$\|y\| = 1, \; f'(\vec{a}, \vec{y}) = \|\nabla f(\vec{a})\| \cdot \cos\theta$

θ angle between $f(\vec{a})$

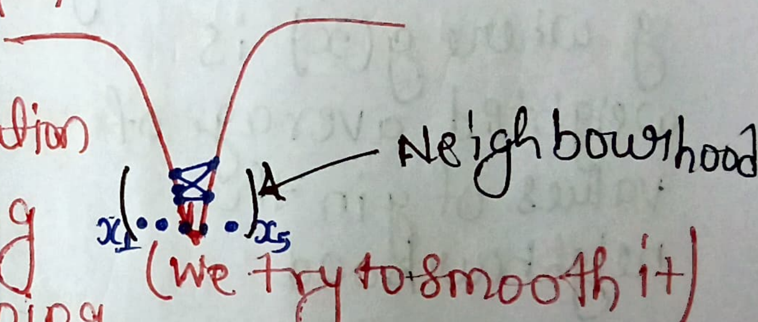
o
o
o
o

⚠ {Iske Baad Kya padhya No IDEA}

Lecture-28 (6 November '24)

Gaussian Smoothing

$f(x)$



→ Tries to find a convex approximation of f by employing Gaussian smothening

→ Help to reach local optima and global optima quickly

we replace

$f(x)$ = weighted average of value in its neighbourhood in its domain

$$g = f(x) = \sum_{i=1}^{5} w_i x_i$$

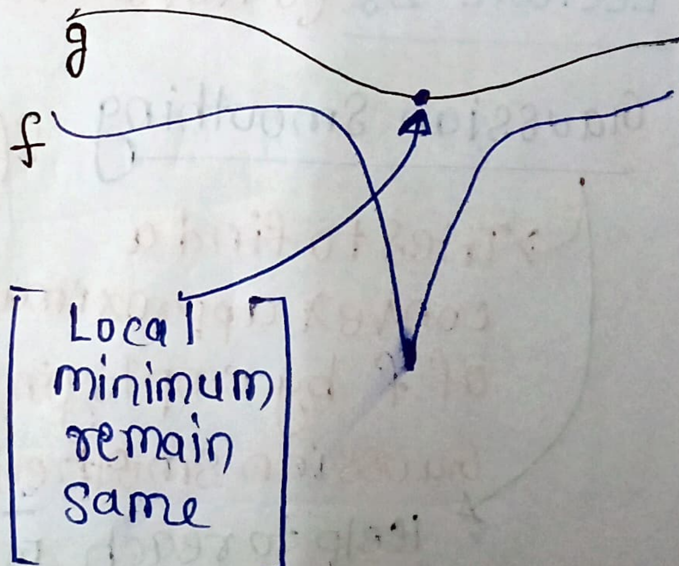
- If values of f(x) are almost same in a range i.e flat then they are not affected much by using weighted average

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{(-x-\mu)^2}{2\sigma^2}}$$

μ: mean

σ^2: valley

- f is replaced by g where g(x) is weighted average of values of y in the neighbourhood



- Has the effect of smoothing out sudden dips or ascents in the value of f

- often finds even a global minimum (as against a local one) even for non-convex function f

- The weights are choosen using Gaussian distribution

Stochastic Gradient Descent

- Effective tool for functions 'f' of the form

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

- Example: least square minimize

$$y = \sum_{i=1}^{d} \theta_i x_i$$

$$\cancel{f\,\text{deviation}} = \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{d} \theta_j x_j\Big)^2$$

Deviation

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{d} \theta_j x_j\Big)^2$$

$$= \sum_{i=1}^{n} f_i(\bar{\theta})$$

$$\boxed{f(\hat{x}_i) = \Big(y_i - \sum_{i=1}^{d} (\theta_i x_i)\Big)^2}$$

$$\nabla f(\bar{\theta}) = \sum_{i=1}^{n} \nabla f_i(\bar{\theta})$$

$$\frac{\partial f_i}{\partial \theta_j} = 2\Big(y_i - \sum_{i=1}^{n} \theta_i x_i\Big)\cdot(-x_j)$$

$$\nabla f_i = \left(\frac{\partial f_i}{\partial \theta_1}, \ldots, \frac{\partial f_i}{\partial \theta_d} \right)$$

▷ Assuming the variance is small then estimated value can be a good approximate

▷ $E[r(\vec{x})] = \frac{\sum_{i=1}^{n} n \nabla f_i(\vec{x})}{n} \doteq \nabla f(\vec{x})$

▷ $r(\vec{x})$ is an unbiased estimator of $\nabla f(\vec{x})$

some vector

Stochastic Gradient Descent Algorithm

$x_0 \leftarrow$ initial guess of $\vec{x}^*$; $k \leftarrow 0$

while $\nabla f(\vec{x}_k) \neq 0$ do

$\gamma^k \leftarrow$ an appropriate estimate of step size γ

Choose uniformly at random $i \in \{1, 2 \ldots n\}$

$\boxed{\vec{x}_{k+1} \leftarrow \vec{x}_k - \gamma_k \cdot n \nabla f_i(\vec{x}_k)}$; $k \leftarrow k+1$;

endwhile

Return $\vec{x}_k$

Random choice

similar to gradient descent

▷ If variance σ of $r(\vec{x})$ is large, one reduces it by considering the arithmatic mean of several independent sample of $r(\vec{a})$

X – random variable

$\mu = E(x)$

$\sigma^2 = \text{variance}(x)$

sample independent $X_1 \dots X_K$, according to X ~~variance of~~

$$Y = \sum_{i=1}^{K} X_i \; ; \; E(y) = \mu$$

$$\boxed{\mathrm{var}(y) = \frac{\mathrm{var}(x)}{K}}$$

▷ If variance is small, $r(\vec{x})$ will be nearly the same as $\nabla f(\vec{x})$

→ If variance is so large the requirement, ~~of random~~ K can become also very large.

▷ $r_K = r$ for each K, for pre determined constant r

→ Here in this method we don't use backtracking line search

▷ $\{\gamma_k\}$ could be a sequence satisfying $\sum_{k \geq 0} \gamma_k = \infty$ but $\sum_{k \geq 0} \gamma_k^2 < \infty$

→ Because backtracking line search is not good for Stochastic Gradient descent:

▷ Ex: $\gamma_k = \frac{c}{k}$ for some constant $c > 0$

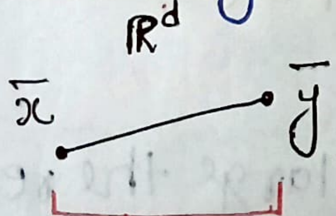
Convex optimization

① $\Omega \subseteq \mathbb{R}^d$

② Ω is a convex set

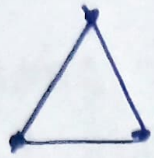
if $\forall\, x, y \in \Omega$, and for any θ lies between $[0,1]$

$$\theta \bar{x} + (1-\theta)\bar{y} \in \Omega$$



can be represented as convex combination of these two points

$$\sum_{i=1}^{k} \theta_i \bar{x}_i \qquad 0 \leq \theta \leq 1, \quad \boxed{\sum_{i=1}^{k} \theta_i = 1}$$

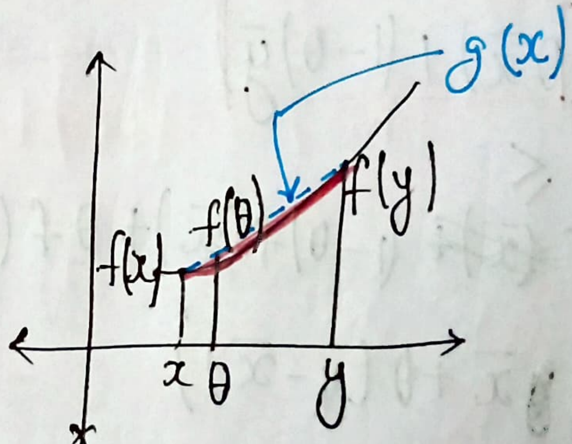


$f : \Omega \rightarrow R$ is convex

→ if Ω is convex

→ for any $\bar{x}, \bar{y} \in \Omega$ $0 \leq \theta \leq 1$

$$f(\theta \bar{x} + (1-\theta)\bar{y}) \leq \theta f(\bar{x}) + (1-\theta) f(\bar{y})$$



Example $\mathbb{R}^3$:

$\Omega = \mathbb{R}^2$

$f(x,y) \triangleq \sqrt{x^2 + y^2}$

Convex bowl can be think of as an example

S : Convex subset of $\mathbb{R}^d$

$f : S \to \mathbb{R}$

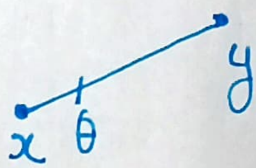
f is convex over S if and only if

$\forall x, y \;\; f(\bar{y}) \geq f(\bar{x}) + \nabla f(\bar{x}) \cdot (\bar{y} - \bar{x})$

<u>Proof:</u>

Assume f is convex

$$f(\overbrace{\boxed{\theta \bar{x} + (1-\theta)\bar{y}}}^{\bar{z}})$$



$$\leq$$

$$\cancel{\theta f(\bar{x}) +} (1-\theta) f(\bar{x}) + \theta f(\bar{y})$$

$$= f(\bar{x} + \theta(\bar{y} - \bar{x}))$$

$$= \frac{f(\bar{x} + \theta(\bar{y} - \bar{x})) - f(\bar{x})}{\theta} \leq - f(\bar{x}) + f(\bar{y})$$

$$\Downarrow$$

Directional Derivative

$$f'(\bar{x}, \bar{y} - \bar{x}) + f(\bar{x}) \leq f(\bar{y})$$

$$= \nabla f(\bar{x}) \cdot (\bar{y} - \bar{x})$$

Assume $f(\bar{y}) \geq f(\bar{x}) + \nabla f((\bar{x}) \cdot (\bar{x}, \bar{y}))$
for all $\bar{x}, \bar{y}$

⋮

(iske baad man nahi kiya
notes banane ka)

$\theta f(\bar{y}) \geq \theta f(\bar{z}) + \theta \nabla f(\bar{z}) \cdot (\bar{y} - \bar{z})$

$+ (1-\theta) f(\bar{x}) \geq (1-\theta) f(\bar{z}) + (1-\theta) \nabla f(\bar{z}) \cdot (\bar{x} - \bar{z})$

$\theta f(\bar{y}) + (1-\theta) f(\bar{x}) \geq f(\bar{z}) + \nabla f(\bar{z})$

(proved).

$\left\{ \begin{array}{l} \theta (\bar{y} - \bar{z}) + (1-\theta)(\bar{x} - \bar{z}) \\ = 0 \end{array} \right\}$

we put $\bar{z}$ value

$f: S \to R$, S: convex subset of $\mathbb{R}$

f is convex if and only if

$\forall x, y \in S \quad \boxed{f(y) \geq f(x) + f'(x) \cdot (y - x)}$

$\nabla^2 f(x) \geq 0 \quad \forall x \in S$ {positive semi-definite}

Proof:

Assume f is convex

Take any $\bar{x} \in S$

Assumption: If $\nabla^{1} f(x)$ is not positive definite

$\exists \bar{p} \cdot \; p^T \nabla^2 f(\bar{x}) p$

h is sufficiently small

$$f(\bar{x} + h\bar{p}) = f(\bar{x}) + \nabla f(\bar{x}) \cdot h\bar{p} + \frac{h^2 \bar{p}^T \nabla^2 f(\bar{y}) \bar{p}}{2}$$

~~for~~ some $\bar{\eta} \in L(\bar{x}, \bar{x} + h\bar{p})$

$$p^T \nabla^2 f(\bar{x}) p = \sum_{i,j=1}^{d} \frac{p_i \, p_j \, \nabla^2 f(\bar{x})}{\partial x_i \, \partial x_j}$$

So this assumption is wrong.

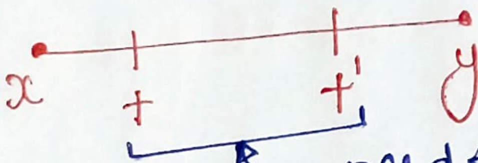
Reverse direction

Proof:

Assume $\nabla^2 f(\bar{x}) \geq 0 \quad \forall \bar{x} \in S$

Take any $x, y \in S$

Proove complexity of f restricted to $L(\bar{x}, \bar{y})$



need to choose t and t' prime ??