CSC 665: Homework 1

Chicheng Zhang

October 12, 2019

Please complete the following set of exercises **on your own**. The homework is due **on Oct 1, 12:30pm, on Gradescope**. You are free to cite existing theorems from the textbook and course notes.

Problem 1

For a random variable Z with mean $\mathbb{E}Z = 0$, we call Z is v-subgaussian, if

$$\psi_Z(t) = \ln \mathbb{E}e^{tZ} \le \frac{vt^2}{2}.$$

Show the following:

1. If Z has Gaussian distribution $\mathrm{N}(0, \sigma^2)$, then Z is σ^2-subgausssian.
2. If Z take values within interval $[a, b]$, then Z is $\frac{(b-a)^2}{4}$-subgaussian.
3. If $Z_1, \ldots, Z_n$ are independent, and each Z_i is v_i subgaussian, then $\sum_{i=1}^n Z_i$ is $\sum_{i=1}^n v_i$-subgaussian.
4. If Z is v-subgaussian, then
$$\mathbb{P}(|Z| \ge t) \le 2\exp\left\{-\frac{t^2}{2v}\right\}.$$

Solution

1. We first show that when Z is standard Gaussian $\mathrm{N}(0, 1)$, Z is 1-subgaussian. To see this, note that
$$\mathbb{E}e^{tZ} = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}+tx} dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx \cdot e^{\frac{t^2}{2}} = e^{\frac{t^2}{2}},$$
where the last inequality uses the fact that $\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-t)^2}{2}\right\}$ is the probability density function of distribution $\mathrm{N}(t, 1)$. Now suppose Z has distribution $\mathrm{N}(0, \sigma^2)$, then Z can be written as σZ_0, where Z_0 has distribution $\mathrm{N}(0, 1)$. This implies that
$$\mathbb{E}e^{tZ} = \mathbb{E}e^{(t\sigma)Z_0} = \exp\left\{\frac{\sigma^2 t^2}{2}\right\},$$
implying that Z is σ^2-subgaussian.
2. This follows directly from Lemma 2 of "concentration of measure (1)" note and the definition of subgaussianity.

3. By the method of moment generating function and the independence of Z_i's, we have:

$$\mathbb{E}e^{tZ} = \mathbb{E}e^{t\sum_{i=1}^n Z_i} = \mathbb{E}\prod_{i=1}^n e^{tZ_i} = \prod_{i=1}^n \mathbb{E}e^{tZ_i} \le \prod_{i=1}^n e^{\frac{v_i t^2}{2}} = e^{\frac{vt^2}{2}}.$$

4. We first show that $\mathbb{P}(Z \ge t) \le 2\exp\left\{-\frac{t^2}{2v}\right\}$. This is because for any $s > 0$,

$$\mathbb{P}(Z \ge t) = \mathbb{P}(e^{sZ} \ge e^{st}) \le e^{-st}\mathbb{E}e^{sZ} \le e^{-st+\frac{v^2 s^2}{2}}.$$

As the choice of s is arbitrary, taking $s = \frac{t}{v^2}$, we have that

$$\mathbb{P}(Z \ge t) \le e^{-\frac{t^2}{2v^2}}.$$

Symmetrically, $\mathbb{P}(Z \le t) \le e^{-\frac{t^2}{2v^2}}$. Therefore,

$$\mathbb{P}(|Z| \ge t) \le \mathbb{P}(Z \ge t) + \mathbb{P}(Z \le t) \le 2\exp\left\{-\frac{t^2}{2v}\right\}.$$

Problem 2

In this exercise we give an alternative proof of the Chernoff bound for Bernoulli random variables: suppose $X_1, \ldots, X_n$ are iid and from Bernoulli(p), define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then,

$$\mathbb{P}(\bar{X} \ge q) \le \exp\left\{-n\,\mathrm{kl}(q,p)\right\}, q \ge p, \tag{1}$$

$$\mathbb{P}(\bar{X} \le q) \le \exp\left\{-n\,\mathrm{kl}(q,p)\right\}, q \le p. \tag{2}$$

1. Show that

$$\mathbb{P}(\bar{X} \ge q) = \sum_{m: m \ge nq} \binom{n}{m} p^m (1-p)^{n-m}.$$

2. Use the elementary inequality that $\binom{n}{m} q^m (1-q)^{n-m} \le 1$, show that for $m \ge nq$,

$$\binom{n}{m} p^m (1-p)^{n-m} \le \left(\frac{p}{q}\right)^{nq} \left(\frac{1-p}{1-q}\right)^{n(1-q)}.$$

3. Use the above two items to conclude that $\mathbb{P}(\bar{X} \ge q) \le (n+1)\exp\left\{-n\,\mathrm{kl}(q,p)\right\}$.

4. Note that compared to Equation 1, the above bound is has an additional factor of n on the right hand side. Use the elementary inequality $\sum_{m \ge nq} \binom{n}{m} q^m (1-q)^{n-m} \le 1$ as a starting point, along with insights you gained from items 1 and 2 to show Equation (1).

5. Repeat the proof for the lower tail bound (Equation (2)).

Solution

1. This simply follows from the observation that

$$\mathbb{P}(\bar{X} \geq q) = \sum_{m \geq nq} \mathbb{P}(\bar{X} = m)$$

and that $\bar{X}$ has distribution $\mathrm{B}(n, p)$.

2. Let $F(r, m) = \binom{n}{m} r^m (1-r)^{n-m}$. Observe that

$$\frac{F(p,m)}{F(q,m)} = \frac{\binom{n}{m} p^m (1-p)^{n-m}}{\binom{n}{m} q^m (1-q)^{n-m}} = (\frac{p}{q})^m \cdot (\frac{1-p}{1-q})^{n-m}.$$

Now, for $m \geq nq$, as $p \leq q$, we have

$$(\frac{p}{q})^m \leq (\frac{p}{q})^{nq}.$$

Similary, as $1 - p \geq 1 - q$ and $n - m \leq n(1-q)$, we have

$$(\frac{1-p}{1-q})^{n-m} \leq (\frac{1-p}{1-q})^{n(1-q)}.$$

This implies that

$$\frac{F(p,m)}{F(q,m)} \leq (\frac{p}{q})^{nq} (\frac{1-p}{1-q})^{n(1-q)}.$$

Item 2 follows as $F(q, m) \leq 1$.

3. Note that

$$\mathbb{P}(\bar{X} \geq q) = \sum_{m: m \geq nq} F(p, m) \leq \sum_{m: m \geq nq} (\frac{p}{q})^{nq} (\frac{1-p}{1-q})^{n(1-q)}.$$

Now, as each summand in the right hand side are the same, and there are at most $(n+1)$ terms, we get that

$$\mathbb{P}(\bar{X} \geq q) \leq (n+1)(\frac{p}{q})^{nq} (\frac{1-p}{1-q})^{n(1-q)} = (n+1) e^{-n \,\mathrm{kl}(q,p)}.$$

4. From item 2 we know that for all $m \geq nq$,

$$\frac{F(p,m)}{F(q,m)} \leq e^{-n \,\mathrm{kl}(q,p)}.$$

Now,

$$\mathbb{P}(\bar{X} \geq q) = \sum_{m: m \geq qn} F(p, m) \leq (\sum_{m: m \geq qn} F(q, m)) \cdot e^{-n \,\mathrm{kl}(q,p)} \leq e^{-n \,\mathrm{kl}(q,p)}.$$

where the last inequality uses the fact that $\sum_{m: m \geq qn} F(q, m) \leq 1$.

5. For $q \leq p$, by the exact same reasoning, we can show that for $m \leq nq$,

$$\frac{F(p,m)}{F(q,m)} \leq (\frac{p}{q})^{nq} (\frac{1-p}{1-q})^{n(1-q)} = e^{-n \,\mathrm{kl}(q,p)}.$$

Now,

$$\mathbb{P}(\bar{X} \leq q) = \sum_{m: m \leq qn} F(p, m) \leq (\sum_{m: m \leq qn} F(q, m)) \cdot e^{-n \,\mathrm{kl}(q,p)} \leq e^{-n \,\mathrm{kl}(q,p)}.$$

where the last inequality uses the fact that $\sum_{m: m \leq qn} F(q, m) \leq 1$.

Problem 3

In this exercise we will use basic concentration inequalities to show that, we can find exponentially many points on the unit sphere in $\mathbb{R}^d$ that are far away from each other. Specifically, consider n random vectors $X_1,\ X_2,\ \ldots,\ X_n$ in $\mathbb{R}^d$, where for each i, $X_i = \frac{1}{\sqrt{d}}(Z_{i,1}, \ldots, Z_{i,d})$. Here $\big\{Z_{i,j}\big\}_{i\in\{1,\ldots,n\},j\in\{1,\ldots,d\}}$'s are all independent and identically distributed, and $Z_{i,j}$ takes value 1 with probability 1/2, and takes value -1 with probabilty 1/2.

1. Check that all X_i's has unit length, i.e. $\|X_i\|_2 = 1$.
2. Use Hoeffding's Inequality to show that for any fixed pair $i, j \in \{1, \ldots, n\}$, $i \neq j$,

$$\mathbb{P}(|\left\langle X_i, X_j\right\rangle| \geq \frac{1}{2}) \leq 2\exp\bigg\{-\frac{d}{8}\bigg\}.$$

3. Suppose $n = \exp\Big\{\frac{d}{32}\Big\}$. Show that with nonzero probability, for all pairs $i, j \in \{1, \ldots, n\}$, $i \neq j$, the angle between X_i and X_j is in $[\frac{\pi}{3}, \frac{2\pi}{3}]$.

Solution

1.

$$\|X_i\|^2 = \sum_{l=1}^{d} (\frac{Z_{i,l}}{\sqrt{d}})^2 = \sum_{l=1}^{d} \frac{1}{d} = 1,$$

where the penultimate equality uses the fact that $Z_{i,j}^2 = 1$ with probability 1.

2. Note that

$$\left\langle X_i, X_j\right\rangle = \sum_{i=1}^{d} \frac{Z_{i,l}Z_{j,l}}{d}.$$

Now, consider $Y_l = Z_{i,l}Z_{j,l}$, are all $Z_{i,l}$'s are independent, Y_l's are also independent. In addition, Y_l is from the Rademacher distribution (take value ±1 with equal probability). Taking $a = -1$, $b = +1$, $\mu = 0$ and $n = d$ in Hoeffding's inequality (Theorem 1 in "concentration of measure (1)" note), we get

$$\mathbb{P}(|\left\langle X_i, X_j\right\rangle| \geq \frac{1}{2}) = \mathbb{P}(|\bar{Y} - \mu| \geq \frac{1}{2}) \leq 2\exp\Bigg\{-\frac{2d(\frac{1}{2})^2}{(1+1)^2}\Bigg\} = 2\exp\bigg\{-\frac{d}{8}\bigg\}.$$

3. Consider all pairs of (i, j) such that $1 \leq i < j \leq n$. There are $\binom{n}{2} = \frac{n(n-1)}{2}$ such pairs. By a union bound, we have

$$\mathbb{P}(\exists i \neq j, |\left\langle X_i, X_j\right\rangle| \geq \frac{1}{2}) = \sum_{1\leq i<j\leq n} \mathbb{P}(|\left\langle X_i, X_j\right\rangle| \geq \frac{1}{2}) \leq \frac{n(n-1)}{2} \cdot 2\exp\bigg\{-\frac{d}{8}\bigg\} < 1.$$

where the last inequality is by the choice of $n = e^{\frac{d}{16}}$. This implies that the complement event, that is, for all $i \neq j$, $|\left\langle X_i, X_j\right\rangle| < \frac{1}{2}$, happens with nonzero probability. Observe that as all X_i's are unit vectors, $|\left\langle X_i, X_j\right\rangle| < \frac{1}{2}$ is equivalent to the angle between X_i and X_j is in $(\frac{\pi}{3}, \frac{2\pi}{3})$. This proves item 3.

Problem 4

Suppose D is a distribution over $[0,1] \times \{-1,+1\}$ such that D_X, the marginal of D over $\mathcal{X} = [0,1]$, is uniform. In addition,

$$P(Y=+1|x) = \begin{cases} 0 & x \le \frac{1}{2}, \\ 1 & x > \frac{1}{2} \end{cases},$$

i.e. the distribution is separable by a threshold classifier with threshold $\frac{1}{2}$. Suppose training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn iid from D. Now consider the following classifier $\hat{h}$:

$$\hat{h}(x) = \begin{cases} Y_i & x = X_i \text{ for some } i \in \{1, \ldots, n\}, \\ -1 & \text{otherwise.} \end{cases}$$

(For simplicity, assume that all X_i's are distinct, which also happens with probability 1.)

1. Calculate $\operatorname{err}(\hat{h}, S)$.
2. Calculate $\operatorname{err}(\hat{h}, D)$. What is the value of $\operatorname{err}(\hat{h}, S) - \operatorname{err}(\hat{h}, D)$?
3. It may be tempting to use following argument to argue the concentration of $\operatorname{err}(\hat{h}, S)$ to $\operatorname{err}(\hat{h}, D)$. Define random variables $Z_i = \mathbf{1}(\hat{h}(X_i) \ne Y_i)$ for all i in $\{1, \ldots, n\}$, therefore, Hoeffding's inequality, with probability $1-\delta$,

   $$|\operatorname{err}(\hat{h}, S) - \operatorname{err}(\hat{h}, D)| \le \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

   Does this contradict the results we got from item 2? Why?

Solution

1. Note that for all training examples (X_i, Y_i), by the definition of $\hat{h}$, $\hat{h}(X_i) = Y_i$.

   Therefore, by the definition of empirical error,

   $$\operatorname{err}(\hat{h}, S) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{h}(X_i) \ne Y_i) = 0.$$

2. Suppose that we are given a training set $(x_i, y_i)_{i=1}^n$. Now, when a new example (X, Y) is drawn, define event $E = \big\{X \in \{x_1, \ldots, x_n\}\big\}$. By union bound,

   $$\mathbb{P}(E) = \mathbb{P}(X \in \{x_1, \ldots, x_n\}) \le \sum_{i=1}^{n} \mathbb{P}(X = x_i) = \sum_{i=1}^{n} 0 = 0,$$

   where the penultimate equality uses the fact that D_X is uniform over $[0,1]$. Therefore,

   $$\begin{aligned} \operatorname{err}(\hat{h}) &= \mathbb{P}_{(X,Y)\sim D}(\hat{h}(X) \ne Y) \\ &= \mathbb{P}_{(X,Y)\sim D}(\Big\{\hat{h}(X) \ne Y\Big\} \cap E) + \mathbb{P}_{(X,Y)\sim D}(\Big\{\hat{h}(X) \ne Y\Big\} \cap \bar{E}) \\ &= \mathbb{P}_{(X,Y)\sim D}(\Big\{\hat{h}(X) \ne Y\Big\} \cap E) + \mathbb{P}_{(X,Y)\sim D}(\{Y = +1\} \cap \bar{E}) \\ &= \mathbb{P}_{(X,Y)\sim D}(\Big\{\hat{h}(X) \ne Y\Big\} \cap E) + \mathbb{P}_{(X,Y)\sim D}(Y = +1) - \mathbb{P}_{(X,Y)\sim D}(\{Y = +1\} \cap E) \end{aligned}$$

where in the first and third equality we use the additivity of probability: for events A and E, $\mathbb{P}(A) = \mathbb{P}(A \cap E) + \mathbb{P}(A \cap \bar{E})$; the second equality uses the fact that when E happens, $\hat{h}(X) = -1$ by the definition of $\hat{h}$.

Now, observe that both $\mathbb{P}_{(X,Y)\sim D}(\left\{\hat{h}(X) \neq Y\right\} \cap E)$ and $\mathbb{P}_{(X,Y)\sim D}(\{Y = +1\} \cap E)$ are at least 0 and at most $\mathbb{P}(E) = 0$, this implies that both terms are identically 0. Therefore,

$$\mathrm{err}(\hat{h}) = \mathbb{P}_{(X,Y)\sim D}(Y = +1) = \mathbb{P}_{(X,Y)\sim D}(X > \frac{1}{2}) = \frac{1}{2}.$$

3. We cannot directly apply Hoeffding's inequality to argue about the concentration of empirical error to generalization error. Consider random variables Z_i's defined in the problem statement. although $\mathrm{err}(\hat{h}, S) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is correct, we don't have $\mathbb{E}Z_i = \mathrm{err}(\hat{h}, D)$. To see this, note Z_i's are in fact identically zero, so that $\mathbb{E}Z_i = 0$; on the other hand, as obtained in item 2, $\mathrm{err}(\hat{h}, D) = \frac{1}{2}$. Recall that to use Hoeffding's inequality to argue about error concentration, we need to choose a classifier h *before* seeing the training examples.

Problem 5

In this exercise, we will unify the analysis of $O(\frac{1}{\epsilon})$-style sample complexity for the realizable case and the $O(\frac{1}{\epsilon^2})$-style sample complexity for the agnostic case, by revisiting the empirical risk minimization algorithm. Suppose $\mathcal{H}$ is a finite hypothesis class, D is a distribution over labeled examples, and S is a training set of size m drawn iid from D. Denote by $\nu^\star = \min_{h\in\mathcal{H}} \mathrm{err}(h, D)$ as the optimal generalization error, and $\hat{h}$ the output of the empirical risk minimzation algorithm.

1. Use Chernoff bound for Bernoulli random variables, show that for a fixed classifier h, with probability $1 - \delta$,
$$\mathrm{kl}(\mathrm{err}(h, S), \mathrm{err}(h, D)) \leq \frac{\ln\frac{2}{\delta}}{m}.$$

2. Use the above reasoning to conclude that with probability $1 - \delta$, for all classifiers h in $\mathcal{H}$,
$$|\,\mathrm{err}(h, S) - \mathrm{err}(h, D)| \leq \sqrt{2\max(\mathrm{err}(h, S), \mathrm{err}(h, D))\frac{\ln\frac{2|\mathcal{H}|}{\delta}}{m}}.$$
(Hint: you can use the fact that $\mathrm{kl}(q, p) \geq \frac{(q-p)^2}{2\max(p,q)}$.)

3. Show that with probability $1 - \delta$, for all classifiers h in $\mathcal{H}$,
$$\mathrm{err}(h, S) \leq \mathrm{err}(h, D) + \sqrt{\mathrm{err}(h, D)\frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m}} + \frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m},$$
$$\mathrm{err}(h, D) \leq \mathrm{err}(h, S) + \sqrt{\mathrm{err}(h, S)\frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m}} + \frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m}.$$
(Hint: you can use the elementary fact that for $A, B, C > 0$, $A \leq B + C\sqrt{A}$ implies $A \leq B + C^2 + C\sqrt{B}$.)

4. Show that with probability $1 - \delta$, $\hat{h}$, the training error minimizer over $\mathcal{H}$, satisfies that
$$\mathrm{err}(\hat{h}, D) \leq \nu^\star + 6\sqrt{\frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m}\nu^\star} + 8\frac{\ln\frac{2|\mathcal{H}|}{\delta}}{m}.$$
(Hint: you may find the following elementary facts useful: for $A, B > 0$, $\sqrt{AB} \leq A + B$, $\sqrt{A + B} \leq \sqrt{A} + \sqrt{B}$. If you get other constants on the right hand side, no worries - you will still get full credit.)

5. Conclude that:

   (a) There exists a function m_A such that $m_A(\epsilon, \delta) = O(\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon^2})$, when $m \geq m_A(\epsilon, \delta)$, for all distributions D, $\text{err}(\hat{h}, D) \leq \nu^\star + \epsilon$ with probability $1 - \delta$.

   (b) There exists a function m_R such that $m_R(\epsilon, \delta) = O(\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon})$, when $m \geq m_R(\epsilon, \delta)$, for all distributions D such that $\nu^\star = 0$, $\text{err}(\hat{h}, D) \leq \epsilon$ with probability $1 - \delta$.

Solution

1. By taking $\epsilon = \frac{\ln \frac{2}{\delta}}{m}$, it suffices to show that for a set of m iid Bernoulli random variables X_i's, each with mean p,

$$\mathbb{P}(\text{kl}(\bar{p}, p) > \epsilon) \leq 2e^{-m\epsilon}.$$

   It can be checked by taking derivatives that $f(q) = \text{kl}(q, p)$ is monotonically decreasing in $(0, p)$, and is monotonically increasing in $(p, 1)$. This motivates the following two definitions for $\epsilon > 0$: define $\text{kl}_+^{-1}(p, \epsilon)$ as the unique value q such that $q > p$ and $\text{kl}(q, p) = \epsilon$; $\text{kl}_-^{-1}(p, \epsilon)$ as the unique value q such that $q < p$ and $\text{kl}(q, p) = \epsilon$. Therefore, we have the following equivalence on events:

$$\left\{\text{kl}(\bar{p}, p) > \epsilon\right\} = \left\{\bar{p} > \text{kl}_+^{-1}(p, \epsilon)\right\} \cup \left\{\bar{p} < \text{kl}_-^{-1}(p, \epsilon)\right\}.$$

   We note that

$$\mathbb{P}(\bar{p} > \text{kl}_+^{-1}(p, \epsilon)) \leq e^{-m\,\text{kl}(\text{kl}_+^{-1}(p,\epsilon),p)} = e^{-m\epsilon},$$

$$\mathbb{P}(\bar{p} < \text{kl}_-^{-1}(p, \epsilon)) \leq e^{-m\,\text{kl}(\text{kl}_-^{-1}(p,\epsilon),p)} = e^{-m\epsilon}.$$

   The item follows from union bound.

2. Fix a classifier h in $\mathcal{H}$. By the fact that $\text{kl}(q, p) \geq \frac{(q-p)^2}{2\max(p,q)}$, along with item 1, we have that with probability $1 - \delta/|\mathcal{H}|$,

$$\frac{(\text{err}(h, S) - \text{err}(h, D))}{2\max(\text{err}(h, S), \text{err}(h, D))} \leq \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{m}.$$

   The item simply follows from algebra (moving the max term to the right hand side and taking square roots on both sides), along with union bound over all h in $\mathcal{H}$.

3. We only prove the first inequality as the proof of the second one is identical. Consider two cases: (1) $\text{err}(h, S) \leq \text{err}(h, D)$. In this case, the inequality holds trivially. (2) $\text{err}(h, S) > \text{err}(h, D)$. In this case, $\max(\text{err}(h, S), \text{err}(h, D)) = \text{err}(h, S)$. Now by the equation in item 2, we have

$$\text{err}(h, S) \leq \text{err}(h, D) + \sqrt{\text{err}(h, S)\frac{2\ln \frac{2|\mathcal{H}|}{\delta}}{m}}.$$

   In this case, the inequality follows by the elementary fact, with $A = \text{err}(h, S)$, $B = \text{err}(h, D)$, $C = \sqrt{\frac{2\ln \frac{2|\mathcal{H}|}{\delta}}{m}}$.

4. Denote by $G = \frac{2\ln \frac{2|\mathcal{H}|}{\delta}}{m}$. First, we have:

$$\text{err}(h^\star, S) \leq \text{err}(h^\star, D) + \sqrt{\text{err}(h^\star, D)G} + G = \nu^\star + \sqrt{\nu^\star G} + G.$$

Now, by optimality of $\hat{h}$, we have $\text{err}(\hat{h}, S) \leq \text{err}(h^\star, S)$, therefore $\text{err}(\hat{h}, S)$ has the same upper bound. Now use the second inequality in item 3, we have the following:

$$\begin{aligned}
\text{err}(\hat{h}, D) &\leq \text{err}(\hat{h}, S) + \sqrt{\text{err}(\hat{h}, S)G} + G \\
&\leq \nu^\star + \sqrt{\nu^\star G} + G + \sqrt{(\nu^\star + \sqrt{\nu^\star G} + G)G} + G \\
&\leq \nu^\star + \sqrt{\nu^\star G} + 2G + \sqrt{2(\nu^\star G + G^2)} \\
&\leq \nu^\star + (1 + \sqrt{2})\sqrt{\nu^\star G} + (2 + \sqrt{2})G \\
&\leq \nu^\star + 6\sqrt{\nu^\star G} + 4G.
\end{aligned}$$

where the second inequality is by plugging in the upper bound on $\text{err}(\hat{h}, S)$; the third inequality is from the fact that $\nu^\star + \sqrt{\nu^\star G} + G \leq 2(\nu^\star + G)$; the fourth inequality uses the simple fact that $\sqrt{2(\nu^\star G + G^2)} \leq \sqrt{2\nu^\star G} + \sqrt{2}G$; the fourth inequality is from simple algebra. Item 4 follows from the definition of G.

5. For the first item, observe that by item 4 and the fact that $\nu^\star \leq 1$, we have that with probability $1 - \delta$,

$$\text{err}(\hat{h}, D) \leq \nu^\star + 6\sqrt{\frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m}} + 8\frac{\ln\frac{2|\mathcal{H}|}{\delta}}{m}.$$

Now, consider $m_1(\epsilon, \delta)$ (resp. $m_2(\epsilon, \delta)$) be the solution of m such that $6\sqrt{\frac{2\ln\frac{2|\mathcal{H}|}{\delta}}{m}} = \frac{\epsilon}{2}$ (resp. $8\frac{\ln\frac{2|\mathcal{H}|}{\delta}}{m} = \frac{\epsilon}{2}$). Note that $m_1(\epsilon, \delta) = O(\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{\epsilon^2})$ and $m_2(\epsilon, \delta) = O(\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{\epsilon^2})$. Now define $m_A(\epsilon, \delta) = m_1(\epsilon, \delta) + m_2(\epsilon, \delta)$; we have that ERM would satisfy (ϵ, δ)-PAC guarantee if $m \geq m_A(\epsilon, \delta)$.

For the second item, observe that by item 4 and the fact that $\nu^\star = 0$, we have that with probability $1 - \delta$,

$$\text{err}(\hat{h}, D) \leq 8\frac{\ln\frac{2|\mathcal{H}|}{\delta}}{m}.$$

Therefore, taking $m_R(\epsilon, \delta) = 8\frac{\ln\frac{2|\mathcal{H}|}{\delta}}{\epsilon}$, ERM would satisfy (ϵ, δ)-PAC guarantee if $m \geq m_R(\epsilon, \delta)$.