Stochastic Gradient Descent:

Efficiently estimate $\nabla f(\omega)$

- Let $G(\omega)$ be an unbiased estimator for $\nabla f(w)$

$$\underbrace{E(G(\omega))}_{\text{Random Vector.}} = \nabla f(\omega)$$

$$G(\omega) = 2\left(\langle \omega, x_{i^*} \rangle - y_{i^*}\right) x_{i^*}$$ where i^* sampled at a random.

or

for $k \ll n$

$$G(\omega) = 2 \sum_{i \in S} \left(\langle \omega, x_i \rangle - y_i\right) x_i$$ where

$S \subseteq$ Input Data

$|S| = k$.

THEOREM (SGD CONVERGENCE):

Representing the optimization variable w_i by x_i

f convex and β - smooth $\eta \leq 1/\beta$

and $\mathrm{var}(G(x)) \leq \sigma^2$

$$\Rightarrow E[f(\bar{x}_k)] \leq f(x^*) + \frac{\|x_0 - x^*\|_2^2}{2\eta k} + \underbrace{\eta\sigma^2}_{\text{error term introduced by using } G}.$$

$$\bar{x}_k = \frac{1}{k}(x_1 + \dots + x_k)$$

So all the needed equations:

Convexity : $\forall x, y \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

β-Smooth : $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|_2^2$

SGD : $x_{k+1} = x_k - \eta G(x_k)$

VECTOR : $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle$

$|\langle u, v \rangle| \leq \|u\| \cdot \|v\|$

UNBIASED ESTIMATOR : $E[G(x)] = \nabla f(x)$

$$\Rightarrow E\left[\|G(x)\|_2^2\right] - \|\nabla f(x)\|_2^2 \leq \sigma^2$$

$$\sigma^2 \geq E\left[\|G(x) - E[G(x)]\|_2^2\right] = E\left[\|G(x)\|_2^2\right] - \|E[G(x)]\|_2^2$$

$$= E\left[\|G(x)\|_2^2\right] - \|\nabla f(x)\|_2^2$$

CLAIM 1:

(SGD function decrease inequality from smoothness alone)

$\forall k$

$$E[f(x_{k+1})] \leq E[f(x_k)] - \frac{\eta}{2} E\left[\|\nabla f(x_k)\|_2^2\right] + \frac{\eta\sigma^2}{2}$$

From β- smoothness

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\beta}{2}\|x_{k+1} - x_k\|_2^2$$

$$= f(x_k) + \langle \nabla f(x_k), -\eta G(x_k) \rangle + \frac{\beta\eta^2}{2}\|G(x_k)\|_2^2 \quad - (1)$$

We also know

$$E\left[\|G(x)\|_2^2\right] - \|\nabla f(x)\|_2^2 \leq \sigma^2$$

From ① , apply expectation

$\mathbb{E}(G(x_k)) = \mathbb{E}(\nabla f(x_k))$

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \eta\, \mathbb{E}[\|\nabla f(x_k)\|_2^2]$$

$$+ \frac{\beta\eta^2}{2}\left[\sigma^2 + \mathbb{E}[\|\nabla f(x)\|_2^2]\right]$$

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \eta\, \mathbb{E}[\|\nabla f(x_k)\|_2^2]$$

$$+ \frac{\beta\eta^2}{2}\mathbb{E}[\|\nabla f(x_k)\|_2^2] + \frac{\beta\eta^2\sigma^2}{2}$$

Given $\beta \leq 1/\eta$

$$\boxed{\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \frac{\eta}{2}\mathbb{E}[\|\nabla f(x_k)\|_2^2] + \frac{\eta}{2}\sigma^2} \quad - ②$$

NOW LET US USE CONVEXITY:

$$f(x_k) \leq f(x^*) + \langle \nabla f(x_k), x_k - x^* \rangle$$

Before that,

$$\|x_{k+1} - x^*\|_2^2 = \|x_k - x^* - \eta G(x_k)\|_2^2 = \|x_k - x^*\|_2^2$$

$$+ \eta^2 \|G(x_k)\|_2^2 - 2\eta \langle x_k - x^*, G(x_k) \rangle$$

$$\left(\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) = \eta^2 \|G(x_k)\|_2^2 - 2\eta \langle G(x_k), x_k - x^* \rangle$$

Apply expectation

$$E\left(\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) = \eta^2 E\left(\|G(x_k)\|_2^2\right) -$$

$$2\eta E \langle (G(x_k)), x_k - x^* \rangle$$

$$E\left[\|G(x)\|_2^2\right] - E\left[\|\nabla f(x_k)\|_2^2\right] \leq \sigma^2 \quad \text{and} \quad E(G(x)) = E(\nabla f(x_k))$$

$$E\left(\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) = \eta^2 \sigma^2 + \eta^2 E\left[\|\nabla f(x_k)\|_2^2\right]$$

$$2\eta E\left(\langle \nabla f(x_k), x_k - x^* \rangle\right)$$

Apply convexity now

$$f(x_k) \leq f(x^*) + \langle \nabla f(x_k), x_k - x^* \rangle$$

$$-\langle \nabla f(x_k), x_k - x^* \rangle \leq f(x^*) - f(x_k)$$

$$-E\langle \nabla f(x_k), x_k - x^* \rangle \leq f(x^*) - E[f(x_k)]$$

$$E\left[\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) \leq \eta^2\sigma^2 + \eta^2 E\left[\|\nabla f(x_k)\|_2^2\right]$$

$$+\left[f(x^*) - E[f(x_k)]\right] 2\eta$$

$$\frac{1}{2\eta} E\left[\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) \leq \frac{\eta}{2}\sigma^2 + \frac{\eta}{2} E\left[\|\nabla f(x_k)\|_2^2\right]$$

$$+ f(x^*) - E[f(x_k)]$$

From 2

$$E[f(x_{k+1})] \leq E[f(x_k)] - \frac{\eta}{2} E[\|\nabla f(x_k)\|_2^2] + \frac{\eta}{2}\sigma^2$$

NO!

$$\frac{1}{2\eta} E\left[\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) \leq f(x^*) - E[f(x_{k+1})] + \eta\sigma^2$$

$$E[f(x_i)] \le f(x^*) - \frac{1}{2\eta} E\left[\|x_i - w^*\|_2^2 - \|x_{i-1} - x^*\|_2^2\right]$$

$$+ \eta\sigma^2.$$

Sum THEM :

$$\sum_{i=1}^{k} E[f(x_i)] \le kf(x^*) - \frac{1}{2\eta}\left(\|x_k - x^*\|_2^2 - \|x_0 - x^*\|_2^2\right)$$

$$+ k\eta\sigma^2$$

$$\le kf(x^*) + \frac{\|x_0 - x^*\|_2^2}{2\eta} + k\eta\sigma^2$$

$$E[f(\bar{x}_k)] \le f(x^*) + \frac{\|x_0 - x^*\|_2^2}{2\eta k} + \eta\sigma^2. \quad - 3$$

Choose η such that

$$\frac{\|x_0 - x^*\|_2^2}{2\eta k} = \eta\sigma^2$$

$$\eta = \frac{\|x_0 - x^*\|}{\sigma\sqrt{2k}}$$

Using this η in (3)

$$E\left[f(\bar{x}_k)\right] \leq f(x^*) + \frac{\|x_0 - x^*\| \sigma \sqrt{2}}{\sqrt{k}}$$

So $\eta \propto 1/\sqrt{k}$.

CONSTRAINED OPTIMIZATION:

Sometime other than

$\arg\min L(x)$

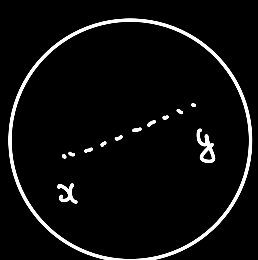
we need x to lie in some bounded region of $\mathbb{R}^d$.

$\Rightarrow$ We can adapt GD / SGD / NAGD to such situations as long as constrained region is a "convex set".

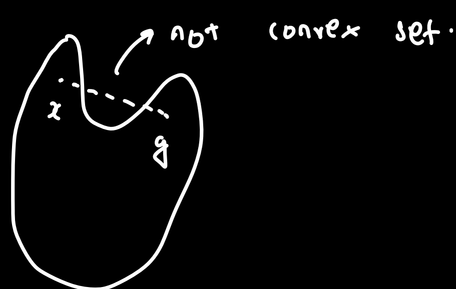
Definition: $C \subseteq \mathbb{R}^d$ is a convex set if

$\forall x, y \in C$, $\dfrac{x+y}{2} \in C$ (mid point also in C)

equivalently $\forall x, y \in C$, $\forall \lambda \in [0,1]$, $\lambda x + (1-\lambda) y \in C$.



C, convex set



PROJECTED GRADIENT DESCENT:

GOAL:

Compute $\underset{x \in C}{\operatorname{argmin}}\ L(x)$

PROJECTION: $\operatorname{proj}_C(y) = \underset{x \in C}{\operatorname{argmin}}\ \|x - y\|_2$

(closest point in C to y)

PGD: $x_{k+1} = \operatorname{Proj}_C(x_k - \eta \nabla f(x_k))$

So, apply gradient descent. If new point x_{k+1} not in C, project it into C.