Analysis of Kernel Methods

Mathematical Analysis of Machine Learning Algorithms
(Chapter 9)

Linear Models with L_2 Regularization

Linear Models in Feature Representation

$$\mathcal{F} = \{f(w, x) : f(w, x) = \langle w, \psi(x) \rangle\}, \tag{1}$$

where $\psi(x)$ is a pre-defined (possibly infinite dimensional) feature vector for the input variable $x \in \mathcal{X}$, and $\langle \cdot, \cdot \rangle$ denotes an inner product in the feature vector space.

Regularized ERM, with L_2 Regularization

$$\hat{w} = \arg\min_{w} \left[\frac{1}{n} \sum_{i=1}^{n} L(\langle w, \psi(X_i) \rangle, Y_i) + \frac{\lambda}{2} \|w\|^2 \right], \tag{2}$$

which employs the linear function class of (1).

Kernel

Given feature map $\psi(x)$, we define its kernel function:

$$k(x, x') = \langle \psi(x), \psi(x') \rangle. \tag{3}$$

Given training data $\{(X_i, Y_i)\}$, we define kernel Gram matrix

$$K_{n \times n} = \begin{bmatrix} k(X_1, X_1) & \cdots & k(X_1, X_n) \\ \cdots & \cdots & \cdots \\ k(X_n, X_1) & \cdots & k(X_n, X_n) \end{bmatrix}. \tag{4}$$

It is easy to check that the kernel Gram matrix $K_{n \times n}$ is always positive-semidefinite.

Kernel Trick

Proposition 1 (Prop 9.1)

Assume that (3) *holds. If w has a representation*

$$w = \sum_{i=1}^{n} \alpha_i \psi(x_i), \quad \alpha = \begin{bmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{bmatrix}, \tag{5}$$

then $f(x) = \langle w, \psi(x) \rangle \in \mathcal{F}$ *of* (1) *satisfies*

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \tag{6}$$

$$\langle w, w \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K_{n \times n} \alpha. \tag{7}$$

The reverse is also true. If $f(x)$ satisfies (6), then with w defined by (5), we have $f(x) = \langle w, \psi(x) \rangle$, and (7) holds.

Proof of Proposition 1

Consider $f(x) = \langle w, \psi(x) \rangle$. If (5) holds, then

$$f(x) = \langle w, \psi(x) \rangle = \sum_{i=1}^{n} \alpha_i \langle \psi(x_i), \psi(x) \rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

Moreover,

$$\langle w, w \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \psi(x_i), \psi(x_j) \rangle.$$

This implies (7). Similarly the reverse direction holds.

Consequence of Kernel Trick

Theorem 2 (Representer Theorem, Thm 9.2)

For real valued functions $f(x)$, the solution of (2) *has the following kernel representation:*

$$\langle \hat{w}, \psi(x) \rangle = \bar{f}(\hat{\alpha}, x), \qquad \bar{f}(\hat{\alpha}, x) = \sum_{i=1}^{n} \hat{\alpha}_i k(X_i, x).$$

Therefore the solution of (2) *is equivalent to the solution of the following finite dimensional kernel optimization problem:*

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \left[\frac{1}{n} \sum_{i=1}^{n} L\left(\bar{f}(\alpha, X_i), Y_i\right) + \frac{\lambda}{2} \alpha^\top K_{n\times n} \alpha \right]. \tag{8}$$

Proof of Theorem 2 (I/II)

Let

$$Q_1(w) = \frac{1}{n}\sum_{i=1}^{n} L(\langle w, \psi(X_i)\rangle, Y_i) + \frac{\lambda}{2}\|w\|^2$$

be the objective function of (2), and let

$$Q_2(\alpha) = \frac{1}{n}\sum_{i=1}^{n} L\left(\bar{f}(\alpha, X_i), Y_i\right) + \frac{\lambda}{2}\alpha^\top K_{n\times n}\alpha$$

be the objective function of (8).
The solution of (2) satisfies the following first order optimality condition:

$$\frac{1}{n}\sum_{i=1}^{n} L_1'(\langle \hat{w}, \psi(X_i)\rangle, Y_i)\psi(X_i) + \lambda\hat{w} = 0.$$

Here $L_1'(p, y)$ is the derivative of $L(p, y)$ with respect to p.

Proof of Theorem 2 (II/II)

We thus obtain the following representation as its solution:

$$\hat{w} = \sum_{i=1}^{n} \tilde{\alpha}_i \psi(X_i),$$

where

$$\tilde{\alpha}_i = -\frac{1}{\lambda n} L_1'(\langle \hat{w}, \psi(X_i) \rangle, Y_i) \qquad (i = 1, \ldots, n).$$

Using this notation, we obtain from Proposition 1 that

$$\langle \hat{w}, \psi(x) \rangle = \bar{f}(\tilde{\alpha}, x), \quad \langle \hat{w}, \hat{w} \rangle = \tilde{\alpha}^\top K_{n \times n} \tilde{\alpha}.$$

This implies that

$$Q_1(\hat{w}) = Q_2(\tilde{\alpha}) \geq Q_2(\hat{\alpha}) = Q_1(\tilde{w}),$$

where the last equality follows by setting $\tilde{w} = \sum_{i=1}^{n} \hat{\alpha}_i \psi(X_i)$. Proposition 1 implies that $Q_2(\hat{\alpha}) = Q_1(\tilde{w})$. It follows that $\tilde{w}$ is a solution of (2), which proves the desired result.

Example 3 (Kernel Ridge Regression, Expl 9.9)

Consider ridge regression in the feature space representation:

$$\hat{w} = \arg\min_{w} \left[\frac{1}{n} \sum_{i=1}^{n} (\langle w, \psi(X_i) \rangle - Y_i)^2 + \frac{\lambda}{2} \langle w, w \rangle \right].$$

The primal kernel formulation is:

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} k(X_i, X_j)\alpha_j - Y_i \right)^2 + \frac{\lambda}{2} \alpha^\top K_{n \times n} \alpha \right].$$

There is also a dual formulation which has the same solution:

$$\hat{\alpha} = \arg\max_{\alpha \in \mathbb{R}^n} \left[-\frac{\lambda}{2} \alpha^\top K_{n \times n} \alpha + \lambda \alpha^\top \mathbf{Y} - \frac{\lambda^2}{4} \alpha^\top \alpha \right],$$

where $\mathbf{Y}$ is the n dimensional vector with Y_i as its component.

Positive-definite Kernel

Definition 4

A symmetric function $k(x, x')$ is called a positive-definite kernel on $\mathcal{X} \times \mathcal{X}$ if for all $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in \mathcal{X}$, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Reproducing Kernel Hilbert Space (RKHS)

Definition 5 (RKHS, Def 9.4)

Given a symmetric positive-definite kernel, we define a function space $\mathcal{H}_0$ of the form

$$\mathcal{H}_0 = \left\{ f(x) : f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) \right\},$$

with inner product defined as

$$\|f(x)\|_{\mathcal{H}}^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j).$$

The completion of $\mathcal{H}_0$ with respect to this inner product, defined as $\mathcal{H}$, is called the reproducing kernel Hilbert space (RKHS) of kernel k.

RKHS Norm is Well-Defined

Proposition 6 (Prop 9.5)

Assume that for all $x \in \mathcal{X}$:

$$\sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m'} \alpha'_i k(x'_i, x),$$

then

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i=1}^{m'} \sum_{j=1}^{m'} \alpha'_i \alpha'_j k(x'_i, x'_j).$$

The result means that even when a function $f(x)$ has two different kernel representations, the RKHS norm $\|f(x)\|_{\mathcal{H}}$ computed using the two representations are identical.

Mercer's Theorem

Theorem 7 (Thm 9.6)

A symmetric kernel function $k(x, x')$ *is positive-definite if and only if there exists a feature map* $\psi(x)$ *so that it can be written in the form of* (3)*. That is,*

$$k(x, x') = \langle \psi(x), \psi(x') \rangle.$$

Moreover, let $\mathcal{H}$ *be the RKHS of* $k(\cdot, \cdot)$*, then any function* $f(x) \in \mathcal{H}$ *can be written uniquely in the form of* (1)*, with*

$$\|f(x)\|_{\mathcal{H}}^2 = \langle w, w \rangle.$$

Example

Example 8

If $x \in \mathbb{R}^d$, then a standard choice of kernel is the RBF (radial basis function) kernel:

$$k(x, x') = \exp\left[\frac{-\|x - x'\|_2^2}{2\sigma^2}\right].$$

It is easy to check that it can be written in the form of (3) using Taylor expansion as:

$$k(x, x') = \exp\left[-\frac{\|x\|_2^2}{2\sigma^2}\right] \exp\left[-\frac{\|x'\|_2^2}{2\sigma^2}\right] \sum_{k=0}^{\infty} \frac{\sigma^{-2k}}{k!}(x^\top x')^k.$$

ERM in RKHS

In general, we can consider abstract ERM problem in any RKHS $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$.

Given an RKHS $\mathcal{H}$, one may consider a norm constrained ERM problem in $\mathcal{H}$ as follows:

$$\hat{f}(\cdot) = \arg\min_{f(\cdot)\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i) \quad \text{subject to } \|f(\cdot)\|_{\mathcal{H}} \leq A. \tag{9}$$

The corresponding soft-regularized formulation with appropriate $\lambda > 0$ is

$$\hat{f}(\cdot) = \arg\min_{f(\cdot)\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i) + \frac{\lambda}{2}\|f(\cdot)\|_{\mathcal{H}}^2\right]. \tag{10}$$

Equivalence Theorem

Theorem 9 (Thm 9.8)

Consider any kernel function $k(x, x')$ *and feature map* $\psi(x)$ *that satisfies* (3). *Let* $\mathcal{H}$ *be the RKHS of* $k(\cdot, \cdot)$. *Then any* $f(x) \in \mathcal{H}$ *can be written in the form*

$$f(x) = \langle w, \psi(x) \rangle, \qquad \|f(x)\|_{\mathcal{H}}^2 = \inf\{\langle w, w \rangle : f(x) = \langle w, \psi(x) \rangle\}.$$

Consequently, the solution of (10)

$$\hat{f}(\cdot) = \arg \min_{f(\cdot) \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i) + \frac{\lambda}{2} \|f(\cdot)\|_{\mathcal{H}}^2 \right]$$

is equivalent to the solution of (2)

$$\hat{w} = \arg \min_{w} \left[\frac{1}{n} \sum_{i=1}^{n} L(\langle w, \psi(X_i) \rangle, Y_i) + \frac{\lambda}{2} \|w\|^2 \right].$$

Example 10 (Expl 9.10)

Consider support vector machines for binary classification, where label $Y_i \in \{\pm 1\}$. Consider the following method in feature space:

$$\hat{w} = \arg\min_{w} \left[\frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - \langle w, \psi(X_i) \rangle Y_i) + \frac{\lambda}{2} \langle w, w \rangle \right].$$

The primal kernel formulation is:

$$\hat{w} = \arg\min_{\alpha} \left[\frac{1}{n} \sum_{i=1}^{n} \max \left(0, 1 - \sum_{j=1}^{n} \alpha_j k(X_i, X_j) Y_i \right) + \frac{\lambda}{2} \alpha^\top K_{n\times n} \alpha \right].$$

The equivalent dual kernel formulation is:

$$\hat{\alpha} = \arg\max_{\alpha \in \mathbb{R}^n} \left[-\frac{\lambda}{2} \alpha^\top K_{n\times n} \alpha + \lambda \alpha^\top \mathbf{Y} \right], \quad \text{subject to } \alpha_i Y_i \in [0, 1/(\lambda n)].$$

Variation of Representer Theorem

Proposition 11 (Prop 9.11)

Let $\mathcal{H}$ be the RKHS of a kernel $k(x, x')$ defined on a discrete set of n points $X_1, \ldots, X_n$. Let $K_{n\times n}$ be the Gram matrix defined on these points in (4), and K^+ be its pseudo-inverse. Then for any function $f \in \mathcal{H}$, we have

$$\|f\|_{\mathcal{H}}^2 = \mathbf{f}^\top K_{n\times n}^+ \mathbf{f}, \qquad \textit{where} \quad \mathbf{f} = \begin{bmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{bmatrix}.$$

Proof

We can express $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$. Let $\alpha = [\alpha_1, \ldots, \alpha_n]^\top$, we have $\mathbf{f} = \mathbf{K_{n \times n}}\alpha$. It follows that

$$\|f\|_{\mathcal{H}}^2 = \alpha^\top K_{n\times n}\alpha = \alpha^\top K_{n\times n} K_{n\times n}^{+} K_{n\times n}\alpha = \mathbf{f}^\top K_{n\times n}^{+}\mathbf{f}.$$

This proves the desired result.

Semi-Supervised Learning Formulation

Corollary 12 (Cor 9.12)

Assume that we have labeled data $X_1, \ldots, X_n$, and unlabeled data $X_{n+1}, \ldots, X_{n+m}$. Let $K = K_{(n+m)\times(n+m)}$ be the kernel Gram matrix of a kernel k on these $m+n$ points, and let $\mathcal{H}$ be the corresponding RKHS. Then (10)

$$\hat{f}(\cdot) = \arg\min_{f(\cdot)\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i) + \frac{\lambda}{2}\|f(\cdot)\|_{\mathcal{H}}^2\right]$$

defined on these data points is equivalent to

$$\hat{f}(\cdot) = \arg\min_{\mathbf{f}\in\mathbb{R}^{n+m}} \left[\frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i) + \frac{\lambda}{2}\mathbf{f}^\top K^{+}\mathbf{f}\right], \quad \mathbf{f} = \begin{bmatrix} f(X_1) \\ \vdots \\ f(X_{n+m}) \end{bmatrix}.$$

Universal Approximation

Definition 13

A kernel $k(x, x')$ is called a universal kernel on $\mathcal{X} \subset \mathbb{R}^d$ (under the uniform convergence topology) if for any continuous function $f(x)$ on $\mathcal{X}$, and any $\epsilon > 0$, there exists $g(x) \in \mathcal{H}$ such that

$$\forall x \in \mathcal{X} : |f(x) - g(x)| \leq \epsilon,$$

where $\mathcal{H}$ is the RKHS of kernel $k(\cdot, \cdot)$.

Theorem 14 (Approximation of Lipschitz Functions, Thm 9.14)

Consider a positive definite translation invariant kernel

$$k(x, x') = h(\|x - x'\|/\sigma),$$

where $\|\cdot\|$ *is a norm on* $\mathbb{R}^d$*. Assume that* $h(\cdot) \in [0, 1]$*, and*

$$c_0 = \int h(\|x\|)dx \in (0, \infty), \qquad c_1 = \int \|x\|h(\|x\|)dx < \infty.$$

Assume that f *is Lipschitz with respect to the norm* $\|\cdot\|$*:* $\exists\gamma > 0$ *such that* $|f(x) - f(x')| \leq \gamma\|x - x'\|$ *for all* $x, x' \in \mathbb{R}^d$*. If*

$$\|f\|_1 = \int |f(x)|dx < \infty,$$

then for any $\epsilon > 0$ *and* $\sigma = \epsilon c_0/(\gamma c_1)$*, there exists* $\psi_\sigma(x) \in \mathcal{H}$*, where* $\mathcal{H}$ *is the RKHS of* $k(\cdot)$*, so that* $\|\psi_\sigma(x)\|_{\mathcal{H}} \leq (c_0\sigma^d)^{-1}\|f\|_1$ *and*

$$\forall x : |f(x) - \psi_\sigma(x)| \leq \epsilon.$$

Approximation Using Polynomials

Theorem 15 (Thm 9.15)

Consider a compact set $\mathcal{X}$ in $\mathbb{R}^d$. Assume that a kernel function $k(x, x')$ on $\mathcal{X} \times \mathcal{X}$ has a feature representation

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

where each $\psi_i(x)$ is a real valued function, and $c_i > 0$. Assume the feature maps $\{\psi_i(x) : i = 1, \ldots\}$ contain all monomials of the form

$$\left\{ g(x) = \prod_{j=1}^{d} x_j^{\alpha_j} : x = [x_1, \ldots, x_d], \alpha_j \geq 0 \right\}.$$

Then $k(x, x')$ is universal on $\mathcal{X}$.

Proof

Let $\mathcal{H}$ be the RKHS of $k(\cdot, \cdot)$. Note that according to Theorem 9, a function of the form $g(x) = \sum_{j=1}^{\infty} w_i \psi_i(x)$ has RKHS norm as

$$\|g\|_{\mathcal{H}}^2 \leq \sum_{i=1}^{\infty} w_i^2 / c_i.$$

It follows from the assumption of the theorem that all monomials $p(x)$ has RKHS norm $\|p\|_{\mathcal{H}}^2 < \infty$. Therefore $\mathcal{H}$ contains all polynomials. The result of the theorem is now a direct consequence of the Stone-Weierstrass theorem.

Example

Example 16 (Expl 9.16)

Let $\alpha > 0$ be an arbitrary constant. Consider the kernel function

$$k(x, x') = \exp(\alpha x^\top x')$$

on a compact set of $\mathbb{R}^d$. Since

$$k(x, x') = \exp(-\alpha) \sum_{i=0}^{\infty} \frac{\alpha^i}{i!} (x^\top x' + 1)^i.$$

It is clear that the expansion of $(x^\top x' + 1)^i$ contains all monomials of order i. Therefore Theorem 15 implies that $k(x, x')$ is universal.

Compositions of Universal Kernels

Theorem 17 (Thm 9.17)

Assume $k(x, x')$ is a universal kernel on $\mathcal{X}$. Let $k'(x, x')$ be any other kernel function on $\mathcal{X} \times \mathcal{X}$, then $k(x, x') + k'(x, x')$ is a universal kernel on $\mathcal{X}$.
Moreover, let $u(x)$ be a real-valued continuous function on $\mathcal{X}$ so that

$$\sup_{x \in \mathcal{X}} u(x) < \infty, \qquad \inf_{x \in \mathcal{X}} u(x) > 0.$$

Then $k'(x, x') = k(x, x')u(x)u(x')$ is a universal kernel on $\mathcal{X}$.

Proof of Theorem 17

Let $k(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}}$ with the corresponding RKHS denoted by $\mathcal{H}$, and let $k'(x, x') = \langle \psi'(x), \psi'(x') \rangle_{\mathcal{H}'}$ with RKHS $\mathcal{H}'$.

$$k(x, x') + k'(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}} + \langle \psi'(x), \psi'(x') \rangle_{\mathcal{H}'}.$$

Using feature representation, we can represent functions in the RKHS of $k(x, x') + k'(x, x')$ by $\langle w, \psi(x) \rangle_{\mathcal{H}} + \langle w', \psi'(x) \rangle_{\mathcal{H}'}$, and thus it contains $\mathcal{H} \oplus \mathcal{H}'$. This implies the first result.

For the second result, we know that $k'(x, x') = \langle \psi(x)u(x), \psi(x')u(x') \rangle_{\mathcal{H}}$, and thus its RHKS can be represented by $\langle w, \psi(x)u(x) \rangle_{\mathcal{H}}$. Since the universality of $k(x, x')$ implies that for any continuous $f(x)$, $f(x)/u(x)$ can be uniformly approximated by $\langle w, \psi(x) \rangle_{\mathcal{H}}$, we obtain the desired result.

Example

Example 18

Consider the RBF kernel function

$$k(x, x') = \exp(-\alpha \|x - x'\|_2^2).$$

Since

$$k(x, x') = \exp(2\alpha x^\top x')u(x)u(x'),$$

where $u(x) = \exp(-\alpha\|x\|_2^2)$, Theorem 17 and Example 16 imply that $k(x, x')$ is universal on any compact set $\mathcal{X} \subset \mathbb{R}^d$.

Property of Universal Kernel

Theorem 19 (Thm 9.19)

Let $k(x, x')$ be a universal kernel on $\mathcal{X}$. Consider n different data points $X_1, \ldots, X_n \in \mathcal{X}$, and let $K_{n \times n}$ be the Gram matrix defined in Theorem 2. Then $K_{n \times n}$ is full-rank.

Generalization Analysis: Constrained RKHS

Consider feature representation

$$f(x) = \langle w, \psi(x) \rangle,$$

with the induced RKHS. Theorem 9 implies the following.

Equivalent Representations

If we define the function class

$$\mathcal{F}(A) = \{f(x) \in \mathcal{H} : \|f\|_{\mathcal{H}}^2 \leq A^2\},$$

then for any feature map that satisfies (3), $\mathcal{F}(A)$ can be equivalently written in the linear feature representation form as:

$$\mathcal{F}(A) = \{f(x) = \langle w, \psi(x) \rangle : \langle w, w \rangle \leq A^2\}. \tag{11}$$

That is, a function with RKHS regularization is equivalent to linear model with L_2 regularization.

Rademacher Complexity

Theorem 20 (The First Inequality of Thm 9.20)

Consider $\mathcal{F}(A)$ defined in (11). We have the following bound for its Rademacher complexity:

$$R(\mathcal{F}(A), \mathcal{S}_n) \leq A\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} k(X_i, X_i)}.$$

Proof of Theorem 20

For convenience, let $\|w\| = \sqrt{\langle w, w \rangle}$. We have

$$
\begin{aligned}
R_\lambda =& \mathbb{E}_\sigma \sup_w \left[\frac{1}{n} \sum_{i=1}^n \sigma_i \langle w, \psi(X_i) \rangle - \frac{\lambda}{4} \langle w, w \rangle \right] \\
=& \mathbb{E}_\sigma \sup_w \left[\langle w, \frac{1}{n} \sum_{i=1}^n \sigma_i \psi(X_i) \rangle - \frac{\lambda}{4} \langle w, w \rangle \right] \\
=& \mathbb{E}_\sigma \frac{1}{\lambda} \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \psi(X_i) \right\|^2 \\
=& \frac{1}{\lambda n^2} \sum_{i=1}^n \|\psi(X_i)\|^2 = \frac{1}{\lambda n^2} \sum_{i=1}^n k(X_i, X_i).
\end{aligned}
$$

This proves the second bound. For the first bound, we note that

$$
R(\mathcal{F}(A), \mathcal{S}_n) \leq R_\lambda + \frac{\lambda A^2}{4} \leq \frac{1}{\lambda n^2} \sum_{i=1}^n k(X_i, X_i) + \frac{\lambda A^2}{4}.
$$

Optimize over $\lambda > 0$, we obtain the desired result.

Lipschitz Loss

Corollary 21 (The First Two Inequalities of Cor 9.21)

Let $\mathcal{G}(A) = \{L(f(x), y) : f(x) \in \mathcal{F}(A)\}$*, where* $\mathcal{F}(A)$ *is defined in* (11)*. If* $L(p, y)$ *is* γ *Lipschitz in p, then*

$$R(\mathcal{G}(A), \mathcal{S}_n) \leq A\gamma \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} k(X_i, X_i)},$$

$$R_n(\mathcal{G}(A), \mathcal{D}) \leq A\gamma \sqrt{\frac{\mathbb{E}_{X \sim \mathcal{D}} k(X, X)}{n}}.$$

Result used in the Proof of Corollary 21

Theorem 22 (Rademacher Comparison, Thm 6.28)

Let $\{\phi_i\}_{i=1}^n$ *be functions with Lipschitz constants* $\{\gamma_i\}_{i=1}^n$ *respectively. That is,* $\forall i \in [n]$*:*

$$|\phi_i(\theta) - \phi_i(\theta')| \leq \gamma_i |\theta - \theta'|.$$

Then

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^n \sigma_i \phi_i(f(Z_i)) \right] \leq \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^n \sigma_i \gamma_i f(Z_i) \right].$$

Proof of Corollary 21

The first inequality follows from Theorem 20 and the Rademacher comparison theorem in Theorem 22.
The second inequality follows from the following derivation:

$$
\begin{aligned}
R_n(\mathcal{G}(A), \mathcal{D}) =& \mathbb{E}_{\mathcal{S}_n} R(\mathcal{G}, \mathcal{S}_n) \leq A\gamma \mathbb{E}_{\mathcal{S}_n} \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} k(X_i, X_i)} \\
\overset{(a)}{\leq}& A\gamma \sqrt{\frac{1}{n^2} \mathbb{E}_{\mathcal{S}_n} \sum_{i=1}^{n} k(X_i, X_i)} \\
=& A\gamma \sqrt{\frac{1}{n} \mathbb{E}_{\mathcal{D}} k(X, X)}.
\end{aligned}
$$

The derivation of (a) used Jensen's inequality and the concavity of $\sqrt{\cdot}$.

Uniform Convergence and Oracle Inequality

Corollary 23 (Cor 9.22)

Assume that $\sup[L(p,y) - L(p',y')] \leq M$, *and* $L(p,y)$ *is* γ *Lipschitz with respect to p. Then with probability at least* $1-\delta$*: for all* $f \in \mathcal{H}$ *with* $\|f\|_{\mathcal{H}} \leq A$*:*

$$\mathbb{E}_{\mathcal{D}} L(f(X), Y) \leq \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i) + 2\gamma A \sqrt{\frac{\mathbb{E}_{\mathcal{D}} k(X,X)}{n}} + M \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Moreover, for (9)*, if we solve it approximately up to sub-optimality of* ϵ'*, then we have with probability at least* $1-\delta$*:*

$$\mathbb{E}_{\mathcal{D}} L(\hat{f}(X), Y) \leq \inf_{\|f\|_{\mathcal{H}} \leq A} \mathbb{E}_{\mathcal{D}} L(f(X), Y) + \epsilon' + 2\gamma A \sqrt{\frac{\mathbb{E}_{\mathcal{D}} k(X,X)}{n}} + M \sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Consistency

In Corollary 23, as $A \to \infty$, we have

$$\inf_{\|f\|_{\mathcal{H}} \leq A} \mathbb{E}_{\mathcal{D}} L(f(X), Y) \to \inf_{\|f\|_{\mathcal{H}} < \infty} \mathbb{E}_{\mathcal{D}} L(f(X), Y).$$

If $k(\cdot, \cdot)$ is a universal kernel, then

$$\lim_{A \to \infty} \inf_{\|f\|_{\mathcal{H}} \leq A} \mathbb{E}_{\mathcal{D}} L(f(X), Y) \to \inf_{\text{measurable } f} \mathbb{E}_{\mathcal{D}} L(f(X), Y).$$

Combine this with the generalization result of kernel method in Corollary 23, we know that as $n \to \infty$, and let $A \to \infty$, the following result is valid.

Consistency

With probability 1,

$$\mathbb{E}_{\mathcal{D}} L(\hat{f}(X), Y) \to \inf_{\text{measurable } f} \mathbb{E}_{\mathcal{D}} L(f(X), Y).$$

Example 24 (Rademacher Complexity Margin Bound)

For binary classification problem with $y \in \{\pm 1\}$, we consider classifier induced by a real valued function $f(x)$: predict $y = 1$ if $f(x) \geq 0$ and $y = -1$ otherwise. If $f(x)$ is taken from an RKHS, then with probability $1 - \delta$, for all $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq A$:

$$\mathbb{E}_{\mathcal{D}} \mathbb{1}(f(X)Y \leq 0) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(f(X_i)Y_i \leq \gamma) + \frac{2A}{\gamma} \sqrt{\frac{\mathbb{E}_{\mathcal{D}} k(X,X)}{n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

It says that if we can find a classifier with a small margin error, then we can achieve a good test classification error.

The bound can be obtained as a direct consequence of Corollary 23, using a loss function $L(p,y) = \min(1, \max(0, 1 - py/\gamma))$, which is γ^{-1} Lipschitz. In this case, $\mathbb{1}(f(x)y \leq 0) \leq L(f(x), y) \leq \mathbb{1}(f(x)y \leq \gamma)$.

Example: SVM Loss

Example 25

For SVM loss, $\gamma = 1$. With hard regularization, we can take $M = (1 + AB)$, where we assume that $k(x, x) \leq B^2$. Consider $\hat{f}$ that solves (9), which we restate here as

$$\hat{f}(\cdot) = \arg \min_{f(\cdot)\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i) \quad \text{subject to } \|f(\cdot)\|_{\mathcal{H}} \leq A,$$

up to an accuracy of $\epsilon' > 0$. From Corollary 23, we obtain with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathcal{D}} L(\hat{f}(X), Y) \leq \inf_{\|f\|_{\mathcal{H}} \leq A} \mathbb{E}_{\mathcal{D}} L(f(X), Y) + \epsilon' + \frac{2AB}{\sqrt{n}} + (1 + AB)\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Vector Valued Functions

We now consider vector valued functions (such as multi-class classification) using kernels.

Feature Space Representation of Vector Valued Functions

Consider $f(x) : \mathcal{X} \to \mathbb{R}^q$ for some $q > 1$. Let $f(x) = [f_1(x), \ldots, f_q(x)]$, then

$$f_\ell(x) = \langle w, \psi(x, \ell) \rangle. \tag{12}$$

Similar to (2), we have the following formulation in feature representation:

$$\hat{w} = \arg\min_{w \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} L(\langle w, \psi(X_i, \cdot) \rangle, Y_i) + \frac{\lambda}{2} \|w\|^2 \right], \tag{13}$$

where $\langle w, \psi(X_i, \cdot) \rangle$ denotes the q-dimensional vector with $\langle w, \psi(X_i, \ell) \rangle$ as its ℓ-th component.

Matrix Kernel for Vector-valued Function

The matrix kernel function can be defined:

$$k_{i,j}(x, x') = \langle \psi(x, i), \psi(x', j) \rangle \qquad (i, j = 1, \ldots, q).$$

and its matrix representation is

$$\mathbf{k}(x, x') = \begin{bmatrix} k_{1,1}(x, x') & \cdots & k_{1,q}(x, x') \\ \vdots & & \vdots \\ k_{q,1}(x, x') & \cdots & k_{q,q}(x, x') \end{bmatrix}.$$

The kernel Gram matrix becomes

$$\begin{bmatrix} \mathbf{k}(X_1, X_1) & \cdots & \mathbf{k}(X_1, X_q) \\ \vdots & & \vdots \\ \mathbf{k}(X_q, X_q) & \cdots & \mathbf{k}(X_q, X_q) \end{bmatrix}.$$

Vector Representer Theorem

Theorem 26 (Thm 9.29)

Consider q-dimensional vector valued function $f(x)$*. Let* $\hat{f}(x) = \langle \hat{w}, \psi(x, \cdot) \rangle$ *with* $\hat{w}$ *being the solution of* (13)*. Then*

$$\hat{f}(x) = \sum_{i=1}^{n} \mathbf{k}(X_i, x)\hat{\alpha}_i,$$

$$\langle \hat{w}, \hat{w} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\alpha}_i^\top \mathbf{k}(X_i, X_j)\hat{\alpha}_j.$$

Therefore the solution of (13) *is equivalent to*

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{R}^{q \times n}} \left[\frac{1}{n} \sum_{i=1}^{n} L\left(\sum_{j=1}^{n} \mathbf{k}(X_i, X_j)\alpha_j, Y_i \right) + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i^\top \mathbf{k}(X_i, X_j)\alpha_j \right].$$

Multi-class Classification Example

Example 27 (Structured SVM Loss, Expl 9.32)

Consider the structured SVM loss function [Tsochantaridis et al., 2005] for q-class classification problem, with $y \in \{1, \ldots, q\}$, and for $f \in \mathbb{R}^q$:

$$L(f, y) = \max_{\ell}[\gamma(y, \ell) - (f_y - f_\ell)],$$

where $\gamma(y, y) = 0$ and $\gamma(y, \ell) \geq 0$. This loss tries to separate the true class y from alternative $\ell \neq y$ with margin $\gamma(y, \ell)$. It is Lipschitz with respect to $\|f\|_1$ with $\gamma_1 = 1$. For problems with $k_{\ell,\ell}(x, x) \leq B^2$ for all x and ℓ, we have from Corollary 9.31 that

$$R(\mathcal{G}, \mathcal{S}_n) \leq \frac{qAB}{\sqrt{n}}.$$

This result employs multi-class Rademacher comparison result in Corollary 9.31, leading to a Rademacher complexity bound of $O(\sqrt{q})$.

Multi-class Classification Example (cont)

Proposition 28 (Prop 9.33)

Consider a loss function $L(f, y)$ that is γ_∞-Lipschitz in p with respect to the L_∞-norm:

$$|L(p, y) - L(p', y)| \leq \gamma_\infty \|p - p'\|_\infty.$$

Let $\mathcal{F} = \{f(x) = [f_1(x), \ldots, f_q(x)] : f_\ell(x) = \langle w, \psi(x, \ell)\rangle,\ \langle w, w\rangle \leq A^2\}$. Assume that $\sup_{x,\ell}\langle \psi(x, \ell), \psi(x, \ell)\rangle \leq B^2$. Let $\mathcal{G} = \{L(f, y) : f \in \mathcal{F}\}$. Then there exists a constant $c_0 > 0$ such that

$$R(\mathcal{G}, \mathcal{S}_n) \leq \frac{c_0 \gamma_\infty AB \ln n \sqrt{\ln(nq)}}{\sqrt{n}}.$$

This result requires the empirical L_∞ covering number estimate of L_2 regularized linear functions in Theorem 5.20.

Summary (Chapter 9)

- Reproducing Kernel Hilbert Space
- Universal Approximation
- Generalization and Rademacher Complexity
- Vector-valued Functions.