Lecture 13:
Gradient Descent for linear regression

Gradient descent for linear regression

- For linear regression, we have $w^* = \arg\min_{w \in \mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}w\|_2^2}_{f(w)}$
- Gradient Descent:
  - Initialize: $w_0 = 0$
  - **For** t=0,1,2,…
    - $w_{t+1} \leftarrow w_t - \eta \cdot \nabla_w f(w_t)$

$$\nabla f(w_t) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t)$$

$$w_{t+1} = w_t + \eta 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}w_t) = (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y}$$

Let the least-squares solution be $w^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\begin{aligned} w_{t+1} - w^* &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})w_t + 2\eta\mathbf{X}^T\mathbf{y} - w^* \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) + 2\eta\mathbf{X}^T\mathbf{y} - 2\eta\mathbf{X}^T\mathbf{X}w^* \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) \end{aligned}$$

Gradient Descent (GD) for Linear Regression (LR)

- We use this analytical derivation of GD for LR
  to understand how the choice of step size η impacts the algorithm

$$w_{t+1} = w_t - \eta \nabla f(w_t) \quad \Longrightarrow \quad w_{t+1} - w^* \; = \; (\mathbf{I} - 2\eta \mathbf{X}^T \mathbf{X})(w_t - w^*)$$

$$= (I - 2\eta X^T X)(I - 2\eta X^T X)(w_{t-1} - w^*)$$

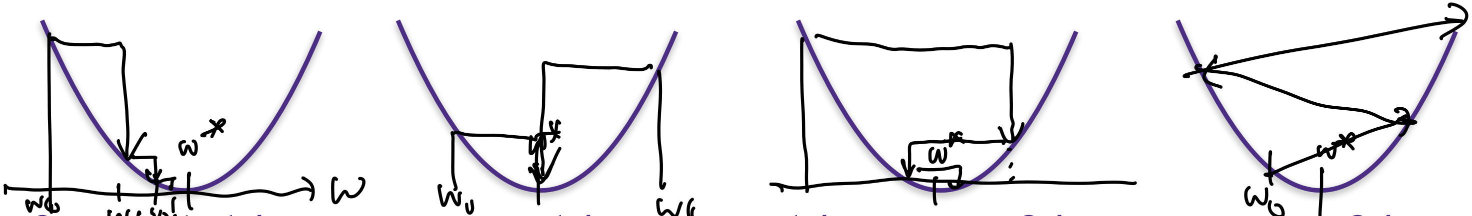
$$\vdots$$

$$= (I - 2\eta X^T X)^{t+1} \cdot \underbrace{(w_0 - w^*)}_{\text{init error}}$$

Gradient descent for linear regression

$$w_{t+1} = w_t - \eta \nabla f(w_t) \implies \begin{aligned} w_{t+1} - w^* &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})(w_t - w^*) \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})^2(w_{t-1} - w^*) \\ &\vdots \\ &= (\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X})^{t+1}(w_0 - w^*) \end{aligned}$$

In one dimension, $\overbrace{2\mathbf{X}^T\mathbf{X}} = \overset{>0}{a}$ is a scalar, and $w_{t+1} - w^* = (1 - \eta a)^{t+1}(w_0 - w^*)$



$0 < \eta < 1/a$ $\quad = 1/2$

$1 > 1 - \eta a > 0$

$\eta = 1/a$

$1 - \eta a = 0$

$1/a < \eta < 2/a$ $\quad = -1/2$

$0 > 1 - \eta a > -1$

$\eta > 2/a$

$-1 > 1 - \eta a$

0 Converges $\frac{1}{a}$ Converges $\frac{2}{a}$ Diverges η

Gradient descent for linear regression

$$w_{t+1} = w_t - \eta \nabla f(w_t) \quad \Longrightarrow \quad w_{t+1} - w^* = (\mathbf{I} - 2\eta \mathbf{X}^T\mathbf{X})^{t+1}(w_0 - w^*)$$

$$\{(q_i, D_{ii})\}_{i=1}^{d}$$

- In multi dimensions, **eigenvalues** of $\boxed{\mathbf{I} - 2\eta \mathbf{X}^T\mathbf{X}}$ are important
- Let the eigenvalue decomposition of $\mathbf{I} - 2\eta \mathbf{X}^T\mathbf{X}$ be $\underline{Q^{-1}DQ}$,
  - Where D is a diagonal matrix with Eigen values $\{D_{ii}\}_{i=1}^{d}$ in the diagonal
  - And Q is an orthogonal matrix, with each Eigen vector as in a row

$$I - 2\eta X^T X = Q^{-1} D Q, \quad D = \begin{bmatrix} D_{11} & & \\ & \ddots & \\ & & D_{dd} \end{bmatrix}, \quad Q = \begin{bmatrix} q_1^T \text{———} \\ q_2^T \text{———} \\ \text{- - —} \end{bmatrix}$$

$$w_{t+1} - w^* = \underbrace{Q^{-1} D \cancel{Q} \cdot \times \cancel{Q}^{-1} D \cancel{Q} \times \cdots \times \cancel{Q}^{-1} D Q}_{t+1} \cdot (w_0 - w^*)$$

$$= Q^{-1} \underbrace{D^{t+1}}_{\begin{array}{c} \| \\ \begin{bmatrix} D_{11}^{t+1} & & \\ & \ddots & \\ & & D_{dd}^{t+1} \end{bmatrix} \end{array}} \cdot Q \cdot (w_0 - w^*) \Rightarrow Q \cdot (w_{t+1} - w^*) = D^{t+1} Q (w_0 - w^*)$$

$$q_1^T (w_{t+1} - w^*) = D_{11}^{t+1} q_1^T (w_0 - w^*)$$

$$q_2^T (\qquad\qquad) = D_{22}^{t+1} q_2^T (\qquad)$$

$$\vdots$$

Gradient descent for linear regression

$$w_{t+1} = w_t - \eta \nabla f(w_t) \quad \Longrightarrow \quad w_{t+1} - w^* \;=\; (\mathbf{I} - 2\eta \mathbf{X}^T \mathbf{X})^{t+1}(w_0 - w^*)$$

- In multi dimensions, **eigenvalues** of $\mathbf{I} - 2\eta \mathbf{X}^T \mathbf{X}$ are important
- Let the eigenvalue decomposition of $\mathbf{I} - 2\eta \mathbf{X}^T \mathbf{X}$ be $Q^{-1}DQ$

  Then, $w_{t+1} - w^* \;=\; (Q^{-1}DQ)^{t+1}(w_0 - w^*)$

$$= \underbrace{Q^{-1}DQQ^{-1}DQ \cdots Q^{-1}DQ}_{t+1 \text{ times}} (w_0 - w^*)$$

$$= Q^{-1}D^{t+1}Q\,(w_0 - w^*)$$

$$Q(w_{t+1} - w^*) \;=\; D^{t+1}Q\,(w_0 - w^*)$$

- This defines a series of equations capturing how the error evolves in Directions defined by the rows of Q, which are the Eigen vectors of $\mathbf{I} - 2\eta \mathbf{X}^T \mathbf{X}$

Gradient descent for linear regression

$$Q(w_{t+1} - w^*) \ = \ D^{t+1} Q\,(w_0 - w^*)$$

- Where eigenvalue decomposition of $\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X}$ is $Q^{-1}DQ$
- Let $Q = \begin{bmatrix} - & q_1^T & - \\ - & q_2^T & - \\ & \vdots & \end{bmatrix}$, then the above multi-dimensional dynamics of GD can be decomposed into multiple 1-d dynamics we saw before
- The eigenvector-eigenvalue pairs $\{(q_i, D_{ii})\}_{i=1}^d$ of the matrix $\mathbf{I} - 2\eta\mathbf{X}^T\mathbf{X}$ determines the behavior of gradient descent
- In direction q_1, the error decreases multiplicatively according to D_{11}

Error in direction q_1 $\longrightarrow$ $q_1^T(w_{t+1} - w^*) = D_{11}^{t+1}\, q_1^T(w_0 - w^*)$

$$q_2^T(w_{t+1} - w^*) = D_{22}^{t+1}\, q_2^T(w_0 - w^*)$$

$$\vdots$$

Gradient descent for linear regression

$$w_{t+1} = w_t - \eta \nabla f(w_t) \implies w_{t+1} - w^* = \boxed{(\mathbf{I} - 2\eta \mathbf{X}^T\mathbf{X})^{t+1}}(w_0 - w^*)$$

$$\implies Q(w_{t+1} - w^*) = D^{t+1} Q(w_0 - w^*)$$

$$q_1^T(w_{t+1} - w^*) = D_{11}^{t+1}\, q_1^T(w_0 - w^*)$$

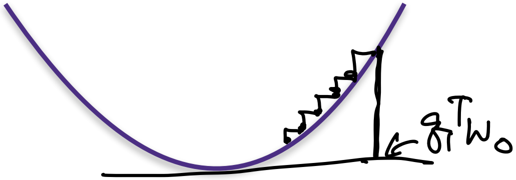
$$q_2^T(w_{t+1} - w^*) = D_{22}^{t+1}\, q_2^T(w_0 - w^*)$$

- For example suppose, the step size η is chosen such that

In direction q_1

$0 < D_{11} < 1$

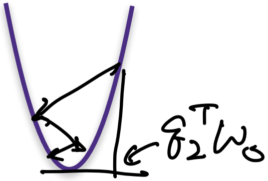
0.9

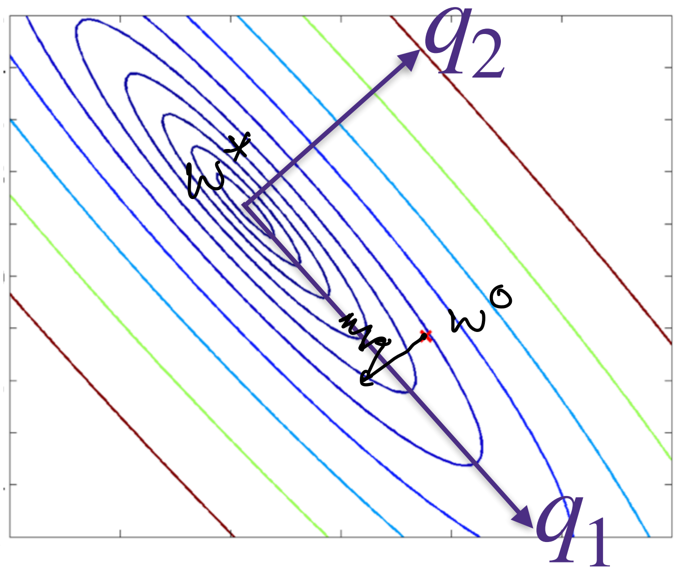


In direction q_2

$-1 < D_{22} < 0$

-0.5





Gradient descent for logistic regression

- Now we know how to find the global minimum of a logistic regression problem, numerically

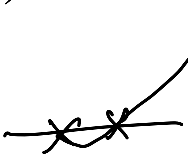
Loss function: Conditional Likelihood

$$\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i | x_i, w) \qquad P(Y = y | x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))$$

$$\nabla f(w) = \sum_{i=1}^n \frac{1}{1 + \exp(-y_i x_i^T w)} \exp(-y_i x_i^T w)(-y_i x_i)$$



$-1 + -x + e^x = 0$

What is known for Gradient descent

Convex $f(\cdot)$

- $f(\cdot)$ is L-smooth if $\|\nabla f(w) - \nabla f(v)\|_2 \le L\|w - v\|_2$ for all $w, v \in \mathbb{R}^d$
- $f(\cdot)$ is μ-strongly convex if $f(w) \ge f(v) + \nabla f(v)^T(w - v) + \frac{\mu}{2}\|w - v\|_2^2$
- For L-smooth functions, with a fixed step size $\eta < 1/L$
  - if $f(w)$ is convex,

$$f(w_t) - f(w^*) \le \frac{\overbrace{\|w_0 - w^*\|_2^2}^{R}}{2\eta t} \simeq \frac{R \cdot L}{2 \cdot t} = O\left(\frac{1}{t}\right)$$

  - if $f(w)$ is μ-strongly convex,

$$f(w_t) - f(w^*) \le \underbrace{\left(1 - \overset{\frac{1}{L}}{\eta}\mu\right)^t}_{} \underbrace{\left(f(w_0) - f(w^*)\right)}_{\Delta} = O\left(e^{-\frac{\mu}{L}\cdot t}\right)$$

- Gradient Descent is oftentimes called full-batch gradient descent to differentiate it from stochastic gradient descent, which uses only a (randomly chosen) subset of training data at each iteration
- In practice, people use Stochastic Gradient Descent (SGD).

Questions?

Model

- Model: $w^Tx+b = f_{w,b}(x)$, CNN, DNN . . .
- loss: Qudratic, Ridge, Lasso, Logistic
- Algorithm: $\nabla_w = 0$, GD, $\boxed{\text{SGD}}$, CGD.

↑

???

- Linear Model
- Quadratic or Ridge

Stochastic Gradient Descent

-What do we use in practice?

Machine Learning Problems

- Given data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$
- Learning a model's parameters: $\frac{1}{n}\sum_{i=1}^n \ell_i(w) = \frac{1}{n}\sum_{i=1}^n \ell(f_w(x_i), y_i)$

- Gradient Descent (GD):

  one update takes cdn operations/time for some constant c>0

$$w_{t+1} \leftarrow w_t - \eta \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(w_t)$$

- Stochastic Gradient Descent (SGD): one update takes cd operations/time

$$w_{t+1} \leftarrow w_t - \eta \boxed{\nabla \ell_{I_t}(w_t)}$$

I_t drawn uniform at random from $\{1, \ldots, n\}$

- SGD is an unbiased estimate of the GD

$$\mathbb{E}[\nabla \ell_{I_t}(w)] = \frac{1}{n} \cdot \sum_{i=1}^{n} \nabla \ell_i(w) = GD.$$

Stochastic Gradient Descent

$\forall(\eta)$

$\mathbb{E}[\quad] \le \min_{\eta} H(\eta) =$

Theorem

Let $w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w)\Big|_{w=w_t}$ I_t drawn uniform at random from $\{1,\ldots,n\}$ so that

$$\mathbb{E}[\nabla \ell_{I_t}(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(w) =: \nabla \ell(w)$$

If $\|w_0 - w_*\|_2^2 \le R$ and $\sup_{w} \max_{i} \|\nabla \ell_i(w)\|_2^2 \le G$ then

$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \le \frac{R}{2T\eta} + \frac{\eta G}{2} \le \sqrt{\frac{RG}{T}} = O\left(\frac{1}{\sqrt{T}}\right) \qquad \eta = \sqrt{\frac{R}{GT}}$$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

(In practice use last iterate)

Convergence rate: $O\left(\frac{1}{\sqrt{T}}\right)$

(Fixed optimal step size)

$$-\frac{R}{2T\eta^2} + \frac{G}{2} = 0$$

Taking the derivative of RHS to zero

We want to show that

$$\mathbb{E}\Big[\ell\Big(\frac{1}{T}\sum_{t=1}^{T} w_t\Big) - \ell(w_*)\Big] \le \mathbb{E}\Big[\frac{1}{T}\sum_{i=1}^{T} \ell(w_t) - \ell(w_*)\Big]$$

← Follows from convexity of $\ell(\cdot)$ and Jensen's inequality (3 slides later)

$$\le \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}\big[\ell(w_t) - \ell(w_*)\big]$$

← Follows from linearity of expectation

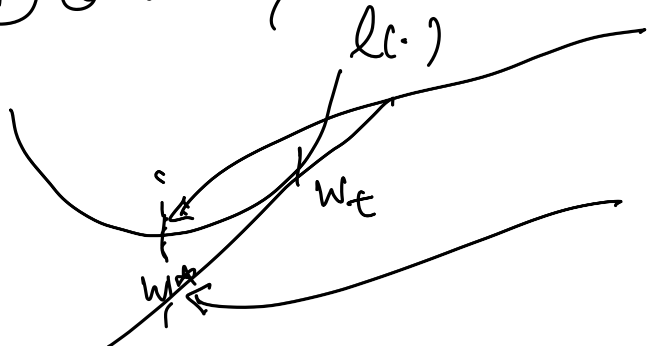
$$\le \frac{R}{2T\eta} + \frac{\eta G}{2}$$

← We are left to show this

Proof $\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||\underbrace{w_t - \eta\nabla\ell_{I_t}(w_t)}_{w_{t+1}} - w_*||_2^2]$

$$= \underbrace{\mathbb{E}[\|w_t - w_*\|_2^2]} - 2\eta\, \mathbb{E}[\overbrace{\nabla \ell_{I_t}(w_t)^T \cdot (w_t - w_*)}^{①}] + \eta^2\, \mathbb{E}[\underbrace{\|\nabla \ell_{I_t}(w_t)\|_2^2}_{G_T}]$$

① Convexity



$$\ell_i(w^*) \ge \ell_i(w_t) + \boxed{\nabla \ell_i(w_t)^T (w_* - w_t)}$$

Stochastic Gradient Descent

Proof

$$\begin{aligned}\mathbb{E}[||w_{t+1} - w_*||_2^2] &= \mathbb{E}[||w_t - \eta\nabla\ell_{I_t}(w_t) - w_*||_2^2] \\ &\underline{\leq \mathbb{E}[\|w_t - w_*\|_2^2] + \eta^2 G - 2\eta(\ell(w_t) - \ell(w_*))}\end{aligned}$$

Stochastic Gradient Descent

Proof

$$\mathbb{E}[||w_{t+1} - w_*||_2^2] = \mathbb{E}[||w_t - \eta \nabla \ell_{I_t}(w_t) - w_*||_2^2]$$

$$= \mathbb{E}[||w_t - w_*||_2^2] - 2\eta \mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*)] + \eta^2 \mathbb{E}[||\nabla \ell_{I_t}(w_t)||_2^2]$$

$$\leq \mathbb{E}[||w_t - w_*||_2^2] - 2\eta \mathbb{E}[\ell(w_t) - \ell(w_*)] + \eta^2 G$$

$$\mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*)] = \mathbb{E}\big[\mathbb{E}[\nabla \ell_{I_t}(w_t)^T (w_t - w_*) | I_1, w_1, \dots, I_{t-1}, w_{t-1}]\big]$$

$$= \mathbb{E}\big[\nabla \ell(w_t)^T (w_t - w_*)\big]$$

$$\geq \mathbb{E}\big[\ell(w_t) - \ell(w_*)\big]$$

$$\sum_{t=1}^{T} \mathbb{E}[\ell(w_t) - \ell(w_*)] \leq \frac{1}{2\eta} \left(\mathbb{E}[||w_1 - w_*||_2^2] - \mathbb{E}[||w_{T+1} - w_*||_2^2] + T\eta^2 G \right)$$

$$\leq \frac{R}{2\eta} + \frac{T\eta G}{2}$$

We have: $\sum_{t=1}^{T} \mathbb{E}[\ell(w_t) - \ell(w_*)] \leq \frac{R}{2\eta} + \frac{T\eta G}{2}$

$$\mathbb{E}[\ell(\bar{w}) - \ell(w_*)] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell(w_t) - \ell(w_*)]$$

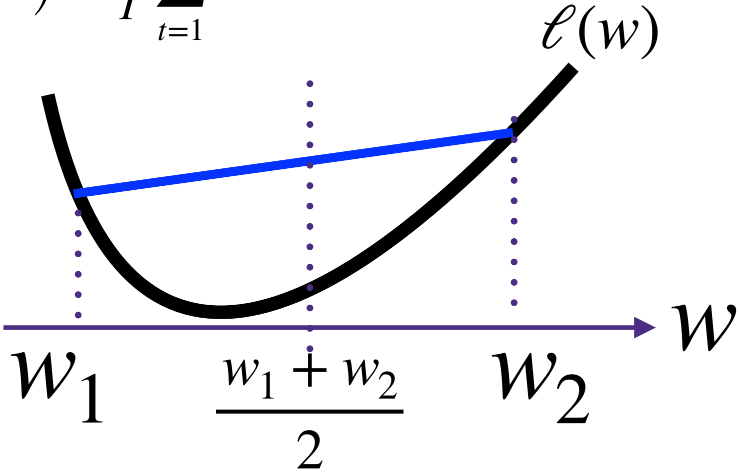
$$\leq \frac{R}{2\eta T} + \frac{\eta G}{2}$$

$$\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

Jensen's inequality:

For any $\{w_1, \ldots, w_T\}$ and a convex function $\ell(\cdot)$, we have

$$\ell\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(w_t)$$



Mini-batch SGD

- Instead of one iterate, average B stochastic gradient together

- Advantages:
  - Smaller variance: the variance of the stochastic gradient is smaller by a factor of $1/\sqrt{B}$
  - Parallelization: each gradient in the mini-batch can be computed in parallel

- If you have regularizer, $\frac{1}{n}\sum_{i=1}^{n}\ell_i(w) + r(w)$, then update with the stochastic gradient of the loss and gradient of the regularizer

Sparsity/Complexity tradeoff

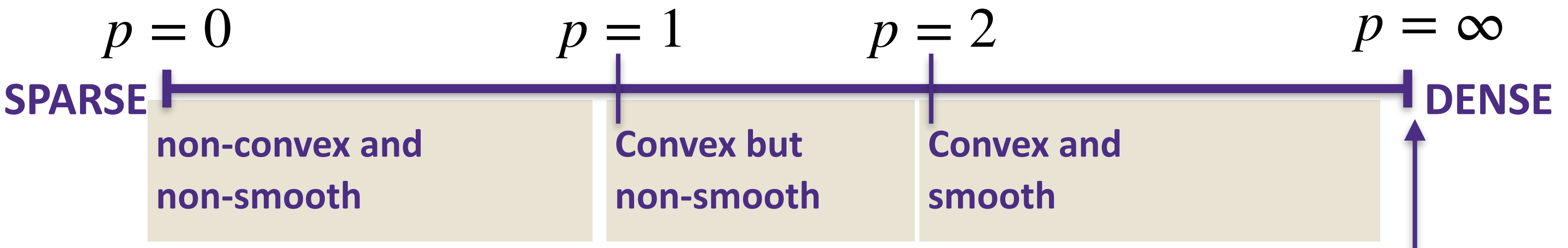
- ℓ_p-norm of a vector is defined as $\|w\|_p \triangleq \left(w_1^p + w_2^p + \cdots + w_d^p\right)^{1/p}$
- Consider regularized least squares problem of minimizing

$$\mathscr{L}(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 \;+\; \lambda \, \|w\|_p^p$$

- This is ridge regression for $p = 2$ and Lasso for $p = 1$

$\|w\|_0 =$ # of non-zero entries

$\|w\|_\infty = \max\{w_i\}$



$p = 0$

$p = 1$

$p = 2$

$p = \infty$

SPARSE

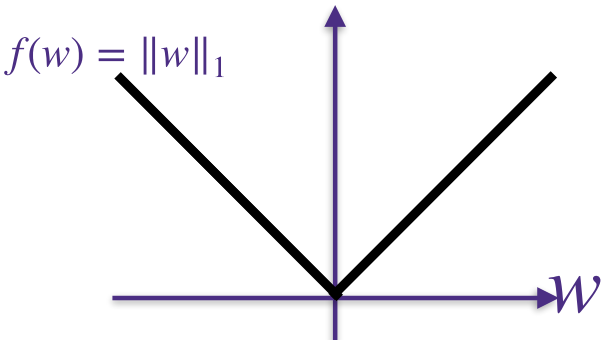
DENSE

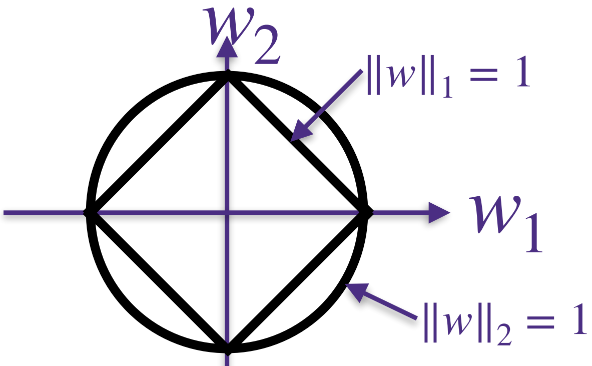
non-convex and non-smooth

Convex but non-smooth

Convex and smooth

convex but non-smooth





Questions?