

recall definition of convexity: $\alpha f(x) + (1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y)$ for $0 \leq \alpha \leq 1$.

the gradient of a function is $\nabla f(x) = \begin{bmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{bmatrix}$, the vector of partial derivatives w.r.t. each component of x .

the directional derivative of f in direction r (at point x) is:

$$\nabla_r f(x) = \lim_{h \rightarrow 0} \frac{f(x+hr) - f(x)}{h} = \left. \frac{\partial f}{\partial h} f(x+hr) \right|_{h=0}$$

using the total derivative

$$= \frac{\partial f}{\partial x_1} r_1 + \frac{\partial f}{\partial x_2} r_2 + \dots + \frac{\partial f}{\partial x_n} r_n = \nabla f(x)^T r.$$

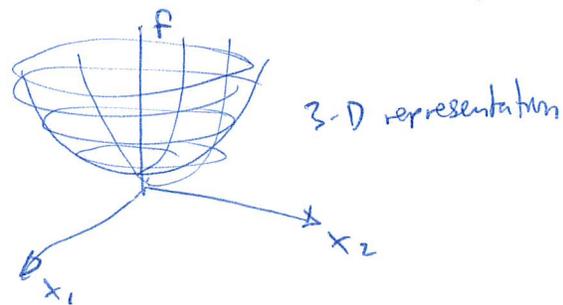
so derivative in a direction equals dot product of gradient with direction. this is maximized when r is aligned with ∇f . ("steepest increase").

★ Interpretation: consider convex function:

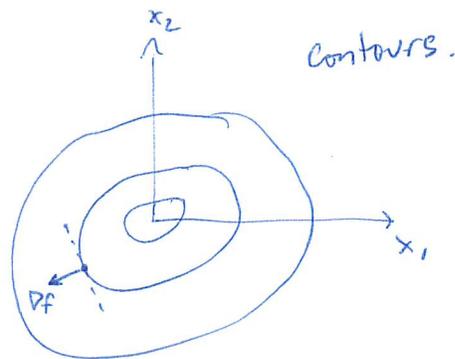
∇f points in direction of steepest increase.

if $\nabla f^T r = 0$ (orthogonal to gradient),

we move along contour (f doesn't change)



★ to minimize f , we can take steps in the direction $-\nabla f$ (steepest decrease)



alternate definition of convexity.

$$\alpha f(y) + (1-\alpha)f(x) \geq f(\alpha y + (1-\alpha)x)$$

rearrange:

$$f(y) \geq f(x) + \frac{f(x + \alpha(y-x)) - f(x)}{\alpha}$$

looks like directional derivative!

limit $\alpha \rightarrow 0$:
(assuming f is differentiable).

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

proof of converse

suppose $f(y) \geq f(x) + \nabla f(x)^T (y-x)$ for all y, x .

then we also have: $f(z) \geq f(x) + \nabla f(x)^T (z-x)$ for all z, x .

take α (first inequality) + $(1-\alpha)$ (second inequality).

$$\alpha f(y) + (1-\alpha)f(z) \geq f(x) + \nabla f(x)^T (\alpha y + (1-\alpha)z - x)$$

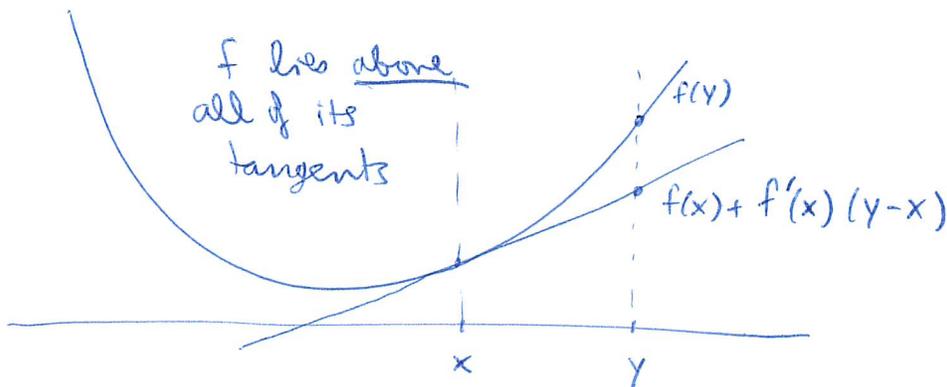
let $x = \alpha y + (1-\alpha)z$ and obtain:

$$\alpha f(y) + (1-\alpha)f(z) \geq f(\alpha y + (1-\alpha)z)$$

So both definitions are equivalent. (when f is differentiable)

* Interpretation (1D)

* in 2D, f lies above the tangent plane.



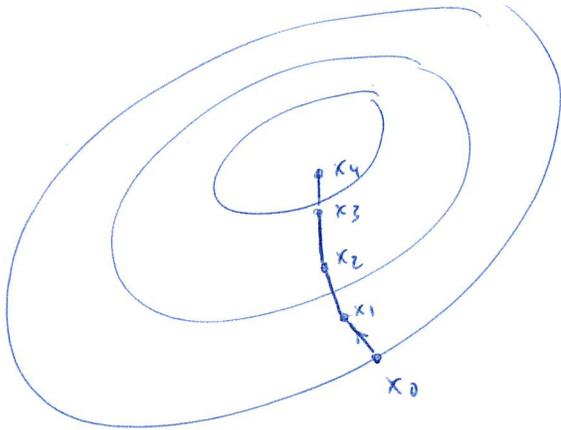
gradient descent

3

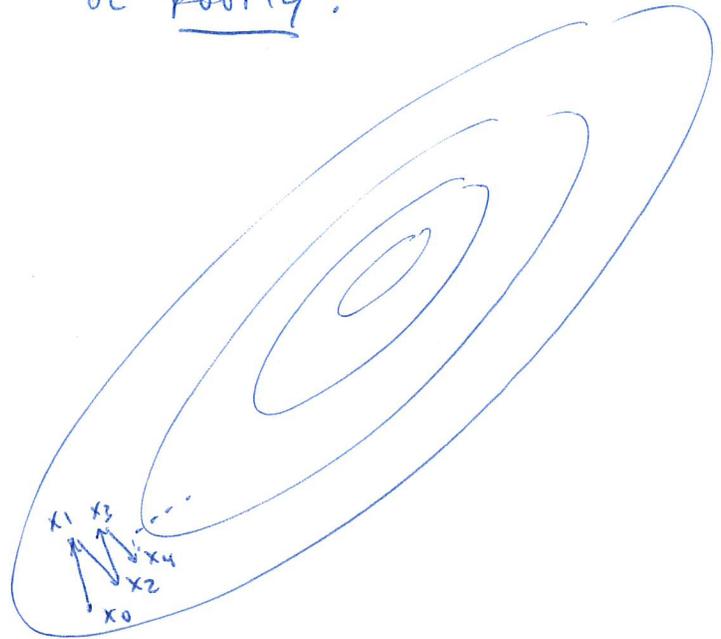
★ basic idea: move in the direction of steepest descent:

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

can work well:



or poorly:



★ what is gradient descent for Least Squares?

$$f(x) = \|y - Ax\|^2 \Rightarrow \nabla f(x) = 2A^T(Ax - y)$$

$$\text{so } x_{k+1} = x_k - \gamma \cdot 2A^T(Ax_k - y).$$

It's simply Landweber with a scaled stepsize!

subgradients i.e. when f is not differentiable.

the gradient of a differentiable f (convex also) satisfies:

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad (\text{for all } x, y).$$

if f is convex but not differentiable, " ∇f " may not make sense.

Define a subgradient of f at x to be any vector v satisfying

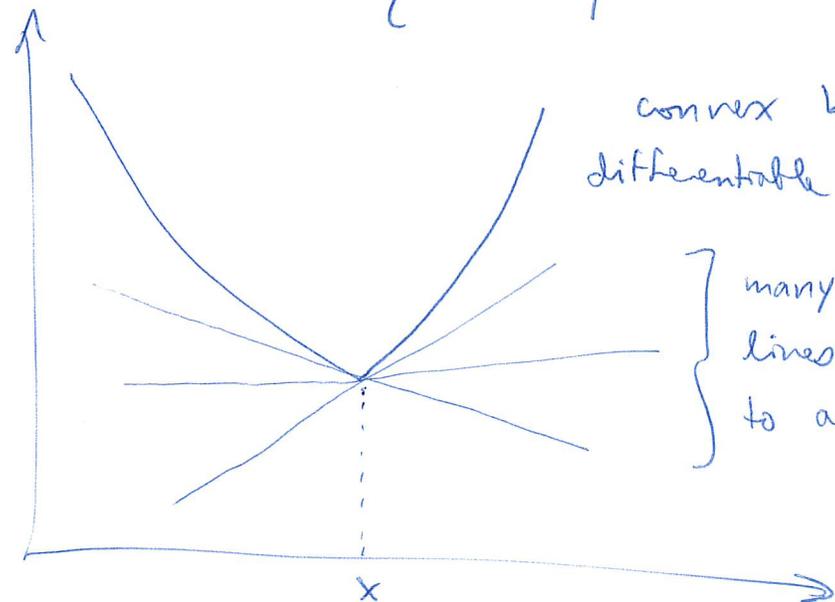
$$f(y) \geq f(x) + v^T (y-x) \quad (\text{for all } y).$$

★ if f is differentiable, the only solution is $v = \nabla f(x)$.

Otherwise, there may be many v 's that work.

The set of all subgradients is called the differential set

$$\partial f(x) = \left\{ v \in \mathbb{R}^n \mid f(y) \geq f(x) + v^T (y-x) \text{ for all } y \in \mathbb{R}^n \right\}$$



convex but not differentiable function.

} many possible "tangent" lines. Each corresponds to a valid subgradient at x .

Note: $\partial(f+g)(x) = \partial f(x) + \partial g(x)$

where the "+" on the right is set addition. i.e.

$$S + T = \{x+y \mid x \in S, y \in T\}.$$

subgradient method

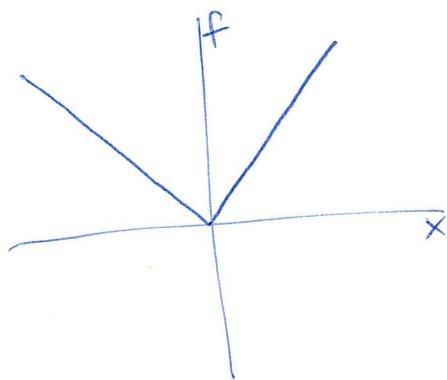
(5)

like gradient descent, but using subgradient instead.

$$x_{k+1} = x_k - \gamma v_k, \text{ where } v_k \in \partial f(x_k).$$

there is no specified way to choose v_k ; so method is not unique.

example: what are the subgradients of $f(x) = |x|$, $x \in \mathbb{R}$?



if $x > 0$, x is differentiable: $\partial f(x) = \{1\}$.

if $x < 0$, x is differentiable; $\partial f(x) = \{-1\}$.

if $x = 0$, x is not differentiable.

v must satisfy: $f(y) \geq f(x) + v^T(y-x)$ for all y .

$$\Leftrightarrow |y| \geq v^T y \text{ for all } y.$$

$$\Leftrightarrow |y| \geq v y \text{ (true for } y=0).$$

$$\Rightarrow 1 \geq v \cdot \text{sign}(y)$$

$$\Rightarrow \boxed{-1 \leq v \leq 1}$$

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

example of a convergence result.

6

Suppose subgradients are bounded: $\|v\| \leq G$ for all $v \in \partial f(x) \forall x$, and f is convex, and we use subgradient descent; then

$$\frac{1}{T} \sum_{k=0}^{T-1} (f(x_k) - f(x_*)) \leq \frac{\|x_0 - x_*\|^2}{2\gamma T} + \frac{\gamma}{2} G^2 \quad \forall T.$$

where x_* is argmin $f(x)$.

Proof: $\|x_{k+1} - x_*\|^2 = \|x_k - \gamma v_k - x_*\|^2$

$$= \|x_k - x_*\|^2 + \gamma^2 \|v_k\|^2 - 2\gamma v_k^T (x_k - x_*)$$

Now: $f(y) \geq f(x) + v^T(y - x)$ for all $v \in \partial f(x)$.

let $y \rightarrow x_*$, $x \rightarrow x_k$, $v_k \in \partial f(x_k)$.

$$\Rightarrow f(x_*) \geq f(x_k) + v_k^T (x_* - x_k)$$

$$\Rightarrow -v_k^T (x_k - x_*) \leq f(x_*) - f(x_k)$$

$$\Rightarrow \|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 + \gamma^2 \|v_k\|^2 + 2\gamma (f(x_*) - f(x_k))$$
$$\leq \|x_k - x_*\|^2 + \gamma^2 G^2 + 2\gamma (f(x_*) - f(x_k))$$

(sum over $k=0, 1, \dots, T-1$).

$$\Rightarrow 2\gamma \sum_{k=0}^{T-1} (f(x_k) - f(x_*)) \leq \gamma^2 T G^2 + \|x_0 - x_*\|^2 - \|x_{k+1} - x_*\|^2$$
$$\leq \gamma^2 T G^2 + \|x_0 - x_*\|^2$$

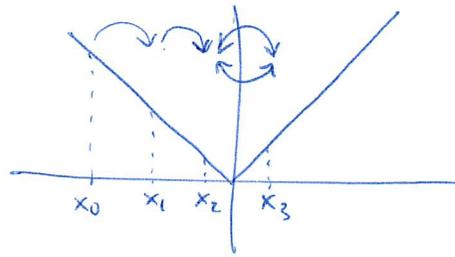
$$\Rightarrow \frac{1}{T} \sum_{k=0}^{T-1} (f(x_k) - f(x_*)) \leq \frac{\|x_0 - x_*\|^2}{2\gamma T} + \frac{\gamma}{2} G^2.$$

$$\frac{1}{T} \sum_{k=0}^{T-1} (f(x_k) - f(x_*)) \leq \frac{\overbrace{\|x_0 - x_*\|^2}^{\text{initial error}}}{2\gamma T} + \frac{\gamma}{2} G^2$$

average distance
to optimality
(in function value)

- ★ very little was assumed (convexity, bounded gradients).
- ★ as $T \rightarrow \infty$ (# iterations grows), first term vanishes.
- ★ second term is fixed. So we can't guarantee convergence to zero. Not surprising, because it's not guaranteed!

ex: $|x|$. has bounded gradients



Never gets to zero!

Note: with typical functions, $\nabla f(x) \rightarrow 0$ as $x \rightarrow x_*$ so gradient steps $x_{k+1} = x_k - \gamma \nabla f(x_k)$ get smaller as we get closer.

for $|x|$, gradients are always ± 1 , so steps are always a fixed size.