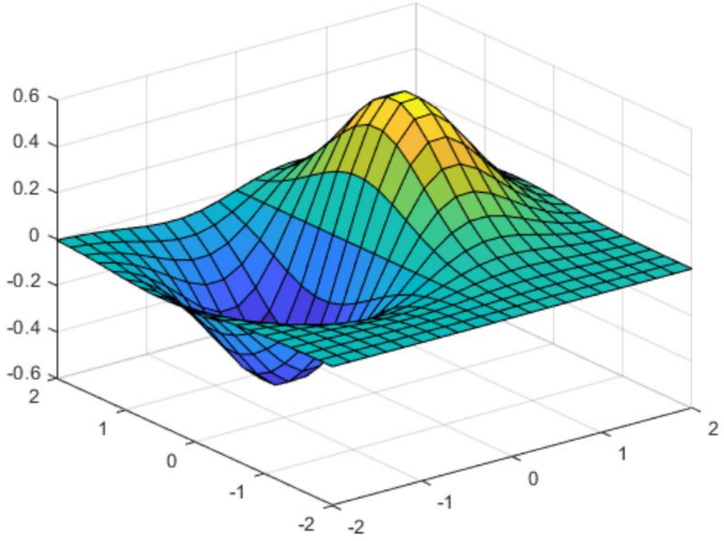
Jak zefektivnit učení NN

Variace SGD



Variace SGD

- Existují jiné způsoby (než SGD + backpropagation) jak minimalizovat C

Variace SGD

- Existují jiné způsoby (než SGD + backpropagation) jak minimalizovat C
- např. metody využívající **Hessovu matici** (matice druhých parciálních derivací) nebo **setrvačnost (momentum)**

Variace SGD - Hessova matice

- Odhlédneme od NN, řešíme úlohu minimalizace $C=C(w)$, w je vektor n proměnných

Variace SGD - Hessova matice

- Odhlédneme od NN, řešíme úlohu minimalizace ***C=C(w)***, ***w*** je vektor ***n*** proměnných
- Taylorovým rozvojem v okolí bodu ***w*** získáme

$$
\begin{aligned}
C(w+\Delta w) &= C(w) + \sum_j \frac{\partial C}{\partial w_j}\Delta w_j + \frac{1}{2}\sum_j\sum_k \Delta w_j \frac{\partial^2 C}{\partial w_j \partial w_k}\Delta w_k + \ldots = \\
&= C(w) + \nabla C \cdot \Delta w + \frac{1}{2}\Delta w^T H \Delta w + \ldots = \\
&= \\
&= C(w) + \nabla C \cdot \Delta w + \frac{1}{2}\Delta w^T H \Delta w + \mathcal{O}\left(\|\Delta w\|^2\right)
\end{aligned}
$$

Variace SGD - Hessova matice

- Odhlédneme od NN, řešíme úlohu minimalizace ***C=C(w)***, ***w*** je vektor ***n*** proměnných
- Taylorovým rozvojem v okolí bodu ***w*** získáme

$$C(w+\Delta w) = C(w) + \sum_j \frac{\partial C}{\partial w_j}\Delta w_j + \frac{1}{2}\sum_j\sum_k \Delta w_j \frac{\partial^2 C}{\partial w_j \partial w_k}\Delta w_k + \ldots =$$

$$= C(w) + \nabla C \cdot \Delta w + \frac{1}{2}\Delta w^T H \Delta w + \ldots =$$

$$=$$

$$= C(w) + \nabla C \cdot \Delta w + \frac{1}{2}\Delta w^T H \Delta w + \mathcal{O}\Big(\|\Delta w\|^2\Big)$$

- Tento výraz minimalizuje $\Delta w = -H^{-1}\nabla C$ (pokud je H pozitivně definitní)

Variace SGD - Hessova matice

- Pravidlo pro update: $w \rightarrow \tilde{w} = w + \eta \Delta w = w - \eta H^{-1} \nabla C$

Variace SGD - Hessova matice

- Pravidlo pro update: $w \rightarrow \tilde{w} = w + \eta \Delta w = w - \eta H^{-1} \nabla C$
- eta je learning rate, podobně jako u SGD

Variace SGD - Hessova matice

- Pravidlo pro update: $w \to \tilde{w} = w + \eta \Delta w = w - \eta H^{-1} \nabla C$
- eta je learning rate, podobně jako u SGD
- Konverguje rychleji, než GD

Variace SGD - Hessova matice

- Pravidlo pro update: $w \rightarrow \tilde{w} = w + \eta \Delta w = w - \eta H^{-1} \nabla C$
- eta je learning rate, podobně jako u SGD
- Konverguje rychleji, než GD
- Díky druhým derivacím se vyhne spoustě problémů, kterými GD trpí

Variace SGD - Hessova matice

- Pravidlo pro update: $w \rightarrow \tilde{w} = w + \eta \Delta w = w - \eta H^{-1} \nabla C$
- eta je learning rate, podobně jako u SGD
- Konverguje rychleji, než GD
- Díky druhým derivacím se vyhne spoustě problémů, kterými GD trpí
- Existují i variace backpropagation algoritmu pro počítání Hessovy matice

Variace SGD - Hessova matice

- Pravidlo pro update: $w \to \tilde{w} = w + \eta \Delta w = w - \eta H^{-1} \nabla C$
- eta je learning rate, podobně jako u SGD
- Konverguje rychleji, než GD
- Díky druhým derivacím se vyhne spoustě problémů, kterými GD trpí
- Existují i variace backpropagation algoritmu pro počítání Hessovy matice
- **Nepraktické - řád Hessovy matice roste v závislosti na počtu parametrů s druhou mocninou**

Variace SGD - Hessova matice

- Pravidlo pro update: $w \to w + \eta \Delta w = w - \eta H^{-1} \nabla C$
- eta je learning rate, podobně jako u SGD
- Konverguje rychleji, než GD
- Díky druhým derivacím se vyhne spoustě problémů, kterými GD trpí
- Existují i variace backpropagation algoritmu pro počítání Hessovy matice
- **Nepraktické - řád Hessovy matice roste v závislosti na počtu parametrů s druhou mocninou**
- Existují “řešení” tohoto problému (truncated/quasi Newton)

Variace SGD - Momentum

- Využití druhých parciálních derivací je výhodné proto, že kromě informace o gradientu získáme i informace o tom, jak se gradient mění

Variace SGD - Momentum

- Využití druhých parciálních derivací je výhodné proto, že kromě informace o gradientu získáme i informace o tom, jak se gradient mění
- Zavedení setrvačnosti vychází ze stejné úvahy, ale chce se vyhnout použití Hessovy matice

Variace SGD - Momentum

- Využití druhých parciálních derivací je výhodné proto, že kromě informace o gradientu získáme i informace o tom, jak se gradient mění
- Zavedení setrvačnosti vychází ze stejné úvahy, ale chce se vyhnout použití Hessovy matice
- Zpřísníme fyzikální analogii GD jako “míče kutálejícího se dolů”

Variace SGD - Momentum

- Využití druhých parciálních derivací je výhodné proto, že kromě informace o gradientu získáme i informace o tom, jak se gradient mění
- Zavedení setrvačnosti vychází ze stejné úvahy, ale chce se vyhnout použití Hessovy matice
- Zpřísníme fyzikální analogii GD jako “míče kutálejícího se dolů”
- Zavedeme “rychlost” změny parametrů (vychází z gradientu), “pozici” (závisí i na rychlosti) a “tření”, které rychlost postupně snižuje

Variace SGD - Momentum

- Matematická formulace:

  Zavedeme vektor rychlostí, jehož složky korespondují s parametry, který optimalizujeme (tj. váhy i biasy) $v = v_1, v_2, \ldots, v_n$ a v GD nahradíme pravidlo pro update vah (biasů) $w \to \tilde{w} = w - \eta \nabla C$ za

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

přičemž μ se nazývá koeficient setrvačnosti

Variace SGD - Momentum

- Matematická formulace:

  Zavedeme vektor rychlostí, jehož složky korespondují s parametry, který optimalizujeme (tj. váhy i biasy) $v = v_1, v_2, \ldots, v_n$ a v GD nahradíme pravidlo pro update vah (biasů) $w \rightarrow \tilde{w} = w - \eta \nabla C$ za

$$v \rightarrow \tilde{v} = \mu v - \eta \nabla C$$
$$w \rightarrow \tilde{w} = w + \tilde{v}$$

přičemž μ se nazývá koeficient setrvačnosti

- Gradient nyní ovlivňuje rychlost ***v*** a ta ovlivňuje změnu vah (biasů)

Variace SGD - Momentum

- Matematická formulace:

  Zavedeme vektor rychlostí, jehož složky korespondují s parametry, který optimalizujeme (tj. váhy i biasy) $v = v_1, v_2, \ldots, v_n$ a v GD nahradíme pravidlo pro update vah (biasů) $w \rightarrow \tilde{w} = w - \eta \nabla C$ za

  $$v \rightarrow \tilde{v} = \mu v - \eta \nabla C$$
  $$w \rightarrow \tilde{w} = w + \tilde{v}$$

  přičemž μ se nazývá koeficient setrvačnosti

- Gradient nyní ovlivňuje rychlost ***v*** a ta ovlivňuje změnu vah (biasů)
- Rychlost roste s tím, jak k ní přičítáme gradient => pokud má stejný směr přes několik batchů, rychlost v tomto směru významně naroste a změna vah v tomto směru zrychluje

Variace SGD - Momentum

$$v \rightarrow \tilde{v} = \mu v - \eta \nabla C$$
$$w \rightarrow \tilde{w} = w + \tilde{v}$$

- V blízkosti minima můžeme snadno “přestřelit”

Variace SGD - Momentum

$$v \rightarrow \tilde{v} = \mu v - \eta \nabla C$$
$$w \rightarrow \tilde{w} = w + \tilde{v}$$

- V blízkosti minima můžeme snadno “přestřelit”
- Další problém může nastat, pokud se gradient rapidně změní - potrvá nám, než na tuto změnu zareagujeme

Variace SGD - Momentum

$$v \rightarrow \tilde{v} = \mu v - \eta \nabla C$$
$$w \rightarrow \tilde{w} = w + \tilde{v}$$

- V blízkosti minima můžeme snadno “přestřelit”
- Další problém může nastat, pokud se gradient rapidně změní - potrvá nám, než na tuto změnu zareagujeme
- Z tohoto důvodu zavádíme koeficient setrvačnosti (“tření” je dáno vztahem $1 - \mu$)

Variace SGD - Momentum

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

- V blízkosti minima můžeme snadno “přestřelit”
- Další problém může nastat, pokud se gradient rapidně změní - potrvá nám, než na tuto změnu zareagujeme
- Z tohoto důvodu zavádíme koeficient setrvačnosti (“tření” je dáno vztahem $1 - \mu$)
- $\mu \in \langle 0; 1 \rangle$

Variace SGD - Momentum

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

- V blízkosti minima můžeme snadno “přestřelit”
- Další problém může nastat, pokud se gradient rapidně změní - potrvá nám, než na tuto změnu zareagujeme
- Z tohoto důvodu zavádíme koeficient setrvačnosti (“tření” je dáno vztahem $1 - \mu$)
- $\mu \in \langle 0; 1 \rangle$
- Pokud $\mu = 1$, není žádné tření a rychlost se plně řídí gradientem. Pokud $\mu = 0$, dostáváme obyčejný GD

Variace SGD - Momentum

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

- V blízkosti minima můžeme snadno “přestřelit”
- Další problém může nastat, pokud se gradient rapidně změní - potrvá nám, než na tuto změnu zareagujeme
- Z tohoto důvodu zavádíme koeficient setrvačnosti (“tření” je dáno vztahem $1 - \mu$)
- $\mu \in \langle 0; 1 \rangle$
- Pokud $\mu = 1$, není žádné tření a rychlost se plně řídí gradientem. Pokud $\mu = 0$, dostáváme obyčejný GD
- Máme další hyperparametr, který je třeba vhodně zvolit

Variace SGD - Momentum

- Momentum + L2 regularizace:

$$v_b \to \tilde{v}_b = \mu v_b - \eta \frac{\partial C}{\partial b} = \mu v_b - \eta \frac{\partial C_0}{\partial b}$$

$$b \to \tilde{b} = b + \tilde{v}_b$$

$$v_w \to \tilde{v}_w = \mu v_w - \eta \frac{\partial C}{\partial w} = \mu v_w - \eta \left(\frac{\partial C_0}{\partial w} + \frac{\lambda}{n} w \right)$$

$$w \to \tilde{w} = w + \tilde{v}_w$$

SGD with momentum: Cvičení

$$v \rightarrow \tilde{v} = \mu v - \eta \nabla C$$
$$w \rightarrow \tilde{w} = w + \tilde{v}$$

- **Cvičení 1: Co se stane, pokud použijeme koeficient setrvačnosti >1?**

SGD with momentum: Cvičení

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

- **Cvičení 1: Co se stane, pokud použijeme koeficient setrvačnosti >1?**
  "Záporné tření" by způsobilo nekontrolovatelný nárůst rychlosti (i v případě, že by byl gradient vždy nulový), což by pravděpodobně vedlo k "přestřelení" minima

SGD with momentum: Cvičení

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

- **Cvičení 2: Co se stane, pokud použijeme koeficient setrvačnosti <0?**

SGD with momentum: Cvičení

$$v \to \tilde{v} = \mu v - \eta \nabla C$$
$$w \to \tilde{w} = w + \tilde{v}$$

- **Cvičení 2: Co se stane, pokud použijeme koeficient setrvačnosti <0?**
  Rychlost by mohla oscilovat (obzvlášť, pokud bude gradient malý) mezi pozitivními a negativními hodnotami, přičemž záporná hodnota by efektivně znamenala pohyb v opačném směru, než k minimu