Machine learning theory

Regression

Hamid Beigy

Sharif university of technology

May 27, 2023

Table of contents

Introduction

1. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ a measurable subset of $\mathbb{R}$ and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$.
2. Learner receives sample $S = \{(x_1, y_m), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ drawn i.i.d. according to $\mathcal{D}$.
3. Let $L : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$ be the loss function used to measure the magnitude of error.
4. The most used loss function is
   - L_2 defined as $L(y, y') = |y' - y|^2$ for all $y, y' \in \mathcal{Y}$.
   - L_p defined as $L(y, y') = |y' - y|^p$ for all $p \geq 1$ and $y, y' \in \mathcal{Y}$.

The regression problem is defined as

Definition (Regression problem)

Given a hypothesis set $H = \{h : \mathcal{X} \mapsto \mathcal{Y} \mid h \in H\}$, regression problem consists of using labeled sample S to find a hypothesis $h \in H$ with small generalization error $\mathbf{R}(h)$ respect to target f:

$$\mathbf{R}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [L(h(x), y)]$$

The empirical loss or error of $h \in H$ is denoted by

$$\hat{\mathbf{R}}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i)$$

If $L(y, y) \leq M$ for all $y, y' \in \mathcal{Y}$, problem is called bounded regression problem.

Generalization bounds

Theorem (Generalization bounds for finite hypothesis sets)

Let $L \leq M$ be a bounded loss function and the hypothesis set H is finite. Then, for any $\delta > 0$, with probability at least $(1-\delta)$, the following inequality holds for all $h \in H$

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + M\sqrt{\frac{\log|H| + \log\frac{1}{\delta}}{2m}}.$$

Proof (Generalization bounds for finite hypothesis sets).

By Hoeffding's inequality, since $L \in [0, M]$, for any $h \in H$, the following holds

$$\mathbb{P}\left[\mathbf{R}(h) - \hat{\mathbf{R}}(h) > \epsilon\right] \leq \exp\left(-2\frac{m\epsilon^2}{M^2}\right).$$

Thus, by the union bound, we can write

$$\begin{aligned}\mathbb{P}\left[\exists h \in H \,\middle|\, \mathbf{R}(h) - \hat{\mathbf{R}}(h) > \epsilon\right] &\leq \sum_{h \in H} \mathbb{P}\left[\mathbf{R}(h) - \hat{\mathbf{R}}(h) > \epsilon\right] \\ &\leq |H| \exp\left(-2\frac{m\epsilon^2}{M^2}\right).\end{aligned}$$

Setting the right-hand side to be equal to δ, the theorem will proved. $\square$

Theorem (Rademacher complexity of μ-Lipschitz loss functions)

Let $L \leq M$ be a bounded loss function such that for any fixed $y' \in \mathcal{Y}$, $L(y, y')$ is μ-Lipschitz for some $\mu > 0$. Then for any sample $S = \{(x_1, y_m), \ldots, (x_m, y_m)\}$, the upper bound of the Rademacher complexity of the family $\mathcal{G} = \{(x, y) \mapsto L(h(x), y) \mid h \in H\}$ is

$$\hat{\mathcal{R}}(\mathcal{G}) \leq \mu \hat{\mathcal{R}}(H).$$

Lemma (Talagrand's Lemma (special case))

Let ϕ be a μ-Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ and $\sigma_1, \ldots, \sigma_m$ be Rademacher random variables. Then, for any hypothesis set H of real-valued functions, the following inequality holds:

$$\hat{\mathcal{R}}(\phi \circ H) \leq \mu \hat{\mathcal{R}}(H).$$

Proof (Rademacher complexity of μ-Lipschitz loss functions).

Since for any fixed y_i, $L(y, y')$ is μ-Lipschitz for some $\mu > 0$, by Talagrand's Lemma, we can write

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{G}) &= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{m}\sigma_i L(h(x_i), y_i)\right] \\ &\leq \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{m}\sigma_i \mu h(x_i)\right] \\ &= \mu\hat{\mathcal{R}}(H).\end{aligned}$$

□

Theorem (Rademacher complexity of L_p loss functions)

Let $p \geq 1$ and $\mathcal{G} = \{\mathbf{x} \mapsto |h(x) - f(x)|^p \mid h \in H\}$ and $|h(x) - f(x)| \leq M$ for all $x \in \mathcal{X}$ and $h \in H$. Then for any sample $S = \{(x_1, y_m), \ldots, (x_m, y_m)\}$, the following inequality holds

$$\hat{\mathcal{R}}(\mathcal{G}) \leq pM^{p-1}\hat{\mathcal{R}}(H).$$

Proof (Rademacher complexity of L_p loss functions).

Let $\phi_p : x \mapsto |x|^p$, then $\mathcal{G} = \{\phi_p \circ h \mid h \in H'\}$ where $H' = \{\mathbf{x} \mapsto h(x) - f(x) \mid h \in H\}$. Since ϕ_p is pM^{p-1}-Lipschitz over $[-M, M]$, we can apply Talagrand's Lemma,

$$\hat{\mathcal{R}}(\mathcal{G}) \le pM^{p-1}\hat{\mathcal{R}}(H').$$

Now, $\hat{\mathcal{R}}(H')$ can be expressed as

$$\begin{aligned}\hat{\mathcal{R}}(H') &= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H}\sum_{i=1}^{m}(\sigma_i h(\mathbf{x}_i) + \sigma_i f(\mathbf{x}_i))\right] \\ &= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i h(\mathbf{x}_i)\right] + \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{m}\sigma_i f(\mathbf{x}_i)\right] = \hat{\mathcal{R}}(H).\end{aligned}$$

Since $\mathbb{E}_\sigma\left[\sum_{i=1}^m \sigma_i f(\mathbf{x}_i)\right] = \sum_{i=1}^m \mathbb{E}_\sigma[\sigma_i]\, f(\mathbf{x}_i) = 0.$ □

Theorem (Rademacher complexity regression bounds)

Let $0 \leq L \leq M$ be a bounded loss function such that for any fixed $y' \in \mathcal{Y}$, $L(y, y')$ is μ-Lipschitz for some $\mu > 0$. Then,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L(h(x), y)] \leq \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i) + 2\mu\mathcal{R}_m(H) + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L(h(x), y)] \leq \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i) + 2\mu\hat{\mathcal{R}}(H) + 3M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Proof (Rademacher complexity of μ-Lipschitz loss functions).

Since for any fixed y_i, $L(y, y')$ is μ-Lipschitz for some $\mu > 0$, by Talagrand's Lemma, we can write

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{G}) &= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sum_{i=1}^{m} \sigma_i L(h(x_i), y_i) \right] \\ &\leq \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sum_{i=1}^{m} \sigma_i \mu h(x_i) \right] \\ &= \mu \hat{\mathcal{R}}(H).\end{aligned}$$

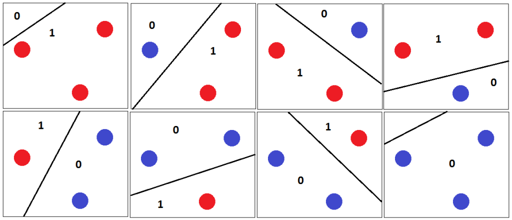
Combining this inequality with general Rademacher complexity learning bound completes proof. □

Pseudo-dimension bounds

1. VC dimension is a measure of complexity of a hypothesis set.

Definition (VC-dimension)

The Vapnik-Chervonenkis (VC) dimension of H, denoted as $VC(H)$, is the cardinality d of the largest set S shattered by H. If arbitrarily large finite sets can be shattered by H, then $VC(H) = \infty$.



2. We define shattering for families of real-valued functions.
3. Let $\mathcal{G}$ be a family of loss functions associated to some hypothesis set H, where

$$\mathcal{G} = \{z = (x, y) \mapsto L(h(x), y) \mid h \in H\}$$

Definition (Shattering)

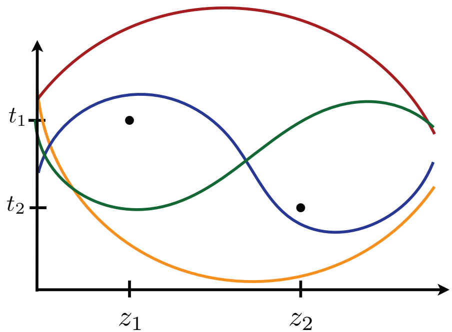
Let $\mathcal{G}$ be a family of functions from a set $\mathcal{Z}$ to $\mathbb{R}$. A set $\{z_1, \ldots, z_m\} \in (\mathcal{X} \times \mathcal{Y})$ is said to be shattered by $\mathcal{G}$ if there exists $t_1, \ldots, t_m \in \mathbb{R}$ such that

$$\left| \left\{ \begin{bmatrix} \operatorname{sgn}\left(g(z_1) - t_1\right) \\ \operatorname{sgn}\left(g(z_2) - t_2\right) \\ \vdots \\ \operatorname{sgn}\left(g(z_m) - t_m\right) \end{bmatrix} \middle| \, g \in \mathcal{G} \right\} \right| = 2^m$$

When they exist, the threshold values $t_1, \ldots, t_m$ are said to witness the shattering.

In other words, S is shattered by $\mathcal{G}$, if there are real numbers $t_1, \ldots, t_m$ such that for $b \in \{0, 1\}^m$, there is a function $g_b \in \mathcal{G}$ with $\operatorname{sgn}\left(g_b(\mathbf{x}_i) - t_i\right) = b_i$ for all $1 \leq i \leq m$.

1. Thus, $\{z_1, \ldots, z_m\}$ is shattered if for some witnesses $t_1, \ldots, t_m$, the family of functions $\mathcal{G}$ is rich enough to contain a function going
   - above a subset A of the set of points $\mathcal{J} = \{(z_i, t_i) \mid 1 \leq i \leq m\}$ and
   - below the others $\mathcal{J} - A$, for any choice of the subset A.



2. For any $g \in \mathcal{G}$, let B_g be the indicator function of the region below or on the graph of g, that is

$$B_g(\mathbf{x}, y) = \operatorname{sgn}\left(g(\mathbf{x}) - y\right).$$

3. Let $B_{\mathcal{G}} = \{B_g \mid g \in \mathcal{G}\}$.

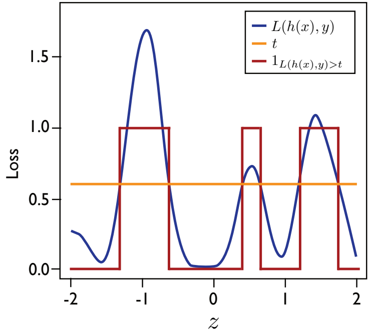
1. The notion of shattering naturally leads to definition of pseudo-dimension.

Definition (Pseudo-dimension)

Let $\mathcal{G}$ be a family of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then, the pseudo-dimension of $\mathcal{G}$, denoted by $Pdim(\mathcal{G})$, is the size of the largest set shattered by $\mathcal{G}$. If no such maximum exists, then $Pdim(\mathcal{G}) = \infty$.

2. $Pdim(\mathcal{G})$ coincides with VC of the corresponding thresholded functions mapping $\mathcal{X}$ to $\{0, 1\}$.

$$Pdim(\mathcal{G}) = VC\left(\{(x, t) \mapsto \mathbb{I}\left[(g(x) - t) > 0\right] \mid g \in \mathcal{G}\}\right)$$



3. Thus $Pdim(\mathcal{G}) = d$, if there are real numbers $t_1, \ldots, t_d$ and 2^d functions g_b that achieves all possible **below/above** combinations w.r.t t_i.

Theorem (Composition with non-decreasing function)

Suppose $\mathcal{G}$ *is a class of real-valued functions and* $\sigma : \mathbb{R} \mapsto \mathbb{R}$ *is a non-decreasing function. Define* $\sigma(\mathcal{G}) = \{\sigma \circ g \mid g \in \mathcal{G}\}$. *Then*

$$Pdim(\sigma(\mathcal{G})) \leq Pdim(\mathcal{G}).$$

Proof (Pseudo-dimension of hyperplanes).

1. For $d \leq Pdim(\sigma(\mathcal{G}))$, suppose set $\left\{\sigma \circ g_b \;\middle|\; b \in \{0,1\}^d\right\} \subseteq \sigma(\mathcal{G})$ shatters a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\} \subseteq \mathcal{X}$ witnessed by $(t_1, \ldots, t_d)$.
2. By suitably relabeling g_b, for all $\{0,1\}^d$ and $1 \leq i \leq d$, we have $\text{sgn}\left(\sigma(g_b(\mathbf{x}_i)) - t_i\right) = b_i$.
3. For all $1 \leq i \leq d$, take $y_i = \min\left\{g_b(\mathbf{x}_i) \;\middle|\; \sigma(g_b(\mathbf{x}_i)) \geq t_i, b \in \{0,1\}^d\right\}$.
4. Since σ is non-decreasing, it is straightforward to verify that $\text{sgn}\left(g_b(\mathbf{x}_i) - t_i\right) = b_i$ for all $\{0,1\}^d$ and $1 \leq i \leq d$

□

A class $\mathcal{G}$ of real-valued functions is **a vector space** if for all $g_1, g_2 \in \mathcal{G}$ and any numbers $\lambda, \mu \in \mathbb{R}$, we have $\lambda g_1 + \mu g_2 \in \mathcal{G}$.

Theorem (Pseudo-dimension of vector spaces)

If $\mathcal{G}$ is a vector space of real-valued functions, then $Pdim(\mathcal{G}) = dim(\mathcal{G})$.

Theorem (VC-dimension of vector spaces)

Let F be a vector space of real-valued functions, g is a real-valued function, and $H = \{sgn(f + g) \mid f \in F\}$. Then $VCdim(H) = dim(F)$.

Proof (Pseudo-dimension of vector spaces).

1. If $B_{\mathcal{G}}$ be class of **below the graph** indicator functions, then $Pdim(\mathcal{G}) = VC(B_{\mathcal{G}})$.
2. But $B_{\mathcal{G}} = \{(\mathbf{x}, y) \mapsto \text{sgn}\left(g(\mathbf{x}) - y\right) \mid g \in \mathcal{G}\}$.
3. Hence, functions $B_{\mathcal{G}}$ are of the form $\text{sgn}\left(g_1 + g_2\right)$, where
   - $g_1 = g$ is a function from vector space
   - g_2 is the fixed function $g_2(\mathbf{x}, y) = -y$.
4. Then, Theorem (VC-dimension of vector spaces) shows that $Pdim(\mathcal{G}) = dim(\mathcal{G})$.

□

Functions that map into some bounded range are not vector space.

Corollary

If $\mathcal{G}$ is a subset of a vector space $\mathcal{G}'$ of real valued functions then $Pdim(\mathcal{G}) \leq dim(\mathcal{G}')$

Theorem (Pseudo-dimension of hyperplanes)

Let $\mathcal{G} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b \mid \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$ be the class of hyperplanes in $\mathbb{R}^n$, then $Pdim(\mathcal{G}) = n + 1$.

Proof (Pseudo-dimension of hyperplanes).

1. It is easy to check that $\mathcal{G}$ is a vector space.
2. Let g_i be the ith coordinate projection $f_i(\mathbf{x}) = x_i$ for all $1 \leq i \leq n$ and $\mathbf{1}$ be identity-1 function. Then $B = \{g_1, \ldots, g_n, \mathbf{1}\}$ is basis of $\mathcal{G}$.
3. Hence, from Theorem (Pseudo-dimension of vector spaces), we obtain $Pdim(\mathcal{G}) = n + 1$

□

A polynomial transformation of $\mathbb{R}^n$ is $g(\mathbf{x}) = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_k\phi_k(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$, where k is an integer and for each $1 \leq i \leq k$, function $\phi_i(\mathbf{x})$ is defined as

$$\phi_i(\mathbf{x}) = \prod_{j=1}^{n} x_j^{r_{ij}}$$

for some nonnegative integers r_{ij} and $r_i = r_{i1} + r_{i2} + \ldots + r_{in}$ and the degree of g as $r = \max_i r_i$.

Theorem (Pseudo-dimension of polynomial transformation)

If $\mathcal{G}$ is a class of all polynomial transformations on $\mathbb{R}^n$ of degree at most r, then $Pdim(\mathcal{G}) = \binom{n+r}{r}$.

Theorem (Pseudo-dimension of all polynomial transformations)

Let $\mathcal{G}$ be class of all polynomial transformations on $\{0,1\}^n$ of degree at most r, then $Pdim(\mathcal{G}) = \sum_{i=0}^{r} \binom{n}{i}$.

Homework: Prove the above Theorems.

Theorem (Generalization bound for bounded regression)

Let H be a family of real-valued functions and $\mathcal{G} = \{z = (\mathbf{x}, y) \mapsto L(h(x), y) \mid h \in H\}$ be a family of loss functions associated to a hypothesis set H. Assume that $Pdim(\mathcal{G}) = d$ and loss function L is non-negative and bounded by M. Then, for any $\delta > 0$, with probability at least $(1-\delta)$ over the choice of an i.i.d. sample S of size m drawn from $\mathcal{D}^m$, the following inequality holds for all $h \in H$

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + M\sqrt{\frac{2d \log \frac{em}{d}}{m}} + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Proof (Generalization bound for bounded regression).

Homework: Prove this Theorem. □

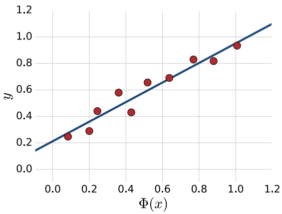
Regression algorithms

Regression algorithms

Linear regression

1. Let $\Phi : \mathcal{X} \mapsto \mathbb{R}^n$ and $H = \{h : \mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b \mid \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$.
2. Given sample S, the problem is to find a $h \in H$ such that

$$h = \min_{\mathbf{w},b} \hat{\mathbf{R}}(h) = \min_{\mathbf{w},b} \frac{1}{m} \sum_{i=1}^{m} \left(\langle \mathbf{w}, \Phi(x_i) \rangle + b - y_i\right)^2$$



3. Define data matrix $\mathbf{X} = \begin{bmatrix} \Phi(\mathbf{x}_1) & \phi(\mathbf{x}_2) & \dots & \phi(\mathbf{x}_m) \\ 1 & 1 & \dots & 1 \end{bmatrix}$.
4. Let $\mathbf{w} = (w_1, \dots, w_n, b)^T$ and $\mathbf{y} = (y_1, \dots, y_m)^T$ be weight and target vectors.
5. By setting $\nabla \hat{\mathbf{R}}(h) = 0$, we obtain

$$\mathbf{w} = (\mathbf{X}\mathbf{X}^T)^\dagger \mathbf{X}\mathbf{y}$$

6. When $\mathbf{X}\mathbf{X}^T$ is invertible, this problem has a unique solution; otherwise there are several solutions.

Theorem

Let $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ *be a PDS kernel,* $\Phi : \mathcal{X} \mapsto \mathbb{H}$ *a feature mapping associated to* K*, and* $H = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\right\}$*. Assume that there exists* $r > 0$ *such that* $K(\mathbf{x}, \mathbf{x}) \leq r^2$ *and* $M > 0$ *such that* $|h(\mathbf{x}) - y| < M$ *for all* $(\mathbf{x}, y \in \mathcal{X} \times \mathcal{Y})$*. Then for any* $\delta > 0$*, with probability at least* $(1 - \delta)$*, each of the following inequalities holds for all* $h \in H$*.*

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + 4M\sqrt{\frac{r^2\Lambda^2}{m}} + M^2\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \frac{4M\Lambda\sqrt{\mathrm{Tr}\left[\mathbf{K}\right]}}{m} + 3M^2\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Proof.

1. By the bound on the empirical Rademacher complexity of kernel-based hypotheses, the following holds for any sample S of size m:

$$\hat{\mathcal{R}}(H) \leq \frac{\Lambda\sqrt{\mathrm{Tr}\,[K]}}{m} \leq \sqrt{\frac{r^2\Lambda^2}{m}}$$

2. This implies that $\mathcal{R}_m(h) \leq \sqrt{\frac{r^2\Lambda^2}{m}}$.
3. Combining these inequalities with the bounds of Theorem Rademacher complexity regression bounds, the Theorem will be proved.

□

Regression algorithms

Kernel ridge regression

1. The following bound suggests minimizing a trade-off between empirical squared loss and norm of the weight vector.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + 4M\sqrt{\frac{r^2\Lambda^2}{m}} + M^2\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

2. Kernel ridge regression is defined by minimization of an objective function

$$\begin{aligned}\min_{\mathbf{w}} F(\mathbf{w}) &= \min_{\mathbf{w}} \left[\lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{m} (\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle - y_i)^2\right] \\ &= \min_{\mathbf{w}} \left[\lambda \|\mathbf{w}\|^2 + \left\|\mathbf{\Phi}^T\mathbf{w} - \mathbf{y}\right\|^2\right]\end{aligned}$$

3. By setting $\nabla F(\mathbf{w}) = 0$, we obtain $\mathbf{w} = (\mathbf{\Phi}\mathbf{\Phi}^T + \lambda\mathbf{I})^{-1}\mathbf{\Phi}\mathbf{y}$.

1. An alternative formulation of kernel ridge regression is

$$\min_{\mathbf{w}} \left\|\mathbf{\Phi}^T\mathbf{w} - \mathbf{y}\right\|^2 \text{ subject to } \|\mathbf{w}\|^2 \leq \Lambda^2$$

$$\min_{\mathbf{w}} \sum_{i=1}^{m} \xi_i^2 \text{ subject to } (\|\mathbf{w}\|^2 \leq \Lambda^2) \wedge (\forall i \in \{1, \ldots, m\}, \xi_i = y_i - \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle)$$

2. By using the Lagrangian method, we obtain

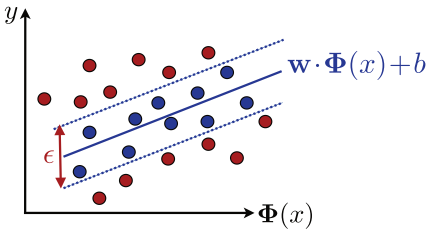
$$\mathbf{w} = \mathbf{\Phi}\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}.$$

3. Note that $(\mathbf{K} + \lambda\mathbf{I})^{-1}$ is invertible.
4. Therefore, the dual optimization problem as well as the primal optimization problem has a closed-form solution.

Regression algorithms

Support vector regression

1. Support vector regression (SVR) algorithm is inspired by SVM algorithm.
2. The main idea of SVR consists of fitting a tube of width $\epsilon > 0$ to the data.



3. This defines two sets of points:
   - points falling inside the tube, which are ϵ-close to the predicted function, not penalized,
   - points falling outside the tube are penalized based on their distance to the predicted function.
4. This is similar to the penalization used by SVMs in classification.
5. Using a hypothesis set of linear functions $H = \{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b \mid \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$, where Φ is the feature mapping corresponding some PDS kernel K.

1. The optimization problem for SVR is

$$\min_{\mathbf{w},b}\left[\frac{1}{2}\lambda\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\left|y_i - (\langle\mathbf{w},\Phi(\mathbf{x}_i)\rangle + b)\right|_\epsilon\right]$$

   where $|.|_\epsilon$ denotes ϵ-insensitive loss

$$\forall y, y' \in \mathcal{Y}, \quad |y' - y|_\epsilon = \max\left(0, |y' - y| - \epsilon\right)$$

2. The use of ϵ-insensitive loss leads to sparse solutions with a relatively small number of support vectors.

1. Using slack variables $\xi_i \geq 0$ and $\xi_i' \geq 0$ for $1 \leq i \leq m$, the problem becomes

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi'} & \left[\frac{1}{2}\lambda \|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i')\right] \\ \text{subject to } & (\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b) \leq \epsilon + \xi_i' \\ & \xi_i \geq 0, \quad \xi_i' \geq 0, \quad \forall i, 1 \leq i \leq m \end{aligned}$$

2. This is a convex quadratic program (QP) with affine constraints.
3. By introducing Lagrangian and applying KKT conditions, the problem will be solved.
4. Let $\mathcal{D}$ be the distribution according to which sample points are drawn.
5. Let $\hat{\mathcal{D}}$ the empirical distribution defined by a training sample of size m.

Theorem (Generalization bounds of SVR)

Let $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a PDS kernel, $\Phi : \mathcal{X} \mapsto \mathbb{H}$ a feature mapping associated to K, and $H = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \Phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\right\}$. Assume that there exists $r > 0$ suh that $K(\mathbf{x}, \mathbf{x}) \leq r^2$ and $M > 0$ such that $|h(\mathbf{x}) - y| < M$ for all $(\mathbf{x}, y \in \mathcal{X} \times \mathcal{Y})$. Then for any $\delta > 0$, with probability at least $(1 - \delta)$, each of the following inequalities holds for all $h \in H$.

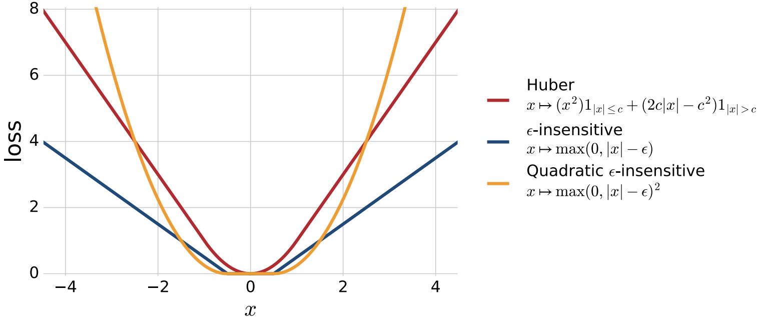
$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[|h(\mathbf{x}) - y|_\epsilon\right] \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\hat{\mathcal{D}}}\left[|h(\mathbf{x}) - y|_\epsilon\right] + 2\sqrt{\frac{r^2\Lambda^2}{m}} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[|h(\mathbf{x}) - y|_\epsilon\right] \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\hat{\mathcal{D}}}\left[|h(\mathbf{x}) - y|_\epsilon\right] + \frac{2\Lambda\sqrt{\mathrm{Tr}\left[\mathbf{K}\right]}}{m} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Proof (Generalization bounds of SVR).

Since for any $y' \in \mathcal{Y}$, the function $y \mapsto |y - y'|_\epsilon$ is 1-Lipschitz, the result follows Theorem Rademacher complexity regression bounds and the bound on the empirical Rademacher complexity of H. □

1. Alternative convex loss functions can be used to define regression algorithms.



2. SVR admits several advantages
   - SVR algorithm is based on solid theoretical guarantees,
   - The solution returned SVR is sparse
   - SVR allows a natural use of PDS kernels
   - SVR also admits favorable stability properties.
3. SVR also admits several disadvantages
   - SVR requires the selection of two parameters, C and ϵ, which are determined by cross-validation.
   - may be computationally expensive when dealing with large training sets.

Regression algorithms

Least absolute shrinkage and selection operator (Lasso)

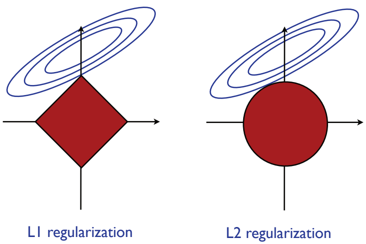
1. The optimization problem for Lasso is defined as

$$\min_{\mathbf{w},b} F(\mathbf{w}) = \min_{\mathbf{w},b} \left[\lambda \|\mathbf{w}\|_1 + C \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i)^2 \right]$$

2. This is a **convex optimization problem**, because
   - $\|\mathbf{w}\|_1$ is convex as with all norms
   - the empirical error term is convex
3. Hence, the optimization problem can be written as

$$\min_{\mathbf{w},b} \left[\sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i)^2 \right] \text{ subject to } \|\mathbf{w}\|_1 \leq \Lambda_1$$

4. The L_1 norm constraint is that it leads to a **sparse solution $\mathbf{w}$**.



Theorem (Bounds of $\hat{\mathcal{R}}(H)$ of Lasso)

Let $\mathcal{X} \subseteq \mathbb{R}^n$ and let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be sample of size m. Assume that for all $1 \leq i \leq m$, $\|\mathbf{x}_i\|_\infty \leq r_\infty$ for some $r_\infty > 0$, and let $H = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \;\middle|\; \|\mathbf{w}\|_1 \leq \Lambda_1\right\}$. Then, the empirical Rademacher complexity of H can be bounded as follows

$$\hat{\mathcal{R}}(H) \leq \sqrt{\frac{2r_\infty^2 \Lambda_1^2 \log(2n)}{m}}$$

Definition (Dual norms)

Let $\|.\|$ be a norm on $\mathbb{R}^n$. Then, dual norm $\|.\|_*$ associated to $\|.\|$ is defined by

$$\forall \mathbf{y} \in \mathbb{R}^n, \quad \|\mathbf{y}\|_* = \sup_{\|\mathbf{x}\|=1} |\langle \mathbf{y}, \mathbf{x} \rangle|$$

For any $p, q \geq 1$ that are conjugate $(\frac{1}{p} + \frac{1}{q} = 1)$, L_p and L_q norms are dual norms.
In particular, L_2 is dual norm of L_2, and L_1 is dual norm of L_∞ norm.

Proof (Bounds of $\hat{\mathcal{R}}(H)$ of Lasso)

1. For any $1 \le i \le m$, we denote by x_{ij}, the jth component of $\mathbf{x}_i$.

$$
\begin{aligned}
\hat{\mathcal{R}}(H) &= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\|\mathbf{w}\|_1 \le \Lambda_1} \sum_{i=1}^{m} \sigma_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right] \\
&= \frac{\Lambda_1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\left\| \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right\|_\infty \right] && \text{(by definition of the dual norm)} \\
&= \frac{\Lambda_1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\max_{j \in \{1,\ldots,n\}} \left| \sum_{i=1}^{m} \sigma_i x_{ij} \right| \right] && \text{(by definition of } \|.\|_\infty) \\
&= \frac{\Lambda_1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\max_{j \in \{1,\ldots,n\}} \max_{s \in \{-1,+1\}} s \sum_{i=1}^{m} \sigma_i x_{ij} \right] && \text{(by definition of } \|.\|_\infty) \\
&= \frac{\Lambda_1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\mathbf{z} \in A} \sum_{i=1}^{m} \sigma_i z_i \right].
\end{aligned}
$$

where A is set of n vectors $\{s(x_{1j}, \ldots, x_{mj}) \mid j \in \{1, \ldots, n\}, s \in \{-1, +1\}\}$.

Proof (Bounds of $\hat{\mathcal{R}}(H)$ of Lasso).

2. For any $\mathbf{z} \in A$, we have $\|\mathbf{z}\|_2 \leq \sqrt{mr_\infty^2} = r_\infty\sqrt{m}$.
3. Thus by Massart's Lemma, since A contains at most $2n$ elements, the following inequality holds:

$$\hat{\mathcal{R}}(H) \leq \Lambda_1 r_\infty \sqrt{m}\frac{2\log(2n)}{m} = \Lambda_1 r_\infty \sqrt{\frac{2\log(2n)}{m}}.$$

□

1. This bounds depends on dimension n is only logarithmic, which suggests that using very high-dimensional feature spaces does not significantly affect generalization.
2. By combining of Theorem (Bounds of $\hat{\mathcal{R}}(H)$ of Lasso) and Rademacher generalization bound, we can prove the following Theorem.

Theorem (Rademacher complexity of linear hypotheses with bounded L_1 norm)

Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $H = \left\{\mathbf{x}_1 \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \;\middle|\; \|\mathbf{w}\|_1 \leq \Lambda_1\right\}$. Let also $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be sample of size m. Assume that there exists $r_\infty > 0$ such that for all $\mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}_i\|_\infty \leq r_\infty$ and $M > 0$ such that $|h(\mathbf{x}) - y| \leq M$ for all $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. Then, for any $\delta > 0$, with probability at least $(1 - \delta)$, each of the following inequality holds for $h \in H$

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + 2r_\infty \Lambda_1 M \sqrt{\frac{2\log(2n)}{m}} + M^2 \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

1. Ridge regression and Lasso have same form as the right-hand side of this generalization bound.
2. Lasso has several advantages:
   - It benefits from strong theoretical guarantees and returns a sparse solution.
   - The sparsity of the solution is also computationally attractive (inner product).
   - The algorithm's sparsity can also be used for feature selection.
3. The main drawbacks are: usability of kernel and closed-form solution.

Regression algorithms

Online regression algorithms

1. The regression algorithms admit natural online versions.
2. These algorithms are useful when we have very large data sets, where a batch solution can be computationally expensive.

Online linear regression

1: Initialize $\mathbf{w}_1$.
2: **for** $t \leftarrow 1, 2, \ldots, T$ **do**.
3: Receive $\mathbf{x}_t \in \mathbb{R}^n$.
4: Predict $\hat{y}_t = \langle \mathbf{w}_t, \mathbf{x}_t \rangle$.
5: Observe true label $y_t = h^*(\mathbf{x}_t)$.
6: Compute the loss $L(\hat{y}_t, y_t)$. s
7: Update $\mathbf{w}_{t+1}$.
8: **end for**

Widrow-Hoff algorithm

1. Widrow-Hoff algorithm uses stochastic gradient descent technique to linear regression objective function.
2. At each round, the weight vector is augmented with a quantity that depends on the prediction error $(\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t)$.

WidrowHoff regression

function $\textsc{WidrowHoff}(\mathbf{w}_0)$
    Initialize $\mathbf{w}_1 \leftarrow \mathbf{w}_0$. ▷ typically $\mathbf{w}_0 = 0$.
    for $t \leftarrow 1, 2, \ldots, T$ **do**.
        Receive $\mathbf{x}_t \in \mathbb{R}^n$.
        Predict $\hat{y}_t = \langle \mathbf{w}_t, \mathbf{x}_t \rangle$.
        Observe true label $y_t = h^*(\mathbf{x}_t)$.
        Compute the loss $L(\hat{y}_t, y_t)$.
        Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - 2\eta \left(\langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t \right) \mathbf{x}_t$. ▷ learning rate $\eta > 0$.
    end for
    return $\mathbf{w}_{T+1}$
end function

1. There are two motivations for the update rule in Widrow-Hoff.
2. The first motivation is that
   - The loss function is defined as
   $$L(\mathbf{w}, \mathbf{x}, y) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$$
   - To minimize the loss function, move in the direction of the negative gradient
   $$\nabla_{\mathbf{w}} L(\mathbf{w}, \mathbf{x}, y) = 2\left(\langle \mathbf{w}, \mathbf{x} \rangle - y\right) \mathbf{x}$$
   - This gives the following update rule
   $$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla_{\mathbf{w}} L(\mathbf{w}_t, \mathbf{x}_t, y_t)$$
3. The second motivation is that we have two goals:
   - We want loss of $\mathbf{w}_{t+1}$ on $(\mathbf{x}_t, y_t)$ be small, which means we want to minimize $(\langle \mathbf{w}_{t+1}, \mathbf{x}_t \rangle - y_t)^2$.
   - We don't want $\mathbf{w}_{t+1}$ be too far from $\mathbf{w}_t$, ie. we don't want $\|\mathbf{w}_t - \mathbf{w}_{t+1}\|$ be too big.

1. Combining these two goals, we compute $\mathbf{w}_{t+1}$ by solving the following optimization problem

$$\mathbf{w}_{t+1} = \arg\min \left(\langle \mathbf{w}_{t+1}, \mathbf{x}_t \rangle - y_t\right)^2 + \|\mathbf{w}_{t+1} - \mathbf{w}_t\|$$

2. Take the gradient of this equation, and make it equal to zero. We obtain

$$\mathbf{w}_{t+1} = \mathbf{w}_t - 2\eta \left(\langle \mathbf{w}_{t+1}, \mathbf{x}_t \rangle - y_t\right) \mathbf{x}_t$$

3. Approximating $\mathbf{w}_{t+1}$ by $\mathbf{w}_t$ on right-hand side gives updating rule of Widrow-Hoff algorithm.
4. Let $L_A = \sum_{t=1}^{T} (\hat{y}_t - y_t)$ be loss of algorithm A.
5. Let $L_{\mathbf{u}} = \sum_{t=1}^{T} (\langle \mathbf{u}, \mathbf{x}_t \rangle - y_t)$ be loss of another regressor denoted by $\mathbf{u} \in \mathbb{R}^n$.
6. We upper bound loss of Widrow-Hoff algorithm in terms of loss of the best vector.

Lemma (Bounds on potential function of Widrow-Hoff algorithm)

Let $\Phi_t = \|\mathbf{w}_t - \mathbf{u}\|_2^2$ *be the potential function, then we have*

$$\Phi_{t+1} - \Phi_t \leq -\eta l_t^2 + \frac{\eta}{1-\eta} g_t^2$$

where

$$l_t = (\hat{y}_t - y) = \langle \mathbf{w}_t, \mathbf{x}_t \rangle - y_t$$

$$g_t = \langle \mathbf{u}_t, \mathbf{x}_t \rangle - y_t$$

So that l_t^2 *denotes the learners loss at round* t*, and* g_t^2 *is* $\mathbf{u}$*'s loss at round* t*.*

Proof (Bounds on potential function of Widrow-Hoff algorithm).

1. Let $\Delta_t = \eta\left(\langle \mathbf{w}_t, \mathbf{x}_t\rangle - y_t\right)\mathbf{x}_t = \eta l_t \mathbf{x}_t$ (update to the weight vector). Then, we have

$$
\begin{aligned}
\Phi_{t+1} - \Phi_t &= \|\mathbf{w}_{t+1} - \mathbf{u}\|_2^2 - \|\mathbf{w}_t - \mathbf{u}\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{u} - \Delta_t\|_2^2 - \|\mathbf{w}_t - \mathbf{u}\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{u}\|_2^2 - 2\left\langle (\mathbf{w}_t - \mathbf{u}), \Delta_t\right\rangle + \|\Delta_t\|_2^2 - \|\mathbf{w}_t - \mathbf{u}\|_2^2 \\
&= -2\eta l_t \left\langle \mathbf{x}_t, (\mathbf{w}_t - \mathbf{u})\right\rangle + \eta^2 l_t^2 \|\mathbf{x}_t\|_2^2 \\
&\leq -2\eta l_t \left(\langle \mathbf{x}_t, \mathbf{w}_t\rangle - \langle \mathbf{u}, \mathbf{x}_t\rangle\right) + \eta^2 l_t^2 && \text{(since } \|\mathbf{x}_t\|_2^2 \leq 1\text{)} \\
&= -2\eta l_t \left[\left(\langle \mathbf{w}_t, \mathbf{x}_t\rangle - y_t\right) - \left(\langle \mathbf{u}, \mathbf{x}_t\rangle - y_t\right)\right] + \eta^2 l_t^2 \\
&= -2\eta l_t (l_t - g_t) + \eta^2 l_t^2 = -2\eta l_t^2 + 2\eta l_t g_t + \eta^2 l_t^2 \\
&\leq -2\eta l_t^2 + 2\eta\left(\frac{l_t^2(1-\eta) + g_t^2/(1-\eta)}{2}\right) + \eta^2 l_t^2 && \text{(by AM-GM)} \\
&= -\eta l_t^2 + \left(\frac{\eta}{1-\eta}\right) g_t^2
\end{aligned}
$$

Proof (Bounds on potential function of Widrow-Hoff algorithm).

2. Arithmetic mean-geometric mean inequality (AM-GM) states:
   for any set of non-negative real numbers, arithmetic mean of the set is greater than or equal to geometric mean of the set.
3. It states for any real numbers $x_1, \ldots, x_n \geq 0$, we have $\frac{x_1+\ldots+x_n}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n}$.
4. For reals $a = l_t^2(1-\eta) \geq 0$ and $b = \frac{g_t^2}{1-\eta} \geq 0$, AM-GM is $\sqrt{ab} \leq \frac{a+b}{2}$.

□

Theorem (Upper bound of loss Widrow-Hoff algorithm)

Assume that for all rounds t we have $\|\mathbf{x}_t\|_2^2 \leq 1$, then we have

$$L_{WH} \leq \min_{\mathbf{u}\in\mathbb{R}^n} \left[\frac{L_{\mathbf{u}}}{1-\eta} + \frac{\|\mathbf{u}\|_2^2}{\eta} \right]$$

where L_{WH} denotes the loss of Widrow-Hoff algorithm.

Proof (Upperbound of loss Widrow-Hoff algorithm).

1. Let $\sum_{t=1}^{T}\left(\Phi_{t+1}-\Phi_t\right)=\Phi_{T+1}-\Phi_1$.
2. By setting $\mathbf{w}_1=0$ and observation that $\Phi_t \geq 0$, we obtain

$$-\|u\|_2^2=-\Phi_1 \leq \Phi_{T+1}-\Phi_1$$

3. Hence, we have

$$\begin{aligned}
-\|u\|_2^2 &\leq \sum_{t=1}^{T}\left(\Phi_{t+1}-\Phi_t\right) \\
&\leq \sum_{t=1}^{T}\left(-\eta l_t^2+\left(\frac{\eta}{1-\eta}\right)g_t^2\right)=-\eta L_{WH}+\left(\frac{\eta}{1-\eta}\right)L_{\mathbf{u}}.
\end{aligned}$$

4. By simplifying this inequality, we obtain $L_{WH} \leq \left(\frac{\eta}{1-\eta}\right)L_{\mathbf{u}}+\frac{\|u\|_2^2}{\eta}$.
5. Since $\mathbf{u}$ was arbitrary, the above inequality must hold for **the best vector**.

□

1. We can look at the average loss per time step

$$\frac{L_{WH}}{T} \leq \min_{\mathbf{u}} \left[\left(\frac{\eta}{1-\eta} \right) \frac{L_{\mathbf{u}}}{T} + \frac{\|u\|_2^2}{\eta T} \right].$$

2. As T gets large, we have

$$\left(\frac{\|u\|_2^2}{\eta T} \right) \rightarrow 0.$$

3. If step-size (η) is very small,

$$\left(\frac{\eta}{1-\eta} \right) \frac{L_{\mathbf{u}}}{T} \rightarrow \min_{\mathbf{u}} \left(\frac{L_{\mathbf{u}}}{T} \right), \qquad \text{Show it.}$$

   which is **the average loss of the best regressor**.

4. This means that the Widrow-Hoff algorithm is performing almost as well as the best regressor vector as the number of rounds gets large.

Summary

- We study the bounded regression problem.
- For unbounded regression, there is the main issue for deriving uniform convergence bounds.
- We defined pseudo-dimension for real-valued function classes.
- We study the generalization bounds based on Rademacher complexity.
- We study several regression algorithms and analysis their bounds.
- We study an online regression algorithms and analysis its bound.

1. Chapter 11 of Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of Machine Learning*. Second Edition. MIT Press.
2. Chapter 11 of Martin Anthony and Peter L. Bartlett (1999). *Learning in Neural Networks : Theoretical Foundations*. Cambridge University Press.

References

References

Anthony, Martin and Peter L. Bartlett (1999). *Learning in Neural Networks : Theoretical Foundations*. Cambridge University Press.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of Machine Learning*. Second Edition. MIT Press.

Questions?