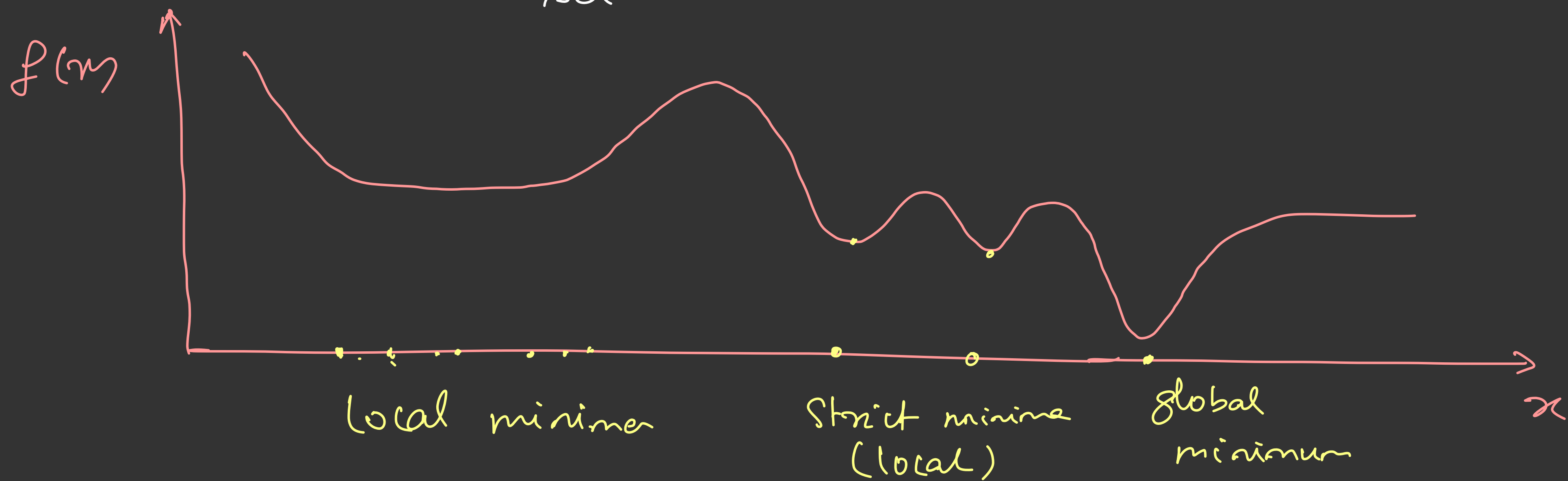
Optimization:

$$\underset{\underline{x} \in \mathcal{X}}{\text{minimize}} \quad f(\underline{x})$$

$f : \mathbb{R}^D \longrightarrow \mathbb{R}$ $\qquad$ $x \in \mathbb{R}^D$ is the design variable

$\mathcal{X} \subseteq \mathbb{R}^D$ Constraint set.



- $\underline{x}^* \in \mathcal{X}$ is a local minima if $\exists\ \epsilon > 0$ s.t. $f(\underline{x}) \geq f(\underline{x}^*)$ $\forall\ \underline{x} \in \mathcal{X}$ with $\| \underline{x} - \underline{x}^* \| \leq \varepsilon$

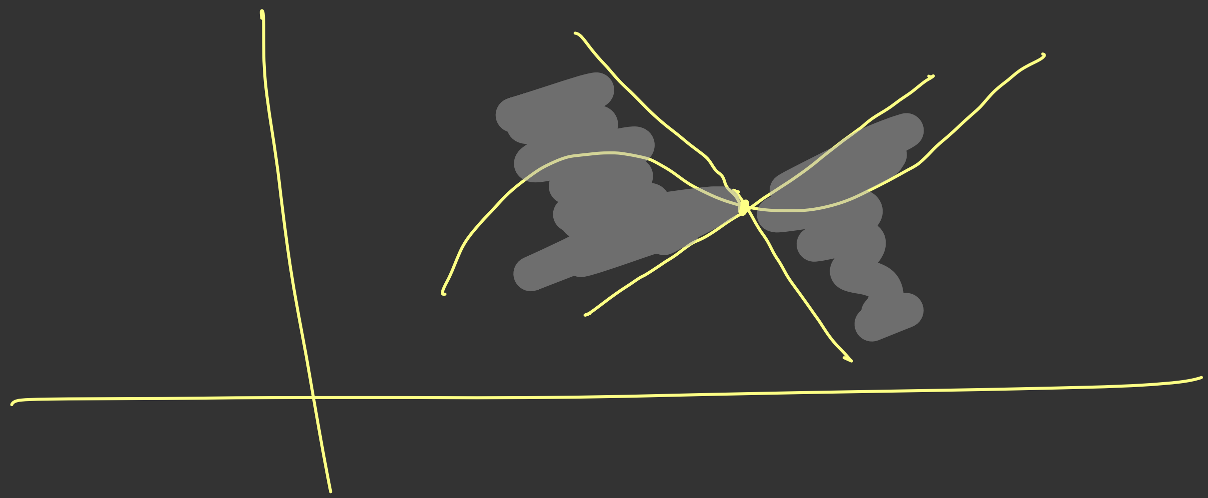
- $\underline{x}^* \in \mathcal{X}$ is a global minimum if $\exists\ \underline{x}^* \in \mathcal{X}$ s.t. $f(\underline{x}) \geq f(\underline{x}^*)$ $\forall\ \underline{x} \in \mathcal{X}$

<u>Lipschitz - continuous function:</u>

$f : \mathbb{R}^D \to \mathbb{R}$ is Lipschitz continuous

if $\exists\ L \geq 0$

$$\| f(\underline{x}) - f(\underline{y}) \| \leq L \, \| x - y \|$$



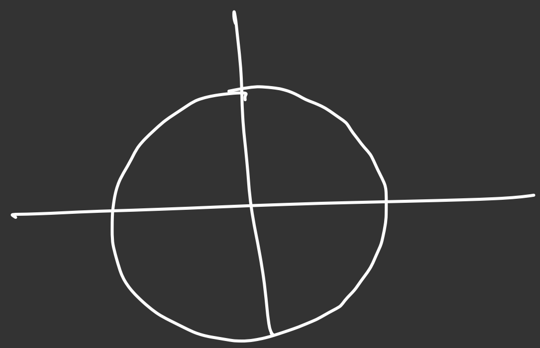
Convex sets:

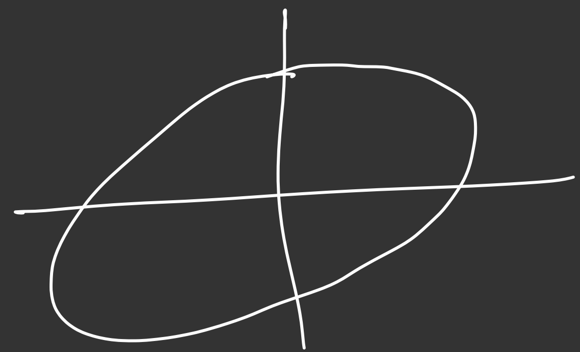
A subset $\mathcal{X} \subseteq \mathbb{R}^d$ is convex if

$$\theta \underline{x}_1 + (1-\theta) x_2 \in \mathcal{X} \quad \forall\ \underline{x}_1, \underline{x}_2 \in \mathcal{X}$$
$$\theta \in [0, 1]$$

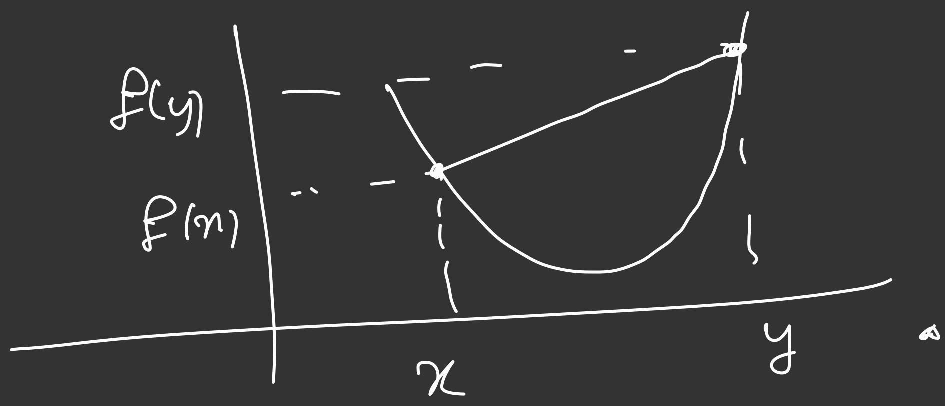
E.g. norm ball

$$\mathcal{X} = \{ \underline{x} \mid \underline{x}^T \Sigma \, \underline{x} \leq 1 \}$$



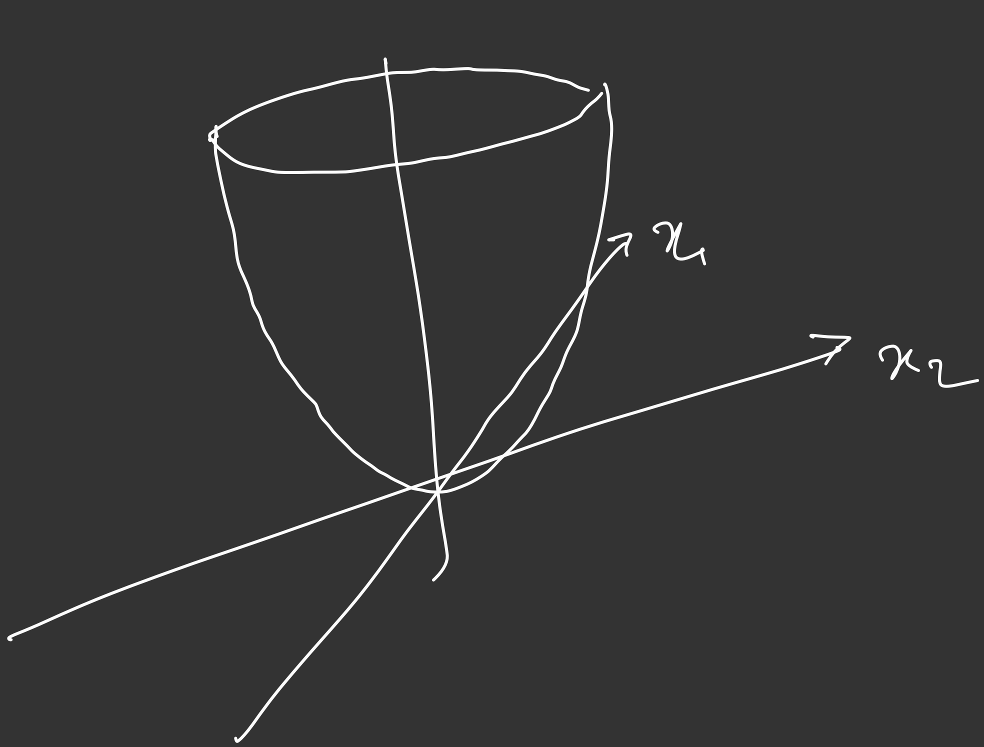


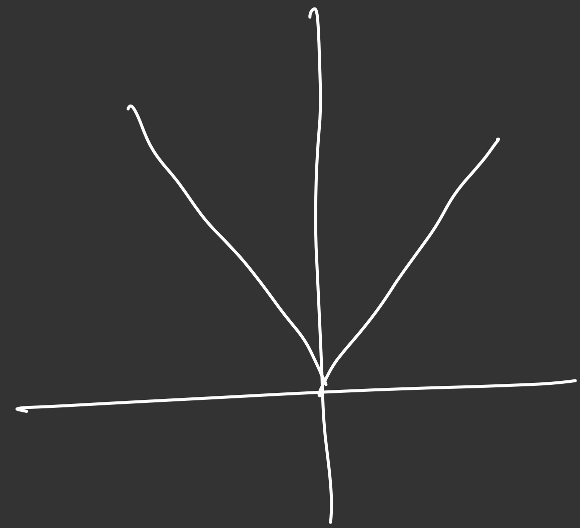
Convex function:



f is convex if: $\forall \, \underline{x}, \underline{y} \in \mathcal{X}$, $\theta \in [0,1]$

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





OR,

$$f(\underline{y}) \geq f(\underline{x}) + \nabla f^T(\underline{x})(\underline{y} - \underline{x})$$

$$\forall \underline{x}, \underline{y} \in X$$

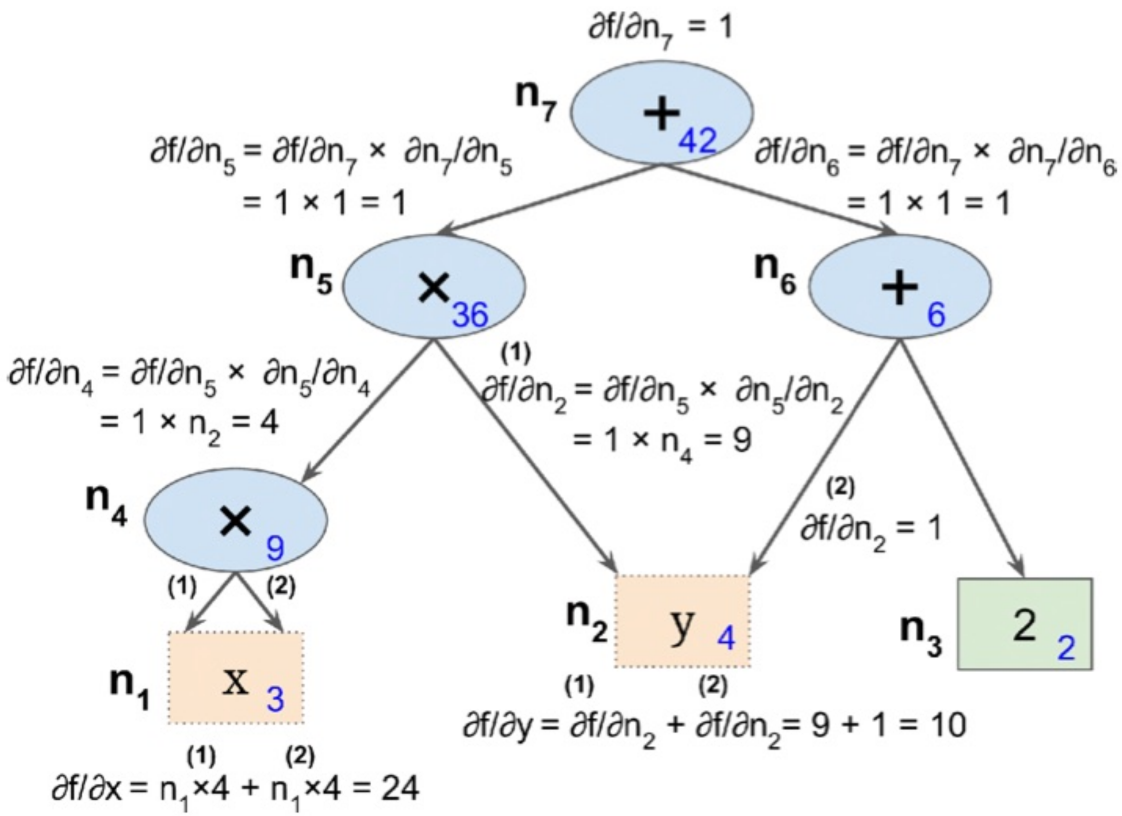
OR,

$$\nabla^2 f(x) \succeq 0$$

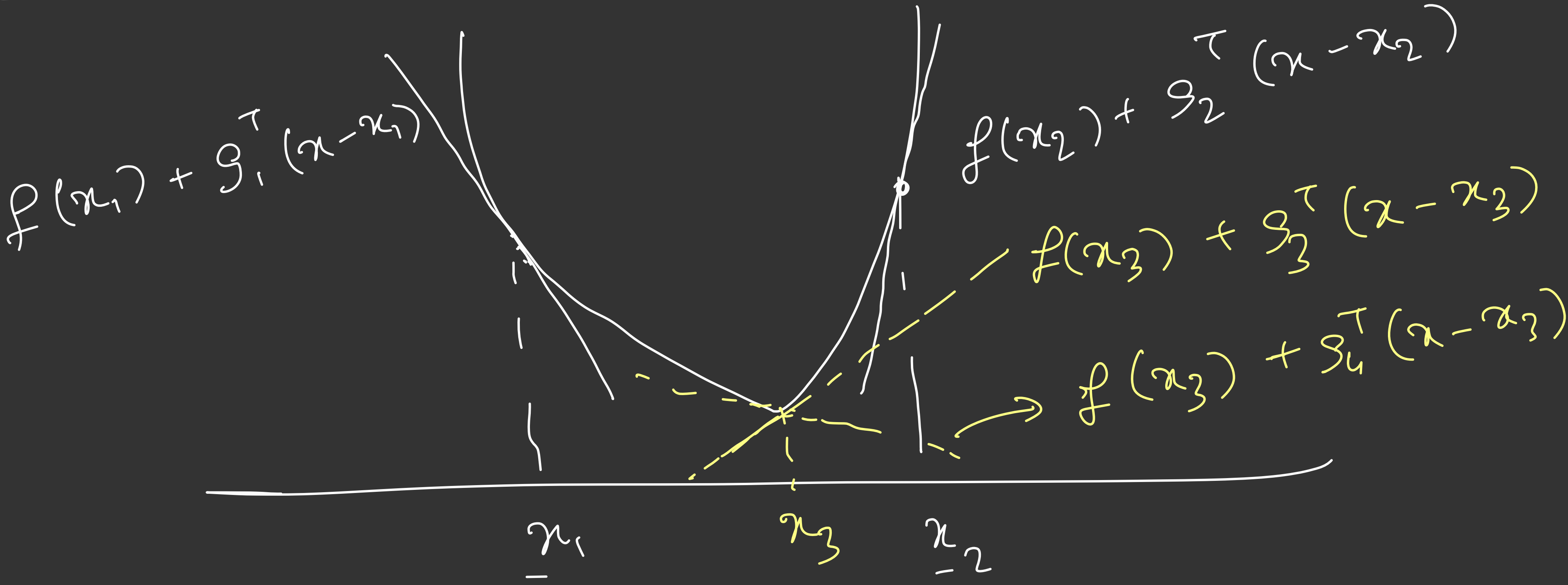
Auto-diff:

Reverse mode auto-diff:

$$f(x, y) = x^2y + y + 2$$



Subgradients:



g is a subgradient of f at x if

$$f(\underline{y}) \geq f(\underline{x}) + \underline{g}^T(\underline{y} - \underline{x}), \quad \forall y$$

Example:

$$f(x) = |x|$$

$$\partial f(\underline{x}) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Iterative descent methods:

$$\min_{\underline{x} \in \mathbb{R}^d} f(x)$$

$$\underline{x}_k \longrightarrow \boxed{\qquad} \longrightarrow \underline{x}_{k+1}$$

$$f(x_{k+1}) < f(x_k)$$

$$k = 0, 1, 2 \ldots$$

$$\underline{x}_{k+1} = x_k + \eta_k \underline{d}_k$$

$$f(\underline{x}_k + \quad d_k) \approx f(x_k) + \nabla f(x_k)^T d_k$$

Steepest descent method:

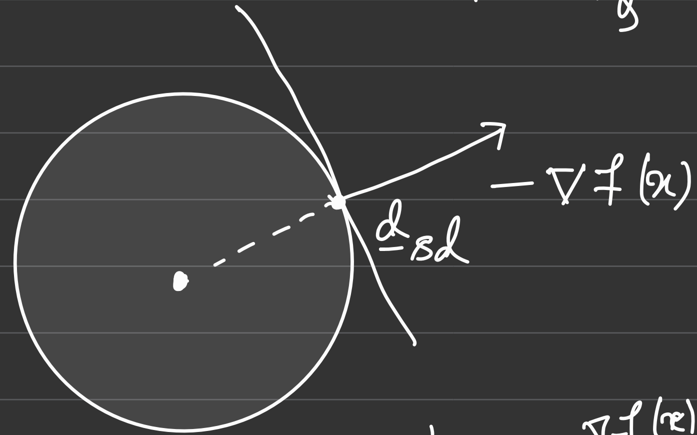
→ descent direction that is as negative as possible ~~that~~ that yields the greatest rate of objective value improvement

$$\underline{d}_{sd} = \arg\min_{\underline{d}} \left\{ \nabla f^T(\underline{x})\, \underline{d} \ : \ \|\underline{d}\| \le 1 \right\}$$

$\|d\|_g \le 1$

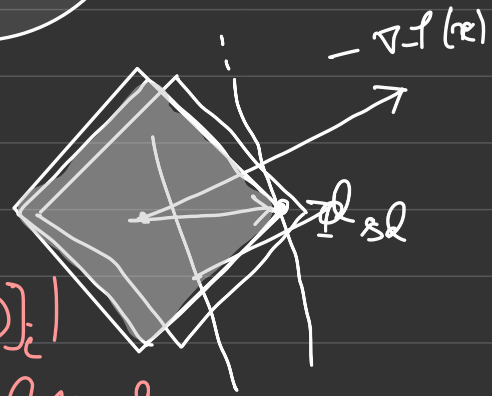
For Euclidean norm:

$$\underline{d}_{sd} = -\nabla f(\underline{x})$$



For l_1-norm:

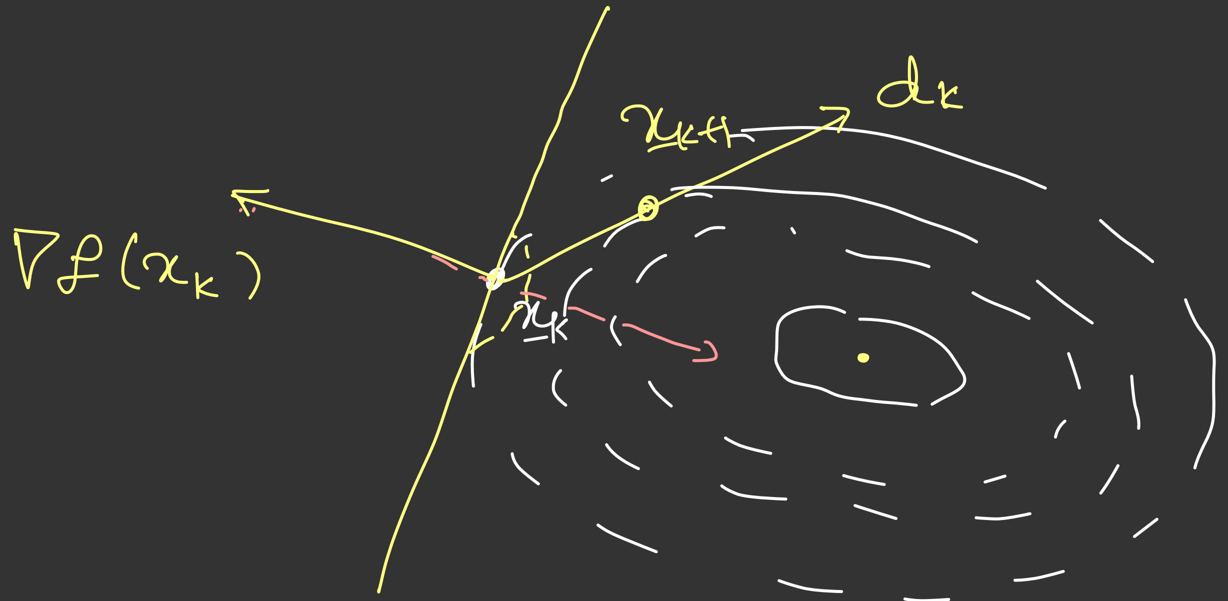
$$\underline{d}_{sd} = \arg\min_{\underline{d}} \left\{ \nabla f^T(\underline{x})\, \underline{d} \ : \ \|\underline{d}\|_1 \le 1 \right\}$$



Let i be any index s.t. $\|\nabla f(\underline{x})\|_\infty = |[\nabla f(x)]_i|$

Then $\underline{d}_{sd} = -\operatorname{sign}\left(\dfrac{\partial f(\underline{x})}{\partial x_i}\right) \underline{e}_i$: coordinate descent.

$$\nabla f^T(x_k)\, d_k < 0$$



$$\underline{d}_k = -\nabla f(x_k)$$

$$\nabla f^T(x_k)\, d_k = -\|\nabla f(x_k)\| < 0$$

Gradient descent:

$$\underline{x}_{k+1} = \underline{x}_k - \eta_k \nabla f(\underline{x}_k)$$

Vanilla analysis: $\quad \underline{x}_{t+1} = \underline{x}_t - \eta \nabla f(\underline{x}_t)$

$t = 0, \ldots, T-1$

Define: $\quad \underline{g}_t = \nabla f(\underline{x}_t)$

so $\quad \underline{g}_t = (x_{t+1} - x_t)/\eta$

Let us relate this vector to our current direction from an optimum $\underline{x}^*$: $x_t - \underline{x}^*$

$$\underline{g}_t^T (\underline{x}_t - \underline{x}^*) = \frac{1}{\eta} (\underline{x}_{t+1} - x_t)^T (\underline{x}_t - x^*)$$

Cosine theorem: $\quad 2 \underline{v}^T \underline{w} = \|\underline{v}\|^2 + \|\underline{w}\|^2 - \|\underline{v} - \underline{w}\|_2^2$

$$\underline{g}_t^T (\underline{x}_t - \underline{x}^*) = \frac{1}{2\eta} \left[\|(\underline{x}_{t+1} - \underline{x}_t)\|_2^2 + \|\underline{x}_t - \underline{x}^*\|_2^2 - \|\underline{x}_{t+1} - x^*\|_2^2 \right] \quad (*)$$

$$= \frac{1}{2\eta} \left[\eta^2 \|\underline{g}_t\|_2^2 + \|\underline{x}_t - x^*\|_2^2 - \|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \right]$$

$$= \frac{\eta}{2} \|\underline{g}_t\|^2 + \frac{1}{2\eta} \left[\|\underline{x}_t - x^*\|_2^2 - \|\underline{x}_{t+1} - \underline{x}^*\|^2 \right]$$

$\circledast$ Sum over the iteration t.

$$\sum_{t=0}^{T-1} \underline{g}_t^T (\underline{x}_t - \underline{x}^*) = \frac{\eta}{2} \sum_{t=0}^{T-1} \|\underline{g}_t\|^2 + \frac{1}{2\eta}\left[\|\underline{x}_0 - \underline{x}^*\|_2^2 - \|\underline{x}_T - \underline{x}^*\|_2^2\right]$$

$5 > 3$

$-5 < -3$

$$\leq \frac{\eta}{2} \sum_{t=0}^{T-1} \|\underline{g}_t\|^2 + \frac{1}{2\eta} \|\underline{x}_0 - \underline{x}^*\|_2^2$$

Now, suppose f is convex : $f(\underline{y}) \geq f(x) + \nabla f^T(x)(y - x)$

$\left. \begin{array}{l} y = x^* \\ x = x_t \end{array} \right\}$ $\Rightarrow$ $f(x_t) - f(x^*) \leq \nabla f^T(x_t)(\underline{x}_t - \underline{x}^*)$

$$= \underline{g}_t^T (\underline{x}_t - \underline{x}^*)$$

- Upper bound on the average error: $f(x_t) - f(x^*)$

$f(x_{T-1}) - f(x_t) \geq \nabla f^T(x_t)(x^* - x_t)$

$$\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \frac{\eta}{2} \sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2\eta} \|x_0 - x^*\|_2^2$$

$f(x_t) - f(x^*) \leq 2\eta \nabla f^T(x_t)(x_t - x)$

- Last iterate is not neccesarily the best one as "fixed" step size can make steps overshoot & increase function value

$\boxed{f_n a := f_{n+t} 1}$

- **For Lipschitz convex functions:**

$f : X \to \mathbb{R}^n$ continuous is called Lipschitz if there exists $B \geq 0$

$$\| f(\underline{x}) - f(\underline{y}) \| \leq B \| \underline{x} - \underline{y} \| \quad \forall\, \underline{x}, \underline{y} \in X$$

$$\iff \| \nabla f(\underline{x}) \| \leq B \quad \forall\, \underline{x} \in X$$

$f : \mathbb{R}^n \to \mathbb{R}$ ✓

- $\| \underline{x}_0 - \underline{x}^* \| \leq R$

Then,
$$\sum_{t=0}^{T-1} \left(f(\underline{x}_t) - f(\underline{x}^*) \right) \leq \frac{\eta}{2} B^2 T + \frac{1}{2\eta} R^2 \quad (**)$$

So, choose η such that $q(\eta) = \frac{\eta}{2} B^2 T + \frac{1}{2\eta} R^2$ is minimized.

$$\frac{d}{d\eta} q(\eta) = 0 \Rightarrow \frac{1}{2} B^2 T - \frac{1}{2\eta^2} R^2 = 0$$

$$\Rightarrow \eta = \frac{R^2}{B^2 \sqrt{T}} \quad \text{and} \quad q(\eta) = RB\sqrt{T}$$

$(**) \div T$

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(f(x_t) - f(\underline{x}^*) \right) \leq \frac{RB}{\sqrt{T}} \approx O\left(\frac{1}{\sqrt{T}}\right)$$

independent of 'n' but R & B does depend on n

So to obtain $\min_{t=0..T-1} f(\underline{x}_t) - f(\underline{x}^*) \leq \varepsilon$

we need $T \geq \frac{R^2 B^2}{\varepsilon^2}$.

Sos. $\varepsilon = 10^{-8}$

R and $B \approx O(d)$??

Second-order method's (Newton's):

$$f(x_1, x_2) = x_1^2 \cdot \frac{1}{100} + x_2^2 \cdot 100$$

$$\underline{x}_k = \begin{bmatrix} -10 \\ -0.1 \end{bmatrix}$$

$$\nabla f(x_t) = \begin{bmatrix} -\frac{1}{5} \\ -20 \end{bmatrix}$$

Check Newton's method with $B = H^{-1}$?
(Hessian)

$$\underline{x}_{k+1} = \underline{x}_k - \eta_k H_k^{-1} \nabla f(\underline{x}_k)$$

Linear regression:

$$f(\underline{w}) = \sum_{n=1}^{N} \left(y_n - \underline{w}^T \cdot \underline{x}_n\right)^2$$

$$\nabla f(\underline{w}) = \sum_{n=1}^{N} \nabla f_n(\underline{w})$$

$$= - \sum_{n=1}^{N} 2\, \underline{x}_n \left(y_n - \underline{w}^T x_n\right)$$

Can we update using one or few examples in each iteration?

Stochastic gradient descent:

$$x \longrightarrow \boxed{} \longrightarrow \underbrace{\tilde{\nabla} f(x)}_{} = \begin{bmatrix} 0 \\ \vdots \\ [\nabla f(x)]_j \\ \vdots \\ 0 \end{bmatrix}$$

$j \sim \text{unif}(1, \ldots d)$ r. v.

$$\underline{x}_{k+1} = \underline{x}_{k+1} - \eta_k \tilde{\nabla} f(x_k)$$

$$\tilde{\nabla} f(x_k) \approx \nabla f(x_k) + \omega$$

Stochastic gradient:

$$\boxed{\mathbb{E}\left[\tilde{\nabla} f(\underline{x}_k)\right] = \nabla f(x_k)}$$

$$\underline{x} \longrightarrow \boxed{} \longrightarrow \underline{\tilde{g}}_x \quad \text{so that } E[\tilde{g}_x] \in \partial f(\underline{x})$$

(random variable)

Example:

① $\underline{\tilde{g}}_x = \nabla f(x) + \underline{w}$; $\underline{w}$ is zero-mean noise

$$E[\tilde{g}_x] = E[\nabla f(x) + \underline{w}] = \nabla f(\underline{x})$$

② Random coordinate descent:

$$\nabla f(\underline{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$$

$$x \longrightarrow \boxed{} \longrightarrow \tilde{g}_x = \begin{bmatrix} 0 \\ \vdots \\ \partial f/\partial x_j \\ \vdots \\ 0 \end{bmatrix} \cdot d \quad ; \quad \underline{x} \in \mathbb{R}^d$$

$j \sim \text{unif}(1, \dots, d)$

$$E_j[\tilde{g}_x] = \sum_{i=1}^{d} d \cdot \begin{bmatrix} 0 \\ \vdots \\ \partial f/\partial x_j \\ \vdots \\ 0 \end{bmatrix} \cdot \frac{1}{d} = \nabla f(\underline{x})$$

Unbiasedness and the vanilla analysis:

Recall: In gradient descent, we could lower bound

$$\underline{g}_t^T(\underline{x}_t - \underline{x}^*) \geq f(\underline{x}_t) - f(\underline{x}^*)$$

but now we cannot as $\tilde{\underline{g}}_t$ may be far from being the true gradient.

- So inequality $f(\underline{x}_t) - f(\underline{x}^*) \leq \tilde{g}_t^T(\underline{x} - x^*)$

(from convexity) may not hold.

We have
$$\underbrace{\mathbb{E}\left[g_t \mid \underline{x}_t = \underline{x}\right]} = \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\underline{x}) = \nabla f(\underline{x})$$

Conditional expectation of g_t given the event $\{\underline{x} = x_t\}$.

$\forall\, \underline{x} \in \mathbb{R}^d$

$$\Rightarrow \mathbb{E}\left[g_t^T (\underline{x} - \underline{x}^*) \,\middle|\, \underline{x}_t = \underline{x} \right] =$$

$$\mathbb{E}\left[\underline{g}_t^T \,\middle|\, \underline{x}_t = \underline{x} \right] (\underline{x} - \underline{x}^*) = \nabla f^T(\underline{x}) (\underline{x} - \underline{x}^*)$$

- $\{ \underline{x}_t = \underline{x} \}$ can occur only for $\underline{x}$ in finite set X

$$\mathbb{E}\left[\underline{g}_t^T (\underline{x}_t - x^*) \right]$$

$$= \sum_{\underline{x} \in X} \mathbb{E}\left[g_t^T (\underline{x} - \underline{x}^*) \,\middle|\, \underline{x}_t = \underline{x} \right] \text{prob}(\underline{x}_t = \underline{x})$$

$$= \sum_{x \in X} \nabla f^T(x) (\underline{x} - \underline{x}^*) \text{ prob}(\underline{x}_t = \underline{x})$$

$$= \mathbb{E}\left[\nabla f^T(\underline{x}_t) (\underline{x}_t - \underline{x}^*) \right]$$

$$\Rightarrow \quad \mathbb{E}\left[\tilde{g}_t^{T}(\underline{x}_t - \underline{x}^*)\right] = \mathbb{E}\left[\nabla f^{T}(x_t)(\underline{x}_t - \underline{x}^*)\right]$$

$$\geq \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*)\right]$$

So the lower bound holds in expectation.

Bounded stochastic gradients:

- Same convergence rate as gradient descent method

Claim: Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, $\underline{x}^*$ be a global minimum of f;

$\| \underline{x}_0 - \underline{x}^* \| \leq R$ and that $\mathbb{E}\left[\| \underline{g}_t \|^2 \right] \leq B^2$ $\forall t$

Then stochastic gradient descent with

constant step size $\eta = \dfrac{R}{B\sqrt{T}}$ yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[f(\underline{x}_t) \right] - f(\underline{x}^*) \leq \frac{RB}{\sqrt{T}}$$

Iteration complexity : $O\left(1/\varepsilon^2 \right)$ $\qquad O\left(\frac{1}{\sqrt{T}} \right)$

Recall our vanilla analysis:

$$g_t^T(x_t - x^*) = \frac{\eta}{2}\|g_t\|^2 + \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$$

Telescoping sum:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{\eta}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\eta}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right)$$

$$\leq \frac{\eta}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\eta}\|x_0 - x^*\|^2$$

Taking expectation on both sides

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\tilde{g}_t^T(x_t - x^*)\right] \leq \frac{\eta}{2}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}\left[\|\tilde{g}_t\|^2\right]}_{\leq B^2} + \frac{1}{2\eta}\underbrace{\|x_0 - x^*\|^2}_{\leq R^2}$$

We have the lower bound:

$$\mathbb{E}\left[\tilde{g}_t^T(x_t - x^*)\right] \geq \mathbb{E}\left[f(x_t) - f(x^*)\right]$$

$$\sum_{t=0}^{T-1} \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*) \right] \leq \frac{\eta}{2} B^2 T + \frac{1}{2\eta} R^2$$

$$= q(\eta)$$

Choose η that minimize the upper bound:

$$\frac{1}{2} B^2 T - \frac{1}{2\eta^2} R^2 = 0$$

$$\eta = \frac{R}{B\sqrt{T}}$$

for which we have $O\left(1/\sqrt{T}\right)$ $\blacksquare$