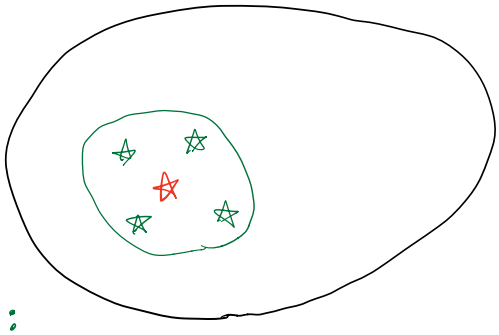
## Lecture Plan

- ☆ Linear regression and the minimum norm estimator
- ☆ Logistic regression and the max-margin classifier
- ☆ Matrix sensing

## Overview: Implicit regularization

We discussed an intriguing question in Lecture 4: why does a NN trained with many more parameters than number of samples generalize well to unseen data?

[C. Zhang et al. ICLR'17] Modern NN contains so many parameters that they are capable of fitting random labels perfectly!




When the # parameters $\gg$ # samples, then there are many solutions in the constrained set!

- Hypotheses: Regularization? Weight decay, Dropout both do not help.
- Alternative hypotheses: the optimization algorithm (SGD) more specifically) has a certain bias towards certain solutions. a.k.a. the inductive bias of the alg.

What kind of inductive biases do we expect from an alg?

Implicit regularization hypothesis. SGD biases towards "simple" solutions that generalizes well.

* Linear regression and the minimum norm estimator

Suppose we have $n$ samples $(x_1, y_1), (x_2, y_2) \cdots, (x_n, y_n)$.

$x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

linear regression: $L(w) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \cdot w - y_i)^2$

Assume: # of parameters in $w = p > n =$ # of samples

- Over-parametrized
- $x_1, x_2, \cdots, x_n$ are linearly independent

Fact: If $\exists$ one solution to $L(w) = 0$, then $\exists$ infinitely many solutions.

Under-determined and feasible linear system

Ex. Verify this fact

Which solution do we want?

Minimum norm estimator

$$\min. \quad \|w\| = \sum_{i=1}^{p} w_i^2$$

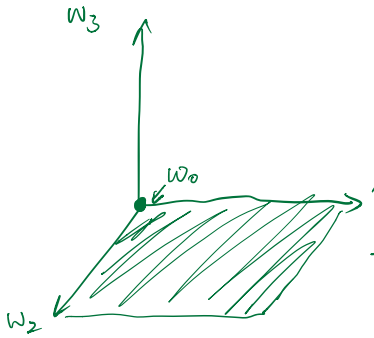
$$s.t. \quad x_i^T w = y_i, \quad \text{for } i = 1, 2, \cdots, n.$$

Claim. Gradient descent with initialization $w_0 = \vec{0}$ converges to the minimum norm estimator.

Proof. (1) $\nabla L(w) = \frac{2}{n} \sum_{i=1}^{n} (x_i^T w - y_i) \cdot x_i$

$$= \sum_{i=1}^{n} a_i \cdot x_i, \qquad \left[\text{let } a_i = \frac{2}{n} \cdot (x_i^T w - y_i)\right]$$

$$\Rightarrow \nabla L(w) \in \text{span}\{x_1, x_2, \cdots, x_n\}.$$




→ span of (1,0,0) and (0,1,0)

GD: $w_{t+1} = w_t - \eta \cdot \nabla L(w_t)$.

If $\nabla L(w_t) \in \text{span}\{x_1, \cdots, x_n\}$,

then $w_{t+1} \in \text{span}\{x_1, \cdots, x_n\}$.

We know that $w_0 = 0 \in \text{span}\{x_1, \cdots, x_n\}$,

therefore, by induction we have that

$w_0, w_1, w_2, \cdots w_\infty$ are all in the span!

(2) Because the problem is convex, GD will converge to the global minimum, i.e. with zero loss

Furthermore, this global minimum is in the span of $\{x_1, x_2, \cdots x_n\}$.

(3) There is a unique solution to the linear system in the span of $\{x_1, \cdots, x_n\}$.

- Prove by contradiction. Suppose that

(i) $w_1, w_2$ are both in span of $\{x_1, \cdots x_n\}$

(ii) $w_1^T x_i = w_2^T x_i = y_i$, for all $i = 1, 2, \cdots, n$.

$$\Rightarrow (w_1 - w_2)^T x_i = 0, \quad \forall\, i = 1, 2, \cdots, n.$$

$\Rightarrow \quad w_1 - w_2 = 0$ !

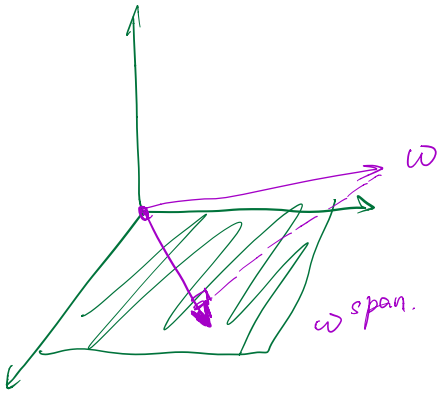
This is because we can write $w_1 - w_2$ as a linear combination of $\{x_1, \dots, x_n\}$.

Since $w_1 - w_2$ is also in the span

$\Rightarrow \quad \| w_1 - w_2 \|^2 = 0$

- The minimum norm estimator is in the span !

Proof by picture :-)




unique

$w^{opt} = w^{span} + w^{\text{out of span}}$

By triangle inequality,

$\| w^{opt} \| \geq \| w^{span} \|$

$\Rightarrow$ $w^{span}$ is the minimum norm estimator.

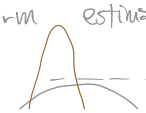
To recap, we have shown ~~that~~ for linear regression, gradient descent converges to the minimum norm estimator.

Question: why is minimum norm estimator a good estimator ? There has been a flurry of research on this question.

- Intuitive answer: the minimum norm estimator adds a strong regularization.

[Benign overfitting in linear regression]

→ Minimum norm estimator in an overparametrized setting.



→ flatten noise

## * Logistic regression and the max-margin classifier

### Logistic regression:

Suppose we have $n$ samples, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.

$x_i \in \mathbb{R}^p$, $y_i \in \pm 1$.

We would like to classify these sample into two classes:

$$\mathcal{L}(w) = \sum_{i=1}^{n} \exp(-y_i \cdot w^T x_i)$$

- If $y_i = +1$, then the loss is smaller when $w^T x_i > 0$ compared to when $w^T x_i < 0$
- If $y_i = -1$.

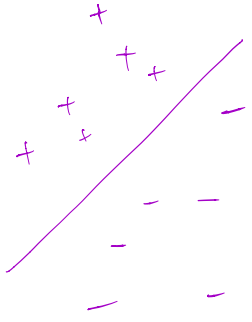
Ex

Gradient descent:

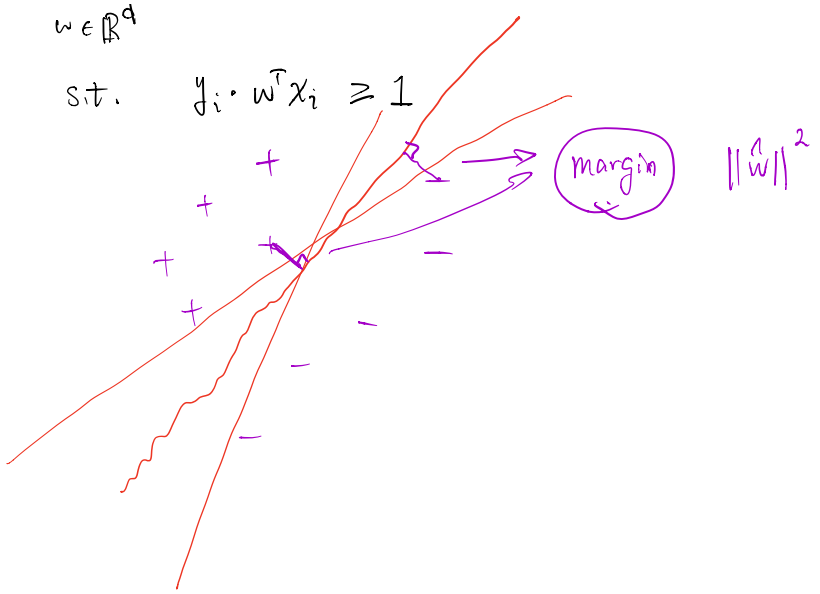
$$w(t+1) = w(t) - \eta \cdot \nabla \mathcal{L}(w(t))$$

Claim. Starting from any initialization $w(0)$, GD converges to the "so-called" max-margin classifier.

Max-margin classifier. Assume that the dataset is separable.



$$\hat{w} = \arg\min \|w\|^2$$




Separable dataset ↑

Is this a reasonable assumption? Not always, but if # parameters $\gg$ # samples, then the assumption always holds

More formally, we are going to show that (under several mild assumptions)

$w(t) = \hat{w} \cdot \log t + \rho(t)$, where $\rho(t)$ is a residual vector whose norm is less than $O(\log\log t)$

As a result, $\frac{w(t)}{\|w(t)\|} \to \frac{\hat{w}}{\|\hat{w}\|}$, as $t \to \infty$.

Remark. (i) The convergence is extremely slow, which is consistent w/ our experience w/ logistic regression in practice.

(ii) The max-margin classifier is the same as SVM.

Proof. Idea: Show that the residual is negligible compared to $\log t$.

(0) setup some notations.

$S$ : support set. "active constraints" from solving SVM.

" $y_i \cdot x_i^T w = 1$ "

$S^c$ : non-support set

$\hat{w}$ : solution of $\eta \cdot \exp(-x_i^T \cdot \hat{w}) = \alpha_i$ for $i \in S$ → Ex. KKT conditions. Lagrangian multiplier of the SVM

↳ assume that rank of $S$ is full, so there is a unique solution to the above linear system.

Let $r(t) = w(t) - \log(t) \cdot \hat{w} - \tilde{w}$.

(1) We show that $\|r(t)\|$ is bounded by some fixed constant. This concludes the proof since $\tilde{w}$ does not grow with $t$.

We consider the learning rate $\eta$ to be extremely small.

In this case, GD = Gradient flow (or PDE).

$$\nabla r(t) = -\nabla \mathcal{L}(w(t)) - \frac{1}{t} \cdot \hat{w}$$

The change in $\|r(t)\|^2$ is

$$\frac{1}{2} \frac{d}{dt} \|r(t)\|^2 = \nabla r(t)^T r(t)$$

$$= \left(-\nabla \mathcal{L}(w(t)) - \frac{1}{t} \cdot \hat{w}\right)^T r(t).$$

Recall that $\mathcal{L}(w(t)) = \sum \exp(- \cdot w(t)^T x_i)$.

$$\Rightarrow \nabla \mathcal{L}(w(t)) = \sum \exp(- \quad \cdot w(t)^T x_i) \cdot (-y_i \cdot x_i)$$ Ex.

$$= \sum_{i=1}^{n} \exp(- x_i^T \cdot w(t)) \cdot x_i^T r(t) - \frac{1}{t} \cdot \hat{w}^T r(t)$$

$$= \sum_{i \in S} \exp(- x_i^T \cdot w(t)) \cdot x_i^T r(t) - \frac{1}{t} \hat{w}^T r(t) \quad (*)$$

$$+ \sum_{i \notin S} \exp(- x_i^T w(t)) \cdot x_i^T r(t) \quad (**)$$

(2) $(*) \le 0$, $(**) \le O(\frac{1}{t^\theta})$, for $\theta = \underset{i \notin S}{\operatorname{argmin}} x_i^T \hat{w} > 1$

$$\Rightarrow \|r(t)\|^2 - \|r(t_1)\|^2 \le C \cdot \int_{t_1}^{t} \frac{dt}{t^\theta} \le C' < \infty.$$

$(*)$ Recall that $r(t) = w(t) - \log(t) \cdot \hat{w} - \tilde{w}$.

Hence $(*) = \sum_{i \in S} \exp(- x_i^T r(t) - \log(t) \cdot x_i^T \hat{w} - x_i^T \tilde{w}) \cdot x_i^T r(t) - \frac{1}{t} \hat{w}^T r(t)$

(i) $= 1$ because of margin condition.

(ii) $\sum_{i \in S} \exp(-\hat{w}^T x_i) \cdot x_i = \hat{w}$

$$= \frac{1}{t} \sum_{i \in S} \exp(- \tilde{w}^T x_i) \cdot \exp(-x_i^T r(t)) \cdot x_i^T r(t)$$

$$- \frac{1}{t} \sum_{i \in S} \exp(-\hat{w}^T x_i) \, x_i^T r(t)$$

$$= \frac{1}{t} \sum_{i \in S} \exp(-\hat{w}^T x_i) \cdot \left( \exp(-x_i^T r(t)) - 1 \right) \cdot x_i^T r(t)$$

$z = x_i^T r(t)$ Ex. $\forall z$

$\Rightarrow \le 0$

$$(**) = \sum_{i \notin S} \exp\left(-x_i^T r(t) - \log(t) \cdot x_i^T \hat{w} - x_i^T \tilde{w}\right) \cdot x_i^T r(t)$$

$z \cdot (\exp(-z) - 1) \leq 0$

$\rightsquigarrow$ (i) $\geq \theta > 1$ because of the margin condition

$$\leq \sum_{i \notin S} \frac{1}{t^\theta} \cdot \exp(-\hat{w}^T x_i) \cdot \exp(-x_i^T r(t)) \cdot x_i^T r(t)$$

$z = x_i^T r(t)$.

Ex. $\forall z$, $\exp(-z) \cdot z \leq 1$.

$$\Rightarrow \leq \frac{1}{t^\theta} \cdot \sum_{i \notin S} \exp(-\tilde{w}^T x_i)$$

→ a fixed constant.

Therefore, we have shown that (*), (**) both are correct.

To recap, we have shown that grad. descent converges to the max-margin classifier!

---

* Matrix sensing

  - The linear regression example is great, but it only works in the linear setting, which is limited.

    Matrix sensing is a non-linear case, where gradient descent also converges to the minimum norm estimator

  - Norm

* NTK

  - Beyond the above two-layer neural nets.