Machine learning theory

Regression

Hamid Beigy

Sharif university of technology

June 1, 2020
| 1. Introduction                           |
| 2. Generalization bounds                 |
| 3. Pseudo-dimension bounds               |
| 4. Regression algorithms                 |
| 5. Summary                               |
Introduction
The problem of regression

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ a measurable subset of $\mathbb{R}$ and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$.

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 $\mathcal{D}$.

Let $L : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$ be the loss function used to measure the magnitude of error.

The most used loss function is

- $L_2$ defined as $L(y, y') = |y' - y|^2$ for all $y, y' \in \mathcal{Y}$,
- or more generally $L_p$ 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 $R(h)$ respect to target $f$:

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

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

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

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 \( h \in H \)

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

Proof (Generalization bounds for finite hypothesis sets).

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

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

Thus, by the union bound, we can write

\[
\mathbb{P}\left[\exists h \in H \mid R(h) - \hat{R}(h) > \epsilon\right] \leq \sum_{h \in H} \mathbb{P}\left[R(h) - \hat{R}(h) > \epsilon\right] \\
\leq |H| \exp\left(-2\frac{m\epsilon^2}{M^2}\right).
\]

Setting the right-hand side to be equal to \( \delta \), the theorem will proved.
Theorem (Rademacher complexity of $\mu$-Lipschitz loss functions)

Let $L \leq M$ be a bounded loss function such that for any fixed $y' \in Y$, $L(y, y')$ is $\mu$-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 $G = \{(x, y) \mapsto L(h(x), y) \mid h \in H\}$ is

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

Proof (Rademacher complexity of $\mu$-Lipschitz loss functions).

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

$$\hat{R}(G) = \frac{1}{m} \mathbb{E}_\sigma \left[ \sum_{i=1}^m \sigma_i L(h(x_i), y_i) \right]$$

$$\leq \frac{1}{m} \mathbb{E}_\sigma \left[ \sum_{i=1}^m \sigma_i \mu h(x_i) \right]$$

$$= \mu \hat{R}(H).$$
Rademacher complexity bounds

Theorem (Rademacher complexity of $L_p$ loss functions)

Let $p \geq 1$ and $\mathcal{G} = \{ x \mapsto |h(x) - f(x)|^p \mid h \in H \}$ 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

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

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 $H' = \{ x \mapsto h(x) - f(x) \mid h \in H' \}$. Since $\phi_p$ is $pM^{p-1}$-Lipschitz over $[-M, M]$, we can apply Talagrand's Lemma,

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

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

$$\hat{R}(H') = \frac{1}{m} \mathbb{E}_\sigma \left[ \sup_{h \in H} \sum_{i=1}^{m} (\sigma_i h(x_i) + \sigma_i f(x_i)) \right]$$

$$= \frac{1}{m} \mathbb{E}_\sigma \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] + \frac{1}{m} \mathbb{E}_\sigma \left[ \sum_{i=1}^{m} \sigma_i f(x_i) \right] = \hat{R}(H).$$

Since $\mathbb{E}_\sigma [\sum_{i=1}^{m} \sigma_i f(x_i)] = \sum_{i=1}^{m} \mathbb{E}_\sigma [\sigma_i] f(x_i) = 0$. 

\[\square\]
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 $\mu$-Lipschitz for some $\mu > 0$. Then,

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

Proof (Rademacher complexity of $\mu$-Lipschitz loss functions).

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

$$
\hat{\mathcal{R}}(G) = \frac{1}{m} \mathbb{E}_\sigma \left[ \sum_{i=1}^{m} \sigma_i L(h(x_i), y_i) \right] 
$$

$$
\leq \frac{1}{m} \mathbb{E}_\sigma \left[ \sum_{i=1}^{m} \sigma_i \mu h(x_i) \right] = \mu \hat{\mathcal{R}}(H).
$$

Combining this inequality with general Rademacher complexity learning bound completes proof.
Pseudo-dimension bounds
VC dimension is a measure of complexity of a hypothesis set.

We define shattering for families of real-valued functions.

Let $G$ be a family of loss functions associated to some hypothesis set $H$, where

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

**Definition (Shattering)**

Let $G$ 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 $G$ if there exists $t_1, \ldots, t_m \in \mathbb{R}$ such that

$$\left| \begin{bmatrix} \text{sgn} (g(z_1) - t_1) \\ \text{sgn} (g(z_2) - t_2) \\ \vdots \\ \text{sgn} (g(z_m) - t_m) \end{bmatrix} \right| = 2^m$$

$g \in G$

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

In other words, $S$ is shattered by $G$, if there are real numbers $t_1, \ldots, t_m$ such that for $b \in \{0, 1\}^m$, there is a function $g_b \in G$ with $\text{sgn} (g_b(x_i) - t_i) = b_i$ for all $1 \leq i \leq m$. 
Thus, \( \{z_1, \ldots, z_m\} \) is shattered if for some witnesses \( t_1, \ldots, t_m \), the family of functions \( G \) is rich enough to contain a function going

1. above a subset \( A \) of the set of points \( J = \{(z_i, t_i) \mid 1 \leq i \leq m\} \) and
2. below the others \( J - A \), for any choice of the subset \( A \).

For any \( g \in G \), let \( B_g \) be the indicator function of the region below or on the graph of \( g \), that is

\[
B_g(x, y) = \text{sgn}(g(x) - y).
\]

Let \( B_G = \{B_g \mid g \in G\} \).
The notion of shattering naturally leads to definition of pseudo-dimension.

**Definition (Pseudo-dimension)**

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

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

$$Pdim(G) = VC \left( \{(x, t) \mapsto \mathbb{1}[ (g(x) - t) > 0 ] \mid g \in G \} \right)$$

Thus $Pdim(G) = d$, 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 $t_i$.
Properties of Pseudo-dimension

Theorem (Composition with non-decreasing function)
Suppose $\mathcal{G}$ is a class of real-valued functions and $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is a non-decreasing function. Let $\sigma(\mathcal{G})$ denote the class $\{\sigma \circ g \mid g \in \mathcal{G}\}$. Then

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

Proof (Pseudo-dimension of hyperplanes).
1. For $d \leq Pdim(\sigma(\mathcal{G}))$, suppose

$$\left\{ \sigma \circ g_b \mid b \in \{0,1\}^d \right\} \subseteq \sigma(\mathcal{G})$$

shatters a set $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ 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 $\text{sgn} \left( \sigma(g_b(x_i)) - t_i \right) = b_i$.
3. For all $1 \leq i \leq d$, take

$$y_i = \min \left\{ g_b(x_i) \mid \sigma(g_b(x_i)) \geq t_i, b \in \{0,1\}^d \right\}$$

4. Since $\sigma$ is non-decreasing, it is straightforward to verify that $\text{sgn} \left( g_b(x_i) - t_i \right) = b_i$ for all $\{0,1\}^d$ and $1 \leq i \leq d$.

\[\square\]
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 $Pdim(\mathcal{G}) = dim(\mathcal{G})$.

**Proof (Pseudo-dimension of vector spaces).**

1. Let $B_\mathcal{G}$ be the class of **below th graph** indicator functions, we have $Pdim(\mathcal{G}) = VC(B_\mathcal{G})$.
2. But $B_\mathcal{G} = \{(x, y) \mapsto \text{sgn} (g(x) - y) \mid g \in \mathcal{G}\}$.
3. Hence, the functions $B_\mathcal{G}$ are of the form $\text{sgn} (g_1 + g_2)$, where
   - $g_1 = g$ is a function from vector space
   - $g_2$ is the fixed function $g_2(x, y) = -y$.
4. Then, Theorem (Pseudo-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 $G = \{ x \mapsto \langle w, x \rangle + b \mid w \in \mathbb{R}^n, b \in \mathbb{R} \}$ be the class of hyperplanes in $\mathbb{R}^n$, then $Pdim(G) = n + 1$.

Pseudo-dimension of hyperplanes.

1. It is easy to check that $G$ is a vector space.
2. Let $g_i$ be the $i$th coordinate projection $f_i(x) = x_i$ for all $1 \leq i \leq n$ and $1$ be identity-1 function. Then $B = \{ g_1, \ldots, g_n, 1 \}$ is basis of $G$.
3. Hence, $Pdim(G) = n + 1$
A polynomial transformation of $\mathbb{R}^n$ is function $g(x) = w_0 + w_1\phi_1(x) + w_2\phi_2(x) + \ldots + w_k\phi_k(x)$ for $x \in \mathbb{R}^n$, where $k$ is an integer and for each $1 \leq i \leq k$, function $\phi_i(x)$ is defined as

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

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

**Proof (Pseudo-dimension of polynomial transformation).**

**Homework:** Prove this Theorem.

**Theorem (Pseudo-dimension of all polynomial transformations)**

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

**Proof (Pseudo-dimension of all polynomial transformations).**

**Homework:** Prove this Theorem.
Theorem (Generalization bound for bounded regression)

Let $H$ be a family of real-valued functions and $G = \{ z = (x, y) \mapsto L(h(x), y) \mid h \in H \}$ be a family of loss functions associated to a hypothesis set $H$. Assume that $\text{Pdim}(G) = d$ 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 $h \in H$:

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

Proof (Generalization bound for bounded regression).

Homework: Prove this Theorem.
Regression algorithms
Let $\Phi : \mathcal{X} \mapsto \mathbb{R}^n$ and $H = \{ h : x \mapsto \langle w, \Phi(x) \rangle + b \mid w \in \mathbb{R}^n, b \in \mathbb{R} \}$.

Given sample $S$, the problem is to find a $h \in H$ such that

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

Define data matrix

$$X = \begin{bmatrix} \Phi(x_1) & \phi(x_2) & \ldots & \phi(x_m) \\ 1 & 1 & \ldots & 1 \end{bmatrix}$$

Let $w = (w_1, \ldots, w_n, b)^T$ be the weight vector and $y = (y_1, \ldots, y_m)^T$ be the target vector.

By setting $\nabla \hat{R}(h) = 0$, we obtain

$$w = (XX^T)^+ Xy$$

When $XX^T$ is invertible, there is a unique solution; otherwise the problem has 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 $H = \{ x \mapsto \langle w, \Phi(x) \rangle \mid \left\| w \right\|_{\mathbb{H}} \leq \Lambda \}$. Assume that there exists $r > 0$ such that $K(x, x) \leq r^2$ and $M > 0$ such that $|h(x) - y| < M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Then for any $\delta > 0$, with probability at least $(1 - \delta)$, each of the following inequalities holds for all $h \in H$.

\[
\begin{align*}
R(h) & \leq \hat{R}(h) + 4M \sqrt{\frac{r^2 \Lambda^2}{m}} + M^2 \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \\
R(h) & \leq \hat{R}(h) + \frac{4M\Lambda \sqrt{\text{Tr}[K]}}{m} + 3M^2 \sqrt{\frac{\log \frac{2}{\delta}}{2m}}
\end{align*}
\]

**Proof.**

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

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

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

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

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

Kernel ridge regression is defined by minimization of an objective function (theoretical analysis)

\[
\min_w F(w) = \min_w \left[ \lambda \|w\|^2 + \sum_{i=1}^{m} (\langle w, \Phi(x_i) \rangle - y_i)^2 \right]
\]

\[
= \min_w \left[ \lambda \|w\|^2 + \|\Phi^T w - y\|^2 \right]
\]

By setting \(\nabla F(w) = 0\), we obtain

\[
w = (\Phi \Phi^T + \lambda I)^{-1} \Phi y
\]

An alternative formulation of kernel ridge regression is

\[
\min_w \|\Phi^T w - y\|^2 \quad \text{subject to} \quad \|w\|^2 \leq \Lambda^2
\]

\[
\min_w \sum_{i=1}^{m} \xi_i^2 \quad \text{subject to} \quad \left(\|w\|^2 \leq \Lambda^2 \right) \land \left( \forall i \in \{1, \ldots, m\}, \xi_i = y_i - \langle w, \Phi(x_i) \rangle \right)
\]
Support vector regression (SVR)

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

This defines two sets of points:
1. points falling inside the tube, which are $\epsilon$-close to the function predicted and thus not penalized,
2. points falling outside the tube, which are penalized based on their distance to the predicted function.

This is similar to the penalization used by SVMs in classification.

Using a hypothesis set of linear functions $H = \{x \mapsto \langle w, \Phi(x) \rangle + b | w \in \mathbb{R}^n, b \in \mathbb{R}\}$, where $\Phi$ is the feature mapping corresponding some PDS kernel $K$.

The optimization problem for SVR is

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

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

$$\forall y, y' \in \mathcal{Y}, \quad |y' - y|_\epsilon = \max (0, |y' - y| - \epsilon)$$
Support vector regression (SVR)

- The $\epsilon$-insensitive loss is defined as

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

- The use of $\epsilon$-insensitive loss leads to sparse solutions with a relatively small number of support vectors.

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

$$
\min_{w, b, \xi, \xi'} \left[ \frac{1}{2} \lambda \|w\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi'_i) \right]
$$

subject to

- $(\langle w, \Phi(x_i) \rangle + b) - y_i \leq \epsilon + \xi_i$
- $y_i - (\langle w, \Phi(x_i) \rangle + b) \leq \epsilon + \xi'_i$
- $\xi_i \geq 0, \quad \xi'_i \geq 0, \quad \forall i, 1 \leq i \leq m$

- This is a convex quadratic program (QP) with affine constraints.

- By introducing Lagrangian and applying KKT conditions, the problem will be solved.
Support vector regression (SVR)

- Let $\mathcal{D}$ be the distribution according to which sample points are drawn.
- 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 $H = \{ x \mapsto \langle w, \Phi(x) \rangle \mid \|w\|_H \leq \Lambda \}$. Assume that there exists $r > 0$ such that $K(x, x) \leq r^2$ and $M > 0$ such that $|h(x) - y| < M$ for all $(x, y \in \mathcal{X} \times \mathcal{Y})$. Then for any $\delta > 0$, with probability at least $(1 - \delta)$, each of the following inequalities holds for all $h \in H$.

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

\[
\mathbb{E}_{(x, y) \sim \mathcal{D}} [ |h(x) - y|_\epsilon ] \leq \mathbb{E}_{(x, y) \sim \hat{\mathcal{D}}} [ |h(x) - y|_\epsilon ] + 2\Lambda \sqrt{\text{Tr}[K]} m + 3M\sqrt{\log \frac{1}{\delta}}
\]

**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 $H$. □
Support vector regression (SVR)

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

SVR admits several advantages

1. SVR algorithm is based on solid theoretical guarantees,
2. The solution returned SVR is sparse
3. SVR allows a natural use of PDS kernels
4. SVR also admits favorable stability properties.

SVR also admits several disadvantages

1. SVR requires the selection of two parameters, \( C \) and \( \epsilon \), which are determined by cross-validation.
2. may be computationally expensive when dealing with large training sets.
The optimization problem for Lasso is defined as

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

This is a **convex optimization problem**, because
1. \(\|w\|_1\) is convex as with all norms
2. the empirical error term is convex

Hence, the optimization problem can be written as

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

The \(L_1\) norm constraint is that it leads to a **sparse solution** \(w\).

---

**L1 regularization**

**L2 regularization**

Figure 11.6

Comparison of the Lasso and ridge regression solutions.
**Theorem (Bounds of $\hat{R}(H)$ of Lasso)**

Let $\mathcal{X} \subseteq \mathbb{R}^n$ and let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be sample of size $m$. Assume that for all $1 \leq i \leq m$, $\|x_i\|_\infty \leq r_\infty$ for some $r_\infty > 0$, and let $H = \{x \mapsto \langle w, x \rangle \mid \|w\|_1 \leq \Lambda_1\}$. Then, the empirical Rademacher complexity of $H$ can be bounded as follows

$$\hat{R}(H) \leq \sqrt{\frac{2r^2_\infty \Lambda^2_1 \log(2n)}{m}}$$

**Definition (Dual norms)**

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

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

For any $p, q \geq 1$ that are conjugate that is such that $\frac{1}{p} + \frac{1}{q} = 1$, the $L_p$ and $L_q$ norms are dual norms of each other.

In particular, the dual norm of $L_2$ is the $L_2$ norm, and the dual norm of the $L_1$ norm is the $L_\infty$ norm.
**Proof (Bounds of $\hat{R}(H)$ of Lasso).**

For any $1 \leq i \leq m$, we denote by $x_{ij}$, the $j$th component of $x_i$.

\[
\hat{R}(H) = \frac{1}{m} E_{\sigma} \left[ \sup_{\|w\|_1 \leq \Lambda_1} \sum_{i=1}^{m} \sigma_i \langle w, x_i \rangle \right] = \frac{\Lambda_1}{m} E_{\sigma} \left[ \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|_\infty \right] \quad \text{(by definition of the dual norm)}
\]

\[
= \frac{\Lambda_1}{m} E_{\sigma} \left[ \max_{j \in \{1, \ldots, n\}} \left| \sum_{i=1}^{m} \sigma_i x_{ij} \right| \right] \quad \text{(by definition of $\|\cdot\|_\infty$)}
\]

\[
= \frac{\Lambda_1}{m} E_{\sigma} \left[ \max_{j \in \{1, \ldots, n\}} \max_{s \in \{-1, +1\}} \left| \sum_{i=1}^{m} \sigma_i x_{ij} \right| \right] \quad \text{(by definition of $\|\cdot\|_\infty$)}
\]

\[
= \frac{\Lambda_1}{m} E_{\sigma} \left[ \sup_{z \in A} \sum_{i=1}^{m} \sigma_i z_i \right].
\]

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

For any $z \in A$, we have $\|z\|_2 \leq \sqrt{mr^2_\infty} = r_\infty \sqrt{m}$.

Thus by Massart's Lemma, since $A$ contains at most $2n$ elements, the following inequality holds:

\[
\hat{R}(H) \leq \Lambda_1 r_\infty \sqrt{m \frac{2 \log(2n)}{m}} = \Lambda_1 r_\infty \sqrt{\frac{2 \log(2n)}{m}}.
\]
This bounds depends on dimension $n$ is only logarithmic, which suggests that using very high-dimensional feature spaces does not significantly affect generalization.

By combining of Theorem Bounds of $\hat{R}(H)$ of Lasso and Rademacher generalization bound, we obtain:

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

Let $\mathcal{X} \subseteq \mathbb{R}^n$ and $H = \{ x_1 \mapsto \langle w, x \rangle \mid \|w\|_1 \leq \Lambda_1 \}$. Let also $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be sample of size $m$. Assume that there exists $r_\infty > 0$ such that for all $x \in \mathcal{X}$, $\|x_i\|_\infty \leq r_\infty$ and $M > 0$ such that $|h(x) - y| \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Then, for any $\delta > 0$, with probability at least $(1 - \delta)$, each of the following inequality holds for $h \in H$

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

Ridge regression and Lasso have same form as the right-hand side of this generalization bound.

Lasso has several advantages:

1. It benefits from strong theoretical guarantees and returns a sparse solution.
2. The sparsity of the solution is also computationally attractive (inner product).
3. The algorithm’s sparsity can also be used for feature selection.

The main drawbacks are: usability of kernel and closed-form solution.
Online regression algorithms

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

Online linear regression

1: Initialize $w_1$.
2: for $t ← 1, 2, \ldots, T$ do.
3: Receive $x_t ∈ \mathbb{R}^n$.
4: Predict $\hat{y}_t = \langle w_t, x_t \rangle$.
5: Observe true label $y_t = h^*(x_t)$.
6: Compute the loss $L(\hat{y}_t, y_t)$.
7: Update $w_{t+1}$.
8: end for
Widrow-Hoff algorithm uses stochastic gradient descent technique to linear regression objective function.

At each round, the weight vector is augmented with a quantity that depends on the prediction error $(\langle w_t, x_t \rangle - y_t)$.

**WidrowHoff regression**

1: function **WIDROWHOF**($w_0$)
2: Initialize $w_1 \leftarrow w_0$. ▷ typically $w_0 = 0$.
3: for $t \leftarrow 1, 2, \ldots, T$ do.
4: Receive $x_t \in \mathbb{R}^n$.
5: Predict $\hat{y}_t = \langle w_t, x_t \rangle$.
6: Observe true label $y_t = h^*(x_t)$.
7: Compute the loss $L(\hat{y}_t, y_t)$.
8: Update $w_{t+1} \leftarrow w_t - 2\eta (\langle w_t, x_t \rangle - y_t) x_t$. ▷ learning rate $\eta > 0$.
9: end for
10: return $w_{T+1}$
11: end function
There are two motivations for the update rule in Widrow-Hoff.

The first motivation is that
1. The loss function is defined as
\[ L(w, x, y) = (\langle w, x \rangle - y)^2 \]
2. To minimize the loss function, move in the direction of the negative gradient
\[ \nabla_w L(w, x, y) = 2 (\langle w, x \rangle - y) x \]
3. This gives the following update rule
\[ w_{t+1} \leftarrow w_t - \eta \nabla_w L(w_t, x_t, y_t) \]

The second motivation is that we have two goals:
1. We want loss on \((x_t, y_t)\) to be small which means that we want to minimize \((\langle w, x \rangle - y)^2\).
2. We don't want to be too far from \(w_t\). That is, we don't want \(||w_t - w_{t+1}||\) to be too big.

Combining these two goals, we compute \(w_{t+1}\) by solving the following optimization problem
\[ w_{t+1} = \text{argmin} \ (\langle w_{t+1}, x_t \rangle - y_t)^2 + ||w_{t+1} - w_t|| \]

Take the gradient of this equation, and make it equal to zero. We obtain
\[ w_{t+1} = w_t - 2\eta (\langle w_{t+1}, x_t \rangle - y_t) x_t \]

Approximating \(w_{t+1}\) by \(w_t\) on right-hand side gives updating rule of Widrow-Hoff algorithm.
Let $L_A = \sum_{t=1}^{T} (\hat{y}_t - y_t)$ be loss of algorithm $A$ and $L_u = \sum_{t=1}^{T} (\langle u, x_t \rangle - y_t)$ be loss of $u \in \mathbb{R}^n$.

We upper bound loss of Widrow-Hoff algorithm in terms of loss of the best vector.

**Theorem (Upper bound of loss Widrow-Hoff algorithm)**

Assume that for all rounds $t$ we have $\|x_t\|_2^2 \leq 1$, then we have

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

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

Before proving this Theorem, we first prove the following Lemma.

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

Let $\Phi_t = \|w_t - u\|_2^2$ 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

$$l_t = (\hat{y}_t - y) = \langle w_t, x_t \rangle - y_t$$

$$g_t = \langle u_t, x_t \rangle - y_t$$

So that $l_t^2$ denotes the learner's loss at round $t$, and $g_t^2$ is $u$'s loss at round $t$. 
Proof (Bounds on potential function of Widrow-Hoff algorithm).

Let $\Delta_t = \eta (\langle w_t, x_t \rangle - y_t) x_t = \eta l_t x_t$ (update to the weight vector). Then, we have

$$
\Phi_{t+1} - \Phi_t = \|w_{t+1} - u\|^2 - \|w_t - u\|^2 \\
= \|w_t - u - \Delta_t\|^2 - \|w_t - u\|^2 \\
= \|w_t - u\|^2 - 2 \langle (w_t - u), \Delta_t \rangle + \|\Delta_t\|^2 - \|w_t - u\|^2 \\
= -2\eta l_t \langle x_t, (w_t - u) \rangle + \eta^2 l_t^2 \|x_t\|^2 \\
\leq -2\eta l_t (\langle x_t, w_t \rangle - \langle u, x_t \rangle) + \eta^2 l_t^2 \\
= -2\eta l_t [(\langle w_t, x_t \rangle - y_t) - (\langle u, x_t \rangle - y_t)] + \eta^2 l_t^2 \\
= -2\eta l_t (l_t - g_t) + \eta^2 l_t^2 = -2\eta l_t^2 + 2\eta l_t g_t + \eta^2 l_t^2 \\
\leq -2\eta l_t^2 + 2\eta \left( \frac{\eta^2(1 - \eta) + g_t^2/(1 - \eta)}{2} \right) + \eta^2 l_t^2 \tag{by AM-GM}
\leq -2\eta l_t^2 + \frac{\eta}{1 - \eta} g_t^2
$$

1. 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.

2. For reals $a \geq 0$ and $b \geq 0$, AM-GM is $\sqrt{ab} \leq \frac{a + b}{2}$, and let $a = l_t^2 (1 - \eta)$ and $b = \frac{g_t^2}{1 - \eta}$.
Proof (Upperbound of loss Widrow-Hoff algorithm).

1. Let $\sum_{t=1}^{T} (\Phi_{t+1} - \Phi_{t}) = \Phi_{T+1} - \Phi_{1}$.

2. By setting $w_1 = 0$ and observation that $\Phi_{t} \geq 0$, we obtain that

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

3. Hence, we have

$$-\|u\|_2^2 \leq \sum_{t=1}^{T} (\Phi_{t+1} - \Phi_{t})$$

$$\leq \sum_{t=1}^{T} (-\eta l_t^2 + \left(\frac{\eta}{1-\eta}\right) g_t^2)$$

$$= -\eta L_{WH} + \left(\frac{\eta}{1-\eta}\right) L_u.$$ 

4. By simplifying this inequality, we obtain

$$L_{WH} \leq \left(\frac{\eta}{1-\eta}\right) L_u + \frac{\|u\|_2^2}{\eta}.$$ 

5. Since $u$ was arbitrary, the above inequality must hold for the best vector.
We can look at the average loss per time step

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

As \( T \) gets large, we have

\[
\left( \frac{\|u\|_2^2}{\eta T} \right) \to 0.
\]

If step-size \( (\eta) \) is very small,

\[
\left( \frac{\eta}{1 - \eta} \right) \frac{L_u}{T} \to \min_u \left( \frac{L_u}{T} \right), \quad \text{Show it.}
\]

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

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.

References

Questions?