Lerning from Data

Maciej Leks

17.02.2016

Problem 1.1:

See problem-1_1.jl

Problem 1.2:

See problem-1_2.jl

Problem 1.3:

For the rest of the problem let's assume that subscript n means $n = n_t$

(a) Let

$$\rho = \min_{1 \leq n \leq N} y_n \mathbf{w}^{*T} \mathbf{x}_n$$

Show that $\rho > 0$

Since w^*perfect for separable data (see the assumption at the begining of the task), hence every example

$$y_n \mathbf{w}^{*T} \mathbf{x}_n \quad > \quad 0$$

because $y_n = sign(\mathbf{w}^{*T} \mathbf{x}_n)$ and $y_n \neq 0$ ($\mathcal{Y} = \{-1, 1\}$)
since

$$y_n w^{*T} x_n \geq \min_{1 \leq n \leq N} y_n \mathbf{w}^{*T} \mathbf{x}_n$$

so

$$\begin{aligned} y_n w^{*T} x_n &\geq \min_{1 \leq n \leq N} y_n \mathbf{w}^{*T} \mathbf{x}_n > 0 \\ y_n w^{*T} x_n &\geq \rho > 0 \end{aligned}$$

(b) To show that $\mathbf{w}_t^T\mathbf{w}^* \geq \mathbf{w}_{t-1}^T\mathbf{w}^* + \rho$ we've already assumed that $\mathbf{w}_0 = \mathbf{0}$ and we've proved that $p > 0$(see 1.3 (a)), hence

_1. The direct proof:

$$\begin{aligned}
\mathbf{w}_t^T\mathbf{w}^* &\geq \mathbf{w}_{t-1}^T\mathbf{w}^* + \rho \\
\mathbf{w}_t^T\mathbf{w}^* &\geq \mathbf{w}_{t-1}^T\mathbf{w}^* + \min_{1\leq n\leq N} y_n\mathbf{w}^{*T}\mathbf{x}_n \\
\mathbf{w}_{t-1}^T\mathbf{w}^* + y_n\mathbf{x}_n^T\mathbf{w}^* &\geq \mathbf{w}_{t-1}^T\mathbf{w}^* + \min_{1\leq n\leq N} y_n\mathbf{w}^{*T}\mathbf{x}_n \\
y_n\mathbf{x}_n^T\mathbf{w}^* &\geq \min_{1\leq n\leq N} y_n\mathbf{w}^{*T}\mathbf{x}_n \\
y_n\mathbf{w}^{*T}x_n &\geq \min_{1\leq n\leq N} y_n\mathbf{w}^{*T}\mathbf{x}_n
\end{aligned}$$

which is always true hence we treat it as the rule enhancement.

_2. To proof $\mathbf{w}_t^T\mathbf{w}^* \geq t\rho$ we use mathematical induction:

_2.1. Verification step: for $t = 1$

$$\begin{aligned}
\mathbf{w}_1^T\mathbf{w}^* &\geq 1\rho \\
\mathbf{w}_1^T\mathbf{w}^* + \mathbf{w}_0^T\mathbf{w}* &\geq \mathbf{w}_0^T\mathbf{w}* + \rho \\
\mathbf{w}_1^T\mathbf{w}^* + 0 &\geq \mathbf{w}_0^T\mathbf{w}* + \rho
\end{aligned}$$

which is true according to _1.

_2.2. Proof step: Induction hypothesis: For some $t \geq 1$ we assume that $\mathbf{w}_t^T\mathbf{w}^* \geq t\rho$ is true.

Induction step: Then for every $t \geq 1$, $\mathbf{w}_{t+1}^T\mathbf{w}^* \geq (t+1)\rho$ must be true, since

$$\begin{aligned}
\mathbf{w}_{t+1}^T\mathbf{w}^* &\geq (t+1)\rho \\
(\mathbf{w}_t + y_t\mathbf{x}_t)^T\mathbf{w}^* &\geq t\rho + \rho \\
\mathbf{w}_t^T\mathbf{w}^* + y_t\mathbf{x}_t^T\mathbf{w}^* &\geq t\rho + \rho
\end{aligned}$$

$\mathbf{w}_t^T\mathbf{w}^*$is going to be changed by $t\rho$ from $\mathbf{w}_t^T\mathbf{w}^* \geq t\rho$:

$$\begin{aligned}
\mathbf{w}_t^T\mathbf{w}^* + y_t\mathbf{w}^{*T}\mathbf{x}_t &\geq t\rho + \rho \\
t\rho + y_t\mathbf{w}^{*T}\mathbf{x}_t &\geq t\rho + \rho \\
y_t\mathbf{w}^{*T}t_n &\geq \rho
\end{aligned}$$

which is true according to _1.

Interpretation of $\mathbf{w}_t^T\mathbf{w}^* \geq t\rho$: It shows that the normalized scalar product between $\mathbf{w}_t$ and the optimal set of weights $\mathbf{w}^*$ will be larger in each iteration.

(c) To show that

$$\|\mathbf{w}_t\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + \|\mathbf{x}_{t-1}\|^2$$

we must remember that $y_{t-1}\mathbf{w}_{t-1}^T\mathbf{x}_{t-1} \leq 0$, because $x_{t-1}$was misclassified (see the hint).

$$\|\mathbf{w}_t\|^2 \quad = \quad \|\mathbf{w}_{t-1} + y_{t-1}\mathbf{x}_{t-1}\|^2 = \|\mathbf{w}_{t-1}\|^2 + 2y_{t-1}\mathbf{w}_{t-1}^T\mathbf{x}_{t-1} + \|y_{t-1}\mathbf{x}_{t-1}\|^2$$

since $y_{t-1}\mathbf{w}_{t-1}^T\mathbf{x}_{t-1} \leq 0$, hence

$$\|\mathbf{w}_t\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + \|y_{t-1}\mathbf{x}_{t-1}\|^2$$

because $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$and $y \in \mathcal{Y} = \{-1, 1\}$, thereby

$$\|\mathbf{w}_t\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + \|\mathbf{x}_{t-1}\|^2$$

Interpretation of $\|\mathbf{w}_t\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + \|\mathbf{x}_{t-1}\|^2$: $\mathbf{w}_t$does not grow too fast and changes
only if $x_{t-1}$is misclassified.

(d) To show that $\|\mathbf{w}_t\|^2 \leq tR^2$, where

$$R = \max_{1 \leq n \leq N} \|\mathbf{x}_n\|$$

we must carry out an induction proof.

Verification step: for $t = 0$

$$\|\mathbf{w}_0\|^2 \leq 0$$

which is true since $w_0 = \mathbf{0}$
for $t = 1$

$$\begin{aligned} \|w_1\|^2 &\leq R^2 \\ \|\mathbf{w}_0 + y_n\mathbf{x}_n\|^2 &\leq R^2 \\ |y_n|\|\mathbf{x}_n\|^2 &\leq \max_{1 \leq n \leq N} \|\mathbf{x}_n\|^2 \\ \|\mathbf{x}_n\|^2 &\leq \max_{1 \leq n \leq N} \|\mathbf{x}_n\|^2 \end{aligned}$$

which is always true.

Proof step: Induction hypothesis: For some $t \geq 1$ we assume that $\|\mathbf{w}_t\|^2 \leq tR^2$ is true.

Induction step: Then for every $t \geq 1$, $\|\mathbf{w}_{t+1}\|^2 \leq (t+1)R^2$ must be true, since

$$\begin{aligned} \|\mathbf{w}_{t+1}\|^2 &\leq (t+1)R^2 \\ \|\mathbf{w}_t + y_n\mathbf{x}_n\|^2 &\leq (t+1)R^2 \\ \|\mathbf{w}_t\|^2 + 2y_n\mathbf{w}_t^T\mathbf{x}_n + |y_n|^2\|\mathbf{x}_n\|^2 &\leq (t+1)R^2 \end{aligned}$$

but we remeber from (c) that $\|\mathbf{w}_t\|^2 + \|\mathbf{x}_n\|^2 \leq \|\mathbf{w}_{t+1}\|^2$, so

$$\|\mathbf{w}_t\|^2 + \|\mathbf{x}_n\|^2 \leq \|\mathbf{w}_{t+1}\|^2 \leq (t+1)R^2$$

, the

$$\|\mathbf{w}_t\|^2 + \|\mathbf{x}_n\|^2 \leq (t+1)R^2$$

, and using our induction hypothesis RHS:

$$\begin{aligned} \|\mathbf{w}_t\|^2 + \|\mathbf{x}_n\|^2 &\leq tR^2 + R^2 \\ tR^2 + \|\mathbf{x}_n\|^2 &\leq tR^2 + R^2 \\ \|\mathbf{x}_n\|^2 &\leq R^2 \\ \|\mathbf{x}_n\|^2 &\leq \max_{1\leq n\leq N} \|\mathbf{x}_n\|^2 \end{aligned}$$

,which is always true what we had to proof.

(e) To show that $\frac{\mathbf{w}_t^T}{\|\mathbf{w}_t\|}\mathbf{w}^* \geq \sqrt{t}\frac{\rho}{R}$ and then proof that $t \leq \frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}$, I need to use $\mathbf{w}_t^T\mathbf{w}^* \geq t\rho$

(from (b)) and $\|\mathbf{w}_t\|^2 \leq tR^2$(from (d)). Hint: $\frac{\mathbf{w}_t^T\mathbf{w}^*}{\|\mathbf{w}_t\|\|\mathbf{w}^*\|} \leq 1$ (comes from Cauchy–Schwarz

inequality). First let show that

$$\begin{aligned} \frac{\mathbf{w}_t^T}{\|\mathbf{w}_t\|}\mathbf{w}^* &\geq \sqrt{t}\frac{\rho}{R} \\ \mathbf{w}_t^T\mathbf{w}^* &\geq \|\mathbf{w}_t\|\sqrt{t}\frac{\rho}{R} \\ t\rho &\geq \|\mathbf{w}_t\|\sqrt{t}\frac{\rho}{R} \qquad (1) \\ t^2\rho^2 &\geq \|\mathbf{w}_t\|^2 t\frac{\rho^2}{R^2} \\ t^2\rho^2 &\geq tR^2 t\frac{\rho^2}{R^2} \qquad (2) \\ 1 &\geq 1 \end{aligned}$$

since in LHS of (1) we use RHS from (b) and in RHS of (2) we use RHS of (b) formula.

Now we prove that:

$$
\begin{aligned}
t &\leq \frac{R^2\|\mathbf{w}^*\|^2}{\rho^2} \\
t\rho^2 &\leq R^2\|\mathbf{w}^*\|^2 \\
\frac{t\rho^2}{R^2} &\leq \|\mathbf{w}^*\|^2 \\
\sqrt{t}\frac{\rho}{R} &\leq \|\mathbf{w}^*\| \\
\|\mathbf{w}^*\| &\geq \sqrt{t}\frac{\rho}{R} \\
\frac{\mathbf{w}^T\mathbf{w}^*}{\|\mathbf{w}_t\|} &\geq \sqrt{t}\frac{\rho}{R} \qquad (3) \\
\sqrt{t}\frac{\rho}{R} &\geq \sqrt{t}\frac{\rho}{R} \qquad (4) \\
1 &\geq 1
\end{aligned}
$$

since in LHS of (3) we use Cauchy–Schwarz inequality and in LHS of (4) we use the result of the first part of (d).

Interpretation of all points: Normalized inner product between $\mathbf{w}_t$and $\mathbf{w}^*$is bounded by 1 (see the hint for (e)). As far as data is linearly separable there is an upper bound of numer of iterations after which $\mathbf{w}_t$ is aligned with $\mathbf{w}^*$(see (e)). So inner product between $\mathbf{w}_t$and $\mathbf{w}^*$grows fast but length of $\mathbf{w}_t$grows slowly (see (c)) that is why the algorithm will converge.

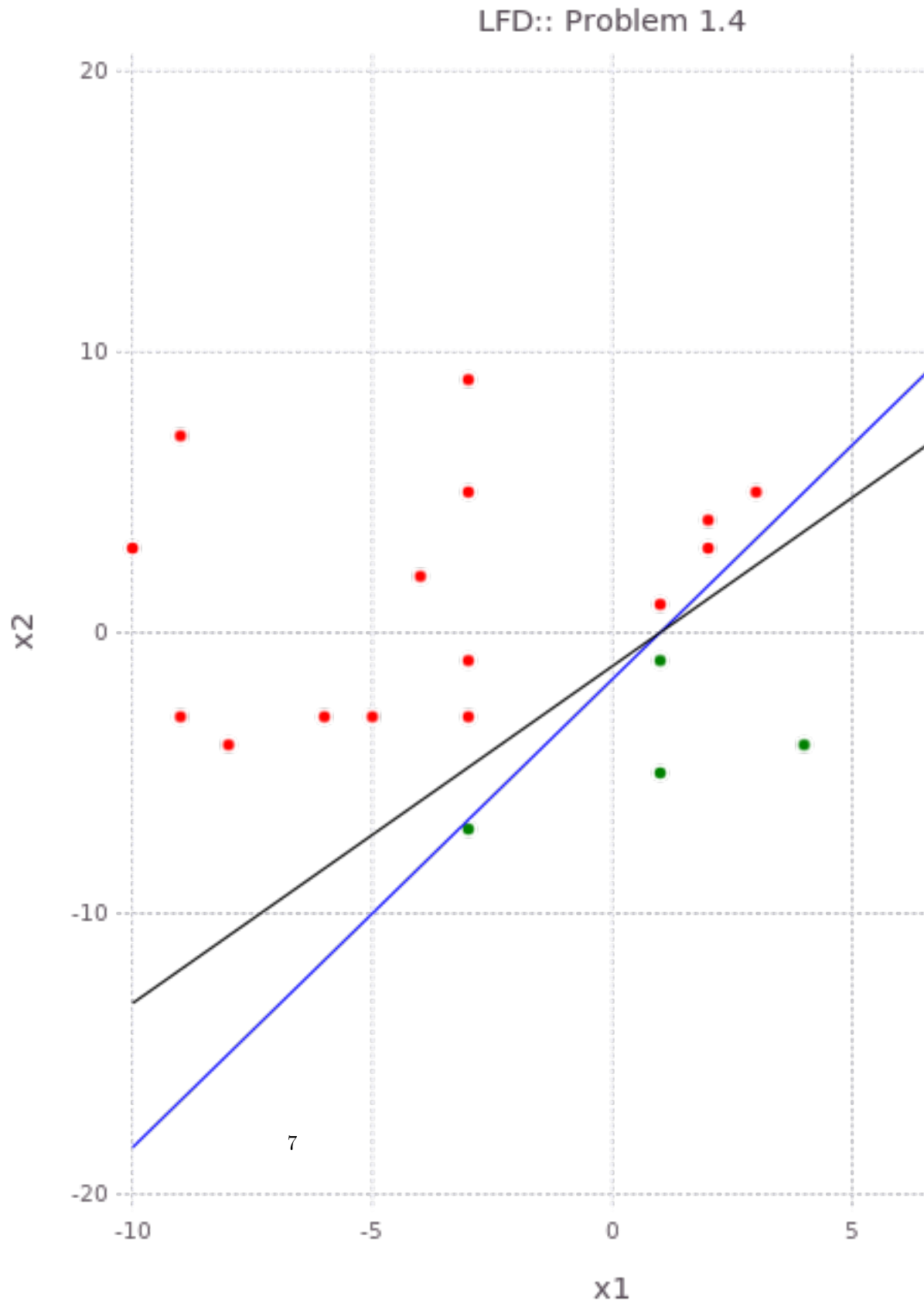
To see PLA algorithm implementation and (a) to (e) sample data look at perceptron-problem-1-3.jl

Problem 1.4:

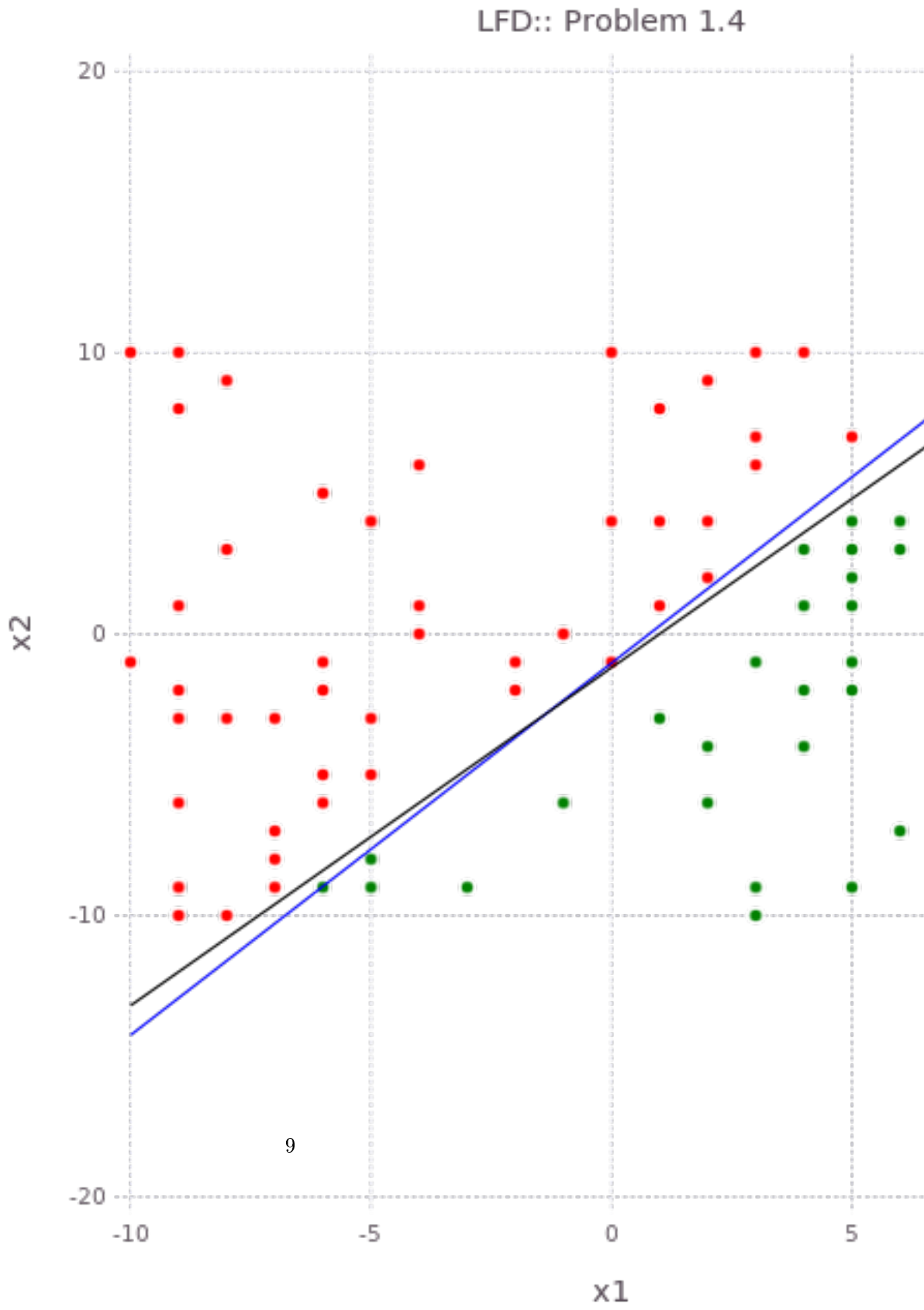
To play with PLA I use code from problem-1-4.jl.

(b) g is close to f but those are not parallel.

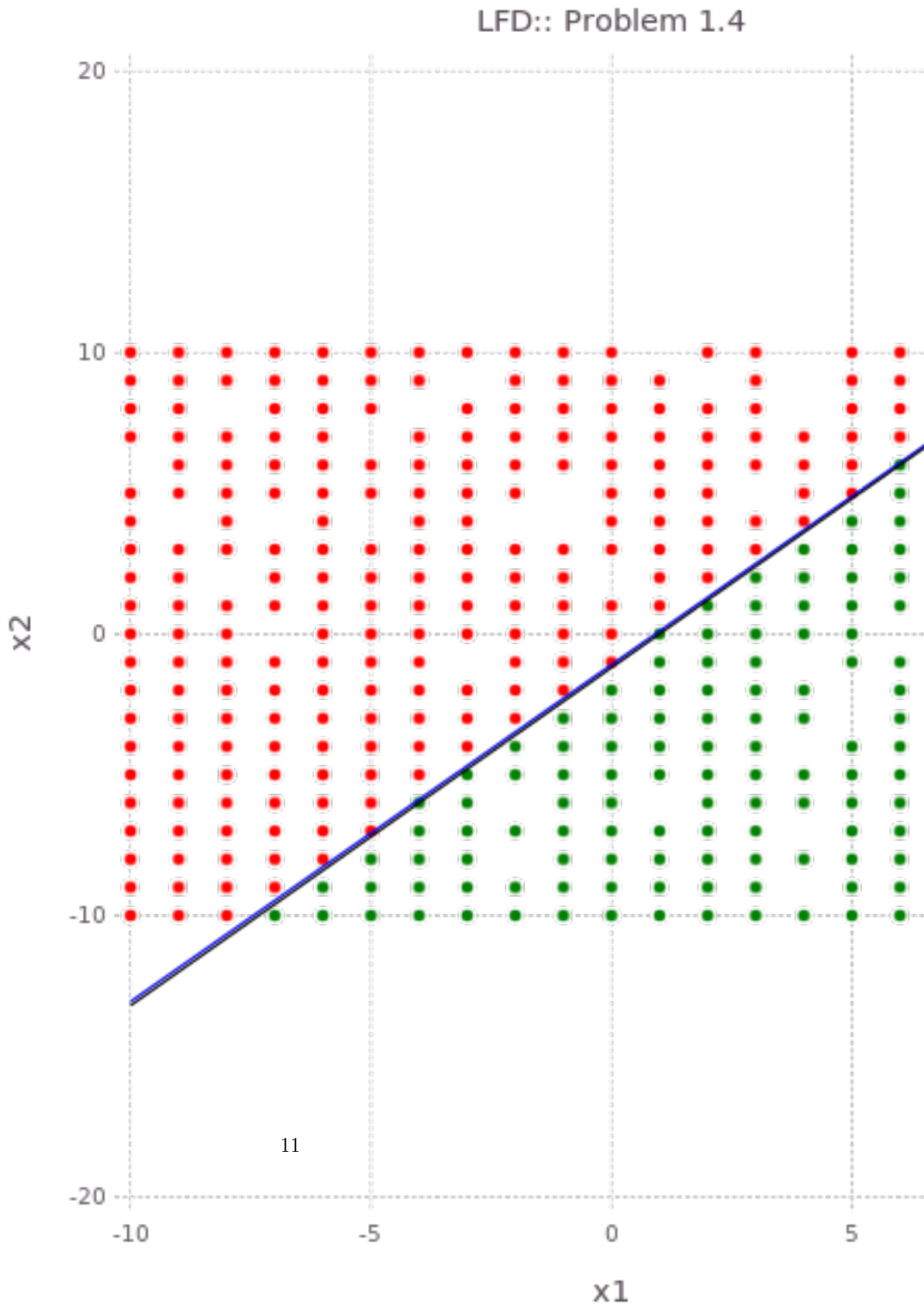
(c) As for (b) lines are not parallel.


LFD:: Problem 1.4
20
10
0
-10
-20
-10
-5
0
5
x1
x2

(d) Lines are not parallel but the angle is much smaller then for 20 samples. It means, that when number of samples grows then g approximates f better.

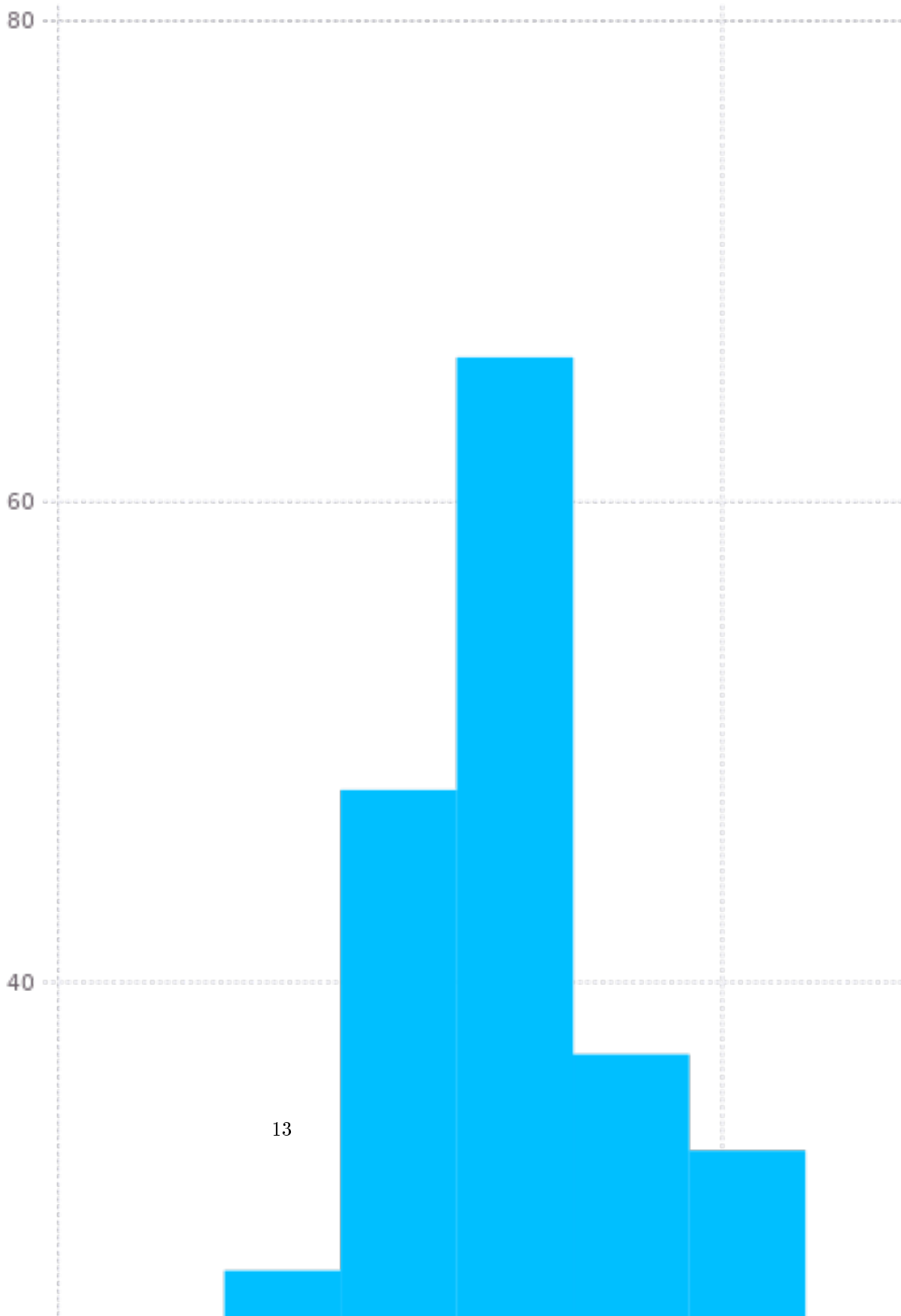

LFD:: Problem 1.4
20
10
0
-10
-20
-10
-5
0
5
x1
x2

(e) For 1,000 samples lines are almost parallel (hence we assume that $sign(.) = 0$ is trated as 1.0 PLA will not cover f by g).g approximates f almost perfectly.


LFD:: Problem 1.4
20
10
0
-10
-20
x2
-10
-5
0
5
x1

(f) In case of 1,000 samples in $\mathbb{R}^{10}$ space PLA needed roughly 3,500 iterations in one experiment (I carried out only one experiment).

(g) The histogram for 300 experiments.

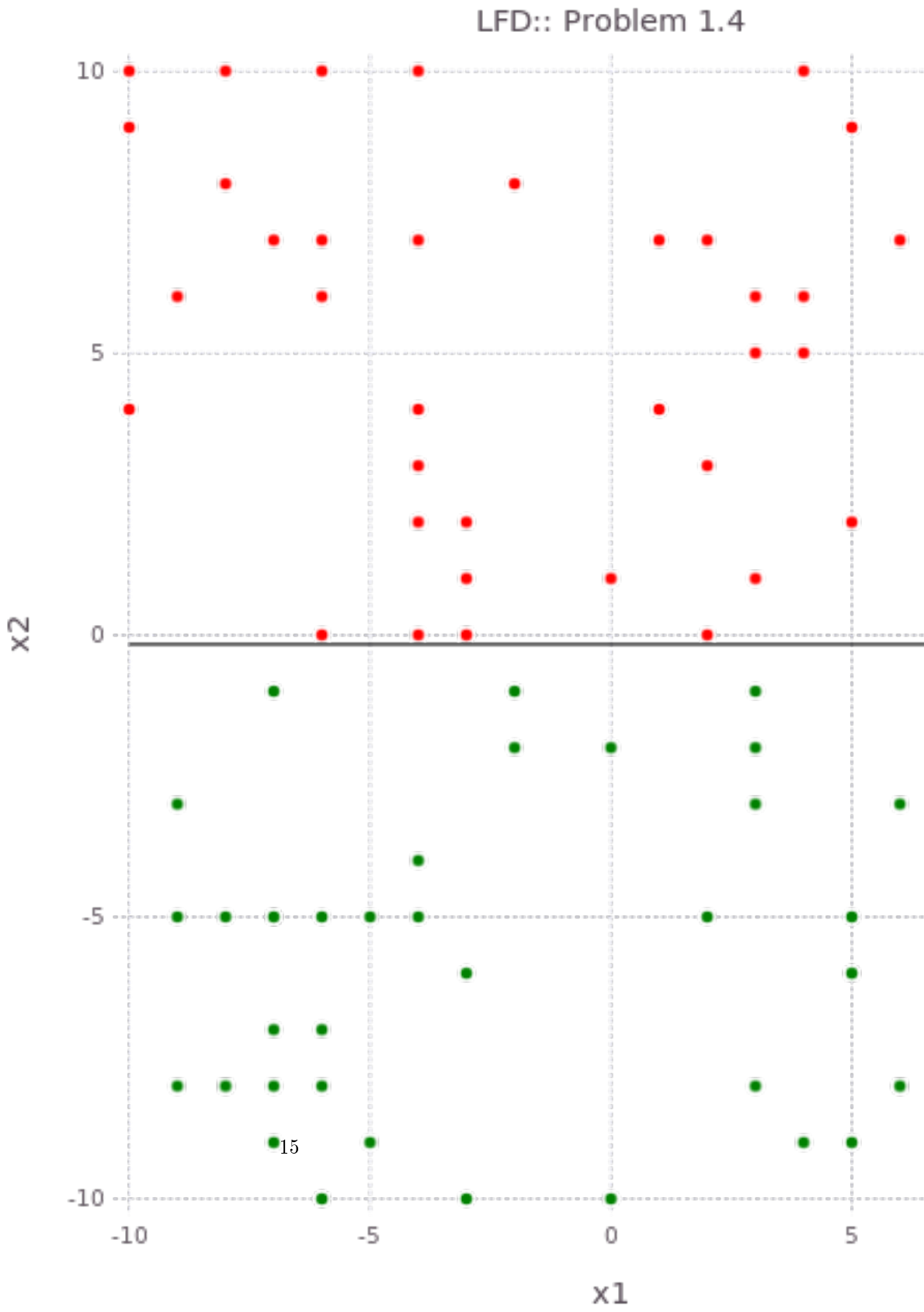

80
60
40

(h) In $\mathbb{R}^2$space PLA converges quickly and the result line (g) is more and more aligned with target function f when number of samples grows. In $d = 10$ space PLA needs more time to converge for the given 1,000 samples. As the histogram depicts (the tail) there are a lot of experiments (~40) where PLA couldn't have been able to converge. For $d = 10$ when sample size is small ($N = 10$, $N = 100$) PLA converges qucikly, but it does not mean it aproximates and generalizes enough (e.g. for $N = 10 => E_{out} \approx 0.65$, $N = 100 => E_{out} \approx 0.95$, since $sign(.) = 0$ is treated as 1.0 and PLA won't ever approximate fully f).

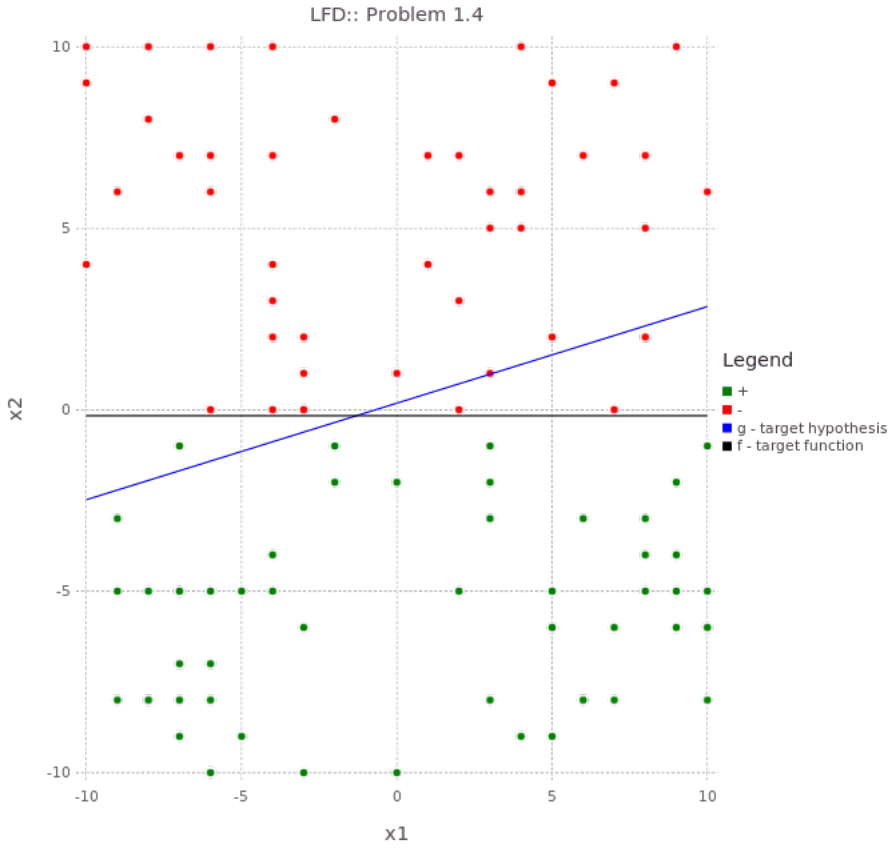
Problem 1.5:

To play with Adaline I use code from problem-1-5.jl.

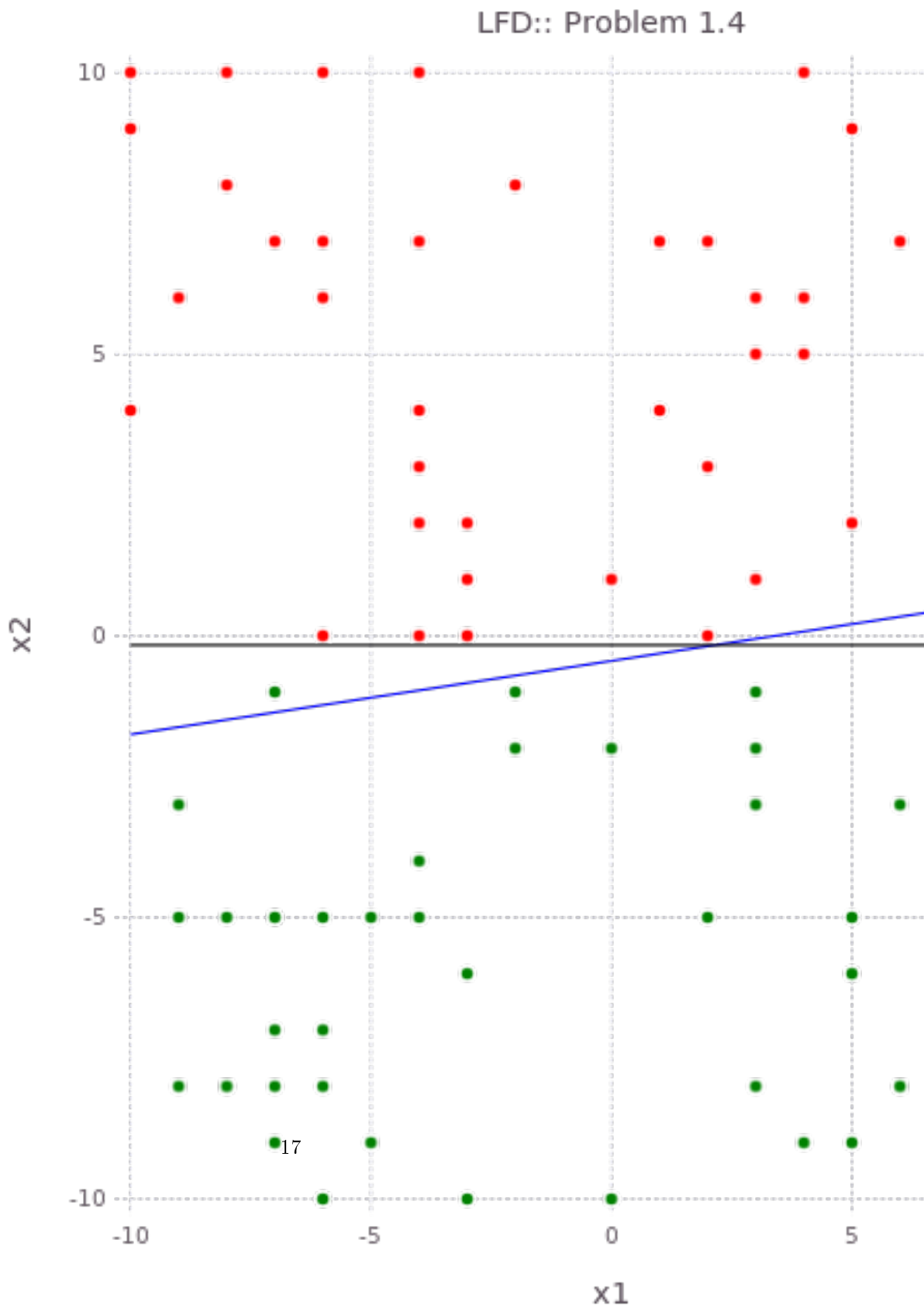
(a) For $\eta = 100$ Adaline out-of-sample error is 1.0 (each test data set is misclassified) and we can't converge as the plot depicts (there is no g)


LFD:: Problem 1.4
10
5
0
-5
-10
x2
-10
-5
0
5
x1

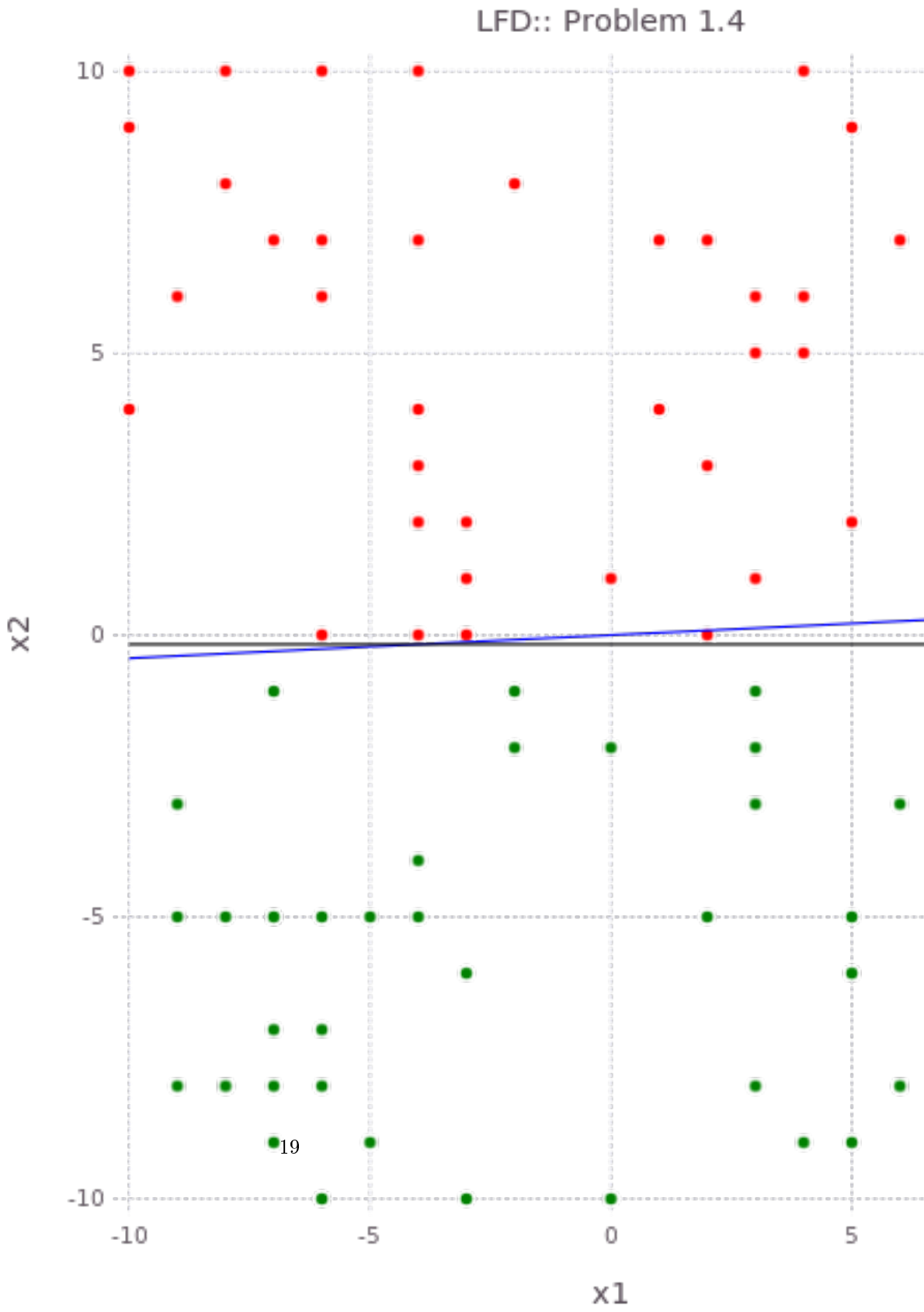
(b) For $\eta = 1.0$ Adaline out-of-sample error is 0.0645 and how well g matches f shows the plot below:



(c) For $\eta = 0.01$ Adaline out-of-sample error is 0.0274 and how well g matches f shows the plot below:


LFD:: Problem 1.4
10
5
0
-5
-10
-10
-5
0
5
x1
x2

(d) For $\eta = 0.0001$ Adaline out-of-sample error is 0.022 and how well g matches f shows the plot below:


LFD:: Problem 1.4
10
5
0
-5
-10
x2
-10
-5
0
5
x1

(e) According to results from (a) to (d) it shows that the smaller η is the better, but as (d) shows when η is too small the E_{out} no longer decreases. The whole algorithm does not behave as good as PLA and needs a lot of iterations.

Problem 1.6:

To see results please run problem-1_6.jl code. From nowμ is probability of getting red marble, and we want to have fraction of red marbles in our sample of 10 marbles to be equal 0.

(a) Ifμ is small then getting no red marbles is very probable in a sample of 10 marbles. When μ grows then getting no red marbles becomes possible but it's by no means probable.

(b) For $1,000$ experiments at a time cumulative probability of not getting red marble for small $\mu = 0.05$ is 1.0, for$\mu = 0.5$ is probable, but for $\mu = 0.8$ is implausible.

(c) For $1,000,000$ experiments at a time cumulative probability of not getting red marble for small and medium $\mu = 0.05, 0.5$ is 1.0, but for the largest value of μ is implausible.

Problem 1.7:

todo: add intuition and explanation

Problem 1.8:

(a) t is a non-negative random variable and α is an random variable value which satisfies the bound $\alpha > 0$
I'm going to proof Markov Inequality:

$$\mathbb{P}[t \geq \alpha] \leq \frac{\mathbb{E}[t]}{\alpha}$$

Proof: Let $\mathbf{1}_{t \geq \alpha}$be an indicator function on t.

$$\mathbf{1}_{t \geq \alpha} := \begin{cases} 1 & t \geq \alpha \\ 0 & t < \alpha \end{cases}$$

We multiply indicator function by α

$$\alpha \mathbf{1}_{\mathrm{t} \geq \alpha} \coloneqq \begin{cases} \alpha & t \geq \alpha \\ 0 & t < \alpha \end{cases}$$

because $\alpha > 0$ and t is a non-negative random variable we can be sure that:

$$\alpha \mathbf{1}_{\mathrm{t} \geq \alpha} \coloneqq \begin{cases} \alpha \leq t, & t \geq \alpha \\ 0 < t, & t < \alpha \end{cases}$$

hence

$$\alpha \mathbf{1}_{\mathrm{t} \geq \alpha} \leq t$$

we could compute expected value of the inequality:

$$\begin{aligned} \mathbb{E}[\alpha \mathbf{1}_{\mathrm{t} \geq \alpha}] &\leq \mathbb{E}[t] \\ \alpha \mathbb{E}[\mathbf{1}_{\mathrm{t} \geq \alpha}] &\leq \mathbb{E}[t] \\ \alpha[1\mathbb{P}[t \geq \alpha] + 0\mathbb{P}[t < \alpha]] &\leq \mathbb{E}[t] \\ \alpha \mathbb{P}[t \geq \alpha] &\leq \mathbb{E}[t] \\ \mathbb{P}[t \geq \alpha] &\leq \frac{\mathbb{E}[t]}{\alpha} \end{aligned}$$

(b) u is any random variable with mean μ and variance δ^2
I'm going to proof Chebyshev Inequality (for any $\alpha > 0$):

$$\mathbb{P}[(u - \mu)^2 \geq \alpha] \leq \frac{\delta^2}{\alpha}$$

To proof the inequality I use Markov Indequality:

$$\mathbb{P}[t \geq \alpha] \leq \frac{\mathbb{E}[t]}{\alpha}$$

but let $t \to (u-v)^2$,since $(u-v)^2$ is still a random variable and it holds Markov Inequaltiy prerequsites that t is a non-negative random variable.

$$\mathbb{P}[(u - v)^2 \geq \alpha] \leq \frac{\mathbb{E}[(u - v)^2]}{\alpha}$$

$$\mathbb{P}[(u - v)^2 \geq \alpha] \leq \frac{Var(u)}{\alpha}$$

$$\mathbb{P}[(u - v)^2 \geq \alpha] \leq \frac{\delta^2}{\alpha}$$

(c) $u_1, ..., u_N$ are iid random variables, each with mean μ and variance δ^2, and $u = \frac{1}{N}\sum_{n=1}^{N} u_n$ (is the same as u_N^-)
I'm going to proof $\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\delta^2}{N\alpha}$ (for any $\alpha > 0$) using general formula of Chebyshev Inequality ($\mathbb{P}[(X-\mu)^2 \geq \alpha] \leq \frac{\delta^2}{\alpha}$).
I replace X by another randoma variable u, hence

$$\mathbb{P}[(u-\mu)^2 \quad \geq \quad \alpha] \leq \frac{Var(u)}{\alpha}$$

while

$$\begin{aligned} Var(u) &= Var(\frac{1}{N}\sum_{n=1}^{N} u_n) \\ Var(u) &= \frac{1}{N^2}(\sum_{n=1}^{N} Var(u_n)) \\ Var(u) &= \frac{1}{N^2}\frac{N\delta^2}{1} = \frac{\delta^2}{N} \end{aligned}$$

, hence

$$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\delta^2}{N\alpha}$$

and when $N \to \infty$then $u \to \mu$ and $\lim_{N\to\infty} \mathbb{P}[|u-\mu| \geq \alpha] = 0$ (Weak Law of Large Numbers)

Problem 1.9:

(a) t is a (finite) random variable, α is a postive constant, and s is a postive parameter. If $T(s) = \mathbb{E}_t[e^{st}]$, I'm going to prove that

$$\mathbb{P}[t \geq \alpha] \leq e^{-s\alpha}T(s)$$

Proof:

$$\begin{aligned} \mathbb{P}[t &\geq \alpha] \leq e^{-s\alpha}T(s) \\ \mathbb{P}[t \geq \alpha] &\leq e^{-s\alpha}\mathbb{E}_t[e^{st}] \end{aligned}$$

To proof the inequality I use Markov Inequality $\mathbb{P}[t \geq \alpha] \leq \frac{\mathbb{E}[t]}{\alpha}$ and monotonicity of the exponent function for every $s > 0$

$$\mathbb{P}[a > b] = \mathbb{P}[e^{sa} > e^{sb}]$$

hence
$\mathbb{P}[t \geq \alpha] = \mathbb{P}[e^{st} \geq e^{s\alpha}] \leq \frac{\mathbb{E}[e^{st}]}{e^{s\alpha}}$

$$\begin{aligned}\mathbb{P}[t \geq \alpha] &\leq \frac{\mathbb{E}_t[e^{st}]}{e^{s\alpha}} \\ \mathbb{P}[t \geq \alpha] &\leq e^{-s\alpha}\mathbb{E}_t[e^{st}] \\ \mathbb{P}[t \geq \alpha] &\leq e^{-s\alpha}T(s)\end{aligned}$$

(b) $u_1, ..., u_N$are iid random variables and $u = \frac{1}{N}\sum_{n=1}^{N} u_n$ (is the same as $\bar{u_N}$). If $U(s) = \mathbb{E}_{u_n}(e^{su_n})$, I'm going to prove that

$$\mathbb{P}[u \geq \alpha] \leq (e^{-s\alpha}U(s))^N$$

Proof: I start with Markov Inequality for u which is a random variable.

$$\mathbb{P}[u \geq \alpha] \leq \frac{\mathbb{E}_u[u]}{\alpha}$$

from (a) I also know that $\mathbb{P}[t \geq \alpha] = \mathbb{P}[e^{sNt} \geq e^{sN\alpha}] \leq \frac{\mathbb{E}[e^{sNt}]}{e^{sN\alpha}}$, hence

$$\mathbb{P}[u \geq \alpha] \leq \frac{\mathbb{E}_u[e^{sNu}]}{e^{sN\alpha}}$$

let me focus on $\mathbb{E}_u[e^{sNu}]$

$$\begin{aligned}\mathbb{E}_u[e^{sNu}] &= \mathbb{E}_{u_n}[e^{sN\frac{1}{N}\sum_{n=1}^{N} u_n}] \\ \mathbb{E}_u[e^{sNu}] &= \mathbb{E}_{u_n}[e^{s(u_1+...+u_n]}]\end{aligned}$$

since $e^{x+y} = e^x e^y$ I get

$$\mathbb{E}_u[e^{sNu}] = \mathbb{E}_{u_n}[(e^{su_n})^N]$$

since $\mathbb{E}[X * X] = \mathbb{E}[X] * \mathbb{E}[X]$ I get

$$\mathbb{E}_u[e^{sNu}] = (\mathbb{E}_{u_n}[e^{su_n}])^N$$

Going back to $\mathbb{P}[u \geq \alpha] \leq \frac{\mathbb{E}_u[e^{sNu}]}{e^{sN\alpha}}$ I get

$$\begin{aligned}\mathbb{P}[u &\geq \alpha] \leq \frac{\mathbb{E}_u[e^{sNu}]}{e^{sN\alpha}} \\ \mathbb{P}[u &\geq \alpha] \leq \frac{(\mathbb{E}_{u_n}[e^{su_n}])^N}{e^{sN\alpha}} \\ \mathbb{P}[u &\geq \alpha] \leq \frac{(\mathbb{E}_{u_n}[e^{su_n}])^N}{(e^{s\alpha})^N} \\ \mathbb{P}[u &\geq \alpha] \leq \left(\frac{\mathbb{E}_{u_n}[e^{su_n}]}{e^{s\alpha}}\right)^N \\ \mathbb{P}[u &\geq \alpha] \leq (e^{-s\alpha}U(s))^N\end{aligned}$$

(c) Let $\mathbb{P}[u_n = 0] = \mathbb{P}[u_n = 1] = \frac{1}{2}$ (fair coin). I have to evaluate $U(s)$ as a function s,and minimize $e^{-s\alpha}U(s)$ with respect of s for fixed α, where $0 < \alpha < 1$. First I'm going to compute $U(s)$ for the given probability distribution.

$$U(s) = \mathbb{E}_{u_n}[e^{su_n}] = e^{s*0}\mathbb{P}[u_n = 0] + e^{s*1}\mathbb{P}[u_n = 1] = \frac{1}{2}(1 + e^s)$$

Now, I'm ready to minimize $e^{-s\alpha}U(s) = \frac{1}{2}e^{-s\alpha} + \frac{1}{2}e^{s(1-\alpha)}$, so I set the derivative equal to 0, yielding

$$\begin{aligned}
(e^{-s\alpha}U(s))' &= (\frac{1}{2}e^{-s\alpha} + \frac{1}{2}e^s e^{-s\alpha})' = (\frac{1}{2}e^{-s\alpha} + \frac{1}{2}e^{s(1-\alpha)})' = -\frac{\alpha}{2}e^{-s\alpha} + (\frac{1}{2} - \frac{\alpha}{2})e^{s(1-\alpha)} = 0 \\
(\frac{1}{2} - \frac{\alpha}{2})e^{s(1-\alpha)} &= \frac{\alpha}{2}e^{-s\alpha} \\
(1-\alpha)e^{s(1-\alpha)} &= \alpha e^{-s\alpha} \\
\frac{(1-\alpha)}{\alpha}\frac{e^{s(1-\alpha)}}{e^{-s\alpha}} &= 1 \\
\frac{(1-\alpha)}{\alpha}e^{s(1-\alpha)}e^{s\alpha} &= \frac{\alpha}{1-\alpha} \\
e^{s(1-\alpha)+s\alpha} &= \frac{\alpha}{1-\alpha} \\
e^s &= \frac{\alpha}{1-\alpha} \\
\ln(e^s) &= \ln(\frac{\alpha}{1-\alpha}) \\
s &= \ln(\frac{\alpha}{1-\alpha})
\end{aligned}$$

$e^{-s\alpha}U(s)$ gets the minimum value for $s = \ln(\frac{\alpha}{1-\alpha})$. Let me compute this min value:

$$e^{-\ln(\frac{\alpha}{1-\alpha})\alpha}U(\ln(\frac{\alpha}{1-\alpha})) = \frac{1}{2}e^{-\ln(\frac{\alpha}{1-\alpha})\alpha} + \frac{1}{2}e^{\ln(\frac{\alpha}{1-\alpha})(1-\alpha)} = \frac{1}{2}(e^{-\ln(\frac{\alpha}{1-\alpha})})^\alpha + \frac{1}{2}(e^{\ln(\frac{\alpha}{1-\alpha})})^{(1-\alpha)} =$$
$$= \frac{1}{2}(\frac{1}{e^{\ln(\frac{\alpha}{1-\alpha})}})^\alpha + \frac{1}{2}(e^{\ln(\frac{\alpha}{1-\alpha})})^{(1-\alpha)} = \frac{1}{2}(\frac{1-\alpha}{\alpha})^\alpha + \frac{1}{2}(\frac{\alpha}{1-\alpha})^{(1-\alpha)} = \frac{1}{2}(\frac{1-\alpha}{\alpha})^\alpha + \frac{1}{2}(\frac{1-\alpha}{\alpha})^{\alpha-1}$$

(e) I must conclude from (c) that , for $0 < \epsilon < \frac{1}{2}$

$$\mathbb{P}[u \geq \mathbb{E}(u) + \epsilon] \leq 2^{-\beta N}$$

where $\beta = 1 + (\frac{1}{2} + \epsilon)\log_2(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon)\log_2(\frac{1}{2} - \epsilon)$ and $\mathbb{E}(u) = \frac{1}{2}$.

To do this let $0 < \alpha = \mathbb{E}(u) + \epsilon < 1$, where $\mathbb{E}(u) = \frac{1}{2}$ in

$$\mathbb{P}[u \geq \alpha] \leq (e^{-s\alpha}U(s))^N$$
$$\mathbb{P}[u \geq \mathbb{E}(u) + \epsilon] \leq (e^{-s(\mathbb{E}(u)+\epsilon)}U(s))^N$$

where $U(s) = \frac{1}{2}(1 + e^s)$

Because we are interested in the tightest bound I assumed that

$$s = \ln(\frac{\alpha}{1-\alpha})$$

For the RHS of the inequality I'm skipping the Nth power and I'm going to use $1 + \frac{1}{1-a} = \frac{a}{1-a}$to simplify the equation.

$$e^{-s(\alpha)}U(s) = \frac{1}{2}e^{-s\alpha}(1 + e^s) = \frac{1}{2}e^{-s\alpha}(1 + e^s) = \frac{1}{2}e^{-ln(\frac{\alpha}{1-\alpha})(\alpha)}(1 + e^{\ln(\frac{\alpha}{1-\alpha})}) = \frac{1}{2}(\frac{\alpha}{1-\alpha})^{-\alpha}(1 + \frac{\alpha}{1-\alpha})$$
$$= \frac{1}{2}\frac{\alpha^{-\alpha}}{(1-\alpha)^{-\alpha}}(\frac{1}{1-\alpha}) = \frac{1}{2}\alpha^{-\alpha}(1-\alpha)^{\alpha}(1-\alpha)^{-1} = \frac{1}{2}\alpha^{-\alpha}(1-\alpha)^{\alpha-1}$$

Using the above equation I'm going to show that

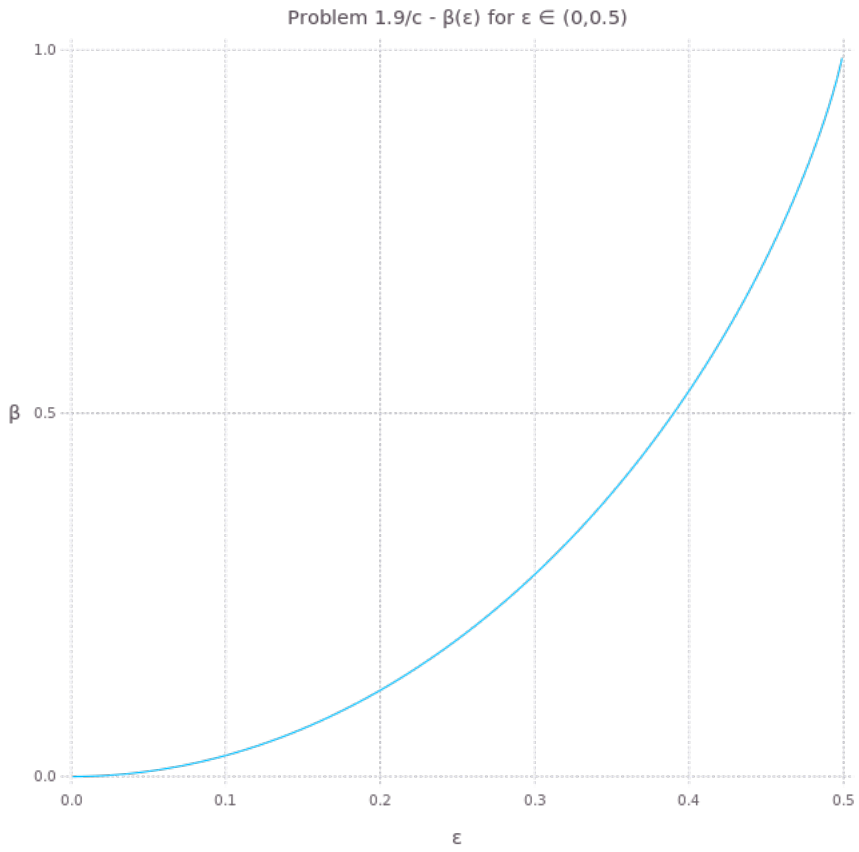
$$2^{-\beta N} = (e^{-s(\alpha)}U(s))^N$$
$$2^{-\beta} = e^{-s(\alpha)}U(s)$$
$$2^{-\beta} = \frac{1}{2}\alpha^{-\alpha}(1-\alpha)^{\alpha-1}$$
$$\log_2(-\beta) = \log_2(\frac{1}{2}\alpha^{-\alpha}(1-\alpha)^{\alpha-1})$$
$$-\beta = \log_2(\frac{1}{2}) + \log_2(\alpha^{-\alpha}) + \log_2((1-\alpha)^{\alpha-1})$$
$$-\beta = -1 - \alpha\log_2(\alpha) + (\alpha-1)\log_2(1-\alpha)$$
$$\beta = 1 + \alpha\log_2(\alpha) + (1-\alpha)\log_2(1-\alpha)$$

Since I assumed that$\alpha = \mathbb{E}(u) + \epsilon = \frac{1}{2} + \epsilon$ I get

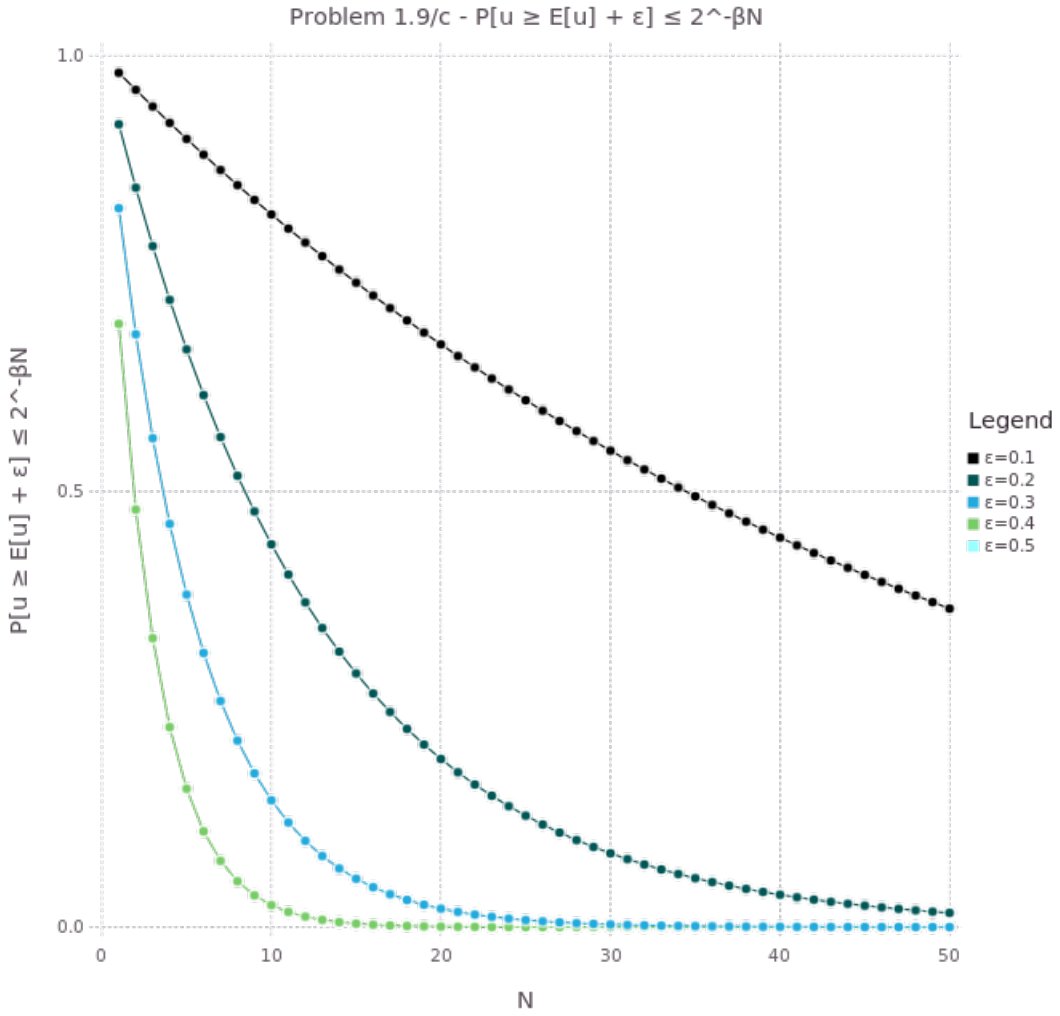
$\beta = 1 + (\frac{1}{2} + \epsilon)\log_2(\frac{1}{2} + \epsilon) + (\frac{1}{2} - \epsilon)\log_2(\frac{1}{2} - \epsilon)$,

what I had to conclude.

Accoring to tha task now I have to show that $\beta > \mathbf{0}$ when $0 < \epsilon < \frac{1}{2}$ what can be seen in the plot:



Hence the bound $2^{-\beta N}$ is expotentially decreasing in N what is shown in the plot below:



Problem 1.10:

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N, \mathbf{x}_{N+1}, ..., \mathbf{x}_{N+M}\}$ and $\mathcal{Y} = \{-1, 1\}$, and target function $f : \mathcal{X} \to \mathcal{Y}$ by definition is uknown to us. Training data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_n)\}$. We define off-training-set-error of a hypothesis h with respect to f by

$$E_{off}(h, f) = \frac{1}{M} \sum_{m=1}^{M} [\![h(\mathbf{x}_{N+m}) \neq f(\mathbf{x}_{N+m})]\!]$$

(a) In this task we say that $f(x) = +1$for all $\mathbf{x}$ and

$$h(x) := \begin{cases} +1, & \text{for } x = x_k \text{and } k \text{ is odd and} 1 \leq k \leq M+N \\ -1, & \text{otherwise} \end{cases}$$

$E_{off}(h, f)$ in this case depends on the number of even numbers in the interval $[N+1, N+2, ..., N+M]$, since h gives the wrong output for an even index k of a datapoint. For any interval of the size M number of even numbers is given by the formula: $\frac{lastEven - firstEven}{2} + 1$, where the first and the last even are numbers in an interval. Because we do not know if M is either even or odd number we must use piecewise function $a(M)$ which give a total number of even numbers in the interval $N + 1 \leq k \leq M$:

$$\#even(k) = \begin{cases} \frac{k}{2} & \text{if } k \text{ is even;} \\ \frac{k-1}{2} & \text{if } k\text{is odd} \end{cases}$$

Info: If we must search for the number of odd numbers then the formula would look like this:

$$\#odd(l) = \begin{cases} \frac{k}{2} & \text{if } k \text{ is even;} \\ \frac{k+1}{2} & \text{if } k\text{is odd} \end{cases}$$

Hence

$$\begin{aligned} E_{off}(h, f) &= \frac{1}{M} \#even(M) \\ E_{off}(h, f) &= \begin{cases} \frac{1}{2} & \text{if } M \text{ is even;} \\ \frac{1}{2} - \frac{1}{2M} & \text{if } M \text{ is odd} \end{cases} \end{aligned}$$

It can be shown that it is the same as $E_{off}(h, f) = \frac{1}{M}(\lfloor \frac{N+M}{2} \rfloor - \lfloor \frac{N}{2} \rfloor)$. Proof to be added.

(b) We must recall that the assumption from (a) does not work here, and we still do not know function f. We also do not know the dimensionality of the datapoint in the input space $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N, \mathbf{x}_{N+1}, ..., \mathbf{x}_{N+M}\}$ but we know that this input space is fixed (all $\mathbf{x}_k$ for $1 \leq k \leq N+M$ are already set). In this case we have $\mathcal{D}$ of size N generated in a deterministic way, and $y_n = f(\mathbf{x}_n)$ $(1 \leq n \leq N)$ is not affected by any noise. So, how many possible $f : \mathcal{X} \to \mathcal{Y}$ can 'generate' $\mathcal{D}$? The subtle point in this case is the assumption: "For a fixed $\mathcal{D}$ of size$N$", which means that$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_n)\}$ is already generated. We can calculate how many possible outputs y_n for $1 \leq n \leq N$ we can get? Only 1. But there are remaining $\{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, ..., \mathbf{x}_{N+M}\}$ datapoints for which we can have 2^M possible values $\{-1, 1\}$. So, the anwer is 2^M.

(c) $h$is fixed on an unknown target function f. We have got off-traing-set of size M. Knowing that $0 \leq k \leq M$ how many of those f (the (b) result) satisfy $E_{off}(h,f) = \frac{k}{M}$?
The whole space of possibilities is hidden in off-training set. In addition we are not interested in the sequence of misclassified datapoints, so $1^k 2^M$is not applicable here. If we are interested only on the total number of such possibilities then the equation will be satisfied by combinations $\binom{M}{k}$. So, the number of f which 'generates' $\mathcal{D}$ and satisfies $E_{off}(h,f) = \frac{k}{M}$ is $\binom{M}{k}$.

(d) In this section of the problem $E_{off}(h,f)$is a random variable for which a probability distribuition is $\mathbb{P}[E(h,f) = \frac{k}{M}] = \frac{1}{M+1}$, because we can get $M+1$values of the random variable $\{\frac{0}{M}, \frac{1}{M}, \frac{2}{M}, ..., \frac{M}{M}\}$and each of them is equally likely in probability, since f is equally likely in probability. So, the expected value of $E_f[E_{off}(h,f)]$ may be computed as follows:

$$E_f[E_{off}(h,f)] = \frac{1}{M+1}\sum_{k=0}^{M}\frac{k}{M} = \frac{1}{M+1}\frac{1}{M}\sum_{k=0}^{M}k = \frac{1}{(M+1)M}\frac{(M+1)M}{2} = \frac{1}{2}$$

$E_f[E_{off}(h,f)]$ does not depend on h and is the same for any h. Since f gives $\{-1,1\}$with the same probability no matter what datapoints it takes, so let take an exact case. The value of f is obtained by flipping a fair coin. We can get two values in such a case: $\{-1,+1\}$. It points out that f is a random variable hence $\mathbb{P}[f(\mathbf{x}_n) = +1] = \mathbb{P}[f(\mathbf{x}_n) = -1] = \frac{1}{2}$. Now let's assume that a given h is going to be a constant function $h(\mathbf{x}_n) = +1$. So, what value $E_f[E_{off}(h,f)]$ is going to take? $E_{off}(h,f)$ is 0 if $f(\cdot) = +1$, and 1when $f(\cdot) = -1$. $E_{off}(h,f) = 0\mathbb{P}[f(\mathbf{x}_n) = +1] + 1\mathbb{P}[f(\mathbf{x}_n) = -1] = \frac{1}{2}$.

This point shows that if you are in a learning situation with a random f, then no matter what you do in-sample, your out-of-sample performance will be very bad.

(e) In this case we still do not know f and the values of f are equally likely in probability, hence the f is a random variable (read carefully the whole point (d) and the description below that point). According to (d) the expected value of $E_f[E_{off}(h,f)]$ does not depend on h, so no matter what algorithm we use, the $E_f[E_{off}(h,f)]$ will be the same, hence $E_f[E_{off}(A_1(D),f)] = E_f[E_{off}(A_2,f)]$.

This point confirms that if you are in a learning situation with a random f, then no matter what you do in-sample, your out-of-sample performance will be very bad.

Problem 1.11

$E_{in}$is defined as $E_{in}(h) = \frac{1}{N}[\![h(x_n) \neq f(x_n)]\!] = \frac{1}{N}[\![h(x_n) \neq y_n]\!]$. In the problem I'm supposed to write down the in-sample error E_{in} that one should minimize to obtain g. Using the risk matrices I should weight the different types of error (false accept, false reject). There are two risk matrices: one for Supermakret example, and the second one for CIA example.

Supermarket	f			CIA	f		
		$+1$	-1			$+1$	-1
h	$+1$	0	1	h	$+1$	0	1000
	-1	10	0		-1	1	0

To calculate in-sample error I need data. But there is no data in the problem, so let me assume that I have one dataset $\mathcal{D} = \{(\mathbf{x}_1, +1), (\mathbf{x}_2, -1), (\mathbf{x}_3, -1), (\mathbf{x}_4, +1)\}$, hence $(N = 4)$ and two constant hipothesis $h_+(\mathbf{x}_n) = +1$, $h_-(\mathbf{x}_n) = -1$, so let $\mathcal{H} = \{h_+, h_-\}$.

In the case of the Supermarket example: $E_{in}(h_+) = \frac{1}{4}(0 + 1 + 1 + 0) = \frac{1}{2}$, $E_{in}(h_-) = \frac{1}{4}(10 + 0 + 0 + 10) = 5$

In the case of CIA example: $E_{in}(h_+) = \frac{1}{4}(0 + 1000 + 1000 + 0) = 500$, $E_{in}(h_-) = \frac{1}{4}(1 + 0 + 0 + 1) = \frac{1}{2}$
Final hypothesis will be different for each of considered examples. In the case of Supermarket it's going to be $h_+$and in the case of CIA it'll be h_-.

Problem 1.12

In this case the author of the problem wants me to find a value which minimize each of the error functions and he calls each value as 'representative' value for a given algorithm. In the problem we have N datapoints and we know that $y_1 \leq y_2 \leq ... \leq y_N$(the crucial part of the problem definition).

(a)

In the case of algorithm that minimizes the in-sample sum of squared deviations to find the hypothesis $h$which can be treated as the final hypothesis g we have:

$$E_{in}(h) = \sum_{n=1}^{N} (h - y_n)^2$$

I have to show that $h_{mean} = \sum_{n=1}^{N} y_n$ minimizes this const function, which means

$$h_{mean} = \arg\min_h E_{in}(h) = \arg\min_h \sum_{n=1}^{N} (h - y_n)^2$$

.
To find the value that minimizes this cost function I must find the derivative

and then find for which value is equal to 0.

$$E_{in}^{'}(h) = (\sum_{n=1}^{N}(h - y_n)^2)' = 2\sum_{n=1}^{N}(h - y_n)$$

Now, let me find the value of h for which $E_{in}^{'}(h)$ equals 0.

$$\begin{aligned} E_{in}^{'}(h) &= 0 \\ 2\sum_{n=1}^{N}(h - y_n) &= 0 \\ \sum_{n=1}^{N} h &= \sum_{n=1}^{N} y_n \\ Nh &= \sum_{n=1}^{N} y_n \\ h &= \frac{1}{N}\sum_{n=1}^{N} y_n = h_{mean} \end{aligned}$$

So, if one uses the sum of squared deviations as the cost function then h_{mean}minimizes that cost function.

(b)

In the case of algorithm that minimizes the in-sample sum of absolute deviations to find the hypothesis $h$which can be treated as the final hypothesis g we have:

$$E_{in}(h) = \sum_{n=1}^{N} |h - y_n|$$

I have to show that h_{median} minimizes this const function, which means

$$h_{median} = \arg\min_{h} E_{in}(h) = \arg\min_{h} \sum_{n=1}^{N} |h - y_n|$$

One should notice that $\frac{d|x|}{dx} = sign(x)$, hence $\frac{d|x-a|}{dx} = sign(x - a)$, which leads to equation

$$E_{in}^{'}(h) = (\sum_{n=1}^{N} |h - y_n|)' = \sum_{n=1}^{N} sign(h - y_n)$$

Now, I must find such h for which $E_{in}^{'}(h)$ equals 0. The sum of all $-1$and all $+1$ in the set $\{sign(h - y_1), sign(h - y_2), ..., sign(h - y_N)\}$ is equal to 0 only if the number of positve items $(+1)$ equals the number of negative items (-1) , which happens when $h = h_{median} = median(\{y_1, y_2, ..., y_N\})$.

(c)

We assume that in the set of items $y_1 \leq y_2 \leq ... \leq y_N$, y_N is an outlier since $y_N = y_N + \epsilon$, where $\epsilon \to \infty$. I must answer what happens to h_{mean}(a) and h_{median}(b) in such a case. Looking at (a) $\frac{1}{N}\sum_{n=1}^{N} y_n = h_{mean}$and (b) $h_{median} = median\{y_1, y_2, ..., y_N\}$ one should notice that for (a) h_{mean}significantly increases, but for the (b) h_{median}does not change. It can be concluded that Mean-Square-Error (a) is sensitive on outliers while Median-Absolute-Error (b) is not. (b) treats every datapoint the same, while (a) emphasizes the extremes - the square of a very small number (smaller than 1) is even smaller, and the square of a big number is even bigger. In the case of (a) outliers may have an impact on the prediction capabilities of final hypothesis. This side effect does not appear in the case of (b) and therefore can lead to better predictions when extreme outliers are present. (b) seems to be more demanding on methods and technques to minimize the cost function.

Extra: Mathematical proof for h_{mean} growth affected by an outlier can be found on `http://stats.stackexchange.com/questions/7032/effect-of-missing-data-and-outliers-o`