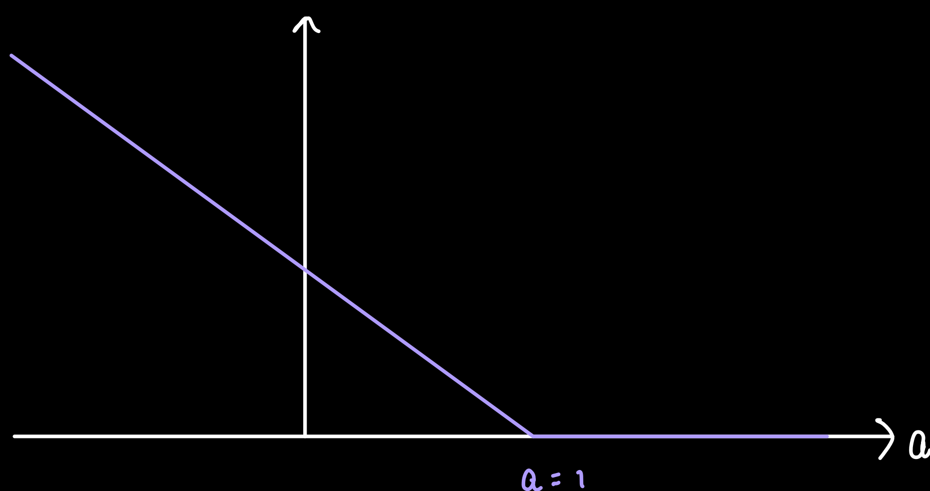
ERM:

$$\arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_\theta(x_i), y_i\right)$$

Notation: The minimizer θ (argument).

HINGE LOSS:

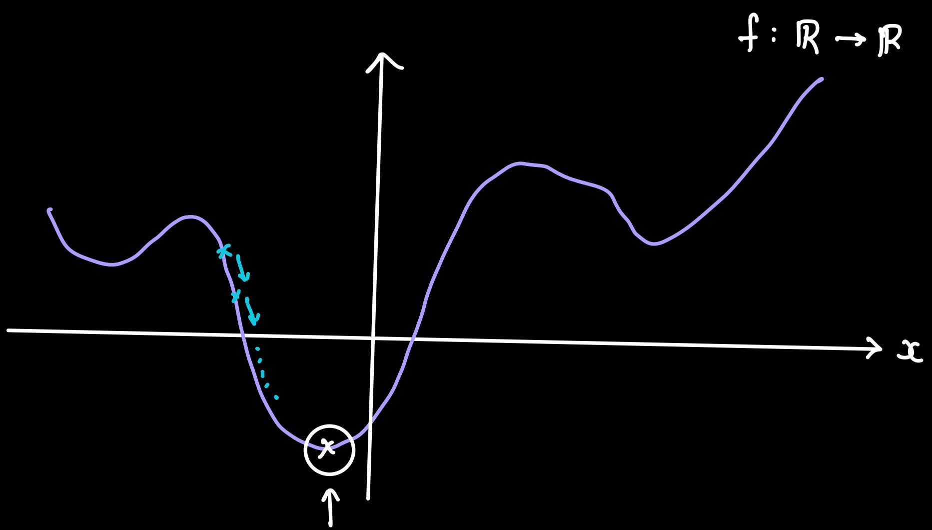
$$\ell(a,b) = \begin{cases} \max(0, 1-a) & \text{if } b > 0 \\ \max(0, 1+a) & \text{if } b < 0 \end{cases}$$



$f : \mathbb{R}^d \to \mathbb{R}$

GOAL: Find $\min_{\theta \in \mathbb{R}^d} f(\theta)$

EXAMPLE:



minimizes.

IDEA:

→ Pick a point

→ Move in a direction that reduces function value.

↓ More than one directions

→ Pick a point to start

→ Move in a direction of steepest descent.

CLAIM:

Steepest descent direction is $-\nabla f(x)$.

$$\nabla f(x) \equiv \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right).$$

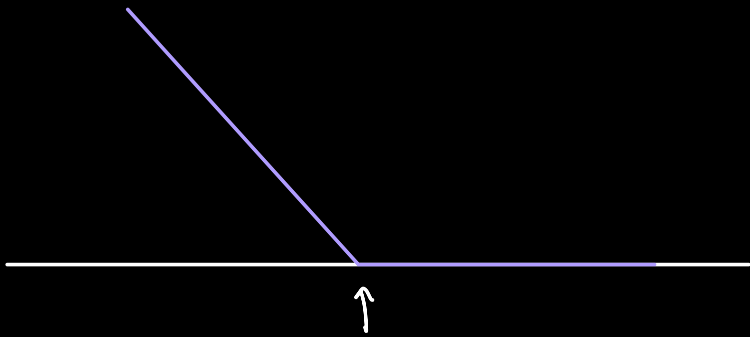
GRADIENT DESCENT ALGORITHM (GD):

1. x_0

2. for $i = 1, \ldots, T$:

$$x_i = x_{i-1} - \underbrace{\eta}_{\text{"step-size"}} \nabla f(x_{i-1})$$

DEALING WITH NON-DIFFERENTIABLE FUNCTIONS:

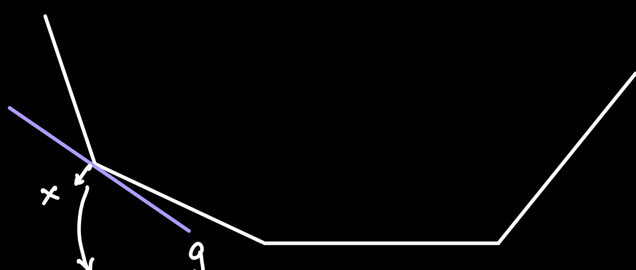


Not differentiable here!

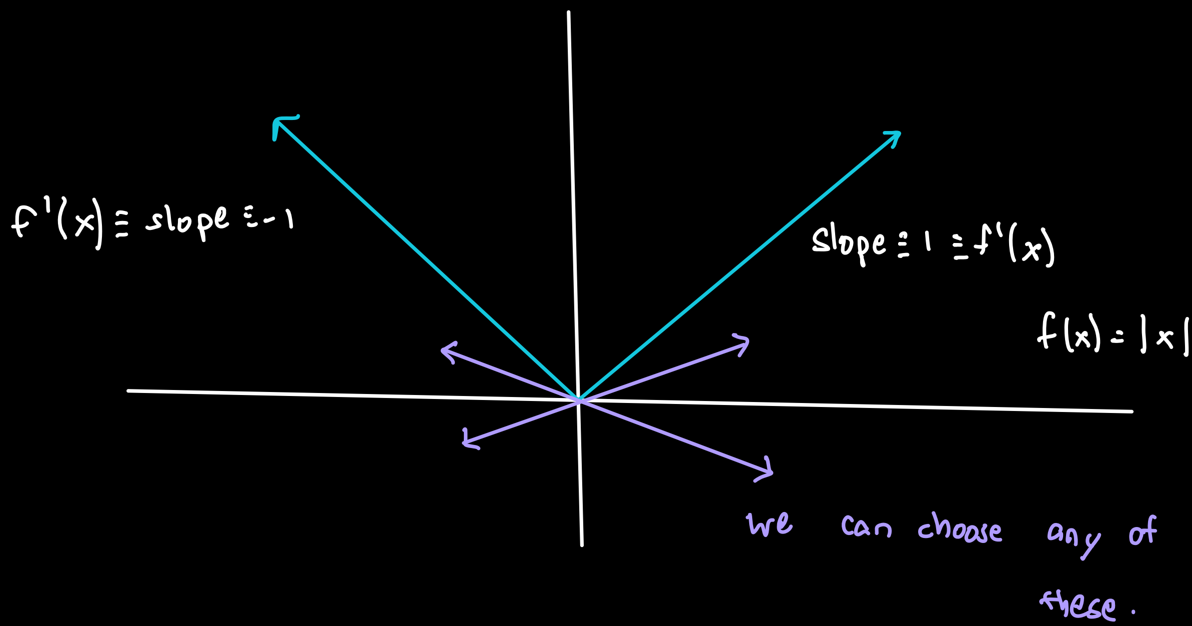
SUB-GRADIENT:

A direction g is a sub-gradient of f at x_0 if $\forall x$

$$f(x_0) + \langle g, x - x_0 \rangle \leq f(x).$$



We want a line such that the function is "above" the line.



What if no subgradient? Use local subgradient. Below $f(x)$ in some vicinity.

Summarize: We can use subgradients as substitute for gradient.

If $f(x) = |x_1| + |x_2| + \dots + |x_n|$

$g = (\mathrm{sign}(x_1), \mathrm{sign}(x_2), \dots, \mathrm{sign}(x_n))$ is a subgradient for f.

Brief idea: $f(x) = |x|$

wherever gradient is defined its $\{-1, 1\}$.

$\mathrm{sign}(x)$.

SUBGRADIENT DESCENT:

→ x_0

→ for $i = 1 \dots T$:

$$x_i = x_{i-1} - \eta g$$

a "sub-gradient" of f at x_{i-1}.

note: when differentiable sub-gradient = gradient (only option)

RECALL: GRADIENT DESCENT ALGORITHM (GD)

$f : \mathbb{R}^d \to \mathbb{R}$

1. x_0

2. for $i = 1, \dots, T$:

$$x_i = x_{i-1} - \eta \nabla f(x_{i-1})$$

WHAT WOULD WE LIKE TO SAY / KNOW ABOUT GD ON f?

$\boxed{1.}$ Does GD get me to the minimum?

$\boxed{2.}$ How many steps would GD take to get to minimum?

3. How to pick the starting point?

$\boxed{4.}$ How to choose the step-size?

$\boxed{5.}$ When do I step?

6. How do I compute ∇f?
(cost)