

Stanford compendium in mathematics and computer science

ADITHYA C. GANESH

December 21, 2018

Abstract

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

Contents

1	Independent + CS229: Statistical Learning	5
1.1	Matrix cookbook	6
1.2	Large-scale distributed training	6
1.3	Hyperparameter optimization	6
1.3.1	Population-based training	6
1.3.2	ENAS	8
1.4	Distillation	8
1.5	ConvNet architectures	8
1.5.1	Recognition	8
1.5.2	Detection	8
1.5.3	Segmentation	10
1.5.4	WaveNet	10
1.5.5	Self-attention networks	10
1.6	PyTorch	10
1.7	Backpropagation: a graph theory perspective	10
1.7.1	Combinatorial explosion	10
1.8	Functional programming \cap Neural networks	11
2	CS236: Deep Generative Models	12
2.1	Variational Autoencoder	12
2.1.1	Reparametrization trick	13
2.1.2	GMVAE	13
2.1.3	IWVAE	14
2.1.4	Questions	14
3	CS168: The Modern Algorithmic Toolbox	15
3.1	Lecture 3	16
3.1.1	Similarity Metrics	16
3.2	Lecture 4	18
3.2.1	Dimensionality reduction for Jacard similarity	19
3.2.2	Dimensionality reduction for Euclidean distance	20
3.3	Lecture 5: Generalization	21
3.4	Lecture 6: Regularization	23
3.5	Lecture 10	25
3.6	Things to review	27
3.7	Key ideas	27
4	EE376A and mathematics directed reading: Information Theory	28
4.1	The Source Coding Theorem	1
4.1.1	A basic example	1
4.2	Maximum Entropy Principle	4
4.3	Core ideas in information theory	4

4.4	Dyadic U and symbol counting	6
4.5	Optimality of Huffman Codes	6
4.6	Channel Capacity	7
4.6.1	Information of Continuous Random Variables	9
4.6.2	Exercises	10
4.6.3	Examples	11
4.7	Constraints and communication theory	13
4.7.1	Joint Asymptotic Equipartition Principle	15
4.8	Channel Capacity Theorem	16
4.8.1	Joint AEP	18
4.8.2	Relation of AEP to Communication Problem	19
4.9	Channel Coding Theorem; Converse Part	21
4.10	Lossy Compression & Rate Distortion Theory	23
4.10.1	Lossy compression problem setting	23
4.10.2	Qualitative analysis of $R(D)$	24
4.10.3	Examples	25
4.11	Method of Types	26
4.12	Strong, Conditional, and Joint Typicality	28
5	CS109: Probability	29
5.1	Key Topics	29
5.2	Theory	35
5.3	Problems to Review	35
5.3.1	Problem Set 1	35
5.3.2	Problem Set 2	35
5.3.3	Problem Set 3	35
5.3.4	Problem Set 4	35
5.3.5	Problem Set 5	35
5.3.6	Problem Set 6	35
5.4	Practice Problems 3-20-17	36
5.5	Practice Final 3-21-17	36
6	MATH113: Matrix Theory	37
6.1	Lecture 1	38
6.1.1	Course logistics	38
6.1.2	Introduction	38
6.1.3	Fields	38
6.2	Lecture 2	40
6.2.1	What it means to read proofs	40
6.2.2	Vector spaces	40
6.3	Lecture 4	43
6.4	Notes on 3.F: Duality	44
6.5	Key Ideas	44
7	MATH120: Group Theory	45
7.1	Lecture 4:	46
7.1.1	Homomorphisms	46
7.1.2	Approach 2: Bottom-up homomorphisms	47
7.2	Lecture 5	49
7.2.1	Order	49
7.3	Lecture 6	50
7.4	Lecture 7	52
7.5	Lecture 8	54
7.5.1	More on conjugation	54

7.5.2	Group actions	55
7.6	Lecture 9	56
7.7	Lecture 11	57
7.7.1	Conjugation / conjugacy classes	57
7.8	Lecture 12: Automorphisms and Sylow's Theorems	58
7.9	Lecture 14	60
7.10	Lecture 15	61
7.11	Lecture 18	63
7.12	Lecture 20	64
7.13	Lecture 23	67
7.14	Lecture 29	67
7.15	Notes on Group Actions	68
7.16	Notes on Irreducibility	69
7.17	Notes on Free Groups	69
7.18	Key Ideas	69
7.18.1	Definitions	69
7.18.2	Propositions and Theorems	70
7.18.3	Examples	70
7.18.4	Ideas	70
7.19	Things to review	70
8	MATH116: Complex Analysis	71
8.1	9-24-18: Introduction	72
8.2	Differential 1-forms	74
8.3	Complex projective line, or Riemann sphere	74
8.4	Riemann surfaces	74
8.5	Key ideas	74
8.5.1	Basic facts	74
8.5.2	Main results	75
8.6	Midterm review sheet	77
8.6.1	Cauchy-Riemann equations	77
8.6.2	Cauchy integral formula + applications	77
8.6.3	Power series	78
8.6.4	Exponential function and logarithm	78
8.6.5	Meromorphic functions	78
8.6.6	Argument principle and Rouché's theorem	79
8.6.7	Computation of integrals using residues	79
8.6.8	Harmonic functions and harmonic conjugates	79
8.6.9	Elementary conformal mappings	80
8.6.10	Properties of fractional linear transformations	80
9	MATH171: Real Analysis	81
9.1	9-24-18: Everything is a set	82
9.1.1	On sets	82
9.1.2	On functions (and cartesian products)	82
9.1.3	On natural numbers	83
9.2	10-1-18: Suprema and infima	83
9.3	Continuity - 10-19	84
9.4	10-22: Connectedness	86
9.5	10-24: Uniform continuity	88
9.5.1	Review	89
9.6	11-05-18: Integrability and FTCs	90
9.7	Sequences and series of functions	92
9.7.1	Pointwise vs. uniform convergence	92

9.8	Key ideas	93
-----	---------------------	----

Chapter 1

Independent + CS229: Statistical Learning

amsmath amssymb fancyhdr todonotes amsthm amsopn amsfonts mathtools libertine hyperref graphicx
[parfill]parskip [margin=1in]geometry
sign Aut GL Ker im Syl
Notes on statistical learning Adithya Ganesh

Contents

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 multiplied 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:

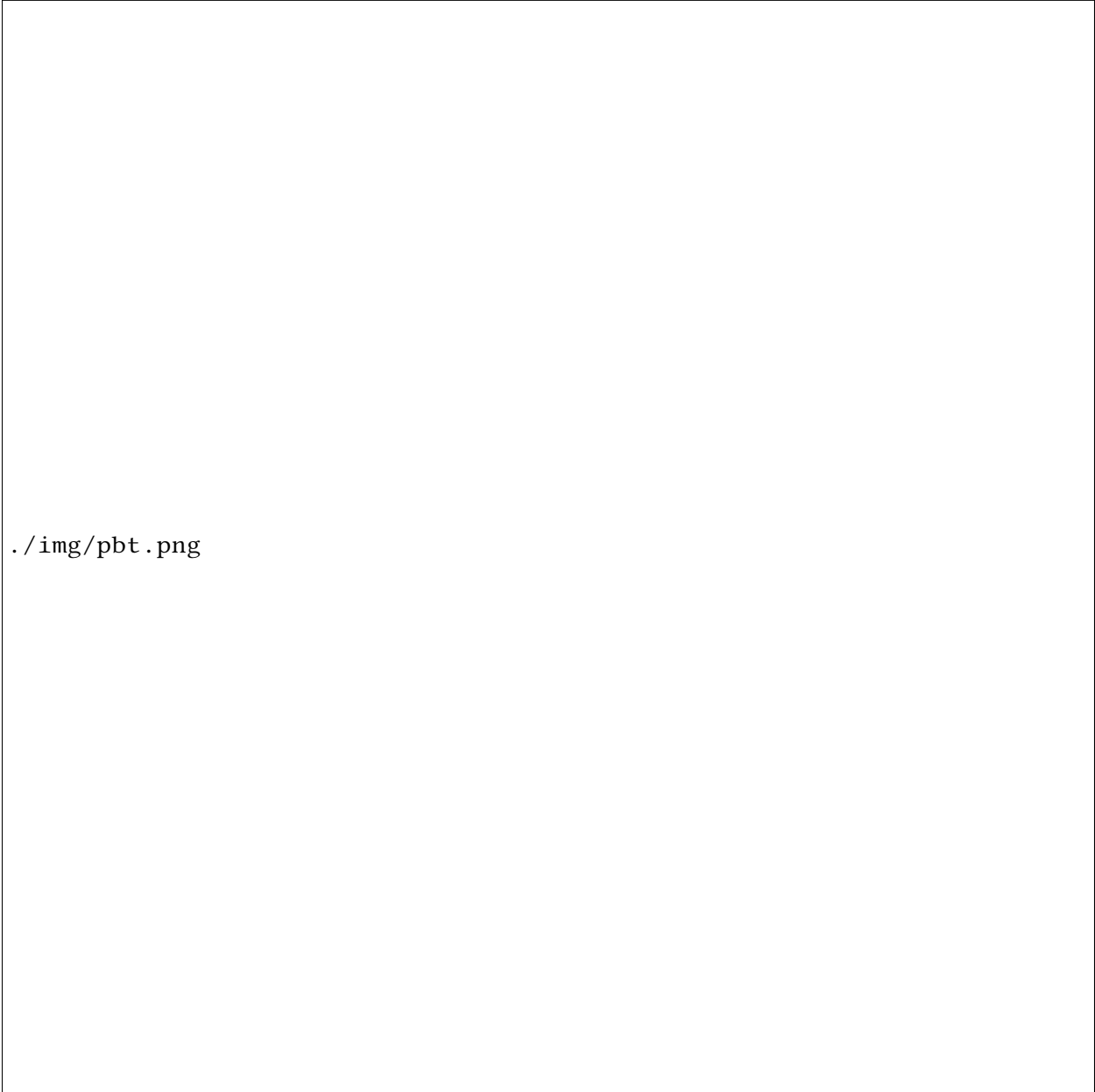
$$\hat{w}_{t+1} = w_t - \hat{\eta} \frac{1}{kn} \sum_{j < k} \sum_{x \in \mathcal{B}_j} \nabla l(x, w_t).$$

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.



`./img/pbt.png`

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 $\pi(\mathbf{m}; \theta)$ and perform SGD on ω to minimize expected loss $\mathbf{m} \sim \pi[\mathcal{L}(\mathbf{m}; \omega)]$. Gradient is computed using Monte Carlo estimate

$$\nabla_{\omega} \mathbf{m} \sim \pi(\mathbf{m}; \theta) [\mathcal{L}(\mathbf{m}; \omega)] \approx \frac{1}{M} \sum_{i=1}^M \nabla_{\omega} \mathcal{L}(\mathbf{m}_i, \omega),$$

where \mathbf{m}_i are sampled from $\pi(\mathbf{m}; \theta)$.

- To train controller parameters θ : fix ω and update θ to maximize the expected reward $\mathbf{m} \sim \pi(\mathbf{m}; \theta) [\mathcal{R}(\mathbf{m}, \omega)]$. 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:

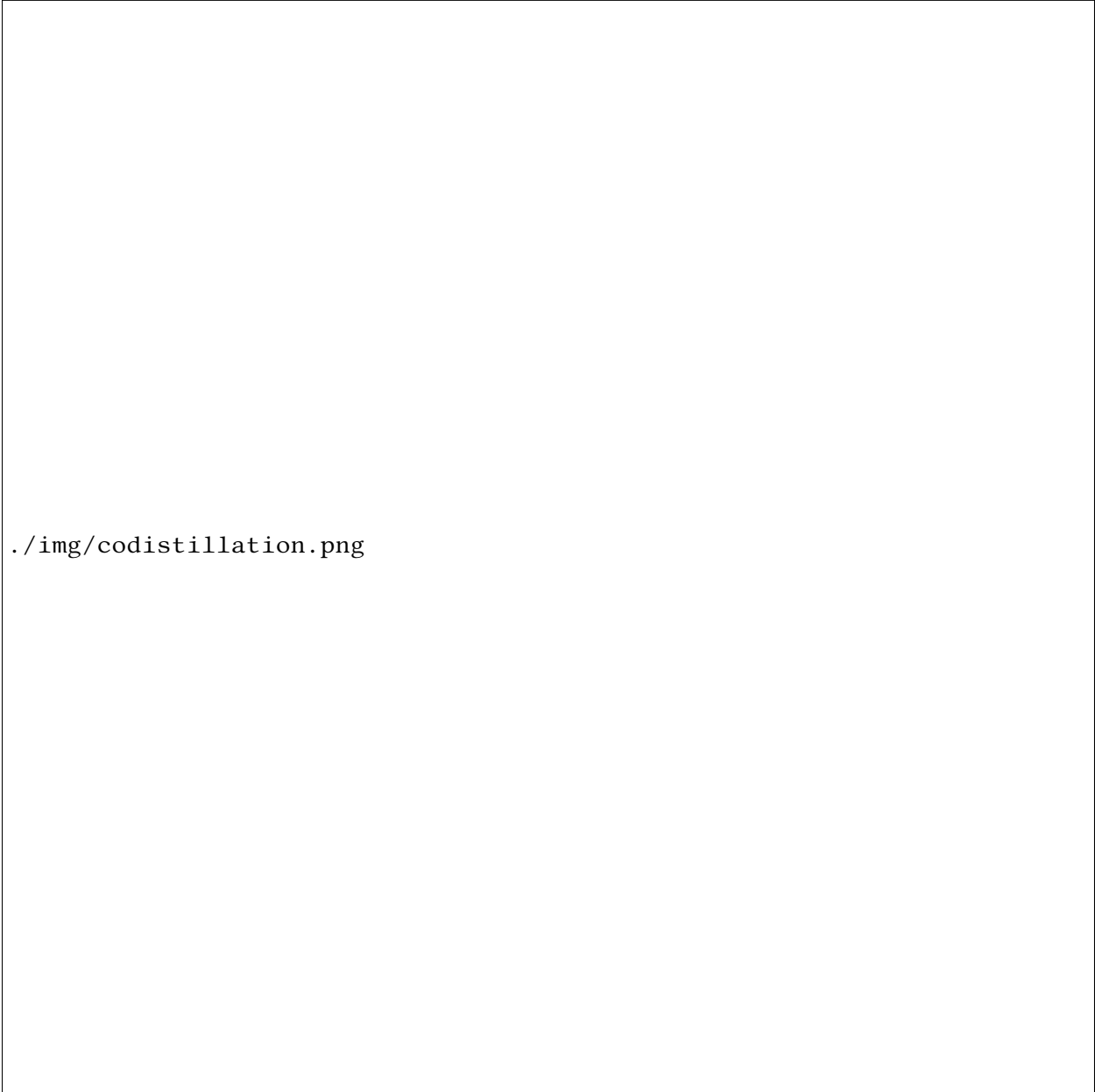
1.5 ConvNet architectures

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



`./img/codistillation.png`

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

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

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

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

- 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 differentiation.” 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.

¹Image source: Chris Olah.

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:

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

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.”

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

Reverse mode differentiation 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 \cap Neural networks

TODO: write up notes on Chris’ article.

Chapter 2

CS236: Deep Generative Models

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 diag Aut GL Ker im Syl argmin

[parfill]parskip [margin=1in]geometry

CS236 - Deep Generative Models Instructor: Stefano Ermon; Aditya Grover; Notes: Adithya Ganesh

2.1 Variational Autoencoder

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

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

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

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0, I) \\ p(\mathbf{x}|\mathbf{z}) &= \text{Ber}(\mathbf{x}|f_{\theta}(\mathbf{z})), \end{aligned}$$

where $f_{\theta}(\mathbf{z})$ 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, $p(\mathbf{z}|\mathbf{x})$).

Indeed, by Bayes' theorem, we can write

$$\begin{aligned}
p(\mathbf{z}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \\
&= \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) d\mathbf{z}}.
\end{aligned}$$

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

We write

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), (\sigma_\phi^2(\mathbf{x}))).$$

Variational inference approximates the posterior with a family of distributions $q_\phi(\mathbf{z}|\mathbf{x})$.

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

The optimal approximate posterior is

$$\begin{aligned}
q_\phi(\mathbf{z}|\mathbf{x}) &=_{\phi} KL(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \\
&=_{\phi} \{\mathbb{E}_q[\log q_\phi(\mathbf{z}|\mathbf{x})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] + \log p(\mathbf{x})\}.
\end{aligned}$$

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

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

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

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

2.1.1 Reparametrization trick

Instead of sampling

$$z \sim \mathcal{N}(\mu, \Sigma),$$

we can sample

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

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,

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

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

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

2.1.3 IWVAE

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

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

where $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$.

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

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

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

2.1.4 Questions

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

Chapter 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]parskip [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)

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

One reason this is useful is that it's 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 $p \rightarrow \infty$, 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 $x \in X$, $\text{part}(X) = \{y \in \mathbb{R}^d \text{ s. t. } D(x, y) = \min_{x' \in X} D(x', y)\}$.

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 diagram look like for L_2 and L_1 ?

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 $x_1, \dots, x_n \in \mathbb{R}^d$, can we find a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ 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 back 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 closer 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 $S = \{x : x \in \mathbb{R}^d\}$.

- 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 is because 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 $\text{dim} = 8$, 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 $x_1, \dots, x_n \in \mathbb{R}^d$. 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 dimensionality.

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

- Pick a uniformly random ordering of the universe U .
- Map set S to $f(S) = \text{“min” element of } S$.

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. \square

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

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

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

$$\begin{aligned} & \left(\frac{\sum_{i=1}^k X_i}{k} \right) \\ &= \frac{1}{k^2} \left(\sum_{i=1}^k X_i \right) \\ &= \frac{1}{k^2} \cdot k(X_1) \geq \frac{1}{k} \end{aligned}$$

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 $r = (r_1, \dots, r_d)$.

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

It turns out that if you pick two angles on the sphere, it doesn't end up being uniform! Similarly, 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 $x_1, \dots, x_n \in \mathbb{R}^d$.
- Known distribution D on \mathbb{R}^d .
- Ground truth label function $f : \mathbb{R}^d \rightarrow \{0, 1\}$.

Problem. Given x_1, \dots, x_n drawn independently from D , and labels $f(x_1), \dots, f(x_n)$, our goal is to output $g : \mathbb{R}^d \rightarrow \{0, 1\}$ 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 $f \in S = \{f_1, f_2, \dots, f_k\}$. That is, f belongs to a set of functions with k elements which is finite.
- 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.

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

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

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

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 approximately correct).

Example. Consider the set of linear classifiers in \mathbb{R}^d . This is defined by a vector $\mathbf{a} = (a_1, a_2, \dots, a_d)$. Then the classifier is just

$$f_{\mathbf{a}}(x) = \left(\sum_{i=1}^d \mathbf{a} \cdot \mathbf{x} \right).$$

Intuition. Note that the vector \mathbf{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 have a set of 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.

Example. For example, we can consider L_2 -regularized least squares. Let $x_1, \dots, x_n \in \mathbb{R}^d$, and $y_1, \dots, y_n \in \mathbb{R}$. In this setting, we want to minimize the following objective:

$$\min_a f(a) = \sum_{i=1}^n (\langle x_i, a \rangle - y_i)^2 + \underbrace{\lambda \|a\|_2^2}_{\text{regularization term}}.$$

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”).

Key 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).

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

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 significance, 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 $n_1 n_2 \cdots n_k$ 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, j th entry. You can refer to a specific element of a k -tensor via A_{i_1, i_2, \dots, 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, \dots, v_k of lengths n_1, \dots, n_k , the tensor product is denoted $v_1 \otimes v_2 \otimes \cdots \otimes v_k$ is the $n_1 \times n_2 \times \cdots \times n_k$ k -tensor A with entry

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

Example. For example, let

$$v_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, v_3 = \begin{pmatrix} 10 \\ 20 \end{pmatrix}.$$

Then $v_1 \otimes v_2 \otimes v_3$ is a $3 \times 2 \times 2$ 3-tensor, that can be thought of as a stack of two 3×2 matrices:

$$M_1 = \begin{pmatrix} -10 & 10 \\ -20 & 20 \\ -30 & 30 \end{pmatrix}, M_2 = \begin{pmatrix} -20 & 20 \\ -40 & 40 \\ -60 & 60 \end{pmatrix}.$$

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, $u_1, \dots, u_r, v_1, \dots, v_r$ and w_1, \dots, 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 $\text{rank}(M - uv^T) = \text{rank}(M) - 1$.
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 $\text{rank}(A - u \otimes v \otimes w) > \text{rank}(A)$.
- For matrices with entries in \mathbb{R} , there is no point in looking for a low-rank decomposition that involves complex numbers, because of $\text{rank}_{\mathbb{R}}(M) = \text{rank}_{\mathbb{C}}(M)$. 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 representation is unique, and can be efficiently recovered.

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

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

Claim: if $\{u_1, \dots, u_k\}, \{v_1, \dots, v_k\}$ 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 $(u_1, \dots, u_k), (v_1, \dots, v_k), (w_1, \dots, w_k)$ 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 $n \times n$ matrices. Let A_x be the weighted sum of these matrices, where the weight given to the i th matrix is x_i . Define A_y analogously.
- Compute the eigen-decompositions of $A_x A_y^{-1} = Q S Q^{-1}$ 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 of the entries of T . The vectors u_1, \dots, u_k are the columns of Q corresponding to nonzero eigenvalues, and the vectors v_1, \dots, 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

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

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-occurrence 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

Chapter 4

EE376A and mathematics directed reading: Information Theory

extsizes

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]parskip [margin=1in]geometry

Information Theory and Statistical Learning Adithya Ganesh

Acknowledgments

Thank you to Yuval Wigderson 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, $\mathcal{A}_X = \{a_1, a_2, \dots, a_i, \dots, a_I\}$ having probabilities $\mathcal{P}_X = \{p_1, p_2, \dots, p_I\}$, 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 $x \bmod 32 \geq 16$?
- Is $x \bmod 16 \geq 8$?
- Is $x \bmod 8 \geq 4$?
- Is $x \bmod 4 \geq 2$?
- Is $x \bmod 2 = 1$.

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 real number. Then

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

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, \dots, h_N , having common mean \bar{h} and common variance σ_h^2 : $x = \frac{1}{N} \sum_{n=1}^N h_n$. Then

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

Theorem (Asymptotic equipartition principle.). For an ensemble of N independent identically distributed (i.i.d.) random variable $X^N \equiv (X_1, X_2, \dots, X_N)$, with N sufficiently large, the outcome $\mathbf{x} = (x_1, x_2, \dots, x_N)$ 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 $N \rightarrow \infty$. 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 $p_i N$ occurrences of the i -th symbol, so that the probability of this "typical" string is roughly

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

so that the information content of a typical string is

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

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 \mathbf{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 have 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 a subset S' having $|S'| \leq 2^{N(H-\beta)}$ and achieving $P(\mathbf{x} \in S') \geq 1 - \delta$ cannot exist.

The probability of the subset S' is

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

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

The 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 $P(\mathbf{x} \notin T_{N\beta})$.

Thus:

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

We can now set $\beta = \frac{\epsilon}{2}$ and N_0 such that $P(\mathbf{x} \in S') < 1 - \delta$, 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 \mathbf{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 $A \rightarrow 00, B \rightarrow 01, C \rightarrow 11$. But too wasteful, since A occurs more frequently.

Approach 2. Better is $A \rightarrow 0, B \rightarrow 01, C \rightarrow 11$. 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:

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

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:

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

Entropy. For any scheme, the value $\bar{L} \geq H(U)$, 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

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

Example 2. Consider a source

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

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 $\mathcal{U} = \{1, 2, \dots, r\}$. Further, suppose that $p(1) \geq p(2) \geq \dots \geq p(r)$ (i.e. they are arranged in descending probabilities).

Let V denote the random variable with $\mathcal{V} = \{1, 2, \dots, r-1\}$ 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. $\bar{l}(1) \leq l(2) \leq \dots \leq l(r-1) \leq l(r) \triangleq l_{max}$ (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 :

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

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.

Examples

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

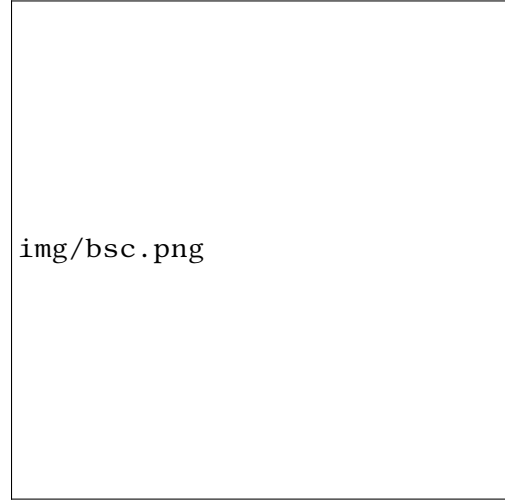
Define alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. A BSC is defined by the PMF:

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

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.

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

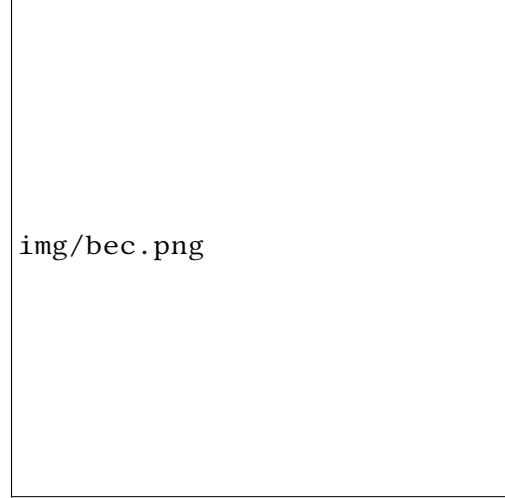
Because only the random noise can't be modeled by conditioning on X , we can simplify the second term:

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

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

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

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.

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

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

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

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 $dy = c dx$. 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 $U \sim \text{Uni}(a, b)$.

- 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:

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

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

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

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. $p \geq E[X_i^2]$.

- 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:

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

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

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:

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

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

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

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

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

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

Therefore, the rate can be bounded by

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

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

$$\begin{aligned} p(x^n) &= \prod_{i=1}^n P_X(x_i), \\ p(y^n) &= \prod_{i=1}^n P_Y(y_i), \\ p(x^n, y^n) &= \prod_{i=1}^n P_{X,Y}(x_i, y_i). \end{aligned}$$

Definition. The set of jointly typical sequences is defined as

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

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

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

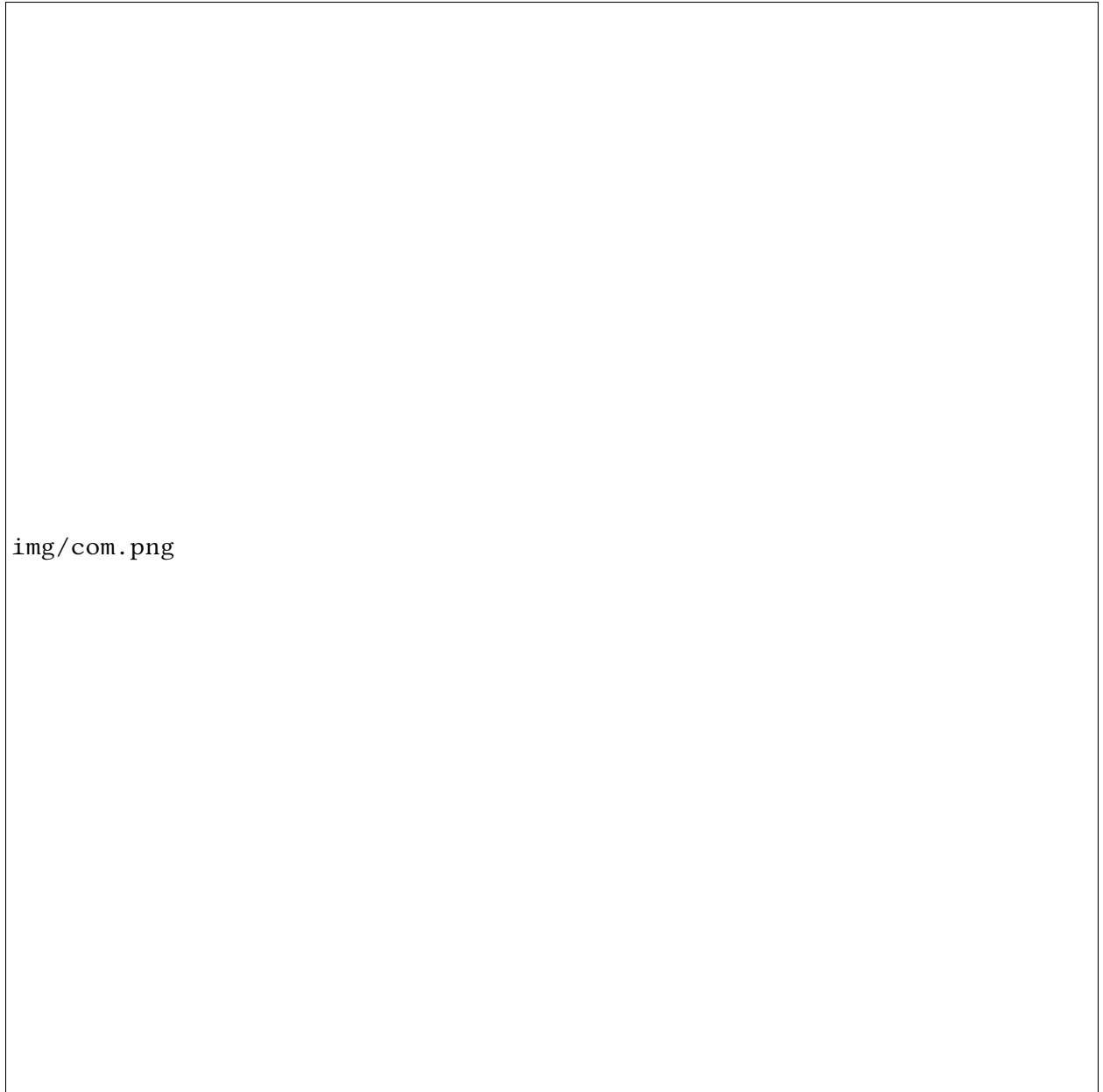
$$(1 - \epsilon)2^{-nI(X, Y) + 3\epsilon} \leq P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X, Y)) \leq 2^{-nI(X, Y) - 3\epsilon},$$

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.



img/com.png

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

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

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 $P_e \rightarrow 0$.
- 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 \times represents the Cartesian product over sets. The jointly typical set

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

Note that Part A of the joint AEP states that:

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

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

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

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

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

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

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

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 $A_\epsilon^{(n)}(X)$.

img/random-cw.pdf

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

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

Further,

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

Applying the previous result with a union bound:

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

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

- Implies that: Joint typically 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:

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

Then, the jointly typical decoding rule states that

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

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

$$\begin{aligned} \mathbb{E}[P_e(c_n)] &= P(\hat{J} \neq J) = \sum_{j=1}^M P(\hat{J} \neq J | J = j) P(J = j) = P(\hat{J} \neq J | J = 1). \\ &\leq P((X^n(1), Y^n) \notin A_\epsilon^{(n)}(X, Y) | J = 1) + \sum_{j=2}^M P((X^n(j), Y^n) \in A_\epsilon^{(n)}(X, Y) | J = 1) \end{aligned}$$

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 $P_e(c_n) \rightarrow 0$. 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 $|c'_n| \geq \frac{1}{2}2^{nR} \implies \text{rate} \geq \frac{\log \frac{1}{2}2^{nR}}{n} = R - \frac{1}{n}$.

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:

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

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

$$\begin{aligned} H(X|Y) &\leq H(X, V|Y) \\ &= H(V|Y) + H(X|V, Y) && \text{(chain rule)} \\ &\leq H(V) + \sum_{v,y} H(X|V=v, Y=y) P(V=v, Y=y) && \text{(conditioning reduces entropy)} \\ &= H(V) + \sum_y \underbrace{H(X|V=0, Y=y) P(V=0, Y=y)}_0 + \sum_y \underbrace{H(X|V=1, Y=y) P(V=1, Y=y)}_{\leq \log(|\mathcal{X}|-1)} \\ &\leq H(V) + P(V=1) \log(|\mathcal{X}| - 1) \\ &= h_2(P_e) + P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

□

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

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

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

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

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.

If $R > C^{(I)}$ then R 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,*

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

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

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

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

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

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

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

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

In conclusion,

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

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.

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

- Compression rate is defined as

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

- 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:

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

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

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

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

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

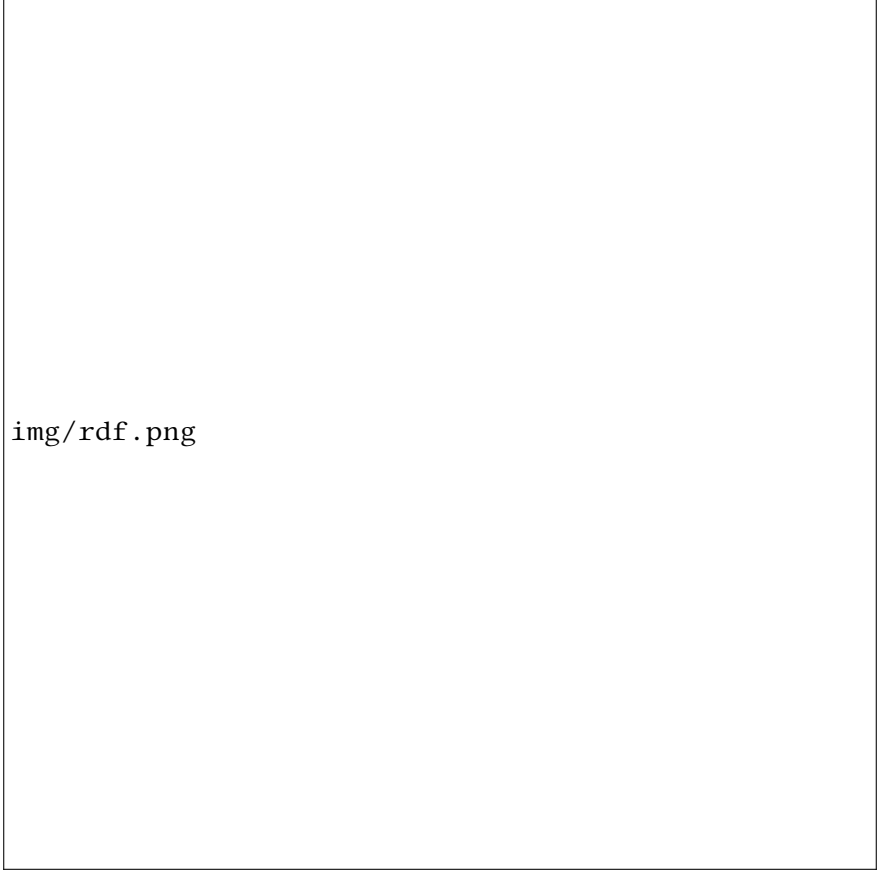
Note that in the setting when you have continuous data, the notion of rate distortion is perhaps more important, because it does not make sense to talk about lossless compression.

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.



img/rdf.png

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, \dots, 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
- $U \oplus_2 V \sim \text{Ber}(D)$.

□

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

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

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 complete 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

$$\begin{aligned} N(a|x^n) &= \sum_{i=1}^n \mathbf{1}_{\{x_i=a\}} \\ P_{x^n}(a) &= \frac{N(a|x^n)}{n}. \end{aligned}$$

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

Definition. \mathbb{P}_n denotes the collection of all empirical distributions of sequences of length n .

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

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.

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

□

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

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

Proof. Proof is straightforward. □

Fill in later

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

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

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 $P \in \mathcal{M}(\mathcal{X})$ if

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

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

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

Chapter 5

CS109: Probability

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

minted

Var Poi Beta Bin Geo Cov

*arg min *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

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{1 \leq i_1 < \dots < i_k \leq n} |A_{i_1} \cap \dots \cap A_{i_k}| \right).$$

5. Computing CDF in terms of Φ :

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

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) $X \sim \text{Bernoulli}(p)$, $0 \leq p \leq 1$. 1 if coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p; & x = 1; \\ 1 - p; & x = 0. \end{cases}$$

$$\mathbb{E}[X] = p; \quad (X) = p(1 - p).$$

(b) $X \sim \text{Binomial}(n, p)$, $0 \leq p \leq 1$. The number of heads in n independent flips of a coin with heads probability p .

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\mathbb{E}[X] = np; \quad (X) = np(1 - p).$$

(c) $X \sim \text{Geometric}(p)$, $p > 0$. The number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1 - p)^{x-1}.$$

$$\mathbb{E}[X] = \frac{1}{p}; \quad (X) = \frac{1 - p}{p^2}.$$

(d) $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$. A probability distribution over the nonnegative integers used for the modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}.$$

$$\mathbb{E}[X] = \lambda; \quad (X) = \lambda.$$

Intuition: let $n \rightarrow \infty$, $p \rightarrow 0$, 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

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

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

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

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

Recall that the definition of e is

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

It follows that

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

Understand how this derivation works with the exponential term

Continuous:

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

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

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

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

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

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

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

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

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

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, \dots, X_n :

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

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

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

Proof - indicator random variables.

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

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

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

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

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

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

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

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

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, \dots . Suppose $\mathbb{E}[X_i] = \mu$ and $(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. For any $\epsilon > 0$:

$$P(|\bar{X} - \mu| \geq \epsilon) \rightarrow 0.$$

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

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

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

Then

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

20. Central Limit Theorem (CLT). Consider I.I.D. random variables X_1, X_2, \dots . Suppose $\mathbb{E}[X_i] = \mu$, and $(X_i) = \sigma^2$. Then

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \rightarrow \mathcal{N}(0, 1); \quad \text{as } n \rightarrow \infty.$$

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

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

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

$$\theta_{MLE} = \arg \max_{\theta} L(\theta).$$

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 :

$$\begin{aligned} \theta_{MAP} &= \arg \max_{\theta} f(\theta|X_1, X_2, \dots, X_n) = \arg \max_{\theta} \frac{f(X_1, X_2, \dots, X_n|\theta)g(\theta)}{h(X_1, X_2, \dots, X_n)} \\ &= \arg \max_{\theta} \frac{(\prod_{i=1}^n f(X_i|\theta))g(\theta)}{h(X_1, X_2, \dots, X_n)} = \arg \max_{\theta} g(\theta) \prod_{i=1}^n f(X_i|\theta). \end{aligned}$$

28. Log a posteriori

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

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

Employ conditional independence assumption:

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

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

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

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 $P(X|\text{parents}(X))$. Graph has no cycles (directed acyclic graph).

34. Showing two distributions are independent. If

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

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-geometric 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

Chapter 6

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

6.1 Lecture 1

6.1.1 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*.

6.1.2 Introduction

Objective. Solving linear equations in an abstract way.

Linear algebra is a useful fundamental 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.

6.1.3 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 $f : \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{F}$. 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 $((x, y), (x', y')) \rightarrow (x + x', y + y')$. There is an additive identity $(0, 0)$, and an additive inverse $(-x, -y)$. We can define multiplication in 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 $+: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{F}$ and $\cdot: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{F}$, satisfying the 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 multiplicative 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 complex 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 multiplying elements of V by elements in \mathbb{K}). Elements in V are called vectors, elements in \mathbb{K} are called scalars.

6.2 Lecture 2

6.2.1 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.

6.2.2 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

$$\begin{aligned}+_v : V \times V &\rightarrow V \\ \cdot_v : \mathbb{K} \times V &\rightarrow V\end{aligned}$$

Aside. The difference between \rightarrow and \mapsto : 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.

$$\begin{aligned}V \times V &\rightarrow V \\ (\vec{v}, \vec{w}) &\mapsto \vec{v} + \vec{w}\end{aligned}$$

Scalar multiplication.

$$\begin{aligned}\mathbb{K} \times V &\rightarrow V \\ (\lambda, \vec{v}) &\mapsto \lambda \vec{v}.\end{aligned}$$

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 identity.* 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

$$\begin{aligned}\lambda(\vec{u} + \vec{v}) &= \lambda\vec{u} + \lambda\vec{v} \\ (\lambda + \mu)\vec{v} &= \lambda\vec{v} + \mu\vec{v}\end{aligned}$$

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 $f : S \rightarrow \mathbb{K}$. We can define addition and scalar multiplication.

- (Addition) Let $f_1, f_2 \in V$, i.e. $f_1 : S \rightarrow \mathbb{K}, f_2 : S \rightarrow \mathbb{K}$. We can define $(f_1 + f_2)(s) := f_1(s) + f_2(s)$.

$$\begin{aligned}f_1 + f_2 &: S \rightarrow \mathbb{K} \\ s &\mapsto f_1(s) + f_2(s).\end{aligned}$$

- (Scalar multiplication) For any $\lambda \in \mathbb{K}$, we can define $\lambda f_1 : S \rightarrow \mathbb{K}$ 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

$$f : \{1, \dots, n\} \rightarrow \mathbb{K}$$

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 $f : \mathbb{N} \rightarrow \mathbb{K}$. So the set of sequences $\{(x_1, \dots) | x_i \in \mathbb{K}\}$ forms a vector space. That is, $\{f : \mathbb{N} \rightarrow \mathbb{K}\}$ 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 $f(2) = 5$.

6.3 Lecture 4

Recall that $U \subset V$ is a vector subspace if

- $0 \in U$
- Closed under addition
- Closed under scalar multiplication

Example. Let $V = k^2$, that is

$$V = \{(a, b) | a, b \in k\}$$

Choose any $v = (a, b)$. Then line

$$L_v = (v) = \{\text{all scalar multiples of } v\}$$

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 $T : U \rightarrow W$ 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 $T : V \rightarrow W$ 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

$$T(\lambda v) = \lambda T(v) = 0.$$

□

Example. Consider the set

$$U : \left\{ f : \mathbb{R} \rightarrow \mathbb{R} \mid f^{(n)} : \mathbb{R} \rightarrow \mathbb{R} \text{ exists and is continuous} \right\}$$

•

6.4 Notes on 3.F: Duality

Definition. A linear functional on V is a linear map from V to \mathbf{F} , i.e. an element of $\mathcal{L}(V, \mathbf{F})$.

Definition. The dual space of V , denoted V' is the vector space of all linear functions on V . In other words, $V' = \mathcal{L}(V, \mathbf{F})$.

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, \dots, v_n is a basis of V , then the dual basis of v_1, \dots, v_n is the list ϕ_1, \dots, ϕ_n of elements of V' where each ϕ_j is the linear functional on V such that

$$\phi_j(v_k) = \begin{cases} 1; & \text{if } k = j \\ 0; & \text{if } k \neq j. \end{cases}$$

6.5 Key Ideas

Chapter 7

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]parskip [margin=1in]geometry

MATH 120 - Groups and Rings Instructor: Church; Notes: Adithya Ganesh

Contents

7.1 Lecture 4:

7.1.1 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:

$$\det GL_2\mathbb{R} \rightarrow \mathbb{R}^x.$$

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:

$$\begin{aligned}\mathbb{R}^x &\rightarrow \mathbb{R}_{>0}^x \\ x &\mapsto |x|.\end{aligned}$$

Since it is true that

$$|xy| = |x||y|,$$

we know that the absolute value is a homomorphism.

Example. Consider the sign function.

$$\begin{aligned}\mathbb{R}^x &\rightarrow \{\pm 1\} \\ x &\mapsto \begin{cases} +1; & \text{if } x > 0 \\ -1; & \text{if } x < 0. \end{cases}\end{aligned}$$

Clearly,

$$(xy) = (x)(y).$$

Example. Consider the map

$$\begin{aligned}\mathbb{R} &\rightarrow_2 \mathbb{R} \\ x &\mapsto \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix}\end{aligned}$$

Is it true that

$$(\cos(x+y), -\sin(x+y), \sin(x+y), \cos(x+y)) = (\cos x - \sin x \sin x \cos x)(\cos y, -\sin y, \sin y, \cos y)$$

$$\begin{pmatrix} \cos(x+y) & -\sin(x+y) \\ \sin(x+y) & \cos(x+y) \end{pmatrix} = \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix} \begin{pmatrix} \cos y & -\sin y \\ \sin y & \cos y \end{pmatrix}$$

This is tricky, but it is

$$\text{rot}(x+y) = \text{rot}(x) \circ \text{rot}(y)$$

Notice that we can also show that there is a homomorphism on rotation without knowing the exact formula for the matrix.

7.1.2 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 $f_1 : \mathbb{Z}_4 \rightarrow GL_2\mathbb{R}$ 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 $f : \mathbb{Z} \rightarrow \{\pm 1\}$ with $f(2) = 1$?

Clearly, there is a trivial homomorphism $f(\text{anything}) = 1$. 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 $f_1 : G \rightarrow H$ and $f_2 : G \rightarrow H$. 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 $G \rightarrow \{\pm 1\}$ with $f(a) = -1$ and $f(b) = -1$?

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 previous setting, we really do have a list of complete coincidences, $x^k = x^l$, only if $k \equiv l \pmod{4}$.

In the future, we will use the following notation, $Z_4 = \langle x | x^4 = 1 \rangle$ 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 $GL_2\mathbb{C}$, 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 $f : G \rightarrow (X)$ in a set, $Perm(X) =$ group of permutations (bijections) with $g : X \rightarrow X$.

7.2 Lecture 5

Notation: given a subset $T \subset G$, the notation

$$\langle T \rangle = \text{subgroup of } G \text{ generated by } T$$

$$\langle T \rangle = \cap H_{\text{all subgroups } H < G}$$

Recall:

$$\text{Perm}(X) = \text{group of bijections } g : X \rightarrow X \text{ under } \circ$$

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 within each cycle.

7.2.1 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 $\langle g \rangle$.

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 $[G : H] = 4$.

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 $[G : H] = 2$.

Example. Let p be a prime number. Suppose $|G| = p$. Then every $g \in G$ has $|g| = 1$ or $|g| = p$. This means that $G \cong \mathbb{Z}_p$.

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

$$|G| = (\# \text{ of cosets}) \cdot |H|$$

7.3 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)* $gHg^{-1} = \{ghg^{-1} | h \in H\}$

Example. Let $G = S_5$, and suppose $H = \{\sigma \in G | \sigma(2) = 2\}$. Let $g = (123)(45)$. Want to find gH , Hg , and gHg^{-1} .

Note that $|G| = 120$, $|H| = 24$. Note that $|gH| = |Hg| = |gHg^{-1}| = 24$. Compute the cosets and the conjugate.

Note that

$$\begin{aligned} gH &= \{\sigma \in S_5 | \sigma(2) = 3\} . \\ Hg &= \{\sigma \in S_5 | \sigma(1) = 2\} . \\ gHg^{-1} &= \{\sigma \in S_5 | \sigma(3) = 3\} . \end{aligned}$$

Also, it is clear that gH and Hg are not subgroups (not preserved under composition). However, gHg^{-1} is a subgroup.

Example. Let $G = \text{Isometries}(\mathbb{R}^2)$, that is, distance preserving bijections in the plane. Let

$$H = \{h \in G \mid h(\mathbf{0}) = \mathbf{0}\}$$

$$g = 90^\circ \text{ rotation around } (1, 0)$$

Compute the cosets and conjugate.

We can compute that

$$gH = \{\gamma \in G \mid \gamma(\mathbf{0}) = g(\mathbf{0})\}$$

$$Hg = \{\gamma \in G \mid \gamma(g) = \mathbf{0}\}$$

$$gHg^{-1} = \{\gamma \in G \mid \gamma(p) = p\}$$

(where $p = g(\mathbf{0})$).

Question 1 on HW2. Answer is $K = \mathbb{Z}$. Look at solutions for details.

On homework, discussed the notion of a kernel. If $f : G \rightarrow Q$, then

$$(f) = \{g \in G \mid f(g) = 1\}.$$

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 $f : G \rightarrow Q$, with $(f) = H$, how can you prove I'm lying?

Answer. Write out where each element maps to:

$$\begin{aligned} e &\rightarrow e \\ (12) &\rightarrow 1 \\ (23) &\rightarrow a \\ (13) &\rightarrow b \\ (123) &\rightarrow c \\ (132) &\rightarrow c^{-1} \end{aligned}$$

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.

7.4 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 $f : G \rightarrow H$, then $\ker(f)$ is always a normal subgroup of G .

Question. Give a normal subgroup $N \triangleleft G$, can we find a group Q and a homomorphism $f : G \rightarrow Q$, with $\ker(f) = N$?

Example. Take $G = \mathbb{Z}$, and let $N = 2\mathbb{Z}$. Does there exist some $f : \mathbb{Z} \rightarrow Q$ with $\ker(f) = 2\mathbb{Z}$?

Let's first establish what this means:

$$f(n) = 1 \text{ if } n \text{ even}$$

$$f(n) \neq 1 \text{ if } n \text{ odd}$$

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 $q^2 = 1$).

Example. Let $G = \mathbb{Z}$, and $N = 10\mathbb{Z}$. We would like some map $f : \mathbb{Z} \rightarrow Q$, with $\ker(f) = 10\mathbb{Z}$.

We know that all of the multiples of 10 must map to the identity in Q . But we know more, we can state that

$$f(m) = f(n) \Leftrightarrow 10 \mid (m - n).$$

The \Leftarrow direction is easy. The \Rightarrow direction is true because if not, the kernel would be bigger.

Note that

$$A = \{\dots, -10, 0, 10, 20\} \text{ map to 1 in } Q$$

$$B = \{\dots, -9, 1, 11, 21\} \text{ map to other element in } Q$$

...

$$J = \{\dots, -1, 9, 19, 29\} \text{ map to a tenth element in } Q.$$

Insight: what if we call $Q = \{A, B, C, \dots, J\}$. Define the group operation $B + C = D$, and we can take any element from these subsets, and get the “answer” D .

So $Q = \mathbb{Z}/10\mathbb{Z}$, which is our quotient group. In other words:

$$10\mathbb{Z} = (\mathbb{Z} \twoheadrightarrow \mathbb{Z}/10\mathbb{Z}).$$

or:

$$N = (G \twoheadrightarrow G/N).$$

Aside on notation: ¹

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 \bar{g} under the notation $g \sim h$ iff $h = gn$.²
- Its group operation is $\bar{g} \cdot \bar{h} = \overline{gh}$. **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: $\bar{1} \cdot \bar{h} = \overline{1h} = \bar{h} = \overline{h \cdot 1} = \bar{h} \cdot \bar{1}$
- Associativity: $\bar{a} \cdot (\bar{b} \cdot \bar{c}) = \bar{a} \cdot \overline{bc} = \overline{a \cdot (b \cdot c)} = \overline{(a \cdot b) \cdot c} = \overline{ab} \cdot \bar{c} = (\bar{a} \cdot \bar{b}) \cdot \bar{c}$.

Several comments: ^{3 4}

Note that there is a canonical surjective homomorphism

$$\pi : G \twoheadrightarrow G/N$$

that takes $g \mapsto \pi(g) = \bar{g}$.

Remark. What is the size of the quotient group? Note that

$$\begin{aligned} |G/N| &= \text{of cosets of } N \text{ in } G \\ &= \text{index}[G : N] \end{aligned}$$

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 $f : G \rightarrow H$ is any homomorphism, then

$$G/(f) \cong \text{Im}(f).$$

To name the map:

$$\psi(\bar{g}) = f(g).$$

¹Note that \twoheadrightarrow indicates a surjective map, and \hookrightarrow indicates an injective map.

²On equivalence classes: $g \sim h \Leftrightarrow h \in \bar{g} \Leftrightarrow \bar{g} = \bar{h}$

³This 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.

⁴(Note that 3.1 in the book is pretty confusing)

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 $\bar{g} \cdot \bar{h} = \overline{gh}$ is well defined. Suppose that $\bar{g} = \bar{a}$ and $\bar{h} = \bar{b}$. Then

- $\bar{g} \cdot \bar{h} = \overline{gh}$
- $\bar{g} \cdot \bar{b} = \overline{gb}$
- $\bar{a} \cdot \bar{b} = \overline{ab}$

We need to check that

$$\overline{gh} = \overline{gb} = \overline{ab}.$$

Suppose $\bar{h} = \bar{b}$. Then there exists $m \in N$ such that $b = hm$. We have

$$(gb) = (gm)m,$$

so $gb \sim gm$, i.e. $\overline{gb} = \overline{gm}$.

Note that $\bar{a} = \bar{g}$, 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 $nb \in Nb = bN$.

7.5 Lecture 8

7.5.1 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

$$1 \rightarrow x \rightarrow y \rightarrow 2$$

$$2 \rightarrow y \rightarrow z \rightarrow 4$$

$$3 \rightarrow w \rightarrow w \rightarrow 3$$

$$4 \rightarrow z \rightarrow x \rightarrow 1.$$

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 $g\sigma g^{-1}$.

Example. Let $\sigma = (1\ 2\ 7)(5\ 8)(3\ 4)$, and let $g(i) = i + 10 \pmod{100}$. Then

$$g\sigma g^{-1} = (11\ 12\ 17)(15\ 18)(13\ 14).$$

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 $\{1\}, \{a\}, \{b\}, \dots$

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 $\alpha_g : G \rightarrow G$ given by conjugation: $\alpha_g(h) = ghg^{-1}$. Is this a homomorphism?

Consider

$$\begin{aligned}\alpha_g(hk) &= ghkg^{-1} \\ \alpha_g(h)\alpha_g(k) &= ghg^{-1}gkg^{-1} = ghkg^{-1}.\end{aligned}$$

So yes, it is a homomorphism. What is the kernel of this function? Just the identity.

If $f : G \rightarrow H$ is a homomorphism, then the following are equivalent:

- f is injective
- $\ker(f) = \{1\}$.

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.

7.5.2 Group actions

We start with the definition.

Definition. An action of a group G on a set X is a homomorphism $\alpha : G \rightarrow \text{Perm}(X)$.

In other words, to each $g \in G$, we can associate a function $\alpha_g : X \rightarrow X$ such that

$$\alpha_g \circ \alpha_h = \alpha_{gh}.$$

This is almost the same as 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 $\text{Perm}(\mathbb{R}^2)$. 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.

Example. Let $G = \mathbb{Z}/7\mathbb{Z}$, and let $X = \{1, 2, 3, 4\}$. Claim: any action of G on X is trivial.

Notation in book. Suppose we have $x \in X$, and we can consider $\alpha_g : X \rightarrow X$. 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.

7.6 Lecture 9

We start by 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 $\{a, b, \bar{a}, \bar{b}\}$, 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

$$n(D_{10}, G) = \{(x, y) | x, y \in G, x^5 = 1, y^2 = 1, yxy^{-1} = x^{-1}\}$$

7.7 Lecture 11

Isomorphism theorems and quotient groups. Suppose you have a homomorphism

$$\alpha : G/N \rightarrow H.$$

Then you can get a homomorphism

$$a : G \rightarrow H$$

with $a(g) = \alpha(\bar{g})$.

Concretely, there is a bijection between homomorphisms $\alpha : G/N \rightarrow H$ and homomorphisms $a : G \rightarrow H$ with $N \leq \ker(a)$.

Suppose we have a group G , let $\bar{G} = G/N$. So we can define a projection π from $G \rightarrow G/N$. Suppose we have some subgroup $M < \bar{G}$. Claim: let $B = \pi^{-1}(M)$ be a subgroup of \bar{G} . Schematically, we can represent this as follows:

This implies that subgroups $\bar{B} \leq \bar{G}$ are in bijection with subgroups $N \leq B \leq G$. Additionally, we have $\bar{B} \cong B/N$.

If we have $N < B < C$ and $B \triangleleft C$, then we have that $\bar{C}/\bar{B} \cong C/B$.

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 $\bar{A} \triangleleft \bar{G} \Leftrightarrow A \triangleleft G$.

7.7.1 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

- $g * (h * x) = (gh) * x$.
- $g * (h x h^{-1}) = (gh) * (gh)^{-1}$.

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: $\{g \in G \mid g * x = x\}$.
- The definition of orbit: $\{g \cdot x \mid g \in G\}$.

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:

- $|G| = \sum \text{size of each conjugacy class}$
- If there are k conjugacy classes in G , pick representatives $1, g_2, g_3, \dots, g_k$. 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

$$|G| = \underbrace{|Z(G)|}_{\text{conjugacy classes of size 1}} + \sum_{i=1}^r [G : C_G(g_i)].$$

7.8 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

$$L = \left\{ a^{k_1} b^{l_1} \dots a^{k_n} b^{l_n} \mid \sum k_n = 0, \sum l_n = 0 \right\}.$$

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. $\{aba^{-1}b^{-1}, aaba^{-1}\} \dots$

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 $f : G \rightarrow G$.

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 $f : G \rightarrow H$ 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

$$\alpha : G \rightarrow (G)$$

with kernel $Z(G)$. You can think of this homomorphism as a group action. Since its kernel is $Z(G)$, we can write

$$Z(G) = \{gzg^{-1} = g\forall g\}.$$

So we can think of G acting on itself by conjugation, with $\alpha_g : G \rightarrow G$ with

$$\alpha_g(h) = ghg^{-1}$$

and

$$\ker(\alpha) = \{z \in G | \alpha_z = id\} = \{z | zhz^{-1} = h\forall h\} = Z(G).$$

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. $(Z_n) \cong (\mathbb{Z}/n\mathbb{Z})^\times$.

Recall $(\mathbb{Z}/n\mathbb{Z})^\times = \{\bar{k} \in \mathbb{Z}/n\mathbb{Z} | \bar{k} \text{ relatively prime to } n\}$, under multiplication.

For example,

$$(Z_8) \cong (\{\bar{1}, \bar{3}, \bar{5}, \bar{7}\}, \times).$$

And here, we have

$$(Z_8) = \{f_1 : x^k \mapsto x^k, f_2 : x^k \mapsto x^{3k}, f_3 : x^k \mapsto x^{5k}, f_4 : x^k \mapsto x^{7k}\}.$$

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).

7.9 Lecture 14

Lemma A. For any G with $|G| = p^k m > 1$ ($p \nmid m$), either

- G has a subgroup $H \leq G$ with $|H| = p^k l < p^k m = |G|$.
- G has a quotient $G \rightarrow \overline{G}$ 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,

$$S \mapsto hSh^{-1} = \{hsh^{-1} | s \in S\}.$$

Write $\Theta_H(S) = \text{orbit of } S, \Theta_H(S) = \{hSh^{-1} | h \in H\}$.

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 $\pi : G \rightarrow \overline{G}$ 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 $p^k m = p^b m |N|$.

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, \dots, 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, \dots, 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 $p \nmid n_p$, which implies that $n_p | m$. Then $n_p | |G|$ follows from Sylow 2(a), since given 2(a), $n_p = \text{size of the orbit under conjugation by } G$. \square

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, \dots, 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

$$\{P_1, \dots, P_k\}, \dots, \{P_k, \dots, P_{n_p}\},$$

which implies that $n_p = 0 + \dots + 0 \pmod{p}$, which contradicts $n_p \equiv 1 \pmod{p}$. \square

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 $c = 1 + 0 + \dots + 0 \pmod{p}$.

Now, assume 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. \square

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.

7.10 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 $|\Theta_Q(R)| = \text{number of } Q \text{ conjugates of } R$.

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 $K = \{g^k h | k \in \mathbb{Z}, h \in H\}$ 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 $K = \{q^k r | k \in \mathbb{Z}, r \in R\}$. Lemma D tells us that this is a subgroup. We now consider its size. Now, since $|q| = r$, we can write

$$K = \{q^k r | k \in \{0, \dots, p-1\}, r \in R\}.$$

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 $gh_1 = h_3g$.

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: $|AB| = \frac{|A||B|}{|A \cap B|}$.⁵

We can write down a helpful definition:

Definition. For any subset $H \leq G$, the normalizer $N_G(H)$ is $N_G(H) = \{g \in G | gHg^{-1} = H\}$.

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 $N \trianglelefteq G, H \trianglelefteq G, N \cap H = \{1\}$. Then every $g \in G = NH$ uniquely 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 $H \rightarrow (N)$.

This is the only information that is necessary to remember what G is.

Question 7 can be rephrased as: suppose you have some action from $H \rightarrow (N)$, 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 .

⁵Note 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.

7.11 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, $\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{Q}(\sqrt{2})^6, \mathbb{F}_p$
- Integers
- Modular stuff: $\mathbb{Z}/n\mathbb{Z}$ (won't write Z_n).
- Polynomials. We can write

$$\mathbb{Z}[x] = \{a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \mid a_i \in \mathbb{Z}, n \geq 0\}$$

More examples: we can write

$$\begin{aligned}\mathbb{Z}[\sqrt{2}] &= \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\} \\ \mathbb{Z}[\sqrt[3]{2}] &= \{a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \mid a, b, c \in \mathbb{Z}\}.\end{aligned}$$

Note that if A and B are rings, $A \times B$ is a ring with

$$\begin{aligned}(a_1, b_1) + (a_2, b_2) &= (a_1 + a_2, b_1 + b_2) \\ (a_1, b_1) \times (a_2, b_2) &= (a_1 a_2, b_1 b_2).\end{aligned}$$

Consider the ring of “functions of something.” For example,

$$C([0, 1]) = \{\text{continuous functions } f : [0, 1] \rightarrow \mathbb{R}\}.$$

Can consider various extensions of this theme:

$$\begin{aligned}C([2, 7]) &= \{\text{continuous functions } f : [2, 7] \rightarrow \mathbb{R}\} \\ \dots &= \{\text{infinitely differentiable functions } f : [0, 1] \rightarrow \mathbb{R}\} \\ \dots &= \{\text{continuous functions } f : [0, 1] \rightarrow \mathbb{C}\} \\ \dots &= \{\text{functions } f : [0, 1] \rightarrow \mathbb{R}\}\end{aligned}$$

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 $f \in C([0, 10])$. We can consider the ring of functions restricted to $[4, 6]$. It turns out that

$$r : C([0, 10]) \rightarrow C([4, 6])$$

is a ring homomorphism.

Grothendieck won a Fields medal for constructing a “something” so you can consider any ring as functions on “something.”

Definition. We say $r \in R$ is a unit if it has a multiplicative inverse. We write

$$R^\times = \{\text{units } r \in R\}.$$

⁶Subfield of \mathbb{R} generated by \mathbb{Q} and $\sqrt{2}$

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 $r \cdot s = 0$.

Example. If $\mathbb{Z}/10\mathbb{Z}$, $r = 4$, $s = 5$, so that $rs = 20 = 0 \in \mathbb{Z}/10\mathbb{Z}$.

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 $f : A \rightarrow B$ 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.

7.12 Lecture 20

We start by discussing the isomorphism theorems for rings.

1. Recall the first isomorphism theorem for rings from last lecture that says

$$(f) \cong R/(f); \quad f : R \rightarrow S.$$

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.⁷

⁷When 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 $f : X \rightarrow Y$ 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>

3. (3rd + 4th) Fix $I \subseteq R$ an ideal. We saw last time that we have this map $R \rightarrow R/I = \overline{R}$. 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 $\overline{f} : R/I \rightarrow C$. What would be great is if there was some function $f : R \rightarrow C$, 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 $0 = \overline{f}(0) = \overline{f}(i) = f(i)$. The theorem is saying in particular that this is all you need! In particular:

$$\{\text{homomorphisms } \overline{f} : R/I \rightarrow C\} \Leftrightarrow \{\text{homomorphisms } f : R \rightarrow C, f(I) = 0\}$$

The rest of today will be spent on definitions, which will help to make a lot of this concrete.⁸

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 $X = \{x_1, \dots, x_k\}$, write

$$(X) = (x_1, \dots, x_k).$$

There are two definitions here that we can state:

- (X) is the smallest ideal containing X , namely

$$X = \bigcap_{\text{ideal } I, X \subseteq I} I$$

- (X) is the set of all linear combinations of arbitrary length:

$$(X) = \{r_1 x_1 + \dots + r_n x_n \mid n \in \mathbb{N}, r_i \in R, x_i \in X\}.$$

Definition. Consider the subring S of R generated of X . We can write

- S is the smallest subring of R containing X :

$$S = \bigcap_{A, X \subseteq A} A.$$

- 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

$$\begin{aligned} & \{\text{homomorphisms } f : R \rightarrow C, f(X) = 0\} \\ &= \{\text{homomorphisms } \overline{f} : R/I \rightarrow C\} \Leftrightarrow \{\text{homomorphisms } f : R \rightarrow C, f(I) = 0\} \end{aligned}$$

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

⁸“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.

many homomorphisms $\phi : A \rightarrow \mathbb{Q}$ are there? Note that $\mathbb{Q} \subset A$ are the constant polynomials, so that $\phi(x) = x$ for any $x \in \mathbb{Q}$ (since you have to take $1 \rightarrow 1$).

This is a great advertisement 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

$$\{ \text{of pairs of numbers } a, b \in \mathbb{Q} \text{ with } a^2 + b^2 = 1 \}.$$

This hints at why rings are useful. It means that we can encode solutions to polynomial equations in terms of homomorphisms from some 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.

$$\phi : \mathbb{F}_n[x, y]/(I) \rightarrow \mathbb{F}_n; \quad \phi(c) = c; \forall c \in \mathbb{F}_n.$$

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 $f(x, y, z, w) = \text{quadratic in the inputs}$. 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 this⁹.

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 $b \in P$.¹⁰

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 $M \subsetneq I \subsetneq R$. Note that maximal ideals are prime.

⁹<http://www.math.uconn.edu/~kconrad/blurbs/gradnumthy/localglobal.pdf>

¹⁰Note 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 .

7.13 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. $I = (m_I)$.

7.14 Lecture 29

Today, we'll discuss: what is $i \in \mathbb{Z}/5^\infty\mathbb{Z}$? Recall that:

$$(\mathbb{Z}/5^k\mathbb{Z})^\times \cong \mathbb{Z}_{5^k-5^{k-1}} \cong \mathbb{Z}_{4 \cdot 5^{k-1}},$$

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 be the 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 $a_1 = 2$.
- $a_2 = a_1 - \frac{f(a_1)}{f'(a_1)} = 2 - \frac{5}{4} = \dots 1112$.

Note, we can write

$$\begin{aligned} -\frac{1}{4} &= \dots 1111 \\ -\frac{5}{4} &= \dots 1110 \\ 2 + \left(-\frac{5}{4}\right) &= \dots 1112 \end{aligned}$$

Similarly, we can write

$$\begin{aligned} \frac{1}{7} &= \dots 1033 \\ \frac{25}{7} &= \dots 103300 \\ -\frac{25}{7} &= \dots 341200 \\ 7 - \frac{25}{7} &= \dots 341212. \end{aligned}$$

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

$$f(a_{k+1}) = f\left(a_k - \frac{f(a_k)}{f'(a_k)}\right) = f(a_k + h) = f(a_k) + hf'(a_k) + O(h^2).$$

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 $(\mathbb{Z}/5^\infty\mathbb{Z})[x] \rightarrow (\mathbb{Z}/5\mathbb{Z})[x]$.

If there exists a_1 such that $\bar{f}(\bar{a}_1) = 0$ and $\bar{f}'(\bar{a}_1) \neq 0$, then we can write

$$\bar{f}(x) \text{ factors as } \bar{f}(x) = (x - \bar{a}_1)\bar{g}(x)$$

$$\bar{g}(x) \text{ not divisible by}$$

$$x - \bar{a}_1.$$

Theorem (Strong Hensel's Lemma). *Given a monic polynomial $f(x) \in \mathbb{Z}/5^\infty\mathbb{Z}[x]$, if $\bar{f}(x)$ factors into monic coprime $\bar{g}_i(x)$, $\bar{f}(x) = \bar{g}_1(x) \dots \bar{g}_k(x)$, then there exists monic coprime $g_i(x) \in (\mathbb{Z}/5^\infty\mathbb{Z})[x]$ such that $f(x) = g_1(x) \dots g_k(x)$.*

7.15 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 $\sigma_g : A \rightarrow A$ 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 :

$$\varphi : G \rightarrow S_A; \quad \text{defined by } \varphi(g) = \sigma_g,$$

called the permutation representation associated to the given action. We note some basic definitions:

1. The *kernel* of an action is $\{g \in G \mid g \cdot a = a\}$.

2. The *stabilizer* on a in G is $\{g \in G \mid g \cdot a = a\}$, 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 .

7.16 Notes on Irreducibility

Eisenstein's. Let p be a prime in and let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. Suppose $p \mid a_i$ for $i \in \{0, 1, \dots, n-1\}$ but $p^2 \nmid a_0$. 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 $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ be a polynomial in $R[x]$. Suppose a_{n-1}, \dots, 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 $\mathbb{Z}[i]$.

7.17 Notes on Free Groups

7.18 Key Ideas

7.18.1 Definitions

Definition. A *binary operation* $*$ on a set G is a function $*$: $G \times G \rightarrow G$.

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

$$GL_n(F) = \{A \mid A \text{ is an } n \times n \text{ matrix with entries from } F \text{ and } \det(A) \neq 0\}.$$

Definition. Let $(G, *)$ and (H, \circ) be groups. A map $\psi : G \rightarrow H$ such that

$$\psi(x * y) = \psi(x) \circ \psi(y) \text{ for all } x, y \in G$$

is called a *homomorphism*.

Definition. The map $\psi : G \rightarrow H$ 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 $f : A \rightarrow B$ 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 g . This homomorphism is called the “natural projection” of G onto G/N .

7.18.2 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 $xy^{-1} \in H$.

7.18.3 Examples

Example. The number of homomorphisms from $\mathbb{Z}_m \rightarrow \mathbb{Z}_n$ is $\gcd(m, n)$.

7.18.4 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 $f(g_1g_2) = f(g_1)f(g_2)$.
2. Studying quotient groups of G is equivalent to the study of the homomorphisms of G .

7.19 Things to review

1. Proof of Sylow’s theorem.
2. More intuition for conjugation.
3. Book sections.

Chapter 8

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]parskip [margin=1in]geometry

MATH 116 - Complex Analysis Instructor: Yakov Eliashberg; Notes: Adithya Ganesh

Contents

8.1 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 $\bar{z} = x - iy$, such that

$$z\bar{z} = x^2 + y^2 = |z|^2.$$

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:

$$e^z = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^n}{n!} + \dots$$

It is easy to check that this definition satisfies the usual properties:

$$e^{z_1+z_2} = e^{z_1} \cdot e^{z_2}; \quad e^{x+iy} = e^x e^{iy}.$$

We can similarly define

$$\begin{aligned} \cos z &= 1 - \frac{z^2}{2} + \frac{z^4}{4!} + \dots \\ \sin z &= z - \frac{z^3}{3!} + \frac{z^5}{5!} + \dots \end{aligned}$$

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 $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. 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 $x = \frac{1}{2}(z + \bar{z})$ and $y = -\frac{i}{2}(z - \bar{z})$. Now, set $A = a + ic$, $B = b + id$, so we can write.

$$\frac{1}{2}(A - iB)z + \frac{1}{2}(A + iB)\bar{z} = \alpha z + \beta \bar{z}.$$

Importantly, αz is complex linear while $\beta \bar{z}$ is complex antilinear (which means $F(\lambda z) = \bar{\lambda}F(z)$).

This proves that any real linear map can be written as a sum of a complex linear map and a complex antilinear map.

8.2 Differential 1-forms

Here, \mathbb{R}_z^2 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 $h \in \mathbb{R}_z^2$. 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 $\lambda_z = \mathbb{R}_z^2 \rightarrow \mathbb{R}$. Thus a 1-form is a function of arguments of 2 kinds: of a point $z \in U$ and a vector $h \in \mathbb{R}_z^2$.

Given a real valued function $f : U \rightarrow \mathbb{R}$ 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 $P, Q : U \rightarrow \mathbb{R}$ 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 is that λ is closed which by definition means $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$.

8.3 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 $\bar{z} = \mu z$ 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 $\{z_2 = 0\}$

8.4 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 or countable set of indices, and for every $\lambda \in \Delta$ a map $\Phi_\lambda : U_\lambda \rightarrow \mathbb{C}$ such that

- $S = \bigcup_{\lambda \in \Delta} U_\lambda$
- The image $G_\lambda = \Phi_\lambda(U_\lambda)$ is an open set in \mathbb{C} .
- The map Φ_λ viewed as a map $U_\lambda \rightarrow G_\lambda$ is one to one.
- For any two sets $U_\lambda, U_\mu, \lambda, \mu \in \Delta$, the images $\Phi_\lambda(U_\lambda \cap U_\mu), \Phi_\mu(U_\lambda \cap U_\mu) \subset \mathbb{C}$ are open and the map

$$h_{\lambda, \mu} = \Phi_\mu \circ \Phi_\lambda^{-1} : \Phi_\lambda(U_\lambda \cap U_\mu) \rightarrow \Phi_\mu(U_\lambda \cap U_\mu) \subset \mathbb{C}$$

8.5 Key ideas

8.5.1 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. $\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}$.

8.5.2 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 on 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 singularity) 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 analytic 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 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

$$z_0 f = \lim_{z \rightarrow z_0} \frac{1}{(n-1)!} \left(\frac{d}{dz} \right)^{n-1} (z - z_0)^n f(z).$$

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

$$\int_C f(z) dz = 2\pi i \operatorname{Res}_{z_0} f.$$

Suppose f is holomorphic on an open set containing a circle C and its interior, except for poles at the points z_1, \dots, z_N inside C . Then

$$\int_C f(z) dz = 2\pi i \sum_{k=1}^N \operatorname{Res}_{z_k} f.$$

Conformal map. A bijective holomorphic function $f : U \rightarrow V$ 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 $F : \Omega \rightarrow \mathbb{D}$ such that

$$F(z_0) = 0; \quad F'(z_0) > 0.$$

Corollary (3.2) Any two proper simply connected open subsets in \mathbb{C} are conformally equivalent.

Mantel's theorem.

8.6 Midterm review sheet

8.6.1 Cauchy-Riemann equations

f is holomorphic iff $u_x = v_y$; $u_y = -v_x$.

Differential operators w.r.t. z and \bar{z} .

$$\frac{\partial}{\partial z} = \frac{1}{2} \left(\frac{\partial}{\partial x} + \frac{1}{i} \frac{\partial}{\partial y} \right)$$

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2} \left(\frac{\partial}{\partial x} - \frac{1}{i} \frac{\partial}{\partial y} \right).$$

8.6.2 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 set Ω , then f has infinitely many complex derivatives in Ω . If $C \subset \Omega$ is a circle 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

$$\begin{aligned} |f^{(n)}(z_0)| &= \left| \frac{n!}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta \right| \\ &= \frac{n!}{2\pi} \left| \int_0^{2\pi} \frac{f(z_0 + Re^{i\theta})}{(Re^{i\theta})^{n+1}} Re^{i\theta} d\theta \right| \\ &\leq \frac{n!}{2\pi} \frac{\|f\|_C}{R^n} 2\pi. \end{aligned}$$

□

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 $R \rightarrow \infty$, 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

$$F(z) = \overline{f(\bar{z})},$$

look at power series expansions, and invoke the symmetry principle. \square

8.6.3 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 .

8.6.4 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 $1 \rightarrow r \in \mathbb{R}$ and then $r \rightarrow z$, so that

$$\begin{aligned} \log z &= \int_1^r \frac{dx}{x} + \int_r^z \frac{dw}{w} \\ &= \log r + \int_0^{\theta} \frac{ire^{it}}{re^{it}} dt \\ &= \log r + i\theta. \end{aligned}$$

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

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \dots$$

8.6.5 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.

8.6.6 Argument principle and Rouché'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 .

Rouché'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 is 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. □

8.6.7 Computation of integrals using residues

Residue limit identity. If f has a pole of order n at z_0 , then

$$z_0 f = \lim_{z \rightarrow z_0} \frac{1}{(n-1)!} \left(\frac{d}{dz} \right)^{n-1} (z - z_0)^n f(z).$$

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

$$\int_\gamma f(z) dz = 2\pi i \sum_{k=1}^N \text{Res}_{z_k} f.$$

Integrals to know.

- $\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi.$
- $\int_{-\infty}^{\infty} \frac{e^{ax}}{1+e^x} dx = \frac{\pi}{\sin \pi a}; 0 < a < 1.$
- $\int_{-\infty}^{\infty} \frac{e^{-2\pi i x \xi}}{\cosh \pi x} dx = \frac{1}{\cosh \pi \xi}.$

8.6.8 Harmonic functions and harmonic conjugates

Definition. A real or 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 $f, g : U \rightarrow \mathbb{R}$ 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. \square

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 $-e^x \cos y + C$.

8.6.9 Elementary conformal mappings

Examples to know:

8.6.10 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.

Chapter 9

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]parskip [margin=1in]geometry

MATH 171 - Real Analysis Instructor: George Schaeffer; Notes: Adithya Ganesh

Contents

9.1 9-24-18: Everything is a set

Administrivia:

- Book: Johnsonbaugh and Pfaffenberger
- (Supplement) Rudin's Principles of Mathematical Analysis
- Exam: likely week 5.

9.1.1 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 $S(x) = x \cup \{x\}$.

- $0 = \emptyset$
- $1 = \{\emptyset\}$
- $2 = \{\{\emptyset\}, \emptyset\}$.
- $3 = \{\{\emptyset\}, \emptyset, \{\{\{\emptyset\}, \emptyset\}\} = \{0, 1, 2\}$.

9.1.2 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 $f : X \rightarrow Y$, where X is the domain(f) and Y is the codomain(f). A function $f : X \rightarrow Y$ 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, $\sin \subseteq \mathbb{R} \times \mathbb{R}$.

9.1.3 On natural numbers

The set \mathbb{N} is equipped with a successor function $S : \mathbb{N} \rightarrow \mathbb{N} : x \mapsto S(x)$. 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 $X \times X \rightarrow X$.

- $+(a, b) = a+b$
- $\cdot(a, b) = ab$

$$\forall a, a + 0 = a. \forall a, b; a + S(b) = S(a + b).$$

9.2 10-1-18: Suprema and infima

Theorems of \mathbb{R} .

- \mathbb{R} is an ordered field.
- There 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 supremum) 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 .

$$\neg[\forall x \in S : x \leq M']$$

$$\exists x \in S : \neg[x \leq M']$$

$$\exists x : S : x > M'$$

$$\boxed{\forall \epsilon > 0, \exists x \in S : x > M - \epsilon}$$

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. $(m + 1)a > b$. \square

Note that the Archimedean principle is true in \mathbb{Q} as well. It inherits AP from \mathbb{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:

$$\begin{aligned} m + 1 &< m + (b - a) \\ &\leq a + (b - a) = b. \end{aligned}$$

Here, we have used the fact that $m \in S$, so $m \leq a$. \square

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 $a < \frac{k}{n} < b$. \square

The irrationals are also dense in the reals - just take the rationals and add $\sqrt{2}$, and follow a similar argument.

9.3 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 a metric 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 topology from d is the subspace topology.

If U is open / closed in S , it need not be closed in X .

¹On HW2: An example of an ordered field, for which the AP fails.

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 $f : X \rightarrow Y$, let $p \in X$. We say that f is continuous at p

1. In the analytic sense if

$$\forall \epsilon > 0, \exists \delta > 0, \forall x \in X : d_X(x, p) < \delta \implies d_Y(f(x), f(p)) < \epsilon.$$

2. In the sequential sense if \forall sequences $(x_n)_n \rightarrow p \in X$, the sequence

$$(f(x_n))_n \rightarrow f(p).$$

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 and 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.

$$\forall \epsilon > 0, \forall \delta > 0; \exists x \in X : d_X(x, p) < \delta \text{ and } d_Y(f(x), f(p)) \geq \epsilon.$$

In particular, we can define $(x_n)_{n=1}^\infty$ so that

$$d_X(x_n, p) < \frac{1}{n} \text{ and } d_Y(f(x_n), f(p)) \geq \epsilon.$$

This means that $x_n \rightarrow p$, but $f(x_n) \not\rightarrow f(p)$.

- $3 \implies 1$. Let $\epsilon > 0$. We know f is topologically continuous, so let $V = B_\epsilon(f(p))$ 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 $B_\delta(p) \subseteq U$.

$$\text{So, } x \in B_\delta(p) \implies x \in U \implies f(x) \in V \implies f(x) \in B_\epsilon(f(p)).$$

- $1 \implies 3$ is similar to the previous, just reverse all the statements.

Definition. The function $f : X \rightarrow Y$ 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 X .

For example, consider $f(x) = x^2$. Note that $f(-1, 1) = [0, 1]$. 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 $g(C) = \{\frac{1}{n} : n \geq 1\}$ is not closed.

Continuous functions don't preserve openness and they don't preserve closedness, but they preserve compactness.

Theorem. If $f : X \rightarrow Y$ 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 $G = \{f^{-1}(V) : V \in H\}$. Now, G covers K . Since f is continuous, it is an open cover² \square

Corollary. (Extreme value theorem). If $f : X \rightarrow \mathbb{R}$ (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 $f(q) = M$.

.

\square

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 remarks 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.

9.4 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 disconnected 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).

Definition (Alternative). A topological space is connected if the only clopen sets are X and \emptyset .

²Recall 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

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 disjoint, 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 $S \rightarrow X$ 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 converge, and since

$$|x_n - y_n| \leq 2^{-n}|x - y|,$$

both sequences converge 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 $(y_n)_n \rightarrow L$, $\exists N$ 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 $y_N \in V', V \cap V' \neq \emptyset$.

□

Example. A disconnected set in \mathbb{R} : \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 $f : X \rightarrow Y$ 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 $f : X \rightarrow Y$ 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 $f : X \rightarrow \mathbb{R}$ be a continuous real-valued 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. □

Example (Incomplete topologist's sine curve). Consider the graph of $\sin(\frac{1}{x})$ for $x \in (0, 1)$. Note that $\sin(\frac{1}{x})$ is continuous on this interval. This is connected.

Example (Midcomplete TSC). Consider $\{\text{Incomplete TSC}\} \cup \{(0, 0)\}$. This will be connected, still. But, it is not path connected. Consider a point (x, y) ; there is no path between (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.

9.5 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 $f : X \rightarrow Y$ be a continuous function with X and Y metric spaces. If X is compact then f is uniformly continuous.

1. Recall that $f : X \rightarrow Y$ 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. $\forall p \in X, \forall (x_n)_n \rightarrow p$ if $f(x_n)_n$ converges $\rightarrow f(p)$.
3. \forall open $U \subseteq Y$ $f^{-1}(U)$ is open in X . “preimage of an open set is open.”

⁴a lot of open sets are closed

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 $\forall p \in X, \forall \epsilon > 0, \exists \delta > 0$ such that $\forall x \in X[\dots]$.

We say that $f : X \rightarrow Y$ (metric spaces) is uniformly continuous if

$$\forall \epsilon > 0, \exists \delta > 0, \forall x, p \in X : d_X(x, p) < \delta \implies d_Y(f(x), f(p)) < \epsilon.$$

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 $p \rightarrow 0$, δ stops working (since it will contain the asymptote at 0).

Example. Let $f(x) = \sin(\frac{1}{x})$ 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

Proof of theorem. Suppose that $f : X \rightarrow Y$ is continuous but not uniformly continuous (where X is compact). Then $\exists \epsilon > 0, \forall \delta > 0, \exists x, p \in X$ such that $d_X(x, p) < \delta$ 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 $\forall n \geq 1$,

$$d_X(x_n, p_n) < \frac{1}{n}; \quad d_Y(f(x_n), f(p_n)) \geq \epsilon$$

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 $(x_{n_i})_i \rightarrow L$. Now, consider $(p_{n_i})_i$; we also have $(p_{n_i})_i \rightarrow L$.

Therefore

$$\lim_{i \rightarrow \infty} f(x_{n_i}) = f(\lim_{i \rightarrow \infty} (x_{n_i})_i) = f(L) = f(\lim_{i \rightarrow \infty} (p_{n_i})_i) = \lim_{i \rightarrow \infty} f(p_{n_i}).$$

(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 $f : [a, b] \rightarrow \mathbb{R}$ is continuous, then it is integrable. The proof of this theorem relies on the notion of uniform continuity.

9.5.1 Review

We now discuss the sequential version of uniform continuity.

Theorem. Let $f : X \rightarrow Y$ 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 .

⁵You can show that \sqrt{x} is uniformly continuous, but not Lipschitz continuous.

Dense sets. Suppose X is a metric space $S \subseteq X$, S is dense if

- $\bar{S} = X$
- 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 $[0, 1] \rightarrow \mathbb{R}$, the rational coefficient polynomials are dense.

9.6 11-05-18: Integrability and FTCs

Recall the Riemann-Darboux integral. Suppose $f : [a, b] \rightarrow \mathbb{R}$ is bounded with $a < b$. Then we can write

$$\int_a^b f = \sup \left\{ \int_a^b \varphi : \varphi \text{ a step fn ; } \varphi \leq f \right\}$$

or

$$\int_a^b f = \inf \left\{ \int_a^b \psi : \psi \text{ a step fn and } \psi \geq f \text{ on } [a, b] \right\}.$$

And if they match, f is integrable and $\int_f = \int_a^b f = \int_a^b f$.

Proposition (Sequential criterion for integrability). Suppose $f : [a, b] \rightarrow \mathbb{R}$ is bounded, $L \in \mathbb{R}$.

Subtext: Wts

$$\int_a^b f = L.$$

Then $\int_a^b f = L$ if and only if $\exists (\psi_n)_n, (\varphi_n)_n \in \text{Step}([a, b])$ for all k , $\psi_k \leq f \leq \varphi_k$; and

$$\int_a^b \psi_n \rightarrow L; \quad \int_a^b \varphi_n \rightarrow L.$$

Sufficient conditions for integrability.

- If $f : [a, b] \rightarrow \mathbb{R}$ 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 $f : [a, b] \rightarrow \mathbb{R}$ is bounded, it is integrable iff $\forall \epsilon > 0$, there exists $\psi, \varphi \in \text{Step}([a, b])$ such that $\varphi \leq f \leq \psi$ and $\int_a^b (\varphi - \psi) = \int_a^b \varphi - \int_a^b \psi < \epsilon$.

(Proof follows from the definition of integrability.)

Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ 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 , $f : \overline{I_k} \rightarrow \mathbb{R}$ 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

$$\varphi = \sum_{k=1}^n f(p_k) \mathbf{1}_{I_k}; \quad \psi = \sum_{k=1}^n f(q_k) \mathbf{1}_{I_k}.$$

Fill in details from notes.

□

Theorem (Lebesgue's Riemann integrability condition). *A function f is integrable if and only if*

$$\lambda(\{p \in [a, b] : f \text{ is discontinuous at } p\}) = 0,$$

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 $f : [a, b] \rightarrow \mathbb{R}$ be bounded, we call $F : [a, b] \rightarrow \mathbb{R}$ 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 $f : [a, b] \rightarrow \mathbb{R}$ be an integrable function $F : [a, b] \rightarrow \mathbb{R} : x \mapsto \int_a^x f$. 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

$$\lim_{x \rightarrow p^+} \frac{F(x) - F(p)}{x - p} = f(p).$$

In these notes we will prove half of it (since the case $x \rightarrow p^-$ is similar.) Make the change of variables $x = p + h$. Then the limit above is equivalent to

$$\lim_{h \rightarrow 0^+} \frac{F(p + h) - F(p)}{h} = \lim_{h \rightarrow 0^+} \left[\frac{1}{h} \int_p^{p+h} f \right],$$

⁶Cannot use this on homework unless you prove it.

where

$$F(x) = \int_a^x f.$$

Let $\delta > 0$ such that $\forall x \in [a, b] : |x - p| < \delta \implies |f(x) - f(p)| < \varepsilon$. 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

$$\left| \lim_{h \rightarrow 0^+} \left[\frac{1}{h} \int_p^{p+h} f \right] - f(p) \right| < \varepsilon.$$

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. \square

9.7 Sequences and series of functions

9.7.1 Pointwise vs. uniform convergence

9.8 Key ideas

Definition of a metric space.

A metric space is a set X together with a distance $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$ 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 $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$. 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 .

$\bar{a}^k \rightarrow \bar{a}$ iff $\bar{a}_j^k \rightarrow a_j$ 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, $x \in X, S \subseteq X$. Then x is interior to S if $\exists \varepsilon > 0, B_\varepsilon(x) \subseteq S$.

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 $\forall \varepsilon > 0, \exists y \neq x$ 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 that $\beta^p = \alpha$.

Proof. Uniqueness is easy once existence is shown. Assume $0 < \alpha < 1$, (since $\alpha = 1$ is easy, and for $\alpha > 1$, just take $\left(\frac{1}{\beta}\right)^n = \frac{1}{\alpha}$).

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 $(s_n)_n$ is increasing is straightforward. Thus, $(s_n)_n$ converges to some limit β .

Finally, we will show that $(s_n^p)_n \rightarrow \alpha$. This will show that β satisfies $\beta^p = \alpha$. Note that

$$\left(\frac{a_n}{2^n} \right)^p \leq \alpha < \left(\frac{a_n + 1}{2^n} \right)^p$$

check this
detail

Thus, it suffices to check that the difference between the left and right hand sides approaches zero as $n \rightarrow \infty$. 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 $k = 0, \dots, p-1$.

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 $\rightarrow 0$ as $n \rightarrow \infty$. □