Statistical foundations of learning (Summer 2025)

Saro Harutyunyan

May 2025

Assignment 1

Exercise 1.1. (a) According to Theorem Risk.2 (Bayes risk = Generalization error of Bayes classifier) from the lecture Statistical Learning Problem and Bayes Risk, the Bayes risk is

$$L^*_{x\sim\mathbf{p}} = \mathbb{E}_{x\sim\mathbf{p}}[\min\{\eta(x), 1-\eta(x)\}] = \alpha$$

since $\min\{\eta(x), 1-\eta(x)\} = \alpha$ for all x. The Bayes classifier is obviously

$$\begin{aligned} h^*(x) &= \begin{cases} +1, & \text{if } x \in \{11, \ldots, 20\}, \\ -1, & \text{else} \end{cases} \\ &= \mathbf{1}\{x \in (0, 11)\} - \mathbf{1}\{x \in (10, 21)\} + \mathbf{1}\{x \in (20, 31)\}. \end{aligned}$$

Hence $h^* \in \mathcal{H}_{int}$.

(b) We will use the notation $p_S = \sum_{x\in S\cap\mathcal{X}} p_x$ for any set S. Let $h \in \mathcal{H}_{ds}$.

Case 1. $h(x) = -1$ for $x < t$ and $h(x) = +1$ else (for some t). Then the risk is

$$\begin{aligned} L_{x\sim\mathbf{p}}(h) &= \mathbb{E}_{x\sim\mathbf{p}}[\mathbb{P}_{y|x}(h(x) \neq y)] \\ &= \mathbb{E}_{x\sim\mathbf{p}}[\mathbb{P}_{y|x}(h(x) \neq y) \cdot (\mathbf{1}\{t \in (0, 10]\} + \mathbf{1}\{t \in (10, 20]\} + \mathbf{1}\{t \in (20, 30]\})] \\ &\geq \begin{cases} p_{[1,t)}\alpha + p_{[t,10]}(1-\alpha) + q_2\alpha + q_3(1-\alpha), & \text{if } t \in (0, 10], \\ q_1\alpha + p_{[11,t)}(1-\alpha) + p_{[t,20]}\alpha + q_3(1-\alpha), & \text{if } t \in (10, 20], \\ q_1\alpha + q_2(1-\alpha) + p_{[21,t)}\alpha + p_{[t,30]}(1-\alpha), & \text{if } t \in (20, 30] \end{cases} \\ &\geq \begin{cases} q_1\alpha + q_2\alpha + q_3(1-\alpha), & \text{if } t \in (0, 10], \\ q_1\alpha + q_2\alpha + q_3(1-\alpha), & \text{if } t \in (10, 20], \\ q_1\alpha + q_2(1-\alpha) + q_3\alpha, & \text{if } t \in (20, 30]. \end{cases} \end{aligned}$$

Case 2. $h(x) = +1$ for $x < t$ and $h(x) = -1$ else (for some t). Then similar to the above steps we find that the risk is

$$L_{x\sim\mathbf{p}}(h) \geq \begin{cases} q_1\alpha + q_2(1-\alpha) + q_3\alpha, & \text{if } t \in (0, 10], \\ q_1(1-\alpha) + q_2\alpha + q_3\alpha, & \text{if } t \in (10, 20] \\ q_1(1-\alpha) + q_2\alpha + q_3\alpha, & \text{if } t \in (20, 30]. \end{cases}$$

Now aggregating our knowledge we find that the Bayes risk is at least the minimum of the six numbers in the RHS of cases 1 and 2.

Clearly this bound is achievable; we just need to take $h^* \in \mathcal{H}_{ds}$ so that it has the correct monotonicity (depending on the minimum in cases 1 and 2) and appropriately adjust the stump $t \in (10k, 10(k+1)] =: I_k$ to the respective end of interval I_k for $k \in \{0, 1, 2\}$.

Exercise 1.2. (a) We consider the model $h(x) = (x^T w)^2$ for $w \in \mathbb{R}^{|w^*|}$. Note that $\mathbb{E}[y|x] = \mathbb{E}[(x^T w^*)^2 + \epsilon|x] = (x^T w^*)^2$. Denote $\Sigma := \mathbb{E}_x[xx^T]$. Then, according to the bias-variance decomposition from the lecture Statistical Learning Problem and Bayes Risk, the risk on an estimator h is

$$L(h) = \mathbb{E}_{x,y}[(h(x) - y)^2] = \mathbb{E}_x[\mathbb{E}_{y|x}\Big((h(x) - \mathbb{E}[y \mid x]) - (\mathbb{E}[y \mid x] - y)\Big)^2].$$

We have

$$\begin{aligned}\mathbb{E}_{y|x}(\mathbb{E}[y \mid x] - y)^2 &= \mathbb{E}_{y|x}[\mathbb{E}[y \mid x]^2 - 2y\,\mathbb{E}[y \mid x] + y^2]\\ &= \mathbb{E}[y \mid x]^2 - 2\,\mathbb{E}[y \mid x]^2 + \mathbb{E}[y^2 \mid x] = \mathbb{E}[y^2 \mid x] - \mathbb{E}[y \mid x]^2,\end{aligned}$$

$$\mathbb{E}_{y|x}[(h(x) - \mathbb{E}[y \mid x])(\mathbb{E}[y \mid x] - y)] = (h(x) - \mathbb{E}[y \mid x])\,\mathbb{E}_{y|x}[\mathbb{E}[y \mid x] - y] = 0.$$

Thus

$$\begin{aligned}L(h) &= \mathbb{E}_{x,\epsilon}[(h(x) - \mathbb{E}[y|x])^2] + \mathbb{E}_{x,\epsilon}[\mathbb{E}[y^2|x] - (\mathbb{E}[y|x])^2]\\ &= \mathbb{E}_{x,\epsilon}[(x^T(w - w^*))^2] + \mathbb{E}_{x,\epsilon}[\mathbb{E}[y^2|x] - (x^T w^*)^4]\\ &= \|w - w^*\|_\Sigma^2 + \mathbb{E}_{x,\epsilon}[2\epsilon(x^T w^*)^2 + \epsilon^2]\\ &= \|w - w^*\|_\Sigma^2 + \sigma^2\end{aligned}$$

where we used the independence of ϵ and x. It is seen that the Bayes predictor is $h^*(x) = (x^T w^*)^2$ with the Bayes risk of σ^2.

(b) We need to prove that the OLS predictor $\widehat{h}(x) = x^T \widehat{w}$, where $\widehat{w} := \widehat{\Sigma}^{-1}\left(\frac{1}{m}\sum_{i=1}^m y_i x_i\right)$ with $\widehat{\Sigma} = \frac{1}{m}\sum_{i=1}^m x_i x_i^T$ (from lecture Risk for Linear Regression), is not consistent w.r.t. the distribution given in the problem. Its risk is

$$\begin{aligned}L(\widehat{h}) &= \mathbb{E}[(x^T\widehat{w} - (x^T w^*)^2 - \epsilon)^2]\\ &= \mathbb{E}[(x^T\widehat{w} - (x^T w^*)^2)^2] - 2\,\mathbb{E}[(x^T\widehat{w} - (x^T w^*)^2)\epsilon] + \mathbb{E}[\epsilon^2]\\ &= \mathbb{E}[(x^T\widehat{w} - (x^T w^*)^2)^2] + \sigma^2\end{aligned}$$

with $\mathbb{E} = \mathbb{E}_{x,\epsilon}$ since

$$\begin{aligned}&\mathbb{E}_{x,x_1,\dots,x_m,\epsilon,\epsilon_1,\dots,\epsilon_m}[(x^T\widehat{w} - (x^T w^*)^2)\epsilon]\\ &= \mathbb{E}_{x,x_1,\dots,x_m}\,\mathbb{E}_{\epsilon,\epsilon_1,\dots,\epsilon_m|x,x_1,\dots,x_m}[x^T\widehat{\Sigma}^{-1}\sum_{i=1}^m x_i((x_i^T w^*)^2 + \epsilon_i)\epsilon]\\ &= \mathbb{E}_{x,x_1,\dots,x_m}\,\mathbb{E}_{\epsilon,\epsilon_1,\dots,\epsilon_m|x,x_1,\dots,x_m}[x^T\widehat{\Sigma}^{-1}\sum_{i=1}^m x_i((x_i^T w^*)^2 + \epsilon_i)]\cdot\mathbb{E}_\epsilon[\epsilon] = 0\end{aligned}$$

(in the last equality we used the independence of ϵ, ϵ_i and x, x_i). It remains to prove that $\liminf\limits_{m\to\infty}\mathbb{E}[(x^T\widehat{w} - (x^T w^*)^2)^2] > 0$ where m is the number of samples:

$$\mathbb{E}_{x,\epsilon}[(x^T\widehat{w} - (x^T w^*)^2)^2] = \mathbb{E}_{x,\epsilon}[(x^T\widehat{w})^2] - 2\,\mathbb{E}_{x,\epsilon}[(x^T\widehat{w})(x^T w^*)^2] + \mathbb{E}_{x,\epsilon}[(x^T w^*)^4]. \tag{1}$$

We have

$$\mathbb{E}_{x,x_1,\ldots,x_m,\epsilon,\epsilon_1,\ldots,\epsilon_m}[(x^T w^*)^2 (x^T \widehat{w})]$$

$$\begin{aligned}
&= \mathbb{E}_{x,x_1,\ldots,x_m} \mathbb{E}_{\epsilon,\epsilon_1,\ldots,\epsilon_m|x,x_1,\ldots,x_m}[(x^T w^*)^2 x^T \widehat{\Sigma}^{-1} \sum_{i=1}^{m} ((x_j^T w^*)^2 + \epsilon_j)] \\
&= \mathbb{E}_{x,x_1,\ldots,x_m}[(x^T w^*)^2 x^T \widehat{\Sigma}^{-1} \sum_{i=1}^{m} (x_j^T w^*)^2] \\
&= \mathbb{E}_x[(x^T w^*)^2 x^T] \, \mathbb{E}_{x_1,\ldots,x_m}[\widehat{\Sigma}^{-1} \sum_{i=1}^{m} (x_j^T w^*)^2] = 0
\end{aligned}$$

since $\mathbb{E}_x[(x^T w^*)^2 x^T] = 0$. Indeed, denote by a^k the k-th coordinate of vector a. Then

$$\begin{aligned}
\mathbb{E}_x[(x^T w^*)^2 x]^k &= \mathbb{E}_x[(x^T w^*)^2 x^k] \\
&= \mathbb{E}_x[x^k((x^k w^{*k})^2 + 2x^k w^{k*} \sum_{i\neq k} x^i w^{*i} + \left(\sum_{i\neq k} x^i w^{*i}\right)^2] \\
&= (w^{*k})^2 \, \mathbb{E}_x[(x^k)^3] + 2\, \mathbb{E}_x (x^k)^2 (w^{*k}) \, \mathbb{E}_x \sum_{i\neq k} x^i w^{*i} + \mathbb{E}_x \, x^k \, \mathbb{E}_x \left(\sum_{i\neq k} x^i w^{*i}\right)^2 \\
&= 0
\end{aligned}$$

since $\mathbb{E}_x[(x^k)^3] = 0$ and $\mathbb{E}_x[x^k] = 0$ as odd moments of standard Gaussian, as well as $\mathbb{E}_x \sum_{i\neq k} x^i w^{*i} = \sum_{i\neq k} \mathbb{E}_x \, x^i w^{*i} = 0$. We also used the fact that the Gaussian vectors x_i and x_j for $i \neq j$ are independent since $\mathrm{cov}(x_i, x_j) = 0$.

Now, using (1),

$$\begin{aligned}
\mathbb{E}_x[(x^T \widehat{w} - (x^T w^*)^2)^2] &= \mathbb{E}[(x^T \widehat{w})^2] + \mathbb{E}[(x^T w^*)^4] \\
&\geq \mathbb{E}[(x^T w^*)^4] \\
&= 3(w^{*T} I w^*)^2 = 3\|w^*\|^4 > 0.
\end{aligned}$$

Exercise 1.3. (a) We have $\forall x \in \mathcal{X}$

$$\widehat{\eta}(x) = \frac{1}{|\{i : |x - x_i| \le \epsilon\}|} \sum_{i:|x_i - x| \le \epsilon} \mathbf{1}\{y_i = 1\}.$$

Then the respective plug-in classifier is $\widehat{h}(x) = \mathbf{1}\left\{\widehat{\eta}(x) \ge \frac{1}{2}\right\}$.

(b) If $\mathcal{X} = \{0, 1\}$ and $\epsilon < 1$ then $\forall x \in \mathcal{X}$

$$\widehat{\eta}(x) = \frac{1}{|\{i : x = x_i\}|} \sum_{i:x = x_i} \mathbf{1}\{y_i = 1\}.$$

Note that for $x = x_i$ the conditional variable $(y_i \mid x)$ has Bernoulli distribution with some parameter p_x. Particularly $(y \mid x)$ has the same distribution $\mathrm{Bern}(p_x)$, so $\eta(x) = \mathbb{P}(y = 1 \mid x) = p_x$. Thus, according to the law of large numbers $\widehat{\eta}(x) \to \eta(x)$ as $m \to \infty$.

(c) We need to prove that $\mathbb{E}_S[L(\widehat{h})] \to L^*$ as $m \to \infty$ where L^* is the Bayes risk and $\widehat{h}$ is the plug-in classifier corresponding to $\widehat{\eta}$ from previous subproblem. Note that

$$\begin{aligned}
\mathbf{1}\{\widehat{h}(x) \ne y\} - \mathbf{1}\{h^*(x) \ne y\} &= \mathbf{1}\{y = 1\}\,\mathbf{1}\{\widehat{h}(x) \ne 1\}\,\mathbf{1}\{h^*(x) = 1\} \\
&\quad - \mathbf{1}\{y = 1\}\,\mathbf{1}\{\widehat{h}(x) = 1\}\,\mathbf{1}\{h^*(x) \ne 1\} \\
&\quad - \mathbf{1}\{y \ne 1\}\,\mathbf{1}\{\widehat{h}(x) \ne 1\}\,\mathbf{1}\{h^*(x) = 1\} \\
&\quad + \mathbf{1}\{y \ne 1\}\,\mathbf{1}\{\widehat{h}(x) = 1\}\,\mathbf{1}\{h^*(x) \ne 1\}.
\end{aligned}$$

After taking expectation w.r.t. $y \mid x$ it becomes

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{\widehat{h}(x) \ne y\} - \mathbf{1}\{h^*(x) \ne y\}] = {} & \eta(x)\,\mathbf{1}\{\widehat{h}(x) \ne 1\}\,\mathbf{1}\{h^*(x) = 1\} \\
& -\eta(x)\,\mathbf{1}\{\widehat{h}(x) = 1\}\,\mathbf{1}\{h^*(x) \ne 1\} \\
& -(1 - \eta(x))\,\mathbf{1}\{\widehat{h}(x) \ne 1\}\,\mathbf{1}\{h^*(x) = 1\} \\
& +(1 - \eta(x))\,\mathbf{1}\{\widehat{h}(x) = 1\}\,\mathbf{1}\{h^*(x) \ne 1\} \\
= {} & (2\eta(x) - 1)\,\mathbf{1}\{\widehat{h}(x) \ne 1\}\,\mathbf{1}\{h^*(x) = 1\} \\
& +(1 - 2\eta(x))\,\mathbf{1}\{\widehat{h}(x) = 1\}\,\mathbf{1}\{h^*(x) \ne 1\} \\
= {} & |1 - 2\eta(x)|\,\mathbf{1}\{\widehat{h}(x) \ne h^*(x)\}.
\end{aligned}$$

Note that for $\widehat{h}(x) \ne h^*(x)$ holds $|\eta(x) - \frac{1}{2}| \le |\eta(x) - \widehat{\eta}(x)|$. Indeed, if $h^*(x) = 1, \widehat{h}(x) \ne 1$ then $\eta(x) \ge \frac{1}{2} \ge \widehat{\eta}(x)$ and same for the other case. Thus

$$\begin{aligned}
\mathbb{E}_S[L(\widehat{h})] - L^* &= 2\,\mathbb{E}_S\left[\left|\eta(x) - \frac{1}{2}\right| \mathbf{1}\{\widehat{h}(x) \ne h^*(x)\}\right] \\
&\le 2\,\mathbb{E}_S[|\widehat{\eta}(x) - \eta(x)|] \xrightarrow[m \to \infty]{} 0
\end{aligned}$$

where the last part is true due to the previous subproblem.

Statistical foundations of learning (Summer 2025)

Saro Harutyunyan

May 2025

Assignment 2

Exercise 1.1. Using the chain rule we can calculate the joint distribution of (X, Y):

$X \backslash Y$	1	2	3
1	0	1/8	1/3
2	1/12	0	1/6
3	1/6	1/8	0

Let $\eta_i(X) := \mathbb{P}(Y = i \mid X)$. Then the risk of a classifier h is

$$
\begin{aligned}
L_{(x,y)}(h) &= \mathbb{E}_{(x,y)}[\mathbf{1}\{h(x) \neq y\}] \\
&= \mathbb{E}_x \, \mathbb{E}_{y|x}[\sum_{i=1}^{3} \mathbf{1}\{y = i\} \, \mathbf{1}\{h(x) \neq i\}] \\
&= \mathbb{E}_x[\sum_{i=1}^{3} \eta_i(x)(1 - \mathbf{1}\{h(x) = y\})] \\
&= 1 - \mathbb{E}_x[\sum_{i=1}^{3} \eta_i(x) \, \mathbf{1}\{h(x) = y\}] \\
&\geq 1 - \mathbb{E}_x[\max_i \eta_i(x)].
\end{aligned}
$$

Obviously the last bound is reachable for $h^*(x) = \arg\max_i \eta_i(x)$.

It remains only to plug in our numbers. The Bayes risk is

$$
\begin{aligned}
L^* &= 1 - \mathbb{E}_x[\max_i \eta_i(x)] \\
&= 1 - \left(\max\left\{0, \frac{1}{8}, \frac{1}{3}\right\} + \max\left\{\frac{1}{12}, 0, \frac{1}{6}\right\} + \max\left\{\frac{1}{6}, \frac{1}{8}, 0\right\} \right) \\
&= 1 - \left(\frac{1}{3} + \frac{1}{6} + \frac{1}{6} \right) = \frac{1}{3}
\end{aligned}
$$

with Bayes classifier $h^*(1) = 3, h^*(2) = 3, h^*(3) = 1$.

Exercise 1.2. We need to prove that $\mathcal{R}_m(X) = \mathcal{R}_m(\mathrm{conv}(X))$ for $X \subset \mathbb{R}^m$ where

$$\mathcal{R}_m(X) = \frac{1}{m}\,\mathbb{E}_\sigma\left[\sup_{x\in X}\langle \sigma, x\rangle\right].$$

To do so we prove that $\sup\limits_{x\in X}\langle \sigma, x\rangle = \sup\limits_{x\in \mathrm{conv}\, X}\langle \sigma, x\rangle$.

Obviously the left supremum is not greater than the right one since $X \subset \mathrm{conv}\, X$.

To prove the converse inequality take any $x \in \mathrm{conv}\, X$. Then $x = \sum_{i=1}^N \lambda_i x_i$ where $x_i \in X$, $\lambda_i \geq 0$ and $\sum_{i=1}^N \lambda_i = 1$. We have

$$\langle \sigma, x\rangle = \sum_{i=1}^N \lambda_i \langle \sigma, x_i\rangle \leq \max_{i\in[N]}\langle \sigma, x_i\rangle \leq \sup_{x'\in X}\langle \sigma, x'\rangle.$$

Hence, after taking supremum in the left side we get $\sup\limits_{x\in \mathrm{conv}\, X}\langle \sigma, x\rangle \leq \sup\limits_{x\in X}\langle \sigma, x\rangle$ as needed.

Exercise 1.3. 1. Note that

$$\begin{aligned}\mathbb{E}_S[L_{S,\alpha}(h)] &= \mathbb{E}_S[\alpha L_{S_1}(h) + (1-\alpha)L_{S_2}(h)] \\ &= \alpha\,\mathbb{E}_{S_1} L_{S_1}(h) + (1-\alpha)\,\mathbb{E}_{S_2} L_{S_2}(h) \\ &= \alpha L_{\mathcal{D}_1} + (1-\alpha)L_{\mathcal{D}_2}.\end{aligned}$$

Hence what we need to prove reduces to the trivial inequality

$$(1-\alpha)(L_{\mathcal{D}_1}(h) - L_{\mathcal{D}_2}(h)) \le (1-\alpha)\sup_{h\in\mathcal{H}} |L_{\mathcal{D}_1}(h) - L_{\mathcal{D}_2}(h)|.$$

2. Note that $Z_i := \mathbf{1}\{h(x_i) \neq y_i\} \in [0,1]$ are independent. Then

$$\begin{aligned}&\mathbb{P}_S\left(\sup_{h\in\mathcal{H}} |L_{S,\alpha}(h) - \mathbb{E}_S[L_{S,\alpha}(h)]| > \epsilon\right) \\ &\le \sum_{h\in\mathcal{H}} \mathbb{P}_S\big(|L_{S,\alpha}(h) - \mathbb{E}_S[L_{S,\alpha}(h)]| > \epsilon\big) \\ &= \sum_{h\in\mathcal{H}} \mathbb{P}_S\big(|\alpha(L_{S_1} - \mathbb{E}_{S_1}[L_{S_1}(h)]) + (1-\alpha)(L_{S_2} - \mathbb{E}_{S_2}[L_{S_2}(h)])| > \epsilon\big) \\ &\le \sum_{h\in\mathcal{H}} \mathbb{P}_S\left(\left|\left(\frac{\alpha}{\beta m}\sum_{i\in[\beta m]} + \frac{1-\alpha}{(1-\beta)m}\sum_{i\in[m]\setminus[\beta m]}\right)\big(\mathbf{1}\{h(x_i)\neq y_i\} - \mathbb{E}[\mathbf{1}\{h(x_i)\neq y_i\}]\big)\right| > \epsilon\right) \\ &\le \sum_{h\in\mathcal{H}} \mathbb{P}_S\left(\left|\left(\frac{\alpha}{\beta m}\sum_{i\in[\beta m]}(Z_i - \mathbb{E}_S Z_i) + \frac{1-\alpha}{(1-\beta)m}\sum_{i\in[m]\setminus[\beta m]}(Z_i - \mathbb{E}_S Z_i)\right)\right| > \epsilon\right) \\ &\le |\mathcal{H}|\cdot 2\exp\left(-\frac{2m^2\epsilon^2}{\displaystyle\sum_{i\in[\beta m]}\frac{\alpha^2}{\beta^2} + \sum_{i\in[m]\setminus[\beta m]}\frac{(1-\alpha)^2}{(1-\beta)^2}}\right) \\ &= 2|\mathcal{H}|\exp\left(-\frac{2m\epsilon^2}{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\right)\end{aligned}$$

where the last inequality is Hoeffding. Now it remains to denote by δ the left hand side of the previous inequality chain and deduce from there the value

$$\epsilon = \sqrt{\frac{1}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}.$$

3. Use the definition of ϵ from the last paragraph. We have

$$\begin{aligned}L_{\mathcal{D}_1}(\widehat{h}) &\le \mathbb{E}_S[L_{S,\alpha}(\widehat{h})] + (1-\alpha)d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_2) \\ &\le L_{S,\alpha}(\widehat{h}) + \epsilon + (1-\alpha)d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_2) \\ &\le L_{S,\alpha}(h^*) + \epsilon + (1-\alpha)d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_2) \\ &\le \mathbb{E}_S L_{S,\alpha}(h^*) + 2\epsilon + (1-\alpha)d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_2) \\ &= (1-\alpha)(L_{\mathcal{D}_2}(h^*) + d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_2)) + 2\epsilon\end{aligned}$$

as needed.

Statistical foundations of learning (Summer 2025)

Saro Harutyunyan

May 2025

Assignment 3

Exercise 2.1. Define $\alpha^T = \begin{pmatrix} \alpha_1 & \dots & \alpha_n \end{pmatrix}$ and let V be the matrix with v_i its i^{th} row. Define

$$\mathcal{F} = \{(x \in \mathbb{R}^n) \mapsto \operatorname{sign}(\alpha^T x + b) \mid \alpha_1, \dots, \alpha_n, b \in \mathbb{R}\},$$
$$\mathcal{G} = \{(x \in \mathbb{R}^d) \mapsto Vx\}.$$

Then $\mathcal{H} = \mathcal{F} \circ \mathcal{G} := \{f \circ g : f \in \mathcal{F}, g \in \mathcal{G}\}$. Note that $\mathcal{G}$ contains only one function (since V is fixed).

1. Suppose $\mathcal{H}$ can shatter $n+2$ points $x_1, \dots, x_{n+2} \in \mathbb{R}^d$. Then $\mathcal{F}$ can shatter $n+2$ points $Vx_1, \dots, Vx_{n+2} \in \mathbb{R}^n$. This is a contradiction with $\operatorname{VCdim}(F) = n+1$ (from the lecture VC-Dimension and Error bounds for 0-1 loss function). Thus

$$\operatorname{VCdim}(\mathcal{H}) \leq \operatorname{VCdim}(\mathcal{F}) \operatorname{VCdim}(\mathcal{G}) = \operatorname{VCdim}(\mathcal{F}) \leq n + 1.$$

2. $\operatorname{VCdim}(\mathcal{H}) = n+1$ iff there are $x_1, \dots, x_{n+1} \in \mathbb{R}^d$ such that the points $Vx_1, \dots, Vx_{n+1} \in \mathbb{R}^n$ do not lie in a hyperplane, i.e. iff $\operatorname{rank} V = n$.

Exercise 2.2. 1. First, $\mathrm{VCdim}(\mathcal{H}_7) \le \lfloor \log_2 |\mathcal{H}_7| \rfloor = 2$. Now we calculate $h_k(x)$ for all $x, k \in \mathcal{X}_7$:

$k \backslash x$	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	0	1	0	1	0	1	0
3	0	0	1	0	0	1	0
4	0	0	0	1	0	0	0
5	0	0	0	0	1	0	0
6	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1

It can be seen that $x_1 = 2, x_2 = 3$ can be shattered by h_1, h_2, h_3, h_4. Thus $\mathrm{VCdim}(\mathcal{H}_7) = 2$.

2. We will prove that the maximal n with $\mathrm{VCdim}(\mathcal{H}_n) = 2$ is 14. Equivalently we need to show that the minimal n with $\mathrm{VCdim}(\mathcal{H}_n) = 3$ is 15. For that we need to find a submatrix of size 8×3 in the above table (but completed to size 15×15) so that all the eight combinations $\{0,1\}^3$ appear there as rows. That is, we seek $x_1, x_2, x_3 \in \mathcal{X}_{15}$ and $k_1, \ldots, k_8 \in \mathcal{X}_{15}$ such that $\{h_{k_i}\}_{i=1}^8$ shatters x_1, x_2, x_3.

   Note that for each x_i there should be at least four of $k_j, j = 1, \ldots, 8$ corresponding to 1. In other words, each x_i should be a multiple of at least four k_j. Inspecting all the numbers from 1 to 15 we find that only $6, 8, 10, 12, 14, 15$ have at least four divisors. Now consider the table corresponding to them:

$k \backslash x$	6	8	10	12	14	15
1	1	1	1	1	1	1
2	1	1	1	1	1	0
3	1	0	0	1	0	1
4	0	1	0	1	0	0
5	0	0	1	0	0	1
6	1	0	0	1	0	0
7	0	0	0	0	1	0
8	0	1	0	0	0	0
9	0	0	0	0	0	0
10	0	0	1	0	0	0
11	0	0	0	0	0	0
12	0	0	0	1	0	0
13	0	0	0	0	0	0
14	0	0	0	0	1	0
15	0	0	0	0	0	1

   If we restrict ourselves to $n \le 14$, then note that the rows corresponding 1 and 2 contain only 1s and we may choose only one of them (for $(1,1,1)$). Thus we will not be able to find a submatrix 8×3 with all the combinations of $\{0,1\}^3$ because for example in the columns of 6 and 8 there will be left only two 1s. Hence we need to consider 15 too. But it can be seen that $10, 12, 15$ can be shattered by $h_j, j \in \{1, 2, 3, 5, 10, 11, 12, 15\}$.

Exercise 2.3. 1. We prove that $\mathrm{VCdim}(\mathcal{H}) = 2k+2$. First, we cannot shatter any $2k+3$ points $0 \le x_1 \le \ldots \le x_{2k+3} \le 1$ since the combination $1, 0, 1, 0, \ldots, 1$ is unreachable (we can have only $k+1$ intervals $[t_{2i}, t_{2i+1})$ for 1s, while we have $k+2$ 1s).

Now, note that we can shatter any $2k+2$ points $0 < x_1 < \ldots < x_{2k+2} < 1$. Indeed, for any labeling $y \in \{0,1\}^{2k+2}$ define $I = \{i : y_i = 1\}$. Then define

$$J = \{[a,b] : \ a-1 \notin I, \ b+1 \notin I, \ c \in I \ \ \forall c \in [a,b] \cap \mathbb{N}, \ a, b \in \mathbb{N}\}.$$

That is, J is the set of segments of consecutive 1s. Obviously $|J| \le k+1$. Indeed, if J consists of n segments $I_1, \ldots, I_n$, then there is at least one zero between I_i and I_{i+1} for any i. Thus $n + (n-1) \le 2k+2$ whence $|J| = n \le k+1$.

Now let $J = \{[a_i, b_i] : i \in [n]\}$ where $b_i < a_{i+1}$ for all i. Then choose $t_{2i} = a_i - \varepsilon$ and $t_{2i+1} = b_i + \varepsilon$ for all $i \in [n]$ for sufficiently small ε. For $i \in [k+1] \setminus [n]$ choose $t_{2i} = t_{2i+1}$. Then the respective function from $\mathcal{H}$ for the chosen t_i attains the values $y_1, \ldots, y_{2k+2}$ on $x_1, \ldots, x_{2k+2}$ as needed.

2. We show that $\mathrm{VCdim}(\mathcal{H}) = \infty$. For any $n \in \mathbb{N}$ choose $x_i = 10^{-i}$ for all $i \in [n]$. For any labels $y_1, \ldots, y_n \in \{\pm 1\}$ put $a = \pi \left(1 + \sum_{i=1}^{n} \frac{(1-y_i)10^i}{2}\right)$. Note that $a = \pi \left(1 + \sum_{i \in I} 10^i\right)$ where $I = \{i : y_i = -1\}$.

   We aim to show that $y_i = \operatorname{sign}\sin(ax_i)$. For a given i, if $y_i = 1$ then

$$\begin{aligned}
\operatorname{sign}\sin(ax_i) &= \operatorname{sign}\sin(\pi(10^{-i} + \sum_{j \in I} 10^{j-i})) \\
&= \operatorname{sign}\sin(\pi(10^{-i} + \sum_{\substack{j \in I \\ j<i}} 10^{j-i})) \\
&= \operatorname{sign}\sin(\pi(10^{-i} + \frac{1}{9})) \\
&= 1 = y_i
\end{aligned}$$

   where we used the fact that $\sum_{\substack{j \in I \\ j<i}} 10^{j-i} < \sum_{i=1}^{\infty} 10^{-i} = \frac{1}{9}$ and that $\pi(10^{-i} + 1/9)$ lies on $(0, \pi/2)$. We deal with the case $y_i = -1$ completely similarly. The only difference is that $\sum_{j \in I} 10^{j-i} = 1 + \sum_{\substack{j \in I \\ j<i}} 10^{j-i}$, so the sign of sin shifts.

Statistical foundations of learning (Summer 2025)

Saro Harutyunyan

May 2025

Assignment 4

Exercise 4.1. (a) By Markov and Paley-Zygmund for any $\theta \in [0,1]$

$$\frac{\mathbb{E}\sqrt{Z}}{\sqrt{\theta\,\mathbb{E}\,Z}} \geq \mathbb{P}(\sqrt{Z} > \sqrt{\theta\,\mathbb{E}\,Z}) = \mathbb{P}(Z > \theta\,\mathbb{E}\,Z) \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} > c(1-\theta)^2.$$

So we can choose $c' = c\sqrt{\theta}(1-\theta)^2$. But note that by Cauchy's AM-GM inequality

$$4\theta(1-\theta)(1-\theta)(1-\theta)(1-\theta) \leq \left(\frac{4\theta + (1-\theta) + (1-\theta) + (1-\theta) + (1-\theta)}{5}\right)^5 = \frac{4^5}{5^5}$$

whence inserting the optimal $\theta = \dfrac{1}{5}$ we find $c' = \sqrt{\dfrac{4^4}{5^5}}c$.

(b) (a) Note that $\mathbb{E}[S]^2 = n^2p^2 > c(np(1-p) + n^2p^2) = c\,\mathbb{E}[S^2]$ for $c = 1/2$ so

$$\mathbb{E}[\sqrt{S}] \geq c'\sqrt{\mathbb{E}[S]} = c'\sqrt{np}.$$

(b) For $Z = Y^2$ we have

$$\mathbb{E}[Z] = \mathbb{E}[Y^2] = \mathbb{E}[\left(\sum_{i=1}^n \sigma_i\right)^2] = \sum_{i=1}^n \mathbb{E}\,\sigma_i^2 + 2\sum_{i<j} \mathbb{E}\,\sigma_i\,\mathbb{E}\,\sigma_j = n$$

$$\begin{aligned}
\mathbb{E}[Z^2] = \mathbb{E}[\left(\sum_{i=1}^n \sigma_i\right)^4] &= \sum_{i=1}^n \mathbb{E}\,\sigma_i^4 + \sum_{\substack{i,j,k,l \\ \text{not all equal}}} \mathbb{E}\,\sigma_i\sigma_j\sigma_k\sigma_l \\
&= n + \sum_{i\neq j} \mathbb{E}\,\sigma_i^2\sigma_j^2 = n + \frac{1}{2}\binom{4}{2}n(n-1) = 3n^2 - 2n.
\end{aligned}$$

Thus $\mathbb{E}[Z]^2 > c\,\mathbb{E}[Z^2]$ for $c = 1/3$ whence

$$\mathbb{E}[Y] = \mathbb{E}[\sqrt{Y^2}] \geq c'\sqrt{\mathbb{E}[Y^2]} = c'\sqrt{n}.$$

Actually I accidentally found $E[Y]$ too and it would be a pity if it is lost, so I present it here. The variable $\#\{i : \sigma_i = 1\}$ is Binomial$(n, 1/2)$ so

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{k=0}^n 2^{-n}\binom{n}{k}|k + (n-k)| = 2\sum_{k=0}^{\lfloor n/2\rfloor} 2^{-n}\binom{n}{k}(n-2k) \\
&= 2n\sum_{k=0}^{\lfloor n/2\rfloor} 2^{-n}\binom{n}{k} - 2\cdot 2\sum_{k=0}^{\lfloor n/2\rfloor} 2^{-n}\binom{n}{k}k.
\end{aligned}$$

We have

$$\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n}{k}k = \sum_{k=1}^{\lfloor n/2\rfloor}\binom{n-1}{k-1}n = n\sum_{k=0}^{\lfloor n/2\rfloor - 1}\binom{n-1}{k} = \begin{cases} n2^{2l-2}, & n = 2l \\ n2^{2l-1} - \dfrac{n}{2}\dbinom{2l}{l}, & n = 2l+1 \end{cases}$$

$$\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n}{k}=\begin{cases}2^{2l-1}+\dfrac{1}{2}\dbinom{2l}{l}, & n=2l\\ 2^{2l}, & n=2l+1.\end{cases}$$

$$\mathbb{E}[Y]=2^{1-n}n\cdot\begin{cases}\dfrac{1}{2}\dbinom{2l}{l}, & n=2l\\ \dbinom{2l}{l}, & n=2l+1.\end{cases}$$

Exercise 4.2. We have

$$\begin{aligned} R_{\mathcal{D}_\mathcal{X},m}(\mathcal{H}) &= \mathbb{E}_{x_1,\ldots,x_m\sim\mathcal{D}_\mathcal{X}}\,\mathbb{E}_{\sigma_1,\ldots,\sigma_m}\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i) \\ &= \mathbb{E}_{x_1,\ldots,x_m\in \mathrm{supp}\mathcal{D}_\mathcal{X}}\,\mathbb{E}_{\sigma_1,\ldots,\sigma_m}\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i) \end{aligned}$$

where $\mathrm{supp}\mathcal{D}_\mathcal{X}$ is the support of $\mathcal{D}_\mathcal{X}$.

WLOG $\mathrm{supp}\mathcal{D}_\mathcal{X} = [d]$. For fixed $x_1,\ldots,x_m \in \mathrm{supp}\mathcal{D}_\mathcal{X}$ define $I_i = \{j : x_j = i\}$ for each $i \in [d]$. We have

$$\begin{aligned} \sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i) &= \sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^d\sum_{j\in I_i}\sigma_j h(x_j) \\ &= \sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^d h(x_i')\sum_{j\in I_i}\sigma_j \end{aligned}$$

where x_i' is any element of $\{x_j : j \in I_i\}$. Since $\mathcal{H}$ shatters $[d]$ the last supremum is at least

$$\frac{1}{m}\sum_{i=1}^d\left|\sum_{j\in I_i}\sigma_j\right|.$$

After taking $\mathbb{E}_{\sigma_1,\ldots,\sigma_m}$ and using part (b) of part (b) of the last exercise it becomes

$$\begin{aligned} \mathbb{E}_{\sigma_1,\ldots,\sigma_m}\frac{1}{m}\sum_{i=1}^d\left|\sum_{j\in I_i}\sigma_j\right| &= \frac{1}{m}\sum_{i=1}^d\mathbb{E}_{\sigma_1,\ldots,\sigma_m}\left|\sum_{j\in I_i}\sigma_j\right| \\ &\geq \frac{1}{m}\sum_{i=1}^d c\sqrt{|I_i|}. \end{aligned}$$

Now note that $Y = |I_1|, |I_2|, \ldots, |I_d|$ have the same distribution. So after taking expectation wrt $x_1,\ldots,x_m \in \mathrm{supp}\mathcal{D}_\mathcal{X}$ it becomes

$$R_{\mathcal{D}_\mathcal{X},m}(\mathcal{H}) \geq \frac{cd}{m}\mathbb{E}_{x_1,\ldots,x_m\in\mathrm{supp}\mathcal{D}_\mathcal{X}}\sqrt{Y}.$$

Note that $Y \sim \mathrm{Binomial}(m, 1/d)$. So using part (a) of part (b) of the previous exercise

$$\mathbb{E}\sqrt{Y} \geq c'\sqrt{\mathbb{E}\,Y} = c'\sqrt{\frac{m}{d}}$$

$$R_{\mathcal{D}_\mathcal{X},m}(\mathcal{H}) \geq \frac{cd}{m}c'\sqrt{\frac{m}{d}} \geq c''\sqrt{\frac{d}{m}}$$

for some $c'' > 0$, as needed.

Exercise 4.3. Write z for (x, y), ℓ_h^z for $\ell(h(x), y)$ and σ for the vector of independent Rademacher variables $\sigma_1, \ldots, \sigma_m$. Along with $S = (z_1, \ldots, z_m)$ we will introduce independent variables $S' = (z'_1, \ldots, z'_m)$ having the same distribution. By Jensen

$$\begin{aligned}
\mathbb{E}_{S,\sigma} \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell_h^{z_i} - \mathbb{E}_{z\sim\mathcal{D}}[\ell_h^z]\right) \right| &= \mathbb{E}_{S,\sigma} \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell_h^{z_i} - \mathbb{E}_{z'_i\sim\mathcal{D}}[\ell_h^{z'_i}]\right) \right| \\
&\le \mathbb{E}_{S,S',\sigma} \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell_h^{z} - \ell_h^{z'_i}\right) \right| \\
&= \mathbb{E}_{S,S'} \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \left(\ell_h^{z_i} - \ell_h^{z'_i}\right) \right|.
\end{aligned}$$

By triangle inequality

$$\sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \left(\ell_h^{z_i} - \ell_h^{z'_i}\right) \right| \le \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \left(\ell_h^{z_i} - \mathbb{E}_z \ell_h^z\right) \right| + \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \left(\ell_h^{z'_i} - \mathbb{E}_z \ell_h^z\right) \right|.$$

Taking $\mathbb{E}_{S,S'}$, taking into account that z_i, z'_i have the same distribution and remembering the above inequalities we get

$$\begin{aligned}
\mathbb{E}_{S,\sigma} \left[\sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell_h^{z_i} - \mathbb{E}_{z\sim\mathcal{D}}[\ell_h^z]\right) \right| \right] &\le 2\, \mathbb{E}_S \left[\sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \left(\ell_h^{z_i} - \mathbb{E}_z \ell_h^z\right) \right| \right] \\
&= 2\, \mathbb{E}_S \left[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \right].
\end{aligned}$$

On the other hand, by triangle inequality

$$\begin{aligned}
\mathbb{E}_{S,\sigma} \left[\sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \left(\ell_h^{z_i} - \mathbb{E}_{z\sim\mathcal{D}}[\ell_h^z]\right) \right| \right] &\ge \mathbb{E}_{S,\sigma} \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \ell_h^{z_i} \right| - \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i\, \mathbb{E}_z \ell_h^z \right| \\
&\ge R_{\mathcal{D},m}(\ell \circ \mathcal{H}) - c\, \mathbb{E}_\sigma \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \right| \\
&\ge R_{\mathcal{D},m}(\ell \circ \mathcal{H}) - \frac{c}{\sqrt{m}}
\end{aligned}$$

where we used as well Cauchy-Bunyakovsky-Schwarz:

$$\mathbb{E}_\sigma \left| \sum_{i=1}^m \sigma_i \right| \le \sqrt{\mathbb{E}_\sigma \left(\sum_{i=1}^m \sigma_i \right)^2} = \sqrt{m}.$$

Statistical foundations of learning (Summer 2025)

Saro Harutyunyan

May 2025

Assignment 5

Exercise 5.1. Using the hint, 1-Lipschitzness of ϕ and Theorem Lip.3 (Contraction principle for Lipschitz functions), as well as Cauchy-Schwarz and Jensen

$$
\begin{aligned}
R_{\mathcal{X}}(\mathcal{H}) &= \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i) \right| \le \frac{2}{n}\,\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i h(x_i) \\
&= \frac{2}{n}\,\mathbb{E}_\sigma \sup_{\substack{v,w_1,\dots,w_m \\ \|v\|\le C_1, \|w_i\|\le C_2}} \sum_{i=1}^n \sigma_i v^T \phi(W x_i) \\
&\le \frac{2}{n}\,\mathbb{E}_\sigma \sup_{\substack{v,w_1,\dots,w_m \\ \|v\|\le C_1, \|w_i\|\le C_2}} \sum_{i=1}^n \sigma_i v^T W x_i \\
&= \frac{2}{n}\,\mathbb{E}_\sigma \sup_{\substack{v,w_1,\dots,w_m \\ \|v\|\le C_1, \|w_i\|\le C_2}} \langle W^T v, \sum_{i=1}^n \sigma_i x_i \rangle \\
&\le \frac{2C_1C_2}{n}\sqrt{m}\,\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i \right\| \\
&\le \frac{2C_1C_2}{n}\sqrt{m}\sqrt{\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i \right\|^2} \\
&= \frac{2C_1C_2}{n}\sqrt{m}\sqrt{\sum_{i=1}^n \|x_i\|^2} \le \frac{2C_1C_2\sqrt{m}}{\sqrt{n}}.
\end{aligned}
$$

Now we prove the hint. Denote $S(\sigma,f,x) := \sum_{i=1}^n \sigma_i f(x_i)$. Then $S(-\sigma,f,x) = -S(\sigma,f,x)$ so we can define

$$
\Sigma_+ := \left\{ \sigma \in \{\pm 1\}^n : \sup_{f\in\mathcal{F}} S(\sigma,f,x) > |\inf_{f\in\mathcal{F}} S(\sigma,f,x)| \right\},
$$

$$
\Sigma_- := \left\{ \sigma \in \{\pm 1\}^n : \sup_{f\in\mathcal{F}} S(\sigma,f,x) < |\inf_{f\in\mathcal{F}} S(\sigma,f,x)| \right\}.
$$

$$
\begin{aligned}
\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right] &= \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} |S(\sigma,f,x)| \\
&= \mathbb{E}_{\sigma\in\Sigma_+} \sup_{f\in\mathcal{F}} |S(\sigma,f,x)| + \mathbb{E}_{\sigma\in\Sigma_-} \sup_{f\in\mathcal{F}} |S(\sigma,f,x)| \\
&= 2\,\mathbb{E}_{\sigma\in\Sigma_+} \sup_{f\in\mathcal{F}} S(\sigma,f,x)
\end{aligned}
$$

$$
\begin{aligned}
2\,\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \sigma_i f(x_i) \right] &= 2\,\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} S(\sigma,f,x) \\
&= 2\,\mathbb{E}_{\sigma\in\Sigma_+} \sup_{f\in\mathcal{F}} S(\sigma,f,x) + 2\,\mathbb{E}_{\sigma\in\Sigma_-} \sup_{f\in\mathcal{F}} S(\sigma,f,x).
\end{aligned}
$$

So the hint is equivalent to $\mathbb{E}_{\sigma\in\Sigma_-} \sup_{f\in\mathcal{F}} S(\sigma, f, x) \geq 0$. This will hold if we, for example, require $\sup_{f\in\mathcal{F}} S(\sigma, f, x) \geq 0$ for any $\sigma \in \{\pm 1\}^n$, or, particularly, that for any $f \in \mathcal{F}$ also $-f \in \mathcal{F}$ holds (for the hypothesis class of the problem this holds). If we do not have any assumption on $\mathcal{F}$ then the inequality of the hint fails for $n = 1$ and $\mathcal{F} = \{f\}$ where $f \equiv 1$.

Exercise 5.2. We have

$$\mathbb{E}_{S\sim\mathcal{D}^m}[(L_{\mathcal{D}}(\mathcal{A}_S)-L_S(\mathcal{A}_S))^2]=\mathbb{E}_S\left[\mathbb{E}_{(x,y)}\left(\ell(\mathcal{A}_S(x),y)-\frac{1}{m}\sum_{i=1}^m\ell(\mathcal{A}_S(x_i),y_i)\right)^2\right]$$
$$=\frac{1}{m^2}\,\mathbb{E}_S\left[\left(\sum_{i=1}^m\mathbb{E}_{(x,y)}\,\ell(\mathcal{A}_S(x),y)-\ell(\mathcal{A}_S(x_i),y_i)\right)^2\right].$$

Using the notations $z=(x,y)$, $z'=(x',y')$, $z_i=(x_i,y_i)$, $\ell_S^z=\ell(\mathcal{A}_S(x),y)$ the last expression is

$$=\frac{1}{m^2}\,\mathbb{E}_S\left[\left(\sum_{i=1}^m\mathbb{E}_z\,\ell_S^z-\ell_S^{z_i}\right)^2\right]$$
$$=\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m\mathbb{E}_{S,z,z'}\left[(\ell_S^z-\ell_S^{z_i})(\ell_S^{z'}-\ell_S^{z_j})\right].$$

For $i=j$ we use $-c\le\ell_S^z-\ell_S^{z_i}\le c$ and $-c\le\ell_S^{z'}-\ell_S^{z_j}\le c$ to get

$$\mathbb{E}_{S,z,z'}\left[(\ell_S^z-\ell_S^{z_i})(\ell_S^{z'}-\ell_S^{z_j})\right]\le c^2$$

$$\frac{1}{m^2}\sum_{i=j=1}^m\mathbb{E}_{S,z,z'}\left[(\ell_S^z-\ell_S^{z_i})(\ell_S^{z'}-\ell_S^{z_j})\right]\le\frac{c^2}{m}.$$

Now define $a=z_i$, $a'=z_j$, $T=S\setminus\{z_i,z_j\}$. Then for $i\ne j$

$$\begin{aligned}&\mathbb{E}_{S,z,z'}\left[(\ell_S^z-\ell_S^{z_i})(\ell_S^{z'}-\ell_S^{z_j})\right]\\&=\mathbb{E}_{T,a,a',z,z'}\left[(\ell_{T,a,a'}^z-\ell_{T,a,a'}^a)(\ell_{T,a,a'}^{z'}-\ell_{T,a,a'}^{a'})\right]\\&=\mathbb{E}_{T,a,a',z,z'}\left[(\ell_{T,a,a'}^z\ell_{T,a,a'}^{z'}-\ell_{T,a,a'}^z\ell_{T,a,a'}^{a'}-\ell_{T,a,a'}^a\ell_{T,a,a'}^{z'}+\ell_{T,a,a'}^a\ell_{T,a,a'}^{a'})\right].\end{aligned}$$

Since a,a',z,z' are i.i.d. we can swap them accordingly:

$$\begin{aligned}&=\mathbb{E}\left[(\ell_{T,z,z'}^a\ell_{T,z,z'}^{a'}-\ell_{T,z,a'}^a\ell_{T,z,a'}^{a'}-\ell_{T,a,z'}^a\ell_{T,a,z'}^{a'}+\ell_{T,a,a'}^a\ell_{T,a,a'}^{a'})\right]\\&=\mathbb{E}\Big[\ell_{T,z,z'}^a(\ell_{T,z,z'}^{a'}-\ell_{T,z,a'}^{a'})+\ell_{T,z,a'}^{a'}(\ell_{T,z,z'}^a-\ell_{T,z,a}^a)+\ell_{T,z,a'}^{a'}(\ell_{T,z,a}^a-\ell_{T,z,a'}^a)\\&\qquad\ell_{T,a,a'}^a(\ell_{T,a,a'}^{a'}-\ell_{T,a,z'}^{a'})+\ell_{T,a,z'}^{a'}(\ell_{T,a,a'}^a-\ell_{T,z',a'}^a)+\ell_{T,a,z'}^{a'}(\ell_{T,z',a'}^a-\ell_{T,z',a}^a)\Big].\end{aligned}$$

But

$$\mathbb{E}\left[\ell_{T,z,z'}^a(\ell_{T,z,z'}^{a'}-\ell_{T,z,a'}^{a'})\right]\le c\,\mathbb{E}\left[|\ell_{T,z,z'}^{a'}-\ell_{T,z,a'}^{a'}|\right]\le c\beta_m$$

etc. whence the conclusion follows.

Exercise 5.3. We prove that there is no such a universal λ. Fix any $\lambda > 0$. Consider the following dataset in $\mathbb{R}^d$: the point $e_1 := (1, 0, \dots, 0)$ of class $+1$, the point $-e_1$ of class -1, and the point $(-1+\epsilon)e_1$ of class $+1$, where $\epsilon \in (0,1)$ will be chosen later.

First we find the solution of hard SVM:

$$\min_{w,b} \|w\|^2 : \quad \begin{cases} 1 \cdot (\langle w, e_1 \rangle + b) \geq 1 \\ -1 \cdot (\langle w, -e_1 \rangle + b) \geq 1 \\ 1 \cdot (\langle w, (-1+\epsilon)e_1 \rangle + b) \geq 1 \end{cases} \iff \begin{cases} w_1 \geq 1 - b \\ w_1 \geq 1 + b \\ w_1 \leq \frac{b-1}{1-\epsilon} \end{cases}$$

$$1 + b \leq w_1 \leq \frac{b-1}{1-\epsilon} \quad \implies \quad b \geq \frac{2-\epsilon}{\epsilon}, \quad w_1 \geq \frac{2}{\epsilon}$$

whence the solution is $w = \dfrac{2}{\epsilon} e_1,\ b = \dfrac{2-\epsilon}{\epsilon}$ (w_1 is the first coordinate of w).

Now soft SVM:

$$\begin{aligned} &\min_{w,b} \frac{1}{3}\Big(\max\{0, 1 - 1 \cdot (\langle w, e_1 \rangle + b)\} \\ &\qquad + \max\{0, 1 - (-1) \cdot (\langle w, -e_1 \rangle + b)\} \\ &\qquad + \max\{0, 1 - 1 \cdot (\langle w, (-1+\epsilon)e_1 \rangle + b)\} \Big) + \lambda \|w\|^2 \\ \iff &\min_{w,b} \frac{1}{3}\Big(\max\{0, 1 - w_1 - b\} \\ &\qquad + \max\{0, 1 - w_1 + b\} \\ &\qquad + \max\{0, 1 + (1-\epsilon)w_1 - b\} \Big) + \lambda \|w\|^2. \end{aligned}$$

When evaluated on the hard SVM solution the soft SVM loss is $4\lambda/\epsilon^2$. However the soft SVM loss is less (for small enough $\epsilon = \epsilon(\lambda)$) when its parameters are $w = 3e_1,\ b = 2$, it is $2 - 3\epsilon + 9\lambda$. Thus the hard and soft SVMs in this case cannot have identical solutions.

Statistical foundations of learning (Summer 2025)

Saro Harutyunyan

May 2025

Assignment 6

Exercise 6.1. We prove the following more general fact: if k is a universal kernel on $\mathcal{X}$ then the normalized kernel $k^* : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with

$$k^*(x,y) := \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}} \quad \forall x,y \in \mathcal{X}$$

is also universal.

Let $\mathcal{H}$ be the RKHS and $\phi : \mathcal{X} \to \mathcal{H}$ be the feature map of k. Define $\alpha(x) := k(x,x)^{-1/2}$ for all $x \in \mathcal{X}$. Then $\alpha\phi : \mathcal{X} \to \mathcal{H}$ is a feature map of k^* so k^* is a kernel by Theorem App.4 (Positive semidefinite (psd) kernel); here $(\alpha\phi)(x) := \alpha(x)\phi(x)$.

Now we show that k^* is universal. Take an $\epsilon > 0$ and a continuous $f : \mathcal{X} \to \mathbb{R}$. Let $c := \|\alpha\|_\infty < \infty$. By universality of k there exists an $h \in \mathcal{H}$ with

$$\left\| h - \frac{f}{\alpha} \right\|_\infty \le \frac{\epsilon}{c}.$$

Thus, using $\langle h, \phi(x)\rangle_{\mathcal{H}} = h(x)$ (follows from the definition of $\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$), we obtain

$$\|\langle h, \alpha\phi\rangle_{\mathcal{H}} - f\|_\infty \le \|\alpha\|_\infty \left\| h - \frac{f}{\alpha} \right\|_\infty \le \epsilon.$$

Now it remains to note that ψ defined as $\psi(x) := \langle h, (\alpha\phi)(x)\rangle_{\mathcal{H}}$ lies in $\mathcal{H}$ (by Riesz representation theorem).

To obtain the problem claim observe that $k(x,y) = \exp(\langle x, y\rangle)$ after normalization becomes exactly $k^*(x) = \exp(-\frac{1}{2}\|x-y\|^2)$.

Exercise 6.2. Suppose the contrary, that ϕ is not injective, i.e. there are $x \neq y$ in $\mathbb{R}^d$ such that $\phi(x) = \phi(y)$. Then for any $f \in \mathcal{H}$

$$f(x) = \langle \phi(x), f \rangle = \langle \phi(y), f \rangle = f(y).$$

In other words any function $f \in \mathcal{H}$ attains equal values at x and y. But then we can find an $\epsilon > 0$, a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and a compact set $C \subseteq \mathbb{R}^d$ such that the inequality

$$\sup_{x \in C} |f(x) - h(x)| \leq \epsilon$$

fails for any $h \in \mathcal{H}$. Indeed, choose $\epsilon = 1/3$, $C = \{x, y\}$ and f to be any continuous function with $f(x) = 0$ and $f(y) = 1$. Then $f(x) = f(y) = c$ so

$$\max(|f(x) - h(x)|, |f(y) - h(y)|) = \max(|c|, |1 - c|) \geq \frac{1}{2} > \frac{1}{3}.$$

Thus we get a contradiction with the universality of k.

Exercise 6.3. 1. The decision boundary is obviously the perpendicular bisector of the segment joining α and β, i.e. $\left\langle x - \frac{\alpha+\beta}{2}, \beta - \alpha \right\rangle = 0$. Hence

$$f(x) = \operatorname{sgn}\left(\langle x, \beta - \alpha \rangle - \left\langle \frac{\alpha+\beta}{2}, \beta - \alpha \right\rangle \right).$$

2. The kernelized classifier is

$$f(x) = \begin{cases} -1, & \text{if} \quad \|\phi(x) - \alpha\|_{\mathcal{H}} \le \|\phi(x) - \beta\|_{\mathcal{H}}, \\ 1, & \text{else} \end{cases}$$

where $\|\phi\|_{\mathcal{H}} = \sqrt{\langle \phi, \phi \rangle_{\mathcal{H}}}$ and $\langle \phi, \phi' \rangle_{\mathcal{H}}$ for $\phi, \phi' \in \mathcal{H}$ is induced by k as defined in the lecture Approximation error ($\mathcal{H}$ is defined here too).

Now we show that the classifier can be evaluated at any x without computing $\phi(x)$:

$$\begin{aligned} & \|\phi(x) - \alpha\|_{\mathcal{H}} \le \|\phi(x) - \beta\|_{\mathcal{H}} \\ \iff \quad & -2\langle \phi(x), \alpha \rangle_{\mathcal{H}} + \langle \alpha, \alpha \rangle_{\mathcal{H}} \le -2\langle \phi(x), \beta \rangle_{\mathcal{H}} + \langle \beta, \beta \rangle_{\mathcal{H}}. \end{aligned}$$

We have

$$\langle \phi(x), \alpha \rangle_{\mathcal{H}} = \left\langle \phi(x), \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \right\rangle_{\mathcal{H}} = \frac{1}{m} \sum_{i=1}^{m} \langle \phi(x), \phi(x_i) \rangle_{\mathcal{H}} = \frac{1}{m} \sum_{i=1}^{m} k(x, x_i)$$

$$\langle \alpha, \alpha \rangle_{\mathcal{H}} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j).$$

Similarly we deal with the expressions with β. Thus we can evaluate the classifier computing only k, without evaluating ϕ.