Welcome!

This course is **not**

- A rigorous introduction to linear algebra, statistics, optimization, machine learning and its applications
- A full 100-unit class with letter grades, lots of homeworks & exams

It **is**

- A special topics course focusing on machine learning methods that use linear algebraic machinery ("spectral techniques")
- A pass/fail 50-unit class, no homeworks or exams (probably)

More like a **tutorial + reading group**
How to Not Fail the Course

- Clearly designed for self-motivated grad/undergrad researchers
  - Implicit assumption: You already know machine learning and just want to learn about the topic.

- Pass/fail judged on participation and **paper presentation**
  - Must have enough substance to give a full lecture to the class and “demonstrate deep understanding”
  - There *might* be a mini quiz towards the end for an extra measurement... So don’t be too comfortable :)

- Logistics
  - Course number: TTIC 41000 (TTIC Room 526)
  - Time: M 3-4:20pm (office hours M 4:30-5pm)
  - Course materials found on the [course website](#)
Overview

Topics

Review on Vector Space

Vector Space
Inner Product Space
Relevance of Spectral Techniques in Machine Learning

- Functional analysis
- Subspace identification (e.g., for parameter estimation)
- Optimization
- Neural networks
What can we say about the **training loss**?

- **Example: semiparametric regression** (Dudeja and Hsu, 2018)

  \[ y = g(u^* \cdot x) + \epsilon \quad x \sim \mathcal{N}(0, I_p), \ \epsilon \sim \mathcal{N}(0, \sigma^2) \]

  \( g : \mathbb{R} \to \mathbb{R} \) unknown smooth function

- **Learning: minimize over unit-length** \( u \in \mathbb{R}^p \)

  \[ R_L(u) = \min_{h \in P_L} E_{x,y} \left[ (y - h(u \cdot x))^2 \right] \]

- By characterizing \( g(z) = \sum_{l=0}^{\infty} a_l^* H_l(z) \) in the Hermite polynomial basis, one can show that

  \[ R_L(u) = \sigma^2 + \sum_{l=1}^{L} (a_l^*)^2 (1 - (u \cdot u^*)^{2l}) \]
Can we recover **low-dimensional** structure from **high-dimensional** observations?

▶ Example: weighted finite automaton (Balle et al., 2014)

\[
f(x_1 \ldots x_N) = \alpha^\top A^{x_1} \cdots A^{x_N} \beta
\]

Unknown function \( f : \mathcal{X}^* \to \mathbb{R} \) maps a sequence of symbols \( x = (x_1 \ldots x_N) \) to a number \( f(x) \).

▶ It is assumed that \( k \ll |\mathcal{X}| \).

▶ Problem: efficiently learn \( f \) from samples of \( (x, f(x)) \).

▶ Model parameters recovered up to rotation by performing rank-\( k \) **singular value decomposition (SVD)** on

\[
\Omega = U \Sigma V^\top
\]

for \( \Omega_{x,y} = f(xy) \).
Can we use decomposition techniques to solve optimization problems?

Example: canonical correlation analysis (CCA) (Hotelling, 1936)

\[(a, b) = \arg \max_{u \in \mathbb{R}^d, v \in \mathbb{R}^{d'}} \text{corr} \left( u^\top X, v^\top Y \right)\]

Find projection vectors to maximize the correlation between random variables \(X, Y\).

Solution given by rank-1 SVD on

\[\mathbf{E} \left[ XX^\top \right]^{-1/2} \mathbf{E} \left[ XY^\top \right] \mathbf{E} \left[ YY^\top \right]^{-1/2} \in \mathbb{R}^{d \times d'}\]
Neural Networks

Most of deep learning is **matrix manipulation**.

- Thus matrix skills are useful even if you only do neural networks.
- Word2vec and language modeling can both be seen as matrix factorization problems (Levy and Goldberg, 2014; Yang et al., 2017)
- Solid background in spectral techniques is just generally useful for various problems in machine learning.

  - For instance, is there a solution to

    \[
    \begin{bmatrix}
    9 & 3 \\
    6 & 5 \\
    0 & 10
    \end{bmatrix}
    \begin{bmatrix}
    x_1 \\
    x_2
    \end{bmatrix}
    =
    \begin{bmatrix}
    1 \\
    2 \\
    -2
    \end{bmatrix}
    \]
Overview

Topics

Review on Vector Space

*Vector Space*

*Inner Product Space*
Vector Space

**Vector space** $V$ over field $\mathbb{F}$ is a set containing 0, equipped with

- **Vector addition** $V \times V \rightarrow V$ denoted $(u, v) \mapsto u + v$ such that
  
  \[
  u + v = v + u \\
  (u + v) + w = u + (v + w) \\
  u + 0 = u
  \]

  and every $u \in V$ has additive inverse $-u \in V$, $u + (-u) = 0$.

- **Scalar multiplication** $\mathbb{F} \times V \rightarrow V$ denoted $(\alpha, u) \mapsto \alpha u$ such that
  
  \[
  \alpha(u + v) = \alpha u + \alpha v \\
  (\alpha + \beta)u = \alpha u + \beta u \\
  \alpha(\beta u) = (\alpha \beta)u \\
  1u = u \\
  0u = 0 \\
  (-1)u = -u
  \]
Vector Space Examples

1. Euclidean space. $\mathbb{R}^d$

\[
(\alpha_1, \ldots, \alpha_d) + (\beta_1, \ldots, \beta_d) := (\alpha_1 + \beta_1, \ldots, \alpha_d + \beta_d)
\]
\[
\gamma(\alpha_1, \ldots, \alpha_d) := (\gamma \alpha_1, \ldots, \gamma \alpha_d) \quad \forall \gamma \in \mathbb{R}
\]

2. Sequence space. $\mathbb{R}^\infty$

\[
(\alpha_1, \alpha_2, \ldots) + (\beta_1, \beta_2, \ldots) := (\alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots)
\]
\[
\gamma(\alpha_1, \alpha_2, \ldots) := (\gamma \alpha_1, \gamma \alpha_2, \ldots) \quad \forall \gamma \in \mathbb{R}
\]

3. Function space. $\{f : \mathcal{X} \to \mathbb{R}\}$

\[
(f + g)(x) := f(x) + g(x)
\]
\[
(\gamma f)(x) := \gamma f(x) \quad \forall \gamma \in \mathbb{R}
\]

4. Polynomial space. $\mathbf{P}_d := \left\{ \sum_{i=0}^{d} \alpha_i x^i : \alpha_i \in \mathbb{R} \right\}$ ($\mathbf{P}_\infty$ denotes all polynomials)
Linear Combination, Span, Independence

- **Linear combination** of \( u_1 \ldots u_n \in V \) with coefficients \( \alpha_1 \ldots \alpha_n \in F \) is the vector

\[
\sum_{i=1}^{n} \alpha_i u_i := \alpha_1 u_1 + \cdots + \alpha_n u_n \in V
\]

- **Span** of \( A \subseteq V \) is the set of all (finite) linear combinations

\[
\text{span} (A) = \left\{ \sum_{i=1}^{n} \alpha_i u_i : u_1 \ldots u_n \in A, \; \alpha_1 \ldots \alpha_n \in F, \; n \in \mathbb{N} \right\}
\]

- \( u_1 \ldots u_n \in V \) are **linearly independent** if

\[
\sum_{i=1}^{n} \alpha_i u_i = 0 \implies \alpha_1 = \cdots = \alpha_n = 0
\]
Subspace of $V$ is a subset $S \subseteq V$ closed under linear combinations.

- A subspace is a vector space itself.
- $V$ and $\{0\}$ are trivial subspaces of $V$.
- Intersection of subspaces is a subspace (what about union?).
- Any nonempty $A \subseteq V$ generates the subspace $\text{span}(A)$. 
“Square-Integrable” Subspaces

- Subspace of $\mathbb{R}^\infty$

$$l^2 := \left\{ (\alpha_1, \alpha_2, \ldots) \in \mathbb{R}^\infty : \sum_{i \in \mathbb{N}} |\alpha_i|^2 < \infty \right\}$$

- Subspace of $\{ f : \mathbb{R} \to \mathbb{R} \}$, with weight function $w : \mathbb{R} \to [0, \infty)$

$$L^2_w([a, b]) := \left\{ f : \mathbb{R} \to \mathbb{R} : \int_a^b |f(x)|^2 w(x) dx < \infty \right\}$$

Denote the unweighted version by $L^2([a, b])$. 
Vector Space of Random Variables

- A random variable \( X \) (real-valued) is just a measurable function from sample space \( \Omega \) to real values.

- Thus the set of all real valued random variables is a vector space (i.e., a subspace of function space).

- We can similarly define the subspace of square-integrable random variables

\[
\text{RV}^2 := \{ X : X \text{ is a random variable such that } \mathbb{E}[X^2] < \infty \}
\]
A **basis** of $V$ is $B \subset V$ such that

- The elements of any finite subset of $B$ are linearly independent, and
- $V = \text{span} (B)$

Equivalently, $B \subset V$ is a basis iff every $u \in V$ can be written as a **finite** and **unique** linear combination of elements in $B$.

**Examples:**

- $\{e_1, e_2\}$ is a basis of $\mathbb{R}^2$. So is $\{(1, 1), (1, 2)\}$.
- $\{1, x, x^2, \ldots\}$ is a basis of $\mathbb{P}_\infty$.
- Is $\{e_1, e_2, \ldots\}$ a basis of $\mathbb{R}^\infty$?
Two Facts Regarding Basis

**Existence.** Every vector space has a basis.

- Try to find a basis for $\mathbb{R}^\infty$ by starting with $B = \{e_1, e_2, \ldots\}$.
- $(1, 1, 1, \ldots) \in \mathbb{R}^\infty$ is not in $\text{span}(B)$, so add it.
- $(1, 2, 3, \ldots) \in \mathbb{R}^\infty$ is not in $\text{span}(B)$, so add it.
- ...
- We will ultimately find a basis given the axiom of choice.

**Dimension.** Every basis of a vector space has the same cardinality.

- $\dim(V)$, the “dimension of vector space $V$”, refers to the (unique) cardinality of a basis of $V$.

\[
\dim(\mathbb{R}^d) = d \quad \dim(\mathbb{P}_\infty) = \aleph_0 \quad \dim(\mathbb{R}^\infty) > \aleph_0
\]
Overview

Topics

Review on Vector Space

Vector Space
Inner Product Space
Inner Product Space

**Inner product space** is vector space $V$ over $\mathbb{R}$ (for now) equipped with $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying

\[
\langle u, u \rangle \geq 0 \\
\langle \alpha u, v \rangle = \alpha \langle u, v \rangle \\
\langle u, v \rangle = \langle v, u \rangle \\
\langle u, u \rangle = 0 \Leftrightarrow u = 0 \\
\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle 
\]

▶ Notion of magnitude $\| \cdot \| : V \to [0, \infty)$ given by

$$
\|u\| := \sqrt{\langle u, u \rangle}
$$

Check that $\|\alpha u\| = |\alpha| \|u\|$ and $\|u\| = 0$ iff $u = 0$.

▶ Notion of distance given by $\|u - v\| = \|v - u\|$
Cauchy-Schwarz Inequality

\[ |\langle u, v \rangle| \leq ||u|| ||v|| \]

**Proof.** True for \( v = 0 \). For any \( v \neq 0 \),

\[
||u - \lambda v||^2 = \langle u - \lambda v, u - \lambda v \rangle
\]

\[
= ||u||^2 - 2\lambda \langle u, v \rangle + \lambda^2 ||v||^2
\]

\[
= ||u||^2 - \frac{\langle u, v \rangle}{||v||^2} \geq 0
\]

by choosing \( \lambda = \langle u, v \rangle / ||v||^2 \).
Triangle Inequality

\[ \|u + v\| \leq \|u\| + \|v\| \]

Proof.

\[
\|u + v\|^2 = \|u\|^2 + 2 \langle u, v \rangle + \|v\|^2 \\
\leq \|u\|^2 + 2 |\langle u, v \rangle| + \|v\|^2 \\
\leq \|u\|^2 + 2 \|u\| \|v\| + \|v\|^2 \\
= (\|u\| + \|v\|)^2
\]

▶ Thus \(\|u\|\) is a norm and \((V, \|u\|)\) a normed space.

▶ Thus \(\|u - v\|\) is a metric and \((V, \|u - v\|)\) a metric space.
Continuity of Inner Product

Fact. A linear function between normed spaces is continuous iff bounded.

\[ \langle u, \cdot \rangle : V \to \mathbb{R} \] is a linear function, and for any \( v \in V \),

\[ \langle u, v \rangle \leq ||u|| ||v|| < \infty \]

Thus \( \langle u, \cdot \rangle \) (or \( \langle \cdot, u \rangle \)) is continuous.

In particular,

\[ \lim_{n \to \infty} \langle u_n, u \rangle = \lim_{n \to \infty} \langle u_n, u \rangle \]

\[ \left\| \lim_{n \to \infty} u_n \right\|^2 = \left\langle \lim_{n \to \infty} u_n, \lim_{m \to \infty} u_m \right\rangle = \lim_{n \to \infty} \langle u_n, u_n \rangle = \lim_{n \to \infty} ||u_n||^2 \]
Inner Product Examples

- Inner product on Euclidean space $\mathbb{R}^d$ (dot product)

$$\langle u, v \rangle = u \cdot v := \sum_{i=1}^{d} u_i v_i$$

- Inner product on square-summable sequences $l^2$

$$\langle u, v \rangle := \sum_{i=1}^{\infty} u_i v_i$$

- Inner product on square-integrable functions $L^2_w([a, b])$

$$\langle f, g \rangle := \int_{a}^{b} f(x) g(x) w(x) dx$$

- Inner product on square-integrable random variables $RV^2$

$$\langle X, Y \rangle := E[XY]$$
For nonzero $u, v \in V$, we define

$$
cos(\theta) := \frac{\langle u, v \rangle}{||u|| ||v||} \in [-1, 1]
$$

- If $u = \alpha v$ for some $\alpha > 0$,
  $$
cos(\theta) = 1 \quad \implies \quad \theta = 0
$$

- If $\langle u, v \rangle = 0$ (i.e., orthogonal, also written $u \perp v$),
  $$
cos(\theta) = 0 \quad \implies \quad \theta = \frac{\pi}{2}
$$

- If $u = \alpha v$ for some $\alpha < 0$,
  $$
cos(\theta) = -1 \quad \implies \quad \theta = \pi$$
Orthogonal Projection

- The **orthogonal complement** of a subspace $S \subseteq V$ is the subspace
  \[ S^\perp := \{ u \in V : \langle u, v \rangle = 0 \ \forall v \in S \} \]

  The **(orthogonal) projection** of nonzero $u \in V$ onto $S$ is $u_S \in S$ such that $u_{S^\perp} := u - u_S \in S^\perp$.

- **Claim 1.** $u_S$ is *unique*, hence the unique decomposition (wrt $S$)
  \[ u = u_S + u_{S^\perp} \]

- **Claim 2.** If $S$ has an *orthonormal* (countable) basis $B$,
  \[ u_S = \sum_{v \in B} \langle v, u \rangle v \]

- **Claim 3.** $u_S \in S$ is the best approximation of $u$ under $\| \cdot \|$.
  \[ u_S = \arg \min_{v \in S} \| u - v \| \]
Aside: An Example Usage in ML

Estimating parameter $\theta \in \mathbb{R}^d$ on data points $x_1 \ldots x_N \in \mathbb{R}^d$ by

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^d} \|\theta\|^2 + \text{Loss}(\langle \theta, x_1 \rangle, \ldots, \langle \theta, x_N \rangle)$$

(e.g., binary support vector machines)

**The Representer Theorem.** The optimal parameter must be a linear combination of the data points,

$$\theta^* = \sum_{i=1}^{N} \alpha_i x_i$$
Gram-Schmidt Process

**Input:** linearly independent \( u_1 \ldots u_n \in V \)

**Output:** \( \tilde{u}_1 \ldots \tilde{u}_n \in V \) such that

\[
\langle \tilde{u}_i, \tilde{u}_j \rangle = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{otherwise}
\end{cases} \quad \forall i, j
\]

\[
\text{span} \left( \{ \tilde{u}_1 \ldots \tilde{u}_i \} \right) = \text{span} \left( \{ u_1 \ldots u_i \} \right) \quad \forall i
\]

**Algorithm:** For \( i = 1 \ldots n \),

\[
\tilde{u}_i \gets u_i - \sum_{j=1}^{i-1} \langle u_i, \tilde{u}_j \rangle \tilde{u}_j
\]

\[
\tilde{u}_i \leftarrow \frac{\tilde{u}_i}{||\tilde{u}_i||}
\]

**Implication:** Any linearly independent set of vectors \( A \subseteq V \) can be made into an orthonormal basis of \( \text{span} \left( A \right) \).
Gram-Schmidt Process: (Countably) Infinite Dimension

**Input:** linearly independent \( u_1, u_2, \ldots \in V \) in \((V, \langle \cdot , \cdot \rangle)\)

**Output:** \( \bar{u}_1, \bar{u}_2, \ldots \in V \) such that

\[
\langle \bar{u}_i, \bar{u}_j \rangle = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{otherwise}
\end{cases} \quad \forall i, j
\]

\[
\text{span} (\{\bar{u}_1 \ldots \bar{u}_i\}) = \text{span} (\{u_1 \ldots u_i\}) \quad \forall i
\]

**Algorithm:** For \( i = 1, 2, \ldots \)

\[
\tilde{u}_i \leftarrow u_i - \sum_{j=1}^{i-1} \langle u_i, \bar{u}_j \rangle \bar{u}_j
\]

\[
\bar{u}_i \leftarrow \frac{\tilde{u}_i}{||\tilde{u}_i||}
\]

**Implication:** Any inner product space with countable dimension has an orthonormal basis.
Example: Legendre Polynomials

Orthonormalize the following basis of $P_\infty$

\[
\begin{align*}
p_0(x) &= 1 \\
p_1(x) &= x \\
p_2(x) &= x^2 \\
&\quad \vdots
\end{align*}
\]

with inner product

\[
\langle p, q \rangle = \int_{-1}^{1} p(x)q(x)dx
\]

to obtain an orthonormal basis of $P_\infty$ called the (normalized) Legendre polynomials.
Example: Legendre Polynomials (Cont.)
Versions of Pythagorean Theorem

- For orthogonal $u_1 \ldots u_n \in V$,
  $$\left\| \sum_{i=1}^{n} u_i \right\|^2 = \sum_{i=1}^{n} \|u_i\|^2$$

- If $B$ is an orthonormal basis of subspace $S$, then for any $u \in S$
  $$\|u\|^2 = \sum_{v \in B} |\langle u, v \rangle|^2$$

- If $u_S \in S$ is the orthogonal projection of $u \in V$ onto subspace $S$,
  $$\|u - u_S\|^2 = \|u\|^2 - \|u_S\|^2$$
Parting Remarks on Orthonormal Basis

Because of algebraic convenience and Gram-Schmidt, we always assume that a basis is orthonormal when the dimension is finite (e.g., $\mathbb{R}^d$) or countably infinite (e.g., $\mathcal{P}_\infty$).

When the dimension is uncountably infinite, that is we cannot express a vector as a finite linear combination (e.g., $l^2$), there may not be an orthonormal basis.

Solution: we will change the definition of an orthonormal basis.