Mathematical Tools for machine learning

Definition : (The realizability assumption) if $\exists f^*$ s.t. $L_{x\sim D}(f^*) = 0$.

$L_{x\sim D}(f^*) = P(f^* \neq f)$

$L_S(f^*)$ = empirical risk $\Rightarrow L_S(f^*) = 0 \geq L_S(f_S^{ERM})$

A formal learning model (chapter 3)

- Waive the realizability assumption : Given a hypothesis class $\mathcal{H}$, we do not assume anymore that $\exists h^* \in \mathcal{H} : L_D(h^*) = \mathbb{P}_{(x,y)}(h^*(x) \neq y) = 0$.

This leads to agnostic PAC learning :

Definition: A hypothesis class $\mathcal{H}$ is called agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm s.t. $\forall (\varepsilon, \delta) \in (0,1)^2$, $\forall$ distribution $\mathcal{D}$ on $X \times Y$, if the algorithm is run on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid examples generated by $\mathcal{D}$, this algorithm returns a predictor h_S (↑ training set) s.t. $\mathbb{P}_{S\sim D^m}(L_D(h_S) \leq \min_{h\in\mathcal{H}} L_D(h) + \varepsilon) \geq 1-\delta$

Remark:
- $m_{\mathcal{H}}(\varepsilon, \delta)$ is viewed as the smallest sample size for which the learning algorithm satisfied.
- In case the realizability assumption, $\min_{h\in\mathcal{H}} L_D(h) = 0$.
  $\Rightarrow$ agnostic PAC = PAC.

- Extending classification to other types of learning
  - Multiclass classification : Consider the problem of classifying some document into 4 categories

    Domain set : $\mathcal{X} = \mathbb{N}_0^p$ with $p > 0$ some integer.

    ↳ $x \in \mathcal{X}$ is a vector storing counts for some specific key words.

    label set : $Y = \{1, 2, \ldots, k\}$ where $k \geq 2$ some integer.

    Training set : $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

A prediction rule h_S is the output of some learning algorithm for a new document with associate feature $x \in \mathcal{X}$ $h_S(x) = y \in \{1, \dots, k\}$ the prediction class

- Regression: find a linear relationship between $y \in \mathcal{Y} = \mathbb{R}$ and covariates $x \in \mathcal{X} = \mathbb{R}^p$ for some $p > 0$.

  A linear model assumes that $\mathbb{E}[Y|X] = \beta_0^T x$. β_0 unknown. the true risk (error) is

  $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(h(x) - y)^2]$: expected squared error.

The measure of error can be generalized to any loss function.

- Generalized loss function

Consider a hypothesis class $\mathcal{H}$ and some domain $\mathcal{Z}$ (so far $= \mathcal{X} \times \mathcal{Y}$).
Also consider a function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty)$
such a function ℓ is called a loss-function. for a given $h \in \mathcal{H}$, the true error (risk) is now defined, as $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$.

the training error (empirical risk) is defined as $\mathcal{D}$

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) \quad \text{where} \quad S = \{z_1, \dots, z_m\}$$

Typical loss functions are:

* 0-1 loss $\ell_{0-1}(h, z) = \ell_{0-1}(h, (x, y)) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{o.w.} \end{cases}$

$$= \mathbb{1}_{\{h(x) \neq y\}}$$

* squared loss $\ell_{sq}(h, z) = (h(x) - y)^2$

Definition (more generalized Agnostic PAC):
$\mathcal{H}$ is agnostic PAC to the domain $\mathcal{Z}$ and loss function if (the same as before), where $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$.

Learning via Uniform Convergence (chapter 4)

Introduction: We know that under realizability assumption, finite classes

is PAC learnable.

Question: How can we show that finite classes are agnostic PAC learnable? and what assumption on the loss function?

Answer: Uniform convergence.

Definition (ε- representative sample)

A training set S is called ε- representative w.r.to. a domain Z, a hypothesis class H, a loss function ℓ and a distribution $\mathcal{D}$ if $\forall h \in \mathcal{H} \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$

Lemma: Assume that a training set S is $(\frac{\varepsilon}{2})$- representative w.r.to. Z, H ℓ and $\mathcal{D}$. Then for any output h_S of $ERM_H(S)$, that is $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$ we have that

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.$$

proof: since S is $(\frac{\varepsilon}{2})$- representative, we have

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\varepsilon}{2} \\ &\leq L_S(h) + \frac{\varepsilon}{2} \quad \forall h \in H \text{ since } h \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h) \\ &\leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= L_{\mathcal{D}}(h) + \varepsilon. \end{aligned}$$

$$\Leftrightarrow L_{\mathcal{D}}(h_S) \leq \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon. \qquad \square$$

Definition (uniform convergence)

We say that a hypothesis class H has the uniform convergence property (w.r.to. Z and ℓ) if there exists a function $m_H^{UC} : (0,1)^2 \to \mathbb{N}$ s.t. $\forall (\varepsilon, \delta) \in (0,1)^2$ $\forall$ distribution $\mathcal{D}$ if S is a training set with $m \geq m_H^{UC}(\varepsilon, \delta)$ example iid $\sim \mathcal{D}$, then

$$\mathbb{P}_{S \sim \mathcal{D}^m}(S \text{ is } \varepsilon\text{- representative}) \geq 1 - \delta$$

$$\Leftrightarrow (\forall h \in H \; |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon)$$

smallest one.

Corollary If a hypothesis class H has the uniform convergence property (w.r.to Z and ℓ) with a function m_H^{UC}, then this function class is agnostic PAC learnable with sample complexity $m_H(\varepsilon, \delta) \leq m_H^{UC}(\frac{\varepsilon}{2}, \delta)$

Furthermore, the ERM predictors is an agnostic PAC learnt for H.

proof: $\mathbb{P}\left(L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon \right) \geq \mathbb{P}\left(L_D(h_S) - L_S(h_S) + L_S(h^*) - L_D(h^*) \leq \varepsilon \right) \geq 1 - \delta$.

(h_S: erm. Probability to get $\forall$ / a $\varepsilon/2$-representative $\forall m \geq m_H^{uc}\left(\frac{\varepsilon}{2}, \delta\right)$)

This means H is agnostic PAC learnable with the ERM paradigm with sample complexity $m_H(\varepsilon, \delta) \leq m_H^{uc}\left(\frac{\varepsilon}{2}, \delta\right)$. □

- Finite classes are agnotic PAC learnable.

Lemma (Hoeffding's inequality) Let $\theta_1, \dots, \theta_m$ be iid random variables s.t. $E[\theta_i] = \mu$, $i = \{1, \dots, m\}$ and $\mathbb{P}(a \leq \theta_i \leq b) = 1$. then

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} (\theta_i - \mu) \right| > \varepsilon \right) \leq 2 \exp\left(- \frac{2 m \varepsilon^2}{(b-a)^2} \right)$$

Auxiliary lemma: Let X be a random variable s.t. $E[X] = 0$ and $\mathbb{P}(X \in [a, b]) = 1$. Then, $\forall \lambda > 0$ we have $E[e^{\lambda X}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$

Proof: Put $f(x) = e^{\lambda x}$, $x \in \mathbb{R}$ for $\lambda \in (0, \infty)$. f is convex on $\mathbb{R}$ and hence $f(\alpha a + (1-\alpha) b) \leq \alpha f(a) + (1-\alpha) f(b)$. for $X \in [a, b]$ chose $\alpha = \frac{b-x}{b-a}$, $\alpha a + (1-\alpha) b = x$

$$\Rightarrow f(x) \leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$$

$$\Leftrightarrow e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}$$

$$\Rightarrow E[e^{\lambda X}] \leq \frac{b - E[X]}{b-a} e^{\lambda a} + \frac{E[X] - a}{b-a} e^{\lambda b}$$

$$= \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}$$

$h = \lambda(b-a)$ and $p = -\frac{a}{b-a}$

Also, define the function

$L(h) = -hp + \log(1 - p + p e^h)$ with $h \in [0, +\infty)$

we have $\frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} = e^{\lambda a}\left(1 - p + p e^{\lambda(b-a)}\right)$

$$= e^{-hp}\left(1 - p + p e^h\right) \quad (\lambda a = -hp)$$

$$= e^{L(h)}$$

In the following, we show that $L(h) \leq \frac{h^2}{8}$

$L(0) = 0$, $L'(h) = -p + \frac{p e^h}{1 - p + p e^h} \Rightarrow L'(0) = 0$.

$$L''(h) = +p \cdot \frac{e^h(1-p+pe^h) - pe^h e^h}{(1-p+pe^h)^2}$$

$$= \frac{p(1-p)e^h}{(1-p+pe^h)^2}$$

$$L''(h) - \frac{1}{4} = \frac{p(1-p)e^h}{(1-p+pe^h)^2} - \frac{1}{4} = \frac{4p(1-p)e^h - (1-p+pe^h)^2}{4(1-p+pe^h)^2} = -\frac{(1-p-pe^h)^2}{-(\cdots)} \leq 0$$

$\forall h \in (0, +\infty)$, $L''(h) \leq \frac{1}{4}$. Using Taylor expansion of L up to the 2nd order we can write:

$L(h) = L(0) + hL'(0) + \frac{h^2}{2}L''(h^*)$ where $h^* \in [0, h]$

$L(0) = L'(0) = 0$

$\Rightarrow L(h) = \frac{h^2}{2}\underbrace{L''(h)}_{\leq \frac{1}{4}} \leq \frac{h^2}{8}$ □

Now, we prove that the Hoeffding's inequality. Recall $\theta_1, \dots, \theta_m$ iid $\in [a, b]$

$X_i = \theta_i - \mu$ where $\mu = \mathbb{E}[\theta_i]$ for $i \in \{1, \dots, m\}$.

Note that $X_1, \dots, X_m$ are iid and $X_i \in [a-\mu, b-\mu]$. Using monotonicity of $x \mapsto e^{\lambda x}$ for $\lambda \in (0, +\infty)$ and the Markov's inequality, we know that $\mathbb{P}(\bar{X} > \varepsilon) = \mathbb{P}(e^{\lambda \bar{X}} > e^{\lambda \varepsilon}) \leq \mathbb{E}[e^{\lambda \bar{X}}] e^{-\lambda \varepsilon}$

where $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$ (the empirical mean of the X_is)

$$\mathbb{E}[e^{\lambda \bar{X}}] = \prod_{i=1}^m \mathbb{E}\left[e^{+\frac{\lambda}{m}X_i}\right] = \left(\mathbb{E}\left[e^{\frac{\lambda}{m}X_i}\right]\right)^m.$$

By the auxiliary result, we have

$$\mathbb{E}\left[e^{\frac{\lambda}{m}X_i}\right] \leq e^{\left(\frac{\lambda}{m}\right)^2 \frac{(b-a)^2}{8}}$$

Hence $\mathbb{P}(\bar{X} > \varepsilon) \leq e^{\frac{\lambda^2(b-a)^2}{8m} - \lambda\varepsilon} \quad \forall \lambda \in (0, +\infty)$

$$\Leftrightarrow \mathbb{P}(\bar{X} > \varepsilon) \leq \inf_{\lambda \in (0, +\infty)} e^{\frac{\lambda^2(b-a)^2}{8m} - \lambda\varepsilon} \Leftrightarrow \lambda^* = \frac{4\varepsilon m}{(b-a)^2} > 0.$$

$$\leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

The same argument can be used to $-X_i$ to show that $\mathbb{P}(-\bar{X} > \varepsilon) \leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$

$$\Rightarrow \mathbb{P}(|\bar{X}| > \varepsilon) \leq 2 \cdot e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

The proof is complete by noting that $\bar{X} = \frac{1}{m}\sum_{i=1}^m \theta_i - \mu_i$. □

$$|\bar{X}|(\omega) = \begin{cases} -\bar{X} \\ \bar{X} \end{cases}$$

Now we go back to the original problem : showing that any finite hypothesis class is Agnostic PAC learnable. In the following, we assume that the loss function, $\ell : H \times Z \to [0, +\infty)$, satisfies $\ell(h, z) \in [0, 1] \quad \forall h, z \in H \times Z$.

If we fix $h \in H$, $\theta = \ell(h, z)$, $S =$ training set $\{z_1, \dots, z_m\}$ are iid, write $\theta_i = \ell(h, z_i)$, $i = 1 \dots m$

$\theta_1 \dots \theta_m$ are also iid with true mean $\mu = E_{z \sim D}[\ell(h, z)]$ (true risk associated to h)

$\hookrightarrow L_S(h) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) = \frac{1}{m} \sum_{i=1}^{m} \theta_i$ (training error)

Note that $\theta_1, \dots, \theta_m \in [0, 1]$ by assumption, using the Hoeffding's inequality $D^m(S : |L_S(h) - L_D(h)| > \varepsilon) \leq 2e^{-2m\varepsilon^2}$ (loss function bounded)

$\Rightarrow D^m(S : \exists h \in H \text{ s.t. } |L_S(h) - L_D(h)| > \varepsilon) \leq |H| \times 2e^{-2m\varepsilon^2}$

we want $|H| \times 2e^{-2m\varepsilon^2} \leq \delta$ (*). If it is true, then $D^m(S : \forall h \in H \ |L_S(h) - L_D(h)| \leq \varepsilon) \geq 1 - \delta$

$$\Leftarrow \quad m \geq \frac{\log\left(\frac{2|H|}{\delta}\right)}{2\varepsilon^2}$$

Corollary Let H be a finite hypothesis class, Z a domain and $\ell : H \times Z \to [0, 1]$ be a loss function. Then, H enjoys the uniform convergence property with $m_H^{UC}(\varepsilon, \delta) \leq \frac{\log\left(\frac{2|H|}{\delta}\right)}{2\varepsilon^2}$.

Furthermore, H is Agnostic PAC learnable with the ERM paradigm.

The sample complexity satisfies

$$m_H(\varepsilon, \delta) \leq m_H^{UC}\left(\frac{\varepsilon}{2}, \delta\right) \leq \frac{2\log\left(\frac{2|H|}{\delta}\right)}{\varepsilon^2}$$

Question what if the class is infinite?

A possible answer "Discretization trick"

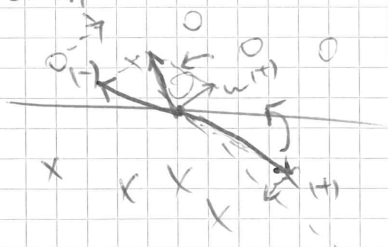
Example $X = \mathbb{R}$ and $Y = \{+1, -1\}$. $H = \{h_\theta, \theta \in \mathbb{R}\}$ where $h_\theta(x) = \text{sign}(x - \theta) \quad x \in \mathbb{R}$. $|H| = +\infty$

Since any real number can be represented by 64 bits, we can "approximate" H by $\tilde{H}$ where $|\tilde{H}| = 2^{64}$. Since we are in the framework of binary classification, the loss function

$$\ell(h, \underset{(x,y)}{z}) = 1_{h_\theta(x) \neq y} \in [0, 1]$$

Bayes estimator which is not consistent. for Jeffrey's Prior.

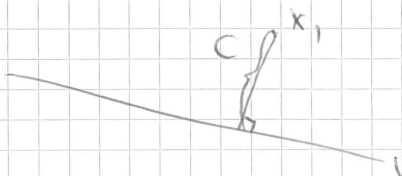
Remark:



Take bias into matrix => all have no bias passing through zero.

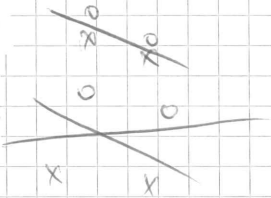
$w^{(t+1)} = w^{(t)} + y_i x_i \quad \langle w, x_i \rangle y_i \le 0 \ \forall i.$

$\Rightarrow \tilde{w} = \sum \alpha_i x_i$

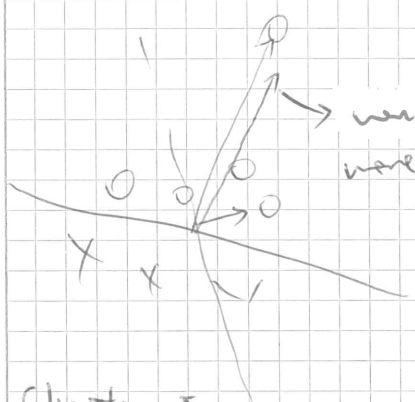


$\langle \frac{w}{\|w\|}, x_i \rangle$ = distance to hyperplane w from $x_1 \ge \frac{1}{B}$

$\Rightarrow B$ large difficult to classify



B small



want get more precise

$\Rightarrow R$ large difficult to classify.

Chapter 5

1 proof of No-Free-Lunch theorem:

Theorem: Let A be any learning algorithm for the task of binary classification with respect to some domain X and the 0-1 loss function. Let $m < \frac{|X|}{2}$ represent the training set size. $\exists$ distribution $\mathcal{D}$ on $X \times \{0,1\}$ such that

1. $\exists f: X \to \{0,1\}$ such that $L_{\mathcal{D}}(f) = 0$
2. $\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \ge \frac{1}{8} \right) > \frac{1}{7}$

→ $|X|$ 不能为有限 fixed!

proof: Consider a subset $C \subset X$ such that $|C| = 2m$.

There are $T = 2^{|C|} = 2^{2m}$ possible functions $C \to \{0,1\}$

Let us denote all these functions by $f_1, \dots, f_T$.

Fix $i \in \{1, \dots, T\}$. Consider $\mathcal{D}_i$ the distribution defined on $C \times \{0,1\}$ by

$$\mathcal{D}_i(\{x,y\}) = \begin{cases} \frac{1}{|C|} & \text{if } y = f_i(x) \quad x \in C \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{D}_i$ draws x from C uniformly (with the same probability $\frac{1}{2m}$), and conditionally on x, assigns $f_i(x)$. As y with probability 1/

The sample complexity $m_{\mathcal{H}}(\varepsilon,\delta) \leq \frac{2\log(\frac{2}{\delta})+1}{\varepsilon^2}$ $\varepsilon=0.1$ and $\delta=0.01$

$m_{\mathcal{H}} \leq 1380$.

Chapter 9 Linear Predictors

1 Introduction consider the following class $L_d = \{ h_{w,\theta} : w \in \mathbb{R}^d, \theta \in \mathbb{R} \}$

where $h_{w,\theta}(x) = \langle w, x \rangle + b$ for $x = (x_1, \dots x_d)^T \in \mathbb{R}^d$

$= w^T x + b$

$w = (w_1, \dots, w_d)^T = \sum_{i=1}^{d} w_i x_i + b$

w: vector of weights, b: bias

Given a function $\phi: \mathbb{R} \to Y$ we can consider also the class

$\phi \circ L_d = \{ \phi \circ h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R} \}$

Example: $X = \mathbb{R}^d$, $Y = \{-1, 1\}$ and $\phi(t) = \operatorname{sign}(t)$

The corresponding class (used for binary classification) is

$\operatorname{sign} \circ L_d = \{ \operatorname{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$

Remark: The bias θ can be incorporated in the vector of weights

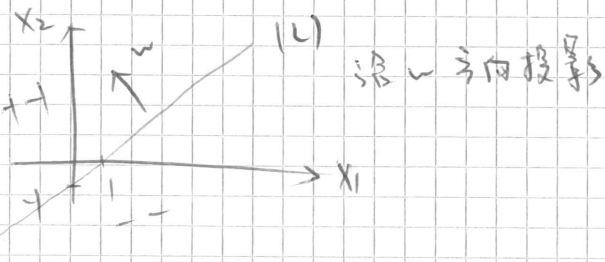
$\langle w, x \rangle + b = (b, w_1, \dots, w_d) \cdot \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix}$

2. Half spaces

Consider again the binary classification problem with $X = \mathbb{R}^d$, $Y = \{-1, 1\}$

$HS_d = \{ \operatorname{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$

Example: $d=2$, $w = (-1, 1)^T$ and $b=1$, $\langle w, x \rangle + b = -x_1 + x_2 + 1$



Remark: It can be shown that the class HS_d is VC of dimension $d+1$. Hence this class is Agnostic PAC learnable using ERM paradigm with sample complexity of order $\frac{d + \log(\frac{1}{\delta})}{\varepsilon}$

Sol 1:

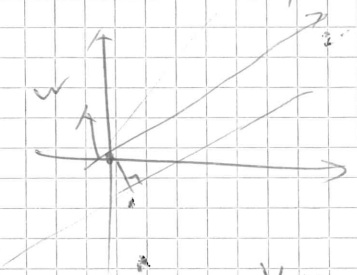
$\overline{w}$ and w^* both a perfect classifier. i.e. $y_i \langle w^{(T)}, x_i \rangle > 0$

Sol 2:

Randomly take one update $\leq 0 \longrightarrow$ add $\|x_i\|^2$ to/a positive guy.

In general, assume $y_i \langle w^*, x_i \rangle \geq 1$

It or stop in finite steps but B not known.



Logistic regression

Interpetation: $\mathbb{P}[y=1 \mid x]$ hopes to close to 0 or 1.

clearly, $L_{D_i}(f_i) \overset{\text{def}}{=} \mathbb{P}_{(x,y)\sim D_i}(f_i(x) \neq y)$

$= 0$

We are going to show for any learning algorithm A which receives a training set S of m examples and outputs a classifier $A(S): C \mapsto \{0,1\}$ if holds that $\max_{1 \leq i \leq T} \mathbb{E}_{S \sim D_i^m}[L_{D_i}(A(S))] \geq \frac{1}{4}$ (*).

相连的分布在 D_i 中 ←

⇕

There exists one distribution, the algorithm fails.

If (*) is true then by Exercise 5-1 the result follows.

Let us prove (*)

For $i \in \{1, \ldots, T\}$, consider a training set S of m i.i.d samples $\sim D_i$. Such a training set S looks like

$$S = \{(x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m))\}$$

$(x_1, \ldots, x_m)$ is a random drew with $|C|^m = (2m)^m$ possible outcomes which are all occuring with the same probability $\frac{1}{|C|^m}$. Put $k = |C|^m$

Also, let us list all possible training sets of sample $\sim D_i$ as $S_1^i, \ldots, S_k^i$.

$\max_{1 \leq i \leq T} \mathbb{E}_{S \sim D_i^m}[L_{D_i}(A(S))] \geq \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{S \sim D_i^m}[L_{D_i}(A(S))]$ ← finite.

↑ 大于平均值

$\frac{1}{k} \sum_{j=1}^{k} L_{D_i}(S_j^i)$ ← discrete.

关于 x fixed

变动 Turn to a fixed j to estimate

$= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$

station #1 $\geq \min_{1 \leq j \leq k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$ A 算法在所有 x 组合据 D_i 分布的最小值.

Now, fix $j \in \{1, \dots, k\}$. A training set S_j is of the form

$S_j = \{(x_1^j, f_j(x_1^j)), \dots, (x_m^j, f_j(x_m^j))\}$

Let $\{v_1, \dots, v_p\} = C \setminus \{x_1^j, \dots, x_m^j\}$ ← These are at most m distinct values in $\{x_1^j, \dots, x_m^j\}$. Since $|C| = 2m$, $p \geq m$

Let $h : C \to \{0,1\}$

$$L_{D_j}(h) \overset{\text{def}}{=} \mathbb{P}_{(x,y)\sim D_j}(h(x) \neq y)$$

$$= \mathbb{P}_{x \sim D_j^x}(h(x) \neq f_j(x))$$

(D_j^x: marginal distribution of x = uniform on C.)

$$= \frac{1}{2m} \sum_{c \in C} \mathbb{1}\{h(c) \neq f_j(c)\}$$ ← expectation of indicators.

$$\geq \frac{1}{2m} \sum_{r=1}^{p} \mathbb{1}\{h(v_r) \neq f_j(v_r)\}$$

$$\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\{h(v_r) \neq f_j(v_r)\}.$$

Replace h by $A(S_j)$, we get: $\frac{1}{T} \sum_{i=1}^{T} \sum_{r=1}^{p} \mathbb{1}\{A(S_j^i)(v_r) \neq f_j(v_r)\}$

$$\leq \frac{1}{T} \sum_{i=1}^{T} L_{D_j}(A(S_j^i))$$

$$= \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\{A(S_j^i)(v_r) \neq f_j(v_r)\}$$

station #2 → $\geq \frac{1}{2} \min_{1 \leq r \leq p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\{A(S_j^i)(v_r) \neq f_j(v_r)\}$

Fix $r \in \{1, \dots, p\}$. The classifier f_j can be constructed in the following way:

- label all elements in $C \setminus \{v_r\}$
- Assign either or $\mathbb{1}$ to v_r

Let us call $\tilde{f}_j^{(0)}$ the first possible classifier and $\tilde{f}_j^{(1)}$ the second one.

Note that $\forall c \in C \setminus \{v_r\}$ we have $\tilde{f}_j^{(0)}(c) = \tilde{f}_j^{(1)}(c) = f_j(c)$

Also $\{f_1, \dots, f_T\} = \{\tilde{f}_1^{(0)}, \dots, \tilde{f}_{\frac{T}{2}}^{(0)}, \tilde{f}_1^{(1)}, \dots, \tilde{f}_{\frac{T}{2}}^{(1)}\}$

$\{(x_1^j, \tilde{f}_j^{(0)}(x_1^{j,i})), \dots, (x_m^j, \tilde{f}_j^{(0)}(x_m^{j,i}))\} = \{(x_1^j, \tilde{f}_j^{(1)}(x_1^j)), \dots, (x_m^j, \tilde{f}_j^{(1)}(x_m^j))\}$

v_r not in the training set $\{x_i\}$ → number i

$= S_j$.

$-\frac{1}{4}$ if $x_1 = 0$, $-\frac{3}{4}$ if $x_1 = 1$ — fixed r!

$$\sum_{i=1}^{T} \mathbb{1}\{A(S_i^j)(v_r) \neq f_j(v_r)\} = \sum_{l=1}^{T/2} \mathbb{1}\{A(S_j^l)(v_r) \neq \underbrace{\tilde{f}_l^{(0)}(v_r)}_{=0}\} + \sum_{l=1}^{T/2} \mathbb{1}\{A(S_j^l)(v_r) \neq \underbrace{\hat{f}_l^{(1)}(v_r)}_{=1}\}$$

$$= \sum_{l=1}^{T/2} \left[\mathbb{1}\{A(S_j^l)(v_r)=0\} + \mathbb{1}\{A(S_j^l)(v_r)=1\}\right]$$

$= \frac{T}{2}$ station #3 ← always $\frac{T}{2}$.

It follows that $\max_{1\le i\le T} E_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_j}(A(S))\right] \ge \frac{1}{2} \min_{1\le j\le k} \min_{1\le i\le p} \frac{1}{T}\cdot\frac{T}{2} = \frac{1}{4}$! □

2. Error decomposition

Let $h_S \in \operatorname{argmin}_{h\in H} L_S(h)$ where H is some hypothesis class

$$L_{\mathcal{D}}(h_S) = \min_{h\in H} L_{\mathcal{D}}(h) + L_{\mathcal{D}}(h_S) - \min_{h\in H} L_{\mathcal{D}}(h)$$

$\min_{h\in H} L_{\mathcal{D}}(h) = \varepsilon_{app}$ = approximation error

$L_{\mathcal{D}}(h) - \min_{h\in H} L_{\mathcal{D}}(h) = \varepsilon_{est}$ = estimation error

Approximation error:
- not connected to any probabilistic statement
- if H is large then we expect that this error is small (could be zero in the realizability case).

Estimation error: Results from replacing the unknown distribution by a training set.
If H is big then we need more examples in the training set.

(eg. if H is finite we know that $m_H(\varepsilon,\delta)$ of order $\frac{2}{\varepsilon^2}\log\left(\frac{2|H|}{\delta}\right)$)

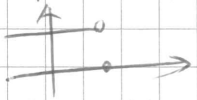
This called the bias-complexity trade-off.

The VC-dimension (Chapter 6)

Question: Can we find infinite hypothesis classes which PAC learnable?

1. Introductory example

Consider the class of thresholds over $\mathbb{R}$: $H = \{h_a : a\in\mathbb{R}\}$

and $h_a(x) = \mathbb{1}_{\{x<a\}}$ 

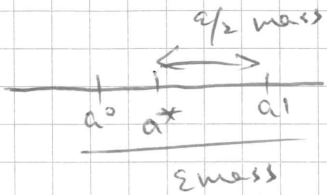
Lemma H is PAC learnable with an ERM rule. The sample complexity is $m_H(\varepsilon,\delta) \le \frac{2}{\varepsilon}\log\left(\frac{2}{\delta}\right)$ for $(\varepsilon,\delta)\in(0,1)^2$

Proof: Let a^* be such that h_{a^*} is a perfect classifier, that is

$L_{\mathcal{D}}(h^*) \overset{\text{def}}{=} \mathbb{P}_{(x,y)\sim\mathcal{D}}(h_{a^*}(x) \neq y) = 0$

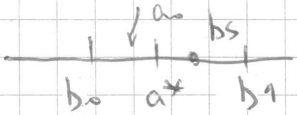
Let $\mathcal{D}_x$ be the marginal distribution of x, assumed to be continuous

consider a_0 and a_1 $\mathbb{P}_{x\sim\mathcal{D}^x}(x \in (a_0, a^*]) = \frac{\varepsilon}{2}$ and $\mathbb{P}_{x\sim\mathcal{D}^x}(x \in (a^*, a_1]) = \frac{\varepsilon}{2}$



Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be some training set with iid examples.

Write $b_0 = \max\{x_i : (x_i, 1) \in S\}$ and $b_1 = \min\{x_i : (x_i, 0) \in S\}$



(we assume that b_0, b_1 are well-defined in $\mathbb{R}$) Note that $b_0 = b_0(S)$ and $b_1 = b_1(S)$

Let b_S be the threshold of an ERM rule $h_S \in \operatorname{argmin}_{a\in\mathbb{R}} L_S(h_a)$

Recall that $L_S(h_S) = 0$ with probability 1 (see chapter 2).

$\Rightarrow \mathbb{1}_{x<b_S} h_S(x_i) = y_i$ for $i \in \{1, \ldots, m\}$

$$\Rightarrow \begin{cases} x_i < b_S \iff y_i = 1 \Rightarrow b_0 < b_S \\ x_i \geq b_S \iff y_i = 0 \Rightarrow b_1 \geq b_S \end{cases} \iff b_0 < b_S \leq b_1$$

We will show that $\{S : b_0(S) \geq a_0 \text{ and } b_1(S) \leq a_1\} \subset \{S : L_{\mathcal{D}}(b_S) \leq \varepsilon\}$

Assume that the event on the left

$L_{\mathcal{D}}(h_S) = \mathbb{P}_{x\sim\mathcal{D}^x}(\mathbb{1}_{x<b_S} \neq \mathbb{1}_{x<a^*})$

$= \mathbb{P}_{x\sim\mathcal{D}^x}(x < b_S \text{ and } x \geq a^*) + \mathbb{P}_{x\sim\mathcal{D}^x}(x \geq b_S \text{ and } x < a^*)$

$\leq \mathbb{P}(a^* \leq x < b_1) + \mathbb{P}(b_0 < x < a^*)$ (on the above event)

$\leq \underbrace{\mathbb{P}(a^* \leq x < a_1)}_{=\frac{\varepsilon}{2}} + \underbrace{\mathbb{P}(a_0 < x < a^*)}_{=\frac{\varepsilon}{2}}$ (by continuity of $\mathcal{D}^*$)

This implies that $\mathbb{P}_{S\sim\mathcal{D}^m}(L_S(h_S) > \varepsilon) \leq \mathbb{P}_{S\sim\mathcal{D}^m}(b_0 < a_0 \text{ or } b_1 > a_1)$

$\leq \mathbb{P}_{S\sim\mathcal{D}^m}(b_0 < a_0) + \mathbb{P}_{S\sim\mathcal{D}^m}(b_1 > a_1)$

$\mathbb{P}_{S\sim\mathcal{D}^m}(b_0 < a_0) = \mathbb{P}_{S\sim\mathcal{D}^m}(\forall x_i \text{ if } x_i < a^* \text{ then } x_i < a_0)$

$= \mathbb{P}_{S\sim\mathcal{D}^m}(\forall i,\ x_i \notin [a_0, a^*))$

Corollary: Let $\mathcal{H}$ be some hypothesis class. Let m be the size of a training set. Assume $\exists\, C \subset X$ such that $|C| = 2m$ and the size $\geq |C|$ is shattered by $\mathcal{H}$. Then, for any learning algorithm A, we can of $\mathcal{H}$. find a distribution $\mathcal{D}$ on $X \times \{0,1\}$ such that

1. $\exists$ classifier f: $L_{\mathcal{D}}(f) = 0$
2. $\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$

Definition: (VC dimension)

The VC dimension of some class $\mathcal{H}$, denoted by $VC\dim(\mathcal{H})$, is the maximal size of a set $C \subset X$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of any size, then we say that $\mathcal{H}$ has an infinite VC-dimension ($VC\dim(\mathcal{H}) = +\infty$)

$\mathcal{H}$ 가 $|C'| = |C|$ 인 모든 C' 을 shatter 하면 $\mathcal{H}$ 가 C 를 shatter 한다.

Theorem Let $\mathcal{H}$ be some hypothesis class with $VC\dim(\mathcal{H}) = +\infty$. Then $\mathcal{H}$ is not PAC learnable.

proof: If $\mathcal{H}$ is such that $VC(\dim(\mathcal{H})) = +\infty$, then for any $m \geq 1$, we can find a set $C \subset X$ such that $|C| = 2m$ and C is shattered by $\mathcal{H}$. Then, the claim follows from the corollary. □

Remarks:

- To show that $VC\dim(\mathcal{H}) \leq d$, it is enough to show that $\mathcal{H}$ does not shatter any set C of size $d+1$.
- To show that $VC\dim(\mathcal{H}) \geq d$, it is enough to find a set C of size d such that C is shattered by $\mathcal{H}$.
- To show that $VC\dim(\mathcal{H}) = d$, then we need to show

$$\begin{cases} \exists C : |C| = d \text{ such that } \mathcal{H} \text{ shatters } C \\ \forall C : |C| = d+1,\ \mathcal{H} \text{ does not shatter } C \end{cases}$$

* If $\mathcal{H}$ shatters C, then it shatters any subset $C' \subset C$.
* If X is finite, then for any $C \subset X$ we have $|\mathcal{H}_C| \leq |\mathcal{H}|$ if $|C|$ such that $|\mathcal{H}| < 2^{|C|}$ ($\Leftrightarrow |C| > \log_2(|\mathcal{H}|)$) then $|\mathcal{H}_C| < 2^{|C|} \Rightarrow \mathcal{H}$ does not shatter C.

Since this true for any $C : |C| > \log_2(|\mathcal{H}|)$

$= \left[\mathbb{P}(x_1 \notin [a_0, a^*])\right]^m \leq \left(1 - \mathbb{P}(x_1 \in [a_0, a^*])\right)^m$

$= \left(1 - \frac{\varepsilon}{2}\right)^m$

Similarly, we can show that $\mathbb{P}_{s\sim\mathcal{D}^m}(b_1 > a_1) = \mathbb{P}(\forall i, x_i \notin [a^*, a_1])$

$= \left[\mathbb{P}(x_1 \notin [a^*, a_1])\right]^m$

$\leq \mathbb{P}(s \text{ not close to } a^*)$ $= \left(1 - \frac{\varepsilon}{2}\right)^m$

Therefore $\mathbb{P}_{s\sim\mathcal{D}^m}(L_{\mathcal{D}}(h_s) > \varepsilon) \leq 2\left(1 - \frac{\varepsilon}{2}\right)^m$

$\leq 2\, e^{-m\frac{\varepsilon}{2}} \leq \delta$

$\Leftrightarrow e^{-\frac{m\varepsilon}{2}} \leq \frac{\delta}{2} \Leftrightarrow \frac{m\varepsilon}{2} \geq \log\left(\frac{2}{\delta}\right)$

$\Leftrightarrow m \geq \frac{2\log\left(\frac{2}{\delta}\right)}{\varepsilon}$

$\Rightarrow m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{2\log\left(\frac{2}{\delta}\right)}{\varepsilon}$

as claimed □

2. The VC dimension

Definition: (restriction of $\mathcal{H}$ to C)

Let $\mathcal{H}$ be a hypothesis class of function from X to $\{0,1\}$. Let $C = \{c_1, \dots, c_s\} \subset X$. The restriction of $\mathcal{H}$ to C is $\mathcal{H}_C = \{(h(c_1), \dots, h(c_s))\}$

Definition: (shattering)

A hypothesis class $\mathcal{H}$ **shatters** a finite set $C \subset X$ if $|\mathcal{H}_C| = 2^{|C|}$

In other words, $\mathcal{H}$ shatters C when $\mathcal{H}_C$ is the set of all possible classifiers that can be defined on C.

Example: ① Consider again the class $\mathcal{H} = \{x \mapsto \mathbb{1}_{x<a} : a \in \mathbb{R}\}$

$C = \{c\}$ with c a fixed real number. There are 2^1 classifiers on C: $f_1(c) = 0$ and $f_2(c) = 1$

0 is obtained by $c - 1 = a$. 1 is obtained by $a = c + 1$

② Take $C = \{c_1, c_2\}$ with $c_1 < c_2$. There are $2^2 = 4$ possible classifiers on C.

if $a \leq c_1$, then $h_a(c_1) = h_a(c_2) = 0$

if $c_1 < a \leq c_2$, then $h_a(c_1) = 1$ and $h_a(c_2) = 0$.

if $c_2 > a$, then $h_a(c_1) = h_a(c_2) = 1$

Hence $|\mathcal{H}_C| = 3 < 4 \Rightarrow \mathcal{H}$ does not shatter $C = \{c_1, c_2\}$ (this is only true for $c_1 < c_2$)

this implies that $VC\dim(H) \leq \log_2(|H|)$

what if $|H| = 2^{|X|}$? In this case, set $VC\dim(H) = |X|$. This means that H does not shatter any subset C of size $|X|+1$ (this is true since such a subset does not exist). Note that $\log_2(|H|) = \log_2(2^{|X|}) = |X|$.

3. Examples:

3.1 The class of thresholds:

$$H = \{x \longmapsto 1_{x<a} : a \in \mathbb{R}\}$$

we have shown that $\begin{cases} H \text{ shatters } C=\{c\} \text{ (for } c\in\mathbb{R}) \\ H \text{ does shatter any } C=\{c_1,c_2\} : c_1 \neq c_2 \end{cases}$

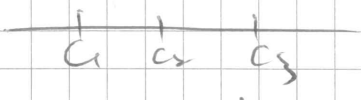
This implies that $VC\dim(H) = 1$

3.2 Class of intervals $h_{a,b}(x)$

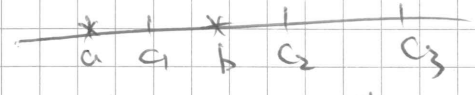
Consider $H = \{x \longmapsto 1_{x\in(a,b)} : -\infty < a \leq b < +\infty\}$

we can show that H shatters $C=\{0,1\}$ (show this!)

Now, consider $C=\{c_1,c_2,c_3\}$ with $c_1<c_2<c_3$



If $a \leq b$ are chosen such that $h_{a,b}(c_1) = 1$ and $h_{a,b}(c_2) = 0$ then we must have $h_{a,b}(c_3) = 0$



This implied that the classifier f such that $f(c_1)=1$, $f(c_2)=0$ and $f(c_3)=1$ is impossible to obtain.

Any $C=\{c_1,c_2,c_3\}$ can not be shattered by H.

This implies that $VC\dim(H) = 2$.

Remark: 3.3 Finite classes.

We have seen that if X is finite (which means that H is finite $VC\dim(H) \leq \log_2(|H|)$

But $\log_2(|H|) - VC\dim(H)$ can be quite large.

$X = \{1,2,\dots,k\}$ $H = \{x \longmapsto 1_{x\leq i} : i \in \{1,\dots,k\}\}$

VC dim(H) = 1 but $\log_2(|H|) = \log_2(|h|)$ which $\to \infty$ if $k \to \infty$

4. The fundamental Theorem of PAC learning:

we have seen that VC dim(H) = +∞ => H is not PAC learnable.

Question: Do we have: VC dim(H) = d <∞ H is PAC learnable?

Theorem: Let H be some hypothesis class: $X \to \{0,1\}$ and let the loss function be the 0-1 loss. Then the following assertions are equivalent:

1. H has the Uniform Convergence property
2. Any ERM rule is a successful agnostic PAC learner for H
3. H is agnostic PAC learnable
4. H is PAC learnable
5. Any ERM rule is a successful PAC learnable for H
6. H is of finite VC dimension.

Proof: (partial)

1 => 2 (chapter 4)

2 => 3 => 4 (trivial)

2 => 5 (trivial)

4 => 6 and 5 => 6 (cosequence of the No-free lunch theorem)

If we show 6 => 1 which needs Sauer's Lemma. then it is proved.

5. Sauer's lemma and the growth of function

Question: when the size of C a growth, how does H_C grow?

$E^{P}[\xi \mid \mathcal{F}_{\tau_m}] = E^{P_0}\left[\xi\left(\int_0^t \sigma_s dB_s\right) \mid \mathcal{F}_{\tau_m}\right]$

$\in C_b$; $\quad = X^{\sigma}$

$A \in \mathcal{F}_t$ ~~$\Leftrightarrow$~~ $A \in \cap \mathcal{F}_{\tau_m}$ $\quad A \cap (\tau_m < t) \in \mathcal{F}_t$.

(Vc-dimension). [6.5]

H Vc dim(H) is the maximal size of a set C such that H shatters C, $|H_C| = 2^{|C|}$

$|c|=m$ 时最多能 shatter 的数量.

(Growth function) $\tau_H(m) = \max_{C \subset X, |C|=m} |H_C|$

(Sauer lemma) H, Vc dim(H) $\le d < \infty \Rightarrow \tau_H(m) \le \sum_{i=0}^{d} \binom{m}{i}$ $(m \ge d)$.

if $m \le \dim(H)$, $\tau_H(m) = 2^m$

proof: $\forall H$, $|H_C| \le |\{B \subseteq C : H \text{ shatters } B\}|$ (*) (like 1_B).

$\le \sum_{i=0}^{d} \binom{m}{i}$

To show (*): $m=1$, $|H_C| = \begin{cases} 1 \\ 2 \end{cases}$ $\quad B = \begin{cases} \emptyset \\ \{c_1\} \end{cases}$ $\quad |\{B \subseteq C : H \text{ shatter } B\}| = 1, 2$.

general m, assume it holds for $k < m$

结定一个 C 看 c_1 是否对 shatter 有影响:

$C = \{c_1, c_2, \cdots, c_m\}$, $|H_C| = |Y_0| + |Y_1|$ → 再加上新加的 c_1 能被 shatter 的点.

$C' = \{c_2, \cdots, c_m\}$ 假设是新加的 c_1 才能被 shatter

only see one value

$Y_0 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in H_C \text{ or } (1, y_2, \cdots, y_m) \in H_C\}$

$Y_1 = \{$ —〃— and —〃— $\}$

(y_1 y_m 同 ✓)

$(Y_1 \subset Y_0)$

$Y_0 = H_{C'}$, $|Y_0| = |H_{C'}| \le |\{B \subseteq C' : H \text{ shatters } B\}|$

$= |\{B \subseteq C : c_1 \notin B \text{ and } H \text{ shatter } B\}|$

$H' \subseteq H$, $H' = \{h \in H, \exists h' \in H, (1-h'(c_1), h'(c_2), \dots, h'(c_m)) = (h(c_1), \dots, h(c_m))\}$

$|Y_1| = |H'_C| \le |\{B \subseteq C : H \text{ shatters } B\}|$

$= |\{B \subseteq C : H' \text{ shatters } B \cup \{c_1\}\}|$

$= |\{B \subseteq C : c_1 \in B \text{ and } H' \text{ shatters } B\}|$

$\le |\{B \subseteq C : c_1 \in B \text{ and } H \text{ shatters } B\}|$

$|H_C| = |Y_0| + |Y_1| \le |\{B \subseteq C : H \text{ shatter } B\}|$ □

Y_0: 相当于新加的点一起也被 shatter.

if $m > d+1$, $\tau_H(m) \le \left(\frac{em}{d}\right)^d$

proof: if $\frac{d}{m} \le 1 \Rightarrow \left(\frac{d}{m}\right)^d < 1$

$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \le \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i$

$\le \left(1 + \frac{d}{m}\right)^d \le \left(1+\frac{d}{m}\right)^m$

$\Rightarrow \sum_{i=0}^{d} \binom{m}{i} \le \left(1+\frac{d}{m}\right)^m \cdot \left(\frac{m}{d}\right)^d$

$\le \left(\frac{em}{d}\right)^d$. □

+ 2× slides!

The Fundamental Theorem of PAC learning

Recall the SGD algorithm runs as follows

- start: $w^{(1)} = 0$
- for $t \in \{1, \dots, T\}$ → generate v_t from a distribution s.t. $\mathbb{E}[v_t | w^{(t)}] \in \partial f(w^{(t)})$
  → update $w^{(t+1)} = w^{(t)} - \underbrace{\eta}_{\text{given step size}} v_t$
- Output: $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$.

Theorem (Convergence of SGD)

Let $B > 0$ and $\rho > 0$. Let f be a convex function with $w^* = \operatorname{argmin}_{w: \|w\| \le B} f(w)$. Assume the SGD algorithm is run for T iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume that v_t satisfies $\|v_t\| \le \rho$ with probability 1. $\forall t \in \{1, \dots, T\}$. Then $\mathbb{E}[f(\bar{w})] - f(w^*) \le \frac{B\rho}{\sqrt{T}}$

proof: skipped.

- **Learning with SGD for risk minimization**

Recall that the main goal is to minimize some risk function. Consider the risk function $L_D(w) = \mathbb{E}_{z \sim D}[\ell(w, z)]$

* Analyzing convergence of SGD for convex and ρ-Lipschitz loss functions:

Assume that $w \mapsto \ell(w, z)$ is convex and ρ-Lipschitz. Assume that it is differentiable ($\forall z \in Z$).

Then by the dominated convergence theorem, the function $w \mapsto L_D(w)$ is differentiable if $\exists$ function $z \mapsto k(z)$ s.t. $k \ge 0$ and $\mathbb{E}_{z \sim D}[k(z)] < +\infty$, and $\|\nabla \ell(w, z)\| \le k(z)$ $\forall w, z$.

Furthermore, we have that $\nabla L_D(w) = \mathbb{E}_{z \sim D}[\nabla \ell(w, z)]$.

Note that if $w \mapsto \ell(w, z)$ is ρ-Lipschitz and differentiable

then $\|\nabla l(w,z)\| \le \rho \ \forall w, z \Rightarrow$ we can take $k(z) = \rho,\ \forall z$.

Define $v_t = \nabla l(w^{(t)}, z)$ where $z \sim \mathcal{D} \perp\!\!\!\perp w^{(t)}$. We know that $\|v_t\| \le \rho$. Let us compute $\mathbb{E}[v_t | w^{(t)}] = \underset{z\sim\mathcal{D}}{\mathbb{E}} [\nabla l(w^{(t)}, z) | w^{(t)}]$

$$= \underset{z\sim\mathcal{D}}{\mathbb{E}} [\nabla l(w^{(t)}, z)]$$

$$= \nabla \underset{z\sim\mathcal{D}}{\mathbb{E}} [l(w^{(t)}, z)]$$

$$= \nabla L_{\mathcal{D}}(w^{(t)})$$ (the only element in $\partial L_{\mathcal{D}}(w^{(t)})$).

Define $v_t \in \partial l(w^{(t)}, z)$ with $z \sim \mathcal{D} \perp\!\!\!\perp w^{(t)}$. We have again $\|v_t\| \le \rho$. (I shall applying properties of subgradient for ρ-Lipschitz function.)

fixed $\forall u$ $\quad l(u,z) \ge l(w^{(t)}, z) + \langle u - w^{(t)}, v_t \rangle$

$$\Rightarrow \underset{z\sim\mathcal{D}}{\mathbb{E}}[l(u,z) | w^{(t)}] \ge \underset{z\sim\mathcal{D}}{\mathbb{E}}[l(w^{(t)}, z) | w^{(t)}] + \mathbb{E}[\langle u - w^{(t)}, v_t \rangle | w^{(t)}]$$

$$\Leftrightarrow \underset{z\sim\mathcal{D}}{\mathbb{E}}[l(u,z)] \ge \underset{z\sim\mathcal{D}}{\mathbb{E}}[l(w^{(t)}, z)] + \langle u - w^{(t)}, \mathbb{E}[v_t | w^{(t)}] \rangle$$

$$\Leftrightarrow L_{\mathcal{D}}(u) \ge L_{\mathcal{D}}(w^{(t)}) + \langle u - w^{(t)}, \mathbb{E}[v_t | w^{(t)}] \rangle.$$

□

The previous inequality holds $\forall u$ and hence $\mathbb{E}[v_t | w^{(t)}] \in \partial L_{\mathcal{D}}(w^{(t)})$. Therefore we can give the SGD algorithm more precisely by the following steps.

* Start: $w^{(t)} = 0$
* For $t = 1, \dots, T$ → generate $z \sim \mathcal{D} \perp\!\!\!\perp w^{(t)}$ ($z^{0} \Rightarrow w^{1}$, z_λ) → pick $v_t \in \partial l(w^{(t)}, z_\lambda)$.
* update $w^{(t+1)} = w^{(t)} + \eta v_t$
* Output $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$

Corollary: Consider a convex - Lipschitz - bounded learning problem with parameters $\rho > 0$ and $B > 0$ ($\mathcal{H} = \{w \in \mathbb{R}^d : \|w\| \le B\}$ and $w \mapsto l(w,z)$ is convex ρ-Lipschitz.)

$\forall \varepsilon > 0$, if we run the SGD algorithm for minimizing $w \mapsto L_{\mathcal{D}}(w)$ (on H) with $T \geq \frac{B^2 \rho^2}{\varepsilon^2}$ and $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then

$$\mathbb{E}[L(\bar{w})] \leq \min_{w \in H} L_{\mathcal{D}}(w) + \varepsilon .$$

- Analyzing convergence of SGD for convex–smooth loss functions.

Theorem Assume that $\forall z$, $w \mapsto \ell(w, z)$ is convex and β-smooth for some $\beta > 0$. Then if we run the SGD algorithm, we have $\forall w^*$ and if $\eta < \frac{1}{\beta}$, then

$$\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \frac{1}{1-\eta\beta}\left(L_{\mathcal{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T}\right)$$

proof: Recall that if f is a β-smooth and non-negative function, then $\|\nabla f(w)\|^2 \leq 2\beta f(w)$. For $t \in \{1, \dots, T\}$, let us denote $z_t = z \sim \mathcal{D}$ at iteration t s.t. $z_t \perp w^{(t)}$, $f_t(w) = \ell(w, z_t)$.

Then, $v_t = \nabla f_t(w^{(t)})$. By convexity of f_t, we have that $f_t(w^{(t)})$ (at $w^{(t)}$)

$$\leq f_t(w^*) + \langle w^{(t)} - w^*, v_t \rangle$$

summing over $t = 1, \dots, T$ yields $\sum_{t=1}^{T} f_t(w^{(t)}) \leq \sum_{t=1}^{T} f_t(w^*) + \sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle$

$$\leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

using Lemma 14.1 ← slide # 29 (week # 11)

Using self-boundness of f_t, it follows that $\sum_{t=1}^{T}\|v_t\|^2 \leq 2\beta \sum_{t=1}^{T} f_t(w^{(t)})$

Hence, $\frac{1}{T}\sum_{t=1}^{T} f_t(w_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(w^*) + \frac{\|w^*\|^2}{2\eta T} + \eta\beta \frac{1}{T}\sum_{t=1}^{T} f_t(w^{(t)})$.

$$\Leftrightarrow (1-\eta\beta)\frac{1}{T}\sum_{t=1}^{T} f_t(w^{(t)}) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(w^*) + \frac{\|w^*\|^2}{2\eta T}$$

$$\Leftrightarrow \frac{1}{T}\sum_{t=1}^{T} f_t(w^{(t)}) \leq \frac{1}{1-\eta\beta}\left(\frac{1}{T}\sum_{t=1}^{T} f_t(w^*) + \frac{\|w^*\|^2}{2\eta T}\right)$$

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f_t(w^{(t)})] \leq \frac{1}{1-\eta\beta}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f_t(w^*)] + \frac{\|w^*\|^2}{2\eta T}\right).$$

Now, $\mathbb{E}[f_t(w^*)] = \mathbb{E}_{z_t \sim \mathcal{D}}[\ell(w^*, z_t)] = L_{\mathcal{D}}(w^*)$ $\forall t$.

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f_t(w^*)] = L_{\mathcal{D}}(w^*)$$

$$\mathbb{E}[f_t(w^{(t)})] = \mathbb{E}[\ell(w^{(t)}, z^{(t)})] = \mathbb{E}[L_{\mathcal{D}}(w^{(t)})]$$

Hence, $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f_t(w^{(t)})] = \frac{1}{T}\sum_{t=1}^{T}[L_{\mathcal{D}}(w^{(t)})]$. □

Now, $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[L_{\mathcal{D}}(w^{(t)})] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(w^{(t)})\right]$

the function $w \mapsto L_{\mathcal{D}}(w) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(w,z)]$ is convex (by convexity of $w \mapsto \ell(w,z)$ for $\forall z$).

Hence, $\frac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(w^{(t)}) \geq L_{\mathcal{D}}\left(\frac{1}{T}\sum_{t=1}^{T} w^{(t)}\right)$ (Jensen's inequality).

we conclude that $\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \frac{1}{1-\eta\beta}\left(L_{\mathcal{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T}\right)$ as claimed.

Corollary: Consider a convex-smooth-bounded learning problem with parameters $\beta>0$ and $B>0$. Assume in addition that $\ell(0,z) \leq 1$ $\forall z\in\mathcal{Z}$

$\forall \varepsilon\in(0,1]$, set $\eta = \frac{1}{\beta(1+\frac{2}{\varepsilon})}$. Then running SGD with $T \geq \frac{12B^2\beta}{\varepsilon^2}$

yields $\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \min_{w\in H} L_{\mathcal{D}}(w) + \varepsilon$ $(H = \{w\in\mathbb{R}^d : \|w\|\leq B\})$

Proof: $1-\eta\beta = \frac{3}{\varepsilon+3} \Leftrightarrow \frac{1}{1-\eta\beta} = 1+\frac{\varepsilon}{3}$

By the previous theorem, we have that

$\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq (1+\frac{\varepsilon}{3})\left(L_{\mathcal{D}}(w^*) + \frac{B^2}{2\eta T}\right)$, where $w^* = \operatorname{argmin}_{w\in H} L_{\mathcal{D}}(w)$

$\leq (1+\frac{\varepsilon}{3}) L_{\mathcal{D}}(w^*) + (1+\frac{\varepsilon}{3})\frac{\varepsilon^2+3\varepsilon}{24}$

$L_{\mathcal{D}}(w^*) = L_{\mathcal{D}}(0) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(0,z)] \leq 1$

$\leq L_{\mathcal{D}}(w^*) + \frac{\varepsilon}{3} + (1+\frac{1}{\varepsilon})\frac{4}{24}\varepsilon$

$\leq L_{\mathcal{D}}(w^*) + \varepsilon.$ □

Regularization and stability (chapter 13)

1 Regularized loss minimization (RLM)

RLM is a learning rule which minimize the criterion $w \mapsto L_S(w) + R(w)$ for a given regularization function R.

Example: If $R(w) = \lambda\|w\|^2 = \lambda\sum_{i=1}^{d} w_i^2$, then we talk about when applying Tikhonov regularization to linear regression with $\ell_{sq}(w) = (\langle w, x\rangle - y)^2$, we obtain the learning rate.

$$\hat{w} = \operatorname{argmin}_w \left(\lambda\|w\|^2 + \frac{1}{2m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2\right)$$

Put $f(w) = \lambda\|w\|^2 + \frac{1}{2m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2$ for $w\in\mathbb{R}^d$

$$= \lambda w^T w + \frac{1}{2m}\sum_{i=1}^{m}(x_i^T w - y_i)^2$$

$$\nabla f(w) = 2\lambda w + \frac{1}{m}\sum_{i=1}^{m} x_i(x_i^T w - y_i) = 0.$$

$$\Leftrightarrow \underbrace{\left[\sum_{i=1}^{m} x_i x_i^T + \lambda m I\right]}_{A} w = \sum_{i=1}^{m} y_i x_i$$

The matrix A is semi-positive definite because for $\forall a \in \mathbb{R}^d$,

$$a^T A a = a^T \sum_{i=1}^{m} x_i x_i^T a = \sum_{i=1}^{m} (x_i^T a)^2 \geq 0$$

$\Rightarrow \exists$ orthogonal matrix $\Omega \in \mathbb{R}^{d \times d}$ such that

$$A = \Omega^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} \Omega \text{ with } \lambda_i \geq 0, \ \forall i \in \{1, \dots, d\}.$$

$$\Rightarrow A + \lambda m I = \Omega^T \begin{pmatrix} \lambda_1 + \lambda m & & \\ & \ddots & \\ & & \lambda_d + \lambda m \end{pmatrix} \Omega$$

Since $\lambda_i + \lambda m > 0 \ \forall i \in \{1, \dots, d\}$, $A + \lambda m I$ is invertible and $\hat{w} = \left[A + \lambda m I\right]^{-1} \cdot \left(\sum_{i=1}^{m} y_i x_i\right)$

2. Stable rules do not overfit.

Let A be a learning algorithm and $S = \{z_1, \dots, z_m\}$ be some training set with m iid example $\sim \mathcal{D}$ (unknown distribution)

Denote $A(S)$ the output of A when it is run on S.

To assess stability of A, we look at the influence of replacing an example z_i by some $z' \sim \mathcal{D}$.

Given S and an additional example $z' \sim \mathcal{D} \perp\!\!\!\perp S$, let $S^{(i)} = \{z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m\}$.

We investigate $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)$

($\leftarrow$ does not observe z_i; $\hookrightarrow$ observe z_i)

Theorem Let $\mathcal{D}$ be some distribution, $S = \{z_1, \dots, z_m\}$ a training set where $z_1, \dots, z_m$ are iid $\sim \mathcal{D}$ and $z' \sim \mathcal{D} \perp\!\!\!\perp S$.

Then, for any learning algorithm we have that

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] = \mathbb{E}_{(S, z', i) \sim \mathcal{D}^{m+1} \times U([m])}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right] \quad (*)$$

where $[m] = \{1, 2, \dots, m\}$.

Proof: $\mathbb{E}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(A(S))\right] = \mathbb{E}_{S \sim \mathcal{D}^m}\left[\mathbb{E}_{z' \sim \mathcal{D}}\left(\ell(A(S), z')\right)\right]$ ($z' \perp\!\!\!\perp S$)

$$= \mathbb{E}_{(S,z')\sim D^{m+1}}\left[\ell(A(S)^{(i)}, z')\right]$$

$$\{z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_m\}$$

$$= \mathbb{E}_{(S,z')\sim D^{m+1}}\left[\ell(A(\{z_1,\dots,z',\dots z_m\}), z_i)\right]$$

$$= \mathbb{E}_{(S,z')\sim D^{m+1}}\left[\ell(A(S^{(i)}), z_i)\right].$$

Since the last expectation does not depend on the index i, it follows that

$$\mathbb{E}_{(S,z')\sim D^{m+1}}[\quad z_i] = \mathbb{E}_{(S,z',i)\sim D^{m+1}\times U[m]}[\quad z_i].$$

On the other hand, we have that

$$\mathbb{E}_{S\sim D^m}[L_S(A(S))] = \mathbb{E}_{S\sim D^m}\left[\frac{1}{m}\sum_{i=1}^m \ell(A(S), z_i)\right] \text{ by definition of}$$

$$= \mathbb{E}_{(S,i)\sim D^m\times U[m]}[\ell(A(S), z_i)] \text{ the training error/rate}$$

same proof as before. □.

When the right term of (*) is small, this is a indication that replacing z_i by "something else" does not influence the algorithm "too much". In other words, we say that the algorithm is stable (a change in a single example does not result in a big change on average.)

Definition (on-average-replace-one-stable)

Let $\varepsilon: \mathbb{N} \to \mathbb{R}$ be a decreasing function, we say that some learning algorithm A is on-average-replace-one-stable with rate $\varepsilon(m)$ if $\forall$ distribution D $\mathbb{E}_{(S,z',i)\sim D^{m+1}\times U[m]}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \varepsilon(m)$

Remark: In view of the previous theorem, when this property holds then $\forall$ distribution D

$$\mathbb{E}_{S\sim D^m}[L_D(A(S)) - L_S(A(S))] \leq \varepsilon(m)$$ ← means you are not overfitting

this means that A does not overfit $\rightarrow$ A "good" learning algorithm should balance between fitting and being stable.

3. Tikhonov regularization as a stabilizer:

In the following, we are going to apply Tikhonov regularization to convex and p-Lipschitz loss function.

Definition: (strongly convex functions) A function f is said to be λ-strongly convex if $\forall u, w$ and $\alpha \in [0,1]$

$$f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2.$$

Lemma 1. The function $f(w) = \lambda\|w\|^2$ is (2λ)-strongly convex.

2. If f is λ-strongly convex and g is convex, then $f+g$ is λ-strongly convex

3. If f is λ-strongly convex and u is the minimizer of f, then $f(w) - f(u) \geq \frac{\lambda}{2}\|w-u\|^2$.

proof: 1 and 2 are easy to show.

3. Let u be the minimizer of f. For $\alpha \in (0,1]$.

$$f(u + \alpha(w-u)) = f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2$$

$$\Leftrightarrow \frac{f(u+\alpha(w-u)) - f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2$$

Since $f(u) = \min_v f(v)$, then $f(u+\alpha(w-u)) \geq f(u)$, and hence

$$0 \leq f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2$$

$$\Leftrightarrow f(w) - f(u) \geq \frac{\lambda}{2}(1-\alpha)\|w-u\|^2 \quad \forall \alpha \in (0,1]$$

Taking the limit as $\alpha \downarrow 0$, we get that $f(w) - f(u) \geq \frac{\lambda}{2}\|w-u\|^2$ □

For a training set $S = \{z_1, \ldots, z_m\}$, consider the RLM Tikhonov $A(S) = \operatorname{argmin}_w (L_S(w) + \lambda\|w\|^2)$. Let us write $f_S(w) = L_S(w) + \lambda\|w\|^2$. Under convexity of $w \mapsto \ell(w, z)$ $\forall z \in Z$ and using the previous lemma, we see that f_S is (2λ)-strongly convex. Also, by the (① + ②) some lemma (③), we have $\forall v$,

$$f_S(v) - f_S(A(S)) \geq \lambda\|A(S) - v\|^2 \quad (*).$$

for any u, v and index $i \in \{1, \dots, n\}$

$$f_S(v) - f_S(u) = L_S(v) + \lambda\|v\|^2 - L_S(u) - \lambda\|u\|^2 = L_{S^{(i)}}(v) - \frac{1}{m}\ell(v, z') + \frac{1}{m}\ell(v, z_i) + \lambda\|v\|^2 - \left(L_{S^{(i)}}(v) - \frac{1}{m}\ell(v, z') + \frac{1}{m}\ell(u, z_i)\right)$$

$$= L_{S^{(i)}}(v) + \lambda\|v\|^2 - \left(L_{S^{(i)}}(u) + \lambda\|u\|^2\right) + \frac{\ell(v, z_i) - \ell(v, z')}{m} + \lambda\|u\|^2 \Big) - \frac{\ell(u, z_i) - \ell(u, z')}{m}$$

Now take $u = A(S)$ and $v = A(S^{(i)})$ since v minimizes $w \mapsto L_{S^{(i)}}(w) + \lambda\|w\|^2$ it follows that.

$$f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}$$

$\geq \lambda\|A(S^{(i)}) - A(S)\|^2$

If the loss function is also ρ-Lipschitz (in addition to convexity), then

$$\text{Term} \leq |\text{Term}| \leq \frac{2\rho}{m}\|A(S^{(i)}) - A(S)\|.$$

Therefore we get $\lambda\|A(S^{(i)}) - A(S)\|^2 \leq \frac{2\rho}{m}\|A(S^{(i)}) - A(S)\|$

$$\Leftrightarrow \quad \|A(S^{(i)}) - A(S)\| \leq \frac{2\rho}{m\lambda}$$

ρ-Lipschitz.

This also implies that $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \left| \quad \right| \leq \frac{2\rho^2}{\lambda m}$.

$\forall S, i, (\forall z' \in \mathcal{D})$. □

Corollary: If the loss function is convex and ρ-Lipschitz, then the Tikhonov RLM rule is on-average-replace-one-stable with rate $\varepsilon(m) = \frac{2\rho^2}{\lambda m}$.

we have $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}$.

4. Controlling the fitting-stability trade-off (for convex and ρ-Lipschitz loss functions).

Let us start with $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = \mathbb{E}_{S\sim\mathcal{D}^m}[L_S(A(S))] + \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))]$

fit. stability.

Let $A(S) = \operatorname{argmin}_w\left(L_S(w) + \lambda\|w\|^2\right)$. we have $\forall w^*$, $L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(w^*) + \lambda\|w^*\|^2$.

$\mathbb{E}_{S\sim\mathcal{D}^m}[L_S(w^*)] = L_{\mathcal{D}}(w^*)$. Hence $\mathbb{E}_{S\sim\mathcal{D}^m}[L_S(A(S))] \leq L_{\mathcal{D}}(w^*) + \lambda\|w^*\|^2$.

$$\Rightarrow \quad \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(w^*) + \lambda\|w^*\|^2 + \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))]$$

Corollary: If the loss function and ρ-Lipschitz, then the Tikhonov RLM rule satisfies

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S))\right] \le L_{\mathcal{D}}(w^*) + \lambda\|w^*\|^2 + \frac{2\rho^2}{\lambda m}.$$

we will take $w^* = \operatorname{argmin}_{w:\|w\|\le B} L_{\mathcal{D}}(w)$

Corollary: Let (H, Z, ℓ) be a convex, Lipschitz bounded learning problem with parameters ρ and B. Also, set $\lambda = \sqrt{\frac{2\rho}{B^2 m}}$. Then, then Tikhonov RLM rule satisfies

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S))\right] \le \min_{w\in H} L_{\mathcal{D}}(w) + \rho B\sqrt{\frac{8}{m}} \quad (H = \{w\in\mathbb{R}^d : \|w\|\le B\}).$$

$L_{\mathcal{D}}(w^*)$.