Lectures 10, 11 Supplementary

24-Feb-10

Rademacher Complexity based learning bounds

Let $T = \sup_{f \in \mathcal{F}} \{ R[f] - R_{emp}^m[f] \}$

$R_{emp}^m[f] \rightarrow \frac{1}{m} \sum_{i=1}^{m} l(f(x_i), y_i)$

Let us denote $z_i = (x_i, y_i)$ & $h_f(z_i) = l(f(x_i), y_i)$

$z_i \rightarrow$ training examples (together with labels)

It is easy to see that T is a function of $z_1, z_2, \ldots, z_m$, which are iid! Let $T = g(z_1, z_2, \ldots, z_m) = \sup_{f \in \mathcal{F}} \{ R[f] - R_{emp}^m[f] \}$

Now let us also assume that $a \leq l(f(x), y) \leq b$

i.e. assume loss function is bounded. $\forall f \in \mathcal{F}$ & $\forall (x, y)$

Then it is easy to see that T (which is a function of iid rvs) satisfies the following property: (known as Bounded difference property)

$$\max_{z_i, z_i'} \left| g(z_1, z_2, \ldots, z_i, z_{i+1}, \ldots z_m) - g(z_1, z_2, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m) \right| \leq \frac{b-a}{m} \quad \forall i$$

Here's why: $\left| g(z_1, \ldots, z_i, \ldots, z_m) - g(z_1, \ldots, z_i', \ldots, z_m) \right|$

$= \left| \left(R[f^*] - R_{emp}^m[f^*] \right) - \left(R[f^*] - \frac{1}{m} \sum_{j \neq i} h_{f^*}(z_j) - \frac{1}{m} h_{f^*}(z_i') \right) \right|$

note this

where $f^* = \arg\sup_{f \in \mathcal{F}} \{ R[f] - R_{emp}^m[f] \}$

$$= \frac{1}{m}\left(h_{f^*}(z_i') - h_{f^*}(z_i)\right) \leq \frac{b-a}{m}$$ (since loss function is bounded between a & b)

Similarly

$$g(z_1, z_2, \ldots, z_i', \ldots, z_m) - g(z_1, \ldots, z_i, \ldots, z_m)$$

$$\leq \left(R[f^*] - \frac{1}{m}\sum_{j\neq i} h_{f^*}(z_j) - \frac{1}{m}h_{f^*}(z_i')\right) - \left(R[f^*] - R_{emp}[f^*]\right)$$

$$= \frac{1}{m}\left(h_{f^*}(z_i) - h_{f^*}(z_i')\right) \leq \frac{b-a}{m}$$ here, $f^* = \underset{f\in\mathcal{F}}{\text{argmax}}$

$\therefore |g(z_1, \ldots, z_i, z_{i+1} \ldots z_m) - g(z_1, \ldots, z_i', \ldots, z_m)| \leq \frac{b-a}{m}$ $\forall z_i, z_i'$

Hence T is a function of m iid rvs satisfies the bounded difference property.

$\longrightarrow$

We saw that Chernoff bounding gives efficient bounds in case of rvs which are in turn sum of independent rvs.

Using similar ideas one can handle generic functions of independent rvs (which satisfy properties like bounded difference) using the McDiarmid's inequality:

McDiarmid's Inequality: Let g be a function of m independent (need not be iid) rvs $z_1, \ldots, z_m$ satisfying Bounded diff property with bound $c\left(= \frac{b-a}{m}\right)$ (in our case)

then, $P[g - E[g] > \varepsilon] \leq e^{\frac{-2\varepsilon^2}{mc^2}}$

(Ill by we can also show that $P[E[g] - g > \varepsilon] \leq e^{-2\varepsilon^2/mc^2}$)

Note that McDiarmid's inequality is an extension of Chernoff's bounding technique from the case of bounding sum of independent rvs to handling generic functions of independent rvs satisfying mild conditions like the bounded difference property.

Proof: $T = g(Z_1, \ldots, Z_m)$ & g satisfies bounded diff. property.

Recall that Chernoff bounds apply to case of sum of independent rvs
So the idea here is to rewrite T as sum of independent rvs so that we can hopefully apply Chernoff techniques.

In view of this, define $V_i = E[T/Z_1, \ldots, Z_i] - E[T/Z_1, \ldots, Z_{i-1}]$

$$V_1 = E[T/Z_1] - E[T] \xrightarrow{\text{fn. of}} Z_1$$

$$V_2 = E[T/Z_1, Z_2] - E[T/Z_1] \xrightarrow{\text{fn. of}} Z_1, Z_2$$

$$V_i = E[T/Z_1, \ldots Z_i] - E[T/Z_1, \ldots, Z_{i-1}] \xrightarrow{\text{fn. of}} Z_1, Z_2, \ldots Z_i$$

$$V_m = E[T/Z_1, \ldots, Z_m] - E[T/Z_1, \ldots, Z_{m-1}] \xrightarrow{\text{fn of}} Z_1, Z_2, \ldots, Z_m$$

also known as Martingale difference

$$\sum_{i=1}^{m} V_i = E[T/Z_1, \ldots Z_m] - E[T]$$

$$= T - E[T]$$

Moreover, V_i, V_j are conditionally independent:

$V_1 \longrightarrow$ fc. of Z_1

$V_2/Z_1 \longrightarrow$ fc. of Z_2

$\vdots$

$V_i/Z_1 Z_2 \cdots Z_{i-1} \longrightarrow$ fc. of Z_i

$\vdots$

$V_m/Z_1 \cdots Z_{m-1} \longrightarrow$ fc. of Z_m

} all are independent! (since Z_i are ind.)

Now the idea is to apply Chernoff bounding technique to $\sum V_i$ noting the conditional independence.

$$P[g - E[g] > \varepsilon] = P\left[\sum V_i > \varepsilon\right] = P\left[e^{s\sum V_i} > e^{s\varepsilon}\right] \quad (s>0$$

$$\leq \frac{E\left[e^{s\sum V_i}\right]}{e^{s\varepsilon}} \quad (\because \text{Markov inequality})$$

$$= \frac{E\left[\prod_{i=1}^{m} e^{sV_i}\right]}{e^{s\varepsilon}}$$

Now we exploit cond. independence to bound mgf:

$$E\left[\prod_{i=1}^{m} e^{sV_i}\right] = E\left[E\left[\prod_{i=1}^{m} e^{sV_i}/Z_1 \cdots Z_{m-1}\right]\right]$$

(total expectation rule)

$$= E\left[E\left[\left(\prod_{i=1}^{m-1} e^{sV_i}\right) e^{sV_m}/Z_1 \cdots Z_{m-1}\right]\right]$$

$$= E\left[\prod_{i=1}^{m-1} e^{sV_i} E\left[e^{sV_m}/Z_1 \cdots Z_{m-1}\right]\right] \quad (\because \text{conditional independence})$$

Now, we will upper bound

$$E\left[e^{sV_m} / Z_1, \cdots, Z_{m-1}\right]$$

In general, we can upper bound:

$$E\left[e^{sV_i} / Z_1, \cdots Z_{i-1}\right] \qquad \forall\ i = 2 \text{ to } m$$

The idea is to note that $V_i / Z_1, \cdots, Z_{i-1}$ is a mean zero r.v & is bounded (with interval lengths c again) and then apply Hoeffding inequality:

Now,

$$E\left[V_i / Z_1, \cdots, Z_{i-1}\right] = E\left[\left(E\left[T / Z_1, \cdots, Z_i\right] - E\left[T / Z_1 \cdots Z_{i-1}\right]\right) / Z_1 \cdots Z_{i-1}\right]$$

$$= E\left[E\left[T / Z_1, \cdots, Z_i\right] / Z_1, \cdots, Z_{i-1}\right] - E\left[E\left[T / Z_1 \cdots Z_{i-1}\right] / Z_1 \cdots Z_{i-1}\right]$$

(↓ total prob. rule; ↓)

$$= E\left[T / Z_1, \cdots, Z_{i-1}\right] - E\left[T / Z_1, \cdots, Z_{i-1}\right]$$

$$\leq 0$$

Also, we know T satisfies bounded difference

$$\max_{Z_i, Z_i'} \left| f(Z_1 \cdots Z_i \cdots Z_m) - g(Z_1, \cdots, Z_i', \cdots Z_m) \right| \leq c$$

$$\max_{Z_i, Z_i'} \left| E\left[T / Z_1 \cdots Z_i\right] - E\left[T / Z_1, \cdots, Z_i'\right] \right| \leq c$$

$$\Rightarrow \max_{Z_i, Z_i'} \left| E\left[T / Z_1 \cdots Z_i\right] - E\left[T / Z_1, \cdots Z_{i-1}\right] - \left(E\left[T / Z_1 \cdots Z_i'\right] - E\left[T / Z_1, \cdots Z_{i-1}\right]\right) \right| \leq c$$

(add-subtract the same term)

$$\Rightarrow \max_{Z_i, Z_i'} \left| \underbrace{E\left[T / Z_1 \cdots Z_i\right] - E\left[T / Z_1, \cdots Z_{i-1}\right]}_{V_i / Z_1 \cdots Z_{i-1}} - \left(\underbrace{E\left[T / Z_1 \cdots Z_i'\right] - E\left[T / Z_1, \cdots Z_{i-1}\right]}_{V_i' / Z_1 \cdots Z_{i-1}}\right) \right| \leq c$$

$$\Rightarrow \max_{z_i, z_i'} \left| V_i/_{z_1 \dots z_{i-1}} - V_i'/_{z_1 \dots z_{i-1}} \right| \le c$$

$$\Rightarrow \max_{z_i} V_i/_{z_1 \dots z_{i-1}} - \min_{z_i'} V_i'/_{z_1 \dots z_{i-1}} \le c$$

Hence $V_i/_{z_1 \dots z_{i-1}}$ ~~is~~ has mean zero and moves in interval of length $\le c$.

$$\therefore E\left[e^{sV_i}/_{z_1 \dots z_{i-1}}\right] \le e^{s^2 c^2/8}$$

$$\Rightarrow E\left[\prod_{i=1}^{m} e^{sV_i}\right] = E\left[\prod_{i=1}^{m-1} e^{sV_i} \, E\left[e^{sV_m}/_{z_1 \dots z_{m-1}}\right]\right]$$

$$\le e^{s^2 c^2/8} E\left[\prod_{i=1}^{m-1} e^{sV_i}\right]$$

$$\le e^{s^2 c^2 m/8}$$ (repeated application of conditional independence)

$$\Rightarrow P\left[g - E[g] > \varepsilon\right] \le \frac{e^{s^2 c^2 m/8}}{e^{s\varepsilon}} \quad (\forall s > 0)$$

$$\therefore P\left[g - E[g] > \varepsilon\right] \le e^{\min_s \frac{s^2 c^2 m}{8} - s\varepsilon} = e^{-2\varepsilon^2/mc^2}$$

Hence McDiarmid's Inequality is proved.

Proof for $P\left[E[g] - g > \varepsilon\right] \le e^{-2\varepsilon^2/mc^2}$ is also similar and is skipped here.

In fact, conl... seem by ||| h... thing happening for changes in bou...

→ Getting back to our problem we have $T = \sup_{f \in \mathcal{F}} \left\{ R[f] - R_{emp}^m[f] \right\}$ and it satisfies bounded diff. prob with $c = \frac{b-a}{m}$ ∴ By McDiarmid's inequality we have:

$$P\left[\sup_{f \in \mathcal{F}} \left\{R[f] - R_{emp}^m[f]\right\} - E\left[\sup_{f \in \mathcal{F}} \left\{R[f] - R_{emp}^m[f]\right\}\right] > \varepsilon\right] \le e^{-\frac{2m\varepsilon^2}{(b-a)^2}} = \delta \text{ (say)}$$

Now re-write ε in terms of δ: $\varepsilon = (b-a)\sqrt{\frac{\ln 1/\delta}{2m}}$

$\Rightarrow \forall f \in \mathcal{F}$, with at least prob. $1-\delta$, we have:

$$R[f] \leq \hat{R}^m_{emp}[f] + E\left[\sup_{f\in\mathcal{F}} \{R[f] - \hat{R}^m_{emp}[f]\}\right] + (b-a)\sqrt{\frac{\ln 1/\delta}{2m}}$$

(New learning bound)

~~key differences with VC type bounds:~~

~~Confidence term~~

difficult to compute

have we will further upper bound this term by a term which reflects capacity of $\mathcal{F}$ (& resembles VC-dim bound)

confidence term (free of complexity terms like VC dim. etc.)

So the task now is to bound the $E[T]$.

$$E[T] = E\left[\sup_{f\in\mathcal{F}} \{R[f] - \hat{R}^m_{emp}[f]\}\right]$$

define $\bar{R}^m_{emp}[f] = \frac{1}{m}\sum_{i=1}^{m} h_f(Z_i')$

Z_i' are also iid copies of Z_i.

$$= E\left[\sup_{f\in\mathcal{F}} \left\{ E\left[\bar{R}^m_{emp}[f]\right] - \hat{R}^m_{emp}[f]\right\}\right]$$

$$= E\left[\sup_{f\in\mathcal{F}} \{ E[\bar{R}^m_{emp}[f] - \hat{R}^m_{emp}[f]\}\right]$$

$$\leq E\left[E\left[\sup_{f\in\mathcal{F}} \{\bar{R}^m_{emp}[f] - \hat{R}^m_{emp}[f]\}\right]\right] \quad (\because \text{Jensen's inequality})$$

(Symmetrization trick)

E wrt Z_i; E wrt Z_i'

Now we will again resort to ~~the~~ Rademacher variables introducing

$$E[T] \leq E\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_i\left(h_f(z_i') - h_f(z_i)\right)\right\}\right]$$

wrt z_i, z_i'

$$= E\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_i \sigma_i\left(h_f(z_i') - h_f(z_i)\right)\right\}\right]$$

wrt z_i, z_i', σ_i ⇒ Rademacher r.vs.

$$\leq E\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_i \sigma_i h_f(z_i')\right\}\right] + E\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_i -\sigma_i h_f(z_i)\right\}\right]$$

wrt z_i', σ_i ; wrt z_i, σ_i

$$= 2E\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_i \sigma_i h_f(z_i)\right\}\right]$$

Now given a loss function, & set of leading functions $\mathcal{F}$,
we get an induced set of loss functions $\rightarrow \{h : l(f(x), y) = h(z) \;\forall f \in \mathcal{F}\}$

↓ let us call this set as $\mathcal{L}$

$$\Rightarrow E[T] \leq 2E\left[\sup_{h\in\mathcal{L}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i)\right\}\right]$$

wrt z_i & σ_i

known as Rademacher Complexity
of
Rademacher average of function class $\mathcal{L}$

denoted by $\mathcal{R}(\mathcal{L})$.

$\therefore$ We have: $R[f] \leq R_{emp}^{m}[f] + 2\mathcal{R}(\mathcal{L}) + (b-a)\sqrt{\frac{\ln 1/\delta}{2m}}$

↓ purely a property of $\mathcal{L}$ & underlying distribution.

Now $\mathcal{R}(\mathcal{L})$ is also difficult to compute as underlying distributio
of $z_1, \ldots z_m$ is unknown.

So we further upper bound it by quantity which is easily computable:

$$\mathcal{R}_m(\mathcal{L}) = \mathbb{E}\left[\sup_{h \in \mathcal{L}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i) \right| \Big/ z_1 \ldots z_m \right]$$

wrt σ_i alone

→ nothing but conditional Rademacher Average

$\mathcal{R}_m(\mathcal{L})$ is easy to compute (atleast numerically) since the distribution of σ_i is known. (estimate)

average computed for given training dataset $z_1 \ldots z_m$

Now we need to bound $\mathcal{R}(\mathcal{L})$ by terms involving $\mathcal{R}_m(\mathcal{L})$

For this we note that

(i) $E[\mathcal{R}_m(\mathcal{L})] = \mathcal{R}(\mathcal{L})$ ($\because$ total prob. rule)

(ii) $\mathcal{R}_m(\mathcal{L})$ is a function of $z_1, \ldots, z_m$ (which are i.i.d) and satisfies bounded difference property.

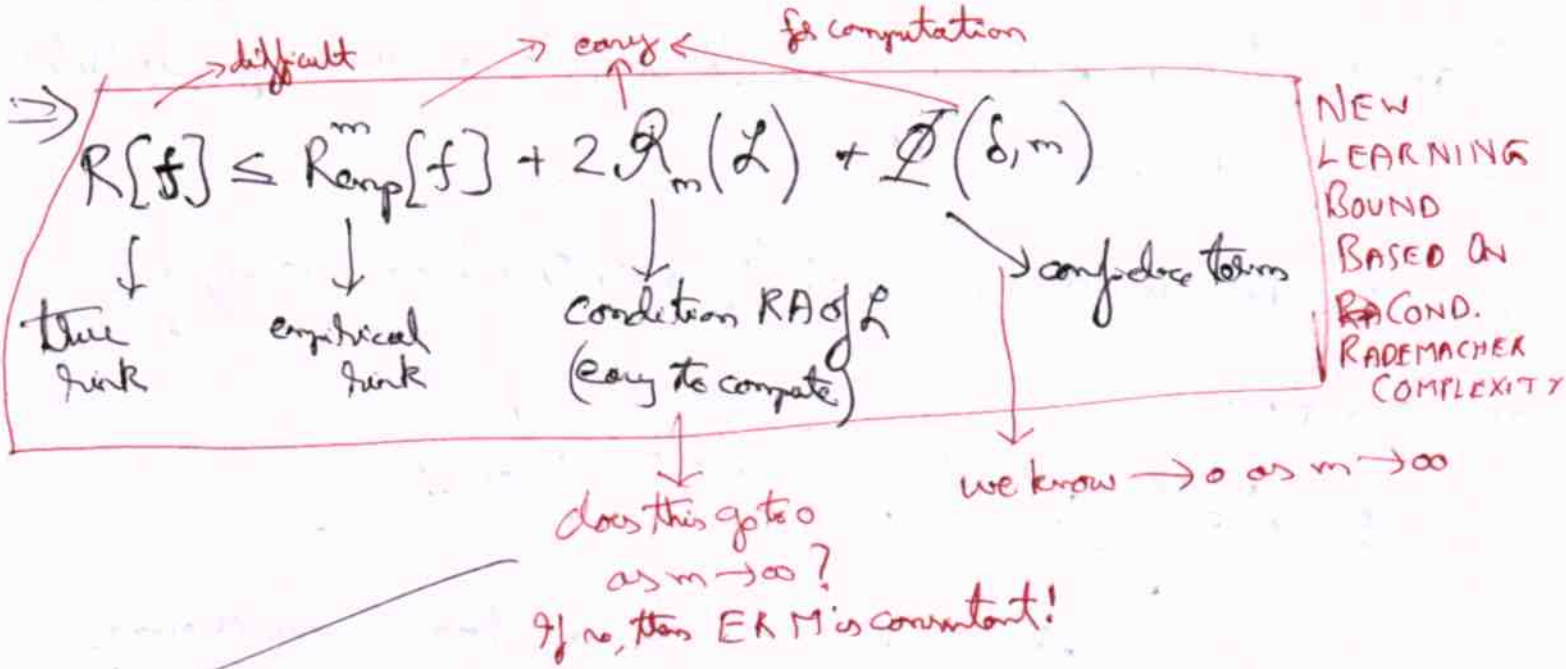
Student must show this

$\therefore$ We can use McDiarmid's inequality again:

$$P\left[\mathcal{R}(\mathcal{L}) - \mathcal{R}_m(\mathcal{L}) > \varepsilon \right] \leq e^{\frac{-2\varepsilon^2 m}{(b-a)^2}} = \delta \text{ (say)}$$

$\mathcal{R}(\mathcal{L})$ → $E[\mathcal{R}_m(\mathcal{L})]$

$$\Rightarrow \mathcal{R}(\mathcal{L}) \leq \mathcal{R}_m(\mathcal{L}) + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}$$



↳ We will now take specific case of binary classification - i.e. $\mathcal{F}$ as set of all functions taking values $\{-1, 1\}$ and let loss function be 0-1 loss.

$$\Rightarrow l(f(x), y) = \frac{1}{2}(1 - yf(x)) = \begin{cases} 1 & \text{if } y \neq f(x) \\ 0 & \text{if } y = f(x) \end{cases}$$

$$\mathcal{R}_m(\mathcal{L}) = E\left[\sup_{h \in \mathcal{L}} \left\{ \frac{1}{m} \sum_i \sigma_i h(z_i) \right\} / z_{1,2} \right]$$

$$= E\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{m} \sum_i \sigma_i \frac{1}{2}(1 - y_i f(x_i)) \right\} / z_{1,2} \right]$$

$$= \frac{1}{2} E\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{m} \sum_i -\sigma_i y_i f(x_i) \right\} / z_{1 \ldots 2m} \right] \quad \text{(II)}$$

(y_i is assigned)

$$= \frac{1}{2} E\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{m} \sum_i \sigma_i f(x_i) \right\} / x_1 \ldots x_m \right]$$

$$= \frac{1}{2} \mathcal{R}_m(\mathcal{F}) \to \text{easily computed given training data points alone.}$$

∴ For binary classification with 0-1 loss we have learning bound:

$$R[f] \leq R_{emp}^{m}[f] + \mathcal{R}_m(\mathcal{F}) + \mathcal{O}(\delta, m)$$

↓

Cond. RA of set of classifiers.

In fact one can show that:

$$\mathcal{R}_m(\mathcal{F}) \leq A\sqrt{\frac{h}{m}}$$ (A: some constant; h: VC dim. of $\mathcal{F}$!)

↳ This recovers our good old bounds depending on VC dim. and leaves extra term of $\left(\log\frac{m}{h}+1\right)$. is removed. (which is very good)

→ Intuitively also (II) says why $\mathcal{R}_m(\mathcal{F})$ is a measure of "complexity" of set of classifiers.

If $\mathcal{F}$ were rich to catch $\sigma_i y_i$ then we would have $\mathcal{R}_m(\mathcal{F}) = \frac{1}{2}$ always $\forall m$

which $\not\to 0$ as $m \to \infty$ ∴ ERM is not consistent in this case

& is expected. (like to VC dim $= \infty$)

→ Now ~~lets compute $\mathcal{R}_m$~~ SRM principle will suggest minimizing Empirical risk as well as $\mathcal{R}_m(\mathcal{F})$ minimize roughly.

Now lets get a bound on $\mathcal{R}_m(\mathcal{F})$ for hyperplane classifiers in feature space induced by a kernel k.

we want nontrivial monitor (previously we got $h \leq \|w\|^2 R^2$)

Cond. R.A. for Canonical hyperplane classifiers in feature space (RKHS)

Suppose we have $\mathcal{F} = \{ f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq W \}$ → RKHS induced by given kernel 'k'

$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$

we wish to find $\mathcal{R}_m(\mathcal{F})$. & atleast upper bound it.

$$\mathcal{R}_m(\mathcal{F}) = E_{\sigma_i}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{m} \sum_i \sigma_i f(x_i) \right\} \right]$$

$f(x_i) = \langle f, k(x_i, \cdot) \rangle_{\mathcal{H}}$

$$= \frac{1}{m} E_{\sigma_i}\left[\sup_{f \in \mathcal{F}} \left\{ \left\langle f, \sum_i \sigma_i k(x_i, \cdot) \right\rangle_{\mathcal{H}} \right\} \right] \quad \left(\because \text{linearity of } \langle \cdot, \cdot \rangle_{\mathcal{H}} \right)$$

$\leq \|f\|_{\mathcal{H}} \left\| \sum_i \sigma_i k(x_i, \cdot) \right\|_{\mathcal{H}}$ by Cauchy-Schwartz inequality.

$$\leq \frac{1}{m} E_{\sigma_i}\left[\sup_{f \in \mathcal{F}} \left\{ \|f\|_{\mathcal{H}} \sqrt{\sum_i \sum_j \sigma_i \sigma_j k(x_i, x_j)} \right\} \right]$$

$$\leq \frac{W}{m} E_{\sigma_i}\left[\sqrt{\sum_i \sum_j \sigma_i \sigma_j k(x_i, x_j)} \right]$$

$$\leq \frac{W}{m} \sqrt{\sum_i \sum_j E[\sigma_i \sigma_j] k(x_i, x_j)} \quad (\because \text{Jensen's inequality})$$

$E[\sigma_i \sigma_j] = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$

$$= \frac{W}{m} \sqrt{\text{trace}(K)}$$ → Gram matrix with kernel 'k' of training data points.