Supplemental Material

Lemma 7. *For the large margin classifier $\hat{\theta}_S$, we have*

$$\mathbf{P}\left[R(\hat{\theta}_S) > \epsilon\right] = \begin{cases} (1-\epsilon)^n + (\epsilon)^n & 0 < \epsilon \le \dfrac{1}{2} \\ (\dfrac{1}{2})^{n-1} & \dfrac{1}{2} < \epsilon < 1 \\ 0 & \epsilon = 1. \end{cases} \tag{29}$$

Proof. The risk is $R(\hat{\theta}_S) = |\hat{\theta}_S|$. Define event $E : \{\exists(x,-1) \in S \wedge \exists(x,+1) \in S\}$. $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon\right]$ can be decomposed into two components depending on if E happens as (40) shows.

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon\right] = \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] + \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E^c\right]. \tag{40}$$

$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E^c\right] = \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E^c\right]\mathbf{P}\left[E^c\right]$. Note that $\mathbf{P}\left[E^c\right] = (\frac{1}{2})^{n-1}$, $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E^c\right]$ is 0 if $\epsilon < 1$ and 1 if $\epsilon = 1$ because $\hat{\theta}_S = \pm 1$ always holds given E^c happens. Thus,

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E^c\right] = \begin{cases} 0 & \text{if } \epsilon < 1 \\ (\dfrac{1}{2})^{n-1} & \text{if } \epsilon = 1. \end{cases} \tag{41}$$

Now we compute $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right]$. Let n_+ be the number of positive points in S. Define $E_i : \{n_+ = i\}$. Note $E_i \cap E_j = \emptyset$ if $i \ne j$ and $E = \cup_{i=1}^{n-1} E_i$, thus

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] = \sum_{i=1}^{n-1}\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E_i\right] = \sum_{i=1}^{n-1}\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E_i\right]\mathbf{P}\left[E_i\right]. \tag{42}$$

$\mathbf{P}\left[E_i\right] = C_i^n(\frac{1}{2})^n$. Note that $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E_i\right] = \mathbf{P}\left[|\frac{x_- + x_+}{2}| \le \epsilon \mid E_i\right]$. To compute it, we first compute $F_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i) = \mathbf{P}\left[-x_- \le \epsilon_1, x_+ \le \epsilon_2 \mid E_i\right]$. Given E_i happens, $\mathbf{P}\left[-x_- \le \epsilon_1 \mid E_i\right] = 1 - (1-\epsilon_1)^{n-i}$ and $\mathbf{P}\left[x_+ \le \epsilon_2 \mid E_i\right] = 1 - (1-\epsilon_2)^i$. Also since $-x_- \le \epsilon_1$ and $x_+ \le \epsilon_2$ are independent given E_i happens, thus

$$F_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i) = \mathbf{P}\left[-x_- \le \epsilon_1, x_+ \le \epsilon_2 \mid E_i\right] = [1-(1-\epsilon_1)^{n-i}][1-(1-\epsilon_2)^i]. \tag{43}$$

Take the derivative of F gives

$$f_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i) = i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}. \tag{44}$$

Note that $|\hat{\theta}_S| \le \epsilon \Leftrightarrow |-x_1 - x_2| \le 2\epsilon$. Therefore, we integrate $f_{-x_-,x_+}(\epsilon_1, \epsilon_2 \mid E_i)$ over the region $|\epsilon_1 - \epsilon_2| \le 2\epsilon$ to obtain $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon \mid E_i\right]$. However, note that $0 \le \epsilon_1, \epsilon_2 \le 1$, thus for $\epsilon > \frac{1}{2}$, the region $|\epsilon_1 - \epsilon_2| \le 2\epsilon$ becomes the whole $[0,1] \times [0,1]$ and the integration is 1. Then (42) becomes $\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] = \sum_{i=1}^{n-1}\mathbf{P}\left[E_i\right] = \mathbf{P}\left[E\right] = 1 - (\frac{1}{2})^{n-1}$. For $\epsilon \le \frac{1}{2}$, by (42) we have

$$\begin{aligned} \mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] &= \sum_{i=1}^{n-1}\mathbf{P}\left[E_i\right]\int_{|\epsilon_1-\epsilon_2| \le 2\epsilon} i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}d\epsilon_2 d\epsilon_1 \\ &= \sum_{i=1}^{n-1} C_i^n(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2| \le 2\epsilon} i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}d\epsilon_2 d\epsilon_1 \\ &= (\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2| \le 2\epsilon} \sum_{i=1}^{n-1} C_i^n i(n-i)(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1}d\epsilon_2 d\epsilon_1. \end{aligned} \tag{45}$$

Note that $C_i^n i(n-i) = n(n-1)C_{i-1}^{n-2}$, (45) becomes

$$\begin{aligned}
\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] &= n(n-1)(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} \sum_{i=1}^{n-1} C_{i-1}^{n-2}(1-\epsilon_1)^{n-i-1}(1-\epsilon_2)^{i-1} d\epsilon_2 d\epsilon_1 \\
&= n(n-1)(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} \sum_{i=0}^{n-2} C_i^{n-2}(1-\epsilon_1)^{n-2-i}(1-\epsilon_2)^{i} d\epsilon_2 d\epsilon_1 \\
&= n(n-1)(\frac{1}{2})^n \int_{|\epsilon_1-\epsilon_2|\le 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 \\
&= n(n-1)(\frac{1}{2})^n [\int_{[0,1]\times[0,1]} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 - \int_{|\epsilon_1-\epsilon_2|> 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1].
\end{aligned} \tag{46}$$

Now we compute the two integration in (46)

$$\begin{aligned}
&\int_{[0,1]\times[0,1]} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = \int_0^1 [-\frac{1}{n-1}(2-\epsilon_1-\epsilon_2)^{n-1}|_0^1] d\epsilon_1 \\
&= \int_0^1 \frac{1}{n-1}[(2-\epsilon_1)^{n-1} - (1-\epsilon_1)^{n-1}] d\epsilon_1 = [-\frac{1}{n(n-1)}(2-\epsilon)^n + \frac{1}{n(n-1)}(1-\epsilon_1)^n] \mid_0^1 = \frac{2^n-2}{n(n-1)}.
\end{aligned} \tag{47}$$

For the second integration, note that it can decomposed as

$$\int_{|\epsilon_1-\epsilon_2|> 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = \int_0^{1-2\epsilon} \int_{\epsilon_1+2\epsilon}^1 (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 + \int_{2\epsilon}^1 \int_0^{\epsilon_1-2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1. \tag{48}$$

Since the two sub-integration's are identical because the two sub regions are symmetric. We only show the computation for the first.

$$\begin{aligned}
&\int_0^{1-2\epsilon} \int_{\epsilon_1+2\epsilon}^1 (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = \int_0^{1-2\epsilon} [-\frac{1}{n-1}(2-\epsilon_1-\epsilon_2)^{n-1}|_{\epsilon_1+2\epsilon}^1] d\epsilon_1 \\
&= \int_0^{1-2\epsilon} [-\frac{1}{n-1}(1-\epsilon_1)^{n-1} + \frac{2^{n-1}}{n-1}(1-\epsilon_1-\epsilon)^{n-1}] d\epsilon_1 \\
&= [\frac{1}{n(n-1)}(1-\epsilon_1)^n - \frac{2^{n-1}}{n(n-1)}(1-\epsilon_1-\epsilon)^n] \mid_0^{1-2\epsilon} \\
&= \frac{2^{n-1}}{n(n-1)}[\epsilon^n + (1-\epsilon)^n] - \frac{1}{n(n-1)}.
\end{aligned} \tag{49}$$

Thus we have

$$\begin{aligned}
&\int_{|\epsilon_1-\epsilon_2|> 2\epsilon} (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 = 2\int_0^{1-2\epsilon} \int_{\epsilon_1+2\epsilon}^1 (2-\epsilon_1-\epsilon_2)^{n-2} d\epsilon_2 d\epsilon_1 \\
&= \frac{2^n}{n(n-1)}[\epsilon^n + (1-\epsilon)^n] - \frac{2}{n(n-1)}.
\end{aligned} \tag{50}$$

Combine (47) and (50), we can compute (46) as follows.

$$\begin{aligned}
\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] &= = n(n-1)(\frac{1}{2})^n [\frac{2^n-2}{n(n-1)} - \frac{2^n}{n(n-1)}(\epsilon^n + (1-\epsilon)^n) + \frac{2}{n(n-1)}] \\
&= \frac{2^n-2}{2^n} - \epsilon^n - (1-\epsilon)^n + (\frac{1}{2})^{n-1} \\
&= 1 - \epsilon^n - (1-\epsilon)^n
\end{aligned} \tag{51}$$

Therefore we have

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon, E\right] = \begin{cases} 1-\epsilon^n-(1-\epsilon)^n & \text{if } \epsilon \le \frac{1}{2} \\ 1-(\dfrac{1}{2})^{n-1} & \text{if } \frac{1}{2} < \epsilon \le 1. \end{cases} \tag{52}$$

Now combine (41) and (52) we have

$$\mathbf{P}\left[|\hat{\theta}_S| \le \epsilon\right] = \begin{cases} 1-\epsilon^n-(1-\epsilon)^n & \text{if } \epsilon \le \frac{1}{2} \\ 1-(\dfrac{1}{2})^{n-1} & \text{if } \frac{1}{2} < \epsilon < 1 \\ 1 & \text{if } \epsilon = 1. \end{cases} \tag{53}$$

which is equivalent to (29). □

Lemma 9. *Let $n = 4m$, where m is an integer. Let S be an n-item iid sample drawn from $p_{\mathbb{Z}}$. $\forall \epsilon > 0, \forall \delta \in (0,1)$, $\exists \mathbb{M}(\epsilon,\delta) = \max\{\frac{3e}{\ln 4-1}\ln\frac{3}{\delta}, (\frac{1}{\epsilon}\ln\frac{3}{\delta})^{\frac{1}{2}}\}$ such that $\forall m \ge \mathbb{M}(\epsilon,\delta)$, $\mathbf{P}\left[R(\hat{\theta}_{B_{ms}(S)}) \le \epsilon\right] > 1-\delta$.*

Proof. Let $S_1 = \{x \mid (x,1) \in S\}$ and $S_2 = \{x \mid (x,-1) \in S\}$ respectively. Then we have $|S_1| + |S_2| = 4m$. Define event $E_1 : \{|S_1| \ge m \wedge |S_2| \ge m\}$. Then we have

$$\mathbf{P}\left[E_1\right] = 1 - 2\sum_{i=0}^{m-1} C_i^{4m}(\frac{1}{2})^{4m}. \tag{54}$$

where we rule out all possible sequences of $4m$ points which lead to $|S_1| < m$ or $|S_2| < m$. By standard result [37] (Lemma A.5) $\sum_{k=0}^{d} C_k^m \le (\frac{em}{d})^d$, we have

$$\mathbf{P}\left[E_1\right] \ge 1 - 2(\frac{4em}{m-1})^{m-1}(\frac{1}{2})^{4m} = 1 - \frac{1}{2}\frac{e^{m-1}}{4^m}(\frac{m}{m-1})^{m-1} \ge 1 - \frac{1}{2}(\frac{e}{4})^m \ge 1 - (\frac{e}{4})^m \tag{55}$$

where the 2nd-to-last inequality follows from the fact that $e \ge (1+\frac{1}{m-1})^{m-1}$. Note that by definition $m \ge \frac{3e}{\ln 4-1}\ln\frac{3}{\delta} > \frac{1}{\ln 4-1}\ln\frac{3}{\delta}$, thus $(\frac{e}{4})^m < \frac{\delta}{3}$ and $\mathbf{P}\left[E_1\right] > 1-\frac{\delta}{3}$. Since $|S_1|+|S_2| = 4m$, then either $|S_1| \ge 2m$ or $|S_2| \ge 2m$. Without loss of generality we assume $|S_1| \ge 2m$. We then divide the interval [0, 1] equally into $N = \lfloor m^2(\ln\frac{3}{\delta})^{-1}\rfloor$ segments. The length of each segment is $\frac{1}{N} = O(\frac{1}{m^2})$ as Figure 4 shows. Note that $m \ge \frac{3e}{\ln 4-1}\ln\frac{3}{\delta} > 3e\ln\frac{3}{\delta}$, thus $N \ge \lfloor 3em\rfloor > 2em > m$.

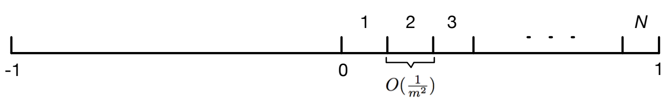

Figure 4: segments

Let N_o be the number of segments that are occupied by the points in S_1. Note that N_o is a random variable. Let E_2 be the event that $N_o \ge m$. Now we lower bound $\mathbf{P}\left[E_2\right]$. This is a variant of the coupon collector's problem: there are N distinct coupons, and in $|S_1|$ trials we want to collect at least m distinct coupons. Note that $\mathbf{P}\left[E_2\right] = 1 - \mathbf{P}\left[E_2^c\right] = 1 - \sum_{i=1}^{m-1}\mathbf{P}\left[N_o = i\right]$. Let T_i be the number of all possible coupon sequences of S_1 such that S_1 occupies exactly i segments (i.e. distinct coupons). We have C_i^n ways of choosing i segments among a total of N. Also, for each choice of i segments, the number of all possible coupon sequences of S_1 such that S_1 fully occupies those i segments without empty is upper bounded by $i^{|S_1|}$. Thus $T_i \le C_i^n i^{|S_1|}$ and we have

$$\mathbf{P}\left[N_o = i\right] = \frac{T_i}{N^{|S_1|}} \le C_i^n(\frac{i}{N})^{|S_1|}. \tag{56}$$

Since $m \ge \frac{3e}{\ln 4-1}\ln\frac{3}{\delta} > \log_2\frac{3}{\delta}$, $|S_1| \ge 2m$, and $N > 2em$, thus

$$\begin{aligned}\mathbf{P}\left[E_2^c\right] &= \sum_{i=1}^{m-1}\mathbf{P}\left[N_o = i\right] \le \sum_{i=1}^{m-1} C_i^n(\frac{i}{N})^{|S_1|} \le \sum_{i=1}^{m-1} C_i^n(\frac{m}{N})^{2m} \\ &< \sum_{i=0}^{m} C_i^n(\frac{m}{N})^{2m} \le (\frac{eN}{m})^m(\frac{m}{N})^{2m} = (\frac{em}{N})^m < (\frac{em}{2em})^m = (\frac{1}{2})^m < \frac{\delta}{3}.\end{aligned} \tag{57}$$

Thus $\mathbf{P}\left[E_2\right] \geq 1 - \frac{\delta}{3}$. Applying union bound, $\mathbf{P}\left[E_1, E_2\right] \geq 1 - \frac{2\delta}{3}$.

Let E_3 be the following event: there exist a point x_2 in S_2 such that $-x_2$, the flipped point, lies in the same segment as some point x_1 in S_1. If E_3 happens, then $|x_1 + x_2| = |x_1 - (-x_2)| \leq \frac{1}{N}$. Note that $\mathbf{P}\left[E_3\right] \geq \mathbf{P}\left[E_1, E_2, E_3\right] = \mathbf{P}\left[E_3 \mid E_1, E_2\right]\mathbf{P}\left[E_1, E_2\right]$. Now we lower bound $\mathbf{P}\left[E_3 \mid E_1, E_2\right]$. Given E_1 and E_2 happen, we have $|S_2| \geq m$ and $N_o \geq m$. Since $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor \leq m^2 (\ln \frac{3}{\delta})^{-1}$, we have

$$\mathbf{P}\left[E_3^c \mid E_1, E_2\right] = (1 - \frac{N_o}{N})^{|S_2|} \leq (1 - \frac{m}{N})^{|S_2|} \leq (1 - \frac{m}{N})^m \leq e^{-\frac{m^2}{N}} \leq \frac{\delta}{3}. \tag{58}$$

Thus, $\mathbf{P}\left[E_3 \mid E_1, E_2\right] = 1 - \mathbf{P}\left[E_3^c \mid E_1, E_2\right] > 1 - \frac{\delta}{3}$. $\mathbf{P}\left[E_3\right] \geq \mathbf{P}\left[E_1, E_2, E_3\right] = \mathbf{P}\left[E_3 \mid E_1, E_2\right]\mathbf{P}\left[E_1, E_2\right] \geq (1 - \frac{\delta}{3})(1 - \frac{2\delta}{3}) > 1 - \delta$. Thus with probability at least $1 - \delta$, there exist $x_2 \in S_2$ and $x_1 \in S_1$ such that $|x_1 + x_2| \leq \frac{1}{N}$.

We now bound $\frac{1}{N}$. $N = \lfloor m^2 (\ln \frac{3}{\delta})^{-1} \rfloor \geq \frac{1}{2} m^2 (\ln \frac{3}{\delta})^{-1}$. Therefore $\frac{1}{N} \leq \frac{2}{m^2} \ln \frac{3}{\delta}$. Recall by definition $m \geq (\frac{1}{\epsilon} \ln \frac{3}{\delta})^{\frac{1}{2}}$, thus $\frac{1}{N} \leq 2\epsilon$.

We now have $|x_1 + x_2| \leq 2\epsilon$. Finally, since $\{s_-, s_+\}$ selected by teacher B_{ms} is the most symmetric pair, it must satisfy $|s_- + s_+| \leq |x_1 + x_2| \leq 2\epsilon$. Putting together, with probability at least $1 - \delta$, $R(\hat{\theta}_{B_{ms}(S)}) = \frac{1}{2}|s_- + s_+| \leq \epsilon$. $\square$