Problem Solutions

e-Chapter 7

Pierre Paquay

Problem 7.1

To solve this problem, we first begin by separating the positive decision region into two components : the lower one corresponding to $x_2 \in [-1, 1]$ and the upper one corresponding to $x_2 \in [1, 2]$. To define the decision region, we need 7 perceptrons, namely

$$h_1(x) = \text{sign}(x_2 - 2),\ h_2(x) = \text{sign}(x_2 - 1),\ h_3(x) = \text{sign}(x_2 + 1),$$

for the horizontal lines, and

$$h_4(x) = \text{sign}(x_1 + 2),\ h_5(x) = \text{sign}(x_1 + 1),\ h_6(x) = \text{sign}(x_1 - 1),\ h_7(x) = \text{sign}(x_1 - 2)$$

for the vertical lines. We are now able to define the lower decision region by $\overline{h_2} h_3 h_4 \overline{h_7}$, and the upper decision region by $\overline{h_1} h_2 h_5 \overline{h_6}$, which means that the total decision region is defined by

$$f = \overline{h_2} h_3 h_4 \overline{h_7} + \overline{h_1} h_2 h_5 \overline{h_6}$$

which actually characterizes a 3-layer perceptron.

Problem 7.2

(a) Let x and x' be two points from the same region. If we consider a set of M hyperplanes defined by $\{x : w_i^T x = 0\}$, we have that

$$(\text{sign}(w_1^T x), \cdots, \text{sign}(w_M^T x)) = (\text{sign}(w_1^T x'), \cdots, \text{sign}(w_M^T x'));$$

or put more simply that $\text{sign}(w_i^T x) = \text{sign}(w_i^T x') = s_i$ for $i = 1, \cdots, M$ where $s_i = \pm 1$. We begin by the case where $s_i = 1$. Here, we know that $w_i^T x > 0$ and $w_i^T x' > 0$, consequently we have that, for $\lambda \in [0, 1]$,

$$w_i^T (\lambda x + (1 - \lambda) x') = \lambda w_i^T x + (1 - \lambda) w_i^T x' > 0$$

and

$$\text{sign}(w_i^T (\lambda x + (1 - \lambda) x')) = 1.$$

Now, we consider the case where $s_i = -1$. Here, we know that $w_i^T x < 0$ and $w_i^T x' < 0$, consequently we have that, for $\lambda \in [0, 1]$,

$$w_i^T (\lambda x + (1 - \lambda) x') = \lambda w_i^T x + (1 - \lambda) w_i^T x' < 0$$

and

$$\text{sign}(w_i^T (\lambda x + (1 - \lambda) x')) = -1.$$

So, in conclusion, the region is actually convex.

(b) A region is defined as the following set

$$\{x : (\text{sign}(w_1^T x), \cdots, \text{sign}(w_M^T x)) = (s_1, \cdots, s_M); s_i \in \{-1, 1\}\};$$

thus a region is characterized by a particular M-uple $(s_1, \cdots, s_M)$. Since there are at most 2^M of such M-uples, we have at most 2^M different regions.

(c) Let $B(N, d)$ be the maximum number of regions created by M hyperplanes in d-dimensional space. Now, consider adding an $(M + 1)$th hyperplane; this hyperplane can obviously be viewed as a $(d - 1)$-dimensional space, so if we project the initial M hyperplanes into this space, we obtain M hyperplanes in a $(d-1)$-dimensional space. These hyperplanes can create at most $B(M, d-1)$ regions in this space, and for each of these regions, we get two regions in the original d-dimensional space. Thus, this means that the $(M+1)$th hyperplane intersects at most $B(M, d-1)$ of the regions created by the M hyperplanes in the d-dimensional space, and so

$$B(M+1, d) \leq B(M, d) + B(M, d-1).$$

Now, we will prove that

$$B(M, d) \leq \sum_{i=0}^{d} \binom{M}{i}$$

by induction. We begin by evaluating the boundary conditions, we have

$$B(M, 1) = M + 1 \leq \sum_{i=0}^{1} \binom{M}{i} = \binom{M}{0} + \binom{M}{1} = M + 1$$

for all M, and

$$B(1, d) = 2 \leq \sum_{i=0}^{d} \binom{1}{i} = \binom{1}{0} + \binom{1}{1} = 2$$

for all d. Now, we assume the statement is true for $M = M_0$ and all d, we will prove that the statement is still true for $M = M_0 + 1$ and all d. We have that

$$\begin{aligned}
B(M_0+1, d) &\leq B(M_0, d) + B(M_0, d-1) \\
&\leq \sum_{i=0}^{d} \binom{M_0}{i} + \sum_{i=0}^{d-1} \binom{M_0}{i} \\
&= \binom{M_0}{0} + \sum_{i=1}^{d} \binom{M_0}{i} + \sum_{i=1}^{d} \binom{M_0}{i-1} \\
&= 1 + \sum_{i=1}^{d} \Bigg[\underbrace{\binom{M_0}{i} + \binom{M_0}{i-1}}_{=\binom{M_0+1}{i}} \Bigg] \\
&= \sum_{i=0}^{d} \binom{M_0+1}{i}.
\end{aligned}$$

We have thus proved the induction step, so the statement is true for all M and d.

Problem 7.3

We begin by proving the following equivalence relation

$$h_m(x) = c_m \Leftrightarrow h_m^{c_m}(x) = +1.$$

The condition is necessary because if $c_m = +1$, we have

$$h_m^{c_m}(x) = h_m(x) = c_m = +1;$$

and if $c_m = -1$, we have

$$h_m^{c_m}(x) = \overline{h}_m(x) = \overline{c}_m = +1.$$

Now the condition is also sufficient because if $c_m = +1$, we have

$$+1 = h_m^{c_m}(x) = h_m(x),$$

which means that $h_m(x) = +1 = c_m$; and if $c_m = -1$, we have

$$+1 = h_m^{c_m}(x) = \overline{h}_m(x),$$

which implies that $h_m(x) = -1 = c_m$.

Now we are able to write that

$$\begin{array}{cc} & x \in r \\ \Leftrightarrow & (h_1(x), \cdots, h_M(x)) = (c_1, \cdots, c_M) \\ \Leftrightarrow & h_m^{c_m}(x) = +1, \ \forall m \\ \Leftrightarrow & \prod_{m=1}^{M} h_m^{c_m}(x) = +1 \\ \Leftrightarrow & t_r(x) = +1. \end{array}$$

The above relation also implies that

$$x \notin r \Leftrightarrow t_r(x) = -1.$$

Now if x is in a positive region ($f(x) = +1$), we know that there exists i such that $x \in r_i$, and consequently that $t_{r_i}(x) = +1$ which means that

$$t_{r_1}(x) + \cdots + t_{r_k}(x) = +1 = f(x).$$

And if x is in a negative region ($f(x) = -1$), we know that $x \notin r_i$ for all i, so $t_{r_i}(x) = -1$ for all i which means that

$$t_{r_1}(x) + \cdots + t_{r_k}(x) = -1 = f(x).$$

Problem 7.4

Since $f = t_{r_1} + \cdots + t_{r_k}$, we may write that

$$f = \mathrm{sign}(k - \frac{1}{2} + \sum_{i=1}^{k} t_{r_i}),$$

which characterizes the penultimate layer of our perceptron. For the layer before, we have that $t_{r_i} = h_1^{c_1^{(i)}} \cdots h_M^{c_M^{(i)}}$, and consequently

$$t_{r_i} = \mathrm{sign}(-M + \frac{1}{2} + \sum_{m=1}^{M} h_m^{c_m^{(i)}});$$

moreover, the previous layer may be characterized with

$$h_m^{c_m^{(i)}} = \mathrm{sign}(c_m^{(i)} w_m^T x).$$

Putting all this together, we obtain the following characterization of a 3-layer perceptron

$$f = \mathrm{sign}(k - \frac{1}{2} + \sum_{i=1}^{k} \mathrm{sign}(-M + \frac{1}{2} + \sum_{m=1}^{M} \mathrm{sign}(c_m^{(i)} w_m^T x)))$$

whose structure is given by $[d, kM, k, 1]$.

Problem 7.5

First, we decompose the unit hypercube $[0,1]^d$ into $1/\epsilon^d$ ϵ-hypercubes (hypercube whose sides have length equal to ϵ), thus we get a grid-like structure of our unit hypercube. Now, if we consider a decision region (which may be composed by disconnected regions) whose boundary surfaces are smooth, this decision region partition the unit hypercube into two regions : one labelled $+1$ and one labelled -1. We now have k_ϵ ϵ-hypercubes labelled $+1$ which are formed by $2d$ hyperplanes each defined by $h_m^{(i)} = \text{sign}(w_m^{(i),T} x)$ where $m = 1, \cdots, 2d$ and $i = 1, \cdots, k_\epsilon$. So, the first layer whose task is to activate the hyperplanes involved in the positive ϵ-hypercubes is characterized by

$$h_m^{(i)} = \text{sign}(w_m^{(i),T} x).$$

Now to activate the positive ϵ-hypercubes H_i themselves we characterize the second layer by

$$t_{H_i} = (h_1^{(i)})^{c_1^{(i)}} \cdots (h_{2d}^{(i)})^{c_{2d}^{(i)}},$$

where the $c_m^{(i)}$ are defined as in Problem 7.3 and 7.4; or

$$t_{H_i} = \text{sign}(-2d + \frac{1}{2} + \sum_{m=1}^{2d} (h_m^{(i)})^{c_m^{(i)}}).$$

And finally to activate all the positive ϵ-hypercubes, we define the MLP output h by

$$h = t_{H_1} + \cdots + t_{H_{k_\epsilon}};$$

or

$$h = \text{sign}(k_\epsilon - \frac{1}{2} + \sum_{i=1}^{k_\epsilon} t_{H_i}).$$

Putting all this together, we obtain the following characterization of a 3-layer perceptron

$$h = \text{sign}(k_\epsilon - \frac{1}{2} + \sum_{i=1}^{k_\epsilon} \text{sign}(-2d + \frac{1}{2} + \sum_{m=1}^{2d} \text{sign}(c_m^{(i)} w_m^{(i),T} x))).$$

Now, it remains to see that the above MLP can aribtrarily closely approximate the initial positive decision region D_+ (and consequently the negative decision region also); to do so, we first note that

$$\text{Vol}(H_i) = \epsilon^d \to 0 \text{ and } k_\epsilon \to \infty$$

when $\epsilon \to 0$. So, the ϵ-hypercubes can be made arbitrarily small, which obviously means that the total volume of the positive ϵ-hypercubes can be made arbitrarily close to the volume of the positive decision region (because of its smoothness). Mathematically, we may write that

$$\text{Vol}(H_1 \cup \cdots \cup H_{k_\epsilon}) = \sum_{i=1}^{k_\epsilon} \epsilon^d \to \text{Vol}(D_+)$$

when $\epsilon \to 0$. This means that the region where our 3-layer perceptron will output $+1$ (resp. -1) converges to the positive (resp. negative) decision region in our unit hypercube.

Problem 7.6

For a specific layer l, if we replace the weight $w_{ij}^{(l)}$ with $w_{ij}^{(l)} + \epsilon$, we need to recompute the corresponding node output of that layer and also the node outputs for the subsequent layers (which are the ones numbered from $l+1$ to L). Consequently, for each weight $w_{ij}^{(l)}$, we have

$$\sum_{k=l+1}^{L} d^{(l)}(d^{(l-1)} + 1) + 1 + \sum_{k=l+1}^{L} d^{(l)}$$

multiplications and θ-evaluations respectively; this means that the computational complexity of obtaining the partial derivatives is overall equal to

$$\begin{aligned}
&\underbrace{2\sum_{l=1}^{L} d^{(l)}(d^{(l-1)}+1)}_{=|W|}\left(\sum_{k=l+1}^{L} d^{(l)}(d^{(l-1)}+1)+1+\sum_{k=l+1}^{L} d^{(l)}\right)\\
\leq\;& 2|W|\left(\underbrace{\sum_{k=1}^{L} d^{(l)}(d^{(l-1)}+1)}_{=|W|}+1+\sum_{k=1}^{L}\underbrace{d^{(l)}}_{\leq d^{(l)}(d^{(l-1)}+1)}\right)\\
\leq\;& 2|W|(2|W|+1)=O(|W|^2)
\end{aligned}$$

since we need to compute the derivatives corresponding to $w_{ij}^{(l)}+\epsilon$ and also to $w_{ij}^{(l)}-\epsilon$.

Problem 7.7

(a) We know that

$$E_{in}=\frac{1}{N}\sum_{i=1}^{N}||y_i-\hat{y}_i||^2,$$

and also that

$$(Y-\hat{Y})(Y-\hat{Y})^T=\begin{pmatrix} y_1^T-\hat{y}_1^T\\ \vdots\\ y_N^T-\hat{y}_N^T\end{pmatrix}(y_1-\hat{y}_1,\cdots,y_N-\hat{y}_N)=\begin{pmatrix}||y_1-\hat{y}_1||^2 & * & *\\ * & \ddots & *\\ * & * & ||y_N-\hat{y}_N||^2\end{pmatrix}.$$

Consequently, we get that

$$E_{in}=\frac{1}{N}\text{trace}((Y-\hat{Y})(Y-\hat{Y})^T).$$

(b) We may write that

$$\begin{aligned}
E_{in} &= \frac{1}{N}\text{trace}(YY^T-ZVY^Y-YV^TZ^T+ZVV^TZ^T)\\
&= \frac{1}{N}\text{trace}(YY^T-2ZVY^T+ZVV^TZ^T),
\end{aligned}$$

since $\text{trace}(A)=\text{trace}(A^T)$. We are now ready to compute the derivatives, we have

$$\begin{aligned}
\frac{\partial E_{in}}{\partial V} &= \frac{1}{N}(-2\underbrace{\frac{\partial\text{trace}(ZVY^T)}{\partial V}}_{=Z^TY}+\underbrace{\frac{\partial\text{trace}(ZVV^TZ^T)}{\partial V}}_{=Z^TZV+Z^TZV=2Z^TZV})\\
&= \frac{1}{N}(2Z^TZV-2Z^TY),
\end{aligned}$$

because of the following identities

$$\frac{\partial\text{trace}(AXB)}{\partial X}=A^TB^T\text{ and }\frac{\partial\text{trace}(AXX^TB)}{\partial X}=BAX+A^TB^TX.$$

We also have

$$\begin{aligned}
&E_{in}\\
&= \frac{1}{N}\text{trace}(YY^T-2(V_0+\theta(XW)V_1)Y^T+(V_0+\theta(XW)V_1)(V_0^T+V_1^T\theta(XW)^T)\\
&= \frac{1}{N}\text{trace}(YY^T-2V_0Y^T-2\theta(XW)V_1Y^T+V_0V_0^T+V_0V_1^T\theta(XW)^T+\theta(XW)V_1V_0^T+\theta(XW)V_1V_1^T\theta(XW)^T)\\
&= \frac{1}{N}\text{trace}(YY^T-2V_0Y^T-2\theta(XW)V_1Y^T+V_0V_0^T+2\theta(XW)V_1V_0^T+V_1V_1^T\theta(XW)^T\theta(XW)),
\end{aligned}$$

since the trace can be permuted in a cycle and $\text{trace}(A) = \text{trace}(A^T)$. The other derivative may be written as

$$\begin{aligned}
\frac{\partial E_{in}}{\partial W} &= \frac{1}{N}(-2\frac{\partial \text{trace}(\theta(XW)V_1Y^T)}{\partial W} + 2\frac{\partial \text{trace}(\theta(XW)V_1V_0^T)}{\partial W} + \frac{\partial \text{trace}(V_1V_1^T\theta(XW)^T\theta(XW))}{\partial W}) \\
&= \frac{1}{N}(-2X^T\theta'(XW)\otimes YV_1^T + 2X^T\theta'(XW)\otimes V_0V_1^T + X^T(\theta'(XW)\otimes[\theta(XW)2V_1V_1^T]) \\
&= 2X^T[\theta'(XW)\otimes(-YV_1^T + V_0V_1^T + \theta(XW)V_1V_1^T)]
\end{aligned}$$

because of the following identities

$$\frac{\partial \text{trace}(\theta(BX)A)}{\partial X} = B^T\theta'(BX)\otimes A^T \text{ and } \frac{\partial \text{trace}(A\theta(BX)^T\theta(BX))}{\partial X} = B^T[\theta'(BX)\otimes(\theta(BX)(A+A^T))].$$

Problem 7.8

(a) By hypothesis, we know that $\{\eta_1, \eta_2, \eta_3\}$ with $\eta_1 < \eta_2 < \eta_3$ is an U-arrangement which means that

$$E(\eta_2) < \min\{E(\eta_1), E(\eta_3)\}.$$

Since $E(\eta)$ is a quadratic curve, we know that it is decreasing (resp. increasing) to the left (resp. right) of its minimum $\bar{\eta}$. So if we assume that $\bar{\eta} < \eta_1$, we get that $E(\eta_1) \le E(\eta_2) \le E(\eta_3)$ which is impossible by definition of an U-arrangement; and if we assume that $\bar{\eta} > \eta_3$, we get that $E(\eta_1) \ge E(\eta_2) \ge E(\eta_3)$ which is also impossible by definition of an U-arrangement. Consequently, we have $\bar{\eta} \in [\eta_1, \eta_3]$.

(b) First, we solve the linear system in a, b, and c below

$$\begin{cases} E(\eta_1) &= a\eta_1^2 + b\eta_1 + c = e_1 \\ E(\eta_2) &= a\eta_2^2 + b\eta_2 + c = e_2 \\ E(\eta_3) &= a\eta_3^2 + b\eta_3 + c = e_3 \end{cases}.$$

Let D be the determinant of the system, which is

$$D = \begin{vmatrix} \eta_1^2 & \eta_1 & 1 \\ \eta_2^2 & \eta_2 & 1 \\ \eta_3^2 & \eta_3 & 1 \end{vmatrix},$$

where $D \neq 0$ since $\eta_1 < \eta_2 < \eta_3$; now we easily get that

$$a = \begin{vmatrix} e_1 & \eta_1 & 1 \\ e_2 & \eta_2 & 1 \\ e_3 & \eta_3 & 1 \end{vmatrix} / D = \frac{(e_1 - e_2)(\eta_1 - \eta_3) - (e_1 - e_3)(\eta_1 - \eta_2)}{D}$$

and

$$b = \begin{vmatrix} \eta_1^2 & e_1 & 1 \\ \eta_2^2 & e_2 & 1 \\ \eta_3^2 & e_3 & 1 \end{vmatrix} / D = \frac{-(e_1 - e_2)(\eta_1^2 - \eta_3^2) + (e_1 - e_3)(\eta_1^2 - \eta_2^2)}{D}.$$

Since the minimum of such a quadratic function is given by $-b/2a$, we finally get

$$\bar{\eta} = \frac{1}{2}\left[\frac{(e_1 - e_2)(\eta_1^2 - \eta_3^2) - (e_1 - e_3)(\eta_1^2 - \eta_2^2)}{(e_1 - e_2)(\eta_1 - \eta_3) - (e_1 - e_3)(\eta_1 - \eta_2)}\right].$$

(c) We enumerate the four cases below.

1. If $\bar{\eta} < \eta_2$:
   - If $E(\bar{\eta}) < E(\eta_2)$, then $\{\eta_1, \bar{\eta}, \eta_2\}$ is a new U-arrangement.

- If $E(\bar{\eta}) > E(\eta_2)$, then $\{\bar{\eta}, \eta_2, \eta_3\}$ is a new U-arrangement.

2. If $\bar{\eta} > \eta_2$:
   - If $E(\bar{\eta}) < E(\eta_2)$, then $\{\eta_2, \bar{\eta}, \eta_3\}$ is a new U-arrangement.
   - If $E(\bar{\eta}) > E(\eta_2)$, then $\{\eta_1, \eta_2, \bar{\eta}\}$ is a new U-arrangement.

(d) If $\bar{\eta} = \eta_2$, by continuity we are always able to find another η_2' close to η_2 such that

$$E(\eta_2') < \min\{E(\eta_1), E(\eta_3)\}.$$

In this case, we can use this new η_2' in place of η_2 and proceed with the algorithm.

Problem 7.9

(a) Since w is uniformly sampled in the unit cube, we may write that

$$\begin{aligned}
\mathbb{P}[E(w) \leq E(w^*) + \epsilon] &= \mathbb{P}\left[\frac{1}{2}(w - w^*)^T H (w - w^*) \leq \epsilon\right] \\
&= \int_{(w-w^*)^T H (w-w^*) \leq 2\epsilon} dw_1 \cdots dw_d \\
&= \int_{x^T H x \leq 2\epsilon} \underbrace{|\det \frac{\partial w}{\partial x}|}_{=1} dx_1 \cdots dx_d
\end{aligned}$$

where we have made the change of variables $x = w - w^*$. As H is positive definite and symmetric, we know that there exists an orthogonal matrix A such that $H = A\text{diag}(\lambda_1^2, \cdots, \lambda_d^2)A^T$. Thus, if we use $y = A^T x$ as a change of variables, we now get that

$$\begin{aligned}
\mathbb{P}[E(w) \leq E(w^*) + \epsilon] &= \int_{x^T H x \leq 2\epsilon} dx_1 \cdots dx_d \\
&= \int_{y^T \text{diag}(\lambda_1^2, \cdots, \lambda_d^2) y \leq 2\epsilon} \underbrace{|\det \frac{\partial x}{\partial y}|}_{=|A|=1} dy_1 \cdots dy_d.
\end{aligned}$$

We now use a third change of variables $z = \text{diag}(\lambda_1, \cdots, \lambda_d)y$, in this case we obtain

$$\begin{aligned}
\mathbb{P}[E(w) \leq E(w^*) + \epsilon] &= \int_{y^T \text{diag}(\lambda_1^2, \cdots, \lambda_d^2) y \leq 2\epsilon} dy_1 \cdots dy_d \\
&= \int_{z^T z \leq 2\epsilon} \underbrace{|\det \frac{\partial y}{\partial z}|}_{=\frac{1}{|\lambda_1 \cdots \lambda_d|} = \frac{1}{\sqrt{\det H}}} dz_1 \cdots dz_d \\
&= \frac{1}{\sqrt{\det H}} \int_{z^T z \leq 2\epsilon} dz_1 \cdots dz_d = \frac{S_d(2\epsilon)}{\sqrt{\det H}}.
\end{aligned}$$

(b) It is clear that

$$\begin{aligned}
\mathbb{P}[E(w_{min}) > E(w^*) + \epsilon] &= \mathbb{P}[(E(w_1) > E(w^*) + \epsilon) \cap \cdots \cap (E(w_N) > E(w^*) + \epsilon)] \\
&= \prod_{i=1}^{N} \mathbb{P}[E(w_1) > E(w^*) + \epsilon] \\
&= (1 - \mathbb{P}[E(w_1) \leq E(w^*) + \epsilon])^N \\
&= \left(1 - \frac{S_d(2\epsilon)}{\sqrt{\det H}}\right)^N.
\end{aligned}$$

We may write that

$$S_d(2\epsilon) = \frac{\pi^{d/2}(2\epsilon^d)}{\Gamma(d/2+1)} \approx \frac{1}{\sqrt{\pi d}}\left(\frac{8e\pi}{d}\right)^{d/2}\epsilon^d,$$

moreover, we also have that

$$\bar{\lambda}^d = \det H.$$

Consequently, we may write that

$$\begin{aligned}
\mathbb{P}[E(w_{min}) > E(w^*) + \epsilon] &= \left(1 - \frac{S_d(2\epsilon)}{\sqrt{\det H}}\right)^N \\
&\approx \left(1 - \frac{1}{\sqrt{\pi d}}\underbrace{\left(\frac{8e\pi}{\bar{\lambda}}\right)^{d/2}}_{\approx \mu^d}\left(\frac{\epsilon^d}{d^{d/2}}\right)\right)^N \\
&\approx \left(1 - \frac{1}{\sqrt{\pi d}}\left(\frac{\mu\epsilon}{\sqrt{d}}\right)^d\right)^N.
\end{aligned}$$

(c) From point (b), we know that

$$\begin{aligned}
\mathbb{P}[E(w_{min}) > E(w^*) + \epsilon] &\approx \left(1 - \frac{1}{\sqrt{\pi d}}\left(\frac{\mu\epsilon}{\sqrt{d}}\right)^d\right)^N \\
&\approx e^{N\ln(1 - \frac{1}{\sqrt{\pi d}}(\frac{\mu\epsilon}{\sqrt{d}})^d)} \\
&\approx e^{-N\frac{1}{\sqrt{\pi d}}(\frac{\mu\epsilon}{\sqrt{d}})^d} \\
&\approx e^{\frac{1}{\sqrt{\pi d}}\log\eta}
\end{aligned}$$

because we have

$$-N\frac{1}{\sqrt{\pi d}}\left(\frac{\mu\epsilon}{\sqrt{d}}\right)^d \approx \frac{1}{\sqrt{\pi d}}\log\eta.$$

In conclusion, we get that

$$\mathbb{P}[E(w_{min}) > E(w^*) + \epsilon] \approx \eta^{\frac{1}{\sqrt{\pi d}}} \geq \eta$$

since $0 \leq \eta \leq 1$; thus we may now write that

$$\mathbb{P}[E(w_{min}) \leq E(w^*) + \epsilon] \leq 1 - \eta.$$