Stochastic Gradient Descent

n huge -> is a problem

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i : \mathbb{R}^m \to \mathbb{R}$$

$$\nabla f = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x)$$

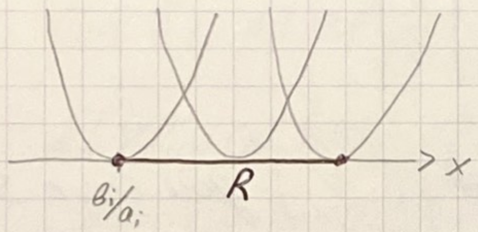
$$\text{SGD}: \quad x^{(k+1)} = x^{(k)} + \alpha^{(k)} \nabla f_{i(k)}(x^{(k)})$$

↳ randomly choosen index

$$\to f(x^{(k+1)}) \overset{?}{<} f(x^{(k)})$$

$\boxed{1}$ Motivation:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2, \quad x \in \mathbb{R}, \; a_i, b_i \in \mathbb{R}$$



$$\nabla f = f' = \sum_{i=1}^{n} a_i(a_i x - b_i) = 0$$

$$\to x^* = \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i^2} \quad \text{global minimum}$$

$$\nabla f_i = f_i' = a_i(a_i x - b_i) = 0 \iff$$

$$\to x_i^* = \frac{b_i}{a_i} \quad \text{global minimum for } f_i = \tfrac{1}{2}(a_i x + b_i)$$

$$R = [\min x_i^*, \max x_i^*]$$

Note that $x^* \in R$ related so that $R = [\frac{b_1}{a_1}, \frac{b_n}{a_n}]$ → $x^* = \frac{\sum a_i b_i}{\sum a_i^2} \leq \frac{b_n}{a_n}$

↳ rearrangement inequality

If the starting point $x^{(0)}$ for SGD is outside of R, then $\nabla f(x^{(0)})$ and $\nabla f_i(x^{(0)})$ have the same signs.

↳ SGD update is 'the same' (in the same direction) as GD

R = region of uncertainty

$\boxed{2}$ Elements of probability theory

X $\underbrace{\text{random variables}}_{\text{'function'}}$ (discrete)

↙ Value of x

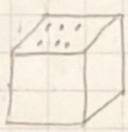
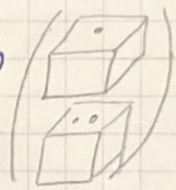
x_1	x_2	x_3	x_4	x_5	$\dots$	x_K
p_1	p_2	p_3	p_4	p_5	$\dots$	p_K

probabilities, $p_i \in [0,1]$

$$P(X = x_i) = p_i \,, \quad p_1 + p_2 + \dots + p_K = 1$$

(a)  $X = \#$(dots on the top after 1 throw)

1	2	3	4	5	6
$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

(b) If  you win 10€, if  you loose 5€

Otherwise you loose 1€. Y = profit.

10	-5	-1
$1/6$	$1/3$	$1/2$

$$\tfrac{1}{3} = \tfrac{1}{6} + \tfrac{1}{6}$$

Expectation of a random variable

$$\mathbb{E}[X] = \sum_{i=1}^{K} x_i p_i$$

'mean value' (if uniform distribution $\Leftrightarrow p_i = \frac{1}{K} \to \mathbb{E}[X] = \frac{1}{K}\sum_{i=1}^{K} x_i$)

(a) $\mathbb{E}[X_a] = \frac{1}{6}[1+2+3+4+5+6] = \frac{21}{6} = 3.5$

(b) $\mathbb{E}[Y] = \frac{1}{6}\cdot 10 + \frac{1}{3}(-5) + \frac{1}{2}(-1) = -\frac{1}{2}$

Properties:

(1) $\mathbb{E}[const] = const$

(2) $\mathbb{E}[cX + c'X'] = c\mathbb{E}[X] + c'\mathbb{E}[X']$

(3) $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$

$Var(X) = \mathbb{E}(X - \mathbb{E}[X])^2$ variance

$\boxed{3}$ Stochastic gradients:

Random variable $g(x)$ s.t. $\mathbb{E}[g(x)] = \nabla f(x)$

Example: Assume that $i(k)$ is choosen uniformly at random from $\{1 \dots n\} \to \mathbb{P}(i(x) = s) = 1/n$

s: some number from $\{1, \dots, n\}$

$$\mathbb{E}[\underbrace{\nabla f_{i(x)}(x)}_{\text{discrete random variable}}] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x) = \nabla f(x)$$

$\nabla f_1(x)$	$\nabla f_2(x)$	$\dots$	$\nabla f_n(x)$
$1/n$	$1/n$	$\dots$	$1/n$

There are at least 2 types to chose $i(k)$:

Theory ▷ uniformly at random

Pratice ▷ uniformly at random without shuffling back $n \gg$ # steps in SGD

Mini-batch appoach

$$x^{(k+1)} = x^{(k)} - \frac{a^{(k)}}{|I^k|} \sum_{j \in I^k} \nabla f_j(x^{(k)})$$

I^k: subset of $\{1..n\}$ chosen at random

Exercise: Suppose $|I_k| = 2$, and it chosen uniformly at random. Show that Variance of $\frac{1}{|I_k|}\sum_{j \in I_k} \nabla f_j(x)$ is smaller than $\nabla f_{i(k)}(x)$

$\boxed{4}$ Convergance: $x^{(k+1)} = x^{(k)} + a^{(k)} \nabla f_{i(k)}(x^{(k)})$

$$\mathbb{E}[\nabla f_{i(k)}(x^{(k)})] = \sum_{s=1}^{n} \nabla f_s (x^{(k)}) \frac{\mathbb{P}(i(k)=s, (i(k-1), i(k-2), \dots i(0)) = (s_{k-1}, s_{k-2}, \dots, 0))}{\mathbb{P}((i(k-1), \dots, i(0)) = (s_{k-1}, \dots, 0))}$$

depends on k random choices

$$= \nabla f(x^{(k)})$$

conditional expectation: X, Y

$$\mathbb{E}[X \mid Y=y] = \sum x_i \frac{\mathbb{P}(X=x_i, Y=y)}{\mathbb{P}(Y=y)}$$

conditioning

$= [$if X & Y are independent, then $\mathbb{P}(X=x_i, Y=y) = \mathbb{P}(X=x_i)\mathbb{P}(Y=y)]$

$= \mathbb{E}[X]$

If $s_0, s_1, \dots, s_{k-1}$ are such that

$$x^{(k)} = x^{(k-1)} - a^{(k-1)} \nabla f_{s_{k-1}}(x^{(k-1)}),\quad x^{(k-1)} = x^{(k-2)} - a^{(k-2)} \nabla f_{s_{k-2}}(x^{(k-2)})$$

(a) G.D.: Assumption $|u^T \nabla^2 f(x) \cdot u| \leq L \|u\|^2$

$$x^{(k+1)} = x^{(k)} - a^{(k)} \nabla f(x^{(k)})$$

By Taylor:

$$f(x^{(k+1)}) = f(x^{(k)} - a^{(k)} \nabla f(x^{(k)}))$$

$$= f(x^{(k)}) - a^{(k)} \langle \nabla f(x^{(k)}), \nabla f(x^{(k)}) \rangle + \frac{1}{2} (a^{(k)} \nabla f(x^{(k)}))^T \underbrace{\nabla^2 f}_{\text{Hessian}}(\underbrace{\xi_k}_{\text{a point in } (x^{(k)}, x^{(k+1)})})(a^{(k)} \nabla f(x^{(k)}))$$

$$\leq f(x^{(k)}) - a^{(k)} \|\nabla f(x^{(k)})\|^2 + \frac{(a^{(k)})^2 L}{2} \|\nabla f(x^{(k)})\|^2$$

$$= f(x^{(k)}) - a^{(k)} \underbrace{\left(1 - \frac{a^{(k)} L}{2}\right)}_{\geq \frac{1}{2}} \underbrace{\|\nabla f^{(k)}\|^2}_{\geq 0}$$

If $a^{(k)} \leq \frac{1}{L}$ then,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{a^{(k)}}{2} \|\nabla f(x^{(k)})\|^2 \quad (*)$$

$f(x^{(k+1)}) < f(x^{(k)}) \to$ We descent!

Note further: assume $a^{(k)} = a \;\forall k$:

$$(\#)\ \frac{1}{2} a \sum_{k=0}^{T-1} \|\nabla f(x^{(k)})\|^2 \underset{(*)}{\leq} \sum_{k=0}^{T-1} f(x^{(k)}) - f(x^{(k-1)}) = f(x^{(0)}) - f(x^{(T)})$$

$$\leq f(x^{(0)}) - f^* \quad \to$$

$f(x^{T}) \geq f^*$ is the local minimum

$$\Rightarrow \min_{k=0,\dots,T} \|\nabla f(x^{(k)})\|^2 \leq \frac{1}{T} \sum_{k=0}^{T-1} \|\nabla f(x^{(k)})\|^2 \underset{(\#)}{\leq} \frac{2(f(x^{(0)})-f^*)}{Ta} \xrightarrow[T\to\infty]{} 0$$

(b) SGD

$$f(x^{(k+1)}) = f(x^{(k)} - a^{(k)}\nabla f_{i(k)}(x^{(k)})) \leq f(x^{(k)}) - a^{(k)}\langle \nabla f(x^{(k)}), \nabla f_{i(k)}(x^{(k)})\rangle + \frac{(a^{(k)})^2 L}{2} \|\nabla f_{i(k)}(x^{(k)})\|^2$$

$$\mathbb{E}[f(x^{(k+1)})] \leq \mathbb{E}[f(x^{(k)})] - a^{(k)}\|\nabla f(x^{(k)})\|^2 + \frac{(a^{(k)})^2 L}{2} \mathbb{E}[\|\nabla f_{i(k)}(x^{(k)})\|^2]$$