Regression with Squared Loss

(X, Y) in $\mathcal{X} \times \mathbb{R}$

Min. mean Square Error (MMSE) prediction:

$$\min_{f: \mathcal{X} \to \mathbb{R}} \underbrace{\mathbb{E}[(Y - f(X))^2]}_{:= L(f)}$$

- optimal predictor if P_{XY} known:

$$f^*(x) = \mathbb{E}[Y \mid X = x]$$

$$L(f) = L(f^*) + \underbrace{\mathbb{E}[(f(X) - f^*(X))^2]}_{\geq 0}$$

- Learning setting: P_{XY} unknown

$(X_1, Y_1), \dots, (X_n, Y_n)$ iid

fix $\mathcal{F}$ (class of fcns $f: \mathcal{X} \to \mathbb{R}$)

$\hat{f}_n$, based on data, in $\mathcal{F}$

Goal: $L(\hat{f}_n) - \inf_{f \in \mathcal{F}} L(f)$ small w.h.p.

Set-up: $\mathcal{X}$ arbitrary

$$\mathcal{P} = \{ P_{XY} : \mathbb{P}[|Y| \leq M] = 1 \}$$

$\mathcal{F}$: subset of RKHS $\mathcal{H}_K$

choice of K is a degree of freedom

e.g. $\mathcal{X} = \mathbb{R}^d$, $K(x, x') = 1 + \langle x, x' \rangle$

$\mathcal{H}_K$ consists of $f(x) = \langle w, x \rangle + b$

(1) Norm-Constrained Predictors

$$\mathcal{F} = \mathcal{F}_\lambda = \{ f \in \mathcal{H}_K : \|f\|_K \le \lambda \}$$

$0 < \lambda < \infty$: fixed parameter

ERM: $$\hat{f}_n = \underset{f \in \mathcal{F}_\lambda}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2}_{L_n(f)}$$

$$L(\hat{f}_n) - L^*(\mathcal{F}_\lambda)$$

$$= L(\hat{f}_n) - \inf_{f \in \mathcal{F}_\lambda} L(f)$$

$$= L(\hat{f}_n) - L(f^*) \qquad f^* = \underset{f \in \mathcal{F}_\lambda}{\operatorname{argmin}} L(f)$$

$$\le 2 \sup_{f \in \mathcal{F}_\lambda} | L_n(f) - L(f) |$$

Notation:

$$\ell(y, u) := (y - u)^2$$

$$\ell \bullet f(x, y) := \ell(y, f(x))$$

$$= (y - f(x))^2$$

$$\ell \bullet \mathcal{F}_\lambda := \{ \ell \bullet f : f \in \mathcal{F}_\lambda \}$$

losses predictors

$$\sup_{f \in \mathcal{F}_\lambda} |L_n(f) - L(f)| = \sup_{f \in \mathcal{F}} |P_n(\ell \circ f) - P(\ell \circ f)|$$

$$= \Delta_n(\underbrace{\ell \circ \mathcal{F}_\lambda}_{\text{induced losses}})$$

where $P_n(h) := \frac{1}{n} \sum_{i=1}^{n} h(z_i)$ $\quad z_i = (x_i, y_i)$

$P(h) := \mathbb{E}(h(z))$

$h = \ell \circ f$ for some $f \in \mathcal{F}_\lambda$

$$g(z^n) = \Delta_n(\ell \circ \mathcal{F}_\lambda)$$

$$= \sup_{f \in \mathcal{F}_\lambda} \left| \frac{1}{n} \sum_{i=1}^{n} \ell \circ f(z_i) - \mathbb{E}[\ell \circ f(z)] \right|$$

Claim: $g(\cdot)$ has bdd diffs (McDiarmid)

$z_1, \dots, z_i, \dots, z_n$ $\quad z_i = (x_i, y_i) \in \mathcal{X} \times [-M, M]$

$z_1, \dots, z_i', \dots, z_n$ $\quad z_i' = (x_i', y_i') \in \mathcal{X} \times [-M, M]$

$$g(z_1, \dots, z_i, \dots, z_n) - g(z_1, \dots, z_i', \dots, z_n)$$

$$\leq \frac{1}{n} \sup_{f \in \mathcal{F}_\lambda} \left| (y_i - f(x_i))^2 - (y_i' - f(x_i'))^2 \right|$$

$$\leq \frac{1}{n} \sup_{x \in \mathcal{X}} \sup_{|y| \leq M} \sup_{f \in \mathcal{F}_\lambda} (y - f(x))^2$$

$$\leq \frac{2}{n} \left(M^2 + \sup_{f \in \mathcal{F}_\lambda} \sup_{x \in \mathcal{X}} |f(x)|^2 \right)$$

$(a+b)^2 \leq 2a^2 + 2b^2$

$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ Mercer kernel

$$C_K := \sup_{x \in \mathcal{X}} \sqrt{K(x,x)} < \infty$$

$f \in \mathcal{H}_K$: $\quad \|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$

Lemma $\quad \|f\|_\infty \leq C_K \|f\|_K$

Proof $\quad f \in \mathcal{H}_K$

$$\begin{aligned} |f(x)| &= |\langle f, K_x \rangle_K| && \text{(repr. prop.)} \\ &\leq \|f\|_K \|K_x\|_K && \text{(Cauchy-Schwarz)} \\ &= \|f\|_K \sqrt{K(x,x)} \\ &\leq C_K \|f\|_K . \end{aligned}$$

∎

$$\sup_{f \in \mathcal{F}_\lambda} \sup_{x \in \mathcal{X}} |f(x)|^2 = \sup_{f : \|f\|_K \leq \lambda} \|f\|_\infty^2 \leq C_K^2 \lambda^2$$

$\Rightarrow g(z^n) = \Delta_n(\ell \circ \mathcal{F}_\lambda)$ has bdd diffs with $c_1 = \dots = c_n = \frac{2}{n}(M^2 + C_K^2 \lambda^2)$

McDiarmid: $\quad \forall t > 0$

$$P\left\{ \Delta_n(\ell \circ \mathcal{F}_\lambda) \geq \mathbb{E}\Delta_n(\ell \circ \mathcal{F}_\lambda) + t \right\} \leq \exp\left(- \frac{n t^2}{2(M^2 + C_K^2 \lambda^2)^2} \right)$$

- $t = \sqrt{\frac{2}{n}(M^2 + C_K^2\lambda^2)^2 \log\left(\frac{1}{\delta}\right)}$

$$= (M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}}$$

$$\Delta_n(\ell \circ \mathcal{F}_\lambda) \leq \mathbb{E}\Delta_n(\ell \circ \mathcal{F}_\lambda) + (M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}}$$

w.p. $\geq 1 - \delta$

- Symmetrization:

$$\mathbb{E}\Delta_n(\ell \circ \mathcal{F}_\lambda) \leq 2\mathbb{E}R_n(\ell \circ \mathcal{F}_\lambda(z^n))$$

$$\ell \circ \mathcal{F}_\lambda(z^n) = \{(\ell \circ f(z_1), \ldots, \ell \circ f(z_n)) : f \in \mathcal{F}_\lambda\}$$

$$= \{((y_1 - f(x_1))^2, \ldots, (y_n - f(x_n))^2) : f \in \mathcal{F}_\lambda\}$$

<u>Claim:</u> for $f \in \mathcal{F}_\lambda$,

$$|y - f(x)| \leq M + C_K\lambda$$

for all $x \in \mathcal{X}$, $|y| \leq M$.

<u>Proof:</u>

$$\begin{aligned} |y - f(x)| &\leq |y| + |f(x)| \\ &\leq M + \|f\|_\infty \\ &\leq M + C_K\lambda. \quad \blacksquare \end{aligned}$$

- Apply contraction principle w/ $\varphi(t) = t^2$ on $[-(M + C_K\lambda), \underbrace{M + C_K\lambda}_{:=A}]$

$$((Y_1 - f(x_1))^2, \dots, (Y_n - f(x_n))^2)$$

$$= (\varphi(\underbrace{Y_1 - f(x_1)}_{\in [-(M + C_k\lambda), M + C_k\lambda]}), \dots, \varphi(Y_n - f(x_n)))$$

$$\Rightarrow R_n(\ell \circ \mathcal{F}_\lambda(z^n)) \le 2 \cdot \overset{\text{Lip-const. of } \varphi \text{ on } [-A,A]}{2A} \cdot \mathbb{E}_\varepsilon\left[\frac{1}{n} \sup_{f \in \mathcal{F}_\lambda} \left|\sum_{i=1}^n (Y_i - f(x_i))\varepsilon_i\right|\right]$$

$$\frac{1}{n} \mathbb{E}_\varepsilon\left[\sup_{f \in \mathcal{F}_\lambda} \left|\sum_{i=1}^n \varepsilon_i (Y_i - f(x_i))\right|\right]$$

$$\le \frac{1}{n} \mathbb{E}_\varepsilon\left[\left|\sum_{i=1}^n \varepsilon_i Y_i\right|\right] + \frac{1}{n} \mathbb{E}_\varepsilon\left[\sup_{f \in \mathcal{F}_\lambda} \left|\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right]$$

$$=: \frac{1}{n}(\mathrm{I} + \mathrm{II})$$

where:

I $\quad \mathbb{E}_\varepsilon \left|\sum_{i=1}^n \varepsilon_i Y_i\right| \qquad\qquad |Y_i| \le M$

$$= \mathbb{E}_\varepsilon \sqrt{\left(\sum_{i=1}^n \varepsilon_i Y_i\right)^2}$$

$$\le \sqrt{\mathbb{E}_\varepsilon \left(\sum_{i=1}^n \varepsilon_i Y_i\right)^2}$$

$$= \sqrt{\sum_{i=1}^n Y_i^2}$$

$$\le M\sqrt{n}$$

$$\text{II} \quad \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]$$

$$= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \left| \sum_{i=1}^n \varepsilon_i \langle f, K_{X_i} \rangle_K \right| \right]$$

$$= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \left| \langle f, \sum_{i=1}^n \varepsilon_i K_{X_i} \rangle_K \right| \right]$$

$$\leq \mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i K_{X_i} \right\|_K \cdot \sup_{f \in \mathcal{F}_\lambda} \|f\|_K$$

$$\leq \lambda \cdot \mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i K_{X_i} \right\|_K$$

$$= \lambda \cdot \mathbb{E}_\varepsilon \sqrt{\langle \sum_{i=1}^n \varepsilon_i K_{X_i}, \sum_{i=1}^n \varepsilon_i K_{X_i} \rangle_K}$$

$$\leq \lambda C_K \sqrt{n}$$

$$\Rightarrow R_n(\ell \circ \mathcal{F}_\lambda) \leq 4(M + C_K \lambda) \cdot \frac{1}{n}(\text{I} + \text{II})$$

$$\leq \frac{4(M + C_K \lambda)^2}{\sqrt{n}}$$

$$\mathbb{E} \Delta_n(\ell \circ \mathcal{F}_\lambda) \leq 2 \mathbb{E} R_n(\ell \circ \mathcal{F}_\lambda)$$

$$\leq \frac{8(M + C_K \lambda)^2}{\sqrt{n}}$$

$\therefore$ w.p. $\geq 1-\delta$,

$$L(\hat{f}_n) \leq L^*(\mathcal{F}_\lambda) + \frac{16\,(M + c_K \lambda)^2}{\sqrt{n}} + (M^2 + c_K^2 \lambda^2)\sqrt{\frac{8\log(1/\delta)}{n}}.$$

Comments:

- ERM over a ball $\mathcal{F}_\lambda$ in $\mathcal{H}_K$
- radius λ: hard constraint on hypothesis space complexity

(2) Norm-penalized predictors

$\mathcal{F} = \mathcal{H}_K$ (entire RKHS)

$$\hat{f}_n = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \left\{ L_n(f) + \gamma \|f\|_K^2 \right\}$$

($\gamma > 0$: tunable parameter)

$$J_\gamma(f) := L(f) + \gamma \|f\|_K^2$$

$$J_{n,\gamma}(f) := L_n(f) + \gamma \|f\|_K^2$$

$$= \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2 + \gamma \|f\|_K^2$$

$$L(f) \leq J_\gamma(f)$$

$$L_n(f) \leq J_{n,\gamma}(f)$$

Observation: suppose $f \in \mathcal{H}_K$ minimizes $J_\gamma(f)$ or $J_{n,\gamma}(f)$. Then $\|f\|_K \leq \frac{M}{\sqrt{\gamma}}$.

Corollary

$$\min_{f \in \mathcal{H}_K} J_\gamma(f) = \min_{f \in \mathcal{F}_{M/\sqrt{\gamma}}} J_\gamma(f)$$

$$\min_{f \in \mathcal{H}_K} J_{n,\gamma}(f) = \min_{f \in \mathcal{F}_{M/\sqrt{\gamma}}} J_{n,\gamma}(f)$$

Proof

Assume $f_\gamma^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} J_\gamma(f)$

$$J_\gamma(f_\gamma^*) \leq J_\gamma(0) = L(0) = \mathbb{E}|Y|^2 \leq M^2$$

$$J_\gamma(f) \geq \gamma \|f\|_K^2 \quad \forall f$$

$$\Rightarrow \quad \|f_\gamma^*\|_K^2 \leq \frac{M^2}{\gamma} \qquad \blacksquare$$

$$J_\gamma(\hat{f}_n) - J_\gamma(f_\gamma^*) \qquad \hat{f}_n, f_\gamma^* \in \mathcal{F}_{M/\sqrt{\gamma}}$$

$$\leq 2 \sup_{f \in \mathcal{F}_{M/\sqrt{\gamma}}} |J_{n,\gamma}(f) - J_\gamma(f)|$$

$$= 2 \sup_{f \in \mathcal{F}_{M/\sqrt{\gamma}}} |L_n(f) + \cancel{\gamma \|f\|_K^2} - L(f) - \cancel{\gamma \|f\|_K^2}|$$

$$= 2 \sup_{f \in \mathcal{F}_{M/\sqrt{\gamma}}} |L_n(f) - L(f)|$$

– already did this w/ $\lambda > 0$ arbitrary
– take $\lambda := M/\sqrt{\gamma}$

$\therefore$ w.p. $\geq 1-\delta$,

- $$J_\gamma(\hat{f}_n) \leq J_\gamma(f^*_\gamma) + \frac{16\left(M + c_K \frac{M}{\sqrt{\gamma}}\right)^2}{\sqrt{n}} + \left(M^2 + c_K^2 \frac{M^2}{\gamma}\right)\sqrt{\frac{8\log(1/\delta)}{n}}$$

- $$L(\hat{f}_n) \leq J_\gamma(\hat{f}_n) \leq J_\gamma(f^*_\gamma) + \cdots$$

$$J_\gamma(f^*_\gamma) = \min_{f \in \mathcal{H}_K} \{L(f) + \gamma \|f\|_K^2\} =: L^*(\mathcal{H}_K) + A(\gamma)$$

where $L^*(\mathcal{H}_K) = \inf_{f \in \mathcal{H}_K} L(f)$, so w.p. $\geq 1-\delta$:

$$L(\hat{f}_n) - \inf_{f \in \mathcal{H}_K} L(f) \leq A(\gamma) + \frac{16 M^2\left(1 + \frac{c_K}{\sqrt{\gamma}}\right)^2}{\sqrt{n}} + M^2\left(1 + \frac{c_K^2}{\gamma}\right)\sqrt{\frac{8\log(1/\delta)}{n}}$$

$\hat{f}_n = \hat{f}_{n,\gamma} = \operatorname{argmin}_{f \in \mathcal{H}_K} J_{n,\gamma}(f)$