Regression with quadratic loss

Maxim Raginsky

October 13, 2015

Regression with quadratic loss is another basic problem studied in statistical learning theory. We have a random couple $Z = (X, Y)$, where, as before, X is an $\mathbb{R}^d$-valued feature vector (or input vector) and Y is the *real-valued* response (or output). We assume that the unknown joint distribution $P = P_Z = P_{XY}$ of (X, Y) belongs to some class $\mathscr{P}$ of probability distributions over $\mathbb{R}^d \times \mathbb{R}$.

The learning problem, then, is to produce a *predictor* of Y given X on the basis of an i.i.d. training sample $Z^n = (Z_1, \ldots, Z_n) = ((X_1, Y_1), \ldots, (X_n, Y_n))$ from P. A predictor is just a (measurable) function $f : \mathbb{R}^d \to \mathbb{R}$, and we evaluate its performance by the *expected quadratic loss*

$$L(f) \triangleq \mathbb{E}[(Y - f(X))^2].$$

As we have seen before, the smallest expected loss is achieved by the *regression function* $f^*(x) \triangleq \mathbb{E}[Y|X = x]$, i.e.,

$$L^* \triangleq \inf_f L(f) = L(f^*) = \mathbb{E}[(X - \mathbb{E}[Y|X])^2].$$

Moreover, for any other f we have

$$L(f) = L^* + \|f - f^*\|^2_{L^2(P_X)},$$

where

$$\|f - f^*\|^2_{L^2(P_X)} = \int_{\mathbb{R}^d} |f(x) - f^*(x)|^2 P_X(dx).$$

Since we do not know P, in general we cannot hope to learn f^*, so, as before, instead we aim at finding a good approximation to the best predictor in some class $\mathscr{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$, i.e., to use the training data Z^n to construct a predictor $\widehat{f}_n \in \mathscr{F}$, such that

$$L(\widehat{f}_n) \approx L^*(\mathscr{F}) \triangleq \inf_{f \in \mathscr{F}} L(f)$$

with high probability.

We will assume that the marginal distribution P_X of the feature vector is supported on a closed subset $\mathsf{X} \subseteq \mathbb{R}^d$, and that the joint distribution P of (X, Y) is such that, with probability one,

$$|Y| \le M \qquad \text{and} \qquad |f^*(X)| \le M. \tag{1}$$

for some constant $0 < M < \infty$. Thus we can assume that the training samples belong to the set $\mathsf{Z} = \mathsf{X} \times [-M, M]$. We will also assume that the class $\mathscr{F}$ is a subset of a suitable reproducing kernel Hilbert space (RKHS) $\mathscr{H}_K$ induced by some Mercer kernel $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$. It will be useful to define

$$C_K \triangleq \sup_{x \in \mathsf{X}} \sqrt{K(x, x)}; \tag{2}$$

we will assume that C_K is finite. The following simple bound will come in handy:

Lemma 1. *For any function $f : \mathsf{X} \to \mathbb{R}$, define the* sup norm

$$\|f\|_\infty \triangleq \sup_{x \in \mathsf{X}} |f(x)|. \tag{3}$$

Then for any $f \in \mathscr{H}_K$ we have

$$\|f\|_\infty \le C_K \|f\|_K.$$

Proof. For any $f \in \mathscr{H}_K$ and $x \in \mathsf{X}$,

$$|f(x)| = |\langle f, K_x \rangle_K| \le \|f\|_K \|K_x\|_K = \|f\|_K \sqrt{K(x,x)},$$

where the first step is by the reproducing kernel property, while the second step is by Cauchy–Schwarz. Taking the supremum of both sides over X, we get (3). □

1 ERM over a ball in RKHS

First, we will look at the simplest case: ERM over a ball in $\mathscr{H}_K$. Thus, we pick the radius $\lambda > 0$ and take

$$\mathscr{F} = \mathscr{F}_\lambda = \left\{ f \in \mathscr{H}_K : \|f\|_K \le \lambda \right\}.$$

The ERM algorithm outputs the predictor

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{F}_\lambda} L_n(f) \equiv \operatorname*{arg\,min}_{f \in \mathscr{F}_\lambda} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2,$$

where $L_n(f)$ denotes, as usual, the empirical loss (in this case, empirical quadratic loss) of f.

Theorem 1. *With probability at least $1 - \delta$,*

$$L(\widehat{f}_n) \le L^*(\mathscr{F}_\lambda) + \frac{16(M + C_K \lambda)^2}{\sqrt{n}} + (M^2 + C_K^2 \lambda^2) \sqrt{\frac{32 \log(1/\delta)}{n}} \tag{4}$$

Proof. First let us introduce some notation. Let us denote the quadratic loss function $(y,u) \mapsto (y-u)^2$ by $\ell(y,u)$, and for any $f : \mathbb{R}^d \to \mathbb{R}$ let

$$\ell \bullet f(x,y) \triangleq \ell(y, f(x)) = (y - f(x))^2$$

Let $\ell \bullet \mathscr{F}_\lambda$ denote the function class $\{\ell \bullet f : f \in \mathscr{F}_\lambda\}$.

Let f_λ^* denote any minimizer of $L(f)$ over $\mathscr{F}_\lambda$, i.e., $L(f_\lambda^*) = L^*(\mathscr{F}_\lambda)$. As usual, we write

$$\begin{aligned}
L(\widehat{f}_n) - L^*(\mathscr{F}_\lambda) &= L(\widehat{f}_n) - L^*(\mathscr{F}_\lambda) \\
&= L(\widehat{f}_n) - L_n(\widehat{f}_n) + L_n(\widehat{f}_n) - L_n(f_\lambda^*) + L_n(f_\lambda^*) - L(f_\lambda^*) \\
&\le 2 \sup_{f \in \mathscr{F}_\lambda} |L_n(f) - L(f)| \\
&= 2 \sup_{f \in \mathscr{F}_\lambda} |P_n(\ell \bullet f) - P(\ell \bullet f)| \\
&= 2 \Delta_n(\ell \bullet \mathscr{F}_\lambda),
\end{aligned} \tag{5}$$

where we have defined the uniform deviation

$$\Delta_n(\ell \bullet \mathscr{F}) \triangleq \sup_{f \in \mathscr{F}} |P_n(\ell \bullet f) - P(\ell \bullet f)|.$$

Next we show that, as a function of the training sample Z^n, $g(Z^n) = \Delta_n(\ell \bullet \mathscr{F}_\lambda)$ has bounded differences. Indeed, for any $1 \le i \le n$, any $z^n \in \mathsf{Z}^n$, and any $z_i' \in \mathsf{Z}$, let $z^n_{(i)}$ denote z^n with the ith coordinate replaced by z_i'. Then

$$\begin{aligned}
\left| g(z^n) - g(z^n_{(i)}) \right| &\le \frac{1}{n} \sup_{f \in \mathscr{F}_\lambda} \left| (y_i - f(x_i))^2 - (y_i' - f(x_i'))^2 \right| \\
&\le \frac{2}{n} \sup_{x \in \mathsf{X}} \sup_{|y| \le M} \sup_{f \in \mathscr{F}_\lambda} \left| y - f(x) \right|^2 \\
&\le \frac{4}{n} \left(M^2 + \sup_{f \in \mathscr{F}_\lambda} \|f\|_\infty^2 \right) \\
&\le \frac{4}{n} \left(M^2 + C_K^2 \lambda^2 \right),
\end{aligned}$$

where the last line is by Lemma 1. Thus, $\Delta_n(\ell \bullet \mathscr{F}_\lambda)$ has the bounded difference property with $c_1 = \ldots = c_n = 4(M^2 + C_K^2\lambda^2)/n$, so McDiarmid's inequality says that, for any $t > 0$,

$$\mathbb{P}\Big(\Delta_n(\ell \bullet \mathscr{F}_\lambda) \ge \mathbb{E}\Delta_n(\ell \bullet \mathscr{F}_\lambda) + t \Big) \le \exp\left(-\frac{nt^2}{8(M^2 + C_K^2\lambda^2)^2} \right).$$

Therefore, letting

$$t = 2(M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}},$$

we see that

$$\Delta_n(\ell \bullet \mathscr{F}_\lambda) \le \mathbb{E}\Delta_n(\ell \bullet \mathscr{F}_\lambda) + 2(M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$. Moreover, by symmetrization we have

$$\mathbb{E}\Delta_n(\ell \bullet \mathscr{F}_\lambda) \le 2\mathbb{E}R_n(\ell \bullet \mathscr{F}_\lambda(Z^n)), \tag{6}$$

where

$$R_n(\ell \bullet \mathscr{F}_\lambda(Z^n)) = \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{f \in \mathscr{F}_\lambda} \left| \sum_{i=1}^n \sigma_i \cdot \ell \bullet f(Z_i) \right| \right] \tag{7}$$

is the Rademacher average of the (random) set

$$\begin{aligned}
\ell \bullet \mathscr{F}_\lambda(Z^n) &= \big\{ (\ell \bullet f(Z_1), \ldots, \ell \bullet f(Z_n)) : f \in \mathscr{F}_\lambda \big\} \\
&= \big\{ ((Y_1 - f(X_1)^2), \ldots, (Y_n - f(X_n))^2) : f \in \mathscr{F}_\lambda \big\}.
\end{aligned}$$

To bound the Rademacher average in (7), we will need to use the contraction principle. To that end, consider the function $\varphi(t) = t^2$. On the interval $[-A, A]$ for some $A > 0$, this function is Lipschitz with constant $2A$, i.e.,

$$|s^2 - t^2| \le 2A|s - t|, \qquad -A \le s, t \le A.$$

Thus, since $|Y_i| \leq M$ and $|f(X_i)| \leq C_K\lambda$ for all $1 \leq i \leq n$, by the contraction principle we can write

$$R_n(\ell \bullet \mathscr{F}_\lambda(Z^n)) \leq \frac{4(M + C_K\lambda)}{n} \mathbb{E}_{\sigma^n}\left[\sup_{f \in \mathscr{F}_\lambda} \left|\sum_{i=1}^n \sigma_i \big(Y_i - f(X_i)\big)\right|\right]. \tag{8}$$

Moreover

$$\begin{aligned}
\mathbb{E}_{\sigma^n}\left[\sup_{f \in \mathscr{F}_\lambda} \left|\sum_{i=1}^n \sigma_i \big(Y_i - f(X_i)\big)\right|\right] &\leq \mathbb{E}_{\sigma^n}\left|\sum_{i=1}^n \sigma_i Y_i\right| + \mathbb{E}_{\sigma^n}\left[\sup_{f \in \mathscr{F}_\lambda} \left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right] \\
&\leq \sqrt{\sum_{i=1}^n Y_i^2} + nR_n(\mathscr{F}_\lambda(Z^n)) \\
&\leq (M + C_K\lambda)\sqrt{n},
\end{aligned} \tag{9}$$

where the first step uses the triangle inequality, the second step uses the result from previous lectures on the expected absolute value of Rademacher sums, and the third step uses (1) and the bound on the Rademacher average over a ball in an RKHS. Combining (6) through (9) (and overbounding (9) slightly), we conclude that

$$\Delta_n(\ell \bullet \mathscr{F}_\lambda) \leq \frac{8(M + C_K\lambda)^2}{\sqrt{n}} + 2(M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}} \tag{10}$$

with probability at least $1 - \delta$. Finally, combining this with (5), we get (4). □

2 Regularized least squares in an RKHS

The observation we had made many times by now is that when the joint distribution of the input-output pair $(X, Y) \in \mathsf{X} \times \mathbb{R}$ is unknown, there is no hope in general to learn the optimal predictor f^* from a finite training sample. Thus, restricting our attention to some hypothesis space $\mathscr{F}$, which is a proper subset of the class of *all* measurable functions $f : \mathsf{X} \to \mathbb{R}$, is a form of *insurance*: If we do not do this, then we can always find some function f that attains zero empirical loss, yet performs spectacularly badly on the inputs outside the training set. When this happens, we say that our learned predictor *overfits*. On the other hand, if our hypothesis space $\mathscr{F}$ consists of well-behaved functions, then it is possible to learn a predictor that achieves a graceful balance between in-sample data fit and out-of-sample generalization. The price we pay is the *approximation error*

$$L^*(\mathscr{F}) - L^* \equiv \inf_{f \in \mathscr{F}} L(f) - \inf_{f : \mathsf{X} \to \mathbb{R}} L(f) \geq 0.$$

In the regression setting, the approximation error can be expressed as

$$L^*(\mathscr{F}) - L^* = \inf_{f \in \mathscr{F}} \|f - f^*\|_{P_X}^2,$$

where $f^*(x) = \mathbb{E}[Y|X = x]$ is the regression function (the MMSE predictor of Y given X).

When seen from this perspective, the use of a restricted hypothesis space $\mathscr{F}$ is a form of *regularization* — a way of guaranteeing that the learned predictor performs well outside the training sample. However, this is not the only way to achieve regularization. In this section, we will analyze another way: *complexity regularization*. In a nutshell, complexity regularization is a modification of the ERM scheme

that allows us to search over a fairly "rich" hypothesis space by adding a *penalty term*. Complexity regularization is a very general technique with wide applicability. We will look at a particular example of complexity regularization over an RKHS and derive a simple bound on its generalization performance.

To set things up, let $\gamma > 0$ be a regularization parameter. Introduce the *regularized* quadratic loss

$$J_\gamma(f) \triangleq L(f) + \gamma \|f\|_K^2$$

and its empirical counterpart

$$J_{n,\gamma}(f) \triangleq L_n(f) + \gamma \|f\|_K^2.$$

Define the functions

$$f_\gamma^* \triangleq \operatorname*{arg\,min}_{f \in \mathscr{H}_K} J_\gamma(f) \tag{11}$$

and

$$\widehat{f}_{n,\gamma} \triangleq \operatorname*{arg\,min}_{f \in \mathscr{H}_K} J_{n,\gamma}(f). \tag{12}$$

We will refer to (12) as the *regularized kernel least squares* (RKLS) algorithm.

Note that the minimization in (11) and (12) takes place in the *entire* RKHS $\mathscr{H}_K$, rather than a subset, say, a ball. However, the addition of the regularization term $\|f\|_K^2$ ensures that the RKLS algorithm does not just select any function $f \in \mathscr{H}_K$ that happens to fit the training data well — instead, it weighs the goodness-of-fit term $L_n(f)$ term against the "complexity" $\|f\|_K^2$, since a very large value of $\|f\|_K^2$ would indicate that f might "wiggle around" a lot and, therefore, overfit the training sample. The regularization parameter $\gamma > 0$ controls the relative importance of the goodness-of-fit and the complexity terms.

We have the following basic bound on the generalization performance of RKLS:

Theorem 2. *With probability at least* $1-\delta$,

$$L(\widehat{f}_{n,\gamma}) - L^* \leq A(\gamma) + \frac{16M^2\left(1+\frac{C_K}{\sqrt{\gamma}}\right)^2}{\sqrt{n}} + 2\left(2M^2 + \frac{C_K^2(M^2+A(\gamma))}{\gamma}\right)\sqrt{\frac{2\log(2/\delta)}{n}} \tag{13}$$

where

$$A(\gamma) \triangleq \inf_{f \in \mathscr{H}_K} \left[L(f) + \gamma \|f\|_K^2\right] - L^*$$

is the regularized approximation error.

Proof. We start with the following lemma:

Lemma 2.

$$L(\widehat{f}_{n,\gamma}) - L^* \leq \delta_n(\widehat{f}_{n,\gamma}) - \delta_n(f_\gamma^*) + A(\gamma), \tag{14}$$

where $\delta_n(f) \triangleq L(f) - L_n(f)$ *for all* f.

Proof. First, an obvious overbounding gives $L(\widehat{f}_{n,\gamma}) - L^* \le J_\gamma(\widehat{f}_{n,\gamma}) - L^*$. Then

$$\begin{aligned}
J_\gamma(\widehat{f}_{n,\gamma}) &= L(\widehat{f}_{n,\gamma}) + \gamma\|\widehat{f}_{n,\gamma}\|_K^2 \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + \underbrace{L_n(\widehat{f}_{n,\gamma}) + \gamma\|\widehat{f}_{n,\gamma}\|_K^2}_{=J_{n,\gamma}(\widehat{f}_{n,\gamma})} \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + J_{n,\gamma}(\widehat{f}_{n,\gamma}) - J_{n,\gamma}(f_\gamma^*) + J_{n,\gamma}(f_\gamma^*) \\
&\le L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + J_{n,\gamma}(f_\gamma^*) \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + L_n(f_\gamma^*) + \gamma\|f_\gamma^*\|_K^2 \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + L_n(f_\gamma^*) - L(f_\gamma^*) + L(f_\gamma^*) + \gamma\|f_\gamma^*\|_K^2 \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + L_n(f_\gamma^*) - L(f_\gamma^*) + J_\gamma(f_\gamma^*).
\end{aligned}$$

This gives

$$\begin{aligned}
L(\widehat{f}_{n,\gamma}) - L^* &\le L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + L_n(f_\gamma^*) - L(f_\gamma^*) + J_\gamma(f_\gamma^*) - L^* \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + L_n(f_\gamma^*) - L(f_\gamma^*) + \inf_{f\in\mathscr{H}}\left[L(f) + \gamma\|f\|_K^2\right] - L^* \\
&= L(\widehat{f}_{n,\gamma}) - L_n(\widehat{f}_{n,\gamma}) + L_n(f_\gamma^*) - L(f_\gamma^*) + A(\gamma),
\end{aligned}$$

and we are done. □

Lemma 2 shows that the excess loss of the regularized empirical loss minimizer $\widehat{f}_{n,\gamma}$ is bounded from above by the sum of three terms: the deviation $\delta_n(\widehat{f}_{n,\gamma}) \triangleq L(\widehat{f}_{n,\gamma}) - L(\widehat{f}_{n,\gamma})$ of $\widehat{f}_{n,\gamma}$ itself, the (negative) deviation $-\delta_n(f_\gamma^*) \triangleq L_n(f_\gamma^*) - L(f_\gamma^*)$ of the best regularized predictor f_γ^*, and the approximation error $A(\gamma)$. To prove Theorem 2, we will need to obtain high-probability bounds on the two deviation terms. To that end, we need a lemma:

Lemma 3. *The functions f_γ^* and $\widehat{f}_{n,\gamma}$ satisfy the bounds*

$$\|f_\gamma^*\|_\infty \le C_K\sqrt{\frac{A(\gamma)}{\gamma}}. \tag{15}$$

and

$$\|\widehat{f}_{n,\gamma}\|_K \le \frac{M}{\sqrt{\gamma}} \qquad \textit{with probability one} \tag{16}$$

respectively.

Proof. To prove (15), we use the fact that

$$A(\gamma) = L(f_\gamma^*) - L^* + \gamma\|f_\gamma^*\|_K^2 \ge \gamma\|f_\gamma^*\|_K^2,$$

which gives $\|f_\gamma^*\|_K \le \sqrt{A(\gamma)/\gamma}$. From this and from (3) we obtain (15).

For (16), we use the fact that $\widehat{f}_{n,\gamma}$ minimizes $J_{n,\gamma}(f)$ over all f. In particular,

$$J_{n,\gamma}(\widehat{f}_{n,\gamma}) = L_n(\widehat{f}_{n,\gamma}) + \gamma\|\widehat{f}_{n,\gamma}\|_K^2 \le J_{n,\gamma}(0) = \frac{1}{n}\sum_{i=1}^n Y_i^2 \le M^2 \qquad \text{w.p. 1},$$

where the last step follows from (1). Rearranging and using the fact that $L_n(f) \ge 0$ for all f, we get (16). □

Now we are ready to bound $\delta_n(\widehat{f}_{n,\gamma})$. For any $R \geq 0$, let $\mathscr{F}_R = \{f \in \mathscr{H}_K : \|f\|_K \leq R\}$ denote the zero-centered ball of radius R in the RKHS $\mathscr{H}_K$. Then Lemma 3 says that $\widehat{f}_{n,\gamma} \in \mathscr{F}_{M/\sqrt{\gamma}}$ with probability one. Therefore, with probability one we have

$$\begin{aligned}
\delta_n(\widehat{f}_{n,\gamma}) &= \delta_n(\widehat{f}_{n,\gamma}) \cdot \mathbf{1}_{\{\widehat{f}_{n,\gamma} \in \mathscr{F}_{M/\sqrt{\gamma}}\}} \\
&\leq \left|\delta_n(\widehat{f}_{n,\gamma})\right| \cdot \mathbf{1}_{\{\widehat{f}_{n,\gamma} \in \mathscr{F}_{M/\sqrt{\gamma}}\}} \\
&\leq \underbrace{\sup_{f \in \mathscr{F}_{M/\sqrt{\gamma}}} \left|\delta_n(f)\right|}_{\equiv \Delta_n(\ell \bullet \mathscr{F}_{M/\sqrt{\gamma}})} \cdot \mathbf{1}_{\{\widehat{f}_{n,\gamma} \in \mathscr{F}_{M/\sqrt{\gamma}}\}} \\
&\leq \Delta_n(\ell \bullet \mathscr{F}_{M/\sqrt{\gamma}}).
\end{aligned}$$

Consequently, we can carry out the same analysis as in the proof of Theorem 1. First of all, the function $g(Z^n) = \Delta_n(\ell \bullet \mathscr{F}_{M/\sqrt{\gamma}})$ has bounded differences with

$$c_1 = \ldots = c_n \leq \frac{4}{n}\left(M^2 + \sup_{f \in \mathscr{F}_{M/\sqrt{\gamma}}} \|f\|_\infty^2\right) \leq \frac{4M^2}{n}\left(1 + \frac{C_K^2}{\gamma}\right)$$

where the last step uses (16) and Lemma 1. Therefore, with probability at least $1 - \delta/2$,

$$\delta_n(\widehat{f}_{n,\gamma}) \leq \Delta_n(\ell \bullet \mathscr{F}_{M/\sqrt{\gamma}}) \leq \frac{8M^2\left(1 + \frac{C_K}{\sqrt{\gamma}}\right)^2}{\sqrt{n}} + 2M^2\left(1 + \frac{C_K^2}{\gamma}\right)\sqrt{\frac{2\log(2/\delta)}{n}}, \tag{17}$$

where the second step follows from (10) with δ replaced by $\delta/2$ and with $\lambda = M/\sqrt{\gamma}$.

It remains to bound $\delta_n(f_\gamma^*)$. This is, actually, much easier, since we are dealing with a single data-independent function. In particular, note that we can write

$$\delta_n(f_\gamma^*) = \frac{1}{n}\sum_{i=1}^n (Y_i - f_\gamma^*(X_i))^2 - \mathbb{E}\left[(Y - f_\gamma^*(X))^2\right] = \frac{1}{n}\sum_{i=1}^n U_i,$$

where $U_i \triangleq (Y_i - f_\gamma^*(X_i))^2 - \mathbb{E}\left[(Y - f_\gamma^*(X))^2\right], 1 \leq i \leq n$, are i.i.d. random variables with $\mathbb{E}U_i = 0$ and

$$|U_i| \leq \sup_{y \in [-M,M]} \sup_{x \in \mathsf{X}} \left(y - f_\gamma^*(x)\right)^2 \leq 2(M^2 + \|f_\gamma^*\|_\infty^2) \leq 2\left(M^2 + \frac{C_K^2 A(\gamma)}{\gamma}\right)$$

with probability one, where we have used (1) and (15). We can therefore use Hoeffding's inequality to write, for any $t \geq 0$,

$$\mathbb{P}\left(-\delta_n(f_\gamma^*) \geq t\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n U_i \leq -t\right) \leq \exp\left(-\frac{nt^2}{8\left(M^2 + C_K^2 A(\gamma)/\gamma\right)^2}\right)$$

This implies that

$$-\delta_n(f_\gamma^*) \leq 2\left(M^2 + \frac{C_K^2 A(\gamma)}{\gamma}\right)\sqrt{\frac{2\log(2/\delta)}{n}} \tag{18}$$

with probability at least $1 - \delta/2$. Combining (17) and (18) with (14), we get (13). □