 1 = \frac{3}{2}$$

Inicializace vah

Jak tedy váhy inicializovat?

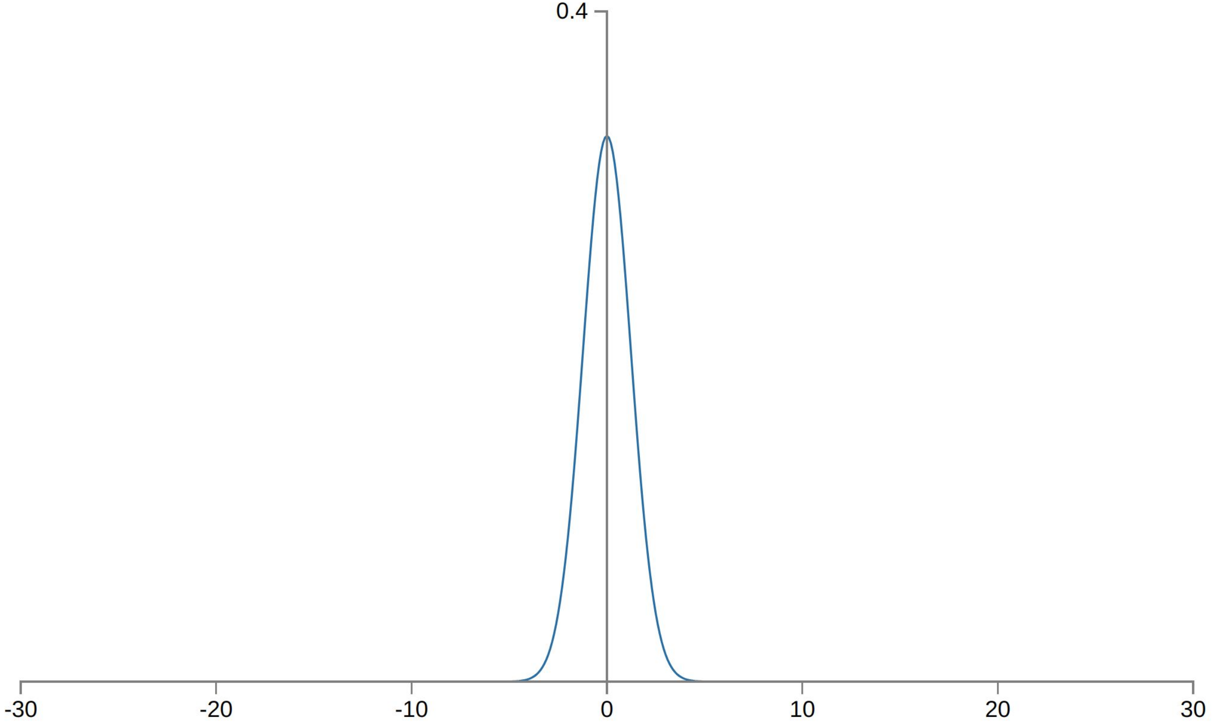
- $z \sim N\left(0, \sqrt{\frac{3}{2}}\right)$



Inicializace vah

Jak tedy váhy inicializovat?

- $z \sim N\left(0, \sqrt{\frac{3}{2}}\right)$
- takové neurony je těžké saturovat a tím pádem je nepravděpodobné, že dojde ke zpomalení učení



Problém

- L2 regularizace dělá něco podobného, jako tato inicializace vah. Nechť lambda není příliš malá.

Problém

- L2 regularizace dělá něco podobného, jako tato inicializace vah. Nechť lambda není příliš malá.
    a. Vysvětlete, proč při použití původní inicializace vah, probíhá během prvních epoch především “weight decay” (zmenšování vah)

Problém

- L2 regularizace dělá něco podobného, jako tato inicializace vah. Nechť lambda není příliš malá.
    a. Vysvětlete, proč při použití původní inicializace vah, probíhá během prvních epoch především “weight decay” (zmenšování vah)

Update vah je dán předpisem $w \rightarrow \left(1 - \frac{\eta\lambda}{n}\right) w - \frac{\eta}{m} \sum_x \frac{\partial C_x}{\partial w}$. Pro velké váhy bude první část výrazu dominovat, neboť váhy byly inicializované náhodně, takže se suma parciálních derivací bude chovat chaoticky.

Problém

- L2 regularizace dělá něco podobného, jako tato inicializace vah. Nechť lambda není příliš malá.
    a. Vysvětlete, proč při použití původní inicializace vah, probíhá během prvních epoch především “weight decay” (změnšování vah)
    b. Nechť $\eta\lambda \ll n$. Ukažte, že váhy se každou epochu zmenší $e^{-\frac{\eta\lambda}{m}}$ krát.

Problém

Nechť $\eta\lambda \ll n$. Ukažte, že váhy se každou epochu zmenší $e^{-\frac{\eta\lambda}{m}}$ krát $w \rightarrow \left(1 - \frac{\eta\lambda}{n}\right) w - \frac{\eta}{m} \sum_x \frac{\partial C_x}{\partial w}$

- Během každé epochy se výraz $\left(1 - \frac{\eta\lambda}{n}\right)$ vynásobí n/m krát. Váha se tedy zmenší $\left(1 - \frac{\eta\lambda}{n}\right)^{\frac{n}{m}}$ krát

Problém

Nechť $\eta\lambda \ll n$. Ukažte, že váhy se každou epochu zmenší $e^{-\frac{\eta\lambda}{m}}$ krát $w \rightarrow \left(1-\frac{\eta\lambda}{n}\right)w - \frac{\eta}{m}\sum_x \frac{\partial C_x}{\partial w}$

- Během každé epochy se výraz $\left(1-\frac{\eta\lambda}{n}\right)$ vynásobí n/m krát. Váha se tedy zmenší $\left(1-\frac{\eta\lambda}{n}\right)^{\frac{n}{m}}$ krát
- Vyjdeme z definice e:

$$e = \lim_{n\to+\infty}\left(1+\frac{1}{n}\right)^n = \lim_{n\to+\infty}\left(1-\frac{1}{n}\right)^{-n}, \text{ protože } \left(1-\frac{1}{n}\right)^{-n} = \left(\frac{n}{n-1}\right)^n = \left(1+\frac{1}{n-1}\right)^n$$

Problém

Nechť $\eta\lambda \ll n$. Ukažte, že váhy se každou epochu zmenší $e^{-\frac{\eta\lambda}{m}}$ krát $w \to \left(1 - \frac{\eta\lambda}{n}\right) w - \frac{\eta}{m} \sum_x \frac{\partial C_x}{\partial w}$

- Během každé epochy se výraz $\left(1 - \frac{\eta\lambda}{n}\right)$ vynásobí n/m krát. Váha se tedy zmenší $\left(1 - \frac{\eta\lambda}{n}\right)^{\frac{n}{m}}$ krát
- Vyjdeme z definice e:

$$e = \lim_{n \to +\infty} \left(1 + \frac{1}{n}\right)^n = \lim_{n \to +\infty} \left(1 - \frac{1}{n}\right)^{-n}, \text{ protože } \left(1 - \frac{1}{n}\right)^{-n} = \left(\frac{n}{n-1}\right)^n = \left(1 + \frac{1}{n-1}\right)^n$$

- přepíšeme

$$\left(1 - \frac{\eta\lambda}{n}\right)^{\frac{n}{m}} = \left(\left(1 - \frac{\eta\lambda}{n}\right)^{-\frac{n}{\eta\lambda}}\right)^{-\frac{\eta\lambda}{m}} \approx e^{-\frac{\eta\lambda}{m}}, \text{ pokud } \eta\lambda \ll n$$

Problém

- L2 regularizace dělá něco podobného, jako tato inicializace vah. Nechť lambda není příliš malá.
    a. Vysvětlete, proč při použití původní inicializace vah, probíhá během prvních epoch především “weight decay” (změnšování vah)
    b. Nechť $\eta\lambda \ll n$. Ukažte, že váhy se každou epochu zmenší $e^{-\frac{\eta\lambda}{m}}$ krát.
    c. Nechť lambda není příliš velká.
    Ukažte, že váhy se přestanou snižovat okolo velikosti $\frac{1}{\sqrt{N_w}}$, přičemž ***Nw*** je celkový počet vah v síti.

Problém

Nechť lambda není příliš velká. Ukažte, že váhy se přestanou snižovat okolo velikosti $\frac{1}{\sqrt{N_w}}$, přičemž ***Nw*** je celkový počet vah v síti.

- Odhadneme váhy $w \approx (N_w)^k$, chceme prozkoumat, kdy bude přínos regularizačního výrazu v účelové funkci $C = C_0 + \frac{\lambda}{2n} \sum_w w^2$ rozumně malý:

Problém

Nechť lambda není příliš velká. Ukažte, že váhy se přestanou snižovat okolo velikosti $\frac{1}{\sqrt{N_w}}$, přičemž ***Nw*** je celkový počet vah v síti.

- Odhadneme váhy $w \approx (N_w)^k$, chceme prozkoumat, kdy bude přínos regularizačního výrazu v účelové funkci $C = C_0 + \frac{\lambda}{2n}\sum_w w^2$ rozumně malý:

$$k = 1: \sum_w w^2 = N_w^3$$

$$k = -\frac{1}{2}: \sum_w w^2 = 1$$

$$k = -1: \sum_w w^2 = \frac{1}{N_w}$$

$$\vdots$$