16B Prof. Anant Sahai

Today: Complex Inner Products
Loose Ends: Why logistic loss (HW 13)
Begin Review

Announcements: Final Exam $\boxed{\text{8 am Fri Dec 17th TBD}}$

0) HW 13 Due Friday
1) HW 14 on different schedule to help with finals prep
Due Wed Dec 8th. Self-grade & Resubmit Due Sat Dec 11th

2) Extra Credit Opportunities
 a) MYOP Contest
 b) 16B Content-illuminating Memes + Lab Videos & Course Ads
 c) Code cleaning & style/robustness improvements

3) RRR week discussions: Review Oriented
based on past exams. Foci to be announced on Piazza

4) Conceptual Questions: Use Week 15 announcement thread
We'll have a vote this week to prioritize if lab
& then we'll get them answered for everyone

Final Exam Scope: Cumulative & Integrative
Majority of points will touch Post-Midterm Material, but could be combined with Pre-midterm ideas.

3 hrs

Get Rest & Sleep.

Mastery of HW, Discussions, Lectures, Notes & Lab
No need to grind past exams to study.
↳ likely have different scope.

Same Policies as Midterm: No Questions.

Why do we need a complex inner product?

A: Orthogonality & Projections are super useful. (i.e. We used these to get Upper-triangularization)

For real vectors $\vec{a}, \vec{b}$ we had $\langle \vec{a}, \vec{b} \rangle = \vec{b}^T \vec{a} = \vec{a}^T \vec{b} = \sum_{i=1}^{n} a_i b_i$

Had nice property $\|\vec{a}\|^2 = \langle \vec{a}, \vec{a} \rangle$

$$\sum_{i=1}^{n} |a_i|^2 = \underbrace{\sum_{i=1}^{n} a_i^2}_{\text{Requires all the } a_i \text{ to be real.}}$$

For complex a_i, we would have $|a_i|^2 = \overline{a_i}\, a_i$

So, the real inner product definition doesn't work for complex vectors

Example: $\vec{a} = \begin{bmatrix} j \\ j \end{bmatrix}$ $\quad \|\vec{a}\|^2 = 2 \quad \vec{a}^T \vec{a} = -2$

We need another definition

To get $\langle \vec{a}, \vec{a} \rangle = \|\vec{a}\|^2$, we could define it in two possible natural ways:

$$\langle \vec{a}, \vec{b} \rangle \stackrel{?}{=} \overline{\vec{b}}^T \vec{a} = \sum_{i=1}^{n} a_i \overline{b_i}$$

$$\stackrel{?}{=} \overline{\vec{a}}^T \vec{b} = \sum_{i=1}^{n} \overline{a_i}\, b_i$$

Not symmetric, but oh well.

Which one should we pick? Should pick the one more convenient for us conceptually.

"Unit tests for our definition"

1) $\langle \vec{a}, \vec{a} \rangle = \|\vec{a}\|^2$

2) If $\|\vec{b}\| = 1$ and $\vec{a} = \alpha \vec{b}$, then $\langle \vec{a}, \vec{b} \rangle = \alpha$.

3) If $\langle \vec{a}, \vec{b} \rangle = 0$, then $\vec{a} \perp \vec{b}$ and Pythagoras holds:
$\|\vec{a}+\vec{b}\|^2 = \|\vec{a}\|^2 + \|\vec{b}\|^2$

4) If $\langle \vec{a}, \vec{b} \rangle = \alpha$ then $\langle \beta \vec{a}, \vec{b} \rangle = \alpha \beta$.

Linearity in first argument.

5) "Easy to remember & use." $\rightarrow$??

Properties (1) & (3) are satisfied by both candidate definitions.

$$\|\vec{a}+\vec{b}\|^2 = \langle \vec{a}+\vec{b}, \vec{a}+\vec{b} \rangle \overset{[\text{Def } 1]}{\underset{[\text{Def } 4]}{=}} \overline{(\vec{a}+\vec{b})}^T (\vec{a}+\vec{b})$$
$$= (\vec{a}+\vec{b})^* (\vec{a}+\vec{b})$$

Define Notation $\vec{b}^* \overset{\text{def}}{=} \bar{\vec{b}}^T$

$$= \vec{a}^* \vec{a} + \vec{b}^* \vec{b} + \vec{a}^* \vec{b} + \vec{b}^* \vec{a} = \|a\|^2 + \|b\|^2$$

Under Def 4. We know $\vec{b}^* \vec{a} = 0 = \sum_{i=1}^{n} a_i \bar{b}_i$

$$\bar{0} = 0 = \overline{\left(\sum_{i=1}^{n} a_i \bar{b}_i\right)} = \sum_{i=1}^{n} \bar{a}_i b_i = 0$$

Def 2 We know $\vec{a}^* \vec{b} = 0 \implies \vec{b}^* \vec{a} = 0$

Property 2 makes the choice,

If $\|\vec{b}\| = 1$ and $\vec{a} = \alpha \vec{b}$, then $\langle \vec{a}, \vec{b} \rangle = \alpha$.

<u>Def 4:</u> $\langle \vec{a}, \vec{b} \rangle = \vec{b}^* \vec{a} = \vec{b}^* \alpha \vec{b} = \alpha \vec{b}^* \vec{b} = \alpha \|\vec{b}\|^2 = \alpha$

<u>Def 2:</u> $\langle \vec{a}, \vec{b} \rangle = \vec{a}^* \vec{b} = \bar{\alpha} \vec{b}^* \vec{b} = \bar{\alpha} \neq \alpha$ if α is complex.

<u>Didn't work</u>

Only Definition 4 works

Any attempt to make it symmetric would break this.

Verify ④:

4) If $\langle \vec{a}, \vec{b} \rangle = \alpha$ then $\langle \beta \vec{a}, \vec{b} \rangle = \alpha\beta$.

$$\langle \beta \vec{a}, \vec{b} \rangle = \vec{b}^* \beta \vec{a} = \beta\, \vec{b}^* \vec{a} = \beta \langle \vec{a}, \vec{b} \rangle = \alpha\beta.$$

Aside: What can/do we mean when we say project $\vec{a}$ onto $\vec{b}$ when $\vec{b}$ is a unit vector?

Meaning 1: We want the scalar coordinate that tells us what multiple of $\vec{b}$ is closest to $\vec{a}$.

$\langle \vec{a}, \vec{b} \rangle$ gives us this scalar if $\|\vec{b}\| = 1$

and $\dfrac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{b}\|^2}$ gives it to us in general.

Meaning 2: We want the vector that is a multiple of $\vec{b}$ closest to $\vec{a}$.

$\vec{b} \langle \vec{a}, \vec{b} \rangle$ if $\|\vec{b}\| = 1$

and $\dfrac{\vec{b} \langle \vec{a}, \vec{b} \rangle}{\|\vec{b}\|^2}$ in general if $\|\vec{b}\| \neq 1$.

What about Property/Test 5: Easy to use.

Whenever we had a transpose $(\cdot)^T$ for the real case, we just use a conjugate transpose $(\cdot)^*$ for the complex case.

How to check? Recall that if $B = [\vec{b}_1\ \vec{b}_2 \cdots \vec{b}_e]$ has orthonormal columns

then $B B^T$ gives us the vector projection onto the subspace spanned by the columns of B.

For the real case.

Q: Does BB^* give us projections for orthonormal B?

If $\vec{x} = \sum_{i=1}^{k} \alpha_i \vec{b}_i$ (So $\vec{x}$ is in the span of columns of B)

then
$$\underbrace{BB^*}_{\text{Gives projection}} \vec{x} = BB^* \sum_{i=1}^{k} \alpha_i \vec{b}_i$$
$$= \sum_{i=1}^{k} BB^* \alpha_i \vec{b}_i$$
$$= \sum_{i=1}^{k} \alpha_i BB^* \vec{b}_i$$
$$= \sum_{i=1}^{k} \alpha_i \vec{b}_i$$

$$BB^* \vec{b}_i = \begin{bmatrix} \vec{b}_1 \cdots \vec{b}_\ell \end{bmatrix} \begin{bmatrix} \vec{b}_1^* \\ \vec{b}_2^* \\ \vdots \\ \vec{b}_\ell^* \end{bmatrix} \vec{b}_i = \begin{bmatrix} \vec{b}_1 \cdots \vec{b}_\ell \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{th} \text{ position}$$
$$= \vec{b}_i$$

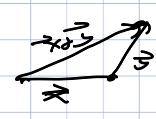
And moreover,

Clearly $B^* \vec{x} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_\ell \end{bmatrix} \leftarrow$ Gives coordinates of projection

$$\underbrace{B \underbrace{B^* \vec{x}}_{\text{Gives coordinates of projection}}}_{\text{Gives the vector projection itself}}$$

Check For Yourself: 1) Triangle Inequality

$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$



2) Cauchy-Schwartz: $|\langle \vec{x}, \vec{y} \rangle| \leq \|\vec{x}\| \cdot \|\vec{y}\|$

<u>Summary</u>: $\langle \vec{a}, \vec{b} \rangle = \vec{b}^* \vec{a}$

& for results, replace $(\cdot)^T$ with $(\cdot)^*$

This gives us: 1) Gram-Schmidt Orthonormalization (Discussion will elaborate)

HW 14 will have you practice

2) Upper-Triangularization in General.

$$A = V \, U \, V^{*}$$

V^* ← conjugate transpose

V: orthonormal basis (complex)

U: upper-triangular has λ's (eigenvalues) on diagonal.

3) If Q is orthonormal & square then $Q^{*}Q = I$

$$\Longrightarrow Q^{-1} = Q^{*}$$

4) Complex matrix SVD $\quad M = U \, \Sigma \, V^{*}$

V^* ← orthonormal.

U: orthonormal

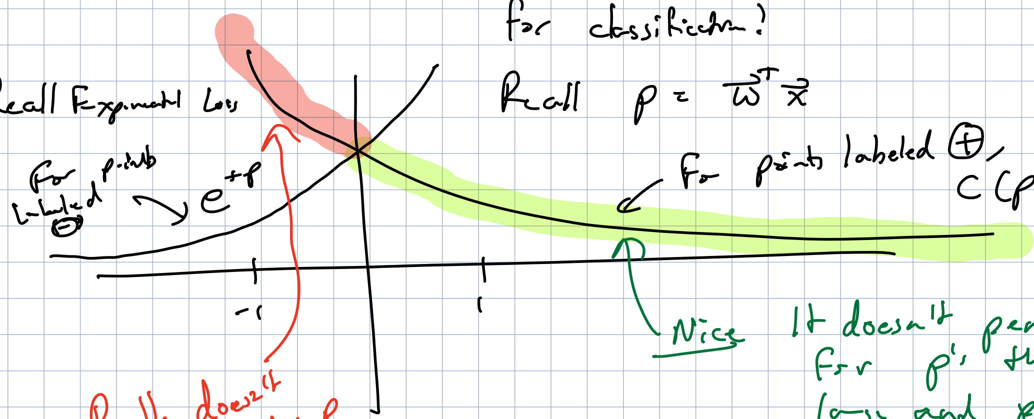
Σ: Real & Diagonal.

Notation: "Unitary" means square orthonormal.

Loose End: (HW 13) What's wrong with the exponential loss for classification?

Recall Exponential Loss

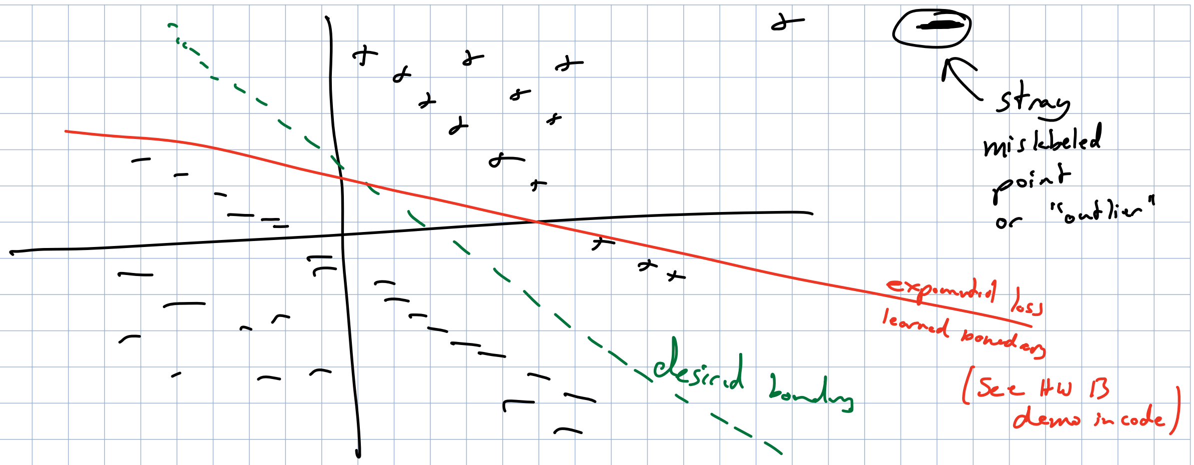
Recall $p = \vec{w}^{T}\vec{x}$



Really doesn't like negative p when the point is labeled ⊕.

Nice It doesn't penalize us for p's that are large and positive if the point is labeled ⊕.

Real-World Problem: Mislabeled Data. Data might not be perfect.



Classification by optimization is a "tug of war" between points.

Exponential loss can result in overpowered points that ruin balance.

Can want to "mellow out" the loss to avoid overpowers mislabeled points.

$e^{-\rho}$ for positive ρ is fine

but is overpowered for negative ρ. ← How to fix this without ruining

Bode Plots suggest log

$$\ln(1+x) \approx \begin{cases} x & \text{if } x \text{ is small} \\ \ln x & \text{if } x \text{ is big.} \end{cases}$$

Bode Plots tell us that $1+x \approx \max(1, x)$

$$\ln(1+e^{-\rho}) \approx \begin{cases} e^{-\rho} & \text{if } \rho \text{ is large +} \\ -\rho & \text{if } \rho \text{ is negative.} \end{cases}$$

is a lot nicer than $e^{-\rho}$ for ρ negative

Gives Logistic loss.

less sensitive to outliers that are mislabeled while preserving nice attributes of exponential loss for properly labeled points.