

## Lecture: Concentration

Date: September 6th, 2023

Author: Eric Wong

**Attribution.** These notes are extremely similar to the beginning lectures of Larry Wasserman's Intermediate Statistics course from CMU (<https://www.stat.cmu.edu/~larry/=stat705/>), with some slight notation tweaks to match the course.

## 1 Concentration Basics

Recall our goal of generalization:

$$\mathbb{P}(R_{\text{emp}}(f, X, Y) - R_{\text{true}}(f) < \epsilon) > 1 - \delta$$

where

$$R_{\text{emp}}(f, X, Y) = \frac{1}{N} \sum_i \ell(f(x_i), y_i)$$

and

$$R_{\text{true}}(f) = \mathbb{E}_{x,y} [\ell(f(x), y)]$$

In other words, we want the empirical average to be close to the mean. This is called *concentration*, i.e. the empirical mean concentrates around the true mean.

### 1.1 Coin flips

Instead of risk, let's consider a much simpler example. Suppose I toss a fair coin  $n$  times, and record  $x_i = 1$  if heads and  $x_i = 0$  otherwise. Consider the average,

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N x_i$$

It is easy to see that  $\mathbb{E}[\hat{\mu}_N] = 1/2$ . How far away is  $\hat{\mu}_N$  from its expectation? For example, if  $x_i = 1$  for all  $N$  flips, then  $\hat{\mu}_N = 1$  and it is very far.

*Concentration of measure phenomenon* says that  $\hat{\mu}_N$  “concentrates” closer to  $\mathbb{E}[\hat{\mu}_N]$ , i.e.

The average of  $N$  i.i.d. variables concentrates within an interval of length roughly  $1/\sqrt{N}$  around the mean.

- Intuitively, if the average is far from the expectation, then many independent variables need to work together which is extremely unlikely.
- The concentration result is actually stronger:  $\hat{\mu}_N$  has an approximately Normal distribution.
- This result underlies pretty much all of statistics and machine learning.

## 1.2 Tail inequalities

- Markov's inequality: for positive random variable  $x \geq 0$  and  $\mathbb{E}[X] = \mu < \infty$  then

$$P(X \geq t) \leq \frac{\mu}{t} = O\left(\frac{1}{t}\right)$$

- Very crude, but no distributional assumption, only non-negativity and finite mean!
- "If mean is small, then it is unlikely to be large."
- Proof: basic probability

$$\mathbb{E}[X] = \int_0^\infty xp(x)dx \geq \int_t^\infty xp(x)dx \geq t \int_t^\infty p(x)dx = t\mathbb{P}(X \geq t)$$

- Chebyshev's inequality: for random variable  $X$  with finite variance  $V(X) = \sigma^2$ , for any  $t > 0$  we have

$$\mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2} = O\left(\frac{1}{t^2}\right)$$

- Proof: apply Markov's inequality

$$\mathbb{P}(|X - \mu| \geq t\sigma) = P(|X - \mu|^2 \geq t^2\sigma^2) \leq \frac{\mathbb{E}[|X - \mu|^2]}{t^2\sigma^2} = \frac{1}{t^2}$$

- With more assumptions (finite variance) we can get a better rate  $1/t^2$  instead of  $1/t$ .

**Weak Law of Large Numbers (almost).** Returning to  $\hat{\mu}_N = \frac{1}{N} \sum_i X_i$  (i.e. the coin flip example), note that this has mean  $\mu$  and variance  $\sigma^2/N$ . Apply Chebyshev's inequality to  $\hat{\mu}_N$  and we get:

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq \frac{1}{t^2}$$

So, with probability at least 0.99 (i.e. by taking  $1/t^2 = 0.01$  for  $t = 10$ ), then the average is within  $10\sigma/\sqrt{N}$  of the expectation. This is something called the Weak Law of Large Numbers. The key property is the  $\frac{1}{\sqrt{N}}$  behavior, with better refinements having dramatically better constants than 10.

- Chernoff Method: introduce a parameter  $t$  and an exponential function to refine the Chebyshev inequality.
- For any  $t > 0$ , we have that

$$\mathbb{P}((X - \mu) \geq u) = P(\exp(t(X - \mu)) \geq \exp(tu)) \leq \frac{\mathbb{E}[\exp(t(X - \mu))]}{\exp(tu)}$$

by applying Markov's inequality.

- Chernoff's bound:

$$\mathbb{P}((X - \mu) \geq u) \leq \inf_{0 \leq t \leq b} \frac{\mathbb{E}[\exp(t(X - \mu))]}{\exp(tu)}$$

where  $b$  is such that  $\mathbb{E}[\exp(tX)]$  (the moment generating function, or mgf) is finite for all  $t \leq b$ .

- This can be rewritten as

$$\mathbb{P}((X - \mu) \geq u) \leq \inf_{0 \leq t \leq b} \exp(-t(u + \mu)) \mathbb{E}[\exp(tX)]$$

which is now in terms of the MGF.

*Aside:* The moment generating function is called such because it can be used to “generate” all the “moments” (i.e. the expected value of  $X^t$  for all integer powers of  $t$ ). Simply write out the Taylor series as

$$M_X(t) = \mathbb{E}[\exp(tX)] = \mathbb{E}\left[1 + tX + \frac{t^2 X^2}{2!} + \dots\right] = 1 + t\mathbb{E}[X] + \frac{t^2 \mathbb{E}[X^2]}{2!} + \dots$$

Then differentiate  $i$  times with respect to  $t$  and set  $t = 0$  to get the  $i$ th moment (i.e.  $\mathbb{E}[X^i]$ ). Fun fact: the form of the MGF specifies the entire distribution (i.e. if you know the MGF then there is only one density it could be). This proof is a bit more technical and can be found in “An Introduction to Probability Theory and Its Applications, Vol. 2” by Feller using Laplace transform theory.

- MGF of a standard normal  $N(0, 1)$ :

$$m_X(t) = \mathbb{E}[\exp(tX)] = \int \exp(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int \frac{1}{\sqrt{2\pi}} e^{tx - \frac{1}{2}x^2} dx$$

- Completing the square gets us

$$\int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2 + tx - \frac{1}{2}t^2 + \frac{1}{2}t^2} dx = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2 + \frac{1}{2}t^2} dx = e^{\frac{1}{2}t^2}$$

- Example: Gaussian tail bound. Suppose  $X \sim N(\mu, \sigma^2)$ . Then, if  $Z$  is standard Normal, then  $X = \sigma Z + \mu$ . Then,

$$\mathbb{E}[\exp(tX)] = \mathbb{E}[\exp(t(\sigma Z + \mu))] = \mathbb{E}[\exp(t\sigma Z) \exp(t\mu)] = \exp(t\mu) m_Z(t\sigma) = \exp(t\mu + \frac{1}{2}t^2\sigma^2)$$

- To apply Chernoff's bound, we compute the minimum over all  $t$ :

$$\inf_{t \geq 0} \exp(-t(u + \mu)) \exp(t\mu + \frac{1}{2}t^2\sigma^2) = \inf_{t \geq 0} \exp(-tu + \frac{1}{2}t^2\sigma^2)$$

which is minimized at  $t = \frac{u}{\sigma^2}$

- Plug this in to get

$$\mathbb{P}((X - \mu) \geq u) \leq \exp(-\frac{u^2}{\sigma^2} + \frac{u^2}{2\sigma^2}) = \exp(-\frac{u^2}{2\sigma^2})$$

- This is a one-sided tail bound. Combining with the other side of the tail bound

$$\mathbb{P}(|X - \mu| \geq u) \leq 2 \exp(-\frac{u^2}{2\sigma^2})$$

- This bound is much tighter than Chebyshevs. For  $\hat{\mu} = \frac{1}{N} \sum_i X_i$ , where  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ , we have  $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/N)$ .
- Then, the Gaussian tail bound for this where  $u = t\sigma/\sqrt{N}$  is

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq t\sigma/\sqrt{N}\right) \leq 2 \exp(-\frac{t^2}{2})$$

- Compare to the WLLN variant from before:

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq \frac{1}{t^2}$$

*Aside:* Both bounds say the deviation goes down at  $\frac{1}{\sqrt{N}}$ . However, Gaussian tail bound goes down with exponentially fast. Previously Chebyshev told us with probability 0.99, the average is within  $10\sigma/\sqrt{N}$ . With the exponential tail bound, with probability 0.99 we have that the average is within

$$\sqrt{2 \ln(1/0.005)}\sigma/\sqrt{N} \approx 3.25\sigma/\sqrt{N}$$

More generally, Chebyshev says:

$$|\hat{\mu} - \mu| \leq \frac{\sigma}{\sqrt{n\delta}}$$

whereas Gaussian tails tell us

$$|\hat{\mu} - \mu| \leq \sigma \sqrt{\frac{2 \ln(2/\delta)}{n}}$$

where the first is polynomial in  $\delta$  and the second is logarithmic.

- The previous Gaussian tail inequality actually applies more generally to a class of random variables known as sub-Gaussian random variables
- Intuitively, a sub-Gaussian distribution is one whose tails decay faster than a Gaussian
- This includes many of the examples we saw before, such as Bernoulli or Beta
- A random variable  $X$  with mean  $\mu$  is sub-Gaussian if there exists a  $\sigma > 0$  such that

$$\mathbb{E}[\exp(t(X - \mu))] \leq \exp(\sigma^2 t^2 / 2)$$

- Note this upper bound is the same as the Gaussian tail bound with zero mean, i.e.

$$\mathbb{E}[\exp(tX)] \leq \exp(t\mu + \frac{1}{2}t^2\sigma^2) = \exp(\frac{1}{2}t^2\sigma^2)$$

if  $X$  has mean zero. .

- Gaussian random variables with variance  $\sigma^2$  trivially satisfy this relation as a  $\sigma$ -sub-Gaussian random variable. Hence, the random variable is *sub*-Gaussian if its moment generating function is dominated by a Gaussian with variance  $\sigma^2$
- You can repeat the same Chernoff procedure for the Gaussian tails to conclude that sub-Gaussians have the same two-sided exponential tail bound (so we won't repeat it here)

$$\mathbb{P}(|X - \mu| \geq u) \leq 2 \exp(-u^2/(2\sigma^2))$$

- Recall that if  $X_1, \dots, X_N \sim \mathcal{N}(\mu, \sigma^2)$  then  $\frac{1}{N} \sum_i X_i \sim \mathcal{N}(\mu, \sigma^2/N)$ . We proved this with properties of Gaussian random variables.
- Similarly, if  $X_1, \dots, X_N$  are  $\sigma$ -sub-Gaussian, then their average  $\frac{1}{N} \sum_i X_i$  is  $\sigma/\sqrt{N}$ -sub-Gaussian.
- Proof:

$$\begin{aligned} \mathbb{E}[\exp(t(\hat{\mu} - \mu))] &= \mathbb{E}[\exp(\frac{t}{N} \sum_i (X_i - \mu))] \\ &= \prod_i \mathbb{E}[\exp(\frac{t}{N} (X_i - \mu))] \leq \prod_i \exp(\frac{t^2}{N^2} \sigma^2 / 2) = \exp(t^2 \sigma^2 / (2N)) \end{aligned}$$

and hence it is  $\sigma/\sqrt{N}$ -sub-Gaussian.

- This directly implies the following two-sided tail bound for the average of sub-Gaussian random variables. Plugging it in gets

$$\mathbb{P}(|\hat{\mu} - \mu| \geq u) \leq 2 \exp(-u^2 N / (2\sigma^2))$$

and substituting  $u = k\sigma/\sqrt{N}$  gets the familiar form from the Gaussian two-sided tail bound:

$$\mathbb{P}(|\hat{\mu} - \mu| \geq k\sigma/\sqrt{N}) \leq 2 \exp(-k^2/2)$$

*Aside: Recall that the property we really cared about for concentration was that the mass in the tails shrinks exponentially. This was formalized in the previous section as “faster than a Gaussian”, and implies an exponentially decaying concentration bound  $2 \exp(-t^2 nd/2)$  as opposed to the Chebyshev concentration bound of  $\frac{1}{t^2}$ , which only assumes finite variance.*

- Hoeffding's bound: a special case of sub-Gaussian random variables is bounded random variables. This will be our final concentration bound.
- Intuition: if a random variable only takes values within a fixed range of  $[a, b]$ , then their tails decay faster than a Gaussian (the tails are zero).
- Bounded random variables are definitely sub-Gaussian, but for what parameter  $\sigma$ ?
- Example: Rademacher random variable, is  $\{+1, -1\}$  with equal probability.
- Then, Rademacher random variables are  $\sigma = 1$ -sub-Gaussian:

$$\mathbb{E}[\exp(tX)] = \frac{1}{2}[\exp(t) + \exp(-t)] = \frac{1}{2} \left[ \sum_{k \geq 0} \frac{t^k}{k!} + \sum_{k \geq 0} \frac{(-t)^k}{k!} \right]$$

$$= \sum_{k \geq 0} \frac{t^{2k}}{(2k)!} \leq \sum_{k \geq 0} \frac{t^{2k}}{2^k k!} = \exp(t^2/2)$$

and therefore we can use the sub-Gaussian tail bound (plug in  $\sigma = 1$ ).

- Can we do this more generally for random variables  $X$  that take on values between some bounded interval  $[a, b]$ ?
- Jensen's inequality: a useful inequality seen in many places (convex optimization).
- Basic 1D definition of convexity: a function  $g$  is convex if

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y)$$

for all  $x, y$  and  $\alpha \in [0, 1]$ . Intuitively, this means that any line connecting two points on  $g$  lies above  $g$ .

- Example:  $g(x) = x^2$  is convex.
- Jensen's inequality states that for a convex function  $g : \mathbb{R} \rightarrow \mathbb{R}$ , then

$$\mathbb{E}[g(x)] \geq g(\mathbb{E}[X])$$

- "A linear function before  $g$  is at most a linear function after  $g$ "
- Proof: Let  $\mu = \mathbb{E}[X]$ , and let  $L_\mu(x) = ax + b$  be the tangent line to the function  $g$  at  $\mu$ , at i.e.  $L_\mu(\mu) = g(\mu)$ . By convexity (\*), we know that  $g(x) \geq L_\mu(x)$  at all  $x$ . Therefore,

$$\mathbb{E}[g(x)] \geq \mathbb{E}[L_\mu(X)] = \mathbb{E}[aX + b] = a\mu + b = L_\mu(\mu) = g(\mu)$$

- Proof of (\*): WLOG suppose  $y > x$ .

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y)$$

$$g(\alpha x + (1 - \alpha)y) - g(x) \leq (\alpha - 1)g(x) + (1 - \alpha)g(y)$$

$$g(\alpha x + (1 - \alpha)y) - g(x) \leq (1 - \alpha)(g(y) - g(x))$$

Note that  $[\alpha x + (1 - \alpha)y] - x = (1 - \alpha)(y - x) \geq 0$ , so divide both sides by this quantity

$$\frac{g(\alpha x + (1 - \alpha)y) - g(x)}{[\alpha x + (1 - \alpha)y] - x} \leq \frac{g(y) - g(x)}{y - x}$$

Take the limit as  $\alpha \rightarrow 1$  and we get

$$g'(x) \leq \frac{g(y) - g(x)}{y - x}$$

$$g'(x)(y - x) + g(x) \leq g(y)$$

so the tangent line lies below  $g$ .

- Next: MGF of bounded random variables. Let  $X$  have zero mean and takes values on the bounded interval  $[a, b]$ .

- Zero mean assumption doesn't matter (can always subtract the mean and use  $Y = X - \mathbb{E}[X]$  instead).
- Let  $X'$  be an independent copy of  $X$ . Then using Jensen's inequality (and the exponential function being convex),

$$\mathbb{E}_X(\exp(tX)) = \mathbb{E}_X(\exp(t(X - \mathbb{E}[X']))) \leq \mathbb{E}_{X,X'}(\exp(t(X - X')))$$

- Furthermore let  $\epsilon$  be a Rademacher random variable, then,  $X - X'$  is identical to the distribution of  $X' - X$ , which is identical to  $\epsilon(X - X')$ . Then, using the Hoeffding bound for Rademacher random variables,

$$\mathbb{E}_X(\exp(tX)) \leq \mathbb{E}_{X,X'}[\exp(t^2(X - X')^2/2)]$$

The goal of this step is to make the bound agnostic to whether  $X > X'$  or vice versa. Using boundedness, we have

$$\mathbb{E}_X(\exp(tX)) \leq \exp(t^2(b - a)^2/2)$$

and so bounded random variables are  $(b - a)$ -sub Gaussian.

- There is a stronger version called Hoeffding's Lemma which has a denominator of 8 instead of 2.
- Concentration bound: Suppose  $X_1, \dots, X_N$  are bounded iid random variables with  $a \leq X_i \leq b$ . Let  $\hat{\mu} = \frac{1}{N} \sum_i X_i$ . Then, applying Markov's followed by the MGF bound, we get

$$\mathbb{P}(\hat{\mu} \geq u) = \mathbb{P}\left(\exp\left(t \sum_i X_i\right) \geq \exp(tNu)\right) \leq e^{-tNu} \mathbb{E}\left[\exp\left(\sum_i tX_i\right)\right]$$

$$\mathbb{P}(\hat{\mu} \geq u) \leq e^{-tNu} \prod_i \mathbb{E}[\exp(tX_i)] \leq e^{-tNu} \exp(Nt^2(b-a)^2/2) = \exp\left(N\left(\frac{(b-a)^2}{2}t^2 - tu\right)\right)$$

The RHS is minimized at

$$t(b-a)^2 - u = 0 \Rightarrow t = \frac{u}{(b-a)^2}$$

and so therefore the one sided bound is

$$\mathbb{P}(\hat{\mu} \geq u) \leq \exp\left(N\left(\frac{(b-a)^2}{2} \frac{u^2}{(b-a)^4} - \frac{u^2}{(b-a)^2}\right)\right) = \exp\left(-\frac{Nu^2}{2(b-a)^2}\right)$$

- The two sided bound is thus

$$\mathbb{P}(|\hat{\mu}| \geq u) \leq 2 \exp\left(-\frac{Nu^2}{2(b-a)^2}\right)$$

- A slightly stronger bound with Hoeffding's Lemma gives

$$\mathbb{P}(|\hat{\mu}| \geq u) \leq 2 \exp\left(-\frac{2Nu^2}{(b-a)^2}\right)$$

*Aside: Hoeffding's Lemma allows us to give concentration inequalities for bounded random variables. This takes the key idea of a Gaussian for concentration (the exponentially decreasing tails) and generalizes it to a broad class of random variables (bounded) that makes the concentration inequality applicable in real-world settings. In many settings, we can assume our data is bounded.*

## 2 Generalization Bound

Finally, we can prove our first generalization bound! We will prove that the empirical estimator  $\hat{f} = \arg \min_f R_{\text{emp}}(f, X, Y)$  has true risk close to the true optimal risk.

$$\mathbb{P} \left( R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) < \epsilon \right) > 1 - \delta$$

where

$$R_{\text{emp}}(f, X, Y) = \frac{1}{N} \sum_i \ell(f(x_i), y_i)$$

and

$$R_{\text{true}}(f) = \mathbb{E}_{x,y} [\ell(f(x), y)]$$

Note this is slightly different from what we looked at earlier, as we want to match the true risk of the *optimal* predictor  $f^* = \arg \min_f R_{\text{true}}(f)$  with the true risk of the estimated predictor  $\hat{f} = \arg \min_f R_{\text{emp}}(f, X, Y)$ . This is an even stronger statement.

To start, we can decompose the difference in risk into three parts:

$$[R_{\text{true}}(\hat{f}) - R_{\text{emp}}(\hat{f})] + [R_{\text{emp}}(\hat{f}) - R_{\text{emp}}(f^*)] + [R_{\text{emp}}(f^*) - R_{\text{true}}(f^*)]$$

- The first term is difficult, as  $\hat{f}$  is a random variable that is not i.i.d.
- The second term is  $\leq 0$  because by definition,  $\hat{f}$  minimizes the empirical risk.
- The third term is an i.i.d. sum since  $f^*$  is deterministic.

To prove generalization, we'll use a concept known as uniform bounds. Upper bounding with absolute values we get

$$\begin{aligned} R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) &\leq |R_{\text{true}}(\hat{f}) - R_{\text{emp}}(\hat{f})| + 0 + |R_{\text{emp}}(f^*) - R_{\text{true}}(f^*)| \\ R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) &\leq \sup_f |R_{\text{emp}}(f) - R_{\text{true}}(f)| + \sup_f |R_{\text{emp}}(f) - R_{\text{true}}(f)| \\ R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) &\leq 2 \cdot \sup_f |R_{\text{emp}}(f) - R_{\text{true}}(f)| \end{aligned}$$

In other words, we are bounding the excess risk (LHS) with the worst case difference between the empirical and true risk over all possible functions  $f$ . The RHS is sometimes called an empirical process in statistics. If we can control this, then we can control the generalization error.

Then, our generalization bound becomes

$$\mathbb{P} \left( R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) \geq \epsilon \right) \leq \mathbb{P} \left( \sup_f |R_{\text{emp}}(f) - R_{\text{true}}(f)| \geq \frac{\epsilon}{2} \right)$$



## 2.1 Generalization for finite function classes, $|\mathcal{F}| < \infty$

We will now prove the following: If a function class is finite,  $|\mathcal{F}| < \infty$ , and loss is bounded ( $0 \leq \ell \leq B$ ), then we have

$$\mathbb{P} \left( R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) < B \sqrt{\frac{2 \log(2|\mathcal{F}|) + 2 \log \delta^{-1}}{n}} \right) > 1 - \delta$$

The proof has three main steps.

1. Hoeffding's inequality and since the loss is bounded, we know that

$$\mathbb{P} \left( |R_{\text{emp}}(f) - R_{\text{true}}(f)| \geq \frac{\epsilon}{2} \right) \leq 2 \exp \left( -\frac{N \epsilon^2}{2B^2} \right)$$

2. Finite function class assumption with union bound:

$$\begin{aligned} P \left( \sup_f |R_{\text{emp}}(f) - R_{\text{true}}(f)| \geq \frac{\epsilon}{2} \right) &= P \left( \bigcup_f \left\{ |R_{\text{emp}}(f) - R_{\text{true}}(f)| \geq \frac{\epsilon}{2} \right\} \right) \\ &\leq \sum_f P \left( \left\{ |R_{\text{emp}}(f) - R_{\text{true}}(f)| \geq \frac{\epsilon}{2} \right\} \right) \\ &\leq 2|\mathcal{F}| \exp \left( -\frac{N \epsilon^2}{2B^2} \right) \end{aligned}$$

3. Finally, connect the uniform convergence bound back to generalization to get

$$\begin{aligned} \mathbb{P} \left( R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) \geq \epsilon \right) &\leq \mathbb{P} \left( \sup_f |R_{\text{emp}}(f) - R_{\text{true}}(f)| \geq \frac{\epsilon}{2} \right) \\ &\leq 2|\mathcal{F}| \exp \left( -\frac{N \epsilon^2}{2B^2} \right) \end{aligned}$$

Setting this equal to  $\delta$  and solving for  $\epsilon$  we get

$$\epsilon^2 = \frac{2B^2}{N} \log(2|\mathcal{F}|\delta^{-1})$$

Plugging this in we get

$$\mathbb{P} \left( R_{\text{true}}(\hat{f}) - R_{\text{true}}(f^*) \geq B \sqrt{\frac{2 \log(2|\mathcal{F}|) + 2 \log \delta^{-1}}{n}} \right) \leq \delta$$

which recovers the end result.