Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning

Instructor: Purushottam Kar
Date: February 15, 2018
Total: 60 marks

ASSIGNMENT 2

Problem 2.1 (Sigmoidal Consistency)**.** Consider a binary classification problem with a distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$. Consider the misclassification loss function $\ell^{0-1}(\hat{y}, y) = \mathbb{I}\{\hat{y} \neq y\}$ for $\hat{y}, y \in \{-1, +1\}$ as well as the sigmoid surrogate loss function $\ell^{\sigma}(\hat{y}, y) = \frac{1}{1+\exp(y\hat{y})}$ for $\hat{y} \in \mathbb{R}$ and $y \in \{-1, +1\}$, which is popularly used to train neural networks. Note that whereas the misclassification loss function expects a proper binary classification as the prediction $\hat{y} \in \{-1, +1\}$, the sigmoid loss can operate with real valued predictions as well i.e. $\hat{y} \in \mathbb{R}$.

For any data point $\mathbf{x} \in \mathcal{X}$, let $\eta(\mathbf{x})$ denote the posterior class probability for the data point $\mathbf{x}$ i.e. $\eta(\mathbf{x}) = \mathbb{P}[Y = 1 \,|\, X = \mathbf{x}]$. Also, for any $\eta \in [0, 1]$ let $Y \sim \eta$ denote a Rademacher variable that takes value $+1$ with probability η and -1 with probability $1 - \eta$. For any loss function ℓ, and any prediction $\hat{y}$, define

$$L^{\ell}(\hat{y}, \eta) := \mathop{\mathbb{E}}_{Y \sim \eta} [\ell(\hat{y}, Y)],$$

as the expected (conditional) ℓ-loss of predicting $\hat{y}$ on a data point whose true label follows the distribution η. Also denote $L^{\ell}(\eta) := \min_{\hat{y}} L^{\ell}(\hat{y}, \eta)$ as the lowest possible conditional ℓ-loss one can hope to achieve on such a data point using a single best prediction. Define the regret of any predictor f with respect to loss function ℓ as

$$\mathcal{R}^{\ell}_{\mathcal{D}}[f] := \mathop{\mathbb{E}}_{(X,Y) \sim \mathcal{D}} \left[L^{\ell}(f(X), \eta(X)) - L^{\ell}(\eta(X)) \right]$$

Show the following results

1. For any $\eta \in [0, 1]$, we have $L^{0-1}(\eta) = L^{\sigma}(\eta) = \min\{\eta, 1 - \eta\}$.
2. For any $\hat{y} \in \{-1, +1\}, \eta \in [0, 1]$, we have $L^{0-1}(\hat{y}, \eta) - L^{0-1}(\eta) = |2\eta - 1| \cdot \mathbb{I}\{\hat{y}(2\eta - 1) < 0\}$.
3. For any $\hat{y} \in \mathbb{R}, \eta \in [0, 1]$, we have $L^{0-1}(\text{sign}(\hat{y}), \eta) - L^{0-1}(\eta) \leq 2\left(L^{\sigma}(\hat{y}, \eta) - L^{\sigma}(\eta)\right)$.
4. Conclude that for any real valued predictor $f \in \mathbb{R}^{\mathcal{X}}$, we have $\mathcal{R}^{0-1}_{\mathcal{D}}[\text{sign} \circ f] \leq 2\mathcal{R}^{\sigma}_{\mathcal{D}}[f]$, where we define $(\text{sign} \circ f)(\mathbf{x}) := \text{sign}(f(\mathbf{x}))$ for any $\mathbf{x} \in \mathcal{X}$.
5. Derive a similar regret transfer bound for the accentuated sigmoid loss $\ell^{\sigma}_{\alpha}(\hat{y}, y) = \frac{1}{1+\exp(\alpha \cdot y\hat{y})}$ for some $\alpha > 0$.

The above shows that achieving small sigmoidal regret automatically ensures small misclassification regret. The sigmoidal loss is easier to work with since it is differentiable but has these nice regret transfer properties as well. Do keep in mind that the minimization over $\hat{y}$ for finding $L^{0-1}(\eta)$ is over Rademacher predictions i.e. over $\hat{y} \in \{-1, +1\}$ whereas the minimization for finding $L^{\sigma}(\eta)$ is over real valued predictions i.e. over $\hat{y} \in \mathbb{R}$. (3+2+3+2+5=15 marks)

Solution. The solution is given in parts below

1. For ℓ^{0-1} we have

$$L^{0-1}(\hat{y}, \eta) = \eta \cdot \mathbb{I}\{\hat{y} \neq 1\} + (1 - \eta) \cdot \mathbb{I}\{\hat{y} \neq -1\} \geq \min\{\eta, 1 - \eta\}$$

since $\mathbb{I}\{\hat{y} \neq 1\} + \mathbb{I}\{\hat{y} \neq -1\} = 1$ and $\mathbb{I}\{\cdot\} \geq 0$. The classification $\hat{y} = 2 \cdot \mathbb{I}\{\eta \geq 0.5\} - 1$ achieves this value. Thus $L^{0-1}(\eta) = \min\{\eta, 1-\eta\}$. For ℓ^σ, we have

$$L^\sigma(\hat{y}, \eta) = \eta \cdot \frac{1}{1+\exp(\hat{y})} + (1-\eta) \cdot \frac{1}{1+\exp(-\hat{y})} \geq \min\{\eta, 1-\eta\},$$

since yet again $\frac{1}{1+\exp(\hat{y})} + \frac{1}{1+\exp(-\hat{y})} = 1$ and both terms are non-negative. This value is not achieved by any finite prediction but if we let $\hat{y} = B(2 \cdot \mathbb{I}\{\eta \geq 0.5\} - 1)$ and let $B \to \infty$, then this value is approached in the limit.

2. If $\hat{y} = 2 \cdot \mathbb{I}\{\eta \geq 0.5\} - 1$ i.e. $\hat{y}(2\eta - 1) \geq 0$, then part 1 shows us that $L^{0-1}(\hat{y}, \eta) = L^{0-1}(\eta)$ and both sides are zero and the result holds. However, if $\hat{y} = 1 - 2 \cdot \mathbb{I}\{\eta \geq 0.5\}$ i.e. $\hat{y}(2\eta - 1) < 0$, then we have $L^{0-1}(\hat{y}, \eta) - L^{0-1}(\eta) = \max\{\eta, 1-\eta\} - \min\{\eta, 1-\eta\} = |2\eta - 1|$ which proves the result.

3. If $\hat{y}(2\eta - 1) \geq 0$ (i.e. it is the "correct sign" – recall that $\hat{y} \in \mathbb{R}$), then we have $\text{sign}(\hat{y}) = 2 \cdot \mathbb{I}\{\eta \geq 0.5\} - 1$. In this case, we have $L^{0-1}(\text{sign}(\hat{y}), \eta) - L^{0-1}(\eta) = 0$. Since $L^\sigma(\hat{y}, \eta) - L^\sigma(\eta) \geq 0$ no matter which $\hat{y}$ we choose, the desired inequality holds. However, if the prediction is not the "correct sign" i.e. $\hat{y}(2\eta - 1) < 0$, then we have

$$\begin{aligned}
L^\sigma(\hat{y}, \eta) &= \eta \cdot \frac{1}{1+\exp(\hat{y})} + (1-\eta) \cdot \frac{1}{1+\exp(-\hat{y})} \\
&= \eta \cdot \frac{1}{1+\exp(\hat{y})} + (1-\eta) \cdot \left(1 - \frac{1}{1+\exp(\hat{y})}\right) \\
&= (2\eta - 1)\frac{1}{1+\exp(\hat{y})} + (1-\eta) \\
&\geq 0.5(2\eta - 1) + (1-\eta)
\end{aligned}$$

where in the last step, we used the fact that the expression $(2\eta - 1)\frac{1}{1+\exp(\hat{y})}$ takes its minimum value, while respecting $\hat{y}(\eta - 0.5) \leq 0$, when $\hat{y} = 0$ (this holds true for all η). Thus, we have, whenever $\hat{y}(2\eta - 1) < 0$,

$$L^\sigma(\hat{y}, \eta) - L^\sigma(\eta) \geq 0.5(2\eta - 1) + (1-\eta) - \min\{\eta, 1-\eta\} = 0.5\,|2\eta - 1|$$

This proves the result when combined with part 2.

4. We have the following relations (in the third step we use part 3)

$$\begin{aligned}
\mathcal{R}^{0-1}_{\mathcal{D}}[\text{sign} \circ f] &= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \left[L^{0-1}(\text{sign}(f(X)), \eta(X)) - L^{0-1}(\eta(X))\right] \\
&= \mathop{\mathbb{E}}_{X} \left[\mathop{\mathbb{E}}_{Y\sim\eta(X)} \left[L^{0-1}(\text{sign}(f(X)), \eta(X)) - L^{0-1}(\eta(X)) \,|\, X\right]\right] \\
&\leq \mathop{\mathbb{E}}_{X} \left[\mathop{\mathbb{E}}_{Y\sim\eta(X)} \left[2(L^\sigma(f(X), \eta(X)) - L^\sigma(\eta(X))) \,|\, X\right]\right] \\
&= 2 \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \left[L^\sigma(\text{sign}(f(X)), \eta(X)) - L^\sigma(\eta(X))\right] = \mathcal{R}^\sigma_{\mathcal{D}}[f]
\end{aligned}$$

5. The accentuated loss function also satisfies $L^\sigma_\alpha(\hat{y}, \eta) \geq \min\{\eta, 1-\eta\}$ for any $\alpha > 0$. It also satisfies the fact that the expression $(2\eta - 1)\frac{1}{1+\exp(\alpha\hat{y})}$ takes its minimum value, while respecting $\hat{y}(\eta - 0.5) \leq 0$, when $\hat{y} = 0$ (this holds true for all η). This gives us $\mathcal{R}^{0-1}_{\mathcal{D}}[\text{sign} \circ f] \leq 2\mathcal{R}^\sigma_{\alpha\mathcal{D}}[f]$ for the accentuated sigmoidal loss function as well.

Problem 2.2 (Grand Statistical Taxation?)**.** The finance ministry wants to restructure the GST brackets. For this, it wishes to find the κ-th percentile annual expenditure for India for some $\kappa \in (0,1)$. If we let $\mathcal{I}$ be the set of all Indians where $|\mathcal{I}| =: N$, and for any Indian $i \in \mathcal{I}$, let $E(i)$ denote his/her expenditure, then the finance ministry wants to find a threshold amount v^κ such that $\frac{1}{N} \cdot |\{i \in \mathcal{I} : E(i) \leq v^\kappa\}| = \kappa$. Since personal expenditure cannot be found from income tax returns (most Indians do not file returns anyway), it seems the only way to do the above is ask each of the $N \approx 1.3 \times 10^9$ Indians what their expenditure is, a daunting task.

However, my expert statistician friend tells me that there is a way out. She instructs me to choose a random set S of $n \ll N$ Indians uniformly at random with replacement, ask them their annual expenditure, and then find an amount $\hat{v}_S^\kappa$ such that $\frac{1}{n} \cdot |\{i \in S : E(i) \leq \hat{v}_S^\kappa\}| = \kappa$.

Do you agree with the sampling strategy and the construction of the estimator $\hat{v}_S^\kappa$ or would you like to change them? For your choice of strategy and estimator, show that for tolerance and confidence parameters ϵ, δ that the finance ministry desires, you can ensure

$$\mathop{\mathbb{P}}_{S}\left[\left|\frac{1}{N} \cdot |\{i \in \mathcal{I} : E(i) \leq \hat{v}_S^\kappa\}| - \kappa\right| > \epsilon\right] \leq \delta$$

You may skip small universal constants like $5, 14, 22$ while stating your result but must give explicit dependence on ϵ, δ. You may also assume that $\kappa \cdot N \in \mathbb{N}$ and $\kappa \cdot n \in \mathbb{N}$.

Hint: Do not attempt to show $v^\kappa \approx \hat{v}_S^\kappa$. Such a result may not even be possible (as both v^κ and $\hat{v}_S^\kappa$ are not unique in general) and as such, may or may not guarantee what we require.

Hint: Any threshold v^κ splits the population $\mathcal{I}$ into two groups in whose relative sizes we are interested. (20 marks)

Solution. Consider a classification problem on the set of Indians. I assign every Indian $i \in \mathcal{I}$ a label $y_i^* \in \{-1, 1\}$. However, I assign every Indian the same label i.e. $y_i^* = -1$ for all $i \in \mathcal{I}$. With respect to every threshold v, I can uniquely associate a classifier $f_v(i) = 2 \cdot \mathbb{I}\{E(i) \leq v\} - 1$. Notice that if $\mathcal{D}$ is the uniform distribution on the set $\mathcal{I}$ then for any threshold $v \in \mathbb{R}$, we have

$$\frac{1}{N} \cdot |\{i \in \mathcal{I} : E(i) \leq v\}| = \mathrm{er}_{\mathcal{D}}^{0-1}[f_v]$$
$$\frac{1}{n} \cdot |\{i \in S : E(i) \leq v\}| = \mathrm{er}_{S}^{0-1}[f_v]$$

Let $\mathcal{F} = \{f_v; v \in \mathbb{R}\}$ denote the set of threshold classifiers. Then, we know that $\mathrm{VC}(\mathcal{F}) = 1$, and using Rademacher complexity bounds, show that if $|S| = n$, then

$$\mathop{\mathbb{P}}_{S}\left[\sup_{v \in \mathbb{R}} \left|\mathrm{er}_{\mathcal{D}}^{0-1}[f_v] - \mathrm{er}_{S}^{0-1}[f_v]\right| > \sqrt{\frac{\log n}{n}} + \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

In particular, setting $v = \hat{v}_S^\kappa$ gives us

$$\mathop{\mathbb{P}}_{S}\left[\left|\mathrm{er}_{\mathcal{D}}^{0-1}[f_{\hat{v}_S^\kappa}] - \kappa\right| > \sqrt{\frac{\log n}{n}} + \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

which proves the desired result since $\mathrm{er}_{\mathcal{D}}^{0-1}[f_{\hat{v}_S^\kappa}] = \frac{1}{N} \cdot |\{i \in \mathcal{I} : E(i) \leq \hat{v}_S^\kappa\}|$ as noted above.

Problem 2.3 (Well . . . the quest for fairness continues)**.** With the proliferation of learning algorithms in various services, the fear of these algorithms making biased, unfair, or even morally dubious decisions, looms. Consider a labeled domain $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ over which a regression problem is to be solved with respect to some bounded loss function $\ell : \mathbb{R} \times \mathcal{Y} \to [0, B]$ and an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Assume that ℓ is L-Lipschitz. I have an algorithm

$\mathcal{A} : \bigcup_{n\geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ that can give, for any training set $S \in (\mathcal{X} \times \mathcal{Y})^n$, a model $\hat{f}_S \in \mathcal{F}$. The algorithm gave me good test performance with respect to ℓ.

However, I am still a bit uneasy since my population $\mathcal{X}$ is actually composed of two equipopulous subpopulations $\mathcal{X}_1$ and $\mathcal{X}_2$. That is to say that $\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$, $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ and $\mathbb{P}_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_1] = \mathbb{P}_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_2] = 0.5$. I am worried that $\hat{f}_S$ may be doing a poor job for individuals in subpopulation $\mathcal{X}_2$ but is yet able to shine in terms of overall ℓ-risk by doing extremely well on subpopulation $\mathcal{X}_1$. If this is true then this is a lawsuit waiting to happen (e.g. the subpopulations could be age based and $\hat{f}_S$ could be mispredicting the regression value for old people but getting them accurately for young people and I may have been charging individuals in $\mathcal{X}$ different amounts of money based on the output of the algorithm)

I do not like being sued and want to know if the above is indeed the case. In particular, I am interested in the following notion of fairness for a predictor $f \in \mathcal{F}$

$$\mathsf{FAIR}(f) := \left| \mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell(f(X), Y) \,|\, X \in \mathcal{X}_1] - \mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell(f(X), Y) \,|\, X \in \mathcal{X}_2] \right|$$

Can you design and analyze an algorithm to estimate $\mathsf{FAIR}(\hat{f}_S)$? You may assume that the model class $\mathcal{F}$ is nice in terms of having "small" Rademacher complexity or VC dimension or covering number, depending on your analysis technique. However, do note that the algorithm $\mathcal{A}$ was not designed to optimize any notion of fairness, including FAIR. Is it possible to estimate $\mathsf{FAIR}(\hat{f}_S)$ using S itself? (25 marks)

Solution. It is indeed possible to estimate $\mathsf{FAIR}(\hat{f}_S)$ using S itself if S is sampled i.i.d. from $\mathcal{D}$ but is inadvisable since it leads to double dipping which should be avoided if more data is available. Below we provide a technique for estimation of fairness uniformly over the class $\mathcal{F}$ which would automatically address the case of $\hat{f}_S$ as well. Define for sake of convenience

$$F(f) := \mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell(f(X), Y) \,|\, X \in \mathcal{X}_1] - \mathbb{E}_{(X,Y)\sim\mathcal{D}} [\ell(f(X), Y) \,|\, X \in \mathcal{X}_2]$$

Notice that $\mathsf{FAIR}(f) = |F(f)|$. For our estimator, select a set of n data points $T = (X^i, Y^i) \sim \mathcal{D}^n$ independently of S. For sake of convenience, define

$$\wp(f, X, Y) := \ell(f(X), Y) \left(\mathbb{I}\{X \in \mathcal{X}_1\} - \mathbb{I}\{X \in \mathcal{X}_2\}\right)$$

as well as define

$$\hat{F}_T(f) := \frac{2}{n} \sum_{i=1}^{n} \wp(f, X^i, Y^i)$$

Our estimator will be

$$\widehat{\mathsf{FAIR}}_T(f) := \left| \hat{F}_T(f) \right|$$

Let $g(T) := \sup_{f\in\mathcal{F}} \left(\hat{F}_T(f) - F(f) \right)$. Since $\hat{F}_T(f)$ is $4B/n$-stable (as ℓ is bounded and the term $\mathbb{I}\left\{X^i \in \mathcal{X}_1\right\} - \mathbb{I}\left\{X^i \in \mathcal{X}_2\right\}$ takes only two values $\{-1, +1\}$ and is thus, bounded), we conclude that $g(T)$ is also $4B/n$-stable. Thus, by an application of McDiarmid's inequality, we get

$$\mathbb{P}_T \left[\sup_{f\in\mathcal{F}} \left(\hat{F}_T(f) - F(f) \right) \geq \mathbb{E} \left[\sup_{f\in\mathcal{F}} \left(\hat{F}_T(f) - F(f) \right) \right] + \epsilon \right] \leq \exp\left(-\frac{n\epsilon^2}{16B^2} \right)$$

Thus, all we are left to do is estimate the expectation term. To do so, notice that we have

$$\begin{aligned}\mathop{\mathbb{E}}_{T}\left[\hat{F}_T(f)\right] &= \frac{2}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[\ell(f(X^i),Y^i)\mathbb{I}\left\{X^i\in\mathcal{X}_1\right\}\right]-\mathbb{E}\left[\ell(f(X^i),Y^i)\mathbb{I}\left\{X^i\in\mathcal{X}_2\right\}\right]\right)\\ &= 2\left(\mathbb{E}\left[\ell(f(X),Y)\,|\,X\in\mathcal{X}_1\right]\cdot\mathbb{P}\left[X\in\mathcal{X}_1\right]-\mathbb{E}\left[\ell(f(X),Y)\,|\,X\in\mathcal{X}_2\right]\cdot\mathbb{P}\left[X\in\mathcal{X}_2\right]\right)\\ &= F(f)\end{aligned}$$

where in the second step, we use the definition of conditional expectation of random variables. Thus, we have, by a standard symmetrization argument,

$$\begin{aligned}\mathop{\mathbb{E}}_{T}\left[\sup_{f\in\mathcal{F}}\left(\hat{F}_T(f)-F(f)\right)\right] &= \mathop{\mathbb{E}}_{T}\left[\sup_{f\in\mathcal{F}}\left(\hat{F}_T(f)-\mathop{\mathbb{E}}_{\tilde{T}}\left[\hat{F}_{\tilde{T}}(f)\right]\right)\right] \le \mathop{\mathbb{E}}_{T,\tilde{T}}\left[\sup_{f\in\mathcal{F}}\left(\hat{F}_T(f)-\hat{F}_{\tilde{T}}(f)\right)\right]\\ &= \mathop{\mathbb{E}}_{T,\tilde{T}}\left[\sup_{f\in\mathcal{F}}\left(\frac{2}{n}\sum_{i=1}^{n}\wp(f,X^i,Y^i)-\frac{2}{n}\sum_{i=1}^{n}\wp(f,\tilde{X}^i,\tilde{Y}^i)\right)\right]\\ &\le 4\mathop{\mathbb{E}}_{\{X^i,Y^i,\epsilon_i\}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\wp(f,X^i,Y^i)\right]\\ &\le 8\sup_{\{X^i,Y^i\}}\mathop{\mathbb{E}}_{\epsilon_i}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(f(X^i),Y^i)\cdot\mathbb{I}\left\{X^i\in\mathcal{X}_1\right\}\right]\end{aligned}$$

where in the last step, we upper bounded the expectation by a supremum, and used the fact that the terms involving $\mathbb{I}\left\{X^i\in\mathcal{X}_2\right\}$ would contribute similarly, hence we changed the multiplier from 4 to 8. Now, the loss function ℓ is L-Lipschitz i.e. for all $Y\in\mathcal{Y}$ we have $|\ell(a,Y)-\ell(b,Y)|\le L\cdot|a-b|$. Since $|\mathbb{I}\{\cdot\}|\le 1$, this means that we also have, for any X,Y

$$|\ell(a,Y)\cdot\mathbb{I}\{X\in\mathcal{X}_1\}-\ell(b,Y)\cdot\mathbb{I}\{X\in\mathcal{X}_1\}|\le L\cdot|a-b|$$

Applying the Talagrand-Ledoux contraction inequality now gives us

$$\mathop{\mathbb{E}}_{T}\left[\sup_{f\in\mathcal{F}}\left(\hat{F}_T(f)-F(f)\right)\right]\le 8L\sup_{\{X^i,Y^i\}}\mathop{\mathbb{E}}_{\epsilon_i}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X^i)\right]\le 8L\cdot\mathcal{R}_n(\mathcal{F})$$

Thus, we get

$$\mathop{\mathbb{P}}_{T}\left[\sup_{f\in\mathcal{F}}\left(\hat{F}_T(f)-F(f)\right)\ge 8L\cdot\mathcal{R}_n(\mathcal{F})+\epsilon\right]\le\exp\left(-\frac{n\epsilon^2}{16B^2}\right)$$

Similarly, we can get

$$\mathop{\mathbb{P}}_{T}\left[\sup_{f\in\mathcal{F}}\left|\hat{F}_T(f)-F(f)\right|\ge 8L\cdot\mathcal{R}_n(\mathcal{F})+\epsilon\right]\le 2\exp\left(-\frac{n\epsilon^2}{16B^2}\right)$$

Now, by an application of triangle inequality, we know that for any $a,b\in\mathbb{R}$ we always have $||a|-|b||\le|a-b|$. This gives us $\left|\widehat{\mathsf{FAIR}}_T(f)-\mathsf{FAIR}(f)\right|\le\left|\hat{F}_T(f)-F(f)\right|$ i.e.

$$\mathop{\mathbb{P}}_{T}\left[\sup_{f\in\mathcal{F}}\left|\widehat{\mathsf{FAIR}}_T(f)-\mathsf{FAIR}(f)\right|\ge 8L\cdot\mathcal{R}_n(\mathcal{F})+\epsilon\right]\le 2\exp\left(-\frac{n\epsilon^2}{16B^2}\right)$$

which completes the argument whenever the class $\mathcal{F}$ has "vanishing" Rademacher complexity.