Additive and Sparse Models

Mathematical Analysis of Machine Learning Algorithms
(Chapter 10)

Additive Models

We consider *additive model* of the following form:

$$f([w, \theta], x) = \sum_{j=1}^{m} w_j \psi(\theta_j, x), \tag{1}$$

where for simplicity, we consider real valued functions

$$\psi(\theta, \cdot) : \mathcal{X} \to \mathbb{R}.$$

Finite Family Model Combination

Assume that Θ has only finite number of m elements $\{\theta_1, \ldots, \theta_m\}$, then (1) can be regarded as a linear model with respect to the model parameter w as:

$$f(w, x) = \sum_{j=1}^{m} w_j \psi_j(x) = w^\top \psi(x),$$

with features $\psi_j(x) = \psi(\theta_j, x)$, and $\psi(x) = [\psi_1(x), \ldots, \psi_m(x)]$.

We further assume that each feature $\psi_j(x)$ is a prediction function

$$\psi_j(x) \in [0, M].$$

Model Complexity

Without regularization

Model complexity is determined by the model dimensionality m.

With L_2 regularization

Function class is

$$\mathcal{F}' = \{f(w, x) = w^\top \psi(x) : \|w\|_2 \leq A\}$$

Rademacher complexity is

$$R_n(\mathcal{F}', \mathcal{D}) \leq A\sqrt{\frac{m}{n}\mathbb{E}_{\mathcal{D}} k(x, x)},$$

$$\text{with kernel} \quad k(x, x) = \frac{1}{m}\sum_{j=1}^{m} \psi_j(x)^2.$$

Model Complexity (L_2 regularization)

From

$$\mathbb{E}_{X\sim\mathcal{D}} k(X,X) \leq \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{X\sim\mathcal{D}} \psi_j(X)^2 \leq M^2,$$

we obtain the following result.

Rademacher Complexity

$$R_n(\mathcal{F}', \mathcal{D}) \leq \sqrt{\frac{m}{n}} AM,$$

which depends linearly on $\sqrt{m}$.

Matching Lower Bound

Proposition 1 (Prop 10.1)

Assume that the m feature functions $\{\psi_j(X)\}$ *are orthonormal when* $X \sim \mathcal{D}$. *Then there exists an absolute constant* $c > 0$ *such that for sufficiently large n,*

$$R_n(\mathcal{F}', \mathcal{D}) \geq c\sqrt{\frac{m}{n}}A,$$

where $\mathcal{F}'$ *is given by* (10.2).

Proposition 1 implies that the factor $\sqrt{m}$ in $R_n(\mathcal{F}', \mathcal{D})$ cannot be removed in general with L_2 regularization. To avoid m dependency, one needs to use small L_2 norm:

$$R_n(\mathcal{F}'', \mathcal{D}) \leq \sqrt{\frac{1}{n}}AM, \quad \mathcal{F}'' = \{f(w, x) : \|w\|_2^2 \leq A^2/m\}.$$

Sparsity

One common assumption to achieve better generalization is to impose an additional sparsity constraint.

Definition 2

The sparsity pattern, or support of a weight vector $w \in \mathbb{R}^m$ is defined as

$$\text{supp}(w) = \{j : w_j \neq 0\},$$

and the L_0 norm of w is defined as

$$\|w\|_0 = |\text{supp}(w)|.$$

Example 3

Consider RBFs of the form

$$\psi(\theta, x) = \exp(-\beta \|x - \theta\|_2^2)$$

for $\beta > 0$. We can use RBFs as basis functions in additive models

$$f([w, \theta], x) = \sum_{j=1}^{m} w_j \exp(-\beta \|x - \theta_j\|_2^2).$$

The corresponding RKHS norm for additive model is

$$\|f([w, \theta], x)\|^2 = m \sum_{j=1} w_j^2. \tag{2}$$

However, even for simple 1-dimensional functions such as $\sum_{j=1}^{m} \exp(-\|x - j\|_2^2)$, the complexity measured by the RKHS norm in (2) can be rather large.

Sparse Learning

For each sparsity pattern $F \subset \{1, \ldots, m\}$, we may define the $|F|$-dimensional sparse function class

$$\mathcal{G}_F = \{\phi(w, z) : \mathrm{supp}(w) \subset F\},$$

where $\phi(w, z) = L(f(w, x), y) = L(w^\top \psi(x), y)$.

We can now consider the following sparse learning method:

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d} \left[\frac{1}{n} \sum_{i=1}^{n} \phi(w, Z_i) + r(F) \right] \quad \text{subject to } \mathrm{supp}(w) \subset F, \quad (3)$$

Generalization error bound only depends logarithmically on m, and linear only with respect to the sparsity.

Oracle Inequality of Sparse Learning

Theorem 4 (Thm 10.4)

Assume $\sup_{w,z,z'}[\phi(w,z)-\phi(w,z')] \le M$. *Let* $\mathcal{S}_n$ *be* n *iid samples from* $\mathcal{D}$. *Then with probability at least* $1-\delta$, *the following bound holds for all* $w \in \mathbb{R}^m$ *and sparsity pattern* F *such that* $\mathrm{supp}(w) \subset F$:

$$\phi(w,\mathcal{D}) \le \phi(w,\mathcal{S}_n) + r(F) + M\sqrt{\frac{\ln(1/\delta)}{2n}},$$

where

$$r(F) \ge 2R(\mathcal{G}_F,\mathcal{D}) + M\sqrt{\frac{|F|\ln(em/|F|)+\ln(|F|+1)^2}{2n}}$$

for all F. *Consider the sparse learning algorithm in* (3). *We have the following oracle inequality. With probability of at least* $1-\delta$:

$$\phi(\hat{w},\mathcal{D}) \le \inf_{w\in\mathbb{R}^m,\mathrm{supp}(w)\subset F}[\phi(w,\mathcal{D})+r(F)] + 2M\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Result used in the Proof of Theorem 4

Theorem 5 (Thm 8.7)

Consider the model selection algorithm in (8.3)*, with*

$$\tilde{R}(\theta, f, \mathcal{S}_n) = \tilde{R}(\theta) \geq 2R_n(\mathcal{F}(\theta), \mathcal{D}) + M(\theta)\sqrt{\frac{\ln(1/q(\theta))}{2n}},$$

where $M(\theta) = \sup_{f,z,z'} \left|\phi(f, z) - \phi(f, z')\right|$*, and* $q(\theta)$ *satisfies* (8.1)*. Then with probability at least* $1 - \delta$*, for all* θ *and* $f \in \mathcal{F}(\theta)$*:*

$$\phi(f, \mathcal{D}) \leq \phi(f, \mathcal{S}_n) + \tilde{R}(\theta) + M(\theta)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Moreover, we have oracle inequality: with probability of at least $1 - \delta$*,*

$$\phi(\hat{f}, \mathcal{D}) \leq \inf_{\theta, f \in \mathcal{F}(\theta)} \left[\phi(f, \mathcal{D}) + \tilde{R}(\theta) + 2M(\theta)\sqrt{\frac{\ln(2/\delta)}{2n}}\right] + \tilde{\epsilon}.$$

Proof of Theorem 4

We note that

$$\binom{m}{s} \leq \frac{m^s}{s!} \leq \frac{m^s}{s^s} \cdot \frac{s^s}{s!} \leq \frac{m^s}{s^s} \cdot e^s = (me/s)^s.$$

We can now consider each $\mathcal{G}_F$ as a model, with

$$q_F = \frac{(|F|+1)^{-2}}{(me/|F|)^{|F|}} \leq \frac{(|F|+1)^{-2}}{\binom{m}{|F|}}.$$

Therefore we have

$$\sum_{F:|F|\geq 1} q_F \leq \sum_{s\geq 1} \frac{1}{(s+1)^2} \sum_{F:|F|=s} \frac{1}{\binom{m}{|F|}} = \sum_{s\geq 1} \frac{1}{(s+1)^2} < 1.$$

For each index F, we may consider $\hat{w}$ as the ERM solution under the constraint $\mathrm{supp}(w) \subset F$. We can thus apply Theorem 5 with models indexed by $F \subset \{1, \ldots, m\}$ to obtain the desired bounds.

Example 6 (Part I/II of Expl 10.5)

Consider the linear binary classification with loss function

$$L(f(w,x),y) = \mathbb{1}(w^\top \psi(x) y \leq 0).$$

We know from the Rademacher complexity of VC-class

$$R(\mathcal{G}_F, \mathcal{S}_n) \leq c_0 \sqrt{\frac{|F|}{n}}$$

for some constant $c_0 > 0$. Therefore we may let

$$r(F) = c' \sqrt{\frac{|F| \ln m}{n}}$$

for a sufficiently large constant c', and solve

$$\hat{w} = \arg \min_{w \in \mathbb{R}^m} \left[\frac{1}{n} \sum_{i=1}^{n} L(w^\top \psi(X_i), Y_i) + \lambda \sqrt{\|w\|_0} \right]$$

with $\lambda = c' \sqrt{\frac{\ln m}{n}}$.

Example 7 (Part II/II of Expl 10.5)

From Theorem 4, we obtain the following oracle inequality. With probability at least $1 - \delta$:

$$\mathbb{E}_{\mathcal{D}} L(f(\hat{w}, X), Y) \leq \inf_{w \in \mathbb{R}^m} \left[\mathbb{E}_{\mathcal{D}} L(f(w, X), Y) + \lambda \sqrt{\|w\|_0} \right] + 2\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

The bound is linear in $\sqrt{\|w\|_0}$, and logarithmic in m since

$$\lambda = c' \sqrt{\frac{\ln m}{n}}.$$

L_1 Regularization

Consider the general situation that Θ is infinite. For notation simplicity, we can define the function class

$$\Psi = \{\psi(\theta, x) : \theta \in \Theta\}.$$

Definition 8 (Convex Hull, Def 5.12)

The convex hull of a function class Ψ is defined as

$$\text{CONV}(\Psi) = \left\{ \sum_{j=1}^{m} w_j \psi(\theta_j, x) : \ m > 0, \ \|w\|_1 = 1, \ w_j \geq 0, \ \theta_j \in \Theta \right\}.$$

We may also include the closure of the finite sum functions above in the convex hull with respect to an appropriately defined topology.

L_1 Norm

The non-negative L_1 regularized additive models are:

$$\mathcal{F}^{+}_{A,L_1}(\Psi) = \{af(x) : a \in [0, A], f(x) \in \text{CONV}(\Psi)\}.$$

We may also consider L_1 regularized additive models as:

$$\mathcal{F}_{A,L_1}(\Psi) = \left\{ \sum_{j=1}^{m} w_j \psi(\theta_j, x) : \ \|w\|_1 \leq A, \ \theta_j \in \Theta, \ m > 0 \right\}.$$

Definition 9

Let $\mathcal{F}_{L_1}(\Psi)$ be point-wise closure of $\cup_{A>0}\mathcal{F}_{A,L_1}(\Psi)$, then

$$\forall f \in \mathcal{F}_{L_1}(\Psi) : \|f\|_1 = \lim_{\epsilon \to 0} \inf \left\{ \|w\|_1 : \sup_x \left| f(x) - \sum_{j=1}^{m} w_j \psi(\theta_j, x) \right| \leq \epsilon \right\}.$$

For notational convenience, we write functions in $\mathcal{F}_{L_1}(\Psi)$ as

$$f(x) = w^\top \psi(x),$$

where $\psi(x)$ is the infinite dimensional vector $[\psi(\theta, x)]_{\theta \in \Theta}$, and $\|f\|_1 = \|w\|_1$.

Rademacher Complexity

Theorem 10 (Thm 10.8)

We have

$$R(\text{CONV}(\Psi), \mathcal{S}_n) = R(\Psi, \mathcal{S}_n).$$

If either $\Psi = -\Psi$ *or* $0 \in \Psi$*, then the following equality holds:*

$$R(\mathcal{F}^{+}_{A,L_1}(\Psi), \mathcal{S}_n) = A \cdot R(\Psi, \mathcal{S}_n).$$

If $\Psi = -\Psi$*, then the following equality holds:*

$$R(\mathcal{F}_{A,L_1}(\Psi), \mathcal{S}_n) = A \cdot R(\Psi, \mathcal{S}_n).$$

Proof of Theorem 10 (I/II)

We prove the second equality. Since $A \cdot \Psi \subset \mathcal{F}^+_{A,L_1}(\Psi, S_n)$, we have

$$A \cdot R(\Psi, S_n) \leq R(\mathcal{F}^+_{A,L_1}(\Psi), S_n).$$

Moreover, consider any function

$$\sum_{j=1}^{m} w_j \psi(\theta_j, x) : \ \|w\|_1 \leq A, \ w_j \geq 0, \ \theta_j \in \Theta$$

and $\sigma_i \in \{\pm 1\}$, we know that under the conditions of the theorem,

$$\begin{aligned}
\sum_{i=1}^{n} \sigma_i \sum_{j=1}^{m} w_j \psi(\theta_j, X_i) &= \sum_{j=1}^{m} w_j \sum_{i=1}^{n} \sigma_i \psi(\theta_j, X_i) \\
&\leq \sum_{j=1}^{m} w_j \sup_{j'} \sum_{i=1}^{n} \sigma_i \psi(\theta_{j'}, X_i) \\
&= \|w\|_1 \sup_{j} \sum_{i=1}^{n} \sigma_i \psi(\theta_j, X_i).
\end{aligned}$$

The first inequality used $w_j \geq 0$.

Proof of Theorem 10 (II/II)

Continue from the previous slide, we have

$$\begin{aligned}\cdots &= \|w\|_1 \sup_j \sum_{i=1}^{n} \sigma_i \psi(\theta_j, X_i) \\ &\leq \|w\|_1 \sup_{\psi \in \Psi} \sum_{i=1}^{n} \sigma_i \psi(X_i) \leq A \sup_{\psi \in \Psi} \sum_{i=1}^{n} \sigma_i \psi(X_i).\end{aligned}$$

The last inequality used the fact that $\|w\|_1 \leq A$ and $\sup_{\psi \in \Psi} \sum_{i=1}^{n} \sigma_i \psi(X_i) \geq 0$. This implies that

$$R(\mathcal{F}^{+}_{A,L_1}(\Psi), \mathcal{S}_n) \leq A \cdot R(\Psi, \mathcal{S}_n),$$

and thus we obtain the second desired equality of the theorem. The proof of the first equality of the theorem is similar. The third equality of the theorem holds because the condition implies that

$$R(\mathcal{F}_{A,L_1}(\Psi), \mathcal{S}_n) = R(\mathcal{F}^{+}_{A,L_1}(\Psi), \mathcal{S}_n).$$

This proves the desired result.

Example 11

Assume that $|\Psi| = N$, then $|\Psi \cup -\Psi| \leq 2N$. From Rademacher complexity of finite function class, we have

$$R(\mathcal{F}_{A,L_1}(\Psi), S_n) \leq A \sup_{\psi \in \Psi} \|\psi\|_{L_2(S_n)} \cdot \sqrt{\frac{2\ln(2N)}{n}}.$$

If $|\psi(x)| \leq B$ for all $\psi \in \Psi$, then

$$R(\mathcal{F}_{A,L_1}(\Psi), S_n) \leq AB\sqrt{\frac{2\ln(2N)}{n}}.$$

Example 12

Assume that Ψ is a binary function class with VC dimension d, then we know that from Example 6.26 that

$$R(\Psi, \mathcal{S}_n) \leq 16\sqrt{\frac{d}{n}}.$$

It follows that

$$R(\mathcal{F}_{A,L_1}(\Psi), \mathcal{S}_n) \leq AR(\Psi, S_n) + AR(-\Psi, \mathcal{S}_n) \leq 32A\sqrt{\frac{d}{n}}.$$

Example 13

In two-layer NN, let Ψ be an L_2-regularized ReLU function class:

$$\Psi = \left\{ \psi(\theta, x) = \max(0, \theta^\top x) : \qquad \|\theta\|_2 \leq \alpha, \|x\|_2 \leq \beta \right\},$$

and the corresponding L_1 regularized two-layer NN can be expressed as

$$f(x) = \sum_{j=1}^{m} w_j \psi(\theta_j, x) : \|w\|_1 \leq A, \|\theta\|_2 \leq \alpha, \|x\|_2 \leq \beta.$$

This function class belongs to $\mathcal{F}_{A,L_1}(\Psi)$. We thus obtain the following bound for L_1 regularized two-layer NN:

$$\begin{aligned} R(\mathcal{F}_{A,L_1}(\Psi), \mathcal{S}_n) \leq & 2AR(\Psi, \mathcal{S}_n) && \text{(Theorem 10)} \\ \leq & 2A\alpha\beta/\sqrt{n}, && \text{(Corollary 9.21)} \end{aligned}$$

where we note that $\max(0, f)$ is 1-Lipschitz in f.

Lasso

We now consider the following hard-constrained L_1 regularized learning:

$$\hat{w} = \arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} L(w^\top \psi(X_i), Y_i) \qquad \|w\|_1 \leq A. \tag{4}$$

Similarly, we may consider the soft-regularized version as:

$$\hat{w} = \arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} L(w^\top \psi(X_i), Y_i) + \lambda \|w\|_1. \tag{5}$$

Generalization Analysis of L_1 Regularization

Corollary 14 (Cor 10.12)

Assume that $L(p, y) \in [0, M]$ is γ Lipschitz with respect to p. For fixed $A > 0$, with probability at least $1 - \delta$: for all $f(x) = w^\top \psi(x)$ such that $\|w\|_1 \le A$:

$$\mathbb{E}_{\mathcal{D}} L(w^\top \psi(X), Y) \le \frac{1}{n} \sum_{i=1}^{n} L(w^\top \psi(X_i), Y_i) + 2\gamma A R_n(\Psi_\pm, \mathcal{D}) + M\sqrt{\frac{\ln(1/\delta)}{2n}},$$

where $\Psi_\pm = \{\psi(x) : \psi(x) \in \Psi \text{ or } -\psi(x) \in \Psi\}$. Moreover, for (4), if we solve it approximately up to sub-optimality of ϵ', then we have with probability at least $1 - \delta$:

$$\mathbb{E}_{\mathcal{D}} L(\hat{w}^\top \psi(X), Y) \le \inf_{\|w\|_1 \le A} \mathbb{E}_{\mathcal{D}} L(w^\top \psi(X), Y) + 2\gamma A R_n(\Psi_\pm, \mathcal{D}) + \epsilon' + M\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Example

Example 15

If Ψ contains m functions $\{\psi_1(x), \ldots, \psi_m(x)\}$, each $|\psi_j(x)| \leq B$, then

$$R_n(\Psi_\pm, \mathcal{D}) \leq B\sqrt{\frac{2\ln(2m)}{n}}.$$

Therefore the bound of Corollary 14 implies the oracle inequality

$$\mathbb{E}_{\mathcal{D}} L(\hat{w}^\top \psi(X), Y) \leq \inf_{\|w\|_1 \leq A} \mathbb{E}_{\mathcal{D}} L(w^\top \psi(X), Y) + 2\gamma AB\sqrt{\frac{2\ln(2m)}{n}} + M\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

This has a logarithmic dependency on m, similar to that of the sparsity constraint in Example 6.

L_1 Penalty Regularization: Uniform Convergence

Corollary 16 (Uniform Convergence in Cor 10.14)

Assume that $L(p, y) \geq 0$ *is* γ *Lipschitz,* $M_0 = \sup_y L(0, y)$*, and* $B = \sup_{x,\psi \in \Psi} |\psi(x)|$*. Consider* $A_0 > 0$*, then with probability at least* $1 - \delta$*: the following inequality holds for all* w*:*

$$\mathbb{E}_{\mathcal{D}} L(w^\top \psi(X), Y) \leq \frac{1}{n} \sum_{i=1}^{n} L(w^\top \psi(X_i), Y_i) + 4\gamma(A_0 + \|w\|_1) R_n(\Psi_\pm, \mathcal{D})$$

$$+ (M_0 + 2\gamma B(A_0 + \|w\|_1)) \left[\sqrt{\frac{\ln\left(2 + \log_2\left(1 + \frac{\|w\|_1}{A_0}\right)\right)}{n}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}} \right].$$

Proof of Corollary 16

Let $A_\theta = 2^\theta A_0$, with $q(\theta) = (1+\theta)^{-2}$ for $\theta = 1, 2, \ldots$, and let $f(x) = w^\top \psi(x)$. Consider $\mathcal{F}(1) = \{w^\top \psi(x) : \|w\|_1 \le A_1\}$, and $\mathcal{F}(\theta) = \{w^\top \psi(x) : A_{\theta-1} \le \|w\|_1 \le A_\theta\}$ for $\theta > 1$. We have

$$R_n(\mathcal{F}(\theta), \mathcal{D}) \le \gamma A_\theta R_n(\Psi_\pm, \mathcal{D}).$$

Given any w, let θ be the smallest number such that $f(w,x) = w^\top \Psi(x) \in \mathcal{F}(\theta)$, then $A_\theta \le 2(A_0 + \|w\|_1)$. Therefore

$$L(f(x), y) \le L(0, y) + \gamma |f(x)| \le M_0 + \gamma A_\theta B \le M_0 + 2\gamma(A_0 + \|w\|_1)B.$$

In Theorem 5, we take $M(\theta) \le M_0 + 2\gamma(A_0 + \|w\|_1)B$, and $1/q(\theta) \le (2 + \log_2(1 + \|w\|_1/A_0))^2$. Let

$$\begin{aligned}\tilde{R}(\theta, f, \mathcal{S}_n) =& 4\gamma(A_0 + \|w\|_1)R_n(\Psi_\pm, \mathcal{D}) \\ &+ (M_0 + 2\gamma B(A_0 + \|w\|_1))\sqrt{\frac{\ln(2 + \log_2(1 + \|w\|_1/A_0))}{n}}. \quad (6)\end{aligned}$$

This implies the desired uniform convergence result.

L_1 Penalty Regularization: Oracle Inequality

Corollary 17 (Oracle Inequality in Cor 10.14)

Assume that $L(p, y) \geq 0$ *is* γ *Lipschitz,* $M_0 = \sup_y L(0, y)$*, and* $B = \sup_{x, \psi \in \Psi} |\psi(x)|$*. Consider* $A_0 > 0$*, and* (5) *with*

$$\lambda \geq 4\gamma R_n(\Psi_\pm, \mathcal{D}) + 2\gamma B \sqrt{\frac{\ln(2 + \log_2(1 + M_0/(\lambda A_0)))}{n}}.$$

We have the following oracle inequality. With probability at least $1 - \delta$*:*

$$\mathbb{E}_{\mathcal{D}} L(\hat{w}^\top \psi(X), Y) \leq \inf_w \left[\mathbb{E}_{\mathcal{D}} L(w^\top \psi(X), Y) + \left(\lambda + 4\gamma B \sqrt{\frac{\ln(2/\delta)}{2n}} \right) \|w\|_1 \right] + \epsilon_n(\delta), \qquad \textit{where}$$

$$\epsilon_n(\delta) = 4\gamma A_0 R_n(\Psi_\pm, \mathcal{D}) + (M_0 + 2\gamma A_0 B) \sqrt{\frac{\ln((2 + \log_2(1 + M_0/(\lambda A_0))))}{n}} + (2M_0 + 4\gamma A_0 B) \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof of Corollary 17

Following the proof of Corollary 16. With the condition of λ, we have $\|\hat{w}\|_1 \leq M_0/\lambda$, and by considering $\|w\|_1 \leq M_0/\lambda$, we can redefine

$$\tilde{R}(\theta) = \tilde{R}(\theta, f, \mathcal{S}_n) = \lambda\|w\|_1 + 4\gamma A_0 R_n(\Psi_{\pm}, \mathcal{D}) + (M_0 + 2\gamma A_0 B)\sqrt{\frac{\ln(2 + \log_2(1 + M_0/(\lambda A_0)))}{n}}.$$

This definition of $\tilde{R}(\theta)$ is an upper bound of (6). We can thus apply Theorem 5 again to obtain the desired oracle inequality, where we also use $2M_0 + 4\gamma A_0 B + 4\gamma B\|w\|_1$ as an upper bound for $2M(\theta)$.

Example 18

Consider (5) with a function class Ψ of finite VC-dimension (or pseudo-dimension) $\mathrm{VC}(\Psi_\pm) = d$, which includes the two-layer neural network as a special case.

Under the assumptions of Corollary 17, we have $R_n(\Psi_\pm, \mathcal{D}) = O(B\sqrt{d/n})$ (see Example 6.26). We can take $A_0 = M_0/(\gamma B)$ and set

$$\lambda = \tilde{O}\left(\gamma B \sqrt{\frac{d}{n}}\right)$$

to obtain

$$\mathbb{E}_{\mathcal{D}} L(\hat{w}^\top \psi(X), Y) \le \inf_w \left[\mathbb{E}_{\mathcal{D}} L(w^\top \psi(X), Y) + \tilde{O}\left(\gamma B \sqrt{\frac{d + \ln(1/\delta)}{n}} \|w\|_1\right)\right] + \tilde{O}\left(M_0 \sqrt{\frac{d + \ln(1/\delta)}{n}}\right).$$

We use the notation $\tilde{O}(\cdot)$ to hide log-factors.

Posterior Averaging as Additive Model

In a continuous additive model, with features $\psi(\theta, x)$ ($\theta \in \Theta$),

$$f(w, x) = \int w(\theta)\psi(\theta, x)\, d\mu(\theta),$$

where $d\mu(\theta)$ is a measure on Θ.

As a special case, we consider $w(\theta)d\mu(\theta)$ as a probability measure. We use $q(\theta)$ instead of $w(\theta)$ to denote the fact that this is a probability measure.

Model Averaging

Our goal is to find a distribution q on Θ (also referred to as *posterior distribution*), such that the additive model is

$$f(q, x) = \int \psi(\theta, x)q(\theta)d\theta = \mathbb{E}_{\theta\sim q(\cdot)}\psi(\theta, x). \tag{7}$$

Entropy Regularization

We consider the entropy regularization to regularize the posterior distribution q:

$$\mathrm{KL}(q||q_0) = \int q(x) \ln \frac{q(\theta)}{q_0(\theta)} d\theta.$$

Corollary 19 (Cor 10.17)

Let $\mathcal{F}_A = \{f(q,x) : \mathrm{KL}(q||q_0) \leq A^2\}$ *be entropy regularized functions of* (7)*. Then*

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

Proof of Corollary 19

We have

$$\mathbb{E}_\sigma \sup_q \left[\frac{1}{n} \sum_{i=1}^n \sigma_i \mathbb{E}_{\theta \sim q} \psi(\theta, X_i) - \frac{\lambda}{2} \mathrm{KL}(q||q_0) \right]$$

$$= \frac{\lambda}{2} \mathbb{E}_\sigma \ln \mathbb{E}_{\theta \sim q_0} \exp \left[\frac{2}{\lambda n} \sum_{i=1}^n \sigma_i \psi(\theta, X_i) \right]$$

$$\leq \frac{\lambda}{2} \ln \mathbb{E}_{\theta \sim q_0} \mathbb{E}_\sigma \exp \left[\frac{2}{\lambda n} \sum_{i=1}^n \sigma_i \psi(\theta, X_i) \right]$$

$$\leq \frac{\lambda}{2} \ln \mathbb{E}_{\theta \sim q_0} \exp \left[\frac{2}{\lambda^2 n^2} \sum_{i=1}^n \psi(\theta, X_i)^2 \right].$$

The first equation follows from Proposition 7.16. The first inequality used Jensen's inequality and the concavity of $\ln(\cdot)$. The second inequality follows from the sub-Gaussian exponential inequality. It follows that

$$R(\mathcal{F}_A, \mathcal{S}_n) \leq \frac{\lambda}{2} A^2 + \frac{1}{\lambda n^2} \sup_\theta \sum_{i=1}^n \psi(\theta, X_i)^2.$$

By optimizing over λ, we obtain the desired bound.

Example: Finite Function Family

Corollary 19 holds for general function classes. In the case of finite family with $\Theta = \{\theta_1, \ldots, \theta_m\}$, and $q_0(\theta) = 1/m$ for all θ, the following inequality always holds:

$$\mathrm{KL}(q||q_0) \leq \ln m$$

for all q. Therefore entropy regularization implies a bound for L_1 regularization with nonnegative constraint $\sum_{j=1}^{m} w_j = 1$ and $w_j \geq 0$. Since this is exactly the convex hull of $\Psi = \{\psi(\theta, x)\}$, Corollary 19 implies that

$$R\left(\mathrm{CONV}(\Psi), S_n\right) \leq \sqrt{\frac{2 \ln m}{n}} \sup_{\theta} \|\psi(\theta, \cdot)\|_{L_2(S_n)},$$

which is identical to the Rademacher complexity of convex hull of finite function class obtained earlier.

Information Theoretical Analysis

We now consider the notations introduced in Section 3.3, where we are interested in minimizing a loss function

$$\phi(w, z) : \Omega \times \mathcal{Z} \to \mathbb{R}.$$

Randomized Prediction

We consider a general randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \Delta(\Omega)$, where $\Delta(\Omega)$ denotes probability measures on Ω.
In this setting, given training data $\mathcal{S}_n$, $\mathcal{A}(\mathcal{S}_n)$ returns a posterior distribution $\hat{q}$ on Ω. It then randomly draws a model from $\hat{q}$ to make prediction.

We want to derive information theoretical generalization bound for an arbitrary randomized learning algorithm.

Information Theoretical Uniform Convergence Bound

Theorem 20 (Expected Generalization Bound in Thm 10.18)

Consider a randomized algorithm $\mathcal{A}$ that returns a distribution $\hat{q}(w|\mathcal{S}_n)$ on the parameter space Ω for each training data $\mathcal{S}_n \in \mathcal{Z}^n$. Then for any data independent distribution q_0 on Ω and $\lambda > 0$:

$$\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{w\sim\hat{q}(\cdot|\mathcal{S}_n)}\Lambda\left(1/(\lambda n), w\right) \leq \mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{w\sim\hat{q}(\cdot|\mathcal{S}_n)}\frac{1}{n}\sum_{i=1}^{n}\phi(w, Z_i) + \lambda\mathbb{E}_{\mathcal{S}_n}\mathrm{KL}(\hat{q}||q_0),$$

where

$$\Lambda(\lambda, w) = -\frac{1}{\lambda}\ln \mathbb{E}_{Z\sim\mathcal{D}}\exp(\ -\lambda\phi(w, Z)).$$

High probability result, referred to as *PAC-Bayes* analysis, can also be obtained (see Theorem 10.18).

Proof of Theorem 20

Let $\Delta(\mathcal{S}_n) = \sup_{\hat{q}} \left[\mathbb{E}_{w\sim\hat{q}(\cdot|\mathcal{S}_n)} \left(\Lambda(1/(\lambda n), w) - \frac{1}{n}\sum_{i=1}^{n} \phi(w, Z_i) \right) - \lambda \mathrm{KL}(\hat{q}||q_0) \right]$,

then

$$\lambda^{-1}\mathbb{E}_{\mathcal{S}_n\sim\mathcal{D}^n}\,\Delta(\mathcal{S}_n) \le \ln \mathbb{E}_{\mathcal{S}_n\sim\mathcal{D}^n} \exp(\lambda^{-1}\Delta(\mathcal{S}_n)) \qquad \text{(Jensen's inequality)}$$

$$= \ln \mathbb{E}_{\mathcal{S}_n\sim\mathcal{D}^n} \exp\left[\ln \mathbb{E}_{w\sim q_0} \exp\left(\lambda^{-1}\Lambda(1/(\lambda n), w) - \frac{\lambda^{-1}}{n}\sum_{i=1}^{n}\phi(w, Z_i) \right) \right]$$

(Proposition 7.16)

$$= \ln \mathbb{E}_{w\sim q_0} \left[\exp\left(\lambda^{-1}\Lambda(1/(\lambda n), w)\right) \mathbb{E}_{\mathcal{S}_n\sim\mathcal{D}^n} \exp\left(-\frac{\lambda^{-1}}{n}\sum_{i=1}^{n}\phi(w, Z_i) \right) \right]$$

$$= \ln \mathbb{E}_{w\sim q_0} \left[\exp\left(\lambda^{-1}\Lambda(1/(\lambda n), w)\right) \left(\mathbb{E}_{Z\sim\mathcal{D}} \exp\left(-\frac{\lambda^{-1}}{n}\phi(w, Z) \right) \right)^n \right]$$

(Independence of Z_i)

$$= \ln \mathbb{E}_{w\sim q_0} \left[\exp\left(\lambda^{-1}\Lambda(1/(\lambda n), w)\right) \left(\exp\left(-\frac{1}{\lambda n}\Lambda(1/(\lambda n), w) \right) \right)^n \right] = 0.$$

Example: Bounded Function Class

Example 21 (Expl 10.19)

Assume that $\phi(w, Z) \in [0, M]$. Then

$$\forall w \in \Omega : \quad -\Lambda(\lambda, w) \leq -\phi(w, \mathcal{D}) + \frac{\lambda M^2}{8}.$$

Then we obtain from Theorem 20 the following generalization bound. For any $\lambda > 0$ and learning algorithm $\hat{q}$:

$$\mathbb{E}\,\mathbb{E}_{w\sim\hat{q}(\cdot|\mathcal{S}_n)}\phi(w, \mathcal{D}) \leq \mathbb{E}\left[\mathbb{E}_{w\sim\hat{q}(\cdot|\mathcal{S}_n)}\phi(w, \mathcal{S}_n) + \lambda \mathrm{KL}(\hat{q}||q_0) + \frac{M^2}{8\lambda n}\right].$$

Mutual Information Bound

Define the mutual information of $\mathcal{A}$ and $\mathcal{S}_n$ as follows:

$$I(\mathcal{A}, \mathcal{S}_n) = \mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{w \sim \hat{q}(\cdot|\mathcal{S}_n)} \ln \frac{\hat{q}(w|\mathcal{S}_n)}{\hat{q}(w)}, \qquad \hat{q}(w) = \mathbb{E}_{\mathcal{S}_n} \hat{q}(w|\mathcal{S}_n).$$

Corollary 22 (Mutual Information Bound, Cor 10.22)

Under the assumptions of Theorem 10.18, we have the following expected generalization bound for all $\lambda > 0$:

$$\mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{\mathcal{A}} \Lambda(\lambda, \mathcal{D}) \leq \mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{\mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \phi(w, Z_i) + \lambda I(\mathcal{A}, \mathcal{S}_n),$$

where $\mathbb{E}_{\mathcal{A}}$ denotes the expectation over the randomization of algorithm $\mathcal{A}$: that is, $w \sim \hat{q}(\cdot|\mathcal{S}_n)$.

Mutual information optimizes the expected KL divergence over prior q_0:

$$I(\mathcal{A}, \mathcal{S}_n) = \inf_{q_0} \mathbb{E}_{\mathcal{S}_n} \mathrm{KL}(\hat{q} || q_0).$$

Example: Bounded Function Class

Example 23

Assume that $\phi(w, z) \in [0, M]$. Then

$$\Lambda(\lambda) \leq \frac{\lambda^2 M^2}{8}.$$

We obtain from Corollary 22:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\phi(w, \mathcal{D}) \leq & \mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\phi(w, Z_i) + \inf_{\lambda>0}\left[\lambda I(\mathcal{A}, \mathcal{S}_n) + \frac{M^2}{8\lambda n}\right] \\ = & \mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\phi(w, Z_i) + M\sqrt{\frac{I(\mathcal{A}, \mathcal{S}_n)}{2n}}.\end{aligned}$$

Gibbs Algorithm

We now consider the Gibbs Algorithm

$$\hat{q}(w|\mathcal{S}_n) \propto q_0(\theta) \exp\left(-\frac{1}{\lambda n}\sum_{i=1}^{n}\phi(w, Z_i)\right). \tag{8}$$

Corollary 24 (Cor 10.25)

The following expected oracle inequality holds for the Gibbs distribution (8)*:*

$$\mathbb{E}_{\mathcal{S}_n}\mathbb{E}_{w\sim\hat{q}}\Lambda(1/(\lambda n), w) \leq \inf_q \left[\mathbb{E}_{w\sim q}\phi(w, \mathcal{D}) + \lambda \mathrm{KL}(q||q_0)\right],$$

where $\Lambda(\cdot)$ *is defined as*

$$\Lambda(\lambda, w) = -\frac{1}{\lambda}\ln \mathbb{E}_{Z\sim\mathcal{D}} \exp(\ -\lambda\phi(w, Z))$$

in Theorem 20.

Proof of Corollary 24

From Proposition 7.16, $\hat{q}$ is the solution of the following regularized empirical risk minimization problem:

$$\hat{q} = \arg\min_{q} \left[\mathbb{E}_{\theta \sim q} \frac{1}{n} \sum_{i=1}^{n} \phi(w, Z_i) + \lambda \mathrm{KL}(q||q_0) \right]. \tag{9}$$

Therefore for any q, we obtain from Theorem 20

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{w \sim \hat{q}} \Lambda(1/(\lambda n), w) \leq & \mathbb{E}_{\mathcal{S}_n} \left[\mathbb{E}_{w \sim \hat{q}} \frac{1}{n} \sum_{i=1}^{n} \phi(w, Z_i) + \lambda \mathrm{KL}(\hat{q}||q_0) \right] \\
\leq & \mathbb{E}_{\mathcal{S}_n} \left[\mathbb{E}_{w \sim q} \frac{1}{n} \sum_{i=1}^{n} \phi(w, Z_i) + \lambda \mathrm{KL}(q||q_0) \right] \\
= & \left[\mathbb{E}_{w \sim q} \phi(w, \mathcal{D}) + \lambda \mathrm{KL}(q||q_0) \right].
\end{aligned}$$

The first inequality used Theorem 20. The second inequality used (9). The last equation used the fact that $Z_i \sim \mathcal{D}$. This implies the result.

Conditional Density Estimation

Corollary 25 ($\alpha = 0.5$ in Cor 10.26)

Consider the following Gibbs algorithm:

$$\hat{q}(w|\mathcal{S}_n) \propto q_0(\theta) \exp\left(0.5 \sum_{i=1}^{n} \ln p(Y_i|w, X_i)\right).$$

Then

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} \mathbb{E}_{w \sim \hat{q}} \mathbb{E}_{X \sim \mathcal{D}} H(p_*(\cdot|X)||p(\cdot|w, X))^2 \\ &\leq \inf_q \left[\mathbb{E}_{w \sim q} \mathbb{E}_{X \sim \mathcal{D}} \mathrm{KL}(p_*(\cdot|X)||p(\cdot|w, X)) + \frac{2\mathrm{KL}(q||q_0)}{n}\right], \end{aligned}$$

where $H(\cdot)$ is the Hellinger distance:

$$H(p(\cdot|X)||p'(\cdot|X))^2 = \mathbb{E}_{Y \sim p(\cdot|X)} \left(\sqrt{\frac{p'(Y|X)}{p(Y|X)}} - 1\right)^2$$

Proof of Corollary 25

The Gibbs algorithm is equivalent to

$$\phi(w, Z) = (\ln p_*(Y|X) - \ln p(Y|w, X))$$

and $\lambda = 2/n$ in Corollary 24.

$$\begin{aligned}
\Lambda(1/(\lambda n), w) &= -2 \ln \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left(\frac{p(Y|w, X)}{p_*(Y|X)} \right)^{0.5} \\
&= -2 \ln \left[1 - 0.5 \mathbb{E}_{X\sim\mathcal{D}} H(p(\cdot|X)||p'(\cdot|X))^2 \right] \\
&\geq \mathbb{E}_{X\sim\mathcal{D}} H(p(\cdot|X)||p'(\cdot|X))^2.
\end{aligned}$$

Moreover,

$$\mathbb{E}_{\mathcal{D}} \phi(w, Z) = \mathbb{E}_X \mathrm{KL}(p_*(\cdot|X)||p(\cdot|w, X)).$$

The desired bound follows directly from Corollary 24.

Boosting

Boosting is a popular method to learn both w_j and θ_j in (1) sequentially for $j = 1, 2, \ldots$

In AdaBoost, we consider binary classification problem $\psi \in \{\pm 1\}$ and $Y_i \in \{\pm 1\}$.
Assume we have an algorithm $\mathcal{A}$ that can learn $\hat{\theta} = \mathcal{A}(\widetilde{\mathcal{S}}_n)$ from any weighted version of data $\widetilde{\mathcal{S}}_n = \{(\rho_i, X_i, Y_i) : i = 1, \ldots, n\}$:

$$\hat{\theta} \approx \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho_i \mathbb{1}(\psi(\theta, X_i) \neq Y_i), \tag{10}$$

where $\rho_i \geq 0$.
The learner $\mathcal{A}$ is often referred to as a weak learner. AdaBoost finds a strong learner using a sequence of fitting with the weak learner $\mathcal{A}$.

AdaBoost

Algorithm 1: AdaBoost

Input: $\mathcal{S}_n$, Ψ

Output: $f^{(T)}(x)$

Let $f^{(0)}(x) = 0$

Let $\rho_1 = \cdots = \rho_n = 1/n$

for $t = 1, 2, \ldots, T$ **do**

  Find θ_t by approximately solving

  $\theta_t \approx \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho_i \mathbb{1}(\psi(\theta, X_i) Y_i \leq 0)$

  Let $r_t = \sum_{i=1}^{n} \rho_i \psi(\theta_t, X_i) Y_i$

  Let $w_t = \frac{1}{2} \ln((1 + r_t)/(1 - r_t))$

  Let $\rho_i = \rho_i \cdot \exp(-w_t \psi(\theta_t, X_i) Y_i)$ for $i = 1, \ldots, n$.

  Normalize ρ_i so that $\sum_{i=1}^{n} \rho_i = 1$

  Let $f^{(t)}(x) = f^{(t-1)}(x) + w_t \psi(\theta_t, x)$

Return: $f^{(T)}(x)$

AdaBoost is Greedy Algorithm

Theorem 26 (Thm 10.29)

Assume that $\Psi = \Psi_{\pm}$, $\psi(\theta, x) \in \{\pm 1\}$, *and* $y \in \{\pm 1\}$. *Then AdaBoost implements the greedy algorithm to minimize the loss function*

$$L(f(x), y) = \exp(-f(x)y).$$

That is, at each time t, AdaBoost (with exact minimization in Line 5 of Algorithm 1) solves the following problem:

$$[w_t, \theta_t] = \arg\min_{w,\theta} \sum_{i=1}^{n} e^{-(f^{(t-1)}(X_i) + w\psi(\theta, X_i))Y_i}.$$

Moreover, the prediction function $f^{(T)}$ *obtained by Algorithm 1 satisfies*

$$\frac{1}{n} \sum_{i=1}^{n} e^{-f^{(T)}(X_i)Y_i} \leq \prod_{t=1}^{T} \sqrt{1 - r_t^2}.$$

Margin Bound for AdaBoost

Corollary 27 (Cor 10.30)

Under the assumptions of Theorem 26, and assume further that Ψ *has VC-dimension* d*. Let* $\|f^{(T)}\|_1 = \sum_{t=1}^{T} w_t$*. Assume that for* $t = 1, \ldots, T$*, we have* $r_t \geq r_0 > 0$ *in Algorithm 1. Then* $\exists C > 0$ *so that with probability at least* $1 - \delta$*:*

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}} \mathbb{1}(f^{(T)}(X)Y \leq 0) \leq \underbrace{\frac{1.5}{n}\sum_{i=1}^{n} \mathbb{1}(f^{(T)}(X_i)Y_i \leq 1)}_{\text{margin error}}$$
$$+ C\frac{(\|f^{(T)}\|_1 + 1)^2 d \ln n \ln(n + \|f^{(T)}\|_1) + \ln(1/\delta)}{n},$$

where margin error is upper bounded by

$$\exp\left(1 - 0.4\sum_{t=1}^{T} \min(1, w_t^2)\right) \leq \exp\left(1 - \frac{T}{10}\min\left(2, \ln\frac{1+r_0}{1-r_0}\right)^2\right).$$

Gradient Boosting for General Loss Function

More generally, we assume that the weak learner (or base learner) $\mathcal{A}$ can approximately solve the least squares regression problem:

$$\min_{w \in \mathbb{R}, \theta \in \Theta} \sum_{i=1}^{n} [w\psi(\theta, X_i) + g_i]^2$$

This leads to the following algorithm.

Algorithm 2: Gradient Boosting

Input: $\mathcal{S}_n$, Ψ, $L(\cdot, \cdot)$

Output: $f^{(T)}(x)$

Let $f^{(0)}(x) = 0$

for $t = 1, 2, \ldots, T$ **do**

  Let $g_i = L_1'(f^{(t-1)}(X_i), Y_i)$ $(i = 1, \ldots, n)$ be the functional gradients

  Solve for $[w_t, \theta_t] = \arg\min_{w \in \mathbb{R}, \theta \in \Theta} \sum_{i=1}^{n} [w\psi(\theta, X_i) + g_i]^2$

  Let $f^{(t)}(x) = f^{(t-1)}(x) + w_t\psi(\theta_t, x)$

Return: $f^{(T)}(x)$

Convergence of Gradient Boosting

Under suitable conditions, gradient boosting can find a solution approaching the minimum of

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(f(X_i), Y_i),$$

where

$$\mathcal{F} = \bigcup \{A \cdot \text{CONV}(\Psi_{\pm}) : A \in \mathbb{R}\}$$

is the linear span of Ψ.

Sparse Recovery

Consider sparse linear regression model

$$Y = X\bar{w} + \epsilon. \tag{11}$$

- X: $n \times p$ design matrix
- Y: n dimensional observation vector.
- ϵ: n-dimensional zero-mean noise vector with independent components.
- Model parameter w: p-dimensional.
- Sparsity: $\|\bar{w}\|_0 \ll p$.

Note that due to the rescaling mentioned above, we will assume that columns of X are bounded in 2-norms. The corresponding proper scaling of ϵ is to assume that each $\sqrt{n}\epsilon_i$ is σ sub-Gaussian.

Lasso and Sparse Recovery

L_1 Regularization

Consider L_1 regularization (Lasso):

$$\hat{w} = \arg\min_{w} Q_{L_1}(w); \quad Q_{L_1}(w) = \frac{1}{2}\|Xw - Y\|_2^2 + \lambda\|w\|_1, \tag{12}$$

where $\lambda > 0$ is an appropriately chosen regularization parameter.

Under appropriate conditions, one can recover the true sparse parameter $\bar{w}$ using Lasso. This is referred to as *sparse recovery*.

- (Support Recovery) Whether Lasso finds the correct feature set: $\mathrm{supp}(\hat{w}) = \mathrm{supp}(\bar{w})$? Moreover, we say the Lasso solution is *sign consistent* if $\mathrm{supp}(\hat{w}) = \mathrm{supp}(\bar{w})$ and $\mathrm{sign}(\hat{w}_j) = \mathrm{sign}(\bar{w}_j)$ when $j \in \mathrm{supp}(\bar{w})$.
- (Parameter Recovery) How good is the parameter estimation, or how small is $\|\hat{w} - \bar{w}\|_2$?

Support Recovery: Irrepresentable Condition

Assumption 28

Let $\bar{F} = \operatorname{supp}(\bar{w})$. *Assume that* $X_{\bar{F}}^\top X_{\bar{F}}$ *is positive definite. Define*

$$\mu = \sup_{j \notin \bar{F}} |X_j^\top X_{\bar{F}} (X_{\bar{F}}^\top X_{\bar{F}})^{-1} \operatorname{sign}(\bar{w})_{\bar{F}}|.$$

Then the condition

$$\mu < 1$$

is referred to as irrepresentable condition*.*

It is known that if elements of $\sqrt{n}X$ have iid standard normal distributions, then for any fixed $\bar{w}$, such that $\|\bar{w}\|_0 = s$, the irrepresentable condition holds with high probability when $n = \Omega(s \ln p)$.

Support Recovery: Theory

Theorem 29 (Thm 10.37)

Assume that the irrepresentable condition holds:

$$\mu = \sup_{j \notin \bar{F}} |X_j^\top X_{\bar{F}} (X_{\bar{F}}^\top X_{\bar{F}})^{-1} \mathrm{sign}(\bar{w})_{\bar{F}}| < 1.$$

Assume that we choose a sufficiently large λ so that

$$\lambda > (1-\mu)^{-1} \sup_{j \notin \bar{F}} |X_j^\top (X_{\bar{F}} (X_{\bar{F}}^\top X_{\bar{F}})^{-1} X_{\bar{F}}^\top - I)\epsilon|.$$

If the weight $\bar{w}$ is sufficiently large:

$$\min_{j \in \bar{F}} |\bar{w}_j| > \|(X_{\bar{F}}^\top X_{\bar{F}})^{-1}\|_{\infty \to \infty} (\lambda + \|X_{\bar{F}}^\top \epsilon\|_\infty),$$

then the solution of (12) *is unique and sign consistent. Here $\|M\|_{\infty \to \infty} = \sup_u [\|Mu\|_\infty / \|u\|_\infty]$ is the maximum absolute row sum of M.*

Parameter Recovery: Condition

Definition 30 (RE)

An $n \times p$ matrix X satisfies the restricted eigenvalue condition $\mathrm{RE}(F, c_0)$ for $F \subset [p]$ if the following quantity is nonzero:

$$\kappa_{\mathrm{RE}}(F, c_0) = \min_{w \neq 0, \|w\|_1 \leq c_0 \|w_F\|_1} \frac{\|Xw\|_2}{\|w\|_2}.$$

A more common condition is *restricted isometry property* (RIP), which is related to RE.

Both RIP and RE holds for all w such that $|\mathrm{supp}(w)| = O(s)$ when $n = \Omega(s \ln p)$.

Sparse Recovery: Theory

Theorem 31 (Thm 10.42)

Let $\bar{F} = \operatorname{supp}(\bar{w})$. *Assume that the columns are normalized so that* $\sup_j \|X_j\|_2 \leq B$ *and (with the corresponding proper scaling) components of* $\sqrt{n}\epsilon$ *are independent zero-mean* σ *sub-Gaussian noise:*

$$\ln \mathbb{E} e^{\lambda \epsilon_i} \leq \lambda^2 \sigma^2 / (2n).$$

Assume that

$$\lambda \geq 2\sigma B \sqrt{\frac{2 \ln(2p/\delta)}{n}}.$$

Then with probability at least $1 - \delta$, *the solution of* (12) *satisfies*

$$\|\hat{w} - \bar{w}\|_2^2 \leq \frac{16\lambda^2 \|\bar{w}\|_0}{\kappa_{\mathrm{RE}}(\bar{F}, 4)^2}.$$

Summary (Chapter 10)

- Additive Model
- Sparsity and L_0 Regularization
- Rademacher Complexity for L_0 Regularization
- L_1 regularization and Rademacher Complexity
- Information Theoretical Analysis with Entropy Regularization
- Boosting and Greedy Algorithm (brief)
- Sparse Recovery (brief)