Lecture notes in Ph.D.

Tingzhou Yu[1]

February 26, 2022

[1] University of Waterloo, tingzhou.yu@uwaterloo.ca, https://uwaterloo.ca/scholar/t229yu,
Github: https://github.com/jerry129y/Lecture-notes

Contents

Conventions

$\mathbb{F}$ denotes either $\mathbb{R}$ or $\mathbb{C}$.

$\mathbb{N}$ denotes the set $\{1, 2, 3, ...\}$ of natural numbers (excluding 0).

Chapter 1

High-dim statistics

This Chapter is based on Spring 2022 STAT 946 - Topics in Statistics: High dim in Statistics (Instructor: Dr. David Saunders, Adam Kolkiewicz) and STAT 946 Fall 2021 (Instructor: Dr. Aukosh Jagannath) at UWaterloo.

References:

- Martin J. Wainwright, High-Dimensional Statistics [34]
- High Dimensional Statistics Lecture Notes, Philippe Rigollet and Jan-Christian Hütte. [1]
- Roman Vershynin, High-Dimensional Probability [2]
- Ramon van Handel, Probability in High Dimension [3]
- Christophe Giraud, Introduction to High-Dimensional Statistics [4]
- Alexandre B. Tsybakov, Introduction to Nonparametric Estimation [5]
- A. S. Bandeira, A. Singer, T. Strohmer, Mathematics of Data Science [6]

[1] https://klein.mit.edu/~rigollet/PDFs/RigNotes17.pdf

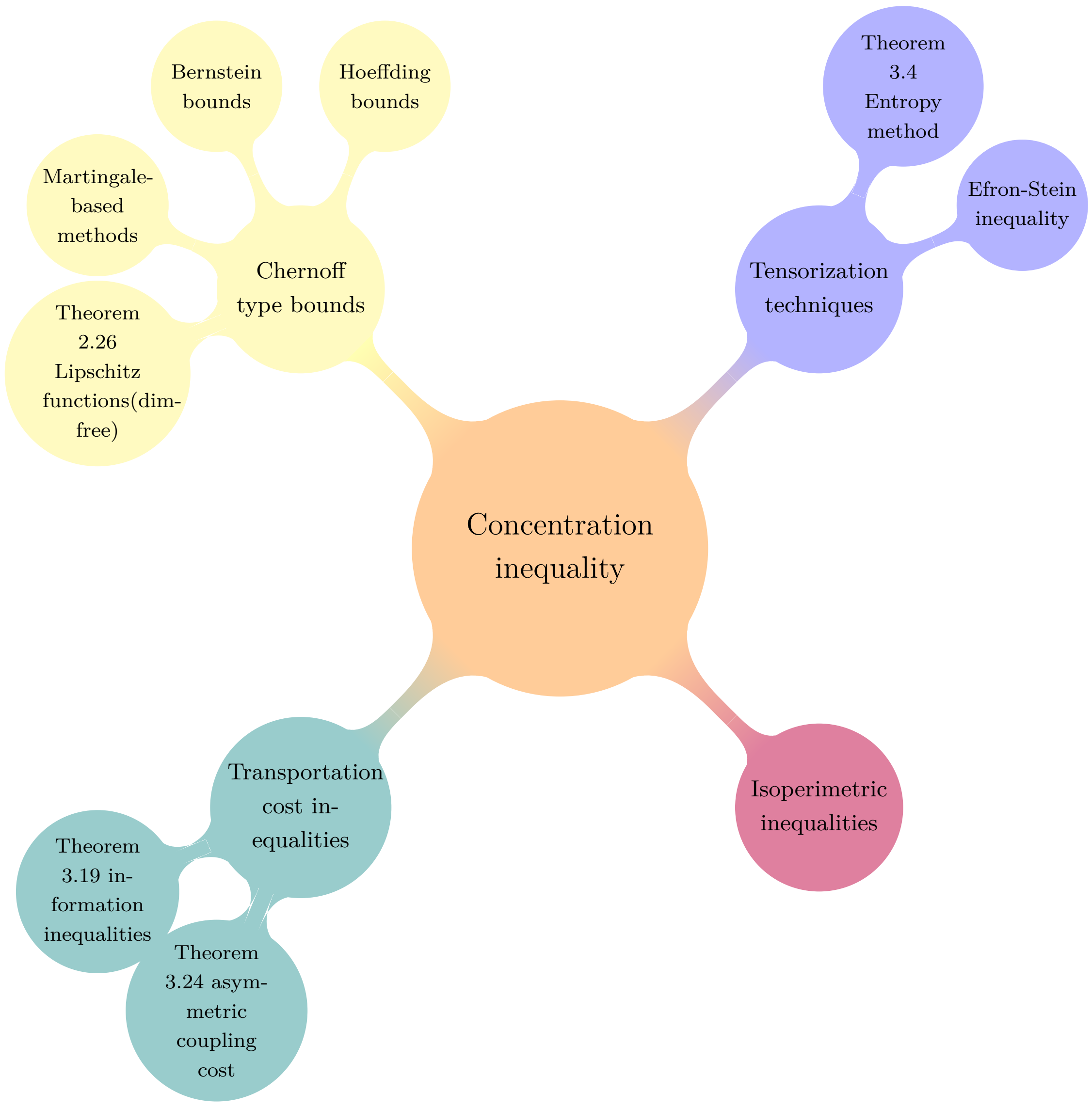
[2] https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf

[3] https://web.math.princeton.edu/~rvan/APC550.pdf

[4] https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/Bookv3.pdf

[5] https://link.springer.com/book/10.1007/b13794

[6] https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf


Concentration inequality
Chernoff type bounds
Bernstein bounds
Hoeffding bounds
Martingale-based methods
Theorem 2.26 Lipschitz functions(dim-free)
Tensorization techniques
Theorem 3.4 Entropy method
Efron-Stein inequality
Transportation cost inequalities
Theorem 3.19 information inequalities
Theorem 3.24 asymmetric coupling cost
Isoperimetric inequalities

- MGF + Chernoff technique:
  - Hoeffding bounds (sub-Gaussian; bounded variables)
  - Bernstein bounds (sub-exponential; conditions on moments)
  - Martingale-based methods (bounds on $f(X)$; f satisfies the bounded difference property; bounds depend on the dimension)
  - Theorem 2.26 (two-sided tail bound on Lipschitz functions of Gaussian i.i.d. variables; bound does not depend on the dimension)
- Entropic techniques
  - Theorem 3.4 (upper tail bound on $f(X)$: f must be separately convex and Lipschitz, X_1^n bounded and i.i.d.; bound does not depend on the dimension)
- Transportation cost inequalities
  - Theorem 3.19 (bounds through information inequalities on Wasserstein distances; f is Lipschitz)
  - Theorem 3.24 (two-sided tail bound through asymmetric coupling cost; f must be convex (or concave) and Lipschitz, X_1^n bounded and i.i.d.; bound does not depend on the dimension)

1.1 Introduction

1.1.1 The Efron-Stein inequality

See https://faculty.math.illinois.edu/~psdey/Math595FA19.html.

1.2 Marchenko-Pastur theorem and BBP transition

Chapter 2

Statistical inference

This Chapter is based on STAT 908 Statistical Inference (Instructor: Dr. Pengfei Li) at UWaterloo.
References:

- Mathematical statistics by Jun shao [30]
- Theoretical Statistics: Topics for a Core Course by R. Keener. [16]
- Asymptotic Statistics by A. W. van der Vaart
- [1] https://arxiv.org/pdf/2010.14863
- Reference lists [23] https://web.stanford.edu/~montanar/OTHER/TALKS/oops_refs.pdf

2.1 Multivariate Statistical Analysis: Multivariate normal distribution

- [1] for basic MND;
- [2] for The sampling distribution of MND.
- [3] for The sampling distribution of MND.
- [4] and [5] for Hypothesis testing of MND.

Definition 2.1.1. Suppose X is a $p-$dimensional random vector. If its probability density function has the followsing form

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

We say X follows a multivariate normal distribution, denoted by $X \sim N(\mu, \Sigma)$.

Given $X \sim N(\mu, \Sigma)$, we can get its moment generating function

$$M(t) = \exp\left(t^T\mu + \frac{1}{2}t^t\Sigma t\right).$$

Remark 2.1.2. This property can be also viewed as the second definition of multivariate normal distribution.

[1] https://www.cnblogs.com/lixddd/p/15432593.html
[2] https://www.cnblogs.com/lixddd/p/15515228.html
[3] https://www.cnblogs.com/lixddd/p/15515228.html
[4] https://www.cnblogs.com/lixddd/p/15518581.html
[5] https://www.cnblogs.com/lixddd/p/15524124.html

If the m.g.f. of a p-dimensional random vector X is $\exp\left(t^T\mu + \frac{1}{2}t^T\Sigma t\right)$, then $X \sim N(\mu, \Sigma)$.

Theorem 2.1.3. If $X \sim N(\mu, \Sigma)$, then $\mathbb{E}(X) = \mu$ and $\operatorname{Var}[X] = \Sigma$.

Proof. Take the first and second order derivative of its mgf. □

Theorem 2.1.4. Suppose that $X \sim N(\mu, \Sigma)$, A is a $m \times p$ non-random matrix, and c is a $m \times 1$ non-random vector. Then

$$AX + c \sim N(A\mu + c, A\Sigma A^T).$$

Corollary 2.1.5.

- Suppose that $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N(\mu, \Sigma)$, in which $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Then

  $$X_1 \sim N_r(\mu_1, \Sigma_{11}), X_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$$

  where r is the dimension of X_1.
- Suppose c is a $p \times 1$ vector, then $c^T X \sim N(c^T\mu, c^T\Sigma c)$.

Theorem 2.1.6. Suppose that we are in the same setting as in Corollary 2.1.5. Then X_1 and X_2 are independent i.f.f. $\operatorname{Cov}(X_1, X_2) = \Sigma_{12} = \Sigma_{21} = 0$.

Corollary 2.1.7. Suppose $X \sim N_p(\mu, \Sigma)$ and $A_1 \in \mathbb{R}^{m\times p}, A_2 \in \mathbb{R}^{s\times p}$, $a_1 \in \mathbb{R}^{m\times 1}$, $a_2 \in \mathbb{R}^{s\times 1}$. Then $A_1X + a_1$ and $A_2X + a_2$ are independent i.f.f. $\operatorname{Cov}(A_1X + a_1, A_2X + a_2) = A_1\operatorname{Var}[X]\, A_2^T = A_1\Sigma A_2^T = 0$.

The next proposition shows that the orthogonal matrix that flips/rotates standard normal does not affect the properties of the distribution.

Proposition 2.1.8. Suppose that $X \sim N_p(0, I)$ and $Y = QX$ with Q is an orthogonal matrix (i.e., $Q^TQ = I$). Then

$$Y = QX \sim N_p(0, I).$$

Proposition 2.1.9 (Additivity property)**.** Suppose that $X_i \overset{\text{i.i.d.}}{\sim} N_p(\mu_i, \Sigma_i)$ for $i = 1, \dots, n$. Then

$$\sum_{i=1}^{n} X_i \sim N_p\left(\sum_{k=1}^{n}\mu_k, \sum_{k=1}^{n}\Sigma_k\right).$$

Definition 2.1.10 (χ^2-distribution)**.** If $X = (X_1, \dots, X_p)^T \sim N_p(0, I_{p\times p})$. Then

$$X^TX = \sum_{k=1}^{p} X_k^2 \sim \chi_p^2.$$

Hence, the quadratic function of normal distribution is related to $\chi-$distribution.

Note that

$$\Sigma = A \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}_{p\times p} A^T$$

where $\lambda_i > 0$.

Define

$$\Sigma^{\frac{1}{2}} = A \begin{bmatrix} \lambda_1^{\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & \lambda_p^{\frac{1}{2}} \end{bmatrix}_{p\times p} A^T$$

and $\Sigma^{\frac{1}{2}}$ is symmetric and is inverse denoted by $\Sigma^{-\frac{1}{2}}$, which is given by

$$\Sigma^{-\frac{1}{2}} = A \begin{bmatrix} \lambda_1^{-\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & \lambda_p^{-\frac{1}{2}} \end{bmatrix}_{p\times p} A^T$$

Theorem 2.1.11. If $X \sim N_p(\mu, \Sigma)$ with Σ being positive definite, then

$$(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_p.$$

Proof. From Theorem 2.1.4, we have $\Sigma^{-\frac{1}{2}}(X-\mu) \sim N_p(0, I)$. Then

$$(X-\mu)^T\Sigma^{-1}(X-\mu) = (\Sigma^{-\frac{1}{2}}(X-\mu))^T\Sigma^{-\frac{1}{2}}(X-\mu) \sim \chi^2_p.$$

□

We will see a generalization result as follows.

Theorem 2.1.12. Suppose that A is a $p \times p$ symmetric matrix with rank r. The following statements hold.

- If $X \sim N_p(0, I)$ and $A^2 = A$, then $X^TAX \sim \chi^2_r$.
- If $X \sim N_p(0, \Sigma)$ with Σ positive definite and $A\Sigma A = A$, then

$$X^TAX \sim \chi^2_r.$$

Proof. (1) Since A is a projection matrix, then $\operatorname{rank}(A) = \operatorname{trace}(A) = r$. Then

$$A = Q\begin{bmatrix} I_r & 0 \\ 0 & 0_{p-r}\end{bmatrix}Q^T.$$

Plug it into X^TAX, it becomes

$$X^TAX = X^TQ\begin{bmatrix} I_r & 0 \\ 0 & 0_{p-r}\end{bmatrix}Q^TX.$$

Let $Y = Q^TX \sim N_p(0, I)$. Hence,

$$X^TAX = Y^T\begin{bmatrix} I_r & 0 \\ 0 & 0_{p-r}\end{bmatrix}Y = \sum_{k=1}^{r} Y_i^2 \sim \chi^2_r.$$

(2) Note that $X^TAX = X^T\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}X$. Let $Y = \Sigma^{-\frac{1}{2}}X \sim N_p(0, I)$ and $B = \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$. Then $X^TAX = Y^TBY$.

Our goal is to use result in (1), so we need to check B is a projection matrix. Clearly, $B^T = B$ and $B^2 = \Sigma^{\frac{1}{2}}A\Sigma A\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} = B$. Moreover, $\operatorname{rank}(B) = rank \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} = \operatorname{rank}(\Sigma A) = \operatorname{rank}(A)$.

It follows from (1) that $X^TAX \sim \chi^2_r$.

□

Remark 2.1.13.

- If $X \sim N_p(\mu, I)$ with $\mu \neq 0$, then $X^TX \sim$ non-central χ^2-distribution.
- If $X \sim \mathcal{N}_p(0, \Sigma)$ and $A = A^T$ with $\operatorname{rank}(A) = r$, Σ positive definite but $A\Sigma A \neq A$, then $X^TAX \sim$ weighted χ^2-distribution.

Definition 2.1.14 (t-distribution)**.** Suppose that $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_n$, then

$$\frac{z}{\sqrt{\frac{x}{n}}} \sim t_n$$

Note that t_n is symmetric and $t_n \to \mathcal{N}(0,1)$ as $n \to \infty$. The t-distribution has a heavier tail than the standard normal.

Example 2.1.15. Given a linear regression model, $Y = X\beta + \varepsilon$ where $Y \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^{p\times 1}$, $\varepsilon \in \mathbb{R}^{n\times 1}$. Assume that data $\{(x_i, y_i)\}_{i=1}^n$ are fixed numbers and $\varepsilon \sim N_n(0, \sigma^2 I_{n\times n})$. Then $Y \sim N_n(X\beta, \sigma^2 I_{n\times n})$ from Theorem 2.1.4. Minimizing mean square error $\mathbb{E}\left[(Y - X\beta)^T(Y - Z\beta)\right]$ to find the 'best' β,

$$\hat{\beta} = (X^TX)^{-1}X^TY \sim N_p(\beta, \sigma^2(X^TX)^{-1})$$

provided by the existence of $(X^TX)^{-1}$. Here $\mathbb{E}\left[\hat{\beta}\right] = (X^TX)^{-1}X^T\mathbb{E}\left[Y\right] = \beta$ (thus an unbiased estimator of β) and $\mathrm{Var}\left[\hat{\beta}\right] = (X^TX)^{-1}X^T\mathrm{Var}\left[\varepsilon\right]X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$.
We have the following results.

- (1) The fitted value is $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY \sim N_n(X\beta, \sigma^2 H)$ where $H = X(X^TX)^{-1}X^T$ (called hat matrix). Since $H = H^T$ and $H^2 = H$ (idempotent), H is a projection matrix. Also, $I - H = (I-H)^T$ and $(I-H)^2 = I - H$, $I - H$ is still a projection matrix.
- (2) The residual is $Y - \hat{Y} = (I-H)Y \sim N_n(0, \sigma^2(I-H))$. Here $\hat{\beta}$ and $Y - \hat{Y}$ are independent because
$$\begin{aligned}\mathrm{Cov}\left(\hat{\beta}, Y - \hat{Y}\right) = \mathrm{Cov}\left(X(X^TX)^{-1}X^TY, (I-H)Y\right) &= X(X^TX)^{-1}X^T\mathrm{Var}\left[Y\right](I-H)\\ &= \sigma^2 X(X^TX)^{-1}X^T(I-H)\\ &= \sigma^2(X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T) = 0\end{aligned}$$
from Corollary 2.1.7.
Similarly, $\hat{Y} = HY$ and $Y - \hat{Y} = (I-H)Y$ are independent since $H(I-H) = H - H^2 = 0$ from Corollary 2.1.7.
- (3) Residual sum of squares (RSS).
Note that $Y - \hat{Y} = Y - HY = (I-H)Y$. Then
$$RSS = \sum_{k=1}^n (y_k - \hat{y}_k)^2 = (Y - \hat{Y})^T(Y - \hat{Y}) = Y^T(I-H)^2Y = (X\beta + \varepsilon)^T(I-H)(X\beta+\varepsilon).$$
Since $(I-H)X = X - X(X^TX)^{-1}X^TX = 0$, then RSS becomes
$$RSS = \varepsilon^T(I-H)\varepsilon = \sigma^2\left(\frac{\varepsilon}{\sigma}\right)^T(I-H)\frac{\varepsilon}{\sigma}$$
Note that $\frac{\varepsilon}{\sigma} \sim N_n(0, I)$, then from Theorem 2.1.12
$$\left(\frac{\varepsilon}{\sigma}\right)^T(I-H)\frac{\varepsilon}{\sigma} \sim \chi^2_{n-p}$$
where $\mathrm{rank}\,(H) = \mathrm{rank}\left(X^TX\right) = p$ then $\mathrm{rank}\,(I-H) = n - p$
Hence,
$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}.$$
- (4) $\hat{\beta}$ and $\hat{\sigma}^2 := \frac{RSS}{n-p}$ are independent, since $\hat{\beta}$ and $Y - \hat{Y}$ are independent.
- (5) Let β_j denote the ith element of β and A^{-1}_{ii} the ith diagonal element of A^{-1}. Then we have
$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{(X^TX)^{-1}_{jj}\hat{\sigma}^2}} \sim t_{n-p}.$$

Indeed, $\hat{\beta}-\beta \sim N(0,\sigma^2(X^TX)^{-1})$ and is independent of $\hat{\sigma}^2$. Then

$$\frac{\hat{\beta}-\beta}{\hat{\sigma}} \sim \frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2\chi^2_{n-p}}{n-p}}} \sim \frac{N(0,(X^TX)^{-1})}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}}$$

So

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{(X^TX)^{-1}_{jj}\hat{\sigma}^2}} \sim \frac{N(0,\sigma^2(X^TX)^{-1}_{jj})}{\sqrt{(X^TX)^{-1}_{jj}\frac{\chi^2_{n-p}}{n-p}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} \sim t_{n-p}.$$

Example 2.1.16. Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $\overline{X} = \frac{1}{n}\sum_{k=1}^n X_k$ and $S_n^2 = \sum_{k=1}^n (X_k - \overline{X}_n)^2$. Then $S_n^2 \sim \chi^2_{n-1}$ and $\overline{X}_n$ and S_n^2 are independent.

2.2 Asymptotic theory

2.2.1 Large-sample theory

Define $\{A_n i.o.\} = \{\omega : \omega$ belongs to infinite infintely many times of events$\} = \cap_{n=1}^\infty \cup_{m\geqslant n} A_m$

Lemma 2.2.1. $X_n \to X$ a.s. iff $\mathbb{P}(|X_n - X| \geqslant, \varepsilon i.o.) = 0$ for every $\varepsilon > 0$.

One way to show that $X_n \to X$ a.s. is that if we have $\sum_{n=1}^\infty \mathbb{P}(|X_n - X| \geqslant \varepsilon) < \infty$, then from Borel-Cantelli Lemma we have $\mathbb{P}(|X_n - X| \geqslant \varepsilon i.o.) = 0$. Apply Lemma 2.2.1, we get the desired result.

Proposition 2.2.2. If $\mathbb{E}(Y_n - Y)^2 \to 0$, then $Y_n \xrightarrow{P} Y$.

Example 2.2.3. Suppose that $X_1, X_2, \dots$ are iid, with mean μ and variance σ^2. Then

$$\mathbb{E}(\bar{X}_n - \mu)^2 = \text{Var}\left[\bar{X}_n\right] = \sigma^2/n \to 0$$

So $\bar{X}_n \xrightarrow{P} \mu$. In fact, $\bar{X}_n \xrightarrow{P} \mu$ even when $\sigma^2 = \infty$ provided $\mathbb{E}|X_i| < \infty$. This result is weak law of large numbers.

Definition 2.2.4. We say $X_n \xrightarrow{D} X$ if $F_n(x) \to F(x)$ for all x, which is continuous of F.

Proposition 2.2.5.

If $X_n \xrightarrow{D} X$ iff $a^T X_n \xrightarrow{D} a^T X$ for all non-zero vectors a.

$X_n \xrightarrow{D} X$ and X has a continuous CDF. Then

$$\lim_{n\to\infty} \sup |F_n(x) - F(x)| = 0$$

Definition 2.2.6 (o_p and O_p). (1) If $a_n/b_n \xrightarrow{P} 0$, then we say $X_n = o_p(a_n)$.

(2) If for any $\varepsilon > 0$, there exists $M(\varepsilon) > 0$ s.t.

$$\mathbb{P}(|\frac{X_n}{a_n}| < M(\varepsilon)) \geqslant 1 - \varepsilon,$$

then we say $X_n = O_p(a_n)$.

Proposition 2.2.7.

- If $X_n \xrightarrow{D} X$, then $X_n = O_p(1)$.
- If $X_n = o_p(a_n)$, then $X_n = O_p(a_n)$.
- If $X_n = O_p(a_n)$ and $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$, then $X_n = o_p(b_n)$.
- If $X_n = O_p(a_n)$, $Y_n = O_p(b_n)$, then $X_n + Y_n = O_p(a_n + b_n)$ and $X_nY_n = O_p(a_nb_n)$.
- If $X_n = O_p(a_n)$, $Y_n = o_p(b_n)$, then $X_nY_n = o_p(a_nb_n)$.

Proof. (3) Apply the last property that if $X_n = O_p(a_n)$,$Y_n = o_p(b_n)$, then $X_nY_n = o_p(a_nb_n)$. From this property, since $X_n = O_p(a_n)$ and $a_n = o_p(b_n)$, then $X_na_n = o_p(a_nb_n)$. This gives the desired result. □

Example 2.2.8. For iid r.v.-s X_i with $\mathbb{E}X_1 = \mu$ and $\operatorname{Var}[X_1] = \sigma^2$ with $\mathbb{E}X_1^2 < \infty$. From C.L.T., $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$. Then

$$\sqrt{n}(\bar{X}_n - \mu) = O_p(1).$$

Hence, $\bar{X}_n - \mu = O_p(n^{-1/2})$.
If $\mu = 0$, $\bar{X} = O_p(n^{-1/2})$. Then $\sum X_i = O_p(n^{1/2})$. Otherwise, $\bar{X}_n = \mu + O_p(n^{-1/2}) = O_p(1) + O_p(n^{-1/2}) = O_p(1)$. Then $\sum X_i = O_p(n)$.

Theorem 2.2.9. $X_n \xrightarrow{D} X$ iff $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions f.

Transformation is an important tool in statistics. If X_n converges to X in some sense, we often need to check whether $g(X_n)$ converges to $g(X)$ in the same sense. The continuous mapping theorem provides an answer to the question in many problems.

Theorem 2.2.10 (Continuous mapping theorem)**.** Let $\{X_n\}$ be a sequence of r.v.-s and $g(x)$ be a continuous function from $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then

- If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.
- If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.
- If $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$.

Proof. (2) For the special case of $X = c$. Since f is continuous at c, given any $\varepsilon > 0$, there exists $\delta > 0$ s.t. $|f(X_n) - f(c)| \le \varepsilon$ whenever $|X_n - c| \le \delta$. Thus,

$$\mathbb{P}(|X_n - c| \le \delta) \le \mathbb{P}(|f(X_n) - f(c)| \le \varepsilon)$$

which implies

$$\mathbb{P}(|f(X_n) - f(c)| \geqslant \varepsilon) \le \mathbb{P}(|X_n - c| \le \delta) \to 0.$$

(3) For any bounded and continuous functions f, $f \circ g$ is also bounded and continuous. Since $X_n \xrightarrow{D} X$, from Theorem 2.2.9

$$\mathbb{E}f(g(X_n)) \to \mathbb{E}f(g(X))$$

Since f is arbitrary, apply Theorem 2.2.9 for $g(X_n)$ again, we have $g(X_n) \xrightarrow{D} g(X)$. □

Remark 2.2.11.

- If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $X_n + Y_n \xrightarrow{a.s.} X + Y$ and $\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{a.s.} \begin{bmatrix} X \\ Y \end{bmatrix}$.
- If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$ and $\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{P} \begin{bmatrix} X \\ Y \end{bmatrix}$.
- If $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{D} Y$, and $\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} X \\ Y \end{bmatrix}$, then $X_n + Y_n \xrightarrow{D} X + Y$.

One counterexample about the sum of r.v.-s are not convergence in distribution: let $X_n = X \sim N(0, 1)$

and $Y_n = X$. Assume that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} -X$. Then $X_n + Y_n = 2X$ does not converge in distribution to $X + Y = 0$.

Theorem 2.2.12 (Slutsky's theorem)**.** Let X, Y be two r.v.-s and two sequences of r.v.-s $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$. Suppose that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$. Then

- $X_n + Y_n \xrightarrow{D} X + c$
- $Y_n X_n \xrightarrow{D} cX$
- $X_n / Y_n \xrightarrow{D} X/c$ for $c \neq 0$.

Proof. (1) Let $t \in \ R$ and $\varepsilon > 0$. Then

$$\begin{aligned} F_{X_n+Y_n}(t) = \mathbb{P}(X_n + Y_n \leq t) &= \mathbb{P}(X_n + Y_n \leq t, |Y_n - c| < \varepsilon) + \mathbb{P}(X_n + Y_n \leq t, |Y_n - c| \geqslant \varepsilon) \\ &\leq \mathbb{P}(X_n \leq t - c + \varepsilon) + \mathbb{P}(|Y_n - c| \geqslant \varepsilon) \end{aligned}$$

where the second term $\mathbb{P}(|Y_n - c| \geqslant \varepsilon) \to 0$.

Also,

$$\begin{aligned} \mathbb{P}(X_n \leq t - c - \varepsilon) &= \mathbb{P}(X_n \leq t - c - \varepsilon, |Y_n - c| \geqslant \varepsilon) + \mathbb{P}(X_n \leq t - c - \varepsilon, |Y_n - c| \leq \varepsilon) \\ &\leq \mathbb{P}(|Y_n - c| \geqslant \varepsilon) + \mathbb{P}(X_n + Y_n \leq t) \end{aligned}$$

Thus,

$$F_{X_n+Y_n}(t) \geqslant \mathbb{P}(X_n \leq t - c - \varepsilon) - \mathbb{P}(|Y_n - c| \geqslant \varepsilon).$$

Choose ε, c s.t. $t - c$, $t - c + \varepsilon, t - c - \varepsilon$ be continuity points of F_X, then it follows from the precious inequalities and hypotheses of theorem

$$F_X(t - c - \varepsilon) \leq \liminf_{n\to\infty} F_{X_n+Y_n}(t) \leq \limsup_{n\to\infty} F_{X_n+Y_n}(t) \leq F_X(t - c + \varepsilon).$$

Let $\varepsilon \to 0$, then

$$\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_X(t - c)$$

where $F_{X+c}(t) = F_X(t - c)$.

(2) If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$, then $Y_n - c = o_p(1)$ and $X_n = O_p(1)$. Apply Theorem 2.2.10, we have

$$X_n Y_n = cX_n + X_n(Y_n - c) = cX_n + o_p(1) \xrightarrow{D} cX.$$

(3) If $\frac{1}{Y_n} \top \frac{1}{c}$, then $\frac{X_n}{Y_n} = X_n \frac{1}{Y_n} \to \frac{1}{c} X$ for $c \neq 0$. □

Theorem 2.2.13 (CLT)**.** Suppose $X_1, X_2, \dots$ are iid with mean μ and variance σ^2. Then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Moreover, for multivariate case we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \Sigma).$$

As an application of this result, let F_n denote the CDF of $\sqrt{n}(\bar{X}_n - \mu)$ and note that

$$\begin{aligned} \mathbb{P}(|\sqrt{n}(\bar{X}_n - \mu)| \geqslant a) &= \mathbb{P}(\mu - \frac{a}{\sqrt{n}} < \bar{X}_n \leq \mu + \frac{a}{\sqrt{n}}) \\ &= F_n(a) - F_n(-a) \to \Phi(a/\sigma) - \Phi(-a/\sigma). \end{aligned}$$

Theorem 2.2.14 (Lyapunov condition)**.** Let $X_1, X_2, \dots$ be a sequence of independent r.v.-s. Assume that $\mathbb{E}X_i = \mu_i$ and $\mathrm{Var}[X_i] = \sigma_i^2$. If

- $\mathbb{E}X_i < \infty$ and $\mathrm{Var}[X_i] < \infty$ exist

- There exists $\delta > 0$ s.t.

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbb{E}\left[|X_k-\mu_k|^{2+\delta}\right]}{(\sum_{k=1}^{n} \sigma_k^2)^{1+\delta/2}} = 0$$

Then

$$\frac{\sum_{k=1}^{n}(X_k-\mu_k)}{\sqrt{\sum_{k=1}^{n}\sigma_k^2}} \xrightarrow{D} N(0,1).$$

Theorem 2.2.15 (Triangular array CLT). Let $\{X_{ni}\}_{i=1,\dots,n,n=1,2,\dots}$ be a sequence of r.v.-s. Suppose that

- For each k, $X_{k1},\dots,X_{kn}$ are independent
- Lyapunov condition are satisfied for $X_{k1},\dots,X_{kn}$.

Then

$$\frac{\sum_{k=1}^{n}(X_{nk}-\mathbb{E}X_{nk})}{\sqrt{\sum_{k=1}^{n}\mathrm{Var}\left[X_{nk}\right]}} \xrightarrow{D} N(0,1).$$

The CLT stated only provides direct information about distribution of averages. Many estimators in statistics are not exactly averages, but can be related to averages in some fashion. In some of these cases, clever use of the CLT still provides a limit theorem for an estimator's distribution. A first possibility would be for variables that are smooth functions of an average and can be written as $f(\bar{X}_n)$. Thy Taylor approximation

$$f(\bar{X}_n) \approx f(\mu) + f'(\mu)(\bar{X}_n-\mu)$$

with CLT motivates the following propostion.

Proposition 2.2.16 (Delta Method). Let $X_1, X_2, \dots$ and Y be random $k-$vectors satisfying

$$a_n(X_n-c) \xrightarrow{D} Y,$$

where $c \in \mathbb{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$. Let f be a function from $C^1([c-\varepsilon, c+\varepsilon])$ for $\varepsilon > 0$. Then

$$a_n(f(X_n)-f(c)) \xrightarrow{D} f'(c)Y.$$

Proof. Note that $X_n - c = O_p(\frac{1}{a_n}) = o_p(1)$. From Mean value theorem, there exists $\xi \in (X_n, c)$ s.t.

$$a_n(f(X_n)-c) = a_n f'(\xi_n)(X_n-c)$$

Since $X_n - c = o_p(1)$ then

$$\mathbb{P}(|\xi_n-\mu| \geqslant \varepsilon) \leq \mathbb{P}(|X_n-\mu| \geqslant \varepsilon) = 0$$

Thus, $\xi_n = c + o_p(1)$.

From Theorem 2.2.10 we have

$$f'(\xi_n) \xrightarrow{P} f'(c)$$

Hence, from Theorem 2.2.12

$$a_n(f(X_n)-c) = f'(\xi_n)a_n(X_n-c) \xrightarrow{D} f'(c)Y.$$

□

Corollary 2.2.17. With the same assumptions in Theorem 2.2.13, if f is differentiable at μ, then

$$\sqrt{n}(f(\bar{X}_n)-f(\mu)) \xrightarrow{D} N(0,[f'(\mu)]^2\sigma^2).$$

Moreover, for multivariate case $X \sim N_k(\mu,\Sigma)$ then

$$\sqrt{n}(f(\bar{X}_n)-f(\mu)) \xrightarrow{D} N(0,[f'(\mu)]^T\Sigma[f'(\mu)]).$$

Example 2.2.18. Let $X_1, \ldots, X_n$ be a iid r.v.-s with mean 0 and variance 1.

- Let $f(x) = \sin(x)$. Using Delta Method,
$$\sqrt{n}(\sin(\bar{X}_n - \sin(0)) \xrightarrow{D} N(0,1).$$
- Let $f(x) = \cos(x)$. Since the first order derivative of $\cos(x)$ at 0 is 0 (not differentiable at 0). We need to expand its for higher order by Taylor's expansion. Note that
$$\cos(\bar{X}_n) - \cos(0) = -\sin(0)\bar{X}_n + \frac{1}{2}(-\cos(0)\bar{X}_n^2 + \frac{1}{3!}\sin(\xi_3)\bar{X}_n^3 = -\frac{1}{2}\bar{X}_n^2 + \frac{1}{6}\sin(\xi_3)\bar{X}_n^3$$
where $\sin(\xi_3) = O_p(1)$ and $(\bar{X}_n)^3 = O_p(n^{-3/2})$. Thus,
$$-2n(\cos(\bar{X}_n) - 1) = n\bar{X}_n^2 + O_p(n^{-1/2}) \xrightarrow{D} \chi_1^2.$$

Example 2.2.19. Let $\{X_n\}$ be a sequence of r.v.-s satisfying
$$\sqrt{n}(X_n - c) \xrightarrow{D} N(0,1)$$
Let $f(x) = x^2$. If $c \neq 0$, then
$$\sqrt{n}(f(X_n) - f(c)) = \sqrt{n}(X_n^2 - c^2) \xrightarrow{D} N(0, 4c^2).$$
If $c = 0$, the first order derivative of f is zero but the second-order is 2. Then
$$f(X_n) - f(0) = \frac{1}{2}f''(0)X_n^2 = X_n^2$$
Thus, $nf(X_n) = nX_n \xrightarrow{D} \chi_1^2$.

Example 2.2.20 (Ratio estimator)**.** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid bivariate r.v.-s with finite 2nd order moments. Let $\mu_x = \mathbb{E}X_1, \mu_y = \mathbb{E}Y_1 \neq 0$ and $\sigma_x^2 = \mathrm{Var}\,[X_1]\,, \sigma_y^2 = \mathrm{Var}\,[Y_1]$. By CLT,
$$\sqrt{n}\left(\begin{bmatrix}\bar{X}_n \\ \bar{Y}_n\end{bmatrix} - \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}\right) \xrightarrow{D} N_2\left(0, \begin{bmatrix}\sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2\end{bmatrix}\right)$$
From Delta Method, let $f(x,y) = \frac{x}{y}$. Then $\frac{\partial f}{\partial x} = y^{-1}$ and $\frac{\partial f}{\partial y} = -xy^{-2}$.
$$\sqrt{n}\left(\frac{\bar{X}_n}{\bar{Y}_n} - \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}\right) \xrightarrow{D} N(0, \sigma^2)$$
where
$$\sigma^2 = [\mu_y^{-1}, -\mu_x\mu_y^{-2}]\begin{bmatrix}\sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2\end{bmatrix}\begin{bmatrix}\mu_y^{-1} \\ -\mu_x\mu_y^{-2}\end{bmatrix} = \frac{\sigma_x^2}{\sigma_y^2} - \frac{\mu_x\sigma_{xy}}{\mu_y^3} + \frac{\mu_x^2\sigma_y^2}{\mu_y^4}.$$

2.3 Parametric likelihood method

Many estimators in statistics are specified implicityly as solutions to equations or as values maximizing some function. In this section we study why these methods work and learn ways to approximate distributions. Although we focus on methods for iid observations. many of the ideas can be extended. The first example concerns maximum likelihood estimation. The maximum likelihood estimator $\widehat{\theta}$ maximizes the likelihood function $L(\cdot)$ or log-likelihood $l(\cdot) = \log L(\cdot)$. And if l is differentiable and the maximum occurs in the interior of the parameter space, then $\widehat{\theta}$ solves $\nabla l(\theta) = 0$.

Setup: given data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ with pdf or pmf $f(x;\theta)$, where $f(x;\theta)$ is known up to θ. We use θ_0 to denote the true value of θ and $\widehat{\theta}_n$ to denote estimator of θ. Let l_n be the log-likelihood function for the first n observations:
$$l_n(\omega) = \log\prod_{i=1}^n f_\omega(X_i) = \sum_{k=1}^n \log f_\omega(X_i).$$
(We use ω as the dummy argument here, reserving θ to represent the true value of the unknown parameter

in the sequel.) Then the MLE estimator $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ from the first n observations will maximize l_n. For regularity, assume $f_\theta(x)$ is continuous in θ.

Definition 2.3.1 (Consistent and Strong consistent)**.** We say $\widehat{\theta}$ is strongly consistent if

$$\widehat{\theta}_n \xrightarrow{a.s.} \theta_0.$$

We say $\widehat{\theta}$ is consistent if

$$\widehat{\theta}_n \xrightarrow{P} \theta_0.$$

The MLE estimator is given by

$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta} \text{ł}_n(\theta)$$

Our hypothesis test is that for parameter space $\Theta = \Theta_0 \cup \Theta_0^c$

$$H_0 : \theta \in \Theta_0,\ H_a : \theta \in \Theta_0^c$$

The likelihood ratio test (LRT) is given by

$$R_n = 2\{\sup_{\theta} l_n(\theta \in \Theta) - \sup_{\theta \in \Theta_0} l_n(\theta)\}$$

which is the unrestricted maximum minus the restricted maximum.

Our main result is the limiting distribution of the MLE as follows.

Theorem 2.3.2. Assume that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta), \theta \in \Theta$. Suppose that the following regularity conditions hold.

- $R_0 : \widehat{\theta_n} \xrightarrow{a.s.} \theta_0$ (i.e., MLE is consistent);
- $R_1 : \theta_0$ is an interior point of Θ and support of X does not dependent on θ;
- $R_2 : f(x;\theta) \in \mathcal{C}^3$ w.r.t. θ for given x;
- R_3 : for each $\theta_0 \in \Theta$, there exist $h(x), H(x)$ (depending on θ_0) s.t. in a small neighborhood of θ_0
$$\left|\frac{\partial f(x;\theta)}{\partial \theta_j}\right| \le h(x),\ \left|\frac{\partial^2 f(x;\theta)}{\partial \theta_j \theta_l}\right| \le h(x)$$
and
$$\left|\frac{\partial^3 \log f(x;\theta)}{\partial \theta_j \partial \theta_k \partial \theta_l}\right| \le H(x)$$
s.t. $\mathbb{E}H(x) < \infty$ and $\mathbb{E}H(x) < \infty$.
- $R_4 : J(\theta_0)$ is positive definite.

Then for any θ in the interior of Θ,

- $\widehat{\theta}_n \xrightarrow{a.s.} \theta$,
- $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, 1/J(\theta))$,
- $R_n \xrightarrow{D} \chi^2_{df}$.

where $J(\theta) = \mathbb{E}\left[((\log f(x;\theta))')^2\right] = -\mathbb{E}\left[(\log f(x;\theta))''\right]$.

Remark 2.3.3. R_1 is the necessary condition for exchanging the order of integration and derivative. R_0 and R_1 ensure $S_n(\widehat{\theta}_n) = 0$. $R_1 - R_3$ ensure that we can exchange the order of integration and derivative. $R_1 - R_4$ ensure that

$$0 = S_n(\theta_0) + \frac{\partial S_n(\theta_0)}{\partial \theta^T}(\widehat{\theta}_n - \theta_0) + o_p(n)(\widehat{\theta}_n - \theta_0).$$

2.3.1 Consistency of the Maximum Likelihood Estimator

Definition 2.3.4 (Kullback-Leibler information)**.** The Kullback-Leibler information is defined as

$$I(\theta,\omega)=\mathbb{E}_\theta \log[f(x;\theta)/f(x;\omega)].$$

It can be viewed as a measure of the information discriminating between θ and ω when θ is the true value of the unknown parameter.

Lemma 2.3.5. If $P_\theta \neq P_\omega$, then $I(\theta,\omega)>0$.

Proof. By Jensen's inequality ($-\log(\cdot)$ is strictly convex function),

$$\begin{aligned}-I(\theta,\omega)&=E_\theta \log[f_\omega(X)/f_\theta(X)]\\ &\le \log E_\theta[f_\omega(X)/f_\theta(X)]\\ &=\log\int_{f_\theta>0}\frac{f(x;\omega)}{f(x;\theta)}f(x;\theta)d\mu(x)\\ &\le \log 1=0.\end{aligned}$$

Strict equality will occur only if $f(x;\omega)/f(x;\theta)$ is constant a.s. □

Based on Wald's idea (1949), we can prove the strong consistency that $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$ under some mild conditions.

2.3.2 Limiting distribution for the MLE

Assume that $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$ and θ_0 is an interior point of Θ. We will prove the second result as in Theorem 2.3.2.

Problem: given $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta)$ with $\theta=\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}$ and $\dim\theta_1=m$, $\dim\theta_2=p-m$. Let the true value is $\theta_0=\begin{bmatrix}\theta_{10}\\ \theta_{20}\end{bmatrix}$.

Our interest:

$$H_0:\theta_1=\theta_{10}\,,H_a:\theta_1\neq\theta_{10}$$

Let

$$\widehat{\theta_n}=\underset{\theta}{\operatorname{argmax}}\, l_n(\theta)$$

and

$$\widetilde{\theta_n}=\underset{\theta\,,\theta_1=\theta_{10}}{\operatorname{argmax}}\, l_n(\theta).$$

We want to show that the limiting distribution of the likelihood ratio test is given by

$$R_n=2\left(l_n(\widehat{\theta_n})-l_n(\widetilde{\theta_n})\right)\xrightarrow{D}\chi^2_m.$$

under condition R_0-R_4.

Here we apply the 2nd-order Taylor expansion for LRT. That is

$$R_n=2\left(\left(l_n(\widehat{\theta_n})-l_n(\theta_0)\right)-\left(l_n(\widetilde{\theta_n})-l_n(\theta_0)\right)\right)$$

We first approximate of $\left(l_n(\widehat{\theta}_n) - l_n(\theta_0)\right)$. Note that

$$\begin{aligned}
l_n(\widehat{\theta}_n) - l_n(\theta_0) &= \frac{\partial l_n(\theta_0)}{\partial \theta}(\widehat{\theta}_n - \theta_0) + \frac{1}{2}\left(\widehat{\theta}_n - \theta_0\right)^T \frac{\partial^2 l_n(\theta_0)}{\partial\theta\partial\theta^T}\left(\widehat{\theta}_n - \theta_0\right) + e_n \\
&= \underbrace{S_n^T(\theta_0)}_{O_p(n^{1/2})} \underbrace{\left(\widehat{\theta}_n - \theta_0\right)}_{O_p(n^{-1/2})} + \frac{1}{2}\underbrace{\left(\widehat{\theta}_n - \theta_0\right)}_{O_p(n^{-1/2})}(-nJ(\theta_0) + \underbrace{o_p(n)}_{O_p(n)})\underbrace{\left(\widehat{\theta}_n - \theta_0\right)}_{O_p(n^{-1/2})} + e_n \\
&= \underbrace{S_n^T(\theta_0)}_{O_p(n^{1/2})} \underbrace{\left(\widehat{\theta}_n - \theta_0\right)}_{O_p(n^{-1/2})} - \frac{n}{2}\underbrace{\left(\widehat{\theta}_n - \theta_0\right)}_{O_p(n^{-1/2})} J(\theta_0)\underbrace{\left(\widehat{\theta}_n - \theta_0\right)}_{O_p(n^{-1/2})} + o_p(1) + e_n
\end{aligned}$$

where $O_p(n^{-1})o_p(n) = o_p(1)$.

Next, we show that $e_n = o_p(1)$. Without loss of generality , assume that θ is 1-dim. Then

$$|e_n| \le \frac{1}{6}\sum_{i=1}^n \left|\frac{\partial^3 \log f(x;\theta)}{\partial\theta^3}\left(\widehat{\theta}_n - \theta_0\right)^3\right| \le \frac{1}{6}\sum_{i=1}^n H(x_i)\left|\widehat{\theta}_n - \theta_0\right|^3 = O_p(n)O_p(n^{-2/3}) = O_p(n^{-1/2}) = o_p(1).$$

Plug in $\widehat{\theta}_n - \theta_0 = \frac{1}{n}J^{-1}(\theta_0)S_n(\theta_0) + o_p(n^{-1/2})$, we obtain

$$\begin{aligned}
l_n(\widehat{\theta}_n) - l_n(\theta_0) &= \frac{1}{n}S_n^T J^{-1} S_n - \frac{n}{2}\left(\frac{1}{n}J^{-1}(\theta_0)S_n(\theta_0) + o_p(n^{-1/2})\right)^T J\left(\frac{1}{n}J^{-1}(\theta_0)S_n(\theta_0) + o_p(n^{-1/2})\right) + o_p(1) \\
&= \frac{1}{n}S_n^T J^{-1}S_n - \frac{1}{2n}S^T J^{-1}S + o_p(1) \\
&= \frac{1}{2n}S_n(\theta_0)^T J^{-1}(\theta_0)S_n(\theta_0) + o_p(1).
\end{aligned}$$

For $\left(l_n(\widetilde{\theta_n}) - l_n(\theta_0)\right)$ part, let $\widetilde{\theta}_n = \begin{bmatrix}\theta_{10}\\ \widetilde{\theta_{2n}}\end{bmatrix}$ where $\widetilde{\theta_{2n}} = \operatorname{argmax}_{\theta_2} l_n(\theta_{10}, \theta_2)$. This means $\widetilde{\theta_{2n}}$ is the MLE of θ_2 under $\theta_1 = \theta_{10}$. Write $l_n(\theta) = l_n(\theta_1, \theta_2)$.

We need to approximate for $\widetilde{\theta_{2n}} - \theta_{20}$. Define

$$S_n(\theta_0) = \begin{bmatrix} S_{1n}(\theta_{10}, \theta_{20}) \\ S_{2n}(\theta_{10}, \theta_{20})\end{bmatrix}$$

Note that

$$S_{2n}(\theta_{10}, \widetilde{\theta_{2n}}) = \frac{\partial l_n(\theta_{10}, \widetilde{\theta_{2n}})}{\partial\theta_2} = 0.$$

Define

$$J(\theta) = \begin{bmatrix} J_{11}(\theta) & J_{12}(\theta) \\ J_{21}(\theta) & J_{22}(\theta)\end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[\frac{f_{\theta_1}(x;\theta)[f_{\theta_1}(x;\theta)]^T}{f(x;\theta)^2}\right] & \mathbb{E}\left[\frac{f_{\theta_1}(x;\theta)[f_{\theta_2}(x;\theta)]^T}{f(x;\theta)^2}\right] \\ \mathbb{E}\left[\frac{f_{\theta_2}(x;\theta)[f_{\theta_1}(x;\theta)]^T}{f(x;\theta)^2}\right] & \mathbb{E}\left[\frac{f_{\theta_2}(x;\theta)[f_{\theta_1}(x;\theta)]^T}{f(x;\theta)^2}\right]\end{bmatrix}$$

where $f_{\theta_j}(x;\theta) = \frac{\partial f(x;\theta)}{\partial\theta_j}$ for $j = 1, 2$.

Similarly, we have

$$\widetilde{\theta_{2n}} - \theta_{20} = \frac{1}{n}J_{22}^{-1}S_{2n}(\theta_0) + o_P(n^{-1/2}).$$

Then we do the same thing as in $l_n(\widehat{\theta}_n) - l_n(\theta_0)$, we get

$$l_n(\widetilde{\theta_n}) - l_n(\theta_0) = \frac{1}{2n}S_{2n}^T(\theta_0)[J_{22}(\theta_0)]^{-1}S_{2n}(\theta_0) + o_p(1).$$

Note that

$$S_{2n}(\theta_0) = [0, I]\begin{bmatrix} S_{1n}(\theta_{10}, \theta_{20}) \\ S_{2n}(\theta_{10}, \theta_{20})\end{bmatrix} = [0_{(p-m)\times m}, I_{(p-m)\times(p-m)}]S_n(\theta_0) = AS_n(\theta_0),$$

where A is of size $(p-m)\times p$ and S_n of size $p \times 1$.

Above all, the LRT becomes

$$\begin{aligned} R_n &= \frac{1}{2n} S_n(\theta_0)^T J^{-1}(\theta_0) S_n(\theta_0) - \frac{1}{2n}(AS_n(\theta_0))^T(\theta_0)[J_{22}(\theta_0)]^{-1} AS_n(\theta_0) + o_p(1) \\ &= \frac{1}{n} S_n(\theta_0)^T \left(J^{-1}(\theta_0) - A^T J_{22}^{-1}(\theta_0) A\right) S_n(\theta_0) + o_p(1) \\ &= \left(\frac{1}{\sqrt{n}} S_n(\theta_0)\right)^T \left(\underbrace{J^{-1}(\theta_0) - A^T J_{22}^{-1}(\theta_0) A}_{=:B}\right) \left(\frac{1}{\sqrt{n}} S_n(\theta_0)\right) + o_p(1) \end{aligned}$$

where $X \sim N(0, J(\theta_0))$.

Apply Theorem 2.1.12, we need to check the condition that

- B is a $p \times P$ symmetric matrix
- what is rank (B)?
- $BJB = B$

If all these conditions hold, then

$$R_n \xrightarrow{D} \chi_m^2.$$

Indeed,

$$BJB = \left(J^{-1}(\theta_0) - A^T J_{22}^{-1}(\theta_0) A\right) \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \left(J^{-1}(\theta_0) - A^T J_{22}^{-1}(\theta_0) A\right) = B - A^T J_{22}^{-1} A + A^T J_{22}^{-1} AJA^T J_{22}^{-1} A = B$$

where note that $AJA^T = J_{22}$.

Also, rank $(B) =$ rank $(B)\, J =$ rank $\left(I - A^T J_{22}^{-1} AJ\right)$. Note that

$$A^T J_{22}^{-1} AJ = \begin{bmatrix} 0 & 0 \\ J_{21}^{-1} J_{21} & I_{(p-m)\times(p-m)} \end{bmatrix}$$

Then

$$\operatorname{rank}(B) = \operatorname{rank}\left(\begin{bmatrix} I_{m\times m} & 0 \\ & I_{(p-m)\times(p-m)} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ J_{21}^{-1} J_{21} & I_{(p-m)\times(p-m)} \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} I_{m\times m} & 0 \\ J_{21}^{-1} J_{21} & 0 \end{bmatrix}\right) = m.$$

Problem: given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta)$ with θ and. Let the true value be θ_0. Given a matrix A of size $m \times p$ with rank $(A) = m \le p$.

Our interest:

$$H_0 : A\theta = 0\,, H_a : A\theta \neq 0.$$

Let

$$\widehat{\theta_n} = \operatorname*{argmax}_{\theta} l_n(\theta)$$

and

$$\widetilde{\theta_n} = \operatorname*{argmax}_{A\theta=0} l_n(\theta).$$

Next, we want to show that

$$R_n = 2\left(l_n(\widehat{\theta}_n) - l_n(\widetilde{\theta_n})\right) \xrightarrow{D} \chi_m^2.$$

For a given A, we can find a matrix B of size $(p-m)\times p$ s.t. $C = \begin{bmatrix} A \\ B \end{bmatrix}_{p\times p}$ has full rank p.

Let $\eta_1 = A\theta, \eta_2 = B\theta$ and $\theta = C^{-1}\eta$. Then it is equivalent to test

$$H_0 : \eta_1 = 0\,, H_a : \eta_1 \neq 0$$

based on data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta) = f(x; C^{-1}\eta) = g(x;\eta)$.

For a non-linear transformation $\eta_1 = g_1(\theta), \eta_2 = g_2(\theta)$, we need to make sure from θ to $[\eta_1, \eta_2]^T$ is 1-1. Then the LRT is still χ_m^2 from the above Problem what we have seen.

Assignments problems 1 and 2 see PDF in IPad.

2.4 Empirical CDF

Compare with parametric model: given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta)$ where θ is unknown and $\dim(\theta) = p < \infty$. We use parametric likelihood method for parametric model.

Next, we will study the non-parameter model and nonparametric likelihood. Consider $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F(x)$ where the CDF $F(x)$ is unknown. Now, our parametric space is $\{F(x) : F(x) \text{ is CDF}\}$, which is an infinite dimensional space.

2.4.1 Empirical CDF and properties

A natural estimator of population CDF $F(x) = \mathbb{E}\left[\mathbf{1}_{-\infty,x}(X)\right]$ is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x)}(X_i).$$

which is called **empirical CDF.** When $X_i^{\prime}s$ are vectors, $I(X_i \leq x) = I(X_{i1} \leq x_1, \ldots, X_{ip} \leq x_p)$, $X_i = [X_{i1}, \ldots, X_{ip}]^T$, and $x = [x_1, \ldots, x_p]^T$.

Note that $F_n(x)$ is a CDF and assign $\frac{1}{n}$ probability to each X_i. So $\sum_{i=1}^{n} I(X_i \leq x) \sim Bin(n, F(x))$. For a given x, $F_n(x)$ is a iid average and hence $F_n(x)$ is a random function or stochastic processes.

Note that the variance of $F_n(X)$ is $\frac{1}{n}F(x)[1 - F(x)]$.

Indeed, note that $\mathbb{E}F_n(X) = \frac{1}{n}\sum\limits_{i=1}^{n} P(X_I \leqslant X) = F(x)$ (the empirical CDF is an unbiased estimate) and $\mathbb{E}F_n(X)^2 = \frac{1}{n^2}\sum\limits_{i,j=1}^{n} P(X_i \leqslant x, X_j \leqslant x)$. Splitting the sum into the parts where $i = j$ and the one where $i \neq j$ we get

$$\mathbb{E}F_n(X)^2 = \frac{1}{n^2}[nF(x) + n(n-1)F(x)^2],$$

since X_i, X_j are independent for $i \neq j$. Hence, the variance of $F_n(X)$ is $\mathbb{E}\left[F_n(X)^2 - (\mathbb{E}F_n(X))^2\right] = \frac{1}{n}[F(x) - F(x)^2]$.

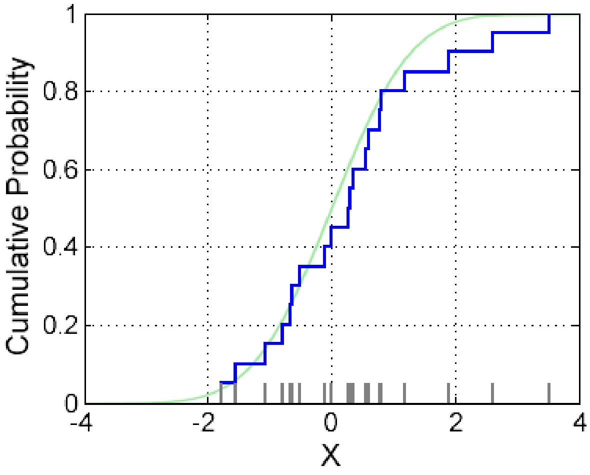


Figure 2.1: Illustration of empirical CDFs. As size n grows, we see F_n approaches F. From SLLN, $F_n(x) \xrightarrow{a.s.} F(x)$. Thus, the estimator $F_n(x)$ is consistent.

Proposition 2.4.1. For each fixed x,

- $F_n \xrightarrow{a.s.} F(x)$;
- $\sqrt{n}(F_n(x) - F(x)) \xrightarrow{D} N(0, F(x)(1-F(x))$;
- $F_n(x) - F(x) = O_p(n^{-1/2})$.

Proof. Since $F_n(x)$ is the sample mean of $Y_i = I(X_i \leq x)$ and $Y_i's$ are iid r.v.-s, they have finite moments to any order. The first property follows from SLLN.

The second property follows from CLT and the third result is directly from second property. Another proof of property 3 can be done using Chebyshev's inequality:

$$\mathbb{P}(\sqrt{n}|F_n(x) - F(x)| \geqslant t) \leq \frac{\text{Var}\,(F(x))}{t^2}$$

and let $t \to \infty$. □

Recall that if $F(x)$ is continuous, then the convergence of $F_n(x)$ at every x implies the uniform convergence in x[6]. That is,

$$D_n = \|F_n - F\|_\infty := \sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

The statistic D_n is called the **Kolmogorov-Smirnov** distance (also a sup-norm) and it is used for the goodness of fit test.

Why are uniform convergence results interesting and important? In statistical settings, a typical use of the empirical CDF is to construct estimators of various quantities associated withe the population CDF. Many such estimation problems can be formulated in a terms of functional γ that maps any CDF F to a real number $\gamma : F \mapsto \gamma(F)$. Given a set of samples distributed according to F, the plug-in principle suggest replacing the unknown F with the empirical CDF F_n, thereby obtaining $\gamma(F_n)$ as an estimate of $\gamma(F)$. For any plug-in estimator $\gamma(F_n)$, an important question is to understand when it is consistent, that is, when does $\gamma(F_n)$ converge to $\gamma(F)$ in probability (or a.s.)?

Indeed, this question can be addressed in a continuous of functional γ at F w.r.t. sup-norm. Given a pair of CDFs F and G, we say the functional γ is **continuous** at F in the sup-norm, if for all $\varepsilon > 0$, there exists $\delta > 0$ s.t. $\|G - F\|_\infty \leq \delta$ implies that $|\gamma(G) - \gamma(F)| \leq \varepsilon$. We can show that given iid samples with CDF F, if γ is continuous in the sup-norm at F, then $\gamma(F_n) \xrightarrow{P} \gamma(F)$.[7]

Hence, for any continuous functional, it reduces the consistency question for the plug-in estimator $\gamma(F_n)$ to the issue of whether or not the random variable $\|F_n - F\|_\infty$ converges to zero. The following classical result (known as the Glivenko-Cantelli theorem, addresses the latter question.

Theorem 2.4.2 (Glivenko-Cantelli theorem). For any distribution, the empirical CDF F_n is a strongly consistent estimator of the population CDF in the uniform norm, that is

$$\|F_n - F\|_\infty \xrightarrow{a.s.} 0.$$

Proof. Indeed, Dvoretsky, Kiefer and Wolfowitz (1956) prove an asymptotic inequality of F_n that there is a constant C independent of F and n s.t.

$$\mathbb{P}(\|F_n - F\|_\infty \geqslant \delta) \leq Ce^{-2n\delta^2}\ , \text{ for all } \delta \geqslant 0$$

Massart (1990) establishes the sharpest possible result with $C = 2$.

Then from Borel-Cantelli Lemma, $\sum_{n=1}^\infty \mathbb{P}(\|F_n - F\|_\infty \geqslant \delta) \leq \sum_{n=1}^\infty Ce^{-2n\delta^2} < \infty$, we have $\mathbb{P}(\|F_n - F\|_\infty \geqslant \delta\,, i.o.) = 0$. Hence, $\|F_n - F\|_\infty \xrightarrow{a.s.} 0$. □

[6] This is from Dini's Theorem: Let (f_n) a sequence of function defined on $[a, b]$ and **increasing respect the variable**, if pointwise limit is continous then the convergence is uniformly. Note that in this case continuity of (f_n)'s isn't required.

[7] [34, Exercise 4.1]

Remark 2.4.3. In fact, we have $\mathbb{P}(\sqrt{n}\,\|F_n - F\|_\infty \geqslant \delta) \leq Ce^{-2\delta^2}$. Then for $\varepsilon > 0$, there exists $M > 0$ s.t. $\mathbb{P}(\sqrt{n}\,\|F_n - F\|_\infty \geqslant M) \leq \varepsilon$. Hence, $\sqrt{n}\sqrt{n}\,\|F_n - F\|_\infty = O_p(1)$, thereby $\sqrt{n}\,\|F_n - F\|_\infty = O_p(n^{-1/2})$.

Let's see some examples of $\gamma(F)$.

Example 2.4.4. Given some integrable function g, we may define the expectation functional γ_g via

$$\gamma_g(F) := \int g(x)\mathrm{d}F(x)$$

For instance, for the function $g(x) = x$, the functional γ_g maps F to $\mathbb{E}X$. Its sample mean $\frac{1}{n}\sum_{i=1}^n X_i$ as an estimate of the mean $\mathbb{E}X$. For any g, the plug-in estimate in given by $\gamma_g(F_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)$, corresponding to the sample mean of $g(X)$.

Example 2.4.5. A special case of example 2.4.4: For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x;\theta)$, recall that $\theta_0 = \operatorname{argmax}_\theta \mathbb{E}\left[\log f(x;\theta)\right]$. Here $\mathbb{E}\left[\log f(x;\theta)\right] = \int \log f(x;\theta)dF$. The plug-in estimator of $\mathbb{E}\left[\log f(x;\theta)\right]$ is $\frac{1}{n}\sum_{i=1}^n \log f(x_i;\theta)$.

2.4.2 Sample moment

In this section, we are interested in estimating of $k-$th raw moment $\mathbb{E}X^k$ and $k-$th central moment $\mathbb{E}\left[(X - EX)^k\right]$. We consider the plug-in estimator of $\mu_k := \mathbb{E}X^k$ by $\widehat{\mu}_k := \frac{1}{n}\sum_{i=1}^n X_i^k$.

We have the following properties.

Proposition 2.4.6.

- $\widehat{\mu}_k \xrightarrow{a.s.} \mu_k$;
- $\mathbb{E}\widehat{\mu}_k = \mu_k$;
- $\operatorname{Var}(\widehat{\mu}_k) = \frac{\mu_{2k} - \mu_k^2}{n}$;
- $\sqrt{n}(\widehat{\mu}_k - \mu_k) \xrightarrow{D} N(0, \mu_{2k} - \mu_k^2)$;
- $\operatorname{Cov}(\widehat{\mu}_k, \widehat{\mu}_l) = \frac{1}{n}(\mu_{k+l} - \mu_k\mu_l)$;
- $\sqrt{n}[\widehat{\mu}_1 - \mu_1, \ldots, \widehat{\mu}_k - \mu_k]^T \xrightarrow{D} N(0, \Sigma)$ where $\Sigma_{ij} = \mu_{i+j} - \mu_i\mu_j$.

Proof. (1) it follows from SLLN; (4) it follows from CLT. □

Next, we want to estimate $\mathbb{E}\left[(X - \mu)^k\right]$, where $\mu = \mathbb{E}X$. Its plug-in estimator is $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^k$. As $k = 1$, $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n) = 0$. For $k \geqslant 2$, it is not iid average since $\bar{X}_n$ is r.v. We approximate $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^k$ by iid average b_k which is defined by

$$b_k = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^k.$$

Note that for $k \geqslant 2$,

- $b_k \xrightarrow{a.s.} \mathbb{E}\left[(X - \mu)^k\right]$, then $b_k = O_p(1)$;
- $\sqrt{n}(b_k - \mathbb{E}\left[(X - \mu)^k\right]) \xrightarrow{D} \mathit{Normal}$ from CLT. Then $b_k - \mathbb{E}\left[(X - \mu)^k\right] = O_p(n^{-1/2})$ and $b_1 = O_p(n^{-1/2})$ as $\mathbb{E}\left[(X - \mu)\right] = 0$.

We approximate $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k$ as follows.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k &= \frac{1}{n}\sum_{i=1}^{n}((X_i-\mu)-(\bar{X}_n-\mu))^k = \frac{1}{n}\sum_{i=1}^{n}((X_i-\mu)-b_1)^k \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^k + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\binom{k}{j}(-b_1)^j(X_i-\mu)^{k-j} \\ &= b_k + \sum_{j=1}^{k}\binom{k}{j}(-b_1)^j(X_i-\mu)^{k-j},\end{aligned}$$

where $b_k \xrightarrow{a.s.} \mathbb{E}\left[(X-\mu)^k\right]$, $b_1 \xrightarrow{a.s.} 0$, and $b_{k-j} \xrightarrow{a.s.} \mathbb{E}\left[(X-\mu)^{k-j}\right]$. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k \xrightarrow{a.s.} \mathbb{E}\left[(X-\mu)^k\right].$$

from continuous mapping theorem.

Next, we want to get limiting distribution of $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k - \mathbb{E}\left[(X-\mu)^k\right]\right)$. Note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k &= b_k + k(-b_1)b_{k-1} + \sum_{j=2}^{k}\binom{k}{j}(-b_1)^j(X_i-\mu)^{k-j} \\ &= b_k - kb_1b_{k-1} + O_p(n^{-1}) \\ &= b_k - kb_1\mathbb{E}\left[(X-\mu)^{k-1}\right] - kb_1(b_{k-1}-\mathbb{E}\left[(X-\mu)^{k-1}\right]) + O_p(n^{-1}).\end{aligned}$$

where $(-b_1)^j = O_p(n^{-j/2})$ for $j \geqslant 2$, and $b_{k-j} = O_p(1)$, so $\sum_{j=2}^{k}\binom{k}{j}(-b_1)^j(X_i-\mu)^{k-j} = O_p(n^{-1})$.

Since $b_{k-1} - \mathbb{E}\left[(X-\mu)^{k-1}\right] = O_p(n^{-1/2})$, then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k = b_k - kb_1\mathbb{E}\left[(X-\mu)^{k-1}\right] + O_p(n^{-1}) = \frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu)^k - k\mathbb{E}\left[(X-\mu)^{k-1}\right](X_i-\mu)\right) + O_p(n^{-1}).$$

So the limiting distribution of

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k - \mathbb{E}\left[(X-\mu)^k\right]\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left((X_i-\mu)^k - \mathbb{E}\left[(X-\mu)^k\right] - k\mathbb{E}\left[(X-\mu)^{k-1}\right](X_i-\mu)\right) + O_p(n^{-1/2}),$$

is $N(0,v)$ [8], where let $m_k = \mathbb{E}\left[(X-\mu)^k\right]$, we have

$$\begin{aligned}\operatorname{Cov}((X-\mu)^k, km_{k-1}(X-\mu)) &= km_{k-1}\operatorname{Cov}((X-\mu)^k, X-\mu) \\ &= km_{k-1}\operatorname{Cov}\left([E[(X-\mu)^{k+1}] - E[(X-\mu)^k]E[X-\mu]\right) \\ &= km_{k-1}m_{k+1},\end{aligned}$$

and

$$\operatorname{Var}((X-\mu)^k) = E[(X-\mu)^{2k}] - (E[(X-\mu)^k])^2 = m_{2k} - m_k^2$$

Then

$$\begin{aligned}v = \operatorname{Var}\left((X_i-\mu)^k - \mathbb{E}\left[(X-\mu)^k\right] - k\mathbb{E}\left[(X-\mu)^{k-1}\right](X_i-\mu)\right) &= Var\left[(X-\mu)^k - kE[(X-\mu)^{k-1}](X-\mu)\right] \\ &= m_{2k} - 2km_{k-1}m_{k+1} - m_k^2 + k^2m_{k-1}^2m_2.\end{aligned}$$

2.4.3 Quantiles

For $0 < p < 1$, define the $p-th$ quantile of F by $\xi_p = \inf\{x : F(x) \geqslant p\}$ and write $\xi_p = F^{-1}(p)$. $F^{-1}(p)$ is well-defined because F is right continuous and increasing, $\{x \in \mathbb{R} : F(x) \geqslant p\}$ is an interval of the form $[a,\infty)$. Thus, the minimum of the set is a.

Note that if F strictly increases from 0 to 1 on an interval (so that the underlying distribution is continuous), then F^{-1} is the ordinary inverse of F.

[8] for the variance of iid term inside summation, apply $\operatorname{Var}(Y+Z) = \operatorname{Var}(Y) + \operatorname{Var}(Z) + 2\operatorname{Cov}(Y,Z)$.

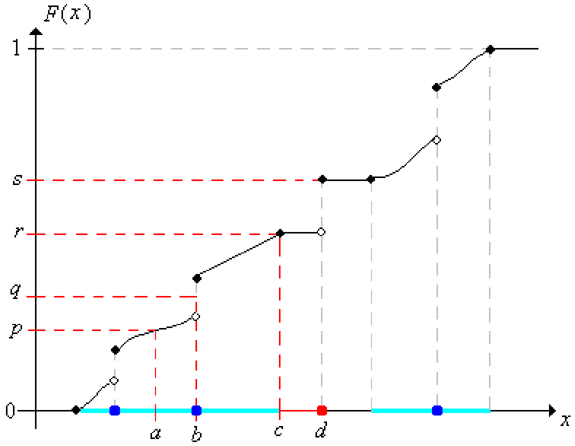


Figure 2.2: Quantiles of various orders

Roughly speaking, a p-th quantile is a value where the graph of the distribution function crosses (or jumps over) p . For example, in the picture Figure 2.3, a is the unique p-th quantile and b is the unique q-th quantile. On the other hand, the r-th quantiles is c, and moreover, d is a quantile for all orders in the interval [r,s] . Note also that if X has a continuous distribution (so that F is continuous) and x is a quantile of order $p \in (0, 1)$, then $F(x) = p$.

Graphically, the five numbers are often displayed as a boxplot or box and whisker plot, which consists of a line extending from the minimum value a to the maximum value b , with a rectangular box from q_1 to q_3 , and whiskers at a , the median q_2 , and b . Roughly speaking, the five numbers separate the set of values of X into 4 intervals of approximate probability 1/4 each.

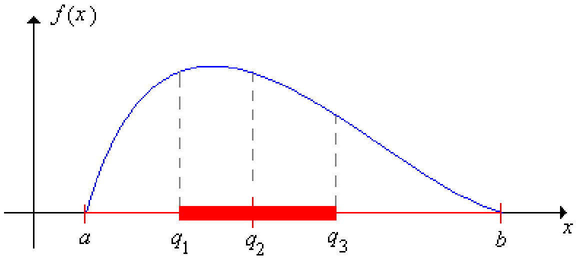


Figure 2.3: The probability density function and boxplot for a continuous distribution.

Other basic properties of the quantile function are given in the following proposition.

Proposition 2.4.7.
- $F^{-1}(p)$ is increasing on p;
- $F^{-1}(F(x)) \leq x$ for all $x \in \mathbb{R}$;
- $F(\xi_p) \geqslant p$ (i.e., $F(F^{-}(p)) \geqslant p$);
- For $x \in \mathbb{R}, p \in (0, 1)$, $F(x) > p$ iff $x \geqslant \xi_p$.

Proposition 2.4.8. Suppose that X has a continuous distribution that is symmetric about a point $a \in \mathbb{R}$. If $a + t$ is a $p-$th quantile of order, then $a - t$ is a quantile of order $1 - p$.

Proof. Note that this is the quantile function version of symmetry result for the distribution function. If $a+t$ is a qantile of order p then (since X has a continuous distribution) $F(a+t) = p$. But then $F(a-t) = 1 - F(a+t) = 1 - p$ so $a - t$ is a quantile of order $1 - p$. □

Remark 2.4.9. If $a = 0$, we have $F^{-1}(p) = t$ and $F^{-1}(1-p) = -t = -F^{-1}(p)$. Then $\xi_{1-p} + \xi_p = 0$.

In [34], we think quantile as a functional Q_p that $Q_p(F) := \inf\{x \in \mathbb{R} : F(x) \geqslant p\}$. The plug-in estimate is given by

$$\widehat{\xi_p} = Q_p(F_n) := \inf\{x \in \mathbb{R} : F_n(x) \geqslant p\} = F^{-1}(p)$$

It is interest to determine in what sense the r.v. $Q_p(F_n)$ approaches $Q_p(F)$ as $n \to \infty$. In this case, $Q_p(F_n)$ is a fairly complicated, nonlinear function of all the variables, so that this convergence does not follow immediately by a classical result such as law of large numbers.

Under the following assumptions, we study $\widehat{\xi_p}$.

- Assumption 1: $f(x) = F'(x)$ is continuous in a neighborhood of $x = \xi_p$;
- Assumption 2: $f(\xi_p) > 0$.

We will show that

- $\widehat{\xi_p} \xrightarrow{a.s.} \xi_p$;
- $\sqrt{n}(\widehat{\xi} - \xi_p) \xrightarrow{D} N\left(0, \frac{p(1-p)}{f(\xi_p)^2}\right)$.

Chapter 3

Statistical learning

References:

- The Elements of Statistical Learning [13] [1]
- Probabilistic Machine Learning: An Introduction [26] [2]
- Pattern Recognition and Machine Learning [3] [3]
- Mathematics of Machine Learning by Prof. Philippe Rigollet (lecture note) [4]
- Statistical Methods for Machine Learning by Larry Wasserman (lecture note) [5]
- An Introduction to Statistical Learning: with Applications in R [15][6]

3.1 Linear Methods for Classification

As explained in [15, Section 4.2] (Why Not Linear Regression?), there are at least two reasons not to perform classification using a regression method:

- a regression method cannot accommodate a qualitative response with more than two classes;
- a regression method will not provide meaningful estimates of $\mathbb{P}(Y|X)$, even with just two classes.

3.1.1 LDA and QDA

Theorem 3.1.1. The true error rate of a classifier h is given by

$$L(h) := \mathbb{P}(h(X) \neq Y).$$

Consider the special case where $Y \in \mathcal{Y} = \{0, 1\}$. Let $r(x) = \mathbb{P}(Y = 1|X = x)$. In this case the Bayes classification rule h^* is given by

$$h^*(x) = \begin{cases} 1, & r(x) > \frac{1}{2} \\ 0, & r(x) \leq \frac{1}{2}. \end{cases}$$

Prove that the Bayes classification rule is optimal, that is, if h is any other classification rule then

[1] https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

[2] https://probml.github.io/pml-book/book1.html

[3] https://cds.cern.ch/record/998831/files/9780387310732_TOC.pdf

[4] https://ocw.mit.edu/courses/mathematics/18-657-mathematics-of-machine-learning-fall-2015/lecture-notes/MIT18_657F15_LecNote.pdf

[5] https://www.stat.cmu.edu/~larry/=sml/

[6] https://www.statlearning.com/

$L(h^*) \leq L(h)$.

Proof.

For a classifier h, we rewrite the true error rate $L(h)$ by

$$\begin{aligned} L(h) &= \mathbb{P}(h(X) \neq Y) = \mathbb{P}(h(X) = 1, Y = 0) + \mathbb{P}(h(X) = 0, Y = 1) \\ &= \mathbb{E}(\mathbb{E}(\mathbf{1}_{\{h(X)=1,Y=0\}}|X)) + \mathbb{E}(\mathbb{E}(\mathbf{1}_{\{h(X)=0,Y=1\}}|X)) \end{aligned}$$

where $\mathbf{1}_A$ is the indicator function over a set A and we write $\mathbb{P}(h(X) = 1, Y = 0) = \mathbb{E}\mathbf{1}_{\{h(X)=1,Y=0\}}$, the second equality is from the disjoint of two events, and the third equality is from the law of total expectation conditioning on X.

Since $h(X)$ is measurable w.r.t X, then we take it away from the inner expectation. So the above equation becomes

$$\begin{aligned} L(h) &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}\mathbb{E}(\mathbf{1}_{\{Y=1\}}|X)) + \mathbb{E}(\mathbf{1}_{\{h(X)=1\}}\mathbb{E}(\mathbf{1}_{\{Y=0\}}|X)) \\ &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}r(X)) + \mathbb{E}(\mathbf{1}_{\{h(X)=1\}}(1 - r(X))) \\ &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}r(X) + \mathbf{1}_{\{h(X)=1\}}(1 - r(X))) \end{aligned} \tag{3.1}$$

where we rewrite $\mathbb{E}(\mathbf{1}_{\{Y=1\}}|X) = \mathbb{P}(Y = 1|X)$ and replace it by $r(X)$ in the second equality.

For a classifier h and the Bayes classifier h^*, using the equality (3.1) we obtain

$$\begin{aligned} L(h) - L(h^*) &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}r(X) + \mathbf{1}_{\{h(X)=1\}}(1 - r(X))) - \mathbb{E}(\mathbf{1}_{\{h^*(X)=0\}}r(X) + \mathbf{1}_{\{h^*(X)=1\}}(1 - r(X))) \\ &= \mathbb{E}[(\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}})r(X) + (\mathbf{1}_{\{h(X)=1\}} - \mathbf{1}_{\{h^*(X)=1\}})(1 - r(X))] \\ &= \mathbb{E}[(\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}})(2r(X) - 1)] \end{aligned}$$

where we use identity $\mathbf{1}_{\{h(X)=1\}} = 1 - \mathbf{1}_{\{h(X)=0\}}$ in the third equality.

There are three cases for the last equality. For $h(X) = h^*(X)$, $L(h) - L(h^*) = 0$. For $h(X) = 1, h^*(X) = 0$, $L(h) - L(h^*) = -\mathbb{E}(2r(X) - 1) = \mathbb{E}(|2r(X) - 1|)$ since $r(X) \leq \frac{1}{2}$. For $h(X) = 0, h^*(X) = 1$, $L(h) - L(h^*) = \mathbb{E}(2r(X) - 1)$. Hence, from the above discussion and definition of h^* we have

$$L(h) - L(h^*) = \mathbb{E}[\mathbf{1}_{\{h(X)\neq h^*(X)\}}|2r(X) - 1|] \geqslant 0$$

which implies $L(h^*) \leq L(h)$. This gives the desired result. □

Theorem 3.1.2. Suppose that $Y \in \{1, \ldots, k\}$ and $\mathbb{P}(X = x|Y = k)$ is Gaussian $N(\mu_k, \Sigma_k)$.

-
- If $\Sigma_k \neq \Sigma_l$ for any k, l, then the Bayes classifier is

  $$h^*(x) = \operatorname*{argmax}_k \delta_k(x)^{(1)}$$

  provided by

  $$\delta_k^{(1)}(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k) + \log(\pi_k).$$
- If $\Sigma_k = \Sigma_l$ for any k, l, then the Bayes classifier is

  $$h^*(x) = \operatorname*{argmax}_k \delta_k^{(2)}(x)$$

  provided by

  $$\delta_k^{(2)}(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log(\pi_k).$$

Exercise 1. If $X|Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X|Y = 1 \sim N(\mu_1, \Sigma_1)$, then the Bayes rule is

$$h(x) = \begin{cases} 1 & \text{if } r_1^2 < r_2^2 + 2\log\frac{\pi_1}{\pi_0} + \log\frac{|\Sigma_0|}{|\Sigma_1|} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

where $r_i^2 = (x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i), i = 0, 1$.

Proof. Let $\mathbb{P}(X = x|Y = k) = f_k(x)$ and $\mathbb{P}(Y = k) = \pi_k$ for $k = 0, 1$. From the Bayes' theorem, we have

$$\mathbb{P}(Y = i|X = x) = \frac{f_i(x)\pi_i(x)}{\sum_{k=0}^{1} f_k(x)\pi_k}, \text{ for } i = 0, 1.$$

Since the Bayes rule is $h^*(x) = \mathbf{1}_{\{\mathbb{P}(Y=1|X=x)>\mathbb{P}(Y=0|X=x)\}}$, we need to simplify $\mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x)$ which is

$$\frac{f_1(x)\pi_1(x)}{\sum_k f_k(x)\pi_k} > \frac{f_0(x)\pi_0(x)}{\sum_k f_k(x)\pi_k}.$$

Note that the denominator can be canceled. Then we have

$$f_1(x)\pi_1(x) > f_0(x)\pi_0(x).$$

Since $X|Y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$ for $i = 0, 1$, the above inequality yields (here we cancel the same term $(2\pi)^{-d/2}$ for both side)

$$|\Sigma_1|^{-1/2} \exp\bigl(-r_1^2/2\bigr) > |\Sigma_0|^{-1/2} \exp\bigl(-r_0^2/2\bigr)$$

where let $r_i^2 = (x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i), i = 0, 1$.

We take logarithm for both side to get

$$-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}r_1^2 + \log\pi_1 > -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}r_0^2 + \log\pi_0.$$

This is just

$$r_1^2 < r_0^2 + \log\frac{|\Sigma_0|}{|\Sigma_1|} + 2\log\frac{\pi_1}{\pi_0}.$$

Hence,

$$h^*(x) = \begin{cases} 1, & \text{if } r_1^2 < r_0^2 + \log\frac{|\Sigma_0|}{|\Sigma_1|} + 2\log\frac{\pi_1}{\pi_0}, \\ 0, & \text{otherwise} \end{cases}$$

where let $r_i^2 = (x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i), i = 0, 1$. □

Exercise 2. Consider a classifier with class conditional densities of the form $N(x|\mu_c, \Sigma_c)$. In LDA, we assume $\Sigma_c = \Sigma$ and in QDA, each Σ_c is arbitrary. Assume that $\Sigma_1 = k\Sigma_2$ for $k > 1$. That is, the Gaussian ellipsoids have the same"shape", but the one for class 1 is "wider". Derive an expression for the decision boundary.

Proof. Here we consider two classes that $Y \in \{1, 2\}$ and We use same notations as class. Let $f_k(x) := \mathbb{P}(X = x|Y = k)$ for $k = 1, 2$. Since class conditional densities of $f_k(x)$ are of the form $\mathcal{N}(x|\mu_c, \Sigma_c)$, which are given by

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right), \ k = 1, 2.$$

In this question, we consider the decision boundary

$$D(h) = \{x : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 2|X = x)\}.$$

From the Bayes' theorem, we have

$$\mathbb{P}(Y = i|X = x) = \frac{f_i(x)\pi_i(x)}{\sum_{k=1}^{2} f_k(x)\pi_k}, \text{ for } i = 1, 2.$$

Using the above equation, the conditional probability equation in decision boundary becomes

$$f_1(x)\pi_1 = f_2(x)\pi_2. \tag{3.3}$$

Plug class conditional densities of $f_k(x)$ into (3.3) and take logarithm for both side, we obtain

$$-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \log\pi_1 = -\frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + \log\pi_2.$$

Since we know that $\Sigma_1 = k\Sigma_2$ for $k > 1$, the above equation becomes

$$\log \frac{|k\Sigma_2|}{|\Sigma_2|} + (x-\mu_1)^T(k\Sigma_2)^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + 2\log\frac{\pi_2}{\pi_1} = 0.$$

Using $|k\Sigma_2| = k^d|\Sigma_2|$ and expanding the above bracket, we get

$$(\frac{1}{k}-1)x^T\Sigma_2^{-1}x + (2\mu_2^T - \frac{2}{k}\mu_1^T)\Sigma_2^{-1}x + \frac{1}{k}\mu_1^T\Sigma_2^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2 + d\log k + 2\log\frac{\pi_2}{\pi_1} = 0.$$

□

Exercise 3. Ex 4.2 in [13].

Proof.

part (a)

We follow the same notations as class. Since there are two classes, assume that $Y \in \{1,2\}$. In LDA, let $\mathbb{P}(X=x|Y=k) = f_k(x)$ and $\mathbb{P}(Y=k) = \pi_k$ for $k=1,2$. We need to compare $\mathbb{P}(Y=1|X=x)$ and $\mathbb{P}(Y=2|X=x)$ in LDA. From the Bayes' theorem, we have

$$\mathbb{P}(Y=i|X=x) = \frac{f_i(x)\pi_i(x)}{\sum_{k=1}^2 f_k(x)\pi_k}, \quad \text{for } i=1,2.$$

To compare $\mathbb{P}(Y=2|X=x) > \mathbb{P}(Y=1|X=x)$ is equivalent to

$$\frac{f_2(x)\pi_2(x)}{\sum_k f_k(x)\pi_k} > \frac{f_1(x)\pi_1(x)}{\sum_k f_k(x)\pi_k}.$$

Note that the denominator can be canceled. Then we have

$$f_2(x)\pi_2(x) > f_1(x)\pi_1(x). \tag{3.4}$$

Since each class density $f_k(x)$ is multivariate Gaussian, then

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right), \; k=1,2$$

where two classes have same covariance matrix Σ.

We plug densities of f_k into (3.4) and take logarithm for both side to get

$$-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \log\pi_1 > -\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \log\pi_2.$$

We expand the above and cancel $x^T\Sigma^{-1}x$, then the above inequality yields

$$-\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + (\mu_2-\mu_1)^T\Sigma^{-1}x + \log\pi_2 - \log\pi_1 > 0$$

Note that we estimate $\pi_1 = \frac{n_1}{n}$ and $\pi_2 = \frac{n_2}{n}$ since the size of class 1 and class 2 are n_1 and n_2 respectively. Using this estimate, we obtain

$$-\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + (\mu_2-\mu_1)^T\Sigma^{-1}x + \log\Big(\frac{n_2}{n}\Big) - \log\Big(\frac{n_1}{n}\Big) > 0$$

Hence,

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\Big(\frac{n_1}{n}\Big) - \log\Big(\frac{n_2}{n}\Big). \tag{3.5}$$

part (b) We label class 1 as C_1 of size n_1 and class 1 as C_2 of size n_2. To minimize the least squares $\sum_{i=1}^n(y_i - \beta_0 - \beta^T x_i)^2$, it suffices to take the derivatives with respect to β_0 and β to zero. We obtain

$$\sum_{i=1}^n (y_i - \beta_0 - \beta^T x_i) = 0 \tag{3.6}$$

and

$$\sum_{i=1}^n (y_i - \beta_0 - \beta^T x_i)x_i = 0 \tag{3.7}$$

So we just need to solve β_0 and β from above equations.

Note that the target coded of y_i as $-n/n_1$ for class 1 and n/n_2 for class 2, we have

$$\sum_{i=1}^{n} y_i = -n_1 \frac{n}{n_1} + n_2 \frac{n}{n_2} = 0. \tag{3.8}$$

Plug (3.8) into (3.6), we obtain

$$n\beta_0 + \beta^T \sum_{i=1}^{n} x_i = 0 \tag{3.9}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(n_1\hat{\mu}_1 + n_2\hat{\mu}_2). \tag{3.10}$$

Using (3.10), the equation (3.9) becomes

$$\beta_0 = (-\frac{n_1}{n}\hat{\mu}_1^T - \frac{n_2}{n}\hat{\mu}_2^T)\beta. \tag{3.11}$$

Next, we try to solve β from equation (3.7). Before that, we need some preparation. Since there are two classes, we estimate the mean as in [13, Chapter 4.3] given by

$$\hat{\mu}_1 = \frac{\sum_{i\in C_1} x_i}{n_1},\ \hat{\mu}_2 = \frac{\sum_{i\in C_2} x_i}{n_2}.$$

where $i \in C_1$ means that y_i is labeled in the first class coded as $-n/n_1$ and $i \in C_2$ means that y_i is labeled in the second class coded as n/n_2.

Then We have

$$\sum_i x_i = \sum_{i\in C_1} x_i + \sum_{i\in C_2} x_i = n_1\hat{\mu}_1 + n_2\hat{\mu}_2. \tag{3.12}$$

Also, We estimate the covariance matrix from our training data as in [13, Chapter 4.3]

$$\hat{\Sigma} = \frac{1}{n-2}\left[\sum_{i\in C_1}(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T + \sum_{i\in C_2}(x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^T\right] = \frac{1}{n-2}\left[\sum_{i=1}^{n} x_i x_i^T - n_1\hat{\mu}_1\hat{\mu}_1^T - n_2\hat{\mu}_2\hat{\mu}_2^T\right]. \tag{3.13}$$

So

$$\sum_{i=1}^{n} x_i x_i^T = (n-2)\hat{\Sigma} + n_1\hat{\mu}_1\hat{\mu}_1^T + n_2\hat{\mu}_2\hat{\mu}_2^T. \tag{3.14}$$

Moreover, we use the target coded of y_i again to get

$$\sum_{i=1}^{n} x_i y_i = \sum_{i\in C_1} x_i y_i + \sum_{i\in C_2} x_i y_i = -\frac{n}{n_1}\sum_{i\in C_1} x_i + \frac{n}{n_2}\sum_{i\in C_2} x_i = -n\hat{\mu}_1 + n\hat{\mu}_2. \tag{3.15}$$

Now we plug (3.11) into equation (3.7) and use equations (3.12), (3.14), and (3.15) for equation (3.7). Thus, we have

$$(n_1\hat{\mu}_1 + n_2\hat{\mu}_2)(-\frac{n_1}{n}\hat{\mu}_1^T - \frac{n_2}{n}\hat{\mu}_2^T)\beta + ((n-2)\hat{\Sigma} + n_1\hat{\mu}_1\hat{\mu}_1^T + n_2\hat{\mu}_2\hat{\mu}_2^T)\beta = n(\hat{\mu}_2 - \hat{\mu}_1). \tag{3.16}$$

After some algebra for LHS of (3.16), note that

$$\begin{aligned}(n_1\hat{\mu}_1 + n_2\hat{\mu}_2)(-\frac{n_1}{n}\hat{\mu}_1^T - \frac{n_2}{n}\hat{\mu}_2^T) + n_1\hat{\mu}_1\hat{\mu}_1^T + n_2\hat{\mu}_2\hat{\mu}_2^T &= \frac{n_1n_2}{n}\hat{\mu}_1\hat{\mu}_1^T + \frac{n_1n_2}{n}\hat{\mu}_2\hat{\mu}_2^T - 2\frac{n_1n_2}{n}\hat{\mu}_1\hat{\mu}_2^T \\ &= \frac{n_1n_2}{n}(\hat{\mu}_1\hat{\mu}_1^T - 2\hat{\mu}_1\hat{\mu}_2^T + \hat{\mu}_2\hat{\mu}_2^T) \\ &= \frac{n_1n_2}{n}(\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)^T.\end{aligned}$$

Hence, equation (3.16) becomes

$$(\frac{n_1n_2}{n}(\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)^T + (n-2)\hat{\Sigma})\beta = n(\hat{\mu}_2 - \hat{\mu}_1).$$

Hence,

$$(\frac{n_1n_2}{n}\hat{\Sigma}_B + (n-2)\hat{\Sigma})\beta = n(\hat{\mu}_2 - \hat{\mu}_1) \tag{3.17}$$

where $\hat{\Sigma}_B = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T$. This gives the desired result.

part (c) Since $\hat{\Sigma}_B\beta = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$ and $(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$ is a scalar, then $\hat{\Sigma}_B\beta$ is in the direction of

$(\hat{\mu}_2 - \hat{\mu}_1)$. Note that equation (3.17) can be rewritten as

$$(n-2)\hat{\Sigma}\beta = n(\hat{\mu}_2 - \hat{\mu}_1) - \frac{n_1 n_2}{n}\hat{\Sigma}_B\beta. \tag{3.18}$$

Since terms $\frac{n_1 n_2}{n}\hat{\Sigma}_B\beta$ and $n(\hat{\mu}_2 - \hat{\mu}_1)$ are in the direction of $(\hat{\mu}_2 - \hat{\mu}_1)$, then the RHS of (3.18) is also in the direction of $(\hat{\mu}_2 - \hat{\mu}_1)$. Thus, β is proportional to $\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$. From equation (3.5), the least squares regression coefficient is identical to the LDA coefficient up to a scalar multiple.

□

Exercise 4. Show that the Naive Bayes Classifier is equivalent to a linear classification rule.

Proof. See https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html.

□

3.1.2 Logistic regression

Exercise 5. Ex 4.4 in [13] for the multi-class logistic regression model.

Proof. For multi-classes logistic regression model, assume that we have K classes and N labels. The response y_{il} is given by that if the data point x_i is from class l where $1 \leq l \leq K-1$, then the $l-$th element of y_i is one and others are zero, and if x_i is from class K, then all elements of y_i are zero. So response y_{il} form a target matrix corresponding to sample $1 \leq n \leq N$ and class $1 \leq k \leq K$. That is

$$y_i = \mathbf{1}_{\{x \text{ is from class } l \text{ and } i = l\}}.$$

From textbook [13, Section 4.4], we know that the posterior probability that x_i comes from class k are given by

$$\mathbb{P}(y = k|X = x) = \frac{\exp\left(\beta_{k0} + \beta_k^T x\right)}{1 + \sum_{l=1}^{K-1}\exp\left(\beta_{l0} + \beta_l^T x\right)},\ k = 1, 2, \ldots, K-1,$$
$$\mathbb{P}(y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1}\exp\left(\beta_{l0} + \beta_l^T x\right)}.$$

Let

$$h_k(x) = \mathbb{P}(y = k|X = x),\ k = 1, 2, \ldots, K$$

The likelihood function for a data point x is given by

$$L(\beta; x) = h_1(x)^{y_1} h_2(x)^{y_2} \cdots h_{K-1}(x)^{y_{K-1}} (h_K(x))^{1 - \sum_{l=1}^{K-1} y_l} \tag{3.19}$$

From the posterior probability $\mathbb{P}(y = k|X = x)$, we have the log-likelihold function for a data point x

$$\ell(\beta; x) = y_1(\beta_{10} + \beta_1^T x) + y_2(\beta_{20} + \beta_2^T x) + \cdots + y_{K-1}(\beta_{(K-1)0} + \beta_{K-1}^T x) + \log(h_K) \tag{3.20}$$

Then sum over the equation (3.20) for all data points x_i, we get the log-likelihood of parameter β, that is,

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{N}\sum_{l=1}^{K-1}[y_{il}\beta_l^T x_i + \log(h_K)] \\
&= \sum_{i=1}^{N}\sum_{l=1}^{K-1}[y_{il}\beta_l^T x_i - \log\left(1 + \sum_{l=1}^{K-1}\exp\left(\beta_{l0} + \beta_l^T x_i\right)\right)]
\end{aligned}$$

where x_i is the $i-$th sample, β_l is a vector of coefficients for the $l-$th class with size $(p+1)$, $\beta = [\beta_1, \beta_2, \ldots, \beta_{K-1}]^T$ is of size $(K-1)(p+1)$.

Next, we compute the derivative of $\ell(\beta)$.

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta_k} &= \sum_{i=1}^{N}\left[y_{ik}x_i^T - \frac{\exp\big(\beta_{k0}+\beta_k^T x_i\big)}{1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x_i\big)}x_i^T\right]\\ &= \sum_{i=1}^{N}(y_{ik}-\mathbb{P}(y=k|X=x_i))x_i^T\\ &= (y_{ik}-h_k(x_i))x_i^T.\end{aligned}$$

Let $y_l = [y_{1l}, y_{2l}, \ldots, y_{Nl}]^T$ and $p_l = [h_l(x_1), h_l(x_2), \ldots, h_l(x_N)]^T$. Then we have

$$\frac{\partial \ell(\beta)}{\partial \beta} = \begin{bmatrix} X^T(y_1-h_1)\\ X^T(y_2-h_2)\\ \vdots\\ X^T(y_{K-1}-h_{k-1})\end{bmatrix}$$

where X is the $N\times(p+1)$ matrix of x_i values.

The Hessian matrix of $\ell(\beta)$ is given by

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k'^T} = -\sum_{i=1}^{N} h_k(x_i)h_{k'}(x_i)x_ix_i^T, \text{ for } k\neq k'$$

and for $k=k'$ we have

$$\begin{aligned}\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k^T} &= -\sum_{i=1}^{N}\left[\frac{\exp\big(\beta_{k0}+\beta_k^T x_i\big)x_i(1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x_i\big)) - \exp\big(\beta_{k0}+\beta_k^T x_i\big)^2 x_i}{(1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x_i\big))^2}x_i^T\right]\\ &= -\sum_{i=1}^{N}[(h_k(x_i)x_i - h_k(x_i)^2x_i)x_i^T]\\ &= -\sum_{i=1}^{N}[h_k(x_i)(1-h_k(x_i))x_ix_i^T].\end{aligned}$$

Write above second order derivative in form of matrix. Let H_k be $N\times N$ diagonal matrices for $k = 1, 2, \ldots, K-1$ with diagonal elements $h_k(x_i)(1-h_k(x_i)$, $i = 1, 2, \ldots, N$. Then we rewrite the second derivative of $\ell(\beta)$ as $k=k'$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k^T} = -X^T H_k X.$$

Let T_k be $N\times N$ diagonal matrices for $k = 1, 2, \ldots, K-1$ with diagonal elements $h_k(x_i)$, $i = 1, 2, \ldots, N$. Then we rewrite the second derivative of $\ell(\beta)$ as $k\neq k'$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k'^T} = -X^T T_k T_{k'} X.$$

Hence, the Hessian matrix of $\ell(\beta)$ is given by

$$\begin{aligned}\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} &= \begin{bmatrix} -X^T H_1 X & -X^T T_1 T_2 X & \cdots & -X^T T_1 T_{K-1} X\\ -X^T T_2 T_1 X & -X^T H_2 X & \cdots & -X^T T_2 T_{K-1} X\\ \vdots & & \ddots & \vdots\\ -X^T T_{K-1} T_1 X & -X^T T_{K-1} T_2 X & \cdots & -X^T H_{K-1} X\end{bmatrix}\\ &= -\hat{X}^T W \hat{X}\end{aligned}$$

where $\hat{X} = X\cdot \mathrm{Id}_{K-1}$, Id_{K-1} is a $(K-1)\times(K-1)$ identity matrix, $\hat{X}$ is a $(K-1)\times(K-1)$ matrix with each block matrix of size $(p+1) \times (p+1)$, and

$$W = \begin{bmatrix} H_1 & T_1T_2 & \cdots & T_1T_{K-1} \\ T_2T_1 & H_2 & \cdots & T_2T_{K-1} \\ \vdots & & \ddots & \vdots \\ T_{K-1}T_1 & T_{K-1}T_2 & \cdots & H_{K-1} \end{bmatrix}$$

Now our Newton-Raphson algorithm for maximizing the log-likelihood is given by

$$\beta^{new} = \beta^{old} + (\hat{X}^T W \hat{X})^{-1} \hat{X}^T \begin{bmatrix} (y_1 - h_1) \\ (y_2 - h_2) \\ \vdots \\ (y_{K-1} - h_{k-1}) \end{bmatrix}$$

Let

$$y - h = \begin{bmatrix} (y_1 - h_1) \\ (y_2 - h_2) \\ \vdots \\ (y_{K-1} - h_{k-1}) \end{bmatrix}$$

Hence, the algorithm can be expressed as

$$\beta^{new} \leftarrow (\hat{X}^T W \hat{X})^{-1} \hat{X}^T W (\hat{X}\beta^{old} + W^{-1}(y-h))$$

So β^{new} is the solution of a non-diagonal weighted least squares problem with a response $(\hat{X}\beta^{old} + W^{-1}(y-h))$. We can still use the Netwon algorithm as an iteratively reweighted least squares algorithm. Let $z = (\hat{X}\beta^{old} + W^{-1}(y-h))$. The iteratively reweighted least squares algorithm is as follows. Set $\beta^0 = 0$, update β^{new} by

$$\beta^{new} \leftarrow \underset{\beta}{\text{argmin}} (z - \hat{X}\beta) W (z - \hat{X}\beta)$$

However, the Hessian maybe not negative definite, Newton-Raphson update cannot perform effective.

Here we implement a improved Newton-Raphson algorithm from paper [11]. Given an intial value β^0, let λ_1 be the largest eigenvalue of Hessian matrix of $\ell(\beta)$ at β^0 defined by $H(\ell(\beta^0))$. Let ε be the step size and let $\alpha = \lambda_1 + \varepsilon \|\frac{\partial \ell(\beta^0)}{\partial \beta}\|_2$. Define the controlling of Hessian matrix H by

$$H_\alpha(\ell(\beta^0)) = \begin{cases} H(\ell(\beta^0)) - \alpha \cdot \text{Id}, & \text{if } \alpha > 0, \\ H(\ell(\beta^0)), & \text{otherwise} \end{cases}$$

where $H_\alpha(\ell(\beta^0))$ is always negative definite.

Update β^{new} by

$$\beta^{new} = \beta^{old} - H_\alpha^{-1}(\ell(\beta^{old})) \frac{\partial \ell(\beta^{old})}{\partial \beta}$$

where we have computed the Hessian and gradient of $\ell(\beta)$ in form of matrix as before. □

3.1.3 SVM

Exercise 6. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable.

Proof. See Bishop 3.4 in https://www.cise.ufl.edu/~anand/fa05/hw1sol_fall05.pdf. □

Exercise 7 (Exercise in [3])**.** In the maximum-margin hyperplane problem, let's τ denotes the value of the margin. Show that

$$\frac{1}{\tau^2} = 2\sum \alpha - \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \alpha_j y_k y_j x_k^T x_j.$$

Chapter 4

Probability

This Chapter is based on STAT 902 - Theory of Probability 2 (Instructor: Dr. Yi shen) at UWaterloo. References:

- Amir Dembo, Probability Theory: STAT310/MATH230 (April 15, 2021)[1]
- Jean-Francois Le Gall, Brownian Motion, Martingales, and Stochastic Calculus[2]
- Timo Seppalainen, Basics of Stochastic Analysis [3]
- Rick Durrett, Theory and Examples (Version 5) [4]
- John B. Walsh, Knowing the Odds: An Introduction to Probability[5]
- Bernt Oksendal, Stochastic Differential Equations [6]

4.1 Weak convergence

Motivated by CLT, we explore here the **convergence in distribution**.

Definition 4.1.1 (Convergence in distribution)**.** We say that r.v.-s X_n converge in distribution to a r.v. X_∞, denoted by $X_n \xrightarrow{D} X_\infty$ if $F_{X_n}(\alpha) \to F_{X_\infty}(\alpha)$ as $n \to \infty$ for each fixed α which is a continuity point of F_{X_∞}.
Similarly, we say that distribution functions F_n **converge weakly** to F_∞, denoted by $F_n \xrightarrow{w} F_\infty$, if $F_n(\alpha) \to F_\infty(\alpha)$ as $n \to \infty$ for each fixed α which is a continuity point of F_∞.

Remark 4.1.2. If the limit r.v. X_∞ has probability density function, or more generally whenever F_{X_∞} is a continuous function, the convergence in distribution of X_n to X_∞ is equivalent to the point-wise convergence of the corresponding distribution functions. Such is the case of the CLT, since the normal r.v. has a density.

Definition 4.1.3 (weak convergence of probability measures)**.** For a topological space $\mathcal{S}$, let $C_b(\mathcal{S})$ denote the collection of all continuous bounded functions on $\mathcal{S}$. We say that a sequence of probability measures ν_n on a topological space $\mathcal{S}$ equipped with its Borel $\sigma-$algebra converges weakly to a probability measure

[1] https://statweb.stanford.edu/~adembo/stat-310b/lnotes.pdf

[2] https://link.springer.com/book/10.1007/978-3-319-31089-3

[3] https://www.math.wisc.edu/~seppalai/courses/735/notes.pdf

[4] https://services.math.duke.edu/~rtd/PTE/PTE5_011119.pdf

[5] https://bookstore.ams.org/gsm-139

[6] https://link.springer.com/book/10.1007/978-3-642-14394-6

ν_∞, denoted $\nu_n \xrightarrow{w} \nu_\infty$, if $\nu_n(h) \to \nu_\infty(h)$ for each $h \in C_b(\mathcal{S})$.

Proposition 4.1.4. The weak convergence of distribution functions is equivalent to the weak convergence of the corresponding laws as probability measures on $(\mathbb{R}, \mathcal{B})$. Consequently, $X_n \xrightarrow{D} X_\infty$ i.f.f. for each $h \in C_b(\mathbb{R})$, we have

$$\lim_{n\to\infty} \mathbb{E}h(X_n) = \mathbb{E}h(X_\infty).$$

Theorem 4.1.5 (Lévy continuity theorem).

Theorem 4.1.6 (CLT). Suppose that X_i i.i.d. r.v. for $i = 1, \dots, n$ with $\mathbb{E}[X_i]^2 < \infty$. Then

$$\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{\sigma^2 n}} \xrightarrow{D} N(0,1)$$

where $\mu = \mathbb{E}X_1$ and $\sigma^2 = \mathrm{Var}[X_1]$.

Proof. Using its characteristic function and continuity theorem. □

4.2 Conditional expectation and martingale

4.2.1 Conditional expectation

Suppose the random variables (r.v.s) X and Z on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are both simple functions. Here we let X take distinct values $x_1, \dots, x_m \in R$ and Z take distinct values $z_1, \dots, z_n \in R$. Without loss of generality , we assume that $\mathbb{P}(Z = z_i) > 0$ for $i = 1, 2, d \dots, n$. Define r.v. $Y := \mathbb{E}[X|Z]$ on the same probability space s.t. $Y(\omega) = \mathbb{E}[X|Z = z_i]$ whenever ω is such that $Z(\omega) = z_i$. Hence, $\mathcal{G}$ is finitely generated and since $Y(\omega)$ is constant on each generator G_i of $\mathcal{G}$, we see that $Y(\omega)$ is measurable on $(\Omega, \mathcal{G})$. Further, since any $G \in \mathcal{G}$ is of the form $G = \cup_{i\in\mathcal{I}} G_i$ for the disjoint sets G_i and some $\mathcal{I} \subset \{1, \dots, n\}$

Partitioning Ω into the discrete collection of $Z-$atoms, namely the sets $G_i = \{\omega : Z(\omega) = z_i\}$, observe that $Y(\omega)$ is constant on each of these sets. The $\sigma-$algebra $\mathcal{G} = \mathcal{F}^Z = \sigma(Z) = \{Z^{-1}(B) : B \in \mathcal{B}\}$ is in this setting merely the collection of all possible unions of various $Z-$atoms. We find that

$$\begin{aligned}\mathbb{E}[Y\mathbf{1}_G] &= \sum_{i\in\mathcal{I}} \mathbb{E}[Y\mathbf{1}_{G_i}] = \sum_{i\in\mathcal{I}} \mathbb{E}[Y\mathbf{1}_{G_i}] = \sum_{i\in\mathcal{I}} \mathbb{E}[X|Z = z_i]\,\mathbb{P}(Z = z_i) \\ &= \sum_{i\in\mathcal{I}} \sum_{j=1}^m x_j \mathbb{P}(X = x_j, Z = z_i) = \mathbb{E}[X\mathbf{1}_G].\end{aligned}$$

Hence, in case X and Z are simple functions and $\mathcal{G} = \sigma(Z)$, we have $Y = \mathbb{E}[X|Z]$ as a r.v. on $(\Omega, \mathcal{G})$ s.t. $\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G]$ for all $G \in \mathcal{G}$.

Proceeding hereafter to consider an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with integrable r.v. X and an arbitrary $\sigma-$algebra $\mathcal{G} \subset \mathcal{F}$, we note that both properties above still make sense, motivating our general definition of the conditional expectation, as the following theorem.

Theorem 4.2.1 (Conditional expectation). Given $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$ a $si-$algebra there exists a r.v. Y called the **conditional expectation** (c.e.) of X given $\mathcal{G}$, denoted by $\mathbb{E}[X|\mathcal{G}]$, s.t. $Y \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ and for any $G \in \mathcal{G}$,

$$\mathbb{E}[(X - Y)\mathbf{1}_G] = 0. \tag{4.1}$$

Moreover, if (4.1) holds for any $G \in \mathcal{G}$ and r.v.s Y and $\widetilde{Y}$, both of which are in $L^1(\Omega, \mathcal{G}, \mathbb{P})$, then $\mathbb{P}(Y = \widetilde{Y}) = 1$. (i,e., the c.e. is uniquely defined for $\mathbb{P}-$a.s. ω.)

Proposition 4.2.2. If $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\sigma-$algebra $\mathcal{H}$ is independent of $\sigma(\sigma(X).\mathcal{G})$, then

$$\mathbb{E}\left[X|\sigma(\mathcal{H}, \mathcal{G})\right] = \mathbb{E}\left[X|\mathcal{G}\right].$$

In particular, for $\mathcal{G} = \{\emptyset, \Omega\}$ this implies that if $\mathcal{H}$ is independent of $\sigma(X)$, then

$$\mathbb{E}\left[X|\mathcal{H}\right] = \mathbb{E}X.$$

Proposition 4.2.3 (Tower rule)**.** Suppose $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\sigma-$algebra $\mathcal{H}$ and $\mathcal{G}$ are s.t. $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$. Then

$$\mathbb{E}\left[X|\mathcal{H}\right] = \mathbb{E}\left[\mathbb{E}\left[X|\mathcal{G}\right]|\mathcal{H}\right].$$

Next, we will introduce the definition of uniform integrability (u.i.).

Let $X \in L^p$. Given a constant $c > 0$, the event $\{|X| \geqslant c\}$ is decreasing in c. This means $\mathbf{1}_{\{|X| \geqslant c\}}$ non-increasing in c. Also, $\mathbf{1}_{\{|X| \geqslant c\}} \to 0$ a.s. as $c \to \infty$ since $\mathbb{P}(\cap_c \{|X| \geqslant c\}) \to 0$. Then for $\mathbb{E}|X|^p < \infty$ we have

$$|X|^p \mathbf{1}_{\{|X| \geqslant c\}} \xrightarrow[c \to \infty]{a.s.} 0.$$

From D.C.T.,

$$\lim_{c \to 0} \mathbb{E}\left[|X|^p \mathbf{1}_{\{|X| \geqslant c\}}\right] = 0$$

Definition 4.2.4 (Uniformly integrable)**.** A sequence of r.v.s X_n is p-th power **uniformly integrable (u.i.)** if

$$\lim_{c \to 0} \sup_n \mathbb{E}\left[|X_n|^p; |X| \geqslant c\right] = 0$$

Several useful lemma are as follows.

Lemma 4.2.5. Let $X \in L^1$, then for any $\varepsilon > 0$, there exists $\delta > 0$ s.t. for any event F with $\mathbb{P}(F) < \delta$, we have

$$\mathbb{E}\left[|X|; F\right] < \varepsilon.$$

Proof. Use contradiction. □

Lemma 4.2.6. If $X \in L^1$, for any $\varepsilon > 0$, there exists a constant k s.t.

$$\mathbb{E}\left[|X|; |X| > k\right] < \varepsilon.$$

Proof. Fix $\varepsilon > 0$, there exists $\delta > 0$ s.t. $\mathbb{P}(|X| > k) \leq \frac{\mathbb{E}|X|}{K}\delta$ for K large enough. Then it follows from Lemma 4.2.5. □

A trivial example of a uniformly integrable family is a collection of r.v.-s that are dominated by an integrable r.v., i.e., $|X_i| \leq Y$ where $\mathbb{E}Y < \infty$. Our first result gives an interesting example that shows that u.i. families can be very large.

Proposition 4.2.7. For any $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, the collection $\{\mathbb{E}[X|\mathcal{H}] : \mathcal{H} \subset \mathcal{F}$ is a $\sigma-$algebra $\}$ is u.i.

Proof. Apply Lemma 4.2.5, Let $Y = \mathbb{E}[X|\mathcal{H}]$ and $A = \{|Y| \geqslant M\} \in \mathcal{H}$ with $\mathbb{P}(A) \leq \delta$. Then $\mathbb{E}[X; A] \leq \varepsilon$

for any $\varepsilon > 0$. Note that

$$\mathbb{E}[|Y|\mathbf{1}_A] = \mathbb{E}\left[\mathbb{E}\left[|X||\mathcal{H}\right]\mathbf{1}_A\right] = \mathbb{E}[|X|\mathbf{1}_A] \leq \varepsilon$$

Since this applies for any $\sigma-$algebra $\mathcal{H} \subset \mathcal{F}$. Then Y is u.i. □

A common way to check u.i. is to use

Theorem 4.2.8. Let $\psi \geqslant 0$ be any function with $\psi(x)/x \to \infty$ as $x \to \infty$, e.g., $\psi(x) = x^p$ with $p > 1$ or $\psi(x) = x \log^+ x$. If $\mathbb{E}\psi(|X_i|) \leq C$ for all $i \in I$. Then $\{X_i\}_{i \in I}$ is u.i.

Theorem 4.2.9. Let $p \geqslant 1$ and assume that $X_n \xrightarrow{P} X$. The following statement are equivalent.

(i) $X_n \xrightarrow{L^p} X$;

(ii) X_n is p-th uniformly integrable;

(iii) $\|X_n\|_p \to \|X\|_p$ as $n \to 0$.

Moreover,

Proof. (i) → (ii), consider $p = 1$.

Fix $\varepsilon > 0$ and N large enough s.t. for $n \geqslant N$

$$\mathbb{E}|X - X_n| < \frac{\varepsilon}{2}.$$

Then pick $\delta > 0$ s.t. $\mathbb{P}(F) < \delta$. From Lemma 4.2.5, for $1 \leq n \leq N$

$$\mathbb{E}\left[|X_n|; F\right] < \varepsilon$$

and due to finitely many n, from triangle inequality we have

$$\mathbb{E}\left[|X|; F\right] < \varepsilon/2.$$

For $n \leq N$, take $F_n := \{|X_n| > K\}$ for some $K > 0$. Take K large enough s.t. $\mathbb{P}(F_n) \leq \delta$, then

$$\mathbb{E}\left[|X_n|; |X_n| > K\right] \leq \varepsilon.$$

For $n > N$, $|X_n| \leq |X_n - X| + |X|$ by triangle inequality. Note that

$$|X_n|\mathbf{1}_{|X_n|>K} \leq |X|\mathbf{1}_{|X_n|>K} + |X - X_n|\mathbf{1}_{|X_n|>K} \leq |X|\mathbf{1}_{|X_n|>K} + |X - X_n|.$$

Taking expectation on both side, since X_n are bounded in L^1, then $\mathbb{P}(F_n) \leq \frac{\sup_n \mathbb{E}|X_n|}{K} < \delta$ for large K. Hence, we get

$$\mathbb{E}\left[|X_n|\mathbf{1}_{|X_n|>K}\right] \leq \mathbb{E}\left[|X|\mathbf{1}_{|X_n|>K}\right] + \mathbb{E}|X - X_n| < \varepsilon$$

(ii) → (i), consider $p = 1$.

Consider the truncation function at k that $\varphi_k(x) := -k\mathbf{1}_{(-\infty,-k]} + x\mathbf{1}_{(-k,k]} + k\mathbf{1}_{(k,\infty)}$. Note that φ is continuous and $|\varphi_k(x) - x| \leq |x|\mathbf{1}_{|x|>k}$.

Since X_n is u.i., for any $n > 0$ we have

$$\mathbb{E}|\varphi_k(X_n) - X_n| < \mathbb{E}|X_n|\mathbf{1}_{|X_n|>k} < \varepsilon/3$$

From Lemma 4.2.6, we have

$$\mathbb{E}|\varphi_k(X) - X| < \mathbb{E}|X|\mathbf{1}_{|X|>k} < \varepsilon/3$$

for k large enough.

Since $X_n \xrightarrow{P} X$,

$$\mathbb{P}(|\varphi_k(X_n) - \varphi_k(X)| \geqslant \varepsilon) \leq \mathbb{P}(|X_n - X| \geqslant \varepsilon) \leq \varepsilon,$$

this means $\varphi_k(X_n) \xrightarrow{P} \varphi_k(X)$ as $n \to \infty$. [7]

[7] If event A holds, then B occurs. This mean $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Since $|\varphi_k(X_n)| \leq k$ for some constant $k > 0$, from Vital's convergence theorem [8]

$$\mathbb{E}|\varphi_k(X_n) - \varphi_k(X)| \leq \varepsilon/3.$$

Hence,

$$\mathbb{E}|X_n - X| \leq \mathbb{E}|\varphi_k(X_n) - X_n| + \mathbb{E}|\varphi_k(X) - X| + \mathbb{E}|\varphi_k(X_n) - \varphi_k(X)| \leq \varepsilon.$$

(i) → (iii), consider $p = 1$.

(iii) → (i), consider $p = 1$.

See Scheffe's theorem [9].

□

4.2.2 Discrete-time Martingale

We first review some notations of general theory of stochastic processes.

Random variables, random vectors, stochastic processes (=random functions) are special cases of the concept of random element.

Definition 4.2.10 (stochastic processes)**.** Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and time $0 < T \leq \infty$, $\{X_t\}_{t\in[0,T]}$ is called stochastic processes (s.p.) if this is a r.v. $X_t : \Omega \to \mathbb{R}$ s.t.

$$(t, \omega) \mapsto X_t(\omega)$$

is measurable w.r.t. $\mathcal{B}([0, T]) \times \mathcal{F}$.

Remark 4.2.11. This one is stronger than to say for any t, $X_t : \Omega \to \mathbb{R}$ is measurable w.r.t. $\mathcal{F}$.

Example 4.2.12. Let U be a r.v. which is uniformly distributed on the interval $[0, 1]$. The probability space on which U is defined is denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. Define two stochastic processes $\{X_t : t \in [0, 1]\}$ and $\{Y_t : t \in [0, 1]\}$ by $X_t(\omega) = 0$ for all $t \in [0, 1]$ and $Y_t(\omega) = 1$ if $t = U(\omega)$ otherwise $Y_t(\omega) = 0$.
Then

$$P(X_t(\omega) \neq Y_t(\omega)) = P(U(\omega) = t) = 0$$

Hence, X_t is a modification of Y_t. However,

$$\mathbb{P}(X_t = Y_t, t \geqslant 0) = 0,$$

which implies X_t and Y_t are not indistinguishable.

A filtrated space is $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=0}^{\infty}, \mathbb{P})$ and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}$. Let $\mathcal{F}_\infty := \sigma(\cup_n \mathcal{F}_n) \subset \mathcal{F}$.

Definition 4.2.13. A stochastic process X_n is **adapted** to the filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$ if X_n is $\mathcal{F}_n-$measurable for all $n = 0, 1, \ldots$. This random process X_t is said to be **progressive** if for every $t \geqslant 0$ the mapping

$$(t, \omega) \mapsto X_t(\omega)$$

is measurable w.r.t. $\mathcal{B}([0, T]) \times \mathcal{F}$.

Note that a progressive process is both adapted and measurable.

Proposition 4.2.14. Let X_t be a random process with values in a metric space. Suppose that X is apated and that he sample pahts of X are right-continous (i.e., for every $\omega \in \Omega, t \mapsto X_t(\omega)$ is right continuous). Then X is progressive.

[8] Let X_n be a seq. of r.v.-s that $|X_n| \leq K \in \mathbb{R}$. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{L^p} X$.

[9] https://math.stackexchange.com/questions/364059/doubt-in-scheffes-lemma

Definition 4.2.15 (Discrete time martingale)**.** A stochastic process $\{M_n\}_{n=0}^{\infty}$ is a martingale (denoted MG) if

- M_n is adapted to the filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$
- $M_n \in L^1(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=0}^{\infty}, \mathbb{P})$, i.e., $\mathbb{E}|M_n| < \infty$
- $\mathbb{E}\left[M_n|\mathcal{F}_{n-1}\right] = M_{n-1}$

Remark 4.2.16. We say $\{M_n\}_{n=0}^{\infty}$ is a supermartingale (submartingale rep.) if the third condition replaced by $\mathbb{E}\left[M_n|\mathcal{F}_{n-1}\right] \leq M_{n-1}$ ($\mathbb{E}\left[M_n|\mathcal{F}_{n-1}\right] \geqslant M_{n-1}$ rep.).

Remark 4.2.17. M_n is a martingale iff $M_n - M_0$ is a martingale

Note that for a martingale M_n we have $\mathbb{E}M_0 = \mathbb{E}M_1 = \cdots = \mathbb{E}M_{n-1} = \mathbb{E}M_n$. For a supermartingale M_n, we have $\mathbb{E}M_n \leq \mathbb{E}M_{n-1} \leq \cdots \leq \mathbb{E}M_0$.

Example 4.2.18. Let X_i be iid r.v. with $Ber(1/2)$. Consider the game: at time n we toss the coin X_i. If it is $+1$ you gain 1 dollar, otherwise you lose. The martingale is "double -or-nothing" strategy. Bet 1 dollar in the first round. If you win, you stop. If you have lost, then double your wager and keeping play this until you have recouped your loss (made 1 dollar).
Let H_n be the wager at time n. So $H_1 = 1$ and $H_n = 2H_{n-1}$. $H_n \in \sigma\{X_1, \ldots, X_{n-1}\}$. Let $M_n = \sum_{k=1}^{n} X_k$. This is a martingale. Let

$$W_n = H \cdot M = \sum_{k=1}^{n} H_k(M_k - M_{k-1})$$

Note that $\mathbb{E}\left[W_n - W_{n-1}|X_{n-1}, \ldots, X_0\right] = \mathbb{E}\left[H_n(M_n - M_{n-1})|X_{n-1}, \ldots, X_0\right] = H_n\mathbb{E}\left[X_n\right] = 0$.

4.2.3 Stopping time

Definition 4.2.19 (Stopping time)**.** Let r.v.-s $T : \Omega \to [0, \infty]$ be a stopping time of the filtration $\mathcal{F}_t$ if $\{T \leq t\} \in \mathcal{F}_t$. Then $\sigma-$field of the past before T is then defined by

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty : t \geqslant 0, A \cap \{T \leq t\} \in \mathcal{F}_t\}$$

Note that

$$\{T = \infty\} = \left(\bigcap_{n=1}^{\infty}\{T \leq n\}\right)^c \in \mathcal{F}_\infty.$$

Proposition 4.2.20. (1) Let T a stopping time, then T is $\mathcal{F}_T$ measurable.
(2) Let S, T be two stopping times s.t. $S \leq T$, Then $\mathcal{F}_S \subset \mathcal{F}_T$.
(3)Let S, T be two stopping time. Then $S \wedge T$ and $S \vee T$ are two stopping time and $\mathcal{F}_{S\wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$. Also, $\{S \leq T\} \in \mathcal{F}_{S\wedge T}$ and $\{S = T\} \in \mathcal{F}_{S\wedge T}$.
(4) If S_n is a monotone increasing (or decreasing) seq of stopping time, then $S = \lim_{n\to\infty} S_n$ is also a stopping time.

For a stopping time T, $X_n^T := X_{T\wedge n}$ is a stopped process. If M_n is a martingale (rep. supermartingale), then $\mathbb{E}M_{T\wedge n} = \mathbb{E}M_0$ (rep. $\mathbb{E}M_{T\wedge n} \leq \mathbb{E}M_0$).

Lemma 4.2.21. Let T be a stopping time. If there exist $N, \varepsilon > 0$ s.t. $\mathbb{P}(T \leq n + N|\mathcal{F}_n) > \varepsilon$ for all n. Then $\mathbb{E}T < \infty$.

Definition 4.2.22 (Predictable)**.** A stochastic process C_n is predictable (or previsible) if C_n is $\mathcal{F}_{n-1}-$measurable for all n.

A martingale transform of X is given by

$$Y_n := (C \cdot X)_n = \sum_{1 \le k \le n} C_k(X_k - X_{k-1}).$$

Theorem 4.2.23. We have following statements about $C \cdot X$

- If C_n is predictable and $0 \le C_n \le k$ for all n, ω and X is a supermartingale, then $C \cdot X$ is also a supermartingale.
- If C_n is predictable and $|C_n| \le k$ for all n, ω and X is a martingale, then $C \cdot X$ is also a martingale.
- Given two processes C_n and X_n, assume that $\mathbb{E}C_n^2 < \infty$ and $\mathbb{E}X_n^2 < \infty$. If X is a supermartingale (rep. martingale), then $C \cdot X$ is also a supermartingale (rep. martingale).

Theorem 4.2.24. Let T be a stopping time. If X is a martingale (rep. supermartingale), then X_T is a martingale (rep. supermartingale).

4.2.4 The convergence of Martingale

Fix $a < b$. Define $C_0 = 0$ and

$$C_n := \mathbf{1}_{C_{n-1}=1}\mathbf{1}_{X_{n-1} \le b} + \mathbf{1}_{C_{n-1}=0}\mathbf{1}_{X_{n-1}<a}.$$

We say C_n is "ON" if $C_n = 1$, otherwise C_n is "OFF" if $C_n = 0$. Informally, the sequence C_n is zero while waiting for the process X_n to enter $(-\infty, a)$ after which time it reverts to one and stays so while waiting for this process to enter (b, ∞). See Figure 4.1.

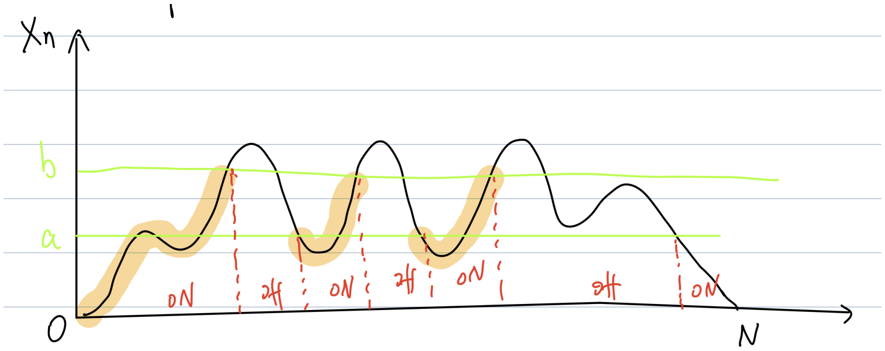


Figure 4.1: Illustration of C_n and upcrossing

It is clear that C_n is predictable since it is $\mathcal{F}_{n-1}-$measurable. Let

$$Y_n := (C \cdot X)_n = \sum_{k=1}^{n} C_k(X_k - X_{k-1})$$

which is the sum of all "ON" paths of MG.

Define $U_N[a, b]$ be the number of upcrossing until time N, which is

$$\max\{k : \exists 0 \le s_1 < t_1 < s_2 < t_2 < \cdots < s_k < t_k \le N, X_{s_i} < a, X_{t_i} > b, i = 1, \ldots, k\}$$

For every upcrossing of the interval $[a, b]$ by X_k contributes to Y_n the difference between the value of X at the end of the upcorssing, which is at least b and its value at the start of the upcrossing, which is at most

a. Thus, each up=crossing increase Y_n by at least $(b-a)$ and if $X_0 < a$ then the first upcrossing must have contributed at least $b - X_0 = (b-a) + (X_0 - a)^-$ to Y_n (Here $(X_0-a)^- = \max\{-(X_0-a), 0\} = -X_0 + a$ if $X_0 - a < 0$). Another contribution to Y_n is by the upcrossing of the interval $[a,b]$ at time n. Since it started at value at most a, its contribution to Y_n is at least $-(X_n - a)^-$. Thus,

$$Y_n \geqslant (b-a)U_n[a,b] + (X_0-a)^- - (X_n-a)^- \geqslant (b-a)U_n[a,b] - (X_n-a)^-$$

Lemma 4.2.25 (Doob's upcrossing lemma)**.** If X_n is a supermartingale, then

$$(b-a)\mathbb{E}U_n[a,b] \leq \mathbb{E}(X_n-a)^-$$

Proof. Since $Y = C \cdot X$ is a super-MG. Then $\mathbb{E}Y_n \leq \mathbb{E}Y_0 = 0$. Hence,

$$\mathbb{E}\left[(b-a)U_n[a,b] - (X_n-a)^-\right] \leq 0.$$

□

Definition 4.2.26 (L^p uniformly bounded)**.** We say $X \in L^p$ if $\mathbb{E}|X|^p < \infty$. A sequence of X_n is bounded in L^p if $\sup_n \|X_n\|_p < \infty$.

Corollary 4.2.27. Let X be a super-MG bounded in L^1. Fix $a < b$, then

$$(b-a)\mathbb{E}U_\infty[a,b] \leq |a| + \sup_n \mathbb{E}|X_n| < \infty.$$

Moreover, $\mathbb{P}(U_\infty[a,b] = \infty) = 0$.

Theorem 4.2.28 (L^1 bounded convergence or Doob's forward convergence theorem)**.** Let X be a super-MG with right-continuous sample paths. Assume that X is bounded in L^1. Then there exists a r.v. $X_\infty \in L^1$ s.t. $\lim_{n\to\infty} X_n =: X_\infty$ exists a.s. and is a.s. finite.

Proof. Note that

$$\{X_n \text{ does not converge }\} = \{\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\} = \cup_{a<b,a,b\in\mathbb{Q}}\{\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n\}$$

this means the upcorssing happens infinitely many times from a to b, which is contained by

$$\bigcup_{a<b,a,b\in\mathbb{Q}} \{U_\infty[a,b] = \infty\}.$$

Since $\mathbb{P}(U_\infty[a,b] = \infty) = 0$, then we have

$$\mathbb{P}\left(\{X_n \text{ does not converge }\}\right) \leq \mathbb{P}\left(\bigcup_{a<b,a,b\in\mathbb{Q}} \{U_\infty[a,b] = \infty\}\right) \leq \sum_{a,b} \mathbb{P}(U_\infty[a,b] = \infty) = 0.$$

Moreover, from Fatou's Lemma

$$\mathbb{E}|X_\infty| = \mathbb{E}\liminf_{n\to\infty} |x_n| \leq \liminf_{n\to\infty} \mathbb{E}|X_n| \leq \sup \mathbb{E}|X_n| < \infty$$

Hence, $|X_\infty| < \infty$ a.s. □

Another approach.

□

Remark 4.2.29. If $X_n \geqslant 0$ is a super-MG, then the condition L^1 bounded is not needed. Because $\mathbb{E}|X_n| = \mathbb{E}X_n \leq \mathbb{E}X_0 < \infty$. It still bounded by $\mathbb{E}X_0 \leq C$.

Remark 4.2.30. Doob's convergence theorem is true for sub-MG.

Theorem 4.2.31 (L^2 bounded convergence)**.** Suppose that $(X_n, \mathcal{F}_n)$ is a MG with $\mathbb{E}X_t^2 \leq K < \infty$ for all $n = 0, 1, \dots$. Then $X_n \to X_\infty$ in L^2 and *a.s.* as $n \to \infty$.

This result holds for both discrete and continuous time MG. By discretization, it is clear that it suffices to prove the result for discrete case.

Proof. Since L^2 bounded (i.e., $\sup_t \mathbb{E}|X_t|^2 < \infty$), it is also L^1 bounded. Then from the Dood's martingale convergence theorem yields $X_t \xrightarrow{a.s.} X$.

Also, it is enough to show that X_n is a Cauchy sequence in L^2.

Indeed, let $Y_n := X_n - X_{n-1}$

$$\mathbb{E}X_n^2 = \mathbb{E}(\sum_{i=0}^{n} Y_i)^2 = \sum_{i=0}^{n} \mathbb{E}(Y_i)^2 \leq C$$

for some constants. Thus,

$$\lim_{n\to\infty} \sum_{i=0}^{n} \mathbb{E}(Y_i)^2 = \sum_{n=1}^{\infty} \mathbb{E}Y_n^2 \leq C$$

Hence, the tail of sum must be zero (otherwise it cannot be bounded) when $m, n \to \infty$

$$\mathbb{E}|X_n - X_m|^2 = \sum_{i=m+1}^{n} \mathbb{E}Y_n^2 \to 0$$

Then $X_n \to X$ in L^2. □

Theorem 4.2.32. Let M be a u.i. martingale. Then $M_\infty := \lim_{n\to\infty} M_n$ exists a.s. and in L^1. Moreover, $M_n = \mathbb{E}\left[M_\infty | \mathcal{F}_n\right]$.

Since M_n is ui, there exists $c > 0$ s.t.

$$\sup_n \mathbb{E}\left[M_n; |M_n| \geqslant c\right] \leq 1.$$

Then for all $n \geqslant 0$

$$\mathbb{E}M_n \leq \mathbb{E}\left[M_n; |M_n| \geqslant c\right] + \mathbb{E}\left[M_n; |M_n| \leq c\right] \leq 1 + c$$

Hence, M_n is L^1 bounded. Hence, from Theorem 4.2.28 we have $M_n \to M$ a.s.

From $M_n \to M$ in probability and ui, $M_n \to M$ in L^1.

Fix n and event $F \in \mathcal{F}_n$,

$$|\mathbb{E}[M_n; F] - \mathbb{E}[M_\infty; F]| \leq \mathbb{E}[|M_n - M_\infty; F] \leq \mathbb{E}[|M_n - M_\infty] \to 0$$

as $n \to \infty$. Hence, $\mathbb{E}[M_n; F] = \mathbb{E}[M_\infty; F]$ for all $F \in \mathcal{F}_n$.

4.2.5 The optional stopping theorem

The optional sampling theorem for bounded stopping time is as follows.

Theorem 4.2.33 (Doob's optional stopping theorem)**.** Let T be a stopping time and X be a supermartingale. Assume that if one of the following conditions holds

- T is bounded
- X_n is bounded and T is finite a.s.
- $\mathbb{E}T < \infty$ and $|X_n - X_{n-1}|$ is bounded.

Then $X_T \in L^1$ and $\mathbb{E}X_T \leq \mathbb{E}X_0$.

Proof. For each of these three conditions, we alwalys have $T < \infty$ a.s. If follows that

$$T \wedge n \to T$$

as $n \to \infty$. Then $X_{T\wedge n} \to X_T$ a.s.

Note that $\mathbb{E}X_{T\wedge n} \leq \mathbb{E}X_0$ for all n (since X_t is a super-MG).

(1) If T is bounded, then for $n > T$ we have

$$\mathbb{E}X_0 \geqslant \mathbb{E}X_{T\wedge n}$$

(2) If X_n is bounded, then $X_{T\wedge n} \leq C$ for a constant $C > 0$. From D.C.T.,

$$\mathbb{E}X_T = \lim \mathbb{E}X_{T\wedge n} \leq \mathbb{E}X_0.$$

(3) If $\mathbb{E}T < \infty$ and $|X_n - X_{n-1}| \leq C$ for a constant $C > 0$. Then

$$|X_{T\wedge n}| = |\sum_{k=1}^{T\wedge n}(X_k - X_{k-1}) \leq (T\wedge n)C \leq TC < \infty.$$

From D.C.T, since X_t is a super-MG

$$\mathbb{E}(X_T - X_0) = \mathbb{E}\left[\lim_{n\to\infty} X_{T\wedge n} - X_0\right] = \lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n} - X_0\right] \leq 0$$

□

Remark 4.2.34. If $X \geqslant 0$ and T is a.s. finite stopping time. Then $EX_T \leq \mathbb{E}X_0$. Indeed, from Fatou's Lemma,

$$\mathbb{E}X_T = \mathbb{E}\liminf_{n\to\infty} X_{T\wedge n} \leq \liminf_{n\to\infty} \mathbb{E}X_{T\wedge n} \leq \liminf_{n\to\infty} \mathbb{E}X_0 = \mathbb{E}X_0.$$

Corollary 4.2.35. Let X_n be a super-MG and for two bounded stopping time S, T adapted to $\{\mathcal{F}_n\}$ with $0 \leq S \leq T \leq N$ for some $N < \infty$. Then we have $\mathbb{E}\left[X_T|\mathcal{F}_S\right] \leq X_S$. Then $X_T \in L^1$ and $\mathbb{E}\left[X_T|\mathcal{F}_s\right] \leq X_S$ a.s.

Proof. Note that $\mathbb{E}X_T \leq \mathbb{E}X_0 \leq \infty$. To show that $\mathbb{E}\left[X_T|\mathcal{F}_s\right] \leq X_S$, it suffices to prove that for all $A \in \mathcal{F}_S$,

$$\mathbb{E}\left[X_T; A\right] \leq \mathbb{E}\left[X_S; A\right]$$

Indeed, noting that $X_T - X_S = \sum_{n=1}^N 1_{S<n\leq T}(X_n - X_{n-1})$. Taking expectation over A,

$$\begin{aligned}\mathbb{E}\left[X_T - X_S; A\right] &= \mathbb{E}\left[\sum_{n=1}^N 1_{S<n\leq T}(X_n - X_{n-1})\right]\\ &= \sum_{n=1}^N \mathbb{E}\left[X_n - X_{n-1}; A\cap S < n \leq T\right]\\ &= \sum_{n=1}^N \mathbb{E}\left[\mathbb{E}\left[X_n - X_{n-1}; A\cap S < n \leq T\right]|\mathcal{F}_{n-1}\right]\\ &= \sum_{n=1}^N \mathbb{E}\left[\mathbb{E}\left[X_n - X_{n-1}|\mathcal{F}_{n-1}\right]\mathbf{1}_{A\cap S<n\leq T}\right] \leq 0.\end{aligned}$$

where the second equality follows from $A \cap \{S < n \leq T\} = A \cap \{S \leq n-1\} \cap \{n-1 < T\} \in \mathcal{F}_{n-1}$. □

Our optional sampling theorem for closed super-MG in discrete time is as follow. We do not assume that stopping is bounded as in Corollary 4.2.35.

Proposition 4.2.36. Suppose that $\{X_n\}$ is a non-negative supermartingale on $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=0,1,2,\ldots}, \mathbb{P})$. Let $X_\infty = 0$. If $S, T : \Omega \to \mathbb{Z}\cup\{\infty\}$ are $\mathcal{F}_n-$measurable stopping time with $S \leq T$. Then

- $\mathbb{E}X_T < \infty$
- $\mathbb{E}(X_T|\mathcal{F}_S) \leq X_S$

Proof. For first part, see Remark 4.2.29.

For second part, it suffices to show that for all $A \in \mathcal{F}_S$

$$\mathbb{E}(X_T; A) \leq \mathbb{E}(X_S; A)$$

Indeed, fixed n

$$\begin{aligned}\mathbb{E}(X_T; A \cap \{T \leq n\}) &= \mathbb{E}(X_{T\wedge n}; A \cap \{T \leq n\}) \\ &\leq \mathbb{E}(X_{T\wedge n}; A \cap \{S \leq n\}) \\ &\leq \mathbb{E}(X_{S\wedge n}; A \cap \{S \leq n\}) = \mathbb{E}(X_S; A \cap \{S \leq n\})\end{aligned}$$

where the last inequality from the bounded of $S \wedge n$ and $T \wedge n$ with $S \wedge n \leq T \wedge n$ and apply Corollary 4.2.35 and $A \cap \{S \leq n\} \in \mathcal{F}_{S\wedge n} \subset \mathcal{F}_n$. [10]

From M.C.T.,

$$\lim_{n\to\infty} \mathbb{E}(X_T; A \cap \{T \leq n\}) = \mathbb{E}(X_T; A \cap \{T < \infty\})$$

and

$$\lim_{n\to\infty} \mathbb{E}(X_S; A \cap \{S \leq n\}) = \mathbb{E}(X_S; A \cap \{S < \infty\})$$

Hence,

$$\mathbb{E}(X_T; A \cap \{T < \infty\}) \leq \mathbb{E}(X_S; A \cap \{S < \infty\}).$$

Also,

$$\mathbb{E}(X_T; A \cap \{T = \infty\}) = \mathbb{E}(X_S; A \cap \{S = \infty\}) = 0.$$

Hence, combing this equality to above we get for all $A \in \mathcal{F}_S$

$$\mathbb{E}(X_T; A) = \mathbb{E}(X_S; A).$$

□

Definition 4.2.37 (Closed supermartingale)**.** Let $\{X_n\}_{n=0,1,\ldots}$ be a supermartingale on $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=0,1,2,\ldots}, \mathbb{P})$. We say $\{X_n\}_{n=0,1,\ldots}$ is **closed** by X_∞ if there exists a r.v. $X_\infty \in L^1$ and $\mathbb{E}\left[X_\infty | \mathcal{F}_n\right] \leq X_n$ a.s. for $n = 0, 1, \ldots$.

The optional stopping theorem for closed super-MG in discrete time as follows.

Theorem 4.2.38. Let $\{X_t\}_{t\geqslant 0}$ be a $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\geqslant 0}, \mathbb{P})$ supermartingale closed by X_∞. If $S, T : \Omega \to \mathbb{R} \cup \{\infty\}$ be two $\{\mathcal{F}\}_{t\geqslant 0}$ stopping times with $S \leq T$. Then

- $\mathbb{E}|X_T| \leq \infty$
- $\mathbb{E}\left[X_T | \mathcal{F}_s\right] \leq X_s$ a.s.

To state our optional sampling theorem for closed supermartingale for continuous time, we need to define negatively indexed supermartingale.

Let $\{X_n\}_{n=0,-1,-2,\ldots}$ associated filtration $\{\mathcal{F}\}_{n=0,-1,-2,\ldots}$. We have $\mathcal{F}_M \subset \mathcal{F}_n$ for $m \leq n$. Similarly, we say X_n is a supermartingale if

- $X_n \in L^1(\Omega, \mathcal{F}, \mathbb{P})$
- $\mathbb{E}(X_n | \mathcal{F}_m) \leq X_m$

Theorem 4.2.39. Let $\{X_n\}_{n=0,-1,-2,\ldots}$ be a negatively indexed supermartingale on $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=0,-1,-2,\ldots}, \mathbb{P})$ s.t. $\sup_{-\infty<n\leq 0} \mathbb{E}X_n < \infty$. Then $\{X_n\}$ is uniformly integrable.

Proof.

□

[10] Because $A \cap \{S \leq n\} \in \mathcal{F}_S$ and $A \cap \{S \leq n\} \in \mathcal{F}_n$ so that $A \cap \{S \leq n\} \in \mathcal{F}_T \cap \mathcal{F}_n = \mathcal{F}_{S\wedge n}$

The optional sampling theorem for closed supermartingale continuous time is as follows.

Theorem 4.2.40. Let $\{X_t\}_{t\geqslant 0}$ be a $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\geqslant 0}, \mathbb{P})$ supermartingale with right-continuous path and closed. If $S, T : \Omega \to \mathbb{R} \cup \{\infty\}$ be two $\{\mathcal{F}\}_{t\geqslant 0}$ stopping times with $S \leq T$. Then

- $\mathbb{E}|X_T| \leq \infty$
- $\mathbb{E}\left[X_T|\mathcal{F}_s\right] \leq X_s$ a.s.

Corollary 4.2.41. A right-continuous, adapted process X_t is a MG iff for every bounded stopping time τ, $X_\tau \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}X_\tau = \mathbb{E}X_0$.

Proof.

(1) "$\Rightarrow$": use Theorem 4.2.33.

(2) "$\Leftarrow$": Take two $\mathcal{F}_t$-stopping time T, S s.t. $S < T$. Set $\tau = S\mathbf{1}_A + T\mathbf{1}_{A^c}$ for a fix $A \in \mathcal{F}_S$. Note that τ is also a $\mathcal{F}_t$-stopping time

$$\begin{aligned}\{\tau \leq t\} &= (A \cap \{S \leq t\}) \bigcup (A^c \cap \{T \leq t\}) \\ &= (A \cap \{S \leq t\}) \bigcup ((A^c \cap \{S \leq t\}) \cap \{T \leq t\}) \in \mathcal{F}_t\end{aligned}$$

Then

$$\mathbb{E}X_0 = \mathbb{E}X_\tau = \mathbb{E}[X_S\mathbf{1}_A + X_T\mathbf{1}_{A^c}]$$

and also

$$\mathbb{E}X_0 = \mathbb{E}X_\tau = \mathbb{E}[X_\tau\mathbf{1}_A + X_\tau\mathbf{1}_{A^c}]$$

subtracting the finite $\mathbb{E}X_T\mathbf{1}_{A^c}$ yields for all $A \in \mathcal{F}_S$

$$\mathbb{E}\left[X_\tau\mathbf{1}_A\right] = \mathbb{E}\left[X_S\mathbf{1}_A\right].$$

□

Corollary 4.2.42. Let X_t be a continuous time and right-continuous MG. A stopped MG $X_{T\wedge t}$ for stopping time T is a MG.

Proof. Let $Y := X_{T\wedge t}$. Y is adapted to $\mathcal{F}_{T\wedge t}$. Let S be a bounded stopping time. Then apply Corollary 4.2.35

$$\mathbb{E}Y_S = \mathbb{E}X_{T\wedge t\wedge S} = \mathbb{E}X_0 = \mathbb{E}Y_0.$$

Hence, from Corollary 4.2.41 Y is a MG. □

Corollary 4.2.43 (Optional stopping time for ui MG)**.** Let X_t be a ui MG with right-continuous sample paths. Let S, T be two stopping times with $S \leq T$. Then X_S and X_T are in L^1 and $X_S = \mathbb{E}[X_T|\mathcal{F}_S]$

Proof. Apply Theorem 4.2.47, X_t is closed. Then this reuslt is directly from Theorem 4.2.40. □

Corollary 4.2.44. Let X_t be a MG with right-continuous sample paths. Let T be stopping time taking value on $\mathbb{R}$ or $\mathbb{Z}$ s.t. $|X_{t\wedge\tau}| \leq C$ a.s. for a constant $C > 0$ and all t. Then $\mathbb{E}X_T = \mathbb{E}X_0$, where $X_T = \lim_{t\to\infty} X_{t\wedge T}$ a.s.

Proof. Since $X_{t\wedge T}$ is bounded, then it is ui. Thus, $X_{t\wedge T} \to X_T$ in L^1 and a.s. from Theorem 4.2.47. Then

$$\mathbb{E}X_0 = \lim_{t\to\infty} \mathbb{E}X_{t\wedge T} = \mathbb{E}X_T.$$

□

4.2.6 Continuous-time Martingale

The definition of continuous-time Martingale is similar as discrete-time. If X_t is a sub-MG, then $(-X_t)$ is a super-MG. So we only consider super-MG. Also, for a super-MG we have $\mathbb{E}X_s \geqslant \mathbb{E}X_t$ with $0 \leq s \leq t$. A simple way to construct a MG is to take a r.v. $Z \in L^1$ and to set $X_t = \mathbb{E}\left[Z|\mathcal{F}_t\right]$ for every $t \geqslant 0$. Let us to turn others important class of examples.

Example 4.2.45. If Z is a real-value process having independent increments w.r.t. $\mathcal{F}_t$ (i.e., $Z_t - Z_s$ is independent of $\mathcal{F}_s$ for $0 \leq s < t$). Then

- if $Z_t \in L^1$ for $t \geqslant 0$, then $X_t = Z_t - \mathbb{E}Z_t$ is a MG;
- if $Z_t \in L^2$ for $t \geqslant 0$, then $Y_t = X_t^2 - \mathbb{E}X_t$ is a MG;
- if for some $\theta \in \mathbb{R}$, we have $\mathbb{E}[\exp(\theta Z_t)] < \infty$ for $t \geqslant 0$, then
$$W_t = \frac{e^{\theta Z_t}}{\mathbb{E}e^{\theta Z_t}}$$
is a MG.

Remark 4.2.46. For a BM $B_t \sim N(0,t)$, we take $Y_t = B_t^2 - t$, which is a MG. Also, B_t is still a MG and $e^{\theta B_t - \frac{\theta^2}{2}t}$ is also a MG (called exponential MG of BM).

The following theorem is about the equivalent of closed and u.i.

Theorem 4.2.47. Let X_t be right-continuous MG. The following statements are equivalent.

- $\lim_{n\to\infty} X_n$ exists in L^1 (also in a.s.)
- X_t is closed by X_∞
- X_t is ui.

Proof. $(i) \Rightarrow (ii)$: Note that let $h \to \infty$ and this gives the desired result (using the fact that the conditional expectation is continuous for the L^1 norm)[11]

$$\mathbb{E}X_t = \lim_{h\to\infty} \mathbb{E}\left[X_{t+h}|\mathcal{F}_t\right] = \mathbb{E}\left[X_\infty|\mathcal{F}_t\right].$$

$(ii) \Rightarrow (iii)$: From proposition 4.2.7, for X_∞, the collection of all r.v.-s $\mathbb{E}[X|\mathcal{H}]$ is u.i. for all sub-$\sigma-$algebra $\mathcal{H} \subset \mathcal{F}$.

$(iii) \Rightarrow (i)$: See Theorem 4.2.32

□

Let X_t be a MG or super-MG with right-continuous sample paths and such that X_t converges a.s. as $t \to \infty$ to a r.v. denoted by X_∞. Then for every stopping time T, we write X_T for the r.v.

$$X_T(\omega) = \mathbf{1}_{T(\omega)<\infty} X_{T(\omega)}(\omega) + \mathbf{1}_{T(\omega)=\infty} X_{T(\infty)}(\omega),$$

where X_T is $\mathcal{F}_T$-measurable.

Indeed, We want to show $\{1_{T=\infty}X_\infty \in B\} \in \mathcal{F}_T$ for any Borel set B. By the definition of $\mathcal{F}_T$, this means we must show $\{1_{T=\infty}X_\infty \in B\} \cap \{T \leq t\} \in \mathcal{F}_t$ for all $t > 0$. On the event $\{T \leq t\}$, if $0 \in B$ $1_{T=\infty}X_\infty = 0$, so

$$\{1_{T=\infty}X_\infty \in B\} \cap \{T \leq t\} = \{T \leq t\}$$

Also, if $0 \notin B$

$$\{1_{T=\infty}X_\infty \in B\} \cap \{T \leq t\} = \emptyset$$

. In either case, $\{1_{T=\infty}X_\infty \in B\} \cap \{T \leq t\} \in \mathcal{F}_t$, so $1_{T=\infty}X_\infty$ is $\mathcal{F}_T$ measurable.

[11]
$$\lim_{n\to\infty} \|E(X_n|\mathcal{G}) - E(X|\mathcal{G})\|_p = \lim_{n\to\infty} \|E(X_n - X|\mathcal{G})\|_p \leqslant \lim_{n\to\infty} \|X_n - X\|_p = 0$$

Theorem 4.2.48 (Optional stopping theorem for MG)**.** Let X_t be a u.i. MG with right-continuous sample paths. Let S, T be two stopping time with $S \leq T$. Then X_S and X_T are in L^1 and

$$X_S = \mathbb{E}\left[X_T|\mathcal{F}_S\right].$$

In particular, for every stopping time S we have

$$X_S = \mathbb{E}\left[X_\infty|\mathcal{F}_S\right]$$

and

$$\mathbb{E}X_S = \mathbb{E}X_\infty = \mathbb{E}X_0.$$

Corollary 4.2.49. Let X_t be a MG with right-continuous sample paths, and let $S \leq T$ be two bounded stopping time. Then X_S and X_T are in L^1 and

$$X_S = \mathbb{E}\left[X_T|\mathcal{F}_S\right].$$

Proof.

□

The next corollary shows that a MG (resp. a ui MG) stopped at an arbitrary stopping time remains a MG (resp. a ui MG). This result play an important role in the continuous semi-MG.

Corollary 4.2.50. Let X_t be a MG with right-continuous sample paths, and let T be a stopping time.

- The process $X_{t\wedge T}$ is still a MG
- Suppose in addition that X_t is ui. Then The process $X_{t\wedge T}$ is still a ui MG and moreover, we have for every $t \geqslant 0$

$$X_{t\wedge T} = \mathbb{E}\left[X_T|\mathcal{F}_t\right].$$

Exercise 8. Let M be a martingale with continuous sample paths with $M_0 = x \geqslant 0$. We assume that $M_t \geqslant 0$ for every $t \geqslant 0$ and $M_t \to 0$ as $t \to \infty$ a.s. Show that for every $y > x$,

$$\mathbb{P}(\sup_{t\geqslant 0} M_t \geqslant y) = \frac{x}{y}.$$

Proof. Let $T_y := \inf\{t : M_t = y\}$. Denote by $X_t := M_{t\wedge T_y}$ the stopped process (this is a martingale). Since $|X_t| \leq y$ is bounded, X_t is uniformly integrable.[12] From optional stopping time theorem,

$$x = \mathbb{E}M_0 = \mathbb{E}X_0 = \mathbb{E}X_{T_y} = \mathbb{E}M_{T_y}$$

Since for any stopping time

$$M_{T_y} = 1_{T_y<\infty}M_{T_y} + 1_{T_y=\infty}M_\infty = 1_{T_y<\infty}M_{T_y}$$

where the second part is zero due to $M_t \to 0$ as $t \to \infty$.

Hence,

$$x = \mathbb{E}M_{T_y} = \mathbb{E}[\mathbf{1}_{T_y<\infty}M_{T_y}] = y\mathbb{E}\mathbf{1}_{T_y<\infty} = y\mathbb{P}(T_y < \infty)$$

where $\mathbb{P}(T_y < \infty) = \mathbb{P}(\sup_{t\geqslant 0} M_t \geqslant y)$. This gives the desired result. □

Remark 4.2.51. It is crucial to always verify the uniform integrability of MG for applying the optional stopping time theorem. In most cases, this is done by verifying that the (stopped) MG is bounded.

[12] Apply Theorem 4.2.8 , it follows from $\mathbb{E}|X_t|^2 \leq \mathbb{E}a^2 = a^2$.

4.3 Brownian motion and stochastic analysis

Let's see our informal definition of Itô's formula, Consider a process that evolves as

$$dX_t = \mu_t\, dt + \sigma_t dB_t$$

An informal way to write this looks like

$$X_{t+dt} - X_t = \mu_t dt + \sigma_t N(0,\, dt)$$

Let $f : \mathbb{R} \to \mathbb{R}$ be a twice-differentiable function. Note that if X_t folows a deterministic smooth trajectory, then we know how $f(X_t)$ evolves: we will have $df(X_t) = f'(X_t)dX_t$. However, if we expand the stochastic version out, we find that

$$\begin{aligned} df(X_t) &= f'(X_t)\, dX_t + \frac{1}{2} f''(X_t)(\, dX_t)^2 \\ &= f'(X_t)[\mu_t\, dt + \sigma_t dB_t] + \frac{f''(X_t)}{2}[\mu_t\, dt + \sigma_t dB_t]^2 \end{aligned}$$

and since $dB_t \sim \sqrt{dt}$ and $\sigma_t^2(dB_t)^2 = \sigma_t^2 dt$. Hence, dropping term $(\mu_t\, dt)^2$ (dominated by dB_t) we get

$$df(X_t) = f'(X_t)[\mu_t\, dt + \sigma_t dB_t] + \frac{f''(X_t)}{2}\sigma_t^2 dt = \left(f'(X_t)\mu_t + \frac{f''(X_t)}{2}\sigma_t^2 \right) dt + f'(X_t)\sigma_t dB_t$$

where the first term is called the drift and second term called a stochastic term.

4.3.1 Stieltjes integral

Definition 4.3.1 (bounded variation)**.** Suppose $F(t)$ is a complex-valued function defined on $[a, b]$. The function F is said to be of **bounded variation** (B.V.) if the variations of F over all partitions are bounded, that is. there exists $M < \infty$ so that

$$\sum_{j=1}^{N} |F(t_j) - F(t_{j-1})| \le M$$

for all partitions $a = t_0 < t_1 < \cdots < t_N = b$.

Remark 4.3.2. A counterexample for a continuous function on $[0, 1]$ but not of bounded variation. See [13]

Theorem 4.3.3. A real-valued function F on $[a, b]$ is of bounded variation i.f.f. F is the difference of two increasing bounded functions.

Theorem 4.3.4. If F is of bounded variation on $[a, b]$, then F is differentiable a.e.

In other words, the quotient

$$\lim_{k \to 0} \frac{F(x + k) - F(x)}{k}$$

exists for a.e. $x \in [a, b]$.

Definition 4.3.5 (absolutely continuous)**.** A function F defined on $[a, b]$ is **absolutely continuous** (a.c.)

[13] https://math.stackexchange.com/questions/1082897/a-continuous-function-on-0-1-not-of-bounded-variation

if for any $\varepsilon > 0$ there exists $\delta > 0$ so that

$$\sum_{k=1}^{N} |F(b_k) - F(a_k)| < \varepsilon, \text{ whenever } \sum_{k=1}^{N} (b_k - a_k) < \delta$$

and the intervals (a_k, b_k) are disjoint for $k = 1, 2, \ldots, N$.

From the definition, it is clear that absolutely continuous functions are continuous. If F is absolutely continuous on a bounded interval, then it is also of bounded variation on the same interval. As a consequence the decomposition of such a function F into two monotonic functions are both continuous. If $F(x) = \int_a^x f(y)dy$ where f is integrable, then F is absolutely continuous.

Theorem 4.3.6. If F is absolutely continuous on $[a, b]$, then $F'(x)$ exists a.e. Moreover, if $F'(x) = 0$ for a.e. x, then F is constant. 0

The next theorem shows that the establishing the reciprocity between differentiation and integration.

Theorem 4.3.7. Suppose F is absolutely continuous on $[a, b]$. Then F' exists a.e. and is integrable. Moreover,

$$F(x) - F(a) = \int_a^x F'(y)dy$$

for all $x \in [a, b]$.
Conversely, if f is integrable on $[a, b]$, then there exists an absolutely continuous function F s.t. $F'(x) = f(x)$ a.e. and in fact, we may take $F(x) = \int_a^x f(y)dy$.

We say that the function F has a jump discontinuity if If $F(x^-) := \lim_{y \to x, y < x} F(y) \neq F(x^+) := \lim_{y \to x, y > x} F(y)$. Let $\{x_n\}_{n=1}^{\infty}$ denote the points where F is discontinuous (since F on $[a, b]$ has at most countably many discontinuities), and let $\alpha_n := F(x_n^+) - F(x_n^-)$. Then

$$F(x_n) = F(x_n^-) + \theta_n \alpha_n$$

for some $\theta_n \in [0, 1]$.

The Stieltjes integral was introduced to provide a generalization of the Riemann integral $\int_a^b f(x)dx$, where the increments dx were replaced by the increments $dF(x)$ for a given increasing function F on $[a, b]$. The following theorem shows that to have a unique correspondence between measures and increasing functions we shall have normalize these functions appropriately.

Theorem 4.3.8. Let F be an increasing function on $\mathbb{R}$ that is right-continuous. Then there is a unique measure μ (also denoted by dF) on the Borel sets $\mathcal{B}$ that is finite on bounded intervals, then F defined by $F(x) = \mu((0, x])$ for $x > 0$, $F(0) = 0$, and $F(x) = -\mu((-x, 0])$ for $x < 0$ is increasing and right-continuous.

Remark 4.3.9.

- The condition that μ be finite on bounded intervals is crucial.
- Two increasing functions F and G give the same measure if $F - G$ is constant.
- When F is r.c. increasing on a bounded interval, we may decompose it into a convex combination of three different increasing functions: a r.c. discrete increasing function, a singular continuous increasing function, and an absolutely continuous increasing function. This implies the Lebesgue-Stieltjes measure associated with r.c. function F can be decomposed into three parts. It is quite difficult to compute the Stieltjes integral when F is singular continuous. So we consider the absolutely continuous case as follows.

- IF F is an absolutely continuous function on $[a,b]$, then
$$\int_a^b f(x)dF(x) = \int_a^b f(x)F'(x)dx$$
for every Borel measurable function f that is integrable w.r.t. $\mu = dF$.
- Suppose F is a pure jump function with jumps $\{\alpha\}_{n=1}^{\infty}$ at points $\{x_n\}$. Then whenever f is continuous and vanishes outside some finite interval we have
$$\int_a^b f(x)dF(x) = \sum_{n=1}^{\infty} f(x_n)\alpha_n = \sum_{n=1}^{\infty} f(x_n)(F(x_n^+) - F(x_n^-)).$$

Other properties of Stieltjes integral are given as follows.

Proposition 4.3.10. If $\int_0^a |f(x)||dF(x)| < \infty$ for $f \in \mathcal{B}([0,a])$, then $g(t) := \int_0^t f(s)dF(s)$ is of r.c. on $[0,a]$.

Proof. It follows from the D.C.T. □

Proposition 4.3.11. If $\int_0^a |f(x)||dF(x)| < \infty$ for $f \in \mathcal{B}([0,a])$, then $g(t) := \int_0^t f(s)dF(s)$ is of B.V. on $[0,a]$.

Proof. One can decompose $f = f^+ - f^-$ for $f^+, f^- \geqslant 0$ and $dF = \mu + \nu$ for ν is a negative measure. Then
$$\int_0^t f(s)dF(s) = (\int_0^t f^+(s)\mu(s) - \int_0^t f^-(s)\nu(s)) - (\int_0^t f^-(s)\mu(s) - \int_0^t f^+(s)\nu(s)).$$
So $g(t)$ is the difference of two increasing functions on $[0,a]$. Hence, $g(t)$ is of B.V. from Theorem 4.3.3. □

Definition 4.3.12 (Absolute continuity of measures)**.** If ν is a signed measure and μ a measure on $(\Omega, \mathcal{F})$, we say that ν is absolutely continuous w.r.t. μ if
$$\nu(E) = 0 \text{ whenever } E \in \mathcal{F} \text{ and } \mu(E) = 0.$$

If $f \in L^1(\Omega, \mu)$, then the singed measure ν defined by
$$\nu(E) = \int_E f d\mu$$
is absolutely continuous w.r.t. μ. We use the shorthand $d\nu = f d\mu$.

To guarantee a converse statement as above was proved in the case of $\mathbb{R}$ by Lebesgue and in the general case by Radon and Nikodym as follows.

Theorem 4.3.13 (Radon-Nikodym)**.** Suppose μ is a $\sigma-$finite positive measure on the measure space $(\Omega, \mathcal{F})$ and ν a $\sigma-$finite signed measure on $\mathcal{F}$. Then there exist unique signed measures ν_a and ν_s on $\mathcal{F}$ s.t. $\nu_a << \nu, \nu_s \perp \mu$ and $\nu = \nu_a + \nu_s$. In addition, the measure ν_a takes the form $d\nu_a = f d\mu$, that is,
$$\nu_a(E) = \int_E f(x)d\mu(x)$$
for some extended $\mu-$integrable function f.

If ν is a.c. w.r.t. μ, then $d\nu = f d\mu$. This is a general result of Theorem 4.3.7.

4.3.2 Gaussian processes

Definition 4.3.14. $Z = (Z_1, \dots, Z_n)$ is a Gaussian vector (or multivariate normal) $\mathcal{N}(\mu_z, \Sigma_z)$ with mean $\mu_z = (\mu_1, \dots, \mu_n)$ and covariance matrix $\Sigma_z(i,j) = \mathrm{Cov}\,(Z_i, Z_j)$ non-singular if

$$f_Z(z) = \frac{1}{(2\pi)^{p/2}|\Sigma_Z|^{1/2}} \exp\left(-\frac{1}{2}(Z-\mu)^T \Sigma_Z^{-1}(Z-\mu)\right).$$

Definition 4.3.15. $\{X_t\}_{t\in[0,T]}$ is called Gaussian process if for any $0 \le t_1 \le \cdots \le t_d \le T$, $(X_{t_1}, \dots, X_{t_d})$ is a Gaussian vector.

Remark 4.3.16. Another way: A stationary process (joint distribution $(X_{t_1}, \dots, X_{t_d})$ is shift invariant under time index) $\{X_t\}_{t\in[0,T]}$ is a Gaussian process if for any $0 \le t_1 \le \cdots \le t_d \le T$, $(X_{t_1}, \dots, X_{t_d})$ is a Gaussian vector.

Theorem 4.3.17. Let $\{X_t\}_{t\in[0,T]}$ be a stochastic process with $X_0 = 0$. If

- the increments $X_{t_k} - X_{s_k}$ are independent for a collection of disjoint intervals $(s_k, t_k) \subset [0, T]$.
- $\mathbb{P}(X_t \in C([0,T]) = 1$ (i.e., a.s. continuous),

then $\{X_t\}_{t\in[0,T]}$ is Gaussian process.

Proof.

□

Remark 4.3.18. The continuity condition is important, without it we can give a counterexample such as the compound Poisson process.

Remark 4.3.19. To characterize a Gaussian process with continuous trajectories. It is enough to know $a(t) := \mathbb{E}X_t$ and $c(s,t) := \mathrm{Cov}\,(X_s, X_t)$.

4.3.3 Brownian motion

Definition 4.3.20 (Brownian motion)**.** A process $\{B_t\}_{t\in[0,T]}$ is called standard 1D Brownian motion (B.M.) or Wiener process if B_t is a Gaussian process with continuous trajectory s.t. $\mathbb{E}B_t = 0$ and $\mathrm{Var}\,[B_t] = t$.

Definition 4.3.21. An adapted process B_t is called a standard 1-dim BM if

- $B_0 = 0$
- B_t has continuous paths
- (independent increments) $B_t - B_s$ is independent of $\mathcal{F}_s$ for $t > s$
- (stationary increments) $B_t - B_s \sim N(0, t-s)$

Definition 4.3.22. An adapted process B_t is a BM iff

- B_t is a Gaussian process
- B_t has continuous paths
- $\mathbb{E}B_t = 0$ and $\mathbb{E}B_tB_s = t \wedge s$ for any $t, s \geqslant 0$.

Theorem 4.3.23. Let X be a $\mathcal{F}_t-$adapted continuous process vanishing at zero. Then X is an $\mathcal{F}_t$ BM iff both X_t and $X_t^2 - t$ are martingales.

Theorem 4.3.24 (Strong Markov property)**.** Let τ be a stopping time w.r.t $\mathcal{F}$. Then

$$B_t^\tau = \{B_{\tau+t} - B_\tau\}_{t\geqslant 0}$$

is BM and independent of $\mathcal{F}_\tau$.

Proposition 4.3.25. Let B_t be a BM. Then the process $X_0 = 0$ and $X_t = tB_{1/t}$ is a BM. Moreover, $\mathbb{P}(\inf_{t>0}\{B_t = 0\} = 0) = 1$.

Theorem 4.3.26 (Reflection principle)**.** Let $a > 0$ and $\tau_a := \inf\{t : B_t = a\}$. Denote $M_t := \sup_{s\in[0,t]} B_s$. Then

$$\mathbb{P}(M_t \geqslant a) = \mathbb{P}(\tau_a \leq t) = 2\mathbb{P}(B_t \geqslant a) = \mathbb{P}(|B_t| \geqslant a).$$

Proof. Since $\{B_t \geqslant a\} \subset \{\tau_a \leq t\}$, then

$$\mathbb{P}(B_t \geqslant a) = \mathbb{P}(\tau_a \leq t, B_t \geqslant a) = \mathbb{P}(\tau_a \leq t, B_t - B_{\tau_a} \geqslant 0)$$

From strong Markov property, since τ_a is a stopping time, then $B_t - B_{\tau_a}$ is independent of $\{\tau_a \leq t\}$-$\mathcal{F}_{\tau_a}$measurable. Thus,

$$\mathbb{P}(\tau_a \leq t, B_t - B_{\tau_a} \geqslant 0) = \mathbb{P}(\tau_a \leq t)\mathbb{P}(B_t - B_{\tau_a} \geqslant 0)$$

where

$$\mathbb{P}(B_t - B_{\tau_a} \geqslant 0) = \mathbb{E}\left[\mathbb{E}(\mathbf{1}_{B_t - B_{\tau_a}\geqslant 0}|\tau_a)\right] = \frac{1}{2}.$$

Hence,

$$\mathbb{P}(B_t \geqslant a) = \frac{1}{2}\mathbb{P}(\tau_a \leq t).$$

□

Remark 4.3.27. Note that

$$\mathbb{P}(B_t \geqslant a) = \mathbb{P}(B_t \geqslant a|\tau_a > t)\mathbb{P}(\tau_a > t) + \mathbb{P}(B_t \geqslant a|\tau_a \leq t)\mathbb{P}(\tau_a \leq t) = 0 + \frac{1}{2}\mathbb{P}(\tau_a \leq t)$$

Hence,

$$\begin{aligned}\mathbb{P}(\tau_a \leq t) = 2\mathbb{P}(B_t \geqslant a) &= 2\mathbb{P}(\sqrt{t}B_1 \geqslant a)\\ &= 2\mathbb{P}(B_1 \geqslant a/\sqrt{t})\\ &= 2\int_{a/\sqrt{t}}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx\end{aligned}$$

Hence, the density function of τ_a is given by

$$f_{\tau_a}(t) = \frac{a}{\sqrt{2\pi t^{3/2}}}e^{-a^2/2t} \sim \frac{c}{t^{3/2}}$$

that is $f_{\tau_a}(t) = o(ct^{-3/2})$. This implies

$$\mathbb{E}\tau_a = \int_0^\infty t\frac{c}{t^{3/2}}dt \geqslant \int_1^\infty \frac{c}{t^{1/2}}dt = \infty$$

and

$$\mathbb{E}(\tau_a)^2 = \infty.$$

Chapter 5

Mixing and hitting times for Markov chains

These are notes from the Online Open Probability School (OOPS) 2020 [1]*. The instructor of this course is Perla Sousi* [2] *(University of Cambridge).*

Mixing times for Markov chains is an active area of research in modern probability and it lies at the interface of mathematics, statistical physics and theoretical computer science. The mixing time of a Markov chain is defined to be the time it takes to come close to equilibrium. There is a variety of techniques used to estimate mixing times, coming from probability, representation theory and spectral theory. In this mini course I will focus on probabilistic techniques and in particular, I will present some recent results (see references below) on connections between mixing times and hitting times of large sets.

This lecture note will cover results from 3 papers.

1. Equivalence (up to constants) between mixing times and hitting times of large sets
2. Hitting times: comparison for different sizes of sets
3. Refined mixing and hitting equivalence

5.1 Basic

Let X be an irreducible Markov chain in a finite state space S. (You can go from any state to any other state in a finite number of steps with positive probability.) Let P be the transition matrix of X. Let $P^t(i,j) = \mathbb{P}_i(X_t = j)$ for all $i,j \in S$ (starting at i, get to j in t steps.

There exists an invariant distribution π, $\pi = \pi P$. If X is also aperiodic, then $P^t(x,y) \to \pi(y)$ as $t \to \infty$, forall x, y.

We use the total variation distance. Let μ and ν be 2 probability distributions on S. Let

$$|\mu - \nu|_{TV} = \max_{A \subset S} |\mu(A) - \nu(A)|.$$

Let

$$d(t) = \max_x \left| P^t(x, \cdot) - \pi \right|_{TV}.$$

[1] Webpage for this course: https://secure.math.ubc.ca/Links/OOPS/abs_Sousi.php.

[2] http://www.statslab.cam.ac.uk/~ps422/

(Take over the worst starting state.). For all $\varepsilon \in (0,1)$,

$$t_{\text{mix}} = \min\{t \geqslant 0 : d(t) \leq \varepsilon\}.$$

Define $t_{\text{mix}}\left(\frac{1}{4}\right)$. X is called **reversible** if from the stationary distribution, running the Markov chain forwards or backwards in time is indistinguishable: for all x, y,

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

I'll mostly talk about the reversible case.

We always consider the lazy verion of the chain. The lazy version of X: either stay with probability $\frac{1}{2}$, or choose a state with respect to the transition matrix. So $P_L = \frac{P+I}{2}$.

Chapter 6

Optimal transport and its application in ML

Topics and Reference:

- Wasserstein barycenter, Fréchet means, empirical and population, Fast computation of Fréchet means. [27] [1]
- Statistical estimation of Monge map, consistency and asymptotic. [14][2]
- Knothe-Rosenblatt transport and trangular maps, Estimation and smoothness class. [31][3],[21]
- Kullback-Leilbler estimatioin and normalizing flows, Asymptotic consistency. [17], [18]
- Sinkhorn divergence, Sinkhorn algorithm, asymptotic convergence. [6], [19]
- Deep learning as measure optimization, Global convergence of measure optimization. [5]
- Asymptotic convergence of gradient flows, Particle limits and time limits. [12]

6.1 Optimal transport

6.1.1 The existence of solution of (KP)

Monge problem(MP). Given two probability measure $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and a cost function $c : X \times Y \to [0, \infty]$, solve

$$\inf \left\{ M(T) := \int c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\} \tag{6.1}$$

where $T_{\#}\mu(A) := \mu(T^{-1}(A))$.

Kantorovich problem(KP). Given $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, and $c : X \times Y \to \mathbb{R}_+$, we consider the problem

$$\inf \left\{ K(\gamma) := \int_{X \times Y} c d\gamma : \gamma \in \Gamma(\mu, \nu) \right\}, \tag{6.2}$$

where $\Gamma(\mu, \nu)$ is the set of the so-called transport plans, i.e.,

$$\Gamma(\mu, \nu) = \{ \gamma \in \mathcal{P}(X \times Y) : (\pi_x)_{\#}\gamma = \mu, (\pi_y)_{\#}\gamma = \nu \},$$

where π_x and π_y are the two projections of $X \times Y$ onto X and Y, and $(\pi_x)_{\#}\gamma(A) := \gamma(\pi_x^{-1}(A))$ is called the

[1] https://arxiv.org/pdf/1806.05500

[2] https://arxiv.org/pdf/1905.05828

[3] https://www.jmlr.org/papers/volume19/17-747/17-747.pdf

pushforward of γ through π_x.

Remark 6.1.1. Should γ be of the form $(id, T)_{\#}\mu$ for a measurable map $T : X \to Y$, the map T would be called the optimal transport map from μ to ν. It is clear that $(id, T)_{\#}\mu \in \Gamma(\mu, \nu)$ iff T pushes μ onto ν.

$$\int_X c(x, T(x))d\mu(x) = \int_{X\times Y} c(x, y)d\gamma(x, y).$$

We will use the "continuity-compactness argument" in [29] to show that a minimum does exist. We use the Weierstrass criterion for the existence of minimizers.

Definition 6.1.2. (lower semi-continuous) On a metric space X, a function $f : X \to \mathbb{R} \cup \{\infty\}$ is said to be lower semi-continuous if for every sequence $x_n \to x$ we have

$$f(x) \leqslant \liminf_n f(x_n).$$

We often use the converging in norm for a sequence $\{x_n\}$ in a normed vector spaces. However, in some cases it is useful to work in topologies on vector spaces that are weaker than a norm topology. One reason for this is that many important modes of convergence are not captured by a norm topology. Another reason (of particular importance in PDE) is that the norm topology on infinite-dimensional spaces is so strong that very few sets are compact or pre-compact in these topologies, making it difficult to apply compactness methods in these topologies. Instead, one often first works in a weaker topology, in which compactness is easier to establish, and then somehow upgrades any weakly convergent sequences obtained via compactness to stronger modes of convergence.

Two basic weak topologies for this purpose are the weak topology on a normed vector space X, and the weak* topology on a dual vector space X'. Compactness in the latter topology is usually obtained from the **Banach-Alaoglu theorem** (and its sequential counterpart), which will be a quick consequence of the Tychonoff theorem (and its sequential counterpart) from the previous lecture.

Remark 6.1.3. For more definition of the strong and weak topologies, see 245B, Notes 11: The strong and weak topologies.

Theorem 6.1.4. (Weierstrass) If $f : X \to \mathbb{R} \cup \{\infty\}$ is lower semi-continuous and X is compact, then there exists $\bar{x} \in X$ such that $f(\bar{x}) = \min\{f(x) : x \in X\}$.

Theorem 6.1.5. (Banach-Alaoglu theorem) The closed unit ball of the dual space of a Banach space is compact in the weak-* topology.

Proof.

□

Theorem 6.1.6. (Sequential Banach-Alaoglu theorem) If $\mathcal{X}$ is separable and ξ_n is a bounded sequence in $\mathcal{X}'$, then there exists a subsequence ξ_{n_k} weakly converging to some $\xi \in \mathcal{X}'$.

We donate by $\mathcal{M}(X)$ the set of finite signed measures on X.

Theorem 6.1.7. Suppose that X is a separable and locally compact metric sapce. Let $\mathcal{X} = C_0(X)$ be the space of continuous function on X vanishing at infinity. Then every element of $\mathcal{X}'$ is represented in a unique way as an element of $\mathcal{M}(X)$: for all $\xi \in \mathcal{X}'$, there exists a unique $\lambda \in \mathcal{M}(X)$ such that

$$< \xi, \phi >= \int \phi d\lambda$$

for all $\phi \in \mathcal{X}$.

We will call it weak convergence and denote it through $\mu_n \rightharpoonup \mu$ iff $\phi \in C_b(X)$ we have

$$\int \phi d\mu_n \to \int \phi d\mu$$

where $C_b(X)$ is the space of bounded continuous functions on X. Note that $C_b(X) = C_0(X) = C(X)$ if X is compact, so the weak-* convergence is same as weak convergence.

Theorem 6.1.8. Let X and Y be compact metric spaces, $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, and $c : X \times Y \to \mathbb{R}$ a continuous function. Then (KP) admits a solution.

Proof. We just need to show that the set $\Gamma(\mu, \nu)$ is compact and $K(\gamma)$ is continuous and apply Weierstrass's theorem. Since $c \in C(X \times Y)$, this gives weak-* continuity of functional $\gamma \mapsto \int c d\gamma$ on $\mathcal{M}(X \times Y)$.

$\Gamma(\mu, \nu)$ is non-empty since $\mu \times \nu \in \Gamma(\mu, \nu)$. By the Banach-Alaoglu theorem, we need to show that $\Gamma(\mu, \nu)$ is weak-* closed subset of $B = \{\gamma : \|\gamma\| \leq 1\}$ which is weak-* compact. Note that a sequence of probability measures $\gamma_n \in \Gamma(\mu, \nu)$ are mass 1, so they are bounded in the dual of $C(X \times Y)$. Hence, this guarantees the existence of a subsequence $\gamma_{\gamma_k} \to \gamma$ converging to a probability γ, We need to check $\gamma \in \Gamma(\mu, \nu)$. Indeed, fix $\phi \in C(X)$,

$$\int_X \phi(x) d\mu(x) = \int_{X \times Y} \phi(x) d\gamma_{n_k}(x, y) \to \int_{X \times Y} \phi(x) d\gamma(x, y)$$

This shows that $(\pi_x)_{\#}\gamma = \mu$. The same may be done for π_y. □

Theorem 6.1.9 (Minimax Theory)**.** We have

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \sup F(\gamma, (\phi, \psi)) = \sup \inf_{\gamma \in \mathcal{M}_+(X \times Y)} F(\gamma, (\phi, \psi))$$

Remark 6.1.10. See KAKUTANI'S fixed point theorem and the minimax theorem in game theory.

6.1.2 Probabilistic interpretation

See here. Convex function.

6.1.3 Duality

The problem (KP) is a linear optimization under convex constraints, given by linear qualities or inequalities. Hence, an important tool will be duality theory, which is typically used for convex problems. The Kantorovich motivation for (KP):

If μ is a distribution of mines producing iron. Let $f(x)$ be production capacity of mine at x which $d\mu(x) = f(x)$. And ν is a distribution of factories consuming iron that $d\nu(y) = g(y)$ and $g(y)$ is the consumption requirements of factory at location y. Let $c(x, y)$ be the transport cost per mass. Then (KP) asks which mine should supply which factory to minize overall transport cost. Now, an independent company might offer to buy iron from the mines and sell iron to the factories say they offer to pay the mine at x, $\phi(x)$ dollars per mass and sell iron to the factory at y for $\psi(y)$ dollars per mass. Their profits are

$$\int_Y \psi(y) d\nu(y) + \int_X \phi(x) d\mu(x).$$

But if for (x, y) we have $\phi(x) + \psi(y) > c(x, y)$, the mine at x could sell and deliver directly to the factory at y for price $\psi(y) - \varepsilon$. The mine's profit would be

$$\psi(y) - \varepsilon - c(x, y) > -\phi(x)$$

for all $\varepsilon > 0$ samll enough.

So, to entice all mines and factories to take the deal, the company must ensure

$$\phi(x) + \psi(y) \le c(x, y)\, everywhere$$

We write the dual optimization problem:

Dual Problem(DP). Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and the cost function $c : X \times Y \to [0, \infty)$. We consider the problem

$$\max\left\{\int_Y \psi(y) d\nu(y) + \int_X \phi(x) d\mu(x) : \phi \in C_b(X), \psi \in C_b(Y) : \phi(x) + \psi(y) \le c(x, y)\right\} \tag{6.3}$$

Theorem 6.1.11. $\sup(DP) = \min(KP)$

6.1.4 The existence of solution to (DP)

Definition 6.1.12 (c-transform, c-concave)**.** Given a function $\phi \in C(X)$, define $\phi^c : Y \to \mathbb{R}$ by

$$\phi^c(y) := \min_{x\in X}(c(x, y) - \phi(x)).$$

Similary, for $\psi \in C(Y)$, define $\psi^c : X \to \mathbb{R}$ by

$$\psi^c(x) := \min_{y\in Y}(c(x, y) - \psi(y)).$$

We say ϕ is **c-concave** if $\phi = \psi^c$ for some ψ.

Lemma 6.1.13. If $(\phi, \psi) \in C(X) \times C(Y)$ and $\phi(x) + \psi(y) \le c(x, y)$. Then (ϕ^{cc}, ϕ^c) satisfies

$$\phi^c(y) \geqslant \psi(y), \phi^{cc}(x) \geqslant \phi^c(x)$$

for all (x, y). Therefore,

$$\int_X \phi d\mu + \int_Y \psi d\nu \le \int_X \phi^{cc} d\mu + \int_Y \phi^c d\nu.$$

Definition 6.1.14 (k-Lip)**.** We say the cost function c is $k-$Lipschitz in X, if

$$c(x_1, y) - c(x_0, y) \le k|x_1 - x_0|$$

for all $x0, x_1, y$.

Theorem 6.1.15. Assume X, Y are compact and c is $k-$Lip. Then there exists a maxmizer (ϕ, ψ) in (DP). It has the form that ϕ and ψ are $c-$concave. In particular

$$\max(DP) = \max_{\phi\in c-concave(X)} \int_X \phi d\mu + \int_Y \phi^c d\nu.$$

Proof.

□

6.1.5 c-cyclical monotonicity

Definition 6.1.16 ($c-$monotone of order 2)**.** We say that a set $S \subset X \times Y$ is c-monotone of order 2, for a given cost function $c : X \times Y \to \mathbb{R}$, if for all $(x_0, y_0), (x_1, y_1) \in S$, we have

$$c(x_0, y_0) + c(x_1, y_1) \leq c(x_0, y_1) + c(x_1, y_0)$$

Definition 6.1.17 (c-cyclical monotone)**.** A subset $S \subset X \times Y$ is **c-cyclical monotone**(c-CM) if for all integers k, all permutation σ, and finite family of points $(x_1, y_1), \dots, (x_N, y_N) \in S$, we have

$$\sum_{i=1}^{N} c(x_i, y_i) \leqslant \sum_{i=1}^{N-1} c(x_i, y_{i+1}) + c(x_N, y_1). \tag{6.4}$$

Definition 6.1.18. On a separable metric space X, the support of a measure γ is defined as the smallest closed set on which γ is concentrated, i.e.,

$$spt(\gamma) := \bigcap \{A : A \text{ is closed and } \gamma(X \setminus A) = 0\}$$

Theorem 6.1.19. Assume that X, Y are compact and c is $k-$Lip. If $\gamma \in \Gamma(\mu, \nu)$ is optimal in (KP). Then $spt(\gamma)$ is $c-$cyclically monotone.

Proof. Let (ϕ, ψ) be a continuous solution to dual problem (DP). We have

$$\phi(x) + \psi(y) \leq c(x, y)$$

and $\phi(x) + \psi(y) = c(x, y)$, $\gamma-$a.e. The set satisfying "=" is closed. So it contains the support of γ.

Let $(x_1, y_1), \dots, (x_N, y_N) \in spt(\gamma)$, we have

$$\begin{aligned}\sum_{i=1}^{N} c(x_i, y_i) &= \sum_{i=1}^{N} (\phi(x_i) + \psi(y_i)) \\ &= \sum_{i=1}^{N-1} (\phi(x_i) + \psi(y_{i+1})) + \phi(x_N) + \psi(y_1) \\ &\leq \sum_{i=1}^{N-1} c(x_i, y_{i+1}) + c(x_N, y_1).\end{aligned}$$

□

Remark 6.1.20. The converse is also true, see [29, Theorem 1.49, p39]

If we admit the duality result $\min(KP) = \max(DP)$ as Theorem 6.1.15(the following theorem), then we have

$$\min(KP) = \max_{\phi \in c-concave(X)} \int_X \phi d\mu + \int_Y \phi^c d\nu.$$

Theorem 6.1.21. Suppose that X and Y are Polish spaces[a] and that $c : X \times Y \to \mathbb{R}$ is uniformly continuous and bounded. Then the problem (DP) admits a solution (ϕ, ϕ^C) and we have $\min(KP) = \max(DP)$

[a] A Polish space is a separable completely metrizable topological space; that is, a space homeomorphic to a complete metric space that has a countable dense subset.

Proof. See [29, Theorem 1.39]. □

6.1.6 1-dimensional solutions

Definition 6.1.22. For $X \subset \mathbb{R}^d$, we say $c : X \times X \to \mathbb{R}$ satisfies the twist condition if c is differentiable w.r.t. x at every point and the map $y \mapsto \nabla_x c(x_0, y)$ is injective for every x_0

This condition is also known in economics as Spence-Mirrlees condition.

Examples of costs satisfying the twist condition:

Example 6.1.23. (i) If n=1,

Theorem 6.1.24. Assume that $X, Y \subset \mathbb{R}$ are compact intervals and cost function c satisfies $\frac{\partial^2 c}{\partial x \partial y} < 0$ throughout $X \times Y$. Then $spt(\gamma)$ is monotone increasing for any optimal γ (i.e. for any $(x_0, y_0), (x_1, y_1) \in spt(\gamma)$, $(x_1 - x_0)(y_1 - y_0) \geqslant 0$.

Proof.

□

$\frac{\partial^2 c}{\partial x \partial y} < 0$ is often called submodularity.

Remark 6.1.25. (1) In 1 dimension, the sets which are $c-$monotone of order 2 are $c-$cyclically monotone. But this is not true in $d \geqslant 2$ dimension. Here is an example:

6.1.7 Uniqueness of Monge solutions

Theorem 6.1.26. Assume that X, Y are compact. If $c \in C(X \times Y)$ satisfies the twist condition and μ is absolutely continuous w.r.t. Lebesgue measure. Then any minimizer γ in (KP) is of Monge form.

Proof.

□

Corollary 6.1.27. Under the assumptions of the Theorem 6.1.26, the solutions γ and T to (KP) and (MP) are unique.

6.1.8 Wasserstein's Distance

Given two probability measures $\mu, \nu \in \mathcal{P}(X)$ and $X \subset \mathbb{R}^n$ be compact, define

$$W_2(\mu, \nu) := \left(\min_{\gamma \in \Gamma} \int_{X \times X} |x - y|^2 d\gamma(x, y) \right)^{1/2} \tag{6.5}$$

Theorem 6.1.28. W_2 defined in 6.5 is a metric on $\mathcal{P}(X)$.

See arXiv:1412.7726

Chapter 7

Statistical mechanics of mean-field disordered systems: a PDE approach

These are notes from the CRM-PIMS Summer School in Probability 2021. Webpage for CRM-PIMS Summer School: https://secure.math.ubc.ca/Links/ssprob21/. The course is taught by Jean-Christophe Mourrat: A PDE approach to mean-field disordered systems and course webpage is http://perso.ens-lyon.fr/jean-christophe.mourrat/Montreal.html.

7.1 Large Deviations and Convex Analysis

Scribe: Tim Banova

We give a primer on some fundamental results from the theory of *large deviations* and *convex analysis*

7.1.1 Motivating example: LDP for IID Bernoulli random variables

Let $(X_n)_{n\in\mathbb{N}}$ denote a collection of independent Bernoulli random variables with parameter $p \in [0,1]$; that is,

$$\mathbb{P}(X_n = 1) = p = 1 - \mathbb{P}(X_n = 0), \qquad n \in \mathbb{N}. \tag{7.1}$$

Define the (rescaled) partial sum via

$$S_N := \frac{1}{N}\sum_{n=1}^{N} X_n, \qquad N \in \mathbb{N}. \tag{7.2}$$

We know by the large large number, the sum S_N converges to its mean p as N goes to infinity. Now, one may ask what is the speed that the probability of S_N converging to a number x other than p tends to zero? This questions is not answered by the central limit theorem because it only tells us typical deviations from p, which are of order $\sqrt{N}$, while the questions is about large deviations, which are of order N.

For this example, we can calculate the probability exactly by the following formula

$$\mathbb{P}(S_N = k/N) = \binom{N}{k} p^k (1-p)^{N-k}, \quad k \in \{0, \dots, N\},\ N \in \mathbb{N}. \tag{7.3}$$

Letting $x := k/N \in (0,1)$, we have that

$$\begin{aligned}\binom{N}{k} &= \exp\Big(N\log(N/e) - Nx\log(Nx/e) + O(\log N) - N(1-x)\log(N(1-x)/e)\Big) \\ &= \exp\Big(-Nx\log x - N(1-x)\log(1-x) + O(\log N)\Big). \qquad (7.4)\end{aligned}$$

Above, we have used *Stirling's formula*, which is

$$n! = \exp\Big(n\log(n/e) + O(\log n)\Big), \quad n \to \infty. \qquad (7.5)$$

Thus, letting

$$I(x) := x\log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}, \quad x \in (0,1), \qquad (7.6)$$

we have that for every $\delta > 0$ and $x \in [\delta, 1-\delta]$ of the form $x = k/N$ for some $k \in \mathbb{Z}$,

$$\mathbb{P}(S_N = x) = \exp\Big(-NI(x) + O(\log N)\Big), \quad N \to \infty. \qquad (7.7)$$

Moreover, note that for any $x \in (p, 1)$, as $N \to \infty$

$$\begin{aligned}\mathbb{P}(S_N \geqslant x) \geqslant \mathbb{P}\Big(S_N = \frac{\lceil Nx\rceil}{N}\Big) &\geqslant \exp\Big(-NI(\lceil Nx\rceil/N) + O(\log N)\Big) \\ &\exp\Big(-NI(x) + O(\log N)\Big) \qquad (7.8)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(S_N \geqslant x) \leqslant \sum_{k=\lfloor Nx\rfloor}^{N} \mathbb{P}(S_n = k/N) &\leqslant (N+1)\mathbb{P}(S_n = \lfloor Nx\rfloor/N) \\ &\leqslant \exp\Big(-NI(x) + O(\log N)\Big) \qquad (7.9)\end{aligned}$$

so that

$$\mathbb{P}(S_N \geqslant x) = \exp\Big(-NI(x) + O(\log N)\Big), \quad N \to \infty \qquad (7.10)$$

Analogously, for every $x \in (0, p)$,

$$\mathbb{P}(S_N \leqslant x) = \exp\Big(-NI(x) + O(\log N)\Big), \quad N \to \infty \qquad (7.11)$$

This leads to an important definition.

Definition 7.1.1 (Large deviations principle for sequences of random variables)**.** For $N \in \mathbb{N}$, let $S_N : \Omega_N \to \mathbb{R}$ be a random variable defined on $(\Omega_N, \mathcal{F}_N, \mathbb{P}_N)$. We say that $(S_N)_{N\in\mathbb{N}}$ *satisfies a large deviations principle (LDP)* with *speed* N and *rate function* I if for all Borel-measurable $A \subseteq \mathbb{R}$,

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}_N(S_N \in A) \leqslant -\inf_{x\in\bar{A}} I(x), \quad \liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}_N(S_N \in A) \geqslant -\inf_{x\in A^\circ} I(x), \qquad (7.12)$$

where $\bar{A}$ and A° denote the closure and interior of A.

Informally, we write this as

$$\mathbb{P}(S_N \simeq x) \simeq \exp(-NI(x)) \qquad (7.13)$$

Exercise 9. Apply (7.10) and (7.11) to show that $(S_N)_{N\in\mathbb{N}}$ defined earlier satisfies a LDP with speed N and rate function I given in (7.6).

7.1.2 Cramér's Theorem: a LDP for i.i.d random variables

In the previous section, we derived (from first principles) a LDP for i.i.d Bernoulli random variables with parameter $p \in [0,1]$. Naturally, this leads us to the following question: when do random variables with an *arbitrary* distribution satisfy a LDP?

Suppose that $(X_n)_{n\in\mathbb{N}}$ are a family of i.i.d random variables with $\mathbb{E}[X_1] = 0$ and $x, \lambda \geqslant 0$.

Then, by Markov's inequality,

$$\mathbb{P}(S_N \geqslant x) = \mathbb{P}(e^{\lambda N S_n} \geqslant e^{\lambda N x}) \leqslant e^{-\lambda N x}\mathbb{E}[e^{\lambda N S_N}], \tag{7.14}$$

where by independence of $(X_n)_{n\in\mathbb{N}}$

$$\mathbb{E}[e^{\lambda N S_N}] = \mathbb{E}[e^{\lambda X_1}]^N. \tag{7.15}$$

Thus, letting

$$\psi(\lambda) := \log \mathbb{E}[\exp(\lambda X_1)] \tag{7.16}$$

denote the *cumulant generating function* of X_1, we have

$$\mathbb{P}(S_N \geqslant x) \leqslant \exp\Big(- N \sup_{\lambda\in\mathbb{R}_+} \big(\lambda x - \psi(\lambda)\big)\Big) \tag{7.17}$$

Define the *Fenchel-Legendre transform* of ψ, $\psi^* : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ via

$$\psi^*(x) := \sup_{\lambda\in\mathbb{R}}(\lambda x - \psi(\lambda)) \tag{7.18}$$

Theorem 7.1.2 (Cramér's Theorem : LDP for IID)**.** Let $(X_n)_{n\in\mathbb{N}}$ be i.i.d random vairables such that for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda X_1}] < \infty$, with cumulant generating function ψ and Legendre transform ψ^*.
Then, $(S_N)_{N\in\mathbb{N}}$, where $S_N := \frac{1}{N}\sum_{k=1}^N X_k$ satisfies a large deviations principle with speed N and rate function ψ^*.

Proof. Apply Theorem 7.1.3 below or see Section 2.2.2 of [7]. □

Exercise 10. Show that if X is a Bernoulli random variable with parameter $p \in [0,1]$, then $\psi^* = I$, where I is given in (7.6).

In fact, we can derive LDPs for arbitrary sequences of random variables (under some fairly strong conditions).

Theorem 7.1.3 (General LDP on $\mathbb{R}$)**.** For $N \in \mathbb{N}$, let $S_N : \Omega_N \to \mathbb{R}$ be a random variable defined on $(\Omega_N, \mathcal{F}_N, \mathbb{P}_N)$ and define

$$\psi_N(\lambda) := \frac{1}{N}\log \mathbb{E}_N[\exp(\lambda N S_N)]. \tag{7.19}$$

Suppose that there exists a $\psi \in C^1(\mathbb{R})$ such that $\psi_N \to \psi$ pointwise. Define

$$\psi^*(x) := \sup_{\lambda\in\mathbb{R}}(\lambda x - \psi(\lambda)). \tag{7.20}$$

Then, S_N satisfies a large deviation principle with speed N and rate function ψ^*.

Remark 7.1.4. This is a special case of the Gärtner–Ellis theorem [7, Theorem 2.3.6].

Proof. We claim that for every $x \in \mathbb{R}$,

$$\limsup_{N\to\infty} \frac{1}{N}\log \mathbb{P}\left(S_N \geqslant x\right) \leq - \inf_{z\geqslant x} \psi^*(x), \tag{7.21}$$

$$\lim_{\varepsilon\to 0}\liminf_{N\to\infty} \frac{1}{N}\mathbb{P}\left(S_N \in [x-\varepsilon, x+\varepsilon]\right) \geqslant -\psi^*(x). \tag{7.22}$$

Denote $m = \psi'(0)$ (Recall that ψ is of C^1). By convexity, $\psi(\lambda) \geqslant m\lambda$. Thus, for all $x \geqslant m$,

$$\sup_{\lambda\leqslant 0}\left(\lambda x - \psi(\lambda)\right)$$

and ψ^* is increasing on $[0, +\infty)$. Moreover, $\psi^*(x) = 0$. Thus, it suffices to prove (7.21) via showing that for all $x \geqslant m$,

$$\limsup_{N\to\infty} \frac{1}{N}\log \mathbb{P}\left(S_N \geqslant x\right) \leq - \inf_{z\geqslant x} \psi^*(x) = \sup_{\lambda\leqslant 0}\left(\lambda x - \psi(\lambda)\right). \tag{7.23}$$

We have

$$\mathbb{P}(S_N \geqslant x) \leq \exp(-\lambda N x) \cdot \mathbb{E}\left[\exp(\lambda N S_N)\right] \leq \exp\left(\lambda x - \psi_N(\lambda)\right). \tag{7.24}$$

Also,

$$\frac{1}{N} \log \exp\left(\lambda x - \psi_N(\lambda)\right) = \frac{\lambda x}{N} - \frac{1}{N}\psi_N^*(\lambda) \to -\psi^*(x) \tag{7.25}$$

as $N \to \infty$, so (7.24) follows from (7.24) and (7.25).

It remains to prove (7.21). TODO: to be continued... □

Example 7.1.5 (Necessity of $\psi \in C^1(\mathbb{R})$). Consider the sequence of random variables $(S_N)_{N \geqslant 1}$ where

$$\mathbb{P}(S_N = 1) = 1/2 = \mathbb{P}(S_N = -1).$$

We have

$$\psi_N(\lambda) = \frac{1}{N} \log\left(\frac{\exp(\lambda N)}{2} + \frac{\exp(-\lambda N)}{2}\right).$$

Since

$$\psi_N(\lambda) = \begin{cases} \lambda + \frac{1}{N}(\log\left(1 + e^{-2\lambda N}\right) - \log 2) \to \lambda & \text{when } \lambda > 0 \\ \\ -\lambda + \frac{1}{N}(\log\left(1 + e^{2\lambda N}\right) - \log 2) \to -\lambda & \text{when } \lambda \leq 0, \end{cases}$$

we derive that

$$\psi(\lambda) = \lim_{N \to \infty} \psi_N(\lambda) = |\lambda|.$$

Note that

$$\psi^*(x) = \sup_{\lambda \in \mathbb{R}} (\lambda x - |\lambda|) = \begin{cases} 0 & x \in [-1, 1] \\ \infty & \text{otherwise} \end{cases}$$

TODO: to be continued...

7.1.3 Some convex analysis

We now develop some of the essentials of the analysis of *convex* functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ on $\mathbb{R}^d$, for $d \geqslant 1$.

Definition 7.1.6 (Convex functions). A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is *convex* if for all $x, y \in \mathbb{R}^d$ and $\alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y). \tag{7.26}$$

Exercise 11. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}^d$ an $\mathbb{R}^d$-random variable (ie. with respect to the Borel σ-algebra on $\mathbb{R}^d$). Show that the map

$$\mathbb{R}^d \to \mathbb{R} \cup \{\infty\}, \quad \lambda \mapsto \log \mathbb{E}[e^{\lambda \cdot X}] \tag{7.27}$$

is convex.

Solution to Exercise 11. Let $x, y \in \mathbb{R}^d$ and $\alpha \in [0, 1]$. Note that by Hölder's inequality, with conjugate exponents $p = \alpha^{-1}$ and $q = (1-\alpha)^{-1}$, we have

$$\mathbb{E}[e^{\alpha x \cdot X} e^{(1-\alpha) y \cdot X}] \leqslant \mathbb{E}[e^{x \cdot X}]^{\alpha} \mathbb{E}[e^{y \cdot X}]^{1-\alpha}. \tag{7.28}$$

Taking logarithms gives the desired inequality.

Definition 7.1.7 (Lower semi-continuous functions). A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is *lower semi-continuous* *(l.s.c.)* if for every sequence $x_n \to x \in \mathbb{R}^d$,

$$f(x) \leqslant \liminf_{n\to\infty} f(x_n). \tag{7.29}$$

Remark 7.1.8. For $\alpha \in I$, let $f_\alpha : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. If for all $\alpha \in I$, f_α is convex (resp., l.s.c.) then $\sup_{\alpha\in I} f_\alpha$ is convex (resp., l.s.c).

Definition 7.1.9 (Convex duals)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The *convex dual (or convex conjugate) of* f is the function $f^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ defined via

$$f^*(\lambda) := \sup_{x\in\mathbb{R}^d} (\lambda x - f(x)), \qquad \lambda \in \mathbb{R} \tag{7.30}$$

Moreover, define the *bidual (or biconjugate)* $f^{**} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ via

$$f^{**}(x) := \sup_{\lambda\in\mathbb{R}} (\lambda x - f^*(\lambda)) \tag{7.31}$$

Note that for any $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^d$,

$$f^*(\lambda) \geqslant \lambda x - f(x), \quad f(x) \geqslant \lambda x - f^*(\lambda) \tag{7.32}$$

so that $f(x) \geqslant f^{**}(x)$. Moreover, the case that $f(x) \leqslant f^{**}(x)$ for all $x \in \mathbb{R}^d$ characterises functions which are convex and l.s.c., by the following theorem.

Theorem 7.1.10 (Fenchel-Moreau)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex and l.s.c if and only if $f^{**} = f$.

7.2 Curie-Weiss Models

Scribe: Tim Banova

7.2.1 Gibbs measures

We give a probabilisitic description of an *isolated system* in which the *energy is fixed.*

In particular, suppose for some $N, K \in \mathbb{N}$ we are given a system of N units, which can each be in one of K states $[K] = \{1, 2, \dots, K\}$.

For each $k \in [K]$, let N_k denote the number of units in state k. To each each state $k \in [K]$, we associate an *energy* $e_k \in \mathbb{R}$, where $e : [K] \to \mathbb{R}$ is the *Hamiltonian* of the system.

Throughout, we assume that there exists $\bar{e} \in \mathbb{R}$ such that

$$\sum_{k\in[K]} N_k e_k = N\bar{e}. \tag{7.33}$$

According to Boltzmann's hypothesis, we postulate that the system is distributed according to the uniform measure.

A natural question arises; for each $k \in [K]$, what is the likelihood of finding a given unit N_k in state k? In other words, what does N_k/N look like?

By a clear counting argument, the number of configurations consistent with $(N_1, \dots, N_k)$ is

$$\binom{N}{N_1}\binom{N-N_1}{N_2}\cdots\binom{N-\sum_{k\in[K-1]} N_k}{N_K} = \frac{N!}{N_1!\cdots N_K!} \tag{7.34}$$

Supposing that $N_k/N \simeq p_k$ where $\sum_{k\in[K]} p_k = 1$, Stirling's formula implies that

$$\log\left(\frac{N!}{N_1!\cdots N_K!}\right) = NS(p) + O(\log N), \tag{7.35}$$

where

$$S(p) := -\sum_{k\in[K]} p_k \log p_k = \sum_{k\in[K]} p_k \log \frac{1}{p_k} \tag{7.36}$$

is the *(Shannon) entropy* of p.

Thus, the system will tend to concentrate on p which maximise the Shannon entropy S, subject to the constraints

$$\sum_{k\in[K]} p_k e_k = \bar{e}, \quad \sum_{k\in[K]} p_k = 1. \tag{7.37}$$

The maximizer p is such that there exists α, β satisfying

$$\partial_{p_k} S(p) = \alpha + \beta e_k, \quad k \in [K], \tag{7.38}$$

or

$$p_k = \frac{1}{Z} \exp(-\beta e_k) \tag{7.39}$$

where

$$Z := \sum_{k\in[K]} \exp(-\beta e_k) \tag{7.40}$$

is called the *partition function*, and β is the *inverse temperature*. Such a $(p_k)_{k\in[K]}$ is called the *(canonical) Gibbs measure* associated with $(e_k)_{k\in[K]}$.

For motivation on the physical terminology, as well as an account of the *macroscopic* thermodynamical theory which this *microscopic* approach describes, refer to [10, Chapter 1].

7.2.2 A brief look at the Ising model

Before introducing the main models of interest, we introducing

7.2.3 The Curie-Weiss model

7.3 Curie-Weiss Model and Generalized Curie-Weiss Model

Scribe: Haotian Gu

7.3.1 Curie-Weiss Model

Let us first recall the definition of the Curie-Weiss model on $\Omega = \{1, \cdots, N\}$, which is obtained by trivializing the geometry of the Ising model. The energy associated with each configuration $\sigma \in \{\pm 1\}^N$, at inverse temperature t and with an external magnetic field h, is

$$H_N(t, h, \sigma) := \frac{t}{N} \sum_{i,j=1}^{N} \sigma_i \sigma_j + h \sum_{i=1}^{N} \sigma_i.$$

Note that here the energy is the opposite of that in [10], which means with more spins align the energy is larger.

The free energy of the system is then defined as

$$F_N(t, h) = \frac{1}{N} \log 2^{-N} \sum_{\sigma\in\{\pm 1\}^N} \exp(H_N(t, h, \sigma)).$$

The expectation for function $f(\sigma)$ with respect to the Gibbs measure is denoted as

$$\langle f\rangle := Z_N^{-1}(t,h)\sum_{\sigma\in\{\pm1\}^N} f(\sigma)\exp(H_N(t,h,\sigma)),$$

where $Z_N(t,h)$ is the partition function of the Gibbs measure:

$$Z_N(t,h)=\sum_{\sigma\in\{\pm1\}^N}\exp(H_N(t,h,\sigma)).$$

What we want to study here is the mean magnetization $m_N := \frac{1}{N}\sum\limits_{i=1}^N \sigma_i$. However, it will be easier to do it in another way: via studying the free energy, from which we can reproduce the information about magnetization.

By treating $2^{-N}\sum_\sigma \mathbf{1}(\frac{1}{N}\sum\limits_{i=1}^N\sigma_i\approx m)$ as the probability of some event under product Bernoulli measures, we know from large deviation result that

$$2^{-N}\sum_\sigma \mathbf{1}(\frac{1}{N}\sum_{i=1}^N\sigma_i\approx m)\approx\exp(-N\psi^*(m)),$$

where $\psi(\lambda)=\log\left(\frac{1}{2}\sum\limits_{\{\pm1\}}\exp(\lambda\sigma_1)\right)=\log\cosh(\lambda)$, and $\psi^*(m)=\sup\limits_\lambda(\lambda m-\psi(\lambda))=\frac{1+m}{2}\log(1+m)+\frac{1-m}{2}\log(1-m)$. Therefore,

$$\begin{aligned}2^{-N}\sum_\sigma\exp(H_N(t,h,\sigma)) &= 2^{-N}\sum_\sigma\exp\left(tN(\frac{1}{N}\sum\sigma_i)^2+hN(\frac{1}{N}\sum\sigma_i)\right)\\ &\approx\sum_m\left(2^{-N}\sum_\sigma\mathbf{1}(\frac{1}{N}\sum_{i=1}^N\sigma_i\approx m)\exp\left(tN(\frac{1}{N}\sum\sigma_i)^2+hN(\frac{1}{N}\sum\sigma_i)\right)\right)\\ &\approx\sum_m\exp\big(tNm^2+hNm\big)\exp(-N\psi^*(m))\\ &=\sum_m\exp\big(N(tm^2+hm-\psi^*(m))\big).\end{aligned}$$

When $N\to\infty$, the sum of m becomes integration over $[-1,1]$, and the maximum of the exponentials dominates in the limit:

Proposition 7.3.1. $\forall t\geqslant 0, h\in\mathbb{R}$,

$$\lim_{N\to\infty}F_N(t,h)=\sup_{m\in[-1,1]}(tm^2+hm-\psi^*(m)).$$

Let $f(t,h)$ be the limit function. Then if it is differentiable in h at some point, then the limit of partial derivatives of F_N w.r.t. h converges to the partial derivative of f as well.

Proposition 7.3.2. If f is differentiable in h at (t,h), then

$$\lim_{N\to\infty}\partial_h F_N(t,h)=\partial_h f(t,h).$$

Proof. Observe that F_N and f are both convex in h(the first obtained using Hölder's inequality, and the second is the supremum of affine functions). Therefore for any $h'\in\mathbb{R}$,

$$F_N(t,h+h')\geqslant F_N(t,h)+\partial_h F_N(t,h)h'.$$

Note that $\partial_h F_N(t,h)=\left\langle\frac{1}{N}\sum\limits_{i=1}^N\sigma_i\right\rangle$, which is bounded for all N, therefore has a convergent subsequence. Along this subsequence(or w.l.o.g. assume the original sequence converges) we have $\lim\limits_{N\to\infty}\partial_h F_N(t,h)=p$ for

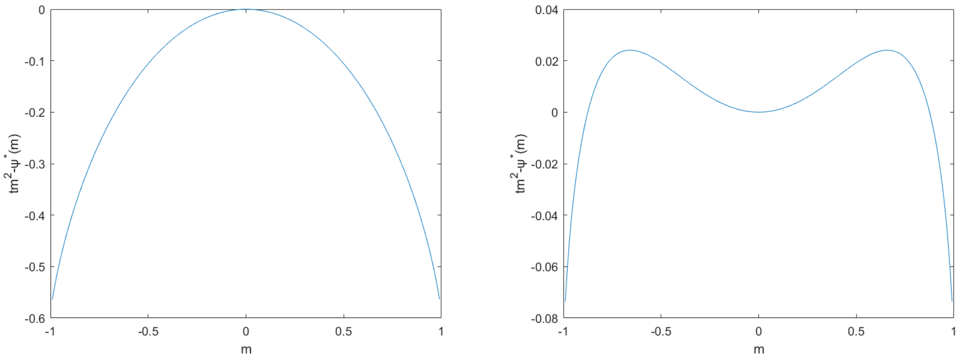


Figure 7.1: Graphs of $tm^2 - \psi^*(m)$ with $t = 0.1$(left) and $t = 0.6$(right).

some p. Hence

$$f(t, h + h') \geqslant f(t, h) + ph'.$$

Meanwhile from the Taylor expansion of f we know

$$f(t, h + h') = f(t, h) + \partial_h f(t, h)h' + o(h'),$$

so $p = \partial_h f(t, h)$ since the inequality holds for both positive and negative h'. □

Note that actually if a sequence of differentiable convex functions f_n converges pointwise to f(thus also convex) and f is differentiable at x, then $f_n'(x) \to f'(x)$.

Now we take a look at the case $h = 0$, which means there is no external magnetic field. Then

$$f(t) = \sup_{m \in [-1,1]} (tm^2 - \psi^*(m)),$$

where $\psi^*(m)$ is a function behaves like a parabola. See Figure 7.1 for the graphs of $tm^2 - \psi^*(m)$ for different t. So this competition between tm^2 and $\psi^*(m)$ leads to the phase transition in the limit behavior of the mean magnetization we are interested in.

Finally, by the so-called "envelope theorem", we know that if f has a unique optimizer $m_0(t, h)$(true for $h = 0$ and small t), the derivative can be taken into the supremum and we obtain $\partial_h f(t, h) = m_0(t, h)$. For more information on the envelope theorem, please refer to SPS3 Problem 1, where JCM has provided a thorough proof to the statement.

7.3.2 Generalized Curie-Weiss Model

The Curie-Weiss model can be generalized in two directions:

1. We use $\xi(\frac{1}{N}\sum_i \sigma_i)$ instead of $(\frac{1}{N}\sum_i \sigma_i)^2$ with given smooth ξ to be the interaction energy, and
2. we use generic measure P_N on $\mathbb{R}^N$ instead of product Bernoulli measure on the space of configurations. In this case, we assume $|\sigma| \leqslant \sqrt{N}$, $P_N(\sigma)$-a.s.

Under these generalizations, the free energy is now

$$F_N(t, h) := \frac{1}{N} \log \int \exp\left(tN\xi(\frac{1}{N}\sum_i \sigma_i) + h\sum_i \sigma_i\right) dP_N(\sigma).$$

We now have the following result.

Theorem 7.3.3. Assume $\forall h \in \mathbb{R}$, $\lim_{N\to\infty} F_N(0, h) = \psi(h)$ and $\psi \in C^1$. Then

$$\lim_{N\to\infty} F_N(t, h) = f(t, h) := \sup_{m \in \mathbb{R}} (t\xi(m) + hm - \psi^*(m)),$$

where $\psi^*(m) = \sup_{h\in\mathbb{R}}(hm - \psi(h))$.

Proof. Assume $|\frac{1}{N}\sum_i \sigma_i| \leqslant (\frac{1}{N}\sum_i \sigma_i^2)^{1/2} \leqslant 1$, P_N-a.s. Then $|\partial_h F_N| = |\left\langle \frac{1}{N}\sum_i \sigma_i \right\rangle| \leqslant 1$, which implies F_N, and therefore ψ, are all 1-Lip. Then we can apply the LDP result(with condition assumed in the context, i.e. $t = 0$ case) to get a lower bound as follows.

Let m_0 be one of the optimization point. Denote U_ϵ to be $(m_0 - \epsilon, m_0 + \epsilon)$. Then

$$F_N(t,h) \geqslant \frac{1}{N}\log\int \exp\left(tN\xi + h\sum_i \sigma_i\right)\mathbf{1}(\frac{1}{N}\sum_i \sigma_i \in U_\epsilon)dP_N(\sigma),$$

which implies (with standard technique)

$$\liminf_{N\to\infty} F_N \geqslant \inf_{m\in U_\epsilon}(t\xi(m) + hm) - \psi^*(m_0).$$

By sending ϵ to 0 we have

$$\liminf_{N\to\infty} F_N(t,h) \geqslant \sup_m(t\xi(m) + hm - \psi^*(m)).$$

On the other hand,

$$F_N(t,h) \leqslant \frac{1}{N}\log\sum_{k=-K}^{K}\int \exp\left(tN\xi + h\sum_i \sigma_i\right)\mathbf{1}(\frac{1}{N}\sum_i \sigma_i \in [\frac{k-1}{K}, \frac{k}{K}])dP_N(\sigma),$$

which by same method gives

$$\limsup_{N\to\infty} F_N(t,h) \leqslant \max_{-K\leqslant k\leqslant K}(\sup_{m\in[\frac{k-1}{K},\frac{k}{K}]}(t\xi(m) + hm) - \inf_{m\in[\frac{k-1}{K},\frac{k}{K}]}\psi^*(m)).$$

Sending K to infinity completes the proof. □

7.4 Hamilton–Jacobi Equations

Scribe: Fu-Hsuan

Recall that in Chapter 7.3, we studied the limiting free energy via Large deviation methods. This approach, however, seems to be inapplicable for the models coming from statistical inference or the spin glass models, which we are interested in. In this chapter, we study the Curie–Weiss model by a Hamilton–Jacobi equation arised from differentiating the finite free energies and then taking limits. This equation doesn't have a unique solution. In fact, we can construct infinitely many solutions that satisfy this solutions. Fortunately, by restricting ourselves to the class of viscosity solutions, the solution is unique.

7.4.1 Revisit of the Curie–Weiss Model

Recall that in Section 7.3.1, the free energy of the centered Curie–Weiss model was defined as

$$F_N(t,h) = \frac{1}{N}\log 2^{-N}\sum_{\sigma\in\{\pm1\}^N}\exp\left(\frac{t}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right). \tag{7.41}$$

For any function $f : \{\pm1\}^N \to \mathbb{R}$, define

$$\langle f(\sigma)\rangle = \frac{\sum_{\sigma\in\{\pm1\}^N} f(\sigma)\exp\left(\frac{t}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right)}{\sum_{\sigma\in\{\pm1\}^N}\exp\left(\frac{t}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + h\sum_{i=1}^{N}\sigma_i\right)}$$

Note that

$$\partial_t F_N = \frac{1}{N}\left\langle \frac{1}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j \right\rangle = \left\langle \left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i\right)^2 \right\rangle \tag{7.42}$$

$$\partial_h F_N = \left\langle \frac{1}{N}\sum_{i=1}^{N}\sigma_i \right\rangle. \tag{7.43}$$

Thus,

$$\begin{aligned}\partial_t F_N - (\partial_h F_N)^2 &= \left\langle \left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i\right)^2 \right\rangle - \left\langle \frac{1}{N}\sum_{i=1}^{N}\sigma_i \right\rangle^2 \\ &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i\right).\end{aligned} \tag{7.44}$$

On the other hand, we have

$$\partial_h^2 F_N = \left\langle \frac{1}{N}\left(\sum_{i=1}^{N}\sigma_i\right)^2 \right\rangle - \frac{1}{N}\left\langle \sum_{i=1}^{N}\sigma_i \right\rangle^2 = N\cdot\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i\right). \tag{7.45}$$

Thus, we conclude from (7.44) and (7.45) that

$$\partial_t F_N - (\partial_h F_N)^2 = \frac{1}{N}\cdot\partial_h^2 F_N. \tag{7.46}$$

Recall that the limit is

$$f(t,h) = \sup_{m\in[-1,1]}(tm^2 + hm - \psi^*(m)). \tag{7.47}$$

Let $m_0(t,h)$ be the minimizer of (7.47). By the envelope theorem (see SPS3.1), at any differentiable point (t,h) of f, we have

$$\partial_t f(t,h) = m_0(t,h)^2 \quad \text{and} \quad \partial_h f(t,h) = m_0(t,h).$$

Thus, at any differentiable point (t,h)

$$\partial_t f(t,h) - (\partial_h f(t,h))^2 = 0. \tag{7.48}$$

The initial condition can be written as

$$F_N(0,h) = \frac{1}{N}\log 2^{-N}\sum_{\sigma\in\{\pm1\}^N}\exp\left(h\sum_{i=1}^{N}\sigma_i\right) = \log\cosh(h) = \psi(h). \tag{7.49}$$

By Rademacher's theorem and the fact f is Lipschitz, f is differentiable almost everywhere. Thus, we conclude from (7.48) and (7.49) that

$$\begin{cases}\partial_t f(t,h) - (\partial_h f(t,h))^2 = 0. \\ f(t=0,\cdot) = \psi,\end{cases} \tag{7.50}$$

almost surely. One may ask if (7.50) characterizes the function f. The answer is no. In fact, the following construction will give us infinitely many solutions.

Example 7.4.1. Note that $(t,h)\mapsto 0$, $(t,h)\mapsto t+h$ and $(t,h)\mapsto t-h$ are all solutions of (7.50). From these solutions, we can construct the following function (See Figure 7.2)

$$\widetilde{f}(t,h) = \begin{cases} t+h & h\in[-t,0] \\ t-h & h\in[0,t] \\ 0 & \text{otherwise.}\end{cases} \tag{7.51}$$

We can put the corner at anywhere or anytime we want, so there are uncountably many solutions of 7.50 which are Lipschitz.

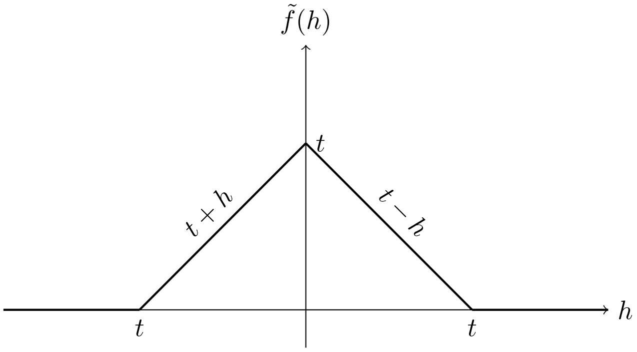


Figure 7.2: Graph of the function $\widetilde{f}$.

One may wonder if stronger regularity assumptions than Lipschitz continuity will solve the uniqueness problem. However, we are not allowed to do so since the desired solution in our mind can have corner singularities. Thus, what we really need is to restrict ourselves to a smaller class of solutions which preserves some properties that we required to be true even after passing to limit.

To be precise, if there exists two solutions f and g of the following PDEs

$$\partial_t f - (\partial_h f)^2 = \varepsilon \cdot \Delta f$$
$$\partial_t g - (\partial_h g)^2 = \varepsilon \cdot \Delta g$$

with intial conditions satisfying

$$f(t=0,\cdot) \leq g(t=0,\cdot),$$

then the inequality holds for all time t. This is called the maximum principle. We want the maximum principle still be valid when we take $\varepsilon \to 0$. In the next section, we will see that this true if we consider the viscosity solutions.

7.4.2 Viscosity solutions

Let $\mathsf{H} \in C^1(\mathbb{R}^d, \mathbb{R})$. The goal of this section is to develop a well-posed theory of the equation

$$\partial_t f - \mathsf{H}(\nabla f) = 0, \tag{HJ}$$

where $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$. Consider the following PDE with the same initial condition of (HJ).

$$\partial_t f_\varepsilon - \mathsf{H}(\nabla f_\varepsilon) = \varepsilon \Delta f_\varepsilon. \tag{HJ$_\varepsilon$}$$

We want to define the solution of (HJ) as the limit of (HJ$_\varepsilon$) when $\varepsilon \to 0$. Suppose that f_ε indeeds converges to a limit f as $\varepsilon \to 0$. We will see that f does retains some property in the same spirit of the maximum principle.

Let $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^d, \mathbb{R})$ and $(t,x) \in (0,\infty) \times \mathbb{R}^d$ be such that $f - \phi$ has a strict local maximum at (t,x).

For each $\varepsilon > 0$, there should be a point $(t_\varepsilon, x_\varepsilon)$ such that $f_\varepsilon - \phi$ has a local maximum at $(t_\varepsilon, x_\varepsilon)$, and $(t_\varepsilon, x_\varepsilon)$ coverges to (t,x) as $\varepsilon \to 0$. Since f_ε is smooth, at $(t_\varepsilon, x_\varepsilon)$ we have

$$\partial_t(f_\varepsilon - \phi) = 0, \quad \nabla(f_\varepsilon - \phi) = 0, \quad \text{and} \quad \Delta(f_\varepsilon - \phi) \leqslant 0. \tag{7.52}$$

Thus, we have

$$(\partial_t \phi - \mathsf{H}(\nabla\phi))(t_\varepsilon, x_\varepsilon) = (\partial_t f_\varepsilon - \mathsf{H}(\nabla f_\varepsilon))(t_\varepsilon, x_\varepsilon) = \varepsilon \cdot \Delta f_\varepsilon(t_\varepsilon, x_\varepsilon) \leqslant \varepsilon \cdot \Delta\phi(t_\varepsilon, x_\varepsilon). \tag{7.53}$$

Since ϕ is smooth, when $\varepsilon \to 0$, we obtain that

$$(\partial_t \phi - \mathsf{H}(\nabla\phi))(t,x) \leqslant 0. \tag{7.54}$$

Therefore, we conclude that if ϕ is smooth and touches f from above, then (7.54) holds at the contact point. Similarly, if ϕ touches f from below, then the converse inequality of (7.54) holds at the contact point. This motivates the following definition.

Definition 7.4.2. Let $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be continuous. We say that f is a *viscosity subsolution* to (HJ) if for any $(t, x) \in (0, \infty) \times \mathbb{R}^d$ and $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^d, \mathbb{R})$ such that $f - \phi$ has a local maximum at (t, x), we have

$$(\partial_t \phi - \mathsf{H}(\nabla \phi))(t, x) \leqslant 0.$$

We say that f is a *viscosity supersolution* to (HJ) if for any $(t, x) \in (0, \infty) \times \mathbb{R}^d$ and $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^d, \mathbb{R})$ such that $f - \phi$ has a local maximum at (t, x), we have

$$(\partial_t \phi - \mathsf{H}(\nabla \phi))(t, x) \geqslant 0.$$

Finally, the function f is called a *viscosity solution* to (HJ) if it is both a subsolution and a supersolution.

Remark 7.4.3. Replacing the locally maximal condition by strictly locally maximal condition yields an equivalent definition. We can also replace $\phi \in C^\infty$ by $\phi \in C^2$. (See SPS.4).

If f is a C^2 fucntion and it solves (HJ) everywhere, then f is a viscosity solution.

Example 7.4.4. Now, we verify that the function $\widetilde{f}$ constructed in Example 7.4.1 is not a viscosity solution. Define $\phi(t, x) = t$. Then,

$$\widetilde{f} - \varphi \leqslant 0 \quad \text{and} \quad \widetilde{f}(t, 0) - \varphi(t, 0) = 0,$$

so $f - \phi$ has a local maximum at $(t, 0)$. If f was a viscosity solution, the we should have

$$(\partial_t \phi - (\partial_x \phi)^2)(t, 0) \leqslant 0.$$

However,

$$(\partial_t \phi - (\partial_x \phi)^2) = 1,$$

which leads to a contradiction.

Back to the Curie–Weiss model, recall that $F_N(0, \cdot)$ does not depend on N, and that $|\partial_h F_N|, |\partial_t F_N| \leqslant 1$. Also, the sequence F_N are all one Lipschitz both in t and x. Therefore, by the Arzelà–Ascoli theorem, the sequence F_N is precompact with respect to the topology induced by local uniform convergence.

Proposition 7.4.5. Recall that F_N is defined as in (7.41). Let f be any subsequential limit of F_N. Then f is a viscosity solution to the PDE with initial condition

$$\begin{cases} \partial_t f - (\partial_h f)^2 = 0 \\ f(0, \cdot) = F_N(0, \cdot) = \psi. \end{cases} \tag{7.55}$$

Proof. Recall that

$$\partial_t F_N - (\partial_h F_N)^2 = \frac{1}{N} \cdot \partial_h^2 F_N.$$

Let $(t, h) \in (0, \infty) \times \mathbb{R}$ and $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R})$ be such that $f - \phi$ has a strict local maximum at (t, h). Since F_N converges to f, there exists $(t_N, h_N) \in (0, \infty) \times \mathbb{R}$ such that for every N sufficiently large, $F_N - \phi$ has a local maximum at (t_N, h_N) and $(t_N, h_N) \to (t, h)$ as $N \to \infty$ (See exercise in PS4.3 for an argument in solution).

Since $t_N > 0$ as N sufficiently large, we have

$$\begin{aligned} (\partial_t \phi - (\partial_h \phi)^2)(t_N, h_N) &= (\partial_t F_N - (\partial_h F_N)^2)(t_N, h_N) \\ &= \frac{1}{N} \partial_h^2 F_N(t_N, h_N) \\ &\leq \frac{1}{N} \partial_h^2 \phi(t_N, h_N). \end{aligned}$$

Since ϕ is smooth, passing to the limit as $N \to \infty$ we obtain

$$(\partial_t \phi - (\partial_h \phi)^2)(t, h) \leq 0.$$

This shows that f is a subsolution. The argument for supersolution is similar. We complete the proof.

□

7.4.3 Comparison principle

Notice that Propostion 7.4.5 already gives us the existence of a viscosity solution. It remains to argue that for a given initial conditon, there exists at most one viscosity solution. We will show the more general result called the comparison principle.

Theorem 7.4.6 (comparison principle). Let u be a subsolution and v be a supersolution to (HJ) such that both u and v are uniformly Lipschitz in the x variable. We have

$$\sup_{\mathbb{R}_+ \times \mathbb{R}^d} (u - v) = \sup_{\{0\} \times \mathbb{R}^d} (u - v)$$

Remark 7.4.7. In particular, if u and v are solutions to (HJ) and $u(0, \cdot) = v(0, \cdot)$, then $u = v$. Indeed, from the Theorem 7.4.6 we have $u \leq v$ and by symmetry we get the result.

Sketch of proof of Theroem 7.4.6. Here we prove the statement in a special case for torus but not $\mathbb{R}^d$. We argue by contradiction. Suppose that for some $T < \infty$ we have

$$\sup_{[0,T] \times \mathbb{R}^d} (u - v) > \sup_{\{0\} \times \mathbb{R}^d} (u - v)$$

For $\varepsilon > 0$, denote $\chi(t) := \frac{\varepsilon}{T-t}$. For $\varepsilon > 0$ sufficiently small, we get

$$\sup_{[0,T] \times \mathbb{R}^d} (u - v - \chi) > \sup_{\{0\} \times \mathbb{R}^d} (u - v - \chi)$$

The main problem is that u and v are not differentiable. The idea is to double the variable as follows. For every $\alpha \geqslant 1$, let

$$\psi_\alpha(t, x, t', x') := u(t, x) - v(t', x') - \frac{\alpha}{2}((t' - t)^2 + |x' - x|^2) - \chi(t).$$

Suppose that the supremum of ψ_α is achieved at some points $(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha)$ and $(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha) \to (t_0, x_0, t'_0, x'_0)$ as $\alpha \to \infty$. We first argue that we must have $t_0 = t'_0$ and $x_0 = x'_0$. Indeed, since u and v are continous on the torus in the space variable, they remain bounded over any bounded set. Then for some constant $C > 0$ we have

$$\frac{\alpha}{2}((t'_\alpha - t_\alpha)^2 + |x'_\alpha - x_\alpha|^2) \leq C$$

Therefore, $t_0 = t'_0$ and $x_0 = x'_0$. (otherwise, $\psi_\alpha(t, x, t', x') \to -\infty$ as $\alpha \to \infty$)

Note that $0 \leq \chi(t_\alpha) \leq C$ for some constant $C > 0$, so we have $t_0 < T$. Next, we claim that we must have $t_0 > 0$. Indeed, observe that

$$\psi_\alpha(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha) \geqslant \sup_{[0,T) \times \mathbb{R}^d} (u - v - \chi) \tag{7.56}$$

and

$$\psi_\alpha(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha) \leq u(t_\alpha, x_\alpha) - v(t'_\alpha, x'_\alpha) - \chi(t_\alpha). \tag{7.57}$$

By continuity, combing (7.56) and (7.57) we obtain

$$\sup_{[0,T) \times \mathbb{R}^d} (u - v - \chi) \leq (u - v - \chi)(t_0, x_0). \tag{7.58}$$

Sine the LHS of (7.58) is strictly greater than $\sup_{\{0\} \times \mathbb{R}^d} (u - v - \chi)$, we must have $t_0 > 0$.

As a consequence, for α suffifently large we have

$$0 < t_\alpha, t'_\alpha < T.$$

Notice that the mapping

$$(t,x) \mapsto u(t,x) - v(t'_\alpha, x'_\alpha) - \frac{\alpha}{2}((t-t'_\alpha)^2 + |x'_\alpha - x|^2) - \chi(t)$$

is maximal at (t_α, x_α). Let $\phi(t,x) := \frac{\alpha}{2}((t'_\alpha - t)^2 + |x'_\alpha - x|^2)$. Since u is a viscosity subsolution to (HJ), we infer that

$$(\partial_t(\phi+\chi) - H(\nabla(\phi+\chi))(t_\alpha, x_\alpha) \le 0. \tag{7.59}$$

From (7.59), we have

$$(\partial_t \phi - H(\nabla \phi))(t_\alpha, x_\alpha) \le -\frac{\varepsilon}{(T-t_\alpha)^2}. \tag{7.60}$$

Similarly, if the mapping

$$(t',x') \mapsto v(t',x') - u(t_\alpha, x_\alpha) + \frac{\alpha}{2}((t'-t_\alpha)^2 + |x'-x_\alpha|^2) + \chi(t_\alpha)$$

is minimal at (t'_α, x'_α). Let $-\widetilde{\phi}(t',x') := \frac{\alpha}{2}((t'-t_\alpha)^2 + |x'-x_\alpha|^2)$. Since v is a viscosity supersolution to (HJ), we get

$$(\partial_t \widetilde{\phi} - H(\nabla \widetilde{\phi}))(t'_\alpha, x'_\alpha) \geqslant 0. \tag{7.61}$$

Notice that $\partial_t \phi(t_\alpha, x_\alpha) = \partial_t \widetilde{\phi}(t'_\alpha, x'_\alpha)$ and $\nabla \phi(t_\alpha, x_\alpha) = \nabla \widetilde{\phi}(t'_\alpha, x'_\alpha)$. Combining (7.60) and (7.60) gives us a contradiction.

□

Remark 7.4.8. The reasons for using function χ are to assume that $\partial_t u - H(\nabla u) \le -\varepsilon$ and detach these optimizatin points from the endpoint.

Theorem 7.4.9 (stronger version of comparison principle)**.** Let $T \in (0,\infty)$ and let u and v be respectively subsolution and supersulotion to (HJ) on $[0,T) \times \mathbb{R}^d$ that are both uniformly L-Lipschitz in the x variable. Define

$$V := \sup\left\{ \frac{|H(p') - H(p)|}{|p'-p|} : |p|, |p'| \le L \right\}.$$

For every $R, M \in \mathbb{R}$ such that $M > 2L$, the mapping

$$(t,x) \mapsto u(t,x) - v(t,x) - M(|x| + Vt - R)_+ \tag{7.62}$$

achieves its supremum at a point in $\{0\} \times \mathbb{R}^d$.

Proof.

Without loss of generality, we assume that u and v are continuous on $[0,T]$. We argue by contradiction. Assume that the supremum of the mapping (7.62) is not achieved on $\{0\} \times \mathbb{R}^d$. We start by replacing

$$(t,x) \mapsto M(|x| + Vt - R)_+$$

by a smooth function.

Let $\theta \in C^\infty(\mathbb{R})$ be such that for every $\omega \in \mathbb{R}$

$$(\omega - \varepsilon_0)_+ \le \theta(\omega) \le \omega + 1.$$

Consider

$$\Phi(t,x) := M\theta((\varepsilon_0 + \sum_{k=1}^d |x_k|^2)^{1/2} + Vt - R).$$

By choosing $\varepsilon_0 \in (0,1]$ sufficiently small, we can make such that

$$\sup_{[0,T)\times\mathbb{R}^d} (u - v - \Phi) > \sup_{\{0\}\times\mathbb{R}^d} (u - v - \Phi).$$

Observe that

$$\partial_t \Phi(t,x) = MV\theta'((\varepsilon_0 + \sum_{k=1}^d |x_k|^2)^{1/2} + Vt - R)$$

and

$$\partial_{x_k}\Phi(t,x) = \frac{Mx_k}{(\varepsilon_0 + \sum_{k=1}^d |x_k|^2)^{1/2}}\theta'((\varepsilon_0 + \sum_{k=1}^d |x_k|^2)^{1/2} + Vt - R).$$

Hence,

$$\partial_t\Phi \geqslant V|\nabla\Phi| \tag{7.63}$$

Notice that

$$\Phi(t,x) \geqslant M(|x| + Vt - R - 1)_+$$

For some $\varepsilon > 0$ to be determined, we set

$$\chi(t,x) := \Phi(t,x) + \frac{\varepsilon}{T-t}.$$

We can choose $\varepsilon > 0$ small enough so that

$$\sup_{[0,T)\times\mathbb{R}^d}(u - v - \chi) > \sup_{\{0\}\times\mathbb{R}^d}(u - v - \chi).$$

For every $\alpha \geqslant 1$, let

$$\Psi_\alpha(t,x,t',x') := u(t,x) - v(t',x') - \frac{\alpha}{2}((t'-t)^2 + |x'-x|^2) - \chi(t,x).$$

Since $M > 2L$, one can check that the supremum of Ψ_α is achieved at a point $(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha)$. Moreover, this point remains in a bounded region, so up to the extaction of a subsequence, we have $(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha) \to (t_0, x_0, t'_0, x'_0)$ as $\alpha \to \infty$. Since for some constants $C > 0$, we have

$$\frac{\alpha}{2}((t'_\alpha - t_\alpha)^2 + |x'_\alpha - x_\alpha|^2) \leq C$$

Thus, we must have $t_0 = t'_0$ and $x_0 = x'_0$. Thanks to the term $\frac{\varepsilon}{T-t}$, we have $t_0 < T$. Arguing as in Theorem 7.4.6, we also have

$$\Psi_\alpha(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha) \geqslant \sup_{[0,T)\times\mathbb{R}^d}(u - v - \chi) \geqslant (u - v - \chi)(t_0, x_0) \tag{7.64}$$

while

$$\Psi_\alpha(t_\alpha, x_\alpha, t'_\alpha, x'_\alpha) \leq u(t_\alpha, x_\alpha) - v(t'_\alpha, x'_\alpha) - \chi(t_\alpha, x_\alpha). \tag{7.65}$$

Thus, by (7.64) and (7.65)

$$(u - v - \chi)(t_0, x_0) \geqslant \sup_{[0,T)\times\mathbb{R}^d}(u - v - \chi). \tag{7.66}$$

Hence, $t_0 > 0$ and thus $t_\alpha > 0$ for every α large enough.

Notice that the mapping

$$(t,x) \mapsto u(t,x) - v(t'_\alpha, x'_\alpha) - \frac{\alpha}{2}((t - t'_\alpha)^2 + |x'_\alpha - x|^2) - \chi(t,x)$$

achieves its maximum at (t_α, x_α). Let $\phi(t,x) := \frac{\alpha}{2}((t - t'_\alpha)^2 + |x'_\alpha - x|^2)$. Hence,

$$\partial_t(\phi + \chi) - H(\nabla(\phi + \chi)) \leq 0 \tag{7.67}$$

at some points (t_α, x_α). Hence,

$$\partial_t\phi(t_\alpha, x_\alpha) + \partial_t\Phi(t_\alpha, x_\alpha) - H(\nabla\phi(t_\alpha, x_\alpha) + \nabla\Phi(t_\alpha, x_\alpha)) \leq -\frac{\varepsilon}{(T - t_\alpha)^2} \tag{7.68}$$

By the Lipschitz property of H,

$$|H(\nabla\phi(t_\alpha, x_\alpha) + \nabla\Phi(t_\alpha, x_\alpha)) - H(\nabla\phi(t_\alpha, x_\alpha))| \leq V|\nabla\Phi(t_\alpha, x_\alpha)|. \tag{7.69}$$

By (7.63) and (7.69), we get

$$(\partial_t\phi - H(\nabla\phi))(t_\alpha, x_\alpha) \leq -\frac{\varepsilon}{(T - t_\alpha)^2}. \tag{7.70}$$

Similarly, the mapping

$$(t', x') \mapsto v(t', x') - u(t_\alpha, x_\alpha) + \frac{\alpha}{2}((t' - t_\alpha)^2 + |x' - x_\alpha|^2) + \chi(t_\alpha, x_\alpha)$$

ahcieves its minimum at (t'_α, x'_α). Let $-\widetilde{\phi}(t', x') := \frac{\alpha}{2}((t' - t_\alpha)^2 + |x' - x_\alpha|^2)$. Thus, we have

$$\partial_t\widetilde{\phi} - H(\nabla\widetilde{\phi}) \geqslant 0 \tag{7.71}$$

at some points (t'_α, x'_α).

Since $\partial_t\phi(t_\alpha, x_\alpha) = \partial_t\widetilde{\phi}(t'_\alpha, x'_\alpha)$ and $\nabla\phi(t_\alpha, x_\alpha) = \nabla\widetilde{\phi}(t'_\alpha, x'_\alpha)$, (7.71) contradicts with (7.70). We complete the proof.

□

Theorem 7.4.9 is indeed stronger than Theorem 7.4.6. Indeed, we can prove Theorem 7.4.6 using Theorem 7.4.9 in the following way.

Another proof of Theorem 7.4.6. We argue by contradiction. Suppose that

$$\sup_{\mathbb{R}_+\times\mathbb{R}^d}(u-v) > \sup_{\{0\}\times\mathbb{R}^d}(u-v)$$

There exist t_0 and x_0 such that

$$(u-v)(t_0,x_0) > \sup_{\{0\}\times\mathbb{R}^d}(u-v).$$

Fix $M = 2L+1$ and $R = |x_0| + Vt_0$ so that

$$\begin{aligned} u(t_0,x_0) - v(t_0,x_0) - M(|x_0| + Vt_0 - R)_+ &= (u-v)(t_0,x_0) \\ &> \sup_{\{0\}\times\mathbb{R}^d}(u-v) \\ &> \sup_{\{0\}\times\mathbb{R}^d}(u - v - M(|x_0| + Vt_0 - R))_+, \end{aligned}$$

which is a contradiction by Theorem 7.4.9. □

To sum up, we have seen that

- The sequence F_N is precompact by the Arzelà–Ascoli theorem.
- Any limit point of the sequence F_N must be a viscosity solution to (7.55).
- There is a unique solution to (7.55).

We conclude that the sequence F_N converges to the unique viscosity solution to (7.55).

The initial condition is

$$0 \le \psi(h) = F_N(0,h) = \log\cosh(h) \sim \frac{h^2}{2}, \quad h \to 0.$$

Then there exists $C < \infty$ such that

$$0 \le \psi(h) \le Ch^2.$$

By Theorem 7.4.6,

$$0 \le f(t,h) \le \frac{Ch^2}{1-4Ct}, \quad t < \frac{1}{4C}.$$

In particular, $\partial_h f(t,0) = 0$ for all $t < \frac{1}{4C}$.

7.4.4 Variational formulas for HJ equations

Let f be the viscosity solution to

$$\begin{cases} \partial_t f - \mathsf{H}(\nabla\phi) = 0 & \text{on } \mathbb{R}_+\times\mathbb{R}^d \\ f(t=0,\cdot) = \psi \end{cases}.$$

We assume throughout that ψ is Lipschitz. Recall that we write

$$\mathsf{H}^*(q) := \sup_{p\in\mathbb{R}}(p\cdot q - \mathsf{H}(p)) \tag{7.72}$$

Theorem 7.4.10 (Hopf-Lax formula)**.** If H is convex then

$$f(t,x) = \sup_{y\in\mathbb{R}^d}\left(\psi(y) - t\mathsf{H}^*\left(\frac{y-x}{t}\right)\right). \tag{7.73}$$

For a proof see []

Theorem 7.4.11 (Hopf formula)**.** If ψ is convex, then

$$f(t,x) = \sup_{p\in\mathbb{R}^d} \inf_{y\in\mathbb{R}^d} \big(\psi(y) + p\cdot(x-y) + t\mathsf{H}(p)\big) . \tag{7.74}$$

Observe that the Hopf-Lax formula can be rewritten as

$$f(t,x) = \sup_{y\in\mathbb{R}^d} \inf_{p\in\mathbb{R}^d} \big(\psi(y) + p\cdot(x-y) + t\mathsf{H}(p)\big) . \tag{7.75}$$

I don't know how the following comment fits in the continuation of the text and I think it is ok just ignore it: The optimization over y of $g(y) := \psi(y) + p\cdot(x-y) + t\mathsf{H}(p)$ imposes that $\nabla\psi(y) = p$ at the optimal y.

When both H and ψ are convex, can we verify that the two formulas (7.75) and (7.74) coincide?

Proposition 7.4.12. Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be two convex functions. We have

$$\sup_{x\in\mathbb{R}^d} \inf_{y\in\mathbb{R}^d} \big(f(x) + g(y) - x\cdot y\big) = \sup_{y\in\mathbb{R}^d} \inf_{x\in\mathbb{R}^d} \big(f(x) + g(y) - x\cdot y\big) \tag{7.76}$$

Proof. Recall the Fenchel-Moreau Theorem: $f(x) = \sup_y \big(x\cdot y - f^*(y)\big)$. We write,

$$\begin{aligned}
\sup_{x\in\mathbb{R}^d} \inf_{y\in\mathbb{R}^d} \big(f(x) + g(y) - x\cdot y\big) &= \sup_{x\in\mathbb{R}^d} \big(f(x) - g^*(x)\big) \\
&= \sup_{x\in\mathbb{R}^d} \sup_{y\in\mathbb{R}^d} \big(x\cdot y - f^*(y) - g^*(x)\big) \\
&= \sup_{y\in\mathbb{R}^d} \sup_{x\in\mathbb{R}^d} \big(x\cdot y - f^*(y) - g^*(x)\big) \\
&= \sup_{y\in\mathbb{R}^d} \big(g(y) - f^*(y)\big) \\
&= \sup_{y\in\mathbb{R}^d} \inf_{x\in\mathbb{R}^d} \big(g(y) + f(x) - x\cdot y\big).
\end{aligned} \tag{7.77}$$

□

7.4.5 Convex Selection Principle

Scribe: Emily Crawford Das

We will rely on a new selection principle that leverages the convexity of F_N (in some loose sense, counterexamples must locally look like the hat function 7.2, so must be neither convex nor concave).

Theorem 7.4.13. (Convex Selection Principle): Let $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be a convex function such that the equation

$$\partial_t f - \mathsf{H}(\nabla f) = 0$$

is satisfied on a dense subset of $\mathbb{R}_+ \times \mathbb{R}^d$, where f is assumed to be uniformly Lipschitz and H is assumed to be locally Lipschitz. Assume also that $f(0,\cdot)$ is of class C^1. Then f is a viscosity solution to the equation. [Clarification: The assumption is that the set $\{(t,x) \in \mathbb{R}_+ \times \mathbb{R}^d : f$ is differentiable at (t,x) and $(\partial_t f - \mathsf{H}(\nabla f))(t,x) = 0\}$ is dense in $\mathbb{R}_+ \times \mathbb{R}^d$.]

We will first prove the following Corollary (assuming the statement of the Theorem).

Corollary 7.4.14. Let $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be a convex function such that $f(0,\cdot) =: \psi$ is of class C^1. Suppose that for every $(t,x) \in (0,\infty) \times \mathbb{R}^d$ and $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$ such that $f - \phi$ has a [strict] local maximum at (t, x), we have $(\partial_t \phi - \mathsf{H}(\nabla \phi))(t, x) = 0$. Then f is the unique viscosity solution to

$$\begin{cases} \partial_t f - \mathsf{H}(\nabla \phi) = 0 & \text{on } \mathbb{R}_+ \times \mathbb{R}^d \\ f(t = 0, \cdot) = \psi \end{cases}.$$

Proof. Let (t, x), ϕ be as in the statement. We have

$f(t+s, x+y) - f(t, x) \leq \phi(t+s, x+y) - \phi(t, x)$
$= \nabla\phi(t, x) \cdot y + \partial_t\phi(t, x)s + \mathcal{O}(|y|^2 + s^2).$

Since f is convex, by the hyperplane separation theorem, there exists $a \in \mathbb{R}, p \in \mathbb{R}^d$ such that

$f(t+s, x+y) - f(t, x) \geqslant as + p \cdot y.$

And, hence, we have

$as + p \cdot y \leq f(t+s, x+y) - f(t, x) \leq \partial_t\phi(t, x)s + y \cdot \nabla\phi(t, x) + \mathcal{O}(|y|^2 + s^2).$

It follows that $a = \partial_t\phi(t, x),\ p = \nabla\phi(t, x)$, and f is differentiable at (t, x) and its derivatives coincide with those of ϕ. [Recalling that $f - \phi$ is maximal at (t, x), we infer that $\partial_t f(t, x) = \partial_t\phi(t, x)$ and $(\nabla f - \nabla\phi)(t, x) = 0$. So $(\partial_t f - \mathsf{H}(\nabla f))(t, x) = 0$.]

In order to show the corollary, it thus suffices to show that the set $\{(t, x) \in (0, \infty) \times \mathbb{R}^d : \exists\ \phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$ such that $f - \phi$ has a local maximum at $(t, x)\}$ is dense in $\mathbb{R}_+ \times \mathbb{R}^d$. For a fixed $(t_0, x_0) \in (0, \infty) \times \mathbb{R}^d$, consider

$$(t, x) \mapsto f(t, x) - \frac{\alpha}{2}((t - t_0)^2 - |x - x_0|^2). \tag{7.78}$$

Let V be a fixed compact neighborhood of (t_0, x_0), and (t_α, x_α) be the maximum of (7.78) over V. It is easy to see that $(t_\alpha, x_\alpha) \to (t_0, x_0)$ as $\alpha \to \infty$, so for α sufficiently large, (t_α, x_α) is a local maximum of (7.78). Since $(t_\alpha, x_\alpha) \to (t_0, x_0)$, we obtain the result. □

7.4.6 Identification of the limit free energy of the Generalized Curie-Weiss model

By definition, we have $F_N(0, \cdot) \to \psi$ as $N \to \infty$, and we assume that ψ is C^1. Let f be a subsequential limit of F_N. [Recall that F_N is precompact by Arzelà-Ascoli Theorem]. Let $(t_0, h_0) \in (0, \infty) \times \mathbb{R}$, and $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$ be such that $f - \phi$ has a strict local maximum at (t_0, h_0). We wish to verify that $(\partial_t\phi - \xi(\partial_h\phi))(t, h) = 0$. There exists $(t_N, h_N) \to (t_0, h_0)$ such that for every N sufficiently large, $F_N - \phi$ has a local maximum at (t_N, h_N). [Recall that $|\partial_t F_N - \xi(\partial_h F_N)| \leq \frac{C}{N}(\partial_h^2 F_N)$]. At (t_N, h_N), we have

$$\partial_t(F_N - \phi) = 0, \partial_h(F_N - \phi) = 0, \text{ and } \partial_h^2(F_N - \phi) \leq 0.$$

So, $|\partial_t\phi - \xi(\partial_h\phi)|(t_N, h_N) \leq \frac{C}{N}(\partial_h^2\phi(t_N, h_N))$; Letting $N \to \infty$, we get $(\partial_t\phi - \xi(\partial_h\phi))(t, h) = 0$, as desired. Hence, by the corollary, any subsequential limit f of F_N is the (unique for a fixed initial condition, ψ) viscosity solution to

$$\begin{cases} \partial_t f - \xi(\partial_h f) = 0 \\ f(t = 0, \cdot) = \psi \end{cases}.$$

Note that by the Hopf formula, the solution to

$$\begin{cases} \partial_t f - \xi(\partial_h f) = 0 \\ f(t = 0, \cdot) = \psi \end{cases}$$

is $f(t, h) = \sup\limits_{m \in \mathbb{R}} (t\xi(m) + hm - \psi^*(m))$, recovering our formula from large deviations. There remains to show that the convex selection is valid. We first state a couple of properties of convex functions.

Definition 7.4.15. For every convex function $f : U \to \mathbb{R}, U \subseteq \mathbb{R}^d$, we define the subdifferential at $x \in U$ by

$$\partial f(x) := \left\{p \in \mathbb{R}^d : \text{ for all } y \in U, f(y) \geqslant f(x) + p \cdot (y - x)\right\}$$

i.e., the set of slopes that give us a supporting hyperplane at x.

Note that this set is not empty.

Proposition 7.4.16. If $x_m \to x$ in the interior of U and $p_m \in \partial f(x_m)$, $p_m \to p$, then $p \in \partial f(x)$.

Proposition 7.4.17. If f is differentiable at x, and x is in the interior of U, then $\partial f(x) = \{\nabla f(x)\}$.

Consequence: If a Lipschitz, convex function $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ satisfies $\partial_t f - \mathsf{H}(\nabla f) = 0$ on a dense set, then it satisfies the equation at every point of differentiability in $(0, \infty) \times \mathbb{R}^d$. Moreover, for every $(t, x) \in \mathbb{R}_+ \times \mathbb{R}^d$, there exists $(a, p) \in \partial f(t, x)$ such that $a - \mathsf{H}(p) = 0$. [Note that on this dense set of points, $\partial f(t, x)$ is a singleton, $\partial f(t, x) = \{(\partial_t f(t, x), \nabla f(t, x))\}$]. We will now prove the Convex Selection Principle.

Proof.

(Step 1): We show that f is a subsolution. Let $(t, x) \in (0, \infty) \times \mathbb{R}^d$ and $\phi \in C^\infty$ be such that $f - \phi$ has a local maximum at (t, x). Then
$f(t', x') - f(t, x) \leq \phi(t', x') - \phi(t, x)$
$= (t' - t)\partial_t \phi(t, x) + (x' - x) \cdot \nabla \phi(t, x) + \mathcal{O}((t' - t)^2 + |x' - x|^2)$.
Since f is convex, we can argue as in the proof of Corollary 7.4.14 and deduce that f is differentiable at (t, x), with $\partial f(t, x) = \{(\partial_t \phi(t, x), \nabla \phi(t, x))\}$ is a singleton. Thus $(\partial_t \phi - \mathsf{H}(\nabla \phi))(t, x) = 0$. Indeed, f is a subsolution.

(Step 2): We give a partial argument showing that f is a supersolution; to be completed in the last step. If ϕ touches f from below at (t, x), then $(\partial_t \phi(t, x), \nabla \phi(t, x)) \in \partial f(t, x)$. It thus suffices to show that, for every $(a, p) \in \partial f(t, x)$, we have $a - \mathsf{H}(p) \geqslant 0$. By definition, we have
$f(t', x') \geqslant f(t, x) + (t' - t)a + (x' - x) \cdot p$.
So the mapping $y \mapsto f(0, y) - y \cdot p$ is bounded from below. In this step, we assume that

$$\inf_{y \in \mathbb{R}^d} \left(f(0, y) - y \cdot p\right)$$

is acheived and show how to conclude. [Note that the point where the infimum is achieved will be a point with slope p].

Call y the optimizer. By the "consequence" preceding this proof, there exists $(b, p') \in \partial f(0, y)$ such that $b - \mathsf{H}(p') = 0$. Since $f(0, \cdot)$ in C^1, we must have $p' = \nabla f(0, y) = p$, and thus $(b, p) \in \partial f(0, y)$ and $b - \mathsf{H}(p) = 0$. Notice that the mapping

$$g : \begin{cases} [0, 1] \to \mathbb{R} \\ \lambda \mapsto f(\lambda(t, x) + (1 - \lambda)(0, y)) \end{cases}$$

is convex. Since $(b, p) \in \partial f(0, y)$, the right derivative at 0 satisfies

$$g'_+(0) \geqslant bt + p \cdot (x - y).$$

Since $(a, p) \in \partial f(t, x)$, the left derivative at 1 satisfies

$$g'_-(1) \leq at + p \cdot (x - y).$$

Since g is convex, we must have $g'_+(0) \leq g'_-(1)$, so $a \geqslant b$. Recalling that $b - \mathsf{H}(p) = 0$, we obtain $a - \mathsf{H}(p) \geqslant 0$, as desired.

(Step 3): We now address the possibility that the infimum is not attained. For every $\varepsilon > 0$, we consider

$$\inf_{y \in \mathbb{R}^d} \left(f(0, y) + \varepsilon|y| - y \cdot p\right)$$

This infimum is acheived at some $y_\varepsilon \in \mathbb{R}^d$, and $|\nabla f(0, y_\varepsilon) - p| \leq \varepsilon$. Moreover,

$$\lim_{\varepsilon \to 0} \inf_{y \in \mathbb{R}^d} \left(f(0, y) + \varepsilon|y| - y \cdot p\right) = \inf_{y \in \mathbb{R}^d} \left(f(0, y) - y \cdot p\right), \text{ and}$$

$$f(0, y_\varepsilon) - y_\varepsilon \cdot p \geqslant \inf_{y \in \mathbb{R}^d} \left(f(0, y) - y \cdot p\right), \text{ so that}$$

$$\lim_{\varepsilon\to 0} \varepsilon|y_\varepsilon| = 0.$$

As in the previous step, there exists $b_\varepsilon \in \mathbb{R}$ such that $(b_\varepsilon, \nabla f(0, y_\varepsilon)) \in \partial f(0, y_\varepsilon)$ and $b_\varepsilon - \mathsf{H}(\nabla f(0, y_\varepsilon)) = 0$. Continuing as before, we find that

$$b_\varepsilon t + \nabla f(0, y_\varepsilon) \cdot (x - y_\varepsilon) \leq at + p \cdot (x - y_\varepsilon).$$

Using $|\nabla f(0, y_\varepsilon) - p| \leq \varepsilon$ and $\lim\limits_{\varepsilon\to 0} \varepsilon|y_\varepsilon| = 0$, we find that $b_\varepsilon \leq a + o(1)$ as $\varepsilon \to 0$. Since $b_\varepsilon - \mathsf{H}(\nabla f(0, y_\varepsilon)) = 0$, and $\nabla f(0, y_\varepsilon) \to p$, we conclude that $a - \mathsf{H}(p) \geqslant 0$.

□

7.5 Statistical inference of low-rank matrices

Scribe: Yang Chu

7.5.1 Brief discussion of Convex Selection Principle

Recall from last lecture, the condition in Corollary 7.4.14 requires $f(0, \cdot) =: \psi$ to be C^1. Similar with the Large Deviation Principle, we show this assumption is necessary by giving a counter example.

Let $P_N = \frac{1}{2}\delta_{(1,\dots,1)} + \frac{1}{2}\delta_{(-1,\dots,-1)}$, $\xi(p) = p^2$.

In this case, the free energy is given by

$$F_N(t, h) = \frac{1}{N} \log\left(\frac{1}{2}\exp(-tN + hN) + \frac{1}{2}\exp(-tN - hN)\right) \to |h| - t$$

as $N \to \infty$ The relevant Hamilton-Jacobi equation is

$$\partial_t f + (\partial_h f)^2 = 0$$
$$f(0, h) = |h|$$

By the comparison principle, we have $f \geqslant 0$. Then $f(t, h) \neq |h| - t$. Note that $g(t, h) := |h| - t$ is convex, and solves the equation almost everywhere.

Remark 7.5.1. One can construct P_N such that the initial condition $f(0, h) = |h|$, but the free energy does converge to the solution of Hamilton-Jacobi equations.

7.5.2 Introduction

For a vector $\bar{x} \in \mathbb{R}^N$, we oberve a noisy version of $\bar{x}\bar{x}^*$, more precisely, $\sqrt{\frac{2t}{N}}\bar{x}\bar{x}^* + W$, where W is a matrix with independent standard Gaussian entries. We want to estimate $\bar{x}$.

Motivation of this model comes from community detection and recommendation systems.

Suppose we have notes $\{1, 2, \dots, N\}$ and two communities $+1$ and -1, indicated by a vector $\bar{x} = (\bar{x}_1, \dots, \bar{x}_N)$ of i.i.d. random variables with values 1 and -1 with equal probability $\frac{1}{2}$. We draw an edge between i and j with probability $\frac{dp}{N}$ if $\bar{x}_i = \bar{x}_j$, or $\frac{dq}{N}$ if $\bar{x}_i \neq \bar{x}_j$. with $p + q = 2$. (so that average degree of a node is d).

Question: if we only oberse the graph of connections, can we revover information about the groups?

For each pair (i, j), we observe $Ber(\frac{dp}{N})$ if $\bar{x}_i = \bar{x}_j$ and $Ber(\frac{dq}{N})$ if $\bar{x}_i \neq \bar{x}_j$, i.e. $Ber(d\frac{p+q}{2N} + d\frac{p-q}{2N}x_i\bar{x}_j)$.

The mean of this random variable is $\frac{d}{N} + d\frac{p-q}{2N}x_i\bar{x}_j$. The variance is $\frac{d}{N}+$ some small constant.

For the Gaussian case instead of Bernoulli, it's similar to oberserve, where W_{ij} are i.i.d. standard Gaussians,

$$d\frac{p-q}{2N}x_i\bar{x}_j + \sqrt{\frac{d}{N}}W_{ij}.$$

Divide both sides by $\sqrt{\frac{d}{N}}$, we have the setting as given in the beginning of this section. In the limit $N \to \infty$, and then $d \to \infty$, fixing $\sqrt{d}(p-q)$ remains $O(1)$.

Remark 7.5.2. More complex models can be imagined, e.g. more communities. This could be encoded by $\bar{X} \in \mathbb{R}^{N\times K}$ for N people and K communities. In this case, we are going to observe a noisy version of

$$\bar{X}^{\otimes 2}A, \text{where } \bar{X}^{\otimes 2} \in \mathbb{R}^{N^2\times K^2} \text{ is the tensor product, and } K \text{ is a fixed } K\times K \text{ matrix.}$$

More generally, $\bar{X}^{\otimes p}A$ (see papers by Reeves).

7.5.3 Approach by viscosity solutions

We will focus on the $x\bar{x}^*$ model (but the methods will be appliable to the general model mentioned above):

$$Y := \sqrt{\frac{2t}{N}}\bar{x}\bar{x}^* + W, \text{ where } \bar{x} \text{ is a vector of iid bounded random variables with law } P_N$$

W is a matrix of independent standard Gaussians. $t \geqslant 0$ denote the signal-to-noise ratio.

Our goal is to understand the "information-theoretic" limit to the possibility of recovering infromation of $\bar{x}$.

Reminder: if $f \in L^2(\mathbb{P})$ and $\mathcal{F}$ is a σ-algebra, then $\inf_{g \text{ is } \mathcal{F} \text{ measurable}} \mathbb{E}[(f-g)^2]$ is achieved for $g = \mathbb{E}[f \mid \mathcal{F}]$.

We are interested in :

$$\inf_{Z:Y\text{-measurable}} \mathbb{E}[|\bar{x}-Z|^2] = \mathbb{E}[|\bar{x}-\mathbb{E}[\bar{x} \mid Y]|^2]$$

Call $\mathbb{E}[|\bar{x}-\mathbb{E}[\bar{x} \mid Y]|^2]$ minimum mean-square error $=: mmse_N(t)$. May also consider $\mathbb{E}[\bar{x}\bar{x}^* - \mathbb{E}[|\bar{x}^* \mid Y]|^2]$, here we write $a \cdot b$ for the entrywise produce, and $|a|^2 = a \cdot a$.

To write down the conditional law of $\bar{x}$ given Y, by Bayes' rule , we informally haveL

$$\begin{aligned}\mathbb{P}[\bar{x} = x \mid Y = y] &= \frac{\mathbb{P}[\bar{x} = x, Y = y]}{\mathbb{P}[Y = y]} \\ &= \frac{dP_n(x)\exp\left(-\frac{1}{2}|y - \sqrt{\frac{2t}{N}}xx^*|^2\right)}{\int dP_N(x')\exp\left(-\frac{1}{2}|y - \sqrt{\frac{2t}{N}}xx^*|^2\right)}\end{aligned}$$

Define $H_N^0 := \sqrt{\frac{2t}{N}}Y\cdot(xx^*) - \frac{t}{N}|xx^*|^2$, so that for any bounded measurable f:

$$\mathbb{E}[f(\bar{x}) \mid Y] = \frac{\int f(x)\exp\left(H_N^0(t,x)\right)dP_N(x)}{\int \exp(H_N^0(t,x'))dP_N(x')} \tag{7.79}$$

This is a Gibbs measure! One difference with Curie-Weiss is that now $H_N^0(t,x)$ is random.

We will use the notation

$$\langle f(x)\rangle := \frac{\int f(x)\exp\left(H_N^0(t,x)\right)dP_N(x)}{\int \exp(H_N^0(t,x'))dP_N(x')} = \mathbb{E}[f(\bar{x}) \mid Y] = \int f(x)dP_{\bar{x}|Y}(x). \tag{7.80}$$

References: [4], [25], [24].

7.5.4 Background

Now we consider rank-one matrix inference. The question is statistical: we only observe a noisy version of a rank-one matrix. Can we recover information about it? This has already been solved by other people. The first that gave a complete solution of this problem is [20] and then [2] gave another proof.

We consider the problem in [25] using the approach in [24].

Problem: Let $\bar{x} = (\bar{x}_1, \dots, \bar{x}_N)$ be a vector of bounded independent random variables with law $P_N =$

$P^{\otimes N}$. For some $t > 0$, we observe

$$Y = \sqrt{\frac{2t}{N}}\bar{x}\,\bar{x}^\top + W \tag{7.81}$$

where $W = (W_{ij})$ are independent standard Gaussians. (We don't assume symmetry.)

Our main goal is to understand the large N behavior of minimum mean square error

$$\mathrm{mmse}_N(t) = \mathbb{E}\left[|\bar{x} - \mathbb{E}[\bar{x}|Y]|^2\right]. \tag{7.82}$$

Informally,

$$\mathbb{P}(\bar{x} = x \text{ and } Y = y) = dP_N(x)\,dy\,\exp\left(-\frac{1}{2}\left|Y - \sqrt{\frac{2t}{N}}xx^\top\right|^2\right).$$

Now compute the conditional probability,

$$\mathbb{P}(\bar{x} = x|Y) = \frac{\exp\left(-\frac{1}{2}\left|Y - \sqrt{\frac{2t}{N}}xx^\top\right|^2\right)\,dP_N(x)}{\int \exp\left(-\frac{1}{2}\left|Y - \sqrt{\frac{2t}{N}}x'x'^\top\right|^2\right)\,dP_N(x')}$$

Expand the exponent and remove $-\frac{1}{2}|Y|^2$ term, which doesn't depend on t, x to get the following.

Define

$$H_N^\circ(t,x) = \sqrt{\frac{2t}{N}}Y\cdot xx^\top - \frac{t}{N}|xx^\top|^2. \tag{7.83}$$

Substituting (7.81) and expanding (7.83) gives

$$H_N^\circ(t,x) = \sqrt{\frac{2t}{N}}x\cdot Wx + \frac{2t}{N}(x\cdot\bar{x})^2 - \frac{t}{N}|x|^4. \tag{7.84}$$

The first term is the important term, which is like the spin glass model. In fact, the conditional law of $\bar{x}|Y$ has the form of a Gibbs measure, with the most important part looking like the spin glass model.

We have

$$\mathbb{E}[f(\bar{x})|Y] = \frac{\int f(x)\exp(H_N^\circ(t,x))\,dP_N(x)}{\int \exp(H_N^\circ(t,x))\,dP_N(x)}. \tag{7.85}$$

The important difference with the Curie-Weiss model is that $H_N^\circ(t,x)$ is still random.

Denote

$$\langle f(x)\rangle := \int f(x)\,dP_{\bar{x}|Y}(x) = \frac{\int f(x)\exp(H_N^\circ(t,x))\,dP_N(x)}{\int \exp(H_N^\circ(t,x))\,dP_N(x)}. \tag{7.86}$$

So we have $\langle f(x)\rangle = \mathbb{E}[f(\bar{x})|Y]$.

In general, we have

$$\mathbb{E}\langle f(x)\rangle = \mathbb{E}[f(\bar{x})], \tag{7.87}$$

and

$$\begin{aligned}\mathbb{E}\langle f(x)g(x')\rangle &= \mathbb{E}[\langle f(x)\rangle\langle g(x)\rangle] && (7.88)\\ &= \mathbb{E}[\langle f(x)\rangle\,\mathbb{E}[g(\bar{x})|Y]] && (7.89)\\ &= \mathbb{E}[\mathbb{E}[\langle f(x)g(\bar{x})\rangle\,|Y]] && (7.90)\\ &= \mathbb{E}\langle f(x)g(\bar{x})\rangle && (7.91)\end{aligned}$$

where x' is an independent copy of x under $\langle\cdot\rangle$.

Proposition 7.5.3 (Nishimori indentity)**.** For any bounded measurable function f, we have

$$\mathbb{E}\langle f(x,x')\rangle = \mathbb{E}\langle f(x,\bar{x})\rangle$$
$$\mathbb{E}\langle f(x,x',x'')\rangle = \mathbb{E}\langle f(x,x',\bar{x})\rangle.$$

Similar identities hold with more replicas.

Proof. By a monotone class argument, it suffices to verify the claim for f of the form

$$f(x, x') = g_1(x)g_2(x')$$

for measurable functions g_1 and g_2.

Note that by (7.91) we have

$$\mathbb{E}\left\langle g_1(x)g_2(x')\right\rangle = \mathbb{E}\left\langle g_1(x)g_2(\bar{x})\right\rangle .$$

Other indentities derived in the same way. We complete the proof. □

Recall that we are interested in the mmse defined as in (7.82). Note that

$$mmse_N(t) = \mathbb{E}[|x - \langle x\rangle|^2] = \mathbb{E}[|\bar{x}|^2 - 2\mathbb{E}\left\langle \bar{x}\cdot x\right\rangle + \mathbb{E}\left\langle x\cdot x'\right\rangle] = \mathbb{E}\left[|\bar{x}|^2\right] - \mathbb{E}\left\langle x\cdot \bar{x}\right\rangle$$

where the last indetity from Proposition 7.5.3. Also, we have

$$\mathbb{E}\left[|\bar{x}|^2\right] = \sum_{i=1}^{N}\mathbb{E}\,\bar{x}_i^2 = N\mathbb{E}\bar{x}_1^2,$$

So we want to understand the asymptotic of $\mathbb{E}\left\langle x\cdot\bar{x}\right\rangle$. In analogy with the Curie-Weiss model, we are interested in the free energy

$$F_N^\circ(t) = \frac{1}{N}\log\int \exp(H_N^\circ(t,x))\,dP_N(x), \tag{7.92}$$

and

$$\overline{F}_N^\circ(t) = \mathbb{E}F_N^\circ(t). \tag{7.93}$$

As will be seen shortly, we have

$$\partial_t\overline{F}_N^\circ(t) = \frac{1}{N^2}\mathbb{E}\left\langle (x\cdot\bar{x})^2\right\rangle \tag{7.94}$$

and $\overline{F}_N^\circ$ should be easier to study than its derivative.

In the context of inference models, a last requirement is that we do not want to destroy the fact that the Gibbs measure is a conditional expectation, since the Nishimori indentity will be essential. So we propose the following: given a parameter $h \geqslant 0$, we observe that

$$\widetilde{Y} = \sqrt{2h}\bar{x} + z$$

where $z = (z_1, z_2, \ldots, z_N)$ is a vector of independent standard Gaussians.

So in total ,we observe $\mathcal{Y} = (Y, \widetilde{Y})$. A similar computtaion as before

$$\mathbb{E}[f(\bar{x})|\mathcal{Y}] = \frac{\int f(x)\exp(H_N(t,h,x))\,dP_N(x)}{\int \exp(H_N(t,h,x))\,dP_N(x)} \tag{7.95}$$

where

$$H_N(t,h,x) = H_N^\circ(t,x) + \sqrt{2h}\widetilde{Y}\cdot x - h|x|^2. \tag{7.96}$$

Substituting (7.84) and expand (7.96) gives

$$\begin{aligned} H_N(t,h,x) &= \sqrt{\frac{2t}{N}}x\cdot Wx + \frac{2t}{N}(x\cdot\bar{x})^2 - \frac{t}{N}|x|^4 + \sqrt{2h}\widetilde{Y}\cdot x - h|x|^2 \\ &= \underbrace{\sqrt{\frac{2t}{N}}x\cdot Wx}_{\text{spin-glass type}} + \frac{2t}{N}(x\cdot\bar{x})^2 - \frac{t}{N}|x|^4 + \underbrace{\sqrt{2h}x\cdot z}_{\text{spin-glass type}} + 2hx\cdot\bar{x} - h|x|^2. \end{aligned}$$

From now on, $\langle\cdot\rangle$ is the Gibbs measure w.r.t. $H_N(t,h,\cdot)$. Now, one may ask why we use the normalizations $\sqrt{2t}$ and $\sqrt{2h}$ in place of t and h. The reason is that $\sqrt{t}$ × 'standard Gaussian' G is natural (think of Brownian motion B_t where $\sqrt{2t}G$ is like B_{2t}).

The main ingredient is **Itô calculus without Itô**. Let G be a standard gaussian, then from Itô calculus we have

$$\mathbb{E}[\exp\Big(\sqrt{2t}G - t\Big)] = 1.$$

Differentiating with respect to t we should get 0, that is,

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{2t}}G-1\right)\exp\left(\sqrt{2t}G-t\right)\right]=0.$$

How to see this without Itô calculus? Indeed, apply integration by parts,

$$\int x\exp\left(\sqrt{2t}x-t\right)e^{-x^2/2}\,dx=\int\sqrt{2t}\exp\left(\sqrt{2t}x-t\right)e^{-x^2/2}\,dx.$$

So for a standard Gaussian G and any $f\in C_c^\infty$,

$$\mathbb{E}[Gf(G)]=\mathbb{E}[f'(G)]. \tag{7.97}$$

Lemma 7.5.4 (Gaussian integration by parts)**.** Let F be a bounded measurable function. Recall that z is the standard Gaussian noise in $\widetilde{Y}$. We have

$$\mathbb{E}\left\langle z\cdot F(x,\bar{x})\right\rangle=\sqrt{2h}\mathbb{E}\left\langle (x-x')\cdot F(x,\bar{x})\right\rangle \tag{7.98}$$

$$\mathbb{E}\left\langle z\cdot F(x,x',\bar{x})\right\rangle=\sqrt{2h}\mathbb{E}\left\langle (x+x'-2x'')\cdot F(x,x',\bar{x})\right\rangle \tag{7.99}$$

$$\mathbb{E}\left\langle (z\cdot F(x,\bar{x}))^2\right\rangle=\sqrt{2h}\mathbb{E}\left\langle ((x-x')\cdot F(x,\bar{x}))(z\cdot F(x,\bar{x}))\right\rangle+\mathbb{E}\left\langle |F(x,\bar{x})|^2\right\rangle \tag{7.100}$$

Proof. For each $i\in\{1,2,\dots,N\}$,

$$\begin{aligned}
\mathbb{E}\left\langle z_i\cdot F_i(x,\bar{x})\right\rangle&=\mathbb{E}\left[\frac{z_i\int F_i(x,\bar{x})\exp(H_N(t,h,x))\,dP_N(x)}{\int\exp(H_N(t,h,x))\,dP_N(x)}\right]\\
&=\mathbb{E}\left[\partial_{z_i}\left(\frac{\int F_i(x,\bar{x})\exp(H_N(t,h,x))\,dP_N(x)}{\int\exp(H_N(t,h,x))\,dP_N(x)}\right)\right]\\
&=\mathbb{E}\left\langle\sqrt{2h}x_iF_i(x,\bar{x})\right\rangle-\mathbb{E}\left[\left\langle F_i(x,\bar{x})\right\rangle\left\langle\sqrt{2h}x_i\right\rangle\right]\\
&=\mathbb{E}\left\langle\sqrt{2h}x_iF_i(x,\bar{x})\right\rangle-\mathbb{E}\left\langle\sqrt{2h}x_i'F_i(x,\bar{x})\right\rangle\\
&=\sqrt{2h}\mathbb{E}\left\langle(x_i-x_i')F_i(x,\bar{x})\right\rangle
\end{aligned}$$

where the second equality from (7.97) and the forth equality from (7.91). Summing over i, we obtain the result.

The second statement is very similar as the first one.

Now for the last statement, for $i\neq j$

$$\begin{aligned}
\mathbb{E}[z_iz_j\left\langle F_i(x,\bar{x})F_j(x,\bar{x})\right\rangle]&=\mathbb{E}[z_i\partial_{z_j}\left\langle F_i(x,\bar{x})F_j(x,\bar{x})\right\rangle]\\
&=\sqrt{2h}\mathbb{E}[z_i(x_i-x_i')\left\langle F_i(x,\bar{x})F_j(x,\bar{x})\right\rangle]
\end{aligned}$$

by the first statement.

For $i=j$, we get

$$\mathbb{E}[z_iz_j\left\langle F_i(x,\bar{x})F_j(x,\bar{x})]\right\rangle=\mathbb{E}\left\langle F_i^2(x,\bar{x})\right\rangle.$$

To sum up, this gives us the desired result. □

We are now ready to compute the derivatives of the free energy

$$F_N(t,j):=\frac{1}{N}\log\int\exp(H_N(t,h,x))dP_N(x),$$

and the derivatives of their expectations

$$\bar{F}_N(t,h):=\mathbb{E}\left[F_N(t,h)\right],\quad t\geqslant 0\text{ and }h\geqslant 0.$$

We have

$$\partial_hF_N=\frac{1}{N}\left\langle\frac{1}{\sqrt{2h}}x\cdot z+2x\cdot\bar{x}-|x|^2\right\rangle. \tag{7.101}$$

Taking expectation, we obtain

$$\begin{aligned}\partial_h \bar{F}_N &= \frac{1}{N}\mathbb{E}\left[\left\langle -x\cdot x' + 2x\cdot\bar{x} - |x|^2\right\rangle\right] && \text{(Gaussian integartion by parts)} \\ &= \frac{1}{N}\mathbb{E}\left\langle x\cdot\bar{x}\right\rangle && \text{(Nishimori identity).} \qquad (7.102)\end{aligned}$$

Next, we have

$$\partial_t F_N = \frac{1}{N}\left\langle \frac{1}{\sqrt{2tN}}W\cdot xx^* + \frac{2}{N}xx^*\cdot\bar{x}\bar{x}^* - \frac{1}{N}|xx^*|^2\right\rangle \tag{7.103}$$

Then by a similar computation as in (7.102), we have

$$\partial_t \bar{F}_N = \frac{1}{N^2}\mathbb{E}\left[\langle -xx^*\cdot x'x'^* + 2xx^*\cdot\bar{x}\bar{x}^*\rangle\right] = \frac{1}{N^2}\mathbb{E}\left[\langle (x\cdot\bar{x})^2\rangle\right]. \tag{7.104}$$

Thus, we derive that

$$\partial_t \bar{F}_N - (\partial_h \bar{F}_N)^2 = \mathrm{Var}\left(\frac{x\cdot\bar{x}}{N}\right) \tag{7.105}$$

which resembles (7.44) appeared in the context of Curie–Weiss model. If we take $P_N = P_1^{\otimes N}$ as we did for the Curie–Weiss model, then

$$\begin{aligned}F_N(0,h) &= \frac{1}{N}\mathbb{E}\left[\log\int \exp\left(\sqrt{2h}x\cdot z - 2hx\cdot\bar{x} + h|x|^2\right)\right] dP_N(x) && (7.106)\\ &= \mathbb{E}\left[\log\int \exp\left(\sqrt{2h}x_1 z_1 - 2hx_1\bar{x}_1 + hx_1^2\right)\right] dP_1(x) && (7.107)\\ &= F_1(0,h). && (7.108)\end{aligned}$$

Remark 7.5.5. One technical difference from the Curie–Weiss model is that we restrict ourselves to $h \geqslant 0$ here.

In order to use the convex selection principle, we have to check that $\bar{F}_N$ is convex in (t,h). Recall that by (7.102), we have

$$N\partial_h \bar{F}_N = \mathbb{E}\langle x\cdot\bar{x}\rangle = \mathbb{E}\left[\frac{\int x\cdot\bar{x}\exp(H_N(t,h,x))dP_N(x)}{\int \exp(H_N(t,h,x))dP_N(x))}\right].$$

Next,

$$\begin{aligned}N\partial_h^2 \bar{F}_N &= \mathbb{E}\left[\left\langle (x\cdot\bar{x})\left(\frac{1}{\sqrt{2h}}x\cdot z + 2x\cdot\bar{x} - |x|^2\right)\right\rangle - \langle x\cdot\bar{x}\rangle\left\langle \frac{1}{\sqrt{2h}}x\cdot z + 2x\cdot\bar{x} - |x|^2\right\rangle\right] \\ &= \mathbb{E}\left[\langle (x\cdot\bar{x})(x\cdot(x-x') + 2x\cdot\bar{x} - |x|^2)\rangle\right] - \mathbb{E}\left[(x\cdot\bar{x})\left(\frac{1}{\sqrt{2h}}x'\cdot z + 2x'\cdot\bar{x} - |x'|^2\right)\right], \qquad (7.109)\end{aligned}$$

where (7.109) is by applying the Gaussian integration formula. Continuing (7.109), we have

$$\begin{aligned}&= \mathbb{E}\left[(x\cdot\bar{x})(2x\cdot\bar{x} - x\cdot x')\right] - \mathbb{E}\left[\langle (x\cdot\bar{x})(x'\cdot(x+x'-2x'') + 2x'\cdot\bar{x} - |x'|^2)\rangle\right] \\ &= 2\mathbb{E}\langle (x\cdot\bar{x})^2\rangle - 4\mathbb{E}\langle (x\cdot\bar{x})(x\cdot x')\rangle + 2\mathbb{E}\langle (x\cdot\bar{x})(x'\cdot x'')\rangle \\ &= 2\mathbb{E}\left[|\langle xx^*\rangle - \langle x\rangle\langle x^*\rangle|^2\right] \geqslant 0.\end{aligned}$$

The joint convexity of F_N can be obtained in the same way, and we leave it as an exercise (see Exercise 3 in SPS8).

Theorem 7.5.6 (approximate Hamilton–Jacobi equation)**.** We have

$$\partial_h \bar{F}_N \geqslant 0 \tag{7.110}$$

and

$$0 \leqslant \partial_t \bar{F}_N - (\partial_h \bar{F}_N)^2 \leqslant \frac{1}{N}\partial_h^2 \bar{F}_N + \mathbb{E}\left[(\partial_h F_N - \partial_h \bar{F}_N)^2\right]. \tag{7.111}$$

Proof. We first show (7.110). By the Nishimori identity, we have

$$\partial_h \bar{F}_N = \frac{1}{N}\mathbb{E}\langle x \cdot \bar{x}\rangle = \frac{1}{N}\mathbb{E}\langle x \cdot x'\rangle = \mathbb{E}\left[\langle x\rangle \cdot \langle x\rangle\right] = \mathbb{E}\left[|\langle x\rangle|^2\right] \geqslant 0.$$

It remains to prove (7.111). By (7.105), we have

$$\partial_t \bar{F}_N - (\partial_h \bar{F}_N)^2 = \mathrm{Var}\left(\frac{x\cdot\bar{x}}{N}\right) = \frac{1}{N^2}\mathbb{E}\langle (x\cdot\bar{x} - \mathbb{E}\langle x\cdot\bar{x}\rangle)^2\rangle \geqslant 0. \tag{7.112}$$

We define

$$H'_N(h,x) := \frac{1}{\sqrt{2h}}x\cdot z + 2x\cdot\bar{x} - |x|^2. \tag{7.113}$$

Note that we have

$$\mathbb{E}\langle H'_N(h,x)\rangle = \mathbb{E}\langle x\cdot\bar{x}\rangle \quad \text{and} \quad \partial_h F_N = \frac{\langle H'_N\rangle}{N}. \tag{7.114}$$

We claim that

$$\mathbb{E}\langle (x\cdot\bar{x} - \mathbb{E}\langle x\cdot\bar{x}\rangle)^2\rangle \leqslant \mathbb{E}\langle (H'_N - \mathbb{E}\langle H'_N\rangle)^2\rangle - \frac{1}{2h}\mathbb{E}\left\langle |x|^2\right\rangle, \tag{7.115}$$

$$\mathbb{E}\langle (H'_N - \mathbb{E}\langle H'_N\rangle)^2\rangle - \frac{1}{2h}\mathbb{E}\left\langle |x|^2\right\rangle \leqslant N\partial_h^2\bar{F}_N + N^2\mathbb{E}\left[(\partial_h F_N - \partial_h\bar{F}_N)^2\right]. \tag{7.116}$$

We start with the proof of (7.115)

□

Comments on Theorem 7.5.6

We are in same position as for our analysis of the Curie–Weiss model, except for the extra term

$$\mathbb{E}\left[(\partial_h F_N - \partial_h \bar{F}_N)^2\right],$$

and for the fact that our domain is now restricted to $\{h \geqslant 0\}$. In SPS9, one can show that for every $M < \infty$, there exists a constant $C < \infty$ such that

$$\mathbb{E}\left[\sup_{[0,M]^2}(F_N - \bar{F}_N)^2\right] \leqslant CN^{-1/3}, \tag{7.117}$$

where the argument is based on classical concentration inequalities. In fact, one can improve (7.117) to the following upper bound. For every $p < \infty$, ε, and $M < \infty$, there exists a constant $C_{p,\varepsilon,M} < \infty$ such that

$$\mathbb{E}\left[\sup_{[0,M]^2}(F_N - \bar{F}_N)^p\right]^{1/p} \leqslant C_{p,\varepsilon,M}N^{-\frac{1}{2}+\varepsilon}. \tag{7.118}$$

However, note that what we need is a concentration inequality for $\partial_h F_N$, not the function itself. In fact, one can construct Curie–Weiss type examples for which the variance of $\partial_h F_N$

$$\mathbb{E}\left[(\partial_h F_N - \partial_h \bar{F}_N)^2\right]$$

is large at some special points, for example, the points where the limit has corners. Thus, similar to the term

$$\frac{1}{N}\partial_h^2 \bar{F}_N,$$

we can only hope to assert that the variance of $\partial_h F_N$ is small at most points.

In view of our Convex Selection Principle, we are interested in showing the following:

Proposition 7.5.7. Let f be any subsequential limit of $\overline{F}_N$, $t, h > 0$, and $\phi \in C_0^\infty$ be such that $f - \phi$ has a strict local maximum at (t,h). [Recall that F_N is precompact by Arzelà-Ascoli Theorem]. We then have

$$\left(\partial_t\phi - (\partial_h\phi)^2\right)(t,h) = 0.$$

Proof. (We omit to denote the subsequence for convenience). Let $t, h > 0$ and $\phi \in C^\infty$ be as in the statement. There exist $(t_N, h_N) \to (t,h)$ such that for N sufficiently large, the function $\overline{F}_N - \phi$ has a local maximum at (t_N, h_N). Then $\partial_h^2 \left(\overline{F}_N - \phi\right)(t_N, h_N) \le 0$. We first show that for every $|h'| \le C^{-1}$,

$$0 \le \overline{F}_N(t_N, h_N + h') - \overline{F}_N(t_N, h_N) - h'\partial_h \overline{F}_N(t_N, h_N) \le C|h'|^2. \tag{7.119}$$

Note that the lower bound is by convexity of $\overline{F}_N$. By Taylor's Formula,

$$\overline{F}_N(t_N, h_N + h') - \overline{F}_N(t_N, h_N) = h'\partial_h \overline{F}_N(t_N, h_N) + \int_0^{h'} (h' - u)\partial_h^2 \overline{F}_N(t_N, h_N + u)\ \mathrm{d}u \tag{7.120}$$

and we also have

$$\phi(t_N, h_N + h') - \phi(t_N, h_N) \geqslant \overline{F}_N(t_N, h_N + h') - \overline{F}_N(t_N, h_N).$$

For N sufficiently large, we have $t_N > 0,\ h_N > 0$, so

$$\partial_t(\overline{F} - \phi)(t_N, h_N) = 0 = \partial_h(\overline{F}_N - \phi)(t_N, h_N).$$

[Notice that we can replace $\overline{F}_N$ by ϕ in (7.120)]. Combining all this, we get

$$\int_0^{h'} (h' - u)\partial_h^2 \overline{F}_N(t_N, h_N + u)\ \mathrm{d}u \le \int_0^{h'} (h' - u)\partial_h^2 \phi(t_N, h_N + u)\ \mathrm{d}u \le C|h'|^2.$$

Using (7.120) once more, we get (7.119). In particular, $\partial_h^2 \overline{F}_N(t_N, h_N) \le C$. [This is obvious since $\partial_h^2(\overline{F}_N - \phi)(t_N, h_N) \le 0$].

We now aim to control $\mathbb{E}\left[\left(\partial_h F_N - \partial_h \overline{F}_N\right)^2\right]$. We want to leverage on the semiconvexity of F_N. For every $|h'| \le C^{-1}$,

$$\partial_h^2 F_N(t_N, h_N + h') \geqslant -C\frac{|z|}{\sqrt{N}}, \text{ so}$$

$$F_N(t_N, h_N + h') \geqslant F_N(t_N, h_N) + h'\partial_h F_N(t_N, h_N) - C|h'|^2\frac{|z|}{\sqrt{N}}.$$

Combining with (7.119), we get, for every $|h'| \le C^{-1}$,

$$h'(\partial_h F_N - \partial_h \overline{F}_N)(t_N, h_N) \le 2\sup_V \left|F_N - \overline{F}_N\right| + C|h'|^2\left(1 + \frac{|z|}{\sqrt{N}}\right),$$

where V is some neighborhood of (t, h). For some deterministic $\lambda \in [0, C^{-1}]$ to be chosen later, we fix

$h' := \lambda \frac{\partial_h F_N - \partial_h \overline{F}_N}{|\partial_h F_N - \partial_h \overline{F}_N|}(t_N, h_N)$, so that

$\lambda \left|\partial_h F_N - \partial_h \overline{F}_N\right|(t_N, h_N) \le 2\sup_V \left|F_N - \overline{F}_N\right| + C\lambda^2\left(1 + \frac{|z|}{\sqrt{N}}\right).$

Squaring and taking the expectation,

$\lambda^2 \mathbb{E}\left[\left(\partial_h F_N - \partial_h \overline{F}_N\right)^2(t_N, h_N)\right] \le 8\mathbb{E}\left[\sup_V \left(F_N - \overline{F}_N\right)^2\right] + C\lambda^4.$

But recall that $\mathbb{E}\left[\sup_V \left(F_N - \overline{F}_N\right)^2\right] \le CN^{-\frac{1}{3}}$. Then fixing $\lambda^4 = N^{-\frac{1}{3}}$, we get

$\mathbb{E}\left[\left(\partial_h F_N - \partial_h \overline{F}_N\right)^2(t_N, h_N)\right] \le CN^{-\frac{1}{6}}.$

Together with the approximate Hamilton-Jacobi equation, this yields

$$0 \le \left(\partial_t \overline{F}_N - (\partial_h \overline{F}_N)^2\right)(t_N, h_N) = \left(\partial_t \phi - (\partial_h \phi)^2\right)(t_N, h_N) \le \frac{C}{N} + \frac{C}{N^{\frac{1}{6}}}.$$

Since ϕ is smooth, we conclude that $\left(\partial_t \phi - (\partial_h \phi)^2\right)(t, h) = 0$, as desired. □

The only thing left to do is address the fact that our system is only defined for $h \geqslant 0$, as opposed to every $h \in \mathbb{R}$ in the case of the Curie-Weiss model. Here we can solve this easily by using a symmetrization argument, i.e., extending the function by reflection, as follows. [Recall that $\partial_h \overline{F}_N \geqslant 0$].

If $f : \mathbb{R}_+^2 \to \mathbb{R}$ is convex and increasing in h, then the mapping

$$\begin{cases} \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R} \\ (t, h) \mapsto f(t, |h|) \end{cases}$$

is convex. In order to see this, for every $\alpha \in [0, 1]$, $t, t' \geqslant 0$, and $h, h' \in \mathbb{R}$, we verify that

$$f(\alpha t + (1 - \alpha)t', |\alpha h + (1 - \alpha)h'|) \le \alpha f(t, |h|) + (1 - \alpha)f(t', |h'|).$$

Up to $(h,h') \leftarrow (-h,-h')$, nothing changes, so we can assume $\alpha h + (1-\alpha)h' \geqslant 0$. By symmetry, we can also assume that $h \leq h'$. If $0 \leq h \leq h'$, then we are okay, since f is convex. The remaining case is if $h \leq 0 \leq h'$. Then,

$$f(\alpha t + (1-\alpha)t', \alpha h + (1-\alpha)h') \leq \\ f(\alpha t + (1-\alpha)t', \alpha|h| + (1-\alpha)h') \leq \alpha f(t,|h|) + (1-\alpha) f(t',h'),$$

by monotonicity and convexity.

So after we have performed this extension of $\overline{F}_N$, we are in position to apply the Convex Selection Principle. We obtain the following:

Theorem 7.5.8. The function $\overline{F}_N$ converges to the unique viscosity solution to

$$\begin{cases} \partial_t f - (\partial_h f)^2 = 0 & \text{on } \mathbb{R}_+ \times \mathbb{R} \\ f(0,h) = \overline{F}_1(0,|h|) =: \psi(h) & (h \in \mathbb{R}) \end{cases}$$

By the Hopf-Lax formula, for every $t, h \geqslant 0$,

$$\lim_{N\to\infty} \overline{F}_N(t,h) = \sup_{h' \geqslant 0} \left(\psi(h') - \frac{(h'-h)^2}{4t} \right) = f(t,h).$$

[When $h \geqslant 0$, it suffices to take the sup over $h' \geqslant 0$].

As in the Curie-Weiss model, we can verify that if $\mathbb{E}[\overline{x}_1] = 0$, then $\psi(h) \underset{h\to 0}{\sim} Ch^2$. Indeed, recall that $\partial_h \overline{F}_1(0,0) = \mathbb{E}\left[\langle x_1 \rangle^2\right] = \mathbb{E}[\overline{x}_1]$, the last equality following from the fact that there is no randomness, we are simply sampling according to P_1. And from there, we find that, for some $t_C \in (0,\infty)$,

$$\begin{cases} \partial_h f(t,0) = 0 & \text{for } t < t_C \\ \partial_h^+ f(t,0) = \partial_h^- f(t,0) > 0 & \text{for } t > t_C \end{cases}.$$

Recall also that if f is differentiable in h at (t,h), then

$$\partial_h f(t,h) = \lim_{N\to\infty} \partial_h \overline{F}_N(t,h) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\langle x \cdot \overline{x} \rangle\right].$$

And similarly for ∂_t we have

$$\begin{cases} \partial_t f(t,0) = 0 & \text{if } t < t_C \\ \partial_t f(t,0) > 0 & \text{if } t > t_C \text{ and } f \text{ differentiable in } t \text{ at } (t,0) \end{cases}.$$

And if f is differentiable in t at $(t,0)$, then

$$\partial_t f(t,0) = \lim_{N\to\infty} \frac{1}{N^2} \mathbb{E}\left\langle (x\cdot\overline{x})^2 \right\rangle = \lim_{N\to\infty} \frac{1}{N^2} \mathbb{E}\left\langle x\, x^* \cdot \overline{x}\, \overline{x}^* \right\rangle = \lim_{N\to\infty} \partial_t \overline{F}_N(t,0).$$

So phase transition for the possiblity to detect “a proportion of the signal” as t increases. This can be identified numerically rather straightforwardly from the Hopf or Hopf-Lax formulas.

To conclude, let us say a few words about more general models. If we think again about the community detection problem, we may want to represent the different community belongings of a given individual as a vector of fixed length K. Then $\overline{x} \in \mathbb{R}^{N\times K}$, and we may want to stick with independent, identically-distributed rows. We observe a noisy version of $\overline{x}^{\otimes 2} A$, for a given matrix $A \in \mathbb{R}^{K^2 \times L}$, where L is another fixed integer, maybe corresponding to the number of different networks we observe, and where $\overline{x}^{\otimes 2} \in \mathbb{R}^{N^2 \times K^2}$ is given by $\left(\overline{x}^{\otimes 2}\right)_{(i,j),(k,l)} = \overline{x}_{ik}\, \overline{x}_{jl}$.

Even more generally, we can take an arbitrary integer, $p \geqslant 1$ and assume that we observe $\overline{x}^{\otimes p} A$ for some $A \in \mathbb{R}^{K^p \times L}$. In this case, not much changes. The nonlinearity, say H, that appears in the equation is not necessarily convex, so our approximate Hamilton-Jacobi equation looks like

$$\left|\partial_t \overline{F}_N - H\left(\nabla \overline{F}_N\right)\right| \leq \frac{C}{N} \Delta \overline{F}_N + C\mathbb{E}\left[\left|\nabla F_N - \nabla \overline{F}_N\right|^2\right].$$

This is essentially like for the generalized Curie-Weiss model, so no problem. The only extra difficulty is that the extra variable h is now a positive, semi-definite matrix, so we cannot do the "cheap symmetrization trick" to revert to an equation posed on the full space. There are a few technicalities related to managing the boundary condition (of Neumann type), but overall not much changes, and in particular, we still get the Hopf formula. [See [4] for more on this].

7.6 Spin Glasses

Scribe: Qiang Wu.

7.6.1 Overview

We start with a brief introduction to some common mean field spin glass models.

Sherrington-Kirkpatrick Model

This model is the most classical mean field spin glass model, it was defined on a complete graph with i.i.d disorder couplings. Fix integer N, for a configuration $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N) \in \{-1, +1\}^N$, the Hamiltonian is given by

$$H_N(\boldsymbol{\sigma}) = \frac{1}{\sqrt{N}} \sum_{i,j=1}^{N} J_{i,j}\sigma_i\sigma_j \tag{7.121}$$

where the disorder coupling $J_{i,j}$ are inpendent standard Gaussians. This model has been understood very well. The limiting free energy was computed in various ways, and the most famous approach is the Parisi formula, which can be used to computed limit of free energy at all temperatures. For the proof about Parisi formula, see [28], and more classical monographs [32, 33].

Bipartite Model

This model has a different structure with the classical SK model, where the spins are now coded into two layers, and the interactions only happen between layers. For simplicity, take a configuration $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \in \mathbb{R}^N \times \mathbb{R}^N$, the Hamiltonian is given by

$$H_N(\boldsymbol{\sigma}) = \frac{1}{\sqrt{N}} \sum_{i,j=1}^{N} J_{i,j}\sigma_{1,i}\sigma_{2,j} \tag{7.122}$$

This model is only partially understood in the high temperature regime, see [8]. Note that in both examples, the Hamiltonian are just some Gaussian random fields indexed by appropriate configurations, so it's useful to look at the covariance structure of $H_N(\boldsymbol{\sigma})$. For SK model, with configuration $\boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbb{R}^N$,

$$\mathbb{E}[H_N(\boldsymbol{\sigma})H_N(\boldsymbol{\tau})] = NR^2_{\boldsymbol{\sigma},\boldsymbol{\tau}} \tag{7.123}$$

where $R_{\boldsymbol{\sigma},\boldsymbol{\tau}} := \frac{1}{N}\sum_{i=1}^{N} \sigma_i\tau_i$ is known as overlap in spin glass literature, and the above expectation $\mathbb{E}$ is with respect to the disorder randomness in J. Similarly, for bipartite model,

$$\mathbb{E}[H_N(\boldsymbol{\sigma})H_N(\boldsymbol{\tau})] = N\frac{\boldsymbol{\sigma}_1 \cdot \boldsymbol{\tau}_1}{N}\frac{\boldsymbol{\sigma}_2 \cdot \boldsymbol{\tau}_2}{N} \tag{7.124}$$

Here we remark that the covariance structure uniquely characterize the spin glass models, for SK model, whose covariance is given by usual quadratic function. There are some other general spin glasses, such as mixed p-spin models, the covariance is given as the polynomials up to order p of overlaps. In general, the discussions in this chapter can be applied to the following more general setting, for $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \cdots, \boldsymbol{\sigma}_D) \in \mathcal{H}_N^D$, where D is an fixed integer, and $\mathcal{H}_N^D$ is a Hilbert space, and for some smooth funciton $\xi : \mathbb{R}^{D\times D} \to \mathbb{R}$, the covariance of Hamiltonian is

$$\mathbb{E}[H_N(\boldsymbol{\sigma})H_N(\boldsymbol{\tau})] = N\xi\left(\left(\frac{\boldsymbol{\sigma}_d \cdot \boldsymbol{\tau}'_d}{N}\right)_{1\le d,d'\le D}\right)$$

This is known as the vector spin glass model. The structure function ξ will appear later in the Hamilton-Jacobi setting.

Before we turn to next section, we will denote the reference measure on the configuration space as P_N. In SK model where the configuration space is a product space, the measure will be simply a product measure of 1 dimensional cases, that is, $P_N = P_1^{\otimes N}$. Similarly for bipartite model, the reference measure is $P_N = \pi_1^{\otimes N} \otimes \pi_2^{\otimes N}$, where π_1, π_2 supported on $\{-1, +1\}$.

Our goal is to study the free energy:

$$\frac{1}{N}\log\int \exp\Big(\sqrt{2t}H_N(\boldsymbol{\sigma}) - Nt\Big)dP_N(\boldsymbol{\sigma}) \tag{7.125}$$

Remark 7.6.1. Note that $\mathbb{E}\exp\big(\sqrt{2t}H_N(\boldsymbol{\sigma}) - Nt\big) = 1$, where $\mathbb{E}$ is with respect to the disorder randomness in Hamiltonian. It forms a mean 1 martingale. One can also regard the Hamiltonian as a Brownian motion and understand it from the stochastic analysis perspective.

7.6.2 Associated Hamilton-Jacobi equation

To find the associated HJ PDE, one has to enrich the corresponding free energy in a propery way. For SK model, we first introduce:

$$F_N(t,h) := -\frac{1}{N}\log\int \exp\Big(\sqrt{2t}H_N(\boldsymbol{\sigma}) - Nt + \sqrt{2h}\boldsymbol{z}\cdot\boldsymbol{\sigma} - Nh\Big)dP_N(\boldsymbol{\sigma}) \tag{7.126}$$

where $\boldsymbol{z}$ is an N-dimenional vector with independent standard Gaussian entries. Recall h plays the role of magentic field like in Curie-Weiss model. We write

$$\bar{F}_N := \mathbb{E}[F_N].$$

Lemma 7.6.2 (Gaussian Integration by Parts)**.** Let Σ be a finite set, $(x(\boldsymbol{\sigma}), y(\boldsymbol{\sigma}))_{\boldsymbol{\sigma}\in\Sigma}$ be a centered Gaussian field, and for every $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \Sigma$, define

$$C(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = \mathbb{E}[x(\boldsymbol{\sigma})y(\boldsymbol{\sigma}')],$$

Let P be a probablity measure on Σ, and define the Gibbs measure

$$\langle f(\boldsymbol{\sigma})\rangle := \frac{\int f(\boldsymbol{\sigma})\exp(y(\boldsymbol{\sigma}))dP(\boldsymbol{\sigma})}{\int \exp(y(\boldsymbol{\sigma}))dP(\boldsymbol{\sigma})}, \tag{7.127}$$

then we have

$$\mathbb{E}\langle x(\boldsymbol{\sigma})\rangle = \mathbb{E}\langle C(\boldsymbol{\sigma},\boldsymbol{\sigma}) - C(\boldsymbol{\sigma},\boldsymbol{\sigma}')\rangle. \tag{7.128}$$

Proof of Lemma 7.6.2 . The proof is similar as we did in the statistical inference case, and also one can try to find the solution in the problem set 11. □

Now let's compute the derivative of $\bar{F}_N$.

$$\begin{aligned}\partial_h \bar{F}_N &= -\frac{1}{N}\mathbb{E}\left\langle \frac{1}{\sqrt{2h}}\boldsymbol{z}\cdot\boldsymbol{\sigma} - N\right\rangle \\ &= -\frac{1}{N}\mathbb{E}\left\langle |\boldsymbol{\sigma}|^2 - \boldsymbol{\sigma}\cdot\boldsymbol{\sigma}' - N\right\rangle \\ &= \mathbb{E}\left\langle \frac{\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'}{N}\right\rangle\end{aligned}$$

The 2nd step is due to the Gaussian integration by parts. Next let's compute

$$\partial_t \bar{F}_N = -\frac{1}{N}\mathbb{E}\left\langle \frac{1}{\sqrt{2t}}H_N(\boldsymbol{\sigma}) - N \right\rangle$$
$$= \mathbb{E}\left\langle \left(\frac{\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'}{N}\right)^2 \right\rangle$$

For the bipartite model, one can compute this similarly, but the free energy will be enriched in a bit different way, more specifically enrich it in each layer as in SK model. So in that case, there will be h_1, h_2 in the corresponding formula (7.126). From those derivatives, it seems that the t-derivative and h-derivative are related.

For SK model, it suggests to consider the following HJ PDE:

$$\partial_t \bar{F}_N - (\partial_h \bar{F}_N)^2 = \text{Var}\left(\frac{\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'}{N}\right)$$

While for bipartite model,

$$\partial_t \bar{F}_N - \partial_{h_1}\bar{F}_N\partial_{h_2}\bar{F}_N = \mathbb{E}\left\langle \left(\frac{\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_1'}{N} - \mathbb{E}\left\langle\frac{\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_1'}{N}\right\rangle\right)\left(\frac{\boldsymbol{\sigma}_2\cdot\boldsymbol{\sigma}_2'}{N} - \mathbb{E}\left\langle\frac{\boldsymbol{\sigma}_2\cdot\boldsymbol{\sigma}_2'}{N}\right\rangle\right)\right\rangle$$
$$\left|\partial_t \bar{F}_N - \partial_{h_1}\bar{F}_N\partial_{h_2}\bar{F}_N\right| \le \frac{1}{2}\text{Var}\left(\frac{\boldsymbol{\sigma}_1\cdot\boldsymbol{\sigma}_1'}{N}\right) + \frac{1}{2}\text{Var}\left(\frac{\boldsymbol{\sigma}_2\cdot\boldsymbol{\sigma}_2'}{N}\right)$$

In the more generic setting with the structure function ξ, we should have a similar form,

$$\left|\partial_t \bar{F}_N - \xi(\nabla \bar{F}_N)\right| \le \text{Var}(\cdots)$$

Note that the convexity of function ξ is important, since it allows us to appeal to the Hopf-Lax formula to represent the solution of the PDE variationally. For SK model, this is true, but for bipartite model, $\xi(x_1, x_2) = x_1x_2$, which is not convex. We further point out that there are 2 main differences w.r.t previous models:

- $\bar{F}_N$ is neither convex nor concave (see excerise 3 in problem set 11).
- The variances of $\frac{\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'}{N}$ are not trivial as t becomes large. We need to refine the enrichment of the free energy.

The enrichment will be encoded by probability measures, for SK model, h will be repaced by a probability measures supported on the positive half line. Actually this is originated from some deep implications about how the Gibbs measure behaves when t is large. In next lecture, we will introduce the so-called Poisson-Dirichelet Cascades, which actully is the limit of the SK model in some sense. When t is small, the variance is trivial. This is because the limiting Gibbs measure is also trivial which concentrates around some constant involving h, but for large t, it has a nontrivial distribution. That's why we need to replace h by some probability measures.

7.6.3 Extreme values for i.i.d random variables

We start as some classical extreme value theory for i.i.d random random vraiables. Let $\zeta > 0$, and $(X_m)_{m\in\mathbb{N}}$ are i.i.d random variable such that $\mathbb{P}[X_n \geqslant y] \sim \frac{1}{y^\zeta}$ for $y \to \infty$, then

$$\mathbb{P}[\max_{1\le i\le n} X_i \le n^{1/\zeta}y] \sim (1 - \frac{1}{ny^\zeta})^n \to \exp\left(-\frac{1}{y^\zeta}\right)$$

as $n \to \infty$. This basically tells us that if the random variables has the above tail decay, then we can normalize the maximum by n^ζ such that it converges in distribution. It's more interesting to understand the extremal process for this collections of random variables. Actually another classical result says $\forall 0 < a < b$, one would have

$$\#\left\{i \leqslant n : \frac{X_i}{n^{1/\zeta}} \in (a,b)\right\} \to \text{Poisson}\left(\int_a^b \frac{\zeta}{x^{1+\zeta}} dx\right)$$

In SK spin glass model, the Hamiltonians can be regrad as some Gaussian random variables with hypercube indices, but they are not independent. One simplification of the SK model is known as Random Energy Model. It is a collection of i.i.d standard Gaussian random variables $\{X_i\}_{1\le i\le 2^N}$. Intutitively the gap between the 1st maximum and 2nd maximum is roughyly $\frac{1}{\sqrt{N}}$. Then we transform X_i by the function

$$\{\exp\left(\beta\sqrt{N}X_i\right)\}_{1\le i\le 2^N}$$

This transform is basically doing some "stretching" on $\{X_i\}$, and now the extremal process looks like the original extremal process with $\zeta = \frac{\sqrt{s\log 2}}{\beta}$. One thing to remark is that that this random energy model behaves in a similar fashion as SK model in the high temperature regime. But in low temperature regime, it's more complicated, the Gibbs measure is hierarchically structured, and the limiting object is more complicated than the Poisson point process, it will become some probbaility cascades.

7.6.4 Poisson-Dirichlet Process

We first recall the definition of the Poisson point process(PPP),

Definition 7.6.3. Let μ be a measure on $\mathbb{R}^d$ without atoms. We say that a random discrete set $\Pi \subset \mathbb{R}^d$ is a Poisson point process with intensity measure μ if

- For any $A_1, \cdots, A_m \subset \mathbb{R}^d$ pairwise disjoint, measurable, the random variables $(|\Pi \cap A_i|)_{1\le i\le n}$ are independent
- For any $A \subset \mathbb{R}^d$ measurable, the law of $|\Pi \cap A|$ is Poisson$(\mu(A))$

Here we introduce a important theorem about PPP.

Theorem 7.6.4. If $\Pi \sim PPP(\mu)$, and $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ measurable function, then

$$f(\Pi) \sim PPP(f(\mu)) \tag{7.129}$$

provided $f(\mu)$ has no atoms.

We will mainly interested in the PPP with intensity measures like $\frac{\zeta}{x^{\zeta+1}}dx1_{x>0}$, for $\zeta \in (0,1)$. Notice that for every $a > 0$ with $\int_a^\infty \frac{\zeta}{x^{\zeta+1}}dx < \infty$. By using different values of a, we can find the largest and 2nd largest values etc. Therefore, one can represent Π as a decreasing sequence. Denote $\Pi = \{u_n, n \geqslant 1\}$ with $u_1 \geqslant u_2 \geqslant \cdots u_n \geqslant \cdots$. Notice also that $\sum_{x\in\Pi} = \sum_{n=1}^\infty u_n < \infty a.s.$. Then

$$\mathbb{E}[\sum_{x\in\Pi} x1_{x\le 1}] = \int_0^1 x\frac{\zeta}{x^{1+\zeta}}dx < \infty$$

The above expectation is finite can be seen from the fact the there are only finitely many very large terms, and summing with many small terms is still finite.

Because of some universality attracted by the PPP, there should be some sort "stability" properties. We will introduce this in the following sense.

Theorem 7.6.5 (Stability Property)**.** Let $(X_n, Y_n)_{n\geqslant 1}$ be i.i.d random vectors with values in $(0,\infty) \times \mathbb{R}^d$, and ν_ζ be the law of Y_1 size biased by X_1^ζ:

$$\nu_\zeta(A) := \frac{\mathbb{E}[X_1^\zeta 1_{Y_1 \in A}]}{\mathbb{E}[X_1^\zeta]}.$$

The PPP $\{(u_n X_n, Y_n); n \geqslant 1\}$ has the same law as $\{(\mathbb{E}[X_1^\zeta]^{1/\zeta} u_n, Y_n'; n \geqslant 1\}$, where Y_n' are i.i.d, and independent of (u_n) with law ν_ζ.

Proof of Theorem 7.6.5.

Let $\hat{\Pi}$ be the PPP with intensity measure $\frac{\zeta}{x^\zeta} dx \otimes P_{(X_1,Y_1)}$, then the first PPP in the statement can be realized as the image of $\hat{\Pi}$ under the mapping $(u,(x,y)) \mapsto (ux, y)$. By using the Theorem 7.6.4, this point process is indeed Poisson and we can compute the corresponding intensity measure. Similarly for the second point processs in the statement. For a more detailed proof, see the handwritten notes or Theorem 2.4 in [28]. □

Corollary 7.6.6. With same notations as above, we have

$$\mathbb{E} \log \sum_{m=1}^{\infty} X_n u_n = \mathbb{E} \log \sum_{n=1}^{\infty} u_n + \frac{1}{\zeta} \log \mathbb{E}[X_1^\zeta] \tag{7.130}$$

Recall the free energy we discussed is very similar to the left handside.

Definition 7.6.7 (Poisson-Dirichlet Process)**.** Let $\zeta \in (0,1)$, $\Pi \sim PPP(\frac{\zeta}{x^{1+\zeta}} dx 1_{x>0})$, and enumerate it in a decreasing way, $\Pi = \{u_n, n \leq 1\}, u_1 \geqslant u_2 \geqslant \cdots$, the random set $\{v_n, n \geqslant 1\}$ with

$$v_n := \frac{u_n}{\sum_{k=1}^\infty u_k}$$

is a Poisson-Dirichlet process with parameter ζ.

Note we already check the denominator is finite, so the above object is well-defined. We also point out the following fact, which will be used to represent the free energy of SK model.

$$\mathbb{E} \log \sum X_n v_n = \frac{1}{\zeta} \log \mathbb{E}[X_1^\zeta].$$

Proposition 7.6.8. Let $\mathbf{v} := (v_n)_{n \geqslant 1}$ be a P.D.(ζ), and conditionally on $\mathbf{v}$. Let α, α' be independent random variable with law $\sum_{n=1}^\infty v_n \delta_n$. We write $\langle \cdot \rangle$ for the law of α, α' given $\mathbf{v}$. Then

$$\mathbb{E}_{\mathbf{v}} \left\langle 1_{\{\alpha = \alpha'\}} \right\rangle = 1 - \zeta \tag{7.131}$$

Proof. Let g_n be independent standard Gaussians, for $t \in \mathbb{R}$, consider

$$\left(u_n \exp\left(t(g_n - \frac{t\zeta}{2}) \right), g_n - t\zeta \right)_{n \geqslant 1}$$

To use the stability property above, we know that $\exp\left(t(g_n - \frac{t\zeta}{2})\right)$ and $g_n - t\zeta$ play the role of X_n, Y_n respectively in the Theorem 7.6.5. By that theorem, it has the same law as

$$(u_n, g_n)_{n \geqslant 1}$$

Define $v_n^t := \frac{u_n \exp\left(t(g_n - \frac{t\zeta}{2})\right)}{\sum_{k=1}^\infty u_k \exp\left(t(g_k - \frac{t\zeta}{2})\right)}$, it has the same law as v_n up to reordering. Using the stability result, we know

$$v_n^t = \frac{u_n \exp(t g_n)}{\sum_{k=1}^\infty u_k \exp(t g_k)},$$

With this result, then $(v_n^t, g_n - t\zeta)_{n \geqslant 1}$ and $(v_n, g_n)_{n \geqslant 1}$ have the same law. Let $\langle \cdot \rangle_t$ be the Gibbs measure w.r.t v_n^t such that $\left\langle 1_{\{\alpha = n\}} \right\rangle_t = v_n^t$. We have $\mathbb{E} \langle g_\alpha \rangle_0 = 0$ Since $t = 0$ implies that there is no correlation between the Gibbs weights and the Gaussians g_n. On the other hand

$$\begin{aligned}\mathbb{E}\left\langle g_\alpha\right\rangle_0 &= \mathbb{E}\left\langle g_\alpha - t\zeta\right\rangle_t \\ &= t\mathbb{E}\left\langle 1 - 1_{\{\alpha=\alpha'\}} - \zeta\right\rangle_t \\ &= t\mathbb{E}\left\langle 1 - \zeta - 1_{\{\alpha=\alpha'\}}\right\rangle_0 = 0\end{aligned}$$

where the first and third equality is due to the invariance principle we discussed. Second step is by Gaussin integration by parts. □

7.7 Poisson-Dirichlet Cascades

We now introduce Poisson-Dirichlet cascades. These are constructed iteratively using Poission-Dirichlet processes as building blocks. Roughly, we do the following:

$$\begin{aligned}&\text{For a fixed } K \geqslant 1 \text{ and } 0 = \zeta_0 < \zeta_1 < \cdots < \zeta_K < \zeta_{K+1} = 1, \\ &(u_n)_{n\geqslant 1} \text{ P.p.p.} \left(\frac{\zeta_1}{x^{\zeta_1+1}\mathbf{1}_{\{x>0\}}}\mathrm{d}x\right), \\ &(u_{1n})_{n\geqslant 1}, (u_{2n})_{n\geqslant 1}, \cdots \text{ independent P.p.p.} \left(\frac{\zeta_2}{x^{\zeta_2+1}\mathbf{1}_{\{x>0\}}}\mathrm{d}x\right), \\ &\cdots \text{ etc., independently on each branch, up to depth } K.\end{aligned}$$

We get weights which can be normalized to form a probability measure. The normalized weights are the Poission-Dirichlet cascade associated with the weights

$$0 = \zeta_0 < \zeta_1 < \cdots < \zeta_K < \zeta_{K+1} = 1.$$

We want to retain the whole hierarchical decomposition of the process, not just the law of the normalized weights. We need some notation. The tree is encoded by

$$\mathcal{A} := \mathbb{N}^0 \cup \mathbb{N}^1 \cup \cdots \cup \mathbb{N}^K,$$

where $\mathbb{N}^0 = \{\phi\}$ and ϕ denotes the root of $\mathcal{A}$, as in the Ulam-Harris tree.

For each $k \in \{0, \cdots, K-1\}$, and $\alpha \in \mathbb{N}^k$, we give ourselves an independent Poisson point process $(u_{\alpha n})_{n\geqslant 1}$ of intensity $\frac{\zeta_{k+1}}{x^{\zeta_{k+1}+1}}\mathbf{1}_{\{x>0\}}\mathrm{d}x$.
For $\ell \leq k \in \{0, \cdots, K\}$ and $\alpha = (n_1, \cdots, n_k) \in \mathbb{N}^k$, we write

$$\alpha|_\ell := (n_1, \cdots, n_\ell).$$

For every $\alpha \in \mathbb{N}^k$, we set

$$\omega_\alpha := \prod_{\ell=1}^k u_{\alpha|_\ell}.$$

The weights that we aim to normalize are the $(\omega_\alpha)_{\alpha\in\mathbb{N}^K}$.

Proposition 7.7.1. With probability 1, we have $\sum\limits_{\alpha\in\mathbb{N}^K} \omega_\alpha < \infty$.

Proof. Clearly, we are ok if $K = 1$. Otherwise, for every $\alpha \in \mathcal{A} \setminus \mathbb{N}^K$, we write $U_\alpha := \sum\limits_{n\in\mathbb{N}} u_{\alpha n}$, so that $\sum\limits_{\alpha\in\mathbb{N}^K} \omega_\alpha = \sum\limits_{\alpha\in\mathbb{N}^{K-1}} \omega_\alpha U_\alpha = \sum\limits_{\alpha\in\mathbb{N}^{K-2}} \omega_\alpha \sum\limits_{n\in\mathbb{N}} u_{\alpha n} U_{\alpha n}$. For each $\alpha \in \mathbb{N}^{K-2}$, $\{u_{\alpha n}U_{\alpha n}, n \in \mathbb{N}\}$ has the same law as $\left\{\mathbb{E}\left[\left(U_{\alpha 1}^{\zeta_{K-1}}\right)^{\frac{1}{\zeta_{K-1}}}\right] u_{\alpha n}, n \geqslant 1\right\}$ and $U_{\alpha 1} = \sum\limits_{n\in\mathbb{N}} u_{\alpha 1 n}$, each of which is P.p.p. $\left(\frac{\zeta_K}{x^{\zeta_K+1}\mathbf{1}_{\{x>0\}}}\mathrm{d}x\right)$. So $\mathbb{E}\left[U_{\alpha 1}^{\zeta_{K-1}}\right] < \infty$. [See Exercise 2 in Problem Set 12]. Then we are left with investigating the finiteness of

$\sum\limits_{\substack{\alpha\in\mathbb{N}^{K-2}\\ n\in\mathbb{N}}} \omega_\alpha u_{\alpha n} = \sum\limits_{\alpha\in\mathbb{N}^{K-1}} \omega_\alpha$. By induction, we are done. □

Definition 7.7.2. Recall that $K \geqslant 1$, $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_K < \zeta_{K+1} = 1$, and we constructed the random weights $(\omega_\alpha)_{\alpha\in\mathbb{N}^K}$. The Poisson-Dirichlet cascade associated with these parameters is the family of normalized weights $\left(\frac{\omega_\alpha}{\sum\limits_{\beta\in\mathbb{N}^K}\omega_\beta}\right)_{\alpha\in\mathbb{N}^K} =: (v_\alpha)_{\alpha\in\mathbb{N}^K}$.

Remark 7.7.3. One can show that $\left\{v_\alpha, \alpha \in \mathbb{N}^K\right\}$ has the law of a Poisson-Dirichlet process with parameter ζ_K. The point of the construction is to provide a hierarchical decomposition of this process, which we can think of as some form of infinite divisibility.

7.8 Enriched Free Energy

We now proceed to define a suitable enriched free energy for the SK and bipartite models. We give ourselves $K \geqslant 1$, $(\zeta_k)_{1\leq k\leq K}$ as above, $(v_\alpha)_{\alpha\in\mathbb{N}^K}$, the associated Poisson-Dirichlet cascade, and $0 = q_{-1} \leq q_0 < q_1 < \cdots < q_K < q_{K+1} = \infty$, and we denote

$$\mu := \sum_{k=0}^{K} (\zeta_{k+1} - \zeta_k)\, \delta_{q_k}.$$

We also give ourselves $(z_\alpha)_{\alpha\in\mathcal{A}}$, independent N-dimensional standard Gaussian random variables, independent of the rest, and set, for every $\alpha \in \mathbb{N}^K$,

$$Z_q(\alpha) := \sum_{k=0}^{K} (2q_k - 2q_{k-1})^{\frac{1}{2}}\, z_{\alpha|_k}.$$

[Note that our naive initial attempt in (7.126), with the random magnetic field $\sqrt{2h}z$, corresponds to the choice of $\mu = \delta_h$.]

This is our refined random field. It is such that, for every α, $\beta \in \mathbb{N}^K$,

$$\mathbb{E}\left[Z_q(\alpha)\cdot Z_q(\beta)\right] = 2Nq_{\alpha\wedge\beta},$$

where $\alpha\wedge\beta := \sup\{k : \alpha|_k = \beta|_k\}$. We define

$$F_N(t,\mu) := -\frac{1}{N}\log\int \sum_{\alpha\in\mathbb{N}^K} \exp\left(\sqrt{2t}H_N(\sigma) - Nt + Z_q(\alpha)\cdot\sigma - Nq_K\right) v_\alpha \,\mathrm{d}P_N(\sigma)$$

and $\overline{F}_N(t,\mu) := \mathbb{E}\left[F_N(t,\mu)\right]$. The associated Gibbs measure is $\langle\,\cdot\,\rangle$, with random variables (σ,α). Independent copies of (σ,α) under $\langle\,\cdot\,\rangle$ are denoted (σ',α'), (σ'',α''), etc. We still have $\partial_t\overline{F}_N = \mathbb{E}\left\langle\left(\frac{\sigma\cdot\sigma'}{N}\right)^2\right\rangle$.

Now, for $k \in \{0,\cdots,K-1\}$,

$$\begin{aligned}
\partial_{q_k}\overline{F}_N &= -\frac{1}{N}\mathbb{E}\left\langle\left(2_{q_k} - 2_{q_{k-1}}\right)^{-\frac{1}{2}} z_{\alpha|_k}\cdot\sigma - \left(2_{q_{k+1}} - 2_{q_k}\right)^{-\frac{1}{2}} z_{\alpha|_{k+1}}\cdot\sigma\right\rangle \\
&= \frac{1}{N}\mathbb{E}\left\langle\left(\mathbf{1}_{\{\alpha|_k=\alpha'|_k\}} - \mathbf{1}_{\{\alpha|_{k+1}=\alpha'|_{k+1}\}}\right)\sigma\cdot\sigma'\right\rangle \\
&= \mathbb{E}\left\langle\frac{\sigma\cdot\sigma'}{N}\mathbf{1}_{\{\alpha\wedge\alpha'=k\}}\right\rangle.
\end{aligned}$$

Moreover, arguing as for Poisson-Dirichlet processes, one can show that

$\mathbb{E}\left\langle\mathbf{1}_{\{\alpha|_k=\alpha'|_k\}}\right\rangle = 1-\zeta_k$, that is, $\mathbb{E}\left\langle\mathbf{1}_{\{\alpha\wedge\alpha'=k\}}\right\rangle = \zeta_{k+1}-\zeta_k$.

So $(\zeta_{k+1}-\zeta_k)^{-1}\,\partial_{q_k}\overline{F}_N = \mathbb{E}\left\langle\frac{\sigma\cdot\sigma'}{N}\middle|\alpha\wedge\alpha' = k\right\rangle$, where the conditional expectation is with respect to $\mathbb{E}\langle\,\cdot\,\rangle$.

Recall also that $(\zeta_{k+1}-\zeta_k)$ is the weight given to δ_{q_k} in the definition of μ. We will see shortly that it makes sense to think of $(\zeta_{k+1}-\zeta_k)^{-1}\,\partial_{q_k}\overline{F}_N(t,\mu)$ as a transport-type derivative, which we denote by $\partial_\mu\overline{F}_N(t,\mu,q_k)$. Instead of our error term being the variance of $\frac{\sigma\cdot\sigma'}{N}$, we can make it be the conditional variance given $\alpha\wedge\alpha'$:

$$\begin{aligned}
&\mathbb{E}\left\langle\left(\frac{\sigma\cdot\sigma'}{N}-\mathbb{E}\left\langle\frac{\sigma\cdot\sigma'}{N}\middle|\alpha\wedge\alpha'\right\rangle\right)^2\right\rangle\\
&=\mathbb{E}\left\langle\left(\frac{\sigma\cdot\sigma'}{N}\right)^2\right\rangle-\mathbb{E}\left\langle\left(\mathbb{E}\left\langle\frac{\sigma\cdot\sigma'}{N}\middle|\alpha\wedge\alpha'\right\rangle\right)^2\right\rangle\\
&=\partial_t\overline{F}_N-\mathbb{E}\left\langle\sum_{k=1}^K\mathbf{1}_{\{\alpha\wedge\alpha'=k\}}\left(\mathbb{E}\left\langle\frac{\sigma\cdot\sigma'}{N}\middle|\alpha\wedge\alpha'=k\right\rangle\right)^2\right\rangle\\
&=\partial_t\overline{F}_N-\sum_{k=1}^K(\zeta_{k+1}-\zeta_k)\left(\partial_\mu\overline{F}_N(t,\mu,q_k)\right)^2.
\end{aligned}$$

We have shown that

$$\partial_t\overline{F}_N-\int\left(\partial_\mu\overline{F}_N\right)^2\mathrm{d}\mu=\mathbb{E}\left\langle\left(\frac{\sigma\cdot\sigma'}{N}-\mathbb{E}\left\langle\frac{\sigma\cdot\sigma'}{N}\middle|\alpha\wedge\alpha'\right\rangle\right)^2\right\rangle,$$

where $\partial_t\overline{F}_N-\int\left(\partial_\mu\overline{F}_N\right)^2\mathrm{d}\mu$ is computed implicitly at (t,μ). The integral can be written more explicitly as

$$\int\left(\partial_\mu\overline{F}_N\right)^2\mathrm{d}\mu=\int\left(\partial_\mu\overline{F}_N(t,\mu,x)\right)^2\mathrm{d}\mu(x).$$

We now motivate the notation

$$\partial\mu\overline{F}_N(t,\mu,q_k)=(\zeta_{k+1}-\zeta_k)^{-1}\,\partial_{q_k}\overline{F}_N(t,\mu).$$

This relates to optimal transport. For two probability measures μ, ν on $\mathbb{R}$, an optimal coupling between μ and ν can easily be realized as follows:

We take $U\sim\mathrm{Unif}([0,1])$ and then $X_\mu:=F_\mu^{-1}(U)$, $X_\nu:=F_\nu^{-1}(U)$, with $F_\mu^{-1}(u)=\inf\{s\geqslant 0:\mu([0,s])\geqslant u\}$, $u\in[0,1]$. For a "smooth" function g on the space of probability measures, we would then want that, as $\nu\to\mu$,

$$g(\nu)=g(\mu)+\mathbb{E}\left[\partial_\mu g\left(\mu,X_\mu\right)\left(X_\nu-X_\mu\right)\right]+o\left(\mathbb{E}\left[\left|X_\nu-X_\mu\right|^2\right]^{\frac{1}{2}}\right).$$

If we specialize this to measures with fixed (ζ_k)'s and moving (q_k)'s, we find that $\partial_\mu g\left(q_k\right)$ should be as defined above. In practice, we do not need to give a more precise sense to this derivative, since we can approximate any measure by a sum of Dirac masses, and use the "explicit" definition of ∂_μ in this case. This is made possible by the Lipschitz continuity of $\overline{F}_N$.

Chapter 8

AMP

References:

- [9] https://arxiv.org/pdf/2105.02180

Chapter 9

Interesting papers

- Replica method and random matrices (I)-(II)
- [22] [1]

[1] http://www.stat.ucla.edu/~ywu/research/documents/BOOKS/MontanariInformationPhysicsComputation.pdf

Bibliography

[1] Jean Barbier. “High-dimensional inference: a statistical mechanics perspective”. In: *arXiv preprint arXiv:2010.14863* (2020).

[2] Jean Barbier, Nicolas Macris. “The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference”. In: *Probability Theory and Related Fields* 174.3-4 (2019), pp. 1133–1185. URL: https://arxiv.org/abs/1705.02780.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Hong-Bin Chen, Jean-Christophe Mourrat, Jiaming Xia. “Statistical inference of finite-rank tensors”. In: *arXiv preprint arXiv:2104.05360* (2021).

[5] Lenaic Chizat, Francis Bach. “On the global convergence of gradient descent for over-parameterized models using optimal transport”. In: *arXiv preprint arXiv:1805.09545* (2018). URL: https://arxiv.org/pdf/1805.09545.

[6] Marco Cuturi. “Sinkhorn distances: Lightspeed computation of optimal transport”. In: *Advances in neural information processing systems* 26 (2013), pp. 2292–2300. URL: http://papers.neurips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf.

[7] Amir Dembo, Ofer Zeitouni. *Large deviations techniques and applications*. Vol. 38. Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2010, pp. xvi+396. ISBN: 978-3-642-03310-0. DOI: 10.1007/978-3-642-03311-7. URL: https://doi.org/10.1007/978-3-642-03311-7.

[8] Partha S. Dey, Qiang Wu. “Fluctuation results for Multi-species Sherrington-Kirkpatrick model in the replica symmetric regime”. In: (Preprint, arXiv:2012.13381). eprint: arXiv:2012.13381.

[9] Oliver Y Feng et al. “A unifying tutorial on Approximate Message Passing”. In: *arXiv preprint arXiv:2105.02180* (2021).

[10] Sacha Friedli, Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017.

[11] Stephen M Goldfeld, Richard E Quandt, Hale F Trotter. “Maximization by quadratic hill-climbing”. In: *Econometrica: Journal of the Econometric Society* (1966), pp. 541–551.

[12] Ziv Goldfeld, Kristjan Greenewald. “Gaussian-smoothed optimal transport: Metric structure and statistical efficiency”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3327–3337. URL: http://proceedings.mlr.press/v108/goldfeld20a/goldfeld20a.pdf.

[13] Trevor. Hastie. *The elements of statistical learning data mining, inference, and prediction*. eng. 2nd ed. Springer series in statistics. New York: Springer, 2009. ISBN: 9780387848587.

[14] Jan-Christian Hütter, Philippe Rigollet. “Minimax estimation of smooth optimal transport maps”. In: *The Annals of Statistics* 49.2 (2021), pp. 1166–1194. URL: https://arxiv.org/pdf/1905.05828.

[15] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.

[16] R.W. Keener. *Theoretical Statistics: Topics for a Core Course.* Springer Texts in Statistics. Springer New York, 2010. ISBN: 9780387938394. URL: https://books.google.co.in/books?id=aVJmcega44cC.

[17] Ivan Kobyzev, Simon Prince, Marcus Brubaker. "Normalizing flows: An introduction and review of current methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). URL: https://arxiv.org/pdf/1908.09257.

[18] Zhifeng Kong, Kamalika Chaudhuri. "The expressive power of a class of normalizing flow models". In: *International Conference on Artificial Intelligence and Statistics.* PMLR. 2020, pp. 3599–3609. URL: http://proceedings.mlr.press/v108/kong20a/kong20a.pdf.

[19] Flavien Léger. "A gradient descent perspective on Sinkhorn". In: *Applied Mathematics & Optimization* (2020), pp. 1–13. URL: https://link.springer.com/content/pdf/10.1007/s00245-020-09697-w.pdf.

[20] Marc Lelarge, Léo Miolane. "Fundamental limits of symmetric low-rank matrix estimation". In: *Probability Theory and Related Fields* 173.3-4 (2019), pp. 859–929. URL: https://arxiv.org/abs/1611.03888.

[21] Youssef Marzouk et al. "An introduction to sampling via measure transport". In: *arXiv preprint arXiv:1602.05023* (2016). URL: https://arxiv.org/pdf/1602.05023.

[22] Marc Mezard, Andrea Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

[23] Andrea Montanari. *Mean field asymptotics in high-dimensional statistics: A few references.* 2020.

[24] J-C Mourrat. "Hamilton–Jacobi equations for finite-rank matrix inference". In: *Annals of Applied Probability* 30.5 (2020), pp. 2234–2260.

[25] Jean-Christophe Mourrat. "Hamilton–Jacobi equations for mean-field disordered systems". In: *Annales Henri Lebesgue* 4 (2021), pp. 453–484.

[26] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction.* MIT Press, 2022. URL: probml.ai.

[27] Victor M Panaretos, Yoav Zemel. "Statistical aspects of Wasserstein distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431. URL: https://arxiv.org/pdf/1806.05500.

[28] Dmitry Panchenko. *The Sherrington-Kirkpatrick model.* Springer Monographs in Mathematics. Springer, New York, 2013, pp. xii+156. ISBN: 978-1-4614-6288-0; 978-1-4614-6289-7. DOI: 10.1007/978-1-4614-6289-7. URL: https://doi-org.proxy2.library.illinois.edu/10.1007/978-1-4614-6289-7.

[29] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63 (2015), p. 94. URL: http://math.univ-lyon1.fr/~santambrogio/OTAM-cvgmt.pdf.

[30] Jun Shao. *Mathematical Statistics.* 2nd. Springer-Verlag New York Inc, 2003.

[31] Alessio Spantini, Daniele Bigoni, Youssef Marzouk. "Inference via low-dimensional couplings". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2639–2709. URL: https://www.jmlr.org/papers/volume19/17-747/17-747.pdf.

[32] Michel Talagrand. *Mean field models for spin glasses. Volume I.* Vol. 54. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Basic examples. Springer-Verlag, Berlin, 2011, pp. xviii+485. ISBN: 978-3-642-15201-6. DOI: 10.1007/978-3-642-15202-3. URL: https://doi-org.proxy2.library.illinois.edu/10.1007/978-3-642-15202-3.

[33] Michel Talagrand. *Mean field models for spin glasses. Volume II.* Vol. 55. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Advanced replica-symmetry and low temperature. Springer, Heidelberg, 2011, pp. xii+629. ISBN: 978-3-642-22252-8; 978-3-642-22253-5.

[34] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108627771.