Advanced Topics in Machine Learning

Phu Sakulwongtana

Contents

1 Convex Optimization **4**

1.1 Introduction . . . 4

1.2 Convex Sets . . . 5

1.2.1 Examples . . . 5

1.2.2 Operators that Preserve Convexity . . . 7

1.3 Convex Functions . . . 8

1.3.1 Properties of Convex Functions . . . 8

1.3.2 Building Convex Functions . . . 10

1.3.3 Other Kinds of Convex Related Functions . . . 11

1.4 Convex Optimization Problems . . . 12

1.4.1 Introductions . . . 12

1.4.2 Equivalent Convex Problems . . . 14

1.4.3 Types of Convex Problems . . . 16

1.4.4 Vector Optimization Problem . . . 21

1.5 Duality . . . 22

1.5.1 Lagragian . . . 22

1.5.2 Dual Problems . . . 24

1.5.3 Techniques of Solving Dual Problems . . . 28

2 RKHS in Machine Learning **31**

2.1 Introduction to RKHS . . . 31

2.1.1 Building a Kernel . . . 31

2.1.2 Further Notions of Kernels and RKHS . . . 33

2.2 Smoothness of RKHS . . . 35

2.2.1 Periodic Case . . . 35

2.2.2 Eigen Expansion Case . . . 37

2.3 More of RKHS . . . 38

2.4 Application of Kernel . . . 39

2.4.1 Kernel PCA . . . 39

2.4.2 Kernel Ridge Regression . . . 42

2.5 Maximum Mean Discrepancy . . . 44
2.5.1 Mean Embedding . . . 44
2.5.2 Algorithm . . . 45
2.5.3 Statistical Testing of MMD . . . 46
2.5.4 Characteristic RKHS . . . 47
2.6 Testing Dependencies . . . 49
2.6.1 Covariance Operators . . . 49
2.6.2 COCO . . . 51
2.6.3 HSIC . . . 53
2.7 Testing Goodness of Fit . . . 56
2.8 Support Vector Machine . . . 60
2.8.1 Introduction . . . 60
2.8.2 Convex Optimization . . . 61

3 Statisical Learning **68**
3.1 Formulating Learning Problem . . . 68
3.1.1 Problem . . . 68
3.1.2 Risk . . . 69
3.1.3 Empirical Risks . . . 70
3.2 Generalization Bound . . . 72
3.2.1 Generalization Error . . . 72
3.2.2 Bound For Generalization . . . 73
3.2.3 Regularization . . . 77
3.3 Tikhonov Regularization . . . 79
3.3.1 Regularized Space . . . 79
3.3.2 Introduction to Convex + Finding Weights . . . 80
3.3.3 Gradient Descent . . . 81
3.3.4 Stability . . . 86
3.4 Early Stopping . . . 90
3.5 Sub-Gradient Methods . . . 92
3.5.1 Introduction to Sub-Gradient . . . 92
3.5.2 Projected Subgradient Method . . . 93
3.5.3 Examples of Stochastic Optimization . . . 97

A Additional Proof **100**
A.1 RKHS in Machine Learning . . . 100
A.1.1 Expansion of Centered Matrix for PCA . . . 100
A.1.2 Centering Kernel Matrix . . . 100
A.1.3 Ridge Regression Expansion . . . 101

A.1.4 Representor Theorem for Ridge Regression 102
A.1.5 MMD Integration 102
A.1.6 Biased Estimate of HSIC Part 2 102
A.2 Experimental Proof 103
A.2.1 Projected Gradient Descent 103

Chapter 1

Convex Optimization

1.1 Introduction

Definition 1.1.1. (Optimization Problem) We have the following optimization problem:

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq b_i \quad i = 1, \ldots, m \end{aligned}$$

where we have

- $x = (x_1, \ldots, x_n)$: Optimization Variable
- $f_0 : \mathbb{R}^n \to \mathbb{R}$: Objective Function
- $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$: Constant Function

The optimal solution x^* has smallest value of f_0 among all vectors that satisfies the constraint.

Definition 1.1.2. (Least Square) We have the following problem:

$$\min \|Ax - b\|_2^2$$

where we have the following analytic solution $x^* = (A^T A)^{-1} A^T b$. There are reliable and efficient algorithm to solve, with the complexity of $\mathcal{O}(n^2 k)$ where $A \in \mathbb{R}^{k \times m}$. The problem is easy to recognize and a few standard technique to increase flexibility.

Definition 1.1.3. (Linear Programming) We have the following problem:

$$\begin{aligned} &\min c^T x \\ &\text{subject to } a_i^T x \leq b_i \quad i = 1, \ldots, m \end{aligned}$$

There is no analytical solution but there are reliable and efficient algorithm to solve with complexity of $\mathcal{O}(n^2 m)$ if $m \geq n$. The problem isn't east to recognize but there are standard tricks to convert problem into a linear program.

Definition 1.1.4. (Convex Optimization Problem) We have the following problem:

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq b_i \quad i = 1, \ldots, m \end{aligned}$$

The objective and constraint functions are convex. This includes a least square and linear program as special case. Trying to solve the convex optimization problem has no analytic solution but we have reliable and efficient algorithm. The time complexity is $\max\{n^3, n^2m, F\}$ where F is the cost of evaluating f_i and their first and second derivative. The problem is hard to recognize, where there are many tricks to covert problem to convex form.

Remark 1. The traditional technique to solve non-convex optimization involves compomise, where:

- Local Optimization Method
  - Find a point that minimize f_0 among feasible point near it.
  - Fast and can handle large problem
  - Require initial guess
  - No information about distance to global optimum.
- Global Optimization Method:
  - Find the global solution
  - Worst case complexity can be exponential with problem size.

These algorithms are based on solving convex subproblem.

1.2 Convex Sets

1.2.1 Examples

Definition 1.2.1. (Line) A line through x_1, x_2 points:

$$x = \theta x_1 + (1 - \theta)x_2$$

where $\theta \in \mathbb{R}$

Definition 1.2.2. (Affine Set) A set that contains a line through any 2 distict points in the set.

Definition 1.2.3. (Line Segment) Between x_1 and x_2 where:

$$x = \theta x_1 + (1 - \theta)x_2$$

with $0 \le \theta \le 1$

Definition 1.2.4. (Convex Set) A set that contains a line segment between any 2 points $x_1, x_2 \in C$ in the set:

$$\theta x_1 + (1 - \theta)x_2 \in C$$

where $0 \le \theta \le 1$

Definition 1.2.5. (Convex Combination) Given points $x_1, x_2, \dots, x_k$, then the convex combination:

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

with $\theta_1 + \theta_2 + \cdots + \theta_k = 1$ where $\theta_i \ge 0$

Definition 1.2.6. (Convex Hull) Set of all convex combination of points in S is called convex hull.

Definition 1.2.7. (Cone (Non-Negative) Combination) Cone Combination of x_1 and x_2 is any points with the form:

$$x = \theta_1 x_1 + \theta_2 x_2$$

with $\theta_1 \ge 0$ and $\theta_2 \ge 0$

Definition 1.2.8. (Convex Cone) Convex Cone is the set that contains all conic combination of points in the set.

Definition 1.2.9. (Hyperplane) Hyperplane is the set of the form $\{x|a^T x = b\}$ where $a \neq 0$

Definition 1.2.10. (Halfspace) Halfspace is the set of the form $\{x|a^T x \leq b\}$ where $a \neq 0$

Definition 1.2.11. (Euclidian Ball) The euclidian with a center x_c and radius r is:

$$B(x_c, r) = \left\{x \middle| \|x - x_C\| \leq r\right\} = \left\{x_c + ru \middle| \|u\|_2 \leq 1\right\}$$

Definition 1.2.12. (Ellipsoid) The set of the form

$$\left\{x \middle| (x - x_c)^T P^{-1} (x - x_c) \leq 1\right\}$$

with P is symmetric positive semi-definite matrices, or we can set

$$\left\{x_c + Au \middle| \|u_1\| \leq 1\right\}$$

where A being square and non-singular.

Definition 1.2.13. A function that satisfies:

- $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$
- $\|tx\| = |t| \|x\|$ for $t \in \mathbb{R}$
- $\|x + y\| \leq \|x\| + \|y\|$

Definition 1.2.14. (Norm Ball) The norm ball is the center x_C and radius r is:

$$\left\{x \middle| \|x - x_C\| \leq 1\right\}$$

Definition 1.2.15. (Norm Cone) We have

$$\left\{(x, y) \middle| \|x\| \leq t\right\}$$

The euclidian norm cone is called second order cone.

Lemma 1.2.1. *The norm balls and cones are convex.*

Definition 1.2.16. (Polyhedra) The solution set of finitely many linear inequalities and equalities:

$$Ax \preceq b \qquad Cx = d$$

The $\preceq$ is component-wise inequality, where $A \in \mathbb{R}^{m \times n}$ and $C \in \mathbb{R}^{p \times n}$. Please note that the polyhedron is intersection of finite number of halfspace and hyperplane.

Definition 1.2.17. $\mathbb{S}^n$ is set of symmetric $n \times n$ matrices.

Definition 1.2.18. (Positive Semi-Definite)

$$\mathbb{S}^n_+ = \left\{X \in \mathbb{S}^n \middle| X \succeq 0\right\}$$

where $X \in \mathbb{S}^n_+ \iff z^T X z \geq 0$ for all z. Note that $\mathbb{S}^n_+$ is convex cone. If we have strictly greater than 0, we have positive definite matrices:

$$\mathbb{S}^n_{++} = \left\{X \in \mathbb{S}^n \middle| X \succ 0\right\}$$

1.2.2 Operators that Preserve Convexity

Proposition 1.2.1. *Intersection of any number of convex sets is convex.*

Proposition 1.2.2. *Suppose* $f : \mathbb{R}^n \to \mathbb{R}^m$ *is affine* ($f(x) = Ax + b$ *with* $A \in \mathbb{R}^{m\times n}$ *and* $b \in \mathbb{R}^m$)*:*

- *The image of convex set under* f *is convex*

$$S \subseteq \mathbb{R}^n \text{ is convex } \implies f(S) = \left\{ f(x) \middle| x \in S \right\}$$

- *The inverse image of* $f^{-1}(C)$ *of a convex set under* f *is convex:*

$$C \subseteq \mathbb{R}^m \text{ is convex } \implies f^{-1}(C) = \left\{ x \in \mathbb{R}^n \middle| f(x) \in C \right\}$$

Proposition 1.2.3. *The perspective function* $P : \mathbb{R}^{n+1} \to \mathbb{R}^n$ *where*

$$P(x, t) = x/t$$

where dom $f = \{(x, t)|t > 0\}$*. The image and inverse image of convex set under perspective are convex.*

Proposition 1.2.4. *A linear fractional function* $f : \mathbb{R}^n \to \mathbb{R}^m$

$$f(x) = \frac{Ax + b}{c^T x + d}$$

where dom $f = \{x|c^T x + d > 0\}$

Definition 1.2.19. (Proper Cone) $\mathcal{K} \subseteq \mathbb{R}^n$ is proper cone if

- $\mathcal{K}$ is closed (Contains Its Boundary)
- $\mathcal{K}$ is solid (Non Empty)
- $\mathcal{K}$ is pointed (Contains No Line)

Definition 1.2.20. (Generalized Ineqality) It is defined by proper cone $\mathcal{K}$, where

$$X \preceq_{\mathcal{K}} Y \iff y - x \in \mathcal{K} \qquad X \prec_{\mathcal{K}} Y \iff y - x \in \operatorname{int} \mathcal{K}$$

The property of generalized inequality is similar to $\leq$ in $\mathbb{R}$. Please note that it isn't a general linear ordering. We can have $X \npreceq_{\mathcal{K}} Y$ and $Y \npreceq_{\mathcal{K}} X$

Definition 1.2.21. (Minimum) The point $x \in S$ is minimum element of S with respected to $\succeq_{\mathcal{K}}$ if

$$y \in S \implies x \preceq_K y$$

Definition 1.2.22. (Minimal) The point $x \in S$ is the minimal element of S with respected to

$$y \in S, y \preceq_{\mathcal{K}} X \implies y = x$$

Theorem 1.2.1. *If* C *and* D *are non-empty disjoint convex set, there exists* $a \neq 0$ *and* b *such that* $a^T x \leq b$ *for* $x \in C$ *and* $a^T x > b$ *for* $x \in D$*. This means that the hyperplane* $\left\{x|a^T x = b\right\}$ *separates* C *and* D*.*

Definition 1.2.23. (Supporting Hyperplane) to a set C at boundary point x_0 such that

$$\left\{ x \middle| a^T x = a^T x_0 \right\}$$

where $a \neq 0$ and $a^T x \leq a^T x_0$ for all $x \in C$

Theorem 1.2.2. *If C is convex, then there exists, a supporting hyperplane at every boundary point of C*

Definition 1.2.24. (Dual Cone) The dual cone of a cone $\mathcal{K}$ is:

$$\mathcal{K}^* = \left\{ y | y^T x \geq 0 \text{ for all } x \in \mathcal{K} \right\}$$

If the cone is a dual of itself is called self-dual. Furtheremore, if dual cone of proper cone is propert, hence defined generalized inequality:

$$y \succeq_{\mathcal{K}^*} 0 \iff y^T x \geq 0 \text{ for all } x \succeq_{\mathcal{K}} 0$$

Proposition 1.2.5. *The minimum element with respected to $\preceq_{\mathcal{K}}$: x is minimum of S iff for all $\lambda \succeq_{\mathcal{K}^*} 0$ is unique minimizer of $\lambda^T z$ over S.*

Proposition 1.2.6. *The minimal element with respected to $\preceq_{\mathcal{K}}$:*

- *If x minimizes $\lambda^T z$ over S for some $\lambda \succeq_{\mathcal{K}^*} 0$ then x is minimal*
- *If x is a minimal element of convex set S then there exists a non-zero $\lambda \succeq_{\mathcal{K}^*} 0$ such that x minimizer $\lambda^T z$ over S*

1.3 Convex Functions

1.3.1 Properties of Convex Functions

Definition 1.3.1. (Convex Function) $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\text{dom}(f)$ is convex:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \text{dom}(f)$ and $0 \leq \theta \leq 1$

Definition 1.3.2. (Concave + Strictly Convex) f is convex if $-f$ is convex. f is strictly convex if $\text{dom}\, f$ is convex and:

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$$

for $x, y \in \text{dom}(f)$ where $x \neq y$ and $0 \leq \theta \leq 1$.

Remark 2. Examples of convex functions in $\mathbb{R}$:

- Affine: $ax + b$ on $\mathbb{R}$ and for any $a, b \in \mathbb{R}$
- Exponential: $\exp(ax)$ for any $a \in \mathbb{R}$
- Power: x^α on $\mathbb{R}_{++}$ for $\alpha \geq 1$ or $\alpha \leq 0$
- Power of Absolute Value: $|x|^p$ on $\mathbb{R}$ with $p \geq 1$
- Negative entropy: $x \log x$ on $\mathbb{R}_{++}$

Examples of concave functions in $\mathbb{R}$:

- Affine: $ax + b$ on $\mathbb{R}$ and for any $a, b \in \mathbb{R}$
- Power: x^α on $\mathbb{R}_{++}$ for $0 \leq \alpha \leq 1$
- Logarithm: $\log x$ on $\mathbb{R}_{++}$

Remark 3. Examples of convex function in $\mathbb{R}^n$:

- Affine Function: $f(x) = a^T x + b$
- Norms: $\|x\|_p$ where

$$\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

for $p \geq 1$ and $\|x\|_\infty = \max_k |x_k|$

Examples of convex function in $\mathbb{R}^{m\times n}$:

- Affine Function: $f(X) = \operatorname{tr}(A^T X) + b = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}X_{ij} + b$
- Special Singular Value:

$$f(X) = \|X\|_2 = \sigma_{\max}(X) = \left(\lambda_{\max}(X^T X)\right)^{1/2}$$

Proposition 1.3.1. *The function $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff the function $g : \mathbb{R} \to \mathbb{R}$ where $g(t) = f(x + tv)$, where* $\operatorname{dom}(g) = \{t | x + tv \in \operatorname{dom} f\}$. *Now we can check the convexity of f by checking convexiy of functions of one variable.*

Remark 4. Let's consider the log-determinant function:

$$\begin{aligned} g(t) &= \log\det(X + tV) = \log\det X + \log\det(I + X^{-1/2}VX^{-1/2}) \\ &= \log\det X + \sum_{i=1}^{n} \log(1 + t\lambda_i) \end{aligned}$$

where λ_i are eigenvalues of $X^{-1/2}VX^{-1/2}$ and therfore g is concave in t for any choice $X \succ 0$ and V hence f is concave.

Definition 1.3.3. (Extended Value Extension) The extended value extension $\tilde{f}$ of f is:

- $\tilde{f}(x) = f(x)$ if $x \in \operatorname{dom}(f)$
- $\tilde{f}(x) = \infty$ if $x \notin \operatorname{dom}(f)$

This would simplify the notation. The condition:

$$0 \leq \theta \leq 1 \implies \tilde{f}(\theta x + (1-\theta)y) \leq \theta \tilde{f}(x) + (1-\theta)\tilde{f}(y)$$

as the inequality in $\mathbb{R} \cup \{\infty\}$ means the same. The domain f is convex.

Proposition 1.3.2. ***(Differentiable)*** *f is differentiable if* $\operatorname{dom}(f)$ *is open and the gradient:*

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_n}\right)$$

exists at each $x \in \operatorname{dom}(f)$

Lemma 1.3.1. *First order condition, a differentiable f with convex domain S is convex iff:*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

For all $x, y \in \operatorname{dom}(f)$. *This means a first order approximation of f is global underestimator.*

Definition 1.3.4. (Twice Differentiable) If f is twice differentiable, if $\operatorname{dom}(f)$ is open then Hessian:

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x \partial y}$$

for $i, j = 1, \ldots, n$ exists at each $x \in \operatorname{dom}(f)$.

Lemma 1.3.2. *For twice differentiable f with convex domain, f is convex iff*

$$\nabla^2 f(x) \succeq 0$$

for all $x \in \text{dom}(f)$. If $\nabla^2 f(x) \succ 0$ for all $x \in \text{dom}(f)$, then f is strictly convex. Note that we can use it to calculate the convexity of the function.

Definition 1.3.5. (α-sublevel Set) α-sublevel set of $f : \mathbb{R}^n \to \mathbb{R}$, which we have:

$$C_\alpha = \left\{ x \in \text{dom}(f) \middle| f(x) \le \alpha \right\}$$

A sublevel set of convex functions are convex but not the converse.

Definition 1.3.6. (Epigraph) The epigraph of $f : \mathbb{R}^n \to \mathbb{R}$:

$$\text{epi}(f) = \left\{ (x, t) \in \mathbb{R}^{n+1} \middle| x \in \text{dom}(f), f(x) \le t \right\}$$

is f is convex iff epi(f) is a convex set.

Definition 1.3.7. (Jensen's Ineqality) If f is convex then for $0 \le \theta \le 1$, we have:

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$$

The extension if f is convex then $f(\mathbb{E}[z]) \le \mathbb{E}[f(z)]$

1.3.2 Building Convex Functions

Proposition 1.3.3. *We have the following opeartors on function that we can use for creating a new convex functions:*

- *Non-negative multiple αf is convex if f is convex and $\alpha > 0$*
- *Sum $f_1 + f_2$ is convex if f_1 and f_2 is convex. This can be extended to infinite sum or integral.*
- *Composition with affine function $f(Ax + b)$ is convex if f is convex.*

Proposition 1.3.4. *If $f_1, \dots, f_m$ are convex then:*

$$f(x) = \max\left\{ f_1(x), \dots, f_m(x) \right\}$$

Proposition 1.3.5. *If $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then:*

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is convex.

Proposition 1.3.6. *Composition of $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ where $f(x) = h(g(x))$. Then f is convex if:*

- *g is convex, h is convex and $\tilde{h}$ is non-decreasing.*
- *g is concave, h is convex and $\tilde{h}$ is non-increasing.*

Proof. Let's consider when the case where $n = 1$ and differentiable g and h:

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

Monotonicity must hold extned value extensions $\tilde{h}$ □

Proposition 1.3.7. *Composition of $g : \mathbb{R}^n \to \mathbb{R}^k$ and $n : \mathbb{R}^k \to \mathbb{R}$:*

$$f(x) = h(g(x)) = h(g_1(x), \ldots, g_k(x))$$

where we have, f is convex: if

- *g_i convex, h convex, $\tilde{h}$ is non-decreasing.*
- *g_i concave, h convex, $\tilde{h}$ is non-increasing.*

Proof. For $n = 1$ and differentiable g, h:

$$f''(x) = g'(x)^T \nabla^2 h(g(x)) g'(x) + \nabla h(g(x))^T g''(x)$$

□

Proposition 1.3.8. *If $f(x, y)$ is convex in (x, y) and C is convex set then:*

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex.

Proposition 1.3.9. *The perspective of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the functtion $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$:*

$$g(x, y) = f(x/t) \cdot t$$

where $\text{dom} = \{(x, y) | x/t \in \text{dom}(f), t > 0\}$. *The g is convex if f is convex.*

1.3.3 Other Kinds of Convex Related Functions

Definition 1.3.8. (Conjugate) Conjugate of a function f is $f^*(y) = \sup_{x \in \text{dom}(f)}(y^T x - f(x))$, then f^* is convex even f isn't.

Definition 1.3.9. (Quasi-Convex) The function $f : \mathbb{R}^n \Rightarrow \mathbb{R}$ is quasi-convex if the domain of f is convex and:

$$S_\alpha = \left\{ x \in \text{dom}(f) \middle| f(x) \leq \alpha \right\}$$

are convex for all α. f is quasi-concave if $-f$ is quasi-convex. and f is quasi-linear if f is quasi-convex and quasi-concave.

Proposition 1.3.10. *Modified Jensen's inequalities: For quasi-convex f, and for $0 \leq \theta \leq 1$:*

$$f(\theta x + (1 - \theta)y) \leq \max\{f(x), f(y)\}$$

Proposition 1.3.11. *For differentiable f with convex domain is quasi-convex iff*

$$f(y) \leq f(x) \implies \nabla f(x)^T (y - x) \leq 0$$

Remark 5. Sum of Quasi-convex functions are not necessary quasi-convex.

Definition 1.3.10. (Log-Concave and Log-Convex Function) A positive function f is log concave if $\log f$ is concave:

$$f(\theta x + (1 - \theta)y) \geq f(x)^\theta f(y)^{1-\theta}$$

for $0 \leq \theta \leq 1$, and f is log convex if $\log f$ is convex.

Proposition 1.3.12. *We have the following results for log-concave:*

- *Twice differentiable f with convex function is* $\log$ *concave iff*

$$f(x)\nabla^2 f(x) \preceq \nabla f(x)\nabla f(x)^T$$

for all $x \in \mathrm{dom}(f)$

- *Product of Log-Concave function is log-concave.*
- *Sum of log-concave function isn't always log-concave.*
- *If $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is log concave then:*

$$g(x) = \int f(x, y)\ \mathrm{d}y$$

is log concave, if the integration exists.

Proposition 1.3.13. *Convolution $f * g$ of log-concave function if f, g is log-concave*

$$(f * g)(x) = \int f(x - y)g(y)\ \mathrm{d}y$$

Proposition 1.3.14. *If $C \subseteq \mathbb{R}^n$ is convex and y is random variable with log-concave probability density function, then:*

$$f(x) = \mathrm{Prob}(x + y \in C)$$

is log-concave.

Proof. We write $f(x)$ as integral of product of log-concave function, where:

$$f(x) = \int g(x + y)p(y)\ \mathrm{d}y \qquad g(u) = \begin{cases} 1 & u \in C \\ 0 & u \in C \end{cases}$$

□

Definition 1.3.11. (K-Convex) The function $f : \mathbb{R}^n \to \mathbb{R}^m$ is K-convex if $\mathrm{dom}(f)$ is convex and:

$$f(\theta x + (1 - \theta)y) \preceq_{\mathcal{K}} \theta f(x) + (1 - \theta)f(y)$$

for $x, y \in \mathrm{dom}(f)$ and $0 \leq \theta \leq 1$.

1.4 Convex Optimization Problems

1.4.1 Introductions

Definition 1.4.1. (Constraint Optimization Problem) The constraint optimization is a problem of the form:

$$\begin{aligned} \min f_0(x)& \\ \text{subject to } f_i(x) \leq 0 &\quad i = 1, \dots, m \\ h_i(x) = 0 &\quad i = 1, \dots, p \end{aligned}$$

where $x \in \mathbb{R}^n$ is optimization variable. $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective. $f_i : \mathbb{R}^n \to \mathbb{R}$ where $i = 1, \dots, m$ be the inequality constraint function. Finally, $h_i : \mathbb{R}^n \to \mathbb{R}$ are equality constraint function. The optimal value is:

$$p^* = \inf \left\{ f_0(x) \middle| f_i(x) \leq 0, i = 1, \dots, m, h_i(x) = 0, i = 1, \dots, p \right\}$$

Definition 1.4.2. (Feasibility) We have the following definitions:

- x is feasible if $x \in \text{dom}(f_0)$ and it satisfies the constraints.
- A feasible x is optimal if $f_0(x) = p^*$.
- X_{opt} is the set of optimal points.

Definition 1.4.3. (Local Optimal) x is locally local if there is $R > 0$ such that x is optimal for:

$$\begin{aligned} &\min f_0(z) \\ &\text{subject to } f_i(z) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad h_i(x) = 0 \quad i = 1, \ldots, p \\ &\qquad\qquad \|z - x\|_2 \leq R \end{aligned}$$

Definition 1.4.4. (Implicit Constraints) The standard form of optimization problem has an implicit constrain:

$$x \in \mathcal{D} = \bigcap_{i=0}^{m} \text{dom}(f_i) \cap \bigcap_{i=1}^{p} \text{dom}(h_i)$$

The constraints $f_i(x) \leq 0$ and $h_i(x) = 0$ are called explicit constriants. The problem is unconstrained if there is no explicit constraints.

Definition 1.4.5. (Feasibility Problem) We can consider a special case of general problem with $f_0(x) = 0$:

$$\begin{aligned} &\min 0 \\ &\text{subject to } f_i(x) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad h_i(x) = 0 \quad i = 1, \ldots, p \end{aligned}$$

where if $p^* = 0$ then the constrains are feasible, and any feasible x is optimal. However, if $p^* = \infty$ if constraints are infeasible.

Definition 1.4.6. (Standard Form of Convex Optimization Problem) We have

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad a_i^T x_i = b_i \quad i = 1, \ldots, p \end{aligned}$$

where $f_0, f_1, \ldots, f_n$ are convex, equality constraints are affine.

Definition 1.4.7. (Quasi-Convex Problem) A Quasi-Convex Problem is when f_0 is quasi-convex (and $f_1, \ldots, f_n$ are convex.), and it is written as

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad Ax = b \quad i = 1, \ldots, p \end{aligned}$$

Proposition 1.4.1. *Any locally optimal point of a convex problem is (globally) optimal.*

Proof. Suppose x is locally optimal but there exists a feasible point y with $f_0(y) < f_0(x)$. We see that x is locally optimal means that there is an $R > 0$ such that z is feasible and

$$\|z - x\|_2 \leq R \implies f_0(z) \geq f_0(x)$$

We then consider $z = \theta y + (1 - \theta)x$ with $\theta = R/(2\|y - x\|_2)$. Since

- $\|y - x\|_2 > R$ so we need $0 \leq \theta \leq 1/2$.
- z is convex combination of feasible points x and y, then z is feasible.

- $\|z-x\|_2 = R/2$ and
$$f_0(z) \leq \theta f_0(y) + (1-\theta) f_0(x) < f_0(x)$$

□

Proposition 1.4.2. *x is optimal iff it is feasible and*

$$\nabla f_0(x)^T (y-x) \geq 0$$

for all feasible y. If we have non-zero $\nabla f_0(x)$ we define a supporting hyperplane to feasible set X at x.

Definition 1.4.8. (Unconstrained Problem) x is optimal iff $x \in \text{dom}(f_0)$ and $\nabla f_0(x) = 0$

Definition 1.4.9. (Equally Constraint Problem) We have the following form:

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } Ax = b \end{aligned}$$

x is optimal iff there exists ν such that $x \in \text{dom}(f)$. $Ax = b$ and $\nabla f_0(x) + A^T \nu = 0$

Definition 1.4.10. (Minimization Over Non-Negative Orthant) We have the following form

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } x \succeq 0 \end{aligned}$$

x is optimal iff $x \in \text{dom}(f_0)$ and $x \succeq 0$

$$\begin{cases} \nabla f_0(x)_i \geq 0 & x_i = 0 \\ \nabla f_0(x)_i = 0 & x_i > 0 \end{cases}$$

1.4.2 Equivalent Convex Problems

Proposition 1.4.3. ***(Eliminating Equality Constraints)*** *These 2 problems are equivalent as one of the the solution can be obtained from the solution of the other:*

$$\begin{aligned} &\min f_0(x) \\ &\text{such that } f_i(x) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad Ax = b \end{aligned}$$

This is equivalent to:

$$\begin{aligned} &\min f_0(Fz + x_0) \\ &\text{such that } f_i(Fz + x_0) \leq 0 \quad i = 1, \ldots, m \end{aligned}$$

where F and x_0 are such that:

$$Ax = b \iff x = Fz + x_0$$

for some z

Proposition 1.4.4. ***(Introducing Equality Constraints)***

$$\begin{aligned} &\min f_0(A_0 x + b) \\ &\text{such that } f_i(A_i x + b_i) \leq 0 \quad i = 1, \ldots, m \end{aligned}$$

is equivalent to

$$\begin{aligned} &\min f_0(y_0) \\ &\text{such that } f_i(y_i) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad y_i = A_i x + b_i \quad i = 0, 1, \ldots, m \end{aligned}$$

Proposition 1.4.5. ***(Introducing Slack Varaible for Linear Inequalities)***

$$\begin{aligned}&\min f_0(x)\\&\text{such that } a_i^T x \le b_i \quad i = 1, \ldots, m\end{aligned}$$

is equivalent to

$$\begin{aligned}&\min f_0(x)\\&\text{such that } a_i^T x + s_i = b_i \quad i = 1, \ldots, m\\&\qquad s_i \ge 0 \quad i = 1, \ldots, m\end{aligned}$$

we minimize over x and s

Proposition 1.4.6. ***(Epigraph Form)*** *Standard Convex Problem is equivalent to*

$$\begin{aligned}&\min t\\&\text{such that } f_0(x) - t \le 0\\&\qquad f_i(x) \le 0 \quad i = 1, \ldots, m\\&\qquad Ax = b\end{aligned}$$

where we minimize over x and t.

Proposition 1.4.7. ***(Minimizer Over Some Variables)***

$$\begin{aligned}&\min f_0(x_1, x_2)\\&\text{such that } f_i(x_1) \le 0 \quad i = 1, \ldots, m\end{aligned}$$

is equivalent to

$$\begin{aligned}&\min \tilde{f}_0(x_1)\\&\text{such that } f_i(x_1) \le 0 \quad i = 1, \ldots, m\end{aligned}$$

where $\tilde{f}_0(x_1) = \inf_{x_2} f_0(x_1, x_2)$

Proposition 1.4.8. *If f_0 is quasi-convex then there exists a famility of functions ϕ_t such that:*

- *$\phi_t(x)$ is convex in x for fixed t*
- *t-sublevel set of f_0 is 0-sublevel set of ϕ_t:*

$$f_0(x) \le t \iff \phi_t(x) \le 0$$

Remark 6. The example of this is:

$$f_0 = \frac{p(x)}{q(x)}$$

where if p is convex and q is concave, and $p(x) \ge 0$ and $q(x) > 0$ on $\mathrm{dom}(f_0)$, we can take $\phi_t(x) = p(x) - tq(x)$

- For $t \ge 0$, ϕ_t is convex in x
- $p(x)/q(x) \le t$ iff $\phi_t(x) \le 0$

Definition 1.4.11. (Bisection Method For Quasi-Convex Optimization) We can consider the feasibility problem, where we have:

$$\phi_t(x) \le 0 \qquad f_i(x) \le 0 \qquad Ax = b$$

Then we can see that, for a fixed t, a convex feasibility problem implies:

- If feasible then $t \ge p^*$

- Otherwise $t \leq p^*$

which leads to binary search-like problem, where:

Algorithm 1 Bisection Method For Quasi-Convex Optimization

Input: $l \leq p^*$ and $u \geq p^*$ and Tolerance $\varepsilon > 0$
while Until Convergence **do**
  $t = (t + u)/2$
  Solve the convex feasibility problem
  if It is Feasible **then**
    u = t
  else
    l = t
  end if
end while

This requires exactly $\lceil \log_2((u - l)/\varepsilon) \rceil$ iterations, when u and l are intial values.

1.4.3 Types of Convex Problems

Definition 1.4.12. (Linear Program)

$$\begin{aligned} &\min c^T x + d \\ &\text{subject to } Gx \preceq h \\ &\qquad\qquad\quad Ax = b \end{aligned}$$

It is an convex problem with affine objective and constraint functions. Feasible set is polyhedron.

Remark 7. The notable problem of LP is Chebshev center of polyhedron, where the Chebshev center of $\mathcal{P} = \left\{x | a_i^T x \leq b_i, i = 1, \ldots, m\right\}$ is the center of largest inscribed ball $B = \{x_c + u | \ \|u\|_2 \leq r\}$. Note that $a_i^T x \leq b$ for all $x \in B$ iff

$$\sup \left\{ a_i^T (x_c + u) \middle| \ \|u\|_2 \leq r \right\} = a_i^T x_c + r \|a_i\|_2 \leq b_i$$

Hence, the x_c and r can be determined by:

$$\begin{aligned} &\max r \\ &\text{subject to } a_i^T x_c + r \|a_i\|_2 \leq b_i \text{ for } i = 1, \ldots, m \end{aligned}$$

Definition 1.4.13. (Linear Fractional Program)

$$\begin{aligned} &\min \frac{c^T x + d}{e^T x + f} \\ &\text{subject to } Gx \preceq h \\ &\qquad\qquad\quad Ax = b \end{aligned}$$

where $\text{dom}(f_0) = \left\{x | e^T x + f > 0\right\}$. This is a quasi-convex optimization, which can be solved by Bisection method. Note that it is equivalent to LP:

$$\begin{aligned} &\min c^T y + dz \\ &\text{subject to } Gy \preceq hz \\ &\qquad\qquad\quad Ay = bz \\ &\qquad\qquad\quad e^T y + fz = 1 \\ &\qquad\qquad\quad z \geq 0 \end{aligned}$$

Definition 1.4.14. (Generalized Fractional Program) where we have

$$f_0(x) = \max_{i=1,\dots,r} \frac{c_i^T x + d_i}{e_i^T x + f_i}$$

where $\operatorname{dom}(f_0) = \{x | e_i^T x + f_i > 0; i = 1, \dots, r\}$. This is also quasi-convex problem, which can be solved by Bisection

Definition 1.4.15. (Quadratic Program)

$$\begin{aligned} &\min (1/2) x^T P x + q^T x + r \\ &\text{subject to } Gx \preceq h \\ &\qquad Ax = b \end{aligned}$$

where $P \in \mathbb{S}_+^n$, therefore,the objective is convex. The examples of quadratic program is least square problem.

Definition 1.4.16. (Linear Program with Random Cost)

$$\begin{aligned} &\min \bar{c}^T x + \gamma x^T \Sigma x = \mathbb{E}[c^T x] + \gamma \operatorname{var}(c^T x) \\ &\text{subject to } Gx \preceq h \\ &\qquad Ax = b \end{aligned}$$

We have c as as random variable with a mean of $\bar{c}$ and covariance of Σ, given this we have $c^T x$ being a random variable with a mean of $c^T x$ and covarance $x^T \Sigma x$. Fianlly, $\gamma > 0$ is risk aversion paramter, which controls the trade-off between expected cost and risk.

Definition 1.4.17. (Quadratic Constrained Quadratic Program)

$$\begin{aligned} &\min \frac{1}{2} x^T P_0 x + q_0^T x + r_0 \\ &\text{subject to } \frac{1}{2} x^T P_i x + q_i^T x + r_i \leq 0 \\ &\qquad Ax = b \end{aligned}$$

where $P_i \in \mathbb{S}_+^n$ where objective and constraints are convex quadratic. If $P_1, \dots, P_m \in \mathcal{S}_{++}^n$ feasible region is intersection of m ellipsoid and an affine set.

Definition 1.4.18. (Second Order Cone Programming)

$$\begin{aligned} &\min f^T x \\ &\text{subject to } \|A_i x + b_i\|_2 \leq c_i^T x + d_i \quad i = 1, \dots, m \\ &\qquad Fx = g \end{aligned}$$

where $A_i \in \mathbb{R}^{n_i \times n}$ and $F \in \mathbb{R}^{p \times n}$. The inequalities are called second order cone constraints: $(A_i x + b_i, c_i^T x + d_i)$ is in second order cone in $\mathbb{R}^{n_i+1}$. For $n_i = 0$, reduces to an LP if $c_i = 0$ reduces to QCQP.

Remark 8. The parameter in the optimization problem are often constraint, for example, in LP:

$$\begin{aligned} &\min c^T x \\ &\text{subject to } a_i^T x \leq b_i \quad i = 1, \dots, m \end{aligned}$$

as there exists an uncertainty in c, a_i, b_i.

Definition 1.4.19. (Deterministic Robust Linear Programming) We can constrain the paramter that must holds for all $a_i \in \mathcal{E}_i$ where:

$$\begin{aligned} &\min c^T x \\ &\text{subject to } a_i^T x \leq b_i \quad \forall a_i \in \mathcal{E}_i \quad i = 1, \dots, m \end{aligned}$$

Definition 1.4.20. (Stochastic Robust Linear Programming) We have a_i as random variables. The constrains must hold with probability η:

$$\begin{aligned}&\min c^T x\\ &\text{subject to } \text{Prob}(a_i^T x \leq b_i) \geq \eta \quad i = 1, \dots, m\end{aligned}$$

Proposition 1.4.9. *We choose an Ellipsoid $\mathcal{E}_i$:*

$$\mathcal{E}_i = \left\{\bar{a}_i + P_i u \middle| \, \|u\|_2 \leq 1\right\}$$

The center is $\bar{a}_i$ with the semi-axis is determined by singular value of P_i. Then the deterministic robust LP (with constraint $\mathcal{E}_i$) is equivalent to:

$$\begin{aligned}&\min c^T x\\ &\text{such that } \bar{a}_i^T x + \left\|P_i^T x\right\|_2 \leq b_i \quad i = 2, \dots, m\end{aligned}$$

This follows from

$$\sup_{\|u\|_2 \leq 1} (\bar{a}_i + P_i u)^T x = \bar{a}_i^T x + \left\|P_i^T x\right\|_2$$

Proposition 1.4.10. *Assume a_i is Guassian with mean $\bar{a}_i$ and covarance $\bar{\Sigma}_i$. We can see that $a_i^T x$ is Gaussian with mean of $\bar{a}_i^T x$ variance $x^T \Sigma_i x$ hence, we have:*

$$\text{Prob}\left(a_i^T x \leq b_i\right) = \Phi\left(\frac{b_i - a_i^{-T} x_i}{\left\|\Sigma_i^{1/2} x\right\|_2}\right)$$

where Φ is CDF with $\|N\|(0,1)$. Given the stochastic robust LP with $\eta \geq 1/2$ is equivalent to SOCP:

$$\begin{aligned}&\min c^T x\\ &\text{such that } \bar{a}_i^T x + \Phi^{-1}(\eta)\left\|\Sigma_i^{1/2} x\right\|_2 \leq b_i \quad i = 1, \dots, m\end{aligned}$$

Definition 1.4.21. (Monomial Function) Monomial function is fuction of the form:

$$f(x) = c x_1^{a_1} c x_2^{a_2} \cdots x_n^{a_n}$$

where $\operatorname{dom} f \in \mathbb{R}_{++}^n$ with $c > 0$, the exponent a_i can be any real number. Note that this can be transformed to:

$$\log f(e^{y_1}, \dots, e^{y_n}) = a^T y + b$$

where $b = \log c$

Definition 1.4.22. (Posynomial Function) Posynomial function is sum of monomials:

$$f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}$$

where $\operatorname{dom} f \in \mathbb{R}_{++}^n$. This can be transformed to:

$$\log f(e^{y_1}, \dots, e^{y_n}) = \log\left(\sum_{k=1}^{K} \exp(a_k^T y + b_k)\right)$$

where $b_k = \log c_k$.

Definition 1.4.23. (Geometric Program)

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq 1 \quad i = 1, \ldots, m \\ &\qquad\qquad h_i(x) = 1 \quad i = 1, \ldots, p \end{aligned}$$

with f_i is posynomial and h_i is monomial. This can be transformed t oconvex problem:

$$\begin{aligned} &\min \log \left(\sum_{k=1}^{K} \exp\left(a_{0k}^T y + b_{0k}\right) \right) \\ &\text{such that } \log \left(\sum_{k=1}^{K} \exp\left(a_{ik}^T y + b_{ik}\right) \right) \leq 0 \\ &\qquad Gy + d = 0 \end{aligned}$$

Definition 1.4.24. (Perron-Frobenius Eigenvalue) This exists in (element-wise) positive $A \in \mathbb{R}^{n \times n}$. It is defined as real, positive eigenvalue of A to spectral radius $\max_i |\lambda_i(A)|$. Note that this determines asymptotic growth/decay rate of A^k as $A^k \sim \lambda_{\text{pf}}^k$ as $k \to \infty$. The alternate characterization:

$$\lambda_{\text{pf}}(A) = \inf \{\lambda | Av \preceq \lambda v \text{ for some } v \succ 0\}$$

Remark 9. We want to minimize $\lambda_{\text{pf}}(A(x))$ where $A(x)_{ij}$ are posynomial of x. This is equivalent to geometric program:

$$\begin{aligned} &\min \lambda \\ &\text{subject to } \sum_{j=1}^{n} A(x)_{ij} v_j / (\lambda v_i) \leq 1 \quad i = 1, \ldots, n \end{aligned}$$

where the variables are λ, v, x.

Definition 1.4.25. (Generalize Inequality Constraints)

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \preceq_{K_i} 0 \quad i = 1, \ldots, m \\ &\qquad Ax = b \end{aligned}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$ convex and $f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$ is K_i-convex with respected to proper cone K_i. This has the same properties as standard convex optimization problem (convex feasible set, local optimum is global etc.)

Definition 1.4.26. (Conic Form Problem) Special case with affine objective and constraints:

$$\begin{aligned} &\min c^T x \\ &\text{subject to } Fx + g \preceq_K 0 \\ &\qquad Ax = b \end{aligned}$$

This extends linear programming (when $K = \mathbb{R}_+^m$) to non-polyhedron cones.

Definition 1.4.27. (Semi-Definite Program)

$$\begin{aligned} &\min c^T x \\ &\text{subject to } x_1 F_1 + x_2 F_2 + \cdots + x_n F_n + G \prec 0 \\ &\qquad Ax = b \end{aligned}$$

with $F_i, G \in \mathbb{S}^k$. This ineqality constraints is called linear matrix ineqality (LMI). By having problems with multiple LMI contraints, for example:

$$\begin{aligned} x_1 \hat{F}_1 + \cdots + x_n \hat{F}_n + \hat{G} \preceq 0 \\ x_1 \tilde{F}_1 + \cdots + x_n \hat{F}_n + \tilde{G} \preceq 0 \end{aligned}$$

is equivalent to single one:

$$x_1 \begin{bmatrix} \hat{F}_1 & 0 \\ 0 & \tilde{F}_1 \end{bmatrix} + x_2 \begin{bmatrix} \hat{F}_2 & 0 \\ 0 & \tilde{F}_2 \end{bmatrix} + \cdots + x_n \begin{bmatrix} \hat{F}_n & 0 \\ 0 & \tilde{F}_n \end{bmatrix} + \begin{bmatrix} \hat{G} & 0 \\ 0 & \tilde{G} \end{bmatrix} \preceq 0$$

Proposition 1.4.11. *Given the LP program:*

$$\begin{aligned} &\min c^T x \\ &\text{such that } Ax \preceq b \end{aligned}$$

is equivalent to SDP program:

$$\begin{aligned} &\min c^T x \\ &\text{such that } \operatorname{diag}(Ax - b) \preceq 0 \end{aligned}$$

Proposition 1.4.12. *Given SOCP*

$$\begin{aligned} &\min f^T x \\ &\text{such that } \|A_i x + b_i\|_2 \le c_i^T x + d_i \quad i = 1, \dots, m \end{aligned}$$

is equivalent to the following SDP:

$$\begin{aligned} &\min f^T x \\ &\text{such that } \begin{bmatrix} (c_i^T x + d_i)I & A_i x + b_i \\ A_i x + b_i & c_i^T x + d_i \end{bmatrix} \succeq 0 \quad i = 1, \dots, m \end{aligned}$$

Proposition 1.4.13. *Given the eigenvalue minimization problem:*

$$\min \lambda_{\max}(A(x))$$

where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$ with given $A_i \in \mathbb{S}^k$. This is equivalent SDP, where:

$$\begin{aligned} &\min t \\ &\text{such that } A(x) \preceq tI \end{aligned}$$

with the variable $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. This follows from $\lambda_{\max}(A) \le t$ iff $A \preceq tI$

Proposition 1.4.14. *Given the matrix norm minimization problem:*

$$\min \|A(x)\|_2 = \left(\lambda_{\max}(A(x)^T A(x))\right)^{1/2}$$

where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$ is equivalent to:

$$\begin{aligned} &\min t \\ &\text{such that } \begin{bmatrix} tI & A(x) \\ A(x) & tI \end{bmatrix} \succeq 0 \end{aligned}$$

Given the variabble $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. We have the constraint follows from:

$$\begin{aligned} \|A\|_2 \le t &\iff A^T A \preceq t^2 I \quad t \ge 0 \\ &\iff \begin{bmatrix} tI & A \\ A^T & tI \end{bmatrix} \succeq 0 \end{aligned}$$

1.4.4 Vector Optimization Problem

Definition 1.4.28. (General Vector Optimization Problem)

$$\begin{aligned} \min\, & f_0(x) \\ \text{such that } & f_i(x) \leq 0 \quad i = 1, \dots, m \\ & h_i(x) = 0 \quad i = 1, \dots, p \end{aligned}$$

The minimization with respected to K. We have vector objective $f_0 : \mathbb{R}^n \to \mathbb{R}^q$ minimized with respected to propert cone $K \in \mathbb{R}^q$.

Definition 1.4.29. (Convex Vector Optimization Problem)

$$\begin{aligned} \min\, & f_0(x) \\ \text{such that } & f_i(x) \leq 0 \quad i = 1, \dots, m \\ & Ax = b \end{aligned}$$

with f_0 is K-convex and $f_1, \dots, f_m$ are convex.

Definition 1.4.30. (Optimality) Set of achievable objective vectors $\mathcal{O} = \{f_0(x) | x \text{ feasible}\}$:

- The feasible x is optimal if $f_0(x)$ is minimum value of $\mathcal{O}$
- The feasible x is pareto optimal if $f_0(x)$ is minimal value of $\mathcal{O}$

Remark 10. The vector optimization problem with $K = \mathbb{R}^d_+$, where

$$f_0(x) = (F_1(x), \dots, F_q(x))$$

we have q different objectives F_i, roughly, we want all f_i to be small. Then the notion of optimality becomes:

- The feasible x^* is optimal if, y is feasible:

  $$f_0(x^*) \preceq f_0(y)$$

  If there exists an optimal point, then the object are non-competing.

- The feasible x^{po} is pareto optimal, if y is feasible:

  $$f_0(y) \preceq f_0(x^{\text{po}}) \implies f_0(x^{\text{po}}) = f_0(y)$$

  If there are multiple pareto optimal value, there is a trade-off between objective.

Definition 1.4.31. (Scalarization) To find a pareto optimal point, we can choose $\lambda \succeq_{K^*} 0$ and have the following scalar problem:

$$\begin{aligned} \min\, & \lambda^T f_0(x) \\ \text{such that } & f_i(x) \leq 0 \quad i = 1, \dots, m \\ & h_i(x) = 0 \quad i = 1, \dots, p \end{aligned}$$

If x is optimal for scalar problem, then it is pareto optimal for vector optimization problems, we have:

$$\lambda^T f_0(x) = \lambda_1 F_1(x) + \cdots + \lambda_q F_q(x)$$

For convex vector optimization problem, we can find (almost) all Pareto optimal point by varying $\lambda \succ_{K^*} 0$.

1.5 Duality

1.5.1 Lagragian

Definition 1.5.1. (Lagragian) Given a standard form of problem:

$$\begin{aligned}
&\min f_0(x) \\
&\text{such that } f_i(x) \leq 0 \quad i = 1, \dots, m \\
&\qquad\qquad h_i(x) = 0 \quad i = 1, \dots, p
\end{aligned}$$

Given the variable $x \in \mathbb{R}^n$, domain D, and optimal value p^*. We have Lagragian to be $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ with domain $L = D \times \mathbb{R}^m \times \mathbb{R}^p$:

$$\mathcal{L}(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

where we have:

- Weight sum of objective and constant functions.
- λ_i is lagragian multiple associated wth $f_i(x) \leq 0$
- v_i is lagragian multiple associated with $h_i(x) = 0$

Definition 1.5.2. (Dual Function) The function $g : \mathbb{R}^m \times \mathbb{R}^D \to \mathbb{R}$:

$$\begin{aligned}
g(\lambda, v) &= \inf_{x \in D} L(x, \lambda, v) \\
&= \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x) \right)
\end{aligned}$$

Note that g is concave and it can be $-\infty$ for some λ, v.

Proposition 1.5.1. *If $\lambda \succeq 0$ then $g(\lambda, v) \leq p^*$*

Proof. If $\tilde{x}$ is feasible and $\lambda \succeq 0$ then:

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, v) \geq \inf_{x \in D} L(x, \lambda, v) = g(\lambda, v)$$

The minimizing over all feasible $\tilde{x}$ gives $p^* \geq g(\lambda, v)$ □

Remark 11. The least norn solution for linear equation, which we have:

$$\begin{aligned}
&\min x^T x \\
&\text{such that } Ax = b
\end{aligned}$$

The lagragian is given by $L(x, v) = x^T x + v^T(Ax - b)$. Let's try to minimize the Lagragian by finding the gradient with respected to x:

$$\nabla_x L(x, v) = 2x + A^T v = 0 \implies x = -(1/2)A^T v$$

Plugging back to L gives us:

$$g(v) = L(-(1/2)A^T v, v) = -\frac{1}{4} v^T A A^T v - b^T v$$

and it is concave function of v. Furtheremore, the lower bound is $p^* \geq -\frac{1}{4} v^T A A^T v - b^T v$ for all v.

Remark 12. If we consider the standard form of LP:

$$\begin{aligned} &\min c^T x \\ &\text{such that } Ax = b \\ &\qquad\qquad x \succeq 0 \end{aligned}$$

The Lagragian is:

$$\begin{aligned} L(x, \lambda, v) &= c^T x + v^T (Ax - b) + x^T x \\ &= -b^T v + (c + A^T v - \lambda)^T x \end{aligned}$$

Note that if L is affine in x, then we have:

$$g(\lambda, v) = \inf_x L(x, \lambda, v) = \begin{cases} -b^T v & \text{if } A^T v - \lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Note that g is linear on affine domain $\{(\lambda, v) | A^T v - \lambda + c = 0\}$, hence concave. Now, the lower bound property: $p^* \geq -b^T v$ if $A^T v + c \succeq 0$

Remark 13. Given the equality constrained norm minimization:

$$\begin{aligned} &\min \|x\| \\ &\text{such that } Ax = b \end{aligned}$$

The dual function is

$$g(v) = \inf_x \|x\| - v^T Ax + b^T v = \begin{cases} b^T v & \text{if } \left\|A^T v\right\|_* \leq 1 \\ -\infty & \text{otherwise} \end{cases}$$

where $\|v\|_* = \sup_{\|u\| \leq 1} u^T v$ is dual norm of $\|\cdot\|$. With the lower bound property: $p^* \geq b^T v$ if $\left\|A^T v\right\|_* \leq 1$

Proposition 1.5.2. *We have* $\inf_x \|x\| - y^T x = 0$ *if* $\|y\|_* \leq 1$ *and* $-\infty$ *otherwise.*

Proof. Then, we have:

- If $\|y\|_* \leq 1$ then $\|x\| - y^T x \geq 0$ for all x with equality if $x \geq 0$
- If $\|y\|_* > 1$ choose $x = tu$ where $\|u\| \leq 1$ and $u^T y = \|y\|_* > 1$:

  $$\|x\| - y^T x = t(\|u\| - \|y\|_*) \to -\infty$$

  as $t \to \infty$.

□

Definition 1.5.3. (Two-Way Partitioning) Given the two way partitioning:

$$\begin{aligned} &\min x^T W x \\ &\text{such that } x_i^2 = 1 \quad i = 1, \ldots, n \end{aligned}$$

This is non-convex problem with a feasible set contains 2^n discrete points. The interpretation is partition $\{1, \ldots, n\}$ in 2 sets. Given the weight W_{ij} is the cost assigning ij into same set and $-W_{ij}$ is the const of defining a different set.

Remark 14. The dual function of two-way partitioning is:

$$\begin{aligned} g(v) &= \inf_x \left(x^T W x + \sum_i v_i (x_i^2 - 1) \right) = \inf_x x^T (W + \operatorname{diag}(v)) x - 1^T v \\ &= \begin{cases} -1^T v & \text{if } w + \operatorname{diag}(v) \succeq 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

Now we have lower bound property $p^* \geq -1^T v$ if $W + \operatorname{diag}(v) \succeq 0$

Proposition 1.5.3. *We have linear programming problem:*

$$\begin{aligned}&\min f_0(x)\\&\text{such that } Ax \preceq b\\&\qquad\qquad Cx = d\end{aligned}$$

Now, consider the dual function:

$$\begin{aligned}g(\lambda, v) &= \inf_{x\in\operatorname{dom} f_0}\Big(f_0(x) + (A^T\lambda + C^Tv)^Tx - b^T\lambda - d^Tv\Big)\\&= -f_0^*(-A^T\lambda - C^Tv) - b^T\lambda - d^Tv\end{aligned}$$

recall the definition of conjugate function $f^*(\cdot)$. The dual if conjugate of f_0 is known.

Remark 15. The example of entropy maximization, we have:

$$f_0(x) = \sum_{i=1}^n x_i \log x_i \qquad f^*(x) = \sum_{i=1}^n \exp(y_i - 1)$$

1.5.2 Dual Problems

Definition 1.5.4. (Lagragian Dual Problem) We have the following problem:

$$\begin{aligned}&\min g(\lambda, v)\\&\text{subject to } \lambda \succeq 0\end{aligned}$$

We find the lower bound on p^* to obtained from Lagragian dual function. Optimal value denote d^*. λ, v are dual feasible if $\lambda \succeq 0$ where $(\lambda, v) \in \operatorname{dom}(g)$. We often simplify by making the implicit constrain $(\lambda, v) \in \operatorname{dom}(g)$ explicit.

Definition 1.5.5. (Weak/Strong-Duality) We consider 2 cases:

- If we have $d^* \le p^*$, this always hold. It can be used to find non-trivial lower bound for difficult problem.
- Otherwise $d^* = p^*$, this doesn't hold in general. We usually hold for convex problem. The conditions that gurantee that gurantee strong duality in convex problem is called constriant qualificaition.

Remark 16. For example, solving the SDP:

$$\begin{aligned}&\min -1^Tv\\&\text{subject to } w + \operatorname{diag}(v) \succeq 0\end{aligned}$$

gives a lower bound for 2 ways partitioning problem

Definition 1.5.6. (Slater's Constraint Qualification) The strong duality holds for convex problem:

$$\begin{aligned}&\min f_0(x)\\&\text{subject to } f_i(x) \le 0 \quad i = 1, \dots, m\\&\qquad\qquad Ax = b\end{aligned}$$

if it is strictly feasible: there exists $x \in \operatorname{int}(D)$

$$f_i(x) < 0 \quad i = 1, \dots, m \quad Ax = b$$

Guarantee that the dual optimum is attained (if $p^* > \infty$). Note that this can be sharpen: $\operatorname{int}(D)$ can be replaced with $\operatorname{relint}(D)$. There exists other types of constraint qualificaition.

Remark 17. **(Linear Programming)** Now, we have inequality for Linear Programming: The primal problem is:

$$\begin{aligned} &\min c^T x \\ &\text{subject to } Ax \preceq b \end{aligned}$$

Together with the dual function:

$$g(\lambda) = \inf_x ((c + A^T\lambda)^T x - b^T\lambda) = \begin{cases} -b^T\lambda & \text{if } A^T\lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Now, the dual problem is:

$$\begin{aligned} &\min -b^T\lambda \\ &\text{subject to } A^T\lambda + c = 0 \\ &\qquad\qquad \lambda \succeq 0 \end{aligned}$$

From Slanter's Constraint, $p^* = d^*$ if $A\tilde{x} \prec b$ for some $\tilde{x}$. In fact $p^* = d^*$ except when primal and dual are infeasible.

Remark 18. **(Quadratic Program)** For quadratic program, where we have primal problem (assuming $P \in \mathbb{S}^n_{++}$):

$$\begin{aligned} &\min x^T P x \\ &\text{subject to } Ax \preceq b \end{aligned}$$

The dual function:

$$g(\lambda) = \inf_x (x^T P x + \lambda^T(Ax - b)) = -\frac{1}{4}\lambda^T A P^{-1} A^T \lambda - b^T\lambda$$

This we have the dual problem to be:

$$\begin{aligned} &\min -\frac{1}{4}\lambda^T A P^{-1} A^T \lambda - b^T\lambda \\ &\text{subject to } \lambda \succeq 0 \end{aligned}$$

From Slater condition $p^* = d^*$ if $A\tilde{x} \prec b$ for some $\tilde{x}$ in fact $p^* = d^*$ always.

Remark 19. **(Non-Convex Problem with Strong Duality)** We have the following non-convex problem:

$$\begin{aligned} &\min x^T A x + 2b^T x \\ &\text{subject to } x^T x \leq 1 \end{aligned}$$

when $A \not\succeq 0$ is non-convex. Given a dual function:

$$g(\lambda) = \inf_x (x^T(A + \lambda I)x + 2b^T x - \lambda)$$

The undbounded below if $A + I\lambda \not\succeq 0$ or $A + I\lambda \succeq 0$ and $b \notin \mathcal{R}(A + I\lambda)$, where it is linear combination of columns. This is minimized by $x = -(A + \lambda I)^\dagger b$ and $g(\lambda) = -b^T(A + I\lambda)^\dagger b - \lambda$. Now the dual problem:

$$\begin{aligned} &\min -b^T(A + I\lambda)^\dagger b - \lambda \\ &\text{subject to } A + \lambda I \succeq 0 \\ &\qquad\qquad b \in \mathcal{R}(A + I\lambda) \end{aligned}$$

is equivalent to:

$$\begin{aligned} &\min -t - \lambda \\ &\text{subject to } \begin{bmatrix} A + I\lambda & b \\ b^T & t \end{bmatrix} \succeq 0 \end{aligned}$$

We can have strong duality although the primal problem isn't convex.

Definition 1.5.7. (Complementary Slackness) Assume strong duality holds, x^* is primal optimal (λ^*, v^*) is dual optimal:

$$\begin{aligned} f_0(x^*) = g(\lambda^*, v^*) &= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} v_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{n} v^* h_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$

Hence the 2 inequalities hold with equality, if:

- x^* minimizes $L(x, \lambda^*, v^*)$
- $\lambda_i^* f_i(x^*) = 0$ for $i = 1, \ldots, m$ (known as complementatry slackness):

$$\begin{aligned} \lambda_i^* > 0 &\implies f_i(x^*) = 0 \\ f_i(x^*) < 0 &\implies \lambda_i^* = 0 \end{aligned}$$

Definition 1.5.8. (KKT Condtion) The following 4 conditions are called KKT condition (for a problem with differentiable f_i and h_i):

- Primal constraints:

$$\begin{aligned} f_i(x) &\leq 0 \text{ for } i = 1, \ldots, m \\ h_i(x) &= 0 \text{ for } i = 1, \ldots, p \end{aligned}$$

- Dual Constraints $\lambda \succeq 0$
- Complementary Slackness: $\lambda_i f_i(x) = 0$ for $i = 1, \ldots, m$
- Gradient of Lagragian with respected to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{n} v_i \nabla h_i(x) = 0$$

The strong duality holds and x, λ, v are optimal, then it must satisfy KKT condition.

Proposition 1.5.4. *If $\tilde{x}, \tilde{\lambda}, \tilde{v}$ satisfy KKT for convex problem, when they are optimal:*

- *From complementatry slackness: $f_0(\tilde{x}) = L(\tilde{x}, \tilde{\lambda}, \tilde{v})$*
- *From the forth condition and convexity: $g(\tilde{\lambda}, \tilde{v}) = L(\tilde{x}, \tilde{\lambda}, \tilde{v})$*

hence $f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{v})$

Proposition 1.5.5. *If slanter's condition is satisfied: x is optimal iff λ, v that satisfies KKT condition:*

- *Recall that slanter implies strong duality and dual optimal is allowed.*
- *The generalies optimality condition $\nabla f(x) = 0$ for unconstrained problems.*

Remark 20. Perturbation and Sensitivity analysis. Consider unperturbed optimization problem and its dual:

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq 0 \quad i = 1, \ldots, m \\ &\qquad\qquad\quad q_i(x) = 0 \quad i = 1, \ldots, p \end{aligned}$$

Its dual is:

$$\max g(\lambda, \nu)$$
$$\text{subject to } \lambda \succ 0$$

Now, the perturbed problem and its dual is:

$$\begin{aligned} &\min f_0(x) \\ &\text{subject to } f_i(x) \leq u_i \quad i = 1, \dots, m \\ &\qquad\qquad g_i(x) = v_i \quad i = 1, \dots, p \end{aligned}$$

and its dual is:

$$\begin{aligned} &\max g(\lambda, \nu) - u^T \lambda - v^T \nu \\ &\text{subject to } \lambda \succeq 0 \end{aligned}$$

We have:

- x as primal variable and u, ν are parameters.
- $p^*(u, v)$ is optimal value as a function of u, v
- We are interested about $p^*(u, v)$ that we can obtain from the solution of unperturbed problems and its dual.

Assume strong duality holds for unperturbed problems and that λ^* and ν are dual optimal for unperturbed problem.

$$\begin{aligned} p^*(u, v) &\geq g(\lambda, \nu^*) - u^T \lambda^* - v^T \nu^* \\ &= p^*(0, 0) - u^T \lambda^* - v^T \nu^* \end{aligned}$$

Given a statistical interpretation:

- If λ_i^* is large, p^* increases greatly if we tighten constraint i $(u_i < 0)$
- If λ_i^* is small, p^* doesn't decrease much if we loosen constraint i $(u_i \geq 0)$
- If ν^* is large and positive: p^* increases greatly if we have $v_i < 0$
- If ν^* is large and negative: p^* increases greatly if we have $v_i > 0$
- If ν^* is small and positive: p^* doesn't decrease much if we take $v_i > 0$
- If ν_i^* is small and negative: p^* doesn't decrease much if we take $v_i < 0$

Lemma 1.5.1. *If $p^*(u, v)$ is differentiable at $(0, 0)$ then:*

$$\lambda_i^* = \frac{\partial p^*(0, 0)}{\partial u_i} \qquad \nu_i^* = -\frac{\partial p^*(0, 0)}{\partial v_i}$$

Proof. For λ_i^* from global sensitivity result:

$$\frac{\partial p^*(0, 0)}{\partial u_i} = \lim_{t \searrow 0} \frac{p^*(te_i, 0) - p^*(0, 0)}{t} \geq -\lambda_i^* \qquad \frac{\partial p^*(0, 0)}{\partial u_i} = \lim_{t \nearrow 0} \frac{p^*(te_i, 0) - p^*(0, 0)}{t} \leq -\lambda_i^*$$

Hence equality. □

1.5.3 Techniques of Solving Dual Problems

Remark 21. We have an equivalent formulations of a problem can lead to very different duals. Reformulating the primal problem can be useful when the duals is difficult to derive. The common reformulations are:

- Introduces new variables and equality constrains
- Make explicit constraint implicit or vice versa
- Transform object or constant functions: Replace $f_0(x)$ by $\phi(f_0(x))$ with ϕ convex and increasing.

Definition 1.5.9. (New Variable and Equality Contraint)

$$\max f_0(Ax+b)$$

This is dual function is constant $g = \inf_x L(x) = \inf_x f_)(Ax+b) = p^*$ but this is useless:

$$\begin{aligned} &\min f_0(y) \\ &\text{subject to } Ax+b-y=0 \end{aligned}$$

Now, its dual is

$$\begin{aligned} &\max b^T\nu - f_0^*(\nu) \\ &\text{subject to } A^T\nu = 0 \end{aligned}$$

As the dual function forms:

$$\begin{aligned} g(\nu) &= \inf_{x,y}(f_0(y) - \nu^T y + \nu^T Ax + b^T\nu) \\ &= \begin{cases} -f_0^*(\nu) + b^T\nu & \text{if } A^T\nu = 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

Remark 22. **(Norm Approximation Problem)** We would like to minimize $\|Ax-b\|$. This is the same as:

$$\begin{aligned} &\min \|y\| \\ &\text{subject to } Ax-b=y \end{aligned}$$

We have the following dual function:

$$\begin{aligned} g(\nu) &= \inf_{x,y}\Big(\|y\| + \nu^T y - \nu^T Ax + b^T\nu\Big) \\ &= \begin{cases} b^T\nu + \inf_x \Big(\|y\| + \nu^T y\Big) & \text{if } A^T\nu \le 1 \\ -\infty & \text{otherwise} \end{cases} \\ &= \begin{cases} b^T\nu & \text{if } A^T\nu = 0, \|\nu\|_* \le 1 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

And, so we have dual of the norm approximation problem is:

$$\begin{aligned} &\max b^T\nu \\ &\text{subject to } A^T\nu = 0 \\ &\qquad\qquad \|\nu\|_* \le 1 \end{aligned}$$

Definition 1.5.10. (Implicit Constraint) Let's consider the linear programming with box constriants, which we have:

$$\begin{aligned} &\min c^T x \\ &\text{subject to } Ax = b \\ &\qquad\qquad -1 \preceq x \preceq 1 \end{aligned}$$

And its dual is

$$\begin{aligned}
&\min -b^T\nu - 1^T\lambda_1 - 1^T\lambda_2 \\
&\text{subject to } c + A^T\nu + \lambda_1 - \lambda_2 \\
&\qquad\qquad \lambda_1 \succeq 0 \\
&\qquad\qquad \lambda_2 \succeq 0
\end{aligned}$$

However, we can simplify by reformulate the box constraint and make the constriant explicit:

$$\begin{aligned}
&\min f_0(x) = \begin{cases} c^T x & \text{if } -1 \succeq x \succeq 1 \\ \infty & \text{otherwise} \end{cases} \\
&\text{subject to } Ax = b
\end{aligned}$$

Now, the dual function becomes:

$$\begin{aligned}
g(\nu) &= \inf_{-1 \preceq x \preceq 1} c^T x + \nu^T(Ax - b) \\
&= -b^T\nu - \left\|A^T\nu + c\right\|_1
\end{aligned}$$

Now, the dual problem is equal to $\max -b^T\nu - \left\|A^T\mu + c\right\|_1$

Definition 1.5.11. (Problems with Generalized Inequalities) We consider the following problem:

$$\begin{aligned}
&\min f_0(x) \\
&\text{subject to } f_i(x) \preceq_{K_i} 0 \quad i = 1, \dots, m \\
&\qquad\qquad h_i(x) = 0 \quad i = 1, \dots, p
\end{aligned}$$

Where $\preceq_{K_i}$ of generalized inequality on $\mathbb{R}^{k_i}$. There are parallels to the scalar case:

- Lagragian multiplier for $f_i(x) \preceq_{K_i} 0$ is a vector $\lambda_i \in \mathbb{R}^{K_i}$
- Lagragian $L : \mathbb{R}^n \times \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_m} \times \mathbb{R}^d \to \mathbb{R}$

$$L(x, \lambda_1, \dots, \lambda_m, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i^T f_i(x) + \sum_{i=1}^{D} \nu_i h_i(x)$$

- Dual Function is $g : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_m} \times \mathbb{R}^D \to \mathbb{R}$ is defined as:

$$g(\lambda_1, \dots, \lambda_m, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda_1, \cdots, \lambda_m, \nu)$$

- Lower bound property: If $\lambda_i \succeq_{K_i^*} 0$ then $g(\lambda_1, \dots, \lambda_m) \leq p^*$
- Dual Problem:

$$\begin{aligned}
&\max f_0(\lambda_1, \dots, \lambda_m, \nu) \\
&\text{subject to } \lambda_i \succeq_{K_i^*} 0 \quad i = 1, \dots, m
\end{aligned}$$

  The weak duality $p^* \geq d^*$. The strong duality is $p^* = d^*$ for some convex problem with constraint optimization (Slater).

Remark 23. To show that the lower bound property is true, we have:

$$\begin{aligned}
f_0(\tilde{x}) &\geq f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i^T f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \\
&\geq \inf_{x \in \mathcal{D}} L(x, \lambda_1, \dots, \lambda_m, \nu) \\
&= g(\lambda_1, \dots, \lambda_m, \nu)
\end{aligned}$$

Mininize over all feasible $\tilde{x}$ will give us $p^* \geq g(\lambda_1, \dots, \lambda_m, \nu)$

Definition 1.5.12. (Semi-Definite Program) The primal SDP is given by $(F_i, G \in \mathbb{S}^k)$

$$\begin{aligned}&\min c^T x\\&\text{subject to } x_1 F_1 + \cdots + x_n F_n \preceq G\end{aligned}$$

The lagrange multiplier is $Z \in \mathbb{R}^K$ where

$$L(x, Z) = c^T x + \operatorname{tr}(Z(x_1 F_1 + \cdots + x_n F_n - G))$$

Dual function is:

$$g(Z) = \inf_x L(x, Z) = \begin{cases} -\operatorname{tr}(GZ) & \text{if } \operatorname{tr}(F_i Z) + c_i = 0 \quad i = 1, \ldots, n \\ -\infty & \text{otherwise} \end{cases}$$

The dual SDP is defined as:

$$\begin{aligned}&\max -\operatorname{tr}(GZ)\\&\text{subject to } Z \succeq 0 \quad \operatorname{tr}(F_i Z) + c_i = 0, i = 1, \ldots, n\end{aligned}$$

$p^* = d^*$ if primal SDP is strictly feasible (there exists x with $x_1 F_1 + \cdots + x_n F_n \prec G$)

Chapter 2

RKHS in Machine Learning

2.1 Introduction to RKHS

2.1.1 Building a Kernel

Definition 2.1.1. (Kernel) Let $\mathcal{X}$ be non-empty set, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$:

$$k(x, x') = \langle \phi(x), \phi(x) \rangle_{\mathcal{H}}$$

Remark 24. For a single kernel, there can be multiple features. For example, the map

$$\phi_1(x) = x \qquad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

corresponds to the same kernel.

Theorem 2.1.1. *Given $\alpha > 0$ and k, k_1, k_2 be kernel on $\mathcal{X}$, then: αk, $k_1 + k_2$, and $k_1 \times k_2$ are kernels.*

Proof. **Scalar Multiplication:** Suppose $k(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{H}}$, with a feature map $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$ and some points $x, x' \in \mathcal{X}$,we can see that

$$\alpha k(x, x') = \left\langle \sqrt{\alpha}\phi(x), \sqrt{\alpha}\phi(x') \right\rangle_{\mathcal{H}}$$

where the new feature map is $\sqrt{\alpha}\phi(\cdot)$

Kernel Addition: Suppose $k_1(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{A}}$ and $k_2(\cdot, \cdot) = \langle \psi(\cdot), \psi(\cdot) \rangle_{\mathcal{B}}$, where $\phi : \mathcal{X} \to \mathcal{A}$ and $\psi : \mathcal{X} \to \mathcal{A}$ are features map. Then, we can see that, for point $x, x' \in \mathcal{X}$:

$$(k_1 + k_2)(x, x') = k_1(x, x') + k_2(x, x') = \langle (\phi \| \psi)(\cdot), (\phi \| \psi)(\cdot) \rangle_{\mathcal{A}}$$

where we define:

$$\phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \\ \vdots \end{bmatrix} \qquad \psi(x) = \begin{bmatrix} \psi_1(x) \\ \psi_2(x) \\ \psi_3(x) \\ \phi_4(x) \\ \vdots \end{bmatrix} \qquad (\phi || \psi)(x) = \begin{bmatrix} \phi_1(x) \\ \psi_1(x) \\ \phi_2(x) \\ \psi_2(x) \\ \vdots \end{bmatrix}$$

Kernel Multiplication: We assume same kernel k_1, k_2. We have

$$\begin{aligned} k_1(x_1,x_2)k(x_1,x_2) &= \Big(\phi^T(x)\phi(x)\Big)\cdot\Big(\psi^T(x)\psi(x)\Big) \\ &= \text{tr}\Big(\phi^T(x)\phi(x)\psi^T(x)\psi(x)\Big) \\ &= \text{tr}\Big(\psi(x)\phi^T(x)\phi(x)\psi^T(x)\Big) \\ &= \text{tr}\Big(\left[\phi(x)\psi^T(x)\right]^T\phi(x)\psi^T(x)\Big) \end{aligned}$$

The feature map for product kernel is $\Phi(\cdot) = \phi(\cdot)\psi^T(\cdot)$, and the inner product is defined as: for matrix A, B:

$$\langle A, B\rangle = \text{tr}(A^T B)$$

□

Proposition 2.1.1. *Let $\mathcal{X}$ and $\tilde{\mathcal{X}}$ be a set, and define a map $A : \mathcal{X} \to \tilde{\mathcal{X}}$ we can define a kernel $k(\cdot,\cdot)$ on $\tilde{\mathcal{X}}$, then:*

$$k(A(\cdot), A(\cdot))$$

is a kernel.

Proof. the new kernel $\tilde{k}$ can be expressed as $\langle\psi(\cdot),\psi(\cdot)\rangle_{\tilde{\mathcal{X}}}$, where $\psi = \phi\circ A$. □

Proposition 2.1.2. *Given the kernel k_1, k_2 (with associated feature map ϕ and ψ ,respectively – note that they don't have to be unique), $k_1 - k_2$ doesn't need to be kernel, nor $|k_1 - k_2|$*

Proof. Given x where $k(x,x') = \langle\phi(x),\phi(x')\rangle_{\mathcal{H}}$, we can see that $(k-k)(x,x) = (|k-k|)(x,x) = 0$, however as the feature map doesn't maps all x to zero vector, it contradicts the definition of inner product as the product can't be zero unless both of the vectors are zero. □

Definition 2.1.2. (Polynomial Kernel) Given theorem 2.1.1, we can construct a polynomial kernel as:

$$k(x,x') = (c + \langle x, x'\rangle)^m$$

and it is valid kernel.

Definition 2.1.3. (Taylor Series Kernel) For $r \in (0,\infty]$ with $a_n \geq 0$ for all $n \geq 0$, we have:

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

for $|z| < r, z \in \mathbb{R}$ and we define $\mathcal{X}$ to be $\sqrt{r}$-ball in $\mathbb{R}^d$, then the Taylor series kernel is defined as:

$$k(x,x') = f(\langle x,x'\rangle) = \sum_{n=0}^{\infty} a_n \langle x, x'\rangle^n$$

Lemma 2.1.1. *Taylor series kernel is kernel*

Proof. There are 2 points we have to proof:

- **Taylor Series Converges:** Let's show that the value of $\langle x, x'\rangle$ is less than or equal to r to make sure that Taylor series converges. This is the application of Cauchy-Schwarz inequality as $|\langle x,x'\rangle| \leq \|x\|\cdot\|x'\| < r$.
- **Taylor Series Kernel is Kernel:** Now, from theorem 2.1.1, we have an addition of kernels and multiplication to scalar, thus being a kernel.

□

Definition 2.1.4. (Exponentiated Quadratic Kernel) We define an exponentiated Quadratic kernel to be

$$k(x, x') = \exp\left(-\gamma^{-2} \|x - x'\|^2\right)$$

Corollary 2.1.1. *Exponentiated Quadratic Kernel is kernel.*

Proof. Let's expand the definition of a square normed, then we have:

$$\begin{aligned}\exp\left(-\gamma^{-2} \|x - y\|^2\right) &= \exp\left(-\gamma^{-2} \left[\|x\|^2 - 2\langle x, y\rangle + \|y\|^2\right]\right) \\ &= \underbrace{\exp\left(-\gamma^{-2} \|x\|^2\right) \exp\left(-\gamma^{-2} \|y\|^2\right)}_{k_1(x,y)} \cdot \underbrace{\exp\left(2\gamma^{-2} \langle x, y\rangle\right)}_{k_2(x,y)}\end{aligned}$$

Thus, we have a product of 2 kernels, where one of them is produced from a feature map $\exp(-\gamma^{-2} \|\cdot\|^2)$ and the other comes from the Taylor series Kernel together with non-negative multiplication. □

Definition 2.1.5. (l_2-Space) The space l_2 comprised of all sequences $a = (a_i)_{i\geq 1}$ for which

$$\|a\|_{l_2}^2 = \sum_{l=1}^{\infty} a_l^2 < \infty$$

Definition 2.1.6. (Infinity Dimension Kernel) Given a sequence of function $(\phi(x)_i)_{i\geq 1}$ in l_2 where $\phi_i : \mathcal{X} \to \mathbb{R}$ being the i-th coordinate of ϕ, then we can define an infinity dimension kernel to be

$$k(x, x') = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')$$

Theorem 2.1.2. *Infinity Dimension Kernel is a kernel.*

Proof. We consider the norm of the kennel, and apply Cauchy Schwarz i.e:

$$\|k(x, x')\| = \left\|\sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')\right\| \leq \|\phi(x)\| \cdot \|\phi(x')\| \leq \infty$$

□

2.1.2 Further Notions of Kernels and RKHS

Definition 2.1.7. (Positive Definite) A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if: for all $a_1, a_2, \ldots, a_n \in \mathbb{R}^n$ and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

The function $k(\cdot, \cdot)$ is strictly positive definite if equality holds when $a_i, a_j \neq 0$.

Theorem 2.1.3. *Let $\mathcal{H}$ be Hilbert space, $\mathcal{X}$ be non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $k(x, y) = \langle\phi(x), \phi(y)\rangle$ is positive definite.*

Proof. For all $a_1, a_2, \ldots, a_n \in \mathbb{R}^n$ and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i\phi(x_i), a_j\phi(x_j)\rangle$$
$$= \left\langle \sum_{i=1}^{n} a_i\phi(x_i), \sum_{j=1}^{n} a_j\phi(x_j) \right\rangle = \left\| \sum_{j=1}^{n} a_j\phi(x_j) \right\|^2 \geq 0$$

□

Definition 2.1.8. (Notion of Function) We will represent a function, throughout the note, as a vector of real numbers; for instance, $f(\cdot) = [f_1 \ f_2 \ f_3]^T$, its evaluation will be based on a feature map $\phi(x)$, as $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$ as $\mathcal{H}$ is space of functions.

Remark 25. Let's consider the example of $f : \mathbb{R}^2 \to \mathbb{R}$ as:

$$f(x) = \langle f, \phi(x)\rangle = f_1x_1 + f_2x_2 + f_3(x_1x_2) \quad \text{where} \quad \phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1x_2 \end{bmatrix}$$

Remark 26. **(Representing Function as Finite Sum of Kernels)** This notion of function can be represented by infinity many feature of f and $\phi(\cdot)$ as the function, which will be shown as:

$$f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} f_l\phi_l(x)$$

As we required that $\sum_{l=1}^{\infty} f_l^2 \leq \infty$ We will assume that f_l can be represented in finite linear combination of the features $\phi_l(x)$:

$$f_l = \sum_{i=1}^{m} \alpha_i\phi_l(x_i)$$

Then, we have:

$$f(x) = \left\langle \sum_{i=1}^{m} \alpha_i\phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

Now, a function with infinite feature can be represented by a finite linear combination of kernels given a certain number of points.

Remark 27. **(Feature Map is also a function)** Let's consider the simpliest case of $m = 1$ with $\alpha_1 = 1$, we have

$$f(x) = k(x_1, x) = \left\langle \underbrace{k(x_1, \cdot)}_{f(\cdot)}, \phi(x) \right\rangle_{\mathcal{H}} = \langle k(x, \cdot), \phi(x_1)\rangle$$

And, so we have a kernel parameterized by x_1, which is a feature map by definition. And thus, we can "swap" the notation around and assigned the coefficient to be $\phi(x_1)$, thus feature map is a function.Please note that, we can write the kernel as

$$k(x, y) = \langle k(x_1, \cdot), k(x_2, \cdot)\rangle_{\mathcal{H}}$$

Now, $k(x, \cdot)$ is called canonical feature map as it is the simpliest, while there are many feature map (potentially infinite) that can construct this kernel. This means that the space of function $\mathcal{H}$ is bigger than all features at single point as it is an combination of functions.

Definition 2.1.9. (Reproducing Property) The features of RKHS have reproducing property, where for all $x \in \mathcal{X}$ and for $f(\cdot) \in \mathcal{H}$:

$$f(x) = \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}}$$

The feature map of every point is a function of kernel $k(\cdot, x) = \phi(x) \in \mathcal{H}$ where for any $x \in \mathcal{X}$, we have:

$$k(x, x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}}$$

2.2 Smoothness of RKHS

2.2.1 Periodic Case

Definition 2.2.1. (Fourier Series) We define a fourier series that represents the function on interval $[-\pi, \pi]$ with periodic boundary as:

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(ilx) = \sum_{l=-\infty}^{\infty} \hat{f}_l(\cos(lx) + \sin(lx))$$

We would like to note that the basis functions are orthogonal to each other as

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \exp(ilx)\overline{\exp(imx)}\, \mathrm{d}x = \begin{cases} 1 & l = m \\ 0 & \text{otherwise} \end{cases}$$

Definition 2.2.2. (Translation Invariance) Translation invariance kernel is kernel that is defined by

$$k(x, y) = k(x - y)$$

Remark 28. Fourier representation of translation invariance kernel is

$$k(x, y) = \sum_{l=-\infty}^{\infty} \hat{k}_l \exp(il(x - y)) = \sum_{l=-\infty}^{\infty} \underbrace{\left[\sqrt{\hat{k}_l} \exp(ilx)\right]}_{\phi_l(x)} \underbrace{\left[\sqrt{\hat{k}_l} \exp(-ily)\right]}_{\overline{\phi_l(y)}}$$

Proposition 2.2.1. *The L_2 inner product of the function can be represented by Fourier series as:*

$$\langle f, g\rangle_{L_2} = \sum_{l=-\infty}^{\infty} \hat{f}_l\overline{\hat{g}_l}$$

Proof. We expand on the definition of inner product in L_2:

$$\begin{aligned}
\langle f, g\rangle_{L_2} &= \int_{-\infty}^{\infty} f(x)\overline{g(x)}\, \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left[\sum_{l=-\infty}^{\infty} \hat{f}_l \exp(ilx)\right] \cdot \overline{\left[\sum_{l=-\infty}^{\infty} \hat{g}_l \exp(ilx)\right]} \\
&= \int_{-\infty}^{\infty} \left[\sum_{l=-\infty}^{\infty} \hat{f}_l \exp(ilx)\right] \cdot \left[\sum_{l=-\infty}^{\infty} \hat{g}_l \exp(-ilx)\right] \\
&= \int_{-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \hat{f}_l\overline{\hat{g}_l}\, \mathrm{d}x + \int_{-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \sum_{j\neq k} \hat{f}_j\overline{\hat{g}_k} \exp(ijx)\overline{\exp(ikx)}\, \mathrm{d}x \\
&= \sum_{l=-\infty}^{\infty} \hat{f}_l\overline{\hat{g}_l}
\end{aligned}$$

□

Definition 2.2.3. (Smooth Dot Product) Recall the coefficient $\hat{k}_l$ from remark 28, we define an inner product in $\mathcal{H}$ to be

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l\overline{\hat{g}_l}}{\hat{k}_l}$$

And so, we define a dot product to be:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \overline{\hat{g}_l}}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}$$

In the case that $\hat{k}_l$ decays fast, we need to have $\hat{f}_l$ to be fast too in order to have bounded sum.

Remark 29. Given the Jacobi-Theta Kernel:

$$k(x,y) = \frac{1}{2\pi}\vartheta\left(\frac{x-y}{2\pi}, \frac{i\sigma^2}{2\pi}\right) \qquad \hat{k}_i = \frac{1}{2\pi}\exp\left(-\frac{\sigma^2 l^2}{2}\right)$$

as it is a Gaussian version of "periodic" kernel. Now given the top hat function, which is a function:

$$f(x) = \begin{cases} 1 & |x| < T \\ 0 & T \le |x| < \pi \end{cases} \qquad \hat{f}_l = \frac{\sin(lT)}{l\pi}$$

We can see that the top hat function isn't in a Gaussian spectrum RKHS. As we can show that $\|f\|_{\mathcal{H}}^2$ won't converge. This is because $|\hat{f}_l|^2$ decays polynomial in l, while $\hat{k}_l$ decays in exponential of l. Thus, the norm doesn't converge.

Proposition 2.2.2. *We can show that*

$$\langle f(\cdot), k(\cdot, z)\rangle_{\mathcal{H}} = f(z)$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ *is defined in 2.2.3. Thus, it has the reproducing property. And, we can show that:*

$$\langle k(\cdot, y), k(\cdot, z)\rangle = k(y, z)$$

Proof. **First Statement:** We consider the following function:

$$g(x) = k(x-z) = \sum_{l=-\infty}^{\infty} \exp(ilx)\underbrace{\hat{k}_l\exp(-ilz)}_{g_l}$$

Now, the dot product is equal to:

$$\langle f(\cdot), g(\cdot)\rangle = \sum_{l=-\infty}^{\infty} \hat{f}_l \frac{\hat{k}_l \exp(ilz)}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(ilz) = f(z)$$

Similarly, we can consider 2 functions $f(x) = k(x-y)$ and $g(x) = k(x-z)$, where

$$f(x) = \sum_{l=-\infty}^{\infty} \exp(ilx)\underbrace{\exp(-ily)\hat{k}_l}_{\hat{f}_l} \qquad g(x) = \sum_{l=-\infty}^{\infty} \exp(ilx)\underbrace{\exp(-ilz)\hat{k}_l}_{\hat{g}_l}$$

Second Statement: And, so the reproducing we have:

$$\begin{aligned}\langle f(\cdot), g(\cdot)\rangle &= \sum_{l=-\infty}^{\infty} \frac{\hat{k}_l\exp(-ily)\overline{\hat{k}_l\exp(-ilz)}}{\hat{k}_l} \\ &= \sum_{l=-\infty}^{\infty} \hat{k}_l \exp(il(z-y)) = k(z-y)\end{aligned}$$

□

Remark 30. Recalling that function can be represented as:

$$f(z) = \sum_{l=-\infty}^{\infty} f_l \overline{\phi_l(z)}$$

Now, recall the function $f(z)$ shown in proposition 2.2.2.

$$\langle f(\cdot), g(\cdot)\rangle = \sum_{l=-\infty}^{\infty} \hat{f}_l \frac{\overline{\hat{k}_l \exp(-ilz)}}{\left(\sqrt{\hat{k}_l}\right)^2}$$

Then, we have

$$f_l = \hat{f}_l / \sqrt{\hat{k}_l} \qquad \phi_l(z) = \sqrt{\hat{k}_l} \exp(-ilz)$$

2.2.2 Eigen Expansion Case

Remark 31. We are going to extension of the definition of RKHS to eigenexpansion as fourier series only gives us the periodic domain $[-2\pi, 2\pi]$

Definition 2.2.4. (Eigenfunction/Eigenvalue) We define a probability measure on $\mathcal{X} = \mathbb{R}$, where we will use Gaussian density:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$$

We define an eigenfunction $e_l(\cdot)$ and eigenvalue λ_l on $k(x, x')$ wrt. to this measure as

$$\lambda_l e_l(x) = \int k(x, x') e_l(x') p(x') \, \mathrm{d}x'$$

Definition 2.2.5. (Eigen-expansion) The eigen-expansion of $k(x, x')$ given eigenfunction e_l and eigenvalue λ_l for $l = 1, 2\dots$ is (it is countable):

$$k(x, x') = \sum_{l=1}^{\infty} \lambda_l(x) e_l(x) e_l(x')$$

where we can show that

$$\int e_i(x) e_j(x) p(x) \, \mathrm{d}x = \begin{cases} 0 & i \neq j \\ 1 & \text{otherwise} \end{cases}$$

Proposition 2.2.3. *The $L_2(p)$ inner product of function $f(x) = \sum_{l=1}^{\infty} \hat{f}_l e_l(x)$ and $g(x) = \sum_{l=1}^{\infty} \hat{f}_m e_m(x)$ is*

$$\langle f, g\rangle_{L_2} = \sum_{l=1}^{\infty} \hat{f}_l \hat{g}_l$$

Proof. We perform similar calculation as fourier series case:

$$\begin{aligned} \langle f, g\rangle_{L_2} &= \int_{-\infty}^{\infty} f(x) g(x) p(x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left[\sum_{l=1}^{\infty} \hat{f}_l e_l(x)\right] \left[\sum_{m=1}^{\infty} \hat{f}_m e_m(x)\right] p(x) \, \mathrm{d}x = \sum_{l=1}^{\infty} \hat{f}_l \hat{g}_l \end{aligned}$$

□

Definition 2.2.6. (Smooth Dot Product 2) We define a smooth dot product (with the norm) to be:

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}\hat{g}}{\lambda_l} \qquad \|f\|^2_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}_l^2}{\lambda_l}$$

Proposition 2.2.4. *We can show that*

$$\langle f(\cdot), k(\cdot, z)\rangle = f(z)$$

Proof. We have

$$\langle f(\cdot), k(\cdot, z)\rangle = \sum_{l=1}^{\infty} \frac{\hat{f}_l \lambda_l e_l(z)}{\lambda_l} = \sum_{l=1}^{\infty} \hat{f}_l e_l(z) = f(z)$$

□

Remark 32. Let's try to find the original definition of function evaluation as in definition 2.1.8. Since we have:

$$\langle f(\cdot), k(\cdot, z)\rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}_l(\lambda_l e_l(z))}{\left(\sqrt{\lambda_l}\right)^2}$$

and so we have $f_l = \hat{f}_l/\sqrt{\lambda_l}$ and $\phi_l(z) = \sqrt{\lambda_l}e_l(z)$, and so we have

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \left[\sum_{j=1}^{\infty} \lambda_j e_i(x_i) e_j(x)\right] = \sum_{l=1}^{\infty} f_l \left[\sqrt{\lambda_j} e_l(x)\right]$$

where $f_l = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_l} e_l(x_l)$. As λ_l decays as e_l becomes rougher, then f_l decays since $\|f\|^2_{\mathcal{H}} < \infty$. This reinforce smoothness.

2.3 More of RKHS

Definition 2.3.1. (Reproducing Kernel Hilbert Space) Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued function on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is reproducing kernel of $\mathcal{H}$ and $\mathcal{H}$ is RKHS if:

- For all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$, then $k(\cdot, x) \in \mathcal{H}$
- For all $x \in \mathcal{X}$, $\langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$

Definition 2.3.2. (Eval Operators) For all $f \in \mathcal{H}, x \in \mathcal{X}$ then we have $\delta_x f = f(x)$

Theorem 2.3.1. ***(Riesz Representation)*** *In Hilbert space $\mathcal{H}$, all bounded linear function f is of form $\langle \cdot, g\rangle_{\mathcal{H}}$ for some $g \in \mathcal{H}$.*

Theorem 2.3.2. *$\mathcal{H}$ is RKHS (δ_x is bounded and linear) iff $\mathcal{H}$ has a reproducing kernel.*

Proof. **(If $\mathcal{H}$ has reproducing kernel, then δ_x is bounded):** Starting with the first direction, we have:

$$\begin{aligned} |\delta_x f| = |f(x)| &= |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| \\ &\le \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \langle k(\cdot, x), k(\cdot, x)\rangle^{1/2}_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \sqrt{k(x, x)}\, \|f\|_{\mathcal{H}} \end{aligned}$$

(If δ_x is bounded, then $\mathcal{H}$ has reproducing kernel): We will utlize riesz representation. As the evaluation operator is bounded and linear, then there exists $f_{\delta_x} \in \mathcal{H}$ such that for all $f \in \mathcal{H}$, we have:

$$\delta_x f = \langle f, f_{\delta_x}\rangle_{\mathcal{H}}$$

We can define $k(\cdot, x) = f_{\delta_x}(\cdot)$ for all $x \in \mathcal{X}$. It is clear that k is reproducing kernel. □

Definition 2.3.3. (Alternative Definition RKHS) $\mathcal{H}$ is an RKHS if the evaluation operator is bounded i.e for all $x \in \mathcal{X}$ there exists $\lambda_x \geq 0$ such that for all $f \in \mathcal{H}$:

$$|f(x)| = |\delta_x| \leq \lambda_x \|f\|_{\mathcal{H}}$$

Remark 33. This definition implies that 2 functions that are identical in RKHS will agree at every point, for all $f, g \in \mathcal{H}$:

$$|f(x) - g(x)| = |\delta_x(f - g)| \leq \lambda_x \|f - g\|_{\mathcal{H}}$$

Theorem 2.3.3. ***(Moore-Aronszajn)*** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive define, then there is unique RKHS $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel k*

2.4 Application of Kernel

Proposition 2.4.1. *Given the sample $(x_i)_{i=1}^m$ from p and $(y_i)_{i=1}^m$ from q. The distance between their mean in a feature space is:*

$$\left\| \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi(y_i) \right\|_{\mathcal{H}}^2 = \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i, y_i)$$

Proof. Let's just expand the definition:

$$\begin{aligned}
&\left\| \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi(y_i) \right\|_{\mathcal{H}}^2 \\
=&\left\langle \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi(y_i), \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi(y_i) \right\rangle \\
=&\frac{1}{m^2}\left\langle \sum_{i=1}^m \phi(x_i), \sum_{i=1}^m \phi(x_i) \right\rangle - \frac{2}{mn}\left\langle \sum_{i=1}^m \phi(x_i), \frac{1}{n}\sum_{i=1}^n \phi(y_i) \right\rangle + \frac{1}{n^2}\left\langle \frac{1}{n}\sum_{i=1}^n \phi(y_i), \frac{1}{n}\sum_{i=1}^n \phi(y_i) \right\rangle \\
=&\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i, y_i)
\end{aligned}$$

□

Remark 34. When we can have $\phi(x) = x$, we distinguish a mean and when we use $\phi(x) = [x, x^2]$, we can distinguish the mean and variance. There is a possibility that we can use kernel to distinguish for 2 distribution. *Please note that, we don't have to explicitly calculate the feature.*

2.4.1 Kernel PCA

Definition 2.4.1. (Centering Matrix) The centering matrix H is defined as

$$I - n^{-1}\mathbf{1}_{n\times n}$$

Definition 2.4.2. (Principle Component Analysis) PCA is a method of finding d-dimensional subspace of a higher dimensional space D that contains the direction in the highest variance. Consider the first principle component:

$$u_1 = \underset{\|u\|\leq 1}{\arg\max}\ \frac{1}{n}\left(u^T\left(x_i - \frac{1}{n}\sum_{j=1}^n x_j\right)\right)^2 = \underset{\|u\|\leq 1}{\arg\max}\ u^T C u$$

where matrix C is defined by:

$$C = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^T = \frac{1}{n}XHX^T$$

where $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$ and H is a centering matrix. To see the expansion please go to appendix A.1.1.

Definition 2.4.3. (Tensor Product) We define tensor product as:

$$(a \otimes b)c = \langle b, c\rangle_{\mathcal{H}}\, a$$

This is analogous to the matrix notation $(ab^T)c = b^T ca$

Definition 2.4.4. (Kernelized Version of PCA) Let's consider the PCA model but with a feature map, starting from the first component:

$$\begin{aligned} f_1 &= \arg\max_{\|f\|_{\mathcal{H}}\le 1} \frac{1}{n}\sum_{i=1}^{n}\left(\left\langle f, \phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_i)\right\rangle\right)^2 \\ &= \arg\max_{\|f\|_{\mathcal{H}}\le 1} \frac{1}{n}\sum_{i=1}^{n}\Big(f(x_i) - \hat{\mathbb{E}}[f]\Big)^2 = \arg\max_{\|f\|_{\mathcal{H}}\le 1} \operatorname{var}(f) \end{aligned}$$

Note that the second equality comes from reproducing property of kernel. We will consider the infinite dimension analogous of covariance:

$$\begin{aligned} C &= \frac{1}{n}\sum_{i=1}^{n}\left(\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right) \otimes \left(\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\otimes\tilde{\phi}(x_i) \end{aligned}$$

Remark 35. We can consider the function:

$$f = \sum_{i=1}^{n}\alpha_i\left(\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right) = \sum_{i=1}^{n}\alpha_i\tilde{\phi}(x_i)$$

Suppose f is constructed as a sum of $f_\parallel + f_\perp$ where $f_\parallel$ is function component that parallels to the $\tilde{\phi}(x_i)$, and $f_\perp$ is function perpendicular to $\tilde{\phi}(x_i)$. However, as we perform inner product, the component $f_\perp$ is gone. Thus, we can write it, in the case of a linear combination.

Proposition 2.4.2. *The matrix equation of kernel PCA is*

$$n\lambda_l\alpha_l = \tilde{K}\alpha_l$$

where $\tilde{K} = HKH$ as H is centering matrix.

Proof. We will start by consider the application of applying C to f:

$$\begin{aligned} Cf &= \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\otimes\tilde{\phi}(x_i)\right)\sum_{j=1}^{n}\alpha_j\tilde{\phi}(x_j) \\ &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\left\langle\tilde{\phi}(x_i), \sum_{j=1}^{n}\alpha_j\tilde{\phi}(x_j)\right\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\left(\sum_{j=1}^{n}\alpha_j\tilde{k}(x_i, x_j)\right) \end{aligned}$$

as $\tilde{k}(x_i, x_j)$ is i, j-entry of the matrix $\tilde{K} = HKH$. To show this please go to appendix A.1.2. Now, we consider the eigenfunction and eigenvalue equation $\lambda_l f_l = Cf_l$, where we will project both side with $\tilde{\phi}(x_q)$:

- Left Hand Side:

$$\left\langle \tilde{\phi}(x_q), f_l \lambda_l \right\rangle = \lambda_l \left\langle \tilde{\phi}(x_q), f_l \right\rangle_{\mathcal{H}} = \lambda_l \sum_{i=1}^{n} \alpha_{li} \tilde{k}(x_q, x_i)$$

- Right Hand Side:

$$\left\langle \tilde{\phi}(x_q), Cf_l \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \tilde{k}(x_q, x_i) \left(\sum_{j=1}^{n} \alpha_{li} \tilde{k}(x_q, x_i) \right)$$

These equation leads to matrix equation $n\lambda_l \tilde{K} \alpha_l = \tilde{K}^2 \alpha_l$, by rearrangement, we get the statement. □

Proposition 2.4.3. *The norm of the function f is equal to*

$$\|f\|_{\mathcal{H}} = n\lambda \|\alpha\|^2$$

Proof. We have the following:

$$\begin{aligned}
\|f\|_{\mathcal{H}} &= \langle f, f \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i), \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i) \right\rangle \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \tilde{k}(x_i, x_j) \\
&= \alpha^T \tilde{K} \alpha = \alpha^T n\lambda \alpha = n\lambda \|\alpha\|^2
\end{aligned}$$

□

Remark 36. Given the norm of the function, we have to set $\alpha \leftarrow \alpha / \sqrt{n\lambda}$ assuming that $\|\alpha\| = 1$.

Proposition 2.4.4. *The projection of a test vector x^* to principle component f is*

$$P_f \phi(x^*) = \langle \phi(x^*), f \rangle f = \left(\sum_{i=1}^{n} \alpha_i \left(k(x^*, x_i) - \frac{1}{n} \sum_{j=1}^{n} k(x^*, x_j) \right) \right) \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i)$$

Proof. We start by expanding the definiton of f and $\tilde{f}$:

$$\begin{aligned}
P_f \phi(x^*) = \langle \phi(x^*), f \rangle f &= \left\langle \phi(x^*), \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i) \right\rangle f \\
&= \sum_{i=1}^{n} \alpha_i \left\langle \phi(x^*), \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right\rangle f \\
&= \left(\sum_{i=1}^{n} \alpha_i \langle \phi(x^*), \phi(x_i) \rangle - \frac{1}{n} \sum_{j=1}^{n} \langle \phi(x^*), \phi(x_j) \rangle \right) f \\
&= \left(\sum_{i=1}^{n} \alpha_i \left(k(x^*, x_i) - \frac{1}{n} \sum_{j=1}^{n} k(x^*, x_j) \right) \right) f
\end{aligned}$$

□

Remark 37. We can consider the application of denoising a hand-written digit. Suppose, we are given a noisy digit x^*:

$$P_d\phi(x^*) = P_{f_1}\phi(x^*) + \cdots + P_{f_d}\phi(x^*)$$

as we can project onto the first d eigenvectors $\{f_l\}_{i=1}^d$ from kernel PCA. The nearby point $y^* \in \mathcal{X}$ as:

$$y^* = \arg\min_{y\in\mathcal{X}} \ \|\phi(y) - P_d\phi(x^*)\|_{\mathcal{H}}^2$$

This is how the image can be denoised, which can be done without the access to feature map.

2.4.2 Kernel Ridge Regression

Definition 2.4.5. (Ridge Regression) Given n training points (in $\mathbb{R}^D$) and labels:

$$X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{D\times n} \qquad y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T$$

We define $\lambda > 0$, and our goal is to find a^*:

$$a^* = \arg\min_{a\in\mathbb{R}^D} \ \left(\left\|y - X^Ta\right\|^2 + \lambda\left\|a\right\|^2\right)$$

Theorem 2.4.1. *We can show that for ridge regression:*

$$a^* = (XX^T + \lambda I)^{-1}Xy$$

Proof. Instead of proving using derivative, we will consider an alternative; that is because when dealing with infinite dimension, derivative is troublesome. Starting expanding the terms:

$$\begin{aligned}
\left\|y - X^Ta\right\|^2 + \lambda\left\|a\right\| &= y^Ty - 2y^TX^Ta + a^TX^TXa + \lambda a^Ta \\
&= y^Ty - 2y^TX^Ta + a^T(XX^T - \lambda I)a \\
&= y^Ty - 2y^TX^T(XX^T + \lambda I)^{-1/2}b + b^Tb \\
&= y^Ty + \left\|(XX^T + \lambda I)^{-1/2}Xy - b\right\|^2 - \left\|y^TX^T(XX^T + \lambda I)^{-1/2}\right\|^2
\end{aligned}$$

where we define $b = (XX^T + \lambda I)^{1/2}a$. To see the expansion, we have appendix A.1.3. Note that matrix b is semi-positive definite, therefore the square is defined. Further, XX^T may not be invertible of $D > n$ but by adding λI will have full rank. To minimize the objective, we have to get:

$$b^* = (XX^T + \lambda I)^{-1/2}Xy \implies a^* = (XX^T + \lambda I)^{-1}Xy$$

□

Definition 2.4.6. Singular Value Decomposition (SVD) We assume $D > n$, and we perform SVD on X such that $X = USV^T$, where:

$$U = \begin{bmatrix} u_1 & \cdots & u_D \end{bmatrix} \qquad S = \begin{bmatrix} \tilde{S} & 0 \\ 0 & 0 \end{bmatrix} \qquad V = \begin{bmatrix} \tilde{V} & 0 \end{bmatrix}$$

where we have:

- U is $D \times D$ matrix where $U^TU = UU^T = I_D$
- S is $D \times D$ where $\tilde{S}$ has n non-zero entry
- V is $n \times D$ where $\tilde{V}^T\tilde{V} = \tilde{V}\tilde{V}^T = I_n$

Theorem 2.4.2. *We can write the solution in a^* by a linear combination of training points:*

$$a^* = \sum_{i=1}^{n} \alpha_i^* x_i$$

where $\alpha_i = \sum_{j=1}^{n} y_j \beta_{ij}$ as β_{ij} is (i,j)-entry of $(X^T X + \lambda In)$

Proof. We start by defining a SVD of $X = USV^T$, then we have:

$$\begin{aligned} a^* &= (XX^T + \lambda I_D)^{-1} Xy = (US^2U^T + \lambda I_D)^{-1} USV^T y \\ &= U(S^2 + \lambda I_D)^{-1} U^T USV^T y \\ &= US(S^2 + \lambda I_D)^{-1} V^T y \\ &= USV^T V (S^2 + \lambda I_D)^{-1} V^T y \\ &= XV(S^2 + \lambda I_D)^{-1} V^T y \\ &= X(X^T X + \lambda I_n)^{-1} y \end{aligned}$$

For the last equality, we have $V(S^2 + \lambda I_D)^{-1} V^T$, and so:

$$\begin{aligned} V(S^2 + \lambda I_D)^{-1} V^T &= \begin{bmatrix} \tilde{V} & 0 \end{bmatrix} \begin{bmatrix} (\tilde{S}^2 + \lambda I_n)^{-1} & 0 \\ 0 & (\lambda I_{D-n}) \end{bmatrix} \begin{bmatrix} \tilde{V}^T \\ 0 \end{bmatrix} \\ &= \tilde{V}(\tilde{S}^2 + \lambda I_n)^{-1} \tilde{V}^T = \tilde{V}(\tilde{S}^2 + \lambda I_n)^{-1} \tilde{V}^{-1} \\ &= \tilde{V}(\tilde{V}(\tilde{S}^2 + \lambda I_n))^{-1} \\ &= (\tilde{V}^T)^{-1}(\tilde{V}(\tilde{S}^2 + \lambda I_n))^{-1} \\ &= (\tilde{V}(\tilde{S}^2 + \lambda I_n)\tilde{V}^T)^{-1} \\ &= (\tilde{V}\tilde{S}^2\tilde{V}^T + \lambda I_n \tilde{V}\tilde{V}^T)^{-1} \\ &= (VS^T V + \lambda I_n)^{-1} = (VSU^T USV^T + \lambda I_n)^{-1} \\ &= (X^T X + \lambda I_n)^{-1} \end{aligned}$$

For the α_i value, we have A.1.4 i.e:

$$X(X^T X + \lambda I_n)^{-1} y = X \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nn} \end{bmatrix} y = \begin{bmatrix} \sum_{i=1}^{n} x_{1i} \sum_{j=1}^{n} y_j \beta_{ij} \\ \sum_{i=1}^{n} x_{2i} \sum_{j=1}^{n} y_j \beta_{ij} \\ \vdots \\ \sum_{i=1}^{n} x_{ni} \sum_{j=1}^{n} y_j \beta_{ij} \end{bmatrix} = \sum_{i=1}^{n} \underbrace{\left(\sum_{j=1}^{n} y_j \beta_{ij} \right)}_{\alpha_i} x_i$$

□

Definition 2.4.7. (Kernel Ridge Regression) We consider the following optimization problem:

$$a^* = \underset{a \in \mathcal{H}}{\arg\min} \left(\sum_{i=1}^{n} (y_i - \langle a, \phi(x_i) \rangle)^2 + \lambda \|a\|_{\mathcal{H}}^2 \right)$$

Corollary 2.4.1. *The kernel ridge regression solution a^* is*

$$a^* = X(K + \lambda I_n)^{-1} y = \sum_{i=1}^{n} \alpha^* \phi(x_i)$$

where K is the gram matrix.

Proof. We can consider a ridge regression with the data matrix:

$$X = \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_n) \end{bmatrix}$$

Please note that $(X^T X)_{ij} = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j)$, given the result in theorem 2.4.2. Or, $X^T X = K$ □

Remark 38. We have the following, in tensor product:

$$XX^T = \sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i)$$

Remark 39. We can see that the smoothness property of RKHS

$$\|f\|_{\mathcal{H}}^2 = \sum_{l=1}^{\infty} \frac{\hat{f}_l^2}{\lambda_l} \qquad \|f\|_{\mathcal{H}}^2 = \sum_{l=1}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}$$

on the left hand side, we eigenvalues based norm and the right hand side is the fourier based norm.

2.5 Maximum Mean Discrepancy

2.5.1 Mean Embedding

Definition 2.5.1. (Feature Map of Probability P) Given P a Borel probability measure on $\mathcal{X}$, define a feature map of probability P to be:

$$\mu_P = \begin{bmatrix} \cdots & \mathbb{E}_P[\phi_i(x)] & \cdots \end{bmatrix}$$

Definition 2.5.2. (Kernel of Probability) For positive definite $k(x, x')$ where:

$$\langle \mu_P, \mu_Q \rangle = \mathbb{E}_{P,Q}[k(x, y)]$$

where $x \sim P$ and $y \sim Q$. We can consider the expectation in an RKHS as $\mathbb{E}_P[f(x)] = \langle f, \mu_P \rangle_{\mathcal{F}}$. And, so μ_P gives the expectation of all RKHS functions.

Remark 40. We can see that the empirical mean embedding is

$$\hat{\mu}_P = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \qquad \text{where} \qquad x_i \sim P$$

Theorem 2.5.1. ***(Existance of Mean Embedding)*** *The element $\mu_P \in \mathcal{F}$ exist, such that*

$$\mathbb{E}_P[f(x)] = \langle f, \mu_P \rangle_{\mathcal{F}}$$

for all $f \in \mathcal{F}$ if $\mathbb{E}_P[\sqrt{k(x,x)}] = \mathbb{E}_P \|\psi(x)\|_{\mathcal{F}} < \infty$

Proof. We will consider the application of Riesz theorem by assuming a linear operator $T_P f = \mathbb{E}_P[f(x)]$ for all $f \in \mathcal{F}$. We will show that this operator is bounded:

$$\begin{aligned} |T_P f| &= |\mathbb{E}_P[f(x)]| \\ &\le \mathbb{E}_P[|f(x)|] \\ &= \mathbb{E}_P[| \langle f, \phi(x) \rangle_{\mathcal{F}} |] \\ &\le \mathbb{E}_P[\|f\|_{\mathcal{F}} \|\phi(x)\|_{\mathcal{F}}] \\ &= \mathbb{E}_P[\sqrt{k(x,x)}] \|f\|_{\mathcal{F}} \end{aligned}$$

By Riesz theorem, since the operator is bounded, then there exists $\mu_P \in \mathcal{F}$ that $T_P f = \langle f, \mu_P \rangle_{\mathcal{F}}$ □

Remark 41. The probability feature map looks like the following:

$$\mu_P(t) = \langle \mu_P, \phi(t) \rangle_{\mathcal{F}} = \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}} = \mathbb{E}_P[k(x, t)]$$

2.5.2 Algorithm

Definition 2.5.3. (Maximum Mean Discrepancy) Maximum Mean Discrepancy (MMD) is the distance between probability feature mean:

$$\text{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{F}}^2$$

Lemma 2.5.1. *We can show that MMD is equal to*

$$\text{MMD}^2(P, Q) = \mathbb{E}_P[k(x, x')] + \mathbb{E}_Q[k(y, y')] - 2\mathbb{E}_{P,Q}[k(x, y)]$$

Proof.

$$\begin{aligned}\|\mu_P - \mu_Q\|_{\mathcal{F}}^2 &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\ &= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} \\ &= \mathbb{E}_P[\mu_P(x)] + \mathbb{E}_Q[\mu_Q(y)] - 2\mathbb{E}_P[\mu_Q(x)] \\ &= \mathbb{E}_P[\mathbb{E}_P[k(x, x')]] + \mathbb{E}_Q[\mathbb{E}_Q[k(y, y')]] + 2\mathbb{E}_P[\mathbb{E}_Q[k(x, y)]]\end{aligned}$$

□

Definition 2.5.4. (Empirical Mean MMD) We have the following unbiased empirical mean MMD:

$$\text{MMD}^2(P, Q) = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{1}{n^2} \sum_{i,j} k(x_i, y_j)$$

Definition 2.5.5. (Integral Probability Metrics) Integral Probability Metrics is divergence measure, which has the form:

$$\sup_{g \in \mathcal{H}} \Big(\mathbb{E}_{x \sim P}[g(x)] - \mathbb{E}_{y \sim Q}[g(y)] \Big)$$

The examples of Integral Probability Metrics are MMD and Wassertein.

Definition 2.5.6. (F-Divergence) F-divergence is divergence measure, which has the form:

$$D_f(P, Q) = \int_{\mathcal{H}} q(x) f\left(\frac{p(x)}{q(x)}\right) \, \mathrm{d}x$$

The example of F-divergence are KL-divergence, Hellinger, and Pearson-Chi Square.

Remark 42. Total Variation can be shown to be both Integral Probability Metrics and F-Divergence. For instance:

$$\text{TV}(p, q) = \sup_{A \in \mathcal{F}} |p(A) - q(A)| = \frac{1}{2} \int \left| \frac{p(x)}{q(x)} - 1 \right| q(x) \, \mathrm{d}x$$

Theorem 2.5.2. *We can show that* MMD *can be represented by:*

$$\text{MMD}(P, Q) = \sup_{\|f\| \leq 1} [\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]]$$

Note that f is unit ball of $\mathcal{F}$.

Proof.

$$\begin{aligned}\sup_{\|f\| \leq 1} [\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]] &= \sup_{\|f\| \leq 1} \langle f, \mu_P \rangle - \langle f, \mu_Q \rangle \\ &= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle\end{aligned}$$

To maximize the dot product, we need f should be in the same direction as $\mu_P - \mu_Q$. Therefore, we set

$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

Thus, the dot product to this is:

$$\sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle = \|\mu_P - \mu_Q\|$$

□

Remark 43. The reason we need a constrain $\|f\| \leq 1$ because the function has to be smooth as too "sharp" will lead to perfect seperation i.e maximizing the MMD.

Corollary 2.5.1. ***(Empirical Witness Function)*** *The empirical witness function is:*

$$f^*(v) = \frac{1}{n}\sum_{i=1}^{n} k(x_i, v) - \frac{1}{n}\sum_{i=1}^{n} k(y_i, v)$$

Proof. Since $f \propto \mu_P - \mu_Q$, and the empirical mean embedding shown in remark 40. we have the following:

$$\begin{aligned} f(v) &\propto \langle \hat{\mu}_P - \hat{\mu}_Q, \phi(v) \rangle \\ &= \left\langle \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) - \frac{1}{n}\sum_{i=1}^{n} \phi(y_i), \phi(v) \right\rangle = \frac{1}{n}\sum_{i=1}^{n} k(x_i, v) - k(y_i, v) \end{aligned}$$

□

2.5.3 Statistical Testing of MMD

Theorem 2.5.3. *When $P \neq Q$, the statistics of empirical MMD is asympototic normal:*

$$\frac{\widehat{\text{MMD}}^2 - \text{MMD}(P,Q)^2}{\sqrt{V_n(P,Q)}} \xrightarrow{D} \mathcal{N}(0,1)$$

where the variance $V_n(P,Q) = \mathcal{O}(n^{-1})$ but affected by kernel. However, when $P = Q$, the statistics has asympototic distribution of:

$$n\widehat{\text{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l [z_l^2 - 2] \qquad \text{where} \qquad \lambda_i \phi_i(x) = \int_{\mathcal{X}} \tilde{k}(x, \tilde{x}) \phi_i(x) \ \mathrm{d}P(x)$$

and $z_l \sim \mathcal{N}(0,2)$

Remark 44. In the perspective of statistical hypothesis testing, we want to find a threshold c_α for which $\widehat{\text{MMD}}^2$ has false positive α. To estimate the c_α, we consider estimating the null-hypothesis $P = Q$, by permuting the items, so that they are uncorrelated.

Definition 2.5.7. (MMD Test Power) Test power is defined as

$$\text{Pr}_1\left(n\widehat{\text{MMD}} > \hat{c}_\alpha\right) \to \Phi\left(\frac{\text{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

where Pr_1 is the probability that $P \neq Q$, and Φ is cumulative distribution function of standard normal distribution.

Remark 45. To find the best kernel, we can find the kernel that minimizes the false negative rate, by maximize the test power. We would like to note the following:

$$\frac{\text{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} = \mathcal{O}(\sqrt{n}) \qquad \frac{c_\alpha}{n\sqrt{V_n(P,Q)}} = \mathcal{O}(n^{-1/2})$$

Then, for a large n, the second term won't matter, and so we can just maximize:

$$\frac{\text{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

which we can use neural network to perform gradient descent on this objective.

2.5.4 Characteristic RKHS

Definition 2.5.8. (Characteristic RKHS) A characteristic RKHS, where $\text{MMD}(P, Q; \mathcal{F}) = 0$ iff $P = Q$

Theorem 2.5.4. *The MMD metrics can be written as, for periodic kernel:*

$$\text{MMD}^2(P, Q; \mathcal{F}) = \sum_{l=-\infty}^{\infty} |\phi_{P,l} - \phi_{Q,l}|^2 \hat{k}_l$$

where $\phi_{P,l}, \phi_{Q,l}$ are fourier coefficient of the probability distributions, while $\hat{k}_l$ is the fourier coefficient of the kernel.

Proof. Let's consider the fourier coefficient of μ_P:

$$\mu_P(x) = \langle \mu_P, k(x, \cdot) \rangle_{\mathcal{F}} = \mathbb{E}_{t \sim P}[k(x-t)] = \int_{-\pi}^{\pi} k(x-t) \ \text{d}P(t)$$

Now, we have

$$\begin{aligned}
\int_{-\pi}^{\pi} k(t-x)P(t) \ \text{d}t &= \int_{-\pi}^{\pi} \left[\sum_{l=-\infty}^{\infty} \hat{k}_l \exp(il(x-t)) \right] \left[\sum_{l=-\infty}^{\infty} \hat{\phi}_{P,l} \exp(ilt) \right] \ \text{d}t \\
&= \int_{-\pi}^{\pi} \left[\sum_{l=-\infty}^{\infty} \hat{k}_l \overline{\exp(ilt)} \exp(ilx) \right] \left[\sum_{l=-\infty}^{\infty} \hat{\phi}_{P,l} \exp(ilt) \right] \ \text{d}t \\
&= \int_{-\pi}^{\pi} \left[\sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{k}_l \overline{\exp(ilt)} \exp(ilx) \hat{\phi}_{P,m} \exp(imt) \right] \ \text{d}t \\
&= \int_{-\pi}^{\pi} \sum_{l=-\infty}^{\infty} \left[\sum_{m \ne l} \hat{k}_l \hat{\phi}_{P,m} \overline{\exp(ilt)} \exp(ilx) \exp(imt) \right] + \left[\sum_{m=l} \hat{k}_l \hat{\phi}_{P,m} \exp(imx) \right] \ \text{d}t \\
&= \sum_{l=-\infty}^{\infty} \int_{-\pi}^{\pi} \left[\sum_{m \ne l} \hat{k}_l \hat{\phi}_{P,m} \overline{\exp(ilt)} \exp(-ilx) \exp(-imt) \right] \ \text{d}t + \sum_{m=l} \hat{k}_l \hat{\phi}_{P,m} \exp(imx) \\
&= \sum_{l=-\infty}^{\infty} \hat{k}_l \hat{\phi}_{P,l} \exp(ilx)
\end{aligned}$$

Thus the fourier coefficient of μ_P is $\hat{\mu}_{P,l} = \hat{k}_l \cdot \hat{\phi}_{P,l}$. This is related to convolution theorem as the convolution in normal domain (time) is equivalent to multiplcation in fourier transformed domain (frequency). We can see that the MMD can be written as:

$$\begin{aligned}
\text{MMD}^2(P, Q; \mathcal{F}) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\
&= \left\| \sum_{l=-\infty}^{\infty} \left(\hat{\phi}_{P,l} - \hat{\phi}_{Q,l} \right) \hat{k}_l \exp(ilx) \right\|_{\mathcal{F}}^2 \\
&= \sum_{l=-\infty}^{\infty} \frac{|\hat{\phi}_{P,l} - \hat{\phi}_{Q,l}|^2 \hat{k}_l^2}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} |\hat{\phi}_{P,l} - \hat{\phi}_{Q,l}|^2 \hat{k}_l
\end{aligned}$$

Recalling the square norm for function f in $\mathcal{F}$ defined in 2.2.3. □

Corollary 2.5.2. *The kernel is characteristic iff none of $\hat{k}_l$ is equal to zero.*

Proof. Suppose the kernel at l' is zero i.e $\hat{k}_{l'} = 0$, then we can find 2 difference distributions P and Q such that its fourier coefficients are equal $\hat{\phi}_{P,l} = \hat{\phi}_{Q,l}$ where $l \neq l'$. Then the MMD will be zero, i.e:

$$\text{MMD}^2(P, Q; \mathcal{F}) = 0$$

where $P \neq Q$ and the kernel isn't characteristic. □

Theorem 2.5.5. ***(Bochner's Theorem)*** *For a translation invariance kernel $k(x-y)$, we have*

$$k(x-y) = \int_{\mathbb{R}^d} \exp(-i(x-y)^T\omega)\ \mathrm{d}\Lambda(\omega)$$

Where the characteristic function of P is equality to

$$\phi_P(\omega) = \int_{\mathbb{R}^d} \exp(ix^T\omega)\ \mathrm{d}P(x)$$

Definition 2.5.9. (Measure Theoretic Integration) We define the following integration, for probability measure P, Q:

$$\int f(s)\ \mathrm{d}(P-Q)(s) = \mathbb{E}_P[f(s)] - \mathbb{E}_Q[f(s)]$$

Theorem 2.5.6. *The Fourier representation MMD for $\mathbb{R}^d$:*

$$\mathrm{MMD}^2(P,Q;\mathcal{F}) = \int |\phi_P(\omega) - \phi_Q(\omega)|^2\ \mathrm{d}\Lambda(\omega)$$

where $\Lambda(w)$ is finite non-negative Borel measure, for translation invariance kernel.

Proof. We have:

$$\begin{aligned}
\mathrm{MMD}^2(P,Q;\mathcal{F}) &= \mathbb{E}_P[k(x,x')] + \mathbb{E}_Q[k(y,y')] - 2\mathbb{E}_{P,Q}[k(x,y)] \\
&= \int \left[\int k(s-t)\ \mathrm{d}(P-Q)(s)\right]\ \mathrm{d}(P-Q)(t) \\
&= \int \left[\iint_{\mathbb{R}^d} \exp(-i(s-t)^T\omega)\ \mathrm{d}\Lambda(\omega)\ \mathrm{d}(P-Q)(t)\right]\ \mathrm{d}(P-Q)(t) \\
&= \int \left[\int_{\mathbb{R}^d} \exp(-is^T\omega)\ \mathrm{d}(P-Q)(s)\right]\left[\int_{\mathbb{R}^d} \exp(it^T\omega)\ \mathrm{d}(P-Q)(t)\right]\ \mathrm{d}\Lambda(\omega) \\
&= \int |\phi_P(\omega) - \phi_Q(\omega)|^2\ \mathrm{d}\Lambda(\omega)
\end{aligned}$$

For the expansion of the first integration please see appendix A.1.5. □

Corollary 2.5.3. *A translation invariance k is characteristic for probability measure on $\mathbb{R}^d$ iff*

$$supp(\Lambda) = \mathbb{R}^d$$

as the support can be zero at most countable set. Furthermore, any continuous, compactly support k is characteristic.

Theorem 2.5.7. *Probability $P = Q$ iff*

$$\mathbb{E}_P[x] = \mathbb{E}_Q[x]$$

for all $f \in C(\mathcal{X})$, the space of bounded continuous function on $\mathcal{X}$.

Definition 2.5.10. (Universal RKHS) A universal RKHS is where $k(x,x')$ is continuous, $\mathcal{X}$ is compact and $\mathcal{F}$ is dense in $C(\mathcal{X})$ with respect to L_∞. This meanse that for any given $\varepsilon > 0$ and $f \in C(\mathcal{X})$, there exists $g \in \mathcal{F}$, such that

$$\|f-g\|_\infty \leq \varepsilon$$

Theorem 2.5.8. *If $\mathcal{F}$ is universal then* $\mathrm{MMD}(P,Q;\mathcal{F}) = 0$ *iff $P = Q$*

Proof. It is clear that if $P = Q$ then $\text{MMD}(P, Q; \mathcal{F}) = 0$. Now, for the converse, let's consider the following inequality:

$$\begin{aligned}
&\Big|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(y)]\Big| \\
&\quad \le \Big|\mathbb{E}_P[f(x)] - \mathbb{E}_P[g(x)]\Big| + \Big|\mathbb{E}_P[g(x)] - \mathbb{E}_Q[g(y)]\Big| + \Big|\mathbb{E}_Q[g(y)] - \mathbb{E}_Q[f(y)]\Big| \\
&\quad \le \Big|\mathbb{E}_P[f(x)] - \mathbb{E}_P[g(x)]\Big| + \Big|\mathbb{E}_Q[g(y)] - \mathbb{E}_Q[f(y)]\Big| \\
&\quad \le \mathbb{E}_P[|f(x) - g(x)|] + \mathbb{E}_Q[|g(y) - f(y)|] \le 2\varepsilon
\end{aligned}$$

For all $f \in C(\mathcal{X})$ and $\varepsilon > 0$, which implies $P = Q$. For the second inequality, we would like to note that (As MMD is equal to zero)

$$\Big|\mathbb{E}_P[g(x)] - \mathbb{E}_Q[g(y)]\Big| = \Big|\left\langle g, \mu_P - \mu_Q \right\rangle\Big| \le \|g\|_{\mathcal{F}} \|\mu_p - \mu_Q\|_{\mathcal{F}} = 0$$

□

2.6 Testing Dependencies

2.6.1 Covariance Operators

Remark 46. We might use MMD to measure the statistical dependence between 2 samples X and Y. However, we will have the following problem:

- We don't have $Q = P_X P_Y$ as we need to have a pair $\{(x_i, y_i)\}_{i=1}^n \sim P_{XY}$.
- What kernel to use for the pair ?

For the first problem, we can simular Q by drawing a pair (x_i, y_j). Also, for the second problem, we can use *product* kernel. But why product ? and is there are more interpretable definition of dependence measure ?

Definition 2.6.1. (Hilbert-Schmidt Operators) Given $\mathcal{F}$ and $\mathcal{G}$, which are seperatable Hilbert space. $(g_j)_{j \in J}$ is an orthogonal basis in $\mathcal{G}$, where J is an index set either finite or countable infinite and:

$$\langle g_i, g_j \rangle = \begin{cases} 0 & i \ne j \\ 1 & i = j \end{cases}$$

Given a linear operators $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$, then Hilbert-Schmidt operator is defined as:

$$\langle L, M \rangle_{\text{HS}} = \sum_{j \in J} \langle Lg_j, Mg_j \rangle_{\mathcal{F}}$$

Please note that the sum is finite if $\|L\|$ and $\|M\|$ are finite, which is by Cauchy-Schwarz. Similarly, we can define a norm to be:

$$\|L\|^2_{\text{HS}} = \sum_{j \in J} \|Lg_i\|^2_{\mathcal{F}}$$

If the norm of L is finite, then L is called Hilbert-Schmidt.

Lemma 2.6.1. *Given a matrix A and B both in $\mathbb{R}^{n \times n}$, then Hilbert-Schmidt inner product is (together with the basis vectors):*

$$\langle A, B \rangle = \sum_{j \in J} \langle Ag_j, Bg_j \rangle = \text{tr}(A^T B)$$

Remark 47. We can consider the covariance operator in finite dimension, which we have:

$$\left\langle C_{xy}, fg^T \right\rangle = \text{tr}(C_{xy}^T (fg^T)) = \text{tr}(g^T C_{xy}^T f) = f^T C_{xy} g = \mathbb{E}_{xy}[f(x)g(y)]$$

Lemma 2.6.2.

$$\|a \otimes b\|_{\text{HS}}^2 = \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{F}}^2$$

Proof.

$$\begin{aligned}\|a \otimes b\|_{\text{HS}}^2 &= \sum_{j \in J} \|(a \otimes b) g_j\|_{\mathcal{F}}^2 = \sum_{j \in J} \left\| \langle b, g_j \rangle_{\mathcal{F}} a \right\|_{\mathcal{F}}^2 \\ &= \sum_{j \in J} \left\langle \langle b, g_j \rangle_{\mathcal{F}} a, \langle b, g_j \rangle_{\mathcal{F}} a \right\rangle_{\mathcal{F}} = \sum_{j \in J} |\langle b, g_j \rangle_{\mathcal{F}}|^2 \langle a, a \rangle_{\mathcal{F}} \\ &= \|a\|_{\mathcal{F}} \sum_{j \in J} |\langle b, g_j \rangle_{\mathcal{F}}|^2 = \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{F}}^2\end{aligned}$$

The last equality is called Parseval's identity. □

Lemma 2.6.3.

$$\langle L, a \otimes b \rangle_{\text{HS}} = \langle a, Lb \rangle_{\mathcal{F}}$$

Proof. Consider the left hand side

$$\begin{aligned}\langle L, a \otimes b \rangle_{\text{HS}} &= \sum_{j \in J} \langle L g_j, (a \otimes b) g_j \rangle_{\mathcal{F}} \\ &= \sum_{j \in J} \left\langle L g_j, \langle b, g_j \rangle_{\mathcal{F}} a \right\rangle_{\mathcal{F}} = \sum_{j \in J} \langle b, g_j \rangle_{\mathcal{F}} \langle a, L g_j \rangle_{\mathcal{F}}\end{aligned}$$

We consider the right hand side

$$\langle a, Lb \rangle_{\mathcal{F}} = \left\langle a, \sum_{j \in J} L g_j \langle b, g_j \rangle_{\mathcal{F}} \right\rangle = \sum_{j \in J} \langle a, L g_j \rangle \langle b, g_j \rangle_{\mathcal{F}}$$

□

Corollary 2.6.1.

$$\langle u \otimes v, a \otimes b \rangle_{\text{HS}} = \langle u, a \rangle_{\mathcal{F}} \langle v, b \rangle_{\mathcal{F}}$$

Proof.

$$\langle u \otimes v, a \otimes b \rangle_{\text{HS}} = \langle a, (u \otimes v) b \rangle_{\mathcal{F}} = \langle a, \langle v, b \rangle_{\mathcal{F}} u \rangle_{\mathcal{F}} = \langle a, u \rangle_{\mathcal{F}} \langle v, b \rangle_{\mathcal{F}}$$

□

Theorem 2.6.1. *There exists $C_{xy} : \mathcal{G} \to \mathcal{F}$ in Hilbert space such that:*

$$\langle C_{xy}, A \rangle_{\text{HS}} = \mathbb{E}_{xy}[\langle \psi(x) \otimes \phi(y), A \rangle_{\text{HS}}]$$

if the kernels associated with ψ and ϕ, k_1 and k_2, respectively are bounded i.e $k_1(x,x) < \infty$ and $k_2(y,y) < \infty$

Proof. We consider Riesz representation thoerem, which we will have to show that $\mathbb{E}_{xy}[\langle \psi(x) \otimes \phi(y), A \rangle_{\text{HS}}]$ is bounded, which:

$$\begin{aligned}\Big| \mathbb{E}_{xy}[\langle \psi(x) \otimes \phi(y), A \rangle_{\text{HS}}] \Big| &\le \mathbb{E}_{xy} \left[\Big| \langle \psi(x) \otimes \phi(y), A \rangle_{\text{HS}} \Big| \right] \\ &\le \mathbb{E}_{xy} \left[\|\psi(x) \otimes \phi(y)\|_{\text{HS}} \cdot \|A\|_{\text{HS}} \right] \\ &= \mathbb{E}_{xy} \left[\|\psi(x) \otimes \phi(y)\|_{\text{HS}} \right] \|A\|_{\text{HS}}\end{aligned}$$

Now, we will show that $\mathbb{E}_{xy} \left[\|\psi(x) \otimes \phi(y)\|_{\text{HS}} \right] < \infty$ is bounded.

$$\begin{aligned}\mathbb{E}_{xy} \left[\|\psi(x) \otimes \phi(y)\|_{\text{HS}} \right] &= \mathbb{E}_{xy} \left[\|\psi(x)\|_{\mathcal{F}} \|\phi(y)\|_{\mathcal{F}} \right] \\ &= \mathbb{E}_x[\sqrt{k_1(x,x)}] \mathbb{E}_y[\sqrt{k_2(y,y)}] < \infty\end{aligned}$$

Thus the Riesz theorem's condition is satisfied. □

Corollary 2.6.2.

$$\langle C_{xy}, f \otimes g \rangle = \mathbb{E}_{xy}[f(x)g(y)]$$

Proof.

$$\begin{aligned}
\langle C_{xy}, f \otimes g \rangle &= \mathbb{E}_{xy}\left[\langle \psi(x) \otimes \phi(x), f \otimes g \rangle\right] \\
&= \mathbb{E}_{xy}\left[\langle \psi(x), f \rangle, \langle \phi(x), g \rangle\right] \\
&= \mathbb{E}_{xy}[f(x)g(y)]
\end{aligned}$$

□

Definition 2.6.2. (Covariance Operator) The covariance operators $C_{xy} : \mathcal{G} \to \mathcal{F}$ is an analogous to covariance matrix of infinite dimension, and it is defiend as:

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbb{E}_{xy}[f(x)g(y)]$$

Definition 2.6.3. (Empirical Covariance Operator) We define an empirical covariance operator as:

$$\hat{C}_{xy} = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i) \otimes \phi(y_i)$$

2.6.2 COCO

Definition 2.6.4. (Constrained Covariance) We have the following covariance problem

$$\begin{aligned}
\text{COCO}(P_{XY}) &= \sup_{\|f\|_{\mathcal{H}} \le 1 \ \|g\|_{\mathcal{H}} \le 1} \text{Cov}[f(x)g(y)] \\
&= \sup_{\|f\|_{\mathcal{H}} \le 1 \ \|g\|_{\mathcal{H}} \le 1} \left\langle f, \tilde{C}_{xy} g \right\rangle \\
&= \sup_{\|f\|_{\mathcal{H}} \le 1 \ \|g\|_{\mathcal{H}} \le 1} \mathbb{E}_{xy}\left[\left(\sum_{j=1}^{\infty} f_j \tilde{\psi}_j(x)\right)\left(\sum_{j=1}^{\infty} g_j \tilde{\phi}_j(x)\right)\right]
\end{aligned}$$

where $\tilde{\psi}(x) = \psi(x) - \mathbb{E}_x\psi(x)$ and $\tilde{\phi}(x) = \phi(x) - \mathbb{E}_x\phi(x)$ and $\tilde{C}$ being a covariance operator with centered feature map. We will use it to determine the dependence between variables. However, please note that covariance isn't the same as dependency.

Definition 2.6.5. (Empircal COCO) We define an empirical COCO problem to be

$$\widehat{\text{COCO}} = \sup_{\|f\|_{\mathcal{H}} \le 1 \ \|g\|_{\mathcal{H}} \le 1} \frac{1}{n}\sum_{i=1}^{n}\left[\left(f(x_i) - \frac{1}{n}\sum_{j=1}^{n} f(x_j)\right)\left(g(y_i) - \frac{1}{n}\sum_{j=1}^{n} g(y_j)\right)\right]$$

Given a sample $\{(x_i, y_i)\}_{i=1}^n$ sample iid from P_{xy}

Theorem 2.6.2. *The empirical $\widehat{\text{COCO}}$ is the largest eigenvalue γ_{max} i.e:*

$$\begin{bmatrix} 0 & \frac{1}{n}\tilde{K}\tilde{L} \\ \frac{1}{n}\tilde{K}\tilde{L} & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

where $\tilde{K} = HKH, \tilde{L} = HLH$ are center kernel matrix

Proof. We consider the following Lagragian:

$$\begin{aligned}
\mathcal{L}(f, g, \lambda, \gamma) = &-\frac{1}{n}\sum_{i=1}^{n}\left[\left(f(x_i) - \frac{1}{n}\sum_{j=1}^{n} f(x_j)\right)\left(g(y_i) - \frac{1}{n}\sum_{j=1}^{n} g(y_j)\right)\right] \\
&+ \frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) + \frac{\gamma}{2}\left(\|g\|_{\mathcal{F}}^2 - 1\right)
\end{aligned}$$

We assume that the function f and g are

$$f = \sum_{i=1}^{n} \alpha_i \tilde{\psi}(x_i) \qquad g = \sum_{i=1}^{n} \beta_i \tilde{\phi}(x_i)$$

Now, consider the smoothness constrain, which we have:

$$\begin{aligned}
\|f\|_{\mathcal{F}}^2 - 1 &= \langle f, f \rangle_{\mathcal{F}} - 1 \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \tilde{\psi}(x_i), \sum_{i=1}^{n} \alpha_i \tilde{\psi}(x_i) \right\rangle - 1 \\
&= \alpha^T \tilde{K} \alpha
\end{aligned}$$

For the covariance, we have

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \left[\left(f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right) \left(g(y_i) - \frac{1}{n} \sum_{j=1}^{n} g(y_j) \right) \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left\langle f, \tilde{\psi}(x_i) \right\rangle_{\mathcal{F}} \left\langle g, \tilde{\phi}(y_i) \right\rangle_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{n} \left\langle \sum_{i=1}^{n} \alpha_i \tilde{\psi}(x_i), \psi(\tilde{x}_i) \right\rangle \left\langle \sum_{i=1}^{n} \beta_i \tilde{\phi}(x_i), \tilde{\phi}(y_i) \right\rangle_{\mathcal{G}} \\
&\quad = \frac{1}{n} \alpha^T \tilde{K} \tilde{L} \beta
\end{aligned}$$

Now, we have the following Lagragian:

$$\mathcal{L}(f, g, \lambda, \gamma) = -\frac{1}{n} \alpha^T \tilde{K} \tilde{L} \beta + \frac{\lambda}{2} (\alpha^T \tilde{K} \alpha - 1) + \frac{\gamma}{2} (\beta^T \tilde{L} \beta - 1)$$

Now, we differentiate that Lagragian with respect to α and β, which we have (respectively) and set to zero:

$$0 = -\frac{1}{n} \tilde{K} \tilde{L} \beta + \lambda \tilde{K} \alpha \qquad 0 = -\frac{1}{n} \tilde{L} \tilde{K} \alpha + \gamma \tilde{L} \beta$$

By multiplying the first equation with α^T and the second one by β^T, we have:

$$0 = -\frac{1}{n} \alpha^T \tilde{K} \tilde{L} \beta + \lambda \alpha^T \tilde{K} \alpha \qquad 0 = -\frac{1}{n} \beta^T \tilde{L} \tilde{K} \alpha + \gamma \beta^T \tilde{L} \beta$$

Subtract both equation, yields:

$$\lambda \alpha^T \tilde{K} \alpha = \gamma \beta^T \tilde{L} \beta$$

when $\lambda \neq 0$ and $\gamma \neq 0$, by complementary slackness we have $\alpha^T \tilde{K} \alpha = \beta^T \tilde{L} \beta = 1$, thus $\lambda = \gamma$. And so, COCO is the largest eigenvalue. □

Definition 2.6.6. (Empirical witness Function) We define the empirical witness function as:

$$f(x) \propto \sum_{i=1}^{n} \alpha_i \left[k(x_i, x) - \frac{1}{n} \sum_{j=1}^{n} k(x_j, x) \right]$$

Remark 48. Even with indepdent variable, COCO won't give us zero at finite sample, since there can be some mild linear dependence found by f, g, which is a bias. Good news, this will be decrease if the sample size is higher.

2.6.3 HSIC

Definition 2.6.7. (Hilbert-Schmidt Indepdent Criterion) We would like to just find the norm of the centered covariance operator i.e

$$\text{HSIC}(P_{XY}; \mathcal{F}, \mathcal{G}) = \|C_{xy} - \mu_x \otimes \mu_y\|_{\text{HS}} = \left\|\tilde{C}_{xy}\right\|_{\text{HS}}$$

Theorem 2.6.3. *MMD with product kernel:*

$$\text{HSIC}^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \text{MMD}^2(P_{XY}, P_X, P_Y; \mathcal{H}_k)$$

where $k((x,y),(x',y')) = k(x,x')l(y,y')$

Proof.

$$\begin{aligned}
\|C_{xy} - \mu_x \otimes \mu_y\|^2_{\text{HS}} &= \langle C_{xy} - \mu_x \otimes \mu_y, C_{xy} - \mu_x \otimes \mu_y\rangle_{\text{HS}} \\
&= \underbrace{\langle C_{xy}, C_{xy}\rangle_{\text{HS}}}_{①} - 2\underbrace{\langle C_{xy}, \mu_x \otimes \mu_y\rangle_{\text{HS}}}_{②} + \underbrace{\langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y\rangle_{\text{HS}}}_{③}
\end{aligned}$$

Let's consider ①, first

$$\begin{aligned}
\langle C_{xy}, C_{xy}\rangle_{\text{HS}} &= \mathbb{E}_{xy}\left[\langle \psi(x) \otimes \phi(y), C_{xy}\rangle_{\text{HS}}\right] \\
&= \mathbb{E}_{xy}\mathbb{E}_{x'y'}\left[\langle \psi(x) \otimes \phi(y), \psi(x') \otimes \phi(y')\rangle_{\text{HS}}\right] \\
&= \mathbb{E}_{xy}\mathbb{E}_{x'y'}\left[\langle \psi(x), \psi(x')\rangle_{\mathcal{F}} \langle \phi(x), \phi(x')\rangle_{\mathcal{G}}\right] \\
&= \mathbb{E}_{xy}\mathbb{E}_{x'y'}\left[k(x,x')k(y,y')\right]
\end{aligned}$$

For the ②, we have:

$$\begin{aligned}
\langle C_{xy}, \mu_x \otimes \mu_y\rangle_{\text{HS}} &= \mathbb{E}_{xy}\left[\langle \psi(x) \otimes \phi(y), \mu_x \otimes \mu_y\rangle_{\text{HS}}\right] \\
&= \mathbb{E}_{xy}\left[\langle \psi(x), \mu_x\rangle_{\mathcal{F}} \langle \phi(y), \mu_y\rangle_{\mathcal{G}}\right] \\
&= \mathbb{E}_{xy}\left[\mathbb{E}_{x'}[k(x,x')\mathbb{E}_{y'}[l(y,y')]]\right]
\end{aligned}$$

Finally, for ③, we have:

$$\begin{aligned}
\langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y\rangle_{\text{HS}} &= \langle \mu_x, \mu_x\rangle_{\mathcal{F}} \langle \mu_y, \mu_y\rangle_{\mathcal{G}} \\
&= \mathbb{E}_x[\mu_x(x)]\mathbb{E}_y[\mu_y(y)] \\
&= \mathbb{E}_{xx'}\left[k(x,x')\right]\mathbb{E}_{yy'}\left[l(y,y')\right]
\end{aligned}$$

Combining them, gives us MMD with product kernel. □

Proposition 2.6.1. *If we define i-th eigenvalue from* COCO *(eigenvevalue of* $\tilde{C}_{XY}$*) as* γ_i*, then we can show that*

$$\text{HSIC}^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma^i$$

Proof. We will proof in finite case first, starting by noting that $\text{HSIC}^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \left\|\tilde{C}_{xy}\right\|^2 = \text{tr}(C^T_{xy}C_{xy})$. Then, we will show the following:

- Trace is sum of eigenvalues. To show this, we cosnider an eigen-decomposition $A = Q\Lambda Q^{-1}$, which Λ is diagonal matrix of eigenvalues. Thus we have

$$\text{tr}(A) = \text{tr}(Q\Lambda Q^{-1}) = \text{tr}(\Lambda Q^{-1}Q) = \text{tr}(\Lambda)$$

- For matrix A^TA is eigenvalue is λ_i^2 where λ_i is the eigenvalue of A. Assume an eigenvector v_i.

$$A^TA = (Q\Lambda Q^{-1})^T(Q\Lambda Q^{-1}) = Q\Lambda^T Q^T A\Lambda Q^T = Q\Lambda^2 Q^T$$

□

Definition 2.6.8. (Unbiased Estimate of $\|C_{xy}\|_{\text{HS}}^2$) The empirical estimator of $\|C_{xy}\|_{\text{HS}}^2$ is

$$\hat{A} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) l(y_i, y_j)$$

Lemma 2.6.4.

$$\left\| \hat{C}_{xy} \right\|_{\text{HS}}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) l(x_i, x_j)$$

Proof.

$$\begin{aligned}
\left\| \hat{C}_{xy} \right\|_{\text{HS}}^2 &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \psi(x_i) \otimes \phi(y_i), \frac{1}{n} \sum_{i=1}^{n} \psi(x_i) \otimes \phi(y_i) \right\rangle \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \psi(x_i) \otimes \phi(y_i), \psi(x_j) \otimes \phi(y_j) \rangle_{\text{HS}} \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{F}} \langle \phi(y_i), \phi(y_j) \rangle_{\mathcal{F}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) l(x_i, x_j)
\end{aligned}$$

□

Definition 2.6.9. (Biased Esimate of $\|C_{xy}\|_{\text{HS}}^2$) The biased estimate of $\|C_{xy}\|_{\text{HS}}^2$ is

$$\hat{A}_b = \left\| \hat{C}_{xy} \right\|_{\text{HS}}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) l(x_i, x_j) = \frac{1}{n^2} \operatorname{tr}(KL)$$

Proposition 2.6.2. *The differences between unbiased estimate and biased estimate is:*

$$\hat{A} - \hat{A}_b = \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^{n} k_{ii} l_{ii} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k_{ij} l_{ij} \right)$$

Proof.

$$\begin{aligned}
\hat{A} - \hat{A}_b &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) l(y_i, y_j) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) l(x_i, x_j) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n-1} \sum_{j \neq i}^{n} k_{ij} l_{ij} - \frac{1}{n} \sum_{j=1}^{n} k_{ij} l_{ij} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} k_{ii} k_{jj} - \frac{1}{n-1} \left[\sum_{j \neq i}^{n} k_{ij} l_{ij} \right] - \frac{1}{n} \left[\sum_{j \neq i}^{n} k_{ij} l_{ij} \right] \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} k_{ii} k_{jj} - \frac{1}{n(n-1)} \left[\sum_{j \neq i}^{n} k_{ij} l_{ij} \right] \right)
\end{aligned}$$

□

Proposition 2.6.3. *The biased estimate of* HSIC^2 *is equal to:*

$$\widehat{\text{HSIC}}^2 = \frac{1}{n^2}\,\text{tr}(KHLH)$$

Proof. We consider the empirical estimate of

$$\begin{aligned}
\left\|\hat{C}_{xy} - \hat{\mu}_x \otimes \hat{\mu}_y\right\|_{\text{HS}}^2 &= \left\langle \hat{C}_{xy} - \hat{\mu}_x \otimes \hat{\mu}_y, \hat{C}_{xy} - \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{\text{HS}} \\
&= \underbrace{\left\langle \hat{C}_{xy}, \hat{C}_{xy} \right\rangle_{\text{HS}}}_{①} - 2\underbrace{\left\langle \hat{C}_{xy}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{\text{HS}}}_{②} + \underbrace{\langle \hat{\mu}_x \otimes \hat{\mu}_y, \hat{\mu}_x \otimes \hat{\mu}_y \rangle_{\text{HS}}}_{③}
\end{aligned}$$

For ①, we use the result from lemma 2.6.4. Let's consider the second one ②:

$$\begin{aligned}
\left\langle \hat{C}_{xy}, \hat{\mu}_x \otimes \hat{\mu}_y \right\rangle_{\text{HS}} &= \left\langle \hat{\mu}_x, \hat{C}_{xy}\hat{\mu}_y \right\rangle_{\text{HS}} \\
&= \left\langle \frac{1}{n}\sum_{a=1}^{n} \psi(x_a), \left(\frac{1}{n}\sum_{b=1}^{n} \psi(x_b) \otimes \phi(x_b)\right)\left(\frac{1}{n}\sum_{c=1}^{n} \phi(y_c)\right) \right\rangle \\
&= \frac{1}{n^3}\left\langle \sum_{a=1}^{n} \psi(x_a), \left(\sum_{b=1}^{n}\sum_{c=1}^{n} \left[\psi(x_b) \otimes \phi(y_b)\right]\phi(y_c)\right) \right\rangle \\
&= \frac{1}{n^3}\left\langle \sum_{a=1}^{n} \psi(x_a), \left(\sum_{b=1}^{n}\sum_{c=1}^{n} \langle \phi(y_b), \phi(y_c)\rangle\, \psi(x_b)\right) \right\rangle \\
&= \frac{1}{n^3}\sum_{b=1}^{n}\sum_{c=1}^{n} l(y_b, y_c) \left\langle \sum_{a=1}^{n} \phi(x_a), \phi(x_b) \right\rangle \\
&= \frac{1}{n^3}\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n} l(y_b, y_c)k(x_a, x_b) = \frac{1}{n^3}\mathbf{1}^T KL\mathbf{1}
\end{aligned}$$

For the expansion please see appendix A.1.6. For ③, we have:

$$\begin{aligned}
\langle \hat{\mu}_x \otimes \hat{\mu}_y, \hat{\mu}_x \otimes \hat{\mu}_y \rangle_{\text{HS}} &= \langle \hat{\mu}_x, \hat{\mu}_x \rangle_{\mathcal{F}} \langle \hat{\mu}_y, \hat{\mu}_y \rangle_{\mathcal{G}} \\
&= \left\langle \frac{1}{n}\sum_{i=1}^{n} \psi(x_i), \frac{1}{n}\sum_{i=1}^{n} \psi(x_i) \right\rangle \cdot \left\langle \frac{1}{n}\sum_{i=1}^{n} \phi(y_i), \frac{1}{n}\sum_{i=1}^{n} \phi(y_i) \right\rangle \\
&= \left(\frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n} k(x_a, x_b)\right)\left(\frac{1}{n^2}\sum_{c=1}^{n}\sum_{d=1}^{n} k(y_c, y_d)\right) \\
&= \frac{1}{n^4}(\mathbf{1}^T K\mathbf{1})(\mathbf{1}^T L\mathbf{1})
\end{aligned}$$

Then we have:

$$\begin{aligned}
\widehat{\text{HSIC}}^2 &= \frac{1}{n^2}\,\text{tr}(KL) - \frac{2}{n^3}\mathbf{1}^T KL\mathbf{1} + \frac{1}{n^4}(\mathbf{1}^T K\mathbf{1})(\mathbf{1}^T L\mathbf{1}) \\
&= \frac{1}{n^2}\left(\text{tr}(KL) - \frac{2}{n}\,\text{tr}(\mathbf{1}^T KL\mathbf{1}) + \frac{1}{n^2}\,\text{tr}(\mathbf{1}^T K\mathbf{1}\mathbf{1}^T L\mathbf{1})\right) \\
&= \frac{1}{n^2}\left(\text{tr}(KL) - \frac{1}{n}\,\text{tr}(\mathbf{1}\mathbf{1}^T KL) - \frac{1}{n}\,\text{tr}(K\mathbf{1}\mathbf{1}^T L) + \frac{1}{n^2}\,\text{tr}(\mathbf{1}\mathbf{1}^T K\mathbf{1}\mathbf{1}^T L)\right) \\
&= \frac{1}{n^2}\,\text{tr}\left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)KL - \frac{1}{n}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)K\mathbf{1}\mathbf{1}^T L\right) \\
&= \frac{1}{n^2}\,\text{tr}\left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(L - \frac{1}{n}\mathbf{1}\mathbf{1}^T L\right)\right) = \frac{1}{n^2}\,\text{tr}\left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)L\right)
\end{aligned}$$

Note that the third equality comes from

$$\mathrm{tr}(\mathbf{1}^T KL\mathbf{1}) = \mathrm{tr}(\mathbf{1}^T L^T K^T \mathbf{1}) = \mathrm{tr}(\mathbf{1}^T LK\mathbf{1}) = \mathrm{tr}(K\mathbf{1}\mathbf{1}^T L)$$

□

Proposition 2.6.4. *The unbiased estimate of* HSIC^2 *is equal to:*

$$\mathrm{HSIC}^2 = \frac{1}{n(n-3)}\left[(K' \odot L')_{++} - \frac{2}{(n-2)}\mathbf{1}^T K'L'\mathbf{1} + \frac{1}{(n-1)(n-2)}\left(\mathbf{1}^T K'\mathbf{1}\right)\left(\mathbf{1}^T L'\mathbf{1}\right)\right]$$

where $(\cdot)_{++}$ *is elementwise sum, and where* K', L' *is this cases are* K *and* L *with zero diagonal entries.*

Theorem 2.6.4. *The asympototic of HSIC when* $P_{XY} = P_x P_y$ *is given by*

$$n\widehat{\mathrm{HSIC}} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2$$

where $z_l \sim \mathcal{N}(0,1)$*, which is sampled iid, and*

$$\lambda_l \psi_l(z_j) = \int h_{ijqr}\psi_l(z_i)\ \mathrm{d}F_{iqr} \qquad h_{ijqr} = \frac{1}{4!}\sum_{(tuvw)}^{(ijqr)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv}$$

Remark 49. We can find the null hypothesis by permuting the set. We will repeat many difference parameters to get the empirical CDF, and the threshold c_α, which is $1-\alpha$ quantile with moment matching:

$$n\,\mathrm{HSIC}_b(z) \sim \frac{x^{\alpha-1}\exp(1 - x/\beta)}{\beta^\alpha \Gamma(\alpha)}$$

as we set

$$\alpha = \frac{\mathbb{E}[\mathrm{HSIC}_b]^2}{\mathrm{var}(\mathrm{HSIC}_b)} \qquad \beta = \frac{\mathrm{var}(\mathrm{HSIC}_b)}{n\mathbb{E}[\mathrm{HSIC}_b]}$$

Note that this moment matching is purely heuristic, and therefore, there is no guarantee for this.

2.7 Testing Goodness of Fit

Remark 50. We would like to compare a sample Q with a distribution P. However, to use MMD:

$$\mathrm{MMD}(P,Q) = \sup_{\|f\|_{\mathcal{H}} \le 1}\Big[\mathbb{E}_Q[f] - \mathbb{E}_P[f]\Big]$$

we could sample from P but that isn't efficient nor possible (if we only know P up to a constant), while we can't also compute $\mathbb{E}_P[f]$ in a closed form.

Definition 2.7.1. (Stein Operator) The operator is defined as:

$$[T_P f](x) = \frac{1}{P(x)}\frac{d}{dx}(f(x)P(x))$$

Lemma 2.7.1. $\mathbb{E}_P[T_P f] = 0$

Proof.

$$\begin{aligned}\int \frac{P(x)}{P(x)}\frac{d}{dx}(f(x))P(x)\ \mathrm{d}x &= \int \frac{d}{dx}(f(x)P(x))\ \mathrm{d}x \\ &= f(x)P(x)\Big|_{-\infty}^{\infty} = 0\end{aligned}$$

□

Definition 2.7.2. (Kernel Stein Discrepancy) We define the metrics as:

$$\text{KSD}(P, Q; \mathcal{F}) = \sup_{\|f\|_{\mathcal{H}} \le 1} \Big[\mathbb{E}_Q[T_P f] - \mathbb{E}_P[T_P f]\Big] = \sup_{\|f\|_{\mathcal{H}} \le 1} \mathbb{E}_Q[T_P f]$$

Lemma 2.7.2. *Stein Operator can be re-written as:*

$$[T_P f](x) = \frac{d}{dx} f(x) + f(x) \frac{d}{dx} \log P(x)$$

Proof. We can write the expression as:

$$\begin{aligned}
\frac{1}{P(x)} \frac{d}{dx}(f(x)P(x)) &= \frac{1}{P(x)} \left[f(x) \frac{d}{dx} P(x) + P(x) \frac{d}{dx} f(x) \right] \\
&= \frac{f(x)}{P(x)} \frac{d}{dx} P(x) + \frac{d}{dx} f(x) \\
&= f(x) \frac{d}{dx} \log P(x) + \frac{d}{dx} f(x)
\end{aligned}$$

□

Remark 51. Consider the fourier transform, $f(x)$ where we have

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(ilx) \qquad \hat{f}_l = \int_{-\pi}^{\pi} f(x) \exp(-ilx) \text{ d}x$$

The fourier representation of the derivative is:

$$\frac{d}{dx} f(x) \xrightarrow{\mathcal{F}} \{(il)\hat{f}_l\}_{i=-\infty}^{\infty}$$

Proposition 2.7.1. *We can show the reproducbility of the differentaible:*

$$\begin{aligned}
\frac{d}{dx} f(x) &= \left\langle f, \frac{d}{dx} k(\cdot, x) \right\rangle \\
\frac{d}{dx} \frac{d}{dx'} k(x - x') &= \left\langle \frac{d}{dx'} k(\cdot, x'), \frac{d}{dx} k(\cdot, x) \right\rangle
\end{aligned}$$

Proof. We will consider the periodic kernel, where $\mathcal{X} = [-\pi, \pi]$, We define:

$$g(y) = \frac{d}{dx} k(x - y) = \sum_{l=-\infty}^{\infty} (il) \hat{k}_l \exp(il(x - y))$$

Since we can see that $g(y)$ is real, we can have:

$$g(y) = \overline{g(y)} = \sum_{l=-\infty}^{\infty} (-il) \hat{k}_l \exp(il(y - x))$$

Let's consider the inner product on the

$$\begin{aligned}
\left\langle f, \frac{d}{dx} k(x, \cdot) \right\rangle &= \langle f, g(\cdot) \rangle_{\mathcal{F}} \\
&= \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{g}}_l}{\hat{k}_l} \\
&= \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \overline{-il\hat{k}_l \exp(il(x - y))}}{\hat{k}_l} \\
&= \sum_{l=-\infty}^{\infty} (il) \hat{f}_l \exp(ilx) = \frac{d}{dx} f(x)
\end{aligned}$$

□

Theorem 2.7.1. *There exists an feature map, where:*

$$\mathbb{E}_{z\sim Q}[T_P f] = \mathbb{E}_{z\sim Q}\langle f, \xi_z\rangle_{\mathcal{F}} = \langle f, \mathbb{E}_{z\sim Q}[\xi_z]\rangle_{\mathcal{F}} \quad \textit{where} \quad \xi_z = k(\cdot, z)\frac{d}{dz}\log p(z) + \frac{d}{dz}k(\cdot, z)$$

If

$$\mathbb{E}_{z\sim Q}\left(\frac{d}{dz}\log p(z)\right)^2 < \infty$$

Proof. We will proof this by Riesz theorem, where we need a boundness. We can consider the Jensen's inequality and Cauchy-Schwarz:

$$\begin{aligned} |\mathbb{E}_{z\sim Q}\langle f, \xi_z\rangle_{\mathcal{F}}| &\le \mathbb{E}_{z\sim Q}|\langle f, \xi_z\rangle_{\mathcal{F}}| \\ &\le \|f\|\,\mathbb{E}_{z\sim Q}\|\xi_z\|_{\mathcal{F}} \end{aligned}$$

We will have to show that this square norm $\|\xi_z\|_{\mathcal{F}}$ is bounded:

$$\begin{aligned} \|\xi_z\|_{\mathcal{F}} &= \langle \xi_z, \xi_z\rangle_{\mathcal{F}} \\ &= \left\langle k(\cdot, z)\frac{d}{dz}\log p(z) + \frac{d}{dz}k(\cdot, z), k(\cdot, z)\frac{d}{dz}\log p(z) + \frac{d}{dz}k(\cdot, z)\right\rangle_{\mathcal{F}} \\ &= \underbrace{\left\langle k(\cdot, z)\frac{d}{dz}\log p(z), k(\cdot, z)\frac{d}{dz}\log p(z)\right\rangle}_{①} + \underbrace{\left\langle \frac{d}{dx}k(\cdot, x), \frac{d}{dx'}k(\cdot, x')\right\rangle\Big|_{x=x'=z}}_{②} \\ &\quad + \underbrace{\left\langle k(\cdot, x)\frac{d}{dx}\log p(x), \frac{d}{dx'}k(\cdot, x')\right\rangle\Big|_{x=x'=z}}_{③} \\ &= c + \left(\frac{d}{dz}\log p(z)\right)^2 c \end{aligned}$$

where we set $k(z, z) = c$. Now, consider each terms: Starting with the first term ①:

$$\left\langle k(\cdot, z)\frac{d}{dz}\log p(z), k(\cdot, z)\frac{d}{dz}\log p(z)\right\rangle = \left[\left(\frac{d}{dz}\log p(z)\right)^2 k(z, z)\right] = \left[\frac{d}{dz}\log p(z)\right]^2 c$$

Now, consider the second part ②:

$$\begin{aligned} \left\langle \frac{d}{dx}k(\cdot, x), \frac{d}{dx'}k(\cdot, x')\right\rangle\Big|_{x=x'=z} &= \sum_{l=-\infty}^{\infty} \frac{\left[-il\hat{k}_l\exp(-ilx)\right]\overline{\left[-il\hat{k}_l\exp(-ilx')\right]}}{\hat{k}_l} \\ &= \sum_{l=-\infty}^{\infty} -(il)^2\hat{k}_l\underbrace{\exp(il(x'-x))}_{\to 1} = \sum_{l=-\infty}^{\infty} l^2\hat{k}_l = c > 0 \end{aligned}$$

For the final part ③, we have:

$$\begin{aligned} \left\langle k(\cdot, z)\frac{d}{dz}\log p(z), \frac{d}{dz}k(\cdot, z)\right\rangle &= \left(\frac{d}{dz}\log p(z)\right)\sum_{l=-\infty}^{\infty} \frac{\left[\hat{k}_l\exp(-ilx)\right]\overline{\left[(-il)\exp(-ilx')\hat{k}_l\right]}}{\hat{k}_l} \\ &= \left(\frac{d}{dz}\log p(z)\right)\sum_{l=-\infty}^{\infty}(il)\hat{k}_l\underbrace{\exp(il(x'-x))}_{\to 1} = 0 \end{aligned}$$

Given the boundness, we have

$$\begin{aligned}\mathbb{E}_{z\sim Q}\|\xi_z\|_{\mathcal{F}} &= \mathbb{E}_{z\sim Q}\sqrt{c+\left(\frac{d}{dz}\log p(z)\right)^2 c} \\ &\le \sqrt{\mathbb{E}_{z\sim Q}\left[c+\left(\frac{d}{dz}\log p(z)\right)^2 c\right]}\end{aligned}$$

Thus, we have the condition that riesz to hold. □

Remark 52. However, the bound condition might not hold. Consider the normal distribution:

$$P(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/x)$$

Then its derivative is $-x$. If q is Cauchy distribution, then the integral is

$$\mathbb{E}_{z\in Q}\left(\frac{d}{dz}\log p(z)\right)^2 = \int_{-\infty}^{\infty} z^2 q(z)\ \mathrm{d}z$$

This is undefined.

Proposition 2.7.2. *The closed form expression of KSD given indepdent $z, z' \sim q$, then:*

$$\mathrm{KSD}(P, Q, \mathcal{F}) = \|\mathbb{E}_{z\in Q}\xi_z\|_{\mathcal{F}}$$

Proof.

$$\begin{aligned}\mathrm{KSD}(P, Q, \mathcal{F}) &= \sup_{\|g\|_{\mathcal{F}}\le 1} \mathbb{E}_{z\sim Q}[(T_P g)(z)] \\ &= \sup_{\|g\|_{\mathcal{F}}\le 1} \mathbb{E}_{z\sim Q}\langle g, \xi_z\rangle_{\mathcal{F}} \\ &= \sup_{\|g\|_{\mathcal{F}}\le 1} \langle g, \mathbb{E}_{z\sim Q}\xi_z\rangle_{\mathcal{F}} = \|\mathbb{E}_{z\sim Q}\xi_z\|_{\mathcal{F}}\end{aligned}$$

□

Proposition 2.7.3. *We can have the following test statistics:*

$$\|\mathbb{E}_{z\sim Q}\xi_z\|^2_{\mathcal{F}} = \mathbb{E}_{z,z'\sim q}h_p(z, z')$$

where we have

$$\begin{aligned}h_p(x, y) =& \frac{d}{dx}\log p(x)\frac{d}{dy}\log p(y)k(x, y) + \frac{d}{dy}\log p(y)\frac{d}{dx}k(x, y) \\ &+ \frac{d}{dx}\log p(x)\frac{d}{dy}k(x, y) + \frac{d}{dx}\frac{d}{dy}k(x, y)\end{aligned}$$

Remark 53. Given an example $\{z_i\}_{i=1}^n$ empirical KSD is

$$\widehat{\mathrm{KSD}}(P, Q; \mathcal{F}) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i}^n h_p(z_i, z_j)$$

when $q = p$ we obtain the estimate of null distribution with wild bootstrap:

$$\widetilde{\mathrm{KSD}}(P, Q; \mathcal{F}) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i}^n \sigma_i\sigma_j h_p(z_i, z_j)$$

when $\{\sigma_i\}_{i=1}^n$ is sampled iid where $\mathbb{E}[\sigma_i] = 0$ and $\mathbb{E}[\sigma_i^2] = 1$

2.8 Support Vector Machine

2.8.1 Introduction

Definition 2.8.1. (Learning Problem) Given a set of paired observation $(x_1, y_1), \ldots, (x_n, y_n)$ either for regression or classification task. We would like to find a function f^* in RKHS $\mathcal{H}$ that satisfies:

$$f^* = \arg\min_{f \in \mathcal{H}} J(f) = \arg\min_{f \in \mathcal{H}} L_y(f(1), \ldots, f(x_n)) + \Omega\left(\|f\|^2_{\mathcal{H}}\right)$$

where Ω is non-decreasing, y is the vector of y_i and loss L that depends on x_i only via $f(x_i)$.

Theorem 2.8.1. *The representor theorem is a solution to:*

$$\min_{f \in \mathcal{H}} \left[L_y(f(x_1), \ldots, f(x_n)) + \Omega\left(\|f\|^2_{\mathcal{H}}\right)\right]$$

which takes the form:

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$$

If Ω is strictly increasing, then the solution must take this form.

Proof. Denote f_S is the projection of f onto the subspace: $\text{span}\{k(x, \cdot) : 1 \le i \le n\}$, such that $f = f_S + f_\perp$ where $f_S = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$. The regularizer is given by $\|f\|^2_{\mathcal{H}} = \|f_\perp\|^2_{\mathcal{H}} + \|f_S\|^2_{\mathcal{H}} \ge \|f_S\|^2_{\mathcal{H}}$. Then by the definition of Ω:

$$\Omega\left(\|f\|^2_{\mathcal{H}}\right) \ge \Omega\left(\|f_S\|^2_{\mathcal{H}}\right)$$

This is clear that this minimize for $f = f_S$. The individual terms $f(x_i)$ in the loss is:

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_S + f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_S, k(x_i, \cdot)\rangle$$

And, so we have $L_y(f(x_1), \ldots, f(x_n)) = L_y(f_S(x_1), \ldots, f_S(x_n))$. Hence, it is clear tha the loss $L(\cdot)$ only depends on the component of f in data subspace:

- Regularizer is minimal when $f = f_S$
- If Ω is non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is minimum. If Ω strictly increasing, as minimum is unique.

□

Definition 2.8.2. (SVM) We will classify 2 clouds of points, where there exists a hyperplane, which linearly separate one cloud from the other without error: The smallest distance each class to the seperating hyperplane $w^T x + b$ is called margin. We can express the problem as follows:

$$\begin{aligned} \min_{w,b} \left(\|w\|^2\right) &= \max_{w,b} \left(\frac{2}{\|w\|}\right) \\ \text{subject to } w^T x_i + b \ge 1 &\quad i : y_i = +1 \\ w^T x_i + b \le 1 &\quad i : y_i = -1 \end{aligned}$$

Please not that we can solve this problem via convex optimization.

Remark 54. To have the sepearting hyperplane, the distance between them

$$d = (x_+ - x_-)^T \frac{w}{\|w\|}$$

Now, we can see that the constraint is:

$$w^T x_+ + b = 1 \qquad w^T x_- + b = -1$$

If we minus themselves together and we have $w^T(x_+ - x_-) = 2$, then it is clear that $d = 2/\|w\|$ as required.

2.8.2 Convex Optimization

Definition 2.8.3. (Convex Set) A set C is convex iff for all $x_1, x_2 \in C$ and any $0 \le \theta \le 1$, which we have:

$$\theta x_1 + (1-\theta)x_2 \in C$$

Definition 2.8.4. (Convex Function) A function f is convex if its domain $\text{dom}(f)$ is a convex set if for all $x, y \in \text{dom}(f)$ and for any $0 \le \theta \le 1$:

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y)$$

The function is strictly convex if the inequality is strict for $x \neq y$.

Definition 2.8.5. (Optimization Problem) The optimization problem on $x \in \mathbb{R}^n$:

$$\begin{aligned} &\min f_0(x) \\ \text{subject to } &f_i(x) \le 0 \quad i = 1, \dots, m \\ &h_i(x) = 0 \quad 1, \dots, p \end{aligned}$$

The point p^* is optimal value. $\mathcal{D}$ assumed non-empty where:

$$\mathcal{D} = \bigcap_{i=0}^{m} \text{dom}(f_i) \cap \bigcap_{i=1}^{m} \text{dom}(h_i)$$

Remark 55. Ideally, we have unconstraint problem:

$$\min f_0(x) + \sum_{i=1}^{m} l_-(f_i(x)) + \sum_{i=1}^{p} l_0(h_i(x)) \quad \text{where} \quad L_- = \begin{cases} 0 & u \le 0 \\ \infty & u > 0 \end{cases}$$

and $l_0(u)$ is indicator of 0.

Definition 2.8.6. (Lagrangian) The Lagragian is the lower bound on the original problem:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i f_i(x)}_{\le l_-(f_i(x))} + \sum_{i=1}^{p} \underbrace{\nu_i h_i(x)}_{\le l_0(h_i(x))}$$

It has a domain $\text{dom}(L) = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. The vector λ and ν are called Lagrange multiplier or dual variable to ensure lower bound, we require $\lambda \succeq 0$.

Definition 2.8.7. (Dual Function) Minimize Lagragian when $\lambda \ge 0$ and $f_i(x) \le 0$. The Lagrange dual function is:

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$

A dual feasible pair (λ, ν) is a pair for which $\lambda \succeq 0$ and $(\lambda, \nu) \in \text{dom}(g)$

Proposition 2.8.1. *For any $\lambda \succeq 0$ and ν, we have $g(\lambda, \nu) \le f_0(x)$ whenever $f_i(x) \le 0$ and $h_i(x) = 0$, including $f_0(x^*) = p^*$*

Proof. Assume $\tilde{x}$ is feasbile i.e $f(\tilde{x}) \le 0$ and $h_i(\tilde{x}) = 0$ and $\tilde{x} \in \mathcal{D}$ and $\lambda \succeq 0$ then:

$$\sum_{i=1}^{n} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \le 0$$

Thus, we have:

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \\ &\le f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \\ &\le f_0(\tilde{x}) \end{aligned}$$

The best lower bound $g(\lambda, \nu)$ on the optimal problem solution p^*. □

Definition 2.8.8. (Lagrange Dual Problem)

$$\begin{aligned} &\max g(\lambda, \nu) \\ &\text{subject to } \lambda \succeq 0 \end{aligned}$$

The dual feasible (λ, ν) with $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$

Definition 2.8.9. (Dual Optimal) The solution (λ^*, ν^*) of the maximal dual and d^* is optimal value. The weak duality holds if $d^* \leq p^*$. However, the strong duality $d^* = p^*$ might not always holds.

Remark 56. If this strong duality holds, we have easy concave dual problem to find p^*. Dual function is a pointwise infininum of affine function of (λ, ν) hence concave in (λ, ν) with convex constraint set $\lambda \succeq 0$

Proposition 2.8.2. *The sufficient condition (non-necessary) for strong duality, which holds if:*

$$\begin{aligned} \min f_0(x) \\ \text{subject to } f_i(x) \leq 0 \quad i = 1, \dots, n \\ Ax = b \end{aligned}$$

as h_i is affine, for convex $f_0, \dots, f_n$. And, Slater's condition holds: if there exists some strictly feasbile points $\tilde{x} \in \operatorname{relint}(\mathcal{D})$ such that: $f_i(\tilde{x}) < 0$ for $i = 1, \dots, m$ where $A\tilde{x} = b$. For the case of affine f_i, the condition is trivial (the inequality constriants no longer strict, reduces to original inequality constraint):

$$f_i(\tilde{x}) \leq 0 \quad i - 1, \dots, m \quad A\tilde{x} = b$$

Proposition 2.8.3. ***(Complementary Slackness)*** *The complementary slackness is the consequence of strong duality, where we have:*

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

which is the condition of complementary slackness, which implies that:

$$\lambda_i^* > 0 \implies f_i(x^*) = 0 \qquad f_i^*(x^*) < 0 \implies \lambda_i^* = 0$$

Proof. Assume the primal is equal to dual then we have x^* solution of original problem and (λ^*, ν^*) is the solution to the dual:

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$

The last inequality comes from x^*, λ^*, ν^* satisfies $\lambda \succeq 0$, $f_i(x^*) \leq 0$, $h_i(x^*) = 0$. □

Definition 2.8.10. (KKT Condition For Global Optimum) Assume function f_i, h_i are differentiable and strong duality, since x^* minimize $L(x, \lambda^*, \nu^*)$ derivative at x^* is zero:

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu^* \nabla h_i(x^*) = 0$$

KKT condition means: we are at global optimum $(x, \lambda, \nu) = (x^*, \lambda^*, \nu^*)$ when:

- Strong Duality Holds (primal problem convex and constraint functions satisfy Slater's condition)

- Primal Feasibility:

$$\begin{cases} f_i(x) \leq 0 & i = 1, \ldots, m \\ h_i(x) = 0 & i = 1, \ldots, p \end{cases}$$

- Dual Feasibility: $\lambda_i \geq 0$ and $i = 1, \ldots, m$
- Complementary Slackness: $\lambda_i f_i(x) = 0$ and $i = 1, \ldots, m$
- Zero Gradient:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda \nabla_i f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

Furthermore, KKT conditions necessary and sufficient for optimality.

Definition 2.8.11. (Optimization Problem for SVM) The problem can be expressed as follows:

$$\begin{aligned} &\max_{w,b} \left(\frac{2}{\|w\|} \right) \\ \text{subject to } &\min(w^T x_i + b) = 1 \quad i : y_i = 1 \\ &\max(w^T x_i + b) = -1 \quad i : y_i = -1 \end{aligned}$$

and we have the classifier to be $y = \text{sign}(w^T x + b)$, where we can re-write it case:

$$\begin{aligned} &\min_{w,b} \|w\|^2 \\ \text{subject to } &y_i(w^T x_i + b) \geq 1 \end{aligned}$$

We allow error points within a margin, or even on the wrong side of the decision boundary. However, ideally, we need the following optimization:

$$\min_{w,b} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \mathbb{I} \left[y_i(w^T x_i + b) < 0 \right] \right)$$

We will replace with convex upper bound, with hinge loss

$$\min_{w,b} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \theta \left(y_i(w^T x_i + b) < 0 \right) \right) \quad \text{where} \quad \theta(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha & 1-\alpha > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now, we replace a hinge loss with simple inequality constraints:

$$\begin{aligned} &\min_{w,b,\xi_i} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right) \\ \text{subject to } &\xi_i \geq 0 \\ &y_i(w^T x_i + b) \geq 1 - \xi_i \end{aligned}$$

Please note that:

- $y_i(w^T x_i + b) \geq 1$ and $\xi_i = 0$. We can minimize if its is correct.
- $y_i(w^T x_i + b) < 1$ and $\xi_i > 0$ takes the value satisfying $y_i(w^T x_i + b) = 1 - \xi_i$. We are able to decrease, which looks like the hinge loss. We can decrease till $1 - \xi_i$ is equal.

Remark 57. The strong duality holds. The optimization problem convex with respect to the variable w, b, ξ turned to ?

$$\min_{w,b,\xi} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right)$$

$$\text{subject to } \xi_i \geq 0$$

$$1 - \xi_i - y_i(w^T x_i + b) \leq 0 \quad i = 1, \dots, n$$

This is clear that $f_0, f_1, \dots, f_n$ are convex. The slater's condition holds. It is trivial since inequality constriants affine and there exists some $\xi_i \geq 0$:

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

Thus the strong duality holds, the problem is differentaible and so KKT holds at global optimum.

Remark 58. C is a hyperparameter that control the trade-off between the margin size and the error. One can try to reduce the error caused by the points in the margin but this might lead to too small margin i.e overfitting.

Remark 59. The Lagragian of the SVM

$$\mathcal{L}(w, b, \xi, \alpha, \lambda)$$
$$= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - (y_i)(w^T x_i + b) - \xi_i) + \sum_{i=1}^{n} \lambda_i (-\xi_i)$$

With dual variable constraint $\alpha_i \geq 0$ and $\lambda_i \geq 0$. Let's minimize the primal variables are:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \implies w = \sum_{i=1}^{n} \alpha_i y_i x_i$$
$$\frac{\partial L}{\partial b} = \sum_{i} y_i \alpha_i = 0$$
$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \implies \alpha_i = C - \lambda_i$$

Note that $\lambda_i \geq 0$ and so $a_i \leq C_i$

Remark 60. We will apply the complementary slackness:

- Non-Margin Support Vector $\alpha_i = C \neq 0$ (Error within the margin):
  - We immediately have $1 - \xi_i = y_i(w^T x_i + b)$
  - From the condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$ (hence we have $\xi_i > 0$)
- Margin Support Vector: $0 \leq \alpha_i \leq C$ (The points on the margin)
  - We again have $1 - \xi_i = y_i(w^T x_i + b)$
  - For $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$ and hence $\xi_i = 0$
- Non Support Vector: $\alpha_i = 0$
  - We have $y_i(w^T x_i + b) > 1 - \xi_i$
  - From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$ hence $\xi_i = 0$

Remark 61. We observe that:

- The solution is sparse: points not on margine or margine error have $\alpha_i = 0$
- The suppor vectors are the points on decision boundary which are margine error contribute.

- The influence of non-margine support vector is bounded since their weight can't exceed C.

We can only remember the points that are critical i.e the first and the second one, which we can remove all the third category point, and still have the same training capability.

Proposition 2.8.4. *The dual of the SVM is given by:*

$$\begin{aligned} g(\alpha, \lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i(1 - (y_i)(w^T x_i + b) - \xi_i) + \sum_{i=1}^n \lambda_i(-\xi_i) \\ &= \frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j x_i^T x_j + C\sum_{i=1}^m \xi_i \\ &\quad - \sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j x_i^T x_j - b\sum_{i=0}^m \alpha_i y_i \\ &\quad + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i\xi_i - \sum_{i=1}^m \underbrace{(c - \alpha_i)}_{\lambda_i}\xi_i \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j x_i^T x_j \end{aligned}$$

We would like to maximize the dual subjected to constraint $0 \leq \alpha_i \leq C$ *where* $\sum_{i=1}^n y_i\alpha_i = 0$. *This is quadratic program. For margin SV, we have* $1 = y_i(w^T x_i + b)$ *to obtain b for any of these or take an average.*

Definition 2.8.12. (Kernelized SVM) We have max margin classifier in RKHS. Given a hinge loss formulation:

$$\min_w \left(\frac{1}{2}\|w\|_{\mathcal{H}}^2 + C\sum_{i=1}^n \theta(y_i, \langle w, k(x_i, \cdot)\rangle_{\mathcal{H}})\right)$$

For RKHS with kernel $k(x, \cdot)$. We use a result of representor theorem:

$$w(\cdot) = \sum_{i=1}^n \beta_i k(x_i, \cdot)$$

For maximizing the margin equivalent to minimize $\|w\|_{\mathcal{H}}^2$: for any RKHS a smoothness constraint holds. The optimization problem becomes:

$$\begin{aligned} &\min_{\beta,\xi} \left(\frac{1}{2}\beta^T K\beta + C\sum_{i=1}^n \xi_i\right) \\ \text{subject to } &\xi_i \geq 0 \\ &y_i\sum_{i=1}^n \beta_j k(x_i, x_j) \geq 1 - \xi_i \end{aligned}$$

This is convex in β and ξ, since $K \succeq 0$, which strong duality holds, where the dual is

$$\begin{aligned} g(\alpha) &= \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j k(x_i, x_j) \\ \text{subject to } w(\cdot) &= \sum_{i=1}^m y_i\alpha_i k(x, \cdot) \quad 0 \leq \alpha_i \leq C \end{aligned}$$

Definition 2.8.13. (ν_i-SVM) We have other kind of SVM, where we have intuitive parameter ν as C is hard to interpret. Let's drop b for simplicity and we have:

$$\begin{aligned} &\min_{w,\rho,\xi} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i\right) \\ \text{subject to } &\rho \geq 0 \\ &\xi_i \geq 0 \\ &y_i w^T x \geq \rho - \xi_i \end{aligned}$$

Now, we are directly adjusting margin width ρ.

Remark 62. We have the following Lagragian:

$$\frac{1}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \nu\rho + \sum_{i=1}^{n}\alpha_i(\rho - y_i w^T x_i - \xi_i) + \sum_{i=1}^{n}\beta_i(-\xi_i) + \gamma(-\rho)$$

for dual variable $\alpha_i \geq 0$, $\beta_i \geq 0$ and $\gamma \geq 0$. Differentiating and setting to zero for each of primal variables w, ξ, ρ:

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i \quad \alpha_i + \beta_i = \frac{1}{n} \quad \nu = \sum_{i=1}^{n}\alpha_i - \gamma$$

from $\beta \geq 0$ we have $0 \leq \alpha_i \leq 1/n$

Remark 63. For complementary slacknes condition, we assume $\rho > 0$ at global solution, hence $\gamma = 0$ and $\sum_{i=1}^{n}\alpha_i = \nu_i$:

- Case of $\xi_i > 0$: Complementary Slackenss state $\beta_i = 0$, hence we have $\alpha_i = n^{-1}$. This denotes this set as $N(\alpha)$, then:

$$\sum_{i\in N(\alpha)}\frac{1}{n} = \sum_{i\in N}\alpha_i \leq \sum_{i=1}^{n}\alpha_i = \nu \quad \text{where} \quad \frac{|N(\alpha)|}{n} \leq \nu$$

- Case of $\xi_i = 0$: where $\beta_i > 0$ then $\alpha_i < n^{-1}$. The set is denoted by $M(\alpha)$. The set of points $n^{-1} > \alpha_i > 0$ is

$$\nu = \sum_{i=1}^{n}\alpha_i = \sum_{i\in N(\alpha)}\frac{1}{n} + \sum_{M(\alpha)}\alpha_i \leq \sum_{i\in M(\alpha)\cup N(\alpha)}\frac{1}{n} \quad \text{where} \quad \nu \leq \frac{|N(\alpha)| + |M(\alpha)|}{n}$$

  and ν is the lower bound based on number of support vector with non-zero weight on margin and margin error.

Remark 64. Let's substute to the Lagragian, as we have:

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \rho\nu - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &\qquad + \sum_{i=1}^{n}\alpha_i\rho - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\left(\frac{1}{n} - \alpha_i\right)\xi_i - \rho\left(\sum_{i=1}^{n}\alpha - \nu\right) \\ &= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \end{aligned}$$

Therefore, the dual is:

$$g(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$\text{subject to } \sum_{i=1}^{n}\alpha_i \geq \nu$$

$$0 \leq \alpha_i \leq \frac{1}{n}$$

Chapter 3

Statisical Learning

3.1 Formulating Learning Problem

3.1.1 Problem

Definition 3.1.1. (Learning Problem) We have the following components for learning problems:

- $\mathcal{X}$: input space.
- $\mathcal{Y}$: output space.
- ρ: unknown distribution on $\mathcal{X} \times \mathcal{Y}$
- $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: loss function that measure discrepancy between $y, y' \in \mathcal{Y}$

We want to minimize the expected risk:

$$\inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \qquad \text{where } \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} l(f(x), y) \ \mathrm{d}\rho(x, y)$$

The relation between $\mathcal{X}$ and $\mathcal{Y}$ are determined by unknown ρ, while we can only access via finite sample.

Remark 65. **(Loss Function for Regression)** The loss function for regression would be in the form of

$$L(y, y') = L(y - y')$$

The examples of this kind of loss is:

- Square Loss: $(y - y')^2$
- Absolute Loss: $|y - y'|$
- ε-sensitive Loss: $\max(|y - y'| - \varepsilon, 0)$

Remark 66. **(Loss Function for Classification)** The loss function for classification would be

$$L(y, y') = L(yy')$$

The examples of this kind of loss is:

- 0-1 Loss: $\mathbf{1}_{-yy'>0}$

- Square loss Loss: $(1 - yy')^2$
- Hinge Loss: $\max(1 - yy', 0)$
- Logistic Loss: $\log(1 + \exp(-yy'))$

Definition 3.1.2. (Realistic Learning Problem) We have the following components:

- $\mathcal{S} = \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n$ be set of finite dataset on $\mathcal{X} \times \mathcal{Y}$
- $\mathcal{F}$ be set of all measurable function $f : \mathcal{X} \to \mathcal{Y}$
- $A : \mathcal{S} \to \mathcal{F}$ be a learning algorithm where $S \mapsto A(S) : \mathcal{X} \to \mathcal{Y}$

We will study the relation between the size of training set and corresponding predictor $f_n = A((x_i, y_n)_{i=1}^n)$.

Remark 67. We can consider the stochastic algorithm. In this case, given a dataset $S \in \mathcal{S}$, the algorithm can be seen as a distribution over $\mathcal{F}$ and its output is simpily one sample of $A(S)$. Note that the deterministic is simpily a Direc's delta distribution.

3.1.2 Risk

Definition 3.1.3. (Excess Risk) We define an excess risk of function f_n as

$$\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)$$

Definition 3.1.4. (Consistency) The algorithm is consistence

$$\lim_{n \to \infty} \mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) = 0$$

Ideally, we want algorithm to behave like this.

Definition 3.1.5. (Notion of Convergence) However, as $f_n = A(S)$ being stochastic or random variable because the training set S is sampled from ρ, there are difference notions of convergence:

- Convergence in expectation
$$\lim_{n \to \infty} \mathbb{E}\left[\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)\right] = 0$$
- Convergence in probability. For all $\varepsilon > 0$:
$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) > \varepsilon\right) = 0$$

Remark 68. We only interested in the risk of our estimator to be the best i.e $\mathcal{E}(f_n) \to \inf_{f \in \mathcal{F}} \mathcal{E}(f)$. However, we don't care about finding the best fucntion f^*, where it is minimizer of expected risk i.e $\mathcal{E}(f^*) = \inf_{f \in \mathcal{F}} \mathcal{E}(f)$

Remark 69. The existence of f^* can be useful in several loss function. As the closer the function f to f^*, the closer the risk $\mathcal{E}(f)$ to $\mathcal{E}(f^*)$:

- For least square function: $l(f(x), y) = (f(x) - y)^2$:
$$\mathcal{E}(f) - \mathcal{E}(f^*) = \|f - f^*\|_{L^2(\mathcal{X}, \rho)}$$
- For any L-Lipschitz loss function, where $|l(z, y) - l(z', y)| \le L \|z - z'\|$, we have:
$$\mathcal{E}(f) - \mathcal{E}(f^*) \le \|f - f^*\|_{L^1(\mathcal{X}, \rho)}$$

This guarantee that the algorithm is consistency when $f \to f^*$.

Definition 3.1.6. (Learning Rate) We can measure the "speed" in which the excess risk goes to zero:

$$\mathbb{E}\left[\mathcal{E}(f_n) - \inf_{f\in\mathcal{F}} \mathcal{E}(f)\right] = \mathcal{O}(n^{-\alpha})$$

where the learning rate is α, which we can compare 2 algorithms via this value.

Definition 3.1.7. (Probabilistic Bound) We would like to consider the following probabilistic bounds on various values:

- **Sample Complexity:** A number $n(\varepsilon, \delta)$ of training points that the algorithm needs to achieve excess risk lower than ε with a least probability $1 - \delta$

$$\mathbb{P}\left(\mathcal{E}(f_{n(\varepsilon,\delta)}) - \inf_{f\in\mathcal{F}} \mathcal{E}(f) \le \varepsilon\right) \ge 1 - \delta$$

- **Error Bound:** An upperbound $\varepsilon(\delta, n)$ on the excess risk f_n, which holds with probability larger than $1 - \delta$:

$$\mathbb{P}\left(\mathcal{E}(f_n) - \inf_{f\in\mathcal{F}} \mathcal{E}(f) \le \varepsilon(\delta, n)\right) \ge 1 - \delta$$

- **Tail Bound:** A lower bound $\delta(\varepsilon, n) \in (0, 1)$ on the probability that f_n will have excess risk larger than ε:

$$\mathbb{P}\left(\mathcal{E}(f_n) - \inf_{f\in\mathcal{F}} \mathcal{E}(f) \le \varepsilon\right) \ge 1 - \delta(\varepsilon, n)$$

3.1.3 Empirical Risks

Definition 3.1.8. (Empirical Risk) Given a finite sample of data $(x_i, y_i)_{i=1}^m$, we can use empirical risk to gather the information about $\mathcal{E}(f)$ as:

$$\mathcal{E}_n(f) = \frac{1}{n}\sum_{i=1}^n l(f(x_i), y_i)$$

Proposition 3.1.1. *The expected empirical risk is expected risk* $\mathbb{E}_{S\sim\rho^n}[\mathcal{E}_n(f)] = \mathcal{E}(f)$.

Proof. We have:

$$\mathbb{E}_{S\sim\rho^n}\left[\frac{1}{n}\sum_{i=1}^n l(f(x_i), y_i)\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{(x_i,y_i)}\left[l(f(x_i), y)\right] = \frac{1}{n}\sum_{i=1}^n \mathcal{E}(f) = \mathcal{E}(f)$$

□

Lemma 3.1.1. *Let's consider an iid variables* $(x_i)_{i=1}^n$ *and let*

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$$

One can show that

$$\text{Var}(\bar{x}_n) = \frac{\text{Var}(x)}{n}$$

Proof. We have:

$$\mathbb{E}\left[(\bar{x}_n-\mu)^2\right]=\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i-\mu\right)^2\right]=\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i-\mu\right)\left(\frac{1}{n}\sum_{i=1}^n x_i-\mu\right)\right]$$
$$=\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i\right)-\frac{2\mu}{n}\sum_{i=1}^n x_i+\mu^2\right]$$
$$=\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\right]-\frac{2\mu}{n}\sum_{i=1}^n \mathbb{E}[x_i]+\mu^2$$
$$=\frac{1}{n^2}\left(n\mathbb{E}[x^2]+(n^2-n)\mu^2\right)-\mu^2=\frac{\mathbb{E}[x^2]+\mu^2}{n}=\frac{\mathrm{Var}(x)}{n}$$

where we have

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\right]=\frac{1}{n^2}\mathbb{E}\Big[x_1x_1+x_1x_2+\cdots+x_1x_n$$
$$x_2x_1+x_2x_2+\cdots+x_2x_n$$
$$\vdots$$
$$x_nx_1+x_nx_2+\cdots+x_nx_n\Big]$$
$$=\frac{1}{n^2}\left(n\mathbb{E}[x^2]+(n^2-n)\mathbb{E}[x]\mathbb{E}[x]\right)$$

□

Proposition 3.1.2. *The expected absolute difference between empirical risk and expected risk is:*

$$\mathbb{E}\left[|\mathcal{E}_n(f)-\mathcal{E}(f)|\right]\leq\sqrt{\frac{\mathrm{var}(l(f(x_i),y_i))}{n}}$$

Proof. Let's apply the lemma 3.1.1 to the empirical risk, after Jensen's ineqalities:

$$\mathbb{E}[|\mathcal{E}_n(f)-\mathcal{E}(f)|]=\mathbb{E}[\sqrt{(\mathcal{E}_n(f)-\mathcal{E}(f))^2}]$$
$$\leq\sqrt{\mathbb{E}[(\mathcal{E}_n(f)-\mathcal{E}(f))^2]}$$
$$=\sqrt{\frac{\mathrm{var}(l(f(x_i),y))}{n}}$$

□

Theorem 3.1.1. ***(Markov's Inequality)*** *Let X be non-negative random variable and $a>0$, then*

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

Proof. We consider the expectation of X:

$$\mathbb{E}[X]=\int_{-\infty}^{\infty}xp(x)\ \mathrm{d}x=\int_0^{\infty}xp(x)\ \mathrm{d}x$$
$$=\int_0^a xp(x)\ \mathrm{d}x+\int_a^{\infty}xp(x)\ \mathrm{d}x$$
$$\geq\int_a^{\infty}xp(x)\ \mathrm{d}x\geq\int_a^{\infty}ap(x)\ \mathrm{d}x$$
$$=aP(X\geq a)$$

□

Theorem 3.1.2. ***(Chebyshev's inequality)*** *Let X be random variable with finite expected value μ and non-zero variance σ^2. For any real number $k > 0$:*

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Proof. We will consider the use of Markov's inequality where the random variable be $|X - \mu|$ and the constant be $k\sigma$, then we have:

$$\mathbb{P}(|X - \mu| \geq k) = \mathbb{P}(|X - \mu|^2 \geq k^2) \leq \frac{\mathbb{E}[|X - \mu|^2]}{k^2} = \frac{\sigma^2}{k^2}$$

□

Proposition 3.1.3. *The probability of expected risk is greater than some number $\varepsilon \geq 0$ is*

$$\mathbb{P}\Big(\mathcal{E}_n(f) - \mathcal{E}(f) \geq \varepsilon\Big) \leq \frac{\text{var}(l(f(x_i), y_i))}{n\varepsilon^2}$$

This follows directly from the Chebyshev's inequality.

3.2 Generalization Bound

3.2.1 Generalization Error

Proposition 3.2.1. *We will consider the bound of the excess risk, where we assume f^* where $\mathcal{E}(f^*) = \inf_{f\in\mathcal{F}} \mathcal{E}(f)$:*

$$\mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}(f^*)\Big] \leq \mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\Big]$$

where $f_n = \arg\min_{f\in\mathcal{F}} \mathcal{E}_n(f)$

Proof. We consider the following risk decomposition:

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}(f^*)\Big] \\
&\quad = \mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}_n(f_n) + \underbrace{\mathcal{E}_n(f_n) - \mathcal{E}_n(f^*)}_{\leq 0} + \mathcal{E}_n(f^*) - \mathcal{E}(f^*)\Big] \\
&\quad \leq \mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\Big] + \mathbb{E}\Big[\mathcal{E}_n(f^*) - \mathcal{E}(f^*)\Big] \\
&\quad = \mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\Big] + 0
\end{aligned}$$

□

Definition 3.2.1. (Generalization Error) We can focus on the generalization error:

$$\mathbb{E}\Big[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\Big]$$

Proposition 3.2.2. *The generalization won't go to zero for some reasonable algorithm (that try to minimize empirical error) as $n \to \infty$*

Proof. We construct such an algorithm. We assume $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, and ρ with dense support. The loss function $l(y, y) = 0$ for all $y \in \mathcal{Y}$. Given a dataset $(x_i, y_i)_{i=1}^n$ such that $x_i \neq x_j$ for all $i \neq j$, if we have $f_n : \mathcal{X} \to \mathcal{Y}$ such that:

$$f_n(x) = \begin{cases} y_i & \text{if } \exists i \in [n] : x_i = x \\ 0 & \text{otherwise} \end{cases}$$

This is clear that the algorithm above have $\mathbb{E}[\mathcal{E}_n(f_n)] = 0$ but $\mathbb{E}[\mathcal{E}(f_n)] = \mathcal{E}(0) \geq 0$. Thus, the generalization error won't go to zero as $n \to \infty$ □

Remark 70. The algorithm constructed is an extream form of memorization, which leads to *overfitting*.

Definition 3.2.2. (Overfitting) An estimator f_n is said to be overfit the training data if for any $n \in \mathbb{N}$:

- $\mathbb{E}[\mathcal{E}(f_n) - \mathcal{E}(f_*)] > C$ for constant $C > 0$
- $\mathbb{E}[\mathcal{E}_n(f_n) - \mathcal{E}(f_*)] \leq 0$

This is where the estimator f_n does better in "practice" than in the real data.

3.2.2 Bound For Generalization

Theorem 3.2.1. ***(Finite Hypothesis Case)*** *For finite $\mathcal{X}$ and $\mathcal{Y}$, we have a space of functions:*

$$\mathcal{F} = \mathcal{Y}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathcal{Y}\}$$

which is also finite, then:

$$\mathbb{E}\left[\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right|\right] \leq |\mathcal{F}|\sqrt{\frac{V_{\mathcal{F}}}{n}}$$

where $V_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \operatorname{var}(l(f(x_i), y))$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right|\right] &\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}}\left|\mathcal{E}_n(f) - \mathcal{E}(f)\right|\right] \\
&\leq \sum_{f \in \mathcal{F}}\left[\left|\mathcal{E}_n(f) - \mathcal{E}(f)\right|\right] \\
&\leq |\mathcal{F}|\sqrt{\frac{V_{\mathcal{F}}}{n}}
\end{aligned}$$

□

Remark 71. Empirical risk minimization still works in finite case as

$$\lim_{n \to \infty} \mathbb{E}\left[\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right|\right] = 0$$

Remark 72. This finite hypothesis case still works when considering the subset $\mathcal{H} \subset \mathcal{F}$ as we have (LHS) and if $f_* \in \mathcal{H}$, we can see that (RHS)

$$\mathbb{E}\left[\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right|\right] \leq |\mathcal{H}|\sqrt{\frac{V_{\mathcal{H}}}{n}} \quad , \quad \mathbb{E}\left[\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_*)\right|\right] \leq |\mathcal{H}|\sqrt{\frac{V_{\mathcal{H}}}{n}}$$

Definition 3.2.3. (Threshold Function) Threshold function of paramter $a \in (-1, 1]$ is

$$f_a(x) = \mathbf{1}_{x \in [a, \infty)}$$

Theorem 3.2.2. ***(Popoviciu's Inequality)*** *For any random varaible X bounded variance $m \leq \sigma^2 \leq M$*

$$\sigma^2 \leq \frac{(M - m)^2}{4}$$

Proof. Setting $g(t) = \mathbb{E}[(X-t)^2]$, then when doing the derivative, we can see that

$$g'(t) = 2t - 2\mathbb{E}[X]$$

when setting to zero, we can see that $t = \mathbb{E}[X]$, which is the minimum as $g''(t) = 2$. Now, setting $t = (M+m)/2$, we have

$$\begin{aligned} \text{var}(X) &\le \mathbb{E}\left[\left(X - \frac{M+m}{2}\right)^2\right] \\ &= \frac{1}{4}\mathbb{E}\Big[((X-m)+(X-M))^2\Big] \\ &\le \frac{1}{4}\mathbb{E}\Big[((X-m)-(X-M))^2\Big] = \frac{(M-m)^2}{4} \end{aligned}$$

□

Remark 73. We consider a binary classification problem $\mathcal{Y} = \{0,1\}$. We know in advanced that the minimizer would be a threshold with parameter a^*. It is clear that the hypothesis space is $\mathcal{F} = \{f_a | a \in \mathbb{R}\} = (-1, 1]$. However, computer can only represent a finite set of number(a), given a precision p, we have:

$$\mathcal{H}_p = \{f_a | a \in (-1,1], a10^p = [a10^p]\}$$

where $[\cdot]$ represents an integer part of the number. For example:

$$\mathcal{H}_1 = \left\{f_a : a \in \left\{-\frac{9}{10}, \cdots, \frac{9}{10}, 1\right\}\right\}$$

We can see that $|\mathcal{H}_p| = 2 \cdot 10^p$, and so we have

$$\mathbb{E}\left[\Big|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\Big|\right] \le |\mathcal{H}_p|\sqrt{\frac{V_\mathcal{H}}{n}} = \frac{10^p}{\sqrt{n}}$$

where the varaince is $V_\mathcal{H} \le 1/4$ via Popoviciu's inequality as our loss is bounded by $[0,1]$. The bound isn't good enough as we need a large n to make the bound being reasonable.

Remark 74. *(Chernoff Bounding Technique)* Given a random varaibel X and $\varepsilon > 0$, we have, for $t > 0$

$$\mathbb{P}(X \ge \varepsilon) = \mathbb{P}(\exp(tX) \ge \exp(t\varepsilon)) \le \frac{\mathbb{E}[\exp(tX)]}{\exp(t\varepsilon)}$$

where we apply the Markov's inequality and use t to make the bound tight.

Lemma 3.2.1. ***(Hoeffding's Lemma)*** *Let X be a random varaible with $\mathbb{E}[X] = 0$ and $a \le X \le b$ with $b > a$. For any $t > 0$, we have*

$$\mathbb{E}[\exp(tX)] \le \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

Theorem 3.2.3. ***(Hoeffding's Ineqality)*** *Consider $X_1, X_1, \dots, X_n$ independent random varaible where $X_i \in [a_i, b_i]$ and let $\bar{X} = 1/n \sum_{i=1}^n X_i$, then*

$$\mathbb{P}\left(\Big|\bar{X} - \mathbb{E}[\bar{X}]\Big| \ge \varepsilon\right) \le 2\exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. Since we have:

$$\mathbb{P}\left(\Big|\bar{X} - \mathbb{E}[\bar{X}]\Big| \ge \varepsilon\right) = 2\mathbb{P}\left(\bar{X} - \mathbb{E}[\bar{X}] \ge \varepsilon\right)$$

Please note that $\mathbb{E}[X_i - \mathbb{E}[X_i]] = 0$, thus we can use Hoeffding lemma, now we have:

$$\begin{aligned}
\mathbb{P}\left(\bar{X} - \mathbb{E}[\bar{X}] \ge \varepsilon\right) &\le \exp(-t\varepsilon)\mathbb{E}\left[\exp\left(\frac{t}{n}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right)\right)\right] \\
&= \exp(-t\varepsilon)\prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{t(X_i - \mathbb{E}[X_i])}{n}\right)\right] \\
&\le \exp(-t\varepsilon)\prod_{i=1}^n \exp\left(\frac{t^2}{8n^2}(b_i - a_i)^2\right) \\
&= \exp\left(\frac{t^2}{8n^2}\sum_{i=1}^n (b_i - a_i)^2 - t\varepsilon\right)
\end{aligned}$$

We will find t that would tighten the bound assuming setting $a = (\sum_{i=1}^n (b_i - a_i)^2)/(8n^2)$ and we have the following equation

$$f(t) = at^2 - t\varepsilon \qquad f'(t) = 2at - \varepsilon$$

which mean $t^* = \varepsilon/(2a)$ plugging back and we have $f(t^*) = -\varepsilon^2/(4a)$, and so:

$$\mathbb{P}\left(\bar{X} - \mathbb{E}[\bar{X}] \ge \varepsilon\right) \le \exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

as required. □

Theorem 3.2.4. *For any $\delta \in (0, 1]$ and bounded loss $0 \le |l(f(x), y)| < M$, for all $f \in \mathcal{H}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have:*

$$\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right| \le \sqrt{\frac{2M^2 \log(2|\mathcal{H}|/\delta)}{n}}$$

for probability of at least $1 - \delta$

Proof. Starting by applying Hoeffding's inequality, for any function f:

$$\mathbb{P}\left(\left|\mathcal{E}_n(f) - \mathcal{E}(f)\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{2n^2\varepsilon^2}{4M^2}\right)$$

Now, let's try to bound the generalization error:

$$\begin{aligned}
\mathbb{P}\left(\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right| \ge \varepsilon\right) &\le \mathbb{P}\left(\sup_{f\in\mathcal{H}}\left|\mathcal{E}_n(f) - \mathcal{E}(f)\right| \ge \varepsilon\right) = \mathbb{P}\left(\bigcup_{f\in\mathcal{H}}\left\{\left|\mathcal{E}_n(f) - \mathcal{E}(f)\right| \ge \varepsilon\right\}\right) \\
&\le \sum_{f\in\mathcal{H}}\mathbb{P}\left(\left|\mathcal{E}_n(f) - \mathcal{E}(f)\right| \ge \varepsilon\right) \le |\mathcal{H}|2\exp\left(-\frac{n^2\varepsilon^2}{2M^2}\right)
\end{aligned}$$

We have used union bound, since at least one of f will achieves a suprenum. To find the form above, we simply set δ to the bound we just derived. □

Remark 75. Recalling the threshold function, our new bound is as follows:

$$\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right| \le \sqrt{\frac{4 + 6p - 2\log\delta}{n}}$$

as $M = 1$ amd $\log 2|\mathcal{H}| = \log 4 \cdot 10^p = \log 4 + p\log 10 \le 2 + 3p$

Proposition 3.2.3. *Let X be a random variable such that $|X| < M$ for some constant $M > 0$, then for any $\varepsilon > 0$, we have*

$$\mathbb{E}[|X|] \le \varepsilon\mathbb{P}(|X| \le \varepsilon) + M\mathbb{P}(|X| > \varepsilon)$$

Proof. Let's consider the expectation of $|X|$, which we have:

$$\begin{aligned}\mathbb{E}[|X|] &= \int_{\varepsilon}^{\infty} p(X)|X| \text{ d}X + \int_{-\infty}^{-\varepsilon} p(X)|X| \text{ d}X + \int_{-\varepsilon}^{\varepsilon} p(X)|X| \text{ d}X \\ &\leq M(P(X > \varepsilon) + P(X < -\varepsilon)) + \varepsilon P(-\varepsilon \leq X \leq \varepsilon) \\ &= MP(|X| > \varepsilon) + \varepsilon P(|X| \leq \varepsilon)\end{aligned}$$

□

Corollary 3.2.1. *Using the proposition above and the generalization bound that we have derived, we have, for any* $\delta \in (0, 1]$*:*

$$\mathbb{E}\left[\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right|\right] \leq (1-\delta)\sqrt{\frac{2M^2 \log(2|\mathcal{H}|/\delta)}{n}} + \delta M$$

Remark 76. The case where $f_* \in \mathcal{H}\backslash\mathcal{H}_p$ for any $p > 0$, then ERM on $\mathcal{H}_p$ will never minimizes the expected risk and tere will be a gap between $\mathcal{E}(f_{n,p}) - \mathcal{E}(f_*)$. As $p \to \infty$, we expect the gap to decrease. However, if p increases too fast:

$$\left|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)\right| \leq \sqrt{\frac{4 + 6p - 2\log\delta}{n}} \to \infty$$

as we can't control the generalization error. We will need to increase p gradually. This process is called regularization.

Proposition 3.2.4. *The error decomposition of excess risk is*

$$\begin{aligned}\mathcal{E}(f_n) - \mathcal{E}(f_*) &= \underbrace{\mathcal{E}(f_n) - \mathcal{E}_n(f_n)}_{\text{Generalization Error}} + \underbrace{\mathcal{E}_n(f_n) - \mathcal{E}_n(f_p)}_{\leq 0} + \underbrace{\mathcal{E}_n(f_p) - \mathcal{E}(f_p)}_{\text{Generalization Error}} + \underbrace{\mathcal{E}(f_p) - \mathcal{E}(f_*)}_{\text{Approximation Error}} \\ &\leq \mathcal{E}(f_n) - \mathcal{E}_n(f_n) + \mathcal{E}_n(f_p) - \mathcal{E}(f_p) + \mathcal{E}(f_p) - \mathcal{E}(f_*)\end{aligned}$$

Lemma 3.2.2. *The approximation error of threshold function is*

$$\mathcal{E}(f_p) - \mathcal{E}(f_*) \leq |a_p - a_*| \leq 10^{-p}$$

Where we assume a distribution on $[-1, 1]$ *together with least square loss* $l = (y - f_a(x))^2$

Proof. We would like to note that, if $b \geq a$, $f_b(x)f_a(x) = f_b(x)$. WLOG, assume that $a_* \geq a_p$

$$\begin{aligned}\mathcal{E}(f_p) - \mathcal{E}(f_*) &= \int_{-1}^{1} (f_*(x) - f_p(x))^2 \text{ d}p(x) \\ &= \int_{-1}^{1} f_*^2(x) \text{ d}p(x) - \int_{-1}^{1} 2f_*(x)f_p(x) \text{ d}p(x) + \int_{-1}^{1} f_p^2(x) \text{ d}p(x) \\ &= \int_{a^*}^{1} p(x) \text{ d}x - 2\int_{a^*}^{1} p(x) \text{ d}x + \int_{a_p}^{1} p(x) \text{ d}x \\ &= \int_{a_p}^{1} p(x) \text{ d}x - \int_{a^*}^{1} p(x) \text{ d}x = \int_{a_p}^{a^*} p(x) \text{ d}x \leq |a^* - a_p|\end{aligned}$$

□

Remark 77. We can find the excess risk of threshold function to be bounded by, following proposition 3.2.4:

$$\mathcal{E}(f_n) - \mathcal{E}(f_*) \leq 2\sqrt{\frac{4 + 6p - 2\log\delta}{n}} + 10^{-p} = \phi(n, \delta, p)$$

This holds with probability greater than $1 - \delta$. We can shoow the precidion to be

$$p(n, \delta) = \underset{p \geq 0}{\arg\min}\ \phi(n, \delta, p)$$

Thus leading to error bound as $\varepsilon(n, \delta) = \phi(n, \delta, p(n, \delta))$.

3.2.3 Regularization

Remark 78. The idea of regularization, which has been discussed early in remark 76, is to parameterize $\mathcal{H}$ where $\mathcal{H} = \bigcup_{\gamma>0} \mathcal{H}_\gamma$ of hypothesis space, where $\mathcal{H}_\gamma \subset \mathcal{H}_{\gamma'}$ iff $\gamma \leq \gamma'$. We perform this to prevent overfitting as we called γ regularization parameter.

Definition 3.2.4. (Regularized Algorithm) Given n training points, the regularized algorithm returns $f_{\gamma,n}$ on $\mathcal{H}_\gamma$, while we let $\gamma = \gamma(n)$ as $n \to \infty$

Proposition 3.2.5. *We can decompose the excess risk as*

$$\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_*) = \underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)}_{\text{Sample Error}} + \underbrace{\mathcal{E}(f_\gamma) - \inf_{f\in\mathcal{H}} \mathcal{E}(f)}_{\text{Approximation Error}} + \underbrace{\inf_{f\in\mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_*)}_{\text{Irreducible Error}}$$

where we let $\gamma > 0$ and $f_\gamma = \arg\min_{f\in\mathcal{H}_\gamma} \mathcal{E}(f)$.

Remark 79. Let's explore the definition of each error:

- **Irreducible Error:** If the irreducible error is zero, then we call $\mathcal{H}$ universal.
- **Approximation Error:** This doesn't depend on the dataset, but it depends on ρ, and we call it bias.
- **Sample Error:** This random quantity depends on data. We can study it by capacity or stability.

We can show, under a mild assumption:

$$\lim_{\gamma\to\infty} \mathcal{E}(f_\gamma) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) = 0$$

Combining this with universal space: $\lim_{\gamma\to\infty} \mathcal{E}(f_\gamma) - \mathcal{E}(f_*) = 0$. Finally, we can have an approximation error to be bounded as:

$$\mathcal{E}(f_\gamma) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq \mathcal{A}(\rho, \gamma)$$

Please note that there will be no rate without any assumption, which is related to no-free launch theorem. If f_* is in Sobolev space $W^{S,2}$ then $\mathcal{A}(\rho,\gamma) = c\gamma^{-s}$

Proposition 3.2.6. *We can decompose the sample error to be:*

$$\begin{aligned}\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma) &= \underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}_n(f_{\gamma,n})}_{\text{Generalization Error}} + \underbrace{\mathcal{E}_n(f_{\gamma,n}) - \mathcal{E}_n(f_\gamma)}_{\text{Excess Empirical Risk } (\leq 0)} + \underbrace{\mathcal{E}_n(f_\gamma) - \mathcal{E}(f_\gamma)}_{\text{Generalization Error}} \\ &\leq \mathcal{E}(f_{\gamma,n}) - \mathcal{E}_n(f_{\gamma,n}) + \mathcal{E}_n(f_\gamma) - \mathcal{E}(f_\gamma)\end{aligned}$$

Remark 80. The generalization error can be controlled by study the empirical process of

$$\sup_{f\in\mathcal{H}_\gamma} |\mathcal{E}_n(f) - \mathcal{E}(f)|$$

as we have shown in theorem 3.2.4 (and union bound).

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}_\gamma} \left|\mathcal{E}_n(f) - \mathcal{E}(f)\right| \geq \varepsilon\right) \leq 2|\mathcal{H}| \exp\left(-\frac{n^2\varepsilon^2}{2M^2}\right)$$

However, it is hard to find empirical risk minimizer for arbitarty $\mathcal{H}_p$ as we need to calculate the expected risk. Good news, in some spaces, it might be easier to do such computation i.e convex or discretization in special dense hypothesis space.

Definition 3.2.5. (Infinite Norm of Function) Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space and $C(\mathcal{X})$ is a space of continous function, we define a norm

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$$

Proposition 3.2.7. *If the loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $l(\cdot, y)$ is uniformly L-Lipschitz. Then, we have*

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \le L \|f_1 - f_2\|_\infty \qquad |\mathcal{E}_n(f_1) - \mathcal{E}_n(f_2)| \le L \|f_1 - f_2\|_\infty$$

Proof. Starting with the first one, which we have:

$$\begin{aligned}
|\mathcal{E}(f_1) - \mathcal{E}(f_2)| &= \left| \int l(f_1(x), y) - l(f_2(x), y) \, \mathrm{d}\rho(x, y) \right| \\
&\le \int \left| l(f_1(x), y) - l(f_2(x), y) \right| \mathrm{d}\rho(x, y) \\
&\le L \int \left| f_1(x) - f_2(x) \right| \mathrm{d}\rho_{\mathcal{X}}(x) \\
&= L \|f_1 - f_2\|_{L^1(\mathcal{X}, \rho_{\mathcal{X}})} \le L \|f_1 - f_2\|_\infty
\end{aligned}$$

For the second one, we have

$$\begin{aligned}
|\mathcal{E}_n(f_1) - \mathcal{E}_n(f_2)| &= \frac{1}{n} \left| \sum_{i=1}^n l(f_1(x_i), y_i) - l(f_2(x_i, y_i)) \right| \\
&\le \frac{1}{n} \sum_{i=1}^n \left| l(f_1(x_i), y_i) - l(f_2(x_i, y_i)) \right| \\
&\le L \frac{1}{n} \sum_{i=1}^n |f_1(x) - f_2(x)| \le L \frac{1}{n} \sum_{i=1}^n \|f_1 - f_2\|_\infty = L \|f_1 - f_2\|_\infty
\end{aligned}$$

□

Remark 81. The function that are closed in $\|\cdot\|_\infty$ have similar expected and empirical risks.

Remark 82. If $\mathcal{H} \subset C(\mathcal{X})$ admits a finite discretization $\mathcal{H}_p = \{h_1, \dots, h_N\}$ with respecte to $\|\cdot\|_\infty$. Then, the generalization error can be controlled by:

$$\begin{aligned}
&\sup_{f \in \mathcal{H}} \left| \mathcal{E}_n(f) - \mathcal{E}(f) \right| \\
&\quad \le \sup_{f \in \mathcal{H}} \left| \mathcal{E}_n(f) - \mathcal{E}_n(h_f) \right| + \left| \mathcal{E}_n(h_f) - \mathcal{E}(h_f) \right| + \left| \mathcal{E}(h_f) - \mathcal{E}(f) \right| \\
&\quad \le 2L \|h_f - f\|_\infty + \sup_{h \in \mathcal{H}_p} \left| \mathcal{E}_n(h) - \mathcal{E}(h) \right|
\end{aligned}$$

where $h_f = \arg\min_{h \in \mathcal{H}_p} \|h - f\|_\infty$. Now, we will only have to control the $\sup_{h \in \mathcal{H}_p} |\mathcal{E}_n(h) - \mathcal{E}(h)|$ since $\mathcal{H}_p$ is finite.

Definition 3.2.6. (Covering Number) We define the covering number of $\mathcal{H}$ of radius $\eta > 0$ as the cardinality of minimal cover of $\mathcal{H}$ with ball of radius η:

$$\mathcal{N}(H, \eta) = \inf \left\{ m \,\middle|\, \mathcal{H} \subseteq \bigcup_{i=1}^m B_\eta(h_i), h_i \in \mathcal{H} \right\}$$

Theorem 3.2.5. *For any $\delta \in [0, 1)$ and $L > 0$ being Lipschitz constant of $l(\cdot, y)$, for all x, y and $|l(f(x), y)| < M$, we have:*

$$\sup_{f \in \mathcal{H}} \left| \mathcal{E}_n(f_n) - \mathcal{E}(f_n) \right| \le \sqrt{\frac{2M^2 \log(2\mathcal{N}(\mathcal{H}, n)/\delta)}{n}}$$

holds with probability $1 - \delta$, and where exists an $\eta(x)$ for which bounds tends to 0 as $n \to \infty$.

Remark 83. The proptotypical results i.e Bias/Variance tradeoff:

$$\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f^*) \leq \underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)}_{<\gamma^\beta n^{-\alpha}\text{(Variance)}} + \underbrace{\mathcal{E}(f_\gamma) - \mathcal{E}(f^*)}_{<\gamma^{-\tau}\text{(Bias)}}$$

We will have to choose $\gamma(n)$ to get best bias-variance tradeoff.

3.3 Tikhonov Regularization

3.3.1 Regularized Space

Definition 3.3.1. (Normed Regularized Space) Let $\mathcal{H}$ be a normed vector space of hypothesis. For $\gamma \geq 0$, we consider

$$\mathcal{H}_\gamma = \left\{ f \in \mathcal{H} \middle| \, \|f\|_\mathcal{H} \leq \gamma \right\}$$

As we have $\mathcal{H}_\gamma = B_\gamma(0) \subset \mathcal{H}$. The empirical risk minimization corresponds to:

$$f_{\gamma,n} = \underset{\|f\|_\mathcal{H} \leq \gamma}{\arg\min} \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$$

Remark 84. If $l(\cdot, y)$ is convex, then empirical risk minimization induces convex program, which we can find the solution in polynomal time.

Definition 3.3.2. (Space of Linear Function) We will focus $\mathcal{H}$ to be a space of linear function. Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$, where

$$\mathcal{H} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| \exists w \in \mathbb{R}^d \text{ s.t. } f(x) = w^T x, \forall x \in \mathbb{R}^d \right\}$$

We will set the norm to be $\|f\|_\mathcal{H} = \|w\|$ as w is the parameter corresponding to f. Thus, we have the empirical risk minimization to be:

$$w_{n,\gamma} = \underset{\|w\|_2 \leq \gamma}{\arg\min} \frac{1}{n} \sum_{i=1}^n l(x_i^T w, y_i)$$

where the empirical risk minimizer being $f_{n,\gamma} : \mathbb{R}^d \to \mathbb{R}$ is defined as $f_{n,\gamma}(x) = x^T w_{n,\gamma}$ for all $x \in \mathbb{R}^d$

Definition 3.3.3. (Non-Linear Function Extension) We expand the space of linear function to richer space of functions using the collection of non-linear function (feature extractor) $\psi_1, \dots, \psi_k : \mathbb{R}^d \to \mathbb{R}$ swhere:

$$\mathcal{H} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| \exists (w_i)_{i=1}^k \in \mathbb{R} \text{ s.t } f(x) = \sum_{i=1}^k \psi_i(x) w_i \ \forall x \in \mathbb{R}^d \right\}$$

we will consider $\|f\|_\mathcal{H} = \|w\|_2$ where $w \in \mathbb{R}^k$. Furthermore, we can construct a non-linear map $\Psi : \mathbb{R}^d \to \mathbb{R}^k$ where $\Psi(x) = (\psi_1(x), \dots, \psi_k(x))$.

Theorem 3.3.1. *The covering number of $\mathcal{H}_\gamma$ is:*

$$\mathcal{N}(\mathcal{H}_\gamma, n) \leq \left(\frac{4\gamma}{\eta} \right)^d$$

for all $\eta > 0$

Proof. For any $\gamma \geq 0$ and $B_\gamma(0) \subset \mathbb{R}^d$, which is a ball of radius γ centered in 0. Then for all $\eta > 0$:

$$\mathcal{N}(B_\gamma(0), \eta) \leq \left(\frac{4\gamma}{\eta}\right)^d$$

But since $\mathcal{H}$, is isomorphic to $\mathbb{R}^d$, we have the sampe covering number. □

Definition 3.3.4. (Tikhonov Regualrization Problem) We define the Tikhonov Regualrization problem to be, instead of constrained optimization problem.

$$w_{\lambda,n} = \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(x_i^T w, y_i) + \lambda \|w\|_{\mathcal{H}}^2$$

We can show that this problem and problem in definition 3.3.2 are the same as we can find $\lambda(\gamma)$ such that $w_{n,\gamma} = w_{\lambda(\gamma),n}$.

Definition 3.3.5. The directional derivative is defined by:

$$\nabla_v f(x) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}$$

Lemma 3.3.1. $\nabla_v f(x) = v^T \nabla f(x)$

Proof. One can use a Taylor's expansion to proof this, but we are going derive it via chain rule. We will prove in 2D but this can be extended easily. Let's define a single variable function $g(t) = f(x + at, y + bt)$. Let's consider $g'(0)$

$$\begin{aligned} g'(t) &= \lim_{h \to 0} \frac{g(t+h) - g(t)}{h} \\ \iff g'(0) &= \lim_{h \to 0} \frac{g(h) - g(0)}{h} \\ &= \lim_{h \to 0} \frac{f(x + ah, y + bh) - f(x, y)}{h} = \nabla_v g(x) \end{aligned}$$

where $v = (a, b)$. Now, we can apply the chain rule, which gives us

$$\nabla_v g(x) = g'(0) = \frac{\partial g}{\partial x}\frac{dx}{dt} + \frac{\partial g}{\partial y}\frac{dy}{dt} = \frac{\partial g}{\partial x} a + \frac{\partial g}{\partial y} b = v^T \nabla g(x)$$

□

3.3.2 Introduction to Convex + Finding Weights

Theorem 3.3.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and S be a convex subset of $\mathbb{R}^n$. Then f is convex iff*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

for all $y, x \in \mathbb{R}$

Proof. ($\implies$) If f is convex. Then we have, by convexity:

$$\begin{aligned} & f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda) f(x) \\ \iff & \frac{f(\lambda y + (1-\lambda)x) - f(x)}{\lambda} = \frac{f(x - (y-x)\lambda) - f(x)}{\lambda} \leq f(y) - f(x) \end{aligned}$$

Then, by setting $\lambda \to 0$, we have

$$\lim_{\lambda \to 0} \frac{f(x - (y-x)\lambda) - f(x)}{\lambda} = \nabla f(x)^T (y - x) \leq f(y) - f(x)$$

By the definiton of directional derivative.

($\Longleftarrow$) We consider 2 points, where we set $z = \lambda y + (1-\lambda)x$:

$$f(y) \geq f(z) + \nabla f(z)^T(y-z) \qquad f(x) \geq f(z) + \nabla f(z)^T(x-z)$$

Then we have:

$$\begin{aligned}
\lambda f(y) + (1-\lambda)f(x) &\geq f(z) + \lambda \nabla f(z)^T(y-z) + (1-\lambda)\nabla f(z)^T(x-z) \\
&= f(z) + \nabla f(z)^T\Big[\lambda(y-z) + (1-\lambda)(x-z)\Big] \\
&= f(z) + \nabla f(z)^T\Big[\lambda y - \lambda^2 y - x\lambda + \lambda^2 x + x - \lambda x - \lambda y + \lambda^2 y - x + 2\lambda x - \lambda^2 x\Big] \\
&= f(z) = f(\lambda y + (1-\lambda)x)
\end{aligned}$$

Thus complete the proof. □

Theorem 3.3.3. *Any differentiable convex function $F : \mathbb{R}^d \to \mathbb{R}$ where $w_* \in \mathbb{R}^d$ is global optimizer iff $\nabla f(w_*) = 0$*

Proof. ($\Longrightarrow$) As the directional derivative measures the rate in which the function grows, we want to find the direction that decrease f the most. It is clear from the dot production that this would be $-\nabla f(x)$. Thus, if $\nabla f(w_*) \neq 0$, then for some $\varepsilon \in \mathbb{R}$, $f(w_* - \varepsilon \nabla f(x)) \leq f(w_*)$, thus contradicts the assumption that w_* is global optimizer.

($\Longleftarrow$) We will show that if $\nabla f(w_*) = 0$ then w_* is global optimizer. Following the theorem 3.3.2, we can see that for all y, we have

$$\begin{aligned}
f(y) &\geq f(w_*) + \nabla f(w_*)^T(y - w_*) \\
&= f(w_*)
\end{aligned}$$

Thus complete the proof. □

Proposition 3.3.1. *If we set $l(f(x), y) = (y - f(x))^2$ then:*

$$\begin{aligned}
w_{\lambda,n} &= \underset{w\in\mathbb{R}^d}{\arg\min}\ \|y - Xw\|_2^2 + n\lambda\|w\|_2^2 \\
&= (X^TX + n\lambda I)^{-1}X^Ty
\end{aligned}$$

where $y \in \mathbb{R}^n$ is a collection of labels, while $\mathbb{R}^{n\times d}$ is the collection of data.

Proof. Since the objective is convex (norm is convex and addition + multiplcation of positive number), we can find the global minima according to theorem 3.3.3 by finding the derivative and set to 0, which we have:

$$\begin{aligned}
&\nabla\Big[\ \|y - Xw\|_2^2 + n\lambda\|w\|_2^2\ \Big] = 2X^TXw - 2X^Ty + 2n\lambda w = 0 \\
\iff & w = (X^TX + n\lambda I)^{-1}X^Ty
\end{aligned}$$

Thus complete the proof. □

Remark 85. The total cost of solving the regression is $\mathcal{O}(nd^2 + d^2)$ and if $d > n$, then the complexity becomes $\mathcal{O}(d^3)$. However, if we use a representor's theorem, then we are able to have $\mathcal{O}(n^3)$.

3.3.3 Gradient Descent

Definition 3.3.6. (Gradient of Weight) In general if $l(\cdot, y) : \mathbb{R} \to \mathbb{R}$ is differentiable, for any $y \in \mathcal{Y}$, then we have:

$$\nabla(\mathcal{E}_n(w) + \lambda\|w\|_2^2) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}l(x_i^Tw, y_i) + 2\lambda w$$

We can solve the minima by setting the above equation to zero.

Remark 86. In most cases, we aren't able to solve the gradient equation analytically, so we need iterated descent optimization, which provided us with $(w^{(k)})_{k\in\mathbb{N}}$ that converges to global minimizer.

Definition 3.3.7. (Gradient Descent Algorithm) Let $F:\mathbb{R}^d\to\mathbb{R}$ be differentiable. Set $w^{(0)}\in\mathbb{R}^d$. For any $k\in\mathbb{N}$, we define $w^{(t+1)}\in\mathbb{R}^d$ as:

$$w_{k+1}=w_k-\gamma\nabla F(w_k)$$

where $\gamma>0$ represents the step size of the descent.

Definition 3.3.8. (Lipschitz Gradient) A function f with Lipschitz gradient with constant L is where, for all $x,y\in\operatorname{dom}(f)$:

$$\|\nabla f(x)-\nabla f(y)\|\le L\,\|x-y\|$$

Lemma 3.3.2. *For 2 points $x,y\in\mathbb{R}^d$ and function $f:\mathbb{R}^d\to\mathbb{R}$ with Lipschitz gradient with constant L, then:*

$$f(y)\le f(x)+\nabla f(x)^T(y-x)+\frac{L}{2}\|y-x\|^2$$

Proof. We consider the function $g(t)=f(x+t(y-x))$ and; therefore, $h'(t)=\langle\nabla f(x+t(y-x)),y-x\rangle$. Following from fundamental theorem of calculus

$$h(1)-h(0)=\int_0^1 h'(t)\ \mathrm{d}t$$

as we have $h(1)=f(y)$ and $h(0)=f(x)$:

$$\begin{aligned}
f(y)&=f(x)+\int_0^1\langle\nabla f(x+t(y-x)),y-x\rangle\ \mathrm{d}t\\
&=f(x)+\langle\nabla f(x),y-x\rangle+\int_0^1\langle\nabla f(x+t(y-x))-\nabla f(x),y-x\rangle\ \mathrm{d}t\\
&\le f(x)+\langle\nabla f(x),y-x\rangle+\int_0^1\|\nabla f(x+t(y-x))-\nabla f(x)\|\cdot\|y-x\|\ \mathrm{d}t\\
&\le f(x)+\langle\nabla f(x),y-x\rangle+L\,\|y-x\|\int_0^1\|t(y-x)\|\cdot\ \mathrm{d}t\\
&=f(x)+\langle\nabla f(x),y-x\rangle+L\,\|y-x\|^2\int_0^1 t\cdot\ \mathrm{d}t\\
&=f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|y-x\|^2
\end{aligned}$$

□

Proposition 3.3.2. *Given the gradient descent algorithm, with update weight of γ:*

$$\left(\frac{1}{\gamma}-\frac{L}{2}\right)\|x_k-x_{k+1}\|^2\le f(x_k)-f(x_{k+1})$$

for all $\gamma>0$

Proof. Using the result from lemma above and definition of gradient descent

$$\begin{aligned}
f(x_{k+1})&\le f(x_k)+\frac{1}{\gamma}\langle\gamma\nabla f(x_k),x_{k+1}-x_k\rangle+\frac{L}{2}\|x_{k+1}-x_k\|^2\\
&=f(x_k)-\frac{1}{\gamma}\|x_{k+1}-x_k\|^2+\frac{L}{2}\|x_{k+1}-x_k\|^2\\
&=f(x_k)-\left(\frac{1}{\gamma}-\frac{L}{2}\right)\|x_{k+1}-x_k\|^2
\end{aligned}$$

Rearrange and we finish the proof. □

Remark 87. We can see that the evaluation $(f(x_k))_{k=1}$ is decreasing iff $\gamma \le 2L$, as the norm is positive.

Lemma 3.3.3. *For convex function f, given that $\gamma \le 2/L$:*

$$\sum_{i=0}^{\infty} \|x_i - x_{i+1}\|^2 \le \frac{2\gamma}{2-\gamma L}\Big(f(x_0) - \min_x f(x)\Big)$$

Proof. We can perform the telescoping sum, assuming that the evaluation of convex function is decreasing, thus having:

$$\begin{aligned}\sum_{i=0}^{\infty} \|x_i - x_{i+1}\|^2 &\le \left(\frac{2\gamma}{2-\gamma L}\right)\sum_{i=0}^{\infty} f(x_i) - f(x_{i+1}) \\ &= \frac{2\gamma}{2-\gamma L}\Big(f(x_0) - \min_x f(x)\Big)\end{aligned}$$

Please note that since f is convex, the minima is global. □

Proposition 3.3.3. *For all $x \in \text{dom}(f)$, we have*

$$2\gamma\Big(f(x_{k+1}) - f(x)\Big) \le \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + (\gamma L - 1)\|x_{k+1} - x_k\|^2$$

Proof. From proposition 3.3.2:

$$\begin{aligned}2\gamma\Big(f(x_{k+1}) - f(x_k)\Big) &\le 2\gamma(f(x_k) - f(x)) - (2-\gamma L)\|x_{k+1} - x_k\|^2 \\ &= 2\gamma(\nabla f(x_k)^T(x_k - x)) - (2-\gamma L)\|x_{k+1} - x_k\|^2 \\ &\le 2\Big((x_k - x_{k+1})^T(x_k - x)\Big) - (2-\gamma L)\|x_{k+1} - x_k\|^2 \\ &= \|x_k - x_{k+1}\|^2 + \|x_k - x\|^2 - \|x - x_{k+1}\|^2 - (2-\gamma L)\|x_{k+1} - x_k\|^2 \\ &= \|x_k - x\|^2 - \|x - x_{k+1}\|^2 - (\gamma L - 1)\|x_{k+1} - x_k\|^2\end{aligned}$$

The first equality comes from $x_k - x_{k-1} = \gamma\nabla f(x_k)$ The second ineqality comes from lemma 3.3.2, where we set $x = x_k$ and $y = x$. The second equality comes from:

$$2u^Tv = \|u\|^2 + \|v\|^2 - \|u-v\|^2$$

□

Theorem 3.3.4. *Suppose that $x_* = \arg\min_x f(x)$ (and it exists) and $\gamma < 2/L$ then, for all $k > 1$:*

$$f(x_k) - \min_x f(x) \le \frac{1}{k}\left[\frac{\|x_0 - x_*\|^2}{2\gamma} + \frac{(\gamma L - 1)_+}{2-\gamma L}\Big(f(x_0) - \min_x f(x)\Big)\right]$$

Proof. We recall proposition 3.3.3, we we set $x = x_*$, which we have:

$$\begin{aligned}\sum_{i=0}^{n}\Big(f(x_{i+1}) - f(x_*)\Big) &\le \frac{1}{2\gamma}\sum_{i=0}^{n}\Big(\|x_i - x_*\|^2 - \|x_{i+1} - x_*\|^2 + (\gamma L - 1)\|x_{i+1} - x_i\|^2\Big) \\ &= \frac{1}{2\gamma}\sum_{i=0}^{n}\Big(\|x_i - x_*\|^2 - \|x_{i+1} - x_*\|^2\Big) + \frac{(\gamma L - 1)_+}{2\gamma}\sum_{i=1}^{n}\|x_{i+1} - x_i\|^2 \\ &\le \frac{1}{2\gamma}\sum_{i=0}^{n}\Big(\|x_i - x_*\|^2 - \|x_{i+1} - x_*\|^2\Big) + \frac{(\gamma L - 1)_+}{2-\gamma L}\sum_{i=1}^{n}\Big(f(x_0) - \min_x f(x)\Big) \\ &= \frac{1}{2\gamma}\Big(\|x_0 - x_*\|^2 - \|x_n - x_*\|^2\Big) + \frac{(\gamma L - 1)_+}{2-\gamma L}\sum_{i=1}^{n}\Big(f(x_0) - \min_x f(x)\Big) \\ &= \frac{\|x_0 - x_*\|^2}{2\gamma} + \frac{(\gamma L - 1)_+}{2-\gamma L}\sum_{i=1}^{n}\Big(f(x_0) - \min_x f(x)\Big)\end{aligned}$$

We use the lemma 3.3.3. For the last equality, we have, a telescoping sum and $\|x_n - x_*\|^2 \geq 0$. Now we can see that

$$\sum_{i=0}^{n} \Big(f(x_{i+1}) - f(x_*)\Big) \geq k\Big(f(x_{i+1}) - f(x_*)\Big)$$

as we shown in lemma 3.3.3 that the evaluation will keep decreasing. Rearrange and we finish the proof. □

Corollary 3.3.1. *It is clear that the best value of γ is $1/L$ and so, the rate in which the gradient descent is:*

$$f(x_k) - \min_x f(x) \leq \frac{L}{2k} \|x_0 - x_*\|^2$$

Definition 3.3.9. (Strongly Convex) The function f is strongly convex with modulus $\mu > 0$ if, for all $x, y \in \text{dom}(f)$:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2$$

Proposition 3.3.4. *For all $x \in \text{dom}(f)$ with f being μ-strongly convex:*

$$f(x) - \min_x f(x) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$$

Proof. We start off by recalling strongly convex function, and minimize both side of ineqalities:

$$\begin{aligned}
\min_y f(y) &\geq \min_y \left(f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2\right) \\
&\geq f(x) + \frac{1}{2\mu} \min_y \left(2\nabla f(x)^T (\mu(y - x)) + \|\mu(y - x)\|^2\right) \\
&= f(x) + \frac{1}{2\mu} \min_y \left(\|\nabla f(x)\|^2 + 2\|\mu(y - x)\|^2 - \|\nabla f(x) - \mu(y - x)\|\right) \\
&= f(x) + \frac{1}{2\mu} \min_y \left(\|\nabla f(x) + \mu(y - x)\|^2 - \|\nabla f\|^2\right) \\
&\geq f(x) + \frac{1}{2\mu} \|\nabla f(x)\|^2
\end{aligned}$$

The last equality can be show as: suppose $a = \mu(y - x)$ and $b = \nabla f(x)$, we have:

$$\begin{aligned}
\|a\|^2 + 2\|b\|^2 - \|a - b\| &= 2a^T a + b^T b - \Big[a^T a - 2a^T b + b^T b\Big] \\
&= a^T a + 2a^T b + b^T b - b^T b = \|a + b\|^2 - \|b\|^2
\end{aligned}$$

Regarrange and we finish the proof. □

Remark 88. From the definition of strongly convex, we can see that

$$\begin{aligned}
f(y) &\geq f(x_*) + \nabla f(x_*)^T (y - x_*) + \frac{\mu}{2} \|y - x_*\|^2 \\
&= f(x_*) + \frac{\mu}{2} \|y - x_*\|^2
\end{aligned}$$

where $x_* = \arg\min_x f(x)$.

Theorem 3.3.5. *For μ-strongly convex function with $\gamma < 2/L$, we have:*

$$f(x_k) - \min_x f(x) \leq \Big(1 - \gamma\mu(2 - \gamma L)\Big)^k \Big(f(x_0) - \min_x f(x)\Big)$$

Proof. First, we will show that

$$f(x_{k+1}) - \min_x f(x) \leq \Big(1 - \gamma\mu(2 - \gamma L)\Big)\Big(f(x_k) - \min_x f(x)\Big)$$

Following proposition3.3.2, we have the following ineqalities:

$$\begin{aligned}
f(x_{k+1}) - \min_x f(x) &\leq f(x_k) - \left(\frac{2-\gamma L}{2\gamma}\right)\|x_k - x_{k-1}\|^2 + \min_x f(x) \\
&= f(x_k) - \left(2\mu\gamma - \gamma^2 L\mu\right)\|\nabla F(x_k)\|^2 - \min_x f(x) \\
&\leq f(x_k) - \min_x f(x) - \left(2\mu\gamma - \gamma^2 L\mu\right)\Big(f(x_k) - \min_x f(x)\Big) \\
&= \left(1 - \left(2\mu\gamma - \gamma^2 L\mu\right)\right)\Big(f(x_k) - \min_x f(x)\Big)
\end{aligned}$$

And so, by repeating the ineqalities, we have the exponential as required. □

Remark 89. For the best value of γ, we should have $\gamma = 2/(\mu + L)$

Definition 3.3.10. (Projected Gradient) The problem such as Tikhonov regularization can be solved using projected gradient descent:

$$w_{k+1} = \Pi_{\mathcal{H}_\gamma}\Big(w_k - \gamma \nabla F(w_k)\Big)$$

where $\Pi_{\mathcal{H}_\gamma} : \mathbb{R}^d \to \mathbb{R}^d$ dentoes the Euclidian projection onto $\mathcal{H}_\gamma$ as

$$\Pi_{\mathcal{H}_\gamma}(w) = \underset{w' \in \mathcal{H}_\gamma}{\arg\min}\ \|w - w'\|_2^2 = \gamma \frac{w}{\|w\|_2}$$

Lemma 3.3.4. *For point $y \in \mathbb{R}$ and $x \in \Omega$:*

$$(y - \Pi_\Omega(y))^T(x - \Pi_\Omega(y)) \leq 0$$

Lemma 3.3.5. *Given the projected gradient descent algorithm, with the update weight of γ:*

$$f(x_k) - f(x_{k+1}) \geq \left(\frac{1}{\gamma} - \frac{L}{2}\right)\|x_{k+1} - x_k\|^2$$

Proof. From lemma we have:

$$\left(x_k - \gamma\nabla f(x_k) - x_{k+1}\right)^T\left(x_k - x_{k+1}\right) \leq 0$$

which implies that

$$\nabla F(x_k)^T(x_{k+1} - x_k) \leq \frac{1}{\gamma}\|x_k - x_{k+1}\|$$

Therefore:

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq f(x_k) + \left(\frac{L}{2} - \frac{1}{\gamma}\right)\|x_{k+1} - x_k\|^2
\end{aligned}$$

By rearranging, we got the statement above. □

Theorem 3.3.6. *The convergence rate of projected gradient is the same as normal gradient descent.*

Remark 90. The gradient step of Tikhonov's regularization:

$$w_{k+1} = w_k - \gamma(X^T X + \lambda I)w_k + \gamma X^T y$$

has the total time complexity as $\mathcal{O}((k+n)d^2)$ operations for k steps. To achieve the same excess risk as ERM, we will need a total time complexity of $\mathcal{O}(nd^2)$.

Proposition 3.3.5. *We can decompose the sample error of the estimator after k iterations:*

$$\begin{aligned}\mathcal{E}(w_k) - \mathcal{E}(w_\gamma) &= \mathcal{E}(w_k) - \mathcal{E}_n(w_k) + \mathcal{E}_n(w_k) - \mathcal{E}_n(w_{\gamma,n}) + \underbrace{\mathcal{E}_n(w_{\gamma,n}) - \mathcal{E}_n(w_\gamma)}_{\leq 0} + \mathcal{E}_n(w_\gamma) - \mathcal{E}(w_\gamma) \\ &\leq \underbrace{\mathcal{E}(w_k) - \mathcal{E}_n(w_k)}_{\text{Sample Error on } \mathcal{H}_\gamma} + \underbrace{\mathcal{E}_n(w_k) - \mathcal{E}_n(w_{\gamma,n})}_{\text{Optimization Error}} + \underbrace{\mathcal{E}_n(w_\gamma) - \mathcal{E}(w_\gamma)}_{\text{Sample Error on } \mathcal{H}_\gamma}\end{aligned}$$

Remark 91. Since we know the generalization error, we can control the optimization error to match this i.e if the generalization error is $\varepsilon(n, \gamma, \delta)$ with probabilistic no less than $1 - \delta$, then we have to perfrom

$$k = \mathcal{O}\left(\frac{1}{\varepsilon(n, \gamma, \delta)}\right)$$

To get the same accurary as empirical risk minimization.

3.3.4 Stability

Definition 3.3.11. (Modified Set) Let Z be a set, for any set $S = \{z_1, \ldots, z_n\} \in Z^n$ for any $z \in Z$ and $i = 1, \ldots, n$ we denote

$$S^{i,z} = \{z_1, \ldots, z_{i-1}, z, z_{i+1}, \ldots, z_n\} \in Z^n$$

Definition 3.3.12. (Uniformed Stability) We denote a dataset $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and for any $f : \mathcal{X} \to \mathcal{Y}$, we denote $l(f, z) = l(f(x), y)$. For an algorithm $\mathcal{A}$ and any dataset $S = (z_i)_{i=1}^n$, we write $f_S = \mathcal{A}(S)$. The algorithm $\mathcal{A}$ is $\beta(n)$-stable with $n \in \mathbb{N}$ and $\beta(n) > 0$, if for all $S \in \mathcal{Z}^n$, $z \in \mathcal{Z}$ and $i = 1, \ldots, n$:

$$\sup_{\bar{z} \in \mathcal{Z}} |l(f_S, \bar{z}) - l(f_{S^{i,z}}, \bar{z})| \leq \beta(n)$$

Theorem 3.3.7. *Let $\mathcal{A}$ be uniform $\beta(n)$-stable algorithm. For any dataset $S \in \mathcal{Z}^n$, define $f_S = \mathcal{A}(s)$, then*

$$|\mathbb{E}_{S\sim\rho^n}[\mathcal{E}(f_S) - \mathcal{E}_n(f_S)]| \leq \beta(n)$$

This means that we can directly control the generalization error with stablility of an algorithm.

Proof. Starting with the empirical risk:

$$\begin{aligned}\mathbb{E}_S[\mathcal{E}_n(f_S)] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n l(f_S, z_i)\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_S[l(f_S, z_i)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_S\mathbb{E}_{z_i'}[l(f_S, z_i)] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_S\mathbb{E}_{z_i'}[l(f_{S^{i,z_i'}}, z_i')] = \mathbb{E}_S\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n l(f_{S^{i,z_i'}}, z_i')\right]\end{aligned}$$

For the expected risk, we have

$$\mathbb{E}_S[\mathcal{E}(f_S)] = \mathbb{E}_S\mathbb{E}_{S'}[l(f_S, z')]] = \mathbb{E}_S\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n l(f_S, z_i')\right]$$

Let's consider the differences:

$$\begin{aligned}|\mathbb{E}_{S\sim\rho^n}[\mathcal{E}(f_S) - \mathcal{E}_n(f_S)]| &= \left|\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n l(f_{S^{i,z_i'}}, z_i') - \frac{1}{n}\sum_{i=1}^n l(f_S, z_i')\right]\right| \\ &\leq \mathbb{E}_{S'}\frac{1}{n}\sum_{i=1}^n \left|l(f_{S^{i,z_i'}}, z_i') - l(f_S, z_i')\right| \leq \beta(n)\end{aligned}$$

□

Lemma 3.3.6. *The norm $\|\cdot\|_{\mathcal{H}}$ of RKHS $\mathcal{H}$ is strongly convex i.e for any $g, h \in \mathcal{H}$ and $\theta \in [0,1]$, we have:*

$$\|\theta g + (1-\theta)h\|_{\mathcal{H}}^2 < \theta \|g\|_{\mathcal{H}}^2 + (1-\theta)\|h\|_{\mathcal{H}}^2$$

Proof. We consider expanding the norm, and then find the differences between the left hand side and the right hand side:

$$\begin{aligned}(\theta g + (1-\theta)h)^T(\theta g + (1-\theta)h) &= \theta^2 g^T g + 2\theta(1-\theta)g^T h + (1-\theta)(1-\theta)h^T h \\ &= \theta^2 g^T g + 2\theta(1-\theta)g^T h + h^T h - 2\theta h^T h + \theta^2 h^T h\end{aligned}$$

Now we will minus it with $\theta g^T g + (1-\theta)h^T h$, which gives us:

$$\begin{aligned}\theta^2 g^T g + 2\theta(1-\theta)g^T h + h^T h - 2\theta h^T h + \theta^2 h^T h - \theta g^T g - (1-\theta)h^T h& \\ = \theta(\theta-1)g^T g + 2\theta(1-\theta)g^T h + \theta(1-\theta)h^T h& \\ = \theta(\theta-1)\|g-h\|_{\mathcal{H}}^2&\end{aligned}$$

Since $\theta < 1$, the inequality holds. □

Lemma 3.3.7. *For any convex function $F' : \mathcal{H} \to \mathbb{R}$ and $F(\cdot) = F'(\cdot) + \lambda\|\cdot\|$. Given the minimizer $f = \arg\min_{f' \in \mathcal{H}} F(f')$, then for some $g \in \mathcal{H}$:*

$$F(g) - F(f) \geq \frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2$$

Proof. By definition of F, we can see that:

$$\begin{aligned}&F(\theta f + (1-\theta)g) \leq \theta F(f) + (1-\theta)F(g) - \lambda\theta(1-\theta)\|f-g\|_{\mathcal{H}}^2 \\ \iff& 2F\left(\frac{f+g}{2}\right) \leq F(f) + F(g) - \frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2 \\ \iff& F(g) - F(f) \geq 2F\left(\frac{f+g}{2}\right) + \frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2 - 2F(f) \\ &\geq \frac{\lambda}{2}\|f-g\|_{\mathcal{H}}^2\end{aligned}$$

Thus complete the proof. □

Theorem 3.3.8. *Let $\mathcal{H}$ be RKHS with associated kernel $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. We can show that for any $S \in \mathcal{Z}^n$, $z' \in \mathcal{H}$ and $i = 1, \dots, i$:*

$$\sup_{z \in \mathcal{Z}} \left| l(f_S, z) - l(f_{S^{i,z'}}, z') \right| \leq \frac{2L^2 k^2}{n\lambda}$$

where $L > 0$ is Lipschitz constant of $l(\cdot, y)$ and $k^2 = \sup_{x \in \mathcal{X}} K(x, x)$

Proof. We consider the following functions:

$$F_1(\cdot) = \mathcal{E}_S(\cdot) + \lambda\|\cdot\|_{\mathcal{H}}^2 \qquad F_2(\cdot) = \mathcal{E}_{S^{i,z'}}(\cdot) + \lambda\|\cdot\|_{\mathcal{H}}^2$$

We will simply the notation $f_1 = f_S$ and $f_2 = f_{S^{i,z'}}$ and by definition, we have:

$$f_1 = \operatorname*{arg\,min}_{f \in \mathcal{H}} F_1(f) \qquad f_2 = \operatorname*{arg\,min}_{f \in \mathcal{H}} F_2(f)$$

Using the lemma above:

$$F_1(f_2) - F_1(f_1) \geq \frac{\lambda}{2}\|f_1 - f_2\|_{\mathcal{H}}^2 \qquad F_2(f_1) - F_2(f_2) \geq \frac{\lambda}{2}\|f_2 - f_1\|_{\mathcal{H}}^2$$

Summing them yields:

$$\begin{aligned}
\lambda \|f_1 - f_2\|_{\mathcal{H}}^2 &\le F_1(f_2) - F_1(f_1) + F_2(f_1) - F_2(f_2) \\
&= \mathcal{E}_S(f_2) - \mathcal{E}_{S^{i,z'}}(f_2) + \mathcal{E}_{S^{i,z'}}(f_1) - \mathcal{E}_S(f_1) \\
&= \frac{1}{n} l(f_2, z_i) - l(f_1, z_i) + l(f_1, z_i') - l(f_2, z_i') \\
&\le \frac{2}{n} \sup_z \Big| l(f_1, z) - l(f_2, z) \Big|
\end{aligned}$$

We can see that l is Lipschitz:

$$\begin{aligned}
\sup_z \Big| l(f_1, z) - l(f_2, z) \Big| &= \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \Big| l(f_1(x), y) - l(f_2(x), y) \Big| \\
&\le L \sup_{x \in \mathcal{X}} \Big| f_1(x) - f_2(x) \Big| \\
&\le Lk \|f_1 - f_2\|_{\mathcal{H}}
\end{aligned}$$

The last equality comes from the fact that $|f(x)| \le \sqrt{k(x,x)} \|f\|_{\mathcal{H}}^2$. Thus, we have

$$\|f_1 - f_2\|_{\mathcal{H}}^2 \le \frac{2Lk}{n\lambda}$$

Plugging this back and we yields the ineqality above. □

Theorem 3.3.9. *The excess risk for Tikhonov regularization is*

$$\mathbb{E}\Big[\mathcal{E}(f_S) - \mathcal{E}(f_*)\Big] \le \mathcal{O}\left(n^{-\frac{s}{s+1}}\right)$$

Proof. We will define $f_\lambda = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2$, and define the following excess risk decomposition:

$$\mathcal{E}(f_S) - \mathcal{E}(f_*) = \mathcal{E}(f_S) - \mathcal{E}_S(f_S) + \mathcal{E}_S(f_S) - \mathcal{E}_S(f_\lambda) + \mathcal{E}_S(f_\lambda) - \mathcal{E}(f_*) + \lambda \|f_\lambda\|_{\mathcal{H}}^2 - \lambda \|f_\lambda\|_{\mathcal{H}}^2$$

Please note that

- $\mathcal{E}(f_S) - \mathcal{E}(f_*) \le \mathcal{E}(f_S) - \mathcal{E}(f_*) + \lambda \|f_S\|_\lambda^2$
- f_S is the minimizer of empirical risk, which means:
$$\mathcal{E}_S(f_S) + \lambda \|f_S\|_{\mathcal{H}}^2 - \mathcal{E}_S(f_\lambda) - \lambda \|f_\lambda\|_{\mathcal{H}}^2 \le 0$$
- $\mathbb{E}_S[\mathcal{E}_S(f_\lambda)] = \mathcal{E}(f_\lambda)$

And, so we have

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(f_S) - \mathcal{E}(f_*)] &\le \mathbb{E}\Big[\mathcal{E}(f_S) - \mathcal{E}_S(f_S) + \mathcal{E}_S(f_S) - \mathcal{E}_S(f_\lambda) + \mathcal{E}_S(f_\lambda) - \mathcal{E}(f_*) + \lambda \|f_\lambda\|_{\mathcal{H}}^2 - \lambda \|f_\lambda\|_{\mathcal{H}}^2 + \lambda \|f_S\|_\lambda^2\Big] \\
&= \mathbb{E}\Big[\mathcal{E}(f_S) - \mathcal{E}_S(f_S) + \underbrace{\mathcal{E}_S(f_S) + \lambda \|f_S\|_\lambda^2 - \mathcal{E}_S(f_\lambda) - \lambda \|f_\lambda\|_{\mathcal{H}}^2}_{\le 0} + \mathcal{E}_S(f_\lambda) - \mathcal{E}(f_*) + \lambda \|f_\lambda\|_{\mathcal{H}}^2\Big] \\
&\le \mathbb{E}\Big[\mathcal{E}(f_S) - \mathcal{E}_S(f_S) + \mathcal{E}_S(f_\lambda) - \mathcal{E}(f_*) + \lambda \|f_\lambda\|_{\mathcal{H}}^2\Big] \\
&= \underbrace{\mathbb{E}\Big[\mathcal{E}(f_S) - \mathcal{E}_S(f_S)\Big]}_{\text{Generalization Error}} + \underbrace{\mathcal{E}(f_\lambda) - \mathcal{E}(f_*) + \lambda \|f_\lambda\|_{\mathcal{H}}^2}_{\text{Interpolation and Approximation Error}}
\end{aligned}$$

Since we know the stability of Tikhonov regualrization, which is $\mathcal{O}(1/(n\lambda))$. If we assume the interpolation and approximation error to be λ^s, for some $s > 0$, then:

$$\mathbb{E}[\mathcal{E}(f_S) - \mathcal{E}(f_*)] \le \mathcal{O}\left(\frac{1}{n\lambda}\right) + \lambda^s$$

We can choose the optimal λ to be $n^{-1/(s+1)}$, and we concluded the proof. □

Remark 92. It is easy to show that $s = 1$ when $f^* \in \mathcal{H}$ and the expected excess risk decrease with rate $\mathcal{O}(n^{-1/2})$

Theorem 3.3.10. ***(McDiarmid's Inequality)*** *Let $F : \mathcal{Z}^n \times \mathcal{Z}^n \to \mathbb{R}$ such that for any $i = 1, \dots, n$, there is $c_i > 0$, where*

$$\sup_{S \in \mathcal{Z}^n, z \in \mathcal{Z}} \left| F(S) - F(S^{i,z}) \right| < c_i$$

Then we have following bounds:

$$\mathbb{P}_{S \sim \rho^n} \left(\left| F(S) - \mathbb{E}_{S' \sim \rho^n}[F(S')] \right| \geq \varepsilon \right) \leq 2 \exp \left(- \frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2} \right)$$

Theorem 3.3.11. *For a $\beta(n)$ uniformly stable algorithm $\mathcal{A}$, where for any $S \in \mathcal{Z}^n$, we have $f_S = \mathcal{A}(S)$, then:*

$$\left| \mathcal{E}_S(f_S) - \mathcal{E}(f_S) \right| \leq \beta(n) + (n\beta(n) + M) \sqrt{\frac{2 \log(2/\delta)}{n}}$$

with probabbilty less than $1 - \delta$, where

$$M \geq \sup_{S \in \mathcal{Z}^n, i=1,\dots,n} |l(S, z_i)|$$

Proof. We would set $F(S)$ to be $\mathcal{E}(f_S) - \mathcal{E}_S(f_S)$, and the apply the McDiarmid's ineqality, which we know that $|\mathbb{E}_{S'} F(S')| \leq \beta(n)$, thus we have:

$$\left| \mathcal{E}_S(f_S) - \mathcal{E}(f_S) \right| \leq \beta(n) + \sqrt{\frac{\sum_{i=1}^n c_i \log(2/\delta)}{2}}$$

Now, to consider the bound, for $F(S) - F(S^{i,z})$

$$\begin{aligned}
\left| F(S) - F(S^{i,z}) \right| &\leq \left| \mathcal{E}(f_S) - \mathcal{E}(f_{S^{i,z}}) \right| + \left| \mathcal{E}_S(f_S) - \mathcal{E}_{S^{i,z}}(f_{S^{i,z}}) \right| \\
&\leq \frac{1}{n} \sum_{j \neq i} \left| l(f_1(x_j), y_j) - l(f_2(x_j), y_j) \right| + \frac{1}{n} \left| l(f_1(x_i), y_i) - l(f_2(x_i'), y_i') \right| + \beta(n) \\
&= \frac{(n-1)\beta(n)}{n} + \frac{2M}{n} + \beta(n) \leq 2\beta(n) + \frac{2M}{n}
\end{aligned}$$

Plugging back, and we have the statement above. □

Proposition 3.3.6. *The value M for Tikhonov's regualrization is:*

$$\sup_{S \in \mathcal{Z}^n, i=1,\dots,n} |l(S, z_i)| \leq kL \sqrt{\frac{c_0}{\lambda}} + c_0$$

where $l(0, y) \leq c_0$ for all $y \in \mathcal{Y}$ as l is L-Lipschitz and $k^2 = \sup_x k(x, x)$

Proof. For the empirical minimizer f_S, we have

$$\mathcal{E}_S(f_S) + \lambda \|f_S\| \leq \mathcal{E}_S(0) \leq c_0$$

This means that, since the loss is negative

$$\|f_S\| \leq \sqrt{\frac{c_0}{\lambda}} - \mathcal{E}_S(f) \leq \sqrt{\frac{c_0}{\lambda}}$$

Then, we have:

$$\begin{aligned}
|l(f_S, z)| &\leq |l(f_S, z) - l(0, z)| - |l(0, z)| \\
&\leq |l(f_S, z) - l(0, z)| - c_0 \\
&\leq kL \|f_S\| + c_0 = kL \sqrt{\frac{c_0}{\lambda}} + c_0
\end{aligned}$$

□

Corollary 3.3.2. *The generalization bound for Tikhonov's regualrization is*

$$\left|\mathcal{E}_S(f_S) - \mathcal{E}(f_S)\right| \leq \frac{2k^2L^2}{n\lambda} + \left(\frac{2k^2L^2}{\lambda} + kL\sqrt{\frac{c_0}{\lambda}} + c_0\right)\sqrt{\frac{2\log(2/\delta)}{n}}$$

with the probabbilty less than $1-\delta$

Remark 93. Or, we have

$$\left|\mathcal{E}_S(f_S) - \mathcal{E}(f_S)\right| \leq \mathcal{O}\left(\frac{1}{\sqrt{n}\lambda}\right)$$

We, now, can find a suitable λ.

3.4 Early Stopping

Remark 94. We consider an iterated algorithm and apply to unregularized ERM with n training points. Let f_n be a solution of ERM and $f_n^{(t)}$ be sequence of function obtained by the gradient descent. We would like to find a spot where the algorithm isn't trained too few or too much.

Remark 95. The intuition here is that every step of gradient descent allows the points to move from previous state in certain amount i,e $f_n^{(t)} \in \mathcal{H}_{r(t)}$ for some radius $r(t)$. To set an early stop means that we regularize the space of $\mathcal{H}$.

Lemma 3.4.1. *For L-Lipschitz, convex, and differentiable function* $f : \mathcal{H} \to \mathbb{R}$*. Then*

$$\|\nabla F(f)\| \leq L$$

for some $f \in \mathcal{H}$*.*

Proof. We consider, where we set $y = x + \nabla F$:

$$L\|\nabla F\| = L\|y - x\| \geq \|f(y) - f(x)\| = \left\|\nabla F^T(y-x)\right\| = \|\nabla F\|^2$$

□

Proposition 3.4.1. *At step t of gradient descent with step size* $\gamma > 0$ *on* F*, we have:*

$$\|f_t\|_{\mathcal{H}} \leq t\gamma L$$

Proof.

$$\|f_t\|_{\mathcal{H}} = \|f_{t-1} - \gamma\nabla F(f_{t-1})\|_{\mathcal{H}} \leq \|f_{t-1}\|_{\mathcal{H}} + \gamma\|\nabla F(f_{t-1})\|_{\mathcal{H}} = \|f_{t-1}\|_{\mathcal{H}} + \gamma L$$

Repeat the process and and we that we have. □

Lemma 3.4.2. *For a function* $F : \mathcal{H} \to \mathbb{R}$ *convex,* M*-smooth with minimizer* $w_* \in \mathcal{H}$*, we have:*

$$F(w) - F(w_*) \geq \frac{1}{2M}\|\nabla F(w)\|_{\mathcal{H}}^2$$

Proof. We consider the lemma 3.3.2

$$\inf_{v\in\mathcal{H}} f(v) \leq \inf_{v\in\mathcal{H}} f(w) + \nabla f(w)^T(v-w) + \frac{L}{2}\|v-w\|_{\mathcal{H}}^2$$

Let's consider the derivative with respect to v:

$$\begin{aligned}
\nabla_v\left[\nabla_w f(w)^T(v-w) + \frac{L}{2}\|v-w\|_{\mathcal{H}}^2\right] &= \nabla_v\left[\nabla_w f(w)^T v - \nabla_w f(w)^T w + \frac{L}{2}(v^Tv - 2v^Tw + w^Tw)\right] \\
&= \nabla_w f(w) - 0 + \frac{L}{2}2v - \frac{L}{2}2w + 0 \\
&= \nabla_w f(w) + L(v-w)
\end{aligned}$$

Setting the derivative to zero gives us:

$$v = w - \frac{1}{L}\nabla_w f(w)$$

Plugging it back, and we have:

$$\begin{aligned}
f(w_*) &\leq f(w) + \nabla f(w)^T \left(w - \frac{1}{L}\nabla_w f(w) - w \right) + \frac{L}{2}\left\| w - \frac{1}{L}\nabla_w f(w) - w \right\|^2_{\mathcal{H}} \\
&= f(w) - \frac{1}{L}\left\|\nabla_w f(w)\right\|^2_{\mathcal{H}} + \frac{1}{2L}\left\|\nabla_w f(w)\right\|^2_{\mathcal{H}} \\
&= f(w) - \frac{1}{2L}\left\|\nabla_w f(w)\right\|^2_{\mathcal{H}}
\end{aligned}$$

Rearrange and we have what required. □

Proposition 3.4.2. *Givne a function $F : \mathcal{H} \to \mathbb{R}$ convex M-smooth, then for all v, w, we have:*

$$\langle \nabla F(w) - \nabla F(v), w - v \rangle_{\mathcal{H}} \geq \frac{1}{M}\left\|\nabla F(w) - \nabla F(v)\right\|^2_{\mathcal{H}}$$

Proof. First, we constructed a function:

$$F_w(z) = F(z) - \langle \nabla_w F(w), z \rangle_{\mathcal{H}} \qquad F_v(z) = F(z) - \langle \nabla_v F(v), z \rangle_{\mathcal{H}}$$

We can see that both functions are M-smooth, as we have:

$$\nabla_z F_w(z) = \nabla_z F(z) - \nabla_w F(w)$$

Furthermore, from this, we can see that $z = w$ is the optima, and same for $F_v(z)$ where $z = v$ is also an optima. Apply the previous lemma, we have:

$$F_w(v) - F_w(w) \geq \frac{1}{2M}\left\|\nabla F_w(v)\right\|^2_{\mathcal{H}} \qquad F_v(w) - F_v(v) \geq \frac{1}{2M}\left\|\nabla F_v(w)\right\|^2_{\mathcal{H}}$$

where:

$$\begin{aligned}
F_w(v) &= F(v) - \langle \nabla_w F(w), v \rangle_{\mathcal{H}} \qquad & F_v(w) &= F(w) - \langle \nabla_v F(v), w \rangle_{\mathcal{H}} \\
F_w(w) &= F(w) - \langle \nabla_w F(w), w \rangle_{\mathcal{H}} \qquad & F_v(v) &= F(v) - \langle \nabla_v F(v), v \rangle_{\mathcal{H}}
\end{aligned}$$

And, so we have:

$$\begin{aligned}
F(v) - F(w) - \langle \nabla_w F(w), v - w \rangle_{\mathcal{H}} &\geq \frac{1}{2M}\left\|\nabla F_w(v)\right\|^2_{\mathcal{H}} \\
F(w) - F(v) - \langle \nabla_v F(v), w - v \rangle_{\mathcal{H}} &\geq \frac{1}{2M}\left\|\nabla F_v(w)\right\|^2_{\mathcal{H}}
\end{aligned}$$

Adding them together, we have:

$$\begin{aligned}
\langle \nabla_w F(w) - \nabla_v F(v), w - v \rangle_{\mathcal{H}} &\geq \frac{1}{2M}\left\|\nabla F_w(v)\right\|^2_{\mathcal{H}} + \frac{1}{2M}\left\|\nabla F_v(w)\right\|^2_{\mathcal{H}} \\
&\geq \frac{1}{M}\left\|\nabla F_w(v) + \nabla F_v(w)\right\|^2_{\mathcal{H}} \\
&= \frac{1}{M}\left\|\nabla F(w) - \nabla F(v)\right\|^2_{\mathcal{H}}
\end{aligned}$$

Thus complete the proof. □

Lemma 3.4.3. *Let $l : \mathcal{H} \to \mathbb{R}$ be convex differentiable and M-smooth. Let $0 \geq \gamma \geq 2/M$ and $G : \mathcal{H} \to \mathcal{H}$ be the gradient step operator: $G(f) = f - \gamma\nabla l(f)$ for $f \in \mathcal{H}$, then:*

$$\left\|G(f) - G(g)\right\|_{\mathcal{H}} \leq \left\|f - g\right\|_{\mathcal{H}}$$

Proof. We have:

$$\begin{aligned}
\Big\|G(f) - G(g)\Big\|_{\mathcal{H}}^2 &= \Big\|f - \gamma\nabla l(f) - g + \gamma\nabla l(g)\Big\|_{\mathcal{H}}^2 \\
&= \Big\|f - g + \gamma\Big(\nabla l(g) - \nabla l(f)\Big)\Big\|_{\mathcal{H}}^2 \\
&= \Big\|f - g\Big\|_{\mathcal{H}}^2 + \Big\|\gamma\Big(\nabla l(g) - \nabla l(f)\Big)\Big\|_{\mathcal{H}}^2 - 2\gamma\Big\langle f - g, \nabla l(f) - \nabla l(g)\Big\rangle \\
&\leq \Big\|f - g\Big\|_{\mathcal{H}}^2 + \gamma^2\|\nabla l(g) - \nabla l(f)\|_{\mathcal{H}}^2 - \frac{2\gamma}{M}\|\nabla l(f) - \nabla l(g)\|_{\mathcal{H}}^2 \\
&= \|f - g\|_{\mathcal{H}}^2 - \gamma\left(\frac{2}{M} - \gamma\right)\|\nabla l(f) - \nabla l(g)\|_{\mathcal{H}}^2 \leq \|f - g\|_{\mathcal{H}}^2
\end{aligned}$$

Since $\gamma(2/M - \gamma) \leq 1$ since $\gamma \in [0, 2/M]$. ☐

Theorem 3.4.1. *Let $l(\cdot, y) : \mathcal{H} \to \mathbb{R}$ be convex, L-Lipschitz and M-smooth uniform. For training set $S \in \mathcal{Z}^n$, let $f_S^{(T)}$ be obtained by applying gradient descent with step size $1/M$ on empirical risk to S. The corresponding algorithm is $\beta(n, T)$-stable where:*

$$\beta(n, T) \leq \frac{2L^2k^2}{M}\frac{T}{n}$$

Proof. Let $S \in \mathcal{Z}^n, z \in \mathcal{Z}$ and $i \in [n]$. We will denote f_t to be function after t iteration with gradient step γ on S. On the other hand, we denote f'_t to be a function after t iteration with same learning on $S^{i,z}$. Recall the result from the proof of theorem 3.3.8, that

$$\sup_{\bar{z}\in\mathcal{Z}}\Big|l(f_T, \bar{z}) - l(f'_T, \bar{z})\Big| \leq Lk\,\|f_T - f'_T\|_{\mathcal{H}}$$

We want to control this value. For any $t \in [n]$ by construction:

$$f_{t+1} = f_t - \gamma\nabla\mathcal{E}_S(f_t) \qquad f'_{t+1} = f'_t - \gamma\nabla\mathcal{E}_{S^{i,z}}(f_t)$$

Then, we have:

$$\begin{aligned}
\|f_{t+1} - f'_{t+1}\|_{\mathcal{H}} &= \left\|f_t - f'_t - \frac{\gamma}{n}\sum_{j\neq i}\Big[\nabla l(f_t, z_j) - \nabla l(f'_t, z_j)\Big] + \frac{\gamma}{n}\Big[\nabla l(f_t, z_i) - \nabla l(f'_t, z)\Big]\right\|_{\mathcal{H}} \\
&\leq \frac{1}{n}\sum_{j\neq i}\Big\|f_t - \gamma\nabla l(f_t, z_j) - f'_t + \gamma\nabla l(f'_t, z_j)\Big\|_{\mathcal{H}}^2 + \frac{1}{n}\|f_t - f'_t\|_{\mathcal{H}} \\
&\quad + \frac{\gamma}{n}\Big(\|\nabla l(f_t, z_i)\|_{\mathcal{H}} + \|\nabla l(f'_t, z)\|\Big) \\
&= \|f_t - f'_t\|_{\mathcal{H}} + \frac{2Lk}{n}\gamma
\end{aligned}$$

The second ineqalities comes from lemma 3.4.1 and lemma 3.4.3. Please note that $\|\nabla l(f_t, z)\|_{\mathcal{H}} \leq Lk$:

$$\|f_{t+1} - f'_{t+1}\|_{\mathcal{H}} \leq \|f_t - f'_t\|_{\mathcal{H}} + \frac{2Lk}{nM} = \frac{2Lk(t+1)}{nM}$$

Setting $t + 1 = T$, and we finish the proof, while setting $\gamma = 1/M$ ☐

3.5 Sub-Gradient Methods

3.5.1 Introduction to Sub-Gradient

Definition 3.5.1. (Convex Function) A function $f : X \to [-\infty, \infty]$ is convex iff, for all $x, y \in X$ and $\lambda \in [0, 1]$:

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y)$$

Definition 3.5.2. (Extended Value Theorem) We can transform the constrained optimization:

$$\min_{\|x\|\leq 1} \|Ax - y\|^2$$

Using the extended value theorem, where this is the same as:

$$\min_{x\in X} f(x) = h(x) + L_{B_1}(x) \qquad \text{where} \qquad L_{B_1}(x) = \begin{cases} 0 & x \in B_1 \\ \infty & x \notin B_1 \end{cases}$$

Definition 3.5.3. (Subdifferential & Subgradient) Let $x \in \text{dom}(f)$, the subdifferential:

$$\partial f(x) = \left\{ u \in X \middle| \forall y \in X : f(y) \geq f(x) + \langle y - x, u\rangle \right\}$$

The subgradient is the element of ∂f at x. Please note that $z = f(y) \geq f(x) + \langle y - x, u\rangle$ is the affine function passing through $(x, f(x))$ with slope u. If $x \notin \text{dom}(f)$, then by definition $\partial f(x) = \emptyset$

Lemma 3.5.1. *Suppose that* $X = X_1 \times \cdots \times X_m$ *and* $f(x_1, \dots, x_m) = f_1(x_1) + \cdots + f_m(x_m)$ *where* $f_i : X_i \to]-\infty, \infty]$*, then we have:*

$$\partial f(x_1, \dots, x_m) = \underbrace{\partial f_1(x_1)}_{\subset X_1} \times \cdots \times \underbrace{\partial_m(x_m)}_{\subset X_m} \subset X$$

Remark 96. Let's consider $X = \mathbb{R}^m$ where $f(x) = \|x\|_1 = \sum_{i=1}^m |x_i|$ where $f_i = |\cdot| : \mathbb{R} \to \mathbb{R}$, then we have:

$$\partial \|\cdot\|_{x_1}(x) = \underbrace{\partial |\cdot| (x_1)}_{\subset \mathbb{R}} \times \cdots \times \underbrace{\partial |\cdot| (x_m)}_{\subset \mathbb{R}} \subset \mathbb{R}^m$$

where, we have:

$$\partial |\cdot| (x) = \begin{cases} \{-1\} & \text{if } t < 0 \\ [-1, 1] & \text{if } t = 0 \\ \{1\} & \text{if } t > 0 \end{cases}$$

Lemma 3.5.2. *For a convex function* $f : \mathbb{R} \to \mathbb{R}$ *(note that it is finite), its subdifferential is:*

$$\partial f(x) = \left[f'_-(x), f'_+(x)\right]$$

However, for infinite value function, its subdifferential is:

$$\partial f(x) = \left[f'_-(x), f'_+(x)\right] \cap \mathbb{R}$$

Remark 97. We have the problem:

$$\min_{x\in C} f(x) \quad \text{where } C \subset X \text{ is closed convex.}$$
$$f : X \to \mathbb{R} \text{ is convex and Lipschitz continuous.}$$

If f is finite every where, then subdifferential is non-empty, while in smooth setting, there is one subgradient, which is the gradient.

3.5.2 Projected Subgradient Method

Definition 3.5.4. (Projected Subgradient Method) The projected subgradient method is given by:

$$x_{k+1} = P_c(x_k - \gamma u_k)$$

where $u_k \in \partial f(x_n)$ and $\gamma_n > 0$.

Remark 98. Projected Subgradient method isn't decending. We will consider $X = \mathbb{R}^2$ where $f(x_1, x_2) = |x_1| + 2|x_2|$ as we have

$$\partial f(1,0) = \{1\} \times [-2,2]$$

it is clear that $(1,2) \in \partial f(1,0)$. Then choosing this subgradient will not lead to any convergence.

Lemma 3.5.3. *We would like to note that: if $u \in \partial f(x)$, then $\|u\| < L$.*

Proof. We consider the following inequalities:

$$\begin{aligned}\langle y - x, u\rangle &\le f(y) - f(x)\\ &\le |f(y) - f(x)|\\ &\le L\|y - x\|\end{aligned}$$

If we were to set, $u = y - x$:

$$\langle y - x, y - x\rangle = \|y - x\|^2 \le L\|y - x\|$$

and by simple rearrangement, we arrived at the statement. □

Lemma 3.5.4.

$$\|x_{k+1} - x_k\| = \|P_C(y_k) - P_c(x_k)\| \le \|y_k - x_k\|$$

Lemma 3.5.5. *For all $k \in \mathbb{N}$ and $x \in C$:*

$$\begin{aligned}2\gamma_k(f(x_k) - f(x)) &\le 2\gamma_k\langle x_k - x, u_k\rangle\\ &\le \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \gamma_k^2 L^2\end{aligned}$$

Proof. The first ineqalities comes from the definition of subgradient:

$$\begin{aligned}2\gamma_k(f(x_k) - f(x)) &\le 2\gamma_k\langle x_k - x, u_k\rangle\\ &= 2\langle x_k - x, \gamma_k u_k\rangle\\ &= \|x_k - x\|^2 + \|\gamma_k u_k\|^2 - \|x_k - x - \gamma_k u_k\|\\ &\le \|x_k - x\|^2 - \|y_k - x\| + \gamma_k^2 L^2\\ &\le \|x_k - x\|^2 - \|x_{k+1} - x_k\|^2 + \gamma_k^2 L^2\end{aligned}$$

□

Theorem 3.5.1. *For all $k \in \mathbb{N}$, we have $f_k = \min_{0\le i\le k} f(x_i)$ and $\bar{x} = \left(\sum_{i=0}^k \gamma_i\right)^{-1}\left(\sum_{i=0}^k \gamma_i x_i\right)$. Then for all $k \in \mathbb{N}$ and $x \in C$:*

$$\max\{f_k, f(\bar{x}_k)\} - f(x) \le \frac{\|x_0 - x\|^2}{2\sum_{i=0}^k \gamma_i} + \frac{L^2}{2}\frac{\sum_{i=0}^k \gamma_i^2}{\sum_{i=0}^k \gamma_i}$$

Proof. We start by summing the lemma:

$$\sum_{i=0}^k 2\gamma_i(f(x_i) - f(x)) \le \sum_{i=0}^k \|x_i - x\|^2 - \|x_{i+1} - x\|^2 + \gamma_i^2 L^2$$

Let's consider with the following, with the convexity of f:

$$f(\bar{x}) = f\left(\frac{\sum_{i=0}^k \gamma_i x_i}{\sum_{i=0}^k \gamma_i}\right) \le \sum_{i=1}^k f(x_i)\gamma_i \Big/ \sum_{i=0}^k \gamma_i$$

And, so we have:

$$\left(\sum_{i=0}^k \gamma_i\right)\max\{f_k, f(\bar{x}_k)\} \le \sum_{i=0}^k \gamma_i f(i)$$

as f_k is always less than $f(i)$ by definition, thus the maximum holds, and, so we can apply the lemma above with telescoping sum:

$$\begin{aligned}\left(\sum_{i=0}^{k} 2\gamma_i\right)\Big[\max\{f_k, f(\bar{x}_k)\} - f(x)\Big] &\leq \sum_{i=0}^{k} 2\gamma_i(f(x_i) - f(x)) \\ &\leq \sum_{i=0}^{k} \|x_i - x\|^2 - \|x_{i+1} - x\|^2 + L^2 \sum_{i=0}^{k} \gamma_i^2 \\ &= \|x_0 - x\|^2 - \|x_{k+1} - x\|^2 + L^2 \sum_{i=0}^{k} \gamma_i^2 \\ &\leq \|x_0 - x\|^2 + L^2 \sum_{i=0}^{k} \gamma_i^2\end{aligned}$$

By rearrange the equation, the statement. □

Corollary 3.5.1. *Suppose that $\sum_{k\in\mathbb{N}} \gamma = \infty$ and $\left(\sum_{i=0}^{k} \gamma_i\right)^{-1} \left(\sum_{i=0}^{k} \gamma_i x_i\right) \to 0$, then it is clear that*

$$f_k \to \inf_c f \qquad f(\bar{x}_k) \to \inf_c f$$

The possible choice: $\gamma_k = \bar{\gamma}/(k+1)^2$ with $\gamma \in [1/2, 1]$. In particular, $\gamma_k = \bar{\gamma}/\sqrt{k+1}$ and $\bar{\gamma}_k = \bar{\gamma}/(k+1)$

Remark 99. The result above doesn't assume that $s_* = \arg\min_c f \neq \emptyset$. As for all $x \in C$, we have:

$$f(\bar{x}_k) \leq f(x) + \frac{\|x_0 - x\|^2}{2\sum_{i=0}^{k} \gamma_i} + \frac{L^2}{2} \frac{\sum_{i=0}^{k} \gamma_i^2}{\sum_{i=0}^{k} \gamma_i}$$

But we can see that $\limsup f(\bar{x}_k) \leq f(x)$ and for all $x \in C$:

$$\limsup f(\bar{x}_k) \leq \inf_c f \leq \liminf_k f(\bar{x}_k) \leq \limsup f(\bar{x}_k)$$

and so they are all equal and will converge to f.

Corollary 3.5.2. *Suppose that $S_* = \arg\min_c f \neq \emptyset$ then the following holds:*

- *Let $k \in \mathbb{N}$ then: set $(\gamma_i)_{0\leq i\leq k} = \frac{\|x_0 - S_*\|}{L\sqrt{k+1}}$ then:*

$$\max\{f_k, f(\bar{x}_k)\} - \min_c f < \frac{Ld(x_0, S_*)}{\sqrt{k+1}}$$

- *Suppose that X is finite dimensional, where $\sum \gamma_k = \infty$ and $\sum \gamma_k^2 < \infty$ then there exists $x_* \in S_*$ such that $x_k \to x_*$*
- *For every $k \in \mathbb{N}$ where $\gamma_k = \bar{\gamma}/(k+1)$, then:*

$$\max\{f_k, f(\bar{x}_k)\} - \min_c f \leq \mathcal{O}\left(\frac{1}{\log(k+1)}\right)$$

- *For every $k \in \mathbb{N}$ where $\gamma_k = \bar{\gamma}/\sqrt{k+1}$, then:*

$$\max\{f_k, f(\bar{x}_k)\} - \min_c f \leq \mathcal{O}\left(\frac{\log(k+1)}{\sqrt{k+1}}\right)$$

- *For every $k \in \mathbb{N}$ where $\gamma_k = \bar{\gamma}/\sqrt{k+1}$, then: $\tilde{f}_k = \inf_{\lfloor k/2 \rfloor \leq i \leq k} f(x_i)$ where:*

$$\tilde{x}_k = \left(\sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i\right)^2 \sum_{i=\lfloor k/2 \rfloor}^{k} \gamma_i x_i$$

Suppose C is bounded then:

$$\max\{f_k, f(\bar{x}_k)\} - \min_c f = \mathcal{O}\left(\frac{1}{\sqrt{k+1}}\right)$$

Definition 3.5.5. (Projected Stochastic Subgradient Method) The algorithm is defined as:

$$x_{k+1} = P_C(x_k - \gamma_k \hat{u}_k)$$

where $\hat{u}_k$ is x-valued random variable such that $\mathbb{E}[\hat{u}_k | x_k] \in \partial f(x_k)$. Now, we have x_k and $f(x_k)$ are random varaible now.

Remark 100. We are going to define a function values $f_k = \min_{0 \leq i \leq k} \mathbb{E}[f(x_i)]$ and $\bar{x}_k = \left(\sum_{i=0}^{k} \gamma_i\right)^2 \left(\sum_{i=0}^{k} \gamma_i x_i\right)$. Together with the assumption that there exists $B > 0$ such that for all $k \in \mathbb{N}$ as $\mathbb{E}[\|\hat{u}_k\|^2] \leq B^2 < \infty$.

Lemma 3.5.6. *For all $k \in \mathbb{N}$ and all points $x \in C$:*

$$2\gamma_n(\mathbb{E}[P(x_n)] - f(x)) \leq \mathbb{E}[\|x_k - x\|^2] - \mathbb{E}[\|x_{k+1} - x\|^2] + \gamma_k^2 B^2$$

Proof. We consider $y_k = x_k - \gamma \hat{u}_k$ and $x_{k+1} = P_C(y_k)$, then we have:

$$\begin{aligned} 2\gamma_k \langle x_k - x, \hat{u}_k \rangle &= 2 \langle x_k - x, x_k - y_k \rangle \\ &= \|x_k - x\|^2 + \|x_k - y_k\|^2 - \|y_k - x\|^2 \\ &\leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \gamma_k^2 \|u_k\|^2 \end{aligned}$$

and so we have:

$$\begin{aligned} 2\gamma_k \langle x_k - x, \mathbb{E}[u_k | x_k] \rangle &= 2 \langle x_k - x, x_k - y_k \rangle \\ &\leq \|x_k - x\|^2 - \mathbb{E}\left[\|x_{k+1} - x\|^2 \,|x_k\right] + \gamma_k^2 \mathbb{E}\left[\|u_k\|^2 \,|x_k\right] \end{aligned}$$

Note that

$$2\gamma_k\Big(f(x_k) - f(x)\Big) \leq 2\gamma_k \langle x_k - x, \mathbb{E}[u_k | x_k] \rangle$$

and so, we have:

$$\begin{aligned} 2\gamma_k\Big(\mathbb{E}[f(x_k)] - f(x)\Big) &\leq \|x_k - x\|^2 - \mathbb{E}\left[\|x_{k+1} - x\|^2\right] + \gamma_k^2 \mathbb{E}\left[\|u_k\|^2\right] \\ &\leq \|x_k - x\|^2 - \mathbb{E}\left[\|x_{k+1} - x\|^2\right] + \gamma_k^2 B^2 \end{aligned}$$

□

Theorem 3.5.2. *For all number $k \in \mathbb{N}$ and for all $x \in C$: we have*

$$\max\{f_k, \mathbb{E}[f(\bar{x}_k)]\} - f(x) \leq \frac{\mathbb{E}[\|x_0 - x\|^2]}{2\sum_{i=0}^{k} \gamma_i} + \frac{B^2}{2} \frac{\sum_{i=1}^{k} \gamma_i^2}{\sum_{i=1}^{k} \gamma_i}$$

Corollary 3.5.3. *Suppose that $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\sum_{i=0}^{n} \gamma_i^2 / \sum_{i=0}^{n} \gamma_i \to 0$ where $\gamma = \bar{\gamma}/(1+k)^2$ with $\alpha \in [1/2, 1]$. Then $f_k \to \inf_C f$ and $\mathbb{E}[f(\bar{x}_k)] \to \inf_C f$*

Corollary 3.5.4. *Suppose that $S_* = \arg\min_c f \neq \emptyset$ and let $D \geq \operatorname{dist}(x_0, S_*)$ then the following holds:*

- *Let $k \in \mathcal{N}$ and set $(\gamma_i)_{1 \le i \le k} = D/(B\sqrt{k+1})$ then:*

$$\max\{f_n, \mathbb{E}[f(\bar{x}_k)]\} - \min_c f \le \frac{BD}{\sqrt{k+1}}$$

- *Set $\gamma_k = \bar{\gamma}/\sqrt{k+1}$ then:*

$$\max\{f_n, \mathbb{E}[f(\bar{x}_k)]\} - \min_c f \le \mathcal{O}\left(\frac{\log(k+1)}{\sqrt{k+1}}\right)$$

3.5.3 Examples of Stochastic Optimization

Remark 101. **(Stochastic Optimization)** We have the following setting:

$$\min_{x \in C} f(x) = \mathbb{E}[f(x,\xi)] = \int_{\mathcal{Z}} F(x,z)\ \mathrm{d}\mu(\mathcal{Z})$$

where ξ is random variable taking values in measurable space $\mathcal{Z}$ with distribution measure $\mu(\mathcal{Z})$ and $F : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ such that:

- $F(\cdot, z)$ is convex and $L(\mathcal{Z})$-Lipschitz continuous and

$$\int_{\mathcal{Z}} L(z)^2\ \mathrm{d}\mu(\mathcal{Z}) < \infty$$

- $F(0,z) \in L^1(\mathcal{Z}, \mu)$
- There exists $\tilde{\nabla} F : X \times \mathcal{Z} \to X$ such that $\tilde{\nabla} F(x,z)$ is subgradient of $F(\cdot, z)$ at X.
- $(\xi_k)_{k \in \mathbb{N}}$ is sequence of independent copies of S.

Remark 102.

$$\begin{aligned} |F(\cdot, x)| &\le |F(x,\cdot) - F(0,\cdot)| + |F(0,\cdot)| \\ &\le L(\cdot)\left\|x\right\| + |F(0,\cdot)| \end{aligned}$$

Thus $F(x,\cdot) \in L^1(z,\mu)$.

Definition 3.5.6. (Projected Gradient Descent) We have the following algorithm:

$$x_{k+1} = P_c(x_k - \gamma_k \underbrace{\tilde{\nabla} F(x_k, \xi_k)}_{\hat{u}_k})$$

Checking the assumption on $\hat{u}_k$:

- $x_k = x_k(\xi_0, \dots, \xi_{k-1})$ as we have x_k and ξ_k are independent that random value.
- We have:

$$\begin{aligned} F(y,z) &\ge F(x,z) + \left\langle y - x, \tilde{\nabla} F(x,z) \right\rangle \\ f(y) &\ge f(x) + \left\langle y - x, \underbrace{\int_{\mathcal{Z}} \tilde{\nabla} F(x,z)\ \mathrm{d}\mu(\mathcal{Z})}_{\mathbb{E}[\tilde{\nabla} F(x,\xi)]} \right\rangle \end{aligned}$$

  for all $x, y \in X$ and $z \in \mathcal{Z}$. And, $\mathbb{E}[\tilde{\nabla} F(x,\xi)] \in \partial f(x)$, or we have

$$\mathbb{E}[\tilde{\nabla} F(x_k, \xi) | x_k] = \int \tilde{\nabla} F(x_k, z)\ \mathrm{d}\mu(z) \in \partial f(x_k)$$

- We have:

$$\mathbb{E}\left[\left\|\tilde{\nabla}F(x_k,\xi_k)\right\|^2\middle| x_k\right] = \int \left\|\tilde{\nabla}F(x_k,z)\right\|^2 \,\mathrm{d}\mu(\mathcal{Z})$$
$$\leq \int L(z)^2 \,\mathrm{d}\mu(\mathcal{Z}) = B^2$$

Definition 3.5.7. (Statistical Learning) Let ξ and η be 2 random values with value in $\mathcal{X}$ and $\mathcal{Y}$ repectively, and let μ be the distribution of (ξ,η). Let $l : \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to\mathbb{R}$ be a convex loss function and $\Phi : \mathcal{X}\to H$ be a feature map:

$$\min_{w\in\mathcal{H}} R(w) = \int_{\mathcal{X}\times\mathcal{Y}} l(x,y,\langle w,\Phi(s)\rangle)\,\mathrm{d}\mu(X,Y)$$
$$= \mathbb{E}[l(\xi,\eta,\langle w,\Phi(s)\rangle)]$$

based on some sequence $(\xi_k,\eta_k)_{k\in\mathbb{N}}$ of independent copies of (ξ,η). We assume:

- $l(x,y,\cdot)$ is 2-Lipschitz continuous and $\mathbb{E}[l(\xi,\eta,0)] < \infty$
- $\mathbb{E}\left[\|\Phi(x)\|^2\right] \leq \infty$ as we have $\mathbb{E}[k(\xi,\xi)] < \infty$

We will now check that the assumption for stochastic optimization holds, where we will set $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, $F : H\times\mathcal{Z}\to\mathbb{R}$ and $F(w,z) = l(x,y,\langle w,\Phi(x)\rangle)$

- Let's consider the $F(\cdot,z) = l(x,y,\langle\cdot,\Phi(x)\rangle)$ and it is convex:

$$|F(w_1,z)-F(w_2,z)| = |l(x,y,\langle w_1,\Phi(x)\rangle) - l(x,y,\langle w_1,\Phi(x)\rangle)|$$
$$\leq 2\,|\langle w_1-w_2,\Phi(x)\rangle|$$
$$\leq 2\underbrace{\|\Phi(x)\|}_{L(z)}\|w_1-w_2\|$$

- We have $F(0,\cdot) = l(\cdot,\cdot,0)\in L^1(\mathcal{Z},\mu)$
- For the subgradient, we have:

$$\partial F(w,z) = \partial\underbrace{l(x,y,\langle w,\Phi(x)\rangle)}_{\subset\mathbb{R}}\underbrace{\Phi(x)}_{\in H}\subset H$$

  as we have $\tilde{l}' : \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to\mathbb{R}$ or we have $\tilde{l}'(x,y,t)\in\partial l(x,y,t)$, thus we have:

$$\tilde{\nabla}F(w,z) = \tilde{l}'(x,y,\langle w,\Phi(x)\rangle)\Phi(x)\in\partial F(w,z)$$

  And so the third condition holds.

- $\xi_k = (\xi_k,\eta_k)$ and so the final assumption holds.

Definition 3.5.8. (Statitical Learning Algorithm) The algorithm:

$$w_{k+1} = w_k - \gamma_k\tilde{l}'(\xi_n,\eta_n,\langle w_k,\Phi(\xi_k)\rangle)\Phi(\xi_k)$$

This isn't practical as $\mathcal{H}$ is ∞-dimension. However, we can have:

$$g_{k+1}(x) = g_k(x) - \gamma_k\tilde{l}'(\xi_k,\eta_k,g_k(\xi_k))K(x,\xi_k)$$

Where $k(x,x') = \langle\Phi(x),\Phi(x')\rangle$ is kernel function.

Remark 103. We let

$$\bar{w}_n = \left(\sum_{i=0}^{k}\gamma_i\right)^{-1}\left(\sum_{i=0}^{k}\gamma_i w_i\right) \qquad \bar{g}_n(x) = \langle\bar{w}_k,\Phi(x)\rangle = \left(\sum_{i=0}^{k}\gamma_i\right)^{-1}\left(\sum_{i=1}^{k}\gamma_i g_i(x)\right)$$

We have:

- The risk of g_k is $R(\bar{w}_k)$ and according to corollary, we have:

$$R(\bar{w}_k) \to \inf_H R$$

provided that $\sum_{k=-\infty}^{\infty} \gamma_k = \infty$ and $(\sum_{i=0}^{\infty} \gamma_i)^{-1}(\sum_{i=0}^{k} \gamma_i^2) \to 0$

- Suppose that $S_* = \arg\min_{\mathcal{H}} R \neq \emptyset$ and let $D \geq d(s_0, S_*)$
  - If $\gamma_k = \bar{\gamma}\sqrt{k+1}$ then:

$$\mathbb{E}[R(\bar{w}_k)] - \min_H R \leq \mathcal{O}\left(\frac{\log(k+1)}{\sqrt{k+1}}\right)$$

  - Let $k \in \mathbb{N}$ and let $(\gamma_i)_{1 \leq i \leq k} = D/(B\sqrt{k+1})$ then:

$$\mathbb{E}[R(\bar{w}_k)] - \min_{\mathcal{H}} R \leq \frac{BD}{\sqrt{k+1}}$$

    Where $B^2 = 4\mathbb{E}\left[\|\phi(\xi))\|^2\right]$

Appendix A

Additional Proof

A.1 RKHS in Machine Learning

A.1.1 Expansion of Centered Matrix for PCA

Smarter way to do it is:

$$X\left(I - \frac{1}{n}\mathbf{1}_{n\times n}\right)X^T = XX^T - \frac{1}{n}X\mathbf{1}_{n\times n}X^T$$

Now, we consider the second one:

$$\begin{aligned}
\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^T &= \sum_{i=1}^{n}x_ix_i^T - \frac{1}{n}X\mathbf{1}x_i - \frac{1}{n}x_i\mathbf{1}^TX^T + \frac{1}{n^2}X\mathbf{1}\mathbf{1}^TX^T \\
&= \left[\frac{1}{n}X\mathbf{1}\mathbf{1}^TX^T + \sum_{i=1}^{n}x_ix_i^T\right] - \left[\frac{1}{n}\sum_{i=1}^{n}X\mathbf{1}x_i^T + x_i\mathbf{1}^TX^T\right] \\
&= \left[\frac{1}{n}X\mathbf{1}\mathbf{1}^TX^T + XX^T\right] - \left[\frac{2}{n}\sum_{i=1}^{n}X\mathbf{1}x_i^T\right] \\
&= \left[\frac{1}{n}X\mathbf{1}\mathbf{1}^TX^T + XX^T\right] - \left[\frac{2}{n}X\mathbf{1}\sum_{i=1}^{n}x_i^T\right] \\
&= \left[\frac{1}{n}X\mathbf{1}\mathbf{1}^TX^T + XX^T\right] - \left[\frac{2}{n}X\mathbf{1}\mathbf{1}^TX^T\right] \\
&= XX^T - \frac{1}{n}X\mathbf{1}\mathbf{1}^TX^T
\end{aligned}$$

Note that for vector $\boldsymbol{a}$ and $\boldsymbol{b}$, we have $\boldsymbol{a}\boldsymbol{b}^T = \boldsymbol{b}\boldsymbol{a}^T$

A.1.2 Centering Kernel Matrix

Please note that

$$\tilde{k}(x_i, x_j) = \left\langle \tilde{\phi}(x_i), \tilde{\phi}(x_j)\right\rangle = \left\langle \phi(x_i) - \frac{1}{n}\sum_{k=1}^{n}\phi(x_k), \phi(x_j) - \frac{1}{n}\sum_{k=1}^{n}\phi(x_k)\right\rangle$$

Let's see that:

$$\begin{aligned}
\tilde{k}(x_i, x_j) &= \left\langle \phi(x_i) - \frac{1}{n}\sum_{k=1}^{n}\phi(x_k), \phi(x_j) - \frac{1}{n}\sum_{k=1}^{n}\phi(x_k) \right\rangle \\
&= \langle \phi(x_i), \phi(x_j) \rangle - \left\langle \phi(x_i), \frac{1}{n}\sum_{k=1}^{n}\phi(x_k) \right\rangle - \left\langle \phi(x_j), \frac{1}{n}\sum_{k=1}^{n}\phi(x_k) \right\rangle + \left\langle \frac{1}{n}\sum_{k=1}^{n}\phi(x_k), \frac{1}{n}\sum_{k=1}^{n}\phi(x_k) \right\rangle \\
&= \underbrace{\langle \phi(x_i), \phi(x_j) \rangle}_{①} \underbrace{- \frac{1}{n}\sum_{k=1}^{n}\langle \phi(x_i), \phi(x_k) \rangle - \frac{1}{n}\sum_{k=1}^{n}\langle \phi(x_j), \phi(x_k) \rangle}_{②} + \underbrace{\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\langle \phi(x_k), \phi(x_l) \rangle}_{③}
\end{aligned}$$

Now, let's consider $\tilde{K} = HKH$, which we have:

$$\begin{aligned}
\tilde{K} &= \left(I - \frac{1}{n}\mathbf{1}_{n\times n}\right) K \left(I - \frac{1}{n}\mathbf{1}_{n\times n}\right) = \left(K - \frac{1}{n}\mathbf{1}_{n\times n}K\right)\left(I - \frac{1}{n}\mathbf{1}_{n\times n}\right) \\
&= K - \frac{1}{n}K\mathbf{1}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}K + \frac{1}{n^2}\mathbf{1}_{n\times n}K\mathbf{1}_{n\times n}
\end{aligned}$$

It is clear that K corresponds to ①, and we can see that:

$$\frac{1}{n}K\mathbf{1}_{n\times n} = \frac{1}{n}\begin{bmatrix} \cdots & \sum_{i=1}^{n}\langle x_1, x_i \rangle & \cdots \\ \cdots & \sum_{i=1}^{n}\langle x_2, x_i \rangle & \cdots \\ & \vdots & \\ \cdots & \sum_{i=1}^{n}\langle x_n, x_i \rangle & \cdots \end{bmatrix} \qquad \frac{1}{n}\mathbf{1}_{n\times n}K = \frac{1}{n}\begin{bmatrix} \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n}\langle x_1, x_i \rangle & \sum_{i=1}^{n}\langle x_2, x_i \rangle & \cdots & \sum_{i=1}^{n}\langle x_n, x_i \rangle \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

And, so the addition of them would lead to the ②. Finally, ③ can be shown easily as we use the result above and multiply by $\mathbf{1}_{n\times n}$.

A.1.3 Ridge Regression Expansion

We will show that

$$-2y^T X^T Cb + b^T b = \|CXy - b\|^2 - \left\|y^T X^T C\right\|^2$$

where $C = (XX^T + \lambda I)^{-1/2}$, please note that $C = C^T$. Let's consider the right handside:

$$\begin{aligned}
\|CXy - b\|^2 - \left\|y^T X^T C^T\right\|^2 &= (CXy - b)^T(CXy - b) - (y^T X^T C^T)^T(y^T X^T C^T) \\
&= (y^T X^T C^T - b^T)(CXy - b) - (y^T X^T C^T)^T(y^T X^T C^T) \\
&= y^T X^T C^T CXy - y^T X^T C^T b - b^T CXy + b^T b - CXyy^T X^T C^T \\
&= (y^T X^T C^T CXy - CXyy^T X^T C^T) - 2y^T X^T C^T b + b^T b \\
&= -2y^T X^T C^T b + b^T b
\end{aligned}$$

A.1.4 Representor Theorem for Ridge Regression

We will assume that

$$
\begin{aligned}
X(X^TX+\lambda I_n)^{-1}y &= X\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nn} \end{bmatrix} y \\
&= \begin{bmatrix} \sum_{i=1}^n x_{1i}\beta_{i1} & \sum_{i=1}^n x_{1i}\beta_{i2} & \cdots & \sum_{i=1}^n x_{1i}\beta_{in} \\ \sum_{i=1}^n x_{2i}\beta_{i1} & \sum_{i=1}^n x_{2i}\beta_{i2} & \cdots & \sum_{i=1}^n x_{2i}\beta_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{di}\beta_{i1} & \sum_{i=1}^n x_{di}\beta_{i2} & \cdots & \sum_{i=1}^n x_{di}\beta_{in} \end{bmatrix} y \\
&= \begin{bmatrix} \sum_{i=1}^n x_{1i}\beta_{i1} & \sum_{i=1}^n x_{1i}\beta_{i2} & \cdots & \sum_{i=1}^n x_{1i}\beta_{in} \\ \sum_{i=1}^n x_{2i}\beta_{i1} & \sum_{i=1}^n x_{2i}\beta_{i2} & \cdots & \sum_{i=1}^n x_{2i}\beta_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{di}\beta_{i1} & \sum_{i=1}^n x_{di}\beta_{i2} & \cdots & \sum_{i=1}^n x_{di}\beta_{in} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^n y_j \sum_{i=1}^n x_{1i}\beta_{ij} \\ \sum_{j=1}^n y_j \sum_{i=1}^n x_{2i}\beta_{ij} \\ \vdots \\ \sum_{j=1}^n y_j \sum_{i=1}^n x_{ni}\beta_{ij} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n \sum_{i=1}^n y_j x_{1i}\beta_{ij} \\ \sum_{j=1}^n \sum_{i=1}^n y_j x_{2i}\beta_{ij} \\ \vdots \\ \sum_{j=1}^n \sum_{i=1}^n y_j x_{ni}\beta_{ij} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n \sum_{j=1}^n y_j x_{1i}\beta_{ij} \\ \sum_{i=1}^n \sum_{j=1}^n y_j x_{2i}\beta_{ij} \\ \vdots \\ \sum_{i=1}^n \sum_{j=1}^n y_j x_{ni}\beta_{ij} \end{bmatrix}
\end{aligned}
$$

The rest will be in main proof.

A.1.5 MMD Integration

We have

$$
\begin{aligned}
&\iint [k(s-t)\ \mathrm{d}(P-Q)(s)]\ \mathrm{d}(P-Q)(t) \\
&\quad = \int \Big[\mathbb{E}_{s\sim P}[k(s-t)] - \mathbb{E}_{s\sim Q}[k(s-t)]\Big]\ \mathrm{d}(P-Q)(t) \\
&\quad = \int \mathbb{E}_{s\sim P}[k(s-t)]\ \mathrm{d}(P-Q)(t) - \int \mathbb{E}_{s\sim Q}[k(s-t)]\ \mathrm{d}(P-Q)(t) \\
&\quad = \Big[\mathbb{E}_{t\sim P}\mathbb{E}_{s\sim P}[k(s-t)] - \mathbb{E}_{t\sim Q}\mathbb{E}_{s\sim P}[k(s-t)]\Big] - \Big[\mathbb{E}_{t\sim P}\mathbb{E}_{s\sim Q}[k(s-t)] - \mathbb{E}_{t\sim Q}\mathbb{E}_{s\sim Q}[k(s-t)]\Big] \\
&\quad = \mathbb{E}_P[k(s-t)] + \mathbb{E}_Q[k(s-t)] - 2\mathbb{E}_{P,Q}[k(s-t)]
\end{aligned}
$$

A.1.6 Biased Estimate of HSIC Part 2

We have

$$
\mathbf{1}^T K = \begin{bmatrix} \sum_{a=1}^n k_{a1} & \sum_{a=1}^n k_{a2} & \cdots & \sum_{a=1}^n k_{an} \end{bmatrix} \qquad L\mathbf{1} = \begin{bmatrix} \sum_{b=1}^n l_{1b} \\ \sum_{b=1}^n l_{2b} \\ \vdots \\ \sum_{b=1}^n l_{nb} \end{bmatrix}
$$

A.2 Experimental Proof

A.2.1 Projected Gradient Descent

Lemma A.2.1. *We would like to note that, for some $y \in \mathbb{R}^d$ and $x \in \Omega$*

$$\|\Pi_\Omega(y) - x\|^2 \leq \|y - x\|^2 - \|y - \Pi_\Omega(y)\|^2$$

Remark 104. The projected gradient descent can be splitted into 2 parts:

$$y_{t+1} = x_t - \gamma \nabla f(x_t)$$
$$x_{t+1} = \Pi_\Omega(y_{t+1})$$