Part I: Understanding the behaviour of sums of independent random variables

To be specific, we analyze how the averages of independent random variables concentrate around their expectation.

1. From the non-asymptotic viewpoint: Concentration Inequalities / Tail Bound

The number of random variables is some fixed finite number

(1) Markov's Inequalities

(2) Chebyshev's Inequalities

(3) Chernoff's Inequalities

- --> σ-sub-Gaussian

(4) For bounded Random Variables: Hoeffding's Bound

- Very useful for consequent proofs in this course
- --> Berstein's Inequalities: for bounded random variable with small variance

(5) Azuma's Inequalities

- For Lipschitz function
- Example: U-statistic
- --> Levy's Inequalities: for Lipschitz function and Gaussian random variable

(6) χ^2 tail bound

- Example: random projections (The Johnson-Lindenstrauss Lemma)
  - note that the bound is not determined by data dimensionality

> **Take-Home Message**
>
> - The average of n i.i.d random variables concentrates within an interval of length roughly $1/\sqrt{n}$ around the mean.

- The bound $\to 0$ as $n \to \infty$

2. From the asymptotic viewpoint

The number of random variables we average $\to \infty$

(1) Basic Concepts of Convergence

1. Almost Sure Convergence
2. Convergence in Probability
3. Convergence in Quadratic Mean
4. Convergence in Distribution

(2) Core Conslusion

1. Laws of Large Numbers (LLNs)
   - Useful tools: 1) Continous Mapping; 2) Slusky's Theorem

   > **Take-Home Message**
   >
   > - Average of i.i.d. random variables converges in probability to its expectation.

2. Central Limit Theorems (CLTs)
   - Useful tools: Delta Method

   > **Take-Home Message**
   >
   > - In most situations, when the number of independent (i.i.d. is not necessary) random variables $\to \infty$, their properly normalized sum converge in distribution to $\mathcal{N}(0, 1)$ even if the original variables themselves are not normally distributed.

(3) Uniform Law and Empirical Process Theory

1. Uniform Convergence of CDF: The Glivenko-Cantelli theorem

   > **Take-Home Message**
   >
   > - Empirical CDF makes sense as $n \to \infty$

2. Generalized to collections of sets: Vapnik-Cervonenkis Theory & VC dimension

   > **Take-Home Message**
   >
   > - Classification error is bounded.
   > - The bound is determined by the VC dimension of the classifier

3. Generalized to classes of functions: Rademacher Theorem & Rademacher Complexity

- Unlike the VC dimension, in the definition of the Rademacher complexity we do not maximize over the locations of points, i.e. in some sense it is not a worst case measure of complexity.
- The Rademacher complexity is measuring the maximum absolute covariance between {f(X1), . . . , f(Xn)} and a vector of random signs {1, . . . , n}. Intuitively, we think of a class F as too large if for many random sign vectors we can find a function in F that is strongly correlated with the random sign vectors.
- The Rademacher theorem straightforwardly implies the VC theorem

> **Take-Home Message**
>
> - Empirical process is bounded with probability at least $1 - \delta$,

Part II: Statistical Estimation

1. Statistical model

1. Taxonomy of statistical model
   - Parametric models
   - Non-parametric models: cannot be described by a **finite** set of parameters
2. Statistics -> sufficient statistics
   - Check sufficiency
     - by definition
     - factorization theorem (Neyman-Fisher factorization criterion): no need to compute the conditional distribution
   - minimal sufficient statistics
   - sufficiency - the partition viewpoint
   - sufficiency - the risk reduction viewpoint: rao-blackwell theorem
     - we can improve the rist of an estimator by conditioning it on sufficient statistics
   - relationship to likelihood:
     - once we have any sufficient statistic, we have everything we need to compute the likelihood function
     - Although minimal sufficient statistics are not unique they induce a unique partition on the possible datasets. This partition is also induced by the likelihood.

> **Take-Home Message**

> - Roughly, once we know the value of the sufficient statistic, the joint distribution no longer has any more information about the parameter θ. One could imagine keeping only the sufficient statistic and throwing away all the data

3. exponential family
    - The partition function is also the moment generating function
    - The likelihood function is concave
    - Minimal representations and minimal sufficiency
    - The maximum entropy duality
    - For exponential families MoM and MLE coincide.

> **Take-Home Message**
>
> - Exponential family distributions possess many useful and pleasant properties, and give us a somewhat unified way to think about "nice" distributions.

2. Point Estimation

(1) methods for evaluating estimators

Mean squared error (MSE): MSE = Bias^2 + Variance

- Unbiased estimators
    - Cramer-Rao bound provides a lower bound (i.e., Fisher Information) on the variance of an unbiased estimator
- Beyond unbiased estimators - decision theory
    - Still, our goal is to minimize the expected loss.

(2) methods of constructing estimators

1. The Method of Moments (MoM)
    - We equate the first k sample moments with the first k theoretical moments. This defines k equations with k unknowns.
2. Maximum Likelihood (MLE)
    - --> Equivariance and the profile likelihood
    - Asymptotic theory of MLE (requires some conditions)
        a. **Consistency**: Does the MLE converge in probability to θ -- analogous to the LLN
        b. **Asymptotic distribution**: What can we say about the distribution of $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$ -- analogous to the CLT

        - --> Influence Functions and Regular Asymptotically Linear Estimators
        - --> Asymptotic Relative Efficiency
3. Bayes estimator
    - minize the average risk.
        - for MSE the posterior mean is the Bayes estimator.
4. Minimax estimator (discussed later)
    - minimizes the worst-case risk
        - the minimax estimator is often difficult to compute directly
        - we can use Bayes estimators to find minimax estimators.
            - Tightly bounding the minimax risk
            - Least favorable prior.
                - Bayes estimators with a constant risk function are minimax

> **Take-Home Message**
>
> - **The method of moments** was popular many years ago because it is often easy to compute.
> - The most popular method for estimating parameters is **maximum likelihood**. The reason is that, under certain conditions, the maximum likelihood estimator is optimal. (explained later)
> - **Bayes estimators** although easy to compute are somewhat subjective (in that they depend strongly on the prior π).
> - **Minimax estimators** although more challenging to compute are not subjective, but do have the drawback that they are protecting against the worst-case which might lead to pessimistic conclusions, i.e. the minimax risk might be much higher than the Bayes risk for a "nice" prior.

3. Hypothesis Testing

1) Basic Concepts of Hypothesis Testing

Power function

- We want the power function $\beta(\theta)$ to be small when $\theta \in \theta_0$
    - keep Type I error (false positive, true H_0 is falsely rejected) rate smaller than α => ensure the **recall** of H_0
- and we want $\beta(\theta)$ to be large when $\theta \in \theta_1$

p-values

- The p-value is the smallest α at which we would reject H_0.
- reject if $p-value < \alpha$ controls the falsely rejection rate at α
- For the Lady Tasting Tea problem, a small p-value means that the lady's judgement is not random guess and we reject the null hypothesis (i.e., the random guess hypothesis)

2) Basic Tests

- The Neyman-Pearson Test
  - $H_0 = \theta_0$, $H_1 = \theta_1$
  - $T = L(\theta_1)/L(\theta_0)$ (likelihood ratio)
  - reject if the ratio is small
  - uniformly most powerful (UMP): $\beta(\theta)$ for $\theta \in \theta_1$ is larger than any other tests
- The Wald Test
  - $H_0 = \theta_0$, $H_1 \neq \theta_0$
  - $T_n = \frac{\hat{\theta}-\theta_0}{\sigma 0}$ or $T_n = \frac{\hat{\theta}-\theta_0}{\hat{\sigma} 0} \rightarrow_d \mathcal{N}(0,1)$
  - reject if $|T| > z_{\alpha/2}$
  - the closer θ_0 is to θ_1, the harder they are to be distinguished
- (Generalized) Likelihood Ratio Test (LRT)
  - $H_0 \in \Theta_0$, $H_1 \notin \Theta_1$
  - $\lambda = \frac{sup\theta_0 \mathcal{L}(\theta)}{sup\theta \mathcal{L}(\theta)}$
  - reject if $|\lambda|$ is small
  - Student's t-test
    - $T_n = \frac{\bar{X}-\theta_0}{S/\sqrt{n}} \rightarrow_d t_{n-1,\alpha/2}$
  - Wilks' phenomenon: We can simplify the LRT by using an asymptotic approximation (Chi-Square Distribution)
    - $-2\log\lambda \rightarrow_d \chi^2_\nu$

3) Advanced tests

- Goodness-of-fit testing: $H_0 : P = P_0$, $H_1 : P \neq P_0$
  - For two multinomial distributions: The chi-square test
- Two-sample Testing: $H_0 : P = Q$, $H_1 : P \neq Q$
  - Kernel-based tests (Maximum Mean Discrepancy) is very popular in machine learning
  - For two multinomial distributions: the Permutation Test

4) Multiple Testing

Control Family-Wise Error Rate (FWER)

FWER = P(falsely reject any null) (= p-value for single test)

1. Sidak correction: requires independence
2. Bonferroni correction
3. Holm's procedure: involves sorting

Control False Discovery Rate (FDR)

$FDR = E[{}^{V}_{maxR,1}]$ where V = # of false rejections, R = # of all rejection

- FWER controls FDR ($FDR \leq FWER$)

1. BH procedure

4. Confidence Intervals

No matter which distribution in P generated the data, the interval guarantees the coverage property

Four methods:

1. Probability Inequalities
2. Inverting a test
3. Pivots
4. Large Sample Approximations

> **Take-Home Message**
>
> - Intuitively, p-values are more informative than an accept/reject decision because it summarizes all the significance levels for which we would reject the null hypothesis.
> - Similarly, a confidence interval is more informative that a test because it summarizes all the parameters for which we would (fail to) reject the null hypothesis.

Part III Advanced Topics

1. Causal Inference

- We measure causality by average treatment effect $\tau = E[Y(1) - Y(0)]$
- However, you never observe it. We observe $Y^{obs} = WY(1) + (1 - W)Y(0)$

- A natural quantity that we can estimate is $\alpha = E(Y(1)|W=1) - E(Y(1)|W=0)$, but $\alpha \neq \tau$ in general (correlation is not causation)There are broadly two ways to try to fix this issue:

1. to randomly assign treatment, i.e. conduct a randomized trial
2. to adjust for confounders.
    - a. The most direct way: we compute the plug-in estimator by regression
    - b. Another way (Horvitz-Thompson estimator/the inverse propensity score estimator): we compute the propensity score by regression

2. Regression

Optimal Regression Function

- This MSE risk(prediction error) is minimized by the conditional expectation $y(x) = E(Y|X=x)$

Low dimensional linear regression

- $y_i = <x_i, \beta^*> + \epsilon_i$
    - Least square: $\hat{\beta} = argmin_\beta \frac{1}{2}||<x_i, \beta^*>||$ --> $\hat{\beta} = (X^T X)^{-1} Xy$, where X is the $n \times d$ design matrix

3. Non-parametric Statistics (taking regression as an example)

One of the most basic ways of doing non-parametric regression is called kernel regression. We will analyze kernel regression in this course since the general case is not very different.

The curse of dimensionality

- For both linear regression and kernel regression, the bound of the risk gets worse as d increases, however in non-parametrics the situation gets exponentially worse.

4. High-dimensional Statistics (taking regression as an example)

How to avoid the aforementioned exponential curse of dimensionality?

1. to make (strong) parametric assumptions – like assuming the regression function is linear.
2. to assume (strong) smoothness assumptions

- In particular, if you assume that the true regression function is β-smooth and β = d (say) then you will observe that the rate of convergence does not degrade as d gets larger (of course, the assumption is increasingly stringent as d gets larger)

3. to assume sparsity, this means that even though we have many covariates, the true regression function only (strongly) depends on a small number of relevant covariates. This is the type of setting we will focus on.
4. More broadly, the main idea is that we want to think about practically relevant structural properties (like smoothness/sparsity) that we can exploit to get around the discouraging worst-case rates of convergence.

What are natural estimators in this case? A couple of popular ones are based on thresholding

- Hard Thresholding Estimator (L0 sparsity)
- Soft Thresholding Estimator (L1 sparsity) Their risks are bounded.

Difference between in the Gaussian Sequence Model and the High-dimensional Regression

- In the Gaussian sequence model (no X) both of these programs had simple closed-form solutions, whereas now this is no longer the case. More importantly, best-subset is computationally intractable but the LASSO is not.

5. Bayes Inference

-	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
Parameter θ	**random variable**	**fixed**
Sample X	random variable	random variable
Use Bayes' theorem?	Yes. To update beliefs.	Yes, if it leads to procedure with good frequentist behavior. Otherwise no.

- Credible set VS confidence set
- Bernstein-von Mises theorem
  - At a high-level the Bernstein-von Mises theorem guarantees us that in fixed-dimensional problems, under the assumption that the prior is continuous, and (strictly) positive in a neighborhood around θ^* , the posterior is close to a Gaussian
- Priors = Regularizers?
  - estimating Bernoulli probabilities with "Laplace smoothing" (i.e. adding psuedo-counts) is just the posterior mean with a Beta prior.

- The LASSO regression estimator or Ridge regression estimator are just the posterior mode with either a Laplace prior or a Gaussian prior (respectively).
- The argument is that "many sensible frequentists procedures (ones with strong guarantees) are just posterior summaries with particular priors."
- What about in a high-dimensional setting?
  - once you leave the realm of the Bernstein-von Mises theorem (fixed d, growing n), things can break down.

6. Computing Sampling from Posterior & Monte Carlo

(1) Direct Sampling

(2) Importance Sampling

sample from P with the help of Q

(3) Markov Chain Monte Carlo (MCMC)

Monte Carlo Integration: for example, estimating the CDF at some position u, i.e., F(u)

In the Bayesian problem we use MCMC

- First, the big picture: In Markov Chain Monte Carlo (MCMC) the goal is to design a Markov Chain, whose limiting distribution is the distribution P under which we want to compute an integral.
- Then, we sample from the Markov Chain, and then use the law of large numbers (adapted to Markov Chains) to argue that our estimate is good.

The Metropolis-Hastings algorithm

- Sample a proposal $y \sim q(y|X_i = x)$ from a *proposal distribution* q.
- Evaluate the ratio $\{\min \frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\}$
- Accept the new sample Y with probability r, and reject it otherwise.

7. Bootstrap

At a high-level the bootstrap is a way to try to estimate the variability (think variance or confidence intervals) of a point estimate, but to do so in a way that avoids difficult analytic calculations. Often we want to estimate the variance of, or derive confidence intervals for an

arbitrary (non MLE, non-average) statistic, and the bootstrap gives a way to do this in many cases without resorting to tedious calculations.

- Bootstrap variance estimate
- Bootstrap Confidence Intervals

8. Model Selection

1. Cross Validation
    - If your goal is prediction, you have a reasonable sample-size and you have a reasonable computation budget use cross-validation.
2. Akaike Information Criterion (AIC):
    - If your goal is prediction, but you either have too small a sample or you have a very low computational budget, you should consider using AIC.
    - $AIC(j) = 2l_j(\hat{\theta}_j) - 2d_j$ (the larger the better)
3. Bayesian Information Criterion(BIC)
    - If you goal is selecting the "true" model you should use BIC.
    - $BIC(j) = 2l_j(\hat{\theta}_j) - d_j\log(n)$

Model selection can affect interpretability, i.e. typically in linear regression for instance

- the model selected by CV will tend to have many small coefficients
- while the model selected by BIC will tend to be much more parsimonious/interpretable.

9. Distances between Distributions

(1) The fundamental statistical distances

1. Total Variation
2. The chi-square divergence
3. Kullback-Leibler divergence
4. Hellinger distance

Relationship of these distances:

1. All of these fundamental statistical distances are special cases of **f-divergences**
2. Le Cam's Lemma
3. why do we need all these different distances?

- Roughly, when we want to compute a lower bound (i.e. understand the fundamental statistical difficulty of our problem), some divergences might be easier to compute than others.
- For instance, it is often the case that for mixture distributions the chi-square is easy to compute, while for many parametric models the KL divergence is natural
- Tensorization

4. Inequalities $TV(P,Q) \leq H(P,Q) \leq \sqrt{KL(P,Q)} \leq \sqrt{\chi^2(P,Q)}$

(2) Other distances

1. Distances from parametric families they are usually only defined for parametric families:
    i. Fisher information distance
    ii. Mahalanobis distance: This is just the Fisher distance for the Gaussian family with known covariance.
2. Integral Probability Metrics
    - The necessary and sufficient conditions for two distributions P, Q to be identical
3. Wasserstein distance
    - The Wasserstein distance has the somewhat nice property of being well-defined between a discrete and continuous distribution, i.e. the two distributions you are comparing do not need to have the same support. This is one of the big reasons why they are popular in ML
4. Maximum Mean Discrepancy