Understanding Machine Learning
Solution Manual

Written by Alon Gonen[*]
Edited by Dana Rubinstein

November 17, 2014

2 Gentle Start

1. Given $S = ((\mathbf{x}_i, y_i))_{i=1}^m$, define the multivariate polynomial

$$p_S(\mathbf{x}) = - \prod_{i \in [m] : y_i = 1} \|\mathbf{x} - \mathbf{x}_i\|^2 \ .$$

   Then, for every i s.t. $y_i = 1$ we have $p_S(\mathbf{x}_i) = 0$, while for every other $\mathbf{x}$ we have $p_S(\mathbf{x}) < 0$.

2. By the linearity of expectation,

$$\begin{aligned}
\mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}[L_S(h)] &= \mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}\left[\frac{1}{m}\sum_{i=1}^m \mathbb{1}_{[h(x_i) \neq f(x_i)]}\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{x_i \sim \mathcal{D}}[\mathbb{1}_{[h(x_i) \neq f(x_i)]}] \\
&= \frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{P}}_{x_i \sim \mathcal{D}}[h(x_i) \neq f(x_i)] \\
&= \frac{1}{m} \cdot m \cdot L_{(\mathcal{D},f)}(h) \\
&= L_{(\mathcal{D},f)}(h) \ .
\end{aligned}$$

[*] The solutions to Chapters 13,14 were written by Shai Shalev-Shwartz

3. (a) First, observe that by definition, A labels positively all the positive instances in the training set. Second, as we assume realizability, and since the tightest rectangle enclosing all positive examples is returned, all the negative instances are labeled correctly by A as well. We conclude that A is an ERM.

   (b) Fix some distribution $\mathcal{D}$ over $\mathcal{X}$, and define $R^\star$ as in the hint. Let f be the hypothesis associated with $R^\star$ a training set S, denote by $R(S)$ the rectangle returned by the proposed algorithm and by $A(S)$ the corresponding hypothesis. The definition of the algorithm A implies that $R(S) \subseteq R^*$ for every S. Thus,

$$L_{(\mathcal{D},f)}(R(S)) = \mathcal{D}(R^\star \setminus R(S)) \ .$$

   Fix some $\epsilon \in (0,1)$. Define R_1, R_2, R_3 and R_4 as in the hint. For each $i \in [4]$, define the event

$$F_i = \{S|_x : S|_x \cap R_i = \emptyset\} \ .$$

   Applying the union bound, we obtain

$$\mathcal{D}^m(\{S : L_{(\mathcal{D},f)}(A(S)) > \epsilon\}) \leq \mathcal{D}^m\left(\bigcup_{i=1}^{4} F_i\right) \leq \sum_{i=1}^{4} \mathcal{D}^m(F_i) \ .$$

   Thus, it suffices to ensure that $\mathcal{D}^m(F_i) \leq \delta/4$ for every i. Fix some $i \in [4]$. Then, the probability that a sample is in F_i is the probability that all of the instances don't fall in R_i, which is exactly $(1-\epsilon/4)^m$. Therefore,

$$\mathcal{D}^m(F_i) = (1-\epsilon/4)^m \leq \exp(-m\epsilon/4) \ ,$$

   and hence,

$$\mathcal{D}^m(\{S : L_{(\mathcal{D},f)}(A(S)) > \epsilon]\}) \leq 4\exp(-m\epsilon/4) \ .$$

   Plugging in the assumption on m, we conclude our proof.

   (c) The hypothesis class of axis aligned rectangles in $\mathbb{R}^d$ is defined as follows. Given real numbers $a_1 \leq b_1, a_2 \leq b_2, \ldots, a_d \leq b_d$, define the classifier $h_{(a_1,b_1,\ldots,a_d,b_d)}$ by

$$h_{(a_1,b_1,\ldots,a_d,b_d)}(x_1,\ldots,x_d) = \begin{cases} 1 & \text{if } \forall i \in [d],\ a_i \leq x_i \leq b_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The class of all axis-aligned rectangles in $\mathbb{R}^d$ is defined as

$$\mathcal{H}^d_{rec} = \{h_{(a_1,b_1,\dots,a_d,b_d)} : \forall i \in [d],\ a_i \le b_i, \}.$$

It can be seen that the same algorithm proposed above is an ERM for this case as well. The sample complexity is analyzed similarly. The only difference is that instead of 4 strips, we have $2d$ strips (2 strips for each dimension). Thus, it suffices to draw a training set of size $\left\lceil \frac{2d\log(2d/\delta)}{\epsilon} \right\rceil$.

(d) For each dimension, the algorithm has to find the minimal and the maximal values among the positive instances in the training sequence. Therefore, its runtime is $O(md)$. Since we have shown that the required value of m is at most $\left\lceil \frac{2d\log(2d/\delta)}{\epsilon} \right\rceil$, it follows that the runtime of the algorithm is indeed polynomial in $d, 1/\epsilon$, and $\log(1/\delta)$.

3 A Formal Learning Model

1. The proofs follow (almost) immediately from the definition. We will show that the sample complexity is monotonically decreasing in the accuracy parameter ϵ. The proof that the sample complexity is monotonically decreasing in the confidence parameter δ is analogous.

   Denote by $\mathcal{D}$ an unknown distribution over $\mathcal{X}$, and let $f \in \mathcal{H}$ be the target hypothesis. Denote by A an algorithm which learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. Fix some $\delta \in (0,1)$. Suppose that $0 < \epsilon_1 \le \epsilon_2 \le 1$. We need to show that $m_1 \overset{\text{def}}{=} m_{\mathcal{H}}(\epsilon_1,\delta) \ge m_{\mathcal{H}}(\epsilon_2,\delta) \overset{\text{def}}{=} m_2$. Given an i.i.d. training sequence of size $m \ge m_1$, we have that with probability at least $1-\delta$, A returns a hypothesis h such that

$$L_{\mathcal{D},f}(h) \le \epsilon_1 \le \epsilon_2\ .$$

   By the minimality of m_2, we conclude that $m_2 \le m_1$.

2. (a) We propose the following algorithm. If a positive instance x_+ appears in S, return the (true) hypothesis h_{x_+}. If S doesn't contain any positive instance, the algorithm returns the all-negative hypothesis. It is clear that this algorithm is an ERM.

   (b) Let $\epsilon \in (0,1)$, and fix the distribution $\mathcal{D}$ over $\mathcal{X}$. If the true hypothesis is h^-, then our algorithm returns a perfect hypothesis.

Assume now that there exists a unique positive instance x_+. It's clear that if x_+ appears in the training sequence S, our algorithm returns a perfect hypothesis. Furthermore, if $\mathcal{D}[\{x_+\}] \le \epsilon$ then in any case, the returned hypothesis has a generalization error of at most ϵ (with probability 1). Thus, it is only left to bound the probability of the case in which $\mathcal{D}[\{x_+\}] > \epsilon$, but x_+ doesn't appear in S. Denote this event by F. Then

$$\mathop{\mathbb{P}}_{S|x \sim \mathcal{D}^m}[F] \le (1-\epsilon)^m \le e^{-m\epsilon} \ .$$

Hence, $\mathcal{H}_{\text{Singleton}}$ is PAC learnable, and its sample complexity is bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \le \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil \ .$$

3. Consider the ERM algorithm A which given a training sequence $S = ((\mathbf{x}_i, y_i))_{i=1}^m$, returns the hypothesis $\hat{h}$ corresponding to the "tightest" circle which contains all the positive instances. Denote the radius of this hypothesis by $\hat{r}$. Assume realizability and let $h^\star$ be a circle with zero generalization error. Denote its radius by $r^\star$.

   Let $\epsilon, \delta \in (0,1)$. Let $\bar{r} \le r^*$ be a scalar s.t. $\mathcal{D}_{\mathcal{X}}(\{x : \bar{r} \le \|\mathbf{x}\| \le r^\star\}) = \epsilon$. Define $E = \{\mathbf{x} \in \mathbb{R}^2 \ : \ \bar{r} \le \|\mathbf{x}\| \le r^\star\}$. The probability (over drawing S) that $L_{\mathcal{D}}(h_S) \ge \epsilon$ is bounded above by the probability that no point in S belongs to E. This probability of this event is bounded above by

$$(1-\epsilon)^m \le e^{-\epsilon m} \ .$$

   The desired bound on the sample complexity follows by requiring that $e^{-\epsilon m} \le \delta$.

4. We first observe that $\mathcal{H}$ is finite. Let us calculate its size accurately. Each hypothesis, besides the all-negative hypothesis, is determined by deciding for each variable x_i, whether x_i, $\bar{x}_i$ or none of which appear in the corresponding conjunction. Thus, $|\mathcal{H}| = 3^d + 1$. We conclude that $\mathcal{H}$ is PAC learnable and its sample complexity can be bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \le \left\lceil \frac{d \log 3 + \log(1/\delta)}{\epsilon} \right\rceil .$$

   Let's describe our learning algorithm. We define $h_0 = x_1 \cap \bar{x}_1 \cap \ldots \cap x_d \cap \bar{x}_d$. Observe that h_0 is the always-minus hypothesis. Let $((\mathbf{a}^1, y^1), \ldots, (\mathbf{a}^m, y^m))$ be an i.i.d. training sequence of size m. Since we cannot produce any information from negative examples, our algorithm neglects them. For each positive example a, we remove from h_i all the literals that are missing in a. That is, if $a_i = 1$, we remove $\bar{x}_i$ from h and if $a_i = 0$, we remove x_i from h_i. Finally, our algorithm returns h_m.

By construction and realizability, h_i labels positively all the positive examples among $\mathbf{a}^1, \ldots, \mathbf{a}^i$. From the same reasons, the set of literals in h_i contains the set of literals in the target hypothesis. Thus, h_i classifies correctly the negative elements among $\mathbf{a}^1, \ldots, \mathbf{a}^i$. This implies that h_m is an ERM.

Since the algorithm takes linear time (in terms of the dimension d) to process each example, the running time is bounded by $O(m \cdot d)$.

5. Fix some $h \in \mathcal{H}$ with $L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon$. By definition,

$$\frac{\mathbb{P}_{X\sim\mathcal{D}_1}[h(X) = f(X)] + \ldots + \mathbb{P}_{X\sim\mathcal{D}_m}[h(X) = f(X))]}{m} < 1 - \epsilon \ .$$

We now bound the probability that h is consistent with S (i.e., that $L_S(h) = 0$) as follows:

$$\begin{aligned}
\underset{S\sim\prod_{i=1}^m \mathcal{D}_i}{\mathbb{P}}[L_S(h) = 0] &= \prod_{i=1}^m \underset{X\sim\mathcal{D}_i}{\mathbb{P}}[h(X) = f(X)] \\
&= \left(\left(\prod_{i=1}^m \underset{X\sim\mathcal{D}_i}{\mathbb{P}}[h(X) = f(X)] \right)^{\frac{1}{m}} \right)^m \\
&\le \left(\frac{\sum_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]}{m} \right)^m \\
&< (1-\epsilon)^m \\
&\le e^{-\epsilon m} \ .
\end{aligned}$$

The first inequality is the geometric-arithmetic mean inequality. Applying the union bound, we conclude that the probability that there exists some $h \in \mathcal{H}$ with $L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon$, which is consistent with S is at most $|\mathcal{H}| \exp(-\epsilon m)$.

6. Suppose that $\mathcal{H}$ is agnostic PAC learnable, and let A be a learning algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot, \cdot)$. We show that $\mathcal{H}$ is PAC learnable using A.

Let $\mathcal{D}, f$ be an (unknown) distribution over $\mathcal{X}$, and the target function respectively. We may assume w.l.o.g. that $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \{0,1\}$, where the conditional probability of y given x is determined deterministically by f. Since we assume realizability, we have $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$. Let $\epsilon, \delta \in (0,1)$. Then, for every positive integer $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, if we equip A with a training set S consisting of m i.i.d. instances which are labeled by f, then with probability at least $1-\delta$ (over the choice of $S|_x$), it returns a hypothesis h with

$$\begin{aligned} L_{\mathcal{D}}(h) &\leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \\ &= 0 + \epsilon \\ &= \epsilon \ . \end{aligned}$$

7. Let $x \in \mathcal{X}$. Let α_x be the conditional probability of a positive label given x. We have

$$\begin{aligned} \mathbb{P}[f_{\mathcal{D}}(X) \neq y | X = x] &= \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot \mathbb{P}[Y = 0 | X = x] + \mathbb{1}_{[\alpha_x < 1/2]} \cdot \mathbb{P}[Y = 1 | X = x] \\ &= \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot (1 - \alpha_x) + \mathbb{1}_{[\alpha_x < 1/2]} \cdot \alpha_x \\ &= \min\{\alpha_x, 1 - \alpha_x\}. \end{aligned}$$

Let g be a classifier[1] from $\mathcal{X}$ to $\{0,1\}$. We have

$$\begin{aligned} \mathbb{P}[g(X) \neq Y | X = x] &= \mathbb{P}[g(X) = 0 | X = x] \cdot \mathbb{P}[Y = 1 | X = x] \\ &+ \mathbb{P}[g(X) = 1 | X = x] \cdot \mathbb{P}[Y = 0 | X = x] \\ &= \mathbb{P}[g(X) = 0 | X = x] \cdot \alpha_x + \mathbb{P}[g(X) = 1 | X = x] \cdot (1 - \alpha_x) \\ &\geq \mathbb{P}[g(X) = 0 | X = x] \cdot \min\{\alpha_x, 1 - \alpha_x\} \\ &+ \mathbb{P}[g(X) = 1 | x] \cdot \min\{\alpha_x, 1 - \alpha_x\} \\ &= \min\{\alpha_x, 1 - \alpha_x\}, \end{aligned}$$

The statement follows now due to the fact that the above is true for every $x \in \mathcal{X}$. More formally, by the law of total expectation,

$$\begin{aligned} L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}_{[f_{\mathcal{D}}(x) \neq y]}] \\ &= \mathbb{E}_{x \sim \mathcal{D}_X}\left[\mathbb{E}_{y \sim \mathcal{D}_{Y|x}}[\mathbb{1}_{[f_{\mathcal{D}}(x) \neq y]} | X = x]\right] \\ &= \mathbb{E}_{x \sim \mathcal{D}_X}[\alpha_x] \\ &\leq \mathbb{E}_{x \sim \mathcal{D}_X}\left[\mathbb{E}_{y \sim \mathcal{D}_{Y|x}}[\mathbb{1}_{[g(x) \neq y]} | X = x]\right] \\ &= L_{\mathcal{D}}(g) \ . \end{aligned}$$

[1] As we shall see, g might be non-deterministic.

8. (a) This was proved in the previous exercise.

   (b) We proved in the previous exercise that for every distribution $\mathcal{D}$, the bayes optimal predictor $f_{\mathcal{D}}$ is optimal w.r.t. $\mathcal{D}$.

   (c) Choose any distribution $\mathcal{D}$. Then A is not better than $f_{\mathcal{D}}$ w.r.t. $\mathcal{D}$.

9. (a) Suppose that $\mathcal{H}$ is PAC learnable in the one-oracle model. Let A be an algorithm which learns $\mathcal{H}$ and denote by $m_{\mathcal{H}}$ the function that determines its sample complexity. We prove that $\mathcal{H}$ is PAC learnable also in the two-oracle model.

   Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0,1\}$. Note that drawing points from the negative and positive oracles with equal provability is equivalent to obtaining i.i.d. examples from a distribution $\mathcal{D}'$ which gives equal probability to positive and negative examples. Formally, for every subset $E \subseteq \mathcal{X}$ we have

$$\mathcal{D}'[E] = \frac{1}{2}\mathcal{D}^+[E] + \frac{1}{2}\mathcal{D}^-[E].$$

   Thus, $\mathcal{D}'[\{x : f(x) = 1\}] = \mathcal{D}'[\{x : f(x) = 0\}] = \frac{1}{2}$. If we let A an access to a training set which is drawn i.i.d. according to $\mathcal{D}'$ with size $m_{\mathcal{H}}(\epsilon/2, \delta)$, then with probability at least $1 - \delta$, A returns h with

$$\begin{aligned}
\epsilon/2 \geq L_{(\mathcal{D}',f)}(h) &= \underset{x\sim\mathcal{D}'}{\mathbb{P}}[h(x) \neq f(x)] \\
&= \underset{x\sim\mathcal{D}'}{\mathbb{P}}[f(x) = 1, h(x) = 0] + \underset{x\sim\mathcal{D}'}{\mathbb{P}}[f(x) = 0, h(x) = 1] \\
&= \underset{x\sim\mathcal{D}'}{\mathbb{P}}[f(x) = 1] \cdot \underset{x\sim\mathcal{D}'}{\mathbb{P}}[h(x) = 0 | f(x) = 1] \\
&+ \underset{x\sim\mathcal{D}'}{\mathbb{P}}[f(x) = 0] \cdot \underset{x\sim\mathcal{D}'}{\mathbb{P}}[h(x) = 1 | f(x) = 0] \\
&= \underset{x\sim\mathcal{D}'}{\mathbb{P}}[f(x) = 1] \cdot \underset{x\sim\mathcal{D}}{\mathbb{P}}[h(x) = 0 | f(x) = 1] \\
&+ \underset{x\sim\mathcal{D}'}{\mathbb{P}}[f(x) = 0] \cdot \underset{x\sim\mathcal{D}}{\mathbb{P}}[h(x) = 1 | f(x) = 0] \\
&= \frac{1}{2} \cdot L_{(\mathcal{D}^+,f)}(h) + \frac{1}{2} \cdot L_{(\mathcal{D}^-,f)}(h).
\end{aligned}$$

   This implies that with probability at least $1 - \delta$, both

$$L_{(\mathcal{D}^+,f)}(h) \leq \epsilon \text{ and } L_{(\mathcal{D}^-,f)}(h) \leq \epsilon.$$

Our definition for PAC learnability in the two-oracle model is satisfied. We can bound both $m^+_{\mathcal{H}}(\epsilon, \delta)$ and $m^-_{\mathcal{H}}(\epsilon, \delta)$ by $m_{\mathcal{H}}(\epsilon/2, \delta)$.

(b) Suppose that $\mathcal{H}$ is PAC learnable in the two-oracle model and let A be an algorithm which learns $\mathcal{H}$. We show that $\mathcal{H}$ is PAC learnable also in the standard model.

Let $\mathcal{D}$ be a distribution over $\mathcal{X}$, and denote the target hypothesis by f. Let $\alpha = \mathcal{D}[\{x : f(x) = 1\}]$. Let $\epsilon, \delta \in (0, 1)$. According to our assumptions, there exist $m^+ \overset{\text{def}}{=} m^+_{\mathcal{H}}(\epsilon, \delta/2), m^- \overset{\text{def}}{=} m^-_{\mathcal{H}}(\epsilon, \delta/2)$ s.t. if we equip A with m^+ examples drawn i.i.d. from $\mathcal{D}^+$ and m^- examples drawn i.i.d. from $\mathcal{D}^-$, then, with probability at least $1 - \delta/2$, A will return h with

$$L_{(\mathcal{D}^+,f)}(h) \leq \epsilon \wedge L_{(\mathcal{D}^-,f)}(h) \leq \epsilon \ .$$

Our algorithm B draws $m = \max\{2m^+/\epsilon, 2m^-/\epsilon, \frac{8\log(4/\delta)}{\epsilon}\}$ samples according to $\mathcal{D}$. If there are less then m^+ positive examples, B returns h^-. Otherwise, if there are less then m^- negative examples, B returns h^+. Otherwise, B runs A on the sample and returns the hypothesis returned by A.

First we observe that if the sample contains m^+ positive instances and m^- negative instances, then the reduction to the two-oracle model works well. More precisely, with probability at least $1 - \delta/2$, A returns h with

$$L_{(\mathcal{D}^+,f)}(h) \leq \epsilon \wedge L_{(\mathcal{D}^-,f)}(h) \leq \epsilon \ .$$

Hence, with probability at least $1 - \delta/2$, the algorithm B returns (the same) h with

$$L_{(\mathcal{D},f)}(h) = \alpha \cdot L_{(\mathcal{D}^+,f)}(h) + (1 - \alpha) \cdot L_{(\mathcal{D}^-,f)}(h) \leq \epsilon$$

We consider now the following cases:

- Assume that both $\alpha \geq \epsilon$. We show that with probability at least $1 - \delta/4$, the sample contain m^+ positive instances. For each $i =\in [m]$, define the indicator random variable Z_i, which gets the value 1 iff the i-th element in the sample is positive. Define $Z = \sum_{i=1}^{m} Z_i$ to be the number of positive examples that were drawn. Clearly, $\mathbb{E}[Z] = \alpha m$. Using Chernoff bound, we obtain

$$\mathbb{P}[Z < (1 - \frac{1}{2})\alpha m] < e^{\frac{-m\alpha}{8}} \ .$$

By the way we chose m, we conclude that

$$\mathbb{P}[Z < m_+] < \delta/4 \ .$$

Similarly, if $1-\alpha \geq \epsilon$, the probability that less than m^- negative examples were drawn is at most $\delta/4$. If both $\alpha \geq \epsilon$ and $1-\alpha \geq \epsilon$, then, by the union bound, with probability at least $1-\delta/2$, the training set contains at least m^+ and m^- positive and negative instances respectively. As we mentioned above, if this is the case, the reduction to the two-oracle model works with probability at least $1-\delta/2$. The desired conclusion follows by applying the union bound.

- Assume that $\alpha < \epsilon$, and less than m^+ positive examples are drawn. In this case, B will return the hypothesis h^-. We obtain
$$L_{\mathcal{D}}(h) = \alpha < \epsilon \ .$$
Similarly, if $(1-\alpha) < \epsilon$, and less than m^- negative examples are drawn, B will return h^+. In this case,
$$L_{\mathcal{D}}(h) = 1-\alpha < \epsilon \ .$$

All in all, we have shown that with probability at least $1-\delta$, B returns a hypothesis h with $L_{(\mathcal{D},f)}(h) < \epsilon$. This satisfies our definition for PAC learnability in the one-oracle model.

4 Learning via Uniform Convergence

1. (a) Assume that for every $\epsilon, \delta \in (0,1)$, and every distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, there exists $m(\epsilon,\delta) \in \mathbb{N}$ such that for every $m \geq m(\epsilon,\delta)$,
$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(A(S)) > \epsilon] < \delta \ .$$
Let $\lambda > 0$. We need to show that there exists $m_0 \in \mathbb{N}$ such that for every $m \geq m_0$, $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \lambda$. Let $\epsilon = \min(1/2, \lambda/2)$. Set $m_0 = m_{\mathcal{H}}(\epsilon,\epsilon)$. For every $m \geq m_0$, since the loss is bounded

above by 1, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \lambda/2]\cdot 1 + \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \leq \lambda/2]\cdot\lambda/2 \\ &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] + \lambda/2 \\ &\leq \epsilon + \lambda/2 \\ &\leq \lambda/2 + \lambda/2 \\ &= \lambda\ .\end{aligned}$$

(b) Assume now that

$$\lim_{m\to\infty}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0\ .$$

Let $\epsilon,\delta\in(0,1)$. There exists some $m_0\in\mathbb{N}$ such that for every $m\geq m_0$, $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))]\leq\epsilon\cdot\delta$. By Markov's inequality,

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] &\leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))]}{\epsilon} \\ &\leq \frac{\epsilon\delta}{\epsilon} \\ &= \delta\ .\end{aligned}$$

2. The left inequality follows from Corollary 4.4. We prove the right inequality. Fix some $h\in\mathcal{H}$. Applying Hoeffding's inequality, we obtain

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[|L_{\mathcal{D}}(h) - L_S(h)| \geq \epsilon/2] \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)\ . \tag{2}$$

The desired inequality is obtained by requiring that the right-hand side of Equation (2) is at most $\delta/|\mathcal{H}|$, and then applying the union bound.

5 The Bias-Complexity Tradeoff

1. We simply follow the hint. By Lemma B.1,

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] &= \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1 - 7/8] \\ &\geq \frac{\mathbb{E}[L_{\mathcal{D}}(A(S))] - (1-7/8)}{7/8} \\ &\geq \frac{1/8}{7/8} \\ &= 1/7\ .\end{aligned}$$

2. We will provide a qualitative answer. In the next section, we'll be able to quantify our argument.

   Denote by $\mathcal{H}_d$ the class of axis-aligned rectangles in $\mathbb{R}^d$. Since $\mathcal{H}_5 \supseteq \mathcal{H}_2$, the approximation error of the class $\mathcal{H}_5$ is smaller. However, the complexity of $\mathcal{H}_5$ is larger, so we expect its estimation error to be larger. So, if we have a limited amount of training examples, we would prefer to learn the smaller class.

3. We modify the proof of the NFL theorem, based on the assumption that $|\mathcal{X}| \geq km$ (while referring to some equations from the book). Choose C to be a set with size km. Then, Equation (5.5) (in the book) becomes

$$L_{\mathcal{D}_i}(h) = \frac{1}{km} \sum_{x \in C} \mathbb{1}_{[h(x) \neq f_i(x)]} \geq \frac{1}{km} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \geq \frac{k-1}{pk} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \, .$$

   Consequently, the right-hand side of Equation (5.6) becomes

$$\frac{k-1}{k} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]}$$

   It follows that

$$\mathbb{E}[L_{\mathcal{D}}(A(S))] \geq \frac{k-1}{2k} \, .$$

6 The VC-dimension

1. Let $\mathcal{H}' \subseteq \mathcal{H}$ be two hypothesis classes for binary classification. Since $\mathcal{H}' \subseteq \mathcal{H}$, then for every $C = \{c_1, \ldots, c_m\} \subseteq \mathcal{X}$, we have $\mathcal{H}'_C \subseteq \mathcal{H}_C$. In particular, if C is shattered by $\mathcal{H}'$, then C is shattered by $\mathcal{H}$ as well. Thus, $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$.

2. (a) We claim that $\text{VCdim}(\mathcal{H}_{=k}) = \min\{k, |\mathcal{X}| - k\}$. First, we show that $\text{VCdim}(\mathcal{H}_{=k}) \leq k$. Let $C \subseteq \mathcal{X}$ be a set of size $k+1$. Then, there doesn't exist $h \in \mathcal{H}_{=k}$ which satisfies $h(x) = 1$ for all $x \in C$. Analogously, if $C \subseteq \mathcal{X}$ is a set of size $|\mathcal{X}| - k + 1$, there is no $h \in \mathcal{H}_{=k}$ which satisfies $h(x) = 0$ for all $x \in C$. Hence, $\text{VCdim}(\mathcal{H}_{=k}) \leq \min\{k, |\mathcal{X}| - k\}$.

   It's left to show that $\text{VCdim}(\mathcal{H}_{=k}) \geq \min\{k, |\mathcal{X}| - k\}$. Let $C = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ be a set with of size $m \leq \min\{k, |\mathcal{X}| - k\}$. Let $(y_1, \ldots, y_m) \in \{0,1\}^m$ be a vector of labels. Denote $\sum_{i=1}^m y_i$

by s. Pick an arbitrary subset $E \subseteq \mathcal{X} \setminus C$ of $k - s$ elements, and let $h \in \mathcal{H}_{=k}$ be the hypothesis which satisfies $h(x_i) = y_i$ for every $x_i \in C$, and $h(x) = \mathbb{1}_{[E]}$ for every $x \in \mathcal{X} \setminus C$. We conclude that C is shattered by $\mathcal{H}_{=k}$. It follows that $\text{VCdim}(\mathcal{H}_{=k}) \geq \min\{k, |\mathcal{X}| - k\}$.

(b) We claim that $\text{VCdim}(\mathcal{H}_{\leq k}) = k$. First, we show that $\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{\leq k}) \leq k$. Let $C \subseteq \mathcal{X}$ be a set of size $k+1$. Then, there doesn't exist $h \in \mathcal{H}_{\leq k}$ which satisfies $h(x) = 1$ for all $x \in C$.

It's left to show that $\text{VCdim}(\mathcal{H}_{\leq k}) \geq k$. Let $C = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ be a set with of size $m \leq k$. Let $(y_1, \ldots, y_m) \in \{0,1\}^m$ be a vector of labels. This labeling is obtained by some hypothesis $h \in \mathcal{H}_{\leq k}$ which satisfies $h(x_i) = y_i$ for every $x_i \in C$, and $h(x) = 0$ for every $x \in \mathcal{X} \setminus C$. We conclude that C is shattered by $\mathcal{H}_{\leq k}$. It follows that $\text{VCdim}(\mathcal{H}_{\leq k}) \geq k$.

3. We claim that VC-dimension of $\mathcal{H}_{n\text{-parity}}$ is n. First, we note that $|\mathcal{H}_{n\text{-parity}}| = 2^n$. Thus,

$$\text{VCdim}(\mathcal{H}_{n\text{-parity}}) \leq \log(|\mathcal{H}_{n\text{-parity}}|) = n \ .$$

We will conclude the tightness of this bound by showing that the standard basis $\{\mathbf{e}_j\}_{j=1}^n$ is shattered by $\mathcal{H}_{n\text{-parity}}$. Given a vector of labels $(y_1, \ldots, y_n) \in \{0,1\}^n$, let $J = \{j \in [n] \ : \ y_j = 1\}$. Then $h_J(e_j) = y_j$ for every $j \in [n]$.

4. Let $\mathcal{X} = \mathbb{R}^d$. We will demonstrate all the 4 combinations using hypothesis classes defined over $\mathcal{X} \times \{0,1\}$. Remember that the empty set is always considered to be shattered.

- $(<, =)$: Let $d \geq 2$ and consider the class $\mathcal{H} = \{\mathbb{1}_{[\|x\|_2 \leq r]} : r \geq 0\}$ of concentric balls. The VC-dimension of this class is 1. To see this, we first observe that if $\mathbf{x} \neq (0, \ldots, 0)$, then $\{\mathbf{x}\}$ is shattered. Second, if $\|\mathbf{x}_1\|_2 \leq \|\mathbf{x}_2\|_2$, then the labeling $y_1 = 0, y_2 = 1$ is not obtained by any hypothesis in $\mathcal{H}$. Let $A = \{\mathbf{e}_1, \mathbf{e}_2\}$, where $\mathbf{e}_1, \mathbf{e}_2$ are the first two elements of the standard basis of $\mathbb{R}^d$. Then, $\mathcal{H}_A = \{(0,0), (1,1)\}$, $\{B \subseteq A \,:\, \mathcal{H} \text{ shatters } B\} = \{\emptyset, \{\mathbf{e}_1\}, \{\mathbf{e}_2\}\}$, and $\sum_{i=0}^{d} \binom{|A|}{i} = 3$.
- $(=, <)$: Let $\mathcal{H}$ be the class of axis-aligned rectangles in $\mathbb{R}^2$. We have seen that the VC-dimension of $\mathcal{H}$ is 4. Let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = (0,0), \mathbf{x}_2 = (1,0), \mathbf{x}_3 = (2,0)$. All the labelings except $(1,0,1)$ are obtained. Thus, $|\mathcal{H}_A| = 7$, $|\{B \subseteq A \,:\, \mathcal{H} \text{ shatters } B\}| = 7$, and $\sum_{i=0}^{d} \binom{|A|}{i} = 8$.

- $(<,<)$: Let $d \geq 3$ and consider the class $\mathcal{H} = \{\text{sign}\langle w, x\rangle : w \in \mathbb{R}^d\}$[2] of homogenous halfspaces (see Chapter 9). We will prove in Theorem 9.2 that the VC-dimension of this class is d. However, here we will only rely on the fact that $\text{VCdim}(\mathcal{H}) \geq 3$. This fact follows by observing that the set $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is shattered. Let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = \mathbf{e}_1, \mathbf{x}_2 = \mathbf{e}_2$, and $\mathbf{x}_3 = (1, 1, 0, \ldots, 0)$. Note that all the labelings except $(1, 1, -1)$ and $(-1, -1, 1)$ are obtained. It follows that $|\mathcal{H}_A| = 6$, $|\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| = 7$, and $\sum_{i=0}^{d} \binom{|A|}{i} = 8$.
- $(=,=)$: Let $d = 1$, and consider the class $\mathcal{H} = \{\mathbb{1}_{[x \geq t]} : t \in \mathbb{R}\}$ of thresholds on the line. We have seen that every singleton is shattered by $\mathcal{H}$, and that every set of size at least 2 is not shattered by $\mathcal{H}$. Choose any finite set $A \subseteq \mathbb{R}$. Then each of the three terms in "Sauer's inequality" equals $|A| + 1$.

5. Our proof is a straightforward generalization of the proof in the 2-dimensional case.

   Let us first define the class formally. Given real numbers $a_1 \leq b_1, a_2 \leq b_2, \ldots, a_d \leq b_d$, define the classifier $h_{(a_1,b_1,\ldots,a_d,b_d)}$ by $h_{(a_1,b_1,\ldots,a_d,b_d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} \mathbb{1}_{[x_i \in [a_i, b_i]]}$. The class of all axis-aligned rectangles in $\mathbb{R}^d$ is defined as $\mathcal{H}_{rec}^d = \{h_{(a_1,b_1,\ldots,a_d,b_d)} : \forall i \in [d],\ a_i \leq b_i, \}$.

   Consider the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_{2d}\}$, where $\mathbf{x}_i = \mathbf{e}_i$ if $i \in [d]$, and $\mathbf{x}_i = -e_{i-d}$ if $i > d$. As in the 2-dimensional case, it's not hard to see that it's shattered. Indeed, let $(y_1, \ldots, y_{2d}) \in \{0, 1\}^{2d}$. Choose $a_i = -2$ if $y_{i+d} = 1$, and $a_i = 0$ otherwise. Similarly, choose $b_i = 2$ if $y_i = 1$, and $b_i = 0$ otherwise. Then $h_{a_1,b_1,\ldots,a_d,b_d}(\mathbf{x}_i) = y_i$ for every $i \in [2d]$. We just proved that $\text{VCdim}(\mathcal{H}_{rec}^d) \geq 2d$.

   Let C be a set of size at least $2d + 1$. We finish our proof by showing that C is not shattered. By the pigeonhole principle, there exists an element $\mathbf{x} \in C$, s.t. for every $j \in [d]$, there exists $\mathbf{x}' \in C$ with $x_j' \leq x_j$, and similarly there exists $\mathbf{x}'' \in C$ with $x_j'' \geq x_j$. Thus the labeling in which x is negative, and the rest of the elements in C are positive can not be obtained.

6. (a) Each hypothesis, besides the all-negative hypothesis, is determined by deciding for each variable x_i, whether x_i, $\bar{x}_i$ or none of which appear in the corresponding conjunction. Thus, $|\mathcal{H}_{con}^d| = 3^d + 1$.

[2] We adopt the convention $\text{sign}(0) = 1$.

(b)

$$\text{VCdim}(\mathcal{H}^d_{con}) \leq \lfloor \log(|\mathcal{H}^d_{con}|) \rfloor \leq 3 \log d \ .$$

(c) We prove that $\mathcal{H}^d_{con} \geq d$ by showing that the set $C = \{\mathbf{e}_j\}_{j=1}^d$ is shattered by $\mathcal{H}^d_{con}$. Let $J \subseteq [d]$ be a subset of indices. We will show that the labeling in which exactly the elements $\{\mathbf{e}_j\}_{j \in J}$ are positive is obtained. If $J = [d]$, pick the all-positive hypothesis h_{empty}. If $J = \emptyset$, pick the all-negative hypothesis $x_1 \wedge \bar{x}_1$. Assume now that $\emptyset \subsetneq J \subsetneq [d]$. Let h be the hypothesis which corresponds to the boolean conjunction $\bigwedge_{j \in J} x_j$. Then, $h(\mathbf{e}_j) = 1$ if $j \in J$, and $h(\mathbf{e}_j) = 0$ otherwise.

(d) Assume by contradiction that there exists a set $C = \{c_1, \ldots, c_{d+1}\}$ for which $|\mathcal{H}_C| = 2^{d+1}$. Define $h_1, \ldots, h_{d+1}$, and $\ell_1, \ldots, \ell_{d+1}$ as in the hint. By the Pigeonhole principle, among $\ell_1, \ldots, \ell_{d+1}$, at least one variable occurs twice. Assume w.l.o.g. that ℓ_1 and ℓ_2 correspond to the same variable. Assume first that $\ell_1 = \ell_2$. Then, ℓ_1 is true on c_1 since ℓ_2 is true on c_1. However, this contradicts our assumptions. Assume now that $\ell_1 \neq \ell_2$. In this case $h_1(c_3)$ is negative, since ℓ_2 is positive on c_3. This again contradicts our assumptions.

(e) First, we observe that $|\mathcal{H}'| = 2^d + 1$. Thus,

$$\text{VCdim}(\mathcal{H}') \leq \lfloor \log(|\mathcal{H}|) \rfloor = d \ .$$

We will complete the proof by exhibiting a shattered set with size d. Let

$$C = \{(1, 1, \ldots, 1) - e_j\}_{j=1}^d = \{(0, 1, \ldots, 1), \ldots, (1, \ldots, 1, 0)\} \ .$$

Let $J \subseteq [d]$ be a subset of indices. We will show that the labeling in which exactly the elements $\{(1, 1, \ldots, 1) - e_j\}_{j \in J}$ are negative is obtained. Assume for the moment that $J \neq \emptyset$. Then the labeling is obtained by the boolean conjunction $\bigwedge_{j \in J} x_j$. Finally, if $J = \emptyset$, pick the all-positive hypothesis $h_\emptyset$.

7. (a) The class $\mathcal{H} = \{\mathbb{1}_{[x \geq t]} : t \in \mathbb{R}\}$ (defined over any non-empty subset of $\mathbb{R}$) of thresholds on the line is infinite, while its VC-dimension equals 1.

(b) The hypothesis class $\mathcal{H} = \{\mathbb{1}_{[x \leq 1]}, \mathbb{1}_{[x \leq 1/2]}\}$ satisfies the requirements.

8. We will begin by proving the lemma provided in the question:

$$\begin{aligned}
sin(2^m\pi x) &= \sin(2^m\pi(0.x_1x_2\ldots)) \\
&= \sin(2\pi(x_1x_2\ldots x_{m-1}.x_mx_{m+1}\ldots)) \\
&= \sin\left(2\pi(x_1x_2\ldots x_{m-1}.x_mx_{m+1}\ldots) - 2\pi(x_1x_2\ldots x_{m-1}.0)\right) \\
&= \sin(2\pi(0.x_mx_{m+1}\ldots).
\end{aligned}$$

Now, if $x_m = 0$, then $2\pi(0.x_mx_{m+1}\ldots) \in (0,\pi)$ (we use here the fact that $\exists k \geq m$ s.t. $x_k = 1$) , so the expression above is positive, and $\lceil sin(2^m\pi x)\rceil = 1$. On the other hand, if $x_m = 1$, then $2\pi(0.x_mx_{m+1}) \in [\pi, 2\pi)$, so the expression above is non-positive, and $\lceil sin(2^m\pi x)\rceil = 0$. In conclusion, we have that $\lceil sin(2^m\pi x)\rceil = 1 - x_m$.

To prove $VC(\mathcal{H}) = \infty$, we need to pick n points which are shattered by $\mathcal{H}$, for any n. To do so, we construct n points $x_1, \ldots, x_n \in [0,1]$, such that the set of the m-th bits in the binary expansion, as m ranges from 1 to 2^n, ranges over all possible labelings of $x_1, \ldots, x_n$:

$$\begin{array}{rcccccccc}
x_1 & = 0\,. & 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\
x_2 & = 0\,. & 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\
 & & & & \vdots & & & & \\
x_{n-1} & = 0\,. & 0 & 0 & 1 & 1 & \cdots & 1 & 1 \\
x_n & = 0\,. & 0 & 1 & 0 & 1 & \cdots & 0 & 1
\end{array}$$

For example, to give the labeling 1 for all instances, we just pick $h(x) = \lceil sin(2^1x)\rceil$, which returns the first bit (column) in the binary expansion. If we wish to give the labeling 1 for $x_1, \ldots, x_{n-1}$, and the labeling 0 for x_n, we pick $h(x) = \lceil sin(2^2x)\rceil$, which returns the 2nd bit in the binary expansion, and so on. We conclude that $x_1, \ldots, x_n$ can be given any labeling by some $h \in \mathcal{H}$, so it is shattered. This can be done for any n, so $\text{VCdim}(\mathcal{H}) = \infty$.

9. We prove that $\text{VCdim}(\mathcal{H}) = 3$. Choose $C = \{1,2,3\}$. The following table shows that C is shattered by $\mathcal{H}$.

1	2	3	a	b	s
-	-	-	0.5	3.5	-1
-	-	+	2.5	3.5	1
-	+	-	1.5	2.5	1
-	+	+	1.5	3.5	1
+	-	-	0.5	1.5	1
+	-	+	1.5	2.5	-1
+	+	-	0.5	2.5	1
+	+	+	0.5	3.5	1

We conclude that $\text{VCdim}(\mathcal{H}) \geq 3$. Let $C = \{x_1, x_2, x_3, x_4\}$ and assume w.l.o.g. that $x_1 < x_2 < x_3 < x_4$ Then, the labeling $y_1 = y_3 = -1$, $y_2 = y_4 = 1$ is not obtained by any hypothesis in $\mathcal{H}$. Thus, $\text{VCdim}(\mathcal{H}) \leq 3$.

10. (a) We may assume that $m < d$, since otherwise the statement is meaningless. Let C be a shattered set of size d. We may assume w.l.o.g. that $\mathcal{X} = C$ (since we can always choose distributions which are concentrated on C). Note that $\mathcal{H}$ contains all the functions from C to $\{0, 1\}$. According to Exercise 3 in Section 5, for every algorithm, there exists a distribution $\mathcal{D}$, for which $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$, but $\mathbb{E}[L_{\mathcal{D}}(A(S))] \geq \frac{k-1}{2k} \underbrace{=}_{k=\frac{d}{m}} \frac{d-m}{2d}$.

    (b) Assume that $\text{VCdim}(\mathcal{H}) = \infty$. Let $\mathcal{A}$ be a learning algorithm. We show that A fails to (PAC) learn $\mathcal{H}$. Choose $\epsilon = \frac{1}{16}$, $\delta = \frac{1}{14}$. For any $m \in \mathbb{N}$, there exists a shattered set of size $d = 2m$. Applying the above, we obtain that there exists a distribution $\mathcal{D}$ for which $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$, but $\mathbb{E}[L_{\mathcal{D}}(A(S))] \geq 1/4$. Applying Lemma B.1 yields that with probability at least $1/7 > \delta$, $L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = L_{\mathcal{D}}(A(S) \geq 1/8 > \epsilon$.

11. (a) We may assume w.l.o.g. that for each $i \in [r]$, $\text{VCdim}(\mathcal{H}_i) = d \geq 3$. Let $\mathcal{H} = \bigcup_{i=1}^{r} \mathcal{H}_i$. Let $k \in [d]$, such that $\tau_{\mathcal{H}}(k) = 2^k$. We will show that $k \leq 4d \log(2d) + 2 \log r$.

    By definition of the growth function, we have

    $$\tau_{\mathcal{H}}(k) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(k) \ .$$

    Since $d \geq 3$, by applying Sauer's lemma on each of the terms $\tau_{\mathcal{H}_i}$, we obtain

    $$\tau_{\mathcal{H}}(k) < rm^d \ .$$

It follows that $k < d \log m + \log r$. Lemma A.2 implies that $k < 4d \log(2d) + 2 \log r$.

(b) A direct application of the result above yields a weaker bound. We need to employ a more careful analysis.

As before, we may assume w.l.o.g. that $\text{VCdim}(\mathcal{H}_1) = \text{VCdim}(\mathcal{H}_2) = d$. Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. Let k be a positive integer such that $k \geq 2d + 2$. We show that $\tau_{\mathcal{H}}(k) < 2^k$. By Sauer's lemma,

$$\begin{aligned}
\tau_{\mathcal{H}}(k) &\leq \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k) \\
&\leq \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{i} \\
&= \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{k-i} \\
&= \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=k-d}^{k} \binom{k}{i} \\
&\leq \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+2}^{k} \binom{k}{i} \\
&< \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+1}^{k} \binom{k}{i} \\
&= \sum_{i=0}^{k} \binom{k}{i} \\
&= 2^k \ .
\end{aligned}$$

12. Throughout this question, we use the convention $\text{sign}(0) = -1$.

(a) It suffices to show that a set $C \subseteq \mathcal{X}$ is shattered by $POS(\mathcal{F})$ if and only if it's shattered by $POS(\mathcal{F} + g)$.

Assume that a set $C = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of size m is shattered by $POS(\mathcal{F})$. Let $(y_1, \ldots, y_m) \in \{-1, 1\}^m$. We first argue that there exists f, for which $f(\mathbf{x}_i) > 0$ if $y_i = 1$, and $f(\mathbf{x}_i) < 0$ if $y_i = -1$. To see this, choose a function $\phi \in \mathcal{F}$ such that for every $i \in [m]$, $\phi(\mathbf{x}_i) > 0$ if and only if $y_i = 1$. Similarly, choose a function $\psi \in \mathcal{F}$, such that for every $i \in [m]$, $\psi(\mathbf{x}_i) > 0$ if and only if $y_i = -1$. Than $f = \phi - \psi$ (which belongs to $\mathcal{F}$ since $\mathcal{F}$ is linearly closed) satisfies the requirements. Next, since $\mathcal{F}$ is closed under multiplication by (positive) scalars, we may assume w.l.o.g. that $|f(\mathbf{x}_i)| \geq |g(\mathbf{x}_i)|$ for every $i \in [m]$. Consequently, for every $i \in [m]$, $\text{sign}((f+g)(\mathbf{x}_i)) = \text{sign}(f(\mathbf{x}_i)) = y_i$. We conclude that C is shattered by $POS(\mathcal{F}+g)$.

Assume now that a set $C = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ of size m is shattered by $POS(\mathcal{F}+g)$. Choose $\phi \in \mathcal{F}$ such that for every $i \in [m]$, $\text{sign}((\phi+g)(\mathbf{x}_i)) = y_i$. Analogously, find $\psi \in \mathcal{F}$ such that for every $i \in [m]$, $\text{sign}((\psi+g)(\mathbf{x}_i)) = -y_i$. Let $f = \phi - \psi$. Note that $f \in \mathcal{F}$. Using the identity $\phi - \psi = (\phi+g) - (\psi+g)$, we conclude that for every $i \in [m]$, $\text{sign}(f(\mathbf{x}_i)) = y_i$. Hence, C is shattered by $POS(\mathcal{F})$.

(b) The result will follow from the following claim:

Claim 6.1. *For any positive integer d, $\tau_{POS(\mathcal{F})}(d) = 2^d$ (i.e., there exists a shattered set of size d) if and only if there exists an independent subset $\mathcal{G} \subseteq \mathcal{F}$ of size d.*

The lemma implies that the VC-dimension of $POS(\mathcal{F})$ equals to the dimension of $\mathcal{F}$. Note that both might equal ∞. Let us prove the claim.

Proof. Let $\{f_1, \dots, f_d\} \in \mathcal{F}$ be a linearly independent set of size d. Following the hint, for each $\mathbf{x} \in \mathbb{R}^n$, let $\phi(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_d(\mathbf{x}))$. We use induction to prove that there exist $\mathbf{x}_1, \dots, \mathbf{x}_d$ such that the set $\{\phi(\mathbf{x}_j) : j \in [d]\}$ is an independent subset of $\mathbb{R}^d$, and thus spans $\mathbb{R}^d$. The case $d = 1$ is trivial. Assume now that the argument holds for any integer $k \in [d-1]$. Apply the induction hypothesis to choose $\mathbf{x}_1, \dots, \mathbf{x}_{d-1}$ such that the set $\{(f_1(\mathbf{x}_j), \dots, f_{d-1}(\mathbf{x}_j)) : j \in [d-1]\}$ is independent. If there doesn't exist $\mathbf{x}_d \notin \{\mathbf{x}_j : j \in [d-1]\}$ such that $\{\phi(\mathbf{x}_j) : j \in [d]\}$ is independent, then f_d depends on $f_1, \dots, f_{d-1}$. This would contradict our earlier assumption.

Let $\{\phi(\mathbf{x}_j) : j \in [d]\}$ be such an independent set. For each $j \in [d]$, let $\mathbf{v}_j = \phi(\mathbf{x}_j)$. Then, since $\mathcal{F}$ is linearly closed,

$$[POS(\mathcal{F})]_{\{\mathbf{x}_1,\dots,\mathbf{x}_d\}} \supseteq \{(\text{sign}(\langle \mathbf{w}, \mathbf{v}_1 \rangle), \dots, \text{sign}(\langle \mathbf{w}, v_d \rangle)) : \mathbf{w} \in \mathbb{R}^d\}$$

Since we can replace $\mathbf{w}$ by $V^\top \mathbf{w}$, where V is the $d \times d$ matrix whose j-th column is $\mathbf{v}_j$, we may assume w.l.o.g. that $\{\mathbf{v}_1, \dots, \mathbf{v}_d\}$ is the standard basis, i.e. $\mathbf{v}_j = \mathbf{e}_j$ for every $j \in [d]$ (simply note that $\langle V^\top \mathbf{w}, \mathbf{e}_j \rangle = \langle \mathbf{w}, \mathbf{v}_j \rangle$). Since the standard basis is shattered by the hypothesis set of homogenous halfspaces (Theorem 9.2), we conclude that $|[POS(\mathcal{F})]_{\{\mathbf{x}_1,\ldots,\mathbf{x}_d\}}| = 2^d$, i.e. $\tau_{POS(\mathcal{F})}(d) = 2^d$. The other direction is analogous. Assume that there doesn't exist an independent subset $\mathcal{G} \subseteq \mathcal{F}$ of size d. Hence, $\mathcal{F}$ has a basis $\{f_j : j \in [k]\}$, for some $k < d$. For each $\mathbf{x} \in \mathbb{R}^n$, let $\phi(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))$. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\} \subseteq \mathbb{R}^n$ be a set of size d. We note that

$$[POS(\mathcal{F})]_{\{\mathbf{x}_1,\ldots,\mathbf{x}_d\}} \subseteq \{(\text{sign}(\langle \mathbf{w}, \phi(\mathbf{x}_1)\rangle), \ldots, \text{sign}(\langle \mathbf{w}, \phi(\mathbf{x}_d)\rangle)) : \mathbf{w} \in \mathbb{R}^d\}$$

The set $\{\phi(\mathbf{x}_j) : j \in [d]\}$ is a linearly dependent subset of $\mathbb{R}^k$. Hence, by (the opposite direction of) Theorem 9.2, the class of halfspaces doesn't shatter it. It follows that $|[POS(\mathcal{F})]_{\{\mathbf{x}_1,\ldots,\mathbf{x}_d\}}| < 2^d$, and hence $\tau_{POS(\mathcal{F})}(d) < 2^d$. □

(c) i. Let $\mathcal{F} = \{f_{\mathbf{w},b} : \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$, where $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$. We note that $HS_n = POS(\mathcal{F})$.

ii. Let $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^n\}$, where $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$. We note that $HHS_n = POS(\mathcal{F})$.

iii. Let $\mathcal{B}_n = \{h_{\mathbf{a},r} : \mathbf{a} \in \mathbb{R}^n, r \in \mathbb{R}_+\}$ be the class of open balls in $\mathbb{R}^n$. That is, $h_{\mathbf{a},r}(\mathbf{x}) = 1$ if and only if $\|\mathbf{x} - \mathbf{a}\| < r$.

We first find a dudley class representation for the class. Observe that for every $\mathbf{a} \in \mathbb{R}^n$, $r \in \mathbb{R}_+$, $h_{\mathbf{a},r}(\mathbf{x}) = POS(s_{\mathbf{a},r}(\mathbf{x}))$, where $s_{\mathbf{a},r}(\mathbf{x}) = 2\langle \mathbf{a}, \mathbf{x} \rangle - \|\mathbf{a}\|^2 + r^2 - \|\mathbf{x}\|^2$. Let $\mathcal{F}$ be the vector space of functions of the form $f_{\mathbf{a},q}(\mathbf{x}) = 2\langle \mathbf{a}, \mathbf{x} \rangle + q$, where $\mathbf{a} \in \mathbb{R}^n$, and $q \in \mathbb{R}$. Let $g(\mathbf{x}) = -\|\mathbf{x}\|^2$. The space $\mathcal{F}$ is spanned by the set $\{f_{\mathbf{e}_j} : j \in [n+1]\}$ (here we use the standard basis of $\mathbb{R}^{n+1}$). Hence, $\mathcal{F}$ is a vector space with dimensionality $n+1$. Furthermore, $POS(\mathcal{F} + g) \supseteq \mathcal{B}_n$. Following the results from the previous sections, we conclude that $\text{VCdim}(\mathcal{B}_n) \leq n + 1$.

Next, we prove that the set $F = \{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_n\}$ (where $\mathbf{0}$ denotes the zero vector, and $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the standard basis) is shattered. Let $E \subseteq F$. If E is empty, then pick $r = 0$ to obtain the corresponding labeling. Assume that $E \neq \emptyset$. Let $\mathbf{a} = \sum_{j:\mathbf{e}_j \in E} \mathbf{e}_j$. Note that $\|\mathbf{a} - \mathbf{0}\| = |E|$, and

$$\|\mathbf{e}_j - \mathbf{a}\|_2 = \begin{cases} |E| - 1 & \mathbf{e}_j \in E \\ |E| + 1 & \mathbf{e}_j \notin E \end{cases}$$

Now, if $\mathbf{0} \in E$, set $r = |E| + 1/2$, and otherwise, set $r = |E| - 1/2$. It follows that $h_{\mathbf{a},r}(\mathbf{x}) = \mathbb{1}_{[x \in E]}$. Hence, F is indeed shattered. We conclude that $\text{VCdim}(\mathcal{B}_n) \geq n + 1$.

iv. A. Let $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^{d+1}\}$, where $f_{\mathbf{w}}(x) = \sum_{j=1}^{d+1} w_j x^{j-1}$. Then $P_1^d = POS(\mathcal{F})$. The vector space $\mathcal{F}$ is spanned by the (functions corresponding to the) standard basis of $\mathbb{R}^{d+1}$. Hence $\dim(\mathcal{F}) = d + 1$, and $\text{VCdim}(P_1^d) = d + 1$.

B. We need to introduce some notation. For each $n \in \mathbb{N}$, let $\mathcal{X}_n = \mathbb{R}$. Denote the product $\prod_{n=1}^{\infty} \mathcal{X}_n$ by $\mathbb{R}^{\infty}$. Denote by $\mathcal{X}$ the subset of $\mathbb{R}^{\infty}$ containing all sequences with finitely many non-zero elements. Let $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathcal{X}\}$, where $f_{\mathbf{w}}(\mathbf{x}) = \sum_{n=1}^{\infty} w_n x^{n-1}$. A basis for this space is the set of functions corresponding to the infinite standard basis of $\mathcal{X}$ (i.e., the set $\{(1, 0, \ldots), (0, 1, 0, \ldots), \ldots\}$). Note that the set $\bigcup_{d \in \mathbb{N}} P_1^d$ of all polynomial classifiers obeys $\bigcup_{d \in \mathbb{N}} P_1^d = POS(\mathcal{F})$. Hence, we have $\text{VCdim}(\bigcup_{d \in \mathbb{N}} P_1^d) = \infty$.

C. Recall that the number of monomials of degree k in n variables equals to the number of ways to choose k elements from a set of size n, while repetition is allowed, but order doesn't matter. This number equals $\binom{n+k-1}{k}$, and it's denoted by $\left(\!\binom{n}{k}\!\right)$. For $\mathbf{x} \in \mathbb{R}^n$, we denote by $\psi(\mathbf{x})$ the vector with all monomials of degree at most d associated with $\mathbf{x}$.

Given $n, d \in \mathbb{N}$, let $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^N\}$, where $N = \sum_{k=0}^{d} \left(\!\binom{n}{k}\!\right)$, and $f_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{N} w_i (\psi(\mathbf{x}))_i$. The vector space $\mathcal{F}$ is spanned by the (functions corresponding to the) standard basis of $\mathbb{R}^N$. We note that $P_n^d = POS(\mathcal{F})$. Therefore, $\text{VCdim}(P_n^d) = N$.

7 Non-uniform Learnability

1. (a) Let $n = \max_{h \in \mathcal{H}}\{|d(h)|\}$. Since each $h \in \mathcal{H}$ has a unique description, we have

$$|\mathcal{H}| \leq \sum_{i=0}^{n} 2^i = 2^{n+1} - 1 \ .$$

It follows that $\text{VCdim}(\mathcal{H}) \leq \lfloor \log |\mathcal{H}| \rfloor \leq n + 1 \leq 2n$.

(b) Let $n = \max_{h \in \mathcal{H}}\{|d(h)|\}$. For $\mathbf{a}, \mathbf{b} \in \bigcup_{k=0}^{n}\{0,1\}^k$, we say that $\mathbf{a} \sim \mathbf{b}$ if a is a prefix of b or b is a prefix of a. It can be seen that $\sim$ is an equivalence relation. If d is prefix-free, then the size of $\mathcal{H}$ is bounded above by the number of equivalence classes. Note that there is a one-to-one mapping from the set of equivalence classes to $\{0,1\}^n$ (simply by padding with zeros). Hence, we have $|\mathcal{H}| \leq 2^n$. Therefore, $\text{VCdim}(\mathcal{H}) \leq n$.

2. Let $j \in \mathbb{N}$ for which $w(h_j) > 0$. Let $a = w(h_j)$. By monotonicity,

$$\sum_{n=1}^{\infty} w(h_n) \geq \sum_{n=j}^{\infty} a = \infty \ .$$

Then the condition $\sum_{n=1}^{\infty} w(h_n) \leq 1$ is violated.

3. (a) For every $n \in \mathbb{N}$, and $h \in \mathcal{H}_n$, we let $w(h) = \frac{2^{-n}}{|\mathcal{H}_n|}$. Then,

$$\sum_{h \in \mathcal{H}} w(h) = \sum_{n \in \mathbb{N}} \sum_{h \in \mathcal{H}_n} w(h) = \sum_{n \in \mathbb{N}} 2^{-n} \frac{|\mathcal{H}_n|}{|\mathcal{H}_n|} = 1 \ .$$

(b) For each n, denote the hypotheses in $\mathcal{H}_n$ by $h_{n,1}, h_{n,2}, \ldots$. For every $n, k \in \mathbb{N}$ for which $h_{n,k}$ exists, let $w(h_{n,k}) = 2^{-n}2^{-k}$

$$\sum_{h \in \mathcal{H}} w(h) = \sum_{n \in \mathbb{N}} \sum_{h \in \mathcal{H}_n} w(h) \leq \sum_{n \in \mathbb{N}} 2^{-n} \sum_{k \in \mathbb{N}} 2^{-k} = 1 \ .$$

4. Let $\delta \in (0,1)$, $m \in \mathbb{N}$. By Theorem 7.7, with probability at least $1-\delta$,

$$\sup_{h \in \mathcal{H}}\{|L_{\mathcal{D}}(h) - L_S(h)|\} \leq \sqrt{\frac{|h| + \ln(2/\delta)}{2m}}.$$

Then, for every $B > 0$, with probability at least $1-\delta$,

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \sqrt{\frac{|h_S| + \ln(2/\delta)}{2m}} \\ &\leq L_S(h_B^*) + \sqrt{\frac{|h_B^*| + \ln(2/\delta)}{2m}} \\ &\leq L_D(h_B^*) + 2\sqrt{\frac{|h_B^*| + \ln(2/\delta)}{2m}} \\ &\leq L_D(h_B^*) + 2\sqrt{\frac{B + \ln(2/\delta)}{2m}} \ , \end{aligned}$$

where the second inequality follows from the definition of h_S. Hence, for every $B > 0$, with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(h_S) - L_D(h_B^*) \leq 2\sqrt{\frac{B + \ln(2/\delta)}{2m}}.$$

5. (a) See Theorem 7.2, and Theorem 7.3.

   (b) See Theorem 7.2, and Theorem 7.3.

   (c) Let $K \subseteq \mathcal{X}$ be an infinite set which is shattered by $\mathcal{H}$. Assume that $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$. We prove that $\text{VCdim}(\mathcal{H}_n) = \infty$ for some $n \in \mathbb{N}$. Assume by contradiction that $\text{VCdim}(\mathcal{H}_n) < \infty$ for every n. We next define a sequence of finite subsets $(K_n)_{n \in \mathbb{N}}$ of K in a recursive manner. Let $K_1 \subseteq K$ be a set of size $\text{VCdim}(\mathcal{H}_1) + 1$. Suppose that $K_1, \ldots, K_{r-1}$ are chosen. Since K is infinite, we can pick $K_r \subseteq K \setminus (\bigcup_{i=1}^{r-1} K_i)$ such that $|K_r| = \text{VCdim}(\mathcal{H}_r) + 1$. It follows that for each $n \in \mathbb{N}$, there exists a function $f_n : K_n \to \{0, 1\}$ such that $f_n \notin \mathcal{H}_n$. Since K is shattered, we can pick $h \in \mathcal{H}$ which agrees with each f_n on K_n. It follows that for every $n \in \mathbb{N}$, $h \notin \mathcal{H}_n$, contradicting our earlier assumptions.

   (d) Let $\mathcal{X} = \mathbb{R}$. For each $n \in \mathbb{N}$, let $\mathcal{H}_n$ be the class of unions of at most n intervals. Formally, $\mathcal{H} = \{h_{a_1,b_1,\ldots,a_n,b_n} : (\forall i \in [n])\ a_i \leq b_i\}$, where $h_{a_1,b_1,\ldots,a_n,b_n}(x) = \sum_{i=1}^{n} \mathbb{1}_{[x \in [a_i,b_i]]}$. It can be easily seen that $\text{VCdim}(\mathcal{H}_n) = 2n$: If $x_1 < \ldots < x_{2n}$, use the i-th interval to "shatter" the pair $\{x_{2i-1}, x_{2i}\}$. If $x_1 < \ldots < x_{2n+1}$, then the labeling $(1, -1, \ldots, 1, -1, 1)$ cannot be obtained by a union of n intervals. It follows that the VC-dimension of $\mathcal{H} := \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$ is ∞. Hence, $\mathcal{H}$ is not PAC learnable, but it's non-uniformly learnable.

   (e) Let $\mathcal{H}$ be the class of all functions from $[0, 1]$ to $\{0, 1\}$. Then $\mathcal{H}$ shatters the infinite set $[0, 1]$, and thus it is not non-uniformly learnable.

6. **Remark:** We will prove the first part, and then proceed to conclude the fourth and the fifth parts.

   First part[3]: The series $\sum_{n=1}^{\infty} \mathcal{D}(\{x_n\})$ converges to 1. Hence, $\lim_{n\to\infty} \sum_{i=n}^{\infty} \mathcal{D}(\{x_i\}) = 0$.

[3] Errata: If $j \geq i$ then $\mathcal{D}(\{x_j\}) \leq \mathcal{D}(\{x_i\})$. Also, we assume that the error of the Bayes predictor is zero.

Fourth and fifth parts: Let $\epsilon > 0$. There exists $N \in \mathbb{N}$ such that $\sum_{n \geq N} \mathcal{D}(\{x_n\}) < \epsilon$. It follows also that $\mathcal{D}(\{x_n\}) < \epsilon$ for every $n \geq N$. Let $\eta = \mathcal{D}(\{x_N\})$. Then, $\mathcal{D}(\{x_k\}) \geq \eta$ for every $k \in [N]$. Then, by applying the union bound,

$$\begin{aligned}
\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[\mathcal{D}(\{x : x \notin S\}) > \epsilon] &\leq \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[\exists i \in [N] : x_i \notin S] \\
&\leq \sum_{i=1}^{N} \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[x_i \notin S] \\
&\leq N(1-\eta)^m \\
&\leq N \exp(-\eta m) \ .
\end{aligned}$$

If $m \geq m_{\mathcal{D}}(\epsilon, \delta) := \left\lceil \frac{\log(N/\delta)}{\eta} \right\rceil$, then the probability above is at most δ.

Denote the algorithm memorize by M. Given $\mathcal{D}, \epsilon, \delta$, we equip the algorithm Memorize with $m_{\mathcal{D}}(\epsilon, \delta)$ i.i.d. instances. By the previous part, with probability at least $1-\delta$, M observes all the instances in $\mathcal{X}$ (along with their labels) except a subset with probability mass at most ϵ. Since h is consistent with the elements seen by M, it follows that with probability at least $1-\delta$, $L_{\mathcal{D}}(h) \leq \epsilon$.

8 The Runtime of Learning

1. Let $x_1 < x_2 < \ldots < x_m$ be the points in S. Let $x_0 = x_1 - 1$. To find an ERM hypothesis, it suffices to calculate, for each $i, j \in \{0, 1, \ldots, m\}$, the loss induced by picking the interval $[x_i, x_j]$, and find the minimal value. For convenience, denote the loss for each interval by $\ell_{i,j}$. A naive implementation of this strategy runs in time $\Theta(m^3)$. However, dynamic programming yields an algorithm which runs in time $O(m^2)$. The main observation is that given $\ell_{i,j}$, $\ell_{i+1,j}$ and $\ell_{i,j+1}$ can be calculated in time $O(1)$. Consider the table/matrix whose (i, j)-th value is $\ell_{i,j}$. Our algorithm starts with calculating $\ell_{1,j}$ for every j and then, fills the missing values, "row by row". Finding the optimal value among the $\ell_{i,j}$ is done in time $O(m^2)$ as well. Thus, the overall runtime of the proposed algorithm is $O(m^2)$.

2. Let $(\mathcal{H}_n)_{n \in \mathbb{N}}$ be a sequence with the properties described in the question. It follows that given a sample S of size m, the hypotheses in $\mathcal{H}_n$ can be partitioned into at most $\binom{m}{n} \leq O(m^{O(n)})$ equivalence classes, such that any two hypotheses in the same equivalence class have the

same behaviour w.r.t. S. Hence, to find an ERM, it suffices to consider one representative from each equivalent class. Calculating the empirical risk of any representative can be done in time $O(mn)$. Hence, calculating all the empirical errors, and finding the minimizer is done in time $O\left(mnm^{O(n)}\right) = O\left(nm^{O(n)}\right)$.

3. (a) Let $\mathcal{A}$ be an algorithm which implements the ERM rule w.r.t. the class HS_n. We next show that a solution of $\mathcal{A}$ implies a solution to the Max FS problem.

   Let $A \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$ be an input to the Max FS problem. Since the system requires strict inequalities, we may assume w.l.o.g. that $b_i \neq 0$ for every $i \in [m]$ (a solution to the original system exists if and only if there exists a solution to a system in which each 0 in $\mathbf{b}$ is replaced with some $\epsilon > 0$). Furthermore, since the system $A\mathbf{x} > \mathbf{b}$ is invariant under multiplication by a positive scalar, we may assume that $|b_i| = |b_j|$ for every $i, j \in [m]$. Denote $|b_i|$ by b.

   We construct a training sequence $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ such that $\mathbf{x}_i = -\text{sign}(b_i)A_{i\to}$ (where $A_{i\to}$ denotes the i-th row of A), and $y_i = -\text{sign}(b_i)$ for every $i \in [m]$.

   We now argue that $(\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ induce an ERM solution w.r.t. S and HS_n if and only if $\mathbf{w}$ is a solution to the Max FS problem. Since minimizing the empirical risk is equivalent to maximizing the number of correct classifications, it suffices to show that $h_{\mathbf{w},b}$ classifies $\mathbf{x}_i$ correctly if and only if the i-th inequality is satisfied using $\mathbf{w}$. Indeed[4],

$$\begin{aligned} h_w(\mathbf{x}_i) = y_i &\Leftrightarrow y_i(\langle w, \mathbf{x}_i\rangle + b) > 0 \\ &\Leftrightarrow \langle w, A_{i\to}\rangle - b_i > 0 \ . \end{aligned}$$

   It follows that $\mathcal{A}$ can be applied to find a solution to the Max FS problem. Hence, since Max FS is NP-hard, ERM_{HS_n} is NP-hard as well. In other words, unless $P = NP$, $\text{ERM}_{HS_n} \notin P$.

   (b) Throughout this question, we denote by $c_i \in [k]$ the color of v_i.

[4] Note that S is linearly separable if and only if it's separable with strict inequalities. That is, S is separable iff there exist $\mathbf{w}, b$ s.t. $y_i(\langle \mathbf{w}, x_i\rangle + b) > 0$ for every $i \in [m]$. While one direction is trivial, the other follows by adjusting b properly. Note however, that this claim doesn't hold for the class of homogenous halfspaces.

Following the hints, we show that $S(G)$ is separable if and only if G is k-colorable.

First, assume that $S(G)$ is (strictly) separable, using $((\mathbf{w}_1, b_1), \ldots, (\mathbf{w}_k, b_k)) \in (\mathbb{R}^n \times \mathbb{R})^k$. That is, for every $i \in [m]$, $y_i = 1$ if and only if for every $j \in [k]$, $\langle \mathbf{w}_j, \mathbf{x}_i \rangle + b_j > 0$. Consider some vertex $v_i \in V$, and its associated instance $\mathbf{e}_i$. Since the corresponding label is negative, the set $C_i := \{j \in [k] : w_{j,i} + b_j < 0\}$ [5] is not empty. Next, consider some edge $(v_s, v_t) \in E$, and its associated instance $(\mathbf{e}_s + \mathbf{e}_t)/2$. Since the corresponding label is positive, $C_s \cap C_t = \emptyset$. It follows that we can safely pick $c_i \in C_i$ for every $i \in [n]$. In other words, G is k-colorable.

For the other direction, assume that G is k-colorable using $(c_1, \ldots, c_n)$. We associate a vector $\mathbf{w}_i \in \mathbb{R}^n$ with each color $i \in [k]$. Define $\mathbf{w}_i = \sum_{j \in [n]} -\mathbb{1}_{[c_j = i]} \mathbf{e}_j$. Let $b_1 = \ldots = b_k = b := 0.6$. Consider some vertex $v_i \in V$, and its associated instance $\mathbf{e}_i$. We note that $\text{sign}(\langle \mathbf{w}_{c_i}, \mathbf{e}_i \rangle + b) = \text{sign}(-0.4) = -1$, so $\mathbf{e}_i$ is classified correctly. Consider any edge (v_s, v_t). We know that $c_s \neq c_t$. Hence, for every $j \in [k]$, either $w_{j,s} = 0$ or $w_{j,t} = 0$. Hence, $\forall j \in [k]$, $\text{sign}(\langle \mathbf{w}_j, (\mathbf{e}_s + \mathbf{e}_t)/2 \rangle + b_j) = \text{sign}(0.1) = 1$, so the (instance associated with the) edge (v_s, v_t) is classified correctly. All in all, $S(G)$ is separable using $((\mathbf{w}_1, b_1), \ldots, (\mathbf{w}_k, b_k))$.

Let $\mathcal{A}$ be an algorithm which implements the ERM rule w.r.t. $\mathcal{H}_k^n$. Then $\mathcal{A}$ can be used to solve the k-coloring problem. Given a graph G, $\mathcal{A}$ returns "Yes" if and only if the resulting training sequence $S(G)$ is separable. Since the k-coloring problem is NP-hard (for $k \geq 3$), we conclude that $\text{ERM}_{\mathcal{H}_k^n}$ is NP-hard as well.

4. We will follow the hint[6]. With probability at least $1 - 0.3 = 0.7 \geq 0.5$, the generalization error of the returned hypothesis is at most $1/|S|$. Since the distribution is uniform, this implies that the returned hypothesis is consistent with the training set.

[5] recall that $w_{j,i}$ denotes the i-th element of the j-th vector.

[6] Errata:

- The realizable assumption should be made for the task of minimizing the ERM.
- The definition of RP should be modified. We don't consider a decision problem ("Yes"/"No" question) but an optimization problem (predicting correctly the labels of the instances in the training set).

9 Linear Predictors

1. Define a vector of auxiliary variables $s = (s_1, \ldots, s_m)$. Following the hint, minimizing the empirical risk is equivalent to minimizing the linear objective $\sum_{i=1}^m s_i$ under the following constraints:

$$(\forall i \in [m]) \;\; \mathbf{w}^T \mathbf{x}_i - s_i \leq y_i \;\; , \;\; -\mathbf{w}^T \mathbf{x}_i - s_i \leq -y_i \qquad (3)$$

It is left to translate the above into matrix form. Let $A \in \mathbb{R}^{2m \times (m+d)}$ be the matrix $A = [X \;\; -I_m; -X \;\; -I_m]$, where $X_{i \rightarrow} = x_i$ for every $i \in [m]$. Let $\mathbf{v} \in \mathbb{R}^{d+m}$ be the vector of variables $(w_1, \ldots, w_d, s_1, \ldots, s_m)$. Define $\mathbf{b} \in \mathbb{R}^{2m}$ to be the vector $\mathbf{b} = (y_1, \ldots, y_m, -y_1, \ldots, -y_m)^T$. Finally, let $\mathbf{c} \in \mathbb{R}^{d+m}$ be the vector $\mathbf{c} = (\mathbf{0}_d; \mathbf{1}_m)$. It follows that the optimization problem of minimizing the empirical risk can be expressed as the following LP:

$$\begin{array}{ll} \min & c^T \mathbf{v} \\ \text{s.t.} & A\mathbf{v} \leq \mathbf{b} \; . \end{array}$$

2. Consider the matrix $X \in \mathbb{R}^{d \times m}$ whose columns are $x_1, \ldots, x_m$. The rank of X is equal to the dimension of the subspace $\text{span}(\{x_1, \ldots, x_m\})$. The SVD theorem (more precisely, Lemma C.4) implies that the rank of X is equal to the rank of $A = XX^\top$. Hence, the set $\{x_1, \ldots, x_m\}$ spans $\mathbb{R}^d$ if and only if the rank of $A = XX^\top$ is d, i.e., iff $A = XX^\top$ is invertible.

3. Following the hint, let $d = m$, and for every $i \in [m]$, let $\mathbf{x}_i = \mathbf{e}_i$. Let us agree that $\text{sign}(0) = -1$. For $i = 1, \ldots, d$, let $y_i = 1$ be the label of x_i. Denote by $\mathbf{w}^{(t)}$ the weight vector which is maintained by the Perceptron. A simple inductive argument shows that for every $i \in [d]$, $\mathbf{w}_i = \sum_{j<i} \mathbf{e}_j$. It follows that for every $i \in [d]$, $\langle \mathbf{w}^{(i)}, \mathbf{x}_i \rangle = 0$. Hence, all the instances $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are misclassified (and then we obtain the vector $w = (1, \ldots, 1)$ which is consistent with $x_1, \ldots, x_m$). We also note that the vector $\mathbf{w}^\star = (1, \ldots, 1)$ satisfies the requirements listed in the question.

4. Consider all positive examples of the form $(\alpha, \beta, 1)$, where $\alpha^2 + \beta^2 + 1 \leq R^2$. Observe that $\mathbf{w}^\star = (0, 0, 1)$ satisfies $y \langle \mathbf{w}^\star, \mathbf{x} \rangle \geq 1$ for all such $(\mathbf{x}, y)$. We show a sequence of R^2 examples on which the Perceptron makes R^2 mistakes.

The idea of the construction is to start with the examples $(\alpha_1, 0, 1)$ where $\alpha_1 = \sqrt{R^2 - 1}$. Now, on round t let the new example be such that the following conditions hold:

(a) $\alpha^2 + \beta^2 + 1 = R^2$

(b) $\langle \mathbf{w}_t, (\alpha, \beta, 1) \rangle = 0$

As long as we can satisfy both conditions, the Perceptron will continue to err. We'll show that as long as $t \leq R^2$ we can satisfy these conditions.

Observe that, by induction, $\mathbf{w}^{(t-1)} = (a, b, t-1)$ for some scalars a, b. Observe also that $\|\mathbf{w}_{t-1}\|^2 = (t-1)R^2$ (this follows from the proof of the Perceptron's mistake bound, where inequalities hold with equality). That is, $a^2 + b^2 + (t-1)^2 = (t-1)R^2$.

W.l.o.g., let us rotate $\mathbf{w}^{(t-1)}$ w.r.t. the z axis so that it is of the form $(a, 0, t-1)$ and we have $a = \sqrt{(t-1)R^2 - (t-1)^2}$. Choose

$$\alpha = -\frac{t-1}{a} \ .$$

Then, for every β,

$$\langle (a, 0, t-1), (\alpha, \beta, 1) \rangle = 0 \ .$$

We just need to verify that $\alpha^2 + 1 \leq R^2$, because if this is true then we can choose $\beta = \sqrt{R^2 - \alpha^2 - 1}$. Indeed,

$$\begin{aligned}\alpha^2 + 1 &= \frac{(t-1)^2}{a^2} + 1 = \frac{(t-1)^2}{(t-1)R^2 - (t-1)^2} + 1 = \frac{(t-1)R^2}{(t-1)R^2 - (t-1)^2} \\ &= R^2 \frac{1}{R^2 - (t-1)} \\ &\leq R^2\end{aligned}$$

where the last inequality assumes $R^2 \geq t$.

5. Since for every $w \in \mathbb{R}^d$, and every $x \in \mathbb{R}^d$, we have

$$\text{sign}(\langle w, x \rangle) = \text{sign}(\langle \eta w, x \rangle) \ ,$$

we obtain that the modified Perceptron and the Perceptron produce the same predictions. Consequently, both algorithms perform the same number of iterations.

6. In this question we will denote the class of halfspaces in $\mathbb{R}^{d+1}$ by $\mathcal{L}_{d+1}$.

   (a) Assume that $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subseteq \mathbb{R}^d$ is shattered by $\mathcal{B}_d$. Then, $\forall \mathbf{y} = (y_1, \ldots, y_d) \in \{-1, 1\}^d$ there exists $B_{\mu,r} \in \mathcal{B}$ s.t. for every i

   $$B_{\mu,r}(\mathbf{x}_i) = y_i \ .$$

   Hence, for the above μ and r, the following identity holds for every $i \in [m]$:

   $$\operatorname{sign}\left((2\mu; -1)^T(\mathbf{x}_i; \|\mathbf{x}_i\|^2) - \|\mu\|^2 + r^2\right) = y_i \ , \tag{4}$$

   where ; denotes vector concatenation. For each $i \in [m]$, let $\phi(x_i) = (x_i; \|x\|_i^2)$. Define the halfspace $h \in \mathcal{L}_{d+1}$ which corresponds to $w = (2\mu; -1)$, and $b = \|\mu\|^2 - r^2$. Equation (4) implies that for every $i \in [m]$,

   $$h(x_i) = y_i$$

   All in all, if $A = \{x_1, \ldots, x_m\}$ is shattered by $\mathcal{B}$, then $\phi(A) := \{\phi(x_1), \ldots, \phi(x_m)\}$, is shattered by $\mathcal{L}$. We conclude that $d + 2 = \text{VCdim}(\mathcal{L}_{d+1}) \geq VC(\mathcal{B}_d)$.

   (b) Consider the set C consisting of the unit vectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$, and the origin $\mathbf{0}$. Let $A \subseteq C$. We show that there exists a ball such that all the vectors in A are labeled positively, while the vectors in $C \setminus A$ are labeled negatively. We define the center $\mu = \sum_{e \in A} e$. Note that for every unit vector in A, its distance to the center is $\sqrt{|A| - 1}$. Also, for every unit vector outside A, its distance to the center is $\sqrt{|A| + 1}$. Finally, the distance of the origin to the center is $\sqrt{A}$. Hence, if $0 \in A$, we will set $r = \sqrt{|A| - 1}$, and if $0 \notin A$, we will set $r = \sqrt{|A|}$. We conclude that the set C is shattered by $\mathcal{B}_d$. All in all, we just showed that $\text{VCdim}(\mathcal{B}_d) \geq d + 1$.

10 Boosting

1. Let $\epsilon, \delta \in (0, 1)$. Pick k "chunks" of size $m_{\mathcal{H}}(\epsilon/2)$. Apply A on each of these chunks, to obtain $\hat{h}_1, \ldots, \hat{h}_k$. Note that the probability that $\min_{i \in [k]} L_{\mathcal{D}}(\hat{h}_i) \leq \min L_{\mathcal{D}}(h) + \epsilon/2$ is at least $1 - \delta_0^k \geq 1 - \delta/2$. Now, apply an ERM over the class $\hat{\mathcal{H}} := \{\hat{h}_1, \ldots, \hat{h}_k\}$ with the training data being the last chunk of size $\left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$. Denote the output hypothesis

by $\hat{h}$. Using Corollary 4.6, we obtain that with probability at least $1-\delta/2$, $L_{\mathcal{D}}(\hat{h}) \leq \min_{i\in[k]} L_{\mathcal{D}}(h_i) + \epsilon/2$. Applying the union bound, we obtain that with probability at least $1-\delta$,

$$\begin{aligned} L_{\mathcal{D}}(\hat{h}) &\leq \min_{i\in[k]} L_{\mathcal{D}}(h_i) + \epsilon/2 \\ &\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \ . \end{aligned}$$

2. Note[7] that if $g(x) = 1$, then $h(x)$ is either 0.5 or 1.5, and if $g(x) = -1$, then $h(x)$ is either -0.5 or -1.5.

3. The definition of ϵ_t implies that

$$\begin{aligned} \sum_{i=1}^{m} D_i^{(t)} \exp(-w_t y_i h_t(x_i)) \cdot \mathbb{1}_{[h_t(x_i)\neq y_i]} &= \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \\ &= \sqrt{\epsilon_t(1-\epsilon_t)} \ . \qquad (5) \end{aligned}$$

Similarly,

$$\begin{aligned} \sum_{j=1}^{m} D_j^{(t)} \cdot \exp(-w_t y_j h_t(x_j)) &= \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} + (1-\epsilon_t) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \\ &= 2\sqrt{\epsilon_t(1-\epsilon_t)} \ . \qquad (6) \end{aligned}$$

The desired equality is obtained by observing that the error of h_t w.r.t. the distribution D_{t+1} can be obtained by dividing Equation (5) by Equation (6).

4. (a) Let $\mathcal{X}$ be a finite set of size n. Let B be the class of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $L(B,T) = B$, and both are finite. Hence, for any $T \geq 1$,

$$\mathrm{VCdim}(B) = \mathrm{VCdim}(L(B,T)) = \log 2^n = n \ .$$

(b)
- Denote by $\mathcal{B}$ the class of decision stumps in $\mathbb{R}^d$. Formally, $\mathcal{B} = \{h_{j,b,\theta} : j \in [d], b \in \{-1,1\}, \theta \in \mathbb{R}\}$, where $h_{j,b,\theta}(\mathbf{x}) = b \cdot \mathrm{sign}(\theta - x_j)$.
  For each $j \in [d]$, let $\mathcal{B}_j = \{h_{b,\theta} : b \in \{-1,1\}, \theta \in \mathbb{R}\}$, where $h_{b,\theta}(\mathbf{x}) = b \cdot \mathrm{sign}(\theta - x_j)$. Note that $\mathrm{VCdim}(\mathcal{B}_j) = 2$.
  Clearly, $\mathcal{B} = \bigcup_{j=1}^{d} \mathcal{B}_j$. Applying Exercise 11, we conclude that

$$\mathrm{VCdim}(\mathcal{B}) \leq 16 + 2\log d \ .$$

[7] Errata: w_1 should be -0.5

- Assume w.l.o.g. that $d = 2^k$ for some $k \in \mathbb{N}$ (otherwise, replace d by $2^{\lfloor \log d \rfloor}$). Let $A \in \mathbb{R}^{k \times d}$ be the matrix whose columns range over the (entire) set $\{0,1\}^k$. For each $i \in [k]$, let $\mathbf{x}_i = A_{i \rightarrow}$. We claim that the set $C = \{\mathbf{x}_i, \ldots, \mathbf{x}_k\}$ is shattered. Let $I \subseteq [k]$. We show that we can label the instances in I positively, while the instances $[k] \backslash I$ are labeled negatively. By our construction, there exists an index j such that $A_{i,j} = \mathbf{x}_{i,j} = 1$ iff $i \in I$. Then, $h_{j,-1,1/2}(\mathbf{x}_i) = 1$ iff $i \in I$.

(c) Following the hint, for each $i \in [Tk/2]$, let $\mathbf{x}_i = \lceil i/k \rceil A_{i,\rightarrow}$. We claim that the set $C = \{\mathbf{x}_i : i \in [Tk/2]\}$ is shattered by $L(B_d, T)$. Let $I \subseteq [Tk/2]$. Then $I = I_1 \cup \ldots \cup I_{T/2}$, where each I_t is a subset of $\{(t-1)k+1, \ldots, tk\}$. For each $t \in [T/2]$, let j_t be the corresponding column of A (i.e., $A_{i,j} = 1$ iff $(t-1)k + i \in I_t$). Let

$$\begin{aligned} h(x) = \operatorname{sign}((&h_{j_1,-1,1/2} + h_{j_1,1,3/2} + h_{j_2,-1,3/2} + h_{j_2,1,5/2} + \ldots \\ &+ h_{j_{T/2-1},-1,T/2-3/2} + h_{j_{T/2-1},1,T/2-1/2} + h_{j_{T/2},-`1,T/2-1/2})(x)) \ . \end{aligned}$$

Then $h(\mathbf{x}_i) = 1$ iff $i \in I$. Finally, observe that $h \in L(B_d, T)$.

5.
- We apply dynamic programming. Let $A \in \mathbb{R}^{n \times n}$. First, observe that filling sequentially each item of the first column and the first row of $I(A)$ can be done in a constant time. Next, given $i, j \geq 2$, we note that $I(A)_{i,j} = I(A)_{i-1,j} + I(A)_{i,j-1} - I(A)_{i-1,j-1}$. Hence, filling the values of $I(A)$, row by row, can be done in time linear in the size of $I(A)$.
- We show the calculation of type B. The calculation of the other types is done similarly. Given $i_1 < i_2, j_1 < j_2$, consider the following rectangular regions:

$$\begin{aligned} A_U &= \{(i,j) : i_1 \leq i \leq \lfloor (i_2 + i_1)/2 \rfloor, j_1 \leq j \leq j_2\} \ , \\ A_D &= \{(i,j) : \lceil (i_2 + i_1)/2 \rceil \leq i \leq i_2, j_1 \leq j \leq j_2\} \ . \end{aligned}$$

Let $b \in \{-1, 1\}$. Let h be the associated hypothesis. That is, $h(A) = b\left(\sum_{(i,j) \in A_L} A_{i,j} - \sum_{(i,j) \in A_R} A_{i,j}\right)$. Let $i_3 = \lfloor (i_1 + i_2)/2 \rfloor, i_4 = \lceil (i_1 + i_2)/2 \rceil$. Then,

$$\begin{aligned} h(A) = &B_{i_2,j_2} - B_{i_3,j_2} - B_{i_2,j_1-1} + B_{i_3,j_1-1} \\ &- (B_{i_3,j_2} - B_{i_1-1,j_2} - B_{i_3,j_1-1} + B_{i_1-1,j_1-1}) \end{aligned}$$

We conclude that $h(A)$ can be calculated in constant time given $B = I(A)$.

11 Model Selection and Validation

1. Let S be an i.i.d. sample. Let h be the output of the described learning algorithm. Note that (independently of the identity of S), $L_{\mathcal{D}}(h) = 1/2$ (since h is a constant function).

   Let us calculate the estimate $L_V(h)$. Assume that the parity of S is 1. Fix some fold $\{(\mathbf{x}, y)\} \subseteq S$. We distinguish between two cases:

   - The parity of $S \setminus \{\mathbf{x}\}$ is 1. It follows that $y = 0$. When being trained using $S \setminus \{\mathbf{x}\}$, the algorithm outputs the constant predictor $h(\mathbf{x}) = 1$. Hence, the leave-one-out estimate using this fold is 1.
   - The parity of $S \setminus \{\mathbf{x}\}$ is 0. It follows that $y = 1$. When being trained using $S \setminus \{\mathbf{x}\}$, the algorithm outputs the constant predictor $h(\mathbf{x}) = 0$. Hence, the leave-one-out estimate using this fold is 1.

   Averaging over the folds, the estimate of the error of h is 1. Consequently, the difference between the estimate and the true error is 1/2. The case in which the parity of S is 0 is analyzed analogously.

2. Consider for example the case in which $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots \subseteq \mathcal{H}_k$, and $|\mathcal{H}_i| = 2^i$ for every $i \in k$. Learning $\mathcal{H}_k$ in the Agnostic-Pac model provides the following bound for an ERM hypothesis h:

   $$L_D(h) \leq \min_{h \in \mathcal{H}_k} L_D(h) + \sqrt{\frac{2(k + 1 + \log(1/\delta))}{m}}.$$

   Alternatively, we can use model selection as we describe next. Assume that j is the minimal index which contains a hypothesis $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$. Fix some $r \in [k]$. By Hoeffding's inequality, with probability at least $1 - \delta/(2k)$, we have

   $$|L_{\mathcal{D}}(\hat{h}_r) - L_V(\hat{h}_r)| \leq \sqrt{\frac{1}{2\alpha m} \log \frac{4}{\delta}}.$$

   Applying the union bound, we obtain that with probability at least $1 - \delta/2$, the following inequality holds (simultaneously) for every $r \in$

$[k]$:

$$\begin{aligned} L_{\mathcal{D}}(\hat{h}) &\leq L_V(\hat{h}) + \sqrt{\frac{1}{2\alpha m}\log\frac{4k}{\delta}} \\ &\leq L_V(\hat{h}_r) + \sqrt{\frac{1}{2\alpha m}\log\frac{4k}{\delta}} \\ &\leq L_{\mathcal{D}}(\hat{h}_r) + 2\sqrt{\frac{1}{2\alpha m}\log\frac{4k}{\delta}} \\ &= L_{\mathcal{D}}(\hat{h}_r) + \sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}\ . \end{aligned}$$

In particular, with probability at least $1-\delta/2$, we have

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(\hat{h}_j) + \sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}\ .$$

Using similar arguments[8], we obtain that with probability at least $1-\delta/2$,

$$\begin{aligned} L_{\mathcal{D}}(\hat{h}_j) &\leq L_{\mathcal{D}}(h^\star) + \sqrt{\frac{2}{(1-\alpha)m}\log\frac{4|\mathcal{H}_j|}{\delta}} \\ &= L_{\mathcal{D}}(h^\star) + \sqrt{\frac{2}{(1-\alpha)m}\log\frac{4|\mathcal{H}_j|}{\delta}} \end{aligned}$$

Combining the two last inequalities with the union bound, we obtain that with probability at least $1-\delta$,

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^\star) + \sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m}\log\frac{4|\mathcal{H}_j|}{\delta}}\ .$$

We conclude that

$$L_D(\hat{h}) \leq L_D(h^*) + \sqrt{\frac{2}{\alpha m} log\frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m}(j + \log\frac{4}{\delta})}\ .$$

Comparing the two bounds, we see that when the "optimal index" j is significantly smaller than k, the bound achieved using model selection is much better. Being even more concrete, if j is logarithmic in k, we achieve a logarithmic improvement.

[8] This time we consider each of the hypotheses in H_j, and apply the union bound accordingly.

12 Convex Learning Problems

1. Let $\mathcal{H}$ be the class of homogenous halfspaces in $\mathbb{R}^d$. Let $\mathbf{x} = \mathbf{e}_1$, $y = 1$, and consider the sample $S = \{(\mathbf{x}, y)\}$. Let $\mathbf{w} = -e_1$. Then, $\langle \mathbf{w}, \mathbf{x} \rangle = -1$, and thus $L_S(h_{\mathbf{w}}) = 1$. Still, $\mathbf{w}$ is a local minima. Let $\epsilon \in (0, 1)$. For every $\mathbf{w}'$ with $\|\mathbf{w}' - \mathbf{w}\| \leq \epsilon$, by Cauchy-Schwartz inequality, we have

$$\begin{aligned}\langle \mathbf{w}', \mathbf{x} \rangle &= \langle \mathbf{w}, \mathbf{x} \rangle - \langle \mathbf{w}' - \mathbf{w}, \mathbf{x} \rangle \\ &= -1 - \langle \mathbf{w}' - \mathbf{w}, \mathbf{x} \rangle \\ &\leq -1 + \|\mathbf{w}' - \mathbf{w}\|_2 \|\mathbf{x}\|_2 \\ &< -1 + 1 \\ &= 0 \ .\end{aligned}$$

   Hence, $L_S(\mathbf{w}') = 1$ as well.

2. **Convexity:** Note that the function $g : \mathbb{R} \to \mathbb{R}$, defined by $g(a) = \log(1 + \exp(a))$ is convex. To see this, note that g'' is non-negative. The convexity of ℓ (or more accurately, of $\ell(\cdot, z)$ for all z) follows now from Claim 12.4.

   Lipschitzness: The function $g(a) = \log(1 + \exp(a))$ is 1-Lipschitz, since $|g'(a)| = \frac{\exp(a)}{1+\exp(a)} = \frac{1}{\exp(-a)+1} \leq 1$. Hence, by Claim 12.7, ℓ is B-Lipschitz

   Smoothness: We claim that $g(a) = \log(1 + \exp(a))$ is 1/4-smooth. To see this, note that

$$\begin{aligned}g''(a) &= \frac{\exp(-a)}{(\exp(-a) + 1)^2} \\ &= \left(\exp(a)(\exp(-a) + 1)^2\right)^{-1} \\ &= \frac{1}{2 + \exp(a) + \exp(-a)} \\ &\leq 1/4 \ .\end{aligned}$$

   Combine this with the mean value theorem, to conclude that g' is 1/4-Lipschitz. Using Claim 12.9, we conclude that ℓ is $B^2/4$-smooth.

   Boundness: The norm of each hypothesis is bounded by B according to the assumptions.

   All in all, we conclude that the learning problem of linear regression is Convex-Smooth-Bounded with parameters $B^2/4$, B, and Convex-Lipschitz-Bounded with parameters B, B.

3. Fix some $(\mathbf{x}, y) \in \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}'\|_2 \leq R\} \times \{-1, 1\}$. Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$. For $i \in [2]$, let $\ell_i = \max\{0, 1 - y\langle \mathbf{w}_i, \mathbf{x}\rangle\}$. We wish to show that $|\ell_1 - \ell_2| \leq R\|\mathbf{w}_1 - \mathbf{w}_2\|_2$. If both $y\langle \mathbf{w}_1, \mathbf{x}\rangle \geq 1$ and $y\langle \mathbf{w}_2, \mathbf{x}\rangle \geq 1$, then $|\ell_1 - \ell_2| = 0 \leq R\|\mathbf{w}_1 - \mathbf{w}_2\|_2$. Assume now that $|\{i : y\langle \mathbf{w}_i, x\rangle < 1\}| \geq 1$. Assume w.l.o.g. that $1 - y\langle \mathbf{w}_1, \mathbf{x}\rangle \geq 1 - y\langle \mathbf{w}_2, \mathbf{x}\rangle$. Hence,

$$\begin{aligned} |\ell_1 - \ell_2| &= \ell_1 - \ell_2 \\ &= 1 - y\langle \mathbf{w}_1, \mathbf{x}\rangle - \max\{0, 1 - y\langle \mathbf{w}_2, \mathbf{x}\rangle\} \\ &\leq 1 - y\langle \mathbf{w}_1, \mathbf{x}\rangle - (1 - y\langle \mathbf{w}_2, \mathbf{x}\rangle) \\ &= y\langle \mathbf{w}_2 - \mathbf{w}_1, \mathbf{x}\rangle \\ &\leq \|\mathbf{w}_1 - \mathbf{w}_2\|\|\mathbf{x}\| \\ &\leq R\|\mathbf{w}_1 - \mathbf{w}_2\| \ . \end{aligned}$$

4. (a) Fix a Turing machine T. If T halts on the input 0, then for every $h \in [0, 1]$,

   $$\ell(h, T) = \langle (h, 1 - h), (1, 0)\rangle \ .$$

   If T halts on the input 1, then for every $h \in [0, 1]$,

   $$\ell(h, T) = \langle (h, 1 - h), (0, 1)\rangle \ .$$

   In both cases, ℓ is linear, and hence convex over $\mathcal{H}$.

   (b) The idea is to reduce the halting problem to the learning problem[9]. More accurately, the following decision problem can be easily reduced to the learning problem described in the question: Given a Turing machine M, does M halt given the input M? The proof that the halting problem is not decidable implies that this decision problem is not decidable as well. Hence, there is mo computable algorithm that learns the problem described in the question.

13 Regularization and Stability

1. **From bounded expected risk to agnostic PAC learning:** We assume A is a (proper) algorithm that guarantees the following: If $m \geq m_{\mathcal{H}}(\epsilon)$ then for every distribution $\mathcal{D}$ it holds that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \ .$$

[9] Errata: "T halts on the input 0" should be replaced (everywhere) by "T halts on the input T"

- Since $A(S) \in \mathcal{H}$, the random variable $\theta = L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ is non-negative. [10] Therefore, Markov's inequality implies that

  $$\mathbb{P}[\theta \geq \mathbb{E}[\theta]/\delta] \leq \frac{\mathbb{E}[\theta]}{\mathbb{E}[\theta]/\delta} = \delta \ .$$

  In other words, with probability of at least $1 - \delta$ we have

  $$\theta \leq \mathbb{E}[\theta]/\delta \ .$$

  But, if $m \geq m_{\mathcal{H}}(\epsilon\,\delta)$ then we know that $\mathbb{E}[\theta] \leq \epsilon\,\delta$. This yields $\theta \leq \epsilon$, which concludes our proof.

- Let $k = \lceil \log_2(2/\delta) \rceil$. [11] Divide the data into $k+1$ chunks, where each of the first k chunks is of size $m_{\mathcal{H}}(\epsilon/4)$ examples.[12] Train the first k chunks using A. Let h_i be the output of A for the i'th chunk. Using the previous question, we know that with probability of at most $1/2$ over the examples in the i'th chunk it holds that $L_{\mathcal{D}}(h_i) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) > \epsilon/2$. Since the examples in the different chunks are independent, the probability that for all the chunks we'll have $L_{\mathcal{D}}(h_i) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) > \epsilon/2$ is at most 2^{-k}, which by the definition of k is at most $\delta/2$. In other words, with probability of at least $1 - \delta/2$ we have that

  $$\min_{i \in [k]} L_{\mathcal{D}}(h_i) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon/2 \ . \tag{7}$$

  Next, apply ERM over the finite class $\{h_1, \ldots, h_k\}$ on the last chunk. By Corollary 4.6, if the size of the last chunk is at least $\frac{8\log(4k/\delta)}{\epsilon^2}$ then, with probability of at least $1 - \delta/2$, we have

  $$L_{\mathcal{D}}(\hat{h}) \leq \min_{i \in [k]} L_{\mathcal{D}}(h_i) + \epsilon/2 \ .$$

  Applying the union bound and combining with Equation (7) we conclude our proof. The overall sample complexity is

  $$m_{\mathcal{H}}(\epsilon/4)\lceil \log_2(2/\delta) \rceil + 8\left\lceil \frac{\log(4/\delta) + \log(\lceil \log_2(2/\delta) \rceil)}{\epsilon^2} \right\rceil$$

[10] One should assume here that A is a "proper" learner.

[11] Note that in the original question the size was mistakenly $k = \lceil \log_2(1/\delta) \rceil$.

[12] Note that in the original question the size was mistakenly $m_{\mathcal{H}}(\epsilon/2)$.

2. **Learnability without uniform convergence:** Let $\mathcal{B}$ be the unit ball of $\mathbb{R}^d$, let $\mathcal{H} = \mathcal{B}$, let $Z = \mathcal{B} \times \{0,1\}^d$, and let $\ell : Z \times \mathcal{H} \to \mathbb{R}$ be defined as follows:

$$\ell(\mathbf{w}, (\mathbf{x}, \boldsymbol{\alpha})) = \sum_{i=1}^{d} \alpha_i (x_i - w_i)^2 \ .$$

   - This problem is learnable using the RLM rule with a sample complexity that does not depend on d. Indeed, the hypothesis class is convex and bounded, the loss function is convex, non-negative, and smooth, since

     $$\|\nabla \ell(\mathbf{v}, (\mathbf{x}, \boldsymbol{\alpha})) - \nabla \ell(\mathbf{w}, (\mathbf{x}, \boldsymbol{\alpha}))\|^2 = 4 \sum_{i=1}^{d} \alpha_i^2 (v_i - w_i)^2 \leq 4\|\mathbf{v} - \mathbf{w}\|^2 \ ,$$

     where in the last inequality we used $\alpha_i \in \{0,1\}$.

   - Fix some $j \in [d]$. The probability to sample a training set of size m such that $\alpha_j = 0$ [13] for all the examples is 2^{-m}. Since the coordinates of α are chosen independently, the probability that for all $j \in [d]$ the above event will not happen is $(1 - 2^{-m})^d \geq \exp(-2^{-m+1} d)$. Therefore, if $d \gg 2^m$, the above quantity goes to zero, meaning that there is a high chance that for some j we'll have $\alpha_j = 0$ for all the examples in the training set.

     For such a sample, we have that every vector of the form $\mathbf{x}_0 + \eta \mathbf{e}_j$ is an ERM. For simplicity, assume that $\mathbf{x}_0$ is the all zeros vector and set $\eta = 1$. Then, $L_S(\mathbf{e}_j) = 0$. However, $L_{\mathcal{D}}(\mathbf{e}_j) = \frac{1}{2}$. It follows that the sample complexity of uniform convergence must grow with $\log(d)$.

   - Taking d to infinity we obtain a problem which is learnable but for which the uniform convergence property does not hold. While this seems to contradict the fundamental theorem of statistical learning, which states that a problem is learnable if and only if uniform convergence holds, there is no contradiction since the fundamental theorem only holds for binary classification problems, while here we consider a more general problem.

3. **Stability and asymptotic ERM is sufficient for learnability:**

[13] In the hint, it is written mistakenly that $\alpha_j = 1$

Proof. Let $h^\star \in \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ (for simplicity we assume that $h^\star$ exists). We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}(h^\star)\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S)) + L_S(A(S)) - L_{\mathcal{D}}(h^\star)\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_S(A(S)) - L_{\mathcal{D}}(h^\star)\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_S(A(S)) - L_S(h^\star)\right] \\
&\le \epsilon_1(m) + \epsilon_2(m)\ .
\end{aligned}$$

□

4. **Strong convexity with respect to general norms:**

   (a) Item 2 of the lemma follows directly from the definition of convexity and strong convexity. For item 3, the proof is identical to the proof for the ℓ_2 norm.

   (b) The function $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_1^2$ is not strongly convex with respect to the ℓ_1 norm. To see this, let the dimension be $d = 2$ and take $\mathbf{w} = \mathbf{e}_1$, $\mathbf{u} = \mathbf{e}_2$, and $\alpha = 1/2$. We have $f(\mathbf{w}) = f(\mathbf{u}) = f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) = 1/2$.

   (c) The proof is almost identical to the proof for the ℓ_2 case and is therefore omitted.

   (d) *Proof.* [14]
   Using Holder's inequality,

   $$\|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_i| \ \le\ \|\mathbf{w}\|_q \|(1,\ldots,1)\|_p\ ,$$

   where $p = (1-1/q)^{-1} = \log(d)$. Since $\|(1,\ldots,1)\|_p = d^{1/p} = e < 3$ we obtain that for every $\mathbf{w}$,

   $$\|\mathbf{w}\|_q \ge \|\mathbf{w}\|_1/3\ .$$

[14] Errata: In the original question, there's a typo: it should be proved that R is (1/3) strongly convex instead of $\frac{1}{3\log(d)}$ strongly convex.

Combining this with the strong convexity of R w.r.t. $\|\cdot\|_q$ we obtain

$$\begin{aligned} R(\alpha\mathbf{w}+(1-\alpha)\mathbf{u}) &\le \alpha R(\mathbf{w})+(1-\alpha)R(\mathbf{u})-\frac{\alpha(1-\alpha)}{2}\|\mathbf{w}-\mathbf{u}\|_q^2 \\ &\le \alpha R(\mathbf{w})+(1-\alpha)R(\mathbf{u})-\frac{\alpha(1-\alpha)}{2\cdot 3}\|\mathbf{w}-\mathbf{u}\|_1^2 \end{aligned}$$

This tells us that R is (1/3)-strongly-convex with respect to $\|\cdot\|_1$. □

14 Stochastic Gradient Descent

1. Divide the definition of strong convexity by α and rearrange terms, to get that

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

   In addition, if $\mathbf{v}$ be a subgradient of f at $\mathbf{w}$, then,

   $$f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w}))-f(\mathbf{w}) \ge \alpha\langle \mathbf{u}-\mathbf{w},\mathbf{v}\rangle ~.$$

   Combining together we obtain

   $$\langle \mathbf{u}-\mathbf{w},\mathbf{v}\rangle \le f(\mathbf{u})-f(\mathbf{w})-\frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 ~.$$

   Taking the limit $\alpha \to 0$ we obtain that the right-hand side converges to $f(\mathbf{w})-f(\mathbf{u})-\frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2$, which concludes our proof.

2. Plugging the definitions of η and T into Theorem 14.13 we obtain

   $$\begin{aligned} \mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] &\le \frac{1}{1-\frac{1}{1+3/\epsilon}}\left(L_{\mathcal{D}}(\mathbf{w}^\star)+\frac{\|\mathbf{w}^\star\|^2(1+3/\epsilon)\epsilon^2}{24B^2}\right) \\ &\le (1+3/\epsilon)\epsilon/3\left(L_{\mathcal{D}}(\mathbf{w}^\star)+\frac{(1+3/\epsilon)\epsilon^2}{24}\right) \\ &= (1+\epsilon/3)\left(L_{\mathcal{D}}(\mathbf{w}^\star)+\frac{\epsilon(\epsilon+3)}{24}\right) \\ &= L_{\mathcal{D}}(\mathbf{w}^\star)+\frac{\epsilon}{3}L_{\mathcal{D}}(\mathbf{w}^\star)+\frac{\epsilon}{3} \end{aligned}$$

   Finally, since $L_{\mathcal{D}}(\mathbf{w}^\star) \le L_{\mathcal{D}}(\mathbf{0}) \le 1$, we conclude the proof.

3. **Perceptron as a sub-gradient descent algorithm:**
   - Clearly, $f(\mathbf{w}^\star) \leq 0$. If there is strict inequality, then we can decrease the norm of $\mathbf{w}^\star$ while still having $f(\mathbf{w}^\star) \leq 0$. But $\mathbf{w}^\star$ is chosen to be of minimal norm and therefore equality must hold. In addition, any $\mathbf{w}$ for which $f(\mathbf{w}) < 1$ must satisfy $1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle < 1$ for every i, which implies that it separates the examples.
   - A sub-gradient of f is given by $-y_i\mathbf{x}_i$, where $i \in \text{argmax}\{1 - y_i \langle \mathbf{w}, \mathbf{x}_i\rangle\}$.
   - The resulting algorithm initializes $\mathbf{w}$ to be the all zeros vector and at each iteration finds $i \in \text{argmin}_i\, y_i\langle \mathbf{w}, \mathbf{x}_i\rangle$ and update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_i \mathbf{x}_i$. The algorithm must have $f(\mathbf{w}^{(t)}) < 0$ after $\|\mathbf{w}^\star\|^2 R^2$ iterations. The algorithm is almost identical to the Batch Perceptron algorithm with two modifications. First, the Batch Perceptron updates with any example for which $y_i\langle \mathbf{w}^{(t)}, \mathbf{x}_i\rangle \leq 0$, while the current algorithm chooses the example for which $y_i\langle \mathbf{w}^{(t)}, \mathbf{x}_i\rangle$ is minimal. Second, the current algorithm employs the parameter η. However, it is easy to verify that the algorithm would not change if we fix $\eta = 1$ (the only modification is that $\mathbf{w}^{(t)}$ would be scaled by $1/\eta$).

4. **Variable step size:** The proof is very similar to the proof of Theorem 14.11. Details can be found, for example, in [1].

15 Support Vector Machines

1. Let $\mathcal{H}$ be the class of halfspaces in $\mathbb{R}^d$, and let $S = ((\mathbf{x}_i, y_i))_{i=1}^m$ be a linearly separable set. Let $\mathcal{G} = \{(\mathbf{w}, b) : \|w\| = 1, (\forall i \in [m])\ y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) > 0\}$. Our assumptions imply that this set is non-empty. Note that for every $(\mathbf{w}, b) \in \mathcal{G}$,

$$\min_{i\in[m]} y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) > 0 \ .$$

   On the contrary, for every $(\mathbf{w}, b) \notin \mathcal{G}$,

$$\min_{i\in[m]} y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \leq 0 \ .$$

   It follows that

$$\arg \max_{\substack{(\mathbf{w},b): \\ \|w\|=1}} \min_{i\in[m]} y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \subseteq \mathcal{G} \ .$$

Hence, solving the second optimization problem is equivalent to the following optimization problem:

$$\arg \max_{(\mathbf{w},b)\in\mathcal{G}} \min_{i\in[m]} y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \ .$$

Finally, since for every $(\mathbf{w}, b) \in \mathcal{G}$, and every $i \in [m]$, $y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) = |\langle \mathbf{w}, \mathbf{x}_i\rangle + b|$, we obtain that the second optimization problem is equivalent to the first optimization problem.

2. Let $S = ((x_i, y_i))_{i=1}^m \subseteq (\mathbb{R}^d \times \{-1, 1\})^m$ be a linearly separable set with a margin γ, such that $\max_{i\in[m]} \|x_i\| \ \le \rho$ for some $\rho > 0$. The margin assumption implies that there exists $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ such that $\|w\| = 1$, and
$$(\forall i \in [m]) \quad y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \ge \gamma \ .$$
Hence,
$$(\forall i \in [m]) \quad y_i(\langle \mathbf{w}/\gamma, \mathbf{x}_i\rangle + b/\gamma) \ge 1 \ .$$
Let $w^\star = \mathbf{w}/\gamma$. We have $\|\mathbf{w}^\star\| = 1/\gamma$. Applying Theorem 9.1, we obtain that the number of iterations of the perceptron algorithm is bounded above by $(\rho/\gamma)^2$.

3. The claim is wrong. Fix some integer $m > 1$ and $\lambda > 0$. Let $\mathbf{x}_0 = (0, \alpha) \in \mathbb{R}^2$, where $\alpha \in (0, 1)$ will be tuned later. For $k = 1, \ldots, m-1$, let $\mathbf{x}_k = (0, k)$. Let $y_0 = \ldots = y_{m-1} = 1$. Let $S = \{(\mathbf{x}_i, y_i) : i \in \{0, 1, \ldots, m-1\}\}$. The solution of hard-SVM is $\mathbf{w} = (0, 1/\alpha)$ (with value $1/\alpha^2$). However, if
$$\lambda \cdot 1 + \frac{1}{m}(1 - \alpha) \le \frac{1}{\alpha^2} \ ,$$
the solution of soft-SVM is $\mathbf{w} = (0, 1)$. Since $\alpha \in (0, 1)$, it suffices to require that $\frac{1}{\alpha^2} > \lambda + 1/m$. Clearly, there exists $\alpha_0 > 0$ s.t. for every $\alpha < \alpha_0$, the desired inequality holds. Informally, if α is small enough, then soft-SVM prefers to "neglect" $\mathbf{x}_0$.

4. Define the function $g : \mathcal{X} \to \mathbb{R}$ by $g(x) = \max_{y\in\mathcal{Y}} f(x, y)$. Clearly, for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$,
$$g(x) \ge f(x, y) \ .$$
Hence, for every $y \in \mathcal{Y}$,
$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y) = \min_{x\in\mathcal{X}} g(x) \ge \min_{x\in\mathcal{X}} f(x, y) \ .$$

Hence,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) \geq \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} f(x,y)\ .$$

16 Kernel Methods

1. Recall that $\mathcal{H}$ is finite, and let $\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\}$. For any two words u, v in Σ^*, we use the notation $u \preceq v$ if u is a substring of v. We will abuse the notation and write $h \preceq u$ if h is parameterized by a string v such that $v \preceq u$. Consider the mapping $\phi : \mathcal{X} \to \mathbb{R}^{|\mathcal{H}|+1}$ which is defined by

   $$\phi(x)[i] = \begin{cases} 1 & \text{if } i = |\mathcal{H}|+1 \\ 1 & \text{if } h_i \preceq x \\ 0 & \text{otherwise} \end{cases}$$

   Next, each $h_j \in \mathcal{H}$ will be associated with $\mathbf{w}(h_j) := (2\mathbf{e}_j; -1) \in \mathbb{R}^{|\mathcal{H}|+1}$. That is, the first $|\mathcal{H}|$ coordinates of $\mathbf{w}(h_j)$ correspond to the vector $\mathbf{e}_j$, and the last coordinate is equal to -1. Then, $\forall x \in \mathcal{X}$,

   $$\langle \mathbf{w}, \phi(x)\rangle = 2[\phi(x)_j] - 1 = h_j(x)\ . \tag{8}$$

2. In the Kernelized Perceptron, the weight vector $\mathbf{w}^{(t)}$ will not be explicitly maintained. Instead, our algorithm will maintain a vector $\boldsymbol{\alpha}^{(t)} \in \mathbb{R}^m$. In each iteration we update $\boldsymbol{\alpha}^{(t)}$ such that

   $$\mathbf{w}^{(t)} = \sum_{i=1}^{m} \alpha_i^{(t)} \psi(\mathbf{x}_i)\ . \tag{9}$$

   Assuming that Equation (9) holds, we observe that the condition

   $$\exists i \text{ s.t. } y_i \langle \mathbf{w}^{(t)}, \psi(\mathbf{x}_i)\rangle \leq 0$$

   is equivalent to the condition

   $$\exists i \text{ s.t. } y_i \sum_{j=1}^{m} \alpha_j^{(t)} K(\mathbf{x}_i, \mathbf{x}_j) \leq 0\ ,$$

   which can be verified while only accessing instances via the kernel function.

   We will now detail the update $\alpha^{(t)}$. At each time t, if the required update is $\mathbf{w}_{t+1} = \mathbf{w}_t + y_i \mathbf{x}_i$, we make the update

   $$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + y_i \mathbf{e}_i\ .$$

A simple inductive argument shows that Equation (9) is satisfied.

Finally, the algorithm returns $\alpha^{(T+1)}$. Given a new instance $\mathbf{x}$, the prediction is calculated using $\mathrm{sign}(\sum_{i=1}^{m} \alpha_i^{(T+1)} K(\mathbf{x}_i, \mathbf{x}))$.

3. The representer theorem tells us that the minimizer of the training error lies in $\mathrm{span}(\{\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_m)\})$. That is, the ERM objective is equivalent to the following objective:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \lambda \| \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i) \|^2 + \frac{1}{2m} \sum_{i=1}^{m} (\langle \sum_{j=1}^{m} \alpha_j \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \rangle - y_i)^2$$

Denoting the gram matrix by G, the objective can be rewritten as

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \lambda \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} + \frac{1}{2m} \sum_{i=1}^{m} (\langle \boldsymbol{\alpha}, G_{\cdot,i} \rangle - y_i)^2 \ . \tag{10}$$

Note that the objective (Equation (10)) is convex[15]. It follows that a minimizer can be obtained by differentiating Equation (10), and comparing to zero. Define $\lambda' = m \cdot \lambda$. We obtain

$$(\lambda' G + G G^T) \boldsymbol{\alpha} - G \mathbf{y} = 0$$

Since G is symmetric, this can be rewritten as

$$G(\lambda' I + G) \boldsymbol{\alpha} = G \mathbf{y} \ .$$

A sufficient (and necessary in case that G is invertible) condition for the above to hold is that

$$(\lambda' I + G) \boldsymbol{\alpha} = \mathbf{y} \ .$$

Since G is positive semi-definite and $\lambda' > 0$, the matrix $\lambda' I + G$ is positive definite, and thus invertible. We obtain that $\boldsymbol{\alpha}^* = (\lambda' I + G)^{-1} \mathbf{y}$ is a minimizer of our objective.

4. Define $\psi : \{1, \ldots, N\} \to \mathbb{R}^N$ by

$$\psi(j) = (\mathbf{1^j}; \mathbf{0^{N-j}}) \ ,$$

[15] The term $\frac{2}{m} \sum_{i=1}^{m} (\langle \alpha, G_{\downarrow i} \rangle - y_i)^2$ is simply the least square objective, and thus it is convex, as we have already seen. The Hessian of $\alpha^T G \alpha$ is G, which is positive semi-definite. Hence, $\alpha^T G \alpha$ is also convex. Our objective is a weighted sum, with non-negative weights, of the two convex terms above. Thus, it is convex.

where $\mathbf{1^j}$ is the vector in $\mathbb{R}^j$ with all elements equal to 1, and $\mathbf{0^{N-j}}$ is the zero vector in $\mathbb{R}^{N-j}$. Then, assuming the standard inner product, we obtain that $\forall (i,j) \in [N]^2$,

$$\langle \psi(i), \psi(j) \rangle = \langle (\mathbf{1^i}; \mathbf{0^{N-i}}), (\mathbf{1^j}; \mathbf{0^{N-j}}) \rangle = \min\{i,j\} = K(i,j) \ .$$

5. We will formalize our problem as an SVM (with kernels) problem. Consider the feature mapping $\phi : \mathcal{P}([d]) \to \mathbb{R}^d$ (where $\mathcal{P}([d])$ is the collection of all subsets of $[d]$), which is defined by

$$\phi(E) = \sum_{j=1}^{d} \mathbb{1}_{[j \in E]} \mathbf{e}_j \ .$$

In words, $\phi(E)$ is the indicator vector of E. A suitable kernel function $K : \mathcal{P}([d]) \times \mathcal{P}([d]) \to \mathbb{R}$ is defined by $K(E,E') = |E \cap E'|$. The prior knowledge of the manager implies that the optimal hypothesis can be written as a homogenous halfspace:

$$\mathbf{x} \mapsto \text{sign}(\langle 2\mathbf{w}, \phi(\mathbf{x}) \rangle - 1) \ ,$$

where $\mathbf{w} = \sum_{i \in I} \mathbf{e}_i$, where $I \subset [d]$, $|I| = k$, is the set of k relevant items. Furthermore, the halfspace defined by $(2\mathbf{w}, 1)$ has zero hinge-loss on the training set. Finally, we have that $\|(2\mathbf{w}, 1)\| = \sqrt{4k+1}$, and $\|(\phi(\mathbf{x}), 1)\| \leq \sqrt{s+1}$. We can therefore apply the general bounds on the sample complexity of soft-SVM, and obtain the following:

$$\begin{aligned} \mathbb{E}_S[L_D^{0-1}(A(S))] &\leq \min_{\mathbf{w}:\|\mathbf{w}\| \leq \sqrt{4k+1}} L_D^{\text{hinge}}(w) + \sqrt{\frac{8 \cdot (4k+1) \cdot (s+1)}{m}} \\ &= \sqrt{\frac{8 \cdot (4k+1) \cdot (s+1)}{m}} \ . \end{aligned}$$

Thus, the sample complexity is polynomial in $s, k, 1/\epsilon$,. Note that according to the regulations, each evaluation of the kernel function K can be computed in $O(s \log s)$ (this is the cost of finding the common items in both carts). Consequently, the computational complexity of applying soft-SVM with kernels is also polynomial in $s, k, 1/\epsilon$.

6. We will work with the label set $\{\pm 1\}$.

Observe that

$$
\begin{aligned}
h(\mathbf{x}) &= \mathrm{sign}(\|\psi(\mathbf{x}) - c_-\|^2 - \|\psi(\mathbf{x}) - c_+\|^2) \\
&= \mathrm{sign}(2\langle \psi(\mathbf{x}), c_+\rangle - 2\langle \psi(\mathbf{x}), c_-\rangle + \|c_-\|^2 - \|c_+\|^2) \\
&= \mathrm{sign}(2(\langle \psi(\mathbf{x}), \mathbf{w}\rangle + b)) \\
&= \mathrm{sign}(\langle \psi(\mathbf{x}), \mathbf{w}\rangle + b) \ .
\end{aligned}
$$

(b) Simply note that

$$
\begin{aligned}
\langle \psi(\mathbf{x}), \mathbf{w}\rangle &= \langle \psi(\mathbf{x}), c_+ - c_-\rangle \\
&= \frac{1}{m_+}\sum_{i:y_i=1}\langle \psi(\mathbf{x}), \psi(\mathbf{x}_i)\rangle + \frac{1}{m_-}\sum_{i:y_i=-1}\langle \psi(\mathbf{x}), \psi(\mathbf{x}_i)\rangle \\
&= \frac{1}{m_+}\sum_{i:y_i=1} K(\mathbf{x}, \mathbf{x}_i) + \frac{1}{m_-}\sum_{i:y_i=-1} K(\mathbf{x}, \mathbf{x}_i) \ .
\end{aligned}
$$

17 Multiclass, Ranking, and Complex Prediction Problems

1. Fix some $(\mathbf{x}, y) \in S$. By our assumption about S (and by the triangle inequality),

$$\|\mathbf{x} - \boldsymbol{\mu}_y\| \le r \quad , \quad (\forall y' \neq y)\ \|\mathbf{x} - \boldsymbol{\mu}_{y'}\| \ge 3r$$

Hence,

$$\|\mathbf{x} - \boldsymbol{\mu}_y\|^2 \le r^2 \quad , \quad (\forall y' \neq y)\ \|\mathbf{x} - \boldsymbol{\mu}_{y'}\|^2 \ge 9r^2$$

It follows that for every $y' \neq y$,

$$\|\mathbf{x}-\boldsymbol{\mu}_{y'}\|^2 - \|\mathbf{x}-\boldsymbol{\mu}_y\|^2 = 2\langle \boldsymbol{\mu}_y, \mathbf{x}\rangle - \|\boldsymbol{\mu}_y\|^2 - (2\langle \boldsymbol{\mu}'_y, \mathbf{x}\rangle - \|\boldsymbol{\mu}_{y'}\|^2) \ge 8r^2 > 0 \ .$$

Dividing by two, we obtain

$$\langle \boldsymbol{\mu}_y, \mathbf{x}\rangle - \frac{1}{2}\|\boldsymbol{\mu}_y\|^2 - (\langle \boldsymbol{\mu}'_y, \mathbf{x}\rangle - \frac{1}{2}\|\boldsymbol{\mu}_{y'}\|^2) \ge 4r^2 > 0 \ .$$

Define $\mathbf{w}$ as in the hint (that is, define $\mathbf{w} = [\mathbf{w}_1\mathbf{w}_2 \ldots \mathbf{w}_k] \in \mathbb{R}^{(n+1)k}$, where each $\mathbf{w}_i$ is defined by $\mathbf{w}_i = [\boldsymbol{\mu}_i, -\|\boldsymbol{\mu}_i\|^2/2]$). It follows that

$$\langle \mathbf{w}, \psi(\mathbf{x}, y)\rangle - \langle \mathbf{w}, \psi(\mathbf{x}, y')\rangle \ge 4r^2 > 0 \ .$$

Hence, $h_{\mathbf{w}}(\mathbf{x}) = y$, so $\ell(\mathbf{w}, (\mathbf{x}, y)) = 0$.

2. We will mimic the proof of the binary perceptron.

Proof. First, by the definition of the stopping condition, if the Perceptron stops it must have separated all the examples. Let $\mathbf{w}^\star$ be defined as in the theorem, and let $B = \|\mathbf{w}^\star\|$. We will show that if the Perceptron runs for T iterations, then we must have $T \leq (RB)^2$, which implies that the Perceptron must stop after at most $(RB)^2$ iterations.

The idea of the proof is to show that after performing T iterations, the cosine of the angle between $\mathbf{w}^\star$ and $\mathbf{w}^{(T+1)}$ is at least $\frac{\sqrt{T}}{RB}$. That is, we will show that,

$$\frac{\langle \mathbf{w}^\star, \mathbf{w}^{(T+1)} \rangle}{\|\mathbf{w}^\star\| \, \|\mathbf{w}^{(T+1)}\|} \geq \frac{\sqrt{T}}{RB} \ . \tag{11}$$

By the Cauchy-Schwartz inequality, the left-hand side of Equation (11) is at most 1. Therefore, Equation (11) would imply that

$$1 \geq \frac{\sqrt{T}}{RB} \quad \Rightarrow \quad T \leq (RB)^2 \ ,$$

which will conclude our proof.

To show that Equation (11) holds, we first show that $\langle \mathbf{w}^\star, \mathbf{w}^{(T+1)} \rangle \geq T$. Indeed, at the first iteration, $\mathbf{w}^{(1)} = (0, \ldots, 0)$ and therefore $\langle \mathbf{w}^\star, \mathbf{w}^{(1)} \rangle = 0$, while on iteration t, if we update using example $(\mathbf{x}_i, y_i)$ (while denoting by $y \in [k]$ a label which satisfies $\langle \mathbf{w}^{(t)}, \psi(\mathbf{x}_i, y_i) \rangle \leq \langle \mathbf{w}^{(t)}, \psi(\mathbf{x}_i, y) \rangle$), we have that:

$$\langle \mathbf{w}^\star, \mathbf{w}^{(t+1)} \rangle - \langle \mathbf{w}^\star, \mathbf{w}^{(t)} \rangle = \langle \mathbf{w}^\star, \psi(\mathbf{x}_i, y_i) - \psi(\mathbf{x}_i, y) \rangle \geq 1 \ .$$

Therefore, after performing T iterations, we get:

$$\langle \mathbf{w}^\star, \mathbf{w}^{(T+1)} \rangle = \sum_{t=1}^{T} \left(\langle \mathbf{w}^\star, \mathbf{w}^{(t+1)} \rangle - \langle \mathbf{w}^\star, \mathbf{w}^{(t)} \rangle \right) \geq T \ , \tag{12}$$

as required.

Next, we upper bound $\|\mathbf{w}^{(T+1)}\|$. For each iteration t we have that

$$\begin{aligned} \|\mathbf{w}^{(t+1)}\|^2 &= \|\mathbf{w}^{(t)} + \psi(\mathbf{x}_i, y_i) - \psi(\mathbf{x}_i, y)\|^2 \\ &= \|\mathbf{w}^{(t)}\|^2 + 2\langle \mathbf{w}^{(t)}, \psi(\mathbf{x}_i, y_i) - \psi(\mathbf{x}_i, y) \rangle + \|\psi(\mathbf{x}_i, y_i) - \psi(\mathbf{x}_i, y)\|^2 \\ &\leq \|\mathbf{w}^{(t)}\|^2 + R^2 \end{aligned} \tag{13}$$

where the last inequality is due to the fact that example i is necessarily such that $\langle \mathbf{w}^{(t)}, \psi(\mathbf{x}_i, y_i) - \psi(\mathbf{x}_i, y)\rangle \le 0$, and the norm of $\psi(\mathbf{x}_i, y_i) - \psi(\mathbf{x}_i, y)$ is at most R. Now, since $\|\mathbf{w}^{(1)}\|^2 = 0$, if we use Equation (13) recursively for T iterations, we obtain that

$$\|\mathbf{w}^{(T+1)}\|^2 \le TR^2 \quad \Rightarrow \quad \|\mathbf{w}^{(T+1)}\| \le \sqrt{T}\, R \ . \tag{14}$$

Combining Equation (9.3) with Equation (14), and using the fact that $\|\mathbf{w}^\star\| = B$, we obtain that

$$\frac{\langle \mathbf{w}^{(T+1)}, \mathbf{w}^\star\rangle}{\|\mathbf{w}^\star\|\,\|\mathbf{w}^{(T+1)}\|} \ge \frac{T}{B\,\sqrt{T}\,R} = \frac{\sqrt{T}}{B\,R} \ .$$

We have thus shown that Equation (9.2) holds, and this concludes our proof. □

3. The modification is done by adjusting the definition of the matrix M that is used in the dynamic programming procedure. Given a pair $(\mathbf{x}, y)$, we define

$$M_{s,\tau} = \max_{(y'_1,\ldots,y'_\tau):y'_\tau = s} \left(\sum_{t=1}^{\tau} \langle w, \phi(\mathbf{x}, y'_t, y'_{t-1})\rangle + \sum_{t=1}^{\tau} \delta(y'_t, y_t) \right) \ .$$

As in the OCR example, we have the following recursive structure that allows us to apply dynamic programming:

$$M_{s,\tau} = \max_{s'} \left(M_{s',\tau-1} + \delta(s, y_\tau) + \langle w, \phi(\mathbf{x}, s, s')\rangle \right) \ .$$

4. The idea: if the permutation $\hat{\mathbf{v}}$ that maximizes the inner product $\langle \mathbf{v}, \mathbf{y}\rangle$ doesn't agree with the permutation $\pi(\mathbf{y})$, then by replacing two indices of $\mathbf{v}$ we can increase the value of the inner product, and thus obtain a contradiction.

Let $V = \{\mathbf{v} \in [r]^r : (\forall i \ne j)\ v_i \ne v_j\}$ be the set of permutations of $[r]$. Let $\mathbf{y} \in \mathbb{R}^r$. Let $\hat{\mathbf{v}} = \mathrm{argmax}_{\mathbf{v}\in V} \sum_{i=1}^r v_i y_i$. We would like to prove that $\hat{\mathbf{v}} = \pi(\mathbf{y})$. By reordering if needed, we may assume w.l.o.g. that $y_1 \le \ldots \le y_r$. Then, we have to show that $\hat{v}_i = \pi(\mathbf{y})_i = i$ for all $i \in [r]$. Assume by contradiction that $\hat{\mathbf{v}} \ne (1, \ldots, r)$. Let $s \in [r]$ be the minimal index for which $v_s \ne s$. Then, $\hat{v}_s > s$. Let $t > s$ be an index such that $\hat{v}_t = s$. Define $\tilde{\mathbf{v}} \in V$ as follows:

$$\tilde{v}_i = \begin{cases} \hat{v}_t = s & i = s \\ \hat{v}_s & i = t \\ \hat{v}_i & \text{otherwise} \end{cases}$$

Then,

$$\begin{aligned}\sum_{i=1}^{r} \tilde{v}_i y_i &= \sum_{i=1}^{r} \hat{v}_i y_i + (\tilde{v}_s - \hat{v}_s) y_s + (\tilde{v}_t - \hat{v}_t) y_t \\ &= \sum_{i=1}^{r} \hat{v}_i y_i + (s - \hat{v}_s) s + (\hat{v}_s - s) t \\ &= \sum_{i=1}^{r} \hat{v}_i y_i + (s - \hat{v}_s)(s - t) \\ &> \sum_{i=1}^{r} \hat{v}_i y_i \ .\end{aligned}$$

Hence, we obtain a contradiction to the maximality of $\hat{\mathbf{v}}$.

5. (a) Averaging sensitivity and specificity, F1-score, F-β-score ($\theta = 0$): Let $\mathbf{y}' \in \mathbb{R}^r$, and let $V = \{-1, 1\}^r$. Let $\hat{\mathbf{v}} = \mathrm{argmax}_{v \in V} \sum_{i=1}^{r} v_i y'_i$. We would like to show that $\hat{\mathbf{v}} = (\mathrm{sign}(y'_i), \ldots, \mathrm{sign}(y'_r))$. Indeed, for every $\mathbf{v} \in V$, we have

$$\sum_{i=1}^{r} v_i y'_i \leq \sum_{i=1}^{r} |v_i y'_i| = \sum_{i=1}^{r} \mathrm{sign}(y'_i) y'_i \ .$$

(b) Recall at k, precision at k: The proof is analogous to the previous part.

18 Decision Trees

1. (a) Here is one simple (although not very efficient) solution: given h, construct a full binary tree, where the root note is ($x_1 = 0$?), and all the nodes at depth i are of the form ($x_{i+1} = 0$?). This tree has 2^d leaves, and the path from each root to the leaf is composed of the nodes ($x_1 = 0$?), ($x_2 = 0$?),...,($x_d = 0$). It is not hard to see that we can allocate one leaf to any possible combination of values for $x_1, x_2, \ldots, x_d$, with the leaf's value being $h(x) = h((x_1, x_2, \ldots, x_d))$.

(b) Our previous result implies that we can shatter the domain $\{0, 1\}^d$. Thus, the VC-dimension is exactly 2^d.

2. We denote by H the binary entropy.

(a) The algorithm first picks the root node, by searching for the feature which maximizes the information gain. The information gain[16] for feature 1 (namely, if we choose $x_1 = 0$? as the root) is:

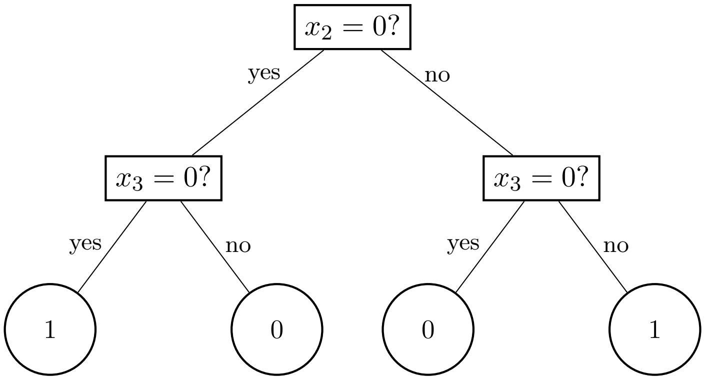
$$H\left(\frac{1}{2}\right) - \left(\frac{3}{4}H\left(\frac{2}{3}\right) + \frac{1}{4}H(0)\right) \approx 0.22$$

The information gain for feature 2, as well as feature 3, is:

$$H\left(\frac{1}{2}\right) - \left(\frac{1}{2}H\left(\frac{1}{2}\right) + \frac{1}{2}H\left(\frac{1}{2}\right)\right) = 0.$$

So the algorithm picks $x_1 = 0$? as the root. But this means that the three examples $((1,1,0),0)$,$((1,1,1),1)$, and $((1,0,0),1)$ go down one subtree, and no matter what question we'll ask now, we won't be able to classify all three examples perfectly. For instance, if the next question is $x_2 = 0$? (after which we must give a prediction), either $((1,1,0),0)$ or $((1,1,1),1)$ will be mislabeled. So in any case, at least one example will be mislabeled. Since we have 4 examples in the training set, it follows that the training error is at least $1/4$.

(b) Here is one such tree:



[16] Here we compute entropy where log is to the base of e. However, one can pick any other base, and the results will just change by a constant factor. Since we only care about which feature has the largest information gain, this won't affect which feature is picked.

19 Nearest Neighbor

1. We follow the hints for proving that for any $k \geq 2$, we have

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i:|C_i\cap S|<k} P[C_i]\right] \leq \frac{2rk}{m} \ .$$

   The claim in the first hint follows easily from the linearity of the expectation and the fact that $\sum_{i:|C_i\cap S|\leq k} \mathbb{P}[C_i] = \sum_{i=1}^{r} \mathbb{1}_{[|C_i|\leq k]} \mathbb{P}[C_i]$. The claim in the second hint follows directly from Chernoff's bounds. The next hint leaves nothing to prove. Finally, combining the fourth hint with the previous hints,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i:|C_i\cap S|<k} \mathbb{P}[C_i]\right] \leq \sum_{i=1}^{r} \max\{8/(me), 2k/m\} \ .$$

   Since $k \geq 2$, our proof is completed.

2. The claim in the first hint follows from

$$\mathbb{E}_{Z_1,\ldots,Z_k} \underset{y\sim p}{\mathbb{P}}[y \neq y'] = p\left(1 - \underset{Z_1,\ldots,Z_k}{\mathbb{P}}[p' > 1/2]\right) + (1-p)\left(\underset{Z_1,\ldots,Z_k}{\mathbb{P}}[p' > 1/2]\right) \ .$$

   The next hints leave nothing to prove.

3. We need to show that

$$\underset{y\sim p}{\mathbb{P}}[y \neq y'] - \underset{y\sim p'}{\mathbb{P}}[y \neq y'] \leq |p - p'| \qquad (15)$$

   Indeed, if $y' = 0$, then the left-hand side of Equation (15) equals $p - p'$. Otherwise, it equals $p' - p$. In both cases, Equation (15) holds.

4. Recall that $\pi_j(\mathbf{x})$ is the j-th NN of $\mathbf{x}$. We prove the first claim. The probability of misclassification is bounded above by the probability that $\mathbf{x}$ falls in a "bad cell" (i.e., a cell that does not contain k instances from the training set) plus the probability that x is misclassified given that $\mathbf{x}$ falls in a "good cell". The next hints leave nothing to prove.

20 Neural Networks

1. Let $\epsilon > 0$. Following the hint, we cover the domain $[-1,1]^n$ by disjoint boxes such that for every x, x' which lie in the same box, we have $|f(\mathbf{x}) - f(\mathbf{x}')| \leq \epsilon/2$. Since we only aim at approximating f to an accouracy of ϵ, we can pick an arbitrary point from each box. By picking the set of representative points appropriately (e.g., pick the center of each box), we can assume w.l.o.g. that f is defined over the discrete set $[-1+\beta, -1+2\beta, \ldots, 1]^d$ for some $\beta \in [0,2]$ and $d \in \mathbb{N}$ (which both depends on ρ and ϵ). From here, the proof is straightforward. Our network should have two hidden layers. The first layer has $(2/\beta)^d$ nodes which correspond to the intervals that make up our boxes. We can adjust the weights between the input and the hidden layer such that given an input $\mathbf{x}$, the output of each neuron is close enough to 1 if the corresponding coordinate of $\mathbf{x}$ lies in the corresponding interval (note that given a finite domain, we can approximate the indicator function using the sigmoid function). In the next layer, we construct a neuron for each box, and add an additional neuron which outputs the constant $-1/2$. We can adjust the weights such that the output of each neuron is 1 if $\mathbf{x}$ belongs to the corresponding box, and 0 otherwise. Finally, we can easily adjust the weights between the second layer and the output layer such that the desired output is obtained (say, to an accuracy of $\epsilon/2$).

2. Fix some $\epsilon \in (0,1)$. Denote by $\mathcal{F}$ the set of 1-Lipschitz functions from $[-1,1]^n$ to $[-1,1]$. Let $G = (V,E)$ with $|V| = s(n)$ be a graph such that the hypothesis class $\mathcal{H}_{V,E,\sigma}$, with σ being the sigomid activation function, can approximate every function $f \in \mathcal{F}$, to an accuracy of ϵ. In particular, every function that belongs to the set $\{f \in \mathcal{F} : (\forall x \in \{-1,1\}^n) \;\; f(\mathbf{x}) \in \{-1,1\}\}$ is approximated to an accuracy ϵ. Since $\epsilon \in (0,1)$, it follows that we can easily adapt the graph such that its size remains $\Theta(s(n))$, and $\mathcal{H}_{V,E,\sigma}$ contains all the functions from $\{-1,1\}^n$ to $\{-1,1\}$. We already noticed that in this case, $s(n)$ must be exponential in n (see Theorem 20.2).

3. Let $C = \{c_1, \ldots, c_m\} \subseteq \mathcal{X}$. We have

$$\begin{aligned}
|\mathcal{H}_C| &= |\{((f_1(\mathbf{c}_1), f_2(\mathbf{c}_2)), \ldots, (f_1(\mathbf{c}_m), f_2(\mathbf{c}_m))) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}| \\
&= |\{((f_1(\mathbf{c}_1), \ldots, f_1(\mathbf{c}_m)), (f_2(\mathbf{c}_1), \ldots, f_2(\mathbf{c}_m))) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}| \\
&= |\mathcal{F}_{1C} \times \mathcal{F}_{2C}| \\
&= |\mathcal{F}_{1C}| \cdot |\mathcal{F}_{2C}| \;.
\end{aligned}$$

It follows that $\tau_{\mathcal{H}}(m) = \tau_{\mathcal{F}_2}(m)\tau_{\mathcal{F}_1}(m)$.

4. Let $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\} \subseteq \mathcal{X}$. We have

$$\begin{aligned}
|\mathcal{H}_C| &= |\{f_2(f_1(\mathbf{c}_1)), \ldots, f_2(f_1(\mathbf{c}_m)) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}| \\
&= \left| \bigcup_{f_1 \in \mathcal{F}_1} \{(f_2(f_1(\mathbf{c}_1)), \ldots, f_2(f_1(\mathbf{c}_m)) : f_2 \in \mathcal{F}_2\} \right| \\
&\le |\mathcal{F}_{1C}| \cdot \tau_{\mathcal{F}_2}(m) \\
&\le \tau_{\mathcal{F}_1}(m)\tau_{\mathcal{F}_2}(m) \ .
\end{aligned}$$

   It follows that $\tau_{\mathcal{H}}(m) = \tau_{\mathcal{F}_2}(m)\tau_{\mathcal{F}_1}(m)$.

5. The hints provide most of the details. We skip to the conclusion part. By combining the graphs above, we obtain a graph $G = (V, E)$ with $V = O(n)$ such that the set $\{a_j^{(i)}\}_{(i,j)\in[n]^2}$ is shattered. Hence, the VC-dimension is at least n^2.

6. We reduce from the k-coloring problem. The construction is very similar to the construction for intersection of halfspaces. The same set of points (namely $\{e_1, \ldots, e_n\} \cup \{(e_i + e_j)/2 : \{i, j\} \in E, i < j\}$) is considered. Similarly to the case of intersection of halfspaces, it can be verified that the graph is k-colorable iff $\min_{h\in\mathcal{H}_{V,E,\mathrm{sign}}} L_s(h) = 0$. Hence, the k-coloring problem is reduced to the problem of minimizing the training error. The theorem is concluded.

21 Online Learning

1. Let $\mathcal{X} = \mathbb{R}^d$, and let $\mathcal{H} = \{h_1, \ldots, h_d\}$, where $h_j(\mathbf{x}) = \mathbb{1}_{[x_j=1]}$. Let $\mathbf{x}_t = \mathbf{e}_t$, $y_t = \mathbb{1}_{[t=d]}$, $t = 1, \ldots, d$. The Consistent algorithm might predict $p_t = 1$ for every $t \in [d]$. The number of mistakes done by the algorithm in this case is $d - 1 = |\mathcal{H}| - 1$.

2. Let $d \in \mathbb{N}$, and let $\mathcal{X} = [d]$ and let $\mathcal{H} = \{h_A : A \subseteq [d]\}$, where

$$h_A(x) = \mathbb{1}_{[x\in A]} \ .$$

   For $t = 1, 2, \ldots$, let $x_t = t$, $y_t = 1$ (i.e., the true hypothesis corresponds to the set $[d]$). Note that at every time t,

$$|\{h \in V_t : h(x_t) = 1\}| = |\{h \in V_t : h(x_t) = 0\}| \ ,$$

(there is a natural bijection between the two sets). Hence, Halving might predict with $\hat{y}_t = 0$ for every $t \in [d]$, and consequently err $d = \log|\mathcal{H}|$ times.

3. We claim that $M_{\text{Halving}}(\mathcal{H}) = 1$. Let $h_{j^\star}$ be the true hypothesis. We now characterize the (only) cases in which Halving might make a mistake:

   - It might misclassify $j^\star$.
   - It might misclassify $i \neq j^\star$ only if the current version space consists only of h_i and $h_{j^\star}$.

   In both cases, the resulting version space consists of only one hypothesis, namely, the right hypothesis $h_{j^\star}$. Thus, $M_{\text{Halving}}(\mathcal{H}) \leq 1$.

   To see that $M_{\text{Halving}}(\mathcal{H}) \geq 1$, observe that if $x_1 = j$, then Halving surely misclassifies x_1.

4. Denote by T the overall number of rounds. The regret of A is bounded as follows:

$$\begin{aligned}\sum_{m=1}^{\lceil \log T \rceil} \alpha\sqrt{2^m} &= \alpha \, \frac{1-\sqrt{2}^{\lceil \log T \rceil + 1}}{1-\sqrt{2}} \\ &\leq \alpha \frac{1-\sqrt{2T}}{1-\sqrt{2}} \\ &\leq \frac{\sqrt{2}}{\sqrt{2}-1}\alpha\sqrt{T} \ .\end{aligned}$$

5. First note that since r is chosen uniformly at random, we have

$$L_{\mathcal{D}}(h_r) = \frac{1}{T}\sum_{t=1}^{T} L_{\mathcal{D}}(h_t) \ .$$

   Hence, by the linearity of expectation,

$$\begin{aligned}\mathbb{E}[L_{\mathcal{D}}(h_r)] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[L_{\mathcal{D}}(h_t)] \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}_{[h_t(x_t)\neq y_t]}] \\ &\leq \frac{M_A(\mathcal{H})}{T} \ ,\end{aligned}$$

where the second equality follows from the fact that h_t only depends on $x_1, \ldots, x_{t-1}$, and x_t is independent of $x_1, \ldots x_{t-1}$.

22 Clustering

1. Let $\mathbb{R}^2 \subseteq X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, where $x_1 = (0,0), x_2 = (0,2), x_3 = (2\sqrt{t}, 0), x_4 = (2\sqrt{t}, 2)$. Let d be the metric induced by the ℓ^2 norm. Finally, let $k = 2$.

   Suppose that the k-means chooses $\mu_1 = \mathbf{x}_1, \mu_2 = \mathbf{x}_2$. Then, in the first iteration it associates $\mathbf{x}_1, \mathbf{x}_3$ with the center $\mu_1 = (\sqrt{t}, 0)$. Similarly, it associates $\mathbf{x}_2, \mathbf{x}_4$ with the center $\mu_2 = (\sqrt{t}, 2)$. This is a convergence point of the algorithm. The value of this solution is $4t$. The optimal solution associates $\mathbf{x}_1, \mathbf{x}_2$ with the center $(0, 1)$, while $\mathbf{x}_3, \mathbf{x}_4$ are associated with the center $(2\sqrt{t}, 1)$. The value of this solution is 4.

2. The K-means solution is:

   $$\mu_1 = 2, C_1 = \{1, 2, 3\}, \mu_2 = 4, C_2 = \{3, 4\} \ .$$

   The value of this solution is $2 \cdot 1 = 2$. The optimal solution is

   $$\mu_1 = 1.5, C_1 = \{1, 2\}, \mu_2 = 3.5, C_2 = \{3, 4\}$$

   whose value is $1 = 4 \cdot (1/2)^2$.

3. For every $j \in [k]$, let $r_j = d(\mu_j, \mu_{j+1})$. Following the notation in the hint, it is clear that $r_1 \geq r_2 \geq \ldots \geq r_k \geq r$. Furthermore, by definition of μ_{k+1} it holds that

   $$\max_{j \in [k]} \max_{x \in \hat{C}_j} d(x, \mu_j) \leq r \ .$$

   The triangle inequality implies that

   $$\max_{j \in [k]} \operatorname{diam}(\hat{C}_j) \leq 2r \ .$$

   The pigeonhole principle implies now that at least 2 of the points $\mu_1, \ldots, \mu_{k+1}$ lie in the same cluster in the optimal solution. Thus,

   $$\max_{j \in [k]} \operatorname{diam}(C_j^*) \geq r \ .$$

4. Let $k = 2$. The idea is to pick elements in two close disjoint balls, say in the plane, and another distant point. The k-diam optimal solution would be to separate the distant point from the two balls. If the number of points in the balls is large enough, then the optimal center-based solution would be to separate the balls, and associate the distant point with its closest ball.

   Here are the details. Let $\mathcal{X}' = \mathbb{R}^2$ with the metric d which is induced by the ℓ_2-norm. A subset $\mathcal{X} \subseteq \mathcal{X}'$ is constructed as follows: Let $m > 0$, and let $\mathcal{X}_1$ be set of m points which are evenly distributed on the sphere of the ball $B_1((2,0))$ (a ball of radius 1 around $(2,0)$) . Similarly, let $\mathcal{X}_2$ be a set of m points which are evenly distributed on the sphere of $B_1((-2,0))$. Finally, let $\mathcal{X}_3 = \{(0,y)\}$ for some $y > 0$, and set $\mathcal{X} = \uplus_{i=1}^{3}\mathcal{X}_i$. Fix any monotone function $f : \mathbb{R}_+ \to \mathbb{R}_+$. We note that for large enough m, an optimal center-based solution must separate between $\mathcal{X}_1$ and $\mathcal{X}_2$, and associate the point $(0,y)$ with the nearest cluster. However, for large enough y, an optimal k-diam solution would be to separate $\mathcal{X}_3$ from $\mathcal{X}_1 \uplus \mathcal{X}_2$.

5. (a) Single Linkage with fixed number of clusters:
      1. Scale invariance: Satisfied. multiplying the weights by a positive constant does not affect the order in which the clusters are merged.
      2. Richness: Not satisfied. The number of clusters is fixed, and thus we can not obtain all possible partitions of $\mathcal{X}$.
      3. Consistency: Satisfied. Assume by contradiction that d, d' are two metrics which satisfy the required properties, but $F(\mathcal{X}, d') \neq F(\mathcal{X}, d)$. Then, there are two points x, y which belong to the same cluster in $F(x, d')$, but belong to different clusters in $F(\mathcal{X}, d)$ (we rely here on the fact that the number of clusters is fixed). Since d' assigns to this pair a larger value than d (and also assigns smaller values to pairs which belong to the same cluster), this is impossible.

   (b) Single Linkage with fixed distance upper bound:
      1. Scale invariance: Not satisfied. In particular, by multiplying by an appropriate scalar, we obtain the trivial clustering (each cluster consists of a single data point).
      2. Richness: Satisfied. Easily seen.
      3. Consistency: Satisfied. The argument is almost identical to the argument in the previous part: Assume by contradiction

that d, d' are two metrics which satisfy the required properties, but $F(\mathcal{X}, d') \neq F(\mathcal{X}, d)$. Then, either there are two points x, y which belong to the same cluster in $F(x, d')$, but belong to different clusters in $F(\mathcal{X}, d)$ or there are two points x, y which belong to different clusters in $F(x, d')$, but belong to the same cluster in $F(\mathcal{X}, d)$. Using the relation between d and d' and the fact that the threshold is fixed, we obtain that both of these events are impossible.

(c) Consider the scaled distance upper bound criterion (we set the threshold r to be $\alpha \max\{d(x, y) : x, y \in \mathcal{X}\}$. Let us check which of the three properties are satisfied:

i. Scale invariance: Satisfied. Since the threshold is scale-invariant, multiplying the metric by a positive constant doesn't change anything.

ii. Richness: Satisfied. Easily seen.

iii. Consistency: Not Satisfied. Let $\alpha = 0.75$. Let $\mathcal{X} = \{x_1, x_2, x_3\}$, and equip it with the metric d which is defined by: $d(x_1, x_2) = 3, d(x_2, x_3) = 7, d(x_1, x_3) = 8$. The resulting partition is $\mathcal{X} = \{x_1, x_2\}, \{x_3\}$. Now we define another metric d' by: $d(x_1, x_2) = 3, d(x_2, x_3) = 7, d(x_1, x_3) = 10$. In this case the resulting partition is $\mathcal{X} = \mathcal{X}$. Since d' satisfies the required properties, but the partition is not preserved, it follows that consistency is not satisfied.

Summarizing the above, we deduce that any pair of properties can be attained by one of the single linkage algorithms detailed above.

6. It is easily seen that Single Linkage with fixed number of clusters satisfies the k-richness property, thus it satisfies all the mentioned properties.

23 Dimensionality Reduction

1. (a) A fundamental theorem in linear algebra states that if V, W are finite dimensional vector spaces, and let T be a linear transformation from V to W, then the image of T is a finite-dimensional subspace of W and

$$\dim(V) = \dim(\text{null}(T)) + \dim(\text{image}(T)).$$

We conclude that $\dim(\text{null}(A)) \geq 1$. Thus, there exists $\mathbf{v} \neq \mathbf{0}$ s.t. $A\mathbf{v} = A\mathbf{0} = \mathbf{0}$. Hence, there exists $\mathbf{u} \neq \mathbf{v} \in \mathbb{R}^n$ such that $A\mathbf{u} = A\mathbf{v}$.

(b) Let f be (any) recovery function. Let $\mathbf{u} \neq \mathbf{v} \in \mathbb{R}^n$ such that $A\mathbf{u} = A\mathbf{v}$. Hence, $f(A\mathbf{u}) = f(A\mathbf{v})$, so at least one of the vectors $\mathbf{u}, \mathbf{v}$ is not recovered.

2. The hint provides all the details.

3. For simplicity, assume that the feature space is of finite dimension[17]. Let X be the matrix whose j-th column is $\psi(\mathbf{x}_j)$. Instead of computing the eigendecomposition of the matrix XX^T (and returning the n leading eigenvectors), we will compute the spectral decomposition of X^TX, and proceed as discussed in the section named "A more Efficient Solution for the case $d \gg m$". Note that $(X^TX)_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, thus the eigendecomposition of $X^\top X$ can be found in polynomial time.

   Let V be the matrix whose columns are the n leading eigenvectors of X^TX, and let D be a diagonal $n \times n$ matrix whose diagonal consists of the corresponding eigenvalues. Denote by U be the matrix whose columns are the n leading eigenvectors of $XX^\top$. We next show how to project the data without maintaining the matrix U. For every $\mathbf{x} \in \mathcal{X}$, the projection $U^\top\phi(x)$ is calculated as follows:

$$U^T\phi(\mathbf{x}) = D^{-1/2}V^TX^T\phi(\mathbf{x}) = D^{-1/2}V^T \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}) \\ \vdots \\ K(\mathbf{x}_m, \mathbf{x}) \end{pmatrix} .$$

4. (a) Note that for every unit vector $\mathbf{w} \in \mathbb{R}^d$, and every $i \in [m]$,

$$(\langle \mathbf{w}, \mathbf{x}_i \rangle)^2 = \text{tr}(\mathbf{w}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w}) \ .$$

   Hence, the optimization problem here coincides with the optimization problem in Equation (23.4) (assuming $n = 1$), which in turn, is equivalent to the objective of PCA. Hence, the optimal solution of our variance maximization problem is the first principal vector of $x_1, \ldots, x_m$.

[17] The proof in the general case is very similar but requires concepts from functional analysis.

(b) Following the hint, we obtain the following optimization problem:

$$\mathbf{w}^\star = \operatorname*{argmax}_{\mathbf{w}:\|\mathbf{w}\|=1\,,\,\langle \mathbf{w},\mathbf{w}_1\rangle=0} \frac{1}{m}\sum_{i=1}^m (\langle \mathbf{w}, \mathbf{x}_i\rangle)^2 = \operatorname*{argmax}_{\mathbf{w}:\|w\|=1\,,\,\langle \mathbf{w},\mathbf{w}_1\rangle=0} \operatorname{tr}(\mathbf{w}^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top \mathbf{w})\ .$$

Recall that the PCA problem in the case $n = 2$ is equivalent to finding a unitary matrix $W \in \mathbb{R}^{d\times 2}$ such that

$$W^\top \frac{1}{m}\sum_{i=1}^m x_i x_i^\top W$$

is maximized. Denote by $\mathbf{w}_1, \mathbf{w}_2$ the columns of the optimal matrix W. We already now that $\mathbf{w}_1$, and $\mathbf{w}_2$ are the two first principal vectors of $x_1, \ldots, x_m$. Note that

$$W^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top W = \mathbf{w}_1^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i x_i^\top \mathbf{w}_1 + \mathbf{w}_2^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top \mathbf{w}_2\ .$$

Since $\mathbf{w}^\star$ and $\mathbf{w}_1$ are orthonormal, we obtain that

$$\mathbf{w}_1^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i x_i^\top \mathbf{w}_1 + \mathbf{w}_2^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top \mathbf{w}_2 \geq \mathbf{w}_1^\top \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i x_i^\top \mathbf{w}_1 + \mathbf{w}^{\star\top} \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top \mathbf{w}^\star\ ,$$

We conclude that $\mathbf{w}^\star = \mathbf{w}_2$.

5. It is instructive to motivate the SVD through the *best-fit subspace* problem. Let $A \in \mathbb{R}^{m\times d}$. We denote the i-th row of A by A_i. Given $k \leq r = \operatorname{rank}(A)$, the best-fit k-dimensional subspace problem (w.r.t. A) is to find a subspace $V \subseteq \mathbb{R}^m$ of dimension at most k which minimizes the expression

$$\sum_{i=1}^m d(A_i, V)^2\ ,$$

where $d(A_i, V)$ is the distance between A_i and its nearest point in V. In words, we look for a k-dimensional subspace which gives a best approximation (in terms of distances) to the row space of A. We will prove now that every $m \times d$ matrix admits SVD which also induces an optimal solution to the best-fit subspace problem.

Theorem 23.1. *(The SVD Theorem) Let $A \in \mathbb{R}^{m \times d}$ be a matrix of rank r. Define the following vectors:*

$$\begin{aligned}
\mathbf{v}_1 &= \arg \max_{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\|=1} \|A\mathbf{v}\| \\
\mathbf{v}_2 &= \arg \max_{\substack{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\|=1 \\ \langle \mathbf{v}, \mathbf{v}_1 \rangle = 0}} \|A\mathbf{v}\| \\
&\vdots \\
\mathbf{v}_r &= \arg \max_{\substack{v \in \mathbb{R}^n : \|v\|=1 \\ \forall i<r, \ \langle \mathbf{v}, \mathbf{v}_i \rangle = 0}} \|A\mathbf{v}\|
\end{aligned}$$

(where ties are broken arbitrarily). Then, the following statements hold:

(a) The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are (right) singular vectors which correspond to the r top singular values defined by

$$\sigma_1 = \|A\mathbf{v}_1\| \geq \sigma_2 = \|A\mathbf{v}_2\| \geq \ldots \geq \sigma_r = \|A\mathbf{v}_r\| > 0 \ .$$

(b) It follows that the corresponding left singular vectors are defined by

$$u_i = \frac{A\mathbf{v}_i}{\|A\mathbf{v}_i\|} \ .$$

(c) Define the matrix $B = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Then, B forms a SVD of A.

(d) For every $k \leq r$ denote by V_k the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Then, V_k is a best-fit k-dimensional subspace of A.

Before proving the SVD theorem, let us explain how it gives an alternative proof to the optimality of PCA. The derivation is almost immediate. Simply use Lemma C.4 to conclude that if V_k is a best-fit subspace, then the PCA solution forms a projection matrix onto V_k.

Proof.
The proof is divided into 3 steps. In steps 1 and 2, we will prove by induction that V_k solves a best-fit k-dimensional subspace problem, and that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are right singular vectors. In the third step we will conclude that A admits the SVD decomposition specified above.

Step 1: Consider the case $k = 1$. The best-fit 1-dimensional subspace problem can be formalized as follows:

$$\min_{\|\mathbf{v}\|=1} \sum_{i=1}^{m} \|A_i - A_i\mathbf{v}\mathbf{v}^T\|^2 \ . \tag{16}$$

The pythagorean theorem implies that for every $i \in [d]$,

$$\begin{aligned} \|A_i\|^2 &= \|A_i - A_i\mathbf{v}\mathbf{v}^T\|^2 + \|A_i\mathbf{v}\mathbf{v}^T\|^2 \\ &= \|A_i - A_i\mathbf{v}\mathbf{v}^T\|^2 + (A_i\mathbf{v})^2 \ , \end{aligned}$$

Therefore,

$$\begin{aligned} \operatorname*{argmin}_{\|\mathbf{v}\|=1} \sum_{i=1}^{m} \|A_i - A_i\mathbf{v}\mathbf{v}^T\|^2 &= \operatorname*{argmin}_{\|\mathbf{v}\|=1} \sum_{i=1}^{m} \|A_i\|^2 - (A_i\mathbf{v})^2 \\ &= \operatorname*{argmax}_{\mathbf{v}:\|\mathbf{v}\|=1} \sum_{i=1}^{m} (A_i\mathbf{v})^2 \\ &= \operatorname*{argmax}_{\mathbf{v}:\|\mathbf{v}\|=1} \|A\mathbf{v}\|^2 \\ &= v_1 \ . \end{aligned} \tag{17}$$

We conclude that V_1 is the best-fit 1-dimensional problem.

Following Lemma C.4, in order to prove that $\mathbf{v}_1$ is the leading singular vector of A, it suffices to prove that $\mathbf{v}_1$ is a leading eigenvector of $A^\top A$. That is, solving Equation (17) is equivalent to finding a vector $\hat{\mathbf{w}} \in \mathbb{R}^m$ which belongs to $\operatorname{argmax}_{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|=1} \|A^\top A\mathbf{w}\|$. Let $\lambda_1 \geq \ldots \geq \lambda_d$ be the leading eigenvectors of $A^\top A$, and let $w_1, \ldots, w_d$ be the corresponding orthonormal eigenvectors. Remember that if $\mathbf{w}$ is a unit vector in $\mathbb{R}^m$, then it has a representation of the form $\sum_{i=1}^{d} \alpha_i\mathbf{w}_i$, where $\sum \alpha_i^2 = 1$. Hence, solving Equation (17) is equivalent to solving

$$\begin{aligned} \max_{\substack{\alpha_1,\ldots,\alpha_d: \\ \sum\alpha_i^2=1}} \|A\sum_{i=1}^{d} \alpha_i\mathbf{w}_i\|^2 &= \max_{\substack{\alpha_1,\ldots,\alpha_d: \\ \sum\alpha_i^2=1}} \sum_{i=1}^{d} \alpha_i\mathbf{w}_i^T A^T A \sum_{j=1}^{d} \alpha_j\mathbf{w}_j \\ &= \max_{\substack{\alpha_1,\ldots,\alpha_d: \\ \sum\alpha_i^2=1}} \sum_{i=1}^{d} \alpha_i\mathbf{w}_i^T \sum_{s=1}^{d} \lambda_s\mathbf{w}_s\mathbf{w}_s^T \sum_{j=1}^{d} \alpha_j\mathbf{w}_j \\ &= \max_{\substack{\alpha_1,\ldots,\alpha_d: \\ \sum\alpha_i^2=1}} \sum_{i=1}^{d} \alpha_i^2\lambda_i \\ &= \lambda_1 \ . \end{aligned} \tag{18}$$

We have thus proved that the optimizer of the 1-dimensional subspace problem is $\mathbf{v}_1$, which is the leading right singular vector of A. It follows from Lemma C.4 that its corresponding singular value is $\sqrt{\lambda_1} = \|A\mathbf{v}_1\| = \sigma_1$, which is positive since the rank of A (which equals to the rank of $A^\top A$) is r. Also, it can be seen that the corresponding left singular vector is $\mathbf{u}_i = \frac{Av_i}{\|A\mathbf{v}_i\|}$.

Step 2: We proceed to the induction step. Assume that V_{k-1} is optimal. We will prove that V_k is optimal as well. Let V_k' be an optimal subspace of dimension k. We can choose an orthonormal basis for V_k', say $z_1, \ldots, z_k$, such that z_k is orthogonal to V_{k-1}. By the definition of V_k' (and the pythagorean theorem), we have that $\sum_{i=1}^k \|Az_i\|^2$ is maximized among all the sets of k orthonormal vectors. By the optimality of V_{k-1}, we have

$$\sum_{i=1}^{k} \|Az_i\|^2 \le \|A\mathbf{v}_1\|^2 + \ldots + \|A\mathbf{v}_{k-1}\|^2 + \|Az_k\|^2 \ .$$

Hence, we can assume w.l.o.g. that $V_k' = \mathrm{Span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}, z_k\})$. Thus, since z_k is orthogonal to V_{k-1}, we obtain that $z_k \in \arg\max_{\substack{v\in\mathbb{R}^d:\|v\|=1 \\ \forall i<r,\ \langle v,v_i\rangle=0}} \|A\mathbf{v}\|$. Hence, V_k is optimal as well. Similar observation to Equation (18) shows that $\mathbf{v}_k$ is a singular vector.

Step 3: We know that V_r is a best-fit r-dimensional subspace w.r.t. A. Since A is of rank r, we obtain that V_r coincides with the row space of A. Clearly, V_k is also equal to the row space of $B = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ forms an orthonormal basis for this space, and $A(\mathbf{v}_i) = B(\mathbf{v}_i)$ for every $i \in [k]$, we obtain that $A = B$. □

6. (a) Following the hint, we will apply the JL lemma on every vector in the sets $\{\mathbf{u} - \mathbf{v} : \mathbf{u}, \mathbf{v} \in Q\}$, $\{\mathbf{u} + \mathbf{v} : \mathbf{u}, \mathbf{v} \in Q\}$. Fix a pair $\mathbf{u}, \mathbf{v} \in Q$. By applying the union bound (over $2|Q|^2$ events), and setting n as described in the question, we obtain that with probability of at least $1 - \delta$, for every $\mathbf{u}, \mathbf{v} \in Q$, we have

$$\begin{aligned} 4\langle W\mathbf{u}, W\mathbf{v}\rangle &= \|W(\mathbf{u}+\mathbf{v})\|^2 - \|W(\mathbf{u}-\mathbf{v})\|^2 \\ &\ge (1-\epsilon)\|\mathbf{u}+\mathbf{v}\|^2 - (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \\ &= 4\langle \mathbf{u}, \mathbf{v}\rangle - 2\epsilon(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2) \ge 4\langle \mathbf{u}, \mathbf{v}\rangle - 4\epsilon \end{aligned}$$

The last inequality follows from the assumption that $u, v \in B(\mathbf{0}, 1)$. The proof of the other direction is analogous.

(b) We will repeat the analysis from above, but this time we will apply the union bound over $2m+1$ events. Also, we will set $\epsilon = \gamma/2$. Hence, the lower dimension is chosen to be $n = \left\lceil \frac{24 \log((4m+2)/\delta)}{\epsilon^2} \right\rceil$.

- We count (twice) each pair in the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \times \{\mathbf{w}^\star\}$, where $\mathbf{w}^\star$ is a unit vector that separates the original data with margin γ. Setting $\delta' = \delta/(2|Q|+1)$, we obtain from above that for every $i \in [m]$, with probability at least $1-\delta$, we have

$$|\langle \mathbf{w}^\star, \mathbf{x}_i \rangle - \langle W\mathbf{w}^\star, W\mathbf{x}_i \rangle| \leq \gamma/2.$$

Since $y_i \in \{-1, 1\}$, we have

$$|y_i \langle \mathbf{w}^\star, \mathbf{x}_i \rangle - y_i \langle W\mathbf{w}^\star, W\mathbf{x}_i \rangle| \leq \gamma/2.$$

Hence,

$$y_i \langle W\mathbf{w}^\star, W\mathbf{x}_i \rangle \geq y_i \langle \mathbf{w}^\star, \mathbf{x}_i \rangle - \gamma/2 \geq \gamma - \gamma/2 = \gamma/2 \ .$$

- We apply the JL bound once again to obtain that

$$\|W\mathbf{w}^\star\|^2 \in [1-\epsilon, 1+\epsilon] \ .$$

Therefore, assuming $\gamma \in (0, 3)$ (hence, $\epsilon \in (0, 3/2)$), we obtain that for every $i \in [m]$, with probability at least $1-\delta$, we have

$$y_i \left(\frac{W\mathbf{w}^\star}{\|W\mathbf{w}^\star\|} \right)^T (W\mathbf{x}_i) \geq \frac{1}{\|W\mathbf{w}^\star\|} \cdot \frac{\gamma}{2} \geq \frac{1}{\sqrt{1+\epsilon}} \cdot \frac{\gamma}{2} \geq \frac{\gamma}{4} \ .$$

24 Generative Models

1. We show that $(\mathbb{E}[\hat{\sigma}])^2 = \frac{M-1}{M}\sigma^2$, thus concluding that σ is biased. We remark that our proof holds for every random variable with finite variance. Let $\mu = \mathbb{E}[x_1] = \ldots = \mathbb{E}[x_m]$ and let $\mu_2 = \mathbb{E}[x_1^2] = \ldots =$

$\mathbb{E}[x_m^2]$. Note that for $i \neq j$, $E[x_i x_j] = \mathbb{E}[x_i]\mathbb{E}[x_j] = \mu^2$.

$$
\begin{aligned}
(\mathbb{E}[\hat{\sigma}])^2 &= \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}[x_i^2] - \frac{2}{m}\sum_{j=1}^{m}\mathbb{E}[x_i x_j] + \frac{1}{m^2}\sum_{j,k}\mathbb{E}[x_j x_k]\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\mu_2 - \frac{2}{m}((m-1)\mu^2 + \mu_2) + \frac{1}{m^2}(m\mu_2 + m(m-1)\mu^2)\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{m-1}{m}\mu_2 - \frac{m-1}{m}\mu^2\right) \\
&= \frac{1}{m}\frac{m(m-1)}{m}(\mu_2 - \mu^2) \\
&= \frac{m-1}{m}\sigma^2 \ .
\end{aligned}
$$

2. (a) We[18] simply add to the training sequence one positive example and one negative example, denoted x_{m+1} and x_{m+2}, respectively. Note that the corresponding probabilities are θ and $1-\theta$. Hence, minimizing the RLM objective w.r.t. the original training sequence is equivalent to minimizing the ERM w.r.t. the extended training sequence. Therefore, the maximum likelihood estimator is given by

$$\hat{\theta} = \frac{1}{m+2}\left(\sum_{i=1}^{m+2} x_i\right) = \frac{1}{m+2}\left(1 + \sum_{i=1}^{m} x_i\right) \ .$$

(b) Following the hint we bound $|\hat{\theta} - \theta^\star|$ as follows.

$$|\theta - \theta^\star| \leq |\hat{\theta} - \mathbb{E}[\hat{\theta}]| + |\mathbb{E}[\hat{\theta}] - \theta^\star| \ .$$

Next we bound each of the terms in the RHS of the last inequality. For this purpose, note that

$$\mathbb{E}[\hat{\theta}] = \frac{1 + m\theta^\star}{m+2} \ .$$

Hence, we have the following two inequalities.

$$|\hat{\theta} - \mathbb{E}[\hat{\theta}]| = \frac{m}{m+2}\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \theta^\star\right| \ .$$

[18] Errata: The distribution is a Bernoulli distribution parameterized by θ

$$|\mathbb{E}[\hat{\theta}] - \theta^\star| = \frac{1-2\theta^\star}{m+2} \leq 1/(m+2) \ .$$

Applying Hoeffding's inequality, we obtain that for any $\epsilon > 0$,

$$\mathbb{P}[|\theta - \theta^\star| \geq 1/(m+2) + \epsilon/2] \leq 2\exp(-m\epsilon^2/2) \ .$$

Thus, given a confidence parameter δ, the following bound holds with probability of at least $1 - \delta$'

$$|\theta - \theta^\star| \leq O\left(\sqrt{\frac{log(1/\delta)}{m}}\right) = \tilde{O}(1/\sqrt{m}) \ .$$

(c) The true risk of $\hat{\theta}$ minus the true risk of θ^* is the relative entropy between them, namely, $D_{RE}(\theta^*||\hat{\theta}) = \theta^*[\log(\theta^*) - \log(\hat{\theta})] + (1 - \theta^*)[\log(1-\theta^*) - \log(1-\hat{\theta})]$. We may assume w.l.o.g. that $\theta^* < 0.5$. The derivative of the log function is $1/x$. Therefore, within $[0.25, 1]$, the log function is 4-Lipschitz, so we can easily bound the second term. So, let us focus on the first term. By the fact that $\hat{\theta} > 1/(m+2)$, we have that the first term is bounded by:

$$\theta^* \log(\theta^*) + \theta^* \log(m+2)$$

We consider two cases:
Case 1: $\theta^* < \sqrt{\epsilon}$. We can bound the first term by $\tilde{O}(m^{-1/4})$.
Otherwise: With high probability, $\hat{\theta} > \theta^* - \epsilon > 0.5\sqrt{\epsilon}$. So, $\log(\theta^*) - \log(\hat{\theta}) < (2/\sqrt{\epsilon}) \cdot \epsilon = 2\sqrt{\epsilon}$. All in all, we got a bound of $\tilde{O}(m^{-1/4})$.

3. Recall that the M step of soft k-means involves a maximization of the expected log-likelihood over the probability simplex in $\mathbb{R}^k$, $\{\mathbf{c} \in \mathbb{R}^k_+ : \sum_{i=1}^k c_i = 1\}$ (while the other parameters are fixed). Denoting $\mathcal{P}_{\theta^{(t)}}$ by $\mathcal{P}$, and defining $\nu_y = \sum_{i=1}^m \mathcal{P}[Y = y|X = \mathbf{x}_i]$ for every $y \in [k]$, we observe that this maximization is equivalent to the following optimization problem:

$$\max_{\mathbf{c}\in\mathbb{R}^k} \sum_{y=1}^k \nu_y \log(c_y) \ \text{ s.t. } \ c_y \geq 0, \sum_y c_y = 1 \ .$$

(a) First, observe that $\nu_y \geq 0$ for every y. Next, note that

$$\sum_y c^\star_y = \frac{\sum_y \nu_y}{\sum_y \nu_y} = 1 \ .$$

(b) Note that

$$\begin{aligned} -D_{RE}(c^\star || c) &= \sum_y c^\star_y \log(c_y / c^\star_y) \\ &= Z_1[\sum_y \nu_y \log(c_y) + Z_2] \ , \end{aligned}$$

where Z_1 is a positive constant and Z_2 is another constant (Z_1, Z_2 depend only on ν_y, which is fixed). Hence, our optimization problem is equivalent to the following optimization problem:

$$\min_{\mathbf{c} \in \mathbb{R}^k} \sum_{y=1}^{k} D_{RE}(\mathbf{c}^\star || \mathbf{c}) \quad \text{s.t.} \quad c_y \geq 0, \sum_y c_y = 1 \ . \tag{19}$$

(c) It is a well-known fact (in information theory) that $\mathbf{c}^\star$ is the minimizer of Equation (19).

25 Feature Selection

1. Given $a, b \in \mathbb{R}$, define $a' = a, b' = b + a\bar{v} - \bar{y}$. Then,

$$\|a\mathbf{v} + b - \mathbf{y}\|^2 = \|a'(\mathbf{v} - \bar{v}) + b' - (\mathbf{y} - \bar{y})\|^2$$

Since the last equality holds also for a, b which minimize the left hand-side, we obtain that

$$\min_{a,b \in \mathbb{R}} \|a\mathbf{v} + b - \mathbf{y}\|^2 \geq \min_{a,b \in \mathbb{R}} \|a(\mathbf{v} - \bar{v}) + b - (\mathbf{y} - \bar{y})\|^2 \ .$$

The reversed inequality is proved analogously. Given $a, b \in \mathbb{R}$, let $a' = a, b' = b - a\bar{v} + \bar{y}$. Then,

$$\|a(\mathbf{v} - \bar{v}) + b - (\mathbf{y} - \bar{y})\|^2 = \|a'\mathbf{v} + b' - \mathbf{y}\|^2 \ .$$

Since the last equality holds also for a, b which minimize the left hand-side, we obtain that

$$\min_{a,b \in \mathbb{R}} \|a(\mathbf{v} - \bar{v}) + b - (\mathbf{y} - \bar{y})\|^2 \geq \min_{a,b \in \mathbb{R}} \|a\mathbf{v} + b - \mathbf{y}\|^2 \ .$$

2. The function $f(w) = \frac{1}{2}w^2 - xw + \lambda|w|$ is clearly convex. As we already know, minimizing f is equivalent to finding a vector w such that $0 \in \partial f(w)$. A subgradient of f at w is given by $w - x + \lambda v$, where $v = 1$ if $w > 0$, $v = -1$ if $w < 0$, and $v \in [-1, 1]$ if $w = 0$. We next consider each of the cases $w < 0, w > 0 < w = 0$, and show that an optimal solution w must satisfy the identity

$$w = \text{sign}(x)[|x| - \lambda]_+ \ . \tag{20}$$

Recall that $\lambda > 0$. We note that $w = 0$ is optimal if and only if there exists $v \in [-1, 1]$ such that $0 = 0 - x + \lambda v$. Equivalently, $w = 0$ is optimal if and only if $|x| - \lambda \leq 0$ (which holds if and only if $[|x| - \lambda]_+ = 0$). Hence, if $w = 0$ is optimal, Equation (20) is satisfied. Next, we note that a necessary and sufficient optimality condition for $w > 0$ is that $0 = w - x + \lambda$. In this case, we obtain that $w = x - \lambda$. Since w and λ are positive, we must have that x is positive as well. Hence, if $w > 0$, Equation (20) is satisfied. Finally, a scalar $w < 0$ is optimal if and only if $0 = w - x - \lambda$. Hence, $w = x + \lambda$. Since w is negative and λ is positive, we must have that x is negative. Also here, if $w < 0$, then Equation (20) is satisfied.

3. (a) Note that this is a different phrasing of Sauer's lemma.

   (b) The partial derivatives of R can be calculated using the chain rule. Note that $R = R_1 \circ R_2$, where R_1 is the logarithm function, and $R_2(w) = \sum_{i=1}^m \exp(-y_i \sum_{j=1}^d w_j h_j(\mathbf{x}_i)) = \sum_{i=1}^m D_i = Z$. Taking the derivatives of R_1 and R_2 yields

$$\frac{\partial R(\mathbf{w})}{\partial w_j} = \sum_{i=1}^m -D_i y_i h_j(x_i) \ .$$

   Note that $-D_i y_i h_j(x_i) = -D_i$ if $h_j(x_i) = y_i$, and $-D_i y_i h_j(x_i) = D_i = 2D_i - D_i$ if $h_j(x_i) \neq y_i$. Hence,

$$\begin{aligned} \frac{\partial R(\mathbf{w})}{\partial w_j} &= \sum_{i=1}^m -D_i \mathbb{1}_{[h_j(x_i)=y_i]} + (2D_i - D_i)\mathbb{1}_{[h_j(x_i)\neq y_i]} \\ &= -1 + 2\sum_{i=1}^m D_i \mathbb{1}_{[h_j(x_i)\neq y_i]} \\ &=: 2\epsilon_j - 1 \end{aligned}$$

Note that if $\epsilon_j \leq 1/2 - \gamma$ for some $\gamma \in (0, 1/2)$, we have that

$$\left| \frac{\partial R(\mathbf{w})}{\partial w_j} \right| \geq 2\gamma \ .$$

(c) We introduce the following notation:

- Let $D_t = (D_{t,1}, \ldots, D_{t,m})$ denote the distribution over the training sequence maintained by AdaBoost.
- Let $\hat{h}_1, \hat{h}_2, \ldots$ be the hypotheses selected by AdaBoost, and let $w_1, w_2, \ldots$ be the corresponding weights.
- Let $\epsilon_1, \epsilon_2, \ldots$ be the weighted errors of $\hat{h}_1, \hat{h}_2, \ldots$.
- Let $\gamma_t = 1/2 - \epsilon_t$ for each t. Note that $\gamma_t \leq \gamma$ for all t.
- Let $H_t = \sum_{q=1}^{t} w_q \hat{h}_q$.

We would like to show that[19]

$$\log(\sum_{i=1}^{m} \exp(-y_i H_{t+1}(x_i))) - \log(\sum_{i=1}^{m} \exp(-y_i H_t(x_i))) \geq \log(\sqrt{1 - 4\gamma^2}) \ .$$

It suffices to show that

$$\frac{\sum_{i=1}^{m} \exp(-y_i H_{t+1}(x_i))}{\sum_{i=1}^{m} \exp(-y_i H_t(x_i))} \geq \sqrt{1 - 4\gamma^2}$$

Note that

$$\begin{aligned} D_{t+1}(i) &= D_1(i) \cdot \frac{\exp(-y_i w_1 \hat{h}_1(x_i))}{Z_1} \cdot \ldots \cdot \frac{\exp(-y_i w_t \hat{h}_t(x_i))}{Z_t} \\ &= \frac{D_1(i) \exp(-y_i H_t(x_i))}{\prod_{q=1}^{t} Z_q} \\ &= \frac{\exp(-y_i H_t(x_i))}{m \prod_{q=1}^{t} Z_q} \ , \end{aligned}$$

where $Z_1, Z_2, \ldots$ denote the normalization factors. Hence,

$$\begin{aligned} \sum_{i=1}^{m} \exp(-y_i H_t(x_i)) &= m \sum_{i=1}^{m} D_{t+1}(i) \prod_{q=1}^{t} Z_q \\ &= m \prod_{q=1}^{t} Z_q \ . \end{aligned}$$

[19] Errata: There was a typo in the question. The inequality that should be proved is $R(\mathbf{w}^{(t+1)}) - R(\mathbf{w}^{(t)}) \geq \sqrt{1 - 4\gamma^2}$.

Hence,

$$\frac{\sum_{i=1}^{m} \exp(-y_i H_{t+1}(x_i))}{\sum_{i=1}^{m} \exp(-y_i H_t(x_i))} = Z_{t+1}$$

Recall that $w_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$. Hence, we have

$$\begin{aligned}
Z_{t+1} &= \sum_{i=1}^{m} D_t(i) \exp(-y_i w_t \hat{h}_t(x_i)) \\
&= \sum_{i:\hat{h}_t(x_i)=y_i} D_t(i) \exp(-w_t) + \sum_{i:\hat{h}_t(x_i)\neq y_i} D_t(i) \exp(w_t) \\
&= \exp(-w_t)(1-\epsilon_t) + \exp(w_t)\epsilon_t \\
&= 2\sqrt{(1-\epsilon_t)\epsilon_t} \\
&= 2\sqrt{(1/2+\gamma_t)(1/2-\gamma_t)} \\
&= \sqrt{(1-4\gamma_t^2)} \\
&\geq \sqrt{(1-4\gamma^2)}\ .
\end{aligned}$$

References

[1] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, 2003.