Solutions to Exercises

Chapter 2

2.1 Two-oracle variant of the PAC model

- Assume that $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the standard PAC model using algorithm $\mathcal{A}$. Consider the distribution $\mathcal{D} = \frac{1}{2}(\mathcal{D}_- + \mathcal{D}_+)$. Let $h \in \mathcal{H}$ be the hypothesis output by $\mathcal{A}$. Choose δ such that:
$$\mathbb{P}[R_{\mathcal{D}}(h) \leq \epsilon/2] \geq 1 - \delta.$$
From
$$\begin{aligned} R_{\mathcal{D}}(h) &= \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq c(x)] \\ &= \frac{1}{2}\big(\mathop{\mathbb{P}}_{x \sim \mathcal{D}_-}[h(x) \neq c(x)] + \mathop{\mathbb{P}}_{x \sim \mathcal{D}_+}[h(x) \neq c(x)]\big) \\ &= \frac{1}{2}(R_{\mathcal{D}_-}(h) + R_{\mathcal{D}_+}(h)), \end{aligned}$$
it follows that:
$$\mathbb{P}[R_{\mathcal{D}_-}(h) \leq \epsilon] \geq 1 - \delta \quad \text{and} \quad \mathbb{P}[R_{\mathcal{D}_+}(h) \leq \epsilon] \geq 1 - \delta.$$
This implies two-oracle PAC-learning with the same computational complexity.
- Assume now that $\mathcal{C}$ is efficiently PAC-learnable in the two-oracle PAC model. Thus, there exists a learning algorithm $\mathcal{A}$ such that for $c \in \mathcal{C}$, $\epsilon > 0$, and $\delta > 0$, there exist m_- and m_+ polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, such that if we draw m_- negative examples or more and m_+ positive examples or more, with confidence $1 - \delta$, the hypothesis h output by $\mathcal{A}$ verifies:
$$\mathbb{P}[R_{\mathcal{D}_-}(h)] \leq \epsilon \quad \text{and} \quad \mathbb{P}[R_{\mathcal{D}_+}(h)] \leq \epsilon.$$
Now, let $\mathcal{D}$ be a probability distribution over negative and positive examples. If we could draw m examples according to $\mathcal{D}$ such that $m \geq \max\{m_-, m_+\}$, m polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, then two-oracle PAC-learning would imply standard PAC-learning:
$$\begin{aligned} &\mathbb{P}[R_{\mathcal{D}}(h)] \\ &\leq \mathbb{P}[R_{\mathcal{D}}(h)|c(x) = 0]\,\mathbb{P}[c(x) = 0] + \mathbb{P}[R_{\mathcal{D}}(h)|c(x) = 1]\,\mathbb{P}[c(x) = 1] \\ &\leq \epsilon(\mathbb{P}[c(x) = 0] + \mathbb{P}[c(x) = 1]) = \epsilon. \end{aligned}$$
If $\mathcal{D}$ is not too biased, that is, if the probability of drawing a positive example, or that of drawing a negative example is more than ϵ, it is not hard to show, using Chernoff bounds or just Chebyshev's inequality, that drawing a polynomial number of examples in $1/\epsilon$ and $1/\delta$ suffices to guarantee that $m \geq \max\{m_-, m_+\}$ with high confidence.
Otherwise, $\mathcal{D}$ is biased toward negative (or positive examples), in which case returning $h = h_0$ (respectively $h = h_1$) guarantees that $\mathbb{P}[R_{\mathcal{D}}(h)] \leq \epsilon$.

To show the claim about the not-too-biased case, let S_m denote the number of positive examples obtained when drawing m examples when the probability of a positive example is ϵ. By Chernoff bounds,

$$\mathbb{P}[S_m \leq (1-\alpha)m\epsilon] \leq e^{-m\epsilon\alpha^2/2}.$$

We want to ensure that at least m_+ examples are found. With $\alpha = \frac{1}{2}$ and $m = \frac{2m_+}{\epsilon}$,

$$\mathbb{P}[S_m > m_+] \leq e^{-m_+/4}.$$

Setting the bound to be less than or equal to $\delta/2$, leads to the following condition on m:

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}$$

A similar analysis can be done in the case of negative examples. Thus, when $\mathcal{D}$ is not too biased, with confidence $1-\delta$, we will find at least m_- negative and m_+ positive examples if we draw m examples, with

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{2m_-}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}.$$

In both solutions, our training data is the set T and our learned concept $L(T)$ is the tightest circle (with minimal radius) which is consistent with the data.

2.5 Triangles

As in the case of axis-aligned rectangles, consider three regions r_1, r_2, r_3, along the sides of the target concept as indicated in figure E.6. Note that the triangle formed by the points $A"$, $B"$, $C"$ is similar to ABC (same angles) since $A"B"$ must be parallel to AB, and similarly for the other sides.

Assume that $\mathbb{P}[ABC] > \epsilon$, otherwise the statement would be trivial. Consider a triangle $A'B'C'$ similar to ABC and consistent with the training sample and such that it meets all three regions r_1, r_2, r_3.

Since it meets r_1, the line $A'B'$ must be below $A"B"$. Since it meets r_2 and r_3, A' must be in r_2 and B' in r_3 (see figure E.6). Now, since the angle $\widehat{A'B'C'}$ is equal to $\widehat{A"B"C"}$, C' must be necessarily above $C"$. This implies that triangle $A'B'C'$ contains $A"B"C"$, and thus $error(A'B'C') \leq \epsilon$.

$$error(A'B'C') > \epsilon \implies \exists i \in \{1,2,3\}\colon A'B'C' \cap r_i = \emptyset.$$

Thus, by the union bound,

$$\mathbb{P}[error(A'B'C') > \epsilon] \leq \sum_{i=1}^{3} \mathbb{P}[A'B'C' \cap r_i = \emptyset] \leq 3(1-\epsilon/3)^m \leq 3e^{-3m\epsilon}.$$

Setting δ to match the right-hand side gives the sample complexity $m \geq \frac{3}{\epsilon}\log\frac{3}{\delta}$.

Chapter 3

3.3 Growth function of linear combinations

(a) $\{\mathcal{X}^+ \cup \{\mathbf{x}_{m+1}\}, \mathcal{X}^-\}$ and $\{\mathcal{X}^+, \mathcal{X}^- \cup \{\mathbf{x}_{m+1}\}\}$ are linearly separable by a hyperplane going through the origin if and only if there exists $\mathbf{w}_1 \in \mathbb{R}^d$ such that

$$\forall \mathbf{x} \in \mathcal{X}^+, \mathbf{w}_1 \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \in \mathcal{X}^-, \mathbf{w}_1 \cdot \mathbf{x} < 0, \text{ and } \mathbf{w}_1 \cdot \mathbf{x}_{m+1} > 0 \tag{E.35}$$

and there exists $\mathbf{w}_2 \in \mathbb{R}^d$ such that

$$\forall \mathbf{x} \in \mathcal{X}^+, \mathbf{w}_2 \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \in \mathcal{X}^-, \mathbf{w}_2 \cdot \mathbf{x} < 0, \text{ and } \mathbf{w}_2 \cdot \mathbf{x}_{m+1} < 0. \tag{E.36}$$

For any $\mathbf{w}_1, \mathbf{w}_2$, the function $f\colon (t \mapsto t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \cdot \mathbf{x}_{m+1}$ is continuous over $[0,1]$. (E.44) and (E.45) hold iff $f(0) < 0$ and $f(1) > 0$, that is iff there exists $\mathbf{w} = t_0\mathbf{w}_1 + (1-t_0)\mathbf{w}_2$ linearly separating $\{\mathcal{X}^+, \mathcal{X}^-\}$ and such at $\mathbf{w} \cdot \mathbf{x}_{m+1} = 0$.

(b) Repeating the formula, we obtain $C(m,d) = \sum_{k=0}^{m-1} \binom{m-1}{k} C(1, d-k)$. Since, $C(1,n) = 2$ if $n \geq 1$ and $C(1,n) = 0$ otherwise, the result follows.

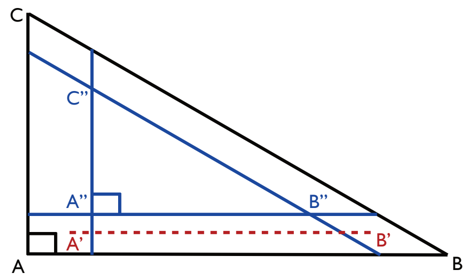

Figure E.5
Rectangle triangles.

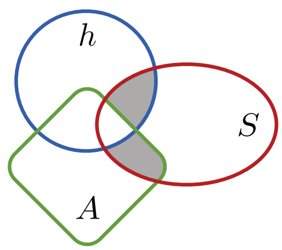

Figure E.6
Illustration of $(h\Delta\mathcal{A})\cap\mathcal{S} = (h\cap\mathcal{S})\Delta(\mathcal{A}\cap\mathcal{S})$ shown in gray.

(c) This is a direct application of the result of the previous question.

3.25 VC-dimension of symmetric difference of concepts

Fix a set $\mathcal{S}$. We can show that the number of classifications of $\mathcal{S}$ using $\mathcal{H}$ is the same as when using $\mathcal{H}\Delta A$. The set of classifications obtained using $\mathcal{H}$ can be identified with $\{\mathcal{S}\cap h\colon h\in\mathcal{H}\}$ and the set of classifications using $\mathcal{H}\Delta A$ can be identified with $\{\mathcal{S}\cap(h\Delta A)\colon h\in\mathcal{H}\}$. Observe that for any $h\in\mathcal{H}$,

$$\mathcal{S}\cap(h\Delta A) = (\mathcal{S}\cap h)\Delta(\mathcal{S}\cap A). \tag{E.37}$$

Figure E.7 helps illustrate this equality in a special case. Now, in view of this inequality, if $\mathcal{S}\cap(h\Delta A) = \mathcal{S}\cap(h'\Delta A)$ for $h, h'\in\mathcal{H}$, then

$$(\mathcal{S}\cap h)\Delta\mathcal{B} = (\mathcal{S}\cap h')\Delta\mathcal{B}, \tag{E.38}$$

with $\mathcal{B} = \mathcal{S}\cap\mathcal{A}$. Since two sets that have the same symmetric differences with respect to a set $\mathcal{B}$ must be equal, this implies

$$\mathcal{S}\cap h = \mathcal{S}\cap h'. \tag{E.39}$$

This shows that ϕ defined by

$$\phi\colon \mathcal{S}\cap\mathcal{H}\to\mathcal{S}\cap(\mathcal{H}\Delta\mathcal{A})$$
$$\mathcal{S}\cap h\mapsto\mathcal{S}\cap(h\Delta\mathcal{A})$$

is a bijection, and thus that the sets $\mathcal{S}\cap\mathcal{H}$ and $\mathcal{S}\cap(\mathcal{H}\Delta A)$ have the same cardinality.

Chapter 5

5.3 Importance weighted SVM

The modified primal optimization problem can be written as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i p_i \\ \text{subject to} & y_i[w\cdot x_i + b] \geq 1 - \xi_i\,. \end{array}$$

The Lagrangian holding for all $w, b, \alpha_i \geq 0, \beta_i \geq 0$ is then

$$\begin{aligned} L(w,b,\alpha) &= \frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i p_i \\ &\quad - \sum_{i=1}^m \alpha_i[y_i(w\cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i\,. \end{aligned} \tag{E.40}$$

Then $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$ are the same as for the regular non-separable SVM optimization problem. We also have $\frac{\partial L}{\partial \xi_i} = Cp_i - \alpha_i - \beta_i$. Thus, to satisfy the KKT conditions we have for all $i \in [m]$,

$$w = \sum_{i=1}^m \alpha_i y_i x_i \tag{E.41}$$

$$\sum_{i=1}^m \alpha_i y_i = 0 \tag{E.42}$$

$$\alpha_i + \beta_i = Cp_i \tag{E.43}$$

$$\alpha_i[y_i(w\cdot x_i + b) - 1 + \xi_i] = 0 \tag{E.44}$$

$$\beta_i \xi_i = 0\,. \tag{E.45}$$

Plugging equation E.79 into equation E.78, we get

$$\begin{aligned} L &= \frac{1}{2}||\sum_{i=1}^m \alpha_i y_i x_i||^2 + C\sum_{i=1}^m \xi_i p_i - \sum_{i,j=1}^m \alpha_i\alpha_j y_i y_j (x_i\cdot x_j) \\ &\quad - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i - \sum \alpha_i \xi_i - \sum_{i=1}^m \beta_i \xi_i\,. \end{aligned} \tag{E.46}$$

Using equation E.81, we can simplify:

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2}||\sum_{i=1}^m \alpha_i y_i x_i||^2\,,$$

meaning that the objective function is the same as in the regular SVM problem. The difference is in the constraints on the optimization. Recall that our dual form holds for $\beta_i \geq 0$. Using again equation E.81, our optimization problem is to maximize L subject to the constraints:

$$\forall i \in [m], 0 \leq \alpha_i \leq Cp_i \wedge \sum_{i=1}^m \alpha_i y_i = 0.$$

5.6 Sparse SVM

(a) Let

$$\mathbf{x}_i' = (y_1\mathbf{x}_i\cdot\mathbf{x}_1, \ldots, y_m\mathbf{x}_i\cdot\mathbf{x}_m)\,.$$

Then the optimization problem becomes

$$\begin{aligned}\min_{\boldsymbol{\alpha},b,\xi} \quad & \frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i\left(\boldsymbol{\alpha}\cdot\mathbf{x}'_i + b\right) \geq 1-\xi \\ & \xi_i, \alpha_i \geq 0, i\in[m]\,,\end{aligned}$$

which is the standard formulation of the primal SVM optimization problem on samples $\mathbf{x}'_i$, modulo the non-negativity constraints on $\boldsymbol{\alpha}_i$.

(b) The Lagrangian of (1) for all $\alpha_i \geq 0, \xi_i \geq 0, b, \alpha'_i \geq 0, \beta_i \geq 0, \gamma_i \geq 0, i \in [m]$ is

$$L = \frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha'_i(y_i(\boldsymbol{\alpha}\cdot\mathbf{x}'_i + b) - 1 + \xi_i) - \sum_{i=1}^{m}\beta_i\xi_i - \sum_{i=1}^{m}\gamma_i\alpha_i\,,$$

and the KKT conditions are

$$\begin{aligned}\nabla_\alpha L = 0 \quad &\Leftrightarrow \quad \boldsymbol{\alpha} = \sum_{i=1}^{m}\alpha'_i y_i \mathbf{x}'_i + \boldsymbol{\gamma} \\ \nabla_b L = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{m}\alpha'_i y_i = 0 \\ \nabla_{\xi_i} L = 0 \quad &\Leftrightarrow \quad \alpha'_i + \beta_i = C\end{aligned}$$

and

$$\begin{aligned}\alpha'_i(y_i(\boldsymbol{\alpha}\cdot\mathbf{x}'_i + b) - 1 - \xi_i) &= 0 \\ \beta_i\xi_i &= 0 \\ \gamma_i\alpha_i &= 0.\end{aligned}$$

Using the KKT conditions on L we get

$$\begin{aligned}L \quad = \quad & \frac{1}{2}\left(\sum_{i=1}^{m}\alpha'_i y_i\mathbf{x}'_i + \boldsymbol{\gamma}\right)\cdot\left(\sum_{j=1}^{m}\alpha'_j y_j \mathbf{x}'_j + \boldsymbol{\gamma}\right) + C\sum_{i=1}^{m}\xi_i \\ & -\sum_{i=1}^{m}\alpha'_i\left(y_i\left(\left(\sum_{j=1}^{m}\alpha'_j y_j\mathbf{x}'_j + \boldsymbol{\gamma}\right)\cdot\mathbf{x}'_i + b\right) - 1 + \xi_i\right) \\ & \cancel{-\sum_{i=1}^{m}\beta_i\xi_i} - \cancel{\sum_{i=1}^{m}\gamma_i\alpha_i} \\ = \quad & -\frac{1}{2}\sum_{i=1}^{m}\alpha'_i y_i \mathbf{x}'_i\cdot\left(\sum_{j=1}^{m}\alpha'_j y_j\mathbf{x}'_j + \gamma\right) + \cancel{\frac{1}{2}\boldsymbol{\gamma}\cdot\boldsymbol{\alpha}} \\ & + \sum_{i=1}^{m}C\xi_i - \alpha'_i\left(y_i b - 1 + \xi_i\right) \\ = \quad & \sum_{i=1}^{m}\alpha'_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha'_i\alpha'_j y_i y_j \mathbf{x}'^{\top}_i\left(\mathbf{x}'_j + \boldsymbol{\gamma}\right) \\ & + \cancel{\sum_{i=1}^{m}(C-\alpha'_i)\xi_i} - \cancel{\sum_{i=1}^{m}\alpha'_i y_i b}.\end{aligned}$$

Thus the dual optimization problem is

$$\max_{\alpha',\gamma} \sum_{i=1}^{m} \alpha'_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha'_i\alpha'_j y_i y_j \mathbf{x}'_i \cdot \left(\mathbf{x}'_j + \boldsymbol{\gamma}\right)$$

$$\text{subject to } \sum_{i=1}^{m} \alpha'_i y_i = 0$$

$$0 \le \alpha'_i \le C, \gamma_i \ge 0, i \in [m]\,.$$

Chapter 6

6.18 Metrics and kernels

(a) If K is an NDS kernel, then by theorem 6.16 the kernel K' defined for any $x_0 \in \mathcal{X}$ by:

$$K'(x,x') = \frac{1}{2}[K(x,x_0) + K(x',x_0) - K(x,x')]$$

is a PDS kernel ($K(x_0,x_0) = 0$). Let $\mathbb{H}$ be the reproducing Hilbert space associated to K'. There exists a mapping $\Phi(x)$ from $\mathcal{X}$ to $\mathbb{H}$ such that $\forall x,x' \in \mathcal{X}, K'(x,x') = \Phi(x)\cdot\Phi(x')$. Then,

$$\begin{aligned} ||\Phi(x) - \Phi(x')||^2 &= K'(x,x) + K'(x',x') - 2K'(x,x') \\ &= \frac{1}{2}[2K(x,x_0) - K(x,x)] + \\ &\quad \frac{1}{2}[2K(x',x_0) - K(x',x')] - \\ &\quad [K(x,x_0) + K(x',x_0) - K(x,x')] \\ &= K(x,x') \end{aligned}$$

It is then straightforward to show that $\sqrt{K}$ is a metric.

(b) Suppose that $K(x,x') = \exp(-|x-x'|^p)$, $x,x' \in \mathbb{R}$, is positive definite for $p > 2$. Then, for any $t > 0$, $\{x_1,\ldots,x_n\} \subseteq X$, $\{c_1,\ldots,c_n\} \subseteq \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j \exp(-t|x_j - x_k|^p) = \sum_{i,j=1}^{n} c_i c_j \exp(-|t^{1/p}x_j - t^{1/p}x_k|^p) \ge 0$$

Thus, by theorem 6.17, $K'(x,x') = |x-x'|^p$ is an NDS kernel. But, $\sqrt{K'}$ is not a metric for $p > 2$ since it does not verify the triangle inequality (take $x = 1$, $x' = 2$, $x'' = 3$), which contradicts part (a).

(c) If $a < 0$ or $b < 0$, $a||x||^2 + b < 0$ for some non-null vectors x. For such values, $K(x,x) = \tanh(a||x||^2 + b) < 0$. The kernel is thus not PDS and the SVM training may not converge to an optimal value. The equivalent neural network may also converge to a local minimum.

6.19 Sequence kernels

(a) $X^* - I$ is a regular language and can be represented by a finite automaton. K can thus be defined by

$$\forall x,y \in \mathcal{X}^*, \quad K(x,y) = [[T \circ T^{-1}]](x,y), \tag{E.47}$$

where T is the weighted transducer shown in figure E.14. Thus, K is a rational kernel and in view of the theorem 6.21, it is positive definite symmetric.

(b) Let M_{X^*-I} be the minimal automaton representing $X^* - I$. The transducer T of figure E.14 can be constructed using M_{X^*-I}. Then, $|T| = |M_{X^*-I}| + 8$. Using composition of weighted transducers, the running time complexity of the computation of the algorithm is:

$$O(|x||y||T \circ T^{-1}|) = O(|x||y||T|^2) = O(|x||y||M_{X^*-I}|^2). \tag{E.48}$$

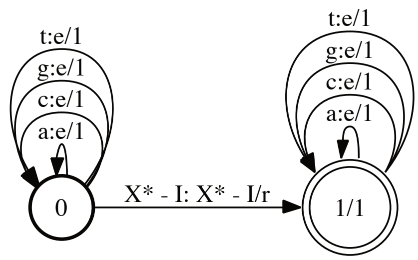

Figure E.7
Weighted transducer T. e represents the empty string, and $r = \rho$. $X^* - I$ stands for a finite automaton accepting $X^* - I$.

(c) The set of strings Y over the alphabet X of length less than n form a regular language since they can be described by:

$$Y = \bigcup_{i=0}^{n-1} X^i. \tag{E.49}$$

Thus, $Y_1 = Y \cap (X^* - I)$ and $Y_2 = (X^* - I) - Y_1$ are also regular languages. It suffices to replace in the transducer T of figure E.14 the transition labeled with $X^* - I : X^* - I/\rho$ with two transitions:

- $Y_1 : Y_1/\rho_1$, and
- $Y_2 : Y_2/\rho_2$,

with the same origin and destination states and with Y_1 and Y_2 denoting finite automata representing them. The kernel is thus still rational and PDS since it is of the form $T' \circ T'^{-1}$.

Chapter 7

7.8 Simplified AdaBoost

(a) As in the standard case, we can show that

$$\widehat{R}(h) \le \prod_{t=1}^{T} Z_t, \tag{E.50}$$

and that

$$Z_t = (1 - \epsilon_t)e^{-\alpha} + \epsilon_t e^{\alpha}. \tag{E.51}$$

By definition of γ and the fact that $e^{\alpha} - e^{-\alpha} > 0$ for all $\alpha > 0$,

$$Z_t = \epsilon_t(e^{\alpha} - e^{-\alpha}) + e^{-\alpha} \tag{E.52}$$
$$\le (1 - \gamma)(e^{\alpha} - e^{-\alpha}) + e^{-\alpha} \tag{E.53}$$
$$= (\frac{1}{2} - \gamma)e^{\alpha} + (\frac{1}{2} + \gamma)e^{-\alpha} = u(\alpha). \tag{E.54}$$

$u(\alpha)$ is minimized for

$$(\frac{1}{2} - \gamma)e^{\alpha} = (\frac{1}{2} + \gamma)e^{-\alpha}, \tag{E.55}$$

that is, for

$$\alpha = \frac{1}{2} \log \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}. \tag{E.56}$$

Tighter bounds on the product of the Z_ts can lead to better values for α.

(b) As in the standard case, at round t, the probability mass assigned to correctly classified points is $p_+ = (1-\epsilon_t)e^{-\alpha}$ and the probability mass assigned to the misclassified points is $p_- = \epsilon_t e^{\alpha}$. Thus,

$$\frac{p_-}{p_+} = \frac{\epsilon_t}{1-\epsilon_t}\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma} \leq \frac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma}\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma} = 1. \tag{E.57}$$

This contrasts with AdaBoost's property.

(c)

$$\begin{aligned} Z_t &\leq (\frac{1}{2}-\gamma)e^{\alpha} + (\frac{1}{2}+\gamma)e^{-\alpha} && \text{(E.58)}\\ &= (\frac{1}{2}-\gamma)\sqrt{\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}} + (\frac{1}{2}+\gamma)\sqrt{\frac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma}} && \text{(E.59)}\\ &= 2\sqrt{(\frac{1}{2}+\gamma)(\frac{1}{2}-\gamma)}. && \text{(E.60)} \end{aligned}$$

Thus, the empirical error can be bounded as follows:

$$\begin{aligned} \widehat{R}_S(h) &\leq \prod_{t=1}^{T} Z_t && \text{(E.61)}\\ &\leq [2\sqrt{(\frac{1}{2}+\gamma)(\frac{1}{2}-\gamma)}]^T && \text{(E.62)}\\ &= (1-4\gamma^2)^{T/2} && \text{(E.63)}\\ &\leq e^{-2\gamma^2 T}. && \text{(E.64)} \end{aligned}$$

(d) If $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m 1_{y_i f(x_i) \leq 0} \leq \frac{1}{m}$, then clearly $\widehat{R}_S(h) = 0$. Using the bound obtained in the previous question, if $e^{-2\gamma^2 T} < \frac{1}{m}$, the empirical error is zero. This can be rewritten as

$$T > \frac{\log m}{2\gamma^2}. \tag{E.65}$$

(e) Using the bound for the consistent case,

$$\mathbb{P}[R(h) > \epsilon] \leq 2\Pi_{\mathcal{C}}(2m)2^{-\frac{m\epsilon}{2}} \leq 2(\frac{2em}{d})^d 2^{-\frac{m\epsilon}{2}}. \tag{E.66}$$

Setting the right-hand side to δ, with probability at least $1-\delta$, the following bound holds for that consistent hypothesis:

$$error_{\mathcal{D}}(\mathcal{H}) \leq \frac{2}{m}\big(d\log_2\frac{2em}{d} + \log_2\frac{2}{\delta}\big), \tag{E.67}$$

with $d = 2(s+1)T\log_2(eT)$ and $T = \left\lfloor \frac{\log m}{2\gamma^2} \right\rfloor + 1$.

The bound is vacuous for $\gamma(m) = O(\sqrt{\frac{\log m}{m}})$. This could suggest overfitting.

7.11 HingeBoost

(a) Since the hinge loss is convex, its composition with affine function of $\boldsymbol{\alpha}$ is also convex and F is convex as as sum of convex functions.

For the existence of one-sided directional derivatives, one can use the fact that any convex function has one-sided directional derivatives or alternatively, that our specific function is the sum of piecewise affine functions, which are also known to have one-sided directional derivatives (think of one-dimensional hinge loss).

(b) Distinguishing different cases depending on the value of $y_i f_{t-1}(x_i) = 1$, it is straightforward to derive the following expressions for all $j \in [N]$:

$$F'_+(\boldsymbol{\alpha}_{t-1}, \boldsymbol{e}_j) = \sum_{i=1}^{m} -y_i h_j(x_i)[1_{y_i f_{t-1}(x_i)<1} + 1_{(y_i h_j(x_i)<0)\wedge(y_i f_{t-1}(x_i)=1)}]$$

$$F'_-(\boldsymbol{\alpha}_{t-1}, \boldsymbol{e}_j) = \sum_{i=1}^{m} -y_i h_j(x_i)[1_{y_i f_{t-1}(x_i)<1} + 1_{(y_i h_j(x_i)>0)\wedge(y_i f_{t-1}(x_i)=1)}].$$

The key here is that when $y_i f_{t-1}(x_i) \neq 1$, each term in the sum will be either 0 or the affine function independent of $y_i h_j(x_i)$. On the other hand, when $y_i f_{t-1}(x_i) = 1$, the sign of $y_i h_j(x_i)$ determines whether the finite differences will extend into the 0 portion of the affine portion of the term.

(c)

HingeBoost$(S = ((x_1, y_1), \ldots, (x_m, y_m)))$

$f \leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
  $r \leftarrow \sum_{i=1}^{m} -y_i h_j(x_i)[1_{y_i f(x_i)<1} + 1_{(y_i h_j(x_i)<0)\wedge(y_i f(x_i)=1)}]$
  $l \leftarrow \sum_{i=1}^{m} -y_i h_j(x_i)[1_{y_i f(x_i)<1} + 1_{(y_i h_j(x_i)>0)\wedge(y_i f(x_i)=1)}]$
  if $(l \leq 0) \wedge (r \geq 0)$ **then**
    $d[j] \leftarrow 0$
  elseif $(l \leq r)$ **then**
    $d[j] \leftarrow r$
  else $d[j] \leftarrow l$
for $t \leftarrow 1$ **to** T **do**
  $k \leftarrow \underset{j\in[N]}{\text{argmin}}\ |d[j]|$
  $\eta \leftarrow \text{argmin}_{\eta\geq 0}\, G(f + \eta h_k)$ ▷ line search
  $f \leftarrow f + \eta h_k$
return f

Chapter 8

8.2 Generalized mistake bound

The bound is unaffected, as shown by the following, using the same definitions and steps as in this chapter:

$$\begin{aligned}
M\rho &\le \frac{\mathbf{v}\cdot\sum_{t\in I} y_t\mathbf{x}_t}{\|\mathbf{v}\|}\\
&= \frac{\mathbf{v}\cdot\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)/\eta}{\|\mathbf{v}\|} \quad \text{(definition of updates)}\\
&= \frac{\mathbf{v}\cdot\mathbf{w}_{T+1}}{\eta\|\mathbf{v}\|}\\
&\le \|\mathbf{w}_{T+1}\|/\eta \quad \text{(Cauchy-Schwarz ineq.)}\\
&= \|\mathbf{w}_{t_m} + \eta y_{t_m}\mathbf{x}_{t_m}\|/\eta \quad (t_m \text{ largest } t \text{ in } I)\\
&= \Big[\|\mathbf{w}_{t_m}\|^2 + \eta^2\|\mathbf{x}_{t_m}\|^2 + \underbrace{\eta y_{t_m}\mathbf{w}_{t_m}\cdot\mathbf{x}_{t_m}}_{\le 0}\Big]^{1/2}/\eta\\
&\le \Big[\|\mathbf{w}_{t_m}\|^2 + \eta^2 R^2\Big]^{1/2}/\eta\\
&\le \Big[M\eta^2R^2\Big]^{1/2}/\eta = \sqrt{M}R. \quad \text{(applying the same to previous } t\text{s in } I).
\end{aligned}$$

8.10 On-line to batch — non-convex loss

(a) We use the following series of inequalities:

$$\begin{aligned}
&\min_{i\in[T]}\big(R(h_i) + 2c_\delta(T-i+1)\big)\\
&\le \frac{1}{T}\sum_{i=1}^{T}\big(R(h_i) + 2c_\delta(T-i+1)\big)\\
&= \frac{1}{T}\sum_{i=1}^{T}R(h_{i-1}) + \frac{2}{T}\sum_{i=0}^{T-1}\sqrt{\frac{1}{2(T-i)}\log\frac{T(T+1)}{\delta}}\\
&< \frac{1}{T}\sum_{i=1}^{T}R(h_{i-1}) + \frac{2}{T}\sum_{i=0}^{T-1}\sqrt{\frac{1}{2(T-i)}\log\Big(\frac{(T+1)}{\delta}\Big)^2}\\
&= \frac{1}{T}\sum_{i=1}^{T}R(h_{i-1}) + \frac{2}{T}\sum_{i=0}^{T-1}\sqrt{\frac{1}{(T-i)}\log\frac{(T+1)}{\delta}}\\
&\le \frac{1}{T}\sum_{i=1}^{T}R(h_{i-1}) + 4\sqrt{\frac{1}{T}\log\frac{T+1}{\delta}}\,.
\end{aligned}$$

The first inequality follows, since the minimum is always less than or equal to the average and the final inequality follows from $\sum_{i=0}^{T-1}\sqrt{1/(T-i)} = \sum_{i=1}^{T}\sqrt{1/i} \le 2\sqrt{T}$.

(b) Coupling the inequality of part (a) with the high probability statement of lemma 8.14 to bound $\frac{1}{T}\sum_{i=1}^{T}R(h_i)$ shows the desired bound.

(c) The square-root terms in part (b) can be bounded further by $6\sqrt{\frac{1}{T}\log\frac{2(T+1)}{\delta}}$.

Now, note that for two events A and B that each occur with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{P}[\neg A \cup \neg B] \le \mathbb{P}[\neg A] + \mathbb{P}[\neg B] \le 2\delta\\
\iff\ &\mathbb{P}[A \wedge B] \ge 1 - 2\delta\,.
\end{aligned}$$

Thus, the probability that both bounds in (b) and (c) hold simultaneously is at least $1-2\delta$; substituting δ with $\delta/2$ everywhere completes the bound.

Chapter 9

9.5 Decision trees. A binary decision tree with n nodes has exactly $n+1$ leaves. Each node can be labeled with an integer from $\{1, \ldots, N\}$ indicating which dimension is queried to make a binary split and each leaf can be labeled with ± 1 to indicate the classification made at that leaf. Fix an ordering of the nodes and leaves and consider all possible labelings of this sequence. There can be no more than $(N+2)^{2n+1}$ distinct binary trees and, thus, the VC-dimension of this finite set of hypotheses can be no larger than $(2n+1)\log(N+2) = O(n \log N)$.

Chapter 11

11.1 Pseudo-dimension and monotonic functions

If for some $m > 0$, there exists $(t_1, \ldots, t_m)$ and a set of points $(x_1, \ldots, x_m)$ that $\mathcal{H}$ shatters, then $\phi \circ \mathcal{H}$ can also shatter it. To see that, note that if for some $h \in \mathcal{H}$,

$$h(x_i) \geq t_i\,,$$

then by the monotonic property of ϕ,

$$\phi(h(x_i)) \geq \phi(t_i)\,.$$

A similar argument holds for the case $h(x_i) < t_i$. Thus, $\phi \circ \mathcal{H}$ can shatter the set of points $(x_1, \ldots, x_m)$ with thresholds $(\phi(t_1), \ldots, \phi(t_m))$, and this proves that $\mathrm{Pdim}(\phi \circ \mathcal{H}) \geq \mathrm{Pdim}(\mathcal{H})$. Since ϕ is strictly monotonic, it is invertible, and a similar argument with ϕ^{-1} can be used to show $\mathrm{Pdim}(\mathcal{H}) \geq \mathrm{Pdim}(\phi \circ \mathcal{H})$.

11.8 Optimal kernel matrix

(a) Using the closed-form solution for the inner maximization problem $\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$, simplifies the joint optimization to a simpler minimization:

$$\min_{\mathbf{K} \succeq 0} \mathbf{y}^\top (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}\,, \quad \text{s.t.} \quad \|\mathbf{K}\|_2 \leq 1\,.$$

Note that for any invertible matrix $\mathbf{A}$, $\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y} \geq \|\mathbf{y}\|^2 \lambda_{\min}(\mathbf{A}^{-1}) = \|\mathbf{y}\|^2 \lambda_{\max}(\mathbf{A})^{-1}$. Thus, it is easy to see that $\min_{\mathbf{K}\succeq 0} \mathbf{y}^\top(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y} \geq \frac{\|y\|^2}{1+\lambda}$ since $\|\mathbf{K}\|_2 = \lambda_{\max}(\mathbf{K}) \leq 1$. We now show $\mathbf{K} = \frac{1}{\|\mathbf{y}\|^2}\mathbf{y}\mathbf{y}^\top$ achieves this lower bound. First, note that $(\frac{1}{\|\mathbf{y}\|^2}\mathbf{y}\mathbf{y}^\top + \lambda\mathbf{I})\mathbf{y} = (1+\lambda)\mathbf{y}$, so $\mathbf{y}$ is an eigenvector of the matrix with eigenvalue $(1+\lambda)$. Since the matrix is invertible, it can be shown that $\mathbf{y}$ is also an eigenvector of $(\frac{1}{\|\mathbf{y}\|^2}\mathbf{y}\mathbf{y}^\top + \lambda\mathbf{I})^{-1}$ with eigenvalue $\frac{1}{1+\lambda}$ (for example, consider the eigen decomposition of the matrix).

(b) The kernel matrix alone is not useful for classifying future unseen points x, which requires computing $\sum_{i=1}^m K(x_i, x)$ and needs access to an underlying kernel *function* that in consistent with the kernel matrix. Finding such a kernel function may be difficult in general, and furthermore the choice of function may not be unique.

Chapter 14

14.1 Tighter stability bounds

(a) No, even as $\beta \to 0$ the generalization bound of theorem 14.2 only guarantees $R(h_S) - \widehat{R}_S(h_S) \leq M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} = O(1/\sqrt{m})$.

(b) In this case, $M = C/\sqrt{m}$ and $M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} = O(1/m)$; thus, it would suffice to have $\beta = O(1/m^{3/2})$ in order to guarantee an $O(1/m)$ generalization bound.

14.2 Quadratic hinge loss stability

We first show that the loss function is σ-admissible. Consider three cases:

- Both $h(x)$ and $h'(x)$ are correct with margin greater than 1, then
$$|L(h(x),y) - L(h'(x),y)| = 0.$$
- Only one hypothesis is correct with large enough margin. Without loss of generality assume $h'(x)$ is correct, then
$$\begin{aligned}|L(h(x),y) - L(h'(x),y)| &= (1-h(x)y)^2\\ &\leq ((1-h(x)y) - (1-h'(x)y))^2 = (h'(x)-h(x))^2\\ &\leq 4\sqrt{M}|h(x)-h'(x)|.\end{aligned}$$
The first inequality follows from the assumption $1 - h'(x)y \leq 0$, and the second inequality follows from the bounded loss assumption, which implies $\forall h \in \mathcal{H}, |h(x)| \leq \sqrt{M}+1 \leq 2M$.
- Finally, we consider the case where both $h(x)$ and $h'(x)$ incur a loss. Without loss of generality assume $(1-h(x)y) \geq (1-h'(x)y)$, then
$$\begin{aligned}|L(h(x),y) - L(h'(x),y)| &= (1-h(x)y)^2 - (1-h'(x)y)^2\\ &= \big((1-h(x)y) + (1-h'(x)y)\big)\big((1-h(x)y) - (1-h'(x)y)\big)\\ &\leq |2 - y(h(x)+h'(x))||y(h(x)-h'(x))| \leq 6\sqrt{M}|h(x)-h'(x)|\,.\end{aligned}$$

Thus, the quadratic hinge loss is σ-admissible with $\sigma = 6\sqrt{M}$. By proposition 14.4, SVM with quadratic hinge loss is stable with $\beta = \frac{36r^2M}{m\lambda}$, and using theorem 14.2 gives the following bound:
$$R(h_S) \leq \widehat{R}_S(h_S) + \frac{36r^2M}{m\lambda} + \Big(\frac{72r^2M}{\lambda} + M\Big)\sqrt{\frac{\log\frac{1}{\delta}}{m}}\,.$$

Chapter 15

15.2 Double centering

(a) Observe that $\|\mathbf{x}_i - \mathbf{x}_j\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top(\mathbf{x}_i - \mathbf{x}_j) = \mathbf{x}_i^\top\mathbf{x}_i + \mathbf{x}_j^\top\mathbf{x}_j - 2\mathbf{x}_i^\top\mathbf{x}_j$ and rearrange terms.

(b) Noting that $\mathbf{X}^* = \mathbf{X} - \frac{1}{m}\mathbf{X}\mathbf{1}\mathbf{1}^\top$ and plugging into the equation $\mathbf{K}^* = \mathbf{X}^{*\top}\mathbf{X}^*$ yields the result.

(c) Note that the scalar form of the equation in (b) is
$$\mathbf{K}^*_{ij} = \mathbf{K}_{ij} - \frac{1}{m}\sum_{k=1}^m \mathbf{K}_{ik} - \frac{1}{m}\sum_{k=1}^m \mathbf{K}_{kj} + \frac{1}{m^2}\sum_k\sum_l \mathbf{K}_{k,l}\,.$$
Substituting with the equation $\mathbf{D}^2_{ij} = \mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij}$ from (a) and simplifying yields the result.

(d) We first observe that $-\frac{1}{2}\mathbf{HDH} = -\frac{1}{2}(\mathbf{D} - \frac{1}{m}\mathbf{D}\mathbf{1}\mathbf{1}^\top - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\mathbf{D} + \frac{1}{m^2}\mathbf{1}\mathbf{1}^\top\mathbf{D}\mathbf{1}\mathbf{1}^\top)$. By inspection, the matrix expression on the RHS corresponds to the scalar expression with four terms on the RHS of the equation in (c).

15.4 Nyström method

(a) For the first part of question, note that $\mathbf{W}$ is SPSD if $\mathbf{x}^\top\mathbf{W}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^l$. This condition is equivalent to $\mathbf{y}^\top\mathbf{K}\mathbf{y} \geq 0$ for all $\mathbf{y} \in \mathbb{R}^m$ where $y_i = 0$ for $l+1 \leq i \leq m$. Since $\mathbf{K}$ is SPSD by assumption, this latter condition holds. For the second part, we write $\widetilde{\mathbf{K}}$ in block form as
$$\widetilde{\mathbf{K}} = \begin{bmatrix}\mathbf{W}\\ \mathbf{K}_{21}\end{bmatrix}\mathbf{W}^\dagger\begin{bmatrix}\mathbf{W} & \mathbf{K}_{21}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{W} & \mathbf{K}_{21}^\top\\ \mathbf{K}_{21} & \mathbf{K}_{21}\mathbf{W}^\dagger\mathbf{K}_{21}^\top\end{bmatrix}.$$

Comparison with the block form of $\mathbf{K}$ then immediately yields the desired result.

(b) Observe that $\mathbf{C} = \mathbf{X}^\top \mathbf{X}'$ and $\mathbf{W} = \mathbf{X}'^\top \mathbf{X}'$. Thus,

$$\widetilde{\mathbf{K}} = \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^\top = \mathbf{X}^\top \mathbf{X}' \big(\mathbf{X}'^\top \mathbf{X}'\big)^\dagger \mathbf{X}'^\top \mathbf{X} = \mathbf{X}^\top \mathbf{U}_{X'} \mathbf{U}_{X'}^\top \mathbf{X} = \mathbf{X}^\top \mathbf{P}_{U_{X'}} \mathbf{X}.$$

(c) Yes. Using the expression for $\widetilde{\mathbf{K}}$ in (b) and the idempotency of orthogonal projection matrices, we can write $\widetilde{\mathbf{K}} = \mathbf{X}^\top \mathbf{P}_{U_{X'}} \mathbf{X} = \mathbf{A}^\top \mathbf{A}$, where $\mathbf{A} = \mathbf{P}_{U_{X'}} \mathbf{X}$.

(d) Since $\mathbf{K} = \mathbf{X}^\top \mathbf{X}$, $\text{rank}(\mathbf{K}) = \text{rank}(\mathbf{X}) = r$. Similarly, $\mathbf{W} = \mathbf{X}'^\top \mathbf{X}'$ implies $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{X}') = r$. The columns of $\mathbf{X}'$ are columns of $\mathbf{X}$, and they thus span the columns of $\mathbf{X}$. Hence, $\mathbf{U}_{X'}$ is an orthonormal basis for $\mathbf{X}$, i.e., $\mathbf{I}_N - \mathbf{P}_{U_{X'}} \in \text{Null}(\mathbf{X})$, and by part (b) of this exercise we have $\mathbf{K} - \widetilde{\mathbf{K}} = \mathbf{X}^\top (\mathbf{I}_N - \mathbf{P}_{U_{X'}})\mathbf{X} = \mathbf{0}$.

(e) Storage of $\mathbf{K}$ requires roughly 3200 TB, i.e.,

$$(20 \times 10^6)^2 \text{ entries} \times 8 \text{ bytes/entry} \times \frac{1\text{TB}}{10^{12} \text{ bytes}} = 3200 \text{ TB}.$$

Storage of $\mathbf{C}$ requires roughly 160 GB, i.e.,

$$(20 \times 10^6 \times 10^3) \text{ entries} \times 8 \text{ bytes/entry} \times \frac{1\text{GB}}{10^9 \text{ bytes}} = 160 \text{ GB}.$$

Note that the computed numbers do not account for the symmetry of $\mathbf{K}$ (doing so would change the storage requirements by less than a factor of two).

Chapter C

C.1 For any $\delta > 0$, let $t = f^{-1}(\delta)$. Plugging this in $\mathbb{P}[X > t] \le f(t)$ yields $\mathbb{P}[X > f^{-1}(\delta)] \le \delta$, that is $\mathbb{P}[X \le f^{-1}(\delta)] \ge 1 - \delta$.

C.2 By definition of expectation and using the hint, we can write

$$\mathbb{E}[X] = \sum_{n \ge 0} n\, \mathbb{P}[X = n] = \sum_{n \ge 1} n(\mathbb{P}[X \ge n] - \mathbb{P}[X \ge n+1]).$$

Note that in this sum, for $n \ge 1$, $\mathbb{P}[X \ge n]$ is added n times and subtracted $n-1$ times, thus $\mathbb{E}[X] = \sum_{n \ge 1} \mathbb{P}[X \ge n]$.

More generally, by definition of the Lebesgue integral, for any non-negative random variable X, the following identity holds:

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}[X \ge t]\, dt.$$

Chapter D

D.2 Estimating label bias. Let $\widehat{p}_+$ be the fraction of positively labeled points in $\mathcal{S} = (x_1, \ldots, x_m)$:

$$\widehat{p}_+ = \frac{1}{m} \sum_{i=1}^m 1_{f(x_i)=+1}$$

Since the points are drawn i.i.d.,

$$\mathbb{E}[\widehat{p}_+] = \frac{1}{m} \sum_{i=1}^m \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m}[1_{f(x_i)=+1}] = \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m}[1_{f(x_1)=+1}] = \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[1_{f(x)=+1}] = p_+.$$

Thus, by Hoeffding's inequality, for any $\epsilon > 0$,

$$\mathbb{P}[|p_+ - \widehat{p}_+| > \epsilon] \le 2e^{-2m\epsilon^2}.$$

Setting δ to match the right-hand side yields the result.

D.3 Biased coins

(a) By definition of the error of Oskar's prediction rule,

$$\begin{aligned} error(f_o) &= \mathbb{P}[f_o(S) \neq x] \\ &= \mathbb{P}[f_o(S) = x_A \wedge x = x_B] + \mathbb{P}[f_o(S) = x_B \wedge x = x_A] \\ &= \mathbb{P}\left[N(S) < \frac{m}{2}\Big| x = x_B\right]\mathbb{P}[x = x_B] + \\ &\quad \mathbb{P}\left[N(S) \geq \frac{m}{2}\Big| x = x_A\right]\mathbb{P}[x = x_A] \\ &= \frac{1}{2}\,\mathbb{P}[N(S) < \frac{m}{2}\Big| x = x_B] + \frac{1}{2}\,\mathbb{P}\left[N(S) \geq \frac{m}{2}\Big| x = x_A\right] \\ &\geq \frac{1}{2}\,\mathbb{P}\left[N(S) \geq \frac{m}{2}\Big| x = x_A\right]. \end{aligned}$$

(b) Note that $\mathbb{P}[N(S) \geq \frac{m}{2} | x = x_A] = \mathbb{P}[B(m,p) \geq k]$, with $p = 1/2 - \epsilon/2$, $k = \frac{m}{2}$, and $mp \leq k \leq m(1-p)$. Thus, by Slud's inequality (section D.5)

$$error(f_o) \geq \frac{1}{2}\,\mathbb{P}\left[N \geq \frac{m\epsilon/2}{\sqrt{1/4(1-\epsilon^2)m}}\right] = \frac{1}{2}\,\mathbb{P}\left[N \geq \frac{\sqrt{m}\epsilon}{\sqrt{1-\epsilon^2}}\right].$$

Using the second inequality of the appendix, we now obtain

$$error(f_o) \geq \frac{1}{4}\Big(1 - \sqrt{1 - e^{-u^2}}\Big),$$

with $u = \frac{\sqrt{m}\epsilon}{\sqrt{1-\epsilon^2}}$, which coincides with (D.29).

(c) If m is odd, since $\mathbb{P}\left[N(S) \geq \frac{m}{2}\Big| x = x_A\right] \geq \mathbb{P}\left[N(S) \geq \frac{m+1}{2}\Big| x = x_A\right]$, we can use the lower bound

$$error(f_o) \geq \frac{1}{2}\,\mathbb{P}\left[N(S) \geq \frac{m+1}{2}\Big| x = x_A\right].$$

Thus, in both cases we can use the lower bound expression with $\lceil m/2 \rceil$ instead of $m/2$.

(d) If $error(f_o)$ is at most δ, then $\frac{1}{4}\left[1 - \left[1 - e^{-\frac{2\lceil m/2\rceil \epsilon^2}{1-\epsilon^2}}\right]^{\frac{1}{2}}\right] < \delta$, which gives

$$e^{-\frac{2\lceil m/2\rceil\epsilon^2}{1-\epsilon^2}} < 1 - (1-4\delta)^2 = 4\delta(2-4\delta) = 8\delta(1-2\delta),$$

and

$$m > 2\left\lceil \frac{1-\epsilon^2}{2\epsilon^2}\log\frac{1}{8\delta(1-2\delta)}\right\rceil.$$

The lower bound varies as $\frac{1}{\epsilon^2}$.

(e) Let f be an arbitrary rule and denote by F_A the set of samples for which $f(S) = x_A$ and by F_B the complement. Then, by definition of the error,

$$\begin{aligned} error(f) &= \sum_{S\in F_A}\mathbb{P}[S \wedge x_B] + \sum_{S\in F_B}\mathbb{P}[S \wedge x_A] \\ &= \frac{1}{2}\sum_{S\in F_A}\mathbb{P}[S|x_B] + \frac{1}{2}\sum_{S\in F_B}\mathbb{P}[S|x_A] \\ &= \frac{1}{2}\sum_{\substack{S\in F_A\\ N(S)<m/2}}\mathbb{P}[S|x_B] + \frac{1}{2}\sum_{\substack{S\in F_A\\ N(S)\geq m/2}}\mathbb{P}[S|x_B] + \\ &\quad \frac{1}{2}\sum_{\substack{S\in F_B\\ N(S)<m/2}}\mathbb{P}[S|x_A] + \frac{1}{2}\sum_{\substack{S\in F_B\\ N(S)\geq m/2}}\mathbb{P}[S|x_A]. \end{aligned}$$

Now, if $N(S) \geq m/2$, clearly $\mathbb{P}[S|x_B] \geq \mathbb{P}[S|x_A]$. Similarly, if $N(S) < m/2$, clearly $\mathbb{P}[S|x_A] \geq \mathbb{P}[S|x_B]$. In view of these inequalities, $error(f)$ can be lower bounded as follows

$$\begin{aligned} error(f) \geq \frac{1}{2} \sum_{\substack{S \in F_A \\ N(S) < m/2}} \mathbb{P}[S|x_B] + \frac{1}{2} \sum_{\substack{S \in F_A \\ N(S) \geq m/2}} \mathbb{P}[S|x_A] + \\ \frac{1}{2} \sum_{\substack{S \in F_B \\ N(S) < m/2}} \mathbb{P}[S|x_B] + \frac{1}{2} \sum_{\substack{S \in F_B \\ N(S) \geq m/2}} \mathbb{P}[S|x_A] \\ = \frac{1}{2} \sum_{S:\, N(S) < m/2} \mathbb{P}[S|x_B] + \frac{1}{2} \sum_{S:\, N(S) \geq m/2} \mathbb{P}[S|x_A] \\ = error(f_o). \end{aligned}$$

Oskar's rule is known as the *maximum likelihood* solution.