=1
Example

August 29, 2019

This document serves as a compendium of my notes at Stanford. Mostly for personal reference, still under construction.

These notes were taken live in lecture, so they definitely contain typos. All errors (typographical or conceptual) are my own.

Contents

1

Independent + CS229: Statistical Learning

1.1 Matrix cookbook

This doesn't matter as much for analytic calculations, but is useful for implementing autodiff on tensors from scratch.

http://www.math.uwaterloo.ca/~hwolkowi//matrixcookbook.pdf

1.2 Large-scale distributed training

Goyal et al. 2017 [?]

- Key contribution: no loss in accuracy when training with large minibatch sizes up to 8192 images.
- Linear scaling rule for adjusting learning rates as a function of mini-batch size. Concretely, when minibatch size is multipled by k, multiply the learning rate by k.
- Warmup scheme that overcomes optimization challenges early in training. Specifically, use a low constant learning rate for the first few epochs of training. For a large minibatch of size kn, train with the low learning rate of η for the first 5 epochs and then return to the target learning rate of $\hat{\eta} = k\eta$.
- Update rule:
$$\hat{w}_{t+1} = w_t - \hat{\eta}\frac{1}{kn}\sum_{j<k}\sum_{x\in\mathcal{B}_j}\nabla l(x, w_t).$$

1.3 Hyperparameter optimization

1.3.1 Population-based training

Jaderberg et al. 2017 [?]

- Key contribution: asynchronous optimization algorithm which uses a fixed computational budget to jointly optimize a population of models and their hyperparameters

- Schedule of hyperparameters (instead of fixing a set for the whole course of training)
- Sequential optimization: run multiple training runs (potentially with early stopping)
- Parallel random/grid search: train multiple models in parallel with different weight initializations + hyperparameters, with the view that one of the models will be optimized the best.
- Population based training: starts like parallel search, randomly sampling hyperparameters and weight initializations. Each training run asynchronously evaluates its performance periodically. Explores new hyperparameters by modifying the better model's hyperparameters, before training is continued.

1.3.2 ENAS

Pham et al. 2018 [?]

- Key contribution: fast + inexpensive approach for automatic model design.
- Controller (trained with policy gradient) discovers neural network architectures by searching for an optimal subgraph within a large computational graph.
- To train the shared parameters ω of the child models: we fix controller policy $\pi(\mathrm{m};\theta)$ and perform SGD on ω to minimize expected loss ${}_{\mathrm{m}\sim\pi}[\mathcal{L}(\mathrm{m};\omega)]$. Gradient is computed using Monte Carlo estimate
$$\nabla_{\omega\,\mathrm{m}\sim\pi(\mathrm{m};\theta)}[\mathcal{L}(\mathrm{m};\omega)] \approx \frac{1}{M}\sum_{i=1}^{M}\nabla_{\omega}\mathcal{L}(\mathrm{m}_i,\omega),$$
where m_i are sampled from $\pi(\mathrm{m};\theta)$.
- To train controller parameters θ: fix ω and update θ to maximize the expected reward ${}_{\mathrm{m}\sim\pi(\mathrm{m};\theta)}[\mathcal{R}(\mathrm{m},\omega)]$. Use Adam optimizer, and compute the gradient with REINFORCE, with a moving average baseline to reduce variance.

1.4 Distillation

Hinton et al. 2015 [?]

- Ensembling: train many different models on the same data and then average their prediction
- Distillation: compress the knowledge in an ensemble into a single model which is much easier to deploy.

Anil et al. 2018 [?]

- Online distillation enables extra parallelism.
- Codistillation algorithm:

1.5 ConvNet architectures

1.5.1 Recognition

- PNASNet-5-Large
  - Similar to NAS, but performs search progressively (starting with models of low complexity).
- NASNet-A-Large
  - Uses a 50-step RNN as a controller to generate cell specifications.
- SENet154
- PolyNet

1.5.2 Detection

- Faster RCNN
- YOLO
- RetinaNet

1.5.3 Segmentation

- FCNet
- DeepLabv4
- Dilated convolutions

1.5.4 WaveNet

Uses dilated convolutions:

- Let F be a discrete function, and k be a discrete filter. The discrete convolution operator $*$ is defined as

  $$(F * k)(\mathrm{p}) = \sum_{\mathrm{s}+\mathrm{t}=\mathrm{p}} F(\mathrm{s})k(\mathrm{t}).$$

  More generally, let l be a dilation factor. The l-dilated convolution can be defined as

  $$(F *_l k)(\mathrm{p}) = \sum_{\mathrm{s}+l\mathrm{t}=\mathrm{p}} F(\mathrm{s})k(\mathrm{t}).$$

- Implemented in TensorFlow as `tf.nn.atrous_conv2d`

1.5.5 Self-attention networks

1.6 PyTorch

1.7 Backpropagation: CS231 intuitions

1.8 Backpropagation: a graph theory perspective

Notes from Chris Olah's post: `http://colah.github.io/posts/2015-08-Backprop/`

Backpropagation is a very common algorithm, and is often referred to as "reverse-mode differentation." At its core, it is a tool for calculating derivatives quickly.

Why are computational graphs a good abstraction? To apply the multivariate chain rule:

1. Sum over all possible paths from one node to the other.
2. Multiply the derivatives on each edge of the path together.

1.8.1 Combinatorial explosion

This is all just standard chain rule. But how do you deal with cases like this?[1]

There are 9 paths in the above diagram. Instead of naively summing over the paths, we can factor them:

$$\frac{\partial Z}{\partial X} = (\alpha + \beta + \gamma)(\delta + \varepsilon + \zeta).$$

There are two algorithms we can leverage here.

1. Forward-mode differentiation. Start at an input to the graph, and move towards the end. Sum all the paths feeding in. The operator here is $\frac{\partial}{\partial X}$; similar to standard calculus.

$$\begin{aligned} \frac{\partial X}{\partial X} &= 1 \\ \frac{\partial Y}{\partial X} &= \alpha + \beta + \gamma \\ \frac{\partial Z}{\partial X} &= (\alpha + \beta + \gamma)(\delta + \varepsilon + \zeta). \end{aligned}$$

2. Reverse-mode differentiation. Start at an output of the graph, and move towards the beginning. At each node, merge all paths which started at that node. The operator here is $\frac{\partial Z}{\partial}$.

   In particular:

$$\begin{aligned} \frac{\partial Z}{\partial Z} &= 1 \\ \frac{\partial Z}{\partial Y} &= \delta + \varepsilon + \zeta \\ \frac{\partial Z}{\partial X} &= (\alpha + \beta + \gamma)(\delta + \varepsilon + \zeta) \end{aligned}$$

[1] Image source: Chris Olah.

What is the difference between forward and reverse mode differentiation? "Forward-mode differentiation tracks how one input affects every node." "Reverse-mode differentiation tracks how every node affects one output."

$$\begin{aligned}\texttt{forward mode: } & \frac{\partial}{\partial X} \\ \texttt{reverse mode: } & \frac{\partial Z}{\partial}\end{aligned}$$

Reverse mode differtiation gives us the derivative of the output w.r.t. every node. This is exactly what we want.

On a large computational graph, this means reverse mode differentiation can get them all in one fell swoop.

In summary: derivatives are ridiculously computationally cheap.

1.9 Functional programming ∩ Neural networks

TODO: write up notes on Chris' article.

2

CS236: Deep Generative Models

2.1 Variational Autoencoder

- Observations: $\mathrm{x} \in \{0, 1\}^d$.
- Latent variables $\mathrm{z} \in \mathbb{R}^k$.
- Goal: learn a latent variable model that satisfies

$$\begin{aligned} p_\theta() &= \int p_\theta(,)\, d \\ &= \int p()p_\theta(|)\, d. \end{aligned}$$

In particular, the VAE is defined by the following generative process:

$$\begin{aligned} p() &= (|0, I) \\ p(|) &= (|f_\theta()), \end{aligned}$$

where $f_\theta()$ is a neural network decoder to obtain the parameters of the d Bernoulli random variables which model the pixels in each image.

For inference, we want good values of the latent variables given observed data (that is, $p(|)$.

Indeed, by Bayes' theorem, we can write

$$\begin{aligned} p(|) &= \frac{p(|)p()}{p()} \\ &= \frac{p(|)p()}{\int p(|)p()\, dz}. \end{aligned}$$

We want to maximize the marginal likelihood $p_\theta()$, but the integral over all possible is intractable. Therefore, we use a variational approximation to the true posterior.

We write

$$q_\phi(|) = (|\mu_\phi(), (\sigma^2_\phi())).$$

Variational inference approximates the posterior with a family of distributions $q_\phi(|)$.

To measure how well our variational posterior $q(|)$ approximates the true posterior $p(|)$, we can use the KL-divergence.

The optimal approximate posterior is

$$\begin{aligned} q_\phi(|) &=_\phi KL(q_\phi(|)||p(|)) \\ &=_\phi \left\{ {}_q \left[\log q_\phi(|)\right] -{}_q \left[\log p(,)\right] + \log p() \right\}. \end{aligned}$$

But this is impossible to compute directly, since we end up getting $p()$ in the divergence.

We then maximize the lower bound to the marginal log-likelihood:

$$\begin{aligned} \log p_\theta() &\geq \text{ELBO}(;\theta,\phi) \\ &=_{q_\phi(|)} \left[\log p_\theta(|)\right] - D_{KL}(q_\phi(|)||p()) \end{aligned}$$

And this ELBO is tractable, so we can optimize it.

2.1.1 Reparametrization trick

Instead of sampling

$$z \sim (\mu, \Sigma),$$

we can sample

$$\begin{aligned} z &= \mu + L; \\ &\sim (0, I); \Sigma = LL^T \end{aligned}$$

Allows for low variance estimates.

2.1.2 GMVAE

Same set up as vanilla VAE, except the prior is a mixture of Gaussians. That is,

$$p_\theta() = \sum_{i=1}^{k} \frac{1}{k}(|\mu_i, (\sigma_i^2))$$

However, the KL term cannot be computed analytically between a Gaussian and a mixture of Gaussians. We can obtain an unbiased estimator, however:

$$\begin{aligned} D_{KL}(q_\phi(|)||p_\theta()) &\approx \log q_\phi(^{(1)}|) - \log p_\theta(^{(1)}) \\ &= \log(^{(1)}|\mu_\phi(), (\sigma^2_\phi())) - \log \sum_{i=1}^{k} \frac{1}{k}(^{(1)}|\mu_i, (\sigma_i^2)). \end{aligned}$$

2.1.3 IWVAE

The ELBO bound may be loose if $q_\phi(|)$ is a poor approximation to $p_\theta(|)$. For a fixed , the ELBO is, in expectation, the log of the unnormalized density ratio

$$\frac{p_\theta(,)}{q_\phi(|)} = \frac{p_\theta(|)}{q_\phi(|)} p_\theta(),$$

where $\sim q_\phi(|)$.

1. Prove that IWAE is a valid lower bound of the log-likelihood.

   $$\begin{aligned} \log p_\theta() &\geq_{^{(1)},\dots,^{(m)}\sim q_\phi(|)} \left(\log \frac{1}{m}\sum_{i=1}^{m} \frac{p_\theta(,^{(i)}}{q_\phi(^{(i)}|)}\right) \\ &\geq_{z^{(1)}\sim q_\phi(|)} \log \frac{p_\theta(,^{(1)})}{q_\phi(^{(1)}|)} \end{aligned}$$

   Jensen states that for convex functions, $f[X] \geq f[X]$. $\log$ is concave. So

2.1.4 Questions

- Why is the reparametrization trick lower variance? (Asked on Piazza.)

3

CS168: The Modern Algorithmic Toolbox

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine

[parfill]parskip [margin=1in]geometry

Theorem Example Definition Remark Claim

latexsym bbm [small,bf]caption2 graphics epsfig amsopn url

sign GL Var

[parfill]parskips [margin=1in]geometry

CS168: Modern Algorithmic Toolbox Instructor: Greg Valiant; Notes: Adithya Ganesh

Contents

3.1 Lecture 3

Administrative updates:

- Mini project 1: due 11:59pm tomorrow
- Mini project 2: posted tonight (due in 8 days)

Core problem. How can we quickly find similar datapoints?

Two variations on this problem:

- One: given a dataset, search for similar pairs within the dataset.
- Two: given a new datapoint, quickly process the query and find points that are similar (nearest neighbor search problem).

Question. How do we define "similarity?"

Motivation / applications.

- Similarity for e.g. documents, webpages, source code. Search engines, for instance, perform a lot of deduplication to ensure that results are not repeated twice.
- Collaborative filtering (think Amazon / Netflix for recommendations). Idea: compute which individuals are similar, or compute which movies / items are similar.
- Machine learning via nearest neighbor search / similarity.

3.1.1 Similarity Metrics

Jacard Similarity. This is a notion that applies between sets / multi-sets S and T.

$$J(S,T) = \frac{|S \cap T|}{|S \cup T||}.$$

In the case of multisets - just count things with redundancy.

Example. Say $S = \{a, b, c, d, e\}$, and $T = \{a, e, f, g\}$, then the Jacard similarity is

$$J(S,T) = \frac{2}{7}.$$

Another context in which we can apply Jacard similarity is to consider the one-hot encoding vector S where S_i represents the number of times i appears.

Then

$$J(S,T) = \frac{\sum_i \min(S_i, T_i)}{\sum_i \max(S_i, T_i)}.$$

Tends to work well in practice especially for sparse data (for example, text and documents).

Euclidean distance. (between vectors in $\mathbb{R}^d$)

$$||x-y||_2 = D_{euc}(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}.$$

One reason this is useful in that its rotationally invariant.

L_1 distance / Manhattan distance.

$$||x-y||_1 = \sum_{i=1}^{d} |x_i - y_i|.$$

Intuition - if walking in a grid, this is the distance no matter how you walk. Also - this is not rotationally invariant. Therefore, if you are ever using L_1 distance, make sure that the basis means something.

More broadly, we can define L_p metrics.

L_p metrics. The L_p distance is defined as

$$||x-y||_p = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p}$$

Note that there are many other notions of similarity.

- Cosine similarity (the angle between two vectors)
- Edit distance (Hamming)

Note that as $p \longrightarrow \infty$, this converges to the max over i over the absolute difference in the components.

Note that there is a subtlety in the corner points (depending on how you define the limit).

Picture:

And note that you can define L_p metrics for p non-integer.

How do you decide between L_1 and L_2? You can reason about this by thinking about the Voronoi diagram of the vectors.

Definition. Given points X in $\mathbb{R}^d$, and some metric D, the Voronoi diagram partitions space into regions (the set of all points that are closest to a single datapoint). Concretely, for $x \in X$, $part(X) = \left\{y \in \mathbb{R}^d \text{ s. t. } D(x,y) = \min_{x' \in X} D(x', y)\right\}$.

Story: John Snow (a doctor) figured out that the people who died of cholera were in a Voronoi cell corresponding to an infected well.

You can think about the Voronoi diagram for different similarity metrics. For L_2, they are going to be straight lines. Question: what do the partitions of the Voronoi diagrm look like for L_2 and L?

Area of math devoted to understanding the difference between different metrics: metric embeddings.

Natural question. How do you map one set of points in one metric to another set of points in a different metric, such that the original distances are equal to the new distances?

Concretely - given $x_1, \ldots, x_n \in \mathbb{R}^d$, can we find a function $f : \mathbb{R}^d \to \mathbb{R}^m$ such that $||x_i - x_j||_1 \approx ||f(x_i) - f(x_j)||_2$?

In many cases, the answer is yes, there exists a function that can do this.

For the rest of the class: let's return to the question of how we find similar objects. We will focus on Euclidean distance.

In two dimensions, Voronoi diagrams are straightforward to construct. Things get much harder in higher dimensions.

At a high level, the number of edges that a cell in the Voronoi diagram will have will scale exponentially with the number of dimensions you are in. Even storing the Voronoi diagram in memory will take, exponential space.

Example. $k - d$ trees (space partitioning data structure). Idea: Use a balanced binary tree that partitions space.

The idea is that each edge in our tree will correspond to a partition in space.

(TODO: Insert image).

How much space does it take to store this data structure? At each node, we only need to store the value of each point.

To find closest point to some new y. There are two steps:

- Go down the tree and figure out which region of space y would fall in $(\log N)$ operations.
- Go bac up, check each case.

Question: do we need to jump to other regions? In 1 dimension, we can just return the datapoint that is in that partition (guaranteed to be the closest point to y).

In higher dimensions - we might need to jump to adjacent regions. Going up the tree, we ask - is it possible that there is a point in the next partition that is closest to us than the next?

Rule of thumb. If dimension $d < 20$, works fairly well. This refers to the intrinsic dimensionality of the dataset (for example if the dataset is higher dimensional, but lies on a lower dimensional subspace), or if the number of points $> 2^d$.

k-d tree data structure.

Given a set of points - pick a dimension.

Given a set $S = \left\{x : x \in \mathbb{R}^d\right\}$.

- Pick a dimension / coordinate i.
- Compute the median of $\{x_i\}$.
- Partition: $S_1 = \{x \in S | x_i < m\}$, and $S_2 = \{x_i \geq m\}$.
- Recurse on S_1 and S_2 and store which dimension we are looking at.

How much back and forth do we need to do in the tree? The number of points that we'll need to check is going to be exponential in the dimension. This i sbecause the number of facets in a Voronoi diagram will tend to scale exponential in the dimension.

Runtime:

Logarithmic in the number of points

And exponential in the dimension of the points.

Does it make sense to sort a subset of the data? Yes, but then you have the question of sorting on which dimension.

3.2 Lecture 4

Review. Last time we talked about $k - d$ trees, which are binary trees that partition space in k dimensions.

The runtime to find a closest point to a new point y is:

$$\log(\text{\# points}) \cdot \underbrace{\exp(\text{dim})}_{\text{number of partitions to look through}}$$

This is one example (broadly) of a phenomenon known as the "curse of dimensionality." For many geometric problems that we care about - the runtime will scale exponentially in the dimension that we're working with.

The kissing number. How many identical spheres can you place around a sphere such that all of them are touching the center sphere?

For $k = 2$, the kissing number is 6. For $k = 3$, the kissing number is 12. Note that in 5 dimensions, the kissing number is unknown! In general, the kissing number will scale exponentially in d.

Question. Why are proving these results so hard to prove? For certain dimensions, (for example dim = 8, it is fairly easy to show results (because of symmetry). But for other, the best packing strategies we know is based on random packing processes.

Note that sphere packing is a very relevant question to ask. You can think of radio stations and wanting to pack them as close together as possible without interference.

Problem: reduce the dimensionality while approximately preserving all pairwise distance.

Suppose you have $x_1, \ldots, x_n \in \mathbb{R}^d$. Suppose you have $y \in \mathbb{R}^d$, and you want to find the closest point in X. What are our options?

- Use $k - d$ tree, and pay $\log(\# \text{ points}) \cdot \exp(\text{dim})$.
- Brute force: $O(nd)$
- Dimensionality reduction + brute force (developed in this lecture).

We will describe a general recipe to perform dimensionality reduction, and this will apply for any similarity metric.

- Find easy / fast way to preserve distances in expectation.
- Repeat it a few times (independent repetitions; this will take you from being good in expectation to being good most of the time).

3.2.1 Dimensionality reduction for Jacard similarity

Consider Jacard similarity, defined earlier. We develop the "MinHash" technique that we can use to reduce dimesionality.

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

- Pick a uniformly random ordering of the universe U.
- Map set S to $f(S) =$ "min" element of S.

This is based on the following claim.

Claim. For any sets S and T, we have

$$\Pr(f(S) = f(T)) = J(S,T).$$

Proof. $f(S) = f(T)$ if and only if the first element is in the intersection. The denominator is the union. The result is clear from here. □

Concretely, suppose we repeat k times. This requires us to pick k random orderings. And then we can ask what fraction of k will satsify $f(T) = f(S)$?

What is the error relative to the true similarity? It is approximately $\frac{1}{\sqrt{k}}$.

Suppose we have independent random variables $X_1, \ldots, X_k \sim \text{Ber}(p)$. We want to know what is the standard deviation of their average?

$$\left(\frac{\sum_{i=1}^k X_i}{k}\right).$$

$$= \frac{1}{k^2}\left(\sum_{i=1}^k X_i\right)$$

$$= \frac{1}{k^2} \cdot k(X_1) \geq \frac{1}{k}$$

Hence the standard deviation is at most $\frac{1}{\sqrt{k}}$. Note: this is true in general; this is how you interpret the statistical significance of election polls.

3.2.2 Dimensionality reduction for Euclidean distance

- Choose a random d dimensional vector $r = (r_1, \ldots, r_d)$.

$$f_r(v) =< v, r >= \sum_{i=1}^{d} v_i r_i.$$

It turns out that if you pick two angles on the sphere, it doesn't end up being uniform! Similarity, if you uniformly pick the coordinates, the resulting distribution is not uniform. Instead, we will pick $r_i \sim \mathcal{N}(0, 1)$, resulting in a uniform distribution that is rotationally invariant.

Fact. Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then:

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Note that this is fairly unique to Gaussians - this isn't true for most distributions.

Claim. For any two vectors $x, y \in \mathbb{R}^d$, we have

$$\mathbb{E}[(f_r(x) - f_r(y))^2] = \mathbb{E}[||x - y||_2^2]$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}[(f_r(x) - f_r(y))^2] &= \mathbb{E}\left\{\left(\sum_{i=1}^{d} r_i x_i - \sum_{i=1}^{d} r_i y_i\right)^2\right\} \\
&= \mathbb{E}\left\{\left(\sum_{i=1}^{d} r_i (x_i - y_i)\right)^2\right\} \\
&= \mathbb{E}\left\{\mathcal{N}(0, \sum_i (x_i - y_i)^2)\right\} \\
&= \left\{\mathcal{N}(0, \sum_i (x_i - y_i)^2)\right\} \\
&= \sum_{i=1}^{n} (x_i - y_i)^2.
\end{aligned}$$

□

Fact. If you repeat $l = \frac{\log n}{\epsilon^2}$ times, then with high probability, all $\binom{n}{2}$ pairwise distances are preserved to within a factor of $1 \pm \epsilon$.

This transformation is referred to as the Johnson-Lindenstrauss transformation.

3.3 Lecture 5: Generalization

General question. How much data is enough?

Broadly, we can think about two different types of data analysis.

- Understanding dataset
- Goal of extrapolating beyond dataset (inference)

Consider the binary classification setting. We can broadly define it as follows:

- Suppose we have some datapoints $x_1, \ldots, x_n \in \mathbb{R}^d$.
- Known distribution D on $\mathbb{R}^d$.
- Ground truth label function $f : \mathbb{R}^d \to \{0, 1\}$.

Problem. Given $x_1, \ldots, x_n$ drawn independently from D, and labels $f(x_1), \ldots, f(x_n)$, our goal is to output $g : \mathbb{R}^d \to \{0, 1\}$ such that "$g \approx f$".

Namely, we want the generalization error to be low, defined this way:

$$\text{generalization error}(g) = \Pr_{x \sim D} [g(x) \neq f(x)] .$$

Can also define the training error as the fraction of the training points on which g disagrees with the true labelling.

Claim. For any function g, the expected training error is equal to the generalization error. [1]

Question. Suppose we find g with training error 0. When does this imply that the generalization error is small?

Factors that influence this question:

- Amount of data (how faithful is the sample?)
- The complexity of the function (# of functions considered).
- Algorithm used to find g

This is often succinctly phrased as "does g generalize?" If answer is no, this implies that you've "overfit" the data.

First, we'll consider the "well-separated finite setting." Here, we will make two enormous assumptions. Assume that:

- The ground truth labelling function $f \in S = \{f_1, f_2, \ldots f_k\}$. That is, f belongs to a set of functions with k elements which is finite.

[1] To be clear: "training error" in this setting refers to datapoints that the functino has not necessarily been trained on. It might be clearer to say: the expected "empirical error" converges to the generalization error.

- All of these functions are well separated. No function in the class is similar to f. For all $f_i \in S$ with $f_i \neq f$, the generalization error of $f_i > \epsilon$. Note that this is sort of a silly assumption, but we will drop both.

Naive “algorithm:” return any g in our set S that has training error 0.

Theorem. Given assumptions 1 and 2, if the number of datapoints $n > \frac{1}{\epsilon}(\log k + \log \frac{1}{\delta})$, then, with probability at least $1 - \delta$, g will generalize.

Some comments: logarithmic function in k and $1/\delta$ is good, but inverse linear function in $\frac{1}{\epsilon}$ is kind of bad.

Proof. We will prove this in two parts.

- First, we will analyze the probability that we are “tricked” by a bad f_i.
- Next, we will union bound over all bad f_i’s.

Consider a bad function f_i. The probability that we are tricked by this function is

$$\begin{aligned}
\Pr\left\{\text{TrainingError}(f_i) = 0\right\} &= \prod_{j=1}^{n} \Pr_{x_j \sim D}(f_i(x_j) = f(x_j)) \\
&\leq (1-\epsilon)^n \\
&< e^{-\epsilon n}.
\end{aligned}$$

The last inequality follows from the inequality $1 + x < e^x$. Proof from Taylor series / plot.

There are at most k “bad” functions. Hence we can apply a union bound, to obtain

$$\delta = \Pr(\text{output bad function}) \leq k e^{-\epsilon n}.$$

Now, the desired result directly follows from solving for failure probability $\frac{1}{\delta}$.

Note that we don’t really need assumption 2. □

Results of this form are generally referred to as being in the “PAC” framework (probably appropximately correct).

```
 Consider the set of linear classifiers in $\RR^d$.  This is defined by a vector $\mbf{a} = (a_1, a_2, \dots, a_d)$.  Then the classifier is just
  \[
    f_a(x) = \sign\left (\sum_{i=1}^{d} \mbf{a} \cdot \mbf{x} \right ).
  \]

  {\it Intuition.} Note that the vector $\mbf{a}$ will be the normal vector to the hyperplane separating datapoints.

  Also note that this is very general because if we don't want a hyperplane through the origin, we can just add another feature.
```

Claim. If we consider the set of linear classifiers, then the error still satisfies the generalization bound, with a few minor tweaks.

Theorem. For linear classifiers, if the number of datapoints $n > \frac{C}{\epsilon}(d + \log \frac{1}{\delta})$, then, with probability $1 - \delta$, g will generalize.

The intuition behind the proof here is that there are an exponential number of "important" directions in d dimensional space.

Important questions to consider:

- How do we find the optimal g?
- What if no function in S has error 0?
- What if you have fewer than the threshold of datapoints, what can you do?

3.4 Lecture 6: Regularization

Note on last part - it is very open ended, at the cutting edge of machine learning research (but don't feel obliged to write pages of analysis).

Punchline from last class. If you a have a set k different functions $\{f_1, \dots, f_k\}$, then the "best" one will generalize if $n > O(\log k)$. If we are classifying in d dimensions - we can approximate by set of $\exp(d)$ linear functions. If $n > d$, expect generalization.

Regularization. A way to express a set of preferences over models. Such a scheme that will take both performance on training data as well as these preferences into account.

```
 For example, we can consider $L_2$-regularized least squares.  Let $x_1, \dots, x_n \in
ting, we want to minimize the following objective:
  \[
      \min_{a}  f(a)  =  \sum_{i=1}^{n}  \left(  \langle  x_i,  a  \rangle  -
y_i \right)^2 + \underbrace{\lambda ||a||_2^{2}}_{\text{regularization term}}.
  \]
```

There are two broad types of regularization, explicit or implicit:

- Explicit regularization (e.g. L_2 regularization, preferring sparse vectors).
- Implicit regularization (algorithm itself has "preferences").

K̂ey question. Why should we regularize; why wouldn't we just return the empirical risk minimizer?

Perspective. You always want roughly $n \approx d$, where generalization might not hold. If you have $n = 1,000,000$, then you want $d \approx 1,000,000$. Otherwise - it's sort of a "waste"; if d is truly 1000 in your dataset, try to construct additional features to learn a model in 1,000,000 dimensional space.

How should we construct additional features? Here are two approaches.

- Polynomial embedding. (For example - quadratic embedding.

  $$x = (x_1, x_2, \dots x_d) \to f(x) = (x_1, \dots, x_d, x_1^2, x_1x_2, x_1x_3, \dots, x_d^2, 1) \in \mathbb{R}^{2d+1+\binom{d}{2}}.$$

  One simple setting in which you need quadratic features to fit a classifier is when you are fitting a circular decision boundary.

- Random projection + non-linearity. You really need the non-linearity, since otherwise you'll just be learning another linear function. Can choose $\sqrt{x}, x^2, \sigma(x)$, or most other "nice" nonlinearities (since they all roughly have similar properties).

Downsides of adding new features:

- One objection could be that working with $\approx d^2$ dimensional points is annoying. But this isn't an issue: you can "implicitly" work over the embedded points without computing the embedding. The area of math devoted to this is referred to as "kernelization" (usually covered in the context of SVMs).
- Real issue: if you need d^2 features, you generally need much more data.

Rule of thumb: if the coordinates actually have signifiance, the polynomial embedding preserves interpretability.

How should you think about regularization? There are two views: the Bayesian view, and the frequentist view.

- Bayesian view. Assume that the true model is drawn from some known "prior" distribution. This allows us to evaluate the "likelihood" / probability of a given candidate model.
- Frequentist view. Goal: argue that if true model has "nice structure," then can find it.

Bayesian approach to regularization. ("Gaussian prior") Suppose we have $x_1, \dots, x_n \in \mathbb{R}^d$, assume that the true label $a^* \in \mathbb{R}^d$ is drawn by choosing each coordinate independently from $\mathcal{N}(0,1)$.

Now, suppose that each label is set as $y_i = \langle x_i, a^* \rangle + z_i$, where noise $z_i \sim \mathcal{N}(0, \sigma^2)$. Now, given $x_1, \dots, x_n, y_1, \dots, y_n$, ask

$$\text{Likelihood}(a) = \Pr(a) \Pr(\text{data}|a).$$

Because we made strong assumptions on the label distributions, we can directly compute these probabilities. Hence

$$\begin{aligned}\text{Likelihood}(a) &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp(-a_i^2/2) \prod_{i=1}^{n} \exp(-(\langle x_i, a \rangle - y_i)^2 / 2\sigma^2) \\ &\propto \exp(-||a||_2^2/2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\langle x_i, a \rangle - y_i)^2)\end{aligned}$$

Maximum likelihood of a is equivalent to minimizing:

$$\sum_{i=1}^{n} (\langle x_i, a \rangle - y_i)^2 + 2\sigma^2 ||a||_2^2.$$

This derivation even told us how to set the regularization constant. Should be 2 times the variance of the noise.

Frequentist approach to sparsity / regularization. Consider a model a^* that is s-sparse (i.e. there are s nonzero coordinates).

Question: why do we care about sparse models?

- One answer is that lots of models are actually sparse. Think of the laws of physics.
- Other view: maybe the world is sloppy, and the best model might not be sparse. But what's most helpful for interpretability is fitting a sparse model.

Question: can we build a regularizer that lets us selectively find sparse models? The obvious choice is

$$f(a) = \sum_{i=1}^{n} (\langle x_i, a\rangle - y_i)^2 + \lambda ||a||_0,$$

where we are using the "0-norm" that computes sparsity.

Claim. If $n > O(s \log(d))$ then the sparsest model that fits the data is "correct."

The number of s-sparse d-dimensional function is just $\binom{d}{s}$, and there are about $\exp(s)$ sparse functions. So there are approximately $d^s \exp(s)$ sparse functions, and we need datapoints around the logarithm of this.

The problem with using sparse models is that it is not differentiable (so finding the minimizer is NP-hard).

So, we can note the following:

- l_0 regularization is great, but computationally intractable.
- Idea: use l_1 regularization as proxy for l_0.

Miraculously, the claim from before still holds for l_1 regularization (proved in the early 2000's, Candes in the stat / math department here).

3.5 Lecture 10

Definition. A $n_1 \times n_2 \times \cdots \times n_k$ k-tensor is a set of $n_1 n_2 \ldots n_k$ numbers which interprets as being arranged in a k-dimensional hypercube.

A 2-tensor is simply a matrix, with $A_{i,j}$ referring to the i, jth entry. You can refer to a specific element of a k-tensor via $A_{i_1,i_2,\ldots,i_k}$.

Note that tensors are very useful in physics, where they are viewed with more geometric intuition.

We can define a notion of rank of a tensor. Note that a matrix M has rank r if it can be written as $M = UV^T$, where U has r columns, and V has r columns.

We can write

$$M = \sum_{i=1}^{r} u_i v_i^T.$$

Here is an informal definition of tensor rank.

- A tensor is rank 1 if all rows of all matrices are multiples of each other.
- A tensor has rank k if it can be written as a sum of k rank 1 tensors.

We can define a tensor product as follows

Definition. Given vectors $v_1, v_2, \ldots, v_k$ of lengths $n_1, \ldots, n_k$, the tensor product is denoted $v_1 \otimes v_2 \otimes \cdots \otimes v_k$ is the $n_1 \times n_2 \times n_k$ k-tensor A with entry

$$A_{i_1,i_2,\ldots,i_k} = v_1(i_1) \cdot v_2(i_2) \cdots v_k(i_k).$$

```
  For example, let
  \[
    v_1 = \mat{1 \\ 2 \\ 3}, v_2 = \mat{-1 \\ 1}, v_3 = \mat{10 \\ 20}.
  \]
   Then $v_1 \times v_2 \times v_3$ is a $3 \times 2 \times 2$ 3-tensor, that can be thought of as a stack of two $3 \times 2$ matrices:
  \[
    M_1 = \mat{-10 & 10 \\ -20 & 20 \\ -30 & 30}, M_2 = \mat{-20 & 20 \\ -40 & 40 \\ -60 & 60}.
  \]
```

More formally, we can define the rank of a tensor as follows.

Definition. A 3-tensor A has rank r if there exists 3 sets of r vectors, $u_1, \ldots, u_r, v_1, \ldots, v_r$ and $w_1, \ldots, w_r$ such that

$$A = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i.$$

Let's go back to the motivation for SVD (the "Spearman experiment").

Suppose there are 1000 students. Construct a 1000×20 matrix where you administer 20 different school tests. He noticed that this is approximated by a rank 2 matrix. Namely, it is approximated by (1000×2) multiplied by 2×20. Question: to what extent is this decomposition unique?

If we can write

$$M = AB,$$

we can also write

$$M = (AX)(X^{-1}B).$$

Let's examine the differences between matrices and tensors.

- For matrices, the best rank-k approximation can be found by iteratively finding the best rank-1 approximation, and then subtracting it off. If uv^T is the best rank 1 approximation of M, then $rank(M - uv^T) = rank(M) - 1$.

  For k-tensors with $k \geq 3$, this is not always true. If $u \otimes v \otimes w$ is the best rank 1 approximation of 3-tensor A, it is possible that $rank(A - u \otimes v \otimes w) > rank(A)$.

- For matrices with entries in $\mathbb{R}$, there is no point in looking for a low-rank decomposition that involves complex numbers, because of $rank_{\mathbb{R}}(M) = rank_{\mathbb{C}}(M)$. For k-tensors, this is not always the case.
- With probability 1, if you pick the entries of an $n \times n \times n$ 3-tensor independently at random from the interval $[0, 1]$, the rank will be on the order of n^2. But we don't know how to describe any construction of $n \times n \times n$ tensors whose rank is greater than $n^{1.1}$ for all n.
- Computing the rank of matrices is easy (via SVD). Computing the rank of 3-tensors is NP-hard.
- If the rank of a 3-tensor is sufficiently small, then its rank can be efficiently computed, its low rank represnetation is unique, and can be efficiently recovered.

Theorem. (Amazing theorem of tensors) Consider a 3-tensor A which has rank k. It can be writen as

$$A = \sum_{i=1}^{k} u_i \otimes v_i \otimes w_i.$$

Claim: if $\{u_1, \ldots, u_k\}$, $\{v_1, \ldots, v_k\}$ are linearly independent, then can efficiently recover this factorization.

We can now discuss the tensor decomposition algorithm (Jenrich's Algorithm). Given an $n \times n \times n$ tensor $A = \sum_{i=1}^{k} u_i \otimes v_i \otimes w_i$ with $(u_1, \ldots, u_k), (v_1, \ldots, v_k), (w_1, \ldots, w_k)$ linearly independent, the following algorithm will output the lists of u's, v's, and w's.

- Choose random unit vectors $x, y \in \mathbb{R}^n$.
- Define the $n \times n$ matrices A_x, A_y, where A_x is defined as follows. Consider A as a stack of n $n \times n$ matrices. Let A_x be the weighted sum of these matrices, where the weight given to the ith matrix is x_i. Define A_y analogously.
- Compute the eigen-decompositions of $A_x A_y^{-1} = QSQ^{-1}$ and $A_x^{-1} A_y = Y^{-1} T Y^T$.
- With probability 1, the entries of diagonal matrix S will be unique, and will be inverses o the entries of T. The vectors $u_1, \ldots, u_k$ are the columns of Q corresponding to nonzero eigenvalues, and the vectors $v_1, \ldots, v_k$ will be the columns of Y, where v_i corresponds to the reciprocal of the eigenvalue to which u_i corresponds.
- Given the u_i's and the v_i's, we can now solve a linear system to find the w_i's, or imagine rotating the whole tensor A and repeating the algorithm to recover the w's.

Why does this work?

Claim. We have that

$$A_x = \sum_{i=1}^{k} \langle w_i, x \rangle u_i v_i^T; \qquad A_y = \sum_{i=1}^{k} \langle w_i, y \rangle u_i v_i^T.$$

We can see this using an SVD / eigendecomposition argument (see lecture notes).

Quick suggestions on where we might encounter tensors:

- Spearman experiment setting.
- NLP setting, where you have a tri-occurence 3 tensor for e.g. words.

- Social network tensor in terms of groups, not just pairs.
- Moment tensor. If you have n-dimensional data, then the covariance is $n \times n$. You can compute third-order and fourth-order moments fairly naturally.

3.6 Things to review

1. Review proof of simple PAC / generalization bound.
2. SVD intuition.

3.7 Key ideas

4

EE376A and mathematics directed reading: Information Theory

extsizes

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine

Theorem Lemma Definition Remark Claim Example Proposition Solution

latexsym bbm [small,bf]caption2 graphics epsfig amsopn url

[parfill]parskip [margin=1in]geometry

sign Aut GL Ker im Syl

[parfill]parskips [margin=1in]geometry

Information Theory and Statistical Learning Adithya Ganesh

Thank you to Yuval Widgerson from the Stanford Mathematics Department for useful discussions.

Contents

4.1 The Source Coding Theorem

Definition. An ensemble X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$ where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ having probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$, with $P(x = a_i) = p_i, p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Definition. We define the Shannon information content of the outcome $x = a_i$ to be

$$h(x = a_i) \equiv \log_2 \frac{1}{p_i}.$$

Definition. We define the entropy of the ensemble to be

$$H(X) = \sum_i p_i \log_2 \frac{1}{p_i}.$$

Intuition. The outcome of a random experiment is guaranteed to be most informative if the probability distribution over outcomes is uniform.

4.1.1 A basic example

What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

Intuitively, the best questions successively divide the 64 possibilities into equal sized sets. One strategy is to ask the following questions.

- Is $x \geq 32$?
- Is $x \mod 32 \geq 16$?
- Is $x \mod 16 \geq 8$?
- Is $x \mod 8 \geq 4$?
- Is $x \mod 4 \geq 2$?
- Is $x \mod 2 = 1$.

The answers to these questions if encoded in binary, give the expansion of x, for example $35 \implies 100011$. If all values of x are equally likely, then the answers to the questions are independent, and each has Shannon information content $\log_2(1/0.5) = 1$ bit.

The Shannon information content in this setting measures the length of a binary file that encodes x.

Similarly, refer to the submarine game example (pg. 71, MacKay).

Definition. The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|,$$

which is a lower bound for the number of binary questions that are guaranteed to identify an outcome from the ensemble X.

Definition. The smallest δ-sufficient subset S_δ is the smaller subset of $\mathcal{A}_x$ satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

Definition. The essential bit content of X is

$$H_\delta(X) = \log_2 |S_\delta|.$$

Theorem (Shannon's source coding theorem). Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left|\frac{1}{N} H_\delta(X^N) - H\right| < \epsilon.$$

Theorem (Chebyshev's inequality 1). Let t be a non-negative real random variable, and let α be a positive rela number. Then

$$P(t \geq a) \leq \frac{\bar{t}}{\alpha}.$$

Theorem (Chebyshev's inequality 2). Let x be a random variable, and let α be a positive real number. Then

$$P((x - \bar{x})^2 \geq \alpha) \leq \sigma_x^2/\alpha.$$

Theorem (Weak law of large numbers). Take x to be the average of N independent random variables $h_1, \ldots, h_N$, having common mean $\bar{h}$ and common variance σ_h^2: $x = \frac{1}{N}\sum_{n=1}^N h_n$. Theno

$$P((x - \bar{h}^2) \geq \alpha) \leq \sigma_h^2/\alpha N.$$

Theorem (Asymptotic equipartition principle.). For an ensemble of N independent identically distributed (i.i.d.) random variable $X^N \equiv (X_1, X_2, \ldots, X_N)$, with N sufficiently large, the outcome $\mathrm{x} = (x_1, x_2, \ldots, x_N)$ is almost certain to belong to a subset of $\mathcal{A}_X^N$ having only $2^{NH(X)}$ members, each having probability "close" to $2^{-NH(X)}$. (The term equipartition is chosen to describe the idea that the members of the typical set have roughly equal probability.)

Proof of source coding theorem.

Verbally, the source coding theorem states that N i.i.d. random variables each with entropy $H(X)$ can be compressed into more than $NH(X)$ with negligible risk of information loss as $N \to \infty$. Conversely, if they are compressed into fewer than $NH(X)$ bits, it is virtually certain that information will be lost.

A long string of N symbols will usually contain about $p_i N$ occurences of the i-th symbol, so that the probability of this "typical" string is roughly

$$P(\mathrm{x})_{typ} \approx p_1^{p_1 N} p_2^{p_2 N} \cdots p_l^{p_l N},$$

so that the information content of a typical string is

$$\log_2 \frac{1}{P(\mathrm{x})} \approx N \sum_i p_i \log_2 \frac{1}{p_i} = NH.$$

First, apply the weak law of large numbers to the random variable $\frac{1}{N} \log_2 \frac{1}{P(x)}$. Define the typical set with parameters N and β as follows:

$$T_{N\beta} = \left\{ x \in \mathcal{A}_X : \left[\frac{1}{N} \log_2 \frac{1}{P(x)} - H \right]^2 < \beta^2 \right\}.$$

For all $x \in T_{N\beta}$, the probability of x satisfies

$$2^{-N(H+\beta)} < P(x) < 2^{-N(H-\beta)}.$$

By the law of large numbers, $P(x \in T_{N\beta}) \geq 1 - \sigma^2/(\beta^2 N)$.

This means that as N increases, the probability that x falls in $T_{N\beta}$ approaches 1, for any β.

Now, we will relate $T_{N\beta}$ to $H_\delta(X^N)$. Our strategy is to show that for any given δ, there is a sufficiently large N such that $H_\delta(X^N) \equiv NH$.

Part 1. $\frac{1}{N} H_\delta(X^N) < H + \epsilon$.

Since the total probability contained by $T_{N\beta}$ can't be larger than 1, we hve that

$$|T_{N\beta}| 2^{-N(H+\beta)} < 1,$$

that is

$$|T_{N\beta}| < 2^{N(H+\beta)}.$$

Setting $\beta = \epsilon$, and choosing N_0 such that $\frac{\sigma^2}{\epsilon^2 N_0} \leq \delta$, then $P(T_{N\beta}) \geq 1 - \delta$, and analyzing the set $T_{N\beta}$ implies

$$H_\delta(X^N) \leq \log_2 |T_{N\beta}| < N(H + \epsilon).$$

Part 2. $\frac{1}{N} H_\delta(X^N) > H - \epsilon$.

We set $\beta = \epsilon/2$, so it suffices to show that that a subset S' having $|S'| \leq 2^{N(H-beta)}$ and achieving $P(\mathrm{x} \in S') \geq 1 - \delta$ cannot exist.

The probability of the subset S' is

$$P(\mathrm{x} \in S') = P(\mathrm{x} \in S' \cap T_{N\beta}) + P(\mathrm{x} \in S' \cap \overline{TN_{N\beta}}),$$

where $\overline{T_{N\beta}}$ denotes the complement of the typical set.

IThe maximum value of the first term is found if $S' \cap T_{N\beta}$ contains $2^{N(H-2\beta)}$ outcomes all with the maximum probability $2^{-N(H-\beta)}$. The maximum value the second term can have is $P(\mathrm{x} \notin T_{N\beta})$.

Thus:

$$P(\mathrm{x} \in S') \leq 2^{N(H-2\beta)} 2^{-N(H-\beta)} + \frac{\sigma^2}{\beta^2 N} = 2^{-N\beta} + \frac{\sigma^2}{\beta^2 N}.$$

We can now set $\beta = \frac{\epsilon}{2}$ and N_0 such that $P(\mathrm{x} \in S') < 1 - \delta$, which shows that S' does not satisfy the desired conditions.

Therefore, for large enough N, the function $\frac{1}{N} H_\delta(X^N)$ is essentially a constant function of δ for $0 < \delta < 1$. In particular, this shows us that regardless of our specific tolerance for error, the number of bits per symbol needed to specify x is H bits.

Figure:

(fill in)

4.2 Maximum Entropy Principle

Due to E.T. Jaynes in 1957, where he explored the correspondence between statistical mechanics and information theory. Take precisely stated prior data or testable information about a probability distribution function. The distribution with maximal entropy is the best choice to encode the prior data.

- The exponential distribution for which the density function is

$$p(x|\lambda) = \begin{cases} \lambda e^{-\lambda x}; & x \geq 0 \\ 0; & x < 0, \end{cases}$$

  is the maximum entropy distribution among all continuous distributions supported in $[0, \infty)$ that have a specified mean of $\frac{1}{\lambda}$.

- The normal distribution $\mathcal{N}(\mu, \sigma^2)$ for which the density function is

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  has maximum entropy among all real-valued distributions supported on $(-\infty, \infty)$ with specified variance σ^2. Therefore: the assumption of normality imposes the minimal prior structural constraint.

To do: watch David Tse talk + talk on Information Theory on deep learning (Stanford).

4.3 Core ideas in information theory

1. Overview
   (a) Compression (lossless vs. lossy)
   (b) Communication (reliable vs. communication with loss [also joint source channel coding)
2. Course goals
   (a) Measures of information (entropy, relative entropy, mutual information, chain rules)
   (b) Compression, storage, communication
   (c) Fundamental limits
   (d) Concrete schemes for compression and communication
   (e) Existence proofs via random constructions (random coding)
   (f) Typical sequences interplay between info theory, probability, and stats

Example 1. Lossless compression.

Consider the source: $U_1, U_2, \dots$ iid $\sim U \in \{A, B, C\}$.

Further, suppose

$$P(U = A) = 0.7, P(U = B) = P(U = C) = 0.15.$$

Approach 1. Consider $A \longrightarrow 00$, $B \longrightarrow 01, C \longrightarrow 11$. But too wasteful, since A occurs more frequently.

Approach 2. Better is $A \longrightarrow 0, B \longrightarrow 01, C \longrightarrow 11$. Note that this is ’prefix code’: no code forms the prefix of another; this makes code easy to decode.

Expected number of bits per source symbol:

$$\overline{L} = 0.7 \cdot 1 + 0.15 \cdot 2 + 0.15 \cdot 2 = 1.3 \text{ bits / symbol}$$

Approach 3. In fact, we can do better. Consider pairs of source symbols. Namely, let us examine

Pair	Probability	Code Word
AA	0.49	0
AB	0.105	100
AC	0.105	111
BA	0.105	101
CA	0.105	1100
BB	0.0225	110100
BC	0.0225	110101
CB	0.0225	110110
CC	0.0225	110111

Satisfies prefix code. Encoding and decoding done in linear time. Later: we will see that this is optimal for these source symbols.

Again, let's compute expected bits per symbol:

$$\overline{L} = \frac{1}{2}(0.49 \cdot 1 + 0.105 \cdot 3 \cdot 3 + 0.105 \cdot 4 + 0.0225 \cdot 6 \cdot 4) = 1.1975 \text{ bits / symbol}$$

Entropy. For any scheme, the value $\overline{L} \geq H(U)$, where the source entropy

$$H(U) = \sum_{u \in \mathcal{U}} P(u) \log_2 \frac{1}{P(u)}.$$

In the above case:

$$H(U) \approx 1.18129.$$

On the other hand for all $\epsilon > 0$, there exists a scheme such that

$$\overline{L} \leq H(U) + \epsilon.$$

Example 2. Consider a source

$$U_1, U_2, \ldots, \quad \text{iid} \ ; \quad P(U_i = 0) = P(U_i = 1) = \frac{1}{2}.$$

Suppose further that a channel flips each bit w.p. $q < \frac{1}{2}$.

Output of channel:

$$Y_i = X_i \bigoplus_2 W_i,$$

where $W_i \sim \text{Ber}(q)$. Note that the source symbol U_i is different from the encoding X_i.

Approach 1. We can let

$$X_i = U_i,$$

we will get probability of error per source bit, $P_e = q$.

Approach 2. Alternatively, can repeat 3 times:

if $U = 0110$, then we can let $X = 000111111000$.

In this case:

$$\text{rate} = \frac{1}{3} \text{ bits / channel use}$$

The upside, is that the probability of error becomes

$$P_e = 3q^2(1-q) + q^3 < q.$$

So probability of error has dropped, at the cost of requiring more space.

4.4 Dyadic U and symbol counting

Lemma. Suppose U is dyadic with $|U| \geq 2$, and let $n_{max} = \max_{u \in \mathcal{U}} n_u$. The number of symbols with $n_u = n_{max}$ is even.

Proof. Observe that

$$\begin{aligned} 1 = \sum_u p(u) &= \sum_u 2^{-n_u} \\ &= \sum_{n=1}^{n_{max}} (\text{\# of letters } u \text{ with } n_u = n) \cdot 2^{-n} \end{aligned}$$

Therefore,

$$\begin{aligned} 2^{n_{max}} &= \sum_{n=1}^{n_{max}} (\text{ of letters } u \text{ with } n_u = n) \cdot 2^{n_{max}-n}. \\ &= \sum_{n=1}^{n_{max}-1} (\text{\# of letters } u \text{ with } n_u = n) \cdot 2^{n_{max}-n} + (\text{\# of letters } u \text{ with } n_u = n_{max}). \end{aligned}$$

By parity, it follows that # of letters u with $n_u = n_{max}$ must be even.

4.5 Optimality of Huffman Codes

Construction of Huffman Codes. Exactly the same as that for dyadic sources. Recall that the procedure identifies the symbols with the smallest probabilities and merges them in a binary tree structure.

Example. (Senary Source) Consider the alphabet

Add D1

u	$p(u)$
a	0.25
b	0.25
c	0.2
d	0.15
e	0.1
f	0.05

Theorem. Huffman code is an optimal prefix code.

(Note that we say an optimal and not "the optimal" because there may be more than one construction. Even within the construction of Huffman, the way we break ties is arbitrary. We can also choose to split the binary tree in one direction via a 1 vs. 0. So we can have many different schemes, though they are all essentially equivalent, in terms of the length function.)

When we use the term "optimality" here, we mean in terms of minimizing the expected length $\bar{l}$.

Proof. Assume without loss of generality that $U \sim P$ over an alphabet $\mathcal{U} = \{1, 2, \ldots, r\}$. Further, suppose that $p(1) \geq p(2) \cdots \geq p(r)$ (i.e. they are arranged in descending probabilities).

Let V denote the random variable with $\mathcal{V} = \{1, 2, \ldots, r-1\}$ obtained from U by merging $r-1$ and r.

Let $\{c(i)\}_{i=1}^{r-1}$ be a prefix code for V. Then we can obtain $\{\tilde{c}\}_{i=1}^{r}$ which is a prefix code that *splitting* the last codeword $c(r-1)$.

Observation. The Huffman code for U is obtained from the Huffman code from V by splitting.

Lemma. Suppose that $\{c(i)\}_{i=1}^{r-1}$ is an optimal prefix code for V. If $\{\tilde{c}(i)\}_{i=1}^{r}$ is obtained from $\{c(i)\}_{i=1}^{r-1}$ by splitting, then $\{\tilde{c}(i)\}_{i=1}^{r}$ is an optimal prefix code for U.

This observation coupled with the lemma directly implies the theorem. We can iterate this argument to merely need establish optimality of Huffman code for binary alphabet ($r = 2$), which is trivially true.

Proof of Lemma. Note there is an optimal prefix code for U that satisfies:

1. $\bar{l}(1) \leq l(2) \leq \cdots \leq l(r-1) \leq l(r) \triangleq l_{max}$ (lengths are in increasing order).
2. $l(r-1) = l(r)$.

   (Otherwise, we would be able to "chop off" the final part of the last code word to achieve $l(r-1) = l(r)$ and improve the code.)
3. The last two code words differ only in the last bit.

   (Otherwise, we can swap out the last code word. This follows since the first $r-1$ codewords comprise a prefix code.) This ensures that the prefix code for U is obtained by splitting on the code for V.

Recall the following:

$$\mathbb{E}l_{split}(U) = \mathbb{E}l(V) + p(r-1) + p(r).$$

Therefore: an optimal prefix code for U is obtained by splitting an optimal prefix code for V. ■

Further reading on lossless compression:

- Shannon-Fano-Elias coding (5.9 of Cover and Thomas)
- Arithmetic coding (13.3)
- Lempel-Ziv coding (13.4)

Note that optimally applying Huffman codes requires working in blocks of symbols. And the table of symbols is exponential in the block length n. The Shannon-Fano-Elias and Arithmetic coding permit constructions that scale gracefully in the block length n. Lempel-Ziv coding is elegant algorithmically, and is guaranteed to be optimal even without the source being memoryless and even without knowing the probability distribution! Indeed, `gzip` at its heart is implemented in terms of the Lempel-Ziv coding scheme.

4.6 Channel Capacity

Given a channel with inputs X and outputs Y:

$$X \to [P(Y|X)] \to Y$$

Define: Channel capacity C is the maximal rate of reliable communication (over memoryless channel characterized by $P(Y|X)$).

Further, recall the following definition:

$$C^{(I)} = \max_{P_X} I(X;Y).$$

Theorem. Channel capacity is limited by maximum mutual information.

$$C = C^{(I)}.$$

Proof: We will see this proof soon.

- This theorem is important because C is challenging to optimize over, whereas $C^{(I)}$ is a tractable optimization problem.

4.6.0.1 Examples

Example I. Channel capacity of a Binary Symmetric Channel (BSC).

Define alphabets $\mathcal{X} = \mathcal{Y} = \{0,1\}$. A BSC is defined by the PMF:

$$P_{Y|X}^{(y|x)} = \begin{cases} p & y \neq x \\ 1-p & y = x. \end{cases}$$

This is equivalent to a channel matrix

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

And the graph representation

This can also be expressed in the form of additive noise.

$$Y = X \bigoplus_2 Z, \text{ where } Z \sim \text{Ber}(p).$$

To determine the channel capacity of a BSC, by the theorem we must maximize the mutual information.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(X \bigoplus_2 Z|X) \end{aligned}$$

Because only the random noise can't be modeled by conditioning on X, we can simplify the second term:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Z) \\ &= H(Y) - h_2(p) \leq 1 - h_2(p). \end{aligned}$$

Taking $X \sim \text{Ber}(\frac{1}{2})$ achieves equality: $I(X;Y) = 1 - h_2(p)$.

Example II. Channel capacity of a Binary Erasure Channel (BEC).

Define alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. Any input symbol X_i has a probability of $1 - \alpha$ of being retained in the output sequence and a probability of α of being erased. Schematically, we have:

Examining the mutual information, we have that

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - [H(X|Y=e)P(Y=e) + H(X|Y=0)P(Y=0) + H(X|Y=1)P(Y=1)] \\ &= H(X) - [H(X) \cdot \alpha + 0 \cdot P(Y=0) + 0 \cdot P(Y=1)] \\ &= (1-\alpha)H(X) \end{aligned}$$

Because the entropy of a binary variable can be no larger than 1:

$$(1-\alpha)H(X) \leq 1 - \alpha$$

Equality is achieved when $H(X) = 1$, that is $X \sim \text{Ber}(\frac{1}{2})$.

4.6.1 Information of Continuous Random Variables

Definition: The relative entropy between two probability density functions f and g is given by

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx.$$

Exercise: Show that $D(f||g) \geq 0$ with equality if and only if $f = g$.

Proof. Observe that that

$$\begin{aligned} D(f||g) &= \int f(x) \log \frac{f(x)}{g(x)} dx \\ &= -\int f(x) \log \frac{g(x)}{f(x)} dx \\ &= -\mathbb{E}\left[\log \frac{g(x)}{f(x)}\right] \\ &\geq -\log \mathbb{E}\left[\frac{g(x)}{f(x)}\right] \\ &= -\log \int f(x) \frac{g(x)}{f(x)}\, dx \\ &= 0. \end{aligned}$$

Equality occurs in the manner of Jensen's when $f = g$.

Definition: The mutual information between X and Y that have a joint probability density function $f_{X,Y}$ is

$$I(X;Y) = D(f_{X,Y}||f_X f_Y).$$

Definition: The differential entropy of a continuous random variable X with probability density function f_X is

$$h(X) = -\int f_X(x) \log f_X(x)\, dx = \mathbb{E}\left[-\log f_X(X)\right]$$

If X, Y have joint density $f_{X,Y}$, the conditional differential entropy is

$$h(X|Y) = -\int f_{X,Y}(x,y) \log f_{X|Y}(x|y)\, dx\, dy = \mathbb{E}[-\log f_{X|Y}(X|Y)],$$

and the joint differential entropy is

$$h(X,Y) = \int f_{X,Y}(x,y) \log f_{X,Y}(x,y)\, dx\, dy = \mathbb{E}[-\log f_{X,Y}(X,Y)].$$

4.6.2 Exercises

Exercise 1. Show that

$$h(X|Y) \leq h(X)$$

with equality iff X and Y are independent.

Proof. This follows since $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \leq f_X(x)$ and log is monotonic. The equality condition is true since we have equality in $f_{X|Y}(x|y) = f_X(x)$ iff X and Y are independent.

Exercise 2. Show that

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y). \end{aligned}$$

Proof.

$$\begin{aligned} I(X;Y) &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \\ &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)} dxdy - \int f_{X,Y}(x,y) \log f_Y(y) dxdy \\ &= \int f_X(x) \left[\int f_{Y|X}(y|x) \log f_{Y|X}(y|x) dy\right] dx - \int f_Y(y) \log f_Y(y) dy \\ &= H(Y) - H(Y|X). \end{aligned}$$

Symmetrically the same can be shown for $I(X;Y) = H(X) - H(X|Y)$. Also

$$\begin{aligned} I(X;Y) &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \\ &= \int f_{X,Y}(x,y) \log f_{X,Y}(x,y) dxdy - \int f_{X,Y}(x,y) \log f_X(x) dxdy - \int f_{X,Y}(x,y) \log f_Y(y) dxdy \\ &= H(X,Y) - H(X) - H(Y). \end{aligned}$$

Exercise 3. Show that

$$h(X + c) = h(X).$$

and

$$h(c \cdot X) = h(X) + \log|c|, c \neq 0.$$

Proof.

Note that

$$h(X + c) = \mathbb{E}[-\log f_X(X + c)] = \mathbb{E}[-\log f_X(X)] = h(X);,$$

since we are integrating over the same probabilities, we are integrating over the same probabilities, the expectation of the log-density is invariant to constant shifts.

Further, note that

$$h(c \cdot X) = \mathbb{E}[-\log f_X]$$

.

To compute $h(c \cdot X)$, we start by considering the density function $p(c \cdot X)$. Set $y = c \cdot X$, yielding $dy = cdx$. We must have

$$\int p(y)\,dy = 1 = \int p(cx) \cdot c\,dx.$$

To satisfy this equality, it follows that $p(y) = \frac{p(x)}{c}$.

Therefore,

$$\begin{aligned} h(Y) &= -\int p(y) \log p(y)\,dy \\ &= -c \int p(cx) \log p(|cx|)\,dx \\ &= -c \int \frac{p(x)}{c} \log \frac{p(x)}{|c|}\,dx \\ &= -\int p(x)[\log p(x) - \log(|c|)]\,dx \\ &= h(X) + \log|c|. \end{aligned}$$

We have introduced the absolute value on c to satisfy the domain of the logarithm function.

4.6.3 Examples

***Example I:* Differential entropy of a uniform random variable $U \sim$ Uni(a, b).**

- Remember that the distribution of a uniform random variable is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The differential entropy is simply:

$$h(X) = \mathbb{E}[-\log f_X(x)] = \log(b-a)$$

- Notice that the differential entropy can be negative or positive depending on whether $b-a$ is less than or greater than 1. In practice, because of this property, differential entropy is usually used as means to determine mutual information rather than by itself.

***Example II:* Differential entropy of a Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$.**

- Remember that the distribution of a Gaussian random variable is $f(X) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}x^2}$.

The differential entropy is:

$$h(X) = \mathbb{E}[-\log f(X)]$$

For simplicity, convert the base to e:

$$\begin{aligned}
h(X) &= \frac{1}{\ln 2}\mathbb{E}[-\ln f(X)] \\
&= \frac{1}{\ln 2}\mathbb{E}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}X^2\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \mathbb{E}\left[\frac{1}{2\sigma^2}X^2\right]\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sigma^2\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi e\sigma^2\right] = \frac{1}{2}\log 2\pi e\sigma^2
\end{aligned}$$

- Per Exercise 3, differential entropies are invariant to constant shifts. Therefore this expression represents the differential entropy of all Gaussian random variables regardless of mean.
- *Claim:* The Gaussian distribution has maximal differential entropy, i.e. for all random variables $X \sim f_X$ with second moment $E[X^2] \leq \sigma^2$ and Gaussian random variable $G \sim \mathcal{N}(0, \sigma^2)$ then $h(X) \leq h(G)$. Equality holds if and only if $X \sim \mathcal{N}(0, \sigma^2)$.

  Proof:

$$\begin{aligned}
0 \leq D(f_X \| G) &= \mathbb{E}\left[\log \frac{f_X(X)}{f_G(X)}\right] \\
&= -h(X) + \mathbb{E}\left[\log \frac{1}{f_G(X)}\right] \\
D(f_X \| G) &= -h(X) + \mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\frac{X^2}{2\sigma^2}}{\ln 2}\right]
\end{aligned}$$

  Because the second moment of X is upper bounded by the second moment of G:

$$\begin{aligned}
0 \leq D(f_X \| G) &\leq -h(X) + \mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\frac{G^2}{2\sigma^2}}{\ln 2}\right] \\
&\leq -h(X) + \mathbb{E}\left[\log \frac{1}{f_G(G)}\right] = -h(X) + h(G)
\end{aligned}$$

Rearranging:

$$h(X) \leq h(G)$$

□

Example III: Channel capacity of an Additive White Gaussian Noise channel (AWGN) that is restricted by power p

- Power constraint upper bounds the second moment of X_i, i.e. $p \geq E\left[X_i^2\right]$.
- Remember that the AWGN channel is a channel in which inputs X_i are corrupted by a sequence of iid additive Gaussian noise terms $W_i \sim \mathcal{N}(0, \sigma^2)$ to produce outputs Y_i.
- The Channel Coding Theorem in this setting states that:

  $$C(p) = \max_{E[X^2] \leq p} I(X;Y)$$

  Where $C(p)$ represents the 'capacity'; the maximal rate of reliable communication when constrained to power p.

4.7 Constraints and communication theory

Note that the encoder is equivalent to a ”codebook” framed as follows:

$$c_n = \{X^n(1), X^n(2), \ldots, X^n(M)\}.$$

Here, the decoder is equivalent to the mapping $\hat{J}(.)$.

In this context, a scheme is defined as an “encoder-decoder” pair. Equivalently, this can be framed in terms of a “codebook-mapping” pair.

Definition. The rate is defined as

$$\text{rate} = \frac{\log M}{n} = \frac{\log |C_n|}{n} \frac{\text{bits}}{\text{channel use}}.$$

where M is the number of messages, and n is the number of channel uses. Note that M is equivalent to the size of the codebook $|C_n|$.

The probability of error can be computed as

$$P_e = P(\hat{J} \neq J).$$

Sometimes we also have a transmission constraint:

$$\frac{1}{n}\sum_{i=1}^{n} \Lambda(X_i) \leq P,$$

where Λ defines a cost function.

Example. The most common physically meaningful cost constraint pertains to the power of an electromagnetic signal. In particular, in wireless communication, we have:

$$\Lambda(x) = x^2.$$

Another example is magnetic storage media, which might have a different cost of encoding.

Recall the notion of capacity, where

$$C = \text{maximal rate of reliable communication.}$$

Further, we had the informational capacity, defined as follows:

$$C^{(I)} = \begin{cases} \max\limits_{P_X} I(X;Y); & \text{without a transmission constraint.} \\ \max\limits_{P_X : \mathbb{E}\Lambda(X) \leq P} I(X;Y); & \text{with a constraint.} \end{cases}$$

Theorem. Recall the channel coding theorem, which states the remarkable fact that

$$C = C^{(I)}.$$

Recall the following results.

1. If $G \sim \mathbb{N}(0, \sigma^2)$, then $h(G) = \frac{1}{2}\log 2\pi e\sigma^2$.
2. If X is any random variable such that $\mathbb{E}[X^2] \leq \sigma^2$ (i.e. the second moment is constrained), then $h(X) \leq h(G)$.

We now go back to example 3 from the previous section

Example III. Consider the additive white Gaussian noise (AWGN) channel, defined as D2

(In particular, when we draw diagrams with perpendicular inputs as we have done here, we mean that X and W are independent.)

And further, suppose that transmission is restricted to a power p. Namely, suppose

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq p.$$

Let $C(P)$ denote the maximal rate of reliable communication constrained to power P. The channel coding theorem states that

$$C(P) = \max_{P_X : \mathbb{E}[X^2] \leq p} I(X;Y)$$

If $\mathbb{E}[X^2] \leq p$, then

$$I(X;Y) = h(Y) - h(Y|X)$$

Since differential entropy is invariant to constant shifts, we can write:

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y - X|X) \\ &= h(Y) - h(W|X) \\ &= h(Y) - h(W). \end{aligned}$$

Since $\mathrm{Var}(Y) = \mathrm{Var}(X) + \mathrm{Var}(W) \leq P + \sigma^2$,

$$\begin{aligned} &\leq h(\mathbb{N}(0, p + \sigma^2)) - h(\mathbb{N}(0, \sigma^2)) \\ &= \frac{1}{2} \log 2\pi e(p + \sigma^2) - \frac{1}{2} \log 2\pi e \sigma^2 \\ &= \frac{1}{2} \log \left(1 + \frac{p}{\sigma^2}\right). \end{aligned}$$

We now try to find a distribution where this bound is achieved. To achieve equality, we require $\mathrm{Var}(Y) = \mathrm{Var}(X) + \mathrm{Var}(W) = p + \sigma^2$.

In particular, let $X \sim \mathbb{N}(0, p)$, which satisfies the equality, i.e.

$$X \sim \mathbb{N}(0, p) \implies C(p) = \frac{1}{2} \log \left(1 + \frac{p}{\sigma^2}\right)$$

Note that $\frac{p}{\sigma^2}$ is known as the signal-to-noise ratio.

Rough geometric intuition:

Note that the power constraint can be expressed as

$$\sqrt{\sum_{i=1}^{n} X_i^2} \leq \sqrt{np}.$$

Think of $X^n(i)$ as points in n-dimensional Euclidean space. Then they lie on a sphere of radius $\sqrt{np}$.

Then, observe that

$$\frac{1}{n} \sum_{i=1}^{n} W_i^2 \approx \sigma^2 \Leftrightarrow \sqrt{\sum_{i=1}^{n} W_i^2} \approx \sqrt{n\sigma^2}.$$

The channel output can be expressed as

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i^2\right] = \sum_{i=1}^{n} \mathbb{E}[X_i^2] + \mathbb{E}[W_i^2] + \underbrace{\mathbb{E}[X_i W_i]}_{=0} \leq np + n\sigma^2.$$

Geometrically, we would like the “noise balls” to be disjoint; i.e. they should not intersect, so we can reliably discern which message point is sent.

We now want to consider bounds on the number of messages we can send. In particular, consider fix

$$\begin{aligned}\text{\# of messages} &\le \frac{\text{Vol(Ball of radius } \sqrt{n(p+\sigma^2)})}{\text{Vol(Vall of radius } \sqrt{n\sigma^2})}.\\ &= \frac{k_n(\sqrt{n(p+\sigma^2)})^2}{k_n(\sqrt{n\sigma^2})^n} = \left(\frac{p+\sigma^2}{\sigma^2}\right)^{n/2} = \left(1+\frac{p}{\sigma^2}\right)^{n/2}\end{aligned}$$

Therefore, the rate can be bounded by

$$\text{rate} = \frac{1}{2}\log\frac{\text{\# of messages}}{n} \le \frac{1}{2}\log\left(1+\frac{p}{\sigma^2}\right).$$

4.7.1 Joint Asymptotic Equipartition Principle

Consider X, Y which have finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, where

$$(X,Y) \sim P_{X,Y}; \quad X \sim P_X; \quad Y \sim P_Y.$$

Here, the pairs

$$(X_i, Y_i); \text{ iid } \sim (X,Y),$$

where

$$p(x^n) = \prod_{i=1}^n P_X(x_i),$$

$$p(y^n) = \prod_{i=1}^n P_Y(y_i).$$

$$p(x^n, y^n) = \prod_{i=1}^n P_{X,Y}(x_i, y_i).$$

Definition. The set of jointly typical sequences is defined as

$$A_\epsilon^n(X,Y) = \{(x^n,y^n) : \left|-\frac{1}{n}\log P(x^n) - H(X)\right| \le \epsilon; \quad \left|-\frac{1}{n}\log P(y^n) - H(Y)\right| \le \epsilon; \quad \left|-\frac{1}{n}\log P(x^n,y^n) - H(X,Y)\right| \le \epsilon\}$$

Part A. If (X^n, Y^n) are formed by iid $(X_i, Y_i) \sim (X,Y)$, then

1. $P((X^n,Y^n) \in A_\epsilon^{(n)}(X,Y)) \longrightarrow 1$, as $n \longrightarrow \infty$ (basically follows directly from the original AEP on each subpart of the definition).
2. $2^{n(H(X,Y)-\epsilon)} \le |A_\epsilon^{(n)}(X,Y)| \le 2^{n(H(X,Y)+\epsilon)}$ (proof left to scribers, basically follows from original AEP).

Part B. If $(\tilde{X}^n, \tilde{Y}^n)$ are formed by iid $(\tilde{X}_i, \tilde{Y}_i) \sim (\tilde{X}, \tilde{Y})$ where $P_{\tilde{X},\tilde{Y}} = P_X P_Y$.

Then:

$$(1-\epsilon)2^{-nI(X,Y)+3\epsilon} \leq P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) \leq 2^{-nI(X,Y)-3\epsilon};$$

for all $\epsilon > 0$ and (analytical details to be covered next time). Requires large n.

Intuition. Suppose $\tilde{X}, \tilde{Y}$ are generated independently, how likely is it to look like it came from a joint distribution? Answer: Exponentially unlikely.

4.8 Channel Capacity Theorem

Recall: the communication problem setting.

Rate of communication: number of bits per channel use, i.e.

$$\texttt{rate} = \frac{\log M}{n} \frac{\texttt{bits}}{\texttt{channel use}}$$

Define probability of error as

$$P_e = P(\hat{J} \neq J).$$

Main result:

$$C = \max_{P_X} I(X;Y).$$

Here, we will not concern ourselves with power / cost constraint.

We will break down this result into two sub-results. Equivalent to:

- Direct part: If $R < \max_{P_X} I(X;Y)$, then R is achievable. This means, that there exist schemes with rate $\geq R$, and $P_e \longrightarrow 0$.
- Converse part: If $R > \max_{P_X} I(X;Y)$, then R is not achievable.

In this section, we will prove the direct part of the theorem.

4.8.1 Joint AEP

Recall the setting. Consider a pair of random variables $(X,Y) \sim P_{X,Y}$ with finite alphabets $\mathcal{X}$, $\mathcal{Y}$. This implies that the pair

$$(X,Y) \text{ has alphabet } \mathcal{X} \times \mathcal{Y},$$

here × represents the Cartesian product over sets. The jointly typical set

$$\begin{aligned} A_\epsilon^{(n)}(X,Y) = \{(X^n, Y^n) : &\left| -\frac{1}{n}\log p(X^n) - H(X) \right| \leq \epsilon, \\ &\left| -\frac{1}{n}\log p(Y^n) - H(Y) \right| \leq \epsilon, \\ &\left| -\frac{1}{n}\log p(X^n, Y^n) - H(X,Y) \right| \leq \epsilon\} \end{aligned}$$

Note that Part A of the joint AEP states that:

- If $(X_i, Y_i) \sim (X, Y)$, then for any $\epsilon > 0$,

$$P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1.$$

- $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}(X,Y)| \leq 2^{nH(X,Y)+\epsilon}$ essentially for all large n.
- Suppose now $\tilde{X}^n \overset{d}{=} X^n$ and $\tilde{Y}^n \overset{d}{=} Y^n$ and $\tilde{X}$ and $\tilde{Y}$ are independent. Then

$$\tilde{X}^n \approx \texttt{uniformly distributed on } A_\epsilon^n(X).$$

$$\tilde{Y}^n \approx \texttt{uniformly distributed on } A_\epsilon^n(Y).$$

  and, since $\tilde{X}^n$ and $\tilde{Y}^n$ are independent, the joint distribution

$$(\tilde{X}^n, \tilde{Y}^n) \approx \texttt{uniformly distributed on } A_\epsilon^n(X,Y).$$

  It follows that

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^n(X,Y)) \approx \frac{|A_\epsilon^n(X,Y)|}{|A_\epsilon^n(X) \times A_\epsilon^n(Y)|} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X;Y)}.$$

- Formally stated, we find that for all $\epsilon > 0$, for sufficient large n,

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^n(X,Y)) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Interpretation of mutual information: quantifies "how unlikely two sequences that are independent appear that they are jointly typical?"

4.8.2 Relation of AEP to Communication Problem

Idea. (Proof of direct part of Communication Theorem)

- Randomly select codewords of the codebook from the typical set $A_\epsilon^{(n)}(X)$.

- Suppose we encode a codeword as $X^n(J)$. Then

$$P(Y^n \texttt{ is jointly typical with } X^n(J)) \approx 1.$$

  Further,

$$P(Y^n \texttt{ is jointly typical with some } X^n(i) \texttt{ for a particular } i \texttt{ not set}) \approx 2^{-nI(X;Y)}.$$

  Applying the previous result with a union bound:

$$P(Y^n \texttt{ is jointly typical with any of the codewords not sent}) \approx \text{very small, provided t}$$

$$R < I(X;Y).$$

- Implies that: Joint typicaly decoding will be reliable, for $R < I(X;Y)$ (i.e. get you very small probability of error).

Proof of direct part. Fix P_X and $R < I(X;Y)$. We need to show that R is an achievable rate for reliable communication. Take $\epsilon > 0$ sufficiently small such that $R < I(X;Y) - 3\epsilon$. Generate a codebook C_n of size $M = \lceil 2^{nR} \rceil$ randomly:

$$\texttt{take } X^n(1), X^n(2), \dots, X^n(m) \texttt{ iid, each iid} \sim P_X.$$

Then, the jointly typical decoding rule states that

$$\hat{J} = (\hat{Y^n}) = \begin{cases} j; & \text{if } (X^n(j), Y^n) \in A_\epsilon^{(n)}(X,Y) \text{ and } (X^n(k), Y^n) \notin A_\epsilon^{(n)}(X,Y) \quad \forall k \neq j \\ e \quad \text{(error)}; & \texttt{otherwise}. \end{cases}$$

Our rough discussion states that with very high probability, we will find the true code word that was sent. Consider one possible codebook c_n and a decoding rule. Let the probability of error be

$$P_e(c_n) = P(\hat{J} \neq J | C_n = c_n) :$$

Then

$$\mathbb{E}[P_e(c_n)] = P(\hat{J} \neq J) = \sum_{j=1}^{M} P(\hat{J} \neq J | J = j) P(J = j) = P(\hat{J} \neq J | J = 1).$$

$$\leq P((X^n(1), Y^n) \notin A_\epsilon^{(n)}(X,Y) | J = 1) + \sum_{j=2}^{M} P((X^n(j), Y^n) \in A_\epsilon^{(n)}(X,Y) | J = 1)$$

In the last inequality, we have used a union bound: either

- the Y sequence is not jointly typical with the message sent,
- or it is jointly typical with one of the other codewords sent.

This last quantity is equal to

$$\begin{aligned} &P((X^n(1), Y^n) \notin A_\epsilon^{(n)}(X,Y) | J = 1) + \sum_{j=2}^{M} P((X^n(j), Y^n) \in A_\epsilon^{(n)}(X,Y) | J = 1) \\ &= P((X^n, Y^n) \notin A_\epsilon^{(n)}(X,Y)) + (M-1) P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) \\ &\leq 2^{-n(I(X;Y) - 3\epsilon - R)}. \end{aligned}$$

Note that in particular there exists a codebook c_n such that $|c_n| \leq 2^{nR}$ and $P_e(c_n) \leq \mathbb{E}[P_e(c_n)]$.

This implies that there exists a sequence of codebooks, $\{c_n\}_{n \geq 1}$ with $|c_n| \geq 2^{nR}$ and vanishing $P_e(c_n) \to 0$. And in particular, this means that R is an achievable rate for reliable communication. ■

There are a couple of problematic aspects of this proof:

- We have shown the existence of the codebooks, but not constructed one explicitly.

- Even if you were to find the codebook, they don't necessarily have good structure (codebook might be exponentially large, and have other undesirable properties.)

Note, our notation of reliability is

$$P_e = P(\hat{J} \neq J) = \sum_{j=1}^{M} P(\hat{J} \neq J | J = j) P(J = j).$$

One can consider a more stringent criterion:

$$P_{max} = \max_{1 \leq j \leq m} P(\hat{J} \neq J | J = j).$$

Exercise. Given c_n with $P_e(c_n)$, there exists a codebook c_n' such that $|c_n'| \geq \frac{1}{2}|c_n|$ and $P_{max}(c_n') \leq 2P_e(c_n)$. In this case, if $|c_n| = 2^{nR}$, then $|c_n'| \geq \frac{1}{2}2^{nR} \implies \texttt{rate} \geq \frac{\log \frac{1}{2}2^{nR}}{n} = R - \frac{1}{n}$.

Next week: we will discuss practical constructions of these codebooks. We still need to prove the converse part as well.

4.9 Channel Coding Theorem; Converse Part

In this section, we will discuss the proof of our main theorem in the communication setting.

Recall the communication setting:

$$J \sim \text{Unif}\{1, 2, \dots, m\} \to \text{encoder } (X_n) \to \text{memoryless channel } P_{Y|X}; Y^n \to \text{decoder } \hat{J}$$

Main result:

$$C = C^{(I)} = \max_{P_X} I(X; Y).$$

Last week, we showed that R is achievable if $R < C^{(I)}$. In this section, we will show the converse, i.e. if $R > C^{(I)}$, then R is not achievable.

Theorem. (Fano's inequality) Let X be a discrete random variable, and $\hat{X}(Y)$ be a guess of X based on Y. Let $P_e = P(X \neq \hat{X})$. Then:

$$H(X|Y) \leq h_2(P_e) + P_e \log(|\mathcal{X}| - 1).$$

Proof. Intuition: Fano's inequality relates the notion of conditional entropy and the probability of error.

Let $V = 1\left\{X \neq \hat{X}\right\}$. By the data processing inequality, we have that

$$\begin{aligned}
H(X|Y) &\leq H(X,V|Y) \\
&= H(V|Y) + H(X|V,Y) && \text{(chain rule)} \\
&\leq H(V) + \sum_{v,y} H(X|V=v,Y=y)P(V=v,Y=y) && \text{(conditioning reduces entropy)} \\
&= H(V) + \sum_{y} \underbrace{H(X|V=0,Y=y)P(V=0,Y=y)}_{0} + \sum_{y} \underbrace{H(X|V=1,Y=y)}_{\leq \log(|\mathcal{X}|-1)} P(V=1,Y=y) \\
&\leq H(V) + P(V=1)\log(|\mathcal{X}|-1) \\
&= h_2(P_e) + P_e \log(|\mathcal{X}|-1)
\end{aligned}$$

□

Remark. Often, the weakened version of Fano's inequality is often used:

$$H(X|Y) \leq 1 + P_e \log|\mathcal{X}|,$$

or equivalently

$$P_e \geq \frac{H(X|Y)-1}{\log|\mathcal{X}|}.$$

Proof. (Proof of converse part of channel coding theorem.)

For any scheme, consider

$$\begin{aligned}
\log M - H(J|Y^n) &= H(J) - H(J|Y^n) \\
&= I(J;Y^n) \\
&= H(Y^n) - H(Y^n|J) \\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1},J) && \text{(by the chain rule)} \\
&\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1},X_i,J) && \text{(conditioning reduces entropy)} \\
&= \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\
& \qquad \text{(memorylessness of the channel, implying that } Y_i - X_i - (Y^{i-1},J)\text{)} \\
&= \sum_{i=1}^{n} I(X_i;Y_i) \\
&\leq nC^{(I)} && \text{(since } C^{(I)} \text{ is the maximal mutual information)}
\end{aligned}$$

Now, consider any scheme with a rate $\frac{\log M}{n} \geq R$. By the weakened version of Fano, we have

$$\begin{aligned} P_e &\geq \frac{H(J|Y^n) - 1}{\log M} \\ &\geq \frac{\log M - nC^{(I)} - 1}{\log M} \\ &\geq 1 - \frac{C^{(I)}}{R} - \frac{1}{nR} \to 1 - \frac{C^{(I)}}{R}. && \text{(as } n \to \infty) \end{aligned}$$

But notice that if $R > C^{(I)}$, then the P_e must be lower bounded by a positive value. So this sequence of schemes cannot have a nonvanishing probability of error.

$$\boxed{\texttt{If } R > C^{(I)} \texttt{ then } R \texttt{ is not achievable.}}$$

This concludes the proof. □

Remark. (Some notes on this proof).

1. Communication with feedback: $X_i(J, Y^{i-1})$. This is a perhaps more powerful encoder - since the encoder can adapt to what it has seen so far. However, one can verify that the proof from before holds verbatim. Therefore,

   $$C = C^{(I)} \qquad \text{(with or without feedback)}$$

   However - P_e will vanish much more quickly, and the resulting schemes will be much more simple.

   Consider the example of communicating to the erasure channel with feedback. Earlier, we found that the capacity is given by

   $$C = 1 - \alpha \frac{\texttt{bits}}{\texttt{channel use}}$$

   With feedback, just repeat each information bit until it gets through. Then, one average, we will need $\frac{1}{1-\alpha}$ channel uses per information bit that we want to send. Hence, the rate achieved will be

   $$1 - \alpha \frac{\texttt{bits}}{\texttt{channel use}}$$

   This protocol has 0 probability of error, since we can just wait until the bit gets through.

2. In the proof of the direct part, we showed mere existence of schemes; i.e. existence of codebooks c_n with size $|c_n| \geq 2^{nR}$ and small P_e. For practical schemes, note that LDPC codes and polar codes are concrete ways to construct these codebooks (see EE388).

3. Note that the proof of the direct part assumed finite alphabets. This carries over to a general case by approximation / quantization.

4. How do communication limits change if we want the maximal probability of error P_{max} to be small, instead of the average probability of error P_e? Recall the definitions:

$$P_e = P(\hat{J} \neq J) = \frac{1}{m} \sum_{j=1}^{m} P(\hat{J} \neq j | J = j).$$

$$P_{max} = \max_{1 \leq j \leq m} P(\hat{J} \neq j | J = j).$$

But: let's look at the "better half" of the codebook. Consider the set of messages

$$|\left\{1 \leq j \leq M : P(\hat{J} \neq j | J = j) \geq 2P_e\right\}| \geq \frac{M}{2} \qquad \text{(by Markov's inequality)}$$

Given c_n with $|c_n| = M$ and P_e, there exists c_n' with $|c_n'| \geq \frac{M}{2}$ and $P_{max} \leq 2P_e$ - just take the messages in this better set. Then:

$$\text{rate of } c_n' \geq \frac{\log \frac{M}{2}}{n} = \frac{\log M}{n} - \frac{1}{n}.$$

If there exist schemes of rate $\geq R$ with $P_e \longrightarrow 0$, then there exist schemes of rate $\geq R - \epsilon$ with $P_{max} \longrightarrow 0$.

In conclusion,

$$C = C^{(I)} \qquad (\text{ under either } P_e \text{ or } P_{max})$$

4.10 Lossy Compression & Rate Distortion Theory

4.10.1 Lossy compression problem setting

- Let U_i iid $\sim U$.
- Let the compressor compress the source to n bits.

$$U_1, U_2, \ldots, U_n \rightarrow \texttt{compressor / encoder} \rightarrow \texttt{decoder} \rightarrow V_1, V_2, \ldots, V_n$$

- Compression rate is defined as

$$\frac{n}{N} \frac{\texttt{bits}}{\texttt{source symbol}}$$

- Specify a distortion criterion d, and we will look at the expected per-symbol distortion; referred to as the "distortion" achieved.

$$D = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)\right]$$

- In lossy compression, we may allow D to be positive, but we want to constrain D.

- In general, there will be a tension between the distortion and the rate. We would like to identify the tradeoff. Of course, if we force distortion $D = 0$, then the best rate is the entropy. More generally, if we agree to incur a positive distortion, we can get away with smaller rate (less than the entropy).
- Concretely, when we parametrize a scheme, we need to specify:

$$\texttt{scheme} = (N, n, \texttt{encoder}, \texttt{decoder}).$$

Definition. A pair (R, D) is said to be achievable if for all $\epsilon > 0$ there exists a scheme such thatits rate

$$\frac{n}{N} \leq R + \epsilon \quad \text{and} \quad \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)\right] \leq D + \epsilon.$$

Definition. The rate distortion function $R(D)$ is defined to be

$$R(D) = \inf\left\{R' : (R', D) \text{ is achievable }\right\}.$$

Note that the rate distortion function is the minimal rate, optimized across all the possible schemes in the world.

Definition. The informational rate distortion function is given by

$$R^{(I)}(D) = \min_{\mathbb{E}d(U,V)\leq D} I(U; V)$$

Note that in the setting when you have continuous data, the notion of rate distortion is perhaps more important, because it does not make sense to talk about lossless compression.

Theorem. (Main result)

$$R(D) = R^{(I)}(D)$$

4.10.2 Qualitative analysis of $R(D)$

What does $R(D)$ look like (qualitatively?) Assume a discrete source, which we can compress losslessly with a rate equal to the entropy.

Note that $R(D)$ is monotone decreasing and convex. This is intuitive – we can always compress at at least the same rate if allowed higher distortion. $R(D)$ takes its maximal value at $D = 0$. On the other end, we see that $R(D)$ reaches its minimal value of 0 at $D_{max} = \min_v \mathbb{E}[d(U, v)]$. If we are willing to accept distortion of D_{max} we can simply encode 0 bits and always decode as v.

Claim. $R(D)$ is convex, i.e. for all $0 \leq \alpha \leq 1$, D_0, D_1, we have that

$$R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1).$$

Proof outline. Consider the “time sharing” scheme for encoding the source symbols $(U_1, \ldots, U_N)$. Take the first αN source symbols and encode them with optimal distortion D_0, and the last $(1-\alpha)N$ source symbols and encode them with optimal distortion D_1. The total expected distortion is $\alpha D_0 + (1-\alpha)D_1$. Then the minimal rate across all schemes is at most the rate for this particular scheme:

$$R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1).$$

4.10.3 Examples

1. Let $U \sim \text{Ber}(p)$ with $p \leq \frac{1}{2}$ and define the Hamming distortion as

$$d(u,v) = \begin{cases} 0; & u = v \\ 1; & u \neq v \end{cases}$$

In this setting U and V take values in $\mathcal{U}$ and $\mathcal{V}$, where $\mathcal{U} = \mathcal{V} = \{0,1\}$. We claim that

$$R(D) = \begin{cases} h_2(p) - h_2(D); & 0 \leq D \leq p \\ 0; & D > p. \end{cases}$$

This function is convex, since in the region $0 \leq D \leq p$, the function takes the value of a constant minus the binary entropy function (which is concave). When $D > p$, we can take the reconstruction to be all zeros.

Conditioning reduces entropy, so we obtain

Proof. Consider the case when $0 \leq D \leq p$. For any U, V such that $U \sim \text{Ber}(p)$ and $\mathbb{E}[d(U,v)] = P(U \neq V) \leq D \leq p \leq 1/2$, consider

$$I(U;V) = H(U) - H(U|V) = H(U) - H(U \oplus_2 V|V)$$

Conditioning reduces entropy, so we obtain

$$\begin{aligned} H(U) - H(U \oplus_2 V|V) &\geq H(U) - H(U \oplus_2 V) \\ &= h_2(p) - h_2(P(U \neq V)) \end{aligned}$$

Equality in the above inequality is achieved when $U \oplus_2 V$ and V are independent.

Since the binary entropy function h_2 is monotonic increasing on the interval $[0, \frac{1}{2}]$, we know that

$$h_2(p) - h_2(P(U \neq V)) \geq h_2(p) - h_2(D).$$

Thus, $I(U;V) \geq h_2(p) - h_2(D)$, implying that

$$\begin{aligned} R(D) &= R^{(I)}(D) \\ &= \min_{\mathbb{E}[d(U,V) \leq D]} I(U;V) \\ &\geq h_2(p) - h_2(D). \end{aligned}$$

To show that equality is achievable, we can demonstrate that the two equality conditions above are satisfied. This is straightforward - essentially we have to find U, V such that

- $U \oplus_2 V$ is independent of V and
- $U \oplus_2 V \sim \mathrm{Ber}(D)$.

□

2. Now, consider $U \sim \mathbb{N}(0, \sigma^2)$. We claim that

$$R(D) = \begin{cases} \frac{1}{2}\log(\sigma^2/D); & 0 < D \leq \sigma^2; \\ 0; \qquad D > \sigma^2. \end{cases}$$

Note that this function is convex, and for allowed distortion D greater than the variance σ^2, we don't need any bits to describe the reconstruction, since it can be taken to be always zero.

Since this is an analog source, the entropy is infinite, so we can't expect to describe it and get zero distortion for a fixed number of bits per source symbol.

We will ccomplete the proof of this result in the next section.

4.11 Method of Types

Notation: Denote $x^n = \{x_1, \dots, x_n\}$ with $x_i \in \mathcal{X} = \{1, \dots, r\}$ and

$$N(a|x^n) = \sum_{i=1}^{n} \mathrm{I}_{\{x_i = a\}}$$
$$\mathrm{P}_{x^n}(a) = \frac{N(a|x^n)}{n}.$$

Definition. The empirical distribution of x^n is the probability vector $(\mathrm{P}_{x^n}(1), \dots, \mathrm{P}_{x^n}(r))$.

Definition. P_n denotes the collection of all empirial distributions of sequences of length n.

Definition. For $\mathrm{P} \in \mathbb{P}_n$, the type class or type of P is $T(P) = \{x^n : \mathrm{P}_{x^n} = P\}$.

Theorem. The number of type classes for sequences of length n, $|\mathbb{P}_n|$, satisfies

$$|\mathbb{P}_n| \leq (n+1)^{r-1}$$

Proof. Every empirical distribution P_{x^n} is determined by a vector $N(1|x^n), N(2|x^n) \dots, N(r-1|x^n)$. This is a vector of length $r-1$, and each element can take up to $n+1$ values. Therefore, there are at most $(n+1)^{r-1}$ possibilities.

Note that for $r \geq 3$ the bound is not tight since we did not include the constraint $\sum_{a=1}^{r-1} N(a|x^n) \geq n$. □

More notation:

- For a probability mass function $Q = \{Q(x)\}_{x \in \mathcal{X}}$, we will write $H(Q)$ to denote $H(X)$ where $X \sim Q$.
- Let $Q(x^n) = \prod_{i=1}^{n} Q(x_i)$. For $S \subset \mathcal{X}^n$, we write $Q(S) = \sum_{x^n \in S} Q(x^n)$.

Theorem. For all x^n, we have $2^{-n[H(P_{x^n})+D(P_{x^n}||Q)}$, where $H(P_{x^n})$ is referred to as the empirical entropy of x^n.

Proof. This is a few straightforward manipulations of definitions.

$$\begin{aligned}
Q(x^n) &= \prod_{i=1}^{n} Q(x_i) \\
&= 2^{\sum_{i=1}^{n} \log Q(x_i)} \\
&= 2^{\sum_{a\in\mathcal{X}} N(a|x^n)\log Q(a)} \\
&= 2^{-n[\sum_{a\in\mathcal{X}} \frac{N(a|x^n)}{n}\log\frac{1}{Q(a)}]} \\
&= 2^{-n\left[\sum_{a\in\mathcal{X}} P_{x^n}(a)\log\left(\frac{1}{Q(a)}\frac{P_{x^n}(a)}{P_{x^n}(a)}\right)\right]} \\
&= 2^{-n[H(P_{x^n})+D(P_{x^n}||Q)]}
\end{aligned}$$

□

Theorem. For all $P \in \mathbb{P}_n$, we have that

$$\frac{1}{(n+1)^{r-1}}2^{nH(P)} \le |T(P)| \le 2^{nH(P)}$$

Proof. Proof is straightforward. □

Fill in later

Theorem. For any probability mass function Q and any empirical distribution $P \in \mathbb{P}_n$,

$$\frac{1}{(n+1)^{r-1}}2^{-nD(P||Q)} \le Q(T(P)) \le 2^{-nD(P||Q)}.$$

That is, on an exponential scale - the probability that the sequence looks like it came from source P if the data is generated iid from distribution Q is very unlikely.

Note that in the expression above, $D(P||Q)$ is between P, the "wrong" source and Q the "true" source. This is different from the cost of mismatch in lossless compression; $D(p||q)$ is such that p is the true source and q is the wrong source.

4.12 Strong, Conditional, and Joint Typicality

Definition. A sequence $x^n \in \mathcal{X}^n$ is strongly δ-typical with respect to a probability mass function $\mathcal{P} \in \mathcal{M}(\mathcal{X})$ if

$$|P_{x^n}(a) - P(a)| \le \delta P(a); \qquad \forall a \in \mathcal{X}.$$

Definition. The strongly δ-typical set of p, $T_\delta(P)$ is defined as the set of all sequences that are strongly δ-typical with respect to P, that is

$$T_\delta(P) = \{x^n : |P_{x^n}(A)\}$$

5

CS109: Probability

graphicx todonotes amsmath amssymb fancyhdr [margin=1.0in]geometry

minted

Var Poi Beta Bin Geo Cov

*arg min *argmax *arg max*

CS 109 — Final Exam Review Adithya Ganesh

5.1 Key Topics

1. Balls and urns
   (a) k distinguishable objects to n distinguishable buckets:
   $$n^k.$$
   (b) k indistinguishable objects to n distinguishable buckets. If each bucket gets a positive number of objects:
   $$\binom{k-1}{n-1}.$$
   (c) If each bucket gets a nonnegative number of objects:
   $$\binom{n-1+k}{n-1}.$$
2. Balls and urns: Ordered vs. unordered set
   (a) Unordered interpretation: k people each get a set of objects
   (b) Ordered interpretation: 1 person gets a series of sets of objects

3. Bayes Theorem
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\neg B)P(\neg B)}.$$
Typically use the second version for computation.

4. Principle of inclusion - exclusion
$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n}(-1)^{k+1}\left(\sum_{1\leq i_1<\cdots<i_k\leq n} |A_{i_1}\cap\cdots\cap A_{i_k}|\right).$$

5. Computing CDF in terms of Φ:
$$P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(Z \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}).$$

6. Expectation properties

   (a) Definition
$$\mathbb{E}[X] = \sum_x x p_X(x).$$
$$\mathbb{E}[X] = \int_x x p(x) dx.$$
   More generally, you can compute
$$\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x)p(x)dx.$$

   (b) Linearity
$$\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)].$$

7. Variance properties

   (a) Definition
$$(X) = \mathbb{E}[(X-\mu)^2]$$

   (b) Key identity
$$(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

   (c) Linear combinations
$$(aX+b) = a^2(X)$$

   (d) Sums
$$(X+Y) = (X) + (Y) + 2(X,Y).$$

   (e) Standard deviation
$$\text{SD}(X) = \sqrt{(X)}.$$

8. Covariance properties

   (a) Definition
   $$(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

   (b) Sum of variance
   $$(X + Y) = (X) + (Y) + 2(X, Y).$$

   (c) If X, Y independent, then $(X, Y) = 0$.

   (d) If X, Y independent, then
   $$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

9. Correlation of X and Y:
   $$\rho(X, Y) = \frac{(X, Y)}{\sqrt{(X)(Y)}}$$

10. Key distributions

    Discrete:

    (a) $X \sim \mathit{Bernoulli}(p), 0 \leq p \leq 1$. 1 if coin with heads probability p comes up heads, zero otherwise.
    $$p(x) = \begin{cases} p; & x = 1; \\ 1 - p; & x = 0. \end{cases}$$
    $$\mathbb{E}[X] = p; \qquad (X) = p(1 - p).$$

    (b) $X \sim \mathit{Binomial}(n, p)$, $0 \leq p \leq 1$. The number of heads in n independent flips of a coinw ith heads probability p.
    $$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$
    $$\mathbb{E}[X] = np; \qquad (X) = np(1 - p).$$

    (c) $X \sim \mathit{Geometric}(p), p > 0$. The number of flips of a coin with heads probability p until the first heads.
    $$p(x) = p(1 - p)^{x-1}.$$
    $$\mathbb{E}[X] = \frac{1}{p}; \qquad (X) = \frac{1 - p}{p^2}.$$

    (d) $X \sim \mathit{Poisson}(\lambda)$, $\lambda > 0$. A probability distribution over the nonnegative integers used for the modeling the frequency of rare events.
    $$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}.$$
    $$\mathbb{E}[X] = \lambda; \qquad (X) = \lambda.$$

Intuition: let $n \to \infty, p \to 0$, and let $np = \lambda$ stay constant. Binomial distribution will converge to this density function.

The binomial in the limit, with $\lambda = np$, when n is large, p is small, and λ is "moderate"

Let X be binomial. Then if $p = \lambda/n$, we obtain

$$P(X = i) = \frac{n!}{i!(n-i)!}p^i(1-p)^{n-i} = \frac{n!}{i!(n-i)!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}$$

$$= \frac{n(n-1)\dots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}.$$

When n is large, p is small, and λ is moderate, we obtain

$$\frac{n(n-1)\dots(n-i+1)}{n^i} \approx 1; \qquad (1-\lambda/n)^n \approx e^{-\lambda}; \qquad (1-\lambda/n)^i \approx 1.$$

Recall that the definition of e is

$$e = \lim_{n\to\infty}(1+1/n)^n.$$

Understand how this derivation works with the exponential term

It follows that

$$P(X = i) \approx \frac{\lambda^i}{i!}e^{-\lambda}.$$

Continuous:

(a) $X \sim Uniform(a, b)$, $a < b$. Equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a}; & a \le x \le b \\ 0; & \text{else.} \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2}; \qquad (X) = \frac{(b-a)^2}{12}.$$

(b) $X \sim Exponential(\lambda)$, $\lambda > 0$. Decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}; & x \ge 0 \\ 0; & \text{else.} \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}; \qquad (X) = \frac{1}{\lambda^2}.$$

(c) $X \sim Normal(\mu, \sigma^2)$. Gaussian distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

$$\mathbb{E}[X] = \mu; \qquad (X) = \sigma^2.$$

11. Moment generating function (MGF) of X:

$$M(t) = \mathbb{E}[e^{tX}].$$

Intuition: uniquely determines the distribution. Can differentiate to compute useful quantities

12. Joint MGF of $X_1, X_2, \ldots, X_n$:

$$M(t_1, t_2, \ldots, t_n) = \mathbb{E}[e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}]$$

13. Markov's inequality. Let X be non-negative RV:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}; \qquad \text{for all } a > 0.$$

Proof - indicator random variables.

14. Chebyshev's Inequality. Let X be an RV with $\mathbb{E}[X] = \mu$, $(X) = \sigma^2$. Then

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}; \qquad \text{for all } k > 0.$$

Proof, apply Markov's Inequality with $a = k^2$.

15. One-sided Chebyshev's Inequality. Let X be an RV with $\mathbb{E}[X] = 0$, $(X) = \sigma^2$. Then

$$P(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Proof. Note that $P(X \geq a) = P(X+b \geq a+b)$, and apply Markov's inequality. Minimize the resulting quadratic as a function of offset b.

Or, if $\mathbb{E}[Y] = \mu$, and $(Y) = \sigma^2$, we obtain

$$P(Y \geq \mathbb{E}[Y] + a) \leq \frac{\sigma^2}{\sigma^2 + a^2}; \qquad \text{for any } a > 0$$

$$P(Y \leq \mathbb{E}[Y] - a) \leq \frac{\sigma^2}{\sigma^2 + a^2} \qquad \text{for any } a > 0.$$

16. Chernoff bound. Let $M(t)$ be an MGF of RV X. Then

$$P(X \geq a) \leq e^{-ta} M(t); \qquad \text{for all } t > 0.$$

$$P(X \leq a) \leq e^{-ta} M(t); \qquad \text{for all } t < 0.$$

Bounds hold for $t \neq 0$, so use t that minimizes $e^{-ta} M(t)$ (i.e. makes bound strictest).

Proof: $P(X \geq a) = P(e^{tX} \geq e^{ta})$, and then apply Markov's inequality.

17. Jensen's Inequality. If $f(x)$ is convex, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Equality when $f''(x) = 0$. Proof: Taylor series of $f(x)$ about μ.

18. Law of Large Numbers. Consider I.I.D. random variables $X_1, X_2, \ldots$. Suppose $\mathbb{E}[X_i] = \mu$ and $(X_i) = \sigma^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. For any $\epsilon > 0$:

    $$P(|\overline{X} - \mu| \geq \epsilon) \longrightarrow 0.$$

    Proof: Apply Chebyshev's inequality on $\overline{X}$. $\mathbb{E}[\overline{X}] = \mu$, $(\overline{X}) = \frac{\sigma^2}{n}$.

    $$P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0.$$

19. Strong Law of Large Numbers. Consider I.I.D. random variables $X_1, X_2, \ldots$. Suppose X_i has distribution F with $\mathbb{E}[X_i] = \mu$.

    Then
    $$P\left(\lim_{n\to\infty}\left[\frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right]\right) = 1.$$

20. Central Limit Theorem (CLT). Consider I.I.D. random variables $X_1, X_2, \ldots$. Suppose $\mathbb{E}[X_i] = \mu$, and $(X_i) = \sigma^2$. Then
    $$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \longrightarrow \mathcal{N}(0,1); \qquad \text{as } n \longrightarrow \infty.$$

    Intuition – the $n\mu$ is for mean normalization, the $\sigma\sqrt{n}$ is for variance normalization. This is why many real world distributions look normally distributed.

21. Method of moments. Let $\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ (sample moments). Set each of these sample moments equal to the ”true” moments.

22. Estimator Bias. Defined as
    $$\mathbb{E}[\hat{\theta}] - \theta.$$

    When bias = 0, estimator is unbiased.

23. Estimator Consistency. Defined as

    $$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1; \qquad \text{for } \epsilon > 0.$$

24. Maximum Likelihood Estimation. Define the likelihood function as

    $$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta),$$

    where this is a product since the X_i are IID. Then

    $$\theta_{MLE} = \text{argmax}_\theta\, L(\theta).$$

25. Log-likelihood
    $$LL(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$

26. Bayesian Estimation. Let θ = model parameters, D = data. Then

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}.$$

We have prior $P(\theta)$ and can compute likelihood $P(D|\theta)$. Posterior $P(\theta|D)$ is assumed to have same parameter form as prior. The term $P(D)$ is a constant that can be ignored (just for integration).

Example: Let $\theta \sim (a, b)$, $D = \{n \text{ heads}, m \text{ tails}\}$. Then maximum a posteriori will give you $(a+n, b+m)$.

27. Maximum A Posteriori (MAP) estimator of θ:

$$\begin{aligned}\theta_{MAP} &= \text{argmax}_\theta\, f(\theta|X_1, X_2, \ldots, X_n) = \text{argmax}_\theta \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots, X_n)} \\ &= \text{argmax}_\theta \frac{\left(\prod_{i=1}^n f(X_i|\theta)\right) g(\theta)}{h(X_1, X_2, \ldots, X_n)} = \text{argmax}_\theta\, g(\theta) \prod_{i=1}^n f(X_i|\theta).\end{aligned}$$

28. Log a posteriori

$$\theta_{MAP} = \text{argmax}_\theta \left(\log(g(\theta)) + \sum_{i=1}^n \log(f(X_i|\theta)) \right).$$

29. Naive Bayes. Estimate probabilities $P(Y)$ and each $P(X_i|Y)$ for all i. Classify as spam or not using $\hat{Y} = \text{argmax}_y\, \hat{P}(\text{X}|Y)\hat{P}(Y)$.

Employ conditional independence assumption:

$$\hat{P}(\text{X}|Y) = \prod_{i=1}^m \hat{P}(X_i|Y).$$

30. Laplace estimate, Naive Bayes.

$$P(X_i = 1|Y = \text{spam}) = \frac{(\text{ spam emails with word } i) + 1}{\text{total spam emails} + 2}.$$

31. Logistic regression. Learn weights β_i to estimate

$$P(Y = 1|\text{X}) = \frac{1}{1+e^{-z}}; \qquad z = \beta^T x.$$

Learn weights β_i from gradient descent.

32. Linear congruential generator. Start with seed number X_0. Next random number is given by

$$X_{n+1} = (aX_n + c) \pmod m.$$

33. Bayesian network. Graphical representation of joint probability distribution. Each node X has a conditional probability $P(X|parents(X))$. Graph has no cycles (directed acylic graph).

34. Showing two distributions are independent. If

$$P(x, y) = P(x)P(y); \qquad \forall x, y.$$

then the two random variables are independent.

5.2 Theory

1.

5.3 Problems to Review

5.3.1 Problem Set 1

1. Classical combinatorics.
2. Balls and urns, and variations.
3. 1.13 - Unordered vs. ordered ways of counting a set for probability.

5.3.2 Problem Set 2

1. Basic applications of Bayes' Theorem.
2. Principle of inclusion - exclusion.
3. Classical combinatorics.

5.3.3 Problem Set 3

1. Infinite summations to compute expectation (typically, arithmetico-geomtric series).
2. CDF of normal in terms of Φ.
3. Binary random variable + sum of expectations.

5.3.4 Problem Set 4

1. Multiple integrals of a density function
2. Independence of two distributions + joint density

5.3.5 Problem Set 5

1. Recursive expectation calculation
2. MGF calculation

5.3.6 Problem Set 6

1. 6.1 - Confidence intervals
2. 6.2 - Maximum likelihood estimation + Jensen for bias

5.4 Practice Problems 3-20-17

1. (See notebook), went through all problems from final review document.

2. PS5.1(a)

3. PS5.4 – the relationship between independence, correlation, and covariance

4. PS5.8, using MGFs to obtain the distribution

5. PS3.3(a)

6. Problem from midterm that uses infinite series

5.5 Practice Final 3-21-17

1. Remember to take $\sqrt{\sigma^2}$ when doing Φ transformation.

2. Review continuity correction

6

MATH53: Differential Equations

adi amsmath

Ker Im

MATH 53 Notes Adithya Ganesh
Lectures: Andrea Ottolini

Contents

6.1 Lecture 1: 6-24-19

Instructor: Andreas Ottolini

Office: 380-381D

Office hours: M-W, 3 - 4pm.

CAs: Evangelie Zachos (OH Tues and Thurs, 3-5pm OH, 380-380M)

Midterm: July 19. Final: August 16.

7 homeworks; lowest score is dropped.

This class is about ordinary differential equations and linear algebra. In some sense, the main part of this course is linear algebra.

What is an ODE? ODE stands for ordinary differential equation, where the unknown is a function. We're dealing with ordinary differential equations, implying that we're dealing with one independent variable.

Example. Let $u(t)$ denote the temperature of an object a ttime t, and k denote the conductivity of the material. Let T be the ambient temperature. Note that there's a lot of simplification, since we're assuming the object uniformly has the same temperature.

Note that Fourier's law in it's difference equation form states that

$$u(t+h) - u(t) = hk(T - u(t)).$$

Intuitively this makes since, since if the ambient temperature is greater than the object temperature, the object's temperature goes up.

If we divide by h and take the limit as $h \longrightarrow 0$, we obtain a differential equation. Indeed, we obtain

$$u'(t) = k(T - u(t)).$$

For the initial condition, we need to know the initial temperature.

Example. Let $x(t)$ denote the position of a spring, whose other end is fixed at 0. Let m denote the mass of the spring, and k denote the elastic constant of the spring.

Hooke's law states that

$$F(t) = -kx(t).$$

Applying Newton's second law, we can rewrite this as

$$mx''(t) = -kx(t).$$

Note that the equation itself isn't enough to guarantee a unique solution; we need to add an initial condition. In this case, we would want to specify $x(0) = x_0, x'(0) = x'_0$.

In general, we would like to know:

- Does a solution exist?
- Is it unique?
- What is the dependence on initial condition?

Formally, now, what is an ordinary differential equation? We are given a function F with the following dependence:

$$F(t, u, u^{(1)}, u^{(2)}, \dots, u^{(n)}) = 0.$$

Finding a solution means finding $u = u(t)$ such that when you plug into $F(t, u(t), \dots, u^{(n)}(t)) = 0$ for all $t \in [0, T]$.

Now for some terminology:

- n is the order of the equation; i.e. the highest derivative that appears in the equation.
- If F does not depend explicitly on t, the system is called autonomous. That is, the law that describes the phenomenon is not dependent on time.
- An n-th order ODE is called linear if

  $$F(t, u, u^{(1)}, u^{(2)}, \dots, u^{(n)}) = a_n(t)u^{(n)} + a_{n-1}(t)u^{(n-1)} + \cdots + a_1(t)u' + a_0(t).$$

  In particular, if an ODE is linear and $a_0(t) = 0$ and $u = u_1+u_2$, then $F(t, u, \dots, u^{(n)}) = F(t, u_1, \dots, u_1^{(n)})+ F(t, u_2, \dots, u_2^{(n)})$.
- An ODE is called linear and homogenous if $a_0(t) = 0$.
- A constant coefficient linear equation is known as a linear, autonomous ODE.

Sidenote: in the above setting, it's possible for the $u^{(i)}$ to be vectors.

Example. We'll consider an example from population dynamics.

Let $p(t)$ denote the population at time t. The population is governed by the following law:

$$p'(t) = rp(t).$$

We can also assume that there's some rate of death (maybe there are wolves that eat rabbits, so the equation becomes

$$p'(t) = rp(t) - a.$$

Let T be the maximum capacity of the environment. Then another formulation is

$$p'(t) = r\left(1 - \frac{p(t)}{T}\right)p(t)$$

Example. Suppose $p(t)$ denotes the prey population, and $q(t)$ denotes the predator population. The Lotko-Volterra equations describe the population of both:

$$\begin{aligned} p'(t) &= \alpha p(t) - \beta p(t)q(t) \\ q'(t) &= -\gamma q(t) + \delta p(t)q(t). \end{aligned}$$

This example is notable because it's the first case where we are studying a system of two differential equations.

Now we'll look at how to solve some easy cases. We'll start with Example 1, say $k = 1, T = 0$. Then we can write

$$u'(t) = -u(t).$$

Intuitively, it looks like $u(t) = e^{-t}$ is a solution. We can use a trick by multiplying both sides by e^t:

If we multiply by e^t, we obtain

$$u'(t)e^t + u(t)e^t = (u(t)e^t)' = 0.$$

Then integrating, we obtain

$$u(t)e^t = C,$$

and in particular, all solutions are $u(t) = Ce^{-t}$. An intuitively this makes sense. Over time, the temperature will decay to 0.

Note that if the start equation is $u'(t) = T - u(t)$, the solution will be $u(t) = ae^{-t} + T$, which makes sense.

For a while, we're going to deal with equations of the form

$$u'(t) = G(t, u).$$

We say that a number u_0 is an equilibrium if $G(t, u_0) = 0$. Because then $u(t) = u_0$ is a solution to the ODE. It's called an equilibrium because it doesn't depend on time.

For example, in the case $G(t, u) = k(T - u)$, an equilibrium is T. Intuitively, if you can analyze the equilibria, you can often draw asymptotic conclusions about solution behavior.

6.2 Lecture 2: 6-25-19

Recall that we define an ODE to be an equation of the form

$$F(t, u, u^{(1)}, u^{(2)}, \ldots, u^{(n)}) = 0.$$

Herein, n is the order of the equation.

Recall that a first order equation is of the form

$$\begin{aligned} F(t, u, u') &= 0 \\ u' &= G(t, u). \end{aligned}$$

Recall that an equilibrium is a number u_0 such that $G(t, u_0) = 0$ for all t. And in particular, this implies that $u(t) = u_0$ is a solution to $u' = G(t, u)$. Often, the first step to understanding an ODE is to understand the equilibrium solution.

Example. Consider the example

$$G(t, u) = -u,$$

which is a special case of $u' = k(T - u)$. Recall that the solution is $u(t) = ce^{-t}$, and $u_0 = 0$ is an equilibrium solution.

For today, we want to identify a heuristic procedure to identify the trajectory of the solution, without doing any work.

Definition. We introduce the notion of a direction field. We don't know $u(t)$, but we know at each point what the derivvative should be.

A direction field is an assignment of a vector to each point of the (t, u) plane, where the vector has slope $G(t, u)$.

Exercise. Draw the direction field corresponding to $G(t, u) = -u$. Note that this will show that the direction field slopes downward to $u = 0$.

Definition. An integral curve is a curve $u = u(t)$ such that $u(t)$ is tangent to the direction field at each time t. Integral curves are solutions.

Example. Consider the example

$$u' = 2u - 3.$$

We can fairly easily plot the direction field in this case. Since the system is autonomous, it suffices to analyze the direction field on a single slice $t = k$.

Note that clearly the equilibrium is at $u = \frac{3}{2}$. Note that $G(t, u) > 0$ implies $u > \frac{3}{2}$, and $G(t, u) < 0$ implies that $u < \frac{3}{2}$.

Problem in section. Consider the problem $u' = -2u + 5$.

The equilibrium is $u = \frac{5}{2}$. The direction field will go down above $\frac{5}{2}$. The direction field will go up below $\frac{5}{2}$.

Note that for an order n equation, you need n initial conditions to uniquely determine the solution.

This is a stable equilibrium, since asymptotically, we will approach the equilibrium.

Note that in general you can't have two solutions that cross (typically).

Example. Consider the problem

$$u' = 2u(3 - u).$$

Note that we can first draw the quadratic, which is an inverted parabola with vertices at $u = 0$ and $u = 3$. In particular, the direction field will have two equilibria at $u = 0$ and $u = 3$. Interesting, $u = 3$ is a stable equilibrium, but $u = 0$ is an unstable equilibrium.

Problem in section. Consider the problem $u' = u^2(1 - u)$. This is a cubic function with vertices at $u = 0$ and $u = 1$.

The direction field looks something like

$$-$$

$$0@u = 1$$

$$+$$

$$0@u = 0$$

$$+$$

Interestingly, the direction field looks as follows:

$$IMG1$$

We can show this by solving it explicitly. Indeed, we'll apply the following lemma.

Lemma. If u_1, u_2 are two solutions of our ODE, then $u_1 - u_2$ solves the associated homogeneous problem.

This is similar to what we do in linear algebra. You find one particular solution of the system, and then you can obtain the general solution from there.

Proof. By hypothesis, $u_1' = t + 2u_1$, $u_2' = t + 2u_2$. Then $v' = u_1' - u_2' = 2u_1 - 2u_2 = 2v$, as claimed.

Example. Alright, we will now solve the problem $u' = t + 2u$. We start by looking for a solution $At + B$. Substituting, we obtain

$$A = t + 2(At + B),$$

that is

$$t(1 + 2A) + 2B - A = 0,$$

so that solving $1 + 2A = 0$ and $2B - A = 0$ gives the solution $-\frac{1}{2}t - \frac{1}{4}$.

By the lemma, since the solution of $v' = 2v$ are ce^{2t}, then, every solution of $u' = t + 2u$ is equal to $u(t) = -\frac{1}{2}t - \frac{1}{4} + ce^{2t}$.

So there is an equilibrium at $u = \frac{1}{4}$.

6.3 Lecture 3: 6-26-19

Today we'll discuss integrating factors. Recall that when we approach the problem

$$y'(t) = T - y(t),$$

we obtain $y(t) = T + ce^{-t}$. Note that the idea is to bring the $y(t)$ terms to one side tand multiply by e^t, and apply the product rule.

Suppose we're given an equation $y'(t) + p(t)y(t) = q(t)$, where p and q are continuous in some interval I. We want to multiply both sides by $\mu(t)$ so that the LHS is of the form $(y(t)\mu(t))'$.

In particular, we need

$$\mu(t)y'(t) + \mu(t)p(t)y(t) = \mu(t)q(t),$$

and we require $\mu'(t) = \mu(t)p(t)$. If this relation holds, then

$$(y(t)\mu(t))' = \mu(t)q(t),$$

so that

$$y(t)\mu(t) = C + \int_0^t \mu(s)q(s)\,ds.$$

In particular, solving for $y(t)$, we obtain the explicit solution

$$y(t) = [\mu(t)]^{-1}\left[C + \int_0^t \mu(s)q(s)\,ds\right]$$

This is satisfactory, but it requires that we be able to

- Solve the integral.

- $\mu(t)$ can't be 0, so that all the steps are invertible.

Now, if $\mu'(t) = p(t)\mu(t)$, if $\mu \neq 0$, we obtain

$$\frac{\mu'(t)}{\mu(t)} = p(t),$$

that is

$$(\ln \mu(t))' = p(t).$$

Integrating, we obtain

$$\mu(t) = c \exp\left(\int_0^t p(s)\, ds\right).$$

We can take $c = 1$, to obtain $\mu(t) = \exp\left(\int_0^t p(s)\, ds\right)$, and in particular

$$y(t) = \exp\left(-\int_0^t p(s)\, ds\right)\left[C + \int_0^t \exp\left(\int_0^t p(s)\, ds\right) q(s)\, ds\right].$$

Let's verify that this works with our previous examples.

Example 1. Consider the problem $y'(t) = T - y(t)$, so that $p(t) = 1, q(t) = T$. Then the integrating factor is $\mu(t) = e^t$, and also

$$\begin{aligned} y(t) &= \exp(-t)\left[C + \int_0^t Te^t\right] \\ &= \exp(-t)\left[C + Te^t\right] \\ &= T + Ce^{-t}, \end{aligned}$$

as claimed.

Example 2. Consider the problem $y'(t) = \sin t - y(t)$.

We can rewrite this in the form

$$y'(t) + y(t) = \sin t.$$

In this setting, $p(t) = 1$, and $q(t) = \sin t$. The integrating factor is $\mu(t) = e^t$. We obtain

$$y'(t)e^t + y(t)e^t = e^t \sin t.$$

Integrating by parts, we obtain

$$\begin{aligned} y(t)e^t &= \int_0^t e^s \sin s \, ds \\ &= C + \frac{e^t}{2} \left(\sin t - \cos t\right) \end{aligned}$$

In particular,

$$y(t) = Ce^{-t} + \frac{1}{2} \left(\sin t - \cos t\right).$$

Example 3. Let's consider the problem

$$u'(t) = \frac{-u(t)}{t+1}.$$

In this case, $p(t) = \frac{1}{t+1}$, and $q(t) = 0$. We obtain $\mu(t) = e^{\int_0^t \frac{1}{s+1} ds} = t + 1$.

Now,

$$u'(t)(t+1) + u(t) = 0,$$

so we obtain

$$(u(t)(t+1))' = 0,$$

and in particular

$$u(t) = \frac{c}{t+1}.$$

Example 4. Consider the problem

$$t^3 y' + 3t^2 y = e^t.$$

Note that it's useful to remember the proof, we can obtain the product rule immediately. Instead of applying the formula, we just obtain

$$(t^3 y)' = e^t,$$

and we are done.

We'll now discuss existence and uniqueness of solutions to differential equations. Suppose p, q are continuous in some interval I. Consider the initial value problem

$$\begin{aligned} y'(t) + p(t)y(t) &= q(t) \\ y(t_0) &= y_0, \end{aligned}$$

for some $t_0 \in I, y_0 \in \mathbb{R}$.

Intuitively, we expect that the solution should exist and be unique (since this is a first order equation, and we have a single parameter).

We know that all solutions to this equation are of the form

$$y(t) = (\mu(t))^{-1} \left[C + \int_0^t \mu(s)q(s)\, ds \right].$$

To determine C, it suffices to just plug t_0 into the above equation, to obtain

$$y_0 = y(t) = (\mu(t_0))^{-1} \left[C + \int_0^{t_0} \mu(s)p(s)\, ds \right].$$

In C, the function on the right hand side is just a line. Clearly, whatever value of y_0 we take, there is a unique intersection with the line.

Example 4. Suppose we have an equation

$$(t^2 - 1)y' + \frac{y}{t} = \ln(t+3) - \arctan t.$$

For which values of t_0, y_0 can we ensure that the IVP associated to the ODE has a unique solution?

We don't need to solve this problem. We simply need to ascertain which values of y and t the function above is continuous.

$$y' + \frac{y}{(t^2-1)t} = \frac{\ln(t+3) - \arctan t}{t^2 - 1},$$

If $t_0 > -3, t_0 \neq -1, 0, 1$, then we can solve the IVP for all choices y_0.

Example 5. Consider the problem

$$y' + py = 0, p \in \mathbb{R}.$$

From the formula before, the solution is $y(t) = ce^{-pt}$. Another approach - sometimes we can attempt to guess at a solution and try to see whether it works.

We can look for the solution of the form $e^{\lambda t}$, check if there is a solution, and go from there. Substituing this into the equation, we can obtain

$$\lambda e^{\lambda t} + pe^{\lambda t} = 0,$$

which implies $\lambda + p = 0$, and thus $\lambda = -p$.

Example 6. We can consider a more general setting, and motivate the use of complex numbers to solve ODEs. Suppose we have the problem

$$y'' + ay' + by = 0,$$

where a and b are real numbers. We start by looking for a solution of the type $e^{\lambda t}$. Substituting, we obtain

$$\lambda^2 e^{\lambda t} + a\lambda e^{\lambda t} + be^{\lambda t} = 0.$$

Dividing by $e^{\lambda t}$, we need $\lambda^2 + a\lambda + b = 0$. If the solutions are real and distinct, λ_1, λ_2, then $e^{\lambda_1 t}, e^{\lambda_2 t}$ are solutions. If they are complex, things are trickier, and we need to develop a theory of complex numbers.

We'll do a review of complex numbers.

Recall that a complex number $z = a + bi$, with $a, b \in \mathbb{R}$. Recall the standard rules for adding and multiplying complex numbers.

If $z_1 = a_1 + b_1 i$, $z_2 = a_2 + b_2 i$, then

$$z_1 + z_2 = (a_1 + a_2) + (b_1 + b_2)i,$$

and

$$z_1 z_2 = (a_1 a_2 - b_1 b_2) + (a_1 b_2 + a_2 b_1)i$$

Recall the geometric interpretation of adding and multiplying two complex numbers.

- Adding two complex numbers: geometrically analogous to vector addition between two complex numbers.
- Multiplying two complex numbers: geometrically analogous to rotating the first complex number by the second; and multiplying the magnitudes.

Let's try to define the exponential e^z. Recall that

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

So we define

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Suppose we plug in $z = ix$, for $x \in \mathbb{R}$. Recall $i^0 = 1$, $i^2 = -1$, $i^3 = -i$, $i^4 = 1$. Givne this,

$$\begin{aligned} e^{ix} &= \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{i^n x^n}{n!} \\ &= 1 + ix - \frac{x^2}{2} - i\frac{x^3}{6} \ldots, \end{aligned}$$

and if you compare this to the power series of sin and cos, you obtain

$$e^{ix} = \cos x + i \sin x.$$

If z is a complex number, $z = a+ib$, we can write $z = \sqrt{a^2+b^2}\left(\frac{a}{\sqrt{a^2+b^2}} + i\frac{b}{\sqrt{a^2+b^2}}\right)$, namely $|z|\left(\cos\theta + i\sin\theta\right) = |z|e^{i\theta}$, which is the polar representation of complex numbers.

Tomorrow, we'll see what happens when we try to construct the logarithm of a complex function. There is a problem - the logarithm is the inverse of the exponential function. But typically, we need the function to be injective to talk about an inverse. Notice that e^{ix} is not injective, since we can replace x with $x + 2\pi$ and obtain the same result.

6.4 Lecture 4: 6-27-19

Let's continue the discussion from before. There's an idea known as the branch cut of the logarithm.

Recall that we define for $z \in \mathbb{C}$, we define

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

If $z \in \mathbb{R}$, we obtain the usual exponential. If $z = ix$, we obtain $e^{ix} = \cos x + i \sin x$. Notice that the map $x \mapsto e^x$ is injective on $\mathbb{R}$.[1]

Using injectivity, we can define a log function as the inverse of the exponential. In particular, saying that $y = \ln x$ is equivalent as saying $e^y = x$.

Problem: for $z \in \mathbb{C}$, the function $z \mapsto e^z$ is not injective. Recall that $e^0 = e^{2\pi i}$, and in general $e^z = e^{z+2\pi i}$.

Since if $z = x + iy$, for $x, y \in \mathbb{R}$, we have

$$\begin{aligned} e^z &= e^{x+iy} \\ &= e^x e^{iy} \\ &= e^x(\cos y + i \sin y). \end{aligned}$$

[1] Recall that $f(x)$ is injective if $f(x) = f(y)$ implies that $x = y$.

When we invert, we have to choose which range of 2π we are using when we are inverting this function. This is exactly the same problem we have when we define $\arcsin$. Note that the square root function actually has the same problem.

A branch cut is a specification of a range $[k, k + 2\pi)$ such that $z \mapsto e^z$ is injective.

Example. What is $\log z$ with branch cut $[0, 2\pi)$. If $z = re^{i\theta}$, then we define

$$\log z = \log(re^{i\theta}) = \log r + i\theta.$$

For example, if $z = -1$, we have $\log z = \log 1 + i\pi = i\pi$.

Example. If $z = i$, then $\log z = \log 1 + i\frac{\pi}{2} = i\frac{\pi}{2}$.

Example. If the branch cut is $[-\pi, \pi)$, we can write $\log -1 = -i\pi$.

Example. What about $z_1^{z_2}$? We can reduce this to computing a log, as follows:

$$\begin{aligned} z_1^{z_2} &= (e^{\log z_1})^{z_2} \\ &= e^{z_2 \log z_1}. \end{aligned}$$

Note that this has the same problem as before - we need to choose a branch cut.

This is why yesterday we looked at

$$1 = 1^i = (e^{0 \cdot i})^i = (e^{2\pi i})^i = e^{-2\pi}.$$

For example, take $z_1 = 1, z_2 = i$, and pick a branch cut to be $[0, 2\pi)$. We obtain

$$\begin{aligned} 1^i &= e^{i \log 1} \\ &= e^{i \log(e^0)} = 1. \end{aligned}$$

But if the branch cut is $[\pi, 3\pi)$, we get

$$\begin{aligned} 1^i &= e^{i \log 1} \\ &= e^{i 2\pi i} \\ &= e^{-2\pi} \end{aligned}$$

Interesting example - you have to be careful when you apply logarithmic identities. For instnace, we might write

$$\begin{aligned} 0 &= \log 1 \\ &= \log((-1)^2) \\ &= \log(-1) + \log(-1) \\ &= 2i\pi. \end{aligned}$$

Example. Consider 4^{i+1} under $[0, 2\pi)$, and under $(-\pi, \pi]$.

Under $[0, 2\pi)$, we have

$$\begin{aligned} 4^{i+1} &= \left(e^{\log 4}\right)^{i+1} \\ &= e^{(i+1)\log 4} \\ &= e^{i\log 4 + \log 4} \\ &= e^{\log 3} \cdot e^{i \log 4} \\ &= 4 \cdot [\cos(\log 4) + i \sin(\log 4)]. \end{aligned}$$

Note that we have to be careful when taking $\log x$ for real x, since sometimes we need to be careful with the branch cuts. Explicitly, we need to write $\ln 4 = \ln 4 + i \cdot 0$.

If we choose $[2\pi, 4\pi)$ as the branch cut, we obtain

$$\begin{aligned} 4^{i+1} &= \left(e^{\log 4}\right)^{i+1} \\ &= \left(e^{\ln 4 + i2\pi}\right)^{i+1} \\ &= e^{\ln 4 - 2\pi} + i(\ln 4 + 2\pi) \\ &= e^{\ln 4 - 2\pi} \left(\cos(\ln 4) + i \sin(\ln 4)\right). \end{aligned}$$

This covers essentially the whole content of homework 1, which consists of:

- Direction fields
- Integrating factors
- Complex numbers
- Existence and uniqueness

Example. Consider the differential equation $ty' - y = t^2 e^{-t}$. Find the general solution.

If we associate this with $y(t_0) = y_0$, which value of t_0 we can guarantee a solution? We first need to put this in canonical form, to obtain

$$y' - \frac{y}{t} = te^{-t},$$

so if $t \neq 0$, we are good.

In this case, we want to multiply by the factor $\mu(t) = \exp\left(\int_1^t -\frac{1}{s}\, ds\right)$, that is $\frac{1}{t}$.

We obtain

$$y'\frac{1}{t} - \frac{y}{t^2} = e^{-t},$$

So,

$$\left(\frac{y}{t}\right)' = e^{-t},$$

that is

$$\frac{y}{t} = -e^{-t} + C,$$

so

$$y = -te^{-t} + tC$$

Example. Consider $ty' + (t+1)y = t$, where $y(\ln 2) = 1$, for $t > 0$.

Find the solution.

We start by dividing out - we get

$$y' + \frac{(t+1)}{t}y = 1,$$

and in particular the integrating factor is $e^{\int_0^t \frac{s+1}{s}\, ds} = \exp\left(\int 1 + \frac{1}{s}\right) = \exp\left(t + \ln t\right)$.

[2]

Ok, then we get

$$y' \exp(t + \ln t) + \left(1 + \frac{1}{t}\right) y \exp(t + \ln t) = \exp(t + \ln t),$$

so that

$$(y \exp(t + \ln t))' = \exp(t + \ln t).$$

Integrating,

$$y \exp(t + \ln t) = e^t(t-1) + C,$$

so that

$$y = \frac{e^t(t-1) + C}{te^t},$$

and

$$y = 1 - \frac{1}{t} + \frac{C}{te^t}.$$

Finally, to obtain the solution, note that at $\ln 2$, we obtain

$$y = 1 - \frac{1}{\ln 2} + \frac{C}{\ln 2 \cdot 2} = 1,$$

[2] Note that we can simplify this as $e^{t+\ln t} = e^t t$.

so that $C = 2$. So the final solution is

$$y = 1 - \frac{1}{t} + \frac{2}{te^t}.$$

Example. Consider the ODE $y' + \frac{2y}{t} = \frac{\cos t}{t^2}$, where $y(\pi) = 0$.

The integrating factor is $\exp\left(\int_0^t \frac{2}{s}\, ds\right)$, that is $e^{2\ln t}$, which is t^2.

Now,

$$t^2 y' + 2ty = \cos t,$$

Integrating,

$$t^2 y = \sin t + C,$$

that is

$$y = \frac{\sin t}{t^2} + \frac{C}{t^2},$$

and we require $C = 0$, so

$$y = \frac{\sin t}{t^2}.$$

6.5 Lecture 5: 6-28-19

To summarize things we should know:

- Direction fields for equations of the form $y' = G(t, y)$.
- Integrating factors, to solve equations like $y' + p(t)y = q(t)$. Also, to solve IVPs given $y(t_0) = y_0$, and further, to understand existence and uniqueness.

  For example: consider the problem

  $$y' + \frac{1}{1+t}y = \frac{1}{t-3},$$

  with $y(0) = 2$. Find, without solving the equation, an interval in which the IVP can be solved.

  The maximal interval in which the IVP can be solved is $(-1, 3)$,
- Complex numbers
  - Understand algebraic representation vs. polar representation.
  - Exponential in $\mathbb{C}$.
  - Definition of $\log z$, $z_1^z 2$, and the idea of a branch cut.

Example. Let's say we have the equation $y'' = -y$, with the initial conditions $y(0) = 1, y'(0) = 0$.

Recall that for a second order ODE, we should expect that we need 2 iniital conditions to guarantee uniqueness.

To solve this, we should guess a solution of the form $e^{\lambda t}$, and try to see if we can find a solution that works.

In particular, we will obtain

$$\lambda^2 e^{\lambda t} = -e^{\lambda t},$$

which implies $\lambda^2 + 1 = 0$, and in particular $\lambda = \pm i$.

Formally, it seems that e^{it} and e^{-it} are solutions. But ideally, we want these solutions to be real. There are a couple of lemmas we can use to obtain a real solution.

Lemma 1. For a homogeneous linear equation of any order, if $y_1(t)$ and $y_2(t)$ are two solutions, then $\alpha y_1(t) + \beta y_2(t)$ is also a solution for complex α, β.

Proof. Just use the fact that the derivative is linear.

Now, by the lemma, if we consider $y(t) = c_1 e^{it} + c_2 e^{-it}$, this is a solution for all choices of $c_1, c_2 \in \mathbb{C}$.

Applying Euler's formula, note that

$$y(t) = c_1(\cos t + i \sin t) + c_2(\cos t + i \sin t).$$

A good choice of c_1, c_2 might satisfy $c_1 = c_2$ for $c_1 \in \mathbb{R}$, and we might take $c_1 = -c_2$ to get another.

Note that $c_1 = c_2 = \frac{1}{2}$, and $c_1 = -c_2 = \frac{1}{2i}$, to obtain $y(t) = \cos t$ and $y(t) = \sin t$, respectively.

So we can conclude that $y(t) = a_1 \cos(t) + a_2 \sin t$ for real a_1, a_2.

Applying the conditions, $y(0) = 1$ requires $a_1 = 1$. And $y'(0) = 0$ requires $a_2 = 0$, so the final solution is $\cos(t)$.

Are the complex solutions "valid"? Well, by definition, we are interested in studying real function solutions to ODEs. But it's not that hard to extend this theory to complex functions. What's harder is developing a theory that works when you are dealing with functions with complex arguments (this is the starting point of complex analysis).

Theorem. Consider the second order equation with constant coefficients

$$y'' + ay' + b = 0,$$

and suppose that one solution λ_1 to the associated equation $\lambda^2 + a\lambda + b = 0$ is $\mathbb{C} \setminus \mathbb{R}$, then we can conclude that:

1. $\overline{\lambda_1}$ is also a solution to $\lambda^2 + a\lambda + b = 0$.
2. If $\lambda_1 = c + id$, then $y_1(t) = e^{ct} \cos dt$, $y_2(t) = e^{ct} \sin dt$.

Proof. (Part 1.) We know that λ_1 is a solution, so $\lambda_1^2 + a\lambda_1 + b = 0$. One thing we can do is express $\lambda_1 = c + id$, and then algebraically show that the conjugate works too, algebraically. Another way is to use the quadratic formula, but this doesn't work for polynomials in general. The best way is just to take the complex conjugate of both sides.

Recall the properties

$$\begin{aligned} z &= \overline{z}, \text{ if } z \in \mathbb{R}, \\ \overline{z \cdot w} &= \overline{z} \cdot \overline{w} \\ \overline{z + w} &= \overline{z} + \overline{w}. \end{aligned}$$

These properties follow pretty easily from the geometric interpretation of complex numbers.

Now, applying this, clearly we obtain

$$\overline{\lambda_1^2 + a\lambda_1 + b} = \overline{\lambda_1}^2 + a\overline{\lambda_1} + b.$$

Proof. (Part 2.) We know that if $y(t) = e^{\lambda_1 t}$, we know that $y(t)$ solves $y'' + ay' + b = 0$.

Now, if we prove that the real part and the imaginary part of $y(t)$ are solutions, we are done. But this follows easily from linearity. This completes the proof.

Example. Let's consider the problem $my''(t) = -ky(t) - \gamma y'(t)$. Take $m = 1, k = 5, \gamma = 4$. So the problem becomes

$$y''(t) + 4y'(t) + 5y(t) = 0,$$

and the solutions to $\lambda^2 + 4\lambda + 5 = 0$ are $-2 \pm i$. So we know that two solutions are $y_1(t) = e^{-2t}\cos t$, $y_2(t) = e^{-2t}\sin t$. We claim that $y(t) = c_1 e^{-2t}\cos t + c_2 e^{-2t}\sin t$ gives all the solutions. We know this intuitively because it's an order 2 equation.

Now, suppose $c_2 = 0, c_1 = 1$, and suppose we are trying to plot $y(t) = e^{-2t}\cos t$. This function looks like a spring that oscillates, but across time it dampens. This makes intuitive sense physically.

We'll do a few more exercises, and then next week we'll start with linear algebra.

Suppose we have the equation $y'' + ay' + by = 0$. What are the possible shapes of this equation? If the associated equation $\lambda^2 + a\lambda + b = 0$ has:

- Complex solutions, if $\lambda_1 = c + id, \lambda_2 = c - id$, and the solution are $y(t) = c_1 e^{ct}\cos dt + c_2 e^{ct}\sin t$.
- Real distinct solutions. If λ_1, λ_2 are the solutions, by direct substitution, $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ are solutions, so $y(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$ is the general solution.
- One real solution λ. If you have a repeated root, you should have $y(t) = ce^{\lambda t} + cte^{\lambda t}$.

To proof this third case, we see that there's a linear algebraic interpretation. We know that in this case $e^{\lambda t}$ is a solution.

What happens if we take $y(t) = te^{\lambda t}$? We obtain $y'(t) = \lambda te^{\lambda t} + e^{\lambda t}$. And further, we get $y''(t) = \lambda^2 e^{\lambda t} + 2\lambda e^{\lambda t}$. So in particular,

$$y''(t) + ay'(t) + by(t) = e^{\lambda t}\left[t(\lambda^2 + a\lambda + b) + 2\lambda + a\right],$$

If $2\lambda + a = 0$, we are done. But using the fact that $a = -(\lambda + \lambda) = -2\lambda$, which proves the desired result.

6.6 Lecture 6: 7-1-19

Recall the problem $y'' + ay' + by = 0$, where $a, b \in \mathbb{R}$.

Recall we need to study the associated polynomial equation

$$\lambda^2 + a\lambda + b = 0,$$

where, we either have:

1. Two distinct roots (real).
2. One multiple root (real).
3. Two complex conjugate roots.

Example. Consider the problem $y'' - 3y' + 2y = 0, y(0) = 5, y'(0) = 2$.

The associated equation is

$$\lambda^2 + 3\lambda + 2 = 0,$$

so that $\lambda = 2, 1$. Two independent solutions for the ODE are

$$y(t) = e^{2t}, e^t,$$

so that all the solutions are given by

$$y(t) = c_1 e^{2t} + c_2 e^t,$$

$c_1, c_2 \in \mathbb{R}$.

Example. Consider the problem $y'' - 2y' + 2y = 0$, so that the associated quadratic is

$$\lambda^2 - 2\lambda + 2 = 0.$$

Indeed, we obtain

$$\lambda = \frac{2 \pm \sqrt{-4}}{2} = 1 \pm i.$$

Two independent solutions are given by

$$e^{(1+i)t}, e^{(1-i)t},$$

but we want real solutions, so we write

$$e^{(1+i)t} = e^t(\cos t + i \sin t).$$

Taking the real part and imaginary part, we get

$$\Re(e^{(1+i)t}) = e^t \cos t,$$

$$\Im(e^{(1+i)t}) = e^t \sin t,$$

which gives us the general solution of

$$y(t) = c_1 e^t \cos t + c_2 e^t \sin t.$$

Example. Consider the problem $y'' - 2y' + y = 0$, so that the associated equation is $\lambda^2 - 2\lambda + 1 = 0$. This implies $y = 1$. In this case, two independent are e^t, te^t, and

$$y(t) = c_1 e^t + c_2 te^t.$$

Example. Recall the Lotko-Volterra model, defined as follows:

$$\begin{aligned} \frac{dx}{dt} &= x(a - by) \\ \frac{dy}{dt} &= y(-c + dx). \end{aligned}$$

Note that this problem as stated is too hard, but we can try to do something simpler. Ideally, we want to solve a problem that is linear in the solution $\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$.

In the linear case, we note that we want to solve something like

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

The "solution" of this equation would be $x(t) = e^{At}x(0)$, but we want to define the exponential of a matrix so that this works.

- Simplest case. Suppose A is diagonal, with

  $$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

  In case, we obtain

  $$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 x(t) \\ \lambda_2 y(t) \end{pmatrix}.$$

  And we obtain the equations

  $$\begin{aligned} x'(t) &= \lambda_1 x(t) \\ y'(t) &= \lambda_2 y(t). \end{aligned}$$

In particular, this implies $x(t) = c_1 e^{\lambda_1 t}$, $y(t) = c_2 e^{\lambda_2 t}$. Finally, we can rewrite this as

$$\mathrm{x}(t) = c_1 e^{\lambda_1 t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_2 e^{\lambda_2 t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

If we consider the associated iniital value problem $\mathrm{x}'(t) = A\mathrm{x}(t)$, with $\mathrm{x}(0) = \mathrm{x}_0$, we just need to find c_1, c_2 such that $\mathrm{x}(0) = \mathrm{x}_0$.

It turns out that we can always reduce to this case in some way.

- More general case. Suppose $A = O\Lambda O^{-1}$, where Λ is diagonal, and O is invertible. Intuitively, this is a change of basis procedure. This means that A is "diagonalizable."

  And suppose we are trying to study the problem

  $$\mathrm{x}'(t) = A\mathrm{x}(t).$$

  Notice that this can be written as

  $$\mathrm{x}'(t) = O\Lambda O^{-1}\mathrm{x}(t),$$

  which is the same as

  $$O^{-1}x'(t) = \Lambda O^{-1}\mathrm{x}(t).$$

  Since O does not depend on t, we can rewrite this as

  $$(O^{-1}\mathrm{x}(t))' = \Lambda O^{-1}\mathrm{x}(t).$$

  Let $\Lambda O^{-1}\mathrm{x}(t) = \mathrm{y}(t)$. In y, we obtain

  $$\mathrm{y}'(t) = \Lambda\mathrm{y}(t).$$

  As before, we can obtain

  $$\mathrm{y}(t) = c_1 e^{\lambda_1 t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_2 e^{\lambda_2 t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

  So, we can compute $\mathrm{x}(t) = O\mathrm{y}(t)$.

Let's try to determine when it is possible to write $A = O\Lambda O^{-1}$. Note that this is equivalent to

$$AO = O\Lambda.$$

Note that

$$AO\mathrm{e}_1 = O\Lambda\mathrm{e}_1 = O(\lambda_1\mathrm{e}_1) = \lambda_1 O\mathrm{e}_1.$$

So the vector $O\mathrm{e}_1$ is an eigenvector with eigenvalue λ_1.

Definition. If A is a matrix, and $\mathrm{v} \neq 0$ has the property $A\mathrm{v} = \lambda\mathrm{v}$ for $\lambda \in \mathbb{R}$, then v is called an eigenvector with eigenvalue λ.

Observation. If A is $O\Lambda O^{-1}$, then $O\mathrm{e}_1$ is an eigenvector with eigenvalue λ_1.

Example. Suppose that $A = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$, and $O = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Then $AO = \begin{pmatrix} 2 & 2 \\ 2 & 0 \end{pmatrix}$. Now, if we let $\mathrm{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathrm{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, we obtain $A\mathrm{v}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ and $A\mathrm{v}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$.

Theorem. If A is an $n \times n$ matrix, and $\mathrm{v}_1, \dots, \mathrm{v}_n$ are n independent eigenvectors with eigenvalue $\lambda_1, \dots, \lambda_n$, then A is diagonalizable. That is, $A = O\Lambda O^{-1}$, where O is a matrix with columns $O = \begin{pmatrix} \mathrm{v}_1 & \dots & \mathrm{v}_n \end{pmatrix}$.[3]

Proof. Multiply on the right by O, to obtain

$$AO = \Lambda O.$$

The left hand side is $A\begin{pmatrix} \mathrm{v}_1 & \mathrm{v}_2 & \dots & \mathrm{v}_n \end{pmatrix} = \begin{pmatrix} A\mathrm{v}_1 & \dots & A\mathrm{v}_n \end{pmatrix}$. And the right side is $\begin{pmatrix} \lambda_1\mathrm{v}_1 & \dots & \lambda_1\mathrm{v}_n \end{pmatrix}$. And we have equality if the two sides are componentwise equal, but this is true by the definition of eigenvector.

A useful criterion is that if the eigenvalues are distinct, then we must have n linearly independent eigenvectors.

Now, notice that $A\mathrm{v} = \lambda\mathrm{v}$ is the same as

$$(A - \lambda I)\mathrm{v} = 0.$$

For this to have non-trivial solutions, we must have $\det(A - \lambda I) = 0$. Note that if A is an $n \times n$ matrix, the determinant is a polynomial of degree n, so the number of solutions is at most n.

So, to find eigenvalues, just look at the roots of $\det(A - \lambda I)$. To find the corresponding eigenvectors once the eigenvalues are known, we need to solve $(A - \lambda I)\mathrm{v} = 0$.

Example. Consider the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Find all eigenvectors and eigenvalues.

We have that

$$A - \lambda I = \begin{pmatrix} -\lambda & 1 \\ 1 & -\lambda \end{pmatrix}.$$

We have that the determinant of this is 0, so that

$$\det(A - \lambda I) = \lambda^2 - 1.$$

So the eigenvalues are $\lambda = \pm 1$. Recall that if $C = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, we have $\det C = ad - bc$, and $C^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Then we need to solve

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0.$$

[3] Note that it is clear that the columns of O must be invertible.

This gives us $-v_1 + v_2 = 0$, $v_1 - v_2 = 0$. In particular, $v_1 = v_2 = 1$. So the first eigenvector is $\begin{pmatrix}1\\1\end{pmatrix}$ with eigenvalue 1.

Now, we need to solve the problem

$$\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}\begin{pmatrix}v_1\\v_2\end{pmatrix} = 0.$$

In this case, we get $v_1 = 1, v_2 = -1$. So the second eigenvector is $\begin{pmatrix}1\\-1\end{pmatrix}$ with eigenvalue -1.

This implies that A is diagonalizable. Let's verify this is the case.

Let

$$O = \begin{pmatrix}\mathrm{v}_1 & \mathrm{v}_2\end{pmatrix} = \begin{pmatrix}1 & 1\\1 & -1\end{pmatrix}.$$

If this is the case, then $A = O\Lambda O^{-1}$. Indeed, we have $O^{-1} = \frac{1}{2}\begin{pmatrix}-1 & -1\\-1 & 1\end{pmatrix}$. We claim that

$$\begin{aligned}\begin{pmatrix}0 & 1\\1 & 0\end{pmatrix} &= \begin{pmatrix}1 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}1 & 0\\0 & -1\end{pmatrix}\cdot\frac{1}{2}\begin{pmatrix}-1 & -1\\-1 & 1\end{pmatrix}\\ &= \begin{pmatrix}1 & -1\\1 & 1\end{pmatrix}\cdot\frac{1}{2}\begin{pmatrix}-1 & -1\\-1 & 1\end{pmatrix}\\ &= \begin{pmatrix}0 & 1\\1 & 0\end{pmatrix},\end{aligned}$$

6.7 Lecture 8: 7-3-19

Phase portaits for linear, homogeneous ODE. (2x2 case, different eigenvalues).

Example 1. Suppose we have the equation $\mathrm{x}'(t) = A\mathrm{x}(t)$, where A is a 2×2 matrix with 2 distinct rea leigenvalues λ_1, λ_2, and $\mathrm{v}_1, \mathrm{v}_2$ associated eigenvectors. This is equivalent to

$$A = O\Lambda O^{-1},$$

where $O = \begin{pmatrix}\mathrm{v}_1 & \mathrm{v}_2\end{pmatrix}$, and $A = \begin{pmatrix}\lambda_1 & 0\\0 & \lambda_2\end{pmatrix}$. We get the solution $\mathrm{x}(t) = c_1 e^{\lambda_1 t}\mathrm{v}_1 + c_2 e^{\lambda_2 t}\mathrm{v}_2$.

We want to plot the solutions - you could make a plot in 3D, but an easier way to visualize is to project this down to 2D space. We plot in $x - y$ space.

Case 1. Consider the case where A is a diagonal matrix. Let $A = \begin{pmatrix}1 & 0\\0 & -1\end{pmatrix}$. Let $\mathrm{y}(t) = \begin{pmatrix}z(t)\\w(t)\end{pmatrix}$, and consider the equation $\mathrm{y}' = A\mathrm{y}$.

We know that the eigenvalues are $1, -1$, and the eigenvectors are $\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}$. Thus the solution is

$$\mathrm{y}(t) = c_1 e^t \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_2 e^{-t} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 e^t \\ c_2 e^{-t} \end{pmatrix}.$$

Note that $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \mathrm{y}(0)$.

- If $c_1 = c_2 = 0$, we get a solution $y(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.
- If $c_2 = 0, c_1 = 1$, we have the solution $\mathrm{y}(t) = \begin{pmatrix} e^t \\ 0 \end{pmatrix}$.
- If $c_1 = -1, c_2 = 0$, we have the solution $\mathrm{y}(t) = \begin{pmatrix} -e^t \\ 0 \end{pmatrix}$.
- Similarly, if $c_2 = 1, c_1 = 0$, we have the solution $\mathrm{y}(t) = \begin{pmatrix} 0 \\ e^{-t} \end{pmatrix}$.
- If $c_2 = -1, c_1 = 0$, we have $\mathrm{y}(t) = \begin{pmatrix} 0 \\ -e^{-t} \end{pmatrix}$.
- If $c_1 = 1, c_2 = 1$, we have $\mathrm{y}(t) = \begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}$. Note that as a function of time, $z(t)w(t) = 1$.

Note that $\mathrm{y}(t) = \begin{pmatrix} c_1 e^t \\ c_2 e^{-t} \end{pmatrix}$.

We can plot the “phase portrait,” which is a plot in the $z - w$ plane of the solution (supressing t).

Example 1’. Suppose we have the problem $\mathrm{x}' = A\mathrm{x}$, where $\mathrm{x}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, and $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

It turns out we can get the phase portrait by transforming the solution of the previous example.

Remember from before: if $\mathrm{y}(t) = O^{-1}\mathrm{x}(t)$, then $\mathrm{y}(t)$ solves $\mathrm{y}'(t) = \Lambda \mathrm{y}(t)$. Therefore $\mathrm{x}(t) = O\mathrm{y}(t)$.

Notice that:

1. Any linear transformation of the plane sends lines to lines.
2. The image of $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is v_1. The image of $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ is v_2.
3. Linear transformations preserve both lines and tangency.

Note that the image is a rotation by 45 degrees counterclockwise.

Remark. In general, for $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, you get the same qualitative picture. Suppose you have

$$z(t) = e^{\lambda_1 t}$$
$$w(t) = e^{\lambda_2 t}.$$

To obtain a relationship between z and w, you need to raise w to the power of λ_1/λ_2. In general, you will get a curve $z(t)(w(t))^{|\frac{\lambda_1}{\lambda_2}|} = c$.

Example 2. Consider the equation $\mathrm{y}' = \Lambda \mathrm{y}$ where $\Lambda = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$.

In this case, $y(t) = c_1 e^{2t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_2 e^t \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 e^{2t} \\ c_2 e^t \end{pmatrix}$. We get $z(t) = e^{2t}$, $w(t) = e^t$, and the relationship if $z = w^2$.

Remark. Note that we're not even considering the case where the eigenvalues are negative, since the phase portrait will be the same, it's just that the arrows are reversed.

- Note that when $\lambda_1, \lambda_2 > 0$, the equilibrium is called an unstable node.
- When $\lambda_1, \lambda_2 < 0$, it's called a stable node.

Example 2'. Suppose $A = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$. First, we need to compute the eigenvalues and eigenvectors.

- Find eigenvalues and eigenvectors. To compute the eigenvalues, we need to solve $\det(A - \lambda I) = 0$, to obtain

  $$\det \begin{pmatrix} 2-\lambda & 0 \\ 1 & 1-\lambda \end{pmatrix} = (2-\lambda)(1-\lambda) = 0,$$

  so $\lambda = 1, 2$. The eigenvectors are $\mathrm{v}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\mathrm{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Example 3. Suppose $A = \begin{pmatrix} -5 & 1 \\ -6 & 0 \end{pmatrix}$. First, determine eigenvalues:

- We have $\det(A - \lambda I) = (-5-\lambda)(-\lambda) + 6 = 0$, so $(\lambda + 3)(\lambda + 2) = 0$, that is $\lambda = -3, -2$.
- Eigenvectors. Computing, for $\lambda = -3$, we have $(A + 3I)\mathrm{v}) = 0$, that is

  $$\begin{pmatrix} -2 & 1 \\ -6 & 3 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

  We get $v_2 = 2v_1$, so $\mathrm{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

  For $\lambda = -2$, we have $(A + 2I)\mathrm{w} = 0$, so $\mathrm{w} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$.

To identify which direction the curve is going, take a point e.g. $(100, 250)$, and try to identify which direction it is "more parallel" to.

On Friday, we will discuss the complex case. Next week, we are going to start on repeated eigenvalues which is pretty tricky.

6.8 Lecture 9: 7-5-19

We are now going to discuss the phase portrait of $\mathrm{x}' = A\mathrm{x}$, where A has complex (conjugate) eigenvalues.

Recalling the geometric interpretation of eigenvalues, we should expect that the linear transformation will scale the eigenvectors. Consider the function f denoting rotation by 90 degrees, counterclockwise.

That is,

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

The eigenvalues are $\lambda = i, -i$, and the eigenvectors are $\begin{pmatrix} 1 \\ -i \end{pmatrix}, \begin{pmatrix} 1 \\ i \end{pmatrix}$.

Theorem. If A has a complex eigenvalue λ, with eigenvector v, then $\overline{\lambda}$ is an eigenvalue, with eigenvector $\overline{\mathrm{v}}$.[4]

Proof. The proof that $\overline{\lambda}$ is an eigenvalue follows from the fact that polynomial roots occur in complex conjugate pairs.

To prove that $\overline{v}$ is also an eigenvector, then

$$A\mathrm{v} = \lambda\mathrm{v}$$

is equivalent to

$$\overline{A\mathrm{v}} = \overline{\lambda\mathrm{v}}$$

which is equivalent to

$$\overline{\mathrm{A}}\overline{\mathrm{v}} = \overline{\lambda}\overline{\mathrm{v}},$$

that is

$$A\mathrm{v} = \overline{\lambda}\overline{\mathrm{v}}.$$

□

Suppose we want to solve $\mathrm{x}' = A\mathrm{x}$. Suppose that A has complex eigenvalues $\lambda_1 = \alpha + i\beta, \lambda_2 = \overline{\lambda_1}$.

We can write the solutions as

$$c_1 e^{\lambda t}\mathrm{v} + c_2 e^{\overline{\lambda} t}\overline{\mathrm{v}}.$$

Now, for example suppose $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Then the solution is $\mathrm{x}(t) = c_1 e^{it}\begin{pmatrix} 1 \\ -i \end{pmatrix} + c_2 e^{-it}\begin{pmatrix} 1 \\ i \end{pmatrix}$.

We would like to take real / imaginary parts of this solution to get the real solutions to the equation.

We will state an important theorem.

[4] For vectors, we define the complex conjugate as taking the complex conjugate of each entry.

Theorem. If $\lambda = \alpha + i\beta$ is an eigenvalue for A with eigenvector v, then $\Re(e^{\lambda t}\mathrm{v})$, $\Im(e^{\lambda t}\mathrm{v})$ are also solutions, and they are independent.

In our example, $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $\mathrm{x}(t) = e^{it}\begin{pmatrix} 1 \\ -i \end{pmatrix}$ is a solution. Then we can apply Euler's formula to obtain

$$\begin{aligned} \mathrm{x}(t) = e^{it}\begin{pmatrix} 1 \\ -i \end{pmatrix} &= (\cos t + i\sin t)\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} + i\begin{pmatrix} 0 \\ -1 \end{pmatrix}\right] \\ &= \begin{pmatrix} \cos t + i\sin t \\ -i\cos t + \sin t \end{pmatrix} \\ &= \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} + i\begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}. \end{aligned}$$

By the theorem, we know that $\begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$ and $\begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}$ are two independent solutions, so

$$\begin{pmatrix} x \end{pmatrix}(t) = c_1\begin{pmatrix} \cos t \\ \sin t \end{pmatrix} + c_2\begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}.$$

Now, let's try to draw the phase portrait. Recall that we would like to "forget" about the t variables and draw the relationship between x and y.

At least for $c_1 = 1, c_2 = 0$, the solution is $\begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$, so the phase portrait is a circle.

For generic c_1, c_2, then

$$\begin{aligned} (c_1\cos t + c_2\sin t)^2 + (c_1\sin t - c_2\cos t)^2 &= c_1^2\cos^2 t + c_2^2\sin^2 t + 2c_1c_2\sin t\cos t + c_1^2\sin^2 t + c_2^2\cos^2 t - 2c_1c_2\sin t\cos t \\ &= c_1^2 + c_2^2. \end{aligned}$$

Example. Suppose we are given the matrix A with complex eigenvalue $\lambda = a + i\beta$, and eigenvector $\mathrm{v} = \mathrm{v}_1 + i\mathrm{v}_2$. Recall that the complex conjugate of the eigenvalue / eigenvector also yields a solution.

In particular

$$\mathrm{x}(t) = e^{(\alpha+i\beta)t}\mathrm{v}$$

is a solution.

By Euler's formula,

$$\begin{aligned} \mathrm{x}(t) &= e^{\alpha t}e^{i\beta t}\mathrm{v} \\ &= e^{\alpha t}(\cos\beta t + i\sin\beta t)(\mathrm{v}_1 + i\mathrm{v}_2) \\ &= e^{\alpha t}(\cos\beta t\mathrm{v}_1 - \sin\beta t\mathrm{v}_2) + ie^{\alpha t}(\cos\beta t\mathrm{v}_2 + \sin\beta t\mathrm{v}_1)) \end{aligned}$$

Therefore, the general solution is a combination of

$$e^{\alpha t}(\cos\beta t \mathrm{v}_1 - \sin\beta t \mathrm{v}_2),$$

and

$$e^{\alpha t}(\cos\beta t \mathrm{v}_2 + \sin\beta t \mathrm{v}_1)$$

What does the phase portrait look like?

- If $\alpha = 0$, the phase portrait will follow the trajectory of an ellipse.
- If $\alpha > 0$, the phase portrait will spiral outward in an "elliptical" fashion.
- If $\alpha < 0$, the phase portrait will spiral inward towards the origin.

6.9 Lecture 10: 7-8-19

We will now discuss the case of repeated eignevalues. Suppose we are trying to solve the system $\mathrm{x}' = A\mathrm{x}$.

And consider the matrix $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$. Solving for eigenvalues, we obtain

$$\det(A - \lambda I) = (2-\lambda)^2 = 0,$$

so the eigenvalue is λ with multiplicity 2. Solving, we find that the eigenvector is $v_2 = \mathrm{v} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Since we only have one eigenvector, we know that we cannot diagonalize it.

Goal. Given a matrix A which cannot be put in diagonal form, find an alternative "canonical form."

Answer. (In 2 × 2 case.) If A is not diagonalizable, A can always be put in the form

$$A = OJO^{-1},$$

where

$$J = \begin{pmatrix} \lambda & 1 \\ \lambda & 0 \end{pmatrix}.$$

How can we solve $\mathrm{x}' = J\mathrm{x}$, where $\mathrm{x}^{(t)} = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, with $J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$.

We can write

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix},$$

implying that

$$x'(t) = \lambda x(t) + y(t)$$
$$y'(t) = \lambda y(t).$$

Solving these equations is relatively easy.

- Equation 2 just reduces to $y(t) = ce^{\lambda t}$.
- Equation 1 just reduces to the integrating factor example. We just write $\mu(t) = \exp(\int p(t))$ as usual.

If $\mathrm{x}' = J\mathrm{x}$, with $J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$, then

$$\mathrm{x}(t) = \begin{pmatrix} c_1 te^{\lambda t} + c_2 e^{\lambda t} \\ c_1 e^{\lambda t} \end{pmatrix} = c_2 e^{\lambda t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_1 \left(te^{\lambda t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + e^{\lambda t} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right).$$

If $e^{\lambda t} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $te^{\lambda t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + e^{\lambda t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ can be shown to be independent, then every solution is given by the previous equation.

To check independence, compute the Wronskian (look at $t = 0$).

If $\mathrm{x}' = A\mathrm{x}$, where $A = OJO^{-1}$, then $\mathrm{y} = O^{-1}\mathrm{x}$ solves $J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$, $\mathrm{y}' = J\mathrm{y}$. We can apply the previous part and solve for y. If $O = \begin{pmatrix} \mathrm{v} & \mathrm{w} \end{pmatrix}$ since $\mathrm{x} = O\mathrm{y}$, the solution for $\mathrm{x}' = A\mathrm{x}$ is $\mathrm{x}(t) = c_1 e^{\lambda t}\mathrm{v} + c_2(te^{\lambda t}\mathrm{v} + e^{\lambda t}\mathrm{w})$.

Suppose $A = OJO^{-1}$. This is equivalent to $AO = OJ$.

Apply the matrix AO to $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, to obtain

$$AO \begin{pmatrix} 1 \\ 0 \end{pmatrix} = O \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

This implies

$$A\mathrm{v} = O\lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda \mathrm{v}.$$

This implies v is an eigenvector relative to λ. Now,

$$\begin{aligned} A\mathrm{w} = O \begin{pmatrix} 1 \\ \lambda \end{pmatrix} &= O \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \lambda \end{pmatrix} \right) \\ &= \mathrm{v} + O \left(\lambda \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = \mathrm{v} + \lambda \mathrm{w}. \end{aligned}$$

So we have

$$(A - \lambda)^2 \mathrm{w} = 0.$$

These are called generalized eigenvectors (in this case of order 2). The order of a generalized eigenvector is the power of $(A - \lambda)$ that we have to apply to obtain it.

Example. Suppose we are solving $\mathrm{x}' = A\mathrm{x}$, $A = \begin{pmatrix} -4 & -9 \\ 4 & 8 \end{pmatrix}$, find the general solution such that $\mathrm{x}(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

- Find eigenvalues. We have

  $$\det(A - \lambda I) = 0,$$

  so that

  $$\begin{pmatrix} -4-\lambda & -9 \\ 4 & 8-\lambda \end{pmatrix} = 0$$

  implying

  $$(\lambda - 2)^2 = 0,$$

  so $\lambda = 2$ with multiplicity 2.

  There are two possible scenarios. Either

  - $A = O \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} O \end{pmatrix}^{-1}$ (impossible, since diagonal matrices commute, implying the false implication that $A = 2I$.
  - Or, $A = OJO^{-1}$, where $J = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$.

To find $O = \begin{pmatrix} \mathrm{v} & \mathrm{w} \end{pmatrix}$, use the result that if v is eigenvector solution of $\lambda = 2$, w solves $(A - \lambda)\mathrm{w} = \mathrm{v}$.

We obtain

$$(A - 2I)\mathrm{v} = 0,$$

implying $\mathrm{v} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$.

So, we solve

$$\begin{pmatrix} -6 & -9 \\ 4 & 6 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}.$$

Since there are infinitely many solutions, we can solve this and get a single w, that is

$$\mathrm{w} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

From the formula, we get the solution:

$$ce^{\lambda t} \begin{pmatrix} 3 \\ -2 \end{pmatrix} + c_2(te^{\lambda t}\mathrm{v} + e^{\lambda t}\mathrm{w}).$$

Theorem. If A is an $n \times n$ matrix that is not diagonalizable, then $A = OJO^{-1}$, where J is composed of "blocks" where each block is either:

- λ,
- $\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$ (generalized eigenvector).
- $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$ (eigenvector).

For example, possible cases include the following:

- $A = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 1 \\ 0 & 0 & \lambda_2 \end{pmatrix}$.
- $A = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$.
- $A = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$.

Example. Suppose $J = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$. Can we solve $\mathrm{x}' = J\mathrm{x}$? Well, if we write $\mathrm{x}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$, we obtain

$$\begin{aligned} x'(t) &= \lambda x(t) + y(t) \\ y'(t) &= \lambda y(t) + z(t) \\ z'(t) &= \lambda z(t). \end{aligned}$$

And we can solve these in the usual way (either integrating factors, or just straight up integrating).

6.10 Lecture 11: 7-9-19

Phase plot for repeated eigenvalue. From yesterday, consider the matrix $A = \begin{pmatrix} -4 & -9 \\ 4 & 8 \end{pmatrix}$. We showed that $A = OJO^{-1}$, where $J = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, $O = \begin{pmatrix} \mathrm{v} & \mathrm{w} \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -2 & -1 \end{pmatrix}$.

Now, given $\mathrm{x}' = A\mathrm{x}$ what does the phase portrait look like?

Remember if $y = O^{-1}\mathrm{x}$, then $\mathrm{y}' = J\mathrm{y}$. If we can plot the phase potrait for y, then using $\mathrm{x} = O\mathrm{y}$, we can plot the phase portrait for x.

Ok, so let's consider the plot for $y' = J\mathrm{y}$, with $\mathrm{y}(t) = \begin{pmatrix} z(t) \\ w(t) \end{pmatrix}$. Then

$$z'(t) = 2z(t) + w(t)$$
$$w'(t) = 2w(t).$$

This implies

$$\begin{aligned} z(t) &= c_2te^{2t} + c_1e^{2t} \\ w(t) &= c_2e^{2t}. \end{aligned}$$

For $c_2 = 0, c_1 = 1$, we obtain $\mathrm{y}(t) = \begin{pmatrix} e^{2t} \\ 0 \end{pmatrix}$.

For $c_1 = 0, c_2 = 1$, we obtain $\mathrm{y}(t) = \begin{pmatrix} te^{2t} \\ e^{2t} \end{pmatrix}$.

Example. Let $A = \begin{pmatrix} 3 & 5 & 2 \\ -7 & 9 & 3 \\ 4 & -4 & 0 \end{pmatrix}$. Find the independent solutions to $\mathrm{x}' = A\mathrm{x}$.

Recall that

$$\begin{aligned} \det(A - \lambda I) &= \det \begin{pmatrix} -3-\lambda & 5 & 2 \\ -7 & 9-\lambda & 3 \\ 4 & -4 & 0 \end{pmatrix} \\ &= (-3-\lambda)\det\begin{pmatrix} 3-\lambda & 3 \\ -4 & -\lambda \end{pmatrix} - 5\det\begin{pmatrix} -4 & 3 \\ 4 & -\lambda \end{pmatrix} + 2\det\begin{pmatrix} -4 & 9-\lambda \\ 4 & -4 \end{pmatrix} \\ &= (-3-\lambda)\left[(3-\lambda)(-\lambda) + 12\right] - 5\left[4\lambda - 12\right] + 2\left[16 - 4(9-\lambda)\right] \\ &= -\lambda^3 + 6\lambda^2 - 12\lambda + 8 \\ &= (2-\lambda)^3. \end{aligned}$$

Recall that we proved that either A is diagonalizable, or that it is similar to a matrix $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$ or $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

Notice that the number of independent eigenvectors is equal to $n = \dim(A - 2I)$.

- If $n = 1$, $J = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$.
- If $n = 2$, $J = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

Note that

$$A - 2I = \begin{pmatrix} -5 & 5 & 2 \\ -7 & 7 & 3 \\ 4 & -4 & -2 \end{pmatrix}.$$

Further, by the rank-nullity theorem

$$3 = \dim(\ker(A - 2I)) + \dim((A - 2I))$$

Here, the form of J is known as the Jordan canonical form.

So far, we have concluded that

$$A = O\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}O^{-1},$$

where $O = \Big(\mathrm{v}, \mathrm{w}_1, \mathrm{w}_2\Big)$.

- To find v, just look at $(A - 2I)\mathrm{v} = 0$.
- Once v is found, w_1 satisfies $(A - 2I)\mathrm{w}_1 = \mathrm{v}$.
- Once w_1 is known, w_2 satisfies $(A - 2I)\mathrm{w}_2 = \mathrm{w}_1$.

We will skip the computation in lecture, but we obtain the following:

$$\mathrm{v} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}; \qquad \mathrm{w}_1 = \begin{pmatrix} -1 \\ 0 \\ -2 \end{pmatrix}; \qquad \mathrm{w}_2 = \begin{pmatrix} 3 \\ 0 \\ 7 \end{pmatrix}.$$

Recall from yesterday, we had seen that if $\mathrm{y}' = J\mathrm{y}$ and $J = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$, it follows that

$$\begin{aligned} y_1' &= \lambda y_1 + y_2 \\ y_2' &= \lambda y_2 + y_3 \\ y_3' &= \lambda y_3. \end{aligned}$$

Then the general solution is given by

$$\mathrm{y}(t) = \begin{pmatrix} c_3\frac{t^2}{2}e^{2t} + c_2te^{2t} + c_1e^{2t} \\ c_3te^{2t} + c_2e^{2t} \\ c_3e^{2t} \end{pmatrix}.$$

Some example solutions include the following:

- $\mathrm{v}_1(t) = \begin{pmatrix} e^{2t} \\ 0 \\ 0 \end{pmatrix}$.
- $\mathrm{v}_2(t) = te^{2t}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + e^{2t}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$.
- $\mathrm{v}_3(t) = \frac{t^2}{2}e^{2t}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + te^{2t}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + e^{2t}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

Note that $\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3$ are independent solutions (check the Wronskian). Since

- $\mathrm{x} = O\mathrm{y}$ satisfies $\mathrm{x}' = A\mathrm{x}$,
- O maps $\begin{pmatrix}1\\0\\0\end{pmatrix} \to \mathrm{v}, \begin{pmatrix}0\\1\\0\end{pmatrix} \to \mathrm{w}_1, \begin{pmatrix}0\\0\\1\end{pmatrix} \to \mathrm{w}_2.$

Three independent solutions for $\mathrm{x}' = A\mathrm{x}$ are given by

$$\begin{aligned}\tilde{\mathrm{v}}_1 &= e^{2t}\mathrm{v}\\ \tilde{\mathrm{v}}_2 &= te^{2t}\mathrm{v} + e^{2t}\mathrm{w}_1\\ \tilde{\mathrm{v}}_3 &= \frac{t^2}{2}e^{2t}\mathrm{v} + te^{2t} + \mathrm{w}_1 + e^{2t}\mathrm{w}_2.\end{aligned}$$

One comment about the phase portrait: recall that we saw from before if $A = \begin{pmatrix}2 & 1\\ 0 & 2\end{pmatrix}$, we get a phase portrait of the following form:

Instead, consider $A = \begin{pmatrix}0 & 1\\ -1 & 1-\varepsilon\end{pmatrix}$. Computing the eigenvalues, we obtain

$$\begin{aligned}\det(A - \lambda I) &= \det\begin{pmatrix}-\lambda & 1\\ -1 & 2-\lambda-\varepsilon\end{pmatrix}\\ &= \lambda^2 + (-2+\varepsilon)\lambda + 1.\end{aligned}$$

Computing the eigenvalues, we obtain

$$\lambda = \frac{2 - \varepsilon \pm \sqrt{-4\varepsilon + \varepsilon^2}}{2}.$$

We can obtain the following phase portraits; the first phase portrait is a limiting case of the second.

6.11 Lecture 12: 7-10-19

Fundamental solution and exponential of a matrix.

Recall the problem $\mathrm{x}' = A\mathrm{x}$. We know how to solve this using $A = O\Lambda O^{-1}$; $A = OJO^{-1}$ (using the diagonal matrix and the Jordan canonical matrix).

The solutions are given by

$$\mathrm{x}(t) = c_1e^{\lambda_1 t}\mathrm{v}_1 + c_2e^{\lambda_2 t}\mathrm{v}_2 + \cdots + c_ne^{\lambda_n t}\mathrm{v}_n.$$

Definition. A fundamental matrix for a system of ODE $\mathrm{x}' = A\mathrm{x}$ is a matrix $X(t)$ such that for any constant vector c, we have $X(t)\mathrm{c}$ is a solution of the system of ODE. This is the same as the columns of $X(t)$ are independent solutions.

Since for all c, $X(t)\mathrm{c}$ is a solution, if we want $x_0 = \mathrm{x}(t_0) = X(t_0)\mathrm{c}$, invert $X(t_0)$ so that $\mathrm{c} = (X(t_0))^{-1}\mathrm{x}_0$. Then $X(t)X^{-1}(t_0)\mathrm{x}_0$ is a solution and it satisfies the initial condition.

Example. If $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \lambda = \pm 1, \mathrm{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathrm{w} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, then $X(t) = \begin{pmatrix} e^t & -e^{-t} \\ e^t & e^{-t} \end{pmatrix}$ is a fundamental matrix.

If $\mathrm{x}' = A\mathrm{x}$ is paired with $\mathrm{x}(0) = x_0$ and a fundamental matrix $X(t)$ is given, then

$$\tilde{X}(t) = X(t)X^{-1}(0)$$

is also a fundamental solution. Therefore, if $\mathrm{c} = \mathrm{x}_0$, then $\tilde{X}(t)\mathrm{c}$ is a solution ot the ODE, and it matches the initial condition since $\tilde{X}(0)\mathrm{c} = I\mathrm{x}_0 = \mathrm{x}_0$.

Definition. Given an $n \times n$ matrix A, its exponential $e^{tA} = \sum_{n=0}^{\infty} \frac{t^n A^n}{n!}$.

Notice that $\tilde{X}(0) = I$, since $e^{0A} = I$. If we also check that the columns of e^{tA} are solutions ot the ODE $\mathrm{x}' = A\mathrm{x}$, we are good.

Equivalently, $e^{tA}\mathrm{c} = \mathrm{y}(t)$ satisfies $\mathrm{y}'(t) = A\mathrm{y}(t)$. Thus,

$$\begin{aligned}
(e^{tA}\mathrm{c})' &= \left(\sum_{n=0}^{\infty} \frac{t^n A^n \mathrm{c}}{n!}\right) \\
&= \sum_{n=1}^{\infty} \frac{t^{n-1j A^n \mathrm{c}}}{(n-1)!} \\
&= A \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!} A^{n-1}\mathrm{c} \\
&= A \sum_{m=0}^{\infty} \frac{t^m}{m!} A^m \mathrm{c} = Ae^{tA}\mathrm{c} \\
&= Ay(t).
\end{aligned}$$

In summary, the $\tilde{X}$ from above is exactly the same as the matrix exponential. We now will look at some examples.

Case 1. If $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, what is e^{tA}? Then

$$e^{tA} = \sum_{n=0}^{\infty} \frac{t^n \Lambda^n}{n!} = \mathrm{diag}(e^{t\lambda_1}, \ldots, e^{t\lambda_n}).$$

Case 2. Suppose A is diagonalizable, i.e. $AO\Lambda O^{-1}$, where Λ is diagonal. To do this, we can write

$$\begin{aligned} e^{tA} &= \sum_{n=0}^{\infty} \frac{t^n A^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n (O\Lambda O^{-1})^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{t^n O \Lambda^n O^{-1}}{n!} \\ &= O \left(\sum_{n=1}^{\infty} \frac{t^n \Lambda^n}{n!} \right) O^{-1} \\ &= O e^{tA} O^{-1} \\ &= O\text{diag}(e^{t\lambda_1}, \ldots, e^{t\lambda_n}) O^{-1}. \end{aligned}$$

Case 3. How do we compute e^{tJ}, where J is an arbitrary Jordan canonical form? We will start by doing the 2×2 case.

Suppose $J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} = \underbrace{\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}}_{\text{call this } \Lambda} + \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{\text{call this } N}$.

Key point. If Λ is diagonal, $N^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, where Λ and N commute. Here N is nilpotent.

And using this, we can apply a sort of binomial theorem for matrices. If A and B commute, then

$$(A + B)^n = \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k}.$$

(Note that this result is not true if A and B don't commute; things are much more complicated in that setting).

Thus,

$$\begin{aligned} e^{tJ} &= \sum_{n=0}^{\infty} \frac{t^n J^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{t^n (\Lambda + N)^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} (A^n + n\Lambda^{n-1} N) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} \Lambda^n + \sum_{n=0}^{\infty} \frac{t^n}{n!} n\Lambda^{n-1} N \\ &= e^{t\Lambda} + tN \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!} \Lambda^{n-1} \\ &= e^{t\Lambda}(1 + tN) \\ &= \begin{pmatrix} e^{\lambda t} & 0 \\ 0 & e^{\lambda t} \end{pmatrix} + t \begin{pmatrix} 0 & t \\ 0 & 0 \end{pmatrix} \begin{pmatrix} e^{\lambda t} & 0 \\ 0 & e^{\lambda t} \end{pmatrix} \\ &= \begin{pmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix}. \end{aligned}$$

Case 4. If $A = OJO^{-1}$, then $e^{tA} = Oe^{tJ}O^{-1}$ (same proof as case 2).

Example (complex eigenvalue case). Suppose $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. What is e^{tA}?

One way is to compute $\lambda_1 = i, \lambda_2 = -i$, v, w, and note that

$$e^{tA} = \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} \begin{pmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix}^{-1}$$

Also, notice that

$$A^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -I.$$

Thus, if we compute the matrix exponential, we can obtain

$$\begin{aligned} e^{tA} &= \sum_{n=0}^{\infty} \frac{t^n A^n}{n!} \\ &= I + tA + \frac{t^2}{2}A^2 + \frac{t^3}{3!}A^3 + \dots \\ &= I - \frac{It^2}{2} + I\frac{t^4}{4!} + \dots + A\left(tI - I\frac{t^3}{3!} + I\frac{t^5}{5!} - \dots\right) \\ &= \cos t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin t \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \end{aligned}$$

In this way, we get two fundamental solutions which are real.

In summary, if $\mathrm{x}' = A\mathrm{x}$, the solution should be $e^{tA}\mathrm{x}_0$.

But there are complications. We need to check which properties are true for a mmatrix.

For example, is it true that $e^{(t+s)A} = e^{tA}e^{sA}$? Answer is yes. Easy way to prove this is to use the existence and uniqueness theorem for ODEs.

What about $e^{t(A+B)} = e^{tA}e^{tB}$? Answer is no, unless A and B commute. In fact, we do have equality here if and only if A and B commute.

It is enough to prove that $\mathrm{x} = e^{tA}e^{tB}$ satisfies $\mathrm{x}' = (A+B)\mathrm{x}$ and $\mathrm{x}(0) = \mathrm{c}$.

6.12 Review: Phase Portraits of Linear Systems

Using reference: http://www.math.psu.edu/tseng/class/Math251/Notes-PhasePlane.pdf.

While labor intensive, it is possible to sketch the phase portrait by hand without having to solve the system (do something similar to the direction field case).

Recall that an equilibrium solution is a constant solution of the system, a point (x_1, x_2) where $\mathrm{x}' = 0$. In general, we will consider systems whose coefficient matrix has nonzero determinant; that is, systems whose origin is the only critical point.

(Asked Andrea); Recall we have the following cases:

- $\lambda_1 < 0\lambda_2$
- $\lambda, \lambda > 0$
- $\lambda_1, \lambda_2 < 0$
- $\lambda_1 = a + ib, \lambda_2 = a - ib$.

Add images

6.13 Lecture 13: 7-11-19

Now, we'll discuss inhomogeneous system of ODE. That is, problems of the form

$$\mathrm{x}' = A\mathrm{x} + \mathrm{b}.$$

Recall that the solution to this is equal to the solution of the homogeneous problem plus a particular solution.

Lemma. If x_1, x_2 are two solutions of $\mathrm{x}' = Ax + b$, then $y = \mathrm{x}_1 - \mathrm{x}_2$ solves $\mathrm{y}' = A\mathrm{y}$.

Proof. Just apply the linearity of the derivative. □

Remark. This idea applies to the case where A, b is nonconstant as well (but of course, we don't know how to solve the problem when A is nonconstant).

To find a particular solution, just note that we can find the equilibrium solution pretty easily. If $\mathrm{x}' = A\mathrm{x}$, then $\mathrm{x} = 0$ is a solution.

Simliarly, if $\mathrm{x}' = A\mathrm{x} + \mathrm{b}$, with the condition A is invertible, we can look for the solution $\mathrm{x}_0 = -A^{-1}\mathrm{b}$.

Therefore, using the lemma, the general solution of $\mathrm{x}' = A\mathrm{x} + \mathrm{b}$ is $\mathrm{x} = -A^{-1}\mathrm{b} + \mathrm{X}(t)\mathrm{c}$, where $X(t)$ is a fundamental matrix and c is an arbitrary vector.

Example. Take $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and $\mathrm{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\mathrm{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Then from before, we have $\lambda = \pm 1$, with $\mathrm{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathrm{w} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$; we obtain the fundamental matrix $X(t) = \begin{pmatrix} e^t & e^{-t} \\ e^t & -e^{-t} \end{pmatrix}$.

Therefore, the final solution to the problem is

$$\begin{aligned} \mathrm{x}(t) &= X(t)\mathrm{c} - A^{-1}\mathrm{b} \\ &= c_1 \begin{pmatrix} e^t \\ e^t \end{pmatrix} + c_2 \begin{pmatrix} e^t \\ -e^{-t} \end{pmatrix} + \begin{pmatrix} -2 \\ -1 \end{pmatrix}. \end{aligned}$$

At $t = 0$, we get

$$\mathrm{x}(0) = c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which gives us $c_1 = \frac{3}{2}, c_2 = \frac{1}{2}$.

This gives us the final solution of

$$\mathrm{x}(t) = \begin{pmatrix} \frac{3}{2}e^t + \frac{1}{2}e^{-t} - 2 \\ \frac{3}{2}e^t - \frac{1}{2}e^{-t} - 1 \end{pmatrix}.$$

For the phase portrait, just shift the phase portrait of $\mathrm{x}' = A\mathrm{x}$ by the vector $-A^{-1}\mathrm{b}$.

What about the general solution for $\mathrm{x}' = A\mathrm{x} + \mathrm{g}$, where g is a function of t? Intuitively, you have a system and there's a force that acts in a nonconstant way on the system (e.g. temperature in some room).

Idea. We just need a particular solution; then just add the solution to the homogeneous problem.

Let's try applying the integrating factor idea from before; this is one reason we introduced the exponential of a matrix. We rewrite the ODE in the form

$$\mathrm{x}' - A\mathrm{x} = g,$$

and multiply by e^{-tA}. We note that e^{-tA} is always invertible because the columns are a solution of the ODE and at time $t = 0$, $e^{-0A} = I$. Alteernatively, we can realize that if $A = O\Lambda O^{-1}$, and that $e^{-tA} = Oe^{-tA}O^{-1}$.

Multiplying on the left, we obtain

$$e^{-tA}\mathrm{x}' - e^{-tA}A\mathrm{x} = e^{-tA}\mathrm{g}.$$

Thus,

$$-e^{tA}\mathrm{x}' - e^{-tA}A\mathrm{x} = e^{-tA}\mathrm{g}.$$

Importantly, e^{-tA} commutes with A, since from the power series expansion A should commute with powers of itself.

This is the same as

$$(e^{-tA}\mathrm{x})' = e^{-tA}\mathrm{g}.$$

Integrating, we obtain

$$e^{-tA}\mathrm{x} = \mathrm{c} + \int^t e^{-sA}\, g(s)\, ds.$$

And in conclusion,

$$\mathrm{x}(t) = e^{tA}\mathrm{c} + e^{tA} \int^t e^{-sA} g(s)\, ds.$$

Remark 1. The solution is $\mathrm{x}(t) = e^{tA}\mathrm{s} + e^{tA} \int^t e^{-sA}\, g(s)\, ds$.

Remark 2. Recall that the integrating factor $\mu(t)$ is non-unique (since you can start the integral from where you want due to invariance under constant multiplication).

Here, instead of e^{At} we use any fundamental matrix $X(t)$, then $X(t)$ is always invertible (using the Wronskian argument), and also $\frac{d}{dt}X(t) = AX(t)$. Here, the previous relation is the same as $(\mathrm{v}', \mathrm{w}') = (A\mathrm{v}, A\mathrm{w})$.

Therefore, $\mathrm{x}(\mathrm{c}) = X(t)\mathrm{c} + X(t) \int^t X^{-1}(s) g(s)\, ds$.

Remark 3. If A is not-constant, but somehow we know a fundamental matrix $X(t)$, then the argument works the same.

Example. Solve the problem $\mathrm{X}' = \mathrm{Ax} + \mathrm{g}$, where $A = \begin{pmatrix} 1 & -4 \\ 2 & -5 \end{pmatrix}$, $\mathrm{x}(0) = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$, $\mathrm{g}(t) = \begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}$. Find $\mathrm{x}(t)$.

First, we find the fundamental matrix, and use the formula from above. To find $X(t)$, we compute the eigenvalues of A. Computing, we obtain

$$\begin{aligned} \det(A - \lambda I) &= (1-\lambda)(-5-\lambda) + 8 = 0 \\ &= (\lambda^2 + 4\lambda + 3) = (\lambda+3)(\lambda+1) = 0. \end{aligned}$$

Thus $\lambda = -1, -3$. Now, if v is relative to $\lambda_1 = -3$, we obtain $\mathrm{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Similarly, if w is relative to $\lambda_2 = -1$, we obtain $\begin{pmatrix} 1 & -4 \\ 2 & -5 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -v_1 \\ -v_2 \end{pmatrix}$, so that $2v_2 - 4v_2 = 0$, and $\mathrm{w} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

Thus, the fundamental matrix is

$$X(t) = \begin{pmatrix} e^{-3t} & 2e^{-t} \\ e^{-3t} & e^{-t} \end{pmatrix}.$$

Computing the inverse, note that

$$X^{-1}(t) = \frac{1}{e^{-2t} - 2e^{-4t}} \begin{pmatrix} e^{-t} & e^{-3t} \\ -2e^{-t} & e^{-3t} \end{pmatrix} = \begin{pmatrix} -e^{3t} & e^{3t} \\ e^t & -e^t \end{pmatrix}.$$

Thus, the final solution is

$$\mathrm{x}(t) = c_1 \begin{pmatrix} e^{-4t} \\ e^{-3t} \end{pmatrix} + c_2 \begin{pmatrix} 2e^{-t} \\ e^{-t} \end{pmatrix} + \begin{pmatrix} e^{-3t} & 2e^{-t} \\ e^{-3t} & e^{-t} \end{pmatrix} \int^t \begin{pmatrix} -e^{3s} & e^{3s} \\ e^s & -e^s \end{pmatrix} \begin{pmatrix} e^s \\ e^{-s} \end{pmatrix} ds.$$

The final result is

$$\begin{aligned} \mathrm{x}(t) &= c_1 \begin{pmatrix} e^{-3t} \\ e^{-3t} \end{pmatrix} + c_2 \begin{pmatrix} 2e^{-t} \\ e^{-t} \end{pmatrix} + \begin{pmatrix} e^{-3t} & 2e^{-t} \\ e^{-3t} & e^{-t} \end{pmatrix} \begin{pmatrix} -\frac{1}{4}e^{4t} + \frac{1}{2}e^{2t} \\ \frac{1}{2}e^{2t} - t \end{pmatrix} \\ &= c_1 \begin{pmatrix} e^{-3t} \\ e^{-3t} \end{pmatrix} + c_2 \begin{pmatrix} 2e^{-t} \\ e^{-t} \end{pmatrix} + \begin{pmatrix} -\frac{1}{4}e^t + \frac{1}{2}e^{-t} + e^t - 2e^{-2t} \\ -\frac{1}{4}e^t + \frac{1}{2}e^{-t} + \frac{1}{2}e^t + te^{-t} \end{pmatrix}. \end{aligned}$$

To solve the initial value problem, just plug in $t = 0$.

In the last part of class, we'll discuss the setting in which A is non-constant. That is, can we solve $\mathrm{x}'(t) = A(t)\mathrm{x}(t)$?

In the scalar case, if $y' = p(t)y$, then $y(t) = ce^{\int^t p(s)\,ds}$.

Guess. For the matrix case, let's try to do the same thing. If $B(t) = \int^t A(s)\,ds$, then is $e^{B(t)}$ is a solution?

The problem is that it does not, which probably indicates a commutativity issue. To see why, notice that we require the following equality, which turns out to be false.

$$\frac{d}{dt}\left(e^{B(t)}\right) = B'(t)e^{B(t)} = A(t)e^{B(t)}.$$

We start by taking the derivative, that is

$$\frac{d}{dt}\left(e^{B(t)}\right) = \sum_{n=0}^{\infty} \frac{d/dt(B(t))^n}{n!}.$$

But, even for $n = 2$, note that $\frac{d}{dt}(B^2(t)) \neq 2B'(t)B(t)$. We actually get

$$\frac{d}{dt}B^2(t) = B'(t)B(t) + B(t)B'(t).$$

(Note, this stuff won't be on exam / homework.)

Remark. Note that this problem is solvable if we have the condition $A(t)A(s) = A(s)A(t)$.

6.14 Lecture 14: 7-12-19

Today, we'll consider non-linear systems of ODEs. Previously, we solved the problem

$$\mathrm{x}' = A\mathrm{x}; \qquad \mathrm{x}' = A\mathrm{x} + b.$$

Now, we want to understand $\mathrm{x}' = f(\mathrm{x})$. Note that if

$$f(x, y) = \begin{pmatrix} f_1(x, y) \\ f_2(x, y) \end{pmatrix} = \begin{pmatrix} x(1 - x - y) \\ y(\frac{3}{4} - y - \frac{1}{2}x) \end{pmatrix}.$$

Then

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} x(1 - x - y) \\ y(\frac{3}{4} - y - \frac{1}{2}x) \end{pmatrix}.$$

Intuitively, we can understand this setting as the dynamics between two competing species, where

- $x(t)$ is the of wolves
- $y(t)$ is the number of foxes

Ideas.

- Understand equilibria of the system; that is find $\begin{pmatrix} x \\ y \end{pmatrix}$ constant vector or $f(x, y) = 0$.
- Understand the nature of the equilibria → understand $\mathrm{x}(t)$ where $\mathrm{x}(0)$ is close to an equilibrium. This tool is called linearization.

More ambitiously, we can understand the global behavior of the system.

What are the equilibria for $\mathrm{x}' = f(x)$. Note that $\begin{pmatrix} x \\ y \end{pmatrix}$ is an equilibrium if

$$\begin{pmatrix} x(1 - x - y) \\ y(\frac{3}{4} - y - \frac{1}{2}x) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

If $x = 0$, we get the equilibria of either $(0,0)$; $(0, \frac{3}{4})$.

Otherwise, $1 - x - y = 0$, which implies $x = 1 - y$, that is

$$y(\frac{3}{4} - y - \frac{1}{2}(1-y)) = 0,$$

giving us the equilibria of either $(1,0)$, $(\frac{1}{2}, \frac{1}{2})$.

Now, we need to understand the phase portrait around the equilibria.

If x_0 is an equilibrium, then $f(\mathrm{x}_0) = 0$. If x is close to x_0, then $f(\mathrm{x}) \approx f(\mathrm{x}_0) + J(\mathrm{x} - \mathrm{x}_0)$, where J is the Jacobian matrix of the function.

Recall that the Jacobian is the matrix where

$$J_{ij} = \frac{\partial f_i}{\partial x_j}.$$

So in this case,

$$J_f(\mathrm{x}_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix} (\mathrm{x}_0).$$

So, around x_0, we have

$$\mathrm{x}' \approx f(\mathrm{x}_0) + J_f(\mathrm{x}_0)(\mathrm{x} - \mathrm{x}_0).$$

In other words, this is using the tangent plane to approximate the multivariate function.

Now, we found four equilibria; and we have to linearize around them to obtain a phase portrait.

First, compute the Jacobian. We have that

$$J_f((x,y)) = \begin{pmatrix} 1 - 2x - y & -x \\ -\frac{y}{2} & \frac{3}{4} - 2y - \frac{x}{2} \end{pmatrix}$$

How does the phase potrait look around $(0,0)$? Should be the same as $\mathrm{x}' = \begin{pmatrix} 1 & 0 \\ 0 & \frac{3}{4} \end{pmatrix} \mathrm{x}$. And we know how to plot the phase portrait for this system. We note that

$$\lambda_1 = 1, \lambda_2 = \frac{3}{4};$$

$$\mathrm{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad \mathrm{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

And we recall that the phase portrait looks like a parabola.

What about (1, 0)? Substituting, we obtain

$$J_f(1,0) = \begin{pmatrix} -2 & -1 \\ 0 & \frac{1}{4} \end{pmatrix}.$$

Thus

$$\mathrm{x}' = \begin{pmatrix} -1 & -1 \\ 0 & \frac{1}{4} \end{pmatrix} \left(\mathrm{x} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right).$$

Solving for the eigenvectors, we have $\lambda_1 = -1, \lambda_2 = \frac{1}{4}$. And for the eigenvectors, we obtain $\mathrm{v} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathrm{w} = \begin{pmatrix} 4 \\ -5 \end{pmatrix}$.

We can plot this as the hyperbolic case (saddle).

The equilibrium $(0, \frac{3}{4})$ is pretty simple. The interesting stuff is what happens at the equilibrium $(\frac{1}{2}, \frac{1}{2})$.

In this case, we have

$$\mathrm{x}' \approx \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{2} \end{pmatrix} \left(\mathrm{x} - \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right)$$

Computing the eigenvalues, we obtain

$$\det(A - \lambda I) = \det \begin{pmatrix} -\frac{1}{2} - \lambda & -\frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{2} - \lambda \end{pmatrix} = \lambda^2 + \lambda + \frac{1}{4} - \frac{1}{8} = \lambda^2 + \lambda + \frac{1}{8}.$$

So the roots are

$$\begin{aligned} \lambda = \frac{-1 \pm \sqrt{1 - \frac{1}{2}}}{2} &= \frac{-1 \pm \frac{\sqrt{2}}{2}}{2} \\ &= \frac{-2 \pm \sqrt{2}}{4}. \end{aligned}$$

Since both eigenvalues are negative, the picture is qualitatively stable. And then, assuming continuity and uniqueness, we should guess that we can “interpolate” between the phase portraits we know how to solve to plot the full phase portrait.

Globally, it turns out that the phase portrait will converge to $(\frac{1}{2}, \frac{1}{2})$, unless you have 0 wolves or 0 foxes.

To show this, we can study the direction field. If we have the problem $\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x(1 - x - y) \\ y(\frac{3}{4} - y - \frac{x}{2}) \end{pmatrix}$.

If the components are positive, we know that the direction field is increasing. We can draw all the 4 cases based on the signs of the components.

If $x(1 - x - y) > 0, y(\frac{3}{4} - y - \frac{x}{2}) > 0$, direction field goes northeast.

If $x' < 0, y' > 0$, direction goes northwest.

If $x' > 0, y' < 0$, direction goes southeast.

If $x' < 0, y' < 0$, direction goes southwest.

Sometimes, it is possible to get complicated oscillations and cycles. But in this case, we happen to know that we won't get these.

In this example, we did two competing species (both predator).

We can consider a more complicated setting where we have prey and predator. That is,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x(1-2y) \\ y(-\frac{3}{4} + \frac{1}{x}) \end{pmatrix}.$$

In this case, y represent the predator (naturally tends to die), and x represents the prey (naturally tends to grow).

To solve for the equilibria, we consider the cases:

- $x = 0, y = 0$.
- $x = 3, y = \frac{1}{2}$ (will be a center).

Computing, we obtain

$$J_f(x,y) = \begin{pmatrix} 1-2y & -2x \\ \frac{y}{4} & \frac{x}{4} - \frac{3}{4} \end{pmatrix}$$

So, at $(0,0)$, $J_f(0,0) = \begin{pmatrix} 1 & 0 \\ 0 & -\frac{3}{4} \end{pmatrix}$. Here, the eigenvalues are $1, -\frac{3}{4}$, with eigenvectors $\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

At $(3, \frac{1}{2})$, $J = \begin{pmatrix} 0 & -6 \\ \frac{1}{8} & 0 \end{pmatrix}$. Computing the eigenvalues, $\lambda^2 = -\frac{3}{4}$, thus $\lambda = \pm\frac{\sqrt{3}}{2}i$. This is the case where $\alpha = 0$, so this will trace out an ellipse.

But note that it could be the case where a small perturbation results in a nonlinearity. In this case, we can actually show that this isn't true, but it is a concern.

Next week, when we discuss separable equations, we will discuss how to obtain the exact shape of these curves.

Let's try to convince ourselves that going circularly around the equilibrium $(3, \frac{1}{2})$ works in terms of the direction field.

The error of the linear approximation is related to the second derivative. So if f is a crazy function, this might have significant errors.

6.15 Lecture 15: 7-15-19

Came in late, summarizing key parts of lecture.

Example. Suppose $\frac{dy}{dt} = \frac{t^2}{1-y^2}$, $y(0) = 0$. Then

$$dy(1-y^2) = dtt^2,$$

so

$$y - \frac{y^3}{3} = \frac{t^3}{3} + C,$$

and plugging in the initial condition, we obtain $C = 0$.

Example. Suppose

$$\frac{dy}{dt} = \frac{3t^2 + 4t + 2}{2(y-1)},$$

with $y(0) = y_0$.

Then

$$2(y-1)dy = (3t^2 + 4t + 2)\, dt,$$

that is

$$y^2 - 2y = t^3 + 2t^2 + 2t + C.$$

Plugging in initial condition, we get

$$y_0^2 - 2y_0 = C,$$

so

$$y^2 - 2y = t^3 + 2t^2 + 2t + y_0^2 - 2y_0.$$

We can solve for y using quadratic formula, that is

$$y^2 - 2y + 1 = t^3 + 2t^2 + 2t + y_0^2 - 2y_0 + 1.$$

Then

$$y(t) = \pm\sqrt{t^3 + 2t^2 + 2t + y_0^2 - 2y_0 + 1} + 1.$$

At least for t small (so that $\sqrt{}$ is defined), we get two solutions to the ODE. Only one matches the initial condition.

Part a). And given $y_0 = 2$, we get $y(t) = 1 \pm \sqrt{t^3 + 2t^2 + 2t + 1}$.

$$y(t) = 1 + \sqrt{1 + t^2 + 2t^2 + 2t},.$$

Part b). And given $y_0 = 0$, we get $y(t) = 1 \pm \sqrt{t^3 + t^2 + 2t + 1}$. In that case $y(t) = 1 - \sqrt{t^3 + 2t^2 + 2t + 1}$.

Part c). And given $y_0 = 1$, we get $y(t) = 1 \pm \sqrt{t^3 + 2t^2 + 2t}$. We have two distinct solutions to the ODE.

6.16 Lecture 16: 7-16-19

This lecture focuses on midterm review.

Let $A = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$. Calculate the eigenvectors and eigenvalues, and e^{tA}.

We obtain $\lambda = \pm 2$, and $\mathrm{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mathrm{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Now,

$$e^{tA} = \begin{pmatrix} e^{2t} & e^{-2t} \\ e^{2t} & -e^{-2t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{-1},$$

we obtain

$$\begin{pmatrix} \frac{1}{2}\left(e^{2t} + e^{-2t}\right) & \frac{1}{2}(e^{2t} - e^{-2t}) \\ \frac{1}{2}(e^{2t} - e^{-2t}) & \frac{1}{2}(e^{2t} + e^{-2t}) \end{pmatrix}.$$

Recall that if we ask for the general solution of $\mathrm{x}' = A\mathrm{x}$, we write down

$$x(t) = c_1 \begin{pmatrix} e^{2t} \\ e^{2t} \end{pmatrix} + c_2 \begin{pmatrix} e^{-2t} \\ -e^{-2t} \end{pmatrix}.$$

Without computing the determinant, we know that this matrix is invertible, since the eigenvalues are distinct.

There are multiple ways to see that this is invertible, since $e^{tA} = Oe^{tA}O^{-1}$.

Let $\mathrm{x}' = A\mathrm{x} + \begin{pmatrix} e^{t} \\ e^{-t} \end{pmatrix}$, where $\mathrm{x}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. What is the solution to this IVP?

We just add $A^{-1}b$ to the solution we just got. That is, take $\frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} e^{t} \\ e^{-t} \end{pmatrix}$.

First, we can determine the general solution from before,

$$\mathrm{x}(t) = c_1 \begin{pmatrix} e^{2t} \\ e^{2t} \end{pmatrix} + c_2 \begin{pmatrix} e^{-2t} \\ -e^{-2t} \end{pmatrix} + X(t) \int^{t} X^{-1}(s) q(s)\, ds.$$

So $X^{-1}(t) = \frac{1}{2}\begin{pmatrix} e^{-2t} & e^{-2t} \\ e^{2t} & e^{-2t} \end{pmatrix}$. And we can compute to get the result.

Example. Let $A = \begin{pmatrix} 7 & -5 & -2 \\ 3 & -1 & -1 \\ 1 & -1 & 2 \end{pmatrix}$, $\mathrm{v} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$, $\mathrm{w} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$.

Compute $A\mathrm{v}$, and compute $(A - 3I)^2\mathrm{w}$.

We have that

$$A\mathrm{v} = \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix},$$

so v is an eigenvector with eigenvalue 2.

Also,

$$\begin{aligned} (A - 3I)^2\mathrm{w} &= \begin{pmatrix} 4 & -5 & -2 \\ 3 & -4 & -1 \\ 1 & -1 & -1 \end{pmatrix}^2 \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \\ &= 0. \end{aligned}$$

Find the Jordan canonical form. To compute it, use the eigenvalues, eigenvectors, and generalized eigenvectors.

Remember when we write $A = OJO^{-1}$, we know that the columns of O are eigenvectors or generalized eigenvectors. To construct J, put the eigenvalues in the diagonal; and if you have a generalized eigenvector, you put a 1 above.

Since $(A - 3I)\mathrm{w} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \neq 0$, w is a generalized eigenvector with eigenvalue 3. Since $(A - 3I)\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} = (A - 3I)^2\mathrm{w} = 0$, we know $\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue 3. So if you have a generalized eigenvector of order k, you can get $k - 1$ generalized eigenvectors (of which the last eigenvector is a a genuine eigenvector).

Now, the final decomposition is

$$A = \begin{pmatrix} 1 & 3 & 2 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 3 & 2 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}^{-1}.$$

You should put the generalized eigenvector last in the columns of O, and in increasing order. The matrix O is invertible, because distinct eigenvalues implies eigenvectors (including generalized ones) are independent.

Proof. Let's say v satisfies $(A - \lambda)\mathrm{v} = 0$ and w satisfies $(A - \lambda)^2\mathrm{w} = 0$, $(A - \lambda)\mathrm{w} \neq 0$. Then, if you have the system

$$c_1\mathrm{v} + c_2\mathrm{w} = 0,$$

and we apply $(A - \lambda)$, we get $c_2(A - \lambda)\mathrm{w} = 0$, so $c_2 = 0$. Thus v, w are linearly independent.

More generally, you have to apply $(A - \lambda_1)(A - \lambda_2)$, (or the powers of products of $(A - \lambda_i)$), and you kill each vector separately. □

We will talk about phase portraits on Thursday.

Example. Let's consider the problem $ty' - y = 1$, with $y(1) = 0$. Solve the IVP.

This is just a linear ODE equivalent to $y' - \frac{y}{t} = \frac{1}{t}$, and $\mu(t) = \exp\left(\int (-\frac{1}{t})\,dt\right)$, so $\mu(t) = \frac{1}{t}$. Then

$$\frac{y'}{t} - \frac{y}{t^2} = \frac{1}{t^2},$$

so

$$(\frac{y}{t})' = \frac{1}{t^2},$$

and

$$\frac{y}{t} = -\frac{1}{t} + C,$$

so that

$$y = -1 + tC.$$

Since $y(1) = 0$, we conclude that $C = 1$ and the solution is $y = -1 + t$. This is the unique solution in the interval $(0, \infty)$ (since there is a discontinuity at $t = 0$).

If $ty' - y = 1, y(0) = 0$, we cannot solve this, since at $t = 0$, we always get -1. There is no solution. And if we set $y(0) = -1$, we have infinitely many solutions (since we can choose any value of C.)

Question. How many solutions does $y' - \frac{e^{-t^2}}{\cos t} y = e^{2t - \cos t}$, $y(37) = 54$. We only need p and q continuous. We only need to check $\cos t \neq -2$, but this never happens. So this problem has only 1 solution on the whole line.

Definition. Recall how to compute the matrix exponential. If e^{tA}, then $e^{tA} = X(t)X^{-1}(0)$ where $X(t)$ is a fundamnetal solution for A. Alternatively, we can observe that $e^{tA} = Oe^{t\Lambda}O^{-1}$. Note that e^{tA} is the uniqu ematrix satisfying $\mathrm{x}' = A\mathrm{x}$ with $\mathrm{X}(0) = Id$.

On Thursday, we'll cover phase portraits / direction fields. Andrea will put practice exam

6.17 Review for midterm

- Direction fields for scalar ODEs.
- Linear scalar ODEs: existence and uniqueness theory, integrating factors.
- Complex numbers (no branch cut).
- Basic notions of linear algebra: determinant, inverse matrix, eigenvalues, eigenvectors, generalized eigenvectors, diagonal/Jordan canonical form, exponential of a matrix.

Review the JCF, and how it is used to solve an ODE / system.

- Linear system of ODEs of the form x'=Ax, x'=Ax+b, for A a constant matrix and b a constant vector. General solution and solution to the initial value problem. Associated phase portraits. Wronskian, fundamental set of solutions, fundamental matrix.

  A fundamental matrix is a matrix valued function $X(t)$ where the columns are linearly independent solutions of the system.

Review phase portraits

- Linear inhomogeneous systems of the form x'=Ax+b(t): solution via variation of parameters (i.e., integrating factors for system of ODEs).

  Recall that the solution is

  $$\mathrm{x}(t) = \mathrm{x}_{hom}(t) + X(t)\int^t X^{-1}(s)q(s)\,ds.$$

  Suppose we have the problem $\mathrm{x}' = A\mathrm{x} + \mathrm{g}(t)$. With the matrix exponential, we obtain

  $$e^{-tA}\mathrm{x}' - e^{-tA}A\mathrm{x} = e^{-tA}\mathrm{g}.$$

  Now, the columns of e^{-tA} solve the problem $A\mathrm{x} = \mathrm{x}'$, so the above equation reduces to

  $$(e^{-tA}\mathrm{x})' = e^{-tA}\mathrm{g},$$

  and therefore

  $$e^{-tA}\mathrm{x} = \int^t e^{-sA}g(s)ds + \mathrm{c}.$$

  In particular,

  $$\mathrm{x} = e^{tA}\mathrm{c} + e^{tA}\int^t e^{-sA}\mathrm{g}.$$

Derive this in the same form as linear ODE

6.18 Lecture 17: 7-17-19

Let $y' = f(t, y)$, where $y(t_0) = y_0$.

Theorem 1. If f is continuous in a neighborhood of (t_0, y_0), then $\exists$ at least one solution to the above equation for $(t - t_0)t$ small enough.

Theorem 2. If $f, \frac{df}{dy}$ are continuous in a neighborhood of (t_0, y_0), then there exists a unique solution for $(t - t_0)$ small enough.

Various examples to guarantee existence / uniqueness.

Remark 1. For $y' + p(t)y = q(t)$, the old theorem said "if p, q are continuous at t_0, the exists a unique solution.

In this case, $f(t, y) = -p(t)y + q(t)$, and $\frac{\partial f}{\partial y} = -p(t)$, so the continiuity of f and it's derivative is equivalent to p, q continuous.

Consider $y' = y^2, y(0) = y_0 \neq 0$. Recall that we can solve this with separation of variables.

Counterexample, consider $y'1$ if $y \geq 0$, $y' = -1$ if $t < 0$. There is no solution that is differentiable (absolute value doesn't work).

Stable / unstable equilibrium. If you start from near the equilibrium, you escape (unstable), and stable means converge to an equilibrium.

- If $f'(c) > 0$, then c is an unstable equilibrium.
- If $f'(c) < 0$, then c is a stable equilibrium.
- If $f'(c) = 0$, we cannot say anything.

Example, if $y' = y(1-y) = f(y)$, you know that the equilibria are solutions to $y(1-y) = 0$, where $y = 0, y = 1$.

Example. Suppose $y' = r\left(1 - \frac{y}{K}\right)y$, where r is the birthrate, where the max capacity is K.

Recall that to integrate do fraction decomposition: $\frac{1}{(1-y)y} = \frac{A}{y} + \frac{B}{1-y}$.

This lecture concludes first order ODEs. Tomorrow we'll do review, Friday there's a midterm, and next week there are second order ODEs.

6.19 Lecture 18: 7-18-19 (Midterm review)

Recall that generalized eigenvectors are associated with repeated eigenvalues.

Problem 1.

Now, to find the fundamental solution of $y' = Jy$, we get the system

$$a'(t) = 2a(t)$$
$$b'(t) = 3b(t) + c(t)$$
$$c'(t) = 3c(t).$$

Solving, we easily obtain $a(t) = c_1e^{2t}$, $c(t) = c_3e^{2t}$. To solve for $b(t)$, we know that $b(t) = f(t) + c_2e^{3t}$, where c_2e^{3t} is a solution to the homogeneous problem, and $f(t)$ is a particular function.

In this case, integrating factor is e^{-3t}, and this gives us

$$(b(t)e^{-3t})' = c_3.$$

Thus $b(t) = c_3te^{3t} + c_2e^{3t}$.

Idea: if $y(t)$ is a solution to $y' = Jy$, then $x(t) = Oy(t)$ is a solution to $x' = Ax$.

Now, if y_1, y_2, y_3 are three independent solutions to $y' = Jy$, then Oy_1, Oy_2, Oy_3 are independent solutions to $x' = Ax$.

Obvious idea, want to check if

$$\begin{pmatrix} e^{2t} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ e^{3t} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ te^{3t} \\ e^{3t} \end{pmatrix}$$

in independent. At time 0, the matrix is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and this is clearly invertible (since determinant is not zero; identity matrix).

There is some subtlety; got to check that O is invertible (and compute the determinant).

Problem 2.

(a) The eigenvalues are $1 \pm \sqrt{a}$.

(b) If $a = -1$, eigenvalues $1 \pm i$, so the phase portrait is an outward spiral.

If $a = 0$, then eigenvalue is 1, you have an unstable equilibrium.

If $a = \frac{1}{4}$, then the eigenvalues are $1 \pm \frac{1}{2} = \frac{1}{2}, \frac{3}{2}$.

Problem 3.

(a) Computation.

(b)

Problem 4.

(a) False.

(b) False.

(c) False. (e.g. $y' = t$).

Recall, stuff like $x' = Ax + g(t)$ will be on the midterm. Use integrating factor for matrices.

Recall that $y' + p(t)y = q(t)$, $y(t_0) = y_0$, you have existence an uniqueness as long as p, q are continuous (in an interval around t_0).

6.20 Midterm practice

6.20.1 Review A

- Direction fields for scalar ODEs.

  Just plug in slopes for scalar ODEs, and draw the vectors with those slopes at each point.

- Linear scalar ODEs: existence and uniqueness theory, integrating factors.

  If $y' + p(t)y = q(t)$, with $y(t_0) = y_0$, we have existence and uniqueness as long as p, q are continuous (in an interval around t_0).

- Complex numbers (no branch cut).

  Pretty simple. The main thing to review is $\log z_1$ and $z_1^{z_2}$. Remember that

$$\log z = \log r + i\theta.$$

Now,

$$\begin{aligned} z_1^{z_2} &= (e^{\log z_1})^{z_2} \\ &= e^{z_2(\log z_1)}. \end{aligned}$$

- Basic notions of linear algebra: determinant, inverse matrix, eigenvalues, eigenvectors, generalized eigenvectors, diagonal/Jordan canonical form, exponential of a matrix.
  - Eigenvalues / eigenvectors satisfy $Av = \lambda v$.
  - Generalized eigenvectors. A vector v is a GE if $(A - \lambda I)^k v = 0$ for some integer k.
  - Diagonalization. Recall that a matrix is diagonalizable iff it has n linearly independent eigenvectors.
  - Any matrix A can be put into JCF with a similarity transform. That is, $A = OJO^{-1}$, where O is invertible, and

    $$J = (J_1, J_2, \ldots, J_q),$$

    where J_i is a block defined as follows:

    $$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}.$$

  - Exponential of a matrix $e^A = 1 + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots$. The easiest way to compute this is diagonalization (when a matrix is diagonalizable).

    Other cases include the nilpotent case (which ends up being a finite sum of matrix powers).

    Alternatively, it is possible to do with JCF decomposition.

Review the JCF, and how it is used to solve an ODE / system.

- Linear system of ODEs of the form x'=Ax, x'=Ax+b, for A a constant matrix and b a constant vector. General solution and solution to the initial value problem. Associated phase portraits. Wronskian, fundamental set of solutions, fundamental matrix.
  - $x' = Ax$ is relatively easy to solve.
    - If eigenvalues are both real, then solution is

      $$c_1 e^{\lambda_1 t} \mathrm{v}_1 + c_2 e^{\lambda_2 t} \mathrm{v}_2.$$

    - If eigenvalues are both complex, the solution is

      $$c_1 e^{\lambda_1 t} \mathrm{v}_1 + c_2 e^{\lambda_2 t} \mathrm{v}_2,$$

      and then have to take real and imaginary parts of this solution to get a linearly independent combination of real solutions.

    - If eigenvalues are repeated but there are two linearly independent vectors, then solution is just

      $$c_1 e^{\lambda t} \mathrm{v}_1 + c_2 e^{\lambda t} \mathrm{v}_2.$$

    - If eigenvalues are repeated but eigenvalues are not linearly independent, then solution is a bit different.

  - To solve $x' = Ax + b$, we have to find a particular solution $-A^{-1}b$, and then add this to the general solution of the homogeneous problem.

  - Phase portraits

    Got to review

  - Fundamental set of solutions; this is a set of n solutions in the $n \times n$ that are linearly independent.

  - Fundamental matrix, this is a matrix whose columns are linearly independent solutions of the system.

  A fundamental matrix is a matrix valued function $X(t)$ where the columns are linearly independent solutions of the system.

  Review phase portraits

- Linear inhomogeneous systems of the form x'=Ax+b(t): solution via variation of parameters (i.e., integrating factors for system of ODEs).

  Recall that the solution is

  $$\mathrm{x}(t) = \mathrm{x}_{hom}(t) + X(t) \int^t X^{-1}(s) q(s)\, ds.$$

  Suppose we have the problem $\mathrm{x}' = A\mathrm{x} + \mathrm{g}(t)$. With the matrix exponential, we obtain

  $$e^{-tA}\mathrm{x}' - e^{-tA}A\mathrm{x} = e^{-tA}\mathrm{g}.$$

  Now, the columns of e^{-tA} solve the problem $A\mathrm{x} = \mathrm{x}'$, so the above equation reduces to

  $$(e^{-tA}\mathrm{x})' = e^{-tA}\mathrm{g},$$

  and therefore

  $$e^{-tA}\mathrm{x} = \int^t e^{-sA} g(s) ds + \mathrm{c}.$$

  In particular,

  $$\mathrm{x} = e^{tA}\mathrm{c} + e^{tA} \int^t e^{-sA}\mathrm{g}.$$

6.20.2 Review B

- Linear scalar ODEs.

  Recall that a linear ODE is given by $y'(t) + p(t)y(t) = q(t)$, and you solve by multiplying by $\mu(t) = \exp(\int p(t))$. Use product rule.

- Complex numbers (how to calculate $\log z$ and $z_1^{z_2}$.

  Recall that

  $$\log z = \log |z| + i\theta,$$

  where θ technically depends on branch cut.

- Generalized eigenvector

  A vector is a generalized EV if $(A - \lambda I)^k v = 0$.

- Diagonalizable vs. defective.

  A matrix is diagonalizable iff it has n linearly independent eigenvectors. Defective otherwise.

- JCF.

  We can factor any matrix as $A = OJO^{-1}$, where J consists of diagonal blocks, each block corresponds to an λ_i, where the blocks look like diagonal with 1s above.

- Matrix exponential.

  Defined as

  $$e^A = 1 + A + \frac{A^2}{2!} + \dots.$$

- Solve defective $x' = Ax$, when there are repeated eigenvalues.

  When repreated eigenvalues; one solution is $c_1 e^{\lambda t}\mathrm{v}$. Other solution is $c_2(te^{\lambda t}\mathrm{v} + e^{\lambda t}\mathrm{w})$.

- Variation of parameters for $x' = Ax + g(t)$.

  To solve this, solution is given by $x_{hom} + X^{-1}\int^t X(s)g(s)\,ds$ (very easy to derive).

- Phase portraits.

Need to review

6.20.3 Review C

- Complex numbers.

  We have

  $$\log z = \log |z| + i\theta.$$

- JCF. Any matrix can be factorized as OJO^{-1}, where J is a Jordan matrix. That is, J consists of blocks $J_1, \dots, J_p$, where each block has a λ_i on the diag, with a 1 right above it. The 1's correspond to a generalized eigenvector.

- Phase portraits.

  There are a couple of cases to consider.

6.21 Lecture 19: 7-22-19

We can define a second order DE as

$$y'' = f(t, y, y').$$

In this setting an IVP is $y'' = f(t, y, y'), y(t_0) = y_0, y'(t_0) = y_0'$.

We can define various terms from before in this setting:

- Linear 2nd order DE is $a(t)y'' + b(t)y' + c(t)y = g(t)$.
- For linear, if y_1, y_2 are solutions, then $y_1 - y_2$ is a solution to the associated homogeneous problem.

How do we solve linear homogenous constant coefficient 2nd order DES?

- Look for $y(t) = e^{\lambda t}$, substitute, and then solve $a\lambda^2 + b\lambda + c = 0$. In general, we expect λ_1, λ_2 two solutions, which means

  $$y(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}.$$

- If $x_1(t) := y(t), x_2(t) := y(t)$, then we can write $\mathrm{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$, and we can set up the system

  $$x'(t) = \begin{pmatrix} x_2(t) \\ -\frac{b}{a}x_2(t) - \frac{c}{a}x_1(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{c}{a} & -\frac{b}{a} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = A\mathrm{x}.$$

If we try this approach, we get the same formula we got from the first approach. We can obtain a characteristic polynomial

$$a\lambda^2 + b\lambda + c = 0,$$

so if λ_1, λ_2 are two solutions, we obtain

$$\mathrm{x}(t) = c_1 e^{\lambda_1 t} \begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix} + c_2 e^{\lambda_2 t} \begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix}.$$

Example. Suppose we have the problem $my''+ky = 0$. That is, $y'' = -\frac{k}{m}y'$, so we can obtain a system

$$\mathrm{x}(t) = \begin{pmatrix} y(t) \\ y'(t) \end{pmatrix}$$

Then this reduces to $\mathrm{x}' = A\mathrm{x}$, where $A = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{pmatrix}$.

Now, the eigenvalues satisfy $\lambda^2 + \frac{k}{m} = 0$, so $\lambda + \pm i\sqrt{\frac{k}{m}}$.

We get that the eigenvectors are

$$\begin{pmatrix} 1 \\ i\sqrt{\frac{k}{m}} \end{pmatrix}, \begin{pmatrix} 1 \\ -i\sqrt{\frac{k}{m}} \end{pmatrix}.$$

Thus we get the solution:

$$e^{iw_0 t}\begin{pmatrix} 1 \\ i\omega_0 \end{pmatrix} = e^{iw_0 t}\begin{pmatrix} 1 \\ i\omega_0 \end{pmatrix};$$

and we can obtain the real solutions, which form a fundamental set:

$$\begin{pmatrix} \cos\omega_0 t \\ -\omega_0\sin\omega_0 t \end{pmatrix}, \begin{pmatrix} \sin\omega_0 t \\ w_0\cos\omega_0 t \end{pmatrix}.$$

So the general solution of $my'' + ky = 0$ is

$$A\cos(\omega_0 t) + B\sin(\omega_0 t); \qquad \omega_0 = \sqrt{\frac{k}{m}},$$

where:

- ω_0 is the frequency.
- $T = \frac{2\pi}{\omega_0}$ is the period.

Usually on physics books the solution is given by $y(t) = R\cos(\omega_0 t - \delta)$, where R is the amplitude and δ is the phase. The question is: how do we relate A and B to R and δ?

Recall that $\cos(a - b) = \cos a\cos b + \sin a\sin b$. Now, equating the values on both sides, we obtain

$$R\cos(\omega_0 t - \delta) = R\cos\omega_0 t\cos\delta + R\sin\omega_0 t\sin\delta = A\cos\omega_0 t + B\sin\omega_0 t,$$

so we can compute $R\cos\delta = A, R\sin\delta = B$, and thus $R = \sqrt{A^2 + B^2}$. Then, δ is determined by

$$\cos\delta = \frac{A}{\sqrt{A^2 + B^2}}; \qquad \sin\delta = \frac{B}{\sqrt{A^2 + B^2}}.$$

Example. Suppose we hve $m = 3$ kg, $k = 12\frac{N}{m}$. Initially, the spring is at distance 4m from the origin to the left and has initial speed $8\frac{m}{s}$.

Find the amplitude and the phase. If you solve $my'' + ky = 0$, you obtain $\omega_0 = 2$, and thus $y(t) = A\cos(2t) + B\sin(2t)$. At time $t = 0$, $-4 = A$, $8 = 2B$, so

$$y(t) = -4\cos(2t) + 4\sin(2t),$$

so $R = 4\sqrt{2}$, and thus $\cos\delta = -\frac{1}{\sqrt{2}}$, $\sin\delta = \frac{1}{\sqrt{2}}$, so $\delta = \frac{3\pi}{4}$.

6.22 Lecture 20: 7-23-19

Consider the problem $my'' = -ky - \gamma y'$. To solve this, we need to turn this into a system of ODEs. We can write

$$\mathrm{x} = \begin{pmatrix} y \\ y' \end{pmatrix},$$

$$\mathrm{x}' = \begin{pmatrix} y' \\ y'' \end{pmatrix} = \begin{pmatrix} y' \\ -\frac{k}{m}y - \frac{\gamma}{m}y' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{\gamma}{m} \end{pmatrix} \begin{pmatrix} y \\ y' \end{pmatrix} = A\mathrm{x}.$$

Then, solving for the roots the characteristic polynomial, we obtain

$$-\lambda\left(-\lambda - \frac{\gamma}{m}\right) + \frac{k}{m} = 0,$$

so $\lambda^2 + \frac{\gamma}{m}\lambda + \frac{k}{m} = 0$. Then,

$$\lambda_{1,2} = \frac{-\frac{\gamma}{m} \pm \sqrt{\frac{\gamma^2}{m^2} - 4\frac{k}{m}}}{2}.$$

From here, we can consider cases on the value of the discriminant.

1. Case 1. $\gamma^2 - 4km < 0$. This is the underdamped oscillator.
2. Case 2. $\gamma^2 - 4km = 0$. In this case, A is defective, so the only eigenvalue is $\lambda = -\frac{\gamma}{2m}$. Then

$$A = \begin{pmatrix} 0 & 1 \\ -\frac{\gamma^2}{4m^2} & -\frac{\gamma}{m} \end{pmatrix}.$$

So, if v is an eigenvector for $-\frac{\gamma}{2m}$, then

$$\left(A + \frac{\gamma}{2m}I\right)\mathrm{v} = 0,$$

and solving this we obtain $\mathrm{v} = \begin{pmatrix} 1 \\ -\frac{\gamma}{2m} \end{pmatrix}$.

Finally, if w is a generalized eigenvector, so $w_1 = 0, w_2 = 1$. Finally, we obtain the general solution

$$\mathrm{x}(t) = c_1 e^{-\frac{\gamma}{2m}t}\begin{pmatrix} 1 \\ -\frac{\gamma}{2m} \end{pmatrix} + c_2\left(te^{-\frac{\gamma}{2m}t}\begin{pmatrix} 1 \\ -\frac{\gamma}{2m} \end{pmatrix} + e^{-\frac{\gamma}{2m}t}\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)$$

This solution is known as the critically damped oscillator. Asymptotically, this reaches 0, but not with 0 speed.

3. Case 3. $\gamma^2 - 4km > 0$. In this case, we'll obtain two negative eigenvalues. In particular, we will get

$$\lambda_{1,2} = -\frac{\gamma}{2m} \pm \frac{1}{2m}\sqrt{\gamma^2 - 4km} < 0.$$

   We can write $\lambda = \mu nu$, where $\mu = \frac{\gamma}{2m}, \nu = \frac{\sqrt{\gamma^2-4km}}{2m}$.

   Solving, the eigenvector relative to $\mu + \nu$ satisfies

$$(A - \mu - \nu)\mathrm{v} = 0,$$

   which is equivalent to $(-\mu - \nu)\, v_1 + v_2 = 0$, so $\mathrm{v} = \begin{pmatrix} 1 \\ \mu + \nu \end{pmatrix}$.

   Simliarly, the eigenvector w relative to $\mu - \nu$ satisfies

$$(A - \mu + \nu)\mathrm{w} = 0,$$

   which is equivalent to $(-\mu + \nu)\, v_1 + v_2 = 0$.

   Solving, we obtain the general solution

$$\mathrm{x}(t) = c_1 e^{(\mu+\nu)t} \begin{pmatrix} 1 \\ \mu + \nu \end{pmatrix} + c_2 e^{(\mu-\nu)t} \begin{pmatrix} 1 \\ \mu - \nu \end{pmatrix}.$$

   This is the overdamped oscillator.

We'll now discuss the nonhomogeneous 2nd order equation. Suppose we have $y'' + y = \sin 2t$. Then we can write $y'' = -y + \sin 2t$, and apply $F = ma$.

We start with the simple case $y'' = -ky + mg$. In this case, we will get the system $Ax + \begin{pmatrix} 0 \\ g \end{pmatrix}$, that is, the solution to the homogeneous problem plus some shift. We can write $A = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{pmatrix}$. The solutions, explicitly, are

$$\mathrm{x}(t) = c_1 \begin{pmatrix} \cos \omega t \\ \sin \omega t \end{pmatrix} + c_2 \begin{pmatrix} -\sin \omega t \\ \cos \omega t \end{pmatrix} - A^{-1}b.$$

that is, one particular solution plus the solution to the homogeneous problem. Inverting, we obtain

$$-A^{-1}b = \begin{pmatrix} \frac{m}{k}g \\ 0 \end{pmatrix}.$$

Importantly, in equilibrium, the second component should always be zero (since the speed is 0). Intuitively, instead of oscillating around 0, you oscillate around other point.

Now, we return to the original problem, $y'' + y = \sin 2t$. One way to solve this is to phrase the problem as a system of first order ODEs, and use variation of parameters.

We now discuss the method undetermined coefficients. By the usual method, the solution will be $y(t) = y_p(t) + c_1 \sin t + c_2 \cos t$, where $c_1 \sin t + c_2 \cos t$ is a solution to $y'' + y = 0$.

In this case, we can plug in $C \sin 2t$, and try to solve for C to obtain the coefficients. Conceretely, if we apply this, we get

$$-4C \sin 2t + C \sin 2t = \sin 2t.$$

Then $-3C - 1 = 0$, so $C = -\frac{1}{3}$. Thus, a particular solution is $-\frac{1}{3} A \sin 2t$.

Example. Suppose $y'' + 2y' + 2y = \cos t$. As before, the general solution is going to be

$$y(t) = x_p + x_{hom}.$$

For the homogeneous problem, note that the characteristic polynomial is $\lambda^2 + 2\lambda + 2 = 0$, so we get $\lambda = -1 \pm i$. Thus, the solution is

$$y(t) = x_p + c_1 e^{-t} \cos t + c_2 e^{-t} \sin t.$$

Now, for the particular solution, suppose that the x_p is of the form $A \cos t + B \sin t$. Then

$$-A \cos t - B \sin t + 2 \left(-A \sin t + B \cos t\right) + 2 \left(A \cos t + B \sin t\right) = \cos t.$$

In particular, this reduces to

$$(B - 2A) \sin t + (A + 2B) \cos t = \cos t.$$

We get the system $A + 2B = 1, B - 2A = 0$, so $A = \frac{1}{5}, B = \frac{2}{5}$. The particular solution is $\frac{1}{5} \cos t + \frac{2}{5} \sin t$.

Of course, there are some weaknesses of this approach; you need an "eigenfunction" of sorts to do this (for example, if the right hand side is $\log t$, this can't really work).

If you have $y'' + y = \sin t$, then the general solution is

$$y(t) = c_1 \cos t + c_2 \sin t + y_p.$$

Now, we have a problem with assuming $y = c \sin t$, since this is already in the general solution. Alternatively, we can try $y_p = c_1 t \sin t$. In this case, we get

$$c_1 t \sin t + c_1 \left(\sin t - t \cos t\right)' = \sin t,$$

and simplifying, we get $2 \cos t = \sin t$. But this doesn't quite work. So alternatively, we can guess $y = ct \cos t$, things work out - you end up with $-\frac{1}{2} t \cos t$.

You can show that every time you have a solution on the RHS that is in the kernel of the LHS, if you multiply by t, this method should work in general.

This isn't in the homework, but it's really helpful to have this method. (And various kinds of oscillations will likely not be on the exam.)

6.23 Lecture 22: 7-25-19

Today, we will discuss Laplace transforms. The broad idea is that it is often easier to solve the ODE in terms of the Laplace transform instead of the original problem.

Recall that when we were studying first order ODEs, we introduced a change of variables from $\mathrm{x}' = Ax$; so $y = -O^{-1}x$, solve the new problem, and then apply the inverse transform.

The Laplace transform is a transform that depends on time.

Definition. If $f(t)$ is a function on $[0, \infty)$, we defined the Laplace transform as

$$\mathcal{L}(F)(s) = \int_0^\infty f(t)e^{-st}\,dt.$$

Question: can you think about the Laplace transform in terms of expectations?

Observation. The Laplace transform is the continuous analogue of power series. If $f(x) = \sum_{n=0}^\infty a_n x^n$, the continuous version is $\int_0^\infty f(t)x^t$, and changing the base of the power from x to e gives $\int_0^\infty f(t)(e^{\ln x})^t$, or, making the substitution $-s = \ln x$, we get $\int_0^\infty f(t)e^{-st}\,dt$.

Example. If $f(t) = 1$, then

$$\mathcal{L}(f)(s) = \int_0^\infty e^{-st}\,dt = -\frac{1}{s}e^{-st}\bigg|_{t=0}^{\infty} = \frac{1}{s},$$

where $s > 0$.

Example. If $f(t) = e^{at}$, we can write

$$\mathcal{L}(f(s)) = \int_0^\infty e^{(a-s)t}\,dt = \frac{1}{s-a}.$$

Proposition. The Laplace transform is linear. Roughly speaking, $\mathcal{L}(c_1F_1(s) + c_2F_s(s)) = c_1\mathcal{L}(F_1(s)) + c_2\mathcal{L}(F_2(s))$ (provided both LHS and RHS are finite).

Proof. Just use the fact that the integral is linear.

Example. Suppose $f(t) = \cos bt$. One way to do this is to write

$$\mathcal{L}(f(s)) = \int_0^\infty \cos bte^{-st}\,dt,$$

and just compute this using integration by parts. Alternatively, we can note that $e^{ibt} = \cos bt + i\sin bt$, and $e^{-ibt} = \cos bt - i\sin bt$.

Then, we can write

$$\begin{aligned}\mathcal{L}(f(s)) &= \int_0^\infty \frac{e^{ibt}+e^{-ibt}}{2}e^{-st}\,dt = \frac{1}{2}\mathcal{L}(e^{ibt}) + \frac{1}{2}\mathcal{L}\left(e^{-ibt}\right) = \frac{1}{2(s-ib)} + \frac{1}{2(s+ib)} \\ &= \frac{s}{(s-ib)(s+ib)} = \frac{s}{s^2+b^2}, \qquad s > 0.\end{aligned}$$

Some important Laplace transforms:

- $\mathcal{L}(1) = \frac{1}{s}, s > 0.$
- $\mathcal{L}(e^{at}) = \frac{1}{s-a}, s > 0.$
- $\mathcal{L}(\cos bt) = \frac{s}{s^2+b^2}, s > 0.$
- $\mathcal{L}(\sin bt) = \frac{b}{s^2+b^2}, s > 0.$
- $\mathcal{L}(e^{at}\sin bt) = \frac{b}{(s-a)^2+b^2}.$
- $\mathcal{L}(e^{at}\cos bt) = \frac{s}{(s-a)^2+b^2}.$

Definition. f is piecewise continuous on $[a,b]$ if f is continuous except for finitely many points, and the left and right limit at those points are finite.

The piecewise continuity ensures that integrals of the form $\int_0^M f(t)e^{-st}\,dt$ make sense. But we still have to be careful to make sure the integral makes sense at infinity.

Definition. We say that f is of exponential order if there exists K, M, a positive such that for $t \geq M$, we have:

$$|f(t)| \leq Ke^{at}.$$

Theorem.

- If f is an exponential order and piecewise continuous on $[0,\infty)$, then $\mathcal{L}(f)(s)$ is well defined for $s > 0$.
- If f, g are piecewise continuous of exponential order, and $\mathcal{L}(f) = \mathcal{L}(g)$ for all s, then $f = g$.

Example. Let $f(t)$ be defined as follows:

$$f(t) = \begin{cases} t^2; \qquad 0 \leq t \leq 1 \\ (t-1)^{-1}; \qquad 1 \leq t \leq 2 \\ 1; \qquad t > 2. \end{cases}$$

Is $f(t)$ piecewise continuous? No, since the limit from the right at $t = 1$ is infinity.

Example. Let $f(t)$ be defined as follows:

$$f(t) = \begin{cases} t; 0 \leq t \leq 2 \\ 1-t; 2 \leq t \leq 4 \\ 3(t^2); t > 4. \end{cases}$$

This is piecewise continuous.

Example. If $f(t) = t^2$ of exponential order? We just need to show $t^3 \leq Ke^{at}$ for $t \geq M$. But $t^2 \leq we^t, t \geq 0$, which we can eaily see by looking at the Taylor series.

Example. Is e^{t^3} of exponential order?

Answer is no. Just need to show that $\lim_{t\to\infty} \frac{e^{t^3}}{e^{at}} = +\infty$, which follows from $\lim_{t\to\infty} e^{t^3-at}$, which is clearly infinite.

To apply Laplace transforms to differential equations, we must study the quantity $\mathcal{L}(f')$.

Important properties of Laplace transforms:

- Laplace transform is linear: we have that

$$\mathcal{L}(c_1f_1 + c_2f_2) = c_1\mathcal{L}(f_1) + c_2\mathcal{L}(f_2).$$

- We have

$$\mathcal{L}(e^{at}f(t))(s) = \mathcal{L}(f(t))(s-a).$$

- Laplace transform of a derivative. We have

$$\mathcal{L}(f'(t))(s) = \int_0^\infty f'(t)e^{-st}\,dt = f(t)e^{-st}|_0^\infty + \int_0^\infty f(t)se^{-st}dt = -f(0) + s\mathcal{L}\left(f(t)\right)(s).$$

- Laplace transform of second derivative. We have

$$\mathcal{L}(f'')(s) = -f'(0) - sf(0) + s^2\mathcal{L}(f).$$

- More generally, we can prove that

$$\mathcal{L}\left(f^n\right)(s) = -f^{(n-1)}(0) - sf^{(n-2)}f(0) - \cdots - s^{n-1}f(0) + s^n\mathcal{L}(f).$$

Now, we will have three different ways to solve second order ODEs.

6.24 Lecture 23: 7-26-19

Recall the main table from before:

- $\mathcal{L}(1) = \frac{1}{s}$
- $\mathcal{L}(e^{at}) = \frac{1}{s-a}$.
- $\mathcal{L}(\cos bt) = \frac{s}{s^2+b^2}$.
- $\mathcal{L}(\sin bt) = \frac{b}{s^2+b^2}$
- $\mathcal{L}(e^{at}\cos bt) = \frac{s-a}{(s-a)^2+b^2}$.
- $\mathcal{L}(e^{at}\sin bt) = \frac{b}{(s-a)^2+b^2}$.

So if we have the problem $y' - y = 0, y(0) = 3$, we can apply the LT on both sides. This implies that

$$-y(0) + s\mathcal{L}(y) - \mathcal{L}(y) = 0,$$

that is

$$\mathcal{L}(y) = \frac{y(0)}{s-1},$$

so

$$y(t) = 3e^t.$$

Example. We can also apply LT's to second order ODEs. If $y'' = y, y(0) = 1, y'(0) = 1$. In this case, the solution is $y(t) = e^t$. The proof is the same idea. Just use the formulae, to obtain

$$\mathcal{L}(y)(s^2 - 1) = 1 + s,$$

so $\mathcal{L}(y) = \frac{1}{s-1}$, that is $y(t) = e^t$.

It's a little more complicated when $y'' = y, y(0) = 2, y'(0) = 0$. Solution ends up being $y(t) = e^t + e^{-t}$, and you can obtain this by using partial fraction decomposition on the result, and inverting each term in the sum.

In the general case, we can derive the solution to $ay'' + by' + cy = 0, y(0) = y_0, y'(0) = y_0'$ using Laplace transforms. We would get

$$\mathcal{L}\left(ay'' + by' + cy\right) = \mathcal{L}(0),$$

and simplifying we get

$$\mathcal{L}(y) = \frac{A + Bs}{as^2 + bs + c}.$$

- If $as^2 + bs + c$ has two distinct real roots, then $\mathcal{L}(y) = \frac{C}{s-\lambda_1} + \frac{D}{s-\lambda_2}$, then $y = Ce^{\lambda_1 t} + De^{\lambda_2 t}$.
- If $as^2 + bs + c$ has complex conjugate roots, say $\mu \pm i\nu$, then $\mathcal{L}(y) = \frac{As+B}{a[(s-\mu)^2+\nu^2]} = \frac{C(s-\mu)}{(s-\mu)^2+\nu^2} + \frac{D\nu}{(s-\mu)^2+\nu^2} = Ce^{\mu t}\cos \nu t + De^{\mu t}\sin \nu t$.
- If $as^2 + bs + c = 0$ has a repeated root, then $as^2 + bs + c = a(s-\lambda)^2$. So

  $$\mathcal{L}(y) = \frac{A + Bs}{a(s-\lambda)^2} = \frac{C}{s-\lambda} + \frac{D}{(s-\lambda)^2}.$$

  This implies $y(t) = Ce^{\lambda t} + Dte^{\lambda t}$.

Now, let's look at the case when $y'' = by' + cy = f(t), y(0) = y_0, y'(0) = y_0'$.

Previously, we saw that we can do this with variation of parameters, and the method of undetermined coefficients.

With Laplace transforms, we have another method that is nice because it can be generalized.

For example, suppose we have $y'' - 2y' + 2y = te^t + 4$.

$$\mathcal{L}(y'') - 2\mathcal{L}(y') + 2\mathcal{L}(y) = \mathcal{L}(te^t) + 4\mathcal{L}(1),$$

which implies

$$-1 - s + s^2\mathcal{L}(y) + 2 - 2s\mathcal{L}(y) + 2\mathcal{L}(y) = \frac{1}{(s-1)^2} + \frac{4}{s}.$$

This implies that

$$\mathcal{L}(y)(s^2 - 2s + 2) = s - 1 + \frac{1}{(s-1)^2} + \frac{4}{s}.$$

Then,

$$\mathcal{L}(y) = \frac{s-1}{(s^2-2s+2)} + \frac{1}{(s-1)^2(s^2-2s+2)} + \frac{4}{s(s^2-2s+2)},$$

and let each term be f_1, f_2, f_3.

Note that

$$f_1 = \frac{s-1}{(s-1)^2+1},$$

so the inverse Laplace transform is $\mathcal{L}^{-1}(f_1) = e^t \cos t$.

Also,

$$f_2 = \frac{1}{(s-1)^2(s^2-2s+2)} = \frac{1}{(s-1)^2} - \frac{1}{(s-1)^2+1},$$

so $\mathcal{L}^{-1}(f_2) = te^t - e^t \sin t$.

Now,

$$f_3 = \frac{4}{s(s^2-2s+2)} = \frac{A}{s} + \frac{Bs+C}{s^2-2s+2} = \frac{(A+B)s^2 + (C-2A)s + 2A}{s(s^2-2s+2)}$$

This implies that $A + B = 0, C - 2A = 0, 2A = 4$. So $A = 2, B = -2, C = 4$.

Then,

$$\mathcal{L}^{-1}(f_3) = 2 + \mathcal{L}^{-1}\left[\frac{-2(s-1)}{(s-1)^2+1} + \frac{2}{(s-1)^2+1}\right] = 2 - 2e^t \cos t + 2e^t \sin t$$

Now, collecting the solutions, we can write

$$y(t) = te^t + 2 - e^t \cos t + e^t \sin t.$$

6.25 Lecture 24: 7-29-19

Suppose we have the quantity $\frac{2s}{(s+2)(s-2)}$. We can do partial fraction decomposition to obtain

$$\begin{aligned}\frac{2s}{(s+2)(s-2)} &= \frac{A}{s+2} + \frac{B}{s-2} \\ &= \frac{As - 2A + Bs + 2B}{(s+2)(s-2)}.\end{aligned}$$

Matching terms, this gives us $A+B=2$, $-2A+2B=0$. This implies $A=B=1$. Alternatively, you can plug in convenient values, e.g. 0 / 1.

Suppose we have the quantity $\frac{8s^2-4s+12}{s(s^2+4)}$. To compute the inverse LT, we want to write

$$\frac{8s^2-4s+12}{s(s^2+4)} = \frac{A}{s} + \frac{B \cdot 2}{s^2+4}.$$

If we have

$$\frac{8s^2-4s+12}{s\left[(s-1)^2+4\right]} = \frac{C}{s} + \frac{A(s-1)}{(s-1)^2+4} + \frac{B}{(s-1)^2+4} = \frac{C(s-1)^2+4C+As(s-1)+Bs}{s\left((s-1)\right)^2+4}.$$

In this case, it's probably easier to match points; to determine an n-th order polynomial, we plug in $n+1$ points.

If you plug in $s=0$, you get $C=12$, $s=1$ implies $16=B$, $s=-1$ implies $0=4C+2A-B$.

Example. Suppose we have the problem $y''+4y=3e^{-2t}$, $y(0)=2, y'(0)=-1$.

Recall that we have three ways to solve this:

- Undertermined coefficients.
- Variation of parameters.
- Laplace transform.

In this case, we can use underdetermined coefficients, since the particular solution is "nice."

Solution to the homogeneous problem is

$$y_{hom}(t) = c_1\cos(2t) + c_2\sin(2t).$$

Guess a particular solution of the form Ae^{-2t}. Then

$$4Ae^{-2t} + 4Ae^{-2t} = 3e^{-2t},$$

so $A = \frac{3}{8}$.

Thus the general solution is

$$y(t) = c_1\cos 2t + c_2\sin 2t + \frac{3}{8}e^{-2t}.$$

To solve this problem with variation of parameters, we need to turn this into a system of ODEs. To obtain the matrix A, recall for a characteristic equation $a\lambda^2+b\lambda+c=0$, the matrix A is given by

$$A = \begin{pmatrix} 0 & 1 \\ -\frac{c}{a} & -\frac{b}{a} \end{pmatrix} \mathrm{x} + \begin{pmatrix} 0 \\ 3e^{-2t} \end{pmatrix},$$

or

$$A = \begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix}.$$

This makes sense, since the first equation reduces to $y' = y'$, and the second is $y'' = -4y + 3e^{-2t}$.

The solution using variation of parameters is given by:

$$x_{hom} + X(t) \int^t X^{-1}(s)\, q(s) ds.$$

Remember that when we linearize a second order ODE, we obtain a vector $\mathrm{x} = (y, y')$. We can find a fundamental solution just by taking, where the second row is the derivative.

$$X = \begin{pmatrix} \cos 2t & \sin 2t \\ -2\sin 2t & 2\cos 2t \end{pmatrix}.$$

Inverting, we need

$$X^{-1} = \frac{1}{2} \begin{pmatrix} 2\cos 2t & -\sin 2t \\ 2\sin 2t & \cos 2t \end{pmatrix}.$$

Interestingly, the determinant is a constant - this makes sense since you can express the solution as complex exponentials (and multiplying by the conjugate makes things cancel).

Now, to compute x_p, we obtain

$$\begin{aligned} x_p &= X(t) \int^t X^{-1}(s) q(s)\, ds \\ &= \begin{pmatrix} \cos 2t & \sin 2t \\ -2\sin 2t & 2\cos 2t \end{pmatrix} \int^t \frac{1}{2} \begin{pmatrix} 2\cos 2s & -\sin 2s \\ 2\sin 2s & \cos 2s \end{pmatrix} \begin{pmatrix} 0 \\ 3e^{-2s} \end{pmatrix} ds. \end{aligned}$$

Now we proceed to the third method. Start by taking LT of both sides to obtain:

$$-y'(0) - sy(0) + s^2 \mathcal{L}(y) + 4\mathcal{L}(y) = 3\mathcal{L}\left\{e^{-2t}\right\}.$$

Thus,

$$1 - 2s + (s^2 + 4)\mathcal{L}(y) = \frac{3}{s+2},$$

this implies

$$\mathcal{L}(y) = \frac{3}{(s+2)(s^2+4)} + \frac{2s-1}{(s^2+4)} = \frac{A}{s+2} + \frac{Bs+C}{(s^2+4)}.$$

To compute this, we just combine everything together, to obtain

$$\frac{3 + (2s-1)(s+2)}{(s+2)(s^2+4)} = \frac{A(s^2+4) + (Bs+C)(s+2)}{(s+2)(s^2+4)}.$$

Plugging in $s = -2$, we get $3 = 8A$, so $A = \frac{3}{8}$. Plugging in $s = 0$, we get $1 = 4A + 2C$. Finally, plug in $s = 1$ to get $6 = 5A + 3B + 3C$. And we can continue to solve this equation to obtain the coefficients.

Suppose we have the function

$$f(t) = \begin{cases} e^t; \geq 0 \leq t \leq 1 \\ 1; t > 1. \end{cases}$$

Question. Can we take the Laplace transform of this?

Answer: yes, since it's clearly piecewise continuous, and of exponential order (best bound is e^{0t}, which shows that the LT is defined after $s > 0$).

Computing, we obtain

$$\begin{aligned} \mathcal{L}(f) &= \int_0^1 e^{(1-s)t}\,dt + \int_1^\infty e^{-st}\,dt \\ &= \frac{1}{1-s}\,e^{(1-s)t}\Big|_0^1 - \frac{1}{s}\,e^{-st}\Big|_1^\infty \\ &= \frac{1}{1-s}\left(e^{(1-s)t} - 1\right) + \frac{1}{s}e^{-s}. \end{aligned}$$

Example. Compute the LT of the function $f(x) = \sin x$, using the power series.

Recall that

$$\mathcal{L}\left(t^n\right) = (-1)^n \frac{d^n}{ds^n}\left(\mathcal{L}(1)\right) = \frac{n!}{s^{n+1}}.$$

Now $f(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$, so

$$\begin{aligned} \mathcal{L}(\sin x) &= \mathcal{L}(x) - \frac{1}{3!}\mathcal{L}(x^3) + \frac{1}{5!}\mathcal{L}(x^5) - \cdots \\ &= \frac{1}{s^2} - \frac{1}{s^4} + \frac{1}{s^6} - \cdots \\ &= \frac{1/s^2}{(1 + 1/s^2)} \\ &= \frac{1}{s^2+1}, \end{aligned}$$

which agrees with our previous computation.

This is pretty useful when we solve ODEs with power series, and in this setting taking the LT in this way is way easier. For instance, see Bessel functions, which are important when studying spherically symmetric systems in physics, that are expressed in terms of power series.

Interestingly, the Laplace transform can be very powerful when you regard s as a complex parameter.

Note: we probably won't have lecture on the Friday of the final.

Example. Suppose we have a system of ODEs which is a second order system, e.g.

$$\begin{cases} x'' + y' + 2x = 0 \\ 2x - y' = \cos t \\ x'(0) = x''(0) = y'(0) = 0. \end{cases}$$

In this case, we can't really apply the methods we have seen so far. One thing we can do is obtain a 3×3 system of ODEs by reducing the second order equation to two first order equations. But this grows large really fast.

One easier way to solve this is to apply Laplace-transform to both of these. Let's define $f(s) = \mathcal{L}(x)$, $g(s) = \mathcal{L}(y)$. We obtain:

$$\begin{aligned} s^2 f(s) + sg(s) + 2f(s) &= 0 \\ 2f(s) - sg(s) &= \frac{s}{s^2+1}. \end{aligned}$$

Adding these equations, we gt

$$(s^2+4)f(s) = \frac{s}{s^2+1},$$

implying

$$\begin{aligned} f(s) = \frac{s}{(s^2+1)(s^2+4)} &= \frac{As+B}{s^2+1} + \frac{Cs+D}{s^2+4} \\ &= \frac{(As+B)(s^2+4)}{(s^2+1)(s^2+4)} + \frac{(Cs+D)(s^2+1)}{(s^2+1)(s^2+4)}. \end{aligned}$$

Matching coefficients, we can put in $s = \pm i, 2i, -2i$ to obtain:

$$\begin{aligned} i &= 3(Ai+B) \\ -i &= 3(-Ai+B) \\ 2i &= -3(2iC+D) \\ -2i &= -3(-2iC+D). \end{aligned}$$

Then we get $B = 0, A = \frac{1}{3}, D = 0, C = -\frac{1}{3}$. Actually, this is redundant, since we can just match real / complex parts.

Thus, $f(s) = \frac{1}{3}\frac{s}{s^2+1} - \frac{1}{3}\frac{s}{s^2+4}$, so $x = \frac{1}{3}\cos t - \frac{1}{3}\cos 2t$.

Now, you just plug in x and solve for y using integrating / Laplace transform.

Tomorrow, we will start by discussing how to solve ODEs using power series. We will discuss Fourier series / Fourier transform and solve some PDEs.

6.26 Lecture 25: 7-30-19

Today, we will talk about series convergence tests. Recall that the region of convergence of $\sum a_n(x-x_0)^n$, the radius is symmetric around x_0. We call the radius of this interval ρ.

Recall the harmonic series. We know that $\sum_{n=1}^{\infty} \frac{1}{n} > \int_1^{\infty} \frac{1}{x}\,dx = \ln(\infty) = \infty$, so the harmonic series divergences.

Theorem. (Ratio test.) Given $\sum_{n=1}^{\infty} a_n$, and $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| = l$, then:

- If $l < 1$, the series converges.
- If $l > 1$, the series divergences.
- If $l = 1$, nothing can be said.

To see that $l = 1$ is inconclusive, recall $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges and $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

Theorem. (Root test.) Given $\sum_{n=0}^{\infty} a_n$ and $\lim_{n\to\infty} a_n^{1/n} = l$, then:

- If $l < 1$, series converges.
- If $l > 1$, series diverges.
- If $l = 1$, nothing can be said.

Remark. If $\frac{a_{n+1}}{a_n} \to l$, then $\rho = \frac{1}{l}$.

Example. Consider $\sum_{n=1}^{\infty} x^n$. In this case, the ratio test gives 1, and the root test also gives 1. The radius of convergence is 1.

Example. Consider $\sum_{n=1}^{\infty} \frac{(-2)^n (x-1)^n}{n}$. In this case, ratio gives 2, so $\rho = \frac{1}{2}$.

Example. Consider $\sum n! x^n$. In this case, $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = n \to \infty$. This implies $l = \infty$, $\rho = \frac{1}{l} = 0$.

Example. Consider $\sum \frac{x^n}{n!}$, so the ratio is $\frac{1}{n} \to 0$, and the $\rho = \infty$. In this case, the sum converges to e^x.

Example. Consider $\sum_{n=1}^{\infty} \frac{x^{2n+1}}{(2n+1)!}(-1)^n$. Theorem "technically" doesn't work because the indices are different. By comparison, consider $\sum b_n x^n$ where $b_n = |a_n|$ for n even, and $b_n = \frac{1}{n!}$ for n odd. Since $\sum b_n x^n$ has radius of convergence ∞, so does $\sum a_n x^n$.

Example. Consider $\sum_{n=1}^{\infty} \frac{3^n}{n^2+1}(x-3)^n$. Then the ratio is

$$\frac{a_{n+1}}{a_n} = 3\frac{n^2+1}{(n+1)^2+1} \to 3,$$

and the radius of convergence is $\rho = \frac{1}{3}$.

Example. Consider $\sum_{n=1}^{\infty} \frac{n!}{n^n}(x-2)^n$. Computing the ratio test, we obtain

$$\frac{a_{n+1}}{a_n} = \frac{n+1}{n+1}\frac{n^n}{(n+1)^n} \to \frac{1}{e},$$

so the radius of convergence is e.

If $f(x) = \sum_{n=0}^{\infty} a_n (x-x_0)^n$, what is $f'(x)$?

Now,

$$f(x) = a_0(x-x_0)^0 + a_1(x-x_0)^1 + \dots$$

so,

$$f'(x) = a_1 + 2a_2(x-x_0) + 3a_3(x-x_0)^2 + \cdots = \sum_{n=0}^{\infty}(n+1)a_{n+1}(x-x_0)^n.$$

We can do the same with integration. If $f(x) = \sum_{n=0}^{\infty} a_n(x-x_0)^n$, then we can compute $\int^x f(t)dt$, that is:

$$\int^x f(t)dt = \sum_{n=0}^{\infty} a_n \int (x-x_0)^n \, dx = \sum_{n=0}^{\infty} \frac{a_n}{n+1}(x-x_0)^{n+1}.$$

6.27 Lecture 26: 7-31-19

Today, we will discuss differential equations via power series.

Consider the problem $y'(x) = y(x), y(0) = y_0$. Recall that the solution is $y(x) = y_0 e^x$, obtained via standard methods.

Assume that $y(x) = \sum_{n=0}^{\infty} a_n x^n$ exists. Then, we can equate both sides, to obtain

$$\sum_{n=0}^{\infty} a_{n+1}(n+1)x^n = \sum_{n=0}^{\infty} a_n x^n.$$

If you equate coefficients, we obtain that $a_n = a_{n+1}(n+1)$, $n = 0, 1, \ldots$. From the initial condition, we obtain that $a_0 = y_0$, and continuing in this fashion, $a_1 = y_0, a_2 = 2y_0, a_3 = 3!y_0$, etc.

Thus,

$$y(x) = y_0 + \frac{y_0}{2!}x + \frac{y_0}{3!}x^2 + \ldots.$$

Example. Suppose $y''(x) = y(x), y(0) = a, y'(0) = b$. Recall that the CP is $\lambda^2 - 1 = 0$, so $\lambda = \pm 1$, so the solutions are e^x, e^{-x}. Thus, $y(x) = c_1 e^x + c_2 e^{-x}$. Then $a = c_1 + c_2, b = c_1 - c_2$, so $c_1 = \frac{a+b}{2}$, $c_2 = \frac{a-b}{2}$.

If y has a power series representation, then

$$y'' = \sum_{n=0}^{\infty} a_{n+2}(n+2)(n+1)x^n,$$

$$\sum_{n=0}^{\infty} a_n x^n.$$

We expect $a_n = a_{n+2}(n+2)(n+1)$.

So,

$$\begin{aligned} 2a_2 &= a_0 \\ 3 \cdot 2a_3 &= a_1 \\ 4 \cdot 3a_4 &= a_2 \\ 5 \cdot 4a_5 &= a_3. \end{aligned}$$

This allows us to solve odd / even equations separately. From initial conditions, $a_0 = a, a_1 = b$. So

$$\begin{aligned} a_{2n} &= (2n)!a \\ a_{2n+1} &= (2n+1)!b. \end{aligned}$$

Thus,

$$y(x) = a(1 + \frac{x^2}{2} + \frac{x^4}{4!} + \dots +) + b(x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots +).$$

Example. Consider $y''(x) = -4y(x)$, $y(0) = 1, y'(0) = 0$, so $y(x) = \cos 2x$.

Then $y(x) = \sum_{n=0}^{\infty} a_n x^n$, so the LHS is

$$y'' = \sum_{n=0}^{\infty} a_n n(n-1)x^{n-2}.$$

Thus,

$$-4y = -4\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty}(-4a_n)x^n.$$

Example. Consider the Airy equation $y'' = xy$, where for $x > 0, y'' = xy$ solutions look like exp. If $x < 0$, $y'' = xy$ solutions oscillate.

If $y(x) = \sum_{n=0}^{\infty} a_n x^n$, we obtain

$$y''(x) = \sum_{n=0}^{\infty} a_{n+2}(n+2)(n+1)x^n.$$

Then, the RHS is

$$xy = \sum_{n=1}^{\infty} a_{n-1}x^n.$$

Since for $n = 0$, $a_2 = 0$, we conclude that $a_2 = a_5 = \dots = 0$.

If $a_0 = a$, then $a_3 = \frac{a}{3\cdot 2}, a_6 = \frac{a}{6\cdot 5\cdot 3\cdot 2}$, etc.

If $a_1 = b$, then $a_4 = \frac{b}{4\cdot 3}, a_7 = \frac{b}{7\cdot 6\cdot 4\cdot 3}$.

And this allows us to write the final solution in terms of these three pieces.

Question - to what extent can we generalize what we did here? In turns out that this trick works directly if we replace the coefficient x with some polynomial $P(x)$. But it also works if we replace x with a function $f(x)$, where $f(x)$ admits a power series representation.

Example: suppose that $y'' + p(x)y' + q(x)y = 0$, $y(0) = a, y'(0) = b$, where $p(x), q(x)$ are analytic (they have a power series expression).

Suppose that $p(x) = \sum_{n=0}^{\infty} p_n x^n$, and $q(x) = \sum_{n=0}^{\infty} q_n x^n$; and suppose $y(x) = a_n x^n$. To solve this equation, you just plug in $y(x)$, and obtain coefficients, and match them.

Now, if we differentiate, we get

$$(2a_2 + 3\cdot 2a_3 x + 4\cdot 3a_4 x^2 + \dots) + (p_0 + p_1 x + p_2 x^2 + \dots)(a_1 + 2a_2 x + 3a_3 x^2 + \dots)$$
$$+(q_0 + q_1 x + q_2 x^2 + \dots)(a_0 + a_1 x + a_2 x^2 + \dots +) = 0.$$

And if you multiply stuff out, we can get equations like

$$2a_2 + p_0a_1 + q_0a_0 = 0$$
$$3 \cdot 2a_3 + 2p_0a_2 + p_1a_1 + q_0a_1 + q_1a_0 = 0.$$

The explicit equation is kind of messy, but it's reasonably practical. If you stop this procedure, you can get a polynomial approximation of a system.

6.28 Lecture 27: 8-1-19

Suppose we have the ODE

$$\begin{aligned} y'' + p(x)y' + q(x)y &= 0 \\ y(x_0) &= y_0 \\ y'(x_0) &= y_0', \end{aligned}$$

and p, q are analytic at x_0 (they have a power series expression around x_0, with radius of convergence ρ_1, ρ_2.

Theorem. With the hypothesis above, there exists a unique solution to the initivla value problem, and it is given by a pwer series whose radius of convergence is at least $\min(\rho_1, \rho_2)$.

Example. If we have the problem

$$\begin{aligned} y'' + \sin xy' + \cos xy &= 0 \\ y(0) &= 1 \\ y'(0) &= 0. \end{aligned}$$

Find $a_0, \ldots, a_4$. Recall $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$, $\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4!}$.

Then you can just expand stuff and equate coefficients. You end up with $a_0 = 1, a_1 = 0, a_2 = -\frac{1}{2}, a_3 = 0, a_4 = \frac{1}{12}$.

Differential equations in nonstandard form. Often, we encounter equations that are written as

$$P(x)y'' + Q(x)y' + R(x)y = 0, y(x_0) = y_0, y'(x_0) = y_0'$$

where P, Q, R are analytic around x_0.

If $P(x_0) \neq 0$, then you can just divide by P and continue as before: you get

$$y'' + \frac{Q(x)}{P(x)}y' + \frac{R(x)}{P(x)}y = 0.$$

But you still have problems at the zeros of P.

If instead $P(x_0) = 0$, x_0 is a singular point.

Theorem. If x_0 is an ordinary point, then we can find a solution using power series with radius of convergence at least $\min\left(\rho(Q/P), \rho(R/P)\right)$.

Determining the ROC of Q/P, where Q, P are polynomial. We will start with an example. Consider $(x^2-2x-3)y''+xy'+4y=0, y(0)=y_0, y'(0)=y_0'$.

In this setting, $\frac{Q}{P}=\frac{x}{x^2-2x-3}, \frac{R}{P}=\frac{4}{x^2-2x-3}$. Now, at least, the radius of convergence of $\frac{Q}{P}, \frac{R}{P}$ must be such that the zeros of P are not contained within the radius of convergence.

Theorem. (This is the only obstruction). That is, the radius of convergence of $\frac{Q}{P}, \frac{R}{P}$ is the size of the largest interval surrounding x_0 that does not contain the zeros of P. In other words, the ROC is $|x_0 -$ closest zero of P after cancellation.

There is some subtlety here, since the quotient only converges on $(-1,1)$ (not $(-1,3)$. The example you should have in mind is $\ln(1+x)=1-x+\frac{x^2}{2}-\frac{x^3}{3}+\ldots$ - turns out that $\ln 3$ is not defined as this series.

Example. Consider the problem $(x^2-1)y''+(x-1)y'+xy=0, y(3)=y_0, y'(3)=y_0'$.

We can write

$$\frac{Q}{P}=\frac{1}{x+1}$$
$$\frac{R}{P}=\frac{x}{(x-1)(x+1)}.$$

For $\frac{Q}{P}$, only problem is at -1, for $\frac{R}{P}$, problems are at $-1,1$. So the ROC of the solution is at least 2 (since this is the minimum of the ROCs of $\frac{Q}{P}$ and $\frac{R}{P}$).

Example. Consider the problem $(1+x^2)y''+xy'+4x^2y=0, y(1)=y_0, y'(1)=y_0'$. Then:

$$\frac{Q}{P}=\frac{x}{1+x^2};\qquad \frac{R}{P}=\frac{4x^2}{1+x^2}.$$

Now, to apply the theorem, it turns out we have to consider zeros in the complex plane. Here, the zeros of look like this:

And the ROC is going to be $\sqrt{2}$, computing the minimum distance to the complex conjugates.

Example. Consider the problem $(x^2+2x+2)y''+y'+y=0, y(0)=y_0, y'(0)=y_0'$. What is a lower bound for the ROC of the solution?

In this case,

$$\frac{Q}{P}=\frac{1}{x^2+2x+2}=\frac{R}{P},$$

and the roots are $-1\pm i$. Thus, the ROC is going to be $\sqrt{2}$ (distance from 0 to $-1\pm i$).

Chebyshev polynomials. We'll now discuss the Chebyshev polynomials, which are useful in numerical analysis. They are defined as solutions to the following ODE:

$$(1-x^2)y''-xy'+\alpha^2y=0, y(0)=y_0, y'(0)=y_0',$$

where $\alpha\in\mathbb{R}$. Following the procedure from before, the ROC of the solution is at least 1.

Now, if we write $y(x) = \sum_{n=0}^{\infty} a_n x^n$, we obtain

$$(1-x^2)\left(\sum_{n=2}^{\infty} a_n n(n-1)x^{n-2}\right) - x\left(\sum_{n=1}^{\infty} a_n n x^{n-1}\right) + \alpha^2 \sum_{n=0}^{\infty} a_n x^n = 0.$$

If you expand this, you get

$$\sum_{n=0}^{\infty} a_{n+2}(n+2)(n+1)x^n - \sum_{n=2}^{\infty} a_{n+1} n(n-1)x^n - \sum_{n=1}^{\infty} a_n n x^n + \alpha^2 \sum_{n=0}^{\infty} a_n x^n = 0.$$

We can consider cases of the value of n to obtain the terms. We get:

- $n = 0$: $2a_2 + \alpha^2 a_0 = 0$
- $n = 1$: $6a_3 - a_1 + \alpha^2 a_1 = 0$
- $n = 2$: $a_{n+2}(n+2)(n+1) - a_n n(n-1) - a_n n + \alpha^2 a_n = 0$.

Now, this implies that

$$a_{n+2} = \frac{a_n(n^2 - \alpha^2)}{(n+2)(n+1)}.$$

Now, say $\alpha = 2, a_1 = 0, a_0 = 1$. This implies $a_1 = a_3 = a_5 = \cdots = 0$. But while $a_2 = -2$, $a_4 = a_6 = \cdots = 0$. So in this case, $y(x) = 1 - 2x^2$.

Tomorrow, we will talk about inner products and Fourier series.

6.29 Lecture 28: 8-2-19

Today, we will discuss Fourier series. For some motivation, note that when solving $y'' + p(x)y' + q(x)y = 0$ when p, q are discontinuous, we can't use the regular power series method. More importantly, we are going to use Fourier series to solve partial differential equations.

Suppose you are given a periodic function $f(x)$ with period $2L$. Then the Fourier series allows us to write

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right).$$

Example. Consider the function

$$f(x) = \begin{cases} 1; & x \in (0, \pi) \\ 0; & x \in (\pi, 2\pi). \end{cases}$$

Then $f(x) = \frac{1}{2} + \frac{2}{\pi} \sum_{n=0}^{\infty} \frac{\sin(2n+1)x}{2n+1}$, and we get a picture like this:

Inner product spaces. Suppose V is a vector space. We can define an inner product on V; it must satisfy these axioms:

- It is linear in the first entry: $\langle c_1 v_1 + c_2 v_2, w\rangle = c_1\langle v, w\rangle + c_2\langle v_2, w\rangle$.
- It is antilinear in the second entry: $\langle v, c_1 w_1 + c_2 w_2\rangle = \overline{c_1}\langle v, w_1\rangle + \overline{c_2}\langle v, w_2\rangle$.
- It satisfies $\langle v, w\rangle = \overline{\langle w, v\rangle}$.
- It satisfies $\langle v, v\rangle = ||v||^2 \geq 0$, $||v|| = 0$ iff $v = 0$.

Some examples of inner products include:

- Dot product on $\mathbb{R}^n$.
- $\langle z_1, z_2\rangle = z_1\overline{z_2}$ on $\mathbb{C}$.

Definition. If $(V, \langle\rangle)$ is an inner product space. We say that v, w are orthogonal if $\langle w, w\rangle = 0$.

Definition. If $(V, \langle\rangle)$ is an inner product space, we say that $v_1, \dots, v_n$ is an orthogonal basis if it is a basis as $\langle v_i, v_j\rangle = 0$ for $i \neq j$. We say that it is an orthonormal basis if $||v_i|| = 1$ for all i.

Example. An example of an orthogonal basis in $\mathbb{R}^2$ is just $v_1 = \begin{pmatrix}1\\1\end{pmatrix}, v_2 = \begin{pmatrix}1\\-1\end{pmatrix}$, but it is not orthonormal.

Now, say that $v_1, \dots, v_n$ is an orthogonal basis. Since it is a basis, if v is arbitrary, we can uniquely write $v = c_1 v_1 + \cdots + c_n v_n$. Suppose we want to know what the c_i's are.

If they are orthogonal, then, we can just write $c_1 = \frac{\langle v, v_1\rangle}{\langle v_1, v_1\rangle}$.

Definition. Let V be the space of periodic functions with period $2L$ such that

$$\int_{-L}^{L} |f(x)|^2 \, dx < +\infty$$

is finite. This is called the space of square integrable functions. Note that every periodic and continuous function belongs to this space.

Example. If $L = \pi$, and f is defined as below,

$$f(x) = \begin{cases} 1; & x \in (0, \pi) \\ 0; & x \in (\pi, 2\pi), \end{cases}$$

then

$$\int_{-\pi}^{\pi} 1 = \pi < \infty.$$

It is straightforward to verify that this is a vector space.

Now, we can define an inner product as follows:

$$\langle f, g\rangle = \int_{-L}^{L} f(x)\overline{g}(x) \, dx \in \mathbb{C},$$

and it is easy to verify that this satisfies the axioms of an inner product. The only slightly tricky one is positivity. Clearly $\langle f, f\rangle = \int_{-L}^{L} |f(x)|^2 \geq 0$. But we have to check that $\langle f, f\rangle = 0$ if and only if $\int_{-L}^{L} |f(x)|^2 =$

0. If the functions are continuous, this is true. If they are not continuous, you have to define “equality” in this space as functions which differ only on a set of measure 0.

Theorem. The set $\left\{1, \sin\left(\frac{n\pi x}{L}\right), \cos\left(\frac{n\pi x}{L}\right), n = 1, 2, \dots\right\}$ is an orthogonal basis for V, the space of integrable $2L$-period functions, where $\langle f, g\rangle = \int_{-L}^{L} f\overline{g}$.

Proof. This is a fairly difficult theorem, since we are working with an infinite dimensional vector space. To prove that this is a basis, we need to use the Stone-Weierstrass theorem.

But we will prove that these are orthogonal with $L = 2\pi$.

- (Case 1.) We know that

$$\int_{-\pi}^{\pi} 1 \cdot \sin nx = \left.\frac{-\cos nx}{n}\right|_{-\pi}^{\pi} = 0.$$

- (Case 2.) Similarly,

$$\int_{-\pi}^{\pi} 1 \cdot \cos nx = \left.\frac{-\sin nx}{n}\right|_{\pi}^{\pi} = 0.$$

- (Case 3.) This is slightly trickier case: we apply Euler’s formula:

$$\begin{aligned}\int_{-\pi}^{\pi} \sin(nx)\sin(mx) = 0 \\ &= \int_{-\pi}^{\pi} \left(\frac{e^{inx} - e^{-inx}}{2i}\right)\left(\frac{e^{imx} - e^{-imx}}{2i}\right) \\ &= \int_{-\pi}^{\pi} e^{i(n+m)x} = \int_{-\pi}^{\pi} \cos((n+m)x)\, dx + i\sin\left((n+m)x\right)\, dx = 0.\end{aligned}$$

□

We can compute the coefficients by applying the inner products with the basis vectors. For example, $a_0 = \frac{1}{L}\int_{-L}^{L} f(x)\, dx$. We will continue to discuss this on Monday.

6.30 Lecture 29: 8-5-19

If $f(x)$ is any function which is $2L$-period and square integrable, then, we can write

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right).$$

Now,

$$\frac{a_0}{2} = \frac{\langle f, 1\rangle}{\langle 1, 1\rangle}; \qquad a_0 = \frac{1}{L}\int_{-L}^{L} f(x)\, dx.$$

And continuing, we have

$$a_n = \frac{\langle f, \cos\left(\frac{n\pi x}{L}\right)\rangle}{\langle \cos\left(\frac{n\pi x}{L}\right), cos\frac{n\pi x}{L}\rangle} = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right).$$

$$b_n = \frac{\langle f, \sin\left(\frac{n\pi x}{L}\right)\rangle}{\langle \sin\left(\frac{n\pi x}{L}\right), \sin\left(\frac{n\pi x}{L}\right)\rangle} = \frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right).$$

Example. If $f(x) = \sin 2x - 3\cos 3x$, what are the a_n's and b_n's? Clearly $a_3 = -3$, $b_2 = 1$, and $a_n = b_n = 0$ otherwise.

Example. Let $L = \pi$, and

$$f(x) = \begin{cases} 1; & x \in (0, \pi) \\ 0; & x \in (-\pi, 0), \end{cases}$$

and extend by periodicity.

In this case, $a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f = \frac{1}{\pi}\int_0^{\pi} 1 = 1$.

Also,

$$a_n = \frac{1}{\pi}\int_0^{\pi} \cos nx\, dx = \frac{1}{n}\sin(n\pi)\Big|_0^{\pi} = 0.$$

Now,

$$b_n = \frac{1}{\pi}\int_0^{\pi} \sin nx\, dx = -\frac{1}{\pi n}\cos nx\Big|_0^{\pi} = \begin{cases} 0; & n \text{ even} \\ \frac{2}{\pi n}; & n \text{ odd.} \end{cases}$$

Thus, the Fourier series looks like:

$$f(x) = \frac{1}{2} + \frac{2}{\pi}\left(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \frac{\sin 7x}{7} + \dots\right).$$

But note that $f(0)$ in the Fourier series actually does not equal $f(0)$: the following theorem clarifies this.

Theorem. If f, f' are piecewise continuous, then the Fourier series of f converges to f at each point where f is continuous, and converge to $\frac{f(x^-)+f(x^+)}{2}$ if f is discontinuous at x.

Example. Suppose $f(x) = x$ on (π, π), extended by periodicity.

Then,

$$a_0 = 0,$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x\cos(nx)\, dx = 0.$$

To compute b_n, we need to integrate by parts. We obtain:

$$b_n = \frac{2}{\pi} \int_0^{\pi} x \sin nx = \frac{2}{n}(-1)^{n+1}.$$

To simplify these computations, we can use the fact that the product of an odd and even function is odd, and the product of two odd functions is even.

Now, this implies that

$$f(x) = 2\left(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \frac{\sin 4x}{4} + \dots\right).$$

Theorem. If f is odd and $2L$ periodic, then $a_n = 0$ for all n, and $b_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right) dx$. If f is even and $2L$ periodic, then $b_n = 0$ for all n, and $a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)\, dx$.

Pythagorean theorem proof. We recall a quick proof of the Pythagorean theorem. Suppose $\langle v_1, v_2\rangle = 0$. What is the length of $v = v_1 + v_2$? Computing, we obtain:

$$\begin{aligned} ||v_1 + v_2||_2^2 &= \langle v, v\rangle = \langle v_1 + v_2, v_1 + v_2\rangle \\ &= \langle v_1, v_1\rangle + \langle v_1, v_2\rangle + \langle v_2, v_1\rangle + \langle v_2, v_2\rangle \\ &= ||v_1||^2 + ||v_2||^2, \end{aligned}$$

where we have used the fact that v_1, v_2 are orthogonal.

Parseval's identity. We'll now discuss a generalized form of the Pythagorean theorem. Suppose $v_1, \dots, v_n, \dots$ is an orthogonal basis, and suppose $v = \sum_{i=1}^{\infty} c_i v_i$. Then,

$$\begin{aligned} ||v||^2 = \langle v, v\rangle &= \langle \sum_{i=1}^{\infty} c_i v_i, \sum_{j=1}^{\infty} c_j v_j\rangle \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_i \overline{c_j}\langle v_i, v_j\rangle = \sum_{i=1}^{\infty} |c_i|^2 ||v_i||^2. \end{aligned}$$

Applying the generalization to Fourier series. Suppose $f(x) = x$, and $f(x) = \sum_{n=1}^{\infty} b_n \sin(nx)$. Then

$$||\sin(nx)||^2 = \int_{-\pi}^{\pi} \sin^2(nx) = \pi.$$

Now,

$$||f||^2 = \int_{-\pi}^{\pi} |x|^2 = \frac{x^3}{3}\Bigg|_{\pi}^{-\pi} = \frac{2\pi^3}{3}.$$

Now, $|b_n|^2 = \frac{4}{n^2}$.

Applying Parseval's identity, we obtain

$$\frac{2}{3}\pi^3 = \sum_{n=1}^{\infty} \frac{4}{n^2}\pi,$$

so we get the nice identity for $\zeta(2)$:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Example. Suppose $L = \pi$, and $f(x) = x(\pi - x)$ on $[0, \pi]$, extend oddly on $[-\pi, \pi]$, and then extend by periodicity .

By the earlier theorem, we know that $a_0 = 0$ for all n, and

$$\begin{aligned} b_n &= \frac{2}{\pi} \int_0^{\pi} x(\pi - x) \sin\left(\frac{n\pi x}{L}\right) = \frac{2}{\pi}\left[-\frac{\cos nx}{n} x(\pi - x)\right]_0^{\pi} + \frac{2}{\pi}\int_0^{\pi} \frac{\cos nx}{n}(\pi - 2x) \\ &= 2\pi \left[\pi \int_0^{\pi} \frac{\cos nx}{n} - 2\int_0^{\pi} \frac{x\cos(nx)}{n}\right] \\ &= -\frac{4}{\pi n}\left[\int_0^{\pi} x\cos(nx)\right] = -\frac{4}{\pi n}\left[-\int_0^{\pi} \frac{\sin nx}{n}\right] = \frac{4}{\pi n^3}\left(-\cos(nx)\right)_0^{\pi} = \frac{4}{\pi n^3}\left(1 - (-1)^n\right). \end{aligned}$$

Interestingly, these Fourier coefficients decay very fast, so we can stop the series quickly, and the remainder is not too big. Intuitively, the more regular a function, the faster its decay. Formally, if $f \in C^k$ (f is k-times differentiable), then the Fourier coefficient decay as $\frac{1}{n^{k+1}}$.

Example. Suppose $f(x)$ is defined as follows:

$$f(x) = \begin{cases} 1; & x \in (0, L) \\ -1; & x \in (-L, 0]. \end{cases}$$

We know that $a_n = 0$ for all n since the function is odd. Now,

$$b_n = \frac{2}{L}\int_0^L \sin\left(\frac{n\pi x}{L}\right) = \frac{2}{n\pi}\left(-\cos\left(\frac{n\pi x}{L}\right)\right)_0^L = \frac{2}{n\pi}\left((-1)^{n+1} + 1\right).$$

Interestingly, this does not depend on L. And this is the first example where we computed a Fourier transform (which works for functions over the whole real line).

6.31 Lecture 30: 8-6-19

Suppose we have a metal-rod with non-uniform temperature[5]. Heat is transferred from regions of higher temperature to regions of lower temperature. Recall the following physical principles:

- Heat energy of a body with uniform properties can be calculated as follows:

  $$E = cmu,$$

  where m is the body mass, u is the temperature, c is the specific heat.

[5] Notes supplemented with https://ocw.mit.edu/courses/mathematics/18-303-linear-partial-differential-equations-fall-2006/lecture-notes/heateqni.pdf

- Fourier's law of heat transfer. The rate of heat transfer is proportional to negative temperature gradient, that is:

  $$\frac{\text{Rate of heat transfer}}{\text{area}} = -K_0 \frac{\partial u}{\partial x},$$

  where K_0 is the thermal conductivity.

- (Conservation of energy). Consumer a uniform rod of length l with non-uniform temperature lying on the x-axis from $x = 0$ to $x = l$. In particular, uniform means that the density ρ, specific heat c, thermal conductivity K_0, cross-sectional area A are all constant. Consider an arbitrary thin slice of the rod of with Δx between x and $x + \Delta x$. suppose that the slice is so thin that the temperature throughout the slice is $u(x, t)$. Thus,

  $$\text{Heat energy of segment} = c\rho A \Delta x u(x, t).$$

Now, by Fourier's law, we have

$$cpA\Delta x u(x, t + \Delta t) - cpA\Delta x u(x, t + \Delta t) = \Delta t A \left(-K_0 u_x\right)_x - \Delta t A(-K_0 u_x)_{x+\Delta x}.$$

Rearranging and taking the limit $\Delta t \to 0, \Delta x \to 0$, we get

$$u_t = \kappa u_{xx},$$

where $\kappa = \frac{K_0}{c\rho}$ is called the thermal diffusivity.

Recall the following types of boundary conditions:

- A Neumann boundary condition specifies the value of a derivative at a boundary.
- A Dirichlet boundary condition specifies the value of a function at a boundary.

Now, in full form, the heat equation is

$$\begin{aligned} u_t &= \alpha^2 u_{xx}; \qquad x \in [0, L], t \geq 0. \\ u(x, 0) &= \phi(x); \qquad x \in [0, L] \text{ (IC)} \\ u(0, t) &= u(L, t) = 0; \qquad t > 0 \text{ (BC)} . \end{aligned}$$

Intuitively, we expect $u \to 0$ as $t \to \infty$.

Recall what we did with ODEs like $x' = Ax$. The first thing was to find v such that $Av = \lambda v$, then look fo ra solution $x(t) = T(t)v$, $T'(t)v = \lambda T(t)v$, so $T(t) = e^{\lambda t}$, and then use linearity.

We will do something similar with the heat equation, and reduce it to studying an ODE.

First, look for solutions $u(x, t) = T(t)X(t)$. Plugging this into the equation, we obtain

$$T'(t)X'(x) = \alpha^2 T(t)X''(x),$$

which implies

$$\frac{T'(t)}{\alpha^2 T(t)} = \frac{X''(x)}{X(x)} = \lambda \in \mathbb{R},$$

where we have used the fact that if a function of time equations a function of x, they must be constant.

We will consider cases on the value of λ.

- Case 1. ($\lambda > 0$) Now, if $X''(x) = k^2X(x)$, we get $X(x) = c_1e^{kx} + c_2e^{-kx}$, and applying the boundary conditions $X(0) = X(L) = 0$, we obtain $c_1 = c_2 = 0$, which is not an interesting solution.
- Case 2. ($\lambda = 0$) In this case, we get $X''(x) = 0$, that is $X(x) = c_1x + c_2$, and plugging in the boundary conditions we get $c_1 = c_2 = 0$, which is also uninteresting.
- Case 3. ($\lambda < 0$). In this case, we get $X''(x) = -k^2X(x)$, so $X(x) = c_1 \sin(kt) + c_2 \cos(kx)$. Plugging in the boundary conditions, we get

$$\begin{aligned} 0 &= X(0) = c_2 \\ 0 &= X(L) = c_1 \sin kL. \end{aligned}$$

  This implies that $kL = n\pi$, that is $k = \frac{n\pi}{L}$ for some $z \in \mathbb{Z}$.

Therefore, for each $n = 1, 2, \ldots$, we found a solution

$$u_n(x,t) = \sin\left(\frac{n\pi x}{L}\right) T_n(t).$$

Now, we just need to determine $T_n(t)$. In particular, we have

$$T_n'(t) = \frac{-\alpha^2 n^2 \pi^2}{L^2} T_n(t),$$

so

$$T_n(t) = \exp\left(\frac{-\alpha^2 n^2 \pi^2}{L^2} t\right),$$

that is

$$u_n(x,t) = \sin\left(\frac{n\pi x}{L}\right) \exp\left(\frac{-\alpha^2 n^2 \pi^2 t}{L^2}\right).$$

This gives a fundamental solution that we can use to compute the rest.

Intuitively, we note that the solution decays faster the more it oscillates at the beginning. That is:

- If $n = 1$, we get $\sin(\pi x)e^{-\pi^2 t}$.
- If $n = 2$, we get $\sin(2\pi x)e^{-4\pi^2 t}$.

By linearity, we know that

$$u(x,t) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi x}{L}\right) \exp\left(-\frac{n^2\pi^2\alpha^2}{L^2} t\right).$$

At $t = 0$, this almost looks like a Fourier series, but it doesn't quite work, since the function is defined only on $[0, L]$, and there are no cosines.

We will do the following procedure to fix this issue.

- Extend the function $\phi(x)$ on $[-L, 0]$ such that the extension is odd .
- Extend periodically on all $\mathbb{R}$. Call this extension $\tilde{\phi}(x)$.

We can write $\tilde{\phi}(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right)$. At $t = 0$, we have $\phi(x) = u(x, 0)$, so we get write

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) = \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{L},$$

which implies that

$$c_n = b_n = \frac{2}{L}\int_0^L \phi(x) \sin\left(\frac{n\pi x}{L}\right).$$

Example. Suppose we have the problem $u_t = u_{xx}$ on $[0, 1]$, where $u(x, 0) = \sin(3\pi x) - 2\sin(5\pi x)$, $u(0, t) = u(1, t) = 0$.

Then, the solution is given by

$$u(x, t) = \sum_{n=1}^{\infty} c_n \sin(n\pi x) \exp\left(-n^2\pi^2 t\right),$$

that is

$$u(x, t) = \sin(3\pi x)e^{-9\pi^2 t} - 2\sin(5\pi x)e^{-25\pi^2 t}.$$

Example. Suppose we have this problem:

$$\begin{aligned} u_t &= u_{xx}; \qquad \text{on } [0, 1) \times (0, +\infty) \\ u(x, 0) &= 1; \qquad x \in [0, 1] \\ u(0, t) &= u(1, t) = 0; \qquad t > 0. \end{aligned}$$

Then, we have

$$c_n = \frac{2}{1}\int_0^1 \sin(n\pi x) = -2\cos n\pi x|_0^1 = \frac{2}{n\pi}\left(1 - (-1)^n\right).$$

Example. Suppose we have

$$\begin{aligned} u_t &= \alpha^2 u_{xx}; \qquad \text{on } [0, L) \times (0, \infty) \\ u(x, 0) &= \phi(x); \qquad x \in [0, L] \\ u(0, t) &= a; u(L, t) = b \end{aligned}$$

In this case, the limiting temperature is a line.

We can define $v(x,t) = u(x,t) + Ax + B$, which satisfies the same equation, but the boundary conditions change. We have

$$\begin{aligned} v(x,0) &= \phi(x) + Ax + B \\ v(0,t) &= u(0,t) + B = A + B \\ v(L,t) &= u(L,t) + B = b + AL + B \end{aligned}$$

,

and if we take $B = -a$, $A = \frac{-b+a}{L}$, we get $v(0,t), v(L,t) = 0$, which reduces the heat equation to the case we are familiar with.

6.32 Lecture 31: 8-7-19

We continue our discussion of the heat equation. Suppose we have the setting

$$\begin{aligned} u_t &= \alpha^2 u_{xx} \\ u(x,0) &= \phi(x) \\ u_x(0,t) &= u_x(L,t) = 0 \text{ Neumann boundary condition.} \end{aligned}$$

As before, if we perform separation of variables, we assume that the solution is of the form $u(x,T) = T(t)X(x)$, and obtain the reduced equation

$$\frac{T'(t)}{\alpha^2 T(t)} = \frac{X''(x)}{X(x)} = \lambda \in \mathbb{R}.$$

As before, we consider cases on the sign of λ.

- (Case 1, $\lambda > 0, \lambda = k^2$). In this case, $X(x) = c_1 e^{kx} + c_2 e^{-kx}$. Plugging in the boundary conditions, we obtain

  $$\begin{aligned} 0 &= u_x(0,t) = (c_1 k - c_2 k)T(t) \\ 0 &= u_x(L,t) = X'(L)T(t) = (c_1 k e^{kL} - c_2 k e^{-kL}). \end{aligned}$$

  Solving this, this forces $c_1 = c_2 = 0$, which is not interesting.

- (Case 2, $\lambda = 0$). In this case, we get $X(x) = c_1 x + c_2$. Now, if we apply the boundary conditions, we get

  $$0 = u_x(0,t) = X'(0)T(t) = c_1 T(t),$$

  so $c_1 = 0$. Also,

  $$0 = u_x(L,t) = X'(L)T(t) = 0,$$

  which means either $X'(L)$ or $T(t) = 0$. Considering each case, we find that constants are solutions.

- (Case 3, $\lambda < 0, \lambda = -k^2$). In this case, we get $X(x) = c_1 \sin(kx) + c_2 \cos(kx)$.

  We know that $u_x(x,t) = (-c_1 k \sin kx + c_2 k \cos kx)T(t)$.

  Plugging in the boundary conditions, we obtain

  $$0 = u_x(0,t) = c_2 k T(t),$$

  so $c_2 = 0$. Similarly,

  $$0 = u_x(L,t) = (-c_1 k \sin kL)T(t).$$

  Thus, $kL = n\pi$ for some $n > 1$, that is $k = \frac{n\pi}{L}$, that is $\lambda_n = -\frac{n^2\pi^2}{L^2}$.

Now, corresponding to $X_n(x) = \cos\left(\frac{n\pi x}{L}\right)$, we obtain

$$T_n'(t) = -\frac{n^2\pi^2}{L^2}\alpha^2 T(t),$$

so that

$$T_n(t) = \exp\left(-\frac{n^2\pi^2\alpha^2}{L^2}t\right).$$

And putting these together, we get

$$u_n(x,t) = \cos\left(\frac{n\pi x}{L}\right)\exp\left(-\frac{n^2\pi^2\alpha^2}{L^2}t\right).$$

Now, in general,

$$u(x,t) = \frac{a_0}{2}\cdot 1 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right)\exp\left(-\frac{n^2\pi^2\alpha^2}{L^2}t\right),$$

and $u(x,t)$ solves the heat equation with Nuemann boundary conditions. To obtain the coefficients a_n, we can use the Fourier series.

If $\tilde{\phi(x)}$ is obtained by evenly extending $\phi(x)$ on $[-L,0]$, and repeating with periodicity on $\mathbb{R}$, then

$$\tilde{\phi(x)} = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos\left(\frac{n\pi x}{L}\right),$$

and

$$c_n = \frac{1}{L}\int_{-L}^{L} \tilde{\phi(x)} \cos\left(\frac{n\pi x}{L}\right) = \frac{2}{L}\int_0^L \phi(x)\cos\left(\frac{n\pi x}{L}\right).$$

Evaluating at time $t = 0$, we conclude that $a_n = c_n$.

Importantly, we know that this gives us all the solutions because the functions $\cos\left(\frac{n\pi x}{L}\right)$ form a basis for square integrable functions (and this assumes that $\phi(x)$ is sufficiently "nice" and square integrable).

There are some minor differences between the Dirichlet and the Neumann case (apart from the fact that the latter contains cosines instead of sines).

In the Neumann case, we note that

$$\lim_{t\to\infty} u(x,t) = \frac{a_0}{2} = \frac{1}{L}\int_0^L \phi(x)dx,$$

which is the average temperature (this intuitively makes sense, and is in some way related to the conservation of energy).

Example. Suppose we have the problem

$$\begin{aligned} u_t &= u_{xx} \\ u_x(0,t) &= u_x(1,t) = 0 \\ u(x,0) &= 3 - 2\cos(2\pi x). \end{aligned}$$

Then, applying the formula directly, we get $u(x,t) = 3 - 2\cos(2\pi x)\, e^{-4\pi^2 t}$.

Example. Suppose we have the problem

$$\begin{aligned} u_t &= \alpha^2 u_{xx} \\ u(x,0) &= x \\ u_x(0,t) &= u_x(1,t) = 0 \end{aligned}$$

We know that

$$u(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right)\exp\left(-\frac{n^2\pi^2\alpha^2}{L^2}t\right).$$

Computing, we obtain

$$a_0 = 2\int_0^1 x = 1.$$

$$a_n = 2\int_0^1 x\cos(n\pi x)dx = \frac{2}{n^2\pi^2}\left(\cos n\pi - 1\right).$$

Example. Suppose we have the equation

$$\begin{aligned} u_t &= \alpha^2 u_{xx} + f(x,t) \\ u(0,t) &= u(L,t) = 0 \\ u(x,0) &= \phi(x). \end{aligned}$$

Suppose $\alpha = 1, L = 1$, and $f(x) = \sin(\pi x)t$, and $\phi(x) = \sin(2\pi x)$.

We will look for a solution $u(x,t) = \sum_{n=1}^{\infty} T_n(t)\sin(n\pi x)$. Then

$$\sum_{n=1}^{\infty} T_n'(t)\sin(n\pi x) = \sum_{n=1}^{\infty} -T_n(t)n^2\pi^2\sin(n\pi x) + \sin(\pi x)t.$$

Since $\sin(n\pi x)$ forms a basis, we can equate coefficients to get an infinite set of ODEs, that is:

$$\begin{aligned} T_1'(t) &= -\pi^2 T_1(t) + t \\ T_n'(t) &= -n^2\pi^2 T_n(t); \qquad n > 1. \end{aligned}$$

From the case when $n > 1$, we obtain that

$$T_n(t) = \exp\left(-n^2\pi^2 t\right).$$

For the case where $n = 1$, we have a first order linear ODE that we can solve with integrating factors. That is,

$$T_1(t) = c_1 e^{-\pi^2 t} + \frac{t}{\pi^2} - \frac{1}{\pi^4}.$$

Now, we can write the full solution as something like this:

$$u(x,t) = \left(e^{-\pi^2 t} + \frac{1}{\pi^2}\left(t - \frac{1}{\pi^2}\right)\right)\sin(\pi x) + \sum_{n=1}^{\infty} c_n \sin(n\pi x)\exp\left(-n^2\pi^2 t\right),$$

(assuming we didn't make computational errors in lecture).

Example. Suppose we have the problem $u_t = u_{xx} + \cos(\pi x) - 1$, $u(x,0) = 1$, $u_x(0,t) = u_x(1,t) = 0$.

Were it not for $\cos(\pi x) - 1$, the solution would be

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} c_n \cos(n\pi x)\exp\left(-n^2\pi^2 t\right).$$

As before, we can guess that the solution is the form $u(x,t) = T_0(t) + \sum_{n=1}^{\infty}\cos(n\pi x)T_n(t)$, so that we end up with

$$T_0'(t) + \sum_{n=1}^{\infty}\cos(n\pi x)T_n'(t) = \sum_{n=1}^{\infty} -n^2\pi^2\cos(n\pi x)T_n(t) + \cos\pi x - 1.$$

And plugging in $t = 0$, we get

$$u(x,0) = 1 = T_0(0) + \sum_{n=1}^{\infty} T_n(0)\cos(n\pi x).$$

Matching terms, we get

$$\begin{aligned} T_0'(t) &= -1 \\ T_0(0) &= 1. \end{aligned}$$

For the $\cos\pi x$ term, we get

$$\begin{aligned} T_1'(t) &= -\pi^2 T_1(t) + 1 \\ T_1(0) &= 0. \end{aligned}$$

All the other equations reduce to

$$\begin{aligned} T_n'(t) &= -n^2\pi^2 T_n(t) \\ T_n(0) &= 0. \end{aligned}$$

Solving, we get

$$\begin{aligned} T_n(t) &= 0, n > 2 \\ T_0(t) &= -t + 1 \\ T_1(t) &= \frac{1}{\pi^2}\left(1 - e^{-\pi^2 t}\right). \end{aligned}$$

6.33 Lecture 32: 8-8-19

Today, we continue our discussion of PDEs, specifically focusing on fluids. Here are a few important equations:

- Inviscid Burger equation in 1D
- Viscous Burger equation in 1D
- Incompressible Euler equation in 3D
- Incompressible Navier Strokes in 3D

Let's consider the equation $u_t + uu_x = 0, u(x,0) = \phi(x)$. We note that this is very different from e.g. the heat equation, since it is not linear.

We will apply the method of characteristics. The idea is to find curves on which the velocity is constant; and reduce the PDE into an infinite family of ODEs.

In this case, the "characteristics" are lines. We can observe that if $\phi(x)$ is decreasing in any region, then we get an intersection of characteristics, which means trouble for the solution.

We can get shocks, e.g. ...

```
https://people.maths.ox.ac.uk/trefethen/pdectb/burgers2.pdf
https://www.iist.ac.in/sites/default/files/people/IN08026/Burgers_equation_inv
http://www.bcamath.org/projects/NUMERIWAVES/Burgers_Equation_M_Landajuela.pdf
```

We can consider a generalization of this equation, which looks like $u_t + uu_x = \nu u_{xx}$, where ν is the viscosity. There is a transformation called the Cole-Hopf transformation, which reduces this into a linear transformation. The idea is that if we write

$$u = -2\nu \frac{1}{v}\frac{\partial v}{\partial x},$$

which reduces to $v_t - \nu v_{xx} = 0$ (which is essentiall the heat equation). This along with the fact that $v \neq 0$ implies that $u(x,t)$ is defined for all time.

This equation is a bit limited because it doesn't consider pressure. It turns out to consider pressure we have to generalize this to additional dimensions (since a theory of pressure is kind of degenerate in the 1-D case).

To generalize this equation, we can consider a function $u(t, x)$, where x, u are now vectors, where $\phi(x_0) = u(x, 0)$. The derivation kind of works the same. We look at

$$\frac{D}{dt}u(x_0 + \phi(x_0)t, t) = 0.$$

To differentiate this, we have to use the multivariable chain rule to obtain

$$u_t + u \cdot \nabla u = 0.$$

This is the inviscid Burger equation in 3D. We can also use the Laplacian to study the viscous Burger equation in 3D. That is,

$$u_t + u \cdot \nabla u = \nu \Delta u.$$

If we add pressure, we get something like

$$u_t = u \cdot \nabla u = \nu \Delta u - \nabla p,$$

where $\nabla \cdot u = 0$ (note the incompressibility condition).

This is the Navier-Stokes equation; if you drop the viscosity term this is the Euler equation.

Now, you need to pair this with an initial condition:

$$\begin{aligned} u(x, 0) &= \phi(x) \\ p(x, 0) &= p(x). \end{aligned}$$

People have been able to prove that if you start with an initial condition is very small, the solution exists for all time. Intuitively, this means that friction is going to dominate (you don't expect lava to have shock waves...ha).

It's possible to define $w = \nabla \times u$, which allows us to simplify this equation. In 2D, the equation reduces to

$$w_t + u \nabla w = \nu \Delta w,$$

and this is "nice" in the sense that is reduces to the heat equation.

In 3D, the equation reduces to

$$w_t + u \cdot \nabla w = \nu \Delta w_1 + w_1 \left(\frac{\partial v_2}{\partial z} + \frac{\partial v_3}{\partial y} \right).$$

Even the numerical solutions are pretty subtle.

Tadashi thinks there is blowup because the model isn't correct.

Most people think that mathematically there is existence / uniqueness.

Terenco Tao has published a lot of papers on blowup in certain cases.

See his recent Nature Physics article: https://www.nature.com/articles/s42254-019-0068-9

6.34 Lecture 33: 8-12-19

Until finals, we will do review and go over some of the practice final problems.

We have $(x^2 - 4x - 5)y'' + x^2 y + y = 0, y(0) = 0, y'(0) = 1$.

- Find the coefficient of the solution via power series. (Just do the standard stuff with differentiation, and solve for coefficients).
- Find a lower bound on the radius of convergence. Recall the theorem that says the ROC is at least the minimum of the ROCs of $\frac{Q}{P}, \frac{R}{P}$, where P, Q, R are the coefficients of the ODE.

Consider the equation $y' = \alpha y(1 - y), \alpha \neq 0$.

- Find and classify the equilibria. Equilibria are $y = 0, 1$, and then you can classify them based on the sign of α.
- Solve the IVP.

6.35 Lecture 34: 8-13-19

Continuing review session.

We have the equation $u_t = u_{xx} + \sin(2\pi x), u(0, t) = u(1, t) = 0$.

(a) Actual problem is: find the eigenfunctions of 2nd derivative with bc $X(0) = 0, X(1) = 0$. That is, we are looking for $X(x)$ such that $X''(x) = \lambda X(x)$.

You end up with $T_n' = -n^2\pi^2 T_n$. Solving, you end up with $u(x, t) = e^{-\pi^2 t}\sin(\pi x) + \frac{1}{4\pi^2}\left(1 - e^{-4n^2 t}\right)\sin(2\pi x)$.

Idea is to equate coefficients since $\sin(n\pi x)$ forms a basis.

(b)

The initial problem $y' = y^{2/5}, y(0) = 0$ has a unique solution - is false. Since there is $y = 0$, you end up with two solutions.

7

MATH104: Matrix Theory

adi amsmath

span diag Ker Im Im Tr rank Proj sgn

MATH104 Notes Adithya C. Ganesh
Instructor: Katerina Velcheva

Contents

7.1 Lecture 4: 6-27-19

Last time, we defined a field, and a vector space. Today, we will define what a vector subspace is. Intuitively, it is a "smaller" vector space inside a bigger vector space.

We will also discuss properties of a vector space.

Recall that if you have a vector space V and a field $\mathbb{F}$, you can define two operations, vector addition and scalar multiplication as follows:

- $+ : V \times V \to V$
- $\times : \mathbb{F} \times V \to V$.

Definition. A vector space $W \subset V$ is a subspace of V if and only if for all $w_1, w_2 \in W$ and $\alpha, \beta \in \mathbb{F}$, we have

$$\alpha w_1 + \beta w_2 \in W.$$

Example. Note that $\mathbb{R}^2$ is a subspace of $\mathbb{R}^3$.

Example. Note that $\mathbb{R}^{n\times n}$ is a vector space over the field $\mathbb{R}$.

Let W be the space of all $n \times n$ symmetric matrices. Note that W is a subspace of $\mathbb{R}^{n\times n}$.

Let A, B be symmetric. Clearly, $\alpha A + \beta B$ is symmetric by linearity.

Example. Let $V = \mathbb{R}^{n\times n}, F = \mathbb{R}, W = \{A | A_{ij} = -A_{ji}\}$, the set of skew-symmetric matrices.

Let A, B be two matrices in W, and consider $\alpha, \beta \in \mathbb{R}$.

We would like to show that

$$(\alpha A + \beta B)_{ij} = -(\alpha A + \beta B)_{ji}.$$

And this follows from the same logic as before.

Example. Let's consider the vector space $V = \mathbb{R}^2$. Let $W = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$.

So show that this isn't a subspace, let

$$w_1 = \begin{pmatrix}1\\0\end{pmatrix} + c_1 \begin{pmatrix}1\\1\end{pmatrix}$$
$$w_2 = \begin{pmatrix}1\\0\end{pmatrix} + c_2 \begin{pmatrix}1\\1\end{pmatrix}.$$

And furthermore,

$$\alpha w_1 + \beta w_2 = (\alpha + \beta) \begin{pmatrix}1\\0\end{pmatrix} + (\alpha c_1 + \beta c_2) \begin{pmatrix}1\\1\end{pmatrix},$$

and we see that $aw_1 + bw_2 \notin W$. Furthermore, there isn't a zero vector in this subspace (if we look at it geometrically).

Definition. Let $v_1, \cdots, v_k$ be vectors. We say that the vectors are linearly dependent if there exists $\alpha_1, \ldots, a_k$ not all 0 such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = 0.$$

We say that vectors are linearly independent if they are not linearly independent.

Example. What does it mean for a matrix to be linearly independent?

Let V be a matrix. We want to solve:

$$V = \begin{pmatrix}v_1 & v_2 & \ldots & v_k\end{pmatrix} \begin{pmatrix}\alpha_1\\ \alpha_2\\ \cdots\\ \alpha_k\end{pmatrix} = 0$$

If the columns of V are linearly independent, then we can conclude that a is the zero vector.

If the columns are linearly dependent, you can find a nontrivial solution a to this problem.

Definition. We define the rank of a matrix V to be the maximum number of linearly independent columns.

Definition. We define the span of a set of vectors $v_1, \ldots, v_k$ to be the set

$$\{V | V = \alpha_1 v_1 + \cdots + \alpha_k v_k, \alpha_1, \ldots, a_k \in \mathbb{F}\}$$

Example. Let $V = \mathbb{R}^2$. What is the span of the vector $\begin{pmatrix}1\\0\end{pmatrix}$? You get $V = \begin{pmatrix}\alpha_1\\0\end{pmatrix}$. Geometrically, this is just the x-axis.

Example. What is the span of the vector $\begin{pmatrix}1\\1\end{pmatrix}$? You get $V = \begin{pmatrix}\alpha_1\\ \alpha_1\end{pmatrix}$, which is the linear $y = x$.

Example. What is the span of $\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}$? This is the whole plane.

Example. What is the span of $\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}5\\0\end{pmatrix}$? Just the x-axis.

Definition. Given a vector space V and a set of vectors $\{v_1, \ldots, v_n\}$, we say that the set is a basis if two conditions hold:

- $v_1, \ldots, v_n$ are linearly independent.
- $\{v_1, \ldots, v_n\} = V$.

Example. The vectors $\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}$ are a basis of the plane $\mathbb{R}^2$.

Definition. The dimension of a vector space V is the number of vectors in a basis.

If $v_1, \ldots, v_n$ form a basis, then for any vector $v \in V$, we can choose constants $c_1, \ldots, c_n$ such that $V = c_1v_1 + c_2v_2 + \cdots + c_nv_n$; note that these must be unique. Otherwise, we can get a contradiction by obtaining different constants $b_1, \cdots, b_n$ such that $V = b_1v_1 + \cdots + b_nv_n$.

7.2 Lecture 5: 7-1-19

Recall the definition of linear dependence. If $v_1, \cdots, v_k$ are linearly dependent, there exists $c_1, \cdots, c_k$ not all 0 such that

$$c_1v_1 + \cdots + c_kv_k = 0.$$

The vectors are linearly independent if the opposite is true.

Recall that the span of $v_1, \ldots, v_k$ is defined as follows:

$$\{v_1, \ldots, v_k\} = \{c_1v_1 + \cdots + c_kv_k | c_i \in \mathbb{R}\}.$$

Recall that $v_1, \ldots, v_n$ is a basis for V if the v_i are linearly independent, and the v_i span V.

Theorem. If $v_1, \ldots, v_m$ form a basis, then any $v \in V$ can be expressed uniquely as

$$v = c_1v_1 + \cdots + c_nv_n, c_i \in \mathbb{F}.$$

Definition. $c_1, c_2, \cdots, c_n$ are the coefficients of v with respect to the basis $v_1, \ldots, v_n$.

Example. Let $V = \mathbb{R}^3$, $\mathbb{F} = \mathbb{R}$. We can define the standard basis as follows:

$$\mathrm{e}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}; \qquad \mathrm{e}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}; \qquad \mathrm{e}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}.$$

Then

$$v = 1\mathrm{e}_1 + 2\mathrm{e}_2 + 3\mathrm{e}_3 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Now, consider a new basis

$$\mathrm{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}; \qquad \mathrm{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}; \qquad \mathrm{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The same vector v can be expressed as

$$v = \mathrm{v}_1 + \mathrm{v}_2 + 3\mathrm{v}_3.$$

Now we will talk a little bit about orthogonality, and sums and intersections of vector spaces.

Given two vectors $x, y \in \mathbb{R}^n$, we can define the inner product between x and y as follows:

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = y^T x.$$

If $x, y \in \mathbb{C}^n$, we define

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i \overline{y}_i = y^H x.$$

Definition. We say that x, y are orthogonal if $\langle x, y \rangle = 0$.

Definition. We say that $v_1, \ldots, v_n$ are mutually orthogonal if $\langle v_i, v_j \rangle = 0$ for any $i \neq j$.

Definition. We say that $v_1, \ldots, v_n$ are orthonormal if the vectors are mutually orthogonal and $\langle v_i, v_i \rangle = 1$.

Example. The standard basis in $\mathbb{R}^n$ is an orthonormal set of vectors.

Example. (Nonexample). The basis from before,

$$\mathrm{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}; \quad \mathrm{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}; \quad \mathrm{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

is not an orthonormal basis, since

$$\langle v_1, v_1 \rangle = 2 \neq 1;$$
$$\langle v_1, v_2 \rangle = 1 \neq 0.$$

Definition. We say that a matrix A is orthogonal if $A^T A = I$. This is the same as requiring that the columns of A are orthonormal vectors.

Lemma. Suppose that $v_1, \ldots, v_n$ are mutually orthogonal in $\mathbb{R}^n$. Then $v_1, \ldots, v_n$ are linearly independent.

Proof. We want to show that if

$$c_1v_1 + \cdots + c_nv_n = 0,$$

we must have $c_i = 0$ for all i.

Now, consider

$$\langle c_1v_1 + \cdots + c_nv_n, v_i \rangle = \langle 0, v_i \rangle = 0.$$

Now, applying the linearity of the inner product, this is equal to:

$$c_1\langle v_1, v_i \rangle + c_2\langle v_2, v_i \rangle + \cdots + c_n\langle v_n, v_i \rangle = 0,$$

which is equal to $c_i\langle v_i, v_i \rangle = 0$. This implies that $c_i = 0$; and this is true for all i.

We will now talk about sums and intersections of subspaces.

Recall that if $W \subset V$, where V is a vector space over $\mathbb{F}$, we have W is a subspace of V iff for all $w_1, w_2 \in W$, $\alpha, \beta \in \mathbb{F}$, we have

$$\alpha w_1 + \beta w_2 \in W.$$

Suppose that $S, R \subset V$ are subspaces. We can define:

1. $S + R = \{s + r | s \in S, r \in R\}$.
2. $S \cap R = \{v | v \in S, v \in R\}$.

Lemma. $S + R$ and $S \cap R$ are subspaces.

Proof. Take $w_1, w_2 \in S + R$. By definition, there exists r_1, r_2, s_1, s_2 such that $w_1 = r_1 + s_1$, and $w_2 = r_2 + s_2$.

Now,

$$w_1 + w_2 = (s_1 + s_2) + (r_1 + r_2).$$

But $(s_1 + s_2) \in S$, and $(r_1 + r_2) \in R$, so $w_1 + w_2 \in S + R$ as claimed.

Similarly,

$$\alpha w_1 = \alpha(s_1 + r_1) = \alpha s_1 + \alpha r_1 \in S + R.$$

Exercise. Show that $S \cap R$ is a subspace.

Remark. Note that $S \cup R$ is not a subspace, the reason is that this isn't closed under addition. For example, the x-axis is a subspace of $\mathbb{R}^2$, and the y-axis is a subspace of $\mathbb{R}^2$. But the union doesn't contain vectors in the interior of the plane.

Definition. We say that $T = S \oplus R$, i.e. T is the direct sum of S and R.

1. We have $T = S + R$.
2. $S \cap R = \{0\}$.

Example. In $\mathbb{R}^3$, we can take

$$S = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

$$R = \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

In this case, $\mathbb{R}^3 = S \oplus R$.

Example. Let $S_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, R_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. Here, $S_2 \cap R_2 = \{0\}$, but $S_2 + R_2 \neq \mathbb{R}^3$.

Example. Let $S_3 = \left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right), R_3 = \left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right)$. Herein, $S_3 + R_3 = \mathbb{R}^3$, but $S_3 \cap R_3 \neq \{0\}$.

Theorem. If $T = S \oplus R$, then every $t \in T$ can be expressed uniquely as $t = s + r$ for $s \in S, r \in R$.

Proof. Suppose for sake of contradiction, that $t = s_1 + r_1 = s_2 + r_2$. Then $s_1 - s_2 = r_2 - r_1$, but this is a contradiction since we assumed the intersection was just the zero vector.

Theorem. If $T = S \oplus R$, then $\dim T = \dim S + \dim R$.

Proof. The idea is that you can obtain a basis of T by choosing a basis of S, a basis of R, and combining them.

Theorem. We have $\dim S + \dim R = \dim S + \dim R - \dim S \cap R$. (Note that the direct sum theorem is a special case of this theorem.)

Note: no lecture on Thursday July 4.

7.3 Lecture 7: 7-3-19

Suppose V is a vector space over $\mathbb{R}$. Further, suppose V has dimension n, meaning that there exists a basis $v_1, \ldots, v_n$. Since this is a basis, for every $v \in V$, we can find constants $c_1, \ldots, c_n$ such that

$$V = c_1 v_1 + \cdots + c_n v_n.$$

Let $\varphi_B : V \longrightarrow \mathbb{R}^n$ be a map such that

$$\varphi_B(v) = \begin{pmatrix} c_1 \\ \ldots \\ c_n \end{pmatrix}.$$

Now, let $V = P^n = \{p | p(t) = a_0 + a_1 t + \cdots + a_n t^n\}$. Here, we can choose our basis to be the monomials, that is

$$\begin{aligned} v_0 &= 1 \\ v_1 &= t \\ v_2 &= t^2 \\ &\dots \\ v_n &= t^n. \end{aligned}$$

Here, we can define a map $\varphi_B : P^n \longrightarrow \mathbb{R}^{n+1}$, such that the map is defined as follows:

$$a_0 + a_1 t + \cdots + a_n t^n \longrightarrow \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{pmatrix}.$$

A general fact. Given two polynomials, if

$$a_0 + a_1 t + \cdots + a_n t^n = b_0 + b_1 t + \cdots + b_n t^n,$$

for all t, then $b_i = a_i$ for all i.

A linear transformation is completely determined by where we send the basis vectors.

Suppose we have a map $L : V \longrightarrow W$, where V, W are vector spaces over $\mathbb{R}$. Further, suppose $\dim V = n$ and $\dim W = m$.

- Let $v_1, \dots, v_n$ be a basis for V.
- Let $w_1, \dots, w_m$ be a basis for W.

Now, let $v = c_1 v_1 + \cdots + c_n v_n$, and write

$$\begin{aligned} L(v) &= L(c_1 v_1 + \cdots + c_n v_n) \\ &= c_1 L(v_1) + c_2 L(v_2) + \cdots + c_n L(v_n). \end{aligned}$$

Now, suppose

$$L(v_i) = a_1^i w_1 + a_2^i w_2 + \cdots + a_m^i w_m$$

Matrix of linear map.

Let $L : V \longrightarrow W$ be a linear map. We would like to find some A_L such that $L(v) = A_L v$. We can define the matrix as

$$A_L = \begin{pmatrix} \phi_{BC}(L(v_2)) & \phi_{BC}(L(v_2)) & \dots & \phi_{BC}(L(v_n)) \end{pmatrix}$$

Diagramatically, we can make a sort of square (commutative diagram):

$$\begin{aligned} L &: V \longrightarrow W \\ \phi_{BD} &: V \longrightarrow \mathbb{R}^n \\ A_L &: R_n \longrightarrow R_m \\ \phi_{BCD}^{-1} &: \mathbb{R}^m \longrightarrow W. \end{aligned}$$

Last time, we discussed the derivative map. That is, given $p(t) = a_0 + a_1 t + \cdots + a_n t^n \in P^n$, we can define a map $L : P^n \longrightarrow P^{n-1}$ known as the derivative map.

Note that we can compute the derivative matrix as follows:

$$A_L = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2 & \dots & 0 \\ \dots \\ 0 & 0 & 0 & \dots & n \end{pmatrix}$$

Change of basis matrix.

Let $v_1, \dots, v_n$ be a basis for V. Let

$$B_1 = \begin{pmatrix} v_1 & v_2 & \dots & v_n \end{pmatrix}.$$

And let

$$B_2 = \begin{pmatrix} w_1 & w_2 & \dots & w_n \end{pmatrix}.$$

We would like to define a linear transformation that takes a vector in B_1 to B_2.

We get another diagram:

- $I_d : V \longrightarrow V$
- $\phi_{B_2} : V \longrightarrow R^n$
- $\phi_{B_1} : V \longrightarrow R^n$.
- $M : \mathbb{R}^n \longrightarrow \mathbb{R}^n$

Here, the change of basis M will take the coordinates in one basis and produce the coordinates in another basis.

If we have a linear transformation $L : V \longrightarrow W$, we can compute the matrix $M_w A_L M_v^{-1}$, to obtain the change of basis matrix. Try to visualize this using the commutative diagram.

7.4 Lecture 8: 7-8-19

We start with continued discussion of the notion of change of basis matrices.

Suppose $S \in V$ is a subspace. Here, let

- V be a vector space over $\mathbb{R}$.
- We define $S^\perp := \{v \in V | \langle v, s \rangle = 0, \forall s \in S\}$.

For example, if $V = \mathbb{R}^2$, and $S = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then $S^\perp = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. To prove this, we need to show $\begin{pmatrix} 0 \\ 1 \end{pmatrix} \subseteq S^\perp$ and $S^\perp \subseteq \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Properties of orthogonal complements.

- $S^\perp$ is a subspace of V.
- $S \oplus S^\perp = V$.
- $(S^\perp)^\perp = S$.
- $R \subseteq S$ is equivalent to $S^\perp \subseteq R^\perp$.
- $(R + S)^\perp = R^\perp \cap S^\perp$.
- $(S \cap R)^\perp = R^\perp + S^\perp$.

7.5 Lecture 9: 7-9-19

Recall that if $S \subset V$, where V is a vector space over $\mathbb{F}$, then

$$S^\perp = \{v \in V | v \perp s, s \in S\}.$$

The most important fact is that we can write

$$V = S \oplus S^\perp.$$

(Think about the geometry in $\mathbb{R}^n$.)

Definition. Given a linear map $A : V \to W$, we can define two relevant spaces.

- The kernel (or null space) of A is defined as follows:

$$A = \{v \in V | Av = 0\}.$$

- The image of A is defined as follows:

$$A = \{Av | v \in V\}.$$

Example. Suppose that our linear transformation A is the projection onto e_1 (the x-axis), that is $\mathrm{Pr}_{e_1} : \mathbb{R}^2 \to \mathbb{R}^2$.

Then:

- $\mathrm{Pr}_{e_1} =$ the y axis.
- $\mathrm{Pr}_{e_1} =$ the x axis.

We can also frame the kernel and the image in terms of the null-space and range. If $A : \mathbb{R}^n \to \mathbb{R}^n$ defines the map $v \mapsto Av$, then

- $A = \{v \in \mathbb{R}^n | Av = 0\} = N(A)$.
- $A = \{Av \in \mathbb{R}^m | v \in \mathbb{R}^n\} = R(A)$.

For example, if our matrix is $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, then

- $A = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
- $A = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

Lemma. If $A = \begin{pmatrix} \mathrm{a}_1 & \mathrm{a}_2 & \dots & \mathrm{a}_n \end{pmatrix}$, then

$$A = R(A) = \{\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_n\}.$$

Proof. To prove the lemma, we will show containment in both directions. Suppose $v \in \{\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_n\}$. Then for $c_i \in \mathbb{R}$, we have

$$\begin{aligned} v &= c_1 a_1 + \dots + c_n a_n \\ &= \begin{pmatrix} \mathrm{a}_1 & \mathrm{a}_2 & \dots & \mathrm{a}_n \end{pmatrix} \begin{pmatrix} c_1 \\ \dots \\ c_n \end{pmatrix}, \end{aligned}$$

which implies that $v \in A$.

The opposite inclusion is equivalent; just reverse the steps. □

Now, given a matrix $A : V \to W$, we have four spaces associated with it. We have:

- $(A) \in W$
- $(A) \in V$
- $(A^T) \in V$
- $(A^T) \in W$.

Theorem. Given a matrix A, we have the following important results relating the spaces associated with it.

- $(A) = (A^T)^\perp$
- $(A^T) = (A)^\perp$.

Proof. We would like to show $(A) = ((A^T))^\perp$. For all $v \in A$, we want to show that v is orthogonal to all vectors in A^T.

Suppose that $x \in A$. This implies $Ax = 0$. Also, for any $y \in \mathbb{R}^m$, we have $y^T Ax = 0$, but also $(A^T y)^T x = 0$. Further, $(A^T y)x = 0$ for all $y \in \mathbb{R}^m$.

This implies X is orthogonal to all vectors of the form $A^T y$ for all $y \in \mathbb{R}^m$. This means that $x \in (A^T)^\perp$.

For the opposite direction, suppose $x \in (A^T)^\perp$ for all $y \in \mathbb{R}^m$. Then $(A^T y)^T x = 0$, which implies $y^T Ax = 0$, that is $Ax = 0$, or $x \in A$.

Hence, $(A^T)^\perp = A$ as claimed. □

Corollary. In other words, (A) is perpendicular to (A^T). This implies that $(A) \oplus (A^T)^\perp = V$, so (A) and (A^T) partition the space V.

7.6 Lecture 10: 7-10-19

Let A be a linear transformation taking $\mathbb{R}^n \to \mathbb{R}^m$. Then

$$\begin{aligned} \mathbb{R}^n &= A \oplus (A)^\perp = A \oplus A^T \\ \mathbb{R}^m &= A \oplus A^T. \end{aligned}$$

Let $S \subset V$ be a subspace. If $V = S \oplus S^\perp$, then $\dim V = \dim S + \dim S^\perp$.

Theorem. Let A be a map from $\mathbb{R}^n \to \mathbb{R}^m$. Further, suppose A_{res} denotes the map from $A^\perp \to \Im A$. Then A_{res} is bijective.

Proof. We start by showing that A_{res} is injective. Suppose $A_{res}v_1 = A_{res}v_2$. This implies $A_{res}(v_1 - v_2) = 0$. Since $v_1, v_2 \in A^\perp$, this implies $v_1 - v_2 \in A^\perp$ and $v_1 - v_2 \in A$. In particular, this means that $v_1 - v_2 = 0$, so $v_1 = v_2$.

Now, we show that A_{res} is surjective. For all $w \in A$, there exists some $v \in \mathbb{R}^n$ such that $Av = w$. Now, take $v_1 \in A$ and $v_2 \in A^\perp$ such that $v = v_1 + v_2$. Now, we know that $w = Av = A(v_1 + v_2) = Av_1 + Av_2 = Av_2 = w$, so $v_2 \in A^\perp$, and A_{res} is surjective.

This proves the desired result. □

Definition. We say that a map is a vector space isomorphism if it is bijective and a linear transformation.

Theorem. Suppose that $L : V \to W$ is a vector space isomorphism. Then:

- Given a basis $v_1, v_2, \dots, v_n$ for V, then $L(v_1), L(v_2), \dots, L(v_n)$ is a basis for W.
- $\dim V = \dim W$.

Now, applying this result to the map between $A \to A^T$, we obtain the result that row rank is equal to column rank.

Also, remarking that $^n = A \oplus A^T$, we can prove the rank nullity theorem (since dimensions add when you use direct sums).

Definition. Suppose $f : S \to R$ is a function. We say that f is invertible if there exists $y : R \to S$ such that

- $f(y(r)) = r; \quad r \in R$
- $g(f(s)) = s; \quad s \in S.$

Lemma. If f is bijective, then it is invertible.

Proof. For all $r \in R$, there exists a unique $s \in S$ such that $f(s) = r$. Define $g(r) = s$. □

Now, suppose $L : V \to W$, and we take L^{-1}. Why is L^{-1} a linear map? Well, we can check it directly:

If

$$L^{-1}(\alpha w_1 + \beta w_2) = \alpha L^{-1}(w_1) + \beta L^{-1}(w_2),$$

then we can show that if we apply L to both sides, then we get equality, applying the linearity of L. Now, since $L(LHS) = L(RHS)$, we must have $LHS = RHS$, because L is injective.

Next lecture, we will discuss pseudoinverses. Given a linear transformation, if a linear transformation is invertible, we can write down the inverse pretty easily. What would happen if the linear transformation is not invertible? If we have a map from $\mathbb{R}^n \to \mathbb{R}^m$, we can redefine a notion of inverse that allows us to go back.

If $w = w_1 + w_2$, where $w_1 \in A, w_2 \in A^T$, then $A^T w = A^\dagger w_1$, where $A^\dagger$ is known as the pseudoinverse. Importantly, the pseudoinverse of a matrix exists always.

Note that everything until tomorrow will be on the midterm (and everything starting on Monday will be on the final).

7.7 Lecture 11: 7-11-19

At this point, we're done with chapters 1, 2, and 3. We'll start discussing the notion of Moore-Penrose pseudoinverses.

Recall that $A \in \mathbb{R}^{n\times n}$ is invertible if there exists $B \in R^{n\times n}$ if

$$AB = BA = I.$$

Problem:

- There are matrices in $R^{n\times n}$ that are not invertible.
- The $n \times m$ matrices are not invertible.

Recall that if $A \in R^{n\times m}$, there are four fundamental spaces related to this matrix.

- (A)
- (A)
- (A^T)
- (A^T).

Recall that we showed that:

- The orthogonal complement of the A is (A^T); the orthogonal complement of (A) is (A^T).
- We have the results:

$$\begin{aligned} \mathbb{R}^n &= A \oplus (A^T) \\ \mathbb{R}^m &= A^T \oplus \Im A. \end{aligned}$$

- Also,

$$A^T \cong A.$$

Why is this useful? We have that $\mathbb{R}^n = A \oplus A^\perp$.

This means that for every $v \in \mathbb{R}^n$, we can find a unique $v_1 \in A, v_2 \in A^T$ such that $v = v_1 + v_2$.

Similarly, since $\mathbb{R}^m = A^T \oplus A$, we can find a unique $w_1 \in A^T, w_2 \in A$ such that $w = w_1 + w_2$.

Definition. Suppose we have a map $A : \mathbb{R}^n \to \mathbb{R}^m$; then the pseudoinverse is $A^+ : \mathbb{R}^m \to \mathbb{R}^n$. Further, suppose $w = w_1 + w_2$, where $w_1 \in A^T; w_2 \in A$. We define:

$$\begin{aligned} A^+(w_1) &= 0 \\ A^+(w_2) &= v_2, \end{aligned}$$

where v_2 is the unique vector $\in A^T$ that satisfies $Av_2 = w_2$.

We can draw a picture explaining this: Add img

Properties.

- If A is invertible, then $A^{-1} = A^+$.
- If B is the pseudoinverse of A, then

$$ABA = A.$$

Proof. Suppose we have some vector $v = v_1 + v_2$, with $v_1 \in A, v_2 \in A^T$. Then

$$\begin{aligned} Av &= A(v_1 + v_2) \\ &= Av_2 \\ &= w_2. \end{aligned}$$

Now, note that

$$\begin{aligned} BAv &= Bw_2 \\ &= v_2. \end{aligned}$$

Now,

$$ABAv = Av_2 = w_2.$$

□

(Similarly, you can prove that $BAB = B$.)

- $(AB)^T = AB$.
- $(BA)^T = BA$.

In the homework, we will have to verify some of these properties.

Example. If $\alpha \in \mathbb{R}$, then

$$a^+ = \begin{cases} \frac{1}{a}; & \alpha \neq 0 \\ 0; & ; \alpha = 0. \end{cases}$$

Example. If A is surjective, then $\Im A$ is $\{0\}$, and

$$AA^+ = I_{m\times m}.$$

so A^+ is the right inverse.

In this setting, $A(v) = A(v_1 + v_2) = w_2$

Example. If A is injective, then $A = \{0\}$, and

$$A^+A = I_{n\times n}.$$

Computationally, if you want to verify that a matrix is the pseudo inverse, it suffices to check the properties from before, i.e.

- $ABA = A$
- $BAB = B$
- $(AB)^T = AB$
- $(BA)^T = BA$.

Theorem. The pseudoinverse satisfies this limit:

$$\begin{aligned} A^+ &= \lim_{\delta\to 0} \left(A^TA + \delta^2 I\right)^{-1} A^T \\ &= \lim_{\delta\to 0} A^T(AA^T + \delta^2 I)^{-1}. \end{aligned}$$

Midterm will have 3 problems:

- Problem from homework.
- Problem from definition.
- Easy computational problem.

7.8 Lecture 12: 7-15-19

The last topic to appear on the midterm will be pseudoinverses. Recall HW4 is due Wednesday, and the exam is Thursday evening. Recall the structure of the midterm:

- Homework problem.
- Problem where you state a definition.
- Relatively similar computational problem.

Recall that if $A : \mathbb{R}^n \to \mathbb{R}^m$, we can decompose

$$\begin{aligned} \mathbb{R}^n &= A \oplus A^T \\ \mathbb{R}^m &= A^T \oplus A. \end{aligned}$$

Recall the defintion of the pseudoinverse $A+ : \mathbb{R}^m \to \mathbb{R}^n$ such that for all $w \in A$, we have

- $A^+(w) = v$ where v is the unique element in A^T such that $Av = w$.
- For all $w \in A^T$, we have $A^+(w) = 0$.

Also, B is the pseudoinverse of A iff

1. $ABA = A$
2. $BAB = B$
3. $(AB)^T = AB$
4. $(BA)^T = BA$.

The pseudoinverse limit definition is given by

$$\begin{aligned} A^+ &= \lim_{\delta \to 0} (A^T A + \delta^2 I)^{-1} A^T \\ &= \lim_{\delta \to 0} A(AA^T + \delta^2 I)^{-1} \end{aligned}$$

Property 1. Recall that if we have a product UAV, where U and V are orthogonal, then

$$U^H U = UU^H = I,$$

and then

$$(UAV)^+ = V^H A^+ U^H.$$

Property 2. Recall that for a special class of diagonalizable matrices, we can express $A = UOU^H$, and thus $A^+ = UO^+U^H$.

The proofs of many of these results are provided in section 4.3.

Property 3. Recall that $(A^{-1})^T = (A^T)^{-1}$. We claim that

$$(A^+)^T = (A^T)^+.$$

To prove this, we will use the limit definition of the pseudoinverse.

Proof. Note that

$$\begin{aligned}(A^T)^+ &= \lim_{\delta\to 0}(AA^T + \delta^2 I)^{-1}A\\ &= \lim_{\delta\to 0}[A^T(A^TA + \delta^2 I)^{-1}]^T\\ &= [\lim_{\delta\to 0} A^T(A^TA + \delta^2 I)^{-1}]^T\\ &= (A^+)^T.\end{aligned}$$

□

Recall the definition of eigenvectors, if $A \in \mathbb{C}^{n\times n}$, we know that v is a right eigenvector if there exists $\lambda \in \mathbb{C}$ such that

$$A\mathrm{v} = \lambda\mathrm{v}.$$

Similarly, w is a left eigenvector if there exists $\mu \in \mathbb{C}$ such that

$$w^H A = uw^H.$$

Recall that the eigenvalues of A are the roots of the characteristic polynomial $\det(A - \lambda I)$.

Also, note that for a 2×2 matrix, the characteristic polynomial is

$$\pi_A(\lambda) = \lambda^2 - \mathrm{Tr}(A)\lambda + \det A.$$

Suppose that $\pi_A(\lambda)$ is the characteristic polynomial of an $n\times n$ matrix, it has exactly n complex roots.

Definition. Note that we can write the characteristic polynomial as

$$\pi_A(\lambda) = (\lambda - \lambda_1)^{a_1} \dots (\lambda - \lambda_m)^{a_m}.$$

The algebraic multiplicity of a root λ_i is a_i, where a_i is defined as in the equation above.

A natural question might be what is the geometric multiplicity? Based on these multiplicites, we get different Jordan canonical forms.

Definition. We can define the geometric multiplicity as

Geometric multiplicity is defined as

$$\dim(A - \lambda_i I)$$

In general, the geometric multiplicity is not equal to the algebraic multiplicity.

To compute the dimension of the kernel, use the rank nullity theorem.

Note that the algebraic multiplicity is ≥ the geometric multiplicity, and the geometric multiplicity is equal to the number of linearly independent eigenvectors.

Theorem. If for all λ_i eigenvalues of A, we have

$$AM(\lambda_i) = GM(\lambda_i),$$

then A is diagonalizable, and there exists V such that $V^H AV = D$, and thus $A = VDV^H$.

In this case, we have n linearly independent eigenvectors, and we can find a basis of $\mathbb{R}^n$ of eigenvectors. In other words, we can find a basis of $\mathbb{R}^n$ where the matrix A is diagonal with respect to the basis of eigenvectors.

7.9 Lecture 13: 7-16-19

Eastern european exams have a theoretical, oral component, and a practical component (problem solving). Homework 4 is due tomorrow at midnight.

Yesterday, given $A \in \mathbb{C}^{n\times n}$, we defined right eigenvectors in terms of the following equality:

$$Ax = \lambda x; \lambda \in \mathbb{C};$$

and similarly left eigenvectors satisfy

$$y^H A = \mu y^H.$$

- Step 1. Find the eigenvalues by finding the roots of the characteristic polynomials.
- Step 2. Solve the linear system in each eigenvalue to obtain the eigenvectors.

Recall that if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then

$$A - \lambda I = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix},$$

and

$$\det(A - \lambda I) = (a - \lambda)(d - \lambda) - bc;$$

and we can solve this system to obtain the eigenvalues.

In general, if A is $n \times n$

$$\det(A - \lambda) = (-\lambda)^n + (A)(-\lambda)^{n-1} + \cdots + \det(A).$$

This is important because the product of the roots are equal to $\det A$, and the sum is equal to (A). If the matrix A is singular, there exists $v \neq 0$ such that $Av = 0 = 0v$.

Example. Consider the matrix $A = \begin{pmatrix} 5 & -4 \\ 2 & -1 \end{pmatrix}$. Then

$$\begin{aligned} \det(A - \lambda I) &= (5 - \lambda)(-1 - \lambda) + 8 \\ &= (\lambda^2 - 4\lambda + 3) \\ &= (\lambda - 3)(\lambda - 1) = 0. \end{aligned}$$

So the eigenvalues are $\lambda = 1, 3$.

Now, to solve for the eigenvectors, we have to find a nontrivial element in the kernel of this matrix.

Definition. If $A \in \mathbb{C}^{n \times n}$, then the spectrum of A is defined as the set of eigenvalues.

Lemma. Eigenvalues are either real or come in conjugate pairs.

To see this, just note that λ is a root of the characteristic polynomial.

Also, note that all the eigenvectors are orthogonal (we will go over tomorrow).

Definition. Given $A \in \mathbb{R}^{n \times n}$, we say that A is diagonalizable if there exists an orthogonal Q, diagonal O such that

$$A = QOQ^H.$$

If A is not diagonalizable, we say that A is defective.

Theorem. If A has n distinct eigenvectors, then A is diagonalizable.

To see this, take D to be the diagonal matrix of eigenvalues, and Q to be the matrix whose columns are the eigenvectors.

Theorem. A is diagonalizable iff for all $\lambda_i \in \text{Spec}(A)$, we have $GM(\lambda_i) = AM(\lambda_i)$.

7.10 Lecture 14: 7-17-19

Proposition. Suppose $v_1, \ldots, v_n$ are eigenvectors for $A \in \mathbb{R}^{n \times n}$ with different eigenvalues. Then, the eigenvectors are linearly independent.

Proof. We can prove this by induction. If $n = 2$, the result is clear, since if $\lambda_1 \neq \lambda_2$, we can write

$$
\begin{aligned}
Av_1 &= \lambda_1 v_1 \\
Av_2 &= \lambda_2 v_2.
\end{aligned}
$$

Assume for sake of contradiction $v_1 = cv_2$, where $c \neq 0$. Now, if $Av_1 = cAv_2 = cA_2v_2$, this implies $A_1v_1 = \lambda_1 cv_2$, which implies $\lambda_1 = \lambda_2$, which is a contradiction.

By induction, assume that $v_1, \ldots, v_j$ are linearly independent with distinct eigenvalues $\lambda_1, \ldots, \lambda_j$, for some $2 < j < n$. By contradiction, assume that for the next eigenvector $v_{j+1}, Av_{j+1} = \lambda_{j+1}v_{j+1}$, and $\lambda_{j+1} \neq \lambda_i, i \in \{1, \ldots, j\}$. Now, if

$$
v_{j+1} = \sum_{i=1}^{j} c_i v_i,
$$

we can write

$$
\begin{aligned}
\lambda_{j+1} \sum_{i=1}^{j} c_i v_i = Av_{j+1} &= A \sum_{i=1}^{j} c_i v_i \\
&= \sum_{i=1}^{j} c_i A v_i \\
&= \sum_{i=1}^{j} c_i \lambda_i v_j.
\end{aligned}
$$

Now, term by term, this is equivalent to saying that

$$
\sum_{i=1}^{j} c_i (\lambda_i - \lambda_{j+1}) v_j = 0.
$$

But since not all the c_i's are 0, this implies that the v_j's are linearly dependent, which is a contradiction. □

Theorem. Let A be a symmetric matrix, with $A^T = A$, with $A \in \mathbb{R}^{n \times n}$, and x, y are eigenvectors with different eigenvalues λ, μ. Then $\langle x, \rangle = 0$.

Proof. Note that

$$
\begin{aligned}
\langle Ax, y \rangle &= (Ax)^T y \\
&= x^T A^T y \\
&= \langle x, A^T y \rangle.
\end{aligned}
$$

Now,

$$\begin{aligned}\lambda\langle x, y\rangle &= \langle \lambda x, y\rangle \\ &= \langle Ax, y\rangle \\ &= \langle x, A^T y\rangle \\ &= \langle x, Ay\rangle \\ &= \langle x, \mu y\rangle \\ &= \mu\langle x, y\rangle.\end{aligned}$$

Since $\lambda \neq \mu$, it follows that $\langle x, y\rangle = 0$. □

Theorem. Let $A \in \mathbb{R}^{n\times n}$. If x, y are right and left eigenvectors with $\lambda \neq \mu$, it turns out that $\langle x, y\rangle = 0$.

Proof. Note that

$$\begin{aligned}\mu\langle x, y\rangle &= \langle \mu x, y\rangle \\ &= \langle Ax, y\rangle \\ &= \langle x, A^T y\rangle \\ &= \langle x, (y^T A)^T\rangle \\ &= \langle x, (\lambda y^T)^T\rangle \\ &= \langle x, \lambda a\rangle \\ &= \lambda\langle x, y\rangle.\end{aligned}$$

□

Theorem. If $A \in \mathbb{R}^{n\times n}$ matrix, then A and A^T have the same eigenvalues.

Proof. Clearly,

$$\begin{aligned}\pi_A(\lambda) &= \det(A - \lambda I) \\ &= \det((A - \lambda I)^T) \\ &= \det(A^T - \lambda I) \\ &= \pi_{A^T}(\lambda).\end{aligned}$$

□

Corollary. The left and right eigenvalues of A are the same.

Theorem. Let $A \in \mathbb{C}^{n\times n}$ with $\lambda_1, \dots, \lambda_n$ distinct and right eigenvectors $x_1, \dots, x_n$, and left eigenvectors $y_1, \dots, y_n$ such that $y_j^H x_i = \delta_{ij}$[1]

Let $D = (\lambda_1, \dots, \lambda_n)$, $X = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}$, $Y = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix}$. Then A is diagonalizable as $A = XD^{X-1} = XDY^H$.

[1] $\delta_{ij} = 1$ if $i = j$, 0 otherwise.

Proof. Then:

- $AX = XD$
- $y^H A = DY^H$
- $y^H X = I$
- $A = XDX^{-1} = XDY^H = \sum_{i=1}^n \lambda_i x_i y_i^H$.

□

Theorem. (Spectral theorem). If $A \in \mathbb{R}^{n\times n}$ and $A^T = A$, then A has n eigenvectors that are orthonormal. Then

- $AQ = QD$
- $A = QDQ^T \sum_{i=1}^n \lambda_i \mu_i \mu_i^H$.

The only thing we didn't show is that n eigenvectors exist. Once we get this result, we can show the fact that A is diagonalizable.

This turns out to be computationally useful in a lot of cases.

For next time, we will continue with similar matrices and matrix exponential.

7.11 Lecture 15: 7-18-19

Definition. We say that a matrix A is similar to matrix B if

$$A = SBS^{-1},$$

for some invertible matrix S.

Example. If A has n linearly independent eigenvectors, then

$$AX = XD,$$

where $X = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}$, and $D = \text{diag}(\lambda_1, \dots, \lambda_n)$. In this case, $A = XDX^{-1}$.

Properties of similar matrices.

- If A and B are similar, then $\det(A) = \det(B)$. This follows directly from the multiplicative property of the determinant.

  Proof. $\det(A) = \det(SBS^{-1})$
  $= \det(S)\det(B)\det(S^{-1})$
  $= \det B$. □

- Similar matrices have the same eigenvalues. To show, we can show that both matrices have the same characteristic polynomial.

Proof. We have

$$\begin{aligned}\det(A-\lambda I) &= \det(SBS^{-1}-\lambda SIS^{-1}) \\ &= \det(S(B-\lambda I)S^{-1}) \\ &= \det(S)\det(B-\lambda I)\det(S^{-1}) \\ &= \det(B-\lambda I).\end{aligned}$$

□

This proves the result. However, it is not true that similar matrices have the same eigenvectors. For example, if $A = XDX^{-1}$, the eigenvectors of D are the standard basis vectors; while the eigenvectors of A are the columns of X.

Theorem. (Cayley-Hamilton). $\pi_A(A) = 0$. For example, $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, and $\pi_(A) = \lambda^2 - 2\lambda + 1$. Now, C-H states that $A^2 - 2A + I = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

We will now discuss matrix exponentials. Recall that we define

$$e^a = 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \dots.$$

In the case of a matrix, we define

$$e^{At} = I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} +$$

While there is no general algorithm to compute this, we can try diagonalizing. For instance, if D is diagonal, we can write

$$\begin{aligned}e^{Dt} &= I + (d_1t, \dots, d_nt) + \dots + (\frac{d_1^n}{n!}, \dots, \frac{d_n^n}{n!}) + \dots \\ &= (e^{d_1t}, e^{d_2t}, \dots, e^{d_nt}).\end{aligned}$$

Claim. Now, if $A = XDX^{-1}$, then $e^{At} = Xe^{Dt}X^{-1}$.

Proof. We have

$$\begin{aligned}e^{At} = I + \sum_{k=1}^{\infty}\frac{A^kt^k}{k!} &= I + \sum_{k=1}^{\infty} XD^kX^{-1}\frac{t^k}{k!} \\ &= X\left[I + \sum_{k=1}^{\infty}\frac{D^kt^k}{k!}\right]X^{-1}.\end{aligned}$$

This result follows. □

Later in the class, we will note that matrix exponentials are useful to solve systems of linear ODEs.

If A is diagonalizable, we can prove the following results:

Claim. $\det(e^A) = e^A$. (HW)

Claim. $(e^A)^{-1} = e^{-A}$. (HW; follows immediately).

Remember A is not always diagonalizable. Recall that we proved that any symmetric matrix is diagonalizable. But we can show that any matrix is similar to a Jordan matrix.

Now, we will discuss probability vectors are Markov matrices.

Motivation. Often, you can have a system with many states. And with some probability, you have users that transition from state i to state j.

For instance, suppose we have smartphone users that either have iPhones, Samsung, Huawei. We can create a state transition matrix as follows:

$$M = \begin{pmatrix} 0.75 & 0.30 & 0.20 \\ 0.15 & 0.60 & 0.10 \\ 0.10 & 0.10 & 0.70 \end{pmatrix}.$$

Definition. A vector $v \in \mathbb{R}^n$ is a probability is $\sum_i v_i = 1$, with $v_i \geq 0$.

Definition. A matrix $M \in \mathbb{R}^{m \times n}$ is a Markov matrix if the columns of M are probability vectors.

Often, we are interested in the long-term behavior of the system.

Suppose we have an initial condition $v_0 = \begin{pmatrix} 1000 \\ 1000 \\ 1000 \end{pmatrix}$. In k years, the distribution will be $v_k = M^k v_0$.

The question is whether this sequence converges to some vector. That is, what is $\lim_{n \to \infty} M^n v_0$? This happens to be related to the eigenvalues of the matrix.

Let's now state a few properties of Markov matrices.

Claim. If M, N are Markov matrices, then MN is a Markov matrix. (Homework).

Claim. If M is a Markov matrix and v is a probability vector, then Mv is a probability vector. (Homework).

Theorem. For all λ_i eigenvalues of a Markov matrix M, we have $|\lambda_i| \geq 1$.

Claim. Any Markov matrix has eigenvalue 1. This is called the principal eigenvalue. (Homework). (Note: the corresponding eigenvector will be the limit.)

Now, suppose that a Markov matrix M is diagonalizable, so that $M = XDX^{-1}$. Now, $M^n = XD^nX^{-1}$. Then,

$$\begin{aligned} \lim_{n \to \infty} M^n &= \lim_{n \to \infty} XD^nX^{-1} \\ &= X \lim_{n \to \infty} D^n X^{-1} \\ &= X(1, 0, \ldots, 0)X^{-1}. & = v_1, \end{aligned}$$

where v_1 is the principal eigenvector.

7.12 Midterm review

7.12.1 Definitions

1. D2.1, Field.
2. D2.3, Vector space.
3. D2.6, subspace.
4. D2.10, linear dependence.
5. D2.12, span.
6. D2.14, basis.
7. D2.19, dimension.
8. D2.21, sum of subspaces.
9. D2.24, direct sum.
10. D3.1, linear maps.
11. D3.3 (image and kernel)
12. D3.9 (orthogonal complement)
13. D3.15 (injective and surjective linear maps)
14. 3.23 (bijective linear maps)
15. 3.16 (columnn and row ranks)
16. 4.1 (pseudoinverse) and its correctness (the bijectivity of the restriction, as proven in 3.17)

7.12.2 Theorems

1. E2.5 P1, 2. Show that $(\mathbb{R}^n, \mathbb{R})$ is a vector space. Show that $(\mathbb{R}^{m\times n}, \mathbb{R})$ is a vector space.
2. E2.8, P1, 2. Show that W (space of symmetric matrices) is a subspace of $\mathbb{R}^{m\times n}$.

   Let W be the space of orthogonal matrices. Then W is not a subspace of $\mathbb{R}^{m\times n}$ (just check scalar multiplication.
3. T2.22. Show that the sum of spaces is a subspace.
4. T2.26. Suppose $T = R \oplus S$. Then every $t \in T$ can be written uniquely in the form $t = r + s$, with $r \in R, s \in S$. Also, $\dim(T) = \dim R + \dim S$.
5. T3.4. Let $A : V \to W$ be a linear map. Then A, A are subspaces.
6. T 3.11, P2, 3. Let $R, S \subseteq \mathbb{R}^n$. Then:
   - $S \oplus S^\perp = \mathbb{R}^n$.

- $(S^{\perp})^{\perp} = S$.

7. T 3.12. Let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Then:
   - $(A)^{\perp} = \Im(A^T)$
   - $(A)^{\perp} = (A^T)$.
8. T3.17. Let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. Then $\dim \Im A = \dim A^{\perp}$. (Restricting a linear mapping to a bijection on fundamental subspaces; row and column ranks are equal).
9. T3.19, P1. Let $A, B \in \mathbb{R}^{n\times n}$. Then
$$0 \leq (A+B) \leq A + B.$$
10. T3.20, P1, 2. Let $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}$. Then
$$(AB) \subseteq A$$
$$(AB) \supseteq (A)$$
11. T3.21, P1, 3. Let $A \in \mathbb{R}^{m\times n}$. Then
$$(A) = (AA^T)$$
$$(A) = (A^T A).$$
12. T4.11. Let $A \in \mathbb{R}^{m\times n}$, and suppose $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{m\times n}$ are orthogonal. Then:
$$(UAV)^{+} = V^T A^{+} U^T.$$

7.12.3 Theoretical HW problems

1. (HW2.1) Determine whether the following are vector spaces over $\mathbb{R}$:
   - $\mathbb{R}$, w.r.t. $+, \cdot$.
   - $\mathbb{C}$, w.r.t. addition and multiplication.
   - , w.r.t. addition and multiplication.
2. (HW2.2)
   - Show that P_2 is a vector space over R.
   - Let $1, t, 2t^2 - 1$ be the Chebyshev polynomials. Show they are a basis.
   - Find components of $P_2(t)$ w.r.t. T_0, T_1, T_2.
3. (HW 2.3). Show that M_n', M_n'' are subspaces of $\mathbb{R}^{n\times n}$. Find dimension by finding a basis.

   Show that $\mathbb{R}^{n\times n}, i\mathbb{R}^{n\times n}$ are not subspaces of $\mathbb{C}^{n\times n}$.
4. (HW 2.4) Show that $L = \{p : p(-t) = p(t)\}, M = \{p : p(-t) = -p(t)\}$ are subspaces, and find their dimensions.

5. (HW 3.1) Show that $\mathbb{R}^{n\times n} = M'_n + \oplus M''_n$. Show that $P_2 = L \oplus M$.

6. (HW 3.4) Show that $M''_n = (M'_n)^\perp$, and that $M'_n = (M''_n)^\perp$.

7. (HW 3.6) Let R, S be subspaces of $\mathbb{R}^n$. Show that $R \subseteq S$ iff $S^\perp \subseteq R^\perp$.

8. (HW 4.1) Let $v_1, \ldots, v_n$ be orthonormal, and $A \in \mathbb{R}^{n\times n}$. Show that $Av_1, \ldots, Av_n$ are orthonormal iff the matrix A is orthogonal.

9. (HW 4.4) Let $A = \begin{pmatrix} 2 & 4 \\ 3 & 6 \end{pmatrix}$.
   - Calculate $B = \lim_{\delta\to 0}(A^TA + \delta I)^{-1}A^T$.
   - Find A, A, A^T, A^T.
   - Show that B is the pseudoinverse of A by erifying Definition 4.1.
   - Show that B is the pseudoinverse of A by verifying the condition of 4.2.

7.12.4 Computational HW problems

1. (HW3.2) For each mapping, state whether it is linear or not.

2. (HW3.3) Find the matrix of A with respect to the bases $U = (u_1, u_2, u_3)$ and $V = (v_1, v_2)$, where u_i, v_i are given.

3. (E3.2, P2) Define $L : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ so that $LX = MX$, for some fixed matrix M.

4. (E 3.2, P3 + HW3 P5). Let L be the differentiation operator w.r.t. the polnynomial p. Is L injective, surjective, bijective?

5. (HW 4.2) Find the four fundamental subspaces of the matrix $A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ (change numbers).

6. (HW 4.4) (See above).

7.12.5 Subtest A

Definitions.

1. Define a field.

2. Define a vector space.

3. Define a subspace.

4. Define the pseudoinverse (and it's correctness, by bijectivity of the restriction).

Theorems.

1. T2.26. Suppose $T = R \oplus S$. Then every $t \in T$ can be written uniquely in the form $t = r + s$, with $r \in R, s \in S$. Also, $\dim(T) = \dim R + \dim S$.

2. T3.4. Let $A : V \to W$ be a linear map. Then A, A are subspaces.

3. T 3.11, P2, 3. Let $R, S \subseteq \mathbb{R}^n$. Then:

- $S \oplus S^\perp = \mathbb{R}^n$.
- $(S^\perp)^\perp = S$.

4. T 3.12. Let $A : \mathbb{R}^n \to \mathbb{R}^n$. Then:
   - $(A)^\perp = \mathfrak{I}(A^T)$
   - $(A)^\perp = (A^T)$.
5. T3.17. Let $A : \mathbb{R}^n \to \mathbb{R}^m$. Then $\dim \mathfrak{I}A = \dim A^\perp$. (Restricting a linear mapping to a bijection on fundamental subspaces; row and column ranks are equal).
6. T3.19, P1. Let $A, B \in \mathbb{R}^{n\times n}$. Then

$$0 \le (A+B) \le A+B.$$

7. T4.11. Let $A \in \mathbb{R}^{m\times n}$, and suppose $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{m\times n}$ are orthogonal. Then:

$$(UAV)^+ = V^T A^+ U^T.$$

Theoretical HW.

1. (HW 2.3). Show that M_n', M_n'' are subspaces of $\mathbb{R}^{n\times n}$. Find dimension by finding a basis.

   Show that $\mathbb{R}^{n\times n}$, $i\mathbb{R}^{n\times n}$ are not subspaces of $\mathbb{C}^{n\times n}$.
2. (HW 3.1) Show that $\mathbb{R}^{n\times n} = M_n' + \oplus M_n''$. Show that $P_2 = L \oplus M$.
3. (HW 3.4) Show that $M_n'' = (M_n')^\perp$, and that $M_n' = (M_n'')^\perp$.
4. (HW 3.6) Let R, S be subspaces of $\mathbb{R}^n$. Show that $R \subseteq S$ iff $S^\perp \subseteq R^\perp$.
5. (HW 4.1) Let $v_1, \ldots, v_n$ be orthonormal, and $A \in \mathbb{R}^{n\times n}$. Show that $Av_1, \ldots, Av_n$ are orthonormal iff the matrix A is orthogonal.
6. (HW 4.4) Let $A = \begin{pmatrix} 2 & 4 \\ 3 & 6 \end{pmatrix}$.
   - Calculate $B = \lim_{\delta\to 0}(A^T A + \delta I)^{-1} A^T$.
   - Find A, A, A^T, A^T.
   - Show that B is the pseudoinverse of A by erifying Definition 4.1.
   - Show that B is the pseudoinverse of A by verifying the condition of 4.2.

Computational HW.

1. (HW3.3) Find the matrix of A with respect to the bases $U = (u_1, u_2, u_3)$ and $V = (v_1, v_2)$, where u_i, v_i are given.
2. (HW 4.2) Find the four fundamental subspaces of the matrix $A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ (change numbers).

7.12.6 Subtest A, Practice solve

Definitions.

1. Define a field.

   A field is a set with $(+, *)$, where:

   - $+$ is commutative, associative.
   - $-x$ exists.
   - 0 exists, with $x + 0 = x$.

   Also,

   - $*$ is commutative, associative.
   - For nonzero x, x^{-1} exists.
   - 1 exists, with $x \cdot 1 = x$.

   Finally, $*$ distributes over addition.

2. Define a vector space.

   A vector space is a set with $(+, *)$, addition, scalar multiplication, where addition is defined over the set, and scalar multiplication is defined over the field.

   - $(V, +)$ is an abelian group.
   - V is closed under addition, and scalar multiplication.
   - There is a 0 element.
   - Additive inverses exist.

3. Define a subspace. A subset of a vector space that is closed under addition and scalar multiplication.

4. Define the pseudoinverse (and it's correctness, by bijectivity of the restriction).

   Defined as

   $$\lim_{\delta \to 0} (A^T A + \delta I)^{-1} A^T.$$

   Alternatively, can verify properties from the book (verify).

Theorems.

1. T2.26. Suppose $T = R \oplus S$. Then every $t \in T$ can be written uniquely in the form $t = r + s$, with $r \in R, s \in S$. Also, $\dim(T) = \dim R + \dim S$.

   $T = R \oplus S$ means $R + S = T$ and $R \cap S = \{0\}$. Suppose some $t = r_1 + s_1 = r_2 + s_2$. Then, $r - 1 - r_2 = s_1 - s_2$, but this contradicts that fact that $R \cap S = \{0\}$, since this forces $r_1 = r_2; s_1 = s_2$.

2. T3.4. Let $A : V \to W$ be a linear map. Then A, A are subspaces. Easy.

3. T 3.11, P2, 3. Let $R, S \subseteq \mathbb{R}^n$. Then:

- $S \oplus S^\perp = \mathbb{R}^n$.
- $(S^\perp)^\perp = S$.

First property - show that intersection is just 0, which is easy. Then, show that they sum to $\mathbb{R}^n$. Can do this with

Note that $S^\perp = \{t \in \mathbb{R}^n | \langle s, t \rangle = 0, s \in S\}$. Not super sure how to conclude.

4. T 3.12. Let $A : \mathbb{R}^n \to \mathbb{R}^n$. Then:

   - $(A)^\perp = (A^T)$
   - $(A)^\perp = (A^T)$.

   If $w \cdot v = 0$, and $Av = 0$, then $A^T v = w$. Not super sure how to conclude.

5. T3.17. Let $A : \mathbb{R}^n \to \mathbb{R}^m$. Then $\dim A = \dim A^\perp$. (Restricting a linear mapping to a bijection on fundamental subspaces; row and column ranks are equal).

   Corollary of the previous part.

6. T3.19, P1. Let $A, B \in \mathbb{R}^{n \times n}$. Then

   $$0 \leq (A + B) \leq A + B.$$

   Not sure.

7. T4.11. Let $A \in \mathbb{R}^{m \times n}$, and suppose $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{m \times n}$ are orthogonal. Then:

   $$(UAV)^+ = V^T A^+ U^T.$$

   Not sure.

Theoretical HW.

1. (HW 2.3). Show that M_n', M_n'' are subspaces of $\mathbb{R}^{n \times n}$. Find dimension by finding a basis.

   Not sure, look at solutions.

   Show that $\mathbb{R}^{n \times n}, i\mathbb{R}^{n \times n}$ are not subspaces of $\mathbb{C}^{n \times n}$.

   This part is easy.

2. (HW 3.1) Show that $\mathbb{R}^{n \times n} = M_n' \oplus M_n''$. Show that $P_2 = L \oplus M$.

   To argue first part, just need to show that intersection is 0, and that the sum sums to all possible matrices.

3. (HW 3.4) Show that $M_n'' = (M_n')^\perp$, and that $M_n' = (M_n'')^\perp$.

   Requires analyzing the term structure.

4. (HW 3.6) Let R, S be subspaces of $\mathbb{R}^n$. Show that $R \subseteq S$ iff $S^\perp \subseteq R^\perp$.

   Not hard, but don't remember the full argument.

5. (HW 4.1) Let $v_1, \dots, v_n$ be orthonormal, and $A \in \mathbb{R}^{n \times n}$. Show that $Av_1, \dots, Av_n$ are orthonormal iff the matrix A is orthogonal.

6. (HW 4.4) Let $A = \begin{pmatrix} 2 & 4 \\ 3 & 6 \end{pmatrix}$.

   - Calculate $B = \lim_{\delta \to 0}(A^T A + \delta I)^{-1} A^T$.

     Easy calculation.

   - Find A, A, A^T, A^T.

     Easy.

   - Show that B is the pseudoinverse of A by verifying Definition 4.1.

     Easy, but annoying (need to recall definition).

   - Show that B is the pseudoinverse of A by verifying the condition of 4.2.

     Easy but annoying (need to recall the criteria).

Computational HW.

1. (HW3.3) Find the matrix of A with respect to the bases $U = (u_1, u_2, u_3)$ and $V = (v_1, v_2)$, where u_i, v_i are given.

   Easy, but annoying. Need to remember matrix product form.

2. (HW 4.2) Find the four fundamental subspaces of the matrix $A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ (change numbers).

   Straightforward. Need to calculate A, A, A^T, A^T.

Review: subtest A.

Definitions.

1. Define a vector space.

   A field F with two operations $+ : V \times V \to V$, $* : F \times V$ that satisfy:

   - $(V, +)$ is an abelian group.
   - Associativity (mult) in field times vector.
   - Distributivity (both dirs).
   - $1 \cdot v = v$.

2. Define the pseudoinverse. Defined as

$$\lim_{\delta \to 0}(A^T A + \delta^2 I)^{-1} A^T.$$

   $B : (A)^\perp \to A$ is bijective. Need to prove this.

Theorems.

1. T 3.11, P2, 3. Let $R, S \subseteq \mathbb{R}^n$. Then:

   - $S \oplus S^\perp = \mathbb{R}^n$.
   - $(S^\perp)^\perp = S$.

   Easiest way to do this is to think about the geometry in $\mathbb{R}^n$. Let v_i be orthonormal basis for S. Then we can write

   $$x_1 = \sum_{i=1}^{k} (x^T v_i) v_i$$
   $$x_2 = x - x_1.$$

   Note that $x_2^T v_j = x^T v_j - x_1^T v_j = x^T v_j - x^T v_j = 0$, so x_2, v_j are orthogonal for any j, and thus to any linear combination of these vectors. Thus, x_2 is orthogonal to S.

   Note that $\dim S = \dim (S^\perp)^\perp$. Now, $W \subset W^{\perp\perp}$. QED.

2. T 3.12. Let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Then:

   - $(A)^\perp = (A^T)$
   - $(A)^\perp = (A^T)$.

   Take $v \in A$. Then $Av = 0$. Implies that $y^T Av$ for all y. But this means $(A^T y)^T x = 0$ for all y, so we have orthogonality.

3. T3.17. Let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. Then $\dim A = \dim A^\perp$. (Restricting a linear mapping to a bijection on fundamental subspaces; row and column ranks are equal).

   Let $T : A^\perp \longrightarrow A$. Then T is bijective (not too hard to see.)

4. T3.19, P1. Let $A, B \in \mathbb{R}^{n \times n}$. Then

   $$0 \le (A + B) \le A + B.$$

   Just concat the bases, and argue that every row of $(A + B)$ works as a linear combination of the respective bases.

5. T4.11. Let $A \in \mathbb{R}^{m \times n}$, and suppose $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{m \times n}$ are orthogonal. Then:

   $$(UAV)^+ = V^T A^+ U^T.$$

   Just verify that it satisfies the four Penrose conditions, namely $ABA = A, BAB = B, (AB)^T = AB, (BA)^T = BA$.

Theoretical HW.

1. (HW 2.3). Show that M_n', M_n'' are subspaces of $\mathbb{R}^{n \times n}$. Find dimension by finding a basis.

   $n(n + 1)/2$, just argue by decomposition.

2. (HW 3.4) Just analyze the term structure. Not hard.

3. (HW 3.6) Let $s' \in S^\perp$. We must have $\langle s, s' \rangle = 0$ for all $s \in S$. Then $\langle s, s' \rangle = 0$ for all $s \in R$ in particular. Hence $s' \in S^\perp$ implies $s' \in R^\perp$. QED.

4. Suppose $v_1, \dots, v_n$ is orthonormal. Show that $Av_1, \dots, Av_n$ are orthonormal if the matrix A is orthogonal.

   Consider $\langle Av_i, Av_j \rangle = v_i^T A^T A v_j$. Argue each way.

Computational HW.

1. Just do $B = V^{-1}AU$, where U is first basis, and V is second basis.

7.12.7 Subtest B

Thms.

1. T3.17. Let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. Then $\dim A = \dim A^\perp$. (Restricting a linear mapping to a bijection on fundamental subspaces; row and column ranks are equal).

HWT.

1. Dimension of M_n'.
2. $v_1, \dots, v_n$ orthonormal, and $A \in \mathbb{R}^{n \times n}$. $Av_1, \dots, Av_n$ ortho if A is ortho.

HWC.

- Find matrix of A w.r.t. bases U and V. $F = V^{-1}AU$.

7.13 Lecture 16: 7-22-19

Recall that if $v, w \in \mathbb{R}^n$, we can define the standard inner product as

$$\langle v, w \rangle = v^T w.$$

Definition. In general, an inner product is a map $V \times V \longrightarrow \mathbb{R}$. It must satisfy:

- Positive definite. So that $\langle v, v \rangle \geq 0$, and $\langle v, v \rangle = 0$ iff $v = 0$.
- Symmetry. We must have $\langle v, w \rangle = \langle w, v \rangle$ for all $v, w \in V$.
- Linearity. We must have $\langle v, \alpha w_1 + \beta w_2 \rangle = \alpha \langle v, w_1 \rangle + \beta \langle v, w_2 \rangle$ for all $v, w_1, w_2 \in V$, $\alpha, \beta \in \mathbb{R}$.

Example. Consider the vector space with P^n, where P^n denotes the space of degree n polynomials with real coefficients. We can define an inner product as follows:

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)\,dx.$$

Here, we note that

$$\langle f, f \rangle \geq 0,$$

and it satisfies all the axioms:

- $\int_{-1}^{1} f^2(x)\,dx = 0.$
- $\int_{-1}^{1} f(x)g(x) = \int_{-1}^{1} g(x)f(x)\,dx.$
- It satisfies linearity, since,

$$\begin{aligned}\langle f, \alpha g_1 + \beta g_2\rangle &= \int_{-1}^{1} f(x)\left[\alpha g_1(x) + \beta g_2(x)\right] \\ &= \alpha \int_{-1}^{1} f(x)g_1(x)\,dx + \beta \int_{-1}^{1} f(x)g_2(x)\,dx \\ &= \alpha\langle f, g_1\rangle + \beta\langle f, g_2\rangle.\end{aligned}$$

Definition. We say that a map $V \times V \longrightarrow \mathbb{C}$ is a complex inner product if the following axioms hold:

- Positive definite. That is, $\langle v, v\rangle \geq 0$ for all $v \in V$, and $\langle v, v\rangle = 0$ iff $v = 0$.
- Hermitian symmetry. For all v, w, we have $\langle v, w\rangle = \langle w, v\rangle$.
- Hermitian linearity. We have

$$\langle v, \alpha w_1 + \beta w_2\rangle = \overline{\alpha}\langle v_1, w\rangle + \overline{\beta}\langle v_2, w\rangle.$$

For example, note that $\langle v, w\rangle = v^H w$ is a complex inner product.

We now discuss vector norms. Given a vector space V over a field $\mathbb{F}$, a norm is a map $|| \cdot || : V \longrightarrow \mathbb{R}$ that satisfies the following properties:

- Positive definite. We must have $||v|| \geq 0$ for all v, and $||v|| = 0$ iff $v = 0$.
- Linearity. This means that $||\alpha v|| = |\alpha|||v||$ for all $v \in V$, $\alpha \in \mathbb{R}$.
- Triangle inequality. We have that $||v + w|| \leq ||v|| + ||w||$.

Here are some examples of norms. If $v \in \mathbb{R}^n$, then

- $||v||_2 = \left(\sum_{i=1}^{n} v_i^2\right)^{\frac{1}{2}}$.
- $||v||_p = \left(\sum_{i=1}^{n} v_i^p\right)^{\frac{1}{p}}$.
- $||v||_1 = \left(\sum_{i=1}^{n} |v_i|\right)$ (also known as the Manhattan norm).
- $||v||_\infty = \max_i |v_i|$.

There are some important inequalities we can state related to these norms (the proofs can be found in the book).

Theorem. (Hölder's Inequality). If $\frac{1}{p} + \frac{1}{q} = 1$, then

$$|v^H w| \leq ||v||_p ||w||_q.$$

A special case of Hölder's Inequality is the Cauchy-Schwarz inequality. That is, when $p = q = 2$.

Theorem. (Cauchy-Schwarz inequality).

$$|v^H w| \leq ||v||_2 ||w||_2.$$

Now, given a vector inner product, we can always define a norm induced by the inner product. Suppose $\langle \cdot, \cdot \rangle$ is an inner product from $V \times V \to \mathbb{R}$. Then we can always define a norm induced by this inner product by

$$||v|| = \sqrt{\langle v, v \rangle}.$$

But, it is not true that all inner products come from a norm.

Definition. Suppose that $| \cdot |_\star$ and $|| \cdot ||_*$ are norms of V. We say that they are equivalent if there exists constants $c_1, c_2 > 0$ such that for any $v \in V$ we have

$$c_1 ||v||_\star \leq ||v||_* \leq c_2 ||v||_\star.$$

For motivation, we note that if norms are equivalent, then they will preserve topological properties in the space. For example, if a function converges in the 1 norm, it will converge in the 2 norm.

Example. We have for n dimensional vectors $v \in \mathbb{R}^n$, we have

$$||v||_2 \leq ||v||_1 \leq \sqrt{n} ||v||_2,$$

so the $||v||_1$ and $||v||_2$ are equivalent.

We note that the right inequality is a consequence of Cauchy-Schwarz.

Example. We have for n dimensional vectors $v \in \mathbb{R}^n$, we have

$$||v||_\infty \leq ||v||_1 \leq n ||v||_\infty.$$

Proof. We have

$$||v||_1 = \sum_{i=1}^n |v_i| \leq \sum_{i=1}^n \max |v_i| = n ||v||_\infty.$$

This constant is sharp because equality is achieved when $v = (1, 1, \ldots, 1)$. The opposite inequality is obvious. Clearly, $\max |v_i| \leq \sum_{i=1}^n |v_i|$. This inequality is sharp by taking a vector $(1, 0, \ldots, 0)$. □

Now, we continue to discuss matrix norms.

Suppose $A \in \mathbb{R}^{m \times n}$. Now, while thinking about the map $v \mapsto Av$, we can define a special type of matrix norm on it.

Definition. Suppose we have $|| \cdot ||_*$ on $\mathbb{R}^n$; and $|| \cdot ||_\star$ on $\mathbb{R}^m$. Then, we can define an induced matrix norm as follows:

$$||A||_{*,\star} = \max_{v \neq 0} \frac{||Av||_\star}{||v||_*}.$$

This is also known as the operator norm.

Example. Suppose $A = I_n$, and we consider the 2-norms on both the domain and the co-domain.

$$||I||_{2,2} = \max_{v \neq 0} \frac{||v||_2}{||v||_2} = 1.$$

Example. Suppose now $A = I_n$, and we use the infinity norm in the domain and the 1-norm in the co-domain. That is,

$$||I||_{\infty,1} = \max_{v \neq 0} \frac{||v||_1}{||v||_\infty} = n.$$

This comes from the inequality $||v||_1 \leq n||v||_\infty$ we proved earlier.

7.14 Lecture 17: 7-23-19

Definition. Let V be a vector space with $V = X \oplus Y$. Define $P_{X,Y} : V \to X \subseteq V$ by

$$P_{X,Y}v = x,$$

for all $v \in V$. Then $P_{X,Y}$ is called the (oblique) projection on X along Y.

Definition. We say that P is an orthogonal projection if $Y = X^\perp$.

Theorem. A linear transformation P is a projection iff $P^2 = P$.

Theorem. P is a projection iff $I - P$ is also a projection.

The reason we care about orthogonal projections is that the matrices associated with them are orthogonal.

Theorem. $P \in \mathbb{R}^{n \times n}$ is the matrix for an orthogonal projection onto (P) iff $P^2 = P = P^T$. (In fact, P is a little more than orthogonal, since we require $P^2 = P$).

We will skip the proof. The first part, that $P^2 = P$ follows from the previous result. The trickiest part is the transpose condition (this is proven in the book, see 7.5).

We will now discuss properties of projections.

Theorem. Let P be a matrix for orthogonal projection. Then

$$||Pv|| \leq ||v||,$$

where $|| \cdot ||$ is any vector name on V.

Proof. This follows from the Pythagorean theorem. By this theorem, if $v \perp w$, then $||v + w||^2 = ||v||^2 + ||w||^2$ (this applies to any norm). We know that $v = Pv + (I - P)v$. Now, $Pv \perp (I - P)v$, and so by this theorem

$$||v||^2 = ||Pv||^2 + ||(I - P)v||^2,$$

implying $||Pv||^2 \leq ||v||^2$. □

Example. Project onto the span of v in $\mathbb{R}^2$, so that $P_v : \mathbb{R}^2 \to \mathbb{R}^2$. We can write

$$P_v(u) = \frac{u \cdot v}{v \cdot v} + v.$$

Suppose tha t$v_1, \ldots, v_n$ are orthogonal, and $X = \{x_1, \ldots, x_n\}$. Then

$${}_X(u) = \sum_{i=1}^{n} {}_{v_i}(u) = \sum_{i=1}^{n} \frac{u \cdot v_i}{v_i \cdot v_i} v_i.$$

Suppose that $v_1, \ldots, v_n$ in $\mathbb{R}^m$, $m \geq n$ are orthonormal. Suppose that $e_1, \ldots, e_m$ is the standard basis. Suppose that $P : \mathbb{R}^m \to \mathbb{R}^m$ is the linear map of orthogonal projections on $\{v_1, \ldots, v_n\}$. How do we find the projection matrix? We just put the projection of the i-th basis vector in the i-th column of the matrix.

Now, we discuss the Gram-Schmidt orthogonalization process.

Suppose that $w_1, \ldots, w_n$ are linearly independent. Then we apply this process by following these steps:

1. Normalize w_1, by writing

$$v_1 = \frac{w_1}{||w_1||}.$$

2. w_2 is not in general orthogonal to w_1. So, we set

$$q_2 = w_2 -_{v_1} w_2,$$

and here q_2 is orthogonal to w_1.

Now, normalize, to obtain $v_2 = \frac{q_2}{||q_2||}$.

3. Keep going, so

$$q_3 = w_3 -_{v_1} w_3 -_{v_2} q_3.$$

Now,

$$v_3 = \frac{q_3}{||q_3||}.$$

4. In general, if we continue this process, we obtain

$$\begin{aligned} q_{k+1} &= w_{k+1} - \sum_{i=1}^{k} {}_{v_i} w_{k+1} \\ v_{k+1} &= \frac{q_{k+1}}{||q_{k+1}||}. \end{aligned}$$

And then, the resulting set of v_i will have the same span of the w_i. To see this, note that the w_i are linearly independent and form a basis; similarly since the v_i are orthogonal, they must span the space for dimensionality reasons.

7.15 Lecture 18: 7-24-19

We define the QR factorization as follows. Given a matrix $A \in \mathbb{R}^{m \times n}$, where $(A) = n$, we can write $A = QR$, where $Q \in \mathbb{R}^{m \times m}$, R is an upper triangular $m \times n$ matrix.

To compute the QR decomposition, we can use Gram Schmidt to compute $q_1, \ldots, q_n \in \mathbb{R}^m$ which are orthonormal. And we can compute scalars $r_{ij} = \langle v_j, q_i \rangle$. Suppose we define $Q^1 = \begin{pmatrix} q_1 & q_2 & \ldots q_n \end{pmatrix}, R^1_{ij} = \begin{pmatrix} r_{ij} \end{pmatrix}$. Then we have that $A = Q^1 R^1$.

Now, suppose that $q_{n+1}, \ldots, q_m$ are orthonormal such that $q_1, \ldots, q_m$ form an orthonormal basis for $\mathbb{R}^m$. Then we can write $Q = \begin{pmatrix} q_1 & q_2 & \ldots & q_m \end{pmatrix}$. So we can write

$$A = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1 R_1.$$

One application of QR decomposition is that they allow you to solve systems of equations. Suppose we have $Ax = b$, this is equivalent to $QRx = b$, so $Rx = Q^{-1}b$.

To compute $q_{n+1}, \ldots, q_m$ - we can look at the Span of $\{q_1, \ldots, q_n\}$ which is not equal to $\mathbb{R}^m$. To pick the rest of the vectors: pick v_{n+1} such that $v_{n+1} \in \mathbb{R}^m$, but v_{n+1} is not in the span. Then: just re-apply Gram-Schmidt to compute the rest of the q_i.

We will now discuss least-squares problems, which are used when we cannot solve a system explicitly. For instance, we might have a system $Ax = b$ that cannot be solved explicitly. The most standard case is linear regression - when we want to minimize the L_2 error of the predictions vs. the data. For example, if we are trying to find a function that takes heights to weights, we might have the dataset,

$$\begin{pmatrix} h = 175 & w = 50 \\ 180 & 70 \\ 164 & \\ 55 & \end{pmatrix}.$$

and we want to take $h(w) = aw + b$, and learn constants a, b.

Pictorially, if the blue dots are the data points, we want to find a line that minimizes the sum of the d_i^2.

Definition. The solution of the least squares problem is x_0 such that $||Ax_0 - v||^2$ is minimized.

We can use projections to relate Ax_0 are v. We note that $Ax_0 \in (A)$. We want to pick v where v is the orthogonal projection onto A. Namely, $Ax_0 =_A v$.

The key point is that the vector $v - Ax_0$ is orthogonal to A. This means that $(v - Ax_0) \in A^T$, so $A^T(v - Ax_0) = 0$. Thus, $A^T V = A^T A x_0$.

Now, suppose that A has linearly independent columns. Then $A^T A$ is invertible. Givne the equation $A^T v = A^T A x_0$, we can find $x_0 = (A^T A)^{-1} A^T v$.

Example (linear regression). Suppose we have a set of points (x_i, y_i) for $1 \leq i \leq N$, and we want to solve $y = \alpha x + \beta$. And we can build a matrix $A = \begin{pmatrix} x_1 & 1 \\ \ldots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix}$.

Example (polynomial regression). Suppose we have $y = \alpha_k x^k + \alpha_{k-1}x^{k-1} + \cdots + \alpha_0$. In this case, the system of equations looks like

$$y_i = \alpha_k x^k + \alpha_{k-1}x^{k-1} + \cdots + \alpha_0.$$

In this case, the unknowns are the α_i's. Here, we can build a bigger matrix, where

$$\begin{pmatrix} x_1^k & x_1^{k-1} & \dots & 1 \\ x_2^k & x_2^{k-1} & \dots & 1 \\ \dots & & & \\ x_n^k & x_n^{k-1} & \dots & 1 \end{pmatrix} \begin{pmatrix} \alpha_k \\ \alpha_{k-1} \\ \dots \\ \alpha_0 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}.$$

Now, this is a system $Ax = v$, you want to find the solution that minimizes the squared error - so this is equivalent to the setting from before.

7.16 Lecture 21: 7-26-19

End of last week, we were doing last least squares problems, and yesterday we discussed SVDs.

Recall that the SVD is like a generalization of the eigenvector / eigenvalue decomposition (since not every matrix has an eigenvector / eigenvalue decomposition).

Remember that we can write $A = UDV^T$, where U, V are orthogonal, and D is diagonal (with added zeros as needed).

To obtain the SVD, we perform the following algorithm:

- Compute eigenvalues for A^TA. We are guaranteed to find n eigenvalues; exactly r are nonzero.
- Compute eigenvectors $v_1, \dots, v_n$ (require orthonormal). The spectral theorem allows us to obtain a full basis of eigenvectors, since A^TA is a symmetric matrix.
- Look at the vectors $Av_1, \dots, Av_n$. Then $||Av_i|| = \lambda_i$, so exactly r of Av_i are not zero. We proved yesterday that the Av_i are orthogonal.

This allows us to compute the SVD matrices. We obtain:

$$\begin{aligned} V &= \begin{pmatrix} v_1 & v_2 & \dots & v_n \end{pmatrix} \\ D &= (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_r}, 0, \dots, 0) \\ U &= \begin{pmatrix} \frac{Av_1}{\sqrt{\lambda_1}} & \frac{Av_2}{\sqrt{\lambda_2}} & \dots & \frac{Av_r}{\sqrt{\lambda_r}} & u_{r+1} & \dots & u_n \end{pmatrix}. \end{aligned}$$

Now, the $u_{r+1}, \dots, u_n$ are redundant (because the the right portion of the matrix gets zeroed out be the zero entries in D).

If we write

$$U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}, D = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}, V = \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},$$

we can compute a reduced SVD as $A = U_1 D_1 V_1^T$.

Importantly, here D is unique; but U and V are not unique. However, the eigenspaces associated with U and V are unique.

Geometrically, this is really cool, because it means that you can express any matrix A as the product of an isometry (V^T), a scaling with D, and another isometry (U). These isometries can be reflections, rotations, or rotoreflection.

Properties of vector norms. Recall that $||\cdot|| : V \to \mathbb{R}$ is a function that satisfies the following properties:

- Positive definite. $\langle v, v\rangle \geq 0$ for all v, with equality iff $v = 0$.
- Linearity. $||av|| = |a|||v||$ for all $v \in V, a \in \mathbb{R}$.
- Triangle inequality. We have that $||v + w|| \leq ||v|| + ||w||$.

Definition. We say that a matrix norm is a norm on the vector space $\mathbb{R}^{m\times n}$.

Let's consider a few examples.

- Frobenius norm, defined as

$$||A||_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}} = (\sum_{k=1}^{r} \sigma_k^2)^{1/2}.$$

- Induced p-norm. We can define:

$$||A||_p = \max_{x\neq 0} \frac{||Ax||_p}{||x||_p}.$$

- Induced pq-norm. We can define

$$||A||_{p,q} = \max_{x\neq 0} \frac{||Ax||_q}{||Ax||_p}.$$

For homework, we will show that

$$||A||_2 = \sigma_1(A),$$

i.e. the largest singular value.

Now, we can prove that different norms are equivalent with matrix norms as well.

Definition. We say that $||\cdot||_\alpha, ||\cdot||_\beta$ on $\mathbb{R}^{m\times n}$ are equivalent if there exists $c_1, c_2 > 0$ such that for all matrices $A \in \mathbb{R}^{m\times n}$, we have

$$c_1||A||_\alpha \leq ||A||_\beta \leq c_2||A||_\alpha.$$

To prove matrix norms are equivalent, we can make heavy use of the equivalence of vector norms.

7.17 Lecture 22: 7-31-19

Note that

$$||\lambda \cdot x||_\star = |\lambda|||x||_\star.$$

Now, we note that we can compute matrix norms over the unit ball instead of the whole vector space, since:

$$\max_{x \neq 0} \frac{||Ax||_\star}{||x||_\star} = \max_{||y||_\star = 1} ||Ay||_\star.$$

Now, if $||A||_1 = \max_{||x|| \neq 0} \frac{||Ax||_1}{||x||_1} = \max_{||x||=1} ||Ax||$. This latter problem is maximizing a continuous function on a compact set - so there must be a maximum.

Claim. We have

$$||A||_1 = \sum_{1 \leq j \leq m} \sum_{i=1}^{n} |a_{ij}|; \qquad \text{max col sum.}$$

Proof. Suppose $A_1, \ldots, A_m$ are the column of A. We want to look at

$$\begin{aligned} ||Ax||_1 &= \max_{||x||=1} \sum_{j=1}^{m} x_j A_j \\ &\leq \sum_{j=1}^{m} |x_j A_j|_1 \\ &= \sum_{j=1}^{m} |x_j|||A_j||_1 \\ &\leq \sum_{j=1}^{m} |x_j| \left(\max_{1 \leq k \leq m} ||A_k|_1 \right) \\ &= \left(\max_{1 \leq k \leq m} ||A_k||_1 \right) ||x||_1. \end{aligned}$$

This shows that for all x, we have:

$$\frac{||Ax||_1}{||x||_1} \leq \max_{1 \leq k \leq n} ||A_k||_1.$$

Since this is true for all x, it must be true for the maximum x, that is:

$$||A||_1 \leq \max_{1 \leq k \leq m} ||A_k||_1.$$

Now, note that for any x_0 fixed, we have

$$\frac{||Ax_0||_1}{||x_1||_1} \leq ||A||_1.$$

Let l be such that $||A_l||_1 \leq ||A_k||$ for all $1 \leq k \leq m$.

Now, take $x_0 = e_l$, the l-ith basis vector. Then:

$$\frac{||Ax_0||_1}{||x_0||_1} = \frac{||Al||_1}{||l||_1} = \max_{1\leq k\leq n} ||A_k||_1 = ||A_l||_1.$$

□

Claim. We have

$$\begin{aligned} ||A||_\infty &= \text{maximum row sum} \\ ||A||_2 &= \sigma_1(A); \qquad \text{maximum singular value} \\ ||A||_{pq} &= \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{m}|a_{ij}|^p\right)^{q/p}\right)^{1/q}. \end{aligned}$$

Definition. We say that a matrix norm is consistent if

$$||AB|| \leq ||A||||B||.$$

Claim. All induced norms are consistent.

Proof. To show this, we need to prove a bunch of triangle inequalities. We have:

$$\frac{||Ax||}{||x||} \leq \max_{y\neq 0}\frac{||Ay||}{||y||} = ||A||.$$

Now, this implies that $||Ax|| \leq ||A|| \cdot ||x||$. To continue, we note that

$$||ABx|| \leq ||A||||Bx|| \leq ||A||||B||||x||.$$

Now, dividing both sides by the norm of x, we obtain for $x \neq 0$,

$$\frac{||ABx||}{||x||} \leq ||A||||B||,$$

thus,

$$\max_{x\neq 0}\frac{||ABx||}{||x||} = ||AB|| \leq ||A||||B||.$$

□

In the HW - problem asks whether Frobenius norm is induced by a vector norm. Hint: show that Frobenius norm is not consistent, therefore it is not induced.

Note that all p-norms are equivalent. This is easy to show, if we just note that

$$\left(\sum |x_i|^p\right)^{1/p} \leq \left(\sum \max |x_i|^p\right)^{1/p} = \left(n||x||_\infty^p\right)^{1/p}$$

Also,

$$||x||_\infty^p = \max |x_i|^p \leq \sum_{i=1}^{n}|x_i|^p = ||x||_p^p.$$

The result follows.

Tomorrow, we will discuss the equivalence of matrix norms.

7.18 Lecture 23: 8-1-19

Yesterday, we showed that

$$||A||_1 = \max_{x\neq 0} \frac{||Ax||_1}{||x||_1} = \text{max col sum}.$$

Today, we note that

$$||A||_\infty = \max_{x\neq 0} \frac{||Ax||_\infty}{||x||_\infty} = \text{max row sum}.$$

Proof. If A_i is the i-th row of A, we have

$$\begin{aligned} ||Ax||_\infty &= \max_{1\leq i\leq n} \langle A_i, x\rangle \\ &= \max_{1\leq i\leq n} \left|\sum_{j=1}^m A_{ij}x_j\right| \\ &\leq \max_{1\leq i\leq n} \sum_{j=1}^m |A_{ij}||x_j| \\ &\leq \max_{1\leq i\leq n} \sum_{j=1}^m |A_{ij}|||x||_\infty. \end{aligned}$$

Then,

$$||A||_\infty \leq \max_{1\leq i\leq n} \sum_{j=1}^n |A_{ij}|.$$

And this is achieved if you take x to be (A_m), where A_m is the maximum row.

(Note: this is part of the homework). □

Now, we will show that

$$||A||_F = \left(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2\right)^{1/2} = \left[\sum_{i=1}^r \sigma_i(A)\right]^{1/2}.$$

Proof. Note that if P is orthogonal, we have $||PA||_F = ||A||_F$. This follows since orthogonal matrices preserve 2-norms.

Now, suppose $A = VDQ^H$ is an SVD. By SVD, we know that V and Q^H are orthogonal. So, $||A||_F = ||D||_F = \left[\sum_{i=1}^r \sigma_i(A)^2\right]^{1/2}$. □

We will prove that the infinity norm is equivalent to the 1 norm.

Proof. Using the results for vectors that

- $||x||_\infty \leq ||x||_1$

- $||x||_1 \leq n||x||_\infty$.

Now, for a fixed vector x, we have that:

$$\frac{||Ax||_\infty}{||x||_\infty} \leq \frac{||Ax||_1}{\frac{1}{n}||x||_1}.$$

Now, we have that

$$\frac{||Ax||_\infty}{||x||_\infty} \leq \frac{n||Ax||_1}{||x||_1} \leq n \cdot \max_{y \neq 0} \frac{||Ay||_1}{||y||_1} = n \cdot ||A||_1.$$

Now, we have equality when $A = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & & & \end{pmatrix}$, i.e. the matrix with the first row all 1s, and the rest of the matrix 0.

Now, we can similarly prove that $||A||_1 \leq m||A||_\infty$, where equality is achieved when $A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & \\ \vdots & & & \\ 1 & 0 & \dots & 0 \end{pmatrix}$, that is, the first column is all 1s, and the rest is 0s.

In the homework, we need to show that 2 norms of matrices are equivalent to infinity norms of matrices. □

7.19 Lecture 24: 8-6-19

Definition. We define a Jordan block of size n for eigenvalue λ as

$$J_n(\lambda) = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}.$$

Theorem. For any complex valued matrix A, there exist X where $A = XJX^{-1}$, where J is a Jordan matrix.

We will discuss an algorithm to compute the JCF.

Example. Suppose $A = \begin{pmatrix} 4 & 1 \\ -1 & 6 \end{pmatrix}$. To diagonalize this, we perform the following steps.

First, we compute the eigenvalues: we get

$$\det(A - \lambda I) = \lambda^2 - 10\lambda + 25 = (\lambda - 5)^2 = 0,$$

so $\lambda = 5$ with multiplicity 2.

Next, we find eigenvectors by solving $Av = \lambda v$, that is $v \in (A - \lambda I)$. Solving, we get the eigenvector $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

If we look at the product

$$A\begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 0 & 5 \end{pmatrix}.$$

Equating columns, we obtain

$$\begin{aligned} Av_1 &= 5v_1 \\ Av_2 &= v_1 + 5v_2, \end{aligned}$$

where the second equality reduces to $(A - 5I)v_2 = v_1$. Here, v_2 is called a generalized eigenvector (or a principal vector) for v_1.

Solving $(A - 5I)v_2 = v_1$, we obtain $v_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$. Then $X = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$.

Note that there is a theorem that states that eigenvectors and principal eigenvectors are linearly independent, so we know that X will be linearly independent.

Definition. We define the right principal eigenvectors of degree k associated with $\lambda \in \Lambda(A)$ is a vector such that

$$\begin{aligned} (A - \lambda I)^k &= 0 \\ (A - \lambda I)^{k-1} &\neq 0. \end{aligned}$$

Example. Suppose $A \in \mathbb{C}^{3\times 3}$, and suppose there is one eigenvalue λ with algebraic multiplicity 3, with only one eigenvector. Then, if we look at

$$A\begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix},$$

we can get the following equations:

$$\begin{aligned} Av_1 &= \lambda v_1 \\ Av_2 &= v_1 + \lambda v_2 \\ Av_3 &= v_2 + \lambda v_3. \end{aligned}$$

This means that v_1 is an eigenvector, $(A-\lambda I)v_2 = v_1$, so v_2 is a generalized eigenvector of degree 2. Lastly, we have $(A - \lambda I)v_3 = v_2$ so v_3 is a generalized eigenvector of degree 3.

Algorithm for computing JCF.

- Find the eigenvalues.
- For each eigenvalues, find the eigenvectors.
- Now:
  - If we have n eigenvectors, we are done.

  - If not: for each linearly independent eigenvector with eigenvalue λ such that $AM(\lambda) \geq GM(\lambda)$, we compute degree 2 generalized eigenvectors.

- If # of eigenvectors + # of generalized eigenvectors of degree $2 = n$, we are done. If the sum is $< n$:
  - Continue: for each generalized eigenvector of degree 2, find (if possible), principle vectors of degree 3.
  - Recurse.

To compute X, just put the generalized eigenvectors in order of increasing degree, and set

$$J = (J_{l_1}(\lambda_1), J_{l_2}(\lambda_2), \ldots, J_{l_k}(\lambda_k)).$$

7.20 Lecture 26: 8-8-19

Theorem. If we have the system $x' = Ax$, $x(t_0) = x_0$, the solution to the IVP is given by

$$x(t) = e^{A(t-t_0)} \cdot x_0.$$

Lemma. We have that $\frac{d}{dt}e^{At} = Ae^{At} = e^{At}A$.

Proof. (Sketch). Note that

$$\begin{aligned} \frac{d}{dt}\left[I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} + \ldots\right] &= A + A^2t + \frac{A^3t^2}{2!} + \frac{A^4t^3}{3!} + \ldots \\ &= A\left[I + At + \frac{A^2t^2}{2!} + \ldots\right] \\ &= Ae^{At}. \end{aligned}$$

(Note that here, we're assuming that we can differentiate term by term, which is not always true.[2].) □

Now, we will verify the theorem. We know that

$$\begin{aligned} \frac{dx(t)}{dt} &= Ae^{A(t-t_0)} \cdot x_0 = Ax(t) \\ X(t_0) &= e^{A \cdot 0} \cdot x_0 = x_0. \end{aligned}$$

This verifies the conditions of the theorem.

To solve this problem:

- First, compute the JCF of the matrix A.
- Use the JCF to compute e^{At}.
- Use the initial condition to the get the solution of the IVP.

[2] For more on this, see https://www.dpmms.cam.ac.uk/ agk22/uniform.pdf

Example. Consider the matrix $A = \begin{pmatrix} -4 & 4 \\ -1 & 0 \end{pmatrix}$, and consider the system

$$\begin{aligned} x_1' &= -4x_1 + 4x_2 \\ x_2' &= -x_1, \end{aligned}$$

with initial conditions

$$\begin{aligned} x_1(2) &= 2 \\ x_2(2) &= 4. \end{aligned}$$

To solve this, we need to compute the matrix exponential of this. We can compute that

$$A = XJX^{-1}$$

$$X = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}; \qquad J = \begin{pmatrix} -2 & 1 \\ 0 & -2 \end{pmatrix}.$$

And we found that

$$\begin{aligned} e^{At} = Xe^{Jt}X^{-1} &= \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} e^{-2t} & te^{-2t} \\ 0 & e^{-2t} \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \\ &= \begin{pmatrix} e^{-2t} - 2te^{-2t} & 4te^{-2t} \\ -te^{-2t} & e^{-2t} + 2te^{-2t} \end{pmatrix}. \end{aligned}$$

And to find the solution, we can compute:

$$x(t) = e^{A(t-t_0)} \cdot x_0,$$

(we will skip this computation for now because it is tedious).

Next week, we will discuss LU decomposition. It's an important idea, but it will not appear on the final.

7.21 Lecture 27: 8-12-19

The idea of LU decomposition is to solve equations of the form $Ax = b$, where $A \in \mathbb{R}^{m \times n}$, b is a vector, and x is a vector of unknowns.

- Case 1. The easiest case is when A is upper triangular.
- Case 2. When we have an equation like this:

  $$A = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 4 & 5 \\ 1 & 2 & 4 \end{pmatrix},$$

  we can just use elementary operations to reduce this to case 1.

- Case 3. Things can a bit more subtle when we have 0s in the coefficients, e.g.

$$A = \begin{pmatrix} 3 & 2 & 7 & 4 \\ -6 & -4 & -14 & -16 \\ 9 & 6 & 23 & 10 \\ 0 & 0 & -2 & -2 \end{pmatrix}.$$

  But we can continue as before, and we end up with

$$A' = \begin{pmatrix} 3 & 2 & 7 & 4 \\ 0 & 0 & 0 & -8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & -2 & -2 \end{pmatrix}.$$

  The idea is to switch rows and columns to ensure that $A'_{2,2}$ is no longer 0s. In this case, we will interchange R_2 and R_3, and C_2 and C_3. Note that we have to be a bit careful, since this will change the order of the solutions.

Definition. The LU-decomposition of A consists of 2 matrices L and U, where L is an invertible lower triangular matrix, and U is an upper triangular matrix.

Definition. We say that a matrix is k-upper triangular if the first k columns are upper triangular.

Definition. We say that the k-step LU decomposition of A is $A = L_k U_k$, where L_k is an invertible k-lower triangular matrix, and U is a k-upper triangular matrix.

Add an overview of LU decomposition algorithm

7.22 Lecture 28: 8-13-19

Recall that given $A \in \mathbb{C}^{m \times n}$, the LU decomposition of A is matrices L and M such that:

- L is a lower triangular matrix with determinant 1 ($L \in \mathbb{C}^{m \times m}$).
- U is an upper triangular matrix with no zero on the diagonal ($U \in \mathbb{C}^{m \times n}$),
- such that $A = LU$.

Definition. We described the r-step LU decomposition were $A = L_r U_r$, where we have .

fill in block matrices stuffj

Algorithmically, we might have something as follows:

- **(Input.)** $A \in \mathbb{C}^{m \times n}$, m, n, r.
- **(Output.)** L_k, U_k, the k step LU factorization where k is the max number $\leq r$ such that the algorithm can continue.[3]

```
U = A;
L = I_m;
if m = 1 # we are done, nothing to eliminate.
  return L, U
  for k = 1, ..., r do
    (step k of Gaussian elimination)
```

[3] This is necessary since we are doing the algorithm without pivoting.

```
    for i = k+1, ..., m do
    \dots
```

Source: page 57 of Kazeev's notes.

7.23 Final review

Definitions.

- Field (ACID)$^2 roughly Vectorspace(ACID, AID)$

- SVD definition, and existence
- projection (square, ortho)
- pseudoinverse (briefly)
- norms, induced norms
- 7.29 part 1 (defn of p-norms)
- Holder's inequality For complex vectors - we have

$$| < x, y > | \leq ||x||_p ||y||_q,$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

- Unitary invariance of 2-norms. If U is unitary invariant, then $||Ux||_2 = ||x||_2$. Unitary: $UU^H = I$.
- Pythag
- 7.46 - 7.47 (operator matrix norm)

Operator norm:

$$|| \cdot ||_{1,\infty} = \max_{x \neq 0} \frac{||Ax||_1}{||x||_\infty}$$

$$||A||_{p,q} = \max_{||X||_q \neq 0} \frac{||Ax||_p}{||x||_q}.$$

- First equation, defn. of Frobenius

It's just L^2 of all entries.

- Least squares problem
- Spectrum (set of all EVs).
- Solution of linear / ODE

Theorems.

- 3.19, P1 If A, B in $R^{n \times n}$, then

  - $0 \leq (A+B) \leq (A) + (B)$.

- 4.11 (pseudoinverse under change of basis) $(UAV)^{+} = V^T A^{+} U^T$. Suppose U, V are orthogonal.
- 5.11 (properties of SVD)
  - Rank A is R, count of SVs.
  - dyadic outer prod.
  - eigenvector relations $Av_i =_i u_i$
  - $R(U_1) = R(A) = N(A^T)^{\perp}$

  <++>

- 7.5 P is the matrix of an orthogonal projection onto $R(P)$ iff $P^2 = P = P^T$. sym + square.
- 9.12 (spectrum of a Hermitian matrix) Recall hermitian : $A = A^H$. All eigenvalues must be real. Take hermitian TP and conclude $\overline{\lambda} = \lambda$.
- 9.13 (orthogonality of the eigenspaces of a Hermitian) If λ, μ are distinct eigenvalues of A with x, z right EVs, then x, z are orthogonal.
- 9.14 (linear independence of eigenvectors - distinct eigenvalues). If λ_i are distict, then right / left eigenvectors are LI.

brief:

- Row / column ranks are equal. Follows from $N(A)^{\perp} = R(A^T)$.
- 3.21, P1, 3 (brief) $R(A) = R(AA^T)$. $N(A) = N(A^T A)$.
- 7.12 / 7.18 (brief) $\langle x, y\rangle_Q = x^T Q y$, weighted inner product. Works for complex case.
- 8.1 (linear least squares problem)

  Let $A \in R^{m\times n}, B^{m\times k}$. Then, the gneeral solution to

  $$\min_X ||AX - B||_2$$

  is of the form

  $$X = A^{+}B + (I - A^{+}A)Y,$$

  where Y is arbitrary. Key idea: if x is full rank, then $x = (A^T A)^{-1} A^T b$.

  It's $(A^T A)^{-1} A^T b$. $(A^T A)^{-1} A^T b$.

- 9.20 Matrix exponential under similarity

  If $X^{-1}AX = \Lambda$, where λ is diag, we have

  $$e^{tA} = X(e^{\lambda_1 t}, \ldots, e^{\lambda_n t})X^{-1}.$$

- 9.25 determinant of a matrix is product of eigenvalues (counted w multiplicity). [x]

Homework 1.

- 3.1 Note any element A can be writte nas $A = A' + A''$. Not hard. Show that the intersection is triv.
- 3.4. We have $M_n'' = (M_n')^\perp$ and $M_n' = (M_n'')^\perp$. Orthogonality of matrix subspaces. Just show orthogonality.
- 3.6. $R \subseteq S$ iff $S^\perp \subseteq R^\perp$. Makes sense.
- 4.1 did this.
- 4.4 fine.
- 5.3. If A, B have the same right EVs, then $A = CDC^{-1}, B = CEC^{-1}$. Change of basis. Then using the prev expressions and commutativity,

$$AB = (CDC^{-1})(CEC^{-1}) = CDEC^{-1} = CEDC^{-1} = (CEC^{-1})(CDC^{-1}) = BA.$$

  makes sense.
- 5.4. Take conj. transposes - show that $-\lambda = \lambda$.
- 5.5. Diagonalize and go.
- 6.1. Verify axioms - $< A, A > \geq 0$ with 0 iff $A = 0$. Linearity
- 7.1 Use normalize unit vectors / scalar.
- 7.3a. Use conj. transpostition.
- 7.4. Max row sum. Show upper / lower bound. $||A||_2$ is σ_1 is harder. Note: 2-norms are unitarily invariant. $||S||_2 = \sigma_1$. Consider various vectors, then you get something that works.
- 8.1a. just use det = prod of diagonal.

Homework 2.

- 4.1. $AA^T = A^T A = I$. Since $v_i \in \mathbb{R}^n$ are orthonormal, $\langle Av_i, Av_j \rangle = v_i^T A^T A v_j$, then use A is orthogonal and get orthogonality.

  Then ineer product of Av_i, Av_j is $v_i^T A^T A v_j$, 0 if $i \neq j$, 1 otherwise. Since $v_i^T (A^T A v_j) = 1$. The v_i are n orthonormal vectors in $\mathbb{R}^n$, so they form a basis.
- 4.4. Compute lim / PE. Conditions (MP): $ABA = A, BAB = B, (AB)^T = AB, (BA)^T = BA$.
- 5.1. CH to compute A^{-1}. Multiply by A^{-1}
- 5.2. Char poly of A is $\det(A - xI) = x^3 + 5x^2 - 8x + 4$. Consider each eigenspace / basis. Defective - does not have a complete basis of EVs (not diagonalizable).
- 6.2 ∞ and 2 is easy. $\infty \leq 2$, but $v = [1, 0, \ldots]$ And $2^2 \leq n\infty^2$. Makes sense.
- 6.3 Need to show $P^2 = P = P^T$ (symmetric and $P^2 = P$, maybe idempotent? not sure).
- 7.5. matrix norms of $2, \infty$ are equiv.

  maybe just recall vector case, and try to generalize
- 8.3 See below.

- 8.4 JNF. The possible JNF blocks correspond to the partitions. Consider diag matrices with 1 on superd. And then just partition stuff.

  Don't need to compute eigenvectors to get JNF.

8

MATH113: Matrix Theory

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine

Theorem Definition Remark Claim Example Lemma Proposition

latexsym bbm [small,bf]caption2 graphics epsfig amsopn url

Ker

[parfill]parskip [margin=1in]geometry

MATH 113 - Matrix Theory Instructor: Michael Kemeny; Notes: Adithya Ganesh

Contents

8.1 Lecture 1

8.1.1 Course logistics

- Website: `web.stanford.edu/~mkemeny/math113.html`
- Grade breakdown
  - 30% homework
  - 30% midterm
  - 40% final
- Office hours: Tues 2pm (382-E).
- Text: Axler, Linear Algebra Done Right.

8.1.2 Introduction

Objective. Solving linear equations in an abstract way.

Linear algebra is a useful fundamnetal framework to have. In some sense linear algebra is really "all we can do." And in higher math classes, we can often reduce problems to linear algebra.

Example. In differential geometry, you will take a linear approximation of the object using a tangent plane. And you can approximate your function using a linear mapping between vector spaces. Calculus is a special case of this kind of setting.

Think of this course as somewhere between philosophy and an applied engineering math class. It is key to understand the proofs and think about theory deeply. A good mental exercise: take a theorem that you think you understand - and try to reproduce the steps of the proof. Also - look for patterns; notice when different theorems use similar arguments.

8.1.3 Fields

A field is a set $\mathbb{F}$, on which it makes sense to add elements, subtract elements, multiply, and divide. Furthermore, these operations should satisfy the usual rules of arithmetic.

In some sense, there are two operations which are more primary (addition / multiplication), and subtraction / division can be viewed as inverse operations.

Definition. A binary operation f is a function $f : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$. That is, for any two elements $x, y \in \mathbb{F}$, we are given a rule to associate a third element $z = f(x, y)$ (note that order matters).

Examples of things we want to be fields or not

- $\mathbb{Q}$, the rational numbers. Clearly we have binary operations $+$ and $\times$. Further, $\mathbb{Q}$ has an additive and multiplicative identity (0 and 1 respectively). We also have an inverse operation to addition. Note: we almost have an inverse operation for multiplication (0 cannot be inverted).
- $\mathbb{N} \cup \{0\}$. This is not a field, because we don't have additive inverses.
- Even if we add additive inverses, $\mathbb{Z}$ is not a field (we don't have multiplicative inverses for nonzero elements).
- Consider $\mathbb{R}^2 := \{(x, y) | x, y \in \mathbb{R}\}$. We can clearly define addition $((x, y), (x', y')) \to (x + x', y + y')$. There is an additive identity $(0, 0)$, and an additive inverse $(-x, -y)$. We can define multiplication is the same way: componentwise. And there is a multiplicative identity $(1, 1)$. But we have a problem: we cannot invert points $(x, 0)$ or $(0, y)$ where $x, y \neq 0$.

We will now formally define fields.

Definition. A field $\mathbb{F}$ is a set together with two binary operations $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ and $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$, satisfying he following axioms.

- There exists additive and multiplicative identities. That is, $\exists\, 0, 1 \in \mathbb{F}$ with $0 \neq 1$ such that

$$x + 0 = x \,\forall x \in \mathbb{F},$$

$$x \cdot 1 = x \,\forall x \in \mathbb{F}.$$

- Existence of inverses. For all $x \in \mathbb{F}$, there exists $-x$ such that

$$x + (-x) = 0.$$

And for all $x \neq 0 \in \mathbb{F}$, there exists x^{-1} such that

$$x \cdot x^{-1} = 1.$$

.

- Commutativity of the operations. We have

$$x + y = y + x \,\forall x, y \in \mathbb{F}$$

$$x \cdot y = y \cdot x \,\forall x, y \in \mathbb{F}$$

.

- Associativity of the operations (this one is somewhat non-obvious). We have

$$x + (y + z) = (x + y) + z \,\forall x, y, z \in \mathbb{F}$$
$$x \cdot (y \cdot z) = (x \cdot y) \cdot z \,\forall x, y, z \in \mathbb{F}$$

- Distributivity of the operations (this can be viewed as a sort of compatibility between addition and multiplication). In particular,

$$x \cdot (y + z) = x \cdot y + x \cdot z.$$

Example. Let's go back to $\mathbb{R}^2$. With a different definition of multiplication on $\mathbb{R}^2$, we can make it a field. We will leave addition as defined earlier (componentwise addition).

Define $(a, b) \cdot (c, d) := (ac - bd, bc + ad)$. Then $(1, 0)$ is a multiplicative identity, since

$$(a, b) \cdot (1, 0) = (a, b).$$

Using this rule, it's straightforward to see that we have multipicative inverses. In particular, observe that

$$(a, b)^{-1} = \frac{1}{a^2 + b^2}(a, -b),$$

for $(a, b) \neq (0, 0)$.

Claim. There exists an element $x \in \mathbb{R}^2$ with this operation with the property

$$x \cdot x = -1.$$

Note: $(0, 1)$ satisfies this, since

$$\begin{aligned}(0, 1) \cdot (0, 1) &= (0^2 - 1^2, 0 + 0)\\ &= -(1, 0) = -1.\end{aligned}$$

Defining $i = (0, 1)$, we can get the cmplex numbers, where we write $x + iy$ for (x, y).

Next time: we will define vector spaces. In words, a vector space V over a field $\mathbb{K}$ will be a set for which we have two operations, vector addition (rule for adding elements of V) and scalar multiplication. (rule for multipliying elements of V by elements in K). Elements in V are called vectors, elements in $\mathbb{K}$ are called scalars.

8.2 Lecture 2

8.2.1 What it means to read proofs

Any proof consists of a sequence of statements. Need to read proofs sequentially - it is critical to understand each statement before you move to the next statement. If you get stuck, think deeply about the statements.

8.2.2 Vector spaces

Last time, we defined fields as sets equipped with two operations: addition and multiplication. The main fields we saw were , $\mathbb{R}$ and $\mathbb{C}$. We defined the field of complex numbers as giving $\mathbb{R}^2$ a special multiplication operation. In the homework you will encounter a field $\mathbb{F}_p$ with p elements, where p is a prime.

A vector space V over a field $\mathbb{K}$ is a set V equipped with two operations, "vector addition" and "scalar multiplication." We call elements $v \in V$ vectors and elements $\lambda \in \mathbb{K}$ scalars. Vector addition will be a rule for adding together two vectors. Scalar multiplication will be a rule for multiplying a scalar by a vector.

Motivating example. Consider the plane in $\mathbb{R}^2$. A vector in $\mathbb{R}^2$ will be a vector space over $\mathbb{R}$.

To add vectors, consider $\vec{v} = (a, b)$ and $\vec{w} = (a', b')$. Then $\vec{v} + \vec{w} = (a + a', b + b')$.

For scalar multiplication, if $\lambda \in \mathbb{R}$ and $\vec{v} = (a, b)$, we have $\lambda\vec{v} = (\lambda a, \lambda b)$.

Geometrically, in vector addition, we move the head of $\vec{w}$ to the tail of $\vec{v}$ to obtain the sum.

If $\lambda > 0$, then geometrically, multiplying by λ has the effect of scaling $\vec{v}$ by a factor of λ. If $\lambda < 0$, then $\lambda\vec{v} = -|\lambda|\vec{v}$, where we scale by λ, and the minus sign changes the direction. Note, we will soon consider objects that cannot be geometrically analyzed, so one shouldn't be too attached to geometric intuition.

Importantly, this extends to any field $\mathbb{K}$. We can define

$$\mathbb{K}^2 = \{(a, b) | a, b \in \mathbb{K}\},$$

with the same addition and multiplication operations:

$$\begin{aligned}(a, b) + (a', b') &= (a + a', b + b')\\ \lambda(a, b) &= (\lambda a, \lambda b).\end{aligned}$$

Note that the above equations, the + on the left side is a different operation than the + on the right side. The left + is vector addition, while the right + is addition over the field $\mathbb{K}$. More concretely, we can define $+_v$ as vector addition and $\cdot_v$ as scalar multiplication. We would obtain

$$\begin{aligned}+_v &: V \times V \longrightarrow V\\ \cdot_v &: \mathbb{K} \times V \longrightarrow V\end{aligned}$$

Aside. The difference between $\longrightarrow$ and $\longmapsto$: we can write $f(x) = \cos(x)$ as

$$\begin{aligned}f &: \mathbb{R} \rightarrow \mathbb{R}\\ x &\mapsto \cos x\end{aligned}$$

This allows us to differentiate the function signature from the formula.

We now proceed to the formal definition.

Definition. Let $\mathbb{K}$ be a field. A vector space V over $\mathbb{K}$ is a set V together with two operations.

Vector addition.

$$\begin{aligned} V \times V &\to V \\ (\vec{v}, \vec{w}) &\mapsto \vec{v} + \vec{w} \end{aligned}$$

Scalar multiplication.

$$\begin{aligned} \mathbb{K} \times V &\to V \\ (\lambda, \vec{v}) &\mapsto \lambda\vec{v}. \end{aligned}$$

Satisfying the following axioms:

- Commutativity of vector addition. For all $\vec{v}, \vec{w} \in V$, we must have $\vec{v} + \vec{w} = \vec{w} + \vec{v}$. (Note: formally, we have not defined operations of $V \times \mathbb{K}$, so we don't have commutativity of scalar multiplication.)
- Associativity of vector addition / scalar multiplication. For all $\vec{u}, \vec{v}, \vec{w} \in V$, we have

$$(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$$

In addition, if we have $\lambda, \mu \in \mathbb{K}$, we have

$$(\lambda\mu)\vec{v} = \lambda(\mu\vec{v})$$

Note that these are different multiplication operators (multiplication in the field, and scalar multiplication in the vector space).

- The usual axioms for addition / multiplication.
  - Existence of an additive identitty. There exists

$$0 \in V;\ \text{such that } \vec{u} + 0 = \vec{u};\ \forall u \in V$$

  - Existence of an additive inverse.

$$\forall \vec{v} \in V,\ \exists - \vec{v} \in V;\ \text{such that } \vec{v} + (-\vec{v}) = 0.$$

  - Existence of a multiplicative identity. There exists some element $1 \in \mathbb{K}$ such that

$$1\vec{v} = \vec{v}; \forall \vec{v} \in V.$$

  - Distributivity. We have

$$\begin{aligned} \lambda(\vec{u} + \vec{v}) &= \lambda\vec{u} + \lambda\vec{v} \\ (\lambda + \mu)\vec{v} &= \lambda\vec{u} + \lambda\vec{v} \end{aligned}$$

for all $\lambda, \mu \in \mathbb{K}, \vec{u}, \vec{v} \in V$.

The bad news: we have six axioms to learn. But the good news: we mostly care about 1 (maybe 2) examples.

Main example. (Finite dimensional vector space) Let $V = \mathbb{K}^n$, the set of ordered n-tuples of elements $(x_1, \dots, x_n) | x_i \in \mathbb{K}, 1 \leq i \leq n$. (If $n \geq 4$, we do not try to visualize this).

Then $\mathbb{K}^n$ is a vector space over $\mathbb{K}$ with addition defined as follows.

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n)$$

Similarly, we can define scalar multiplication as follows:

$$\lambda(x_1, \dots, x_n) = (\lambda x_1, \dots, \lambda x_n).$$

Exercise. Check that the axioms hold.

More general example. (Infinite dimensional vector space) Let $\mathbb{K}$ be a field, and let S be any set. Let V be the set of functions $f : S \to \mathbb{K}$. We can define addition and scalar multiplication.

- (Addition) Let $f_1, f_2 \in V$, i.e. $f_1 : S \to \mathbb{K}$, $f_2 : S \to \mathbb{K}$. We can define $(f_1 + f_2)(s) := f_1(s) + f_2(s)$.

$$\begin{aligned} f_1 + f_2 : S &\to \mathbb{K} \\ s &\mapsto f_1(s) + f_2(s). \end{aligned}$$

- (Scalar multiplication) For any $\lambda \in \mathbb{K}$, we can define $\lambda f_1 : S \to \mathbb{K}$ as

$$\lambda f_1(s) = \underbrace{\lambda f_1}_{\text{multiplication in } \mathbb{K}} (s).$$

Remark. Importantly, note that the first example is a special case of the second example. Take $S = \{1, \dots, n\}$. Then you can think of an n-tuple as a function

$$f \{1, \dots, n\} \to \mathbb{K}$$

where we can think of an n-tuple as a function:

$$\begin{aligned} (x_1, \dots, x_n) &\in \mathbb{K}^n \\ f(x_i) &\mapsto x_i \end{aligned}$$

One advantage of this formulation is that we can think of an infinite sequence $(x_1, x_2, \dots)$ as a function $f : \mathbb{N} \to \mathbb{K}$. So the set of sequences $\{(x_1, \dots,) | x_i \in \mathbb{K}\}$ forms a vector space. That is, $\{f : \mathbb{R} \to \mathbb{K}\}$ is also a vector space.

Concretely, consider $k = \mathbb{R}$, and $n = 2$. If we have the vector $(3, 5)$, we can consider this as a function f with $f(1) = 3$ and $f(2) = 5$.

8.3 Lecture 4

Recall that $U \subset V$ is a vector subspace if

- $0 \in U$
- Closed under addition
- Closed under scalar multiplication

```
  Let $V = k^2$, that is
  \[
    V = \left\{ (a, b) | a, b \in k \right\}
  \]

  Choose any $v = (a, b)$.  Then line
  \[
    L_v = \span(v) = \left\{ \text{all scalar multiples of $v$} \right\}
  \]
  is a subspace.
```

It is easy to show that the additive identity exists, and that it satisfies closure under addition and scalar multiplication.

Definition. Let V, W be k vector spaces. Let $T : U \to W$ be a linear map. This means that

- $T(a+b) = T(a) + T(b)$ "linearity"
- $T(\lambda a) = \lambda T(a)$ "homogeneity"

Then we define the kernel of T (or nullspace) as the subset of V such that

$$T = \{v \in V | T(v) = 0\}.$$

Lemma. Let $T : V \to W$ be a linear map. Then we have $T(0) = 0$.

Proof. Use the fact that

$$0 = 0 + 0$$

So

$$T(0) = T(0) + T(0)$$

Subtracting, we get $T(0) = 0$. □

Note that if T is a linear map, then T is injective, i.e. if $v, w \in V$ such that $T(v) = T(w)$ then $v = w$, if and only if $T = \{0\}$.

So intuitively, the kernel is a kind of "measure" of how far T is from being injective.

Proposition. $T \subset V$ is a subspace.

Proof. We will prove each axiom sequentially.

- Note that $0 \in T$ since $T(0) = 0$ from the lemma.
- Now, need to check closure under addition. By linearity, note that

$$T(v + w) = T(v) + T(w) = 0 + 0 = 0.$$

- Now, need to check closure under scalar multiplication. By homogeneity, note that

$$T(\lambda v) = \lambda T(v) = 0.$$

□

```
  Consider the set
  \[
   U: \left\{ f: \RR \to \RR | f^{(n)} : \RR \to \RR \text{ exists and is continuous }\right\}
  \]
  \begin{itemize}
    \item
  \end{itemize}
```

8.4 Notes on 3.F: Duality

Definition. A linear functional on V is a linear map from V to F, i.e. an element of $\mathcal{L}(V, \mathrm{F})$.

Definition. The dual space of V, denoted V' is the vector space of all linear functions on V. In other words, $V' = \mathcal{L}(V, \mathrm{F})$.

Note that $\dim V' = \dim V$. This follows from 3.61, which states that $\dim \mathcal{L}(V, W) = \dim(V)\dim(W)$.

Definition. If $v_1, \ldots, v_n$ is a basis of V, then the dual basis of $v_1, \ldots, v_n$ is the list $\phi_1, \ldots, \phi_n$ of elements of V' where each ϕ_j is the linear functional on V such that

$$\phi_j(v_k) = \begin{cases} 1; & \text{if } k = j \\ 0; & \text{if } k \neq j. \end{cases}$$

8.5 Key Ideas

9

MATH120: Group Theory

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine

Theorem Definition Remark Claim Example Proposition Solution

latexsym bbm [small,bf]caption2 graphics epsfig amsopn url

[parfill]parskip [margin=1in]geometry

sign Aut GL Ker im Syl

[parfill]parskips [margin=1in]geometry

MATH 120 - Groups and Rings Instructor: Church; Notes: Adithya Ganesh

Contents

9.1 Lecture 4:

9.1.1 Homomorphisms

Two ways in which a homomorphisms can arise.

- Define a function completely, and ask if its a homomorphism.
-

Example. Consider the group $GL_2\mathbb{R}$. Consider the function called the determinant:

$$\det GL_2\mathbb{R} \to \mathbb{R}^x.$$

Fix matrix

$$\det \begin{pmatrix} a & b \\ cd & \end{pmatrix} = ad - bc.$$

We know the value of the function unambiguously. The way to determine whether this is a homomorphism is to ask whether

$$\det(AB) = \det A \det B.$$

Side comment on notation. Note that $\mathbb{R}^x$ should be viewed as the nonzero elements of $\mathbb{R}$ as a group under multiplication.

$$\mathbb{R}^x = \mathbb{R} - \{0\}, \times$$
$$\mathbb{C}^x = \mathbb{C} - \{0\}, \times.$$

What about $\mathbb{Z}^x$? Clearly the nonzero elements of $\mathbb{Z}$ under multiplication is not a group.

So concretely,

$$\mathbb{R}^x = \{\text{elements of } \mathbb{R} \text{ with multiplicative inverses in } \mathbb{R}\}$$

Generalizing this to $\mathbb{Z}^x$, we know $\mathbb{Z}^x = \{1, -1\}$ all have inverses.

Another example:

$$(\mathbb{Z}/8\mathbb{Z})^x = \{1, 3, 5, 7\}$$

In general,

$$(\mathbb{Z}/n\mathbb{Z})^x = \{m \text{ such that } n \text{ and } m \text{ are relatively prime}\}$$

Example. Consider the absolute value function:

$$\mathbb{R}^x \longrightarrow \mathbb{R}^x_{>0}$$
$$x \longmapsto |x|.$$

Since it is true that

$$|xy| = |x||y|,$$

we know that the absolute value is a homomorphism.

Example. Consider the sign function.

$$\mathbb{R}^x \longrightarrow \{\pm 1\}$$
$$x \longmapsto \begin{cases} +1; & \text{if } x > 0 \\ -1; & \text{if } x < 0. \end{cases}$$

Clearly,

$$(xy) = (x)(y).$$

Example. Consider the map

$$\mathbb{R} \longrightarrow_2 \mathbb{R}$$
$$x \longmapsto \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix}$$

Is it true that

$$(cos(x + y), -\sin(x + y), sin(x + y), cos(x + y)) = (cosx - sinxsinxcosx)(cosy, -siny, siny, cosy)$$

$$\begin{pmatrix} \cos x + y & -\sin(x + y) \\ \sin(x + y) & \cos(x + y) \end{pmatrix} = \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix} \begin{pmatrix} \cos y & -\sin y \\ \sin y & \cos y \end{pmatrix}$$

This is tricky, but it is

$$\mathrm{rot}(x + y) = rot(x) \circ \mathrm{rot}(y)$$

Notice that we can also show that there is a homomorphism on rotation without knoing the exact formula for the matrix.

9.1.2 Approach 2: Bottom-up homomorphisms

In this setting, partially define the map. Define the function on generators for a group. Then, ask if there exists a homomorphism (alternative phrasing: ask if it extends to a homomorphism).

Example. Let $G = \mathbb{Z}_4 = \{1, x, x^2, x^3\}$ with $x^4 = 1$.

Questions we can ask

Is there a homomorphism from $f_1 : \mathbb{Z}_4 \to GL_2\mathbb{R}$ with $f(x) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$?

Is there a homomorphism with $f_2(x) = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$, and $f_3(x) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

Clearly f_1 and f_3 with matrix multiplication operations end up being homomorphisms (since the matrix raised to fourth powers are the identity). But f_2 is not: since if you raise it to the fourth power, you do not get the identity.

Key observation:

- Whether or not there is a homomorphism, there is at most one. i.e. If there is one, it's unique. Why? Because if it is a homomorphism, we must have

$$f_1(x) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix};$$

$$f_1(x^2) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^2.$$

$$\dots$$

We also know that

$$f_1(x^4) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^4$$

and

$$f_1(x^5) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^5$$

But $x = x^5$, so $f(x)$ must equal $f(x^5)$!

But the element C from f_3 has order 2! We must have $C^4 = 1$ to obtain a well defined homomorphism, and this is fine).

There is only one homomorphism for each question above.

Question: is there a homomorphism $f : \mathbb{Z} \to \{\pm 1\}$ with $f(2) = 1$?

Clearly, there is a trivial homomorphism $f(anything) = 1$. We also have $f'(k) = (-1)^k$.

If G is generated by $\{x, y, z\}$ and you pick $p, q, r \in H$, there's at most one homomorphism.

Suppose you know $f_1 : G \to H$ and $f_2 : G \to H$. You want to know if $f_1 = f_2$. Just need to check if $f_1(x) = f_2(x)$ for all x.

This is very much like the theorem in linear algebra that says the value of a linear transformation is determined by its value on a basis.

Key observation 2. In general, its hard to know if there is a homomorphism or not. Suppose we had asked, instead that

Example. Let $G =$ subgroup of $GL_2\mathbb{C}$ generated by $a = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$ and $b = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Is there a homomorphism from $G \to \{\pm 1\}$ with $f(a) = -1$ and $(b) = -1$?

Problem: we don't have a list of all the coincidences in the group G. For example, suppose we had to know $a^2bab = -1$ and $b^{-1}aba^2b = 1$. Without the list of coincidences, can't check whether this extends to a valid homomorphism.

In the previou ssetting, we really do have a list of complete concidences, $x^k = x^l$, only if $k \equiv l \pmod 4$.

In the future, we will use the following notation, $Z_4 =< x|x^4 = 1 >$ which is called a group presentation. The only relation you need to know and all other relations follow from that.

Q: With enough computation, is it possible to systematically enumerate these coincidences?

A: For this subgroup of GL2 C, the answer is yes (since there are only 8 elements). But for an infinite group, this isn't in general possible, because you run into the halting problem. This is broadly referred to as the "word problem".

Frequent setting: you can have homomorphisms f with $f : G \to (X)$ in a set, $Perm(X) =$ group of permutations (bijections) with $g : X \to X$.

9.2 Lecture 5

Notation: given a subset $T \subset G$, the notation

$$< T >= \text{subgroup of } G \text{ generated by } T$$

$$< T >= \cap H_{\text{all subgroups } H < G}$$

Recall:

$$\text{Perm}(X) = \text{group of bijections } g : X \to X \text{ under O}$$

The symmetric group S_n is defined as

$$S_n = \text{Perm}(\{1, 2, \dots, n\})$$

Note the idea of cycle decomposition. Suppose we have a setting where

$$g(1) = 4$$
$$g(2) = 3$$
$$g(3) = 2$$
$$g(4) = 5$$
$$g(5) = 1$$

We can also draw out a diagram.

Alternatively, we can decompose this into cycles. Can write this as

$$g = (1\ 4\ 5)(2\ 3)$$

This is called the "cycle decomposition" of g, where we express the group element as a product of disjoint cycles.

Usually we drop length-1 cycles. For example, in a setting with

$$h = (1\ 2)(3)(4)(5),$$

we would usually write

$$h = (1\ 2)$$

Note that symmetric groups are not abelian! The order of function composition definitely matters. However - disjoint cycles do commute with each other.

What is the order of a permutation $\sigma \in S_n$?

For example, the order of $\sigma = (2\ 3)$ is two. In general, the order of a cycle decomposed permutation is the LCM of the cycle lengths.

Cycle decomposition is unique up to ordering of each cycle + ordering withing each cycle.

9.2.1 Order

HW question. Recall the optional homework question - had to show that if $|g| = 2$, then $|G|$ is even.

More general statement. More generally, if $|g| = k$, then $|G| \equiv 0 \pmod{k}$. (If g is a generator of the group G, then we know that $|G| = k$ exactly).

Remark. Any element g corresponds to a cyclic subgroup $< g >$.

Even more general. (Lagrange's Theorem) If H is any subgroup of G, then $|G|$ is divisible by $|H|$.

Notation. The index of H in G is the whole number $|G|/|H|$. So, for example, if $|G| = 100$, and $H < G$, with $|H| = 25$, then the index $[G \cdot H] = 4$.

Importantly, this makes sense even when we have infinite groups. Consider $G = \mathbb{Z}$, with $H < G =$ the multiples of 2. We can still say that $[G \cdot H] = 2$.

```
 Let $p$ be a prime number.  Suppose $|G| = p$.  Then every $g \in G$ has $|g| = 1$ or $|g|
```

Outline of proof. Define the following equivalence relation. Given a subgroup $H < G$, we say $x \sim y$ iff there exists $h \in H$ such that $y = xh$.

This is an equivalence relation exactly because H is a subgroup. Note that:

- Reflexivity holds. (since $1 \in H$)
- Symmetry holds. (since H is closed under inverses)

- Transitivity holds. (since H is closed under multiplication)

We will say that the equivalence class of x is called xH is called its left coset (of H in G). In particular,

$$\begin{aligned} xH &= \{y|x \sim y\} \\ &= \{y|\exists h \text{ s.t. } y = xh\} \\ &= \{xh|h \in H\} \end{aligned}$$

The key to the argument, which we will show on Friday, is that $|xH| = |H|$. Therefore, G can be partitioned into a bunch of cosets (which are all the same size); and hence

$$|G| = (\# \text{ of cosets}) \cdot |H|$$

9.3 Lecture 6

Problem setting. Let G be a group, $H < G$, and let $g \in G$. We will discuss three notions of translations of H by g.

- Left coset. $\{gH = \{gh|h \in H\}\}$
- Right coset. $\{Hg = \{hg|h \in H\}\}$
- Conjugate (of H by g) $gHg^{-1} = \left\{ghg^{-1}|h \in H\right\}$

```
  Let $G = S_5$, and suppose $H = \left\{ \sigma \in G | \sigma(2) = 2 \right\}.$  Let $g = (
1}$.

   Note that $|G| = 120$, $H = 24$.  Note that $|gH| = |Hg| = |gHg^{-
1}| = 24$.  Compute the cosets and the conjugate.
```

Note that

$$gH = \{\sigma \in S_5|\sigma(2) = 3\}.$$

$$Hg = \{\sigma \in S_5|\sigma(1) = 2\}.$$

$$gHg^{-1} = \{\sigma \in S_5|\sigma(3) = 3\}.$$

Also, it is clear that gH and Hg are not subgroups (not preserved under composition). However, gHg^{-1} is a subgroup.

```
  Let $G = \text{Isometries}(\RR^2)$, that is, distance preserving bijec-
tions in the plane.  Let
  \begin{align*}
    H &= \left\{ h \in G | h(\mbf{0}) = \mbf{0} \right\} \\
    g &= 90^{\circ} \text{ rotation around }  (1, 0)
  \end{align*}

  Compute the cosets and conjugate.
```

We can compute that

$$\begin{aligned} gH &= \{\gamma \in G | \gamma(0) = g(0)\} \\ Hg &= \{\gamma \in G | \gamma(q) = 0\} \\ gHg^{-1} &= \{\gamma \in G | \gamma(p) = p\} \end{aligned}$$

(where $p = g(0)$).

Question 1 on HW2. Answer is $K = \mathbb{Z}$. Look at solutions for details.

On homework, discussed the notion of a kernel. If $f : G \longrightarrow Q$, then

$$(f) = \{g \in G | f(g) = 1\}.$$

We can ask a question. Can every subgroup $H < G$ be the kernel of something?

Consider an example of $G = S_3 = \{e, (12), (23), (13), (123), (132)\}$. Set $H = \{e, (12)\}$.

Question. If I tell you I have a homomorphism $f : G \longrightarrow Q$, with $(f) = H$, how can you prove I'm lying?

Answer. Write out where each element maps to:

$$\begin{aligned} e &\longrightarrow e \\ (12) &\longrightarrow 1 \\ (23) &\longrightarrow a \\ (13) &\longrightarrow b \\ (123) &\longrightarrow c \\ (132) &\longrightarrow c^{-1} \end{aligned}$$

Note that

$$(12)(23) = (123)$$

so

$$f(12)f(23) = f(123).$$

Hence $1a = c$, that is $a = c$. Similarly, $(13)(12) = (123)$ implies $b = c$. Finally, $(23)(13) = (123)$ implies $ab = c$. This implies $a \cdot a = a$, which gives $a = 1$.

That means, H was not the kernel. That means, the kernel was the entire group.

Proposition. If $H < (F)$, then $gHg^{-1} < (f)$ also, for all $g \in G$.

Proof. Note that

$$\begin{aligned} f(ghg^{-1}) &= f(g)f(h)f(g)^{-1} \\ &= f(g)1f(g)^{-1} = 1. \end{aligned}$$

Therefore, this shows that if $K = (f)$, we must have $gKg^{-1} = K$ for all $g \in G$. □

Definition. We say that a subgroup $K < G$ is normal if $gKg^{-1} = K$ for all G. Notation: we write $K \triangleleft G$.

Note that the kernel of any homomorphism is always a normal subgroup.

This is completely unlike linear algebra. You need a normal subgroup to be the kernel of something.

For normal subgroups, left cosets equal right cosets, since $gKg^{-1} = K$ implies $gK = Kg$. This is not true otherwise.

9.4 Lecture 7

Recall the definition of a normal subgroup.

Definition. A subgroup $N < G$ is normal if $gN = Ng$ for all $g \in G$.

(Obviously, if G is abelian, then every subgroup of G is normal.)

Recall, that the coset gN is the equivalence class of g under the equivalence relation $G \sim h$ if $h = gn$ for some $n \in N$.

Last week, we saw that if you have some homomorphism $f : G \to H$, then (f) is always a normal subgroup of G.

Question. Give a normal subgroup $N \triangleleft G$, can we find a group Q and a homomorphism $f : G \to Q$, with $(f) = N$?

```
Take $G = \ZZ$, and let $N = 2 \ZZ$.  Does there exist some $f: \ZZ \to Q$ with $\Ker(f) =

 Let's first establish what this means:
 \[
   f(n) = 1 \text{ if $n$ even}
 \]
 \[
   f(n) \neq 1 \text{ if $n$ odd}
 \]

Now, from Friday, call $f(1) = q$.  Then we know that $f(n) = q^n$.  We must have
 \[
   f(-1) = q^{-1} = q \neq 1
 \]
 \[
   f(0) = q^{0} = 1
 \]
 \[
   f(1) = q \neq 1
 \]
 \[
   f(2) = q^{2} = 1
 \]
 \[
   f(3) = q^{3} = q \neq 1.
 \]
 Therefore, the only possible group has two elements, $1$ and $q$ (with $q^2 = 1$). \\
 Let $G = \ZZ$, and $N = 10 \ZZ$.  We would like some map $f: \ZZ \to Q$, with $\Ker(f) = 10

   We know that all of the multiples of $10$ must map to the identity in $Q$.  But we know more, we can state that
  \[
    f(m) = f(n) \Leftrightarrow 10 | (n-m).
  \]
 The $\Leftarrow$ direction is easy.  The $\Rightarrow$ direction is true because if not, the kernel would be bigger.

  Note that
  \[
    A = \left\{ \dots, -10, 0, 10, 20 \right\} \text{ map to 1 in $Q$}
  \]
  \[
     B = \left\{ \dots, -9, 1, 11, 22 \right\} \text{ map to other element in $Q$}
  \]
  \[
    \dots
  \]
  \[
    J = \left\{ \dots, -1, 9, 19, 29 \right\} \text{ map to a tenth element in $Q$}.
  \]

  Insight: what if we call $Q = \left\{ A, B, C, \dots, J \right\}$.  Define the group operation $B+C = D$, and we can take any element from these subsets, and get the ``answer'' $D$.

  So $Q = \ZZ / 10 \ZZ$, which is our quotient group.  In other words:
  \[
    10 \ZZ = \Ker (\ZZ \twoheadrightarrow \ZZ / 10 \ZZ).
  \]
  or:
  \[
    N = \Ker(G \twoheadrightarrow G / N).
  \]



 Aside on notation: \footnote{Note that $\twoheadrightarrow$ indicates a surjective map, and $\hookrightarrow$ indicates an injective map.}
```

We now formally define quotient groups.

Definition. Given a group G and a normal subgroup $N \triangleleft G$, the quotient group G/N is defined by:

- its elements are the cosets gN (note that left = right cosets for a normal subgroup). In other words, these are equivalence classes $\overline{g}$ under the notation $g \sim h$ iff $h = gn$.[1]
- Its group operation is $\overline{g} \cdot \overline{h} = \overline{gh}$. We need to check that this is well defined! This is critical, and this is where it matters that N is normal.

  Check that the quotient group is a group.

  - Identity: $\overline{1} \cdot \overline{h} = \overline{1h} = \overline{h} = \overline{h \cdot 1} = \overline{h} \cdot \overline{1}$
  - Associativity: $\overline{a} \cdot (\overline{b} \cdot \overline{c}) = \overline{a} \cdot \overline{bc} = \overline{a \cdot (b \cdot c)} = \overline{(a \cdot b) \cdot c} = \overline{ab} \cdot \overline{c} = (\overline{a} \cdot \overline{b}) \cdot \overline{c}$.

  Several comments: [2] [3]

  Note that there is a canonical surjective homomorphism

  $$\pi : G \twoheadrightarrow G/N$$

  that takes $g \mapsto \pi(g) = \overline{g}$.

  Remark. What is the size of the quotient group? Note that

  $$|G/N| = \text{ of cosets of } N \text{ in } G$$

  $$= \text{index}[G : N]$$

  If $|G|$ is finite, then

  $$|G/N| = |G|/|N|$$

  Now, we can still define the index of two groups that are infinite. Easy example:

  $$[\mathbb{Z} : 10\mathbb{Z}] = |\mathbb{Z}/10\mathbb{Z}| = 10.$$

  $$|\mathbb{Z}|/|10\mathbb{Z}| = \infty/\infty.$$

  Theorem. (First isomorphism theorem.) If $f : G \to H$ is any homomorphism, then

  $$G/(f) \cong \text{Im}(f).$$

  To name the map:

  $$\psi(\overline{g}) = f(g).$$

  This theorem seems really simple - but its subtle, because its not super clear if its obvious or if we need to prove it.

  Checking that $\overline{g} \cdot \overline{h} = \overline{gh}$ is well defined. Suppose that $\overline{g} = \overline{a}$ and $\overline{h} = \overline{b}$. Then

  - $\overline{g} \cdot \overline{h} = \overline{gh}$
  - $\overline{g} \cdot \overline{b} = \overline{gb}$

[1] On equivalence classes: $g \sim h \Leftrightarrow h \in \overline{g} \Leftrightarrow \overline{g} = \overline{h}$

[2] This is somehow similar to compiled languages. Check once that the quotient group is well defined, and know that can write "loose" notation that can't go wrong.

[3] (Note that 3.1 in the book is pretty confusing)

- $\overline{a} \cdot \overline{b} = \overline{ab}$

We need to check that

$$\overline{gh} = \overline{gb} = \overline{ab}.$$

Suppose $\overline{h} = \overline{b}$. Then there exists $m \in N$ such that $b = hm$. We have

$$(gb) = (gb)m,$$

so $\sim gh$, i.e. $\overline{gb} = \overline{gh}$.

Note that $\overline{a} = \overline{g}$, i.e. there exists $n \in N$ such that $a = gn$. Therefore -

$$ab = gnb = gbn'.$$

Therefore $ab \sim gb$ so $\overline{ab} = \overline{gb}$.

For the above, we have used normality, since we know that $nb \in Nb = bN$.

9.5 Lecture 8

9.5.1 More on conjugation

Comment from office hours. It was brought up that we don't really have that many examples of abelian groups. D_{2n} is generally non-abelian (D_2 is the only abelian case).

One of the main topics today will be conjugation. Earlier, we saw that gNg^{-1} is a notion of "translation" by N. More explicitly, let us analyze a vs gag^{-1}.

Where might you have seen an equation like $b = gag^{-1}$? One place is linear algebra — if A and B are matrices that represent the same linear transformation, but in different bases, then you get $B = CAC^{-1}$ where C is the change of basis matrix. In this setting A and B are "the same," but in different coordinate systems.

Suppose we have two sets $\{x, y, z, w\}$ and $\{1, 2, 3, 4\}$ with the bijection f where $x \mapsto 1$, $y \mapsto 2$, $z \mapsto 4$, $w \mapsto 3$. Suppose we had a permutation $\sigma \in \text{Perm}(\{x, y, z, w\})$ where

$$\sigma = (xyz)(w).$$

The claim: the earlier bijection lets us "turn" this into a bijection of the set $\{1, 2, 3, 4\}$. We can remap the permutations to obtain

$$1 \to x \to y \to 2$$
$$2 \to y \to z \to 4$$
$$3 \to w \to w \to 3$$
$$4 \to z \to x \to 1.$$

This composition can be expressed as $f\sigma f^{-1}$.

More commonly, suppose we have some $g \in S_n$, and some $\sigma \in S_n$. We can consider a conjugation $g\sigma g^{-1}$.

```
Let $\sigma = \text{(1 2 7)(5 8)(3 4)}$, and let $g(i) = i+10 \pmod{100}$.  Then
\[
   g \sigma g^{-1} = \text{(11 12 17)(15 18)(13 14)}.
\]
```

Definition. Let G be a group and let $a, b \in G$. We say that a, b are conjugates (in G) if there exists $g \in G$, such that

$$b = gag^{-1}$$

Notes:

- This is an equivalence relation.
- The equivalence classes are called conjugacy classes.
- Intuitively, the conjugacy classes group elements with the “same structure.”
- If G is abelian, then conjugacy classes are just $\{1\}$, $\{a\}, \{b\}, \ldots.$

Question. When are two permutations $\sigma, \tau \in S_n$ conjugates?

Any permutation with a cycle decomposition in a “3-2-2” pattern is conjugate to $\sigma = (1\ 2\ 7)(5\ 8)(3\ 4)$, for example $(14\ 3\ 2)(7\ 8)(10\ 11)$.

Answer. If their cycle decompositions have the same number of cycles of each length.

Proposition. Every $A \in GL_2\mathbb{C}$ is conjugate to some B of the form $B = \begin{pmatrix} x & y \\ 0 & z \end{pmatrix}$.

If you apply this B to $e_1 = [1\ 0]$, you get $Be_1 = xe_1$, i.e. e_1 is an eigenvector. So this holds because every A has an eigenvector.

Note. Fix some element $g \in G$. Consider the function $\alpha_g : G \longrightarrow G$ given by conjugation: $\alpha_g(h) = ghg^{-1}$. Is this a homomorphism?

Consider

$$\alpha_g(hk) = ghkg^{-1}$$
$$\alpha_g(h)\alpha_g(k) = ghg^{-1}gkg^{-1} = ghkg^{-1}.$$

So yes, it is a homomorphism. What is the kernel of this function? Just the identity.

If $f : G \longrightarrow H$ is a homomorphism, then the following are equivalent:

- f is injective
- $(f) = \{1\}$.

This implies that α_g is actually an isomorphism from G to itself.

Question. Do the size of conjugacy classes divide the order of the group? A: Yes, but not for the same reason as in Lagrange’s theorem.

9.5.2 Group actions

We start with the definition.

Definition. An action of a group G on a set X is a homomorphism $\alpha : G \to \mathrm{Perm}(X)$.

In other words, to each $g \in G$, we can associate a function $\alpha_g : X \to X$ such that

$$\alpha_g \circ \alpha_h = \alpha_{gh}.$$

This is almost the same is a group of functions, but the only difference is that nobody says that this homomorphism has to be injective.

Note: this is discussed in section 1.7 in the text, but it uses a different framework.

Concretely, suppose we have a group D_8, with a map to $\mathrm{Perm}(\mathbb{R}^2)$. Since for example "rotation by 90 degrees" can be viewed as a translation in the plane. So — this is an action of D_8 on $\mathbb{R}^2$.

The action is a realization of the group as functions on the plane. The important thing here is how the homomorphism is defined (not the given D_8 or $\mathbb{R}^2$).

But note that group actions don't always have natural geometric interpretations.

```
   Let $G = \ZZ / 7 \ZZ$, and let $X = \left\{ 1, 2, 3, 4 \right\}$. Claim: any action of $G$ on $X$ is trivial.
```

Notation in book. Suppose we have $x \in X$, and we can consider $\alpha_g : X \to X$. Then we can view $\alpha_g(x) \in X$. One thing the book points out is that you can write $\alpha_g(x)$ as $g \cdot x$. But the $\alpha_g(x)$ notation is arguably a bit clearer.

9.6 Lecture 9

We start be defining the notion of a product group. If A and B are groups, then we can define a group on

$$A \times B = \{(a, b) | a \in A, b \in B\}.$$

The operation is defined component wise:

$$(a, b) \cdot (\alpha, \beta) = (a\alpha, b\beta).$$

Comments on question 6. Can we find some group K such that $n(K, G) = |G|^2$. As many reasoned correctly, we must have K have two generators. The reason $\mathbb{Z}^2$ does not work is because although it is generated by two elements $(1, 0)$ and $(0, 1)$, this group is abelian. Now, the number of homomorphisms $n(\mathbb{Z}^2, G)$, is

$$n(\mathbb{Z}^2, G) = | \{(g, h) | g \in G, h \in G, gh = hg\} | \leq |G|^2.$$

(Mentioning $\mathbb{Z}^2$ and arguing why it is wrong is much better than turning in a "false" proof that $\mathbb{Z}^2$ works.)

Now, we want to build some $K = F_2$, defined as the "free group on two generators." (Free, here is a semi-technical term, see the idea of a "free module.") We want:

- F_2 is generated by two elements.
- a and b don't satisfy any relations except those that are forced by the group axioms.

Now, how do we turn this vague desire into an actual group? The challenge, as it turns out, is mainly in establishing associativity.

Let $F_2 = \langle a, b \rangle$; this is what we hope to be able to write.

- First try, let F_2 be the set of finite strings on the alphabet $\left\{a, b, \overline{a}, \overline{b}\right\}$, with the operation = concatenation. (Similar to Kleene closure.) Problem: there is no inverses, since $(ab)BA = abBA$.
- Second possibility: define an equivalence relation on $\{a, b, A, B\}^*$, where $waAu \sim wu$, $wAau \sim wu$, $wbBu \sim wu$, $wBbu \sim wu$, and if $w \sim w'$ and $u \sim u'$, then $wu \sim w'u'$.

  Possible problem. How to show that we haven't identified too many things together?
- Third possibility: define a string to be "reduced" if it has no aA, Aa, bB, Bb substrings. Define our F_2 to be the set of all reduced words $w \in \{a, b, A, B\}^*$. The operation here is
  - Concatenate, then
  - Delete aA, Aa, bB, Bb substrings until the string is reduced.

  Issue 1. Need to show this operation results in a unique reduced string.

  Issue 2. This assumes you have associativity.

  Question: can you just define the operation from left to right? Answer: yes, this gives you uniqueness, but you have to show associativity.

Comment on this — you have to do a decent amount of work to rigorously show that you have $aba^{-1} \neq baab$. The idea of "conservation of difficulty.'

This is covered in section 6.3. Note that all the future homeworks will be hard (won't depend on chapters 1-6).

Consider $\mathbb{Z} = \langle x \rangle$. We also saw $\mathbb{Z}_5 = \langle x | x^5 = 1 \rangle$. Now, we can write $F_2 = \langle r, s \rangle$.

Note that we can write $D_{10} = \langle r, s | r^5 = 1, s^2 = 1, srs^{-1} = r^{-1} \rangle$. It is easy to check that all of these equalities hold, but the important take-away here is that all elements in D_{10} is a consequence of these three elements. Note that this fact implies that

$$n(D_{10}, G) = \left\{(x, y) | x, y \in G, x^5 = 1, y^2 = 1, yxy^{-1} = x^{-1}\right\}$$

9.7 Lecture 11

Isomorphism theorems and quotient groups. Suppose you have a homomorphism

$$\alpha : G/N \to H.$$

Then you can get a homomorphism

$$a : G \to H$$

with $a(g) = \alpha(\overline{g})$.

Concretely, there is a bijection between homomorphisms $\alpha : G/N \to H$ and homomorphisms $a : T \to H$ with $N < \ker(a)$.

Suppose we have a group G, let $\overline{G} = G/N$. So we can define a projection π from $G \to G/N$. Suppose we have some subgroup $M < \overline{G}$. Claim: let $B = \pi^{-1}(M)$ be a subgroup of $\overline{G}$. Schematically, we can represent this as follows:

This implies that subgroups $\overline{B} \leq \overline{G}$ are in bijection with subgroups $N \leq B \leq G$. Additionally, we have $\overline{B} \cong B/N$.

If we have $N < B < C$ and $B \triangleleft C$, then we have that $\overline{C}/\overline{B} \cong C/B$.

This is a statement of the second / fourth isomorphism theorems $(-\epsilon)$.

Other way to write this isomorphism is to say $(C/N)/(B/N) \cong C/B$. One other thing we can check is that $\overline{A} \triangleleft \overline{G} \Leftrightarrow A \triangleleft G$.

9.7.1 Conjugation / conjugacy classes

Remember that we say that a is conjugate to b in G if there exists some g so that $b = gag^{-1}$.

Note that we can think of conjugation as a group action of the group G on the set $X = G$. Our definition here is

$$g * x := gxg^{-1}.$$

To check this is an action, we just need to check

- $g * (h * x) = (gh) * x$.
- $g * (hxh^{-1}) = (gh) * (gh)^{-1}$.

Key thing we gain from thinking of this as a group action is that the conjugacy class of x is an orbit of x.

Corollary. The size of the conjugacy class of X divides $|G|$. More explicitly, the size of this conjugacy class equals the size of $|G|/|\text{stabilizer of } x|$. Definition: the stabilizer in this setting is called the "centralizer" subgroup (notation is $C_G(x)$).

- The definition of stabilizer: $\{g \in G | g * x = x\}$.
- The definition of orbit: $\{g \cdot x | g \in G\}$.

We are often interested in the size of some orbit. But instead, we can compute the fixed points. Note that each element x will have different centralizers.

This ends up implying the class equation. Let G be a finite group. Then we can write:

- $|G| = \sum$ size of each conjugacy class
- If there are k conjugacy classes in G, pick representatives $1, g_2, g_3, \ldots, g_k$. Then we can write

$$|G| = \sum_{i=1}^{k} [G : C_G(g_i)].$$

- If there are r conjugacy classes of size > 1, say $g_1, \dots, g_r$, then

$$|G| = \underbrace{|Z(G)|}_{\text{conjugacy classes of size 1}} + \sum_{i=1}^{r} [G : C_g(g_i)].$$

9.8 Lecture 12: Automorphisms and Sylow's Theorems

5B. Show that the commutator subgroup is not finitely generated. Call L the commutator subgroup of F_2, so that

$$L = \left\{ a^{k_1} b^{l_1} \dots a^{k_n} b^{l_n} | \sum k_n = 0, \sum l_n = 0 \right\}.$$

The reason we call this L is to think about languages in computer sciences. Indeed, there is a notion of regular language (meaning it can be recognized by a finite state machine). This is a classic example of a language that is not regular.

Suppose L were finitely generated by a set, e.g. $\{aba^{-1}b^{-1}, aaba^{-1}\}$...

The idea is that you can build a finite state automaton. This is not a DFA (but you can convert a non-deterministic automaton to a deterministic automaton.

Pumping Lemma. Idea: if you can produce longer and longer words, then it can't be regular. Importantly, to work on higher level math, you need to be able to "chunk" simple systems and apply "sub"-theorems.

Definition. An automorphism of a group G is an isomorphism $f : G \to G$.

Intuitively, we can think of the analogy bijection : permutation :: isomorphism : automorphism.

Definition. (G) is the group of automorphisms of G under composition.

Proposition. Let G be a group. Then $G/Z(G) \cong$ a subgroup of (G).

This is a strange statement, but it tells you a lot about how to prove it. When you see G/N is isomorphic to a subgroup H, immediately, you should think — I should produce a homomorphism $f : G \to H$ with $\ker(f) = N$. Since the first isomorphism theorem says that

$$G/\ker(f) \cong (f) \leq H.$$

Back to the proposition — we are looking for some homomorphism α with

$$\alpha : G \to (G)$$

with kernel $Z(G)$. You can think of this homomorphism as a group action. Since its kernel is $Z(G)$, we can write

$$Z(G) = \left\{ zgz^{-1} = g \forall g \right\}.$$

So we can think of G acting on itself by conjugation, with $\alpha_g : G \to G$ with

$$\alpha_g(h) = ghg^{-1}$$

and

$$\ker(\alpha) = \{z \in G | \alpha_z = id\} = \left\{z | zhz^{-1} = h \forall h\right\} = Z(G).$$

Then, by the 1st isomorphism theorem

$$G/Z(G) = G/\ker(\alpha) \cong (\alpha) \leq (G).$$

Recall Euler's totient function, defined as

$$\phi(n) = \text{ of } 1 \leq k \leq n \text{ that are relatively prime to } n$$

Proposition. $(\mathrm{Z}_n) \cong (\mathbb{Z}/n\mathbb{Z})^\times$.

Recall $(\mathbb{Z}/n\mathbb{Z})^\times = \left\{\overline{k} \in \mathbb{Z}/n\mathbb{Z} | \overline{k} \text{ relatively prime to } n\right\}$, under multiplication.

For example,

$$(Z_8) \cong (\left\{\overline{1}, \overline{3}, \overline{5}, \overline{7}\right\}, \times).$$

And here, we have

$$(Z_8) = \left\{f_1 : x^k \mapsto x^k, f_2 : x^k \mapsto x^{3k}, f_3 : x^k \mapsto x^{5k}, f_4 : x^k \mapsto x^{7k}\right\}.$$

Theorem. (Non-obvious) Let p be an odd prime. Then (Z_{p^k}) is cyclic of order $\phi(p^k) = p^k - p^{k-1}$.

Theorem. (Non-obvious) For $p = 2$, note that (Z_{2^k}) is not cyclic, but its "almost cyclic":

$$(Z_{2^k}) \cong Z_2 \times Z_{2^{k-2}}.$$

We will now change gears and discuss Sylow's theorem. We may not get to motivate why this is important, but it is very powerful.

Before class, for $G = S_4$, Church worked out the number of subgroups of S_4 of size k.

- $k = 1$, number of subgroups $N = 1$.
- $k = 2$, $N = 9$.
- $k = 3$, $N = 4$.
- $k = 4$, $N = 4$
- $k = 6$, $N = 0$.
- $k = 8$, $N = 3$
- $k = 12$, $N = 1$
- $k = 24$, $N = 1$.

The Sylow theorem is concerned with $k = 3$ and $k = 8$, since $|S_4| = 24 = 8 \cdot 3 = 2^3 \cdot 3$. Also, note the definition:

Definition. A group P is called a p-group if $|P| = p^k$ for some $k \geq 0$.

Definition. Given a group G and a prime p, write $|G| = p^a \cdot m$ with $p \nmid m$. A subgroup $P \geq G$ is called a Sylow p-subgroup if $|P| = p^a$.

Definition. Let ${}_p(G)$ denote the set of Sylow p-subgroups of G. Let $n_p(G)$ denote the number of Sylow p-subgroups of G.

Theorem. (Sylow's Theorem). Fix G and P.

(1) G has at least one Sylow p-subgroup (i.e. $n_p(G) \geq 1$).

(2a) Any two Sylow p-subgroups are conjugate. If $P_1 \leq G, P_2 \leq G$, with $|P_1| = |P_2| = p^a$, then there exist g such that $P_2 = gP_1g^{-1}$. In particular, they are all isomorphic to the others!

(2b) Any p-subgroup is contained in some Sylow subgroup. If $Q \leq G$ and $|Q| = p^b$, there exists some $P \leq G$ with $|P| = p^a$ and $Q \leq P$.

(3) We know that $n_p(G) \equiv 1 \pmod{p}$ and $n_p(G)$ divides m.

For example, if $|G| = 11 \cdot 5^{100}$. Then the number $n_p(G) \equiv 1 \pmod{5}$ and divides 11. This tells us $n_5(G) = 1$ or 11.

If $|G| = 7 \cdot 5^{100}$. This tells us $n_5(G) \equiv 1 \pmod{5}$ and it divides 7, so $n_5(G) = 1$.

Corollary. G has a unique Sylow p-subgroup if and only if G has a normal Sylow p-subgroup if and only if $n_p(G) = 1$.

Why are these equivalent? If there is only one subgroup that that size, then imagine conjugating it. Clearly $|P| = p^k$, and $|gPg^{-1}| = p^k$, which implies that they are the same subgroup, and that it is normal. (The converse is straightforward too).

9.9 Lecture 14

Lemma A. For any G with $|G| = p^k m > 1$ ($p \nmid m$), either

- G has a subgroup $H \not\trianglelefteq G$ with $|H| = p^k l < p^k m - |G|$.
- G has a quotient $G \longrightarrow \overline{G}$ with $\overline{G} = p^b m < p^k m = |G|$.

For any $H \leq G$, we can consider action of H on the set of subsets of G by conjugation.

That is,

$$S \mapsto hSh^{-1} = \left\{hsh^{-1} | s \in S\right\}.$$

Write $\Theta_H(S)$ = orbit of S, $\Theta_H(S) = \left\{hSh^{-1} | h \in H\right\}$.

Proposition. (B.) Let R be a p subgroup of G. Then let Q be any p subgroup of G.

- $|\Theta_Q(R)| = 1$, if and only if $Q \subseteq R$.
- Otherwise, $|\Theta_Q(R)| \equiv 0 \pmod{p}$.

From this, we are going to prove Sylow's Theorem.

Proof. (Proof of Sylow (i) using Lemma A.)

By induction on $|G|$. Base case $|G| = 1$. Inductive step, consider Lemma A.

- If there exists $H \leq G$, with $|H| = p^k l < p^k m = |G|$. By induction, H has a p subgroup with $|P| = p^k$, so there exists a p-Sylow subgroup.
- If $\pi : G \to G$ with $|\overline{G}| = p^b m < p^k m = |G|$. Let $N = \ker(\pi)$, and set $\overline{G} \cong G/N$. Then $|\overline{G}| = |G|/|N|$, and we can write $p^k m = p^k m/|N|$.

  By induction, $\overline{G}$ has a p-Sylow subgroup $\overline{P}$. Set $P = \pi^{-1}(\overline{P})$. Then

$$|P| = |\overline{P}||N| = p^b \cdot p^{k-b} = p^k.$$

□

We now proceed to prove Sylow (3) using Prop B:

Proof. Let $P_1, P_2, \ldots, P_{n_p}$ be all the p-Sylow subgroups of G, and suppose P is a p-Sylow.

Consider the p-orbits on $P_1, \ldots, P_{n_p}$ acting by conjugation. Apply Proposition B with $Q = P$ and $R = P_i$. Now,

- $|\Theta_p(P_i)| = 1$ if and only if $P \subseteq P_i$ but $P = P_i$ b/c $|P| = |P_1| = p^k$.

Now, we just need to know that n_p divides $|G|$. Because then we have $n_p | p^k m$ and $p \not| n_p$, which implies that $n_p | m$. Then $n_p || G|$ follows from Sylow 2(a), since given 2(a), n_p = size of the orbit under conjugation by G. □

We now proceed to prove Sylow (2b):

Proof. Let Q be any p-subgroup of G. Suppose for a contradiction that Q is not contained in any p-Sylow subgroup.

Consider the Q-orbits on $P_1, \ldots, P_{n_p}$. By proposition B, this assumption implies that all of the orbits have size $\equiv 0 \pmod p$.

Examining the list, we can break this into

$$\{P_1, \ldots, P_k\}, \ldots, \{P_k, \ldots, P_{n_p}\},$$

which implies that $n_p = 0 + \cdots + 0 \pmod p$, which contradicts $n_p \equiv 1 \pmod p$. □

Note that we can also get a corollary ("2b + ϵ"). In fact, the number of p-subgroups contained in Q is $\equiv 1 \pmod p$.

We now will prove (2a) using Proposition B.

Proof. Let $c =$ of conjugates of P, and let $P_1, P_2, \dots, P_c$ be the G-conjugates of P, with $P = P_1$. Similarly, we can look at any group acting on the list $P_1, \dots, P_c$.

If we first consider P acting on this list by conjugation, then we get that it splits up into something like:

$$\{P_1\}, \{\dots\}, \dots, \{\dots, P_k\},$$

where $c = 1 + 0 + \cdots + 0 \pmod{p}$.

Now, asume for contradiction that L is a p-Sylow that is not conjugate to P. Then L is not in this list. Look at the L-orbits; the only possible size 1 orbits would be if e.g. $L = P_7$ (L has to be equal some element in this list). But we just assumed that L is not in this list, so there is no size 1 orbits, but this gives $c \equiv 0 \pmod{p}$, which is a contradiction. □

Aside: we can use this to show that every matrix mod p whose order is a power of p has an eigenvector with eigenvalue 1.

9.10 Lecture 15

Lemma C. If R is a p-Sylow subgroup of G, with $G = p^k m$, and $q \in G$ has $|q| = p^b$, then $qRq^{-1} = R$ iff $q \in R$ (conjugation is the trivial map).

Proposition B. Let R be a p-Sylow subgroup of G, and let Q be any p-subgroup of G. Then $||\Theta_Q(R)| = 1$ iff $Q \subseteq R$, otherwise $|\Theta_Q(R)| \equiv 0 \pmod{p}$, and $|\Theta_Q(R)| =$ number of Q conjugates of R.

Note that Lemma C implies Proposition B. (This is not that hard of a proof).

Lemma D. For any G, and any $H \leq G$, if $gHg^{-1} = H$, then $K = \left\{g^k h | k \in \mathbb{Z}, h \in H\right\}$ is a subgroup of G, and it's $\langle g, H\rangle$.

Proof that Lemma D implies Lemma C. (In the case where $|q| = p$.) Set $K = \left\{q^k r | k \in \mathbb{Z}, r \in R\right\}$. Lemma D tells us that this is a subgroup. We now consider its size. Now, since $|q| = r$, we can write

$$K = \left\{q^k r | k \in \{0, \dots, p-1\}, r \in R\right\}.$$

Suppose that $q \notin R$, for a contradiction. This implies that $q^k \notin R$ for $k \in \{1, \dots, p-1\}$. This implies that the cosets $R, qR, q^2R, \dots, q^{p-1}R$ are all disjoint (this statement requires some thought, think about it). This implies that the size of the set $K = p \cdot |R| = p^{k+1}$. No subgroup of b has size p^{k+1}, since it does not divide the order of the group.

Proof of Lemma D. We just need to check that this set is closed under multiplication / inverses. Choose two elements $a, b \in K$. We can write $a = q^k h_1$, $b = q^l h_2$. Then we can write $ab = q^k h_1 q^l h_2$. Now, $h_3 = q^{-l} h_1 q^l \in H$. Some substitution / algebra gives us $ab = q^{k+l} h_2 h_3 \in K$, so that K is closed under multiplication.

Broader point of Lemma D. For all $h_1 \in H, g$, there exists some h_3 such that $gh_1 = h_3 q$.

2nd Isomorphism Theorem Now, suppose $A \leq G, B \leq G$, and suppose $aBa^{-1} = B$ for all $a \in A$. Then the set

$$AB = \{ab | a \in A, b \in B\}$$

is a subgroup and it's $\langle A, B\rangle$. Note that the proof ends up being exactly the same as before.

And furthermore:

- $B \trianglelefteq AB$
- $A \cap B \trianglelefteq A$
- $AB/B \cong A/A \cap B$.

Note: $|AB| = \frac{|A||B|}{|A \cap B|}$.[4]

We can write down a helpful definition:

Definition. For any subset $H \leq G$, the normalizer $N_G(H)$ is $N_G(H) = \left\{g \in G | gHg^{-1} = H\right\}$.

Proposition. (5.4.9) Suppose you have two normal subgroups $N \triangleleft G$, $H \trianglelefteq G$, $N \cap H = 1$. Then $NH \cong N \times H$.

Corollary. If $|G| = 15$, then $G \cong Z_3 \times Z_5$.

Proof sketch. We know that there's a bijection between NH and $N \times H$, just by writing $\{nh\} \mapsto (n, h)$. Need to show: for all $n \in N$ and $h \in H$, we need to show that n and h commute (implies that this bijection is an isomorphism).

We know this because $hn = nh$ is the same as $n^{-1}h^{-1}nh = 1$. We can write

$$\underbrace{n^{-1}}_{\in N}\underbrace{h^{-1}nh}_{\in N} = \underbrace{n^{-1}h^{-1}n}_{\in H}\underbrace{h}_{h \in H} \in \{1\}.$$

Now, suppose $N \trianglelefteq G$, $H \leq G$, $N \cap H = \{1\}$. Then every $g \in G = NH$ uniuqely written as

$$g = nh$$

where

$$G = NH \text{ is a bijection with } N \times H$$

but it is not an isomorphism because $nhnh^{-1}$ can be viewed as $H \to (N)$.

This is the only infomration that is necessary to remember what G is.

Question 7 can be rephrased as: suppose you have some action from $H \to (N)$, you can define a group G whose elements are pairs (n, h) with

$$(n_1, h_1)(n_2, h_2) = (n_1(h_1 * n_2), h_1 h_2),$$

where we are thinking of $*$ as the action. This is the semi direct product of N and H.

[4] Note that this is true even without the assumption that $aBa^{-1} = B$ for all a, i.e. even when AB is not a subgroup, it still has this size.

9.11 Lecture 18

In this lecture, we'll discuss examples of rings to have in mind. Note that the operations can in principle be "strange" and not be usual addition / multiplication (see Question 0 on homework), but typically the operations will be canonical. "Main" examples of rings are like: $\mathbb{Q}, \mathbb{Z}, \mathbb{Z}/3\mathbb{Z}, \mathbb{Z}/10\mathbb{Z}$.

- Fields, $\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{Q}(\sqrt{2})$[5], $\mathbb{F}_p$
- Integers
- Modular stuff: $\mathbb{Z}/n\mathbb{Z}$ (won't write Z_n).
- Polynomials. We can write

$$\mathbb{Z}[x] = \left\{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 | a_i \in \mathbb{Z}, n \geq 0\right\}$$

More examples: we can write

$$\mathbb{Z}[\sqrt{2}] = \left\{a + b\sqrt{2} | a, b \in \mathbb{Z}\right\}$$
$$\mathbb{Z}[\sqrt[3]{2}] = \left\{a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 | a, b, c \in \mathbb{Z}\right\}.$$

Note that if A and B are rings, $A \times B$ is a ring with

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$$
$$(a_1, b_1) \times (a_2, b_2) = (a_1 a_2, b_1 b_2).$$

Consider the ring of "functions of something." For example,

$$C([0,1]) = \{\text{continuous functions } f : [0,1] \to R\}.$$

Can consider various extensions of this theme:

$$\begin{aligned} C([2,7]) &= \{\text{continuous functions } f : [2,7] \to R\} \\ \ldots &= \{\text{infinitely differentiable functions } f : [0,1] \to R\} \\ \ldots &= \{\text{continuous functions } f : [0,1] \to \mathbb{C}\} \\ \ldots &= \{\text{functions } f : [0,1] \to R\} \end{aligned}$$

It turns out that these examples are the "most canonical" in some sense (but this requires theory to explain). One important idea is to consider rings that are restrictions of larger rings. For example, let $f \in \mathbb{C}([0,10])$. We can consider the ring of functions restricted to $[4,6]$. It turns out that

$$r : C([0,10]) \to C([4,6])$$

is a ring homomorphism.

Grothendieck won a Fields medal for constructing a "something" so you can consider any ring as functions on "something."

[5] Subfield of $\mathbb{R}$ generated by $\mathbb{Q}$ and $\sqrt{2}$

Definition. We say $r \in R$ is a unit if it has a multiplicative inverse. We write

$$R^{\times} = \{\text{units } r \in R\}.$$

Definition. A ring R is a field if $0 \neq 1$ and $R^{\times} = R - \{0\}$, i.e. every nonzero element is a unit.

Recall key idea from linear algebra. Suppose you have an equation $ax + by = 0$, where $a \neq 0$. In a field, this would imply that $x + a^{-1}by = 0$. This is useful, but it isn't possible in general (even in the integers).

Now, over the integers, suppose we had the equation $2x + 3y = 0$. We could not write $x + \frac{3}{2}y = 0$. However, we can divide in certain cases; if $10x = 10y$, then $x = y$.

Definition. We say $r \in R$ is a zero-division if $r \neq 0$ and there exists $s \neq 0 \in R$ such that $r \cdot s = 0$.

```
  If $\ZZ / 10 \ZZ$, $r = 4$, $s = 5$, so that $rs = 20 = 0 \in \ZZ / 10 \ZZ$. \\
```

Definition. A commutative ring R is a domain if $0 \neq 1$ and it has no zero-divisors.

Proposition. If R is a domain, if $a \neq 0$, then $ax = ay$ implies $x = y$.

In the examples we described, all the fields are domains. Also, a given $r \in R$ can't be both a unit and a zero-divisor. To see this, suppose we had $rs = 0$, with $s \neq 0$, but also $r^{-1}r = 1$. Then left multiplying by r^{-1}, we get $r^{-1}rs = 0$, implying $s = 0$, which is a contradiction.

We now turn to the definition of a ring homomorphism.

Definition. If A, B are rings, a function $f : A \to B$ is a (ring) homomorphism if:

- $f(1) = 1$
- $f(a + b) = f(a) + f(b)$
- $f(ab) = f(a)f(b)$.

Definition. Let R be a ring, and suppose $A \subseteq R$ is a subset of R. We say A is a subring of R if

- $1 \in A$
- A is a subgroup under addition
- A is closed under multiplication

Caution: the book lies. (About point 1 of the previous two definitions). E.g. the book will say $2\mathbb{Z}$ is a subring, but it is not, because it doesn't contain 1.

9.12 Lecture 20

We start be discussing the isomorphism theorems for rings.

1. Recall the first isomorphism theorem for rings from last lecture that says

$$(f) \cong R/(f); \quad f : R \to S.$$

2. Suppose A is a subring of R, and I is an ideal of R. Then
$$A + I = \{a + i | a \in A, i \in I\} \text{ is a subring}$$
$$A \cap I \text{ is an ideal of } A$$
$$\frac{A + I}{I} \cong \frac{A}{A \cap I}.$$
This basically parallels the second isomorphism theorem for groups, except that they call it N instead of I. Don't worry too much about the raw intuition of this one, focus on seeing lots of examples.[6]

3. (3rd + 4th) Fix $I \subseteq R$ an ideal. We saw last time that we have this map $R \longrightarrow R/I = \overline{R}$. We claim that there is a correspondence between ideals $J \subseteq R$ containing I and ideals $\overline{J} \subseteq \overline{R}$. In particular, we can write
$$R/J \cong \overline{R}/\overline{J} = (R/I)/(J/I).$$
It's also true that subrings $S \subseteq R$ containing I are in bijection with subrings $\overline{S} \subseteq \overline{R}$, where $\overline{S} = S/I$.

Here's another (easier) way to keep track of what this is saying. Suppose you want to define some homomorphism $\overline{f} : R/I \longrightarrow C$. What would be great is if there was some function $f : R \longrightarrow C$, so we can just write $\overline{f}(\overline{r}) = f(r)$. The question is: when is this actually well defined? We need that $\overline{f}$ needs to be equal for all representatives mod I. In particular, we need $0 = \overline{f}(0) = \overline{f}(\overline{i}) = f(i)$. The theorem is saying in particular that this is all you need! In particular:

$$\left\{\text{homomorphisms } \overline{f} : R/I \longrightarrow C\right\} \Leftrightarrow \{\text{homomorphisms } f : R \longrightarrow C, f(I) = 0\}$$

The rest of today will be spent on definitions, which will help to make a lot of this concrete. [7]

Let's now talk about generators.

Definition. Suppose you have a subset $X \subseteq R$. We write (X) to denote the ideal of R generated by X. If X is finite, with $X = \{x_1, \ldots, x_k\}$, write
$$(X) = (x_1, \ldots, x_k).$$

There are two definitions here that we can state:

- (X) is the smallest ideal containing X, namely
$$X = \bigcap_{\text{ideal } I, X \subseteq I} I$$

[6] When do we see the second isomorphism theorem? We used the group analog a lot with identities like $|HN| = \frac{|H||N|}{|H \cap N|}$. Say we were thinking about vector spaces. We would say something like $\frac{V+W}{W} \cong \frac{V}{V \cap W}$, where V, W are subspaces of X. Suppose we had some $f : X \to Y$ with $\ker(f) = W$, then both sides are isomorphic to $f(V)$. In particular, we can obtain $f(V + W) = f(V)$. Check out https://math.stackexchange.com/questions/1738334/intuition-about-the-second-isomorphism-theorem

[7] "Comment that is maybe too enlightened": even if we hadn't defined this previous bijection, you could take this bijection as the definition of R/I; and the answer is that you don't need to know exactly which set it is, just need to know where things map.

- (X) is the set of all linear combinations of arbitrary length:

$$(X) = \{r_1 x_1 + \cdots + r_n x_n | n \in \mathbb{N}, r_i \in R, x_i \in X\}.$$

Definition. Consider the subring S of R generated of X. We can write

- S is the smallest subring of R containing X:

$$S = \bigcap_{A, X \leq A} A.$$

- You can also write

$$S = \left\{ \sum_{i=1}^{n} \prod_{j=0}^{m_i} x_{ij} \mid x_{ij} \in X, m_i \geq 0 \right\}.$$

In particular, we can take the empty product to that 1 is in the subring S..

Further, note that if I is generated by some subset X, so that $I = (X)$, we can define an equality

$$\{\text{homomorphisms } f : R \to C, f(X) = 0\}$$
$$= \left\{\text{homomorphisms } \overline{f} : R/I \to C\right\} \Leftrightarrow \{\text{homomorphisms } f : R \to C, f(I) = 0\}$$

Example. Consider the following ring. Let $R = \mathbb{Q}[x, y]$, that is linear combinations of $x^i y^j$. We say that I is a principal ideal if it is generated by one element. In particular, let $I = (x^2 + y^2 - 1)$, and $A = R/I$. Question: how many homomorphisms $\phi : A \to \mathbb{Q}$ are there? Note that $\mathbb{Q} \subset A$ are the constrant polynomials, so that $\phi(x) = x$ for any $x \in \mathbb{Q}$ (since you have to take $1 \to 1$).

This is a great advertisements for the bijections mentioned previously! It is really hard to write up an explicit homomorphism from first principles. But it turns out that the answer is

$$\left\{ \text{ of pairs of numbers } a, b \in \mathbb{Q} \text{ with } a^2 + b^2 = 1 \right\}.$$

This hints at why rings are useful. It means that we can encode solutions to polynomial equations in terms of homomorphisms from sonme ring. Just like group theory in some sense is based on understanding symmetric groups, ring theory is based on understanding solutions to equations.

Question: can you apply this argument to encode solutions to equations over other fields? Yes. One caveat is that if you are working over F_n for some composite n, you have to specify the constant mapping explicitly, i.e.

$$\phi : \mathbb{F}_n[x, y]/(I) \to \mathbb{F}_n; \qquad \phi(c) = c; \forall c \in \mathbb{F}_n.$$

This starts to hint at why algebraic geometry, number theory, and ring theory are fundamentally intertwined. Note that there's a fantastic theorem proved by Hasse.

Hasse Local-Global Primes. Let's say you have a function $f(x, y, z, w) =$ quadratic in the inputs. Hasse says that $f(x, y, z, w) = 0$ has a solution in the integers if and only if $f(x, y, z, w) = 0$ has a solution in $\mathbb{Z}/p\mathbb{Z}$, $\mathbb{Z}/p^2\mathbb{Z}\dots$ for all primes p, and has a solution in the reals. Check out Keith Conrad's article on this[8].

We continue to some basic definitions.

Definition. An ideal $P \subsetneq R$ is a prime ideal if $a \cdot b \in P$ implies $a \in P$ or $b \in P$.[9]

Definition. Let $M \subseteq R$ with $M \neq R$ is a maximal ideal if it's maximal. That is, there doesn't exist I with $M \subsetneq I \subsetneq I$. Note that maximal ideals are prime.

[8] http://www.math.uconn.edu/ kconrad/blurbs/gradnumthy/localglobal.pdf

[9] Note that this is a property of prime numbers. $p = (5)$ is prime, since if $ab \equiv 0 \pmod 5$ then $a \equiv 0 \pmod 5$ or $b \equiv 0 \pmod 5$. Note that this isn't the same as not having factors other than 1 and p.

9.13 Lecture 23

Fix a field F and let $R = F[x]$. Claim: R is a PID. Given ideal $I \subseteq R$, let $m_I \in I$ be the monic polynomial in I of smallest degree. Then I is generated by m_I, i.e. $I = (m_I)$.

9.14 Lecture 29

Today, we'll discuss: what is $i \in \mathbb{Z}/5^\infty\mathbb{Z}$? Recall that:

$$(\mathbb{Z}/5^k\mathbb{Z})^\times \equiv \mathbb{Z}_{5^k - 5^{k-1}} \equiv \mathbb{Z}_{4 \cdot 5^{k-1}},$$

so there exists four solutions to $x^4 = 1$ in $\mathbb{Z}/5^k\mathbb{Z}$, that is, $1, -1, a \equiv 2 \pmod 5, a \equiv 3 \pmod 5$.

We want: an algorithm / procedure to compute a. Question: how would Newton compute $\sqrt{2} \in R$? One application of his calculus is an algorithm to compute roots of polynomials. Suppose we are trying to find the roots of $f(x) = 0$.

- Start with an initial guess, e.g. $a_1 = 10$.
- Take a linear approximation to the function $f(x)$. This is a line through $(a_1, f(a_1))$, with slope $f'(a_1)$. Set the next guess to the be x-intercept, $a_2 = a_1 - \frac{f(a_1)}{f'(a_1)}$.
- Repeat the update rule $a_k = a_{k-1} - \frac{f(a_{k-1})}{f'(a_{k-1})}$ until convergence.

Coming back to the infinite integers, we try to find solutions of $x^2 = -1$.

- Initial guess $a_1 = 2$.
- $a_2 = a_1 - \frac{f(a_1)}{f'(a_1)} = 2 - \frac{5}{4} = \ldots 1112$.

Note, we can write

$$-\frac{1}{4} = \ldots 1111$$
$$-\frac{5}{4} = \ldots 1110$$
$$2 + (-\frac{5}{4}) = \ldots 1112$$

Similarly, we can write

$$\frac{1}{7} = \ldots 1033$$
$$\frac{25}{7} = \ldots 103300$$
$$-\frac{25}{7} = \ldots 341200$$
$$7 - \frac{25}{7} = \ldots 341212.$$

To show that this converges, we just need to argue that this will give us one more digit each time.

Suppose $f(x) \in (\mathbb{Z}/5^\infty\mathbb{Z})[x]$. Take the initial guess a_1 such that $f(a_1) \equiv 0 \pmod 5$ and $f'(a_1) \not\equiv 0 \pmod 5$. Claim: if you define $a_{k+1} = a_k - \frac{f(a_k)}{f'(a_k)}$, then

$f(a_k) \equiv 0 \pmod{5^k}$ (i.e. the last k digits are 0).

Note: this implies a_{k+1} and a_k have the same last k digits. In the context of convergence, this is like saying that the difference between two successive terms is getting smaller and smaller.

The convergence test for a series is just: do the terms go to 0? But this isn't true in calculus, since the harmonic series diverges.

Let's do what Newton would do and plug in $f(a_{k+1})$. We are hoping that $f(a_{k+1}) \equiv 0 \pmod{5^{k+1}}$. We can write

$$f(a_{k+1}) = f(a_k - \frac{f(a_k)}{f'(a_k)}) = f(a_k + h) = f(a_k) + hf'(a_k) + O(h^2).$$

We know here that $h \equiv 0 \pmod{5^k}$, so $h^2 \equiv 0 \pmod{5^{2k}}$, and certainly $h^2 \equiv 0 \pmod{5^{2k+1}}$.

Therefore,

$$f(a_{k+1}) \equiv f(a_k + h) \equiv f(a_k) + hf'(a_k) \pmod{5^{k+1}}$$
$$f(a_k) - \frac{f(a_k)}{f'(a_k)} f'(a_k) \equiv 0 \pmod{5^{k+1}}.$$

Note that this argument works for any prime p, not just 5.

And furthermore, this implies that a_k is a well defined element $a_\infty \in \mathbb{Z}/p^\infty\mathbb{Z}$ with $f(a_\infty) = 0$. This is called Hensel's Lemma.

Here's another formulation of Hensel's Lemma (which happens to be a bit stronger). Consider some $f(x) \in (\mathbb{Z}/5^\infty\mathbb{Z})[x]$. Something we could do is to drop everything after the last coefficient. There's a ring homomorphism from $(\mathbb{Z}/5^\infty\mathbb{Z})[x] \to (\mathbb{Z}/5\mathbb{Z})[x]$.

If there exists a_1 such that $\overline{f}(\overline{a_1}) = 0$ and $\overline{f}'(\overline{a_1}) \neq 0$, then we can write

$$\overline{f}(x) \text{ factors as } \overline{f}(x) = (x - \overline{a_1})\overline{g}(x)$$
$$\overline{g}(x) \text{ not divisible by}$$

x - $\overline{a_1}$.

Theorem (Strong Hensel's Lemma). Given a monic polynomial $f(x) \in \mathbb{Z}/5^\infty\mathbb{Z}[x]$, if $\overline{f}(x)$ factors into monic coprime $\overline{g_i}(x)$, $\overline{f}(x) = \overline{g}_1(x) \ldots \overline{g}_k(x)$, then there exists monic coprime $g_i(x) \in (\mathbb{Z}/5^\infty\mathbb{Z})[x]$ such that $f(x) = g_1(x) \ldots g_k(x)$.

9.15 Notes on Group Actions

Let G be a group acting on a nonempty set A. Recall that a group action must satisfy the following properties:

- $g_1 \cdot (g_2 \cdot a)$ for all $g_1, g_2 \in G$, $a \in A$ and
- $1 \cdot a = a$ for all $a \in A$.

Note that for each $g \in G$, the map $\sigma_g : A \to A$ defined by $a \mapsto g \cdot a$ is a permutation of A. To see this, note that σ_g has a two sided inverse (follows from the first condition above). Note also that there is a homomorphism associated to an action of G on A:

$$\varphi : G \to S_A; \qquad \text{defined by } \varphi(g) = \sigma_g,$$

called the permutation representation associated to the given action. We note some basic definitions:

1. The kernel of an action is $\{g \in G | g \cdot a = a\}$.
2. The stabilizer on a in G is $\{g \in G | g \cdot a = a\}$, denoted by G_a.
3. An action is faithful if its kernel is the identity.

The kernel of an action is a normal subgroup of G. An action of G on A may also be viewed as a faithful action of the quotient group $G/\ker\varphi$ on A.

9.16 Notes on Irreducibility

Eisenstein's. Let p be a prime in and let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$. Suppose $p | a_i$ for $i \in \{0, 1, \ldots, n-1\}$ but $p^2 \not| a_0$. Then $f(x)$ is irreducible in both $[x]$ and $[x]$.

Generalized Eisenstein's. Let P be a prime ideal of the integral domain R and let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial in $R[x]$. Suppose $a_{n-1}, \ldots, a_1, a_0$ are elements of P and suppose a_0 is not an element of P^2. Then $f(x)$ is irreducible in $R[x]$.

(p 312). $x^{n-1} + \cdots + x + 1$ is irreducible if and only if n is prime.

(p 315). $x^n - p$ is irreducible over $\mathbb{Z}[i]$.

9.17 Notes on Free Groups

9.18 Key Ideas

9.18.1 Definitions

Definition. A binary operation $*$ on a set G is a function $* : G \times G \to G$.

Definition. A group is an ordered pair $(G, \star)$ where G is a set and $\star$ is a binary operation on G satisfying the following axioms:

1. $\star$ is associative
2. There exists an element e in G, such that $a \star e = e \star a = a$ for all $a \in G$.
3. For all $a \in G$, there exists an element $a^{-1} \in G$ such that $a \star a^{-1} = a^{-1} \star a = e$.

Definition. A group $(G, \star)$ is called abelian if $a \star b = b \star a$ for all $a, b \in G$.

Definition. Let F be a field. Then $GL_n(F)$ is

$$GL_n(F) = \{A | A \text{ is an } n \times n \text{ matrix with entries from } F \text{ and } \det(A) \neq 0\}.$$

Definition. Let $(G, *)$ and $(H, \circ)$ be groups. A map $\psi : G \longrightarrow H$ such that

$$\psi(x * y) = \psi(x) \circ \psi(y) \text{ for all } x, y \in G$$

is called a homomorphism.

Definition. The map $\psi : G \longrightarrow H$ is called an isomorphism if

1. ψ is a homomorphism
2. ψ is a bijection

Definition. A group H is cyclic if H can be generated by a single element.

Definition. A function $f : A \longrightarrow B$ is injective if $f(x) = f(y)$ implies $x = y$. f is surjective if for all $b \in B$, there exists some $a \in A$ with $f(a) = b$.

Definition. A subgroup N is called normal if it is invariant under conjugation. In other words:

- For all g, $gH = Hg$.
- For all g, $gNg^{-1} = N$.
- There is some homomorphism on G for which N is the kernel. Intuition: consider the map $\pi(g) = gN$ for all ga This homomorphism is called the "natural projection" of G onto G/N.

9.18.2 Propositions and Theorems

Proposition. A subset H of a group G is a subgroup if and only if:

1. $H \neq \emptyset$ and
2. for all $x, y \in H$, we have $xy^{-1} \in H$.

9.18.3 Examples

```
The number of homomorphisms from $\ZZ_m \to \ZZ_n$ is $\gcd(m, n)$.
```

9.18.4 Ideas

1. To show a mapping is a homomorphism, first show that the mapping is well-defined ($b_1 = b_2$ implies $f(b_1) = f(b_2)$). Then, show that f is a homomorphism, that is $f(g_1 g_2) = f(g_1)f(g_2)$.
2. Studying quotient groups of G is equivalent to the study of the homomorphisms of G.

9.19 Things to review

1. Proof of Sylow's theorem.
2. More intuition for conjugation.
3. Book sections.

10

MATH116: Complex Analysis

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine

Theorem Definition Remark Claim Example Proposition Solution

latexsym bbm [small,bf]caption2 graphics epsfig amsopn url

[parfill]parskip [margin=1in]geometry

sign res Aut GL Ker im Syl

[parfill]parskips [margin=1in]geometry

MATH 116 - Complex Analysis Instructor: Yakov Eliashberg; Notes: Adithya Ganesh

Contents

10.1 9-24-18: Introduction

We can build up complex numbers with a few basic axioms.

1. $(1, 0)$ - unit.
2. $(0, 1)^2 = -(1, 0)$.
3. Bi-linear in z_1, z_2 (i.e. linear with respect to each argument).

Suppose $z = x + iy$. We define the conjugation operator as $\overline{z} = x - iy$, such that

$$z\overline{z} = x^2 + y^2 = |z|^2.$$

We can also express z in polar coordinates, so that

$$z = x + iy = r(\cos\phi + i\sin\phi).$$

We can extend the Taylor series of the exponential function on the real line to the complex plane by defining:

$$e^z = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^n}{n!} + \ldots.$$

It is easy to check that this definition satisfies the usual properties:

$$e^{z_1+z_2} = e^{z_1} \cdot e^{z_2}; \qquad e^{x+iy} = e^x e^{iy}.$$

We can similarly define

$$\cos z = 1 - \frac{z^2}{2} + \frac{z^4}{4!} + \ldots.$$

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} + \ldots.$$

We can combine these formulae to obtain $e^{iy} = \cos y + i\sin y$ (Euler).

Combining this with the previous definition, we can write

$$re^{i\phi} = r(\cos\phi + i\sin\phi).$$

Now, if $z = re^{i\phi}$, we can write $z^{-1} = \frac{1}{r}e^{-i\phi}$. This gives you a very natural geometric interpretation of inversion (conjugation + scaling).

Note that it is straightforward to derive trigonometric identities from Euler's formula; for example it is easy to see that

$$(\cos\phi + i\sin\phi)^n = \cos n\phi + i\sin n\phi.$$

Linear functions. Suppose we have a linear map $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$. We can write this as

$$F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

And furthermore, the following axioms must be satisfied:

- $F(z_1 + z_2) = F(z_1) + F(z_2)$.
- $F(\lambda z) = \lambda F(z)$.

One question: we could have either $\lambda \in \mathbb{R}$ (termed a real linear map) or $\lambda \in \mathbb{C}$ (termed a complex valued linear map).

If F is a complex linear map, we must have $F(iz) = iF(z)$ (i.e. the matrix has to commute). Furthermore, we must have $F(z) = F(z \cdot 1) = zF(1) = c$, where $c = a + ib$. So

$$F(z) = (a + ib)(x + iy) = (ax - by) + (ay + bx)i.$$

Also,

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax - by \\ ay + bx \end{pmatrix}.$$

It follows that F is complex if and only if $a = d$ and $b = -c$.

If $z = x + iy$, we can write $x = \frac{1}{2}(z + \overline{z})$ and $y = -\frac{i}{2}(z - \overline{z})$. Now, set $A = a + ic$, $B = b + id$, so we can write.

$$\frac{1}{2}(A - iB)z\frac{1}{2}(A + iB)\overline{z} = \alpha z + \beta\overline{z}.$$

Importantly, αz is complex linear while $\beta\overline{z}$ is complex antilinear (which means $F(\lambda z) = \overline{\lambda}F(z)$.

This proves that any real linear map can be written as a sum of a complex linear map and a complex antilinear map.

10.2 Differential 1-forms

Here, $\mathbb{R}^2_z$ denotes the space $\mathbb{R}^2$ with the origin shifted to the point z. A differential 1-form is a function of arguments of 2 kinds: of a point $z \in U$ and a vector $h \in \mathbb{R}^2_z$. It depends linearly on h and arbitrarily (but usually continuously and even differentiably) on z.

We will need only 1-forms on domains in $\mathbb{R}^2$. A differential 1-form λ on a domain $U \subset \mathbb{R}^2$ is a field of linear functions $\lambda_z = \mathbb{R}^2_z \to \mathbb{R}$. Thus a 1-form is a function of arguments of 2 kinds: of a point $z \in U$ and a vector $h \in \mathbb{R}^2_z$.

Given a real valued function $f : U \to \mathbb{R}^2$ on U, its differential df is an example of a differential form: $d_z(f)(h) = \frac{\partial f}{\partial x}h_1 + \frac{\partial f}{\partial y}h_2$. In particular, differentials dx and dy of the coordinate functions x, y are differential 1-forms. Any other differential form can be written as a linear combination of dx and dy:

$$\lambda = Pdx + Qdy,$$

where $P, Q : U \to \mathbb{R}$ are functions on the domain U.

A differential 1-form λ is exact if $\lambda = df$. The function f is called the primitive of the 1-form λ. The necessary condition for exactness if that λ is closed which by definition means $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$.

10.3 Complex projective line, or Riemann sphere

Consider the space $\mathbb{C}^n$. Similar to the real case, one can projectivise $\mathbb{C}^n$. $\mathbb{C}P^n$ is defined as the space of all complex lines through the origin. For us, the one-dimensional complex projective space is most relevant, the complex projective line ($\mathbb{C}P^1$).

Any vector $z = (z_1, z_2) \in \mathbb{C}^2$ generates the 1-dimensional complex subspace (complex line) denoted as

$$l_z = (z) = \{\lambda z : \lambda \in \mathbb{C}\}.$$

This line l_z can be viewed as a point of $\mathbb{C}P^1$. Any proportional vector $\overline{z} = \mu z$ generates the same line. Fix an affine line $L_1 = \{z_2 = 1\} \subset \mathbb{C}^2$. Any line from $\mathbb{C}P^1$ except $\{z_2 = 0\}$

10.4 Riemann surfaces

A Riemann surface is a 1-dimensional complex manifold. A set S is called a Riemann surface if there exist subsets $U_\lambda \subset X, \lambda \in \Delta$, where Δ is a finite of countable set of indices, and for every $\lambda \in \Delta$ a map $\Phi_\lambda : U_\lambda \to$ such that

- $S = \bigcup_{\lambda \in \Delta}$
- The image $G_\lambda = \Phi_\lambda(U_\lambda)$ is an open set in $\mathbb{C}$.
- The map Φ_λ viewed as a map $U_\lambda \to G_\lambda$ is one to one.
- For any two sets $U_\lambda, U_\mu, \lambda, \mu \in \Delta$, the images $\Phi_\lambda(U_\lambda \cap U_\mu), \Psi_\mu(U_\lambda \cap U_\mu) \subset$ are open and the map

$$h_{\lambda,\mu} = \Phi_\mu \circ \Phi_\lambda^{-1} : \Phi_\lambda(U_\lambda \cap U_\mu) \to \Phi_\mu(U_\lambda \cap U_\mu) \subset^n$$

10.5 Key ideas

10.5.1 Basic facts

1. $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$.
2. $\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}$
3. $\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}$.

10.5.2 Main results

Cauchy's integral formulas.

Suppose f is holomorphic on an open set that contains the closure of a disc D. If C denotes the boundary circle of this disc with the positive orientation, then

$$f(z) = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{z - \zeta} \, d\zeta.$$

Cauchy's integral formulas for derivatives.

Let f be holomorphic on an open set Ω, then f has infinitely many complex derivatives in Ω. Moreover, if $C \subset \Omega$ is a circle whose interior is also contained in Ω, then

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z)^{n+1}} \, d\zeta.$$

Cauchy-Riemann. f is analytic iff $u_x = v_y$, $u_y = -v_x$.

C^1 class. Every holomorphic function is a domain U is of class C^1, i.e. its derivative continuously depends on the point of U.

Cauchy's theorem.

Liouville's theorem. If f is entire and bounded, then f is constant.

Singularities and poles. A point singularity (or isolated singularities) of f is a $z_0 \in \mathbb{C}$ such that f is defined in a neighborhood of z_0 but not at the point z_0 itself. A zero for the holomorphic function f is z_0 such that $f(z_0) = 0$. By analtic continuation, the zeros of a non-trivial holomorphic function are isolated. A function F defined in a deleted neighborhood of z_0 has a pole at z_0 if the function $\frac{1}{f}$, defined to be to be zero at z_0, is holomorphic in a full neighborhood of z_0.

Pole power series representation. If f has a pole of order n at z_0, then

$$f(z) = \frac{a_{-n}}{(z - z_0)^n} + \frac{a_{-n+1}}{(z - z_0)^{n-1}} + \cdots + \frac{a_{-1}}{(z - z_0)} + G(z),$$

where G is a holomorphic function in a neighborhood of z_0.

Residue at a pole. The residue of f at that pole is defined as the coefficient a_{-1}, so that ${}_{z_0}f = a_{-1}$. In particular, if f has a pole of order n at z_0, then

$${}_{z_0}f = \lim_{z \to z_0} \frac{1}{(n-1)!} \left(\frac{d}{dz} \right)^{n-1} (z - z_0)^n f(z).$$

Residue formula, and corollary. Suppose that f is holomorphic in an open set containing a circle C and its interior, except for a pole at z_0 inside C. Then

$$\int_C f(z)\,dz = 2\pi i_{z_0} f.$$

Suppose f is holomorphic on an open set containing a circle C and its interior, except for poles at the points $z_1, \cdots, z_N$ inside C. Then

$$\int_C f(z)\,dz = 2\pi i \sum_{k=1}^{N} {}_{z_k} f.$$

Conformal map. A bijective holomorphic function $f : U \to V$ is called a conformal map or biholomorphism.

Riemann mapping theorem. Suppose Ω is proper and simply connected. If $z_0 \in \Omega$, then there exists a unique conformal map $F : \Omega \to \mathbb{D}$ such that

$$F(z_0) = 0; \qquad F'(z_0) > 0.$$

Corollary (3.2) Any two proper simply connected open subsets in $\mathbb{C}$ are conformally equivalent.

Mantel's theorem.

10.6 Midterm review sheet

10.6.1 Cauchy-Riemann equations

f is holomorphic iff $u_x = v_y; u_y = -v_x$.

Differential operators w.r.t. z and $\overline{z}$.

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} + \frac{1}{i}\frac{\partial}{\partial y}\right)$$
$$\frac{\partial}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} - \frac{1}{i}\frac{\partial}{\partial y}\right).$$

10.6.2 Cauchy integral formula + applications

Suppose f is holomorphic on an open set that contains the closure of a disc D. If C is the boundary circle, then for any $z \in D$:

$$f(z) = \frac{1}{2\pi i}\int_C \frac{f(\zeta)}{\zeta - z}\,d\zeta.$$

n-th derivative. If f is holomorphic in an open se TΩ, then f has infinitely many complex derivatives in Ω. If $C \subset \Omega$ is a cricle whose interior is only contained in Ω, then for all z in the interior of C:

$$f^{(n)}(z) = \frac{n!}{2\pi i}\int_C \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\,zeta.$$

Cauchy inequality + quick proof. If f is holomorphic in an open set that contains the closure of a disc D centered at z_0 and of radius R, then

$$|f^{(n)}(z_0)| \leq \frac{n!||f||_C}{R^n}.$$

Proof. By the Cauchy integral formula, we obtain

$$\begin{aligned} |f^{(n)}(z_0)| &= \left|\frac{n!}{2\pi i}\int_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}}\,d\zeta\right| \\ &= \frac{n!}{2\pi}\left|\int_0^{2\pi} \frac{f(z_0 + Re^{i\theta})}{(Re^{i\theta})^{n+1}} Rie^{i\theta} d\theta\right| \\ &\leq \frac{n!}{2\pi}\frac{||f||_C}{R^n} 2\pi. \end{aligned}$$

□

Liouville's theorem. If f is entire and bounded, then f is constant.

Proof. By Cauchy inequality, we obtain

$$|f'(z_0)| \leq \frac{B}{R},$$

where B is some bound for f. Taking $R \to \infty$, we obtain the desired result. □

Quick proof of FTA. Suppose P has no roots. Then $\frac{1}{P(z)}$ is bounded and entire. But then $\frac{1}{P(z)}$ is constant, which is a contradiction.

Schwarz reflection principle. Suppose that f is a holomorphic function in Ω^+ that extends continuously to I and such that f is real-valued on I. Then there exists a function F holomorphic in all of Ω such that $F = f$ on Ω^+.

Proof. For $z \in \Omega^-$, define $F(z)$ by

$$F(z) = \overline{f(\overline{z})},$$

look at power series expansions, and invoke the symmetry principle. □

10.6.3 Power series

Suppose f is holomorphic in an open set Ω. If D is a disc centered at z_0 and whose closure is contained in Ω, then f has a power series expansion at z_0 :

$$f(Z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n,$$

where $a_n = \frac{f^{(n)}(z_0)}{n!}$.

Analytic continuation. Suppose f and F are analytic in regions Ω, Ω' with $\Omega \subset \Omega'$. If the two functions agree on the smaller set Ω, then F is an analytic continuation of f into the region Ω', and is uniquely determined by f.

In particular, suppose f and g are holomorphic in a region Ω and $f(z) = g(z)$ for all z in some non-empty open subset of Ω. Then $f(z) = g(z)$ throughout Ω.

10.6.4 Exponential function and logarithm

Complex logarithm. Write

$$\log z = \log r + i\theta;$$

principal branch when $|\theta| < \pi$. Constructively, we can write

$$\log_{\Omega}(z) = F(z) = \int_{\gamma} f(w)\, dw,$$

where γ is any curve connecting 1 to z. Standard path of integration is to take $1 \longrightarrow r \in \mathbb{R}$ and then $r \longrightarrow z$, so that

$$\begin{aligned}\log z &= \int_1^r \frac{dx}{x} + \int_{\eta} \frac{dw}{w}\\ &= \log r + \int_0^{\theta} \frac{ire^{it}}{re^{it}}\, dt\\ &= \log r + i\theta.\end{aligned}$$

Note that in general

$$\log(z_1 z_2) \neq \log z_1 + \log z_2.$$

Taylor expansion for $\log(1+x)$. For the principal branch, we can write

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \dots.$$

10.6.5 Meromorphic functions

Definition of a meromorphic function. A function f on an open set Ω is meromorphic if there exists a sequence of points $z_0, z_1, \dots$ that has no limit points in Ω and such that

- f is holomorphic in $\Omega \setminus \{z_0, z_1, \dots\}$
- f has poles at the points $\{z_0, z_1, \dots\}$.

Casorati-Weierstrass. Suppose f is holomorphic in the punctured disc $D_r(z_0) \setminus \{z_0\}$ and has an essential singularity at z_0. Then the image of $D_r(z_0) - \{z_0\}$ under f is dense in the complex plane.

10.6.6 Argument principle and Rouche's theorem

Argument principle. Suppose f is meromorphic in an open set containing a circle C and its interior. If f has no poles and never vanishes on C, then

$$\frac{1}{2\pi i} \int_C \frac{f'(z)}{f(z)}\, dz = Z - P,$$

where Z is the number of zeros inside C, and P is the number of poles inside C.

Rouche's theorem. Suppose that f and g are holomorphic in an open set containing a circle C and its interior. If $|f(z)| < |g(z)|$ for all $z \in C$, then f and $f+g$ have the same number of zeros inside C.

Proof. Let $f_t(z) = f(z) + tg(z); t \in [0,1]$. Argue that

$$n_t = \frac{1}{2\pi i} \int_C \frac{f_t'(z)}{f_t(z)}\, dz$$

is constant; and in particular that $n_0 = n_1$. □

Open mapping theorem. If f and holomorphic and nonconstant in a region Ω, then f is open.

Maximum modulus principle. If f is a nonconstant holomorphic function in a region Ω, then f cannot attain a maximum in Ω.

Proof. Immediate from open mapping theorem. □

10.6.7 Computation of integrals using residues

Residue limit identity. If f has a pole of order n at z_0, then

$$_{z_0} f = \lim_{z \to z_0} \frac{1}{(n-1)!} \left(\frac{d}{dz}\right)^{n-1} (z - z_0)^n f(z).$$

Residue theorem. Suppose that f is holomorphic in an open set containing a toy contour γ and its interior, except for poles at the points z_i inside γ. Then

$$\int_{\gamma} f(z)\,dz = 2\pi i \sum_{k=1}^{N} {}_{z_k} f.$$

Integrals to know.

- $\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi.$
- $\int_{-\infty}^{\infty} \frac{e^{ax}}{1+e^x}\,dx = \frac{\pi}{\sin \pi a}; 0 < a < 1.$
- $\int_{-\infty}^{\infty} \frac{e^{-2\pi i x \xi}}{\cosh \pi x}\,dx = \frac{1}{\cosh \pi \xi}.$

10.6.8 Harmonic functions and harmonic conjugates

Definition. A real of complex valued C^2-smooth function f on a domain $U \subset \mathbb{C}$ is harmonic if

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0.$$

Unique determination. let $f, g : U \to \mathbb{R}$ be two harmonic functions which extend continuously to the boundary ∂U. Suppose that $f = g$ on ∂U. Then $f = g$ on U.

Proof. Suppose for $a \in U$ we have $f(a) > g(a)$; then consider $f - g$ and apply maximum modulus principle; contradiction. □

Log-composition. If f is a holomorphic function then $h(z) = \ln |f(z)|$ is harmonic.

Harmonic conjugate. The harmonic conjugate to a function $u(x, y)$ is a function $v(x, y)$ such that $u + iv$ is analytic.

Example. The harmonic conjugate of $u(x, y) = e^x \sin y$ is $-e^x \cos y + C$.

10.6.9 Elementary conformal mappings

Examples to know:

10.6.10 Properties of fractional linear transformations

Fractional linear transformations are mappings of the form

$$z \mapsto \frac{az + b}{cz + d}.$$

They always map circles and lines to circles and lines.

11

MATH171: Real Analysis

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine

Theorem Lemma Definition Remark Claim Example Proposition Solution

latexsym bbm [small,bf]caption2 graphics epsfig amsopn url

[parfill]parskip [margin=1in]geometry

sign Aut GL Ker im Syl

[parfill]parskips [margin=1in]geometry

MATH 171 - Real Analysis Instructor: George Schaeffer; Notes: Adithya Ganesh

Contents

11.1 9-24-18: Everything is a set

Administrivia:

- Book: Johnsonbaugh and Pfaffenberger
- (Supplement) Rudin's Principles of Mathematical Analysis
- Exam: likely week 5.

11.1.1 On sets

One motivation for analysis is a problem identified in 1901: Russell's Paradox. Consider

$$R = \{x : x \notin X\} = \text{ the set of all sets that do not contain themselves}$$

Problem: does the set contain itself? Either $R \in R$ or $R \notin R$, but neither is possible.

Rules for what is isn't a set: Zermelo-Frankel axioms.

In particular, under ZF: Can't build $\{x : x \text{ has property } P\}$. You must say

$$\{x \in S : x \text{ has property } P \text{ where } S \text{ is already a set. }\}$$

But going further: the collection of all sets is itself not a set.

Axioms of choice (the Cartesian product of a collection of non-empty sets is non-empty).

We can define the natural numbers in the framework of sets. If x is a set we can define its successor as $S(x) = x \cup \{x\}$.

- $0 = \emptyset$
- $1 = \{\emptyset\}$
- $2 = \{\{\emptyset\}, \emptyset\}$.
- $3 = \{\{\emptyset\}, \emptyset, \{\{\{\emptyset\}, \emptyset\}\}\,. = \{0, 1, 2\}$.

11.1.2 On functions (and cartesian products)

Cartesian Product. Let X and Y be sets. Then we can write

$$X \times Y = \{(x, y) : x \in X, y \in Y\} .$$

How do we define ordered pairs? $(x, y) \neq \{x, y\} = \{y, x\}$ doesn't work, since order matters.

Instead, we want to say

$$(x, y) = \{x, \{x, y\}\} .$$

What is a function? We can write $f : X \to Y$, where X is the domain(f) and Y is the codomain(f). A function $f : X \to Y$ is a subset of $X \times Y$ satisfying the following:

- $\forall x \in X, \exists y \in Y : (x, y) \in f$.
- $\forall x \in X, \forall y, y' \in Y : (x, y) \wedge (x, y') \in f \implies y = y'$.

As a set, for example, $\sin \subseteq \mathbb{R} \times \mathbb{R}$.

11.1.3 On natural numbers

The set $\mathbb{N}$ is equipped with a sucessor function $S : \mathbb{N} \to \mathbb{N} : x \mapsto S(x)$. There are a few rules attached to this, namely the Peano axioms:

- $\forall x \in \mathbb{N} : S(x) \neq 0$
- S is "injective": If $S(x) = S(y) \implies x = y$.
- Axiom of induction: If $K \subseteq \mathbb{N}$ satisfying
  - $0 \in K$
  - $\forall x \in K, S(x) \in K$.

$\mathbb{N}$ has two binary operations, +, $\cdot$, addition and multiplication.

A binary operation on X is a function $X \times X \to X$.

- + (a, b) = a+b

- $\cdot(a, b) = ab$

$\forall a, a + 0 = a. \forall a, b; a + S(b) = S(a + b)$.

11.2 10-1-18: Suprema and infima

Theorems of R.

- $\mathbb{R}$ is an ordered field.
- Tere are lots of ordered fields: $\mathbb{Q}$.

- Least upper bound axiom: If $S \subseteq \mathbb{R}$ is nonempty and bounded above, then S has a least bound $\in \mathbb{R}$.

Definition. Let $S \subseteq \mathbb{R}, M \in \mathbb{R}$. We say that M is an upper bound on S is $\forall x \in S : x \leq M$. M is the least upper bound (or the supremeum) of S is $\forall M' < M$, M' is not an upper bound on S.

Furthermore: $M = \sup(S)$ if

- M is an upper bound on S:
- $\forall M' < M : M'$ is not an upper bound on S.

$$\begin{gathered}\neg[\forall x \in S : x \leq M'] \\ \exists x \in S : \neg[x \leq M'] \\ \exists x : S : x > M' \\ \boxed{\forall \epsilon > 0, \exists x \in S : x > M - \epsilon}\end{gathered}$$

Easy two step process for proving $M = \sup(S)$.

The "greatest lower bound" axiom is equivalent to the "least upper bound" axiom. Note that for convenience $\sup(\text{unbounded above } S) = +\infty$ and $\sup(\emptyset) = -\infty$.

Consequences of the axioms in $\mathbb{R}$.

Archimedean Property. $\forall a, b \in \mathbb{R}, a, b > 0$, then $\exists n \in \mathbb{N}$ such that $na > b$.

Proof. Let $S = \{n \in \mathbb{N} : na \leq b\}$; which implies $n \leq \frac{b}{a}$. S is nonempty because $0 \in S$. S is bounded above by $\frac{b}{a}$. By LUBA: S has a supremum $m = \sup(S)$. $m + 1 \notin S$ and $m + 1 \in \mathbb{N}$ (left as an easy exercise).

Why is $m + 1 \notin S$? Otherwise $m + 1 \leq m$. So $\neg[(m + 1)a \leq b]$, i.e. $(m + 1)a > b$. □

Note that the Archimedean principle is true in $\mathbb{Q}$ as well. It inherits AP from R. There is also an independent proof just using the construction of $\mathbb{Q}$ as fractions.[1]

Theorem. The rational numbers form a dense subset of $\mathbb{R}$.

We start by explaining the definition: $\forall a, b \in \mathbb{R} : a < b \implies [\exists r \in \mathbb{Q} : a < r < b]$. We now mention a lemma that will help us prove the theorem.

Lemma. If $a < b \in \mathbb{R}$ and $b - a > 1$, then $\exists n \in \mathbb{Z} : a < n < b$.

Proof. Let $S = \{n \in \mathbb{Z} : n \leq a\}$.

By LUBA, we let $m = \sup(S)$. Note that $m \in \mathbb{Z}$. $m + 1 \notin S$. We can easily verify that $a < m + 1 < b$. The first inequality follows from $m + 1 \notin S$, and for the second, note that:

$$\begin{aligned} m + 1 &< m + (b - a) \\ &\leq a + (b - a) = b. \end{aligned}$$

[1] On HW2: An example of an ordered field, for which the AP fails.

Here, we have used the fact that $m \in S$, so $m \leq a$. □

Proof. (Main Theorem.) We know $a < b$. By the Archimedean Principle, since $b - a > 0$ and $1 > 0$, there must exist $n \in \mathbb{N}$ such that $n(b - a) > 1$. This implies that $nb - na > 1$. Also $na < nb$.

By the lemma, there exists an integer $k \in \mathbb{N}$ with $na < k < nb$. Dividing by n, we obtain the fraction $\frac{k}{n}$ which satisfies $a < \frac{k}{n} < b$. □

The irrationals are also dense in the reals - just take the rationals and add $\sqrt{2}$, and follow a similar argument.

11.3 Continuity - 10-19

If (X, τ) is a topological space and $S \subseteq X$, we can give S a topology

$$\tau_S = \{U \cap S : \text{ where } U \in \tau_X\}.$$

This is a subspace / induced / inherited / relative topology on S.

Note: $[0, 1]$ w/ the subspace topology from $\mathbb{R}$.

If (X, d_X) is ametric space, $S \subseteq X$, then S is also a metric space, where $d_S = d_X|_{S \times S}$ (restricted for $S \times S$).

If X is a metric space and $S \subseteq X$, then the topoloy from d is the subspace topology.

If U is open / closed in S, it need not be closed in X.

However, if K is compact in S, then K is compact in X.

Self explanatory (?) We say that a topological space X is compact if it is a compact set in its own topology.

Continuous functions. Let (X, d_X) and (Y, d_Y) be metric spaces, let $f : X \to Y$, let $p \in X$. We say that f is continuous at p

1. In the analytic sense if

$$\forall \epsilon > 0, \exists > 0, \forall x \in X : d_X(x, p) < \implies d_Y(f(x), f(p)) < \epsilon.$$

2. In the sequential sense if $\forall$ sequences $(x_n)_n \to p \in X$, the sequence

$$(f(x_n))_n \to f(p).$$

3. In the topological sense if $\forall$ open $V \subseteq Y$, if $f(p) \in V$, then there exists an open $U \subseteq X$ such that $p \in U$ and $f(U) \subseteq V$.

We will show that all of these are equivalent.

Notes on these definitions.

- For the topological sense: can switch "open" for "closed."

- Also, (iii) is the definition of "continuous' at p" for general topological spaces. It doesn't require a metric on X or Y.
- In (i), δ can depend on both p an ϵ.

Outline of proof of equivalence.

- $1 \implies 2$ (easy; definition pushing).
- $2 \implies 1$ is slightly harder. We will prove this by contrapositive. We'll show: if f is not analytically continuous at p, then it's not sequentially cts.

  $$\forall \varepsilon > 0, \forall \delta > 0; \exists x \in X : d_X(x, p) < p \text{ and } d_Y(f(x), f(p)) \geq \varepsilon.$$

  In particular, we can define $(x_n)_{n=1}^{\infty}$ so that

  $$d_X(x_n, p) < \frac{1}{n} \text{ and } d_Y(f(x_n), f(p)) \geq \varepsilon.$$

  This means that $x_n \to p$, but $f(x_n) \not\to f(p)$.
- $3 \implies 1$. Let $\varepsilon > 0$. We know f is topologically continuous, so let $V = B_\varepsilon(f(p))$ is open. Then there exists an open $U \ni p$ such that $f(U) \subseteq V$. Because U is open, p is interior to U, so $\exists \delta$ such that $B_\delta(p) \subseteq U$.

  So, $x \in B_\delta(p) \implies x \in U \implies f(x) \in V \implies f(x) \in B_\varepsilon(f(p))$.
- $1 \implies 3$ is similar to the previous, just reverse all the statements.

Definition. The function $f : X \to Y$ is continuous if it is continuous at all $p \in X$.

Translated definitions to "everywhere continuous."

- Analytic continuity is easy.
- Sequential continuity: $\forall$ convergent sequences $(x_n)_n$, $(f(x_n))_n$ is convergent and $\lim f(x_n) = f(\lim x_n)$.
- Topological continuity: f is continuous if for all open $V \subseteq Y$, $f^{-1}(V)$ is open in U.

For example, consider $f(x) = x^2$. Note that $f\left((-1, 1)\right)) = [0, 1)$. If f is continuous and U is open $f(U)$ may not be open.

Also, look at $g(x) = \frac{1}{x}$ on $(0, \infty)$. Consider $C = \mathbb{Z}_{>0}$, and then look at the set $g(C) = \left\{\frac{1}{n} : n \geq 1\right\}$ is not closed.

Continuous functions don't preserve openness and they don't preserve closedness, but they preserve compactness.

Theorem. If $f : X \to Y$ is continuous, (for (X, Y) topological spaces) and $K \subseteq X$ is compact in X, then $f(K)$ is compact in Y.

Proof. Need to show that every open cover of A has a finite subcover. Let H be an open cover of $f(K)$. Let $G = \left\{f^{-1}(V) : V \in H\right\}$. Now, G covers K. Since f is continuous, it is an open cover[2] □

Corollary. (Extreme value theorem). If $f : X \to \mathbb{R}$ (X is a compact topological space, f is continuous). Then f achieves a maximum and minimum on X. Meaning, there exists $p, q \in X$ such that $\forall x \in X, f(p) \leq f(x) \leq f(q)$.

Proof. $f(X)$ is a compact $\subseteq \mathbb{R}$, so $f(X)$ is closed and bounded (Heine-Borel). Provided $X \neq \emptyset, f(X) \neq \emptyset$. Now,

- $m = \inf(f(X)) \in f(X)$ and $M = \sup(f(X)) \in f(X)$
- By definition, since $m, M \in f(X)$, $\exists p, q \in X : f(p) = m$ and $f(q) = M$.

. □

The topological proof is surprisingly fast. Indeed, you can prove this using the sequential definition of continuity using the Bolzano-Weierstrass; but it is tricky.

Some remarmks on Heine-Borel:[3].

Next week: we'll discuss the notion of "connectedness."

Take home exam: goes out after class on Wed, have until Friday to finish.

11.4 10-22: Connectedness

Exam: released on Wednesday after class. You'll have 3 hours + extra time to submit. Can start at any time until midnight - ϵ. Open textbooks (J&P, Rudin), + notes.

Definition. Let X be a topological space. X is called disconneted if there are nonempty, disjoint open sets U and U' of X such that $X = U \cup U'$. If X is not disconnected, it's called connected.

Definition. If X is a topological space and $S \subseteq X$, then S is "connected" if S is connected as a topological space (with respect to the subspace topology).

[2] Recall that a topological space X is called compact if each of its open covers has a finite subcover. That is, X is compact if for every collection C of open subsets of X such that

$$X = \bigcup_{x \in C} x,$$

there is a finite subset F of C such that

$$X = \bigcup_{x \in F} x.$$

[3] $f(X)$ is compact $\subseteq \mathbb{R}$, so $f(X)$ is closed and bounded (true in any metric space). The other direction requires being a subset of $\mathbb{R}^n$, which requires Heine-Borel

Definition (Alternative). A topological space is connected if the only clopen sets are X and $\emptyset$.

We now ask: what are the connected subsets of $\mathbb{R}$?

Lemma. $S \subseteq \mathbb{R}$ is connected iff S is an interval. $\forall x, y \in S, \forall z \in \mathbb{R}, x < z < y \implies z \in S$.

Proof. We start with the forward direction. Assume S is not an interval. Then there exists some $z \in \mathbb{R}$ so that $x < z < y$ but $z \notin S$. Let $U = (-\infty, z)$ and let $U' = (z, \infty)$. Then it is easy to check that $(S \cap U) \cup (S \cap U')$ is a disconnection of S.

Need to check:

- $S \cap U$ is open in S, because U and U' are open in $\mathbb{R}$.
- The sets $S \cap U$, $S \cap U'$ are disjointed, because U, U' are disjoint.
- $S = (S \cap U) \cup (S \cap U')$.

If (X, τ) is a topological space and $S \subseteq X$, the subspace topology on S is

$$\tau_S = \{S \cap U : U \in \tau\}.$$

Also, τ_S is the coarsest topology such that $S \to X$ inclusion is continuous.

Now, we move on to the reverse direction. Let S be an interval, so that

$$S = (S \cap U) \cup (S \cap U'),$$

where U and U' are open in $\mathbb{R}$, $S \cap U$ and $S \cap U'$ are nonempty. Want to show that they are not disjoint. Let $V = S \cap U, V' = S \cap U'$.

Let $x \in V$ and $y \in V'$. Without loss of generality, assume $x < y$. Also, $\frac{x+y}{2} \in S$, since S is an interval.

We will construct sequences $(x_n)_n$ and $(y_n)_n$ as follows.

- $x_0 = x$ and $y_0 = y$.
- Let $\alpha_{n+1} = \frac{x_n + y_n}{2}$. If $\alpha_{n+1} \in V$, then $x_{n+1} = \alpha_{n+1}$ and $y_{n+1} = y_n$. Otherwise $\alpha_{n+1} \in V'$ and $x_{n+1} = x_n$ and $y_{n+1} = \alpha_{n+1}$.
- $(x_n)_n$ is an increasing sequence in V, and $(y_n)_n$ is a decreasing sequence in V. They are also bounded, since they are termwise bounded by each other.
- $(x_n)_n$ and $(y_n)_n$ both convergence, and since

$$|x_n - y_n| \leq 2^{-n}|x - y|,$$

both sequences convergence to the same limit L; $x < L < y \implies L \in S$. Therefore, $L \in V$ or $L \in V'$.

- Suppose $L \in V$. Then $L \in U$. L is an interior to U (since U is open). In particular, $\exists \epsilon > 0$ such that $B_\epsilon(L) \subseteq U$. By the convergence of $(y_n)_n \to L$, $\exists N$ such that

$$|y_n - L| < \epsilon; \forall n \geq N$$

In particular, $y_N \in B_\epsilon(L) \subseteq U \implies y_n \in V$. So since $y_N \in V'$, $V \cap V' \neq \emptyset$.

□

```
A disconnected set in $\RR$: $\mathbb{Q}$ is disconnected.
```

Consider a dramatic example: the Cantor set.

- In homework, proved that every open ball is closed in an ultrametric space.
- The Cantor set is "totally disconnected[4]" Turns out that every ultrametric space is topologically equivalent to the Cantor set.

Last time: Let $f : X \to Y$ be a continuous function of topological spaces. If X is compact, then $f(X)$ is compact. This implies the Extreme Value Theorem.

This time: Let $f : X \to Y$ be a continuous function of topological spaces. If X is connected, then $f(X)$ is connected.

Proof. Suppose $f(X)$ is disconnected. then

$$f(X) = (f(X) \cap V) \cup (f(X) \cap V').$$

Let $W = f(X) \cap V$, $W' = f(X) \cap V'$. Let $U = f^{-1}(W)$ and $U' = f^{-1}(W')$.

- $U \cap U' = X$ (by definition of preimage).
- U and U' are open, because $f^{-1}(W) = f^{-1}(f(X) \cap V) = f^{-1}(V)$. Further, $f^{-1}(V)$ is open because f is continuous and V is open.
- They're nonempty because W, W' are nonempty $\subseteq f(X)$.
- They're disjoint because if $x \in U \cap U'$, then $f(x) \in W \cap W'$; but W and W' are disjoint.

□

Corollary. (Intermediate value theorem.) Let X be a connected topological space, and let $f : X \to \mathbb{R}$ be a continuous real-valed function. If there are $p, q \in X$ and $c \in \mathbb{R}$ such that $f(p) < c < f(q)$, $\exists \xi \in X$ such that $f(\xi) = c$.

Proof. $f(X)$ is an interval. □

```
[Incomplete topologist's sine curve]
 Consider the graph of $\sin \left( \frac{1}{x} \right)$ for $x \in \left( 0, 1 \right)$.
tinuous on this interval.  This is connected.
```

[4] a lot of open sets are closed

```
[Midcomplete TSC]
 Consider $\left\{ \text{Incomplete TSC} \right\} \cup \left\{ (0, 0) \right\}$.  This wi
nected, still.  But, it is not path connected.  Consider a point $(x, y)$; there is no path
tween $(x, y)$ and $(0, 0)$.

   Interestingly, this is a converse to the Intermediate Value Theorem.  Has the intermediate value property, but it is not continuous at $0$.
```

11.5 10-24: Uniform continuity

Exam will be ready at 12:30pm. Have 3 hours + 30 extra minutes to scan + upload.

Today: just one proof, and then Q&A time / review.

Theorem. Let $f : X \to Y$ be a continuous function with X and Y metric spaces. If X is compact then f is uniformly continuous.

1. Recall that $f : X \to Y$ is continuous if $\forall p \in X, [\forall \varepsilon > 0, \exists \delta > 0, \forall x \in X$

$$d_X(x,p) < \delta \implies d_Y(f(x), f(p)) < \varepsilon.$$

2. $\forall p \in X, \forall (x_n)_n \to p$ if $f(x_n)_n$ converges $\to f(p)$.
3. $\forall$ open $U \subseteq V f^{-1}(V)$ is open in X. "preimage of an open set is open."

Importantly, 1, 2, 3, work in metric spaces, and 3 works in topological space.

Continuity.

In general, when we talk about continuity, we are discussing conditions of the form $\forall p \in X, \forall \varepsilon > 0, \exists > 0$ such that $\forall x \in X[\dots]$.

We say that $f : X \to Y$ (metric spaces) is uniformly continuous if

$$\forall \varepsilon > 0, \exists > 0, \forall x, p \in X : d_X(x,p) < \delta \implies d_Y(f(x), f(p)) < \epsilon.$$

The salient difference here is that δ only depends on ε, and no longer depends on p.

Obvious implication. If f is uniformly continuous, it is continuous. The converse, however is false.

Example. Let $f(x) = \frac{1}{x}$ on $(0, +\infty)$. This is a continuous function (on the interval). It is not uniformly continuous.

Consider some point $(p, f(p))$. Suppose we have a range $(f(p) - \epsilon, f(p) + \epsilon)$. Need a delta such that whenever $x \in (p - \delta, p + \delta)$, $f(x) \in (f(p) - \epsilon, f(p) + \epsilon)$. As $p \to 0$, δ stops working (since it will contain the asymptote at 0).

Example. Let $f(x) = \sin\left(\frac{1}{x}\right)$ on $(0, +\infty)$. This function is continuous, but not uniformly so. No matter how small we make δ, there is some point p close to 0 so that $f((p - \delta, p + \delta)) = [-1, 1]$.

Aside: the difference between Lipschitz continuity and uniform continuity. [5]

[5] You can show that $\sqrt{x}$ is uniformly continuous, but not Lipschitz continuous.

Proof of theorem. Suppose that $f : X \to Y$ is continuous but not uniformly continuous (where X is compact). Then $\exists \varepsilon > 0, \forall > 0, \exists x, p \in X$ such that $d_X(x,p) <$ and $d_Y(f(x), f(p)) \geq \epsilon$. We want to use the above to build two sequences $(x_n)_n$ and $(p_n)_n$ such that $\forall n \geq 1$,

$$d_X(x_n, p_n) < \frac{1}{n}; \qquad d_Y(f(x_n), f(p_n)) \geq \varepsilon$$

We have not used compactness yet. By sequential compactness: some subsequence of (x_n) converges. Some subsequence of $(x_n)_n$ converges to $L \in X$, so that $(x_{n_i})_i \to L$. Now, consider (p_{n_i}); we also have $(p_{n_i})_i \to L$.

Therefore

$$\lim_{i\to\infty} f(x_{n_i}) = f(\lim(x_{n_i})_i) = f(L) = f(\lim_{i\to\infty}(p_{n_i})_i) = \lim_{i\to\infty} f(p_{n_i}).$$

(for the first equality we have used continuity). But this is a contradiction, since our sequences are actually far apart.

Recall the theorem that states that if $f : [a,b] \to \mathbb{R}$ is continuous, then it is integrable. The proof of this theorem relies on the notion of uniform continuity.

11.5.1 Review

We now discuss the sequential version of uniform continuity.

Theorem. Let $f : X \to Y$ be continuous metric spaces. Then the following are equivalent:

- Let f is uniformly continuous.
- If $(x_n)_n$ is a Cauchy sequence, then $f(x_n)_n$ is also Cauchy.

Theorem. If X is a metric space and $K \subseteq X$ is compact and $(x_n)_n$ is a sequence in K, it has a convergent subsequence whose limit is in K.

Dense sets. Suppose X is a metric space $S \subseteq X$, S is dense if

- $\overline{S} = X$
- Every nonempty open U overlaps with S.
- For all $x \in X$ and $\forall \epsilon > 0$, $\exists y \in S$ such that $d_X(x,y) < \epsilon$.

For example, $\mathbb{Q}$ is dense in $\mathbb{R}$ (that is, all real numbers can be arbitrarily well approximated by rational numbers).

If X is a metric space with a countable dense set, then we call X separable. This is good because this means that computers can deal with such sets quite well (e.g. floating point).

Theorem. In the space of continuous functions $[0,1] \to \mathbb{R}$, the rational coefficient polynomials are dense.

11.6 11-05-18: Integrability and FTCs

Recall the Riemann-Darboux integral. Suppose $f : [a,b] \to \mathbb{R}$ is bounded with $a < b$. Then we can write

$$\int_{\underline{a}}^{b} = \sup\left\{\int_a^b \varphi : \varphi \text{ a step fn }; \varphi \leq f\right\}$$

or

$$\int_a^{\overline{b}} = \inf\left\{\int_a^b \psi : \psi \text{ a step fn and } \psi \geq f \text{ on } [a,b]\right\}.$$

And if they match, f is integrable and $\int_f = \int_{\underline{a}}^b f = \int_a^{\overline{b}} f$.

Proposition (Sequential criterion for integrability). Suppose $f : [a,b] \to \mathbb{R}$ is bounded, $L \in \mathbb{R}$.

Subtext: Wts

$$\int_a^b = L.$$

Then $\int_a^b f = L$ if and only if $\exists(\psi_n)_n, (_n)_n \in \text{Step}([a,b])$ for all k, $\psi_k \leq f \leq \varphi_k$; and

$$\int_a^b \psi_n \to L; \qquad \int_a^b \varphi_n \to L.$$

Sufficient conditions for integrability.

- If $f : [a,b] \to \mathbb{R}$ is continuous, it is integrable.
- Piecewise continuous.
- Monotonic functions.
- "Piecewise monotone and/or continuous..."
- Thomae's function shows that the converse is not true.

Proposition (Cauchy criterion for integrability.). If $f : [a,b] \to \mathbb{R}$ is bounded, it is integrable iff $\forall\varepsilon > 0$, there exists $\psi, \varphi \in \text{Step}([a,b])$ such that $\varphi \leq f\psi$ and $\int_a^b(\varphi - \psi) = \int_a^b \varphi - \int_a^b \psi < \varepsilon$.

(Proof follows from the definition of integrability.)

Theorem. If $f : [a,b] \to \mathbb{R}$ is continuous, then it is integrable.

Proof. Strategy: for any interval I we want $\varphi(I) - \psi(I)$ to be small.

Let $\varepsilon > 0$. Pick $\delta > 0$ such that $\forall x, y \in [a,b]$, we have

$$|x - y| < \delta \implies |f(x) - f(y)| < \frac{\varepsilon}{b-a}.$$

Pick a partition of $[a,b]$ into disjoint intervals $\{I_k\}_{k=1}^n$ with $|I_k| < \delta$. For each I_k, $f : \overline{I_k} \to \mathbb{R}$ achieves its min and max at $p_k, q_k \in \overline{I}_k$ respectively with

$$f(p_k) \leq f(x) \leq f(q_k)$$

for all $x \in I_k$.

Now, let

$$\varphi = \sum_{k=1}^{n} f(p_k) 1_{I_k}; \qquad \psi = \sum_{k=1}^{n} f(q_k) 1_{I_k}.$$

Fill in details from notes.

□

Theorem (Lebesgue's Riemann integrability condition). A function f is integrable if and only if

$$\lambda\left(\{p \in [a,b] : f \text{ is discontinuous at p }\}\right) = 0,$$

that is the set of discontinuities has Lebesgue measure 0 (alternatively, f is almost everywhere continuous).

[6] This theorem implies:

- f is continuous implies it is integrable.
- Monotone functions are continuous.

Let $f : [a,b] \to \mathbb{R}$ be bounded, we call $F : [a,b] \to \mathbb{R}$ a [continuous] antiderivative of f if it is continuous on $[a,b]$, differentiable on (a,b), and $\forall p \in (a,b) : F'(p) = f(p)$.

Theorem (The fundamental theorem of calculus). Let $f : [a,b] \to \mathbb{R}$ be an integrable function $F : [a,b] \to \mathbb{R} : x \mapsto \int_a^x f$. Then:

1. F is Lipschitz continuous.
2. If f is continuous at $p \in (a,b)$, then F is differentiable at p and $F'(p) = f(p)$.
3. If f is continuous on all of $[a,b]$, then F is an antiderivative of f.

Proof. The first statement is proven by the corresponding theorem for the upper / lower integrals. Statement (3) follows from statement (2).

Hence, it suffices to prove (2).

If f is continuous at p, then

$$\lim_{x \to p^+} \frac{F(x) - F(p)}{x - p} = f(p).$$

In these notes we will prove half of it (since the case $x \to p^-$ is similar.) Make the change of variables $x = p + h$. Then the limit above is equivalent to

$$\lim_{h \to 0^+} \frac{F(p+h) - F(p)}{h} = \lim_{h \to 0^+} \left[\frac{1}{h} \int_p^{p+h} f \right],$$

[6] Cannot use this on homework unless you prove it.

where

$$F(x) = \int_a^x f.$$

Let $\delta > 0$ such that $\forall x \in [a, b] : |x - p| < \delta \implies \left|f(x) - f(p)\right| < \varepsilon$. If $|x - p| < \delta$ then

$$f(p) - \varepsilon < f(x) < f(p) + \varepsilon.$$

So, for $h < \delta$, we have

$$\begin{aligned}\int_p^{p+h} [f(p) - \varepsilon] &\leq \int_p^{p+h} f(x) \\ &\leq \int_p^{p+h} [f(p) + \varepsilon] .\end{aligned}$$

This implies that for $h < \delta$,

$$h(f(p) - \varepsilon) \leq \int_p^{p+h} f \leq h(f(p) + \varepsilon);$$

that is

$$f(p) - \varepsilon \leq \frac{1}{h} \int_p^{p+h} f \leq f(p) + \varepsilon.$$

The punch line is that

$$\left|\lim_{h \to 0^+} \left[\frac{1}{h} \int_p^{p+h} f\right] - f(p)\right| < \varepsilon.$$

In particular, the limit above is equal to $f(p)$. This proves the FTC for the derivative of the integral; Wednesday we will do the integral of the derivative. □

11.7 Sequences and series of functions

11.7.1 Pointwise vs. uniform convergence

11.8 Key ideas

Definition of a metric space.

A metric space is a set X together with a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ that satisfies

- $d(x, x) = 0$.
- $d(x, y) = d(y, x)$.
- $d(x, y) + d(y, z) = d(x, z)$.

These three results together imply $d(x, y) \geq 0$.

Theorem: Cauchy-Schwarz.

Let $a_1, \cdots, a_n, b_1, \ldots, b_n \in \mathbb{R}$. Then

$$\left[\sum_{k=1}^{n} a_k b_k\right]^2 \leq \left[\sum_{k=1}^{n} a_k^2\right]\left[\sum_{k=1}^{n} b_k^2\right].$$

To recall inequality, just recall that $||u||||v|| \cos\theta = u \cdot v$.

Convergence in a metric space.

A sequence $(x_n)_n$ is convergent to L if $\forall \varepsilon > 0, \exists N$ such that $n > N \implies |x_n - L| < \varepsilon$.

Cauchy-ness in a metric space.

$\forall \varepsilon > 0, \exists N$ s.t. $\forall m, n \geq N, d(x_n, x_m) < \varepsilon$.

Complete metric space, and an example.

A metric space is called complete if Cauchy $\implies$ convergent. $\mathbb{R}$ is a complete metric space.

Convergence in $\mathbb{R}^n$.

$\overline{a}^k \to \overline{a}$ iff $\overline{a}_j^k \to a_j$ for all j.

Similar proof for Cauchy in $\mathbb{R}^n$.

Topological space.

A topological space is a (X, τ) where X is a set and $\tau \subseteq P(x)$ satisfies

- $\emptyset, X \in \tau$
- If $\mathcal{F} \subseteq \tau$, then $\bigcup_{S \in \mathcal{F}} S \in \tau$ (an arbitrary union of sets in τ is in τ).
- If $U_1, U_2 \in \tau$, then $U_1 \cap U_2 \in \tau$ (the intersection of any sets in τ are in τ).

Intuition `https://math.stackexchange.com/a/523794`. Broadly: a topology defines a notion of nearness on a set.

Interior / adherent sets.

Let (X, d) be a metric space, $x \in X, S \in X$. Then X is interior to S if $\exists \varepsilon > 0, B_\varepsilon \subseteq S$.

X is adherent to S if $\forall \varepsilon > 0, B_\varepsilon(x) \cap S \neq \emptyset$.

Limit point / isolated point.

x is a limit point of S if $\forall \epsilon > 0$, $\exists y \neq x$ such that $y \in B_\varepsilon(x) \cap S$.

x is an isolated point of S if $\exists \varepsilon > 0$, s.t. $B_\varepsilon(x) \cap S = \{x\}$.

Open / closed sets.

S is open if every $x \in S$ is interior to S.

S is closed if it contains all its adherent (or limit) points.

Perfect / bounded / dense sets.

S is perfect if it closed and contains no isolated points.

S is bounded if $\exists p \in X$ and $M \geq 0$ so that $\forall x \in S, d(x,p) \leq M$.

S is dense if $\forall x \in X$, x is adherent to S.

Cover of a set.

A cover of a set is a set of subsets whose union equals the original set. If $C = \{U_\alpha ; \alpha \in A\}$ is an indexed family of sets U_α then C is a cover of X if

$$X \subseteq \bigcup_{\alpha in A} U_\alpha$$

Compactness.

A subset $K \subseteq X$ of a topological space is called compact if $\forall \mathcal{G} \subseteq T$, with

$$K \subseteq \bigcup \mathcal{G},$$

$\exists$ a finite $\mathcal{G}'$ such that $K \subseteq \bigcup \mathcal{G}'$.

That is, each cover of K has a finite subcover.

Prove that compact sets are closed.

Bolzano-Weierstrass Theorem.

Heine-Borel Theorem.

Inverse powers.

If $\alpha > 0$ and $p \geq 1$ is an integer, there exists a unique $\beta > 0$ such tha $\beta^p = \alpha$.

Proof. Uniqueness is easy once existence is shown. Assume $0 < \alpha < 1$, (since $\alpha = 1$ is easy, and for $\alpha > 1$, just take $\left(\frac{1}{\beta}\right)^n = \frac{1}{\alpha}$.

check this detail

Consider two sequences $(a_n)_n$ and $(s_n)_n$ with

$$a_n = \max\left\{k \in \mathbb{Z} : \left(\frac{k}{2^n}\right)^p \leq \alpha\right\}.$$

and $s_n = \frac{a_n}{2^n}$. We define these because $(s_n)_n$ consists of better binary approximations to $\alpha^{1/p}$. Need to check that a_n is well defined, but that is not too difficult.

Claim: $(s_n)_n$ is increasing and bounded above. It is bounded above by 1, since $s_n > 1$ would imply $s_n^p = (a_n/2^n)^p > 1 > \alpha$, which contradicts the definition of a_n. To see that is (s_n) is increasing is straightforward. Thus, $(s_n)_n$ converges to some limit β.

Finally, we will show that $(s_n^p)_n \to \alpha$. This will show that β satisfies $\beta^p = \alpha$. Note that

$$\left(\frac{a_n}{2^n}\right)^p \leq \alpha < \left(\frac{a_n+1}{2^n}\right)^p$$

Thus, it suffices to check that the difference between the left and right hand sides approaches zero as $n \to \infty$. Note that

$$(a_n+1)^p - a_n^p = \sum_{k=0}^{p-1} a_n^k,$$

so since $a_n \leq 2^n$, we have $a_n^k \leq 2^{nk} \leq 2^{n(p-1)}$, when $k = 0, \ldots, p-!$.

In particular,

$$\left(\frac{a_n+1}{2^n}\right)^p - \left(\frac{a_n}{2^n}\right)^p \leq \frac{p \cdot 2^{n(p-1)}}{2^{np}} = \frac{p}{2^n}.$$

Thus, the left hand quantity $\to 0$ as $n \to \infty$. □

12

POLISCI101Z: Introduction to International Relations

adi

POLISCI 101Z: Introduction to International Relations Adithya Ganesh

Contents

12.1 Lecture 1: 6-24-19

This lecture will focus on the main themes typically discussed in international relations.

Example headline dealing with Iran deal - Iran would agree to not work on nuclear weapons, and US would remove economic sanctions. Trump decided to withdraw from the deal.

Another example headline - "EU Decarbonization Plan for 2050 Collapses After Polish Veto." Recall the Paris Climate Accord, in which ...

Headline 3 - "WTO warns of rising trade barriers ahead of G20 summit."

Headline 4 - "UN Report: Record number of South Sudanese face critical lack of food." Some 60% of the Sudanese population are at risk for not having enough food (it's about 7 million people).

As we study these topics, we will think like scientists and ethicists.

Key idea: the problem of international anarchy. There is no common power that performs the standard functions of a domestic government.

Example functions that a world government performs:

- Preserve peace.
- Protect environment.
- Regulate economy.
- Redistribute income.

This course examines the causes of (and solutions to) four international problems.

- War.
- Environment.
- Trade.
- Poverty.

12.1.1 Unit 1 - War

Some recent wars:

- War that arose due to US invasion of Iraq in 2003.
- India-Pakistan war in 1999. Over the Cargill sector in Kashmir.
- Civil war in Libya, which led to the overthrow of the Libyan regime.
- Russian tanks rolling into the invasion of Ukraine.

Major wars of the twentieth century:

- Assassination of archduke Franz Ferdinand by a Serbian nationalist; many viewed as the trigger for World War 1.
- WW2: 15 million soldiers died.
- Korean War: 5 million people died.

If you ask during what periods major powers were fighting against each other, this is a rarity (despite the significance of the previous inter-state conflicts).

Key questions:

- Why does war occur?
- Are some countries war-prone?
- How can leaders prevent war?

Ethical question:

- Under what condition might war be justified?

12.1.2 Unit 2 - Environment

CO2 production is growing, which is pretty worrisome. But some problems appear to be getting better - CFC production has fallen a lot since 1987, when the Montreal Protocol was inroduced.

Key question:

- When do countries enter (and honor) environment agreements?
- What ethical issues should enter the debate?

12.1.3 Unit 3 - International Trade

Many people believe in the advantages of free trade. But not everyone agrees (some people think socialism is better).

There is a global trend toward freer trade. Average traffics across time have declined since 1985.

Note - agriculture is a major point of contention in trade debates - since countries don't want their farmers to have to deal with foreign competition.

12.1.4 Unit 4 - Poverty and Capital Flows

Globally, there are 746 million people in extreme poverty (in 2013).

Some key questions:

- Why do countries give aid, and to whom?
- Is the IMF necessary, and is it driven by poltical concerns?
- What obligation do rich nations have to poorer ones?

Reading load is about 100 pages / week. Lectures reinforce and supplement the readings. Attend lectures (M, T, W only), review slides (posted after lectures).

Section timeline: register on Canvas. Sections begin this week.

12.2 Lecture 2: 6-25-19

Recall that international anarchy means that there isn't a centralized government to manage the world; this creates issues that we may take for granted in a domestic context.

12.2.1 Democracy and War

Is there a connection between democracy and war? If so, what is the definition.

Quotes from world leaders:

- "Democracies don't attack each other…Ultimately the best strategy to insure our security and to build a durable peace is to support the advance of democracy everywhere." (1994 State of the Union Address, Bill Clinton).
- "The reason why I'm so strong on democracy is democracies don't go to war with each other. And the reason why is the people of most oscieties don't like war, and they understand what war means." (George Bush, Nov. 2004).
- "…America has never fought a war against a democracy, and our cloests friends are governments that protect the rights of their citizens." (2009 Nobel Acceptance Speech, Barack Obama).

Plans for next 3 lectures:

- June 25: Theories about democracy and war
- June 26: How to evaluate evidence
- July 1: Evidence about democracy and war

Casual hypotheses have three parts:

- Dependent variable
- Independent variable
- Connecting logic

Clinton’s dependent variable is peace, and the independent variable is whether the country is a democracy.

Hypothesis should be

- General (eliminate the proper nouns)
- Falsifiable (could be proven wrong)

We will consider two types of mechanics that could connect democracy to peace.

- Structural
- Normative

Structural mechanisms:

- Democracy -> constraints on the executive
- Constraints reduce the propensity for war

Democratic structures could contribute to peace by:

- Empowering voters
- Delaying mobilization
- Conveying information

Empower voters arguments:

- Leaders want to remain in office
- Voters will remove belligerent leaders
- Thus, democratic leaders have electoral incentives to be peaceful.
- Autocratic leaders don’t have the same constraints

This idea is due to Kant, Perceptual Peace, 1795.

Can critique this argument. 1. isn’t necessarily true. 2. some voters want war (e.g. Nazi Germany, also Philippines). 4. Some autocrats may be overthrown if they are sufficiently poorly viewed by the people.

Some case studies to look into:

- Putin’s Russia, invading Ukraine
- Vietnam war, when Nixon got elected
- Eisenhower’s election during the Cold War.

There are also electoral autocracies, e.g. Mexico in the 80s.

Note the “Delay Mobilization” mechanism.

- Democracies have checks and balances (sometimes called veto players).

- These checks delay decisions to use force.
- Delay reduces the propensity for war ...
  - Affording time to negotiate
  - Reducing the risk of surprise

Two critiques: democracy may not lead to delay; and delay may not reduce the risk of war.

“Convey Information” mechanism.

- Democracy increases transparency about intentions and capabilities.
- Democracy can increase credibility.
- By raising transparency and credibility, democracy prevents misperceptions that could lead to war.

Tomorrow, we’ll discuss normative mechanisms.

12.3 Lecture 3: 6-26-19

In this lecture, we’ll talk about normative mechanisms that could connect democracy with peace.

Normative Mechanisms.

- At home: Democratic leaders solve disputes peacefully. Autocratic leaders use violence.
- Abroad: leaders “externalize” domestic norms; they apply the same norms they use at home.

Some terminology.

- Suppose there are two types of polticial regimes.
- There are three kinds of dyads: DD, DA, AA.

Version 1: “unconditional externalization.” we expect that DD is peaceful, AA is war-prone, DA is typically peaceful.

Vesrion 2: conditional externalization.

- DD is peaceful.
- DA is slightly less peaceful,
- And AA is war-prone.

Version 3: democratic crusade. Democratic states try to export their norms.

- DD: peaceful.
- AA: slightly less peaceful.
- DA: war-prone.

Nuances:

- Democratic norms take time to develop.

- Democratic norms are not upheld everywhere.

When we test theories, we have three steps.

- Collect data.
- Describe the data.
- Analyze the relationships between variables.

Ideally, we'd like to run experiments, but often running experiments is difficult. There are a couple of strategies to use when we can't run experiments:

- Select all known cases (may be experiments)
- Take a random sample (may be inefficient?)
- Choose extreme values of the IV (e.g. Turkmenistan is not democratic).

Some things to avoid:

- Choosing only one extreme of the IV.
- Selecting based on a value of the DV.
- Picking cases that "prove" your point.

Steps 2: describe the data.

There are some simple things:

- Find the range.
- Compute the mean.

Approval of Bush on Iraq graph.

Important statistic: MoE refers to the variance in the result that could emerge based on repeated sampling.

There are a number of ways to analyze relationships between variables.

- Cross-tabulation (just taking two variables, and arraying them in a table).
- Scatterplot
- Regression

Nuclear weapons experiment, (Herrmann, Tetlock and Visser, APSR 1999).

They said: think about a country that has savagely attacked its neighbor, a long-time friend of the United States. The attacker …

- Version 1: has no nuclear weapons.
- Version 2: has nuclear weapons that give it the capacity to kill millions of people in a single airstrike.

Interestingly, more people wanted to use force when the country had nuclear weapons.

Interesting case study - think critically about the quality of the data. The “Safe Celebration Study” (Sept 2004) - do they tailgate safely or not?

- Surveyed 986 college students
- 9/10 tailgate safely

Some concerns...

- Funded by Anheuser-Busch
- Relies on self-reported behavior
- Excluded students under 21

12.4 Lecture 5: 7-1-19

The focus of today: focusing on evidence about democracy and war.

Typically, when analyzing theories, we follow three steps:

1. Collecting data
2. Describing data
3. Analyze relationships.

Recall from before, the dependent variable is war vs. peace, and the independent variable is regime type, in this case democracy.

How do we define war?

Russett defines wars as “large scale institutional violence.” Typically, when we talk about war, we refer to wars between sovereign states.

- Before WWI, territory was sovereign if it received diplomatic missions from UK and France.
- After WWI, could also demonstrate sovereignty by being a member of League of Nations or the UN.

Authors exclude colonial, civil, tribal wars. The dataset we will consider is called the Correlates of War dataset.

- Each war must include at least 1000 battle fatalities.
- This excludes:
  - Mere declarations
  - Accidents
  - Rogue commanders
  - Unresisted invasions

Class survey: how to operationalize democracy:

- Free and fair elections; (Fixed periods; 2 or more different parties)
- Checks and balances / separation of powers.
- Potentially, civilian control of military.
- Policy matches what people want.
- Civil participation - vote.
- Free speech / free religion.

How does Russett operationalize democracy?

1. Creates an index (polity index), based on
    - Political participation, executive recruitment, diffuse power
    - Excludes civil and economic liberties.
2. Divides the index into segments $(-100, -25)$ autocratic; $(-25, 30)$ is anocratic, $(30, 100)$ is democratic.
3. Requires stability ($\geq$ 30 for 3 years).

How do Farber and Gowa operationalize democracy? Two variables, democ / autoc.

- If democ score ≥ 6, then it's a democracy.
- If autoc score ≥ 5, then it's an autocracy.
- Otherwise anocratic.

They omit interruptions, transitions, and interregnums.

After operationalizing variables, the authors obtained samples. They collected dyad-years (pairs of countries, corresponding to particular years).

The authors focused on certain dyads and years.

Russett focused on:

- 1946 - 1986
- Only "politically relevant" dyads.

Farber and Gowa focused on:

- 1816 - 1980.
- Exececpt WWI and WWII.
- All dyads, not just politically relevant ones.

First, note that war is a rare event. In Russett's data, of 29081 dyad-years, he finds war in 32.

Scond, note that democracy is rare, too.

- Before 1914, only 16% of countries were democratic.

- 1914-1945, 38% of countries were democratic.

Are DD dyads less likely to fight?

In Russett's data, 32 instances of nondemocratic dyad wars. There were 0 democratic dyad wars.

If democracy were unrelated to war, how many wars would we expect in each type of dyad?

- Null hypothesis, war is unrelated to a countries state as a democracy.
- Under the null hypothesis, the rate of war overall should be equivalent to the rate of war in the democratic dyads.

We can use the chi-squared test to determine whether this difference occurred by chance.

1. Scalculated chi-squared, which measures how much the table we observed differed from what we expected.
2. Use the chi-squared to find the probability of seeing a difference that large by chance alone.

Formula:

$$\chi^2 = \sum_{k=1}^{n} \frac{(x_i - m_i)^2}{m_i}.$$

Recall Russett's table:

	Democratic	Nondemocratic
War	0	32
Peace	3878	25171

Calculating, we find the that χ^2 statistic is 5. To find out how significant this is, use a χ^2 table or a calculated. If the null were true, mere chance would produce a pattern this strong only 3% of the time. Russett concluded: the pattern almost certainly did not arise by chance alone.

12.5 Lecture 6: 7-8-19

Case study, Israel and Palestine.

Timeline (focusing on territory).

- 1923, British Mandate.
- 1947, UN Partition Plan.
- 1948: State of Israel declared.
- 1948: Arab-Israeli War.

Outcome of Arab-Israeli war: Israel gained control of 78% of the territory. The armistice line was called the green line.

- Israel took Sinai, Gaza, Golan, West Bank, and East Jerusalem.
- UN Security Council passed Resolution 242. Called for Israel to withdraw.

Who controls the territories now?

- Sinai: Returned to Egypt as part of 1978 Camp David accords.
- Gaza: Israel disengaged in 2005, but still controls borders.
- Others: Israel still occupies Golan, West Bank, and East Jerusalem.

PLO Goals.

- Sovereign state.
- Based on 1967 borders.
- Capital in East Jerusalem.

Israeli Goals.

- Jewish state
- Democratic state
- In Holy Land

Israel wants security.

Historically, there have been many conflicts In Israel's history from 1948 - 2006.

There is some common ground.

- Many Palestinians accept Israel, reject vioence.
- The Israeli center-left wants a 2-state solution
- Most of the international community wants a 2-state solution.

Let's apply the framework we've learned. Perhaps conflict persists because of problems with:

- Divisibility
- Information
- Commitment

Obstacles to dividing territory.

- In theory, territory is divisible.
- In practice, division in difficult.

There are Jewish settlements in the West Bank.

Water resources in the West Bank.

In theory, Jerusalem could be divided.

- Could split or put under international control.
- Not a new idea:

  - UN partition plan called for corpus separatum
  - Jordan controlled E Jerusalem 1949 - 1967.

In practice, dividing Jerusalem would be hard.

- Israel says Jerusalem can't be divided (1980 law).
- Palestinians say Jerusalem must be their capital.
- Holy sites sit together and may be "indivisible."
- Settlers complicate the issue.

Holy Sites in the Old City.

Both sides have reasons to delay.

- For Israel, delay would allow more settlements and walls, changing the de facto division of territory.
- For Palestinians, delay would cause demoraphic patterns to shift in their favor.

Democracy is making bargaining difficult.

- We have discussed whether democracy promotoes or impedes peace.

Potential veto players.

- Israel. In 2009, centrist party (Kadima) won the most seats, but a right-wing coalition took control. Right-wing Likud won in 2013, 2015, 2019.
- Palestinians. In 2006, Hamas won a majority in the Palestinian parliament, now controls Gaza.

Would the parties keep an agreement?

- A commitment is credible if the actor has an interest in carrying it out.
- Does each side think the other has an interest in carrying out an agreement?

Many Israelis don't trust the Palestinians. Why?

- Palestinians have been attacking from Gaza.
- PA might not be able to control extremists.
- PA might not be willing to control extremists.

Many Palestinians don't trust Israel. Why?

- Israel insists new Palestinian state be disarmed.
- Israeli settlements are signals of negative intent.
- Israel makes regular incurions into "Area A."

How to ensure commitment? Some have proposed third-party enforcement.

But. . . would third-party promises be credible?

In our class survey:

- If a democracy invaded a neighbor, only 47% would support US intervention (and only 13% strongly).
- If a civil war broke out, only 46% would support US intervention (only 7% "strongly").

Would Trump side with Israel?

- US Embassy in Jerusalem
- David Friedman, US Ambassador to Israel.

For reflection.

- Why haven't Israel and the Palestinians reached a bargained solution?
- What steps would you recommend to promote peace?

12.6 Lecture 7: 7-9-19

Has interstate war become obsolete?

On the x axis is the year, and on the y axis is the historical percentage of states involved in war. Using the same criteria as in the Correlates of War dataset.

Some impressive zeros since 1946.

- There have been no wars between West European countries. This is pretty unusual.
- There have been no wars between developed countries.[1]
- No wars between US and USSR / Russia.
- No wars between great powers since 1953.

A puzzle. If interstate war has been declining, what could explain this trend?

How could we make war less likely?

- Increase the costs of war.
- Enforce commitments.
- Reduce uncertainty.
- Eliminate contentious issues.

Since 1945, what might have caused the following:

- Increased the costs of war?
  - Nuclear weapons
  - Common market.
- Enforced commitments?
  - The UN was formed.

[1] Need to determine definitions of developed country.

- Reduced uncertainty?
    - The UN reduce uncertainty.
- Eliminated contentious issues?

A few possiblities.

- Democracy.
- Trade.
- Nuclear weapons.
- International organizations.

Democratic peace.

Statistically, democracy has been spreading; whether you think about it in terms of the number of democracies or the percentage of democratic countries in the world.

Democracy could contribute to peace by. . .

- Sensitizing leaders to the costs of war.
- Making commitments more credible.
- Increasing transparency / information.

Commercial peace.

- World trade has soared since 1950s.

Idea that trade contributes to peace:

- Montesquieu wrote: "The natural effect of commerce is to bring peace."
- JS Mill wrote that: "Trade is the principal guarantee of the peace of the world" and "is rapidly rendering war obsolete"

Trade could increase the costs of war.

- Trade is mutually beneficial (due to the theory of competitive advantage; developed by Adam Smith, David Ricardo, and other influential writers). When two countries trade, it helps both countries.
- War disrupts commerce, imposing economic costs.

Trade could reduce uncertainty.

- Countries can use trade sanctions to signal resolve without fighting.
- Economic exchange might also contribute to information and mutual understanding.

Trade could eliminate contentious issues.

- In the past, countries conquered territory to gain access to economic resources.
- Today, countries trade for those resources.

- By reducing the need to take territory, trade reduces one historically common motive for war.

Problems with these arguments

- Unequal trade can be a source of tension.
- Trade causes reallocation of resources away from declining sectors to more efficient ones. But it might create loss of opportunity for people in these declining sectors, and thus violence.
- Trade policies can be a source of tension.

Testable hypothesis.

- For each dyad, measure either the volume of trade or how much each depends on trade with the other.
- Hypothesis: war should be less common in dyads with high trade, than in dyads wiht low trade.

Looking at data, trade is correlated with peace. But wait, the relationship could be spurious; there could be some factor x that leads to both trade and peace.

Examples:

- Democracy
- Capitalism
- Human rights conditions
- Affinity / good relations; democracy; alliances; geography; etc.

Wait! Causation could run the other way.

We hypothesized that peace implies trade; but it could be the other directions.

Nuclear weapons.

It could be that the existence of nuclear weapons is contributing to peace.

Many non-nuclear states are under a "nuclear umbrella."

Nuclear weapons. . .

- Increase the cost of war
- Reduce uncertainty about capabilities
- Increase uncertainty about resolve

Theory: nuclear deterrence.

- Deterrence: a strategy of preventing an attack through the threat of a retaliatory action that would cause unacceptable damage.
- Nuclear deterrence: preventing an attack throuhg the threat of nuclear retaliation.

To deter effectively.

- Able to relatiate. This requires survivable "second strike" forces.

- Willing to reliate. Would leaders actually push the nuclear button?

Evidence for a nuclear peace.

- There has never been a nuclear war.
- Nuclear weapons have been used only once, against a non-nuclear state (US-Japan 1945).
- Only one war between nuclear states (India-Pakistan 1999), and it was minor.

But non-nuclear dyads have become more peaceful, too.

International organizations.

Today, we have more IOs than ever.

- United Nations
- Regional organizations, many focused on the topics of security and the maintenance of peace.

International organizations such as the UN can…

- Increase the cost of war, e.g. by banding countries together by fighting a country that is belligerent.
- Enforce commitments

But the UN has taken action in relatively feew wars.

Of 39 interstate wars since 1945…

- UN applied military sanctions in only 2 (Korea, Persian Gulf war).
- UN applied economics sanctions in 3 (Persian Gulf, Yugoslavia).

Of 168 civil wars from 1946-277…

- Military sanctions: 2
- Economic sanctions: 13.

If the UN were successful in preventing war…

We might expect:

- Most dyads would not erupt into wars.
- If a dyad does erupt into war…
  - The UN was unable to prevent a war
  - And might not be able to end the war
  - So UN involvement might not make sense.

Implication.

- To many, the UN looks like a failure.
- But perhaps its success is mostly invisible!

12.7 Lecture 8: 7-10-19

Agenda.

- What causes civil war?
- Why are civil wars so prevalent today?
- When do civil wars end?

Interestingly, interstate war has become less common, but civil war and insurgency (uprising) have not.

What is a civil war?

- A civil war is a violent conflict between the stae and non-state armed groups for political control.
- Battle-death threshold
  - >1000 battle-deaths total (Fearon and Latin 2003)
  - 25 battle deaths (UCDP PRIO)

Types of civil wars.

There are many ways to categorize civil wars, e.g.

- Political aims
- Ideological cleavages
- Foreign intervention

What causes civil wars today?

We will discuss 2 different classes of explanations for civil war onset:

- Motive based explanations
  - Grievance
  - Greed
- Opportunity based explanations

Ethnic / Islamist tensions have grown (see graph.)

These tensions drive groups to rebel.

- Connecting logic: groups rebel because they "hate" the state authority due to differences in identity and cultures.
- Hypothesis: Ethnic and religiously-motivated groups are more likely ot cause civil war.

Test: Ethnic hatreds?

- Note that $P(\text{War}|\text{Ethnic Grievance}) = 0.108$.
- Note that $P(\text{War}|\text{Non-ethnic grievance}) = 0.073$.

- The p-value is roughly 0.07.

Test: religious hatreds.

- $P(\text{War}|\text{Islamist}) = 0.095$
- $P(\text{War}|\text{Non-Islamist}) = 0.089$.

Is "greed" to blame?

- The state controls lots of "prizes"
- Chronic poverty means few opportunity costs to fighting.
- Groups rebel because they can profit from war.
- e.g. Blood Diamond case in Sierra Leone.

Lots of group try and fail to rebel.

- Greed and grievance are insufficient to explain civil war onset by themselves.
- The state is typically good at destroying (or negotiating with) emerging threats.
- Some groups have better opportunities to rebel.

Groups have different opportunities to rebel.

- Groups today are relatively weak; so must be careful in when and how they choose to rebel.
- Strategic choices:
  - Regular fighting (conventional military engagements)
  - Insurgent fighting (Guerrilla and terrorism tactics)

Historical civil wars were fought between conventional forces.

- e.g. The Confederacy during the US civil war thought of itself as a fully functional independent government.

Modern civil wars involve insurgency.

Why does insurgency matter?

- Insurgency allows rebel group to avoid detection / destruction by the state.
- Insurgency still puts pressure on state to meet rebel demands.
- Opportunities to conduct insurgency make civil war more likely.

Operationalizing our Hypothesis.

What conditions make insurgency more likely?

- Areas where it's harder to detect rebel groups
- Areas where it's harder to destroy rebel groups.

Factors that may influence the presence of rebel groups.

- Low population vs. high population density area.
- Existence of roads.
- Presence of civilian shields (e.g. hospitals, schools).
- Low income / low education areas.
- Rebel detection tends to be harder in rough terrain.
- Rebel destruction is harder in weaker states (e.g. countries below the median GDP tend to be more likely to fight).

What causes civil war?

- Motive and opportunity are not sufficient
- War requires:
  - Conflicting preferences
  - Opportunity to negotiate (or fight)
  - Bargaining failure

Why might opportunities for insurgency cause bargaining failures?

- Rough terrain → information problems. e.g. State fails to identify an emerging insurgent threat and miscalculates response.
- Weak governance → commitment problems. e.g. State identifies a threat, but can't credibly commit to acommodate its demands.

Why have civil wars become more common?

Are conflicts breaking gout more frequently? Not exactly. The increase in civil war today is due to the accumulation of ongoing civil conflicts.

Civil wars last a long time.

- Median duration of conflicts, 1946-2007.
  - Civil: 20 months
  - Interstate: 3.4 months

What are the obstacles to ending civil war?

Disarmament dilemma prolongs civil war.

Logic of disarmament dilemma.

- Settlement requires rebels to disarm.
- But the government can exploit disarmament.
- Fearing exploitation, rebels might prefer to keep fighting.

What resolves the disarmament dilemma?

- Power-sharing (e.g. in Iraq)
- Independence (but this leads to slippery slope-style problems)
- Ceasefire? (e.g. in Syria civil war)

But: in nearly all of these cases, fighting resumes. The core problem is that these are not credible commitments.

What makes civil war settlements credible?

- Third party enforcement
- Why?
  - Raises the costs of cheating.
  - Monitor compliance.
  - Threaten military force or economic sanctions for non-compliance.

Third-party = peacekeepers.

The UN frequently serves as a third-party to oversee the enforcement of a civil war settlement.

Does peacekeeping work?

If you control for selection effect, then yes.

Midterm format.

Midterm is 50 minutes. Two types of questions.

- T/F or multiple choice questions (probably 20).
- Short answer (choose 4 identification questions out of 6).
- Covers reading, lectures, and discussions.

Final exam is up to 3 hours.

Review session on Tuesday.

12.8 Lecture 9: 7-15-19

Note that we have a midterm on Wednesday, and there will be a review session.

Two aspects of the ethics of war:

- Groups for war (Jus ad bellum)
- Conduct in war (Jus in bello)

We will describe three different ethical traditions to this question:

- Political realism
- Consequentialism

    - Utilitarianism
    - Case study: Hiroshima
- Non-consequentialism
    - Christian Just War
    - Islamic Just War

Political Realism.

Realists claim that the real world is a violent world. And states in IR are like gladiators.

In this state war…

- "It is better to seek salvation via the sewer." - Bismarck (Have to play the same game, and be brutal in war.)
- "Notions of right and wrong have no place." - Hobbes

Example: Melian Dialogue (416 BC).

This dialogue takes place during the Peloponnesian war.

"the strong do what they can and the weak suffer what they must"

The Athenians crush the Melians.

Some implications.

- A realist would say that moral behavior is dangerous and irresponsible.
- A realist would say that moral rhetoric is hypocritical and pointless.

Utilitarianism and war.

- Consequentialist: the moral value of an action/institution lies in its consequences
- Hedonist: pleasure/happiness is the ultimate good

Utilitarian principle.

An action or policy is morally right if it produces the greatest balance of happiness over unhappiness.

Applying utilitarianism to nuclear weapons.

U.S. dropped two nuclear bombs in 1945:

- Hiroshima (Aug. 6)
- Nagasaki (Aug. 9)

Hundreds of thousands died.

- Hiroshima (in 1945, 140K died; next 5 years, 60K died).
- Nagasaki (in 1945, 70K died; next 5 years, 70K died).

Total of 340K.

How might a utilitarian analyze the decision?

Did it save lives? (Use counterfactual reasoning)

- Look at U.S. casualities that may have been saved
- Look at the positive good created by ending the war

Claim: the alternative was invasion.

There were two

Invasion would have killed. . .

- Truman says 500K Americans would have died.
- Churchill claims that 1M soldiers would have died.

Utilitarians would conclude that the atomic bombing was justified because it saved lives.

Utilitarian counter-arguments.

- War might have ended without invasion
- Invasion might not have been so deadly
- There are more humane ways to use the bomb

War might have ended without invasion.

Evidence:

- Japanese were getting weaker
- Soviets were about to tip the scales
- U.S. could have relaxed its demands

Invasion might not have been so deadly.

- Martin Bernstein: at most, 46K Americans would have died. According to Bernstein, the 500K / 1M figures from before were post-hoc justification (and not necessarily true).

More humane way.

Why not just show them how powerful this weapon is?

But, US worried about:

- Failure / embarrassment
- Disclosure to enemy (they might have information about military power that could be used against you)

But: even after bomb was dropped on Hiroshima, Japan didn't surrender.

Warning shot.

Drop a bomb in e.g. Tokyo Bay, A Forest, Mt. Fuji.

Tactical strike.

Could we do a tactical strike against a military target? But, they needed a target 3 miles in diameter.

Allow evacuation.

But, U.S. feared

- Bomb could fail.
- Japan could intercept.
- POWs as human shields.

Did dropping the bomb save lives?

- Relative to non-nuclear alternatives?
- Relative to other nuclear options?

Divine command and war.

How authoritative are divine commands?

- Texts often ambiguous
- They require interpretation

Christian tradition.

- Originated with St. Augustine
- Between skepticism and pacifism

Grounds for war.

- Just cause
- Last resort
- Chance of success (interestingly tied in with utilitarian tradition)

Conduct in war.

- Discrimination (acceptable to kill military, but not civilians.)
- Proportional tactics (goal of the war is to solve an injustice; can't use force that isn't proportionate to the injustice).

Islamic just war thinking shares similar conclusions.

Key concepts.

- dar al-Islam (house of Islam, territory under the control of Islam, the Zone of Peace) versus dar al-harb (non-Muslim world)
- Jihad is striving, toiling for God (somewhat misrepresented in literature / media).

Grounds for war (Islam.)

- Just cause (e.g. there is a debate about whether you can fight a war to propagate Islam)
- Last resort

Jihad against Jews and Crusaders (1998)

- Highly controversial (signed by bin Laden)
- Says the US has been occupying holy places, killing Muslims in Iraq.
- Calls for military action vs. US allies

Conduct in war.

- Discrimination. The Quran says "fight in God's cause against those who wage war against you, but do not transgress God's limits"
- Proportional tactics

For reflection: apply ethics to a contemporary issue.

- U.S. counterterrorism policy
- The ongoing war in Syria
- Preemptive strike vs North Korea
- Israeli-Palestinian relations

12.9 Lecture 10: Midterm review

Format.

- Exam: in class (be on time).
- True / false questions (20) + ID questions (4 of 6).
- Writing utensils.
- Exam is closed book, but you get a reading list.

Identification question.

- Clearly and succinctly define the term
- Contextualize the term
- Consider how it is important to IR
- Where possible, cite readings
- Write legibly

Write for the full time (10 minutes, 2 paragraphs).

Example: "Insurgency."

- Concise definition
- Historical example
- Use of reading example

Underline key concepts so graders don't miss them.

Statistics and experiment design.

Chi-squared test.

- Statistical test for categorical data
- Measures how far the observations deviated from what we expected under the null hypothesis of no relationship
- Tells us how likely it is that the observed difference between the categories arose by chance

Reliability vs. validity.

Used to describe measurements.

Validity. Does the masure capture the concept?

Reliability. Would different people's measurements of the concept produce the same results?

We can think of four cases (Valid vs. invalid × reliable vs. unreliable).

Unreliable but valid (maybe centered around the right thing, but high variance).

Statistical terminology.

- Independent variable
- Dependent variable
- Hypothesis
- Connecting logic
- Spurious Relationship (a relationship that you detect between two variables that might be caused by a confounder. e.g. correlation but not causation)
- Operationalization (when you want to measure a concept, and you define a concrete measurable variable)
- Null hypothesis
- Reliability
- Validity
- Cross-tabulation
- Regress
- Scatterplot

- Chi-squared test
- Empirical evidence
- Sampling error (when due to pure chance, your sample differ from the population)
- Marging of error
- Non-sampling error (e.g. when you non-randomly sample your units, or measure something incorrectly)
- Selection effect (e.g. in Fortna reading, when your sample is biased due to some selection reason. For example, when looking at peacekeeping, sample is necessarily more violent.)

Democratic peace theory.

Structural reasons why democracy are more likely to stay peaceful.

- Empower voters, delay mobilization, transparent
- Assumptions needed:
  - voters dislike war
  - political leaders care about staying in office
  - veto players, transparency

Normative.

- States externalize about the norms that that are used to resolve disputes at home
- Democracies have solve domestic disputes through peaceful methods and autocracies solve domestic disputes through violent methods
- Unconditional vs. conditional
- Democratic crusade - export democratic norms to autocracies

Recall the readings:

- Russett (claims that normative theories of democratic peace are weaker, thinks it is a good thing democracies delay mobilization)

Is Democratic Peace Theory Right?

- Since 1946, only one possible case of two democracies fighting a war (Kargin War 1999, between India and Pakistan)
  - Is this by chance?
  - Are common interests (e.g. opposite to the Soviet Union a better explanation for why democracies have not fought each other?); due to Farber and Gowa (1995).

Bargaining theory.

The literature / readings are good. Go back to the section notes (combined section on July 3rd).

Puzzle of War.

- War is costly -> must be deals that both states prefer to war (bargaining range).
- Example: Mexican-American war (settlement would have saved casualties, but it fell apart).
- Conflicting preferences are not sufficient for war ->

Need three things for war to break out:

- Conflicting preferences
- Opportunity to fight / negotiate (e.g. it's unlikely that Sweden and Nicaragua will fight because of the distance)
- Bargaining failure

What is the bargaining range?

- Let's assume that Sate A and state B value some territory at 100. Both State A and B have a 50 pct chance of winnning the war and would pay $20 for fighting.
- Calculating the payoff each side can expect from war.

$$\text{War value} = \text{Prob. Win} \times \text{Gain of war} - \text{Costs of war}$$

- Bargaining range for each player is the intersection between deals that post prefer to war.

How does the BR change with costs of war?

- The bargaining range expands / shrinks with the costs of war.
- e.g. if war becomes more costly, both side would prefer a negotiated settlement.
- Bargaining range expands as the costs of war increase.

BR range shifts to the left or right depending on who is more / less likely to win.

Causes of bargaining failure.

- Issue indivisbility
- Information problems
- Commitment problems
  - Preventive wars
  - Preemptive wars

Issue indivisibility.

- Issues that cannot be divided into a range of potential settlements (e.g. King Solomon Baby problem).

Information problems.

- State *A* sometimes has incomplete information about state B's capabilities / resolve / costs of fighting.

- State *B* has an incentive to misrepresent this informaiton in order to negotiate a better deal for itself.

- Private information + incentives to misrepresent increase risk of underestimating opponent's willingness to go to war.

- Analogy. Poker game and bluffing.

- Examples: WMDs in war on terror (Iraq War, 2003).

Commitment problem.

Idea: states may end up in war because one or both sides are not able to commit to abide by the terms of an agreement that would allow them to avoid fighting.

States may have an incentive to defect from their agreement.

Analogy: Stanford Honor Code

There can be many kinds of commitment problems that lead to war, but two common examples are preventive and preemptive wars.

Preventive wars.

Wars that occur in the context of large power shifts.

Bargaining range is more favorable to *A* today, but will be more favorable to *B* in the future.

Example: U.S. - China wars. China's economy will eclipse US in 5 years, and thus there's an incentive to fight a war earlier.

Examples: 2003 Iraq War (WMDs), Iran-Iraq War (1981)

Preemptive wars.

- These are wars that occur in response to large first-strike advantages.

Examples. Rare, but 1967 Six Day War, and Pearl Harbor are plausible examples.

Lake (2010/11).

Uses Bargaining theory to examine outbreak of 2003 Iraq War and evaluates how useful it is.

Critiques following assumptions of bargaining mode:

- States are unitary actors.

- Bargaining is modeled as a two-player game

- Bargaining theory does not include the costs of enforcing a settlement

- States are rational actors.

Israel-Palestine case study.

Israeli Goals:

- Jewish state

- Democratic state

- In the Holy Land

PLO Goals:

- Sovereign State
- Based on 1948 Borders (aka “Green Line”)
- Capital in East Jerusalem

Focus: think about this as an example of the bargaining problem.

Issue Indivisbility?

- In theory, territory is divisible, but in practice it is very difficult.
- Complicating issues:
  - Israel’s Jerusalem law insists that Jerusalem by unified, but Palestinians insist on East Jerusalem as capital of Palestine.
  - Israel Settlements in West Bank and East Jerusalem
    - Changes the de facto division of division of territory
    - Gives incentive to delay.
  - Water Resources in West Bank (very few water resources).

Commitment problems.

- Israelis do not trust Palestinians (Palestinians have been attacking from Gaza).
- PA might not be willing or able to control extremists
- Israel insists Palestinian state to be disaremd
- Israeli settlements are signal of negative intent.

Ethics of Warfare.

Four theories of ethics and war.

Political realism.

- Thucydides, The Melian Dialogue

Utilitarianism.

- Pick the option that produces the greatest net happiness. (Critique: Holt).

Christian Just War Theory

- Rights to go to war
  - Self defense, last resort, chance of success
- Rights during war
  - Discriminate between combatants and civilians, proportional tactics

Islamic War Theory.

- Rights to go to war. Just cause (including propagating Islam, last resort).
- Rights during war (Discriminate between combatans and civilians, proportional tactics)

12.10 Lecture 11: 7-22-19

In survey data, people say they are quite concerned about environmental issues. The main reason:

- Pollution is individually rational.
- Coercion could solve the problem, but that's difficult in a condition of anarchy.

Example: overgrazing (leads to loss of nutrients).

What happens when too many animals graze?

- Desertification
- Soil erosion
- Invasive weeds

Thus, the land eventually becomes unstable.

There are a couple of cases to consider when thinking about the tragedy of the commons.

- I own the grass (then I'll overgraze)
- If you own the grass (then I'll not overgraze)
- If the grass is common property (then I'll overgraze).

Concretely, the pursuit of individual self-interest leads to a collectively bad outcome.

Identify international environmental problems that have a similar logic. Several examples, including air pollution, overfishing, aquaculture, etc.

We can model this broad setting using game theory. Consider a game between two countries: A and B. Assume the following:

- Each can contribute to reducing pollution.
- Both contributing would be better than neither.
- But: contributing is costly.

Numerical example. Assume:

- If both countries contribute, they produce a public good worth $4 per country.
- If one contributes, the contributor produces a public good worth $2 per country.
- If neither contributes, no public good is produced.
- It costs $3 for each to contribute.

It follows that:

- If they both contribute, they each get 1.
- If only one contributes, one gets 2, and another gets -1.
- If neither contributes, they both get 0.

We can summarize this in a matrix, where rows indicate action of each player, and entries indicate net payoffs per agent. This is analogous to prisoner's dilemma.

Importantly, the prisoner's dilemma:

- Does not depend on lack of communication.
- Does not arise from uncertainty.
- Does depend on a lack of mutual concern.

Tomorrow, we will discuss potential ways to solve the game.

12.11 Lecture 12: 7-23-19

Last time, we emphasized that there isn't a technical solution to the prisoner's dilemma (see the Hardin reading). There are four potential solutions to this broad class of problems we will discuss:

- Technology
- Coercion
- Reciprocity
- Domestic pressure

We will also discuss how international agreements can reinforce these mechanisms.

Science (or nature) could decrease costs and/or increase benefits of contributing. This could align individual and collective incentives.

Let's consider the game from yesterday, with a change of payoffs. What is A's cost of contributing fell from $3 to $1. Then, we might get a table like

- A, B both contribute. A gets $4 - 1 = 3$, B gets $4 - 3 = 1$.
- A contributes, B doesn't. A gets $2 - 1 = 1$, B gets $2 - 0 = 2$
- A does not contribute, B contributes. A gets 2, B gets -1.
- If A and B don't contribute, A gets 0, B gets 0.

With these new payoffs, A should contribute, and B should not contribute, so we have a new equilibrium.

Now, if the value of the collective good rose, assume 8 if one contribute, 4 if one contributes. Then we get the following results:

- A, B both contribute. A gets 7, B gets 5.

- *A* contributes, *B* doesn't. *A* gets 3, *B* gets 4.

- *A* does not contribute, *B* contributes. *A* gets 4, *B* gets 1.

- If *A* and *B* don't contribute, *A* gets 0, *B* gets 0.

In this case, the payoff for contributing is always higher (regardless of what the other country does), so the equilibrium is *A* and *B* both contributing.

Example: the ozone layer (blocks harmful UV radiation). In this setting, science shifted the benefits / costs. The benefits of action increased as people learned about the dangers of CFCs. The costs of action decreased as companies developed subtitutes for CFCs. Political response: 1987 Montreal Protocols: countries pledged that they would phase out over time the use of CFCs with the goal of protecting the ozone layer.

Kofi Annan stated that the Montreal Protocol is the most successful international agreement that he had ever encountered.

Broader significance of the Montreal Protocol:

- First treaty to address a global environmental threat

- Embodied the principle of "differentiated responsibilities" (maybe richer countries can implement these changes more quickly).

- Acted without scientific certainty (precautionary principle).

- Can strengthen the treaty without formal amendments - (ratcheting provision).

For more on consts and benefits, see Sprinz and Vaahtoranta.

Second way to solve international problems: coercion by a strong state.

- The PD involves a commitment problem: states could promise to contribute, but the promise would not be credible.

- A strong state could solve the commitment problem by punishing shirkers and/or rewarding contributors.

What is an enforcer imposed a \$2 cost on cheaters? Then the game would go this way:

- *A* contributes, *B* contributes, *A* and *B* get 1.

- *A* contributes, *B* does not contribute, so *A* gets $2 - 3 = -1$, *B* gets $2 - 2 = 0$.

- *A* doesn't contribute, *B* contributes, *A* gets $2 - 2 = 0$, *B* gets -1.

- *A* and *B* don't contribute; so *A* gets -2, *B* gets -2.

Some problems with this approach:

- Would the strong state actually punish? Punishment is costly to both the target of the sanctions, but also the country that is imposing the sanction.

- Would other countries join in the punishment?

Example. Whaling moratorium: in 1985, moratorium came into effect. Whale catch by Japan, Norway, and Iceland has falling drastically since then.

Limits of the agreement. Japan and others have exploited a loophole: countries may whale for "scientific purposes." Now, violators are US allies, e.g. Japan, Norway, Iceland. Would US really punish these countries?

But recently, in July 2019, Japan withdrew from the IWC; it promptly resumed commercial whaling.

Strategies of reciprocity. This is the third class of solutions to the problem.

If the game is played repeatedly - the incentives change. With repetition, players can use strategies of reciprocity. You condition your move based on what others do in previous plays of the game. There are various versions of this:

- "I will cooperate only as long as you cooperate."
  - e.g. I will limit my fishing if other countries do so.
  - e.g. I will restrain my use of fossil fuels if other countries do the same.
- If leaders care enough about the future, this strategy could sustain cooperation.

Let's see how reciprocity would work. Recall the matrix looks like

$$\begin{pmatrix} 1,1 & -1,2 \\ 2,-1 & 0,0 \end{pmatrix} \begin{pmatrix} 1,1 & -1,2 \\ 2,-1 & 0,0 \end{pmatrix} \begin{pmatrix} 1,1 & -1,2 \\ 2,-1 & 0,0 \end{pmatrix}$$

- "I will cooperate in every period, but if you ever defect on me, I will never copperate with you again." (grim trigger strategy). Then:
  - Payoff to copperating: $1+1+1+\dots$
  - Payoff to defecting: $2+0+0+\dots$

Lessons from the repeated prisoner's dilemma:

- Defection is profitable in the short term.
- With strategies of reciprocity, the long-term benefits of ongoing cooperation can outweight the short-term incentive to defect.

International agreements can facilitate reciprocity by...

- Setting clear expectations
- Monitoring behavior
- Coordinating punishments

Mobilize domestic interests.

Another way to solve environmental problems: use international greements to mobilize domestic groups.

- Give groups the right to sue in domestic courts.
- Create benchmarks for "naming and shaming"
- Foster international linkages among groups.
- Change preferencces/beliefs of ordinary citizens.

But, would citizens support an agreement? Tried to assess whether people would enter into hypothetical agreements (see Bechtel and Scheve (2013)).

Next assignment: policy memo. Goal: work with another student to advise the U.S. government about a major problem involving:

- Environment (unit 2), or
- Trade (unit 3), or
- Poverty/aid (unit 4)

Four-page memo: should have four parts:

- Executive summary
- Problem
- Solution
- Political feasibility

Deadlines:

- Mon, July 29 at noon (Partner with another student and send names to TAs).
- Fri, 8/2 at noon: send TAs two sentences: problem and recommendation.
- Wed, 8/14 at 4:30: submit memo.

On 8/7: no lecture (work with partner on policy memo).

12.12 Lecture 13: 7-24-19

Today: we will discuss the ethics of climate change. Before the Paris agreement, we had the Kyoto Protocol of 1997.

- Annex 1 countries committed to binding reductions in GHG emissions relative to 1990.
- Non-annex I countries had no binding commitments

Kyoto didn't work - no limits on China, India, other developing countries.

Lack of political support in rich countries:

- US never ratified, then withdrew signature
- Canada failed to reach target, withdrew
- Japan missed target

Deepest cuts came from collapse of USSR - because of economic collapse.

Main features of Paris agreement -

- Goal: prevent temperature from rising more than $2°C$ by 2100 (relative to pre-industrial levels). Ideally, no more than $1.5°$.

- Method: intended nationally determined contributions (INDCs).

Paris attempted to solve the problmes of Kyoto

- All countries submitted INDCs;
- Rich countries help LDCs shoulder cost
- Regular monitoring and reporting
- Commitment to ratchet up. Initial pledge; and the ability to increase commitment

Could Paris work? Factors:

- Technology
  - Innovation could decrease costs of contribution
  - Increasing evidence of harm could spur action
- Coercion
  - Rich countries will help pay for mitigation
  - Will they punish countries that cheat?
- Reciprocity
  - Agreement calls for monitoring and ratcheting
  - These provisions could facilitate reciprocity
- Domestic politics
  - Agreement could sway the domestic public
  - Violations could prompt shaming and lawsuits

US is...

- Withdrawing from Paris
- Increasing oil / gas drilling
- Imposing tariffs on solar panel imports

Question - do we have a moral responsibility to address climate change?

Utilitarian argument:

- Climate change is bad.
- Climate change can be reasonably averted

Problem 1: scientific uncertainty:

- Negative feedback mechanisms, such as cloud cover
- Positive feedback mechanisms, such as methane from thawing permafrost and the ice albedo feedback

Problem 2: how much to weigh the future?

- Solving climate change requires immediate sacrifices in exchange for future benefits.
- How should we weight current versus sfuture payoffs?
- Need to compute a discount rate to calculate things effectively.

Problem 3: contingency

- Assume no single nation can cause or prevent climate change
- If others don't act; my country shouldn't act (efforts are futile)

Some non-utilitarian approaches to climate justice.

- Corrective justice
- Egalitarian justice
- Shared responsbilities

Corrective justice. Broadly just means - "you broke it, you buy it"

This would imply that US, China, UK should repair climate change. These are standard objections:

- The harm was unintentional. (reply: if unintentional, countries shouldn't pay punitive damages, but they should still pay compensation)
- Current generation shouldn't pay for the sins of previous generations. (reply: current generations should pay, becaue they are the beneficiaries of exploitation by previous generations)

Posner - discusses more objections.

Egalitarian justice.

Idea: give each person an equal share of the atmosphere. Or an equal share of the remaining carbon budget.

Problem: rich are emitting far more than their equal share. To implement egalitarian justice, you could:

- Require rich countries to cut emissions
- Or: allow emissions trading: rich could buy pollution rights from the poor. or: cap and trade

Shared reponsibilities.

(from Goodin's article).

- Shared rights
  - Sovereign countries have right to exploit resources. Can't intervene unless there is a transboundary impact
- Shared duties
  - Shared duty not to pollute
  - Intervention is supererogatory

- Shared responsbilities (the idea he most strongly supports)
    - Countries have a duty not to pollute, a duty to "pick up the slack," and also to intervene against others

Question: no individual can have a discernible effect on global climate change. Given this fact, do you, personally, have a moral obligation to reduce your emissions of CO2?

12.13 Lecture 14: 7-29-19

Free trade refers to unregulated economic activity. Protectionism is when you impose barriers.

Many American oppose free trade (65% argue for more restrictions).

If you ask economists, 95% of economists in the US support free trade, while 88% of economists worldwide support free trade.

Classic case for FT.

- FT increases overall economic welfare (increases the size of the pie).
- The pie gets bigger in both countries. Importantly, both sides gain, even when one country is better at making everything.

Intuitively - it might be the case that Lebron James is the world's best lawnmower, but it doesn't make sense to get Lebron James to mow everyone's lawn.

This lays out a case for collaboration, even if you're really smart.

This is an old argument, dating back to Adam Smith, Ricardo.

Consider a simple economic model.

That is, suppose there are:

- 2 countries (France and Switzerland)
- 2 goods (wine and cheese)
- 1 factor of product (labor).

Assumptions:

- Each country has 1 million workers
- Production process is linear

Key concepts:

- Absolute advantage (you can create more product at same amount of time).
- Opportunity costs

We will consider 2 cases:

- Each country has AA in one good

- Each country has AA in both goods

The second case is kind of surprising, which shows that trade is beneficial even when one country has AA in both goods.

France has AA in wine, Switzerland has AA in cheese.

Suppose:

- France: Wine for 100 labor units, cheese for 50 (red).
- Switzerland: Wine for 50 labor units, cheese for 100 (blue).

You can draw a graph as follows:

Under autarky, each country can consume only what it produces.

What price would be acceptable to both sides?

- France: In trade, refuse to pay more than 100w for 50c.
- Swiss: In trade, refuse to pay more than 100c for 50w.

In particular, there is a wide range of acceptable prices. Red: range that is acceptable to France; blue: range that is acceptable to Switzerland.

Trade would both to consume more. Suppose the trading price is 100/100. Then..., the CPF looks like (it moves outward).

New case: France has AA in both goods. In this setting, suppose that:

- France can produce 100 wine, 50 cheese.
- Swiss can produce 10 wine, 20 cheese.

It turns out that understanding this case requires us to understand absolute vs. comparative advantage.

- Absolute advantage: lower absolute cost of producing x.
- Comparative advtange: lower opportunity cost of producing x.

Importantly, Switzerland Has a comparative advantage in cheese. France has to give up 200 wine to make 100 cheese, while Switzerland has to give up 50 wine to make 100 cheese.

Now, obviously - Switzerland gains from trade.

But also, France gains too:

- With trade: France could make 100m wine, then trade 20m wine for 20m cheese.
- Result: consume 80m wine, 20m cheese.

Without trade:

- If France consumed 80m wine, it could only consume 10m cheese.
- If France consumed 20m cheese, it could only consume 60m wine.

Other arguments for free trade:

- Trade increases welfare via economies of scale. In some industries production costs fall as output increases (due to learning, technology, machinery).
- Trade increases welfare due to competitions. Firms will feel compelled to cut costs, enhance their products, and improve their services.

Conclusions:

- FT increases aggregate welfare, even when one country has an AA in all goods.
- Puzzle: why do countries have trade barriers?

12.14 Lecture 15: 7-30-19

Even though economists are in favor of free trade, protectionism is widespread. Why is this?

Tariffs. There are two types:

- Ad valorem tariff (tax is a % of the good's value; e.g. sales tax).
- Specific tariff (tax is a fixed amount per unit).

Interestingly, rich countries have lower tariffs, on average. Note that:

- High income countries (3.6% avg pct tariff)
- Middle income countries (9.2% avg pct tariff)
- Low income countries (12.1% avg pct tariff).

But: tariffs vary within income groups. E.g. South Korea has 8.9% tariff, while U.S. has 2.9% tariff.

Also, tariffs have historically varied over time.

There are also nontariff barriers (NTBs).

- Quotas / licenses: quantitative limits on imports.
- product standards: block imports that don't meet standards (sanitary, environmental, medical, etc).

Subsidies are another type of NTB (they help domestic producers beat foreign competition).

Currency policies can be NTBs.

- Restrict access to foreign currency
- Devalue your currency (if a country cheapens its currency, foreigners will buy more from that country, and consumers in the country will buy less from foreigners).

China as currency manipulator. e.g. Romney says that companies have shut down and people have lost their jobs because China has not played by the same rules. Would prevent Chinese consumers from buying stuff from the U.S. and making Chinese

How could China keep its currency low?

- China could print new currency, use it to buy dollars and U.S. debt and hold them as financial assets (China holds $1.1T$ of US debt).

- Consequences: supply of Chinese currency rises (lowering its value), demand for the U.S. dollra rises (raising its value).

What evidence is there that this is happening?

Before 2010, value of Yuan is completely flat relative to the USD. Eight year low right before 2019.

The problem of substitutability: there is more than one way to block an import. Makes it hard to ensure that a country is practicing free trade.

Question: why might protectionism be in the national interest?

- National defense
- Infant industries
- Market power
- Industrialization

P = protectionism. Some people say that protectionism means that it is important for national security.

e.g. steel tariffs - administration notes that US should build steel at home. Similar for aluminum.

Justification for automobile tariffs? Commerce dept concluded that imports of autos / certain auto parts posed a threat to US national security.

Rebuttal -

- What wouldn't qualify? (it's a slippery slope).
- There are alternatives: buying from allies, and stockpiling for emergencies.

Infant industries:

- Shield young firms from foreign competition until they can succeed on their own
- Example - blocking Google / Twitter In China led to Baidu / Weibo

Rebuttal:

Might not work:

- Can government pick winners?
- Can the infact be weaned?

There are alternatives:

- Let prviate investors support the industry
- If you must help an infant, subsidize it

Market power:

(Krugman is famous for this)

Talked about how countries that strategically use subsidies to help one country gain at the expense of another.

Example. Suppose:

- 2 regions (US and Europe)
- 2 firms (Boeing and Airbus)
- Only 1 firm can remain profitable

Initial payoffs: numbers in cells are profits, in $ millions.

	Airbus Produce	Airbus Abstain
Boeing Produce	(-5, -5)	(100, 0)
Boeing Abstain	(0, 100)	(0, 0).

There is no dominant strategy here. If Boeing produces, Airbus should abstain. If Boeing abstains, Airbus should produce.

Interestingly: subsidies give Airbus a dominant strategy. Government might add +10 if they produce, but no plus if they abstain.

	Airbus Produce	Airbus Abstain
Boeing Produce	(-5, -5+10)	(100, 0)
Boeing Abstain	(0, 100+10)	(0, 0).

Knowing this: Boeing will stay out, and Airbus will take the profits.

Rebuttal:

- Requires a super wise government
- Applies to only a few industries

Tomorrow, will discuss argument that protectionism could help LDCs industrialize (shift from primary to secondary products).

12.15 Lecture 16: 7-31-19

Some people argue that protectionism could help LDCs industrialize (shift from primary to secondary products).

- Primary: agricultural production
- Secondary: industrial production

Why is it better to focus on industrial production?

Claim 1. Industrial goods have better prospects

- Engel's law: as income rises, % spent on food will fall, while % spent on non-food (industrial) items will rise.
- Technology: synthetic substitutes can reduce the demand for primary goods.

Claim 2. Prices of primary goods are more volatile.

Why?

- Business cycles in rich countries
- Unpredictable weather

LDCs used protectionism to address these problems.

Import substitution industrialization:

- Latin America, 1930s-1960s
- High barriers on final products
- Allow inputs to enter freely

Did ISI work?

Domestic policies.

Key ideas about domestic policies:

- Protection actually helps some domestic groups
- Their influence dpeends on political institutions.

Consider the following model, with these assumptions.

- Two products: shirts and cars.
- Two factors of production: labor and capital (means machines, factories, etc.)
- Two countries with different factor andowments. One country - lots of labor; another country - lots of capital.

As from before, recall that trade leads to specialization.

- Under autarky: you make both shirts and cars.
- Under free trade: specialize according to comparative advantage

Specialization will cause certain industries to expand, others to contract.

e.g. country with a lot of labor will focus on shirts; country with a lot of capital will specialize on capital.

We will consider two theories on who wins domestically. See Stolper-Samuelson and Ricardo-Viner.

Stolper-Samuelson:

- Assumption: both factors of production are highly mobile (labor and capital). e.g. this means that people who make shirts can move. And, the equipment that is used to make shirts, can be reconfigured to make BMWs (questionable assumption).
- Prediction: trade -> class conflict

In a labor abundant country, labor-intensive industries will grow. This leads to a shortage of labor in shirt industry, so wages rise.

This leads to surplus capital in shirt industry - so value of capital falls.

In a capital abundant country, capital-intensive industries will grow. Labor will move from shirts to cars, and capital will move from shirts to capital.

- Causes a shortage of capital in auto industry -> value of capital rises.
- Causes a surplus of workers, so wages fall.

In this setting, FT helps capitalists, hurts workers.

Political implications of SS theory:

- In labor-abundant countries:
  - Workers should favor free trade
  - Capitalists should favor protectionism
- In capital-abundant countries:
  - Capitalists should favor free trade
  - Workers should favor protectionism
- Implies a fight between classes (labor vs capital).

Objection to Stolper-Samuelson: assumes that factors of production can be redeployed easily. True in some cases, not in others.

Ricardo-Viner theory: a different way of thinking about this process.

- Assumption: some factors are fully mobile.
- Prediction: trade → conflicts between industries, rather than classes.

Example. Suppose that capital is hard to move.

- In a labor abundant country, trade → redeployment of labor, but some capital remains stuck. Can't move capital from car to shirt.
- How will this affect domestic groups?
  - Shirt capitalists will win
  - Car capitalists will lose
  - Effect on workers is harder to predict (depends on how much their wages change, etc).

The opposite is true in a capital-abundant country. Car capitalists will win, shirt capitalists will lose.

Political implications of RV:

- If capital is immobile, trade will hurt some capitalists while helping others.
- If labor is immobile, trade will hurt some workers while helping others.

- Thus: battle is between industries, rather than classes.

Puzzle:

- FT increases economic welfare (overall size of the pie).
- So: why don't winners compensate the losers?

One approach: trade adjustment assistance. Try to help workers who lose jobs because of foreign trade.

consider writing about this in policy memo

Ways to help:

- Training
- subsidies
- healthcare
- job search allowance / relocation allowance.

Another approach: tax reform. e.g. Scheve/Slaughter recommend cutting the payroll taxes of workers who earl less than the national median.

Puzzle: why do countries pursue protectionism instead of a combined policy of free trade + trade adjustment?

Briefly, we'll discuss domestic institutions and trade policy.

In U.S., Congress typically sets U.S. tariffs.

- Constitution empowers Congress to:
  - impose import duties
  - regulate commerce with foreign nations
- president needs 2/3 approval for treaty.

Smoot-Hawley tariffs of 1930 - brought protectionism to highest level in US history.

Reciprocal Trade Agreements Act of 1934:

Authorized the president to make reciprocal tariff reductions without congressional arppoval. Democrats largely supported this, while Republicans argued for protectionism.

Puzzle: why did Congress delegate to the President?

12.16 Lecture 20: 8-12-19

Structure of exam:

- 20 true false (10 min)
- 6 of 8 ID's (10 min each)
- 1 essay (30 min each)

Review session is Thursday @ 6:30pm.

Recall the course focuses on four international problems that arise in the context of anarchy.

- Unit 1: War
- Unit 2: Environment
- Unit 3: Trade
- Unit 4: Poverty

Puzzle of foreign aid.

- No world government redistributes income
- But many governments voluntarily send aid. Why is this?

Reasons countries might give foreign aid: stability, morality, reduce human rights.

Measuring government aid. We will talk about ODA (official development assistance).

Before 1945 countries did not give ODA. Instead, rich countries engaged with poor through trade, colonialism, conquest.

ODA emerged after World War II.

- In 1948, the US launched the MArshall Plan to rebuild economics of Western Europe.
- 18 countries received aid. More than half went to U.K., France, West Germany.
- No aid to Spain (Franco) or Eastern Bloc (Soviets).

By the 1960s, most rich democracies had their own aid programs. Since the 1960s, we can look at a graph of ODA in billions of 2015 US dollars. As of last year, US gave about $120B in aid.

ODA may be given:

- Directly from the donor government to a recipient country (bilateral aid)
- Indirectly via an international organization (multilateral aid).

Many international organizations channel aid:

- UN (UNICEF, IFA, WHO, UNHCR)
- World Bank, etc.

Some countries give most of their aud multilaterally.

Sometimes, aid can be tied:

- Tied aid must be used to buy oods / services from the donor country.
- United aid does not have this requirement.

About 40% of US aid tends to be tied aid.

Motives for giving aid. Could be two main categories of reasons:

- Altruistic (give for humanitarian reasons)
- Egoistic (give for selfish reasons)

To infer leaders, we could study what leaders say, e.g.

- Speeches
- Interviews
- Memoirs
- Diaries
- Letters

But to get more insight, study behavior, e.g. whether they are egoistic / selfish.

If donors were egoistic, they would:

- Give little aid
- Favor "important" countries
- Deliver aid bilaterally
- Tie their aid

On the other hand, if the donors were altruistic, they would:

- Have large aid budgets
- Help less important countries
- Deliver aid multilaterally
- Not tie their aid.

Let's evaluate US aid according to these criteria (USAID).

- Federal budget is $ 3.8T. About 1% of international budget is international affairs (of which a fraction is foreign aid). U.S. gives less than 0.2% of its national income.
- U.S. gives econ aid to strategically important countries. e.g. Afghanistan, Pakistan, Jordan. But overall, Sub-Saharan countries are the biggest recipients of U.S. development assistance.
- U.S. aid responds to political shifts - fell after the Cold War, but then has greatly risen. Rose a lot after 9/11.
- Interesting, U.S. aid changes a lot based on security council membership. Interesting research paper: how much is a seat on the security council worth?
- U.S. gives less than 30% multilaterally.
- Well, it's tied a lot, so not great (the highest percentage out of any donor countries we are talking about).

What about other donors? Longstanding goal: give 0.7% of national income. Most countries have not met this target. Only countries that have met: Denmark, Sweden, Luxembourg, Normay.

From readings: aid goes disproportionately to:

- Former colonies
- Military / political allies
- Culturally similar countries

See ("Who Gets Aid", short article).

But…most generous donors send aid elsewhere. e.g. Sweden.

Generous donors don't always give multilaterally, but they do avoid tying their aid.

If foreign aid reflected concerns about thepoor - countries with the most domestic social spending would also give the most ODA. This is similar to the theme of norm externalization discussed in unit 1.

As Lumdaine predicted - domestic aid is positively related to foreign aid.

12.17 Lecture 21: 8-13-19

Today, we will focus on:

- Aid from international organizations.
- Effectiveness of foreign aid.

Example: international monetary fund. Need to revisit the 19th century to understand it.

From 1870 - 1914, there was the Gold Standard, which is an exchange rate system. The idea was:

- Each country pegged to gold.
- Participants committed to "convertibility."
- One advantage of this standard - it minimized exchange fluctuations (so exchange range would be constant).

The effects of WWI:

- Governments printed money to pay for the war
- The gold standard collapsed (due to inflation).
- It was restored in the 1920s, but…

However, the gold standard collapsed again during the great depression.

Common practices during the depression:

- Competitive devaluation (to make goods cheaper).
- Exchange control.
- Protectionist barriers

Origins of the IMF: established at Bretton Woods, 1944, to prevent the mistakes of the Depression. Some inspiration came from White and Keynes.

The Bretton Woods System:

- Country A → USA → (@ $35 / oz) gold.
- Country B → USA → (@ $35 / oz) gold.
- Country C → USA → (@ $35 / oz) gold.
- That is, all countries peg their exchange rate between its currency and the U.S. dollar.
- The U.S. completes the system by establishing an exchange rate with gold.

The Bretton Woods System:

- Stabilized exchange rates
- Constrained monetary policy.

IMF played the role of emergency "creditor."

- Each member paid a quota to IMF (calibrated based on size of economy)
- IMF lent to countries in need
- But it imposed conditions on borrowers.

IMF played the role of "referree"

- Judged whether pegs should be changed
- Monitored convertibility of currencies

US policies in late 1960s led to

- Inflation, overvalued dollar
- Shrinking trade surplus

Therefore, a country would have two options:

- Deflate (politically unpopular)
- Devaluate (required others to revalue)

Collapse of Bretton Woods

- Nixon chose devaluation
- Countries moved to floating rates
- IMF needed to reinvent itself.

What activities does the IMF undertake today?

- Loans
- Surveillance (bilateral and global)

- Technical assistance (money doctor)

Interestingly, during the recent crisis, the biggest borrowers were in Europe! e.g. Greece, Portugal, Ireland.

Question: are multilateral donors less political?

- They include many countries and diverse interests.
- But, their management structure favors the US:
  - IMF has weighted voting
  - World Bank usually has a U.S. director

About 9 countries form a majority in the IMF (due to weighted voting).

IMF and WB give more to countries that:

- Are U.S. military allies
- Vote with U.S. in the UN
- Serve on the WB governing board
- Rotate onto the Security Council

For reflection: should aid be given bilaterally or multilaterally?

Rest of lecture - talking about effects of foreign aid.

Argument (from Jeff Sachs) - the poor face a dilemma.

- They can't afford to invest.
- Without investments, they can't escape poverty.
- Thus, the poor are stuck in a poverty trap.

What kind of investments could help?

- Human investments (health, nutrition, education)
- Business: machines
- Infrastructure: roads, power, water.
- Natural: e.g. land
- Knowledge: science / tech.

It's argued that aid can provide investment.

- Aid to govts → public investments
- Aid ot familiies → household investments.
- Microfinance → business investments,

thereby promoting growth and reducing poverty. For more on this, see Jeff Sachs' The Development Challenge.

But - aid doesn't always succeed. e.g. Zambia (source Easterly).

Potential reasons:

- Bad inidvidual choices
- Bad government policies
- Bad political institutions

12.18 Lecture 22: 8-14-19

Today, we will talk about ethics and foreign aid.

Trajectory of international poverty over time:

- In 2010, about 2.4B people in developing countries were living on less than $2 per day.
- These estimates were adjusted for "purchasing power parity".

Percentage of world population living in poverty has declined, but total number has stayed the same.

- Percentage ha sdeclined, especially in Asia.
- But: percent has not changed in Africa.

Recent improvement:

- Since 2010, economy has improved
- 900M lives on less than $1.90 a day
- 2.1B lives on less than $3.10 a day.
- Still: the problems remain enormous.

Three ethical perspectives that we can think of:

- Utilitarian
- Libertarian
- Rawlsian

Utilitarian case for aid:

- We ought to prevent suffering and death if we can do so at low moral cost.

According to Singer, it is obligatory to feed victims of famine.

How much ought we give? Keep giving, until we "reach the level of marginal utility." That is, give until you reach marginal utility (when giving more hurts you more than it helps).

Do you think we have an obligation to...

- Prevent suffering if we do so at low moral cost (yes).
- Give until we reach the level of marginal utility (this is complex. I thin it is quite hard to predict the dynamics of "what helps").

Rebuttal?

- Unrealistic: few would give that much
- Too impartial: ignores special obligations
- Ineffective: aid does not reduce poverty.
- Counterproductive: fosters dependency.
- My objection: Marginal utility is complex - there is a time dependence.

A different utilitarian view: "Lifeboat Ethics" (Hardin; wrote about tragedy o fthe commons)

- Limited carrying capacity
- Any more would sink the boat.

The Malthusian dilemma (overshoot and collapse)

Malthus - known for his pessimism about the future of humanity. Major contributions to economic thought: Principles of population.

Key idea: if you give aid, population might go up, and gradually crash after.

Rebuttals:

- Not close to carrying capacity
- Technology might change limits
- You could promote population control (birth control, etc.)
- Why not sacrifice ourselves?

Libertarian perspective, part 1

A government that taxes its citizens to provide foreign aid is coercing its people.

Statement from US libertarian party - "Individuals should not be coerced via taxes into funding a foreign nation or group."

Libertarian perspective, part 2

Individuals have no obligation to give.

- I acquired my property justly
- Anything I give is ‘pure "charity"

Rebuttals:

- Assumes property was acquired jtsly

- Other values may outweigh freedom.
- Need money to exercise liberty.

Rawlsian perspective (developed by Beitz)

- International distribution of resources is morally arbitrary: a matter of brute luck. (e.g. poverty is concentrated in the topics).

In this situation, what would be just?

- Beitz imagines international original position (veil of ignorance)

Review of unit 4:

- Aid from govts.
- Aid from IO iEffects of aid
- Ethics of aid.

Tomz - undergrad research program.

12.19 Lecture 23: 8-15-19

Today: talking about Ethics for aid. Interestingly, citiziens think foreign aid is a high fraction of the budget, but in reality, it is not very high (<1%).

Question: How do we think about Trump's proposed cuts in the context of a few different ethical frameworks?

Two types of frameworks:

- Consequentialist:
  - Singer (utilitarian). Marginal utility - should give until you reach the threshold of marginal utility.
  - Hardin: lifeboat ethics. Shouldn't have foreign aid, because there is a limited carrying capacity (lifeboat ethics).
- Non-consequentialist
  - Rawlsian. Veil of ignorance - probably pro aid. Distributive justice. Prefer policy that supports the least advantaged. Caveat - if aid is going to not least advantaged
  - Libertarian: Libertarian would agree with cuts. They believe individuals should decide where the money you own should go.

Logistics:

- Review session (6:30 - ?)
- Extra OH (Sat 12pm - 2pm CoHo)
- Final Exam (Sat 7pm - 9pm)

12.20 Review for midterm

Outline:

- Memorize statistical terms (Recall sampling vs. non-sampling error).
- Democratic peace theory (structural vs. normative; Russett).
- Bargaining theory (bargaining range, and how it changes based on costs of war).
- Bargaining failure (issue indivisibility, information problems, commitment problems).
- Ethics of warfare (Political realism, Utilitarianism, Christian Just war theory, Islamic war theory).
- Case studies
    - Kargil War (1999, India / Pakistan, only case of two democracies fighting a war).
    - Iraq War (2003, WMDs, information problem; preventive wars).
    - Iran-Iraq wwar (1981, preventive war)
    - 1967 Six Day war (preemptive war)
    - Pearl Harbor (preemptive war).
    - Israel Palestine case study.
        - Israel goals (Jewish state; Democratic state; in the Holy Lands).
        - PLO goals (sovereign state, based on 1948 borders, capital in east Jerusalem).
        - Think about this as example of bargaining failure. (Issue indivibislity, commitment, information problems).
- Reading overivews
    - Farber and Gowa
    - Fortna (selection effect in peacekeeping)
    - Russett (democratic peace theory)
    - Lake (use bargaining theory to examine 2003 Iraq War).

Practice A.

(Raw reading list).

- Hoover and Donovan, The Elements of Social Scientific Thinking
- Russett, "Democratic Norms and Culture?"
- Russett, "The Fact of Democratic Peace"
- Farber and Gowa, "Polities and Peace"
- Frieden, Lake, Schultz, World Politics: Interests, Interactions, Institutions

- Lake, “Two Cheers for Bargaining Theory” (Iraq War)
- Beauchamp, “Everything you need to know about Israel-Palestine”
- Council on Foreign Relations, “Crisis Guide: The Israeli-Palestinian Conflict”
- Mueller, “War has almost ceased to exist: an assessment”
- Pinker, “Violence vnaquished”
- Fazal, “The reports of war’s demise have been exaggerated
- Fearon and Laitin, ”Ethnicity, Insurgency, and Civil War“
- Walter, ”The Critical Barrier to Civil War Settlement“
- Fortna, ”Does Peacekeeping Keep Peace¿‘
- Thucydides, ”The Melian Dialogue“
- Holt, ”Morality, Reduced to Arithmetic“
- Crawford, ”Just War Theory and the U.S. Counterterror War“
- Cornell, ”Jihad: Islam’s Struggle for Truth“

(Read list with summaries).

- Hoover and Donovan, The Elements of Social Scientific Thinking.

  Defines various terms, and experiment design in social science. Regression, Pearson’s r, Sample bias, selection, etc.

- Russett, “Democratic Norms and Culture?”

  Notes that there are structural and normative reasons that democracies are more peaceful overall.

  Three structural reasons:

  - Empowering voters
  - Delay mobilization
  - Convey information

  Normative models:

  - Unconditional externalization
  - Conditional externalization
  - Democratic crusade

- Russett, “The Fact of Democratic Peace”

  (see above). Did analysis on dyads. Russett uses χ^2 and concluded that there is a relationship.

- Farber and Gowa, “Polities and Peace”

More nuanced analysis on dyad-years. Probability of war by regime type and time period, number of dyad-years.

No statistically significant relationship between democracy and war before 1914.

Peace after 1945 (Cold War) coincides with common interests among a large number of states.

Farber and Gowa say that the relationship is spurious! Democracies tend to have common interests, which has led to recent peace. It's not the case that democracies are inherently More likely to be peaceful.

- Frieden, Lake, Schultz, World Politics: Interests, Interactions, Institutions

  Textbook that broadly describes bargaining range / costs of war. Incomplete Information.

- Lake, "Two Cheers for Bargaining Theory" (Iraq War)

  Bargaining theory as one possible explanation of the Iraq War, shows it is an inadequate explanation of the Iraq War. Two player games vs. multiple actors. States don't quite act rationally.

- Beauchamp, "Everything you need to know about Israel-Palestine"

  1923: British Mandate

  1947: UN Partition Plan

  1948: Arab-Israeli War

  1967: Six Day War

  PLO Goals:

  - Sovereign state
  - Based on 1967 borders
  - Capital in east Jerusalem
  - Solution for refugees

  Israeli Goals:

  - Jewish State
  - Democratic state
  - In Holy Land

  Israel wants security (given history of conflicts).

  Why does conflict occur? Problems of:

  - Divisibility (dividing Jerusalem is hard in practice)
  - Information
  - Commitment (history of conflict)

- Council on Foreign Relations, “Crisis Guide: The Israeli-Palestinian Conflict”

  (similar)

- Mueller, “War has almost ceased to exist: an assessment”

  Reasons for peace:

  - Democratic peace
  - Commercial peace
  - Nuclear peace
  - IOs

- Pinker, “Violence vanquished”

  Notes the same idea as before. But also, civil wars are more frequent. Less bad news: civil wars tend to kill fewer people.

- Fazal, “The reports of war’s demise have been exaggerated”

  Advances in battlefield medicinea / preventive care. Military evaluation practices.

- Fearon and Laitin, ”Ethnicity, Insurgency, and Civil War“

  Civil war is due to the steady accumulation of conflicts since the 50s / 60s, rather than a sudden change. Civil wars tend to last a long time.

- Walter, ”The Critical Barrier to Civil War Settlement“

  Civil wars rarely end, because of indivisibility, commitment, various issues. Fill in

- Fortna, ”Does Peacekeeping Keep Peace¿‘

  Selection effect - peacekeeping does help, but we note that the conflicts that end up requiring peacekeeping tend to be selected for.

- Thucydides, ”The Melian Dialogue“

  Athenians crush the Melians.

- Holt, ”Morality, Reduced to Arithmetic“

  Criticizes the “save lives” argument for dropping the bomb.

- Crawford, ”Just War Theory and the U.S. Counterterror War“

  Argues that U.S. Counterterror policy is unethical.

- Cornell, ”Jihad: Islam’s Struggle for Truth“

  Argues that the term Jihad has been misrepresented to be associated with violent acts, when it actually means “striving on behalf of God”

Subtest A.

Readings.

- Russett "Democratic Norms and Culture"
- Farber and Gowa, "Polities and Peace" (example of common interest: opposition ot Soviet Union)
- Lake, "Two Cheers for Bargaining Theory"
- Beauchamp, Israel-Palestine
- Mueller, "War has almost ceased to exist: an assessment"
- Fearon and Laitin, "Ethnicity, Insurgency, and Civil War"
- Walter, "The Critical Barrier to Civil War Settlement"

Ideas / examples.

- Historical case studies
- Ethics of warfare (political realism, utilitarianism, Christian vs. Islamic just war theory)
- Statistical terms
- Preemptive vs. Preventive wars
- Causes of bargaining failure
- Spurious relationship (relationship
- Validity vs. reliability
- Sampling vs. non-sampling error
- Civil war - settlement problems + disarmament dilemma (Mexican standoff). Peacekeeping works well for civil war settlements.
- Jus ad bellum vs. jus in bello.
- Utilitarian counterarguments to nuclear weapons
- Christian tradition vs. Islamic tradition, jus ad bellum vs. jus in bello

Christian tradition:

- Grounds for war (religion)
  - Just case
  - Last resort
  - Chance of success
- Conduct in war (religion)
  - Discrimination
  - Proportional tactics

Islamic tradition:

- Dar al-Islam vs. dar al-harb
- Just cause (propagate Islam)
- Last resort

Conduct in war is similar to Christianity

- Discrmination
- Proportional tactics

Example of bargaining failure: Mexican-American war.

Subtest A, practice solve.

- Russett “Democratic Norms and Culture”

  Structural and normative models:

  S:

  - Delay mobilization
  - Empower voters
  - Convey information

  N:

  - Conditional review
  - Unconditional
  - Democratic crusades

- Farber and Gowa, “Polities and Peace” (example of common interest: opposition ot Soviet Union)

  Argues that democracy / peace is not statistically significant before 1914, and that they can only obtain a correlation after 1945. Argues that this is due to interests, rather than polities. Correlates of war dataset.

- Lake, “Two Cheers for Bargaining Theory”

  Uses bargaining theory to analyze Iraq war. Points out several shortcomings of the model:

  - *n*-player games vs. two player games
  - Assumes rational actors
  - Imperfect information?

- Beauchamp, Israel-Palestine

  Conflict between Israel and Palestine

  - Israel wants to fight PLO over Jerusalem. Holy site for both.

  - Bargaining failure:
    - * Indivisible (Jerusalem for holiness)
    - * Commitment (lack of trust)
    - * Information (lack of information)

- Mueller, “War has almost ceased to exist: an assessment”

  Four key reasons war has almost ceased to exist:

  - Democratic peace
  - Commericial peace
  - Nuclear peace (deterrence).
  - IOs

- Fearon and Laitin, “Ethnicity, Insurgency, and Civil War”

  Note that the counts of civil wars have increased since 1945, but this is not due to an increase in conflict rate, but rather a steady accumulation of ongoing conflicts (since civil wars tend to last longer). The reasons for why are below (see Walter).

- Walter, “The Critical Barrier to Civil War Settlement”

  - Issue indivisibility
  - Disarmament dilemma (commitment problem)
  - Information problem (insurgents who are hard to track down)

Ideas / examples.

- Historical case studies

  - Kargill War (India and Pakistan) - the only war between two democratic states since 1945.
  - Mexican American War (failure of bargaining; outcomes for both parties were much worse than the available deal)
  - Palestine conflict
  - Iran-Iraq war (not sure)
  - Iraq war 2003 (information problem)
  - Hiroshima / Nagasaki (utilitarian argument)
  - Pearl Harbor (preemptive war)

- Ethics of warfare (political realism (Bismarck), utilitarianism, Christian vs. Islamic just war theory)

  Realists say war is reality, don’t be moral.

  Utilitarians: greatest good for greatest number (Mill).

- Preemptive vs. Preventive wars

  Preemptive: fighting for first strike advantage

  Preventive: fighting because it'll prevent later conflict.

- Causes of bargaining failure

  Indivisibility, Information problems, commitment problems,

- Spurious relationship

  A apparent but false relationship between two variables that is due to a confounder.

- Validity vs. reliability

  Does the measure model the concept: valid? Does the measure stay consist across time: reliable?

- Sampling vs. non-sampling error

  Sampling: error due to sample being off just by chance

  Non-sampling error: other forms of error.

- Civil war - settlement problems + disarmament dilemma (Mexican standoff).

  Settlement is hard (Walter), because of divisibility, Information (terrain), and commitment (Mexican standoff).

- Jus ad bellum vs. jus in bello.

  Jus ad bellum: ethics of going to war

  jus in bello: conduct in war

- Utilitarian counterarguments to nuclear weapons

  Are more humane alternatives possible? Is it possible to just demonstrate the power without killing civilians?

  Maybe flesh this out further

- Christian tradition vs. Islamic tradition, jus ad bellum vs. jus in bello

  Christian: Self defense, last resort, chance of success. Jus in bello: discrimination, and proportionality.

  Islam: Just cause and last resort. Jus in bello: discrimination and proportionality.

  dar al-Islam (house of Islam), dar al-harb (the world outside of Islam).

Subtest B.

Readings.

- Russett, “Democratic Peace Theory”
- Lake, “Two Cheers for Bargaining Theory”
- Beauchamp, Israel-Palestine
- Walter, “The Critical Barrier to Civil War Settlement”

Ideas / examples.

- Historical case studies
- Preemptive vs. Preventive wars
- Causes of bargaining failure
- Sampling vs. non-sampling error
- Utilitarian counterarguments to nuclear weapons
- Originator of democratic peace?

Subtest B.

Readings.

- Russett, “Democratic Peace Theory”

  Structural norms (voter empowerment, delay mobilization, and information)

- Lake, “Two Cheers for Bargaining Theory”

  Criticizes the bargaining model. Notes that imperfect information (WMDs), commitment problems (No guarantee of regime change after US and Iraq agree to peace). Also, *n*-player games, and rational actor assumption.

- Beauchamp, Israel-Palestine

  This conflict won’t resolve because of indivisibility of Jerusalem and commitment issues.

- Walter, “The Critical Barrier to Civil War Settlement”

  Information problems (insurgency), commitment problems (disarmament dilemma), and indivisibility.

Ideas / examples.

- Historical case studies

- Preemptive vs. Preventive wars

  Preemptive: first strike advantage.

  Preventive war (U.S. China war). 2003 Iraq War.

- Sampling vs. non-sampling error

  Sampling: when due to pure chance, sample differs from the population. Non-sampling error: when you randomly sample units or measure something incorrectly.

- Utilitarian counterarguments to nuclear weapons

  More humane ways to conduct. Warning shot. Bomb a non-populated area. Show them the capacity of the weapon. Invasion may not have been that deadly.

- Originator of democratic peace? Kant.

Subtest C.

- Historical case studies
  - Kargil War (
  - Iraq War (2003)
  - Iran-Iraq war
  - 1967 Six Day war
  - Pearl Hearbor
- Israel vs. Palestine
- preemptive vs. preventive.

Subtest C - practice solve.

- Historical case studies
  - Kargil War (first example of a democracy-democracy conflict, India vs. Pakistan).
  - Iraq War (2003). Information problem, preventive war. Crawford war on terror reading.
  - Iran-Iraq war (preventive war, Iraq wanted to invade Iran following the Iranian revolution, but failed).
  - 1967 Six Day war (Preemptive war)
  - Pearl Hearbor (preemptive war)
- Israel vs. Palestine

  Roughly: Jerusalem hard to divide, Israel and Palestine have commitment problems.
- preemptive vs. preventive.

  Preventive: war that occurs in the context of a large power shift. Preemptive: war in which first strike advantage is critical.

Subtest C - practice solve 2.

- Iran Iraq war: example of preventive war.
- Iraq war (2003): preventive war.
- Six day war (1967), Pearl Harbor (example of pre-emptive war).

Subtest D.

Describe the history of the Israel-Palestine conflict.

Israel declared a state in 1948, Arab-Israeli war happened. Israel gained control of 78% of territory.

PLO goals:

- Sovereign state
- Based on 1967 borders
- Wants capital in East Jerusalem
- Solution for refugees.

Israeli Goals

- Jewish state
- Democratic state
- In Holy Land
- Security.
- West bank control by Palestinian authority
- Jerusalem, home to holy sites, Islam / Jewish.
- Gaza strip controlled by Hamas.

Subtest D.

Describe the history of the Israeli Palestine conflict.

Arab-Israel war happened in 1967.

PLO goals:

- Sovereign state
- 1967 Borders
- Wants capital in East jerusalem

Israeli Goals

- Jewish state
- Democratic state
- Holy land
- Security

West bank controlled by Palestinian authority Jerusalem, holy sites. Gaza strip: controlled by Hamas.

12.20.1 Hoover, 11-35

- "by tinkering with the meanings of concepts, one can play with the foundations of human understanding and social control."
- Variable: a name for something that is thought to influence a particular state of being in something else.

12.20.2 Hoover, 38 - 46

Theory - "a set of related propositions that attempt to explain, and sometimes to predict, a set of events"

Model - "an implication of greater order and system in a theory"

Paradigm - "larger frame of understanding, shared by a wider community of scientists"

Laws / axioms (there are few in social science).

Induction: building theory through the accumulation and summation of a variety of inquiries.

Deduction: using the logic of a theory

12.20.3 Russett, Democratic Norms and Culture

- Democracies are not always peaceful. It ignores the rally around the flag effect.
- "When a player employing a conditionally cooperative strategy like tot-for-tat is confronted by someone playing a consistently noncooperative strategy, noncooperation dominates."

The cultural / normative model.

- Violent conflicts between democracies will be rare.
- Violent conflicts between nondemocracies, and between democracies and nondemocracies will be more frequent.

12.20.4 Farber and Gowa

- No statistically significant relations between democracy and war before 1914.
- Only after 1945 that the probability of war is significantly lower between democracies than between members of other pairs of states.
- Correlates of war dataset.

12.20.5 Russett 1993 - Factor of Democratic Peace

- Democratically organized

12.21 Section 1: 6-27-19

Writing polisci.

- Response paper: 2 total, 1 before July 12.
- 2 p. 2x-spaced paper / email Iris by Wed 7pm.

Possible options for a response paper:

- Critique the reading.
- Propose alternate explanation to describe the phenomenon.
- Arbitrate between different arguments.
- Apply to a historical / current case.

Russett (1993). His main puzzle is:

Q. What is the relationship between democracy and war?

It's a conditional normative mechanism.

- Democratic states have peaceful norms of conflict resolution.
- Autocrats have violent norms of conflict resolution.
- Democratic states believe autocratic states are untrustworthy.
- Democratic state will pursue violent norms against autocrats.

12.22 Section 2: 7-3-19

Data analysis assignment: due July 16.

Recall that the Iraq War was a conflict bwetween US and Iraq, 2003 - 2011. Fought over the belligerent nature of the regime under Saddam Hussein. Also, the US demanded Iraq to shut down WMD program, but they didn't. But it turned out that WMD program actually didn't exist.

Why do wars occur? 2 Explanations.

1. Behavioral theories of war.
    - Assumption: Leaders fallible to cognitive biases.
    - Leaders misinterpret information aka perceive an action incorrectly
    - Leaders are irrationally overconfident about capabilities.
2. Bargaining model of war.
    (a) Assumption: States are rationalm try to maximize expected payoffs.

    (b) If war is costly: there always should be some deal that is less costly than war itself.

    (c) Expected payoff of fighting:

    $$\text{(probability of fighting)} \times \text{(payoff if win)} + \text{(probability of losing)} \times \text{(payoff if lose)} - \text{cost of fighting}$$

    There are some limitations of this model; if the resource is indivisible, then you won't be able to reason about the expected value.

Recall the class example of bargaining:

- Two players, Iris and Zuhad, can split \$100.
- If they fight, winner takes all (payoff = 100).
- Iris pays \$20 in medical bills, aka cost of fighting = 20; Zuhad pays \$40.
- Players have an equal chance of winning pr(win) = 0.5.

Expected value for Iris:

- $0.5 \times 100 + 0.5 \times 0 - 20 = 30$

This means that Iris prefers any deal that gives her at least 30.

Expected value for Zuhad:

- $0.5 \times 100 + 0.5 \times 0 - 40 = 10$

So we get a bargaining range between $[30, 90]$ that should be acceptable to both parties.

If bargaining range still exists, why do wars happen?

1. Issue indivisibility (e.g. Israel-Palestine Conflict)
2. Information problem
    - Poker game problem; people bluff and misrepresent their capabilities
    - e.g. Berlin crisis

3. Commitment problems
    - Stanford Honor Code (aka can't trust people to do what you want them to do)
    - Example: Iran Nuclear Deal

Bargaining Failures.

- Information problem: Hussein had incentives to conceal information to US re: weapons program.
- Commitment problem: US-Iraq agreement was reached, no guarantees US wouldn't use power to induce regime change.

Behavioral.

- States not unitary actors - domestic political actors.
- Self-delusions and Bush's gut feeling about Hussein.
- Inability to estimate costs of war
- Cognitive biases, self-delusions, failure to update beliefs with new facts.

12.23 Section 3: 7-11-19

Agenda.

- Data analysis tips.
- Recap: end interstate war?
- Application: GP Rivarly.

Passed out review sheet. Recall that the response paper must be turned in before next Wednesday.

The structure of the data analysis:

- Intro: Preview your results
- Connecting logic
    - Reasons / evidence (e.g. empirical support) ⟶ Identify the observable implications of connecting logic.
    - Analyze hidden assumptions.
    - Note, since all the data has to do with public opinion data, hypothesis should probably include the word belief.
- Results
    - Look at Marginal / Conditional probabilities.
    - χ^2 test.
- Analysis
    - What do the results mean?

  - Sampling error (is the sample representative of the world's population distribution?)
  - Non-sampling error (problems that could occur by chance; e.g. differences between Stanford students and broader population; question wording; could prime the respondent to answer in a certain way).
  - Example of question wording: asking whether we want intervention in Iraq could be interpreted as asking whether we want to intervene in 2003 Iraq; vs. asking whether we want to intervene in ISIS.
- Conclusion.
  - Found some difference / no difference in the subpopulations associated with IV vs. not.

Recall that there are four reasons we've seen decline in interstate war:

- Democratic peace
- Commercial peace (increase in globalization)
- Nuclear peace
  - e.g. India / Pakistan; even though there
- IOs (International Organizations)

However, recently - people argue that we have seen a return to great power rivaly.

Two possible adversaries:

- Concerns about China
  - (e.g. US and China are currently in the midst of a brewing trade war)
  - Concerns that something will happen in Taiwan; US is obligated to defend Taiwan (see Taiwan American relations Act, etc.)
- Concerns about Russia.
  - Russia withdrew from a nuclear arms agreement.
  - Russia embarked on a nuclear modernization program (would increase the size of the arsenal, and the destructive power of the weapons).
  - Concerns among many that more nuclear weapons don't make the world more stable, but they increase the likelihood of accidents.

Question. What are the odds?

- What are the odds of a US China war; vs. what are the odds of a US Russia war?

Factors for China.

- Nuclear peace
- Commerical peace
- Democratic vs. nondemocratic peace

- Brinkmapship / accident?

Important factor: Freedom of Navigation Operations; complicated maritime law.

US economy is currently larger than China; but China's economy will be larger in 5 years. So it may make sense to wage war earlier (preventive war).

If China / US got into war, it's possible that other countries would be involved.

US is agresssively building the Navy because of vulnerabilities in the Pacific.

The Thuycidides trap (On the fear of a large nation overpowering another one leading to an increased probability of war).

Factors for Russia.

- Nuclear peace
- Historically, US has had more historical conflict with Russia.
- How Russia fights war:
  - Insurgent fighting (as opposed to regular fighting on a battlefield). US has less infrastructure to fight an insurgent war.

Consider survey data, what is the threat of Russia? We can analyze survey data.

- IV: Partisanship.
- DV: Perceived threat of Russia.

If we have a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

where the entries are frequencies representing

- a: Republican; not threat.
- b: Republican; threat.
- c: Democrat; not threat.
- d: Democrat; threat.

We can compute various quantities:

$$\begin{aligned} Pr(\text{Rus Threat}) &= \frac{b+d}{a+b+c+d} \\ Pr(\text{Threat} \mid \text{Republican}) &= \frac{b}{a+b} \\ Pr(\text{Threat} \mid \text{Democrat}) &= \frac{d}{c+d}. \end{aligned}$$

Given the data, we have

$$\begin{aligned}
Pr(\text{Rus Threat}) &= \frac{238+113}{32+72+238+113} = 0.77 \\
Pr(\text{Threat} \mid \text{Democrat}) &= \frac{238}{32+238} = 0.88 \\
Pr(\text{Threat} \mid \text{Republican}) &= \frac{113}{72+113} = 0.61.
\end{aligned}$$

Even without doing a χ^2 test, we can see that there's a large divergence in opinion.

If we do the test, we find that the p value is less than 0.5.

Do we think that the poll has sampling error?

Representatives

- English only
- Only 869 pepole
- Online

Question wording:

- Scale
- "Imminent"
- "Threat"

Questi

12.24 Section 4: 7-18-19

Midterm. If you say things that are wrong; they will take off points.

Should the US use drones? Pros and cons.

Pros of drones:

- Drone strikes make US safer by decimating terrorists.

  upwards of 3500 militants killed.
- Drones kill fewer civilians. PCT of fatatliesi
- Drones make US military personnel safer.

  Less room for human error.
- Drone strikes are cheaper than engaging in ground / manned aerial combat.

  $5B$ allocated for Drones, only about 1% of the entire military budget.
- Drone strikes are legal under internationa l law.

  Aritcle 51, sel defense. anticipitaroy self defense.

- Drone strikes ar elegal under US law.
- Drones limit the scope / scale of military action.
- Subject to strict review process.
- Cannot risk falling behind rest of the world.
- Drone pilots have a lower risk of PTSD than pilots of manned aircraft.
- Majority of Americans support drone strikes.

Cons:

- Drone strikes create more terrorists than they kill.
- Drone strikes target individuals who may not be terrorists / combatans.
- Kill large numbers of civilians / traumatize populations.
- Kill low value targets
- Violate internation law.
- Secretive, prevent citizens from holding accountable
- Violate sovereignty of other countries (without permission).
- Allow US to be emotionally disconnected from horrors of war. May propagate war.
- US drone strikes give cover for others to engage in human rights abuses.
- Extremely unpopular in affected countries.
- Drone operators have stress.

Team antidrone.

- Lot of accidents happen.
- Drone fails to target military leaders In the past. Scuceeded in killing 14 military leaders of terorrism, but more civilians.
- Anti-US sentiment.
- Killing military leaders is not helpful in solving the true issue.

According to a July 18, 2013 survey by Pew Research, 61% of Americans supported drone strikes in Pakistan, Yemen, and Somalia.

Core arguments.

- Accidents. Crawford mention that accidents can make it possible to sidestep responsibility. Big challenge - how do you develop metrics that quantify the impact of drone strikes (got to measure civilian attitudes / radicalization).
- Drones reduce the risk for PTSD. Anti-drone mentioned the psychological effect impacts civilians at well (terrorizing relative to other military options).

- Proportionality. Effective in getting outcomes we want, but comes at a larger cost. e.g. Haven't had an attack from Al Qaeda since 2001, ISIS hasn't really done anything in a while.
- Thinking in analogies (compare to WW2). Reason that civilian casualities were very high is that firebombing Dresden was a good way to attack Nazis.

12.25 Section 5: 7-25-19

Policy memo.

Topic should be:

- Related to trade / environment / poverty.
- US-focus.
- Feasibility - needs to be tractable and you should be discuss why. Should be sufficiently limited to that it's possible to find research and support.

Research:

- Scholarly articles, books, think-tanks, NGOs. (e.g. see Heritage / Cato / Brettonwoods).

Writing:

- Emphasize broadcasting in advance, and be brief.

Pieces:

- Exec summary
- Problem
- Solution
- Feasibility
- Conclusion

12.26 Section 7: 8-8-19

Announcement:

- 2nd response paper due Wed 8/14 by 7pm
- policy memo due 8/14
- final exam review 8/15 (6:30 - 7:20)
- final exam, Saturday 8/17 (7pm - 10pm)

Agenda:

- Protectionism / tariffs
- US-China trade war

- Ethics of trade

Recap: free trade would help everyone, but states can't credibly commit to not be protectionists. Can model this with a prisoner's dilemma.

Two broad models:

- Stolper-Samuelson: trade -> class conflict. (asusmptions: labor and capital are highly mobile).
- Ricardo-Viner: trade -> conflicts between industries (rather than classes). (assumptions: some factors are fully mobile).

Why protect?

- National interest
- Domestic politics (SS and RV)
- International politics -> optimal tariffs

Four broad ways to solve the commitment problem:

- Change costs / benefits
- Third-party / outside actor
- Strategies of reciprocity (tit for tat or grim trigger). Alexrod reading talks about why tit for tat is better.
- Domestic pressure. (Valhontra reading).

US-China trade war. It starts with a commitment problem, which states that:

- US and China can't commit to FT because tariffs can shift the terms of trade.

  $$\text{ToT} = \frac{\text{price of exports}}{\text{price of imports}}.$$

  Example: consider the ratio $\frac{\text{price of soybeans}}{\text{price of iPhones}}$. If numerator increases while denominator decreases, trade becomes more favorable to the U.S.

Last week: trade talks in Shanghai.

Why is there a US-China trade war? Five underlying factors.

- Trade deficit (US is importing a lot more than they are exporting; roughly US imports $500B more than it exports).
- China currency manipulation / fair market (WTO accession / manipulation). China opened up its economy in 1979. Transformation is huge - almost more than US from 1800 to now.

  One condition to join WTO is that you need a free and open market. But U.S. says this is not the case.
- IP theft: China takes IP property -> joint venture. In order for a foreign company to enter into Chinese market, they have to form a joint venture (and join with local players).

  CFIUS: needs to review to make sure that China doesn't acquire / merge investment.

- 2025 Initiative (Made in China). Goal is to invest in AI / advanced technologies; creates a national security risk. (Huawei controversy).
- Great power rivalry: Power transition war? China may surpass the U.S. no later than 2050.

Solution	Who?	How?
Change costs	China	nuclear options w.r.t. U.S. debt. Would drive up interest rates, kill US
Change benefits	US	Voters love free trade. Trump might change his mind
Reciprocity	Both US and China	committing to de-escalate and reduce tariffs
International organizations	WTO	Arbitrate / initiate dispute settlement mechanism

12.27 Review for final

Prisoner's dilemma in int'l politics.

- Prisoner's dilemma can be a useful representation of many types of problems

Reeated prisoner's dilemma

- States often engage in these interactions repeatedly
- Using strategies of reciprocity (e.g. Grim trigger, tit-for-tat), states can sustain cooperation if the long term benefits from cooperating outweigh the short term benefit a state can get from defecting. (Recall Axelrod's tournament)

Strategies of reciprocity are most likely to work when...

- Players value the future
- Reward for defecting is small
- etc.

Also - recall shadow of the future.

Recall trade: key terms (see the slides).

Engel's Law: as a country gets richer, smaller proportion goes to commodities.

General system of preferences (GSP)

WTO is about nondiscrimination / reciprocity.

Stolper-Samuelson asssumes factors of production are mobile across sectors.

Ricard-Viner does not assume factors of production are mobile across sectors.

12.28 Review for final 2

Need to go over all readings. Everything in the class is fair game. Essay should be 30 minutes.

ID:

- Explain issue

- Contextualize
- Provide example
- Cite readings
- Explain significance for IR

Game theory:

- Prisoner's dilemma (defect); example of a commitment problem.
- If opponent is cooperating, you should defect (i.e. defecting is a dominant strategy; so Defect; defect is the Nash equilibrium).
- Repeated prisoner's dilemma. Need to use strategies of reciprocity (e.g. Axelrod's tournament).
- Prisoner's dilemma in int'l politics.
  - Classic application of PD: tragedy of commons, optimal tariff argument.
  - The barrier to settlement of civil war
  - Commitment problems generally as a cause of war.

Repeated PD:

- When games are iterated, you need to be cooperative.

When does reciprocity not work:

- When players value the future
- Reward for defecting is small
- Punishment for cheating is long and severe.

International institutions

- Help states use reciprocity to sustain cooperation
  - Set clear expectations (e.g. what counts as "defecting"?)
  - monitoring behavior (need to know whether defection has occurred).
  - Coordinating punishments (helps avoid echo chambers)
- Examples: GATT / WTO, Paris Climate Agreement, etc.

Key ideas in trade:

- Free trade
- Protectionism
- Autarky (when a state is sustainaible without international trade)
- Comparative advantage (when individual can produce an activity more efficiently than other activity).

- Absolute advantage (when individual produces an activity more efficiently than another group)
- Economies of scale (when scale allows you to do things more efficiently; e.g. bulk discounts).
- Ad valorem (tax is % of good's value) vs specific tariff (fixed amount per unit)
- Nontariff barriers
  - Quotas (limits on imports)
  - Product standards (need to regulate sanitary / evnrionmental / medical quality of goods).
- Infant industries
  - New industries - want to encourage them ideally
- Strategic trade policy
  - Sometimes good to be protectionist, sometimes good to be freer.
- LDCs and Industrialization.
  - Least developed countries.
  - Industrialization is good - because of technology, capital, automation, etc.
- Stolper-Samuelson.
  - Factors of production (labor / capital) are highly mobile.
  - In a labor abundant country, labor-intensive industries will grow.
  - Result: trade -> class conflict.
  - Summary: Factors mobile -> Class conflict.
- Ricardo Viner.
  - Factor of productions are not mobile
  - Some factors are fully mobile. Then leads to industry conflict.
  - S: Factors mobile -> Industry conflict.
- Trade adjustment assistance
  - Reduce damaging impact of imports.
- Tax reform
  - Generally, reform taxes.
- Smoot-Hawley 1930
  - Law which implemented protectionist trade policies in the US.
  - Highest level in US history at time.
- Reciprocal trade agreements, 1934

  - Make reciprocal tariff reductions without congressional approval.
- optimal tariff
  - a country that is a large importer of a particular commodity can shift the economic burden of an import tariff from domestic consumers to foreign (Chicago PR)
- GATT: agreement to promot trade by reducing tariffs / quotas.
- Nonscirmination: human right s/ trade?
- Most-favored nation: granted the most favorable trading terms available by another country
- preferential trade agreements
- generalized system of preferences (GSP): economic growth bc. duty free.
- Escape calsuses: allows escape trade?
- WTO dispute process
  - once a complaint has been filed in WTO (multilateral dispute resolution).
  - If informal consultations fail, panel is formed automatically.
  - Panel findings
- Trragedy of commeons.
- Kyoto - failed
- Paris - reasonable, prevent global temps from rise >2 deg C.
- Four broad solutions: CRDT (coercion, reciprocity, domestic, tech).
- Foreign aid
  - IMF: Vreeland notes that no one entity controls IMF.
  - Aid can fail - bc bad individual choices.
  - Bad govt. / political inst.
- Ethics
  - Realism
  - Utilitarian
  - Just war
- Ethics and environ
  - Utilitarian
  - Corrective
  - Egaliatian
  - Shared respo.

- Trade
    - Utilitarian
    - Rawlsian (original position) difference principle.
    - Kapstein - RAwls
- Libertarian - maximize freedom.
- Hardin - lifeboat ethics

Reading review:

- Easterly: utopian nightmare
    - Helping prolonbgs the true nightmare. Have not gotten around to the most needy countries.
- Singer: utilitarian
- Hardin - ethics
    - Lifeboat ethics relates to population dynamics - since there's a carrying capacity, you'll overshoot.
- Armstrong: distributive justice.
    - Beitz builds on Rawls
- Sweatshops, NY Times.
- Kapstein: Rawls, globalization.
- McGee: rights. Trade shouldn't violate property, contract, or association rights.
- Beitz: international DP (inequalities should be arranged internationally so they benefit the least advantaged).
- Goodin - shared responsibilities.
- Just war
    - Christian / Islamic
    - Crawford / Cornell
    - Holt: Morality reduced to arithmeitc.