Lecture Notes on Probabilistic Learning and Inference

Qiang Liu
UT Austin
lqiang@cs.utexas.edu

Fall 2018

Contents

1 Maximum Likelihood Estimation and KL Divergence **2**

2 Concentration Inequalities, Law of Large Numbers, Central Limit Theorem **5**

2.1 Central Limit Theorem 7

3 Asymptotics of Maximum Likelihood Estimator **10**

4 Expectation Maximization **15**

4.1 EM as KL Minimization 16

5 Integral Probability Measures (IPM) and GANs **19**

5.1 Integral Probability Metrics (IPM) 19

6 f-Divergence **22**

6.1 Dual Representation of f-Divergence 23

7 Reproducing Kernel Hilbert Space **27**

7.1 Bounded Evaluation Functional and Riesz Representation Theorem 30

7.2 Random Features 32

7.3 Nonparametric Learning and Finite Representer Theorem 33

7.4 Generalization and Rademacher Complexity 35

7.5 Maximum Mean Discrepancy 36

7.6 Empirical Estimation of MMD: U-Statistics and V-Statistics 38

7.7 Energy Distance 38

7.8 Applications 38

8 Bayesian Inference **39**

8.1 Bayesian Estimators and Admissibility 39

8.2 PAC Bayesian Bounds 43

9 Variational Inference Using Parameteric Families **47**

9.1 Black Box Optimization 50

9.1.1 Evolutionary Strategy 50

9.1.2 Policy Gradient for Reinforcement Learning 51

10 Stein Variational Methods **52**

10.1 Stein's Method: Overview 52

A Measure Theory, Probability Measures, Random Variables **55**

B Rademacher Complexity **58**

B.1 Sub-Gaussian Random Variables 61

C Convex Conjugate **62**

D Learning and Inference as Approximating Probabilities **64**

1 Maximum Likelihood Estimation and KL Divergence

Notation *Unless specified otherwise, all the distributions are assumed to be defined on $\mathcal{X} := \mathbb{R}^d$, where d denotes the dimension. To avoid measure theoretical terminologies, we represent all the distributions on $\mathbb{R}^d$ using their density functions; for distributions exists no density functions, Delta functions are used. For example, $q(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}_i)$ represents the atomic measure that concentrates probability around $\boldsymbol{x}_i$.*

Given data $\{\boldsymbol{x}_i\}$, we want to learn a distribution $p(\vec{x})$ to match the data as close as possible. Assume we already have a family of distribution that we want to select from:

$$\mathcal{P} = \{p_\theta(\boldsymbol{x}) \colon \ \theta \in \Theta\},$$

where this set of distributions is indexed by some parameter θ. An example is the family of Gaussian distributions, $\mathcal{P} = \{\mathcal{N}(\boldsymbol{x};\ \mu, \Sigma) \colon \mu \in \mathbb{R}^d,\ \ \Sigma \succ 0\}$, where $\theta = \{\mu,\ \Sigma\}$ is the parameter, consisting of both the means and variances.

The data can be equivalently presented using its empirical distribution:

$$\hat{q}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i).$$

As a general framework, we can formulate the search of the optimal p_θ that fits the data distribution q best as an optimization problem:

$$\min_\theta D(\hat{q}_n \,||\, p_\theta),$$

where D is some notion of "distance" or "divergence" measure, that satisfies $D(q \,||\, p) \geq 0$ and $p = q$ iff $D(q \,||\, p) = 0$.

The most basic way to achieve this is using Kullback-Leibler (KL) divergence: For any two distributions p and q, their KL divergence is defined as

$$\mathrm{KL}(q \,||\, p) = \mathbb{E}_q \left[\log \left(\frac{q(x)}{p(x)} \right) \right].$$

Theorem 1.1. *1) For any p and q, we have* $\mathrm{KL}(q \,||\, p) \geq 0$.

2) $\mathrm{KL}(q \,||\, p) = 0$, *if and only if* $q = p$.

Proof. Recall Jensen's inequality: if $h(z)$ is a convex function, we have $\mathbb{E}[h(z)] \geq h(\mathbb{E}[z])$, for any random variable z. If h is strictly convex, and $\mathbb{E}[h(z)] = h(\mathbb{E}[z])$, then z must be deterministic. Note that $-\log(x)$ is a strictly convex function.

Denote by S_q the support of q, that is, $S_q = \{x\colon\ q(x) > 0\}$. We have

$$\mathbb{E}_q \left[\frac{p(x)}{q(x)} \right] = \int_{S_q} q(x) \frac{p(x)}{q(x)} dx = \int_{S_q} p(x) dx = \mathbb{E}_p[\mathbb{I}(x \in S_q)] \leq 1.$$

Therefore, we have $\log(\mathbb{E}_q[p(x)/q(x)]) \leq 0$. Hence

$$\begin{aligned}
\mathrm{KL}(q \,||\, p) &= \mathbb{E}_q \left[-\log \left(\frac{p(x)}{q(x)} \right) \right] \\
&\geq \mathbb{E}_q \left[-\log \left(\frac{p(x)}{q(x)} \right) \right] + \log \left(\mathbb{E}_q \left[\frac{p(x)}{q(x)} \right] \right) \qquad \text{//Lemma 1.2} \\
&\geq 0 \qquad \text{//Jensen's inequality with } h(x) = -\log(x).
\end{aligned}$$

2) We just need to prove that $\mathrm{KL}(q \,||\, p) = 0$ implies $p = q$. Because $-\log$ is strictly convex, when $\mathrm{KL}(q \,||\, p) = 0$, we must have $p(x)/q(x) = c$ for some constant c. Hence, $\mathrm{KL}(q \,||\, p) = \mathbb{E}_q[-\log(p(x)/q(x))] = -\log c$. But $\mathrm{KL}(q \,||\, p) = 0$, we have $c = 1$. This suggests $p(x)/q(x) = 1$ for $x \sim S_q$, and there is no probability mass of p outside of S_q because $\int_{x \notin S_q} p(x)dx = \int_\Omega p(x)dx - \int_{S_q} p(x)dx = \int_\Omega p(x)dx - \int_{S_q} q(x)dx = 1 - 1 = 0$. This completes the proof. □

Problem 1.1. *1) The χ-square divergence between p and q is defined by* $\chi^2(q \,||\, p) = \mathbb{E}_p \left[\left(\frac{q(x)}{p(x)} - 1 \right)^2 \right]$.

Prove

$$\mathrm{KL}(q \,||\, p) \leq \chi^2(q \,||\, p).$$

2) Prove that $\mathrm{KL}(q \,||\, p)$ *is a convex function of (p, q), that is, for any (p_1, q_1) and (p_2, q_2) and $\alpha \in [0, 1]$, we have*

$$\mathrm{KL}(\alpha q_1 + (1-\alpha) q_2 \,||\, \alpha p_1 + (1-\alpha) p_2) \leq \alpha \mathrm{KL}(q_1 \,||\, p_1) + (1-\alpha) \mathrm{KL}(q_2 \,||\, p_2).$$

Remark p is said to be *absolutely continuous* w.r.t. q (denoted by $p \ll q$), if there exists a measurable function f, such that $p(x) = f(x)q(x)$. This is equivalent to that $q(x) = 0 \implies p(x) = 0$. When p is *absolutely continuity* w.r.t. Lebesgue measure, we simply say that p is *absolutely continuity*; these are distributions that have (non-delta) density functions.

The following result plays an important role for *importance sampling*, which we will discuss later.

n:pqmean}

Lemma 1.2. *If p is absolutely continuous w.r.t q, that is, $p \ll q$, we have*

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\right] = 1.$$

Proof. Because $p \ll q$, we can write $p(x) = f(x)q(x)$ for all $x \in \mathcal{X}$. We have

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\right] = \mathbb{E}_{x\sim q}\left[f(x)\right] = \int q(x)f(x)dx = \int p(x)dx = 1.$$

□

Maximum Likelihood When $\hat{q}_n$ is an empirical distribution of the dataset $\{\boldsymbol{x}_i\}$, we have $\mathrm{KL}(\hat{q}_n \,||\, p_\theta) = +\infty$ according the definition. However, it turns out that the "the infinite part is a constant" and it is still possible to minimize KL divergence in this case. Note that

$$\mathrm{KL}(q \,||\, p_\theta) = -\mathbb{H}[q] - \mathbb{E}_q[\log p_\theta],$$

where $\mathbb{H}[q] := -\mathbb{E}_q[\log q]$ is the entropy of q. When $q = \hat{q}_n$, we have $\mathbb{H}[q] = -\infty$ and hence causing $\mathrm{KL}(q \,||\, p) = +\infty$. However, observe that $\mathbb{H}[q]$ does not depend on the parameter θ we want to optimize. Minimizing KL divergence is hence equivalent to maximizing the negative of the second term:

$$\hat{\theta}_n := \arg\max_\theta \left\{ L(\theta) := \mathbb{E}_{\hat{q}_n}[\log p_\theta] = \frac{1}{n}\sum_{i=1}^n \log p_\theta(\boldsymbol{x}_i) \right\},$$

where $L(\theta)$ is called the log-likelihood function, and $\hat{\theta}_n$ is called the maximum likelihood estimator.

Problem 1.2. *Consider Gaussian family with*

$$p_\theta(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

where $\theta = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Given data $\{\boldsymbol{x}_i\}_{i=1}^n$, please derive that the maximum likelihood estimator of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\Sigma}$ is constrained to be in the set of positive definite matrices in the optimization.

Problem 1.3. *1) Assume the true data distribution q is the uniform distribution on interval $[-1, 1]$ and p_θ is Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, with $\theta = [\mu, \sigma]$, $\sigma \geq 0$. When given an iid observation $\{x_i\}_{i=1}^n \sim q$, please derive the MLE estimator $\hat{\theta}_n$. Calculate the limit $\lim_{n\to\infty} \theta_n$ as we have inifinite number of data.*

2) Assume the data distribution is q is standard normal $\mathcal{N}(0, 1)$ and set $p_\theta(x) = Uniform([a, b])$, where $\theta = [a, b]$ is the parameter. So this case, we want to find a best uniform distribution to find with data generated from Gaussian distribution. When given an iid observation $\{x_i\}_{i=1}^n \sim q$, please derive the MLE estimator $\hat{\theta}_n$. Calculate the limit $\lim_{n\to\infty} \theta_n$ as we have inifinite number of data.

2 Concentration Inequalities, Law of Large Numbers, Central Limit Theorem

Theorem 2.1 (Markov Inequality). *Let X be a non-negative random variable on $\mathbb{R}$ with finite mean $\mu := \mathbb{E}[X] < \infty$, and $a > 0$ is any constant, we have*

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. Define function $\sigma(X) = a\mathbb{I}(X \geq a)$. It is clear that $\sigma(X) \leq X$ for any X. We have

$$\mathbb{E}[X] \geq \mathbb{E}[\sigma(X)] = \mathbb{E}[a\mathbb{I}(X \geq a)] = a\Pr(X \geq a),$$

where we note that $\mathbb{E}[\mathbb{I}(X \geq a)] = \Pr(X \geq a)$. □

Theorem 2.2 (Chebyshev Inequality). *Let X be any random variable on $\mathbb{R}$ with finite mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[(X-\mu)^2]$. We have*

$$\Pr(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof. Take $|X - \mu|$ as a non-negative random variable and apply Markov inequality. □

Problem 2.1 (Tightness of Markov Inequality). *1) For a given $\mu, a \in \mathbb{R}_+$, solve the following optimization:*

$$\max_X \{ \Pr(X \geq a) \quad s.t. \quad \mathbb{E}[X] = \mu \},$$

where the optimization is overall all possible non-negative random variables.

2) Why is it necessary to require that X is non-negative?

3) Solve

$$\max_X \{\Pr(|X - \mu| \geq \epsilon) \quad s.t. \quad \mathbb{E}[X] = \mu, \quad \mathbb{E}[(X-\mu)^2] = \sigma^2, \}$$

where the optimization is overall all possible random variables on $\mathbb{R}$.

Definition 2.3 (Convergence of Random Variables). *A sequence of random variables $S_1, \ldots, S_n$ on $\mathbb{R}$ is said to **converge in probability** to a real number a if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr(|S_n - a| \geq \epsilon) \to 0.$$

This is denoted by $S_n \xrightarrow{p} a$.

*$S_1, \ldots, S_n$ is said to **converge almost surely** to a if*

$$\Pr(\lim_{n\to\infty} S_n = a) = 1.$$

This is denoted by $S_n \xrightarrow{a.s.} a$.

convergence almost surely implies Convergence in probability, but the other way is not always true.

Theorem 2.4 (Weak Law of Large Numbers (LLN) with Finite Variance). *Assume $X_1, \dots, X_n$ is a set of identical and independently (iid) random variables with finite mean μ and variance σ^2. Define*

$$S_n = \frac{1}{n}\left(X_1 + \cdots + X_n\right).$$

We have for any $\epsilon > 0$,

$$\Pr(|S_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Therefore, S_n converges to μ in probability.

Proof. Assume $\mu = 0$ without loss of generality. Note that

$$\begin{aligned}\mathrm{var}(S_n) &= \frac{1}{n^2}\mathbb{E}[(X_1 + \cdots + X_n)^2] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[X_i^2] \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Applying Chebyshev inequality gives the result. □

The most general form of law of large number theorem does not require to assume finite variance. The proof for the case of infinite variance can be done using characteristic function. Check yourself.

We also have strong LLN which establish the convergence of almost surely (see Section 1.7 [6]).

Theorem 2.5 (Strong Law of Large Numbers). *Let $X_1, X_2, \cdots$ be pairwise independent identically distributed random variables with $\mathbb{E}[|X_i|] < \infty$. Let $\mathbb{E}[X_i] = \mu$, and $S_n = (X_1 + \cdots + X_n)/n$. Then $S_n \xrightarrow{a.s.} \mu$ as $n \to \infty$.*

In many problems in statistical learning theory, we need uniform law of large numbers. See Lemma 2.4 of Newey and McFadden [17] and Theorem 2 of Jennrich [12].

iformlaw}

Theorem 2.6 (Uniform Law of Large Numbers). *Let $X_1, X_2, \dots$ are i.i.d. random variables. Let $\{g(x;\theta) : \theta \in \Theta\}$ is a family of functions such that*

1) Θ is a compact subset of a finite dimensional Euclidean space.

2) g is continuous in θ for each x.

3)We have $|g(x;\theta)| \leq M(x)$ for some M with $\mathbb{E}[M(X_1)] < \infty$.

Then we have

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n g(X_i;\ \theta) - g(X;\ \theta)\right| \xrightarrow{a.s.} 0.$$

Basically, the idea of proving the uniform law is by covering. If Θ contains only finite number of elements, then the result follows trivially. If Θ has infinite number of elements, we then approximately cover Θ with a finite subset.

todo Specifically, for any $\epsilon_0 > 0$, because Θ is compact, there exists a finite set $\{\theta_i\}$, such that for any $\theta \in \Theta$, there exists a θ_i, such that $||\theta - \theta_i|| \leq \epsilon$. If $\{g(x;\theta)\}$ is equicontinuous, in that we have $||g(x;\theta) - g(x;\theta')||_\infty \leq$ whenever $||\theta - \theta'|| \leq \delta_0$, then we can show the convergence by ...

Generally, A class of functions $\mathcal{F}$ is called a **Glivenko-Cantelli class** with respect to a probability measure P if

{equ:fsup}
$$\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{a.s.} 0. \tag{1}$$

where $X, X_1, \cdots$ are i.i.d from P. A function class is a **universal Glivenko-Cantelli class** if (1) holds for any measure P.

The trivial case when (1) is when number of functions in $\mathcal{F}$ is finite. Mathematical conditions of Glivenko-Cantelli class has been a central topic in statistics and learning theory, and closely relates to the key notion of overfiting and generalizability. For example, one fundamental result in learning theory is that if $\mathcal{F}$ has a finite Vapnik-Chervonenkis (VC) dimension, then it is a universal Glivenko-Cantelli class. See Vapnik [27], van de Geer [26].

2.1 Central Limit Theorem

Definition 2.7 (Convergence in Distribution). *1) A sequence $X_1, X_2, \cdots$ of real-valued random variables in $\mathbb{R}$ is said to **converge in distribution, or converge weakly, or converge in law** to a random variable X if*

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every number $x \in \mathbb{R}$ at which F is continuous. Here F_n and F are the cumulative distribution functions of random variables X_n and X, respectively.

2) In $\mathbb{R}^d$, weak convergence is defined by the following two equivalent statements:

a) $\lim_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ for all bounded, continuous functions.

b) $\lim_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ for all bounded, Lipschitz functions.

Theorem 2.8 (Central Limit Theorem). *Assume $X, X_1, \ldots, X_n$ is a set of identical and independently (iid) random variables with finite mean μ and variance σ^2. Define*

$$S_n = \frac{1}{n}\left(X_1 + \cdots + X_n\right).$$

We have

$$\sqrt{n}(S_n - \mu) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2),$$

where $\xrightarrow{d}$ denotes convergence in distribution.

Proof. Recall that the characteristic function of a random variable X is defined to be

$$\phi_X(t) = \mathbb{E}[\exp(itX)].$$

where i is the imaginary unit, recall that $\exp(iz) = \cos(z) + i\sin(z)$.

By Levy's continuity theorem (see Section 2.3, Durrett [6]), if Z is a random variable whose characteristic function $\phi_Z(t)$ is continuous at the origin $t = 0$, then S_n converges in distribution to S if and only if the sequence ϕ_{S_n} converges pointwise to ϕ_Z. Therefore, we just need to prove that ϕ_{S_n} converges to the characteristic function of normal distribution $Z \sim \mathcal{N}(0, \sigma^2)$, which equals $\phi_Z(t) = \exp(-\frac{\sigma^2 t^2}{2})$.

It is known that $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$ and $\phi_{aX}(t) = \phi_X(at)$. Without loss of generality, assume $\mu = 0$. We have

$$\phi_{S_n}(t) = \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{n}}\sum_{j=1}^n X_j\right)\right] = \prod_{j=1}^n \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{n}}X_j\right)\right] = \left(\phi_X\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

By Taylor expansion,

$$\begin{aligned}
\phi_X\left(\frac{t}{\sqrt{n}}\right) &= \mathbb{E}[\exp\left(it\sqrt{n}X\right)] \\
&= \mathbb{E}\left[1 + \frac{it}{\sqrt{n}}x + \frac{(itx)^2}{2n} + o\left(\frac{1}{n}\right)\right] \\
&= 1 + \frac{it}{\sqrt{n}}\mathbb{E}[X] - \frac{t^2}{2n}\mathbb{E}[X^2] + o\left(\frac{1}{n}\right) \\
&= 1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{1}{n}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty}\phi_{S_n}(t) &= \lim_{n\to\infty}\left(\phi_X\left(\frac{t}{\sqrt{n}}\right)\right)^n \\
&= \lim_{n\to\infty}\left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \\
&= \exp\left(-\frac{t^2\sigma^2}{2}\right) \qquad \text{//Recall that } e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n. \\
&= \phi_Z(t),
\end{aligned}$$

Therefore, $S_n \xrightarrow{d} Z$, which follows $\mathcal{N}(0, \sigma^2)$. □

Problem 2.2. *Recall that the median m of a random variable X is the minimum number that satisfies $\Pr(X \le m) \ge 1/2$. Like mean, median also has a central limit theorem. Unlike mean, CLT for median does not require finite variance. This allows us to construct more robust median estimation.*

1) Assume $\hat{m}_n$ is the median of $X_1, \ldots, X_{2n+1}$ i.i.d. drawn from a non-negative smooth density f on $\mathbb{R}$, whose true median is m. Please prove

$$\sqrt{n}(\hat{m}_n - m) \sim \mathcal{N}\left(0, \frac{1}{8f(m)^2}\right).$$

2) Generalize the result to derive a CLT for general quantiles (recall that the β-quantile q_β of random variable X is the minimum number that satisfies $\Pr(X \le q_\beta) \ge \beta$, for $\beta \in [0, 1]$).

Remark 2.1. *Check out Stein's method for an alternative proof of central limit theorem* http://www.math.nus.edu.sg/~lhychen/files/IMS4-pp-1-59.pdf.

Remark 2.2. *When X has infinite variance, but with finite $E[|X|^\alpha]$ for $1 < \alpha < 2$, we can find proper sequences a_n and b_n, such that $b_n(S_n - a_n)$ converges to some limit distribution called* stable distributions *(of which Gaussian is a special case). This draws connection to heavy tail distributions and extreme value statistics. Check generalized CLT https://amir.seas.harvard.edu/files/amir/files/gclt_evd_amir_acre_2017_v2.pdf.*

3 Asymptotics of Maximum Likelihood Estimator

Given $\{\boldsymbol{x}_i\}_{i=1}^n$ i.i.d. drawn from an unknown q_*, and a parametric family $\{q_\theta \colon \theta \in \Theta\}$. The maximum likelihood estimator is defined to be

$$\hat{\theta}_n = \arg\max_\theta \Big\{ L_n(\theta) := \mathbb{E}_{x\sim\hat{q}_n}[\log q_\theta(x)] \Big\}. \tag{2}$$

Note that $\hat{\theta}_n$ depends on the random sample $\{\boldsymbol{x}_i\}_{i=1}^n$ and is hence a random variable.

Recall that our fundamental goal is to minimize the KL divergence between the true and estimated distributions:

$$\mathrm{KL}(q_* \,||\, q_\theta) = -\mathbb{H}[q_*] - \mathbb{E}_{x\sim q_*}[\log q_\theta(x)].$$

Because the entropy term does not depend on the parameter θ. We can be evaluate the performance of θ by the second term, which is called the *expected testing likelihood*:

$$L(\theta) = \mathbb{E}[L_n(\theta)] = \mathbb{E}_{x\sim q_*}[\log q_\theta(x)].$$

Theorem 3.1 (Regret of MLE)**.** *Assume data $\{\boldsymbol{x}_i\}_{i=1}^n$ is i.i.d. drawn from an unknown distribution q_*. Let $\hat{\theta}_n$ be the MLE estimator in (2). Define the regret:*

$$\begin{aligned} R_n &:= \sup_{\theta\in\Theta} L_*(\theta) - L_*(\hat{\theta}_n) \\ &= \mathrm{KL}(q_* \,||q_{\hat{\theta}_n}) - \inf_{\theta\in\Theta}(\mathrm{KL}(q_* \,||q_\theta)). \end{aligned}$$

Then we have

$$R_n \le 2\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \log q_\theta(\boldsymbol{x}_i) - \mathbb{E}_{q_*}[\log q_\theta(\boldsymbol{x})]\right|.$$

Assume the function set $\mathcal{F} := \{\log q_\theta \colon \theta \in \Theta\}$ is a Glivenko-Cantelli class w.r.t. q_* in the sense of (1). Then we have R_n converges to zero with probability one, i.e., $R_n \xrightarrow{p} 0$.

Proof. For any $\theta \in \Theta$, we have $L_n(\hat{\theta}_n) \ge L(\theta)$ by the definition of $\hat{\theta}_n$. Therefore,

$$L_*(\theta) - L_*(\hat{\theta}_n) := \big(L_*(\theta) - L_n(\theta)\big) + \underbrace{\big(L_n(\theta) - L_n(\hat{\theta}_n)\big)}_{\le 0} + \big(L_n(\hat{\theta}_n) - L_*(\hat{\theta}_n)\big)$$

$$\le \big|L_*(\hat{\theta}_n) - L_n(\hat{\theta}_n)\big| + \big|L_n(\theta) - L_*(\theta)\big|$$

$$\le 2\sup_{\theta\in\Theta}\big|L_n(\theta) - L_*(\theta)\big|$$

$$= 2\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \log q_\theta(\boldsymbol{x}_i) - \mathbb{E}_{q_*}[\log q_\theta(\boldsymbol{x})]\right|.$$

The remaining part follows the definition of Glivenko-Cantelli class. □

Theorem 2.6 shows that $\mathcal{F} = \{\log q_\theta(x) \colon \theta \in \Theta\}$ is a Glivenko-Cantelli class if (1) Θ is compact, (2) $\log q_\theta(x)$ is continuous on θ for each x and (3) $|g(x,\theta)| \le M(x)$ for some M with $\mathbb{E}_{\boldsymbol{x}\sim q_*}[M(\boldsymbol{x})] < \infty$. Also, the set of bounded, Lipschitz functions is a universal Glivenko-Cantelli class. Read more about the importance of uniform bounds in learning theory in Chapter 2 of [27].

If the model is correctly specified, such that $q_* = q_{\theta_*} \in \mathcal{F}$, then it is expected that $\hat{\theta}_n$ converges to θ_* almost surely under some further assumptions. In this case, we say that $\hat{\theta}_n$ is a **consistent estimator**. One of the basic requirement for consistency is that the parameters has to be **identifiable**, meaning that there exists no $\theta_1 \neq \theta_2$ with $q_{\theta_1} = q_{\theta_2}$.

Theorem 3.2 (Consistency of MLE)**.** *Assume*

(1) Theorem 3.1 holds;

(2) Θ is compact; $q_\theta(x)$ is continuous w.r.t. θ for every x;

(3) the model is identifiable, that is, there exists no $\theta_1 \neq \theta_2$, such that $\log q_{\theta_1}(x) = \log q_{\theta_2}(x)$ for all x;

(4) there is no model specification, that is, $q_* = q_{\theta_*}$ and $\theta_* \in \Theta$.

Then we have

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_*.$$

Proof. Because there is no model misspecification, we have $\inf_{\theta\in\Theta} \mathrm{KL}(q_{\theta_*} \,||\, q_\theta) = 0$. By Theorem 3.1, we have

$$\mathrm{KL}(q_{\theta_*} \,||\, q_{\hat{\theta}_n}) \xrightarrow{a.s.} 0,$$

which suggests that $q_{\hat{\theta}_n} \xrightarrow{d} q_{\theta_*}$ almost surely.

Because Θ is compact, if $\hat{\theta}_n$ does not converge to θ_*, there must exists a convergent subsequent $\{\hat{\theta}_{n_k}\}_{k=1}^\infty$, whose limit, denote by θ_∞, is different from θ_*. For this subsequence, we have

$$\lim_{k\to\infty} \mathrm{KL}(q_{\theta_\infty} \,||\, q_{\hat{\theta}_{n_k}}) = \mathrm{KL}(q_{\theta_\infty} \,||\, q_{\theta_\infty}) = 0,$$

which is true because $\mathrm{KL}(q_{\theta_\infty} \,||\, q_\theta)$ is a continuous function on θ. This suggests that $q_{\hat{\theta}_{n_k}} \xrightarrow{d} q_{\theta_\infty}$ as $k \to \infty$. But because the limit of convergence in law is unique, we must have $q_{\theta_*} = q_{\theta_\infty}$, which contradicts with $\theta_\infty \neq \theta_*$ and the identifiability assumption. □

Theorem 3.3 (Asymptotic Normality of MLE)**.** *Assume $\theta_n \xrightarrow{a.s.} \theta_*$. Assume $\nabla_\theta \log q_\theta(x)$ yields a Taylor expansion around θ_*:*

$$\nabla_\theta \log q_\theta(x) = \nabla_\theta \log q_{\theta_*}(x) + \nabla^2_{\theta,\theta} \log q_{\theta_*}(x)(\theta - \theta_*) + R(\theta, \theta_*; x)$$

where $R(\theta, \theta_*; x)$ is the residual term and satisfies $|R(\theta, \theta_*; x)| \le M||\theta - \theta_*||^2$ for some constant M. Define the Fisher information matrix to be

$$I(\theta_*) = \mathrm{cov}_{q_*}\left(\nabla_\theta \log q_{\theta_*}(x)\right)$$

which we assume is invertible around θ_*. We have

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, I(\theta_*)^{-1}).$$

Proof. Note that at the fixed point, we have

$$\mathbb{E}_{\hat{q}_n}[\nabla_\theta \log q_{\hat{\theta}_n}(\boldsymbol{x})] = 0.$$

Using Taylor expansion, we have

$$\mathbb{E}_{\hat{q}_n}[\nabla_\theta \log q_{\theta_*}(\boldsymbol{x}) + \nabla^2_{\theta,\theta} \log q_{\theta_*}(\boldsymbol{x})(\hat{\theta}_n - \theta_*) + R(\hat{\theta}_n, \theta_*; \boldsymbol{x})] = 0.$$

This gives

$$\sqrt{n}\hat{J}(\theta_*)(\hat{\theta}_n - \theta_*) - \sqrt{n}\mathbb{E}_{\hat{q}_n}[R(\hat{\theta}_n, \theta_*; x)] = \hat{s}(\theta_*),$$

where we define the empirical Hessian matrix $\hat{J}(\hat{\theta}_*) := -\mathbb{E}_{\hat{q}_n}[\nabla^2_{\theta,\theta} \log q_{\theta_*}(\boldsymbol{x})]$ and $\hat{s}(\theta_*) := \sqrt{n}\mathbb{E}_{\hat{q}_n}[\nabla_\theta \log q_{\theta_*}(\boldsymbol{x})]$ By Lemma 3.4 and central limit theorem, we have

$$\hat{s}(\theta_*) := \sqrt{n}\mathbb{E}_{\hat{q}_n}[\nabla_\theta \log q_{\theta_*}(x)] \xrightarrow{d} \mathcal{N}(0, I(\theta_*)).$$

Therefore,

$$\sqrt{n}\hat{J}(\theta_*)(\hat{\theta}_n - \theta_*) + \mathbb{E}_{\hat{q}_n}[R(\hat{\theta}_n, \theta_*;\ x)] \xrightarrow{d} \mathcal{N}(0,\ I(\theta_*)).$$

Because $\hat{\theta}_n \xrightarrow{a.s.} \theta_*$, and $R(\theta, \theta_*; x) = o(\theta_n - \theta_*)$, we have

$$\hat{J}(\theta_*\sqrt{n}((\hat{\theta}_n - \theta_*) - \mathbb{E}_{\hat{q}_n}[R(\hat{\theta}_n, \theta_*;\ x)]) \xrightarrow{a.s.} J(\theta_*)\sqrt{n}(\hat{\theta}_n - \theta_*),$$

where $J(\theta_*) = -\mathbb{E}_{q_*}[\nabla^2_{\theta,\theta} \log q_{\theta_*}(x)]$, which equals $I(\theta_*)$ according to Lemma 3.4. Therefore,

$$I(\theta_*)\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{a.s.} \mathcal{N}(0,\ I(\theta_*)).$$

Because $I(\theta_*)$ is assumed to be inevitable, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{a.s.} \mathcal{N}(0,\ I(\theta_*)^{-1}).$$

□

{lem:IJ}

Lemma 3.4. *Let $\{q_\theta(x)\colon \theta \in \Theta\}$ is a set of second-order differentiable PDFs on $\mathbb{R}^d$, whose support $S = \{x \in \mathbb{R}^d\colon q_\theta(x) > 0\}$ does not depend on θ. We have*

$$\mathbb{E}_{q_\theta}[\nabla_\theta \log q_\theta(x)] = 0$$

$$\mathrm{cov}_{q_\theta}[\nabla_\theta \log q_\theta(x)] = -\mathbb{E}_{q_\theta}[\nabla^2_{\theta,\theta} \log q_\theta(x)].$$

Proof.

$$\begin{aligned}
\mathbb{E}_{q_\theta}[\nabla_\theta \log q_\theta(x)] &= \int_S q_\theta(x)\nabla_\theta \log q_\theta(x)dx \\
&= \int_S \nabla_\theta q_\theta(x)dx \\
&= \nabla_\theta \int_S q_\theta(x)dx \\
&= \nabla_\theta(1) = 0.
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}_{q_\theta}[\nabla_\theta \log q_\theta(x)] + \mathbb{E}_{q_\theta}[\nabla^2_{\theta,\theta} \log q_\theta(x)] &= \mathbb{E}_{q_\theta}[\nabla_\theta \log q_\theta(x) \nabla_\theta \log q_\theta(x)^\top + \nabla^2_{\theta,\theta} \log q_\theta(x)] \\
&= \int_S \nabla_\theta \log q_\theta(x) \nabla_\theta q_\theta(x)^\top dx + \nabla^2_{\theta,\theta} \log q_\theta(x) q_\theta(x) dx \\
&= \int_S \nabla_\theta \left(\nabla_\theta \log q_\theta(x)^\top q_\theta(x) \right) dx \\
&= \int_S \nabla_\theta \left(\nabla_\theta q_\theta(x)^\top \right) dx \\
&= \nabla^2_{\theta,\theta} \int_S q_\theta(x) dx \\
&= \nabla^2_{\theta,\theta}(1) = 0.
\end{aligned}$$

□

:crbound}

Theorem 3.5 (Cramer-Rao Bound (1D)). *Let $\hat\theta_n = \hat\theta_n(\boldsymbol{x})$ be any estimator of a parameter θ based on i.i.d. data $\boldsymbol{x} = \{x_i\}_{i=1}^n$ from q_θ. Define the bias of $\hat\theta_n$ to be*

$$b_n(\theta) = \mathbb{E}_{q_\theta}[\hat\theta_n(\boldsymbol{x})] - \theta.$$

Note that an estimator $\hat\theta_n$ is nothing but a function over data $\boldsymbol{x}$.

I) Assume the support S of q_θ does not depend on θ. For simplicity, consider the 1D case when $\theta \in \mathbb{R}$. We have

$$\mathrm{var}_{q_\theta}(\hat\theta_n(\boldsymbol{x})) \geq \frac{(1 + \nabla_\theta b_n(\theta))^2}{nI(\theta)}, \qquad \text{where} \quad I(\theta) := \mathrm{var}_{q_\theta}(\nabla_\theta \log q_\theta(x)).$$

By the bias-variance decomposition of the mean square error (MSE), we have

$$\mathbb{E}[(\hat\theta_n - \theta)^2] = \mathrm{var}_{q_\theta}(\hat\theta_n(\boldsymbol{x})) + (b_n(\theta))^2 \geq \frac{(1 + \nabla_\theta b_n(\theta))^2}{nI(\theta)} + (b_n(\theta))^2.$$

2) In particular, if $\hat\theta_n(\boldsymbol{x})$ is an unbiased estimator, that is, $\mathbb{E}_{q_\theta}[\hat\theta_n(\boldsymbol{x})] = \theta$, then we have

$$\mathbb{E}_{q_\theta}[(\hat\theta_n(\boldsymbol{x}) - \theta)^2] = \mathrm{var}_{q_\theta}(\hat\theta_n(\boldsymbol{x})) \geq \frac{1}{nI(\theta)}.$$

Proof. Taking $s(\boldsymbol{x}) = \sum_{i=1}^n \nabla_\theta \log q_\theta(x_i)$. We have from Lemma 3.4 that $\mathbb{E}_{q_\theta}[s(\boldsymbol{x})] = 0$. For notation, we denote by $\mathbb{E}[\cdot]$ and $\mathrm{var}(\cdot)$ the expectation and variance under q_θ below. By Cauchy Schwarz inequality:

$$\mathrm{var}(s(\boldsymbol{x})) \times \mathrm{var}(\hat\theta_n(\boldsymbol{x})) \geq (\mathbb{E}[s(\boldsymbol{x})(\hat\theta_n(\boldsymbol{x}) - \mathbb{E}(\hat\theta_n(\boldsymbol{x})))])^2 = (\mathbb{E}[s(\boldsymbol{x})\hat\theta_n(\boldsymbol{x})])^2,$$

where the last step is because $\mathbb{E}[s(\boldsymbol{x})] = 0$. Denote by $q_\theta(\boldsymbol{x}) = \prod_{i=1}^n q_\theta(x_i)$, and hence $s(\boldsymbol{x}) = \nabla_\theta \log q_\theta(\boldsymbol{x})$. We note that

$$\begin{aligned}
\mathbb{E}[s(\boldsymbol{x})\hat\theta_n(\boldsymbol{x})] &= \mathbb{E}[\nabla_\theta \log q_\theta(\boldsymbol{x}) \hat\theta_n(\boldsymbol{x})] \\
&= \int_{S^n} \nabla_\theta q_\theta(\boldsymbol{x}) \hat\theta_n(\boldsymbol{x}) d\boldsymbol{x} \\
&= \nabla_\theta \int_{S^n} q_\theta(\boldsymbol{x}) \hat\theta_n(\boldsymbol{x}) d\boldsymbol{x} \\
&= \nabla_\theta \mathbb{E}_{q_\theta}[\hat\theta_n(\boldsymbol{x})] \\
&= \nabla(\theta + b_n(\theta)) \\
&= 1 + \nabla_\theta b_n(\theta),
\end{aligned}$$

and

$$\mathrm{var}(s(\boldsymbol{x})) = \mathrm{var}(\sum_{i=1}^{n} \nabla_\theta \log q_\theta(x_i)) = n\mathrm{var}_{q_\theta}(\nabla_\theta \log q_\theta(x)) = nI(\theta).$$

Therefore,

$$\mathrm{var}(\hat{\theta}_n) \geq \frac{(\mathbb{E}[s(\boldsymbol{x})\hat{\theta}_n(\boldsymbol{x})])}{\mathrm{var}(s(\boldsymbol{x}))} = \frac{(1+\nabla_\theta b_n(\theta))^2}{nI(\theta)}.$$

□

Problem* 3.1. *Assume $\hat{\theta}_n$ is an unbiased estimator. Construct an biased estimator by $\tilde{\theta}_n = \lambda\theta_n$, where λ is constant (sometimes known as shrinkage parameter). By finding a proper λ, we can trade-off bias and variance and improve over the unbiased estimator.*

1. *Find the optimal λ such that the MSE $\mathbb{E}[(\tilde{\theta}_n - \theta)^2]$ is minimized.*

2. *Find the optimal λ such that lower bound of MSE in Theorem 3.5 is minimized.*

The Cramer-Rao bound ensures that no unbiased estimator can achieve asymptotically lower variance than the MLE. Stronger results, which we will not prove in this class, in fact show that no estimator, biased or unbiased, can asymptotically achieve lower mean-squared-error than $1/(nI(\theta))$, except possibly on a small set of special values $\theta \in \Theta$. However, rather surprisingly, one can find estimators that is strictly better than MLE even for the estimation of mean of Gaussian distribution. See Stein paradox https://en.wikipedia.org/wiki/Stein%27s_example.

4 Expectation Maximization

The goal is to learn latent variable models of form

$$p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz,$$

where z is some unseen latent variable that generates x through conditional distribution $p_\theta(x|z)$. We want to estimate θ given only observing $\{x_i\}_{i=1}^n$.

Nothing prevents us from applying MLE to this case, which yields an optimization of form

$$\max_\theta \left\{ L(\theta) := \frac{1}{n}\sum_{i=1}^n \log \int p_\theta(x_i|z)p_\theta(z)dz \right\}. \tag{3}$$

Unfortunately, people find it is difficult to solve this optimization directly (e.g., using gradient descent). Expectation maximization (EM) is a specialized algorithm for optimizing objectives like (3).

Main Idea If we observe both $\{x_i\}$ and their related latent variables $\{z_i\}$, we can simply maximize the joint distribution on (x, z):

$$\max_\theta \sum_{i=1}^n \log(p_\theta(x_i|z_i)p_\theta(z_i)).$$

This can be much easier to solve than maximizing the marginal likelihood (3).

However, we do not observe $\{z_i\}$ in practice. The idea of EM is to iteratively "impute" the missing values z_i, using the posterior distribution $p_{\theta_{old}}(z_i|x_i)$ from the parameter θ_{old} at the last iteration. Because θ_{old} may be inaccurate. We can repeat this procedure as θ improves until it converges.

The procedure of EM works as follows: Starting from some initial value θ_0, and perform iterative update by

$$\theta_{t+1} = \arg\max \left\{ Q(\theta \mid \theta_t) := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{z_i\sim p_{\theta_t}(\cdot|x_i)} \left[\log(p_\theta(x_i|z_i)p_\theta(z_i))\right] \right\}, \tag{4}$$

where the conditional expectation denotes drawing z_i from the posterior $p_{\theta_t}(z_i|x_i)$, where θ_t is the parameter from the last step t, that is,

$$p_{\theta_t}(z_i|x_i) = \frac{p_{\theta_t}(x_i|z_i)p_{\theta_t}(z_i)}{p_{\theta_t}(x_i)}.$$

Problem 4.1. *1) Assume $\log(p_\theta(x|z)p_\theta(z))$ is differentiable on θ. Prove that $Q(\theta \mid \theta_t)$ satisfies the following key properties:*

$$\begin{aligned} &i) \quad Q(\theta \mid \theta_t) - Q(\theta_t \mid \theta_t) \le L(\theta) - L(\theta_t) \qquad \forall\theta \ \text{ and } \ \theta_t \\ &ii) \quad \nabla_\theta Q(\theta \mid \theta_t) = \nabla_\theta L(\theta), \quad \text{when} \quad \theta = \theta_t. \end{aligned}$$

*$Q(\theta \mid \theta_t)$ is called a **minorization function** of $L(\theta)$ when the proprieties above holds.*

2) Prove that the EM update in (4) never decreases the object $L(\theta)$, that is,

$$L(\theta_{t+1}) \ge L(\theta_t), \quad \forall t.$$

In addition, if $\theta_{t+1} = \theta_t$, we must have $\nabla_\theta L(\theta_t) = 0$, meaning that the fixed point of EM must be a stationary point of $L(\theta)$.

The general procedure of successively maximizing a minorization function is known as **minorization-maximization (MM)** algorithm (it is called majorization-minimization for minimization problems). So EM is a special case of MM. The results above show that MM provides a monotonic ascending algorithm and converges to a stationary point of the objective function. However, note that when $L(\theta)$ is non-convex, the stationary point could be a local optima, or even a saddle point.

Problem 4.2. *Consider the mixture of linear regression:*

$$p_\theta(y|x) = \sum_{i=1}^m w_i \mathcal{N}(y|\ \mu_i^\top x, \sigma_i^2).$$

where $y = \mu_i x + \sigma_i \xi$ with probability w_i for $i = 1, \dots, m$, with ξ denoting Gaussian noise. The parameters of the model are $\theta = [w_i, \mu_i, \sigma_i]$, which should satisfies $\sum_i w_i = 1$, $w_i \geq 0$, $\sigma_i \geq 0$. Please derive the EM update rule for learning θ.

4.1 EM as KL Minimization

We provide an alternative interpretation of EM as KL divergence minimization. This derivation makes it natural to derive variational EM, an approximation of EM when calculating the posterior expectation w.r.t. $p(z_i \mid x_i)$ in (4). Note that MLE is equivalent to minimizing KL divergence:

$$\min_\theta \mathrm{KL}(q^X \ ||\ p_\theta^X), \tag{5}$$

where we use the superscript X to denote that they are distributions on X, to distinguish with the joint and conditional distribution shown in the sequel. Note that q^X denotes the data distribution that we observe. Our result is based on the following chain rule of KL divergence.

Theorem 4.1 (Chain Rule of KL Divergence)**.**

$$\mathrm{KL}(q^{X,Z} \ ||\ p^{X,Z}) = \mathrm{KL}(q^X \ ||\ p^X) \ +\ \mathbb{E}_{x\sim q^X}\left[\mathrm{KL}(q^{Z|x} \ ||\ p^{Z|x})\right], \tag{6}$$

where $q^{X,Z} = q^{X,Z}(x,z)$ is a joint distribution on (x,z), and $q^X(x) = \int q^{X,Z}(x,z)dz$ and $q^{Z|x}(z) = \frac{q^{X,Z}(x,z)}{q^X(x)}$ denote the marginal and conditional distributions. Note that $\mathrm{KL}(q^{Z|x} \ ||\ p^{Z|x})$ is a function of x, which is why an expectation on $x \sim q^X$ is needed in the formula.

If we fix the data distribution q^X, and minimize the both sides of (6) over *all possible conditional distributions* $q^{Z|X} = \{q^{Z|x} : \forall x\}$, we have

$$\begin{aligned}
\min_{q^{X|Z}} \mathrm{KL}(q^{X,Z} \ ||\ p_\theta^{X,Z}) &= \min_{q^{Z|X}} \left\{\mathrm{KL}(q^X \ ||\ p_\theta^X) \ +\ \mathbb{E}_{x\sim q^X}\left[\mathrm{KL}(q^{Z|x} \ ||\ p_\theta^{Z|x})\right]\right\} \\
&= \mathrm{KL}(q^X \ ||\ p_\theta^X) \ +\ \min_{q^{Z|X}} \left\{\mathbb{E}_{x\sim q^X}\left[\mathrm{KL}(q^{Z|x} \ ||\ p_\theta^{Z|x})\right]\right\} \\
&= \mathrm{KL}(q^X \ ||\ p_\theta^X),
\end{aligned}$$

where the second term is eliminated in the last step because the optimum is achieved when $p^{Z|X} = q^{Z|X}$. Intuitively, we may view $q^{Z|X}$ as an imputation distribution for sampling the hidden variable Z given observation X.

Therefore, the optimization of θ in (5) can be re-framed into a joint optimization on θ and the imputation distribution $q^{Z|X}$:

$$\min_\theta \mathrm{KL}(q^X \,||\, p_\theta^X) = \min_\theta \min_{q^{Z|X}} \mathrm{KL}(q^{X,Z} \,||\, p_\theta^{X,Z}).$$

In practice, we can start from some initialization and alternate between the optimization of θ and $q^{Z|X}$.

Given θ_t, the optimal $q^{Z|X}$ simply equals $p_{\theta_t}^{Z|X}$:

$$q_t^{Z|X} = \arg\min_{q^{Z|X}} \mathrm{KL}(q^{X,Z} \,||\, p_{\theta_t}^{X,Z}) = p_{\theta_t}^{Z|X},$$

Given $q_t^{Z|X}$, the update rule of θ should be

$$\begin{aligned}\theta_{t+1} &= \arg\min_\theta \mathrm{KL}(q^X q_t^{Z|X} \,||\, p_\theta^{X,Z}) \\ &= \arg\max_\theta \mathbb{E}_{(x,z)\sim q^X q_t^{Z|X}}[\log p_\theta^{X,Z}(x,z)],\end{aligned}$$

which can be easily seen to be equivalent to the EM update in (4). Because the alternative minimization of θ and $q^{Z|X}$ decrease the KL divergence monotonically, the monotonic property of EM is obvious from this perspective.

Variational EM When the model is too complex and it is intractable to calculate or evaluate the expectation of the posterior $p_{\theta_t}^{Z|X}$, EM algorithm can not be implemented directly. Variational EM is an approximation of EM algorithm, which restricts the optimization of $q^{Z|X}$ to a simple parametric family, whose optimization can be carried out numerically.

Specifically, assume $q^{Z|X} = q_\beta^{Z|X}$ has some simple form indexed by some parameter β (e.g., $q_\beta^{Z|X}$ can be a conditional Gaussian distribution). We have

$$\min_\theta \mathrm{KL}(q^X \,||\, p_\theta^X) = \min_\theta \min_{q^{Z|X}} \mathrm{KL}(q^X q^{Z|X} \,||\, p_\theta^{X,Z}) \le \min_\theta \min_\beta \mathrm{KL}(q^X q_\beta^{Z|X} \,||\, p_\theta^{X,Z}),$$

where the right side is larger because the the minimization of $q^{Z|X}$ is restricted on a smaller set indexed by parameter β. Here the idea is that we can not minimize the exact marginal KL divergence due to intractability, and instead minimize the upper bound on the right side as a practical surrogate. This is of course not ideal (and loss the theoretical guarantees we had for MLE), but it trades accuracy for faster, practical algorithm. Alternatively update of θ and β yields the following algorithm:

With fixed θ_t, update of β is

$$\begin{aligned}\beta_t &= \arg\min_\beta \mathbb{E}_{x\sim q^X}\left[\mathrm{KL}(q_\beta^{Z|x} \,||\, p_{\theta_t}^{Z|x})\right] \\ &= \arg\max_\beta \mathbb{E}_{x\sim q^X}\left[\mathbb{E}_{z\sim q_\beta^{Z|x}}[\log p_{\theta_t}^{X,Z}(x,z) - \log q_\beta^{Z|x}(z|x)]\right] \\ &= \arg\max_\beta \mathbb{E}_{x\sim q^X}\left[\mathbb{E}_{z\sim q_\beta^{Z|x}}[\log p_{\theta_t}^{X,Z}(x,z)] + \mathbb{H}(q_\beta^{Z|x})\right],\end{aligned}$$

where $\mathbb{H}(q_\beta^{Z|x}) = -\mathbb{E}_{z\sim q_\beta^{Z|x}}[\log q_\beta^{Z|x}(z|x)]$ denotes the conditional entropy.

Update of θ is the same as before:

$$\begin{aligned}\theta_{t+1} &= \arg\min_{\theta} \mathrm{KL}(q^X q_{\beta_t}^{Z|X} \;||\; p_\theta^{X,Z}) \\ &= \arg\max_{\theta} \mathbb{E}_{x\sim q^X}\left[\mathbb{E}_{z\sim q_{\beta_t}^{Z|x}}[\log p_\theta^{X,Z}(x,z)]\right].\end{aligned}$$

Variational Auto-encoder (VAE) VAE is the application of variational EM when p_θ^X is specified as a deep generative models. In this case, assume x is some complex, high dimensional object (such as image or text), which is associated with some unseen, lower dimensional latent representation z. The latent z is assumed to be generated by p_θ^Z, which could be something simple, such as a fixed standard Gaussian distribution $\mathcal{N}(0,1)$. Given z, the conditional distribution $p_\theta^{X|z}$ is often specified to be conditional Gaussian distribution:

$$p_\theta^{X|z}(x|z) = \mathcal{N}(x;\; \mu_\theta(z), \Sigma_\theta(z)),$$

where ξ is standard Gaussian noise and where $\mu_\theta(z)$ and $\Sigma_\theta(z)$ are two (likely highly complex) nonlinear functions (such as neural networks) that specify the conditional mean and variance of x given the value of z. When x are images, μ_θ is typically a convolutional neural network. The covariance matrix $\Sigma_\theta(z)$ is often assumed to be a diagonal matrix for simplicity. The distribution of p_θ^X is hence

$$p_\theta^X(x) = \int p_\theta^{X|Z}(x|z)p_\theta^Z(z)dz = \int \exp\left(-\frac{1}{2}(x-\mu_\theta(z))^\top \Sigma_\theta(z)^{-1}(x-\mu_\theta(z)) - \frac{1}{2}||z||_2^2\right) dz.$$

The map from z to x is called the decoder. When doing variational EM, we also need to specify $q_\beta^{Z|x}$, which can be viewed as an "encoder". An example of q_β can be

$$q_\beta^{Z|x}(z|x) = \mathcal{N}(z;\; \tilde{\mu}_\beta(x), \tilde{\sigma}_\beta^2(x)),$$

where $\tilde{\mu}_\beta$ and $\tilde{\sigma}_\beta^2(x)$ can be another two neural networks, which maps x to z.

When implementing VAE, we need to use Monte Carlo sampling to further approximate the expecatation involved with the posterior distribution $q_\beta^{Z|X}$. In addition, we need to be careful how to estimate the gradient when optimizing β. A technique, called reparametrization trick, need to be used in order to optimize β more efficiently. These issues will be covered later in the class.

Problem 4.3. *The optimization of θ and β is performed using gradient descent in practice. Write down the gradient descent update formula of θ and β for VAE. Check the original VAE paper by Kingma and Welling [14].*

5 Integral Probability Measures (IPM) and GANs

In MLE and KL divergence, the distributions being learned are specified by their density functions. However, in many practical cases, distributions of interest may be singular and have no finite density function (that is, they are not absolutely continuous w.r.t. Lebesgue measure). This happens, for example, when all the data concentrates on a low dimensional manifold of a higher dimensional space, so that the density function is either infinite (on the manifold), or zero (outside the manifold). MLE and KL divergence are **undefined** (not just inefficient or computationally intractable). We need other tools.

Generative adversarial networks (GANs) provide a general approach for learning singular distributions without valid density functions. The idea is to parameterize distributions by their generative mechanisms, instead of density functions.

Specifically, let P_θ be the distribution of random variables generated by $X = g_\theta(\xi)$, where ξ is a "random seed" generated from some given, fixed distribution P_0 (such as Gaussian or uniform), and $g_\theta(\xi)$ is some parametric function (e.g., a neural network) with parameter θ that converts ξ to the distribution we want. For notation, we may write $\mathsf{P}_\theta = g_\theta \sharp \mathsf{P}_0$, and P_θ is considered as the pushforward measure obtained by transferring ("pushing forward") measure P_0 by function $g_\theta(\cdot)$. So the class of distributions we are interested in is defined by

$$\mathcal{P}_\Theta := \{\mathsf{P}_\theta = g_\theta \sharp P_0 : \quad \theta \in \Theta\}.$$

The GAN Problem Assume we observe a set of data $\{x_i\}_{i=1}^n$ i.i.d. drawn from an unknown distribution P_*. We want to find that best θ, such that P_θ approximates P_*.

This is a rather natural idea. In fact, all the random variables in digital computers are generated in a similar fashion. The randomness in computers starts from some basic random number generator programs which generates samples from certain basic distributions, typically a uniform distribution (which is approximated by pseudo random numbers that appear random, but are actually deterministic); various transforms are then applied on the basic distribution to obtain the samples from more complex distributions of interest.

For example, for one-dimensional $\mathbb{R}$-valued distributions with cumulative probability function (CDF) $F(x)$, one common approach is to apply transform $F^{-1}(\xi)$, where $\xi \sim \text{Uniform}([0,1])$ and F^{-1} is the inverse function of F. The GAN problem above can be viewed as "learning random number generators", where the transform function g_θ is learned empirically from data.

Since KL divergence can not be used, we need alternative divergence measures that can be defined and computed without accessing density functions. One of the most natural approach is to use integral probability metric (IPM), which we now introduce.

5.1 Integral Probability Metrics (IPM)

Given two probability measures Q and P on a domain $\mathcal{X}$, and a class of functions $\mathcal{F}$ defined on the same domain. Assume $\mathcal{F}$ is *even* in that $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$. The $\mathcal{F}$-based integral probability measure ($\mathcal{F}$-IPM) of P and Q is defined to be

{equ:ipm}
$$D_{\mathcal{F}}(\mathsf{Q},\ \mathsf{P}) = \sup_{f\in\mathcal{F}}\{|\mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f|\} = \sup_{f\in\mathcal{F}}\{\mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f\}, \tag{7}$$

where the absolute value can be dropped because $\mathcal{F}$ is even. It is obvious to see that $D_{\mathcal{F}}(\mathsf{Q},\ \mathsf{P})$ is symmetric and satisfies the triangle inequality,

$$\begin{aligned} &D_{\mathcal{F}}(\mathsf{Q},\ \mathsf{P}) = D_{\mathcal{F}}(\mathsf{P},\ \mathsf{Q}) \\ &D_{\mathcal{F}}(\mathsf{Q}_1,\ \mathsf{Q}_2) \le D_{\mathcal{F}}(\mathsf{Q}_1,\ \mathsf{P}) + D_{\mathcal{F}}(\mathsf{P},\ \mathsf{Q}_2). \end{aligned} \tag{8}$$

This is in contrast with KL divergence which is not symmetric and does not satisfy triangle inequality.

Functions that satisfy (8) are called *semi-distances*. In addition, if $\mathcal{F}$ is chosen properly or sufficiently rich such that

$$D_{\mathcal{F}}(\mathsf{Q},\ \mathsf{P}) = 0 \quad \text{implies} \quad \mathsf{Q} = \mathsf{P},$$

then $D_{\mathcal{F}}$ is a distance (or probability metric) on the set of distributions. A sufficient condition for $D_{\mathcal{F}}(\mathsf{Q},\ \mathsf{P})$ to be a distance is when the linear span of functions in $\mathcal{F}$ is dense in the set of bounded continuous functions under $\|\cdot\|_\infty$, that is, for any bounded continuous function f^* and any $\epsilon > 0$, there exists $\{a_i\} \subseteq \mathbb{R}$, $\{f_i\} \subseteq \mathcal{F}$ and $n \in \mathbb{N}$, such that

$$\sup_{x \in \mathcal{X}} \left| \sum_i a_i f_i(x) + a_0 - f^*(x) \right| \le \epsilon.$$

See Theorem 2.2 of Zhang et al. [29]. This is a fairly mild requirement. An example of functions spaces that satisfies this conditions is

$$\mathcal{F} = \{\sigma(\theta^\top x + b) \colon [\theta, b] \in \mathbb{R}^{d+1}\},$$

for any σ that are not polynomial, when $\mathcal{X} \subset \mathbb{R}^d$ is compact (see Theorem 2.3 of Zhang et al. [29]). This includes moment generating functions (when $\sigma(t) = \exp(t)$), and typical neural networks with a **single** neuron(!). Of course, the "discriminative power" would be larger if we have a larger set of functions.

IPM includes a large number of probability metrics as special cases, depending on the choice of $\mathcal{F}$.

1-Wasserstein Distance 1-Wasserstrein distance is the case when $\mathcal{F}$ includes all Lipschitz functions, that is,

$$D_{Wass}(\mathsf{Q} \,||\, \mathsf{P}) = \sup_f \left\{ \mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f \quad s.t. \quad ||f||_{Lip} \le 1 \right\},$$

where the Lipschitz norm is defined to be

$$||f||_{Lip} = \sup_{x \neq y} \frac{f(x) - f(y)}{\mathsf{d}(x, y)},$$

and $\mathsf{d}(x, y)$ is a notion of distance between x and y (such as the L2 or L1 distance).

Theorem 5.1. *$D_{Wass}(\mathsf{Q} \,||\, \mathsf{P})$ is equivalent to the following definition motivated by optimal transport:*

$$D_{Wass}(\mathsf{Q} \,||\, \mathsf{P}) = \inf_\gamma \mathbb{E}_\gamma[\mathsf{d}(x, y)],$$

where $\inf_\gamma$ is over all distributions γ on (X, Y), such that $\gamma^X = \mathsf{P}$ and $\gamma^Y = \mathsf{Q}$. Here γ^X and γ^Y denotes the marginal distribution of γ on X and Y, respectively.

Proof. The optimization is equivalent to

$$D_{Wass}(\mathsf{Q} \,||\, \mathsf{P}) = \sup_f \left\{ \mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f \quad s.t. \quad f(x) - f(y) \le \mathsf{d}(x,y), \quad \forall x, y \in \mathcal{X} \right\},$$

Using the Lagrange multiplier, the optimization is equivalent to

$$\begin{aligned} D_{Wass}(\mathsf{Q} \,||\, \mathsf{P}) &= \sup_f \left\{ \inf_\gamma \inf_{c \ge 0} \mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f + c\mathbb{E}_\gamma \left[\mathsf{d}(x,y) - f(x) + f(y)\right] \right\} \\ &= \inf_\gamma \inf_{c \ge 0} \sup_f \left\{ \mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f + c\mathbb{E}_\gamma \left[\mathsf{d}(x,y) - f(x) + f(y)\right] \right\} \qquad \text{//assume strong duality} \\ &= \inf_\gamma \inf_{c \ge 0} \left\{ c\mathbb{E}_\gamma \left[\mathsf{d}(x,y)\right] + \sup_f \{ \mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f + c\mathbb{E}_\gamma [-f(x) + f(y)] \} \right\} \end{aligned}$$

This is equivalent to

$$D_{Wass}(\mathsf{Q} \,||\, \mathsf{P}) = \inf_\gamma \inf_{c \ge 0} \left\{ c\mathbb{E}_\gamma \left[\mathsf{d}(x,y)\right] \quad s.t. \quad \mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f + c\mathbb{E}_\gamma [-f(x) + f(y)] = 0, \ \ \forall f \right\},$$

where the constraint of $\mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f + c\mathbb{E}_\gamma[-f(x) + f(y)] = 0$ for all f is equivalent to $c\gamma^X = \mathsf{P}$ and $c\gamma^Y = \mathsf{Q}$, which implies $c = 1$ for normalization. Therefore, we have

$$D_{Wass}(\mathsf{Q} \,||\, \mathsf{P}) = \inf_\gamma \left\{ \mathbb{E}_\gamma \left[\mathsf{d}(x,y)\right] \quad s.t \quad \gamma^X = \mathsf{P}, \quad \gamma^Y = \mathsf{Q} \right\}.$$

This completes the proof. □

Neural IPM and Wasserstein GAN In practice, it is difficult (although possible in principle) to optimize over the set of Lipschitz functions. A more practical approach is to take $\mathcal{F}$ to be a parametric family of functions, that is, $\mathcal{F} = \{f_\beta \colon \beta \in \mathcal{B}\}$, where Θ is a finite dimensional parameter space.

$$D_{NN}(\mathsf{Q} \,||\, \mathsf{P}) = \sup_{\beta \in \mathcal{B}} \left\{ \mathbb{E}_{\mathsf{P}} f_\beta - \mathbb{E}_{\mathsf{Q}} f_\beta \right\}.$$

Obviously, the parametric set needs to be constructed such that the norm of f_β can not be arbitrarily large.

We can apply $D_{NN}(\mathsf{Q} \,||\, \mathsf{P})$ to solve the GAN problem. Denote by Q_n the empirical measure of the observation $\{x_i\}_{i=1}^n$, that is, $\mathsf{Q}_n(dx) := \sum_i \delta(x - x_i)dx/n$ We find P_θ to approximate Q_n by solving the following minimax problem:

$$\min_{\theta \in \Theta} \max_{\beta \in \mathcal{B}} \left\{ \mathbb{E}_{\mathsf{Q}_n}[f_\beta(x)] - \mathbb{E}_{\mathsf{P}_\theta}[f_\beta(x)] \right\} = \min_{\theta \in \Theta} \max_{\beta \in \mathcal{B}} \left\{ \frac{1}{n} \sum_{i=1} \left(f_\beta(x_i)\right) - \mathbb{E}_{\xi \sim \mathsf{P}_0} \left[f_\beta\left(g_\theta(\xi)\right)\right] \right\}$$

This algorithm is called **Wasserstein-GAN** [1]. In practice, $\mathbb{E}_{\mathsf{Q}_n}$ is approximated by subsampling from the dataset, $\mathbb{E}_{\mathsf{P}_\theta}$ is approximated by drawing $\{\xi_i\}$ from P_0 and evaluating the empirical averaging over $x_i = g_\theta(\xi_i)$. And θ and β are updated alternatively by performing stochastic gradient descent. In the terminology of GAN, and g_θ is known as the generator, which generates the data from the model; and f_β is known as the discriminator, which attempts to identify the difference of the empirical and model

distributions. The set β should be chosen properly to penalize the Lipschitz or other norm of the function f. For example, Arjovsky et al. [1] did this by clipping the norm of β in gradient descent updates, while Gulrajani et al. [11] introduces a penalty term of the form $(||\nabla f_\beta||_2 - 1)^2$.

More theoretical issues of W-GAN can be found in a line of recent works. See, for example, Arora et al. [2], Zhang et al. [29] and references therein for discussions on generalization and discriminative powers of the minimax loss function. Also, understanding the convergence of this procedure is an important but challenging task [see e.g., 16].

6 f-Divergence

Let P and Q be two probability distributions over a space Ω such that P is absolutely continuous with respect to Q. Then, for a convex function f such that $f(1) = 0$, the f-divergence of P from Q is defined as

$$D_f(\mathsf{P} \parallel \mathsf{Q}) = \int_\Omega f\left(\frac{d\mathsf{P}}{d\mathsf{Q}}\right) d\mathsf{Q} = \mathbb{E}_\mathsf{Q}\left[f\left(\frac{d\mathsf{P}}{d\mathsf{Q}}\right)\right].$$

If P and Q are both absolutely continuous with respect to a reference measure μ on Ω, then their probability densities p and q satisfy $d\mathsf{P} = pd\mu$ and $d\mathsf{Q} = qd\mu$. In this case the f-divergence can be written as

$$D_f(\mathsf{P} \parallel \mathsf{Q}) = D_f(p \parallel q) = \mathbb{E}_\mathsf{Q}\left[f\left(\frac{p(x)}{q(x)}\right)\right].$$

We may not use P and p interchangeably to represent the same distribution.

Remark: Please do not confuse the f in f-divergence with the f used in IPM. I should have used some other notation.

Theorem 6.1. *Assume $f\colon \mathbb{R}_+ \cup \{0\} \to \mathbb{R} \cup \{\pm\infty\}$ is a strictly convex function with $f(1) = 0$. If P is absolutely continuous w.r.t. Q, then*

$$D_f(\mathsf{P} \parallel \mathsf{Q}) \geq 0,$$

and $D_f(\mathsf{P} \parallel \mathsf{Q}) = 0$ implies $\mathsf{P} = \mathsf{Q}$.

Proof. Because P is absolutely continuous on Q, we have $\mathbb{E}_{x\sim q}[p(x)/q(x)] = 1$ following Lemma 1.2. Using Jensen's inequality, we have

$$D_f(\mathsf{P} \parallel \mathsf{Q}) = \mathbb{E}_\mathsf{Q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] - f(1) = \mathbb{E}_\mathsf{Q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] - f\left(\mathbb{E}_\mathsf{Q}\left[\frac{p(x)}{q(x)}\right]\right) \geq 0.$$

In addition, since f is strictly convex, the inequality is tight iff $p(x)/q(x) = const$ almost surely, which means $\mathsf{P} = \mathsf{Q}$. □

Problem* 6.1. *What if P is not absolutely continuous on Q? Do we need any additional condition on f to ensure that the non-negativity and discriminativeness of f-divergence? Do some study.*

Many common divergences, such as KL-divergence, Hellinger distance, and total variation distance, are special cases of f-divergence, coinciding with a particular choice of f. For example, we have

$$D_f(\mathsf{P} \,||\, \mathsf{Q}) = \begin{cases} \mathrm{KL}(\mathsf{Q} \,||\, \mathsf{P}) & \text{if } f(t) = -\log t \\ \mathrm{KL}(\mathsf{P} \,||\, \mathsf{Q}) & \text{if } f(t) = t \log t \\ \mathrm{TV}(\mathsf{P} \,||\, \mathsf{Q}) & \text{if } f(t) = |t-1| \\ \mathrm{H}(\mathsf{P} \,||\, \mathsf{Q}) & \text{if } f(t) = (\sqrt{t}-1)^2, \end{cases}$$

where $\mathrm{H}(\mathsf{P} \,||\, \mathsf{Q})$ denotes the Helinger distance, defined by

$$\mathrm{H}(\mathsf{P} \,||\, \mathsf{Q}) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.$$

f-divergence and IPM have very different properties. The total variation distance $\mathrm{TV}(\mathsf{P} \,||\, \mathsf{Q})$ is the only divergence that belong to both f-divergence and IPM. See Sriperumbudur et al. [24] for more discussion regarding f-divergence and IPM.

Problem* 6.2. *1) Given a convex function f, find another convex function $\tilde{f}$, such that $D_f(\mathsf{P} \,||\, \mathsf{Q}) = D_{\tilde{f}}(\mathsf{Q} \,||\, \mathsf{P})$.*

2) Under what condition $D_f(\mathsf{P} \,||\, \mathsf{Q})$ is symmetric?

6.1 Dual Representation of f-Divergence

Recall that continuous convex functions have the dual representation

$$f(x) = \sup_t \{t^\top x - f^*(t)\},$$

where f^* is the convex conjugate function of f, which is also a convex function. If f is differentiable and strictly convex, the optimal t is achieved when $\nabla f(t) = x$. See Section C in Appendix for more discussion on convex conjugate.

Using this we can rewrite f-divergence to a very useful dual representation.

Theorem 6.2. *Let f be a differentiable, strictly convex function on $\mathbb{R}_+$. Denote by f^* the convex conjugate of f, we have*

$$D_f(\mathsf{P} \,||\, \mathsf{Q}) = \sup_\phi \left\{\mathbb{E}_\mathsf{P}[\phi(x)] - \mathbb{E}_\mathsf{Q}[f^*(\phi(x))]\right\}. \tag{9}$$

where the sup *is overall possible functions ϕ such that the objective above is finite. In addition, the optimality is achieved when*

$$\frac{p(x)}{q(x)} = f'(\phi(x)).$$

Proof. Plugging the dual representation of f into the definition of f-divergence, we have

$$\begin{aligned} D_f(\mathsf{P} \,||\, \mathsf{Q}) &= \mathbb{E}_{\mathsf{Q}}\left[f\left(\frac{p(x)}{q(x)}\right)\right] \\ &= \mathbb{E}_{\mathsf{Q}}\left[\sup_{\phi(x)\colon \forall x}\left(\frac{p(x)}{q(x)}\phi(x) - f^*(\phi(x))\right)\right] \\ &= \sup_{\phi(x)\colon \forall x} \mathbb{E}_{\mathsf{Q}}\left[\left(\frac{p(x)}{q(x)}\phi(x) - f^*(\phi(x))\right)\right] \\ &= \sup_{\phi(x)\colon \forall x} \left\{\mathbb{E}_{\mathsf{P}}\left[\phi(x)\right] - \mathbb{E}_{\mathsf{Q}}\left[f^*(\phi(x))\right]\right\}. \end{aligned}$$

Following the property of convex conjugate in Section C, the optimality is achieved when $p(x)/q(x) = f'(t(x))$. □

Problem* 6.3. *The dual form provides a convenient tool for proving some mathematical properties of f-divergence that are more difficult to see thought its original definition.*

1) Using the dual form to prove that $D_f(p \,||\, q)$ is a convex function of $[p, q]$. (hint: maximum of a set of linear functions is convex.)

2) Let $T\colon \mathcal{X} \to \mathcal{X}$ be any map. Prove that

$$D_f(T\sharp\mathsf{P} \,||\, T\sharp\mathsf{Q}) \le D_f(\mathsf{P} \,||\, \mathsf{Q}).$$

Recall that $T\sharp\mathsf{P}$ is the pushforward measure of P obtained through transform T.

In addition, when $T\colon \mathcal{X} \to \mathcal{X}$ is an one-to-one map, we have

$$D_f(T\sharp\mathsf{P} \,||\, T\sharp\mathsf{Q}) = D_f(\mathsf{P} \,||\, \mathsf{Q}).$$

f-Divergence as Regularized IPM It is unclear directly from (9) to see intuitively why the dual presentation should measure the discrepancy between P and Q. This is in contrast with the IPM in (7), which has a clear meaning as a measuring the maximum discrepancy between expectations of a class of functions.

In addition, unlike IPM in (7), f-divergence (9) is asymmetric and the optimization of ϕ is over arbitrary functions (that keep the objective function finite and defined), while IPM constraints the optimization inside a function class $\mathcal{F}$ (of bounded norm in some sense) to prevent the optimization in (7) to diverge to infinite. It is unclear what the role of f^* is in f-divergence. To see this, let us rewrite

$$D_f(\mathsf{P} \,||\, \mathsf{Q}) = \sup_{\phi} \left\{\mathbb{E}_{\mathsf{P}}[\phi(x)] - \mathbb{E}_{\mathsf{Q}}[\phi(x)] - \Phi_{f^*,q}[\phi]\right\}, \tag{10}$$

with

$$\Phi_{f^*,\mathsf{Q}}[\phi] = \mathbb{E}_{\mathsf{Q}}[f^*(\phi(x)) - \phi(x)],$$

where $\Phi_{f^*,\mathsf{Q}}[\phi]$, as we show in the next theorem, can be viewed as as a complexity regularization term on ϕ that prevents the solution goes to infinite. This suggests that f-divergence can be viewed as a regularized variant of maximum mean discrepancy; this explains why is ok to optimize ϕ over the set of arbitrary functions. In contrast, IPM measures the maximum mean discrepancy while constraining the test functions within a function class $\mathcal{F}$.

Theorem 6.3 (Lemma B.1. of Zhang et al. [29]). *1) Let f be a convex function with $f(1) = 0$ and f^* the convex conjugate of f. We have*

$$\Phi_{f^*,\mathsf{Q}}[\phi] \geq 0, \quad \forall \phi.$$

In addition, if f is strictly convex, we have $\Phi_{f^*,\mathsf{Q}}[\phi] = 0$ if and only if $\phi(x) = c$ almost surely under $x \sim q$ for some constant c.

Proof. 1) Since $f^*(t) = \sup_s\{ts - f(s)\}$, we have $f^*(t) \geq ts - f(s)$ for any s and t. Taking $s = 1$, we have $f^*(t) \geq t$, which implies that $\Phi_{f^*,\mathsf{Q}}[\phi] \geq 0$, $\forall \phi$.

If $\Phi_{f^*,\mathsf{Q}}[\phi] = 0$, we have $f^*(\phi(x)) - \phi(x) = 0$ almost surely under $x \sim q$. If f is strictly convex, then $f^*(t) - t$ is also strictly convex, which suggests that there exists only a single c such that $f^*(c) - c = 0$. This suggests that $\phi(x) = c$ almost surely under q. □

Problem* 6.4. *Propose a condition on f (or f^*), such that $\Phi_{f^*,\mathsf{Q}}[\phi]$ is equivalent to the L2 norm $\mathbb{E}_{\mathsf{Q}}[(\phi(x))^2]$, that is, there exists $\alpha \geq \beta > 0$, such that $\beta\mathbb{E}_{\mathsf{Q}}[(\phi(x))^2] \leq \Phi_{f^*,\mathsf{Q}}[\phi] \leq \alpha\mathbb{E}_{\mathsf{Q}}[(\phi(x))^2]$.*

The dual representation involves optimization overall all possible functions, which is not numerically tractable. However, it is possible to construct the optimization to a proper function class to derive approximation. We give two examples below.

Density Ratio Estimation [18] Assume $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^m$ two samples drawn from two unknown distributions P and Q, respectively. How to estimate the density ratio $r(x) := p(x)/q(x)$, without constructing density approximations?

This problem can be addressed using f-divergence, by noting that the optimal solution of the optimization in (9) solves $p(x)/q(x) = f'(\phi^*(x))$. Therefore, we can approximately solve the optimization, and derive the density ratio from the optimal ϕ^*. The optimization can be approximated by

$$\max_{\phi\in\mathcal{H}} \left\{\mathbb{E}_{\mathsf{P}_n}[\phi(x)] - \mathbb{E}_{\mathsf{Q}_m}[f^*(\phi(x))]\right\},$$

where P_n and Q_m are empirical measures of the two samples, and $\mathcal{H}$ is a class of functions on which numerical optimization can be performed (it is taken to be an RKHS (see Section 7) in Nguyen et al. [18]).

f-GAN Given a collection of data $\{x_i\}_{i=1}^n$, whose empirical measure is denoted by Q_n, we can formulate the learning of generative models P_θ by minimizing the dual form of f-divergence:

$$\min_{\theta\in\Theta}\max_{\beta\in\mathcal{B}} \left\{\mathbb{E}_{\mathsf{Q}_n}[\phi_\beta(x)] - \mathbb{E}_{\mathsf{P}_\theta}[f^*(\phi_\beta(x))]\right\},$$

where the optimization of ϕ is restricted to a parametric set $\{\phi_\beta \colon \beta \in \mathcal{B}\}$, which is typically taken to be a neural network. The objective here yields a "neural-f-divergence", which is a lower bound of the f-divergence.

Problem 6.1. *Write down the convex function f and the corresponding dual representation of* $\mathrm{KL}(\mathsf{Q} \,||\, \mathsf{P})$ *and* $\mathrm{KL}(\mathsf{P} \,||\, \mathsf{Q})$*, and the following Jensen-Shanon divergence (JSD):*

$$\mathrm{JSD}(\mathsf{P} \,||\, \mathsf{Q}) = \frac{1}{2}\mathrm{KL}\left(\mathsf{P} \,||\, \frac{\mathsf{P}+\mathsf{Q}}{2}\right) + \frac{1}{2}\mathrm{KL}\left(\mathsf{Q} \,||\, \frac{\mathsf{P}+\mathsf{Q}}{2}\right).$$

Derive the original GAN by Goodfellow et al. [8] using Jensen-Shanon divergence (note that there is a transform between the discriminator in the original GAN and the ϕ in our formulation).

Problem 6.2. *Prove that*

$$\mathrm{KL}(\mathsf{Q} \,||\, \mathsf{P}) = \sup_{\phi}\{\mathbb{E}_{\mathsf{Q}}[\phi(x)] - \log(\mathbb{E}_{\mathsf{P}}[\exp(\phi(x))])\}.$$

This dual representation is different from the one we obtained from f-divergence. Compare these two representations, and do some analysis on which one might be better (and for what purposes).

7 Reproducing Kernel Hilbert Space

{sec:rkhs}

Positive Definite Kernels A two-variable function $K(x, x')\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a kernel if it is symmetric in x and x', that is, $K(x, x') = K(x', x)$. The kernel is said to be positive definite if the matrix $\boldsymbol{K} := [K(x_i, x_j)]_{ij}$ is positive semi-definite for any $\{x_i\} \subseteq \mathcal{X}$ of finite size. A typical example of kernel is the Gaussian RBF kernel:

usskernel}

$$K(x, x') = \exp\left(-\frac{1}{2h^2}\left\|x - x'\right\|_2^2\right), \tag{11}$$

where h is a positive scaling factor called bandwidth. Another example is the inner product kernels:

{mercer}

$$K(x, x') = \sum_{\ell} \phi_\ell(x)\phi_\ell(x'), \tag{12}$$

where $\{\phi_\ell\colon \forall \ell\}$ is a set of (potentially infinitely many) functions called feature map sometimes. For example, when we obtain polynomial kernels when $\{\phi_\ell\}$ are polynomials.

In general, we may intuitively think the kernel $K(x, x')$ as a measure of similarity between x and x' (even though $K(x, x')$ does not have to be positive for every $x, x' \in \mathcal{X}$). Mercer's theorem guarantees that any continuous positive definite kernel on a compact domain $\mathcal{X}$ can be represented as an inner product kernel like (12), with ϕ_ℓ being orthogonal to each other.

Inner Product and Hilbert Space An inner product space is a linear space on which a notion of inner product is defined. Let $\mathcal{F}$ be a linear space, a two-variable operator $\langle \cdot, \cdot \rangle_{\mathcal{F}}\colon \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ can be called an inner product if it satisfies that following axioms, for any $f, g, h \in \mathcal{F}$ and $a \in \mathbb{R}$,

1. Symmetric: $\langle f,\ g\rangle_{\mathcal{F}} = \langle f,\ g\rangle_{\mathcal{F}}$.
2. Linearity: $a\langle f,\ g\rangle_{\mathcal{F}} = \langle af,\ g\rangle_{\mathcal{F}}$, $\langle f + g,\ h\rangle_{\mathcal{F}} = \langle f, g\rangle_{\mathcal{F}}\ +\ \langle g, h\rangle_{\mathcal{F}}$.
3. Positive-definiteness: $\langle f,\ f\rangle_{\mathcal{F}} \geq 0$, *and* $\langle f,\ f\rangle_{\mathcal{F}} = 0$ iff $f = 0$.

The pair $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$, or simply $\mathcal{F}$, is called an inner product space. With the inner product, we can define a norm by

$$\|f\|_{\mathcal{F}} = \sqrt{\langle f,\ f\rangle_{\mathcal{F}}},$$

which then implies a distance between f and g via $||f - g||_{\mathcal{F}}$. Therefore, an inner product space is also a normed space. The definition of inner product and norm implies Cauchy-Schwarz inequality.

m:cauchy}

Theorem 7.1 (**Cauchy-Schwarz Inequality**). *Following the definition of inner product and norm above, we have*

$$\langle f,\ g\rangle_{\mathcal{F}} \leq \|f\|_{\mathcal{F}} \cdot \|g\|_{\mathcal{F}}, \qquad \forall f, g \in \mathcal{F}.$$

In addition, the equality is achieved when $f = g$.

Proof. Define $f_0 = f/\|f\|_{\mathcal{F}}$ and $g_0 = g/\|g\|_{\mathcal{F}}$. Note that

$$\|f_0 - g_0\|_{\mathcal{F}}^2 = \langle f_0 - g_0,\ f_0 - g_0\rangle_{\mathcal{F}} = \|f_0\|_{\mathcal{F}}^2 + \|g_0\|_{\mathcal{F}}^2 - 2\langle f_0,\ g_0\rangle_{\mathcal{F}} = 1 + 1 - 2\frac{\langle f,\ g\rangle_{\mathcal{F}}}{\|f\|_{\mathcal{F}} \cdot \|g\|_{\mathcal{F}}}. \tag{13}$$

The result then follows by $\|f_0 - g_0\|_{\mathcal{F}}^2 \geq 0$. □

A normed space $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is called **complete** if every Cauchy sequence converges to a well defined limit that is within that space. Specifically, a sequence $\{f_i\}_{i=1}^{\infty} \subset \mathcal{F}$ is called a Cauchy sequence if for every positive real number $\epsilon > 0$, there is a positive integer N such that for all positive integers $m, n \geq N$, we have $||f_n - f_m||_{\mathcal{F}} \leq \epsilon$. We call $\mathcal{F}$ complete if, for every Cauchy sequence, there exists an element $f_* \in \mathcal{F}$, such that $\lim_{n\to\infty} ||f_n - f_*|| = 0$.

If an inner product space is complete, then it is called a **Hilbert space**. Given an inner product space one can construct an (unique) Hilbert space by adding all the limit points of all Cauchy sequence. The Hilbert space obtained this way is called the **completion** of the inner product space, which is technically the equivalent class of all the Cauchy sequence of the original space. An incomplete inner product space is called a **pre-Hilbert** space, since its completion with respect to the norm induced by the inner product is a Hilbert space. The result below shows that every inner product space can be completed to a Hilbert space.

{mpletion}

Theorem 7.2. *Let $(\mathcal{H}^o, \langle\cdot,\cdot\rangle_{\mathcal{H}^o})$ is an inner product (or pre-Hilbert) space, then there exists an Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}^o})$ and a map $\mathcal{U}\cdot\mathcal{H}^o \to \mathcal{H}$ such that*

- $\langle \mathcal{U}f, \mathcal{U}g\rangle_{\mathcal{H}} = \langle f, g\rangle_{\mathcal{H}^o}$.
- $\mathcal{U}\mathcal{H}^o := \{\mathcal{U}f : f \in \mathcal{H}^o\}$ *is dense in* $\mathcal{H}$. *If* $\mathcal{H}^o$ *is complete, then* $\mathcal{H}^o = \mathcal{H}$.
- $\mathcal{U}: \mathcal{H}^o \to \mathcal{U}\mathcal{H}^o$ *is an one-to-one and linear map.*

*$\mathcal{H}$ is called the **completion** of $\mathcal{H}^o$.*

Proof (brief sketch). Construct $\mathcal{H}$ to be the equivalent class of all Cauchy sequences in $\mathcal{H}^o$,

$$\mathcal{H} = \{\{f_i\}_{i=1}^{\infty} : \{f_i\}_{i=1}^{\infty} \quad \text{is a Cauchy sequence in } \mathcal{H}^o\},$$

where two Cauchy sequences are viewed the same if they have the same limit. For two Cauchy sequences $\{f_i\}_{i=1}^{\infty}$ and $\{g_i\}_{i=1}^{\infty}$, define their inner product to be $\lim_{i\to\infty}\langle f_i, g_i\rangle_{\mathcal{H}^o}$, which can be shown to exist and satisfy the basic properties of inner product. □

{def:rkhs}

Definition 7.3 (Reproducing Kernel Hilbert Space). *A Hilbert space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS), if there exists a kernel $K(x, x')$, such that*

1. *$K(x,\cdot) \in \mathcal{H}$ for every $x \in \mathcal{X}$;*
2. *It satisfies the reproducing property:*

$$f(x) = \langle f(\cdot),\ K(x,\cdot)\rangle_{\mathcal{H}}, \qquad \forall f \in \mathcal{H},\quad x \in \mathcal{X}.$$

*$K(x, x')$ is called the **reproducing kernel** of $\mathcal{H}$.*

This basic definition turns out to be very powerful, and many good things can be derived from it.

Fact 1: If a kernel $K(x, x')$ is the reproducing kernel of some RKHS, then $K(x, x')$ must be symmetric and positive definite. To see this, applying the reproducing property on $K(x,\cdot)$, we have

{equ:krep}

$$K(x, x') = \langle K(x,\cdot),\ K(x',\cdot)\rangle_{\mathcal{H}}. \tag{14}$$

This implies the symmetricity $K(x,x') = K(x',x)$. To see the positive definiteness, note that for any $\{a_i\} \subset \mathbb{R}$, $\{x_i\} \subset \mathcal{X}$, we have

$$\sum_{ij} a_i K(x_i,x_j) a_j = \langle \sum_i a_i K(x_i,\cdot),\ \sum_i a_i K(x_i,\cdot)\rangle_{\mathcal{H}} = \left\| \sum_i a_i K(x_i,\cdot) \right\|_{\mathcal{H}}^2 \geq 0.$$

Fact 2: Every positive definite kernel $K(x,x')$ has an unique RKHS whose reproducing kernel is $K(x,x')$. In fact, the RKHS of $K(x,x')$ consists of the closure of the linear span of kernel functions $\{K(x,\cdot)\colon x \in \mathcal{X}\}$, with a proper definition of inner product that satisfies the reproducing property. We now describe the procedure of constructing RKHS from $K(x,x')$, and prove the uniqueness.

Definition 7.4. *Define the linear span of $\{K(x,\cdot)\colon x \in \mathcal{X}\}$ to be*

$$\mathcal{H}^o := \text{span}\{K(x,\cdot)\colon x \in \mathcal{X}\} = \left\{ f(\cdot) = \sum_{i=1}^n a_i K(\cdot,x_i)\colon \quad \{a_i\} \subset \mathbb{R},\ \{x_i\} \subset \mathcal{X},\ n \in \mathbb{N} \right\}.$$

$\mathcal{H}^o$ can be equipped with an inner product defined as follows: for $f(x) = \sum_i a_i K(x,x_i) \in \mathcal{H}^o$ and $g(x) = \sum_i b_i K(x,x_i) \in \mathcal{H}^o$,

{inner}

$$\langle f,\ g\rangle_{\mathcal{H}^o} = \sum_{ij} a_i b_j K(x_i,x_j). \tag{15}$$

This definition is unique: even when f has two different representations, e.g., $f(x) = \sum_i a_i K(x,x_i) = \sum_i a'_i K(x,x_i)$, the inner product only depends on f itself instead of its representations, since $\langle f,\ g\rangle_{\mathcal{H}^o} = \sum_j b_j f(x_j)$. It is easy to verify that $\langle \cdot,\ \cdot\rangle_{\mathcal{H}^o}$ is a valid inner product when $K(x,x')$ is positive definite.

Denote by $\overline{\mathcal{H}^o}$ the completion of the inner product space $\mathcal{H}^o$ following Theorem 7.2. Note that $\overline{\mathcal{H}^o}$ is a Hilbert space consisting of all the limit points of Cauchy sequence of $\mathcal{H}^o$.

Theorem 7.5 (Existence and Uniqueness of RKHS)**.** *Let $K(x,x')$ be a positive definite kernel. Then*

i) $\overline{\mathcal{H}^o}$ is an RKHS of kernel $K(x,x')$.

ii) If $\mathcal{H}$ is an RKHS of kernel $K(x,x')$, then $\mathcal{H} = \overline{\mathcal{H}^o}$.

Proof (brief sketch). i) The key step of showing $\overline{\mathcal{H}^o}$ is an RKHS is to prove the reproducing property. For $f(\cdot) = \sum_i a_i K(\cdot,x_i) \in \mathcal{H}^o$, and setting $g(\cdot) = K(x,\cdot)$, following the definition in (15),

$$\langle f(\cdot),\ K(x,\cdot)\rangle_{\mathcal{H}^o} = \langle f,\ g\rangle_{\mathcal{H}^o} = \sum_i a_i K(x_i,\ x) = f(x).$$

The result extends to $\overline{\mathcal{H}^o}$ by taking the limits.

ii) To show that any RKHS $\mathcal{H}$ equals $\overline{\mathcal{H}^o}$ requires to show that (1) $\mathcal{H}^o$ consists the same set of functions, and (2) it must have the same inner product structure.

For (1), note that we already have $\overline{\mathcal{H}^o} \subseteq \mathcal{H}$ following definition of RKHS. If $\overline{\mathcal{H}^o}$ is a strict subspace of $\mathcal{H}$, then there exists an orthogonal complement $\mathcal{H}^{o\perp}$, such that $\langle f,g\rangle = 0$ for any $f \in \mathcal{H}^{o\perp}$ and $g \in \overline{\mathcal{H}^o}$. Since we have $g(\cdot) := K(x,\cdot) \in \overline{\mathcal{H}^o}$ for all $x \in \mathcal{X}$, we have following the reproducing property:

$$f(x) = \langle f,\ K(x,\cdot)\rangle_{\mathcal{H}} = 0.$$

This suggests that all the elements in $\mathcal{H}^{o\perp}$ equals zero, which hence implies $\mathcal{H} = \overline{\mathcal{H}^o}$.

For (2), consider $f(\cdot) = \sum_i a_i K(\cdot, x_i) \in \mathcal{H}^o$ and $g(\cdot) = \sum_i b_i K(\cdot, x_i) \in \mathcal{H}^o$. Following the reproducing property of $\mathcal{H}$, we have

$$\langle f, g\rangle_{\mathcal{H}} = \sum_j b_j \langle f, K(x_j, \cdot)\rangle_{\mathcal{H}} = \sum_j b_j f(x_j) = \sum_{ij} a_i b_j K(x_i, x_j),$$

which shows that $\langle f, g\rangle_{\mathcal{H}} = \langle f, g\rangle_{\mathcal{H}^o}$ following the definition of $\langle\cdot,\cdot\rangle_{\mathcal{H}^o}$ in (15). Extending this to the limit points of $\mathcal{H}^o$ completes the proof.

□

An immediate result of the existence of RKHS this is that every positive definite kernel can be decomposed into an inner product of some feature map.

Theorem 7.6. *A kernel $K(x, x')$ is positive definite if and only if there exists a Hilbert space $\mathcal{H}$, and a map (called the **feature map**) $\boldsymbol{\phi}\colon \mathcal{X} \to \mathcal{H}$, such that*

$$K(x, x') = \langle \boldsymbol{\phi}(x),\ \boldsymbol{\phi}(x')\rangle_{\mathcal{H}}. \tag{16}$$

Proof. Kernels of form (16) is obviously positive definite. On the other hand, for every positive definition kernel, we can simply take $\mathcal{H}$ to its RKHS, and $\boldsymbol{\phi}(x) = K(x, \cdot)$, and (16) follows (14). □

Note that the feature map $\boldsymbol{\phi}(x)$ constructed this way is "function-valued". In machine learning, (16) is related to what is called **kernel trick**, which is the idea of lifting the original feature vectors x to a higher, possibly infinite, dimensional feature map $\boldsymbol{\phi}(x)$, for which all the computation and evaluation can be performed by only using the kernel $K(x, x')$ (instead of the potentially infinite dimensional feature map $\boldsymbol{\phi}(x)$). You can learn more about kernel method from a range of textbooks and reviews, such as Scholkopf and Smola [23].

7.1 Bounded Evaluation Functional and Riesz Representation Theorem

For any $f \in \mathcal{H}$ and $x \in \mathcal{X}$, we can define L_x to be the evaluation operator at x:

$$L_x f = f(x), \quad \forall f \in \mathcal{H}, x \in \mathcal{X},$$

where L_x is viewed as an operator acted on functions in $\mathcal{H}$ and returns the value of f at point x. It is easy to see that L_x is a linear functional, that is, $L_x(f + g) = L_x f + L_x g$. The norm of a linear functional is defined by $||L_x|| := \sup_{f\in\mathcal{H}} \frac{|L_x f|}{||f||_{\mathcal{H}}}$. A linear functional is bounded if its norm is finite, that is, $||L_x|| < \infty$.

Fact 3: RKHS can be equivalently defined as Hilbert spaces on which all the evaluation operators L_x are bounded.

Theorem 7.7 (Reproducing kernel is equivalent to bounded L_x)**.** *A Hilbert space $\mathcal{H}$ is a RKHS following Definition 7.3 if and only if all its evaluation operators L_x are bounded linear functionals.*

Proof. First, if a $\mathcal{H}$ has reproducing kernel $K(x, x')$, we have

$$L_x f = f(x) = \langle f,\ K(x,\cdot)\rangle_{\mathcal{H}} \le \|f\|_{\mathcal{H}} \|K(x,\cdot)\|_{\mathcal{H}} = \|f\|_{\mathcal{H}} K(x,x) < \infty,$$

where we use $\|K(x,\cdot)\|^2_{\mathcal{H}} = \langle K(x,\cdot), K(x,\cdot)\rangle_{\mathcal{H}} = K(x,x)$, following the reproducing property.

The other direction is more technical, and requires what is called **Riesz representation theorem**, which we state without proof.

Theorem 7.8 (Riesz representation)**.** *If L is a bounded linear functional on a Hilbert space $\mathcal{H}$, then there is a unique vector $g_L \in \mathcal{H}$ such that*

$$Lf = \langle g_L,\ f\rangle_{\mathcal{H}} \qquad \text{for all } f \in \mathcal{H}.$$

This says that bounded linear operators on Hilbert spaces can be expressed as inner products. Applying Riesz representation theorem to L_x, we have that there exists $\phi_x \in \mathcal{H}$, such that $L_x f = f(x) = \langle f,\ \phi_x\rangle_{\mathcal{H}}$. Define $K(x,x') = \langle \phi_x,\ \phi_{x'}\rangle_{\mathcal{H}}$. We just need to show that $K(x,x')$ is the reproducing kernel of $\mathcal{H}$. To see it, first note that

$$\phi_x(x') = L_{x'}\phi_x = \langle \phi_x, \phi_{x'}\rangle_{\mathcal{H}} = K(x,x').$$

In other words, $\phi_x(\cdot) = K(x,x')$. Therefore, for any $f \in \mathcal{H}$,

$$f(x) = \langle f,\ \phi_x\rangle_{\mathcal{H}} = \langle f,\ K(x,\cdot)\rangle_{\mathcal{H}}.$$

This shows $\mathcal{H}$ is the RKHS with reproducing kernel $K(x,x')$. □

Problem 7.1. *1) Prove that the RKHS of the Gaussian RBF kernel in (11) on domain $\mathcal{X} = \mathbb{R}^d$ does not include any linear function $f(x) = a^\top x + b$, except $f(x) = 0$.*

2) Construct a new kernel so that its RKHS includes the union of all linear functions and all the functions in the RKHS of the Gaussian RBF kernel.

Problem* 7.1. *For differentiable functions on $(0,1)$, we may define the following inner product*

$$\langle f, g\rangle_{\mathcal{H}} = \int_0^1 f'(x)g'(x)dx,$$

where f' denotes the derivative of f. Let $\mathcal{H}$ be the Hilbert space with this inner product, consisting of the closure of infinite differentiable functions supported in compact subsets of $(0,1)$. Prove that $\mathcal{H}$ is an RKHS with kernel $k(x,y) = \min(x,y)$. Use this fact to solve the following optimization for non-linear regression:

$$\min_f \sum_i (f(x_i) - y_i)^2 + \lambda \int_0^1 (f'(x))^2 dx.$$

(hint: Denote by $k'_x(x,y) = \mathbb{I}(x \le y)$ the derivative of k w.r.t. x). To prove the reproducing property, we note that

$$f(y) = \int f'(x)k'_x(x,y)dx = \int_0^y f'(x)dx = f(y).$$

)

7.2 Random Features

Random features provide a powerful tool for approximating kernels in big data settings. It also provides an alternative interpretation of RKHS that is very useful for both scalable kernel learning and theoretical understanding.

In many case, we can decompose the kernel into the following form

$$K(x,x') = \int_{\mathcal{W}} \phi(x,w)\phi(x',w)d\mu(w), \tag{17}$$

where μ is a measure on some domain $\mathcal{W}$, and $\phi(\cdot,\cdot)$ is a measurable function of x and w. In cases when μ is a probability measure, (17) is viewed as a random feature expansion of $K(x,x')$, because we can draw $\{w_i\}_{i=1}^m$ i.i.d. from μ and approximate $K(x,x')$ by

$$\hat{K}_m(x,x') = \frac{1}{m}\sum_{i=1}^m \phi(x,w_i)\phi(x',w_i).$$

This approach is widely used in large scale kernel learning. A major class of random features are Fourier features (see Rahimi and Recht [20]).

Theorem 7.9 (Bochner [21]). *A kernel is called stationary if it has a form of $K(x,y) = K(x-y)$, for a one-variable function $K(\cdot)$. A continuous stationary kernel on $\mathbb{R}^d$ is positive definite if and only if $K(\cdot)$ is the Fourier transform of a non-negative measure.*

Using this result, we can decompose continuous stationary kernels into expectations of cosine random features,

$$\phi(x,w) = \sqrt{2}\cos(w_1^\top x + w_0), \qquad w = [w_1, w_0].$$

We obtain different kernels by taking different distribution μ on $[w_1, w_0]$. For example, Gaussian RBF kernel with unit bandwidth $K(x,x') = \exp(-||x-x'||_2^2/2)$ can be obtained by $w_1 \sim \mathcal{N}(0,1)$ and $w_0 \sim ([0,1])$. See Figure 1 of Rahimi and Recht [20].

We need some basic definition before explaining the relation between RKHS and random features.

Definition 7.10 ($\mathcal{L}_2(\mu)$ Space). *Given a measure μ on domain $\mathcal{X}$, the $\mathcal{L}_2(\mathcal{X},\mu)$ space is defined to be the set of square integral functional under measure μ:*

$$\mathcal{L}_2(\mathcal{X},\mu) = \left\{ f\colon f \text{ is measurable and } \|f\|^2_{\mathcal{L}_2(\mathcal{X},\mu)} := \int f(x)^2 d\mu(x) < \infty \right\}, \tag{18}$$

where $\|f\|_{\mathcal{L}_2(\mathcal{X},\mu)}$ is called the $\mathcal{L}_2(\mathcal{X},\ \mu)$ norm of f, so $\mathcal{L}_2(\mathcal{X},\ \mu)$ is the space of functions with finite $\mathcal{L}_2(\mathcal{X},\ \mu)$ norm. We simply write $\mathcal{L}_2(\mu)$ or $\mathcal{L}_{2,\mu}$ when it is obvious what domain it is defined on, and $\mathcal{L}_2(\mathcal{X})$ or $\mathcal{L}_2$ when μ is the Lebesgue measure.

$\mathcal{L}_2(\mu)$ can be turned into a Hilbert space with the following definition of inner product

$$\langle f,\ g\rangle_{\mathcal{L}_2(\mu)} = \int f(x)g(x)d\mu(x).$$

Note that f and g are viewed as identical in $\mathcal{L}_2(\mu)$ if $f = g$ almost surely under μ, that is, $\mu(\{x \in \mathcal{X}\colon f(x) \neq g(x)\}) = 0$.

aturerkhs}

Theorem 7.11. *For positive definite kernel $K(x,x')$ of form (17), its RKHS consists functions of form*

$$\mathcal{H} = \left\{ f(x) = \int \phi(x,w)\varrho_f(w)\mu(dw), \quad \forall \varrho_f \in \mathcal{L}_2(\mu), \quad \|\varrho_f\|_{\mathcal{L}_2(\mu)} < \infty \right\}.$$

For simplicity, assume $\{\phi(\cdot,w)\colon w \in \mathcal{W}\}$ is linearly independent, that is, $\int \phi(\cdot,w)\rho_f(w)d\mu(w) = 0$ implies $\|\varrho_f\|_{\mathcal{L}_2(\mu)} = 0$, so that the map from ϱ_f to f is one-to-one. Then the inner product and norm on $\mathcal{H}$ can be represented by that in $\mathcal{L}_2(\mu)$:

{equ:fh}

$$\langle f,\ g\rangle_{\mathcal{H}} = \langle \varrho_f,\ \varrho_g\rangle_{\mathcal{L}_2(\mu)}, \qquad \|f\|_{\mathcal{H}} = \|\varrho_f\|_{L_2(\mu)}. \tag{19}$$

where $f = \int \phi(x,w)\varrho_f(w)\mu(dw)$ and $g = \int \phi(x,w)\varrho_g(w)\mu(dw)$. Note that this establishes an ismorphism between $\mathcal{L}_2(\mu)$ and $\mathcal{H}$. TODO

Proof (sketch). Obviously, $\mathcal{H}$ defined in this way is a Hilbert space. We just show that $K(x,x')$ is a reproducing kernel of $\mathcal{H}$ defined in this way. To see it, note that

$$f(x) = \int \phi(x,w)\varrho_f(w)\mu(dw) = \langle \phi(x,\cdot),\ \varrho_f(\cdot)\rangle_{\mathcal{L}_2(\mu)} = \langle K(x,\cdot),\ f(\cdot)\rangle_{\mathcal{H}},$$

where the last step follows the definition of $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ in (19) and the feature expansion of kernel in (17). □

{prob:nn1}

Problem* 7.2. *Consider one-layer neural networks of form*

$$f(x;\ w) = \sum_i \sigma(w_i^\top x),$$

where σ is the activation function (such as ReLU $\sigma(t) = \max(0,t)$) and w_i is the weight vector of the i-th neuron. Assume we have continuously infinite numbers of neurons, whose weights follow a measure ρ. Such networks can be represented

$$f(x;\ \rho) = \int \sigma(w^\top x)\rho(dw)$$

Does the set of neural networks $\{f(x;\ \rho)\colon \forall\rho\}$ form an RKHS? If not, could you construct a RKHS which includes (a reasonably large) subset of such infinite neural networks? Read Bach [3].

7.3 Nonparametric Learning and Finite Representer Theorem

RKHS can be used as the model class for nonparametric learning and estimation. We consider the nonparameteric regression problem as illustrative examples. Given observation $\{x_i,y_i)\}_{i=1}^n$, and we want to find a function f, such that $f(x_i) \approx y_i$. Instead of assuming f to form certain parametric form like typical parametric regression approaches, we assume f is an element of an RKHS, yielding the following infinite dimensional optimization problem.

nelregress}

$$\min_{f\in\mathcal{H}} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{20}$$

where we search for the best f within RKHS $\mathcal{H}$ to minimize the mean square loss function, with a penalty term that enforces the RKHS norm to be small, which controls the complexity of f.

The remarkable fact is that this optimization, despite being infinite dimensional, can be solved numerically with a very simple form. This is thanks to the **finite representer theorem** [13, 22], which suggests that the optimal solution of (20) has to in the linear span of kernels $K(x, x_i)$ evaluated on the data points $\{x_i\}$, that is,

$$f(x) = \sum_{i=1}^{n} a_i K(x, x_i),$$

where $\{a_i\}$ is a set of parameters to be decided. Therefore, we just need to search the optimal $\{a_i\}$, which can be done by plugging this presentation into the original optimization:

$$\min_{\{a_i\}} \sum_{i=1}^{n} (\sum_{j=1}^{n} a_j K(x_i, x_j) - y_i)^2 + \lambda \sum_{ij=1}^{n} a_i K(x_i, x_j) a_j.$$

This can be rewritten into a matrix form

$$\min_{\boldsymbol{a}} \|\boldsymbol{K}\boldsymbol{a} - \boldsymbol{y}\|_2^2 + \lambda \boldsymbol{a}^\top \boldsymbol{K} \boldsymbol{a},$$

where $\boldsymbol{a} = [a_i]_{i=1}^n$, $\boldsymbol{K} = [K(x_i, x_j)]_{ij}$, $\boldsymbol{y} = [y_i]_{i=1}^n$. The optimal solution is

$$\hat{\boldsymbol{a}} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y},$$

where $\boldsymbol{I}$ is the identity matrix of the same size as $\boldsymbol{K}$.

Let us now formally introduce a version of finite representer theorem. The key idea is to decompose f into the sum of functions in $span(K(\cdot, x_i) \colon i = 1, \dots n)$ and its orthogonal complementary, and show that the orthogonal complementary can not contribute to decrease the data-related loss, and hence should be set to be zero due to the RKHS norm penalty.

Theorem 7.12. *Consider the optimization of the following form:*

$$\min_{f \in \mathcal{H}} \left\{ L(f(x_1), \dots, f(x_n)) \ + \ \Phi(\|f\|_{\mathcal{H}}) \right\},$$

which depends on f only through its evaluate $f(x_i)$ at a finite number of points $\{x_i\}_{i=1}^n$ (through an arbitary function $L(\cdots)$), and its RKHS norm $\|f\|_{\mathcal{H}}$. Assume Φ is a strictly increasing function so that functions with smaller norm is favored. Then the optimal solution must have a form of

$$f(x) = \sum_{i=1}^{n} a_i K(x, x_i),$$

for some coefficient $\{a_i\}_{i=1}^n \in \mathbb{R}^n$.

Proof. Note that any f in $\mathcal{H}$ can be decomposed into

$$f(x) = \sum_{i} a_i K(x, x_i) + f_\perp(x),$$

where $f_\perp$ is a function in the orthogonal complement of the subspace spanned by $\{K(\cdot, x_i) \colon i = 1, \dots, n\}$. We just need to show that $f_\perp = 0$ at the optimal point.

We start with observing that we must have $f_\perp(x_i) = 0$ for $i = 1, \dots, n$. This is because of the reproducing property:

$$f_\perp(x_i) = \langle f_\perp,\ K(\cdot, x_i)\rangle_{\mathcal{H}} = 0,$$

which equals 0 because $f_\perp$ is orthogonal to all $K(\cdot, x_i)$ by assumption.

Therefore, different choices of $f_\perp$ does not influence the data term $L(f(x_1), \dots, f(x_n))$; for the RKHS norm penalty, note that

$$\|f\|^2_{\mathcal{H}} = ||\sum_{i=1}^n a_i K(x, x_i)||^2_{\mathcal{H}} + \|f_\perp(x)\|^2_{\mathcal{H}},$$

which is minimized when $f_\perp = 0$. Therefore, we must have $f_\perp = 0$ at the optimality, which suggests that the optimal solution should have the form of $f(x) = \sum_{i=1}^n a_i K(x, x_i)$. □

Since the loss function L is arbitrary, its application is not restricted to simple mean square loss. Consider the case of **kernel classification**: Given data $\{x_i, y_i\}_{i=1}^n$, where we have binary labels $y_i = \{0, 1\}$, which is assumed to be generated by

$$p(y|x;\ f) = \frac{\exp(yf(x))}{1 + \exp(yf(x))}.$$

Maximum likelihood estimation of f within RKHS yields

$$\max_{f\in\mathcal{H}} \sum_{i=1} \log p(y_i|x_i;\ f) + \lambda \|f\|^2_{\mathcal{H}},$$

which again has a solution of form $f(x_i) = \sum_i a_i K(x, x_i)$. Unlike the case of regression, the optimal $\{a_i\}$ in this case should be defined to numerical algorithms, such as gradient descent.

Computation Complexity of kernel regression is $O(n^3)$, which is slow when the data size n is very large (a.k.a. big data). Developing fast kernel methods for big data settings is a well studied topic in machine learning and various other areas. Typical ideas includes finding a small set "anchor points" to obtain more compact representations, using random feature approximation, etc. See Scholkopf and Smola [23].

The choice of kernels is another important decision that we need to face when using kernel methods. A default choice is obviously the Gaussian RBF kernel $K(x, x') = \exp(-\|x - x'\|_2^2/h^2)$, whose bandwidth can be chosen using the so called **"median trick"**, which amounts to take h to be proportional to the median of the pairwise distance of the dataset, that is, $h = c \times \text{med}(\{\|x_i - x_j\| : i \neq j\})$, and c is some scaling constant that can be adjusted. This allows the bandwidth to adapt with the dataset.

Gaussian RBF kernel does not work well for complex, high dimensional data. People have investigated specialized kernels for structured objects such as strings, graphs, even distributions, etc. People also studied leveraging deep neural networks to design kernels that leverage the power of deep learning.

7.4 Generalization and Rademacher Complexity

RKHS appears to include "a lot" of functions. Should we worry about overfitting. It turns out this is not a problem. The set of RKHS functions is actually "very small", in that its Rademacher complexity is only $O(1/\sqrt{n})$, the same rate as typical parametric classes.

Theorem 7.13. *Consider the Rademacher complexity of the unit ball $\mathcal{B} = \{f \in \mathcal{H}: \quad \|f\|_{\mathcal{H}} \le 1\}$ of RKHS $\mathcal{H}$ of kernel $K(x, x')$:*

$$\hat{R}_{X_n}(B) = \mathbb{E}_\sigma[\sup_{f\in\mathcal{H}} \sum_{i=1}^n \sigma_i f(x_i)]$$

7.5 Maximum Mean Discrepancy

As another major application, RKHS is widely used as the discriminator set in IPM, which again, yields a simple closed form. Define

$$D_{\mathcal{H}}(\mathsf{P},\ \mathsf{Q}) = \max_{f\in\mathcal{H}} \{\mathbb{E}_{\mathsf{P}}[f] - \mathbb{E}_{\mathsf{Q}}[f]: \quad \|f\|_{\mathcal{H}} \le 1\}, \tag{21}$$

where $D_{\mathcal{H}}(\mathsf{P},\ \mathsf{Q})$ is often known as (kernel) maximum mean discrepancy (MMD) [9].

Theorem 7.14. *Let $\mathcal{H}$ be the RKHS of kernel $K(x, x')$. For MMD defined in (21), we have*

$$D_{\mathcal{H}}(\mathsf{P},\ \mathsf{Q}) = \sqrt{\mathbb{E}[K(X, X') + K(Y, Y') - 2K(X, Y)]},$$

where (X, X') and (Y, Y') are i.i.d. random variables drawn from P and Q, respectively. The optimal f that solves (21) is

$$f^*(\cdot) = \frac{1}{D_{\mathcal{H}}(\mathsf{P},\ \mathsf{Q})}\mathbb{E}[K(X, \cdot) - K(Y, \cdot)].$$

Proof (using reproducing property). Recall a basic fact of Hilbert space following Cauchy-Schwarz inequality: the solution of

$$\max_{f\in\mathcal{H}} \langle f,\ g\rangle_{\mathcal{H}}, \quad s.t. \quad \|f\|_{\mathcal{H}} \le 1$$

equals $f^* = g/\|g\|_{\mathcal{H}}$, and the optimal value is $\langle f^*, g\rangle_{\mathcal{H}} = \|g\|_{\mathcal{H}}$.

For our problem, using the reproducing property, we have

$$\begin{aligned}\mathbb{E}[f(X) - f(Y)] &= \mathbb{E}[\langle f,\ K(X, \cdot) - K(Y, \cdot)\rangle_{\mathcal{H}}] \\ &= \langle f,\ \ \mathbb{E}[K(X, \cdot) - K(Y, \cdot)]\rangle_{\mathcal{H}}.\end{aligned}$$

Taking $g(\cdot) = \ \mathbb{E}[K(X, \cdot) - K(Y, \cdot)]$, we get

$$D_{\mathcal{H}}(\mathsf{P},\ \mathsf{Q}) = \sup_{\|f\|_{\mathcal{H}}\le 1} \{\mathbb{E}[f(X) - f(Y)]\} = \sup_{\|f\|_{\mathcal{H}}\le 1} \{\langle f,\ g\rangle_{\mathcal{H}}\} = \|g\|_{\mathcal{H}}.$$

So we just need to calculate the RKHS norm.

$$\begin{aligned}\|g\|_{\mathcal{H}}^2 &= \|\mathbb{E}[K(X, \cdot) - K(Y, \cdot)]\|_{\mathcal{H}}^2 \\ &= \big\langle \mathbb{E}[K(X, \cdot) - K(Y, \cdot)],\ \ \mathbb{E}[K(X', \cdot) - K(Y', \cdot)]\big\rangle_{\mathcal{H}} \\ &= \mathbb{E}\left[\big\langle K(X, \cdot) - K(Y, \cdot),\ \ K(X', \cdot) - K(Y', \cdot)\big\rangle_{\mathcal{H}}\right] \\ &= \mathbb{E}\left[\langle K(X, \cdot), K(X', \cdot)\rangle_{\mathcal{H}} + \langle K(Y, \cdot), K(Y', \cdot)\rangle_{\mathcal{H}} - \langle K(X, \cdot), K(Y', \cdot)\rangle_{\mathcal{H}} - \langle K(Y, \cdot), K(X', \cdot)\rangle_{\mathcal{H}}\right] \\ &= \mathbb{E}\left[K(X, X') + K(Y, Y') - K(X, Y') - K(X', Y)\right] \\ &= \mathbb{E}\left[K(X, X') + K(Y, Y') - 2K(X, Y)\right].\end{aligned}$$

The form of optimal $f^* = g/\|g\|_{\mathcal{H}}$ can be found accordingly. □

Proof (using random features). Assume $K(x,x') = \int \phi(x,w)\phi(x',w)\mu(dx)$ for some function $\phi\colon \mathcal{X}\times\mathcal{W}\to\mathbb{R}$ and a measure μ on $\mathcal{W}$. Following Theorem 7.11, every $f\in\mathcal{H}$ can be represented by $f(x) = \int \phi(x,w)\varrho_f(w)\mu(dw)$ with some $\varrho_f\in\mathcal{L}_2(\mathcal{W},\mu) := \mathcal{L}_2(\mu)$. Therefore,

$$\mathbb{E}_{\mathsf{P}} f - \mathbb{E}_{\mathsf{Q}} f = \int g(w)\varrho_f(w)\mu(dw),$$

where g is defined by

$$g(w) := \mathbb{E}[\phi(X,w)] - \mathbb{E}[\phi(Y,w)],$$

with $X\sim\mathsf{P}$ and $Y\sim\mathsf{Q}$. Therefore,

$$D_{\mathcal{H}}(\mathsf{P},\ \mathsf{Q}) = \sup_{\varrho\in\mathcal{L}_2(\mu)} \left\{ \int g(w)\varrho(w)\mu(dw) \quad s.t. \quad \|\varrho\|_{\mathcal{L}_2(\mu)} \le 1 \right\} = \|g\|_{\mathcal{L}_2(\mu)}.$$

We just need to calculate the $\mathcal{L}_2(\mu)$ norm of g:

$$\begin{aligned}
\|g\|^2_{\mathcal{L}_2(\mu)} &= \int (\mathbb{E}[\phi(X,w)] - \mathbb{E}[\phi(Y,w)])^2\mu(dw) \\
&= \int (\mathbb{E}[\phi(X,w)] - \mathbb{E}[\phi(Y,w)])(\mathbb{E}[\phi(X',w)] - \mathbb{E}[\phi(Y',w)])^2\mu(dw) \\
&= \mathbb{E}\left[\int (\phi(X,w) - \phi(Y,w))(\phi(X',w) - \phi(Y',w))^2\mu(dw)\right] \\
&= \mathbb{E}\left[K(X,X') + K(Y,Y') - K(X,Y') - K(X',Y)\right] \\
&= \mathbb{E}\left[K(X,X') + K(Y,Y') - 2K(X,Y)\right].
\end{aligned}$$

□

Problem* 7.3. *Consider the infinite neural network in Problem 7.2:*

$$f(x;\ \rho) = \int \sigma(w^\top x)\rho(dw) = \mathbb{E}_\rho[\sigma(w^\top x)],$$

where we assume ρ is a probability measure. Consider learning the optimal ρ (in the space of all distributions), by minimizing the measure square loss:

$$\min_\rho L(\rho) := \mathbb{E}[(f(x;\ \rho) - y)^2].$$

Assume x is drawn from some distribution μ, and $y = f(x;\ \rho^*) + \sigma\xi$, where ρ^* is the unknown true parameter, ξ is an independent standard Gaussian noise and σ is a variance parameter. Prove that

$$L[\rho] = D_{\mathcal{H}}(\rho,\ \rho^*)^2 + \sigma^2.$$

where $D_{\mathcal{H}}(\rho,\ \rho^*)$ denotes the MMD under RKHS $\mathcal{H}$, whose kernel is

$$K(w,w') = \mathbb{E}_{x\sim\mu}[\sigma(w^\top x)\sigma(w'^\top x)].$$

Note that the kernel is defined on the weights (w,w'), instead of x.

7.6 Empirical Estimation of MMD: U-Statistics and V-Statistics

7.7 Energy Distance

7.8 Applications

Two sample tests

Density ratio estimation, domain adaptation, transfer learning.

MMD-GAN, herding.

8 Bayesian Inference

Given an observation x, which assumed to be drawn from a model $p(x \mid \theta)$, where θ is an unknown parameter. Bayesian inference provides a general approach for estimating the unknown θ, as well as quantifying its uncertainty. In Bayesian methods, we assign a prior $\pi(\theta)$ on θ, and calculate the posterior distribution of θ given the observation x:

$$p(\theta \mid x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \propto p(x|\theta)\pi(\theta),$$

where $p(\theta)$ is the marginal distribution of the data, which serves a normalization constant for the posterior distribution of θ.

$$p(\theta) = \int p(x|\theta)\pi(\theta)d\theta.$$

When the observation is $D = \{x_i\}_{i=1}^n$ i.i.d. drawn from $p(x|\theta)$, we have

$$p(\theta \mid D) \propto \left[\prod_{i=1}^n p(x|\theta)\right] \pi(\theta).$$

We can see that the importance of prior π diminishes as the number of data points n increases.

The idea is that the posterior distribution $p(\theta \mid x)$ summarizes all the information regarding θ given observation x, and all the queries regarding θ can be answered from it. For example, we may estimate $\phi(\theta)$ for any function ϕ by the posterior expectation:

$$\mathbb{E}[\phi(\theta) \mid x].$$

We may also construct interval estimates from the posterior distribution. For example, let ψ_α be the α-quantile of $p(\theta|x)$, then $[\psi_{\alpha/2},\ \psi_{1-\alpha/2}]$ can be used a α-credible interval of θ. We should distinguish credible intervals with confidence intervals in frequentist statistics.

We should compare Bayesian methods with frequentist methods: Bayesian methods the estimated parameter θ as a random variable and the observation x as fixed, , whereas frequentist methods treat the data x as random variables and the parameter as a (unknown) fixed value.

Check this interesting blog post, and also this lecture note.

8.1 Bayesian Estimators and Admissibility

Given a distribution family $p(x \mid \theta)$, $\theta \in \Theta$. An estimator (or decision rule) is any measurable function $\delta \colon \mathcal{X} \to \Theta$. The estimator can be deterministic, or random (in which case δ is a distribution conditioning on x). Assume we are interested in constructing an estimator $\delta(x)$ to minimize certain loss function: $L(\delta(x),\ \theta)$. The expected loss function is

$$R(\theta, \delta) = \int p(x|\theta)L(\delta(x), \theta)dx = \mathbb{E}_{x\sim p(\cdot|\theta)}[L(\delta(x), \theta)].$$

Note that the expected loss depends $R(\theta, \delta)$ on the (unknown) parameter θ; two different estimators δ_1 and δ_2 may perform better than the other on different θ. Therefore, the problem of finding the best estimator is a fundamentally a **multi-objective optimization problem**. This fact that the expected

loss depends on the unknown θ is one of the main sophistication of statistical estimation theory. Bayesian estimators can be viewed as resolving this issue using a simple *weighed-sum* approach, which minimizes a weighted sum of loss of different θ, with each preference weight of θ coinciding with its prior $\pi(\theta)$.

We call an estimator δ_1 is **as good as** estimator δ_2, if we have $R(\theta, \delta_1) \le R(\theta, \delta_2)$ for all $\theta \in \Theta$. We call δ_1 **dominates** δ_2 if δ_1 is as good as δ_2, and there exists at least some $\theta \in \Theta$, such that $R(\theta, \delta_1) < R(\theta, \delta_2)$. An estimator δ is called **admissible** if it is not dominated by any other estimator. This is the same as the **Pareto optimality** or Pareto efficiency in multi-objective optimization.

An estimator δ^* is called a **Bayesian estimator**, if there exists a prior π (which is a measure on Θ), such that

$$\delta^* = \arg\min_{\delta} \mathbb{E}_{\theta\sim\pi}[R(\theta,\ \delta)].$$

Under mild conditions, we can show that **Bayesian estimators with positive priors are admissible**; and **all admissible estimators are Bayesian estimators with some prior**, a result known as **complete class theorem** in the literature.

An alternative approach is to consider the **minimax estimators**, defined

$$\delta^* = \arg\min_{\theta} \big\{ \sup_{\theta\in\Theta} R(\theta, \delta) \big\}.$$

Unlike Bayesian estimators which can be automatically derived from Bayesian rules. It is much more difficult to construct minimax estimators (which can only done in a case by case fashion). In addition, being an minimax estimator does not guarantee admissibility.

Let us now discuss the admissibility of Bayesian estimators. It is illustrative to first consider the special case when the space of parameters Θ has finite number of elements.

Theorem 8.1 (**Bayesian estimators with positive priors are admissible**). *Assume $\Theta = \{\theta_1, \dots, \theta_K\}$ is a finite set, and the prior π is positive on all the elements of Θ, then Bayesian estimator with prior π is admissible.*

Proof. If δ_2 dominates δ, we have $R(\theta, \delta_2) \le R(\theta, \delta)$ for all θ and $R(\theta, \delta_2) < R(\theta, \delta)$ for some θ. Because the prior π is strictly positive on all θ, we must have $\mathbb{E}_{\theta\sim\pi}[R(\theta, \delta_2)] < \mathbb{E}_{\theta\sim\pi}[R(\theta, \delta)]$. This contradicts with the assumption that δ is the Bayesian estimator with prior π. □

Theorem 8.2. *Consider the case when $\Theta \subseteq \mathbb{R}^d$. Define the support of a prior π to be the set of points θ such that $\pi(\{\theta' \colon \|\theta' - \theta\| \le \epsilon\}) > 0$ for any $\epsilon > 0$. Assume the support of π is Θ (that is, π is strictly positive on Θ), and $R(\theta, \delta)$ is continuous on θ for any δ. Then the Bayesian estimator of π is admissible.*

Proof. If there exists an δ' that dominates δ, we have $R(\theta, \delta) - R(\theta, \delta') \ge 0$ for all $\theta \in \Theta$, and can find a θ_* such that $\delta_{\theta_*} := R(\theta_*, \delta) - R(\theta_*, \delta') > 0$. Due to the continuous of $R(\theta, \delta) - R(\theta, \delta')$ on θ, we can find $\epsilon > 0$, such that $R(\theta, \delta) - R(\theta, \delta') > \frac{1}{2}\delta_{\theta_*}$ for all $\theta' \in B_{\theta_*}(\epsilon) := \{\theta' \in \Theta \colon \|\theta' - \theta_*\| \le \epsilon\}$. This gives

$$\begin{aligned}\mathbb{E}_{\theta\sim\pi}[R(\theta, \delta)] - \mathbb{E}_{\theta\sim\pi}[R(\theta, \delta')] &\ge \int_{B_{\theta_*}(\epsilon)} (R(\theta, \delta) - R(\theta, \delta'))\pi(d\theta) \\ &\ge \frac{1}{2}\delta_{\theta_*} \times \pi(B_{\theta_*}(\epsilon)) > 0.\end{aligned}$$

But this contradicts with the assumption that δ is the Bayesian estimator with prior π. □

Problem* 8.1 (Unique Bayesian estimators are admissible). *If an estimator δ is the **unique** minimizer of $\mathbb{E}_{\theta\sim\pi}[R(\theta,\ \delta)]$ for some prior π, then δ is admissible.*

Theorem 8.3 (Admissible Estimators are Bayesian Estimators). *Consider the case when $\Theta = \{\theta_1, \ldots, \theta_K\}$ is finite. For any admissible estimator δ, there exists a prior π, such that δ is the Bayesian estimator with prior π.*

Proof. Visualize the risk body. TODO: plot a figure. □

Problem* 8.2 (James-Stein Estimator). *Assume $X \sim \mathcal{N}(\theta, \sigma^2 I)$ where $\theta \in \mathbb{R}^d$ is an unknown mean parameter and the variance σ^2 is assumed to be known. Recall that the maximum likelihood estimator (given an observation of X) is $\delta_0(x) = x$ in this case.*

*Let us consider instead a more general **Shrinkage estimator**:*

$$\delta_a(x) = x - a\frac{x}{\|x\|_2^2},$$

where a is a non-negative coefficient, which controls how much we want to shrink the estimation towards zero. We want to investigate the problem of choosing the optimal a to achieve the optimal mean square error:

$$\mathrm{MSE}(\delta_a) = \mathbb{E}_{X\sim p_\theta}[(\theta - \delta(X))^2],$$

where $p_\theta = \mathcal{N}(\theta, \sigma^2 I)$. Prove that

1) When $d = 1$ or 2, the optimal a equals zero ($a^* = 0$), that is, no shrinkage should be applied.

*2) When $d \geq 3$, the optimal shrinkage coefficient is $a^* = \sigma^2(d-2)$, and $\mathrm{MSE}(\delta_{a^*})$ is strictly smaller than $\mathrm{MSE}(\delta_0)$. This suggests that the maximum likelihood estimator $\delta_0(x) = x$ is **NOT** an admissible estimator in this case. The estimator is known as James-Stein estimator in this case:*

$$\hat{\theta}_{\mathrm{JS}} = x - \sigma^2(d-2)\frac{x}{\|x\|_2^2}.$$

Proof. Taking $g(x) = -a\frac{x}{\|x\|_2^2}$ in theorem 8.4, we have

$$\partial_{x_i} g(x) = -a\frac{1}{\|x\|_2^2} + a\frac{2x_i^2}{\|x\|_2^4}.$$

This gives

$$\nabla \cdot g(x) = -a\frac{d-2}{\|x\|_2^2}.$$

Therefore,

$$\mathrm{MSE}(\delta_a) = d\sigma^2 + \left(a^2 - 2a\sigma^2(d-2)\right)\mathbb{E}[\|x\|_2^{-2}].$$

When $d \leq 2$, we can see that $\mathrm{MSE}(\delta_a) \geq d\sigma^2 = \mathrm{MSE}(\delta_0)$.

When $d > 2$, this minimum is achieved when $a = \sigma^2(d-2)$, in which case we have

$$\mathrm{MSE}(\delta_a) = d\sigma^2 - \sigma^4(d-2)^2\mathbb{E}[\|x\|_2^{-2}] < d\sigma^2 = \mathrm{MSE}(\delta_0).$$

□

{thm:sure}

Theorem 8.4 (Stein's unbiased risk estimate (SURE)). *Let $\mu \in \mathbb{R}^d$ be an unknown parameter and let $x \in \mathbb{R}^d$ be a measurement vector whose components are independent and distributed normally with mean μ and variance σ^2. Suppose $\delta(x)$ is an estimator of μ from x, and can be written $\delta(x) = x + g(x)$, where $g(x) = [g_1(x), \dots, g_d(x)]^\top \in \mathbb{R}^d$ is (weakly) differentiable. Then,*

$$\mathrm{MSE}(\delta) = d \times \sigma^2 + \mathbb{E}[\|g(X)\|_2^2 + 2\sigma^2 \nabla \cdot g(X)],$$

where $X \sim \mathcal{N}(\mu, \sigma^2 I)$, and $\nabla \cdot g(x) = \sum_{i=1}^d \partial_{x_i} g_i(x)$ and $\|g(x)\|_2^2 = \sum_{i=1}^d g_i(x)^2$.

Proof.

$$\begin{aligned}
\mathrm{MSE}(\delta) &= \mathbb{E}[\|\delta(X) - \mu\|_2^2] \\
&= \mathbb{E}[\|X + g(X) - \mu\|^2] \\
&= \mathbb{E}[\|X - \mu\|^2 + \|g(X)\|_2^2 + 2g(X)(X - \mu)] \\
&= d\sigma^2 + \mathbb{E}[\|g(X)\|_2^2] + 2\sigma^2 \mathbb{E}[g(X)(X - \mu)].
\end{aligned}$$

The result then follows Stein's identity $\mathbb{E}[g(X)(X - \mu)] = \sigma^2 \mathbb{E}[\nabla \cdot g(X)]$, for $X \sim \mathcal{N}(\mu, \sigma^2 I)$. □

Problem 8.1. *Please point out any typo and error you find in the note (use the page and line numbers to locate the places).*

Theorem 8.5 (Stein's Identity). *Assume $p(x)$ is a differentiable density function, we have*

$$\mathbb{E}_{x \sim p}[f(x) \nabla_x \log p(x) + \nabla f(x)] = 0.$$

Proof. Let $B_r = \{x \in \mathbb{R}^d \colon \|x\|_2 \le r\}$ be the Euclidean ball with radius r. Assume $p(x)f(x)$ is continuosly differentiable. By Stokes theorem,

$$\int_{\partial B_r} p(x) f(x) dx = \int_{B_r} \nabla \cdot (p(x) f(x)) dx \int_{B_R} \nabla p(x) f(x)$$

□

Theorem 8.6 (Stein's Identity for Normal Distributions). *Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a $\mathbb{R}^d$ Gaussian random variable, and $f \colon \mathbb{R}^d \to \mathbb{R}$ is continuously continuous. We have*

$$\mathbb{E}[X f(X) + f'(X)] = 0.$$

Proof. Note that $f(x) = \int_{-\infty}^x f'(t) dt$, we have

$$\mathbb{E}[\int X \mathbb{I}(Y \le X) f'(Y) dY] = \int \mathbb{E}[X \mathbb{I}(Y \le X) f'(Y)] =$$

□

8.2 PAC Bayesian Bounds

Recall the standard setting of empirical risk minimization:

$$\min_\theta L_S(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(x_i,\ \theta),$$

where $S := \{x_i\}_{i=1}^n$ is an i.i.d. sample from data distribution $\mathcal{D}$. Denote by $L_\mathcal{D}(\theta) = \mathbb{E}_{X\sim\mathcal{D}}[\ell(X,\ \theta)]$ the expected loss. The standard generalization bound, based on Rademacher complexity is: with probability at least $1-\delta$,

$$\sup_{\theta\in\Theta}\left\{L_\mathcal{D}(\theta) - L_S(\theta)\right\} \le 2\mathbb{E}_{S\sim\mathcal{D}^n}[R(\mathcal{F}_\Theta,\ S)] + 2c\sqrt{\frac{\log(2/\delta)}{n}},$$

where $\mathcal{F}_\Theta = \{f(x) = \ell(x,\theta),\ \theta\in\Theta\}$, and $R(\mathcal{F}_\Theta, S)$ is its Rademacher complexity, and $c = \sup_{x,\theta}|\ell(x,\theta)|$.

PAC-Bayesian bounds are different types of generalization that works for randomized estimators. In this case, a randomized estimator is a distribution $\rho(\cdot|S)$ of θ, given data S. whose training and testing loss is $\mathbb{E}_{\theta\sim\rho}[L_S(\theta)]$ and $\mathbb{E}_{\theta\sim\rho}[L_\mathcal{D}(\theta)]$, respectively. PAC-Bayesian inequalities provides a uniform bound between the difference between training and testing losses, which, surprisingly, does not depend on the complexity of the hypothesis class!

{pac:Bayes}

Theorem 8.7. *Assume $\ell(x,\theta)$ is λ-Sub-Gaussian, uniformly in θ, in that*

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\exp\left(\ell(x,\theta) - \mathbb{E}_{x\sim\mathcal{D}}[\ell(x,\theta)]\right)\right] \le \exp\left(\frac{\lambda x^2}{2}\right), \qquad \forall\theta\in\Theta.$$

This is achieved, for example, when $\ell(x,\theta)$ is bounded $\sup_{x,\theta}|\ell(x,\theta)| \le \lambda$. should not it be $\lambda = \sigma^2$?

Let π be prior distribution on θ, We have with probability at least $1-\delta$, we have the following bound holds uniformly for all distributions ρ that are absolutely continuous w.r.t. π (so that $\mathrm{KL}(\rho\ ||\ \pi) < \infty$),

$$\mathbb{E}_\rho[L_\mathcal{D}(\theta)] \le \mathbb{E}_\rho[L_S(\theta)] + \sqrt{\frac{2\lambda(\mathrm{KL}(\rho\ ||\ \pi) + \log(1/\delta))}{n}}, \qquad \forall\rho.$$

The proof of this result relies on a key variational representation of KL divergence.

{em:KLvar}

Lemma 8.8. *Assume ρ is absolutely continuous w.r.t. π, we have*

$$\mathrm{KL}(\rho\ ||\ \pi) = \sup_f\left\{\mathbb{E}_{x\sim\rho}[f(x)] - \log\mathbb{E}_{x\sim\pi}[\exp(f(x))]\right\}.$$

Proof. Define the following f-tilted density function:

$$\pi^f(x) = \frac{\pi(x)f(x)}{\mathbb{E}[\exp(f(x))]}.$$

We have

$$\begin{aligned}\mathrm{KL}(\rho\ ||\ \pi^f) &= \mathbb{E}_{x\sim\rho}[\log\rho(x) - \log\pi(x) - f(x) + \log\mathbb{E}[\exp(f(x))]] \\ &= \mathrm{KL}(\rho\ ||\ \pi) - \left(\mathbb{E}_{x\sim\rho}[f(x)] - \log\mathbb{E}[\exp(f(x))]\right).\end{aligned}$$

Because $\mathrm{KL}(\rho\ ||\ \pi^f) \ge 0$, we have

$$\mathrm{KL}(\rho\ ||\ \pi) \ge \left(\mathbb{E}_{x\sim\rho}[f(x)] - \log\mathbb{E}[\exp(f(x))]\right).$$

In addition, the inequality is tight when $f(x) = \rho(x)/\pi(x)$, so that $\pi^f = \rho$ and hence $\mathrm{KL}(\rho\ ||\ \pi^f) = 0$. □

Proof of Theorem 8.7. Define $\Delta(\theta) := \Delta(\theta;\ S) = L_{\mathcal{D}}(\theta) - L_S(\theta)$. Applying Lemma 8.8, we have

$$\mathbb{E}_{\theta\sim\rho}[\alpha\Delta(\theta)] \le \mathrm{KL}(\rho \,||\, \pi) + \log \mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta))],$$

and hence

$$\mathbb{E}_{\theta\sim\rho}[\Delta(\theta)] \le \frac{1}{\alpha}\mathrm{KL}(\rho \,||\, \pi) + \frac{1}{\alpha}\log \mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta))],$$

where α is a positive number that we will decide later. We just need to show that the second term $\frac{1}{\alpha}\log \mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta))]$ is small. This is true because $\Delta(\theta) = L_{\mathcal{D}}(\theta) - L_S(\theta)$ is the difference of the empirical and expected loss and should decay with the sample size. We can bound $\exp(\alpha\Delta(\theta))$ using Markov inequality.

Recall Markov inequality:

$$\Pr\left(Z \ge \frac{\mathbb{E}[Z]}{\delta}\right) \le \delta.$$

Define $Z := \mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta;\ S))]$, which is a random variable due to the randomness in the data S. Applying Markov inequality on Z, we have with probability at least $1-\delta$

$$\mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta))] \ge \frac{\mathbb{E}_{S\sim\mathcal{D}^n}\left[\mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta))]\right]}{\delta}.$$

We have, with probability $1-\delta$,

$$\mathbb{E}_{\rho}[\Delta(\theta)] \le \frac{1}{\alpha}\mathrm{KL}(\rho \,||\, \pi) + \frac{1}{\alpha}\log\left(\frac{\mathbb{E}_{S\sim\mathcal{D}^n}\left[\mathbb{E}_{\theta\sim\pi}[\exp(\alpha\Delta(\theta))]\right]}{\delta}\right), \qquad \forall\rho.$$

Note that this bound holds true for all ρ uniformly! This is because the term we apply Markov inequality does not depend on ρ (but depends on the fixed "prior" π).

For notation, define $\bar{\ell}(x,\theta) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(x,\theta)] - \ell(x,\theta)$. We have for any θ,

$$\begin{aligned}
\mathbb{E}_{S\sim D^n}[\exp(\Delta(\theta;\ S))] &= \mathbb{E}_{S\sim D^n}\left[\exp\left(\frac{\alpha}{n}\sum_{i=1}^n \bar{\ell}(x_i,\theta)\right)\right] \\
&= \left(\mathbb{E}_{x\sim D}\left[\exp\left(\frac{\alpha}{n}\bar{\ell}(x,\theta)\right)\right]\right)^n \\
&\le \left(\exp\left(\frac{\lambda}{2}\left(\frac{\alpha}{n}\right)^2\right)\right)^n \qquad \text{// Applying Sub-Gaussianity} \\
&= \exp(\lambda\alpha^2/(2n))
\end{aligned}$$

Therefore, we have $\log \mathbb{E}_{\theta\sim\pi}[\mathbb{E}_{S\sim\mathcal{D}^n}[\exp(\alpha\Delta(\theta;\ S))]] \le \lambda\alpha^2/(2n)$. Hence, w.r.t. $1-\delta$

$$\mathbb{E}_{\rho}[\Delta(\theta)] \le \frac{1}{\alpha}\mathrm{KL}(\rho \,||\, p) + \frac{\lambda\alpha}{2n} + \frac{1}{\alpha}\log(1/\delta). \tag{22}$$

Taking the optimal α:

$$\mathbb{E}_{\rho}[\Delta(\theta)] \le \inf_{\alpha\ge 0}\frac{1}{\alpha}(\mathrm{KL}(\rho \,||\, p) + \log(1/\delta)) + \frac{\lambda\alpha}{2n} = \sqrt{\frac{2\lambda}{n}(\mathrm{KL}(\rho \,||\, p) + \log(1/\delta))}.$$

Corollary 8.9. *Under the condition of Theorem 8.7, for any $\delta \in (0,1)$ and $\beta > 0$, we have with probability at least $1-\delta$,*

$$\mathbb{E}_\rho[L_{\mathcal{D}}(\theta)] \le \mathbb{E}_\rho[L_S(\theta)] + \frac{\beta}{\sqrt{n}}\left(\mathrm{KL}(\rho \,||\, \pi) + \log(1/\delta) + \frac{\lambda}{2\beta^2}\right), \qquad \forall \rho.$$

Proof. If we take $\alpha = \sqrt{n}/\beta$ in (22) for any $\beta > 0$, we have

$$\mathbb{E}_\rho[\Delta(\theta)] \le \frac{\beta}{\sqrt{n}}\mathrm{KL}(\rho \,||\, p) + \frac{\lambda}{2\beta\sqrt{n}} + \frac{\beta}{\sqrt{n}}\log(1/\delta).$$

□

There is a been a throughout literature on PAC-Bayesian bounds and its application. The results above can be found **??**. See Guedj [10] for a good recent overview.

α-Divergence α-divergence is a special class of f-divergence,

$$D_\alpha(\rho \,||\, \pi) = \frac{1}{\alpha(\alpha-1)}\Big(\int \rho(x)^\alpha \pi(x)^{1-\alpha} dx - 1\Big),$$

where α is a number in $\mathbb{R}\{0,1\}$, we obtain Helinger divergence when $\alpha = 1/2$, and χ^2-divergence when $\alpha = 2$ or -1. In addition, α-divergence includes KL divergence as limit when α approaches to 0, or 1:

$$\lim_{\alpha\to 1} D_\alpha(\rho \,||\, \pi) = \mathrm{KL}(\rho \,||\, \pi), \qquad \lim_{\alpha\to 0} D_\alpha(\rho \,||\, \pi) = \mathrm{KL}(\pi \,||\, \rho).$$

We can derive a variational form of α-divergence similar to that of (8.8) using Holder's inequality, and hence derive a generalization of PAC-Bayesian bound based on α-divergence.

Lemma 8.10. *For any pair of positive numbers α and β in $(1,+\infty)$ that satisfies $1/\alpha + 1/\beta = 1$,*

$$\mathbb{E}_\rho[\Delta(\theta)] = \int \rho(\theta)\Delta(\theta)d\theta = \int \pi(\theta)\frac{\rho(\theta)}{\pi(\theta)}\Delta(\theta)d\theta \le \left(\int \pi(\theta)\Delta(\theta)^\beta\right)^{1/\beta}\left(\int \frac{\rho(\theta)}{\pi(\theta)}^\alpha \pi(\theta)d\theta\right)^{1/\alpha}.$$

This is rewrite to

$$\mathbb{E}_\rho[\Delta(\theta)] \le ||\Delta||_{\beta,\pi} \times I_\alpha(\rho \,||\, \pi),$$

where $I_\alpha(\rho \,||\, \pi) = \left(\int \rho(x)^\alpha \pi(x)^{1-\alpha} dx\right)^{1/\alpha}$.

We just need to bound $||\Delta||_{\beta,\pi}$. This can be done using the moment bounds of Sub-Gaussian random variables. Assume $\ell(x,\theta)$ is λ-Sub-Gaussian, uniformly for θ, we have

$$\mathbb{E}_{S\sim\mathcal{D}^n}[||\Delta||^\beta_{\beta,\pi}] = \mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{\theta\sim\pi}\left[\Delta(\theta;\ S)^\beta\right] \le C_\beta\left(\frac{2\lambda}{n}\right)^{(\beta-1)/2}$$

where $C_\beta := \frac{\beta\Gamma(\beta/2)}{\sqrt{\pi}}$. Therefore, with probability at least $1-\delta$, we have

$$||\Delta||^\beta_{\beta,\pi} \le \frac{\mathbb{E}_{S\sim\mathcal{D}^n}[||\Delta||^\beta_{\beta,\pi}]}{\delta} \le \frac{1}{\delta}C_\beta\left(\frac{2\lambda}{n}\right)^{(\beta-1)/2},$$

which implies that

$$\mathbb{E}_\rho[\Delta(\theta)] \le \frac{1}{\delta^{1/\beta}}C_\beta^{1/\beta}\left(\frac{2\lambda}{n}\right)^{(\beta-1)/2\beta} I_\alpha(\rho \,||\, \pi).$$

Lemma 8.11. *Let X be a λ-Sub-Gaussian random variable in that $\mathbb{E}[\exp(t(X - \mathbb{E}[X]))] \le \exp(\lambda t^2/2)$ for $t \in \mathbb{R}$. We have*

$$\mathbb{E}[|X|^\beta] \le \frac{\beta\Gamma(\beta/2)}{\sqrt{\pi}}(2\lambda)^{(\beta-1)/2}.$$

Proof. Let us first consider the the case when X is positive valued. Let $\bar{F}(x) = \Pr(X \ge x)$ the tail probability of X. Since X is λ-Sub-Gaussian, we have

$$\bar{F}(x) \le \exp\left(-\frac{x^2}{2\lambda}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}[X^\beta] &= \int_0^\infty x^\beta d\bar{F}(x) \\
&= \beta \int_0^\infty \bar{F}(x) x^{\beta-1} dx \qquad \text{//integration by parts} \\
&\le \beta \int_0^\infty \exp\left(-\frac{x^2}{2\lambda}\right) x^{\beta-1} dx \qquad \text{//tail bound} \\
&= \frac{\beta\Gamma(\beta/2)}{2\sqrt{\pi}}(2\lambda)^{(\beta-1)/2}.
\end{aligned}$$

For general X, we have $X = \max(X, 0) - \max(-X, 0)$. Applying and combing the bound on each part gives the result. □

https://en.wikipedia.org/wiki/Sub-Gaussian_distribution

http://lear.inrialpes.fr/people/harchaoui/teaching/2013-2014/ensl/m2/lecture6.pdf

We should be able to derive similar bounds from IPM, and RKHS???

Proof. □

9 Variational Inference Using Parameteric Families

{sec:para}

In parametric variaional inference, we approximate the target distribution p with a simpler distribution from a parametric family q_θ, indexed with a parameter θ:

$$\min_\theta \{\mathrm{KL}(q_\theta \,||\, p) := \mathbb{E}_{x\sim q_\theta}[\log q_\theta(x) - \log p(x)]\}\,.$$

A key point is that this optimization does not depend on the normalization constant of $p(x)$. To see this, assume $p(x) = \exp(f(x))/Z$, where $Z = \int \exp(f(x))dx$, we have $\mathrm{KL}(q_\theta \,||\, p) = \mathbb{E}_{x\sim q_\theta}[\log q_\theta(x) - f(x)] + \log Z$. where $\log Z$ is a constant that is irrelavant to the optimization of θ, Therefoe, the optimal θ is equivalently

$$\hat\theta = \arg\max_\theta \mathbb{E}_{x\sim q_\theta} f(x) + H(q_\theta).$$

This can be viewed as an entropy regularized optimization. Maximizing the expectation of $\mathbb{E}_{x\sim q_\theta}[f(x)]$ under q_θ, with an entropy regularization on $H(q_\theta)$.

We will consider gradient descent algorithm for solving it

$$\theta \leftarrow \theta - \epsilon \nabla_\theta L(\theta),$$

where ϵ is a step-size. The key question is then how to estimate the gradient. We will introduce two major techniques for gradient estimation, including the score function method and the reparameterization trick. The score function gradient estimator provides an almost universal tool, which does not require to know gradient information of f, and applicable even when f is not differentiable, or x is discrete valued (meanwhile, it does not requires that $\log q_\theta$ exists and differentiable). Reparameterization trick, on the other hand, relies on taking derivative on f, which is often found more efficient when it is available, because it leverages the gradient information of f.

Theorem 9.1. *Assume the support of q_θ does not depend on θ, that is, $\mathrm{supp}(q_) = \mathcal{X}$, and $\mathcal{X}$ does not change with θ. Assume $\log q_\theta(x)$ is differentiable w.r.t. θ on for $x \in X$.*

i) For any function f, we have $L(\theta) := \mathbb{E}_{q_\theta}[f(x)]$ is differentiable, and

efunction}

$$\mathbb{E}_{x\sim q_\theta}[f(x)] = \mathbb{E}_{x\sim q_\theta}[f(x)\nabla_\theta \log q_\theta(x)]. \tag{23}$$

*This is known as the **score function gradient estimator**.*

ii) For $L(\theta) := \mathrm{KL}(q_\theta \,||\, p)$, denote by $f(x) = \log q_\theta(x) - \log p(x)$. We have

u:KLscore}

$$\nabla_\theta \mathrm{KL}(q_\theta \,||\, p) = \nabla_\theta \left(\mathbb{E}_{x\sim q_\theta}[f(x)]\right) \tag{24}$$

$$= \mathbb{E}_{x\sim q_\theta}\left[\left(\log \frac{q_\theta(x)}{p(x)}\right)\nabla_\theta \log q_\theta(x)\right], \tag{25}$$

where the idea is that we do not differentiate θ through f.

Proof. i) Note $\mathbb{E}_{x\sim q_\theta}[f(x)] = \int_{\mathcal{X}} q_\theta(x) f(x) dx$. Because $\mathcal{X}$ does not dependent on θ, we have

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{x\sim q_\theta}[f(x)] &= \int_{\mathcal{X}} \nabla_\theta q_\theta(x) f(x) dx \\
&= \int_{\mathcal{X}} q_\theta(x) \nabla_\theta \log q_\theta(x) f(x) dx \\
&= \mathbb{E}_{x\sim q_\theta}\left[f(x)\nabla_\theta \log q_\theta(x)\right],
\end{aligned}$$

where we used the famous fact of $\nabla_\theta \log q_\theta(x) = \nabla_\theta q_\theta(x)/q_\theta(x)$.

ii) For $\mathrm{KL}(q_\theta \,||\, p)$, using chain rule, we have

$$\nabla_\theta \mathrm{KL}(q_\theta \,||\, p) = \nabla_\theta \left(\mathbb{E}_{x\sim q_\theta}[f(x)]\right) \; + \; \mathbb{E}_{x\sim q_\theta}[\nabla_\theta \log q_\theta(x)],$$

and we need to show that the second term equals zero:

$$\mathbb{E}_{x\sim q_\theta}[\nabla_\theta \log q_\theta(x)] = \int_{\mathcal{X}} q_\theta(x) \frac{\nabla_\theta q_\theta(x)}{p(x)} dx = \int_{\mathcal{X}} \nabla_\theta q_\theta(x) dx = \nabla_\theta (\int q_\theta(x)) dx = 0.$$

See also Lemma 3.4. □

Remark 9.1. *Both (23) and (24) do not hold when the support of $q_\theta(x)$ changes with θ. To give an example, assume $q_\theta(x) = \dfrac{\mathbb{I}(x \in [0,\theta])}{\theta}$, so that $\nabla_\theta \log q_\theta(x) = -\frac{1}{\theta}$ for $x \in [0,\theta]$. We have*

$$\nabla_\theta \mathbb{E}_{q_\theta}[f] = \nabla_\theta \int_0^\theta \frac{1}{\theta} f(x) dx = \int_0^\theta \frac{-1}{\theta^2} f(x) dx + \frac{f(\theta)}{\theta} = \mathbb{E}_{q_\theta}\left[f(x)\nabla_\theta \log q_(x)\right] + \frac{f(\theta)}{\theta},$$

where the second term due to the dependency of support on θ. The score function formula only takes the first term into account.

Following this result, we can construct an unbiased estimator of the gradient by drawing $\{x_i\}_{i=1}^n$ from q_θ and estimate the gradient by

$$\nabla_\theta \mathbb{E}_{q_\theta}[f] \approx G_{score}(\theta) := \frac{1}{n} \sum_{i=1}^n f(x_i) \nabla_\theta \log q_\theta(x_i). \tag{26}$$

proxscore}

Running gradient descent of θ with this Monte Carlo gradient estimation yields a general perhaps inference algorithm. [cite, black box, evaluation strategy, policy gradient, cross entropy].

The formula above can be further generalized by adding an arbitrary "baseline":

$$\mathbb{E}_{x\sim q_\theta}[f(x)] = \mathbb{E}_{x\sim q_\theta}[(f(x) - b)\nabla_\theta \log q_\theta(x)],$$

where b is any real number, whose choice does not change the expectation because $\mathbb{E}_{q_\pi}[\nabla_\theta \log q_\theta(x)] = 0$. Different baselines, however, does influence the variance of the estimator. It is often convenient to choose $b = \bar{f} := \sum_{i=1}^n f(x_i)/n$ in (26), so that we have

$$\nabla_\theta \mathbb{E}_{q_\theta}[f] \approx G_{score}(\theta) := \frac{1}{n} \sum_{i=1}^n (f(x_i) - \bar{f}) \nabla_\theta \log q_\theta(x_i).$$

Therefore, for points x_i that is better than average ($f(x_i) - \bar{f} > 0$), the gradient update increases its likelihood by moving along $\nabla_\theta \log q_\theta(x_i)$, while for points that are worse than the average ($f(x_i) - \bar{f} < 0$), update increases its likelihood by moving along $-\nabla_\theta \log q_\theta(x_i)$.

What is interesting is that the score function estimator is not only way to estimate gradient. There is another important class of gradient estimators, known as the reparameterization trick, which provides significantly different, often much better, estimation.

Theorem 9.2. *Assume $x \sim q_\theta$ is realized by $x = g_\theta(\xi)$, where ξ follows a distribution q_0 that is independent of θ. Function $f(x)$ is differetiable w.r.t. x. We have*

$$\nabla_\theta \mathbb{E}_{x\sim q_\theta}[f(x)] = \mathbb{E}_{x=g_\theta(\xi),\xi\sim q_0}\left[\nabla_x f(x)\nabla_\theta g_\theta(\xi)\right].$$

This estimator is known as ***the reparametrization trick****, or* ***pathwise gradient estimator****.*

Proof. Note that

$$\mathbb{E}_{x\sim q_\theta}[f(x)] = \mathbb{E}_{\xi\sim q_0}[f(g_\theta(\xi))].$$

Applying chain rule:

$$\mathbb{E}_{x\sim q_\theta}[f(x)] = \mathbb{E}_{\xi\sim q_0}[\nabla_x f(g_\theta(\xi))\nabla_\theta g_\theta(\xi)].$$

□

Compared with the score function estimator in (23), the reparameterized gradient estimator requires that $f(x)$ is differentiable w.r.t. x and q_θ can be "reparameterized", but does not require that the support of q_θ to be fixed, nor the existence of the density q_θ (so the probability mass of q_θ can concentrate on a low dimensional space of $\mathbb{R}^d$, similar to GAN).

The reparameterization trick implies a different Monte Carlo gradient estimator, which yields

$$\nabla_\theta L(\theta) \approx G_{rep}(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla_x f(x_i)\nabla_\theta g_\theta(\xi_i),$$

where $\{\xi_i, x_i\}_{i=1}^n$ is i.i.d. drawn by $\xi_i \sim q_0$ and $x_i = g_\theta(\xi_i)$. Compared with the score function estimator in (**??**), the reparameterization estimation is often shown to perform much better (for having smaller variance). Intuitively, we can also see this from the fact that the reparameterized gradient depends on the gradient $\nabla_x f(x)$, which provides gradient information on directions for exploring the landscape of $f(x)$.

Why is reparameterization trick (when it is available) is "in general" better than score function? Let us try to understand this by some problems.

Problem* 9.1 (Comparing Variance of Score-function and Reparameterization Gradient Estimators)**.** *In this problem, we illustrate that reparameterized gradient estimators work better for "smooth" functions with bounded Lipschitz norm (or bounded gradient), while score function estimators tend to work better for bounded, but highly oscillated functions (which perhaps appear less commonly in practice).*

1) Consider linear functions $f(x) = a^\top x + b$. Compare the variance of $G_{rep}(\theta)$ and $G_{score}(\theta)$ and decide which of them has smaller variance.

2) Consider function $f(x) = \frac{1}{\omega}\sin(\omega^2 x)$, and assume ω is very large (e.g., $\omega \to +\infty$). Compare the variance of $G_{rep}(\theta)$ and $G_{score}(\theta)$ and decide which of them has smaller variance.

3) Assume $f(x)$ is L_f-Lipschitz w.r.t x and $g_\theta(\xi)$ is L_g-Lipschitz w.r.t θ for every ξ, so that $\|\nabla_x f(x)\|_2 \le L_f$ and $\|\nabla_\theta g_\theta(\xi)\|_2 \le L_f$ and for all x, ξ, and θ. Prove that

$$\mathbb{E}[\|G_{rep}(\theta)\|_2^2] \le \frac{1}{n}L_f^2 L_g^2.$$

On the other hand, show that the variance of $G_{score}(\theta)$ can be arbitrarily large in this case.

4) Assume f is bounded, that $|f(x)| \le B_f$, and denote by I_θ the Fisher information matrix of q_θ:

$$I_\theta = \mathbb{E}_{x\sim q_\theta}[(\nabla_\theta \log q_\theta(x))^2] = -\mathrm{cov}_{x\sim q_\theta}[(\nabla_\theta \log q_\theta(x))^2].$$

We have

$$\mathbb{E}[\|G_{score}(\theta)\|_2^2] \le \frac{1}{n}\text{trace}(I_\theta).$$

On the other hand, show that the variance of $G_{rep}(\theta)$ can be arbitarily large in this case.

Problem* 9.2. *In this problem, we illustrate that reparameterized estimators work better when q_θ has a small variance, while score function estimators work better when the variance of q_θ is large.*

Assume $q_\theta = \mathcal{N}(\theta,\ \sigma^2 I)$, where σ^2 is the variance.

1) Show (under proper conditions) that when $\sigma \to 0^+$, the variance of $G_{score}(\theta)$ goes to infinite, while that of $G_{rep}(\theta)$ goes to zero.

2) When $\sigma \to +\infty$, show that the variance of $G_{rep}(\theta)$ goes to infinite, while that of $G_{score}(\theta)$ goes to zero.

Problem* 9.3 (Variational Inference with Mixture of Uniform Distributions)**.** *Define* $\text{Unif}(\mu,\ \sigma)$ *to be the uniform distribution on set $\{x\colon\ \|x-\mu\| \le \sigma\}$. Consider approximating p using a mixture of*

$$q_\theta = \sum_{i=1}^m w_i \text{Unif}(\mu_i,\ \sigma_i),$$

where $\theta = [w_i,\ \mu_i,\ \sigma_i]_{i=1}^m$. Design a Monte Carlo gradient descent for estimating the optimal θ. What might be the potential advantage of such approximation compared with, say, mixture of Gaussian distributions?

9.1 Black Box Optimization

The score function trick provides a highly generic framework for deriving gradient-free optimization algorithms, which has been widely applied on various fields. We introduce two examples, evolutionary strategy and policy gradient.

9.1.1 Evolutionary Strategy

Evolutionary strategy (ES) is a set of derivative-free, black-box optimization techniques motivated by ideas of evolution. Mathematically, it is viewed as applying the score function gradient estimation for optimization.

Let $f(x)$ be an non-convex objective function, whose gradient can not be directly accessed. Instead of directly solving $\max_x f(x)$, we look for a distribution q_θ, indexed by some parameter θ, which generates the maxima of $f(x)$. The optimal θ is obtained by

$$\max_\theta \mathbb{E}_{x\sim q_\theta}[f(x)] = \int f(x) q_\theta(x) dx.$$

If $\{q_\theta\}$ includes δ_{x^*}, where x^* is the global optima of $f(x)$, then the optimization of θ is equivalent to the original optimization on x. A typical choice would be $q_\theta = \mathcal{N}(\mu,\ \Sigma)$, where $\theta = [\mu,\ \Sigma]$.

Using Monte Carlo estimation of score function estimator, we have obtain an iterative algorithm:

$$\theta \leftarrow \theta + \epsilon \frac{1}{m}\sum_{i=1}^m f(x_i)\nabla_\theta \log q_\theta(x_i), \qquad \{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} q_\theta.$$

In the literature, ES often refers to the case when we optimize μ, while fixing Σ to be a small diagonal matrix. The case when we optimize μ and Σ jointly is called covariance matrix adaptation evolution strategy (CMA-ES).

9.1.2 Policy Gradient for Reinforcement Learning

Reinforcement learning is the problem of finding optimal policies for sequential decision making problems, typically under unknown enviorment. This involves black box optimization for which score function gradient can be applied to derive policy gradient.

RL is often formulated using Markov decision process. At each time step t, the RL agent observes a state variable s_t, and takes an action a_t according to some policy $\pi(\cdot|s_t)$, which is a distribution conditional on s_t. The agent receives an incremental reward r_t, and the state variable transit to the next step s_{t+1}. Assume this process repeats for $T+1$ time steps, and gives a trajectory $\boldsymbol{\tau} = \{s_t, a_t, r_t\}_{t=0}^T$. The total reward associated with trajectory $\boldsymbol{\tau}$ is often defined as a discounted sum of the local reward:

$$R(\boldsymbol{\tau}) = \sum_{t=0}^{T} \gamma_t r_t, \tag{27}$$

where γ_t is a discount factor for time step t. The goal is to find an optimal policy π, typically from some parametric set $\{\pi_\theta : \theta \in \Theta\}$, to maximize the expected reward function:

$$\max_\theta \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}[R(\boldsymbol{\tau})],$$

where we use p_{π_θ} to denote the distribution of the trajectory $\boldsymbol{\tau}$ when we use policy π_θ, whose density has the form of

$$p_\theta(\{s_t, a_t\}) = p_0(s_0)\prod_{t=0}^{T} T(s_{t+1} \mid s_t,\ a_t)\pi_\theta(a_t \mid s_t),$$

where p_0 denotes the initial distribution of s_0 and $T(\cdot|s,a)$ is the transition probability. Therefore, using the score function trick, it is not difficult to see that

$$\nabla_\theta \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}[R(\boldsymbol{\tau})] = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}\left[(R(\boldsymbol{\tau}) - b)\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\right],$$

where $R(\boldsymbol{\tau})$ is any baseline constant. This gradient can be estimated by Monte Carlo using samples from policy π_θ. This algorithm is called REINFORCE [28]. What is nice is that this formula does not require to explicitly estimate the unknown transition model $T(\cdot|s,a)$. It is hence called a model-free algorithm. In comparison, methods that explicitly estimate the transition models and leverage it to estimate and optimize reward are called model-based methods.

However, further simplification is can be constructed by exploiting the Markov structure and the additive structure of the reward function $R(\boldsymbol{\tau})$ in (27). In particular, note that for any $t' < t$, we have

$$\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}\left[r_{t'}\nabla_\theta \log \pi_\theta(a_t|s_t)\right] = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}\left[r_{t'}\mathbb{E}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \mid s_t\right]\right] = 0, \qquad \text{for any } t' < t.$$

Therefore, we can rewrite the gradient into

$$\nabla_\theta \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}[R(\boldsymbol{\tau})] = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_\theta}}\left[(Q_t - b_t)\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\right].$$

where $Q_t = \sum_{t' \geq t} \gamma_{t'} r_{t'}$, and b_t is any baseline function that does not depend on a_t and all the subsequent trajectory (that is, $\{a_t\} \cup \{a_{t'}, s_{t'}, r_{t'}\}_{t'>t}$). This result is often called policy gradient theorem. See the Sutton & Barto book [25], and Deisenroth et al. [5] (Section 2.4).

10 Stein Variational Methods

Stein's method is a technique from probability theory for bounding the distance between probability measures using differential and difference operators. Although the method was initially designed as a technique for proving central limit theorems, its key idea has recently applied in machine learning for developing practical computational tools for probabilistic learning and inference. Recent applications include variational inference, generative modeling, variance reduction, goodness of fit tests, among many others. We start with an introduction of the Stein's method in its original form, and then discuss its application to probabilistic inference and learning.

10.1 Stein's Method: Overview

Theorem 10.1. *Let $S = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$, where X_i are i.i.d. with mean zero and finite first three order moments: $\sigma^2 = \mathbb{E}[X_1^2] < \infty$, and $\mathbb{E}[|X_1|^3] < \infty$. Let f be a second-order differentiable function and has bound second order derivative ($\|f''\|_\infty < \infty$). We have*

$$\left| \mathbb{E}[Sf(S) - \sigma^2 f'(S)] \right| \leq \frac{1}{\sqrt{n}} \|f''\|_\infty \mathbb{E}[|X_1|^3].$$

Proof. Construct $S_1 = S - \frac{X_1}{\sqrt{n}} + \frac{\tilde{X}_1}{\sqrt{n}}$, where $\tilde{X}_1$ is another independent random copy of X_1. Note that S and S_1 shares the same distribution as S (and hence (S, S') is an exchangeable pair), and satisfies $S - S_1 = \frac{1}{\sqrt{n}}(X_1 - \tilde{X}_1)$, and S_1 is independent with X_1, so that $\mathbb{E}[X_1 f(S_1)] = 0$ for any f. The fact that we are able to construct such an exchange pair is the key of the proof.

$$\begin{aligned}
\mathbb{E}[Sf(S)] &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[X_i f(S)] \\
&= \sqrt{n} \mathbb{E}[X_1 f(S)] \\
&= \sqrt{n} \mathbb{E}[X_1 (f(S) - f(S_1))].
\end{aligned}$$

Applying Taylor expansion, we have

$$f(S) - f(S_1) = f'(S_1)(S - S_1) \ + \ R(S, S_1).$$

where $R(S, S_1) = \int_0^1 \left[f'(S_1 + u(S - S_1)) - f'(S_1) \right] (S - S_1) du$.

Therefore

$$\begin{aligned}
\mathbb{E}[Sf(S)] &= \sqrt{n}\mathbb{E}[X_1(f(S)-f(S_1))] \\
&= \sqrt{n}\mathbb{E}[X_1 f'(S_1)(S-S_1)] + \Delta \qquad //\text{Define } \Delta = \sqrt{n}\mathbb{E}\left[X_1 R(S,S_1)\right] \\
&= \mathbb{E}[X_1(X_1-\tilde{X}_1)f'(S_1)] + \Delta \qquad //S-S_1 = (X_1-\tilde{X}_1)/\sqrt{n} \\
&= \mathbb{E}[X_1^2 f'(S_1) - X_1\tilde{X}_1 f'(S_1)] + \Delta \\
&= \mathbb{E}[X_1^2]\mathbb{E}[f'(S_1)] - \mathbb{E}[X_1]\mathbb{E}[\tilde{X}_1 f'(S_1)] + \Delta \qquad //\text{By independence between } X_1 \text{ and } \tilde{X}_1,\ S_1 \\
&= \sigma^2\mathbb{E}[f'(S_1)] + \Delta \\
&= \sigma^2\mathbb{E}[f'(S)] + \Delta.
\end{aligned}$$

This suggests that

$$\mathbb{E}[Sf(S) - \sigma^2 f'(S)] = \Delta.$$

We just need to bound Δ.

$$\begin{aligned}
|\Delta| &= \left|\sqrt{n}\mathbb{E}\left[X_1 R(S,S_1)\right]\right| \\
&= \left|\mathbb{E}\left[|X_1|\int_0^1 \left[f'(S_1+u(S-S_1)) - f'(S_1)\right](S-S_1)du\right]\right| \\
&\le \mathbb{E}\left[|X_1|\int_0^1 \left\|f''\right\|_\infty u(S-S_1)^2 du\right] \\
&\le \left(\int_0^1 u du\right)\cdot \mathbb{E}\left[|X_1|\left\|f''\right\|_\infty (S-S_1)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[|X_1|\left\|f''\right\|_\infty (S-S_1)^2\right] \\
&= \frac{1}{2\sqrt{n}}\left\|f''\right\|_\infty \mathbb{E}\left[|X_1|(X_1-\tilde{X}_1)^2\right] \\
&= \frac{1}{2\sqrt{n}}\left\|f''\right\|_\infty \mathbb{E}\left[|X_1|^3 - 2|X_1|X_1\tilde{X}_1 + |X_1\tilde{X}_1|^2\right] \\
&\le \frac{1}{2\sqrt{n}}\left\|f''\right\|_\infty \mathbb{E}\left[|X_1|^3 + |X_1\tilde{X}_1|^2\right] \qquad //\mathbb{E}[|X_1|X_1\tilde{X}_1] = \mathbb{E}[|X_1|X_1]\mathbb{E}[\tilde{X}_1] = 0 \\
&\le \frac{1}{\sqrt{n}}\left\|f''\right\|_\infty \mathbb{E}\left[|X_1|^3\right],
\end{aligned}$$

where the last step is based on Generalized mean inequality: $\mathbb{E}[|X|^\alpha]^{1/\alpha} \le \mathbb{E}[|X|^\beta]^{1/\beta}$ for any $\alpha < \beta$,

$$\mathbb{E}[|X_1\tilde{X}_1|^2] = \mathbb{E}[|X_1|]\mathbb{E}[|\tilde{X}_1|^2] = \mathbb{E}[|X_1|]\mathbb{E}[|X_1|^2] \le \mathbb{E}[|X_1|^3]^{1/2}\mathbb{E}[|X_1|^3]^{2/3} = \mathbb{E}[|X_1|^3].$$

□

See more https://arxiv.org/pdf/1109.1880.pdf

References

[1] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

[2] Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*.

[3] Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681.

[4] Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*.

[5] Deisenroth, M. P., Neumann, G., Peters, J., et al. (2013). A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142.

[6] Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

[7] Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[9] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

[10] Guedj, B. (2019). A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*.

[11] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777.

[12] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

[13] Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

[14] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[15] Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213.

[16] Liu, S., Bousquet, O., and Chaudhuri, K. (2017). Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553.

[17] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

[18] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

[19] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279.

[20] Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

[21] Rudin, W. (2017). *Fourier analysis on groups*. Courier Dover Publications.

[22] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

[23] Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[24] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009). On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

[25] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[26] van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

[27] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

[28] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

[29] Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. (2017). On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*.

A Measure Theory, Probability Measures, Random Variables

A **measurable space** $(\Omega, \mathcal{A})$consists of a set Ω, and $\mathcal{A}$ is a *σ-algebra* on Ω, which is a collection of subsets of Ω that includes the empty set $\emptyset$ and Ω itself, and is closed under complement, and is closed under countable unions (this definition implies that it also includes the empty subset and that it is closed under countable intersections). Each element of $\mathcal{A}$ is called a **measurable set**, or an **event**.

In most cases we consider, Ω is a finite Euclidean space $\mathbb{R}^d$, or its subset, for which we always assume $\mathcal{A}$ is generated from open sets (or, equivalently, from closed sets) through the operations of countable union, countable intersection, and relative complement. This σ-algebra, denoted by $\mathcal{B}(\mathbb{R}^d)$, is the smallest possible σ-algebra that includes all open sets (and hence called **Borel σ-algebra**) on $\mathbb{R}^d$. We will simply use $\mathbb{R}^d$ to denote measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

A **measure** μ on $(\Omega, \mathcal{A})$ is a function from $\mathcal{A}$ to $\mathbb{R}_+ \cup \{0, +\infty\}$, which maps each measurable set $A \in \mathcal{A}$ to a non-negative number that represents the "volume" or "size" of A. A measure should satisfy a few intuitive properties, including $\mu(\emptyset) = 0$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$ when $\{A_i\}$ are pairwise disjoint. We may use P or Q to denote probability measures.

A measure is called a **finite measure** if $\mu(\Omega) < \infty$. If a measure satisfies $\mu(\Omega) = 1$, it is called a **probability measure**, **(probability) distribution law**, or simply **(probability) distribution**. In this case, the triplet $(\Omega, \mathcal{A}, \mu)$ is called a **probability space**.

Lebesgue Measure on $\mathbb{R}^d$ (equipped with its Borel σ-algebra) is the typical measure we encounter in multiple dimensional integration, which assigns to every rectangle its d-dimensional volume in the ordinary sense.

The **(unit) Dirac measure** at $\delta \in \Omega$, denoted by δ_a is defined to be $\delta_a = \mathbb{I}(x \in A)$, that is,

$$\delta_a(A) = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{if } a \notin A. \end{cases}$$

Dirac measure can be equipped with any σ-algebra that includes $\{a\}$. We may denote δ_a by δ when $a = 0 \in \Omega$.

When the size Ω is finite, the **counting measure** (equipped with power set σ-algebra $\mathcal{A} = 2^\Omega$) assigns each subset of Ω the number of elements in it, that is, for $A = \{a_1, \dots, a_m\} \subset \Omega$, we have $\mu(A) = m$.

A **measurable function** is a function from a measurable space $(\Omega_1, \mathcal{A}_1)$ to measurable space $(\Omega_1, \mathcal{A}_1)$ such that the preimage of any measurable set is measurable, that is, if $A \in \mathcal{A}_2$, we must have $f^{-1}(A) \in \mathcal{A}_1$, where f^{-1} denotes the inverse map of f.

Given a measure μ and a measurable function f on a measurable space $(\Omega, \mathcal{A})$, one can define the **Lebesgue integral** $\int_\Omega f(x)d\mu(x)$, which may be abbreviated as $\int_\Omega f d\mu$ or $\int f d\mu$ The rigorous definition of Lebesgue integral is based on decomposing f as limits of weighted sums of indicator functions, e.g., $f(x) \approx \sum_i w_i \mathbb{I}(x \in A)$ (here $\mathbb{I}(z)$ denotes the 0/1-indicator function, which equals 0 if $z = 0$ and 1 otherwise), and elementary rules like $\mu(A) = \int \mathbb{I}(x \in A)d\mu$, and $\int af(x)d\mu = a \int f(x)d\mu$ for $a \in \mathbb{R}$. When μ is Lebesgue measure, we simple write $\int f(x)d\mu(x) = \int f(x)dx$.

A function f is called *Lebesgue-integrable* w.r.t. μ if $\int |f|d\mu < \infty$.

A **random variable** X is a measurable function from a probability space $(\Omega, \mathcal{A}, \mathsf{P})$ (called the sample space) to another measurable space (called the state space). The state space usually taken to be a real number of vector $\mathbb{R}^d$ with the Borel σ-algebra, so $X \colon \Omega \to \mathbb{R}^d$. For simplicity, let us always consider a $\mathbb{R}^d$-valued random variables.

For an $\mathbb{R}^d$-valued random variable X, every element $\omega \in \Omega$ is mapped to a value $X(\omega) \in \mathbb{R}^d$. For any measurable set B in $\mathbb{R}^d$, its preimage $X^{-1}(\omega) := \{\omega \in \Omega | X(\omega) \in B\}$ is a measurable set (and event) in $(\Omega, \mathcal{A}, \mu)$ by definition. Here X^{-1} denotes the inverse map of X (remember that X is a measurable function technically). $X^{-1}(\omega)$ is often also written as $\{X \in B\}$, and is sometimes just called "the event that $X \in B$."

Each $\mathbb{R}^d$-random variable X is characterized by its probability distribution or law, denoted by P_X, on $\mathbb{R}^d$, defined by $\mathsf{P}_X(B) = \mathsf{P}(X \in B)$.

The statistical properties of the random variable X is fully characterized by its law P_X. Notice here that P_X is a measure (in fact a probability measure) on $\mathbb{R}^d$, as opposed to the original measure P, which is a measure on $(\Omega, \mathcal{A})$. In most cases, the original probability space $(\omega, \mathcal{A}, \mathsf{P})$ remains in the background, hidden or unused, and one works directly with the much more tangible probability space on $\mathbb{R}^d$ (which is technically $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathsf{P}_X)$).

For an $\mathbb{R}$-valued random variable X, The function $F_X \colon \mathbb{R} \to [0, 1]$, defined by

$$F_X(x) = \mathsf{P}_X((-\infty, x]) = \mathsf{P}(X \le x),$$

is called the cumulative distribution function (CDF) of X. Here we use the lower case x to denote a deterministic value, and upper case X a random variable. F is also simply referred as the **distribution function**, but it should not be confused with the distribution, or law of X, which is P_X. Meanwhile, the probability law P_X of a random variable X is uniquely determined by the CDF F_X.

When the image (or range) $X(\Omega) := \{X(\omega) \colon \omega \in \Omega\}$ of X is finite or countably infinite, the random variable is called a **discrete random variable**, and its distribution can be described by a probability mass function $p_X(x) = \mathsf{P}(X = x)$, which assigns a probability to each value in the image of X.

If the image is uncountably infinite then X is called a continuous random variable. The **probability density function (PDF)** of X, if it exits, is the function p_X that satisfies

$$\mathsf{P}(X \in B) = \int_B p_X(t)dt,$$

for all measurable set B.

B Rademacher Complexity

Let P be a probability measure, and P_n be an empirical measure of an i.i.d. sample $S = \{x_1, \dots, x_n\}$ of size n drawn from P. Define

$$\hat{U}_{n,P}(\mathcal{F};\ S) := \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_P[f(x)] \right|,$$

and $\hat{U}_n$ is called empirical representiveness of function class $\mathcal{F}$ under measure P. Taking the expectation:

$$U_{n,P}(\mathcal{F}) := \mathbb{E}_P[\hat{U}_{n,P}(\mathcal{F};\ S)] = \mathbb{E}_P\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_P[f(x)] \right| \right],$$

where U_n is called the representiveness of function class $\mathcal{F}$ under measure P. Note that Markov inequality, $U_n(\mathcal{F}) \to 0$ implies that $\hat{U}_n(\mathcal{F}) \xrightarrow{p} 0$.

Rademacher complexity provides a powerful approach to upper bound the representiveness. Define the empirical Rademacher complexity to be

$$\hat{R}_{n,P}(\mathcal{F}, S) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right],$$

where σ_i are i.i.d. random variables with $\Pr(\sigma_i = 1) = \Pr(\sigma_i = -1) = 1/2$; they are called Rademacher random variables in this context. Expectation $\mathbb{E}_\sigma[\cdot]$ is w.r.t. $\{\sigma_i\}$. The Rademacher complexity is

$$R_{n,P}(\mathcal{F}) = \mathbb{E}_P[\hat{R}_{n,P}(\mathcal{F}, S)] = \mathbb{E}_P\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right],$$

where $\mathbb{E}_P[\cdot]$ is w.r.t. $S = \{x_i\}$.

Lemma B.1. *The expected representativeness is upper bounded by Rademacher complexity:*

$$U_{n,P}(\mathcal{F}) \le 2R_{n,P}(\mathcal{F}).$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\{x_i\}\sim P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_P[f(x)]\right|\right] &= \mathbb{E}_{\{x_i\}\sim P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(x_i) - \mathbb{E}_P[f(x_i')])\right|\right] \\
&\le \mathbb{E}_{\{x_i,x_i'\}\sim P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(x_i) - f(x_i'))\right|\right] \\
&= \mathbb{E}_{\{x_i,x_i'\}\sim P}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(x_i) - f(x_i'))\right|\right] \qquad //x_i,\ x_i' \text{ are symmetric} \\
&\le \mathbb{E}_{\{x_i,x_i'\}\sim P}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i')\right|\right)\right] \\
&\le \mathbb{E}_{\{x_i,x_i'\}\sim P}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i')\right|\right] \\
&= 2\mathbb{E}_{\{x_i\}\sim P}\mathbb{E}_\sigma\left[\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right|\right] \\
&= 2R_{n,P}(\mathcal{F}).
\end{aligned}$$

□

By Markov inequality, we can already provide a high probability bound of $\hat{U}_{n,P}(\mathcal{F};\ S)$ using Rademacher complexity:

$$\Pr\left(\hat{U}_{n,P}(F;\ S) \ge \epsilon\right) \le \frac{U_{n,P}(\mathcal{F})}{\epsilon} \le \frac{2R_{n,P}(\mathcal{F})}{\epsilon}.$$

This bound, however, can be significantly improved when the function class $\mathcal{F}$ is uniformly bounded, that is, $||f||_\infty \le M$, $\forall f \in \mathcal{F}$ for some $M < \infty$. This can be achieved using **McDiarmid inequality**.

Theorem B.2 (McDiarmid Inequality)**.** *Assume function $\rho(x_1, \dots, x_n)$ satisfies*

$$\sup_{x_1,\dots,x_n;x_i'} |\rho(x_1, \dots, x_i, \dots, x_n) - \rho(x_1, \dots, x_i, \dots, , x_n)| \le c_i, \qquad \forall i = 1, \dots, n,$$

where $c_i < \infty$ is some constant. Effectively, this suggests that $\rho(x_1, \dots, x_n)$ is 1-Lipschitiz w.r.t. the weighted 0/1 distance $d_0(x, x') := \sum_i c_i \mathbb{I}(x_i - x_i')$.

Then when $x_1, \dots, x_n$ are i.i.d. drawn from some distribution P, we have

$$\Pr\left(\rho(x_1, \dots, x_i, \dots, x_n) - \mathbb{E}_P[\rho(x_1, \dots, x_i', \dots, x_n)] \ge \epsilon\right) \le \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

Proof. Using Martingale method.

$$\rho(\boldsymbol{x}) - \mathbb{E}[\rho(\boldsymbol{x}')] = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{x}'\sim P}[\rho([\boldsymbol{x}_{1:i}; \boldsymbol{x}'_{i+1:n}]) - \rho([\boldsymbol{x}_{1:i-1}; \boldsymbol{x}'_{i:n}])] := \sum_{i=1}^n \Delta_i(\boldsymbol{x}_{1:i})$$

where each $|\Delta_i(\boldsymbol{x}_{i:i})| \le c_i$ and

$$\mathbb{E}[\Delta_i \mid \boldsymbol{x}_{1:i-1}] = 0.$$

We have

$$\begin{aligned}\mathbb{E}[\exp(t(\rho(\boldsymbol{x}) - \mathbb{E}[\rho(\boldsymbol{x})]))] &\le \mathbb{E}[\exp(t\sum_{i=1}^{n}\Delta_i)] \\ &= \mathbb{E}[\mathbb{E}[\exp(t\Delta_n + t\sum_{i=1}^{n-1}\Delta_i \mid \boldsymbol{x}_{1:n-1}]] \\ &\le \exp(\frac{c_2 t^2}{2})\end{aligned}$$

TODO. □

quadratic}

Lemma B.3. *Let X be a random variable on $\mathbb{R}$ with finite moment generating function $\mathbb{E}[\exp(tX)]$ for some $t \in \mathbb{R}_+$, we have*

$$\Pr(X \ge \epsilon) \le \inf_{t\ge 0} \frac{\mathbb{E}[\exp(tX)]}{\exp(t\epsilon)}.$$

Proof. Applying Markov inequality on $\exp(tX)$. □

Lemma B.4. *If random variable X is bounded in interval $[a, b] \subset \mathbb{R}$, we have*

$$\phi(t) := \log \mathbb{E}[\exp(tX)] \le \mu t + \frac{(a-b)^2 t^2}{8}.$$

Combining with Lemma B.3, we have

$$\Pr(X - \mu \ge \epsilon) \le \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right).$$

Proof. Taking the derivative:

$$\begin{aligned}\phi'(t) &= \frac{\mathbb{E}[\exp(tX)X]}{\mathbb{E}[\exp(tX)]} = \mathbb{E}[\tilde{X}] \\ \phi''(t) &= \frac{\mathbb{E}[\exp(tX)X^2]}{\mathbb{E}[\exp(tX)]} - \left(\frac{\mathbb{E}[\exp(tX)X]}{\mathbb{E}[\exp(tX)]}\right)^2 = \mathrm{var}(\tilde{X}),\end{aligned}$$

where $\tilde{X}$ is the random variable with the exponentially tiltled law: $\tilde{P}(dx) = \frac{\exp(tx)P(dx)}{\mathbb{E}[\exp(tX)]}$, where P is the law of X. Because X is bounded in $[a, b]$, $\tilde{X}$ must also be contained in $[a, b]$. Therefore, we have

$$\phi''(t) = \mathrm{var}(\tilde{X}) \le (a-b)^2/4.$$

On the other hand, note that $\phi(0) = 0$ and $\phi'(t) = \mathbb{E}[X] = \mu$. We have

$$\phi'(t) = \phi'(0) + \int_0^t \phi''(t)dt \le \mu + (a-b)^2 t/4.$$

$$\phi(t) = \phi(0) + \int_0^t \phi'(t)dt \le \int_0^t (\mu + (a-b)^2 t/4)dt = \mu t + (a-b)^2 t^2/8.$$

And hence by Lemma B.3

$$\Pr(X - \mu \ge \epsilon) \le \inf_{t\ge 0} \exp(\phi(t) - t(\epsilon + \mu)) \le \inf_{t\ge 0} \exp\left(\frac{(a-b)^2 t^2}{8} - t\epsilon\right) = \exp\left(-\frac{2\epsilon^2}{(a-b)^2}\right)$$

□

B.1 Sub-Gaussian Random Variables

Definition B.5. *A random variable X on $\mathbb{R}$ is called sub-Gaussian with variance proxy σ^2 if*

$$\mathbb{E}[\exp(t(X - \mathbb{E}X))] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right)$$

In this case, we write $X \sim \mathrm{SubG}(\sigma^2)$.

Theorem B.6. *Let $X \sim \mathrm{SubG}(\sigma^2)$. Then for any $t > 0$, it holds*

$$\max\{\Pr(X \geq t), \quad \Pr(X \leq -t)\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

Proof. □

C Convex Conjugate

convexity}

We introduce some background on convex conjugate. For any function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$, its convex conjugate f^* is defined by

$$f^*(x) = \sup_t \{t^\top x - f(t)\}.$$

We may apply the definition to get the double conjugate f^{**}:

$$f^{**}(x) = \sup_t \{t^\top x - f^*(t)\}.$$

By Fenchel-Moreau theorem, $f = f^{**}$ if and only if f is convex and lower semi-continuous.

Theorem C.1. *For any function f, we have $f^{**} \le f^*$. In addition, if f is convex and lower semi-continuous, we have $f^{**} = f$, that is,*

$$f(x) = \sup_t \left\{ t^\top x - f^*(t) \right\}.$$

Proof.

$$\begin{aligned}
f^{**}(x) &= \sup_t \left\{ t^\top x - f^*(t) \right\} \\
&= \sup_t \left\{ t^\top x - \sup_s \left\{ t^\top s - f(s) \right\} \right\} \\
&= \sup_t \inf_s \left\{ t^\top x - t^\top s + f(s) \right\} \\
&\le \inf_s \sup_t \left\{ t^\top x - t^\top s + f(s) \right\} \\
&= \inf_s \left\{ \sup_t t^\top (x - s) + f(s) \right\} \\
&= \inf_s \{ f(s) \colon \quad s.t. \;\; x = s \} \\
&= f(x),
\end{aligned}$$

where in the last two steps, we use the fact that

$$\sup_t t^\top (x - s) = \begin{cases} +\infty & \text{if } x \ne s \\ 0 & \text{if } x = s. \end{cases}$$

Denote by $L(s,t) = t^\top (x - s) + f(s)$. Then $f^{**} = f$ is equivalent to

$$\sup_t \inf_s L(s,t) = \inf_s \sup_t L(s,t),$$

which is expected to be true when $L(s,t)$ is continuous, and convex on s and concave on t. The rigorous proof is more technical. □

For *differentiable* convex functions, there is an simpler elementary argument for convex conjugacy. Note that for convex functions f, all the tangent lines are beneath the curve of f, that is,

$$f(x) \ge f(y) + (x - y)^\top \nabla f(y), \quad \forall x, y.$$

Meanwhile, the equality is achieved when $x = y$. This suggests that

$$f(x) = \sup_y \left\{ f(y) + (x - y)^\top \nabla f(y) \right\}.$$

This simple representation is fact equivalent to the dual representation shown above. To see this, note that

$$\begin{aligned} f(x) &= \sup_y \left\{ f(y) + (x - y)^\top \nabla f(y) \right\} \\ &= \sup_y \left\{ x^\top \nabla f(y) - (y^\top \nabla f(y) - f(y)) \right\} \end{aligned}$$

Define $t = \nabla f(y)$ and assume ∇f is an one-to-one map and invertible. Denote by $y = g(t) = \nabla f^{-1}(t)$. We have

$$\begin{aligned} f(x) &= \sup_y \left\{ x^\top \nabla f(y) - (y^\top \nabla f(y) - f(y)) \right\} \\ &= \sup_t \left\{ x^\top t - (g(t)^\top t - f(g(t))) \right\}. \end{aligned}$$

from which we can read that $f^*(t) = g(t)^\top t - f(g(t)))$ and the optimality is achieved when $x = y$, which is equivalent to $\nabla f(x) = t$.

D Learning and Inference as Approximating Probabilities

Probabilistic modeling provides a predominant framework for reasoning under certainty. Essentially all learning and inference problems can be reduced to *matching or approximating probabilities from one form to other one*. The forms of probability that appears mostly falls into one of the three categories.

Empirical Data Empirical observation is often available to us through a set of data points $\{\boldsymbol{x}_i\}$, where each $\boldsymbol{x}_i$ is vector, representing an objective, in some space. For example, in computer vision, we observe a set of images, so that each x_i represent an image, with each coordinate representing a pixel. In natural language, each x_i is a sequence of words. In speech processing, each x_i represents a wave signal, which is a time series. We can represent data as an empirical distribution:

$$p(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\boldsymbol{x} - \boldsymbol{x}_i),$$

where δ denotes a Dirac delta function. This assigns a uniform probability on each data point $\boldsymbol{x}_i$, while zero probability everywhere else.

For statistical learning, we are often interested in constructing *smoother* distributions to represent the data, so that we can assign non-zero probabilities to data points that we do not observe, based on its similarity with points inside the data set. The ability of building useful models for unseen data is called *generalization*.

Simple, Tractable Distributions There exists a set of classical distributions for which (almost) everything is computationally tractable. For example, Gaussian distribution

It is tractable to calculate the density function,

$$p(\boldsymbol{x};\ \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi\det(\Sigma))^{d/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Not that the density is properly normalized, with $\int p(\boldsymbol{x};\ \boldsymbol{\mu}, \Sigma)d\boldsymbol{x} = 1$.

It is tractable to draw sample from Gaussian. Let $\boldsymbol{\xi} \sim \mathcal{N}(0,\ I)$, then

$$\boldsymbol{x} := \Sigma^{1/2}\xi + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu},\ \Sigma).$$

It is tractable to calculate the polynomial moments and moment generating function:

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}, \qquad \mathrm{cov}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^\top] = \Sigma,$$

and

$$\mathbb{E}[\exp(\boldsymbol{t}^\top\boldsymbol{x})] = \exp\left(\boldsymbol{t}^\top\mu + \frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t}\right).$$

But for more complex functions f, it may be intractable to calculate $\mathbb{E}_p[f]$, even when p is Gaussian.

Density function. The distribution is available through its density function $p(x)$.

Simulator. The distribution is available through a simulator that generates samples from p. However, the density of p is not available.

A special case of simulator is empirical observation. The distribution is observed through an empirical data $\{x_i\}$, viewed as an empirical distribution $p(x) = \sum_{i=1}^n \delta(x - x_i)/n$. More generally, one may also observe a set of weighted sample $\{x_i, w_i\}$, where $w_i \in \mathbb{R}$, $w_i \geq 0$, $\sum_i w_i = 1$, and the empirical distribution $p(x) = \sum_{i=1}^n w_i \delta(x - x_i)$. This appears in, for example, reinforcement learning.

Inference and learning then reduces to transforming distributions from one form to another one. Including

1. Model estimation (learning). Given an empirical data $\{x_i\}$, viewed as an empirical distribution $p(x) = \sum_{i=1}^n \delta(x - x_i)/n$. We want to find a parametric (or nonparametric model).
2. Bayesian inference. Given a distribution in the form of density $p(x)$, a sample based representation.
3. Black-box optimization and reinforcement learning. In many cases, we have access to an black box function $J(z)$ which represents the reward given a input parameter z, and we want to find the optimal z to optimize J. A particular case of this reinforcement learning.

   The nonconvex optimization problem can be viewed as special cases of sampling problem by using simulated annealing trick.

	(Normalized) Density	Simulator	Data
(Unormalized) Density	Parametric VI	Amortized SVGD	MCMC, SVGD
Simulator	MLE	N/A	N/A
Data	MLE (CD for unnormalized densities)	GAN	Compression? (Herding)

Input	Output	Metric	Algorithm	Reference	Note
Sample	Generator	(Neural) JS divergence (Neural) f divergence (Neural) IPM Maximum mean discrepancy Energy distance	GAN f-GAN W-GAN MMD-GAN Cramer-GAN	Goodfellow et al. [8] Nowozin et al. [19] Arjovsky et al. [1] Li et al. [15], Dziugaite et al. [7] Bellemare et al. [4]	
Sample	Unnormalized	KL divergence KL divergence (approximate) Composite KL divergence Fisher divergence Stein discrepancy pseudolikelihood	MLE MCMC-MLE / CD / Variational Bayesian Composite Likelihood Contrastive divergence (CD) Stein CD CD with parallel Gibbs		
Unormalized	Sample	KL divergence Stein discrepancy Maximum mean discrepancy Wasserstein distance Energy distance	SVGD / Langevin / reversible MCMC Stein points Herding Wasserstein Variational Gradient Descent ?		
Unormalized	Generator	KL divergence Neural Stein discrepancy MMD Wasserstein	Amortized SVGD / Amortized MCMC / Adversarial Bayesian Operator VI ?		
Unormalized	Tractable	KL divergence $\chi^2 - divergence$ f-divergence	(parametric) VI adaptive importance sampling, etc		