Notes on the Offset Rademacher Complexity Paper [Liang et al., 2015]

Tomas Vaškevičius

October 25, 2019

1 Set Up

We observe n data points $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ sampled i.i.d. from some unknown distribution P. Let $\mathcal{H}$ be a closed and convex *hypothesis class* containing functions mapping $\mathcal{X}$ to $\mathbb{R}$. Further, let $\ell(\hat{y}, y) = (\hat{y} - y)^2$ denote the square loss function and for every $h \in \mathcal{H}$ let $\ell_h(x, y) = \ell(h(x), y)$. We measure the quality of a hypothesis h with respect to the unknown distribution P with the *population loss* defined as

$$P\ell_h = \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(h(X), Y)]\,.$$

The best function in class $\mathcal{H}$ will be denoted by $h^\star$ and given by

$$h^\star = \mathop{\mathrm{argmin}}_{h\in\mathcal{H}} P\ell_h$$

assuming that the above argmin exists. For an estimator $\hat{h}$ we are interested in upper-bounding the *excess risk* given by

$$\mathcal{E}(\hat{h}) = P\ell_{\hat{h}} - P\ell_{h^\star}.$$

Since the distribution P is unknown we cannot evaluate any of the above expressions. Instead, we can replace the distribution P with its empirical counterpart

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i,y_i}$$

where δ_{x_i,y_i} denotes point mass on the sample (x_i, y_i). We will study the *empirical risk minimization* (ERM) estimator $\hat{h}$ given by

$$\hat{h} = \mathop{\mathrm{argmin}}_{h\in H} P_n\ell_h = \mathop{\mathrm{argmin}}_{h\in H} \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2.$$

To summarize, we will study the *excess risk* of an *empirical risk minimization* algorithm performed over a closed and *convex* class $\mathcal{H}$. In particular, we will present a novel complexity measure of $\mathcal{H}$ introduced by Liang et al. [2015] which is based on a modification of

Rademacher complexities and is applicable to estimators obtained by performing empirical risk minimization.

Remark 1. *We will only present a subset of results presented in Liang et al. [2015]. In what follows, we will only consider excess loss in expectation and assume boundedness of the functions and observations. See the paper for extensions to high-probability results, relaxation of boundedness assumptions and relaxation of convexity of $\mathcal{H}$.*

2 A Motivating Example

Consider the following problem of bounding excess loss of bounded linear predictors.

Problem 1 (Excess Risk of Bounded Linear Regression)**.** *Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$, $\mathcal{Y} = [-Y, Y]$ and $\mathcal{H} = \{\langle x, \cdot\rangle : x \in R^d, \|x\|_2 \leq B\}$. We want to obtain an upper-bound on*

$$\mathbb{E}\left[P\ell_{\hat{h}} - P\ell_{h^\star}\right].$$

A classical approach for controlling the excess risk goes as follows. First consider the decomposition

$$P\ell_{\hat{h}} - P\ell_{h^\star} = (P\ell_{\hat{h}} - P_n\ell_{\hat{h}}) + (P_n\ell_{\hat{h}} - P_n\ell_{h^\star}) + (P_n\ell_{h^\star} - P\ell_{h^\star}). \tag{1}$$

Since the second term is always non-positive and the third term is equal to 0 in expectation, we obtain

$$\mathbb{E}\left[P\ell_{\hat{h}} - P\ell_{h^\star}\right] \leq \mathbb{E}\left[(P - P_n)\ell_{\hat{h}}\right] \leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}(P - P_n)\ell_{\hat{h}}\right].$$

Hence the expected excess risk is upper-bounded by the supremum of the empirical process $(P - P_n)$ indexed by $\mathcal{H}$.

Let $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of i.i.d. $\{-1, +1\}$ (Rademacher) random variables independent of the data. Let $R_n \coloneqq \frac{1}{n}\sum_{i=1}^n \varepsilon_i \delta_{(x_i,y_i)}$ so that R_n is a *Rademacher process* indexed by $\mathcal{H}$. Denote $\ell \circ \mathcal{H} = \{\ell_h : h \in \mathcal{H}\}$. A classical symmetrization argument then shows that

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}(P - P_n)\ell_{\hat{h}}\right] \leq 2 \mathop{\mathbb{E}}_{(x_i,y_i),\varepsilon_i}\left[\sup_{h\in\mathcal{H}} R_n\ell_{\hat{h}}\right] =: 2\mathfrak{R}(\ell \circ \mathcal{H})$$

where $\mathfrak{R}(\ell \circ \mathcal{H})$ is called the *Rademacher complexity* of $(\ell \circ \mathcal{H}, P)$. Moving from $P - P_n$ to R_n allows to work conditionally on data, that is, with finite dimensional objects which is a considerable simplification in most settings. The cost of moving from $P - P_n$ to R_n is only a multiplicative factor.

Denote $\mathop{\mathbb{E}}_{\varepsilon}[\cdot] = \mathop{\mathbb{E}}_{\varepsilon}[\cdot \mid (X_i, Y_i)]$. Going back to problem 1 we can now upper-bound the excess risk

as follows:

$$
\begin{aligned}
\mathbb{E}\big[P\ell_{\hat{h}} - P\ell_{h^\star}\big] &\le 2\mathfrak{R}(\mathcal{H}) \\
&= 2\mathbb{E}\left[\mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell_h(x_i,y_i)\right]\right] \\
&\le 4(Y+B)\,\mathbb{E}\left[\mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)\right]\right] && \text{since } \ell(\cdot,y) \text{ is } 2(Y+B)\text{-Lipschitz} \\
&= 4(Y+B)\mathbb{E}\left[\mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{\|w\|_2\le B}\frac{1}{n}\left\langle w, \sum_{i=1}^{n}\varepsilon_i x_i\right\rangle\right]\right] \\
&\le 4(Y+B)B\,\mathbb{E}\left[\mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n}\left\|\sum_{i=1}^{n}\varepsilon_i x_i\right\|_2\right]\right] && \text{by Cauchy-Schwarz inequality} \\
&\le 4(Y+B)B\,\mathbb{E}\left[\mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\left\|\sum_{i=1}^{n}\varepsilon_i x_i\right\|_2^2\right]^{1/2}\right] && \text{by Jensen's inequality} \\
&= 4(Y+B)B\,\mathbb{E}\left[\mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\sum_{i=1}^{n}\|x_i\|_2^2\right]^{1/2}\right] \\
&\le \frac{4(Y+B)B}{\sqrt{n}}. && (2)
\end{aligned}
$$

The below bound scales as $\Omega(1/\sqrt{n})$. However, it is possible to obtain bounds scaling as $O(1/n)$, however, with different dependence on the constants B and Y and linear dependence on d as shown in Shamir [2015].

3 Offset Rademacher Complexity

The sub-optimality of the bound given in Equation (2) comes from inability to fully take advantage of $\hat{h}$ being an empirical risk minimizer. While in Equation (1) we exploit the fact that $P_n\ell_{\hat{h}} - P_n\ell_{h^\star} \le 0$, when controlling the supremum of $P - P_n$ over $\ell \circ \mathcal{H}$ we do not take into account, that our algorithm will never pick hypotheses with large empirical error. In particular, our algorithm should implicitly only consider subsets of $\mathcal{H}$ which decrease as the sample size n increases. This is achieved by *offset Rademacher complexity* defined below[1] defined below.

Definition 1 (Offset Rademacher Complexity)**.** *We will call* $2R_nh - cP_nh^2$ *an* offset Rademacher process *(with parameter* $c > 0$*) indexed by* $\mathcal{H}$*. The* offset Rademacher complexity *of a class* $\mathcal{H}$

[1] For an alternative approach (*local Rademacher complexities*) developed in 2000-2005 see Bartlett et al. [2005] and references therein. See also a closely related paper Mendelson [2014] fixing boundedness and correct scaling with respect to the noise level issues.

is given by the expected supremum of the offset Rademacher process as follows:

$$\mathfrak{R}^c(\mathcal{H}) = \mathop{\mathbb{E}}_{(x_i,y_i),\varepsilon_i}\left[\sup_{h\in\mathcal{H}}(2R_n h - P_n h^2)\right] = \mathop{\mathbb{E}}_{(x_i,y_i),\varepsilon_i}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\left\{2\varepsilon_i h(x_i) - c h(x_i)^2\right\}\right].$$

Note that for $c = 0$ we recover the Rademacher process/complexity. When $c > 0$ due to non-negativity of quadratic terms, the offset Rademacher process is point-wise upper-bounded by the usual Rademacher process. Finally, note that the term $-cP_nh^2$ intuitively penalizes hypotheses with large empirical error, thus excluding them from the consideration in our upper-bounds. The following theorem shows that $\mathfrak{R}^c(\mathcal{H})$ can be used to upper-bound the excess risk, which is the main result of these notes.

Theorem 1. *Let $\mathcal{H}$ be a closed and convex hypothesis class. Further, suppose that $\mathcal{Y} = [-Y, Y]$ and our functions are bounded on $\mathcal{X}$, that is, $\sup_{h\in\mathcal{H},x\in\mathcal{X}} h(x) \le B$. Denote $\mathcal{H} - h^\star = \{h - h^\star : h \in \mathcal{H}\}$. Then, the following holds for an empirical risk minimizer $\hat{h}$:*

$$\mathbb{E}\left[P\ell_{\hat{h}} - P\ell_{h^\star}\right] \le (8B + 2Y)\mathfrak{R}^c(\mathcal{H} - h^\star)$$

with $c = \frac{1}{12(B+Y)}$.

We defer the proof to Section 4 and come back to problem 1. To obtain a bound on the excess risk, we now simply need to upper-bound the offset Rademacher complexity. This is achieved by the following lemma.

Lemma 1. *Consider the setting of problem 1. Then, for any $c > 0$ we have*

$$\mathfrak{R}^c(\mathcal{H} - h^\star) \le \frac{d}{cn}.$$

Proof. Condition on the data and let $\frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T} = \Sigma$, which will be assumed to be invertible. Then

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{h\in\mathcal{H}}(R_n(h - h^\star) - cP_n(h - h^\star)^2)\right] \\
&\le \mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{w\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\left\{\varepsilon_i\left\langle w, x_i\right\rangle - c\left\langle w, x_i\right\rangle^2\right\}\right] && \text{sup over a larger class} \\
&= \mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{w\in\mathbb{R}^d}\left\{\left\langle w, \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_i\right\rangle - w^\mathsf{T}(c\Sigma)\,w\right\}\right]
\end{aligned}$$

. The expression in the curly brackets can be seen to equal the Fenchel-Legendre transform of the function $f(w) = w^\mathsf{T}(c\Sigma)w$ applied to the vector $\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i$. The Fenchel-Legendre

transform of f is equal to $f^\star(w) = w^\mathsf{T}(c\Sigma)^{-1}w$. Hence we can carry on as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\varepsilon}\left[\sup_{w\in\mathbb{R}^d}\left\{\left\langle w, \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\right\rangle - w^\mathsf{T}\left(c\Sigma\right)w\right\}\right] \\
&= \mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\left(\sum_{i=1}^n \varepsilon_i x_i\right)^\mathsf{T}(c\Sigma)^{-1}\left(\sum_{i=1}^n \varepsilon_i x_i\right)\right] \\
&= \mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\sum_{i=1}^n x_i^\mathsf{T}(c\Sigma)^{-1}x_i\right] \\
&= \mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\sum_{i=1}^n \operatorname{trace}\left(x_i^\mathsf{T}(c\Sigma)^{-1}x_i\right)\right] \\
&= \mathop{\mathbb{E}}_{\varepsilon}\left[\frac{1}{n^2}\operatorname{trace}\left((c\Sigma)^{-1}n\Sigma\right)\right] \\
&= \frac{d}{nc}.
\end{aligned}$$

□

Theorem 1 together with Lemma 1 yield an $O(1/n)$ upper-bound for problem 1 as shown in the following corollary.

Corollary 1. *Combining results of Theorem 1 and Lemma 1 we obtain*

$$\begin{aligned}
\mathbb{E}\left[P\ell_{\hat{h}} - P\ell_{h^\star}\right] &\leq (8B+2Y)\mathfrak{R}^{1/(12(B+Y))}(\mathcal{H}-h^\star) \\
&\leq 96\frac{(B+Y)^2 d}{n}.
\end{aligned}$$

Remark 2. *Despite obtaining a bound which scales as $O(1/n)$ the above result is sub-optimal in terms of scaling with B and Y. As shown in Shamir [2015], the correct scaling (possibly up to a $\log(1+n/d)$ factor in the middle term) is*

$$\Theta\left(Y^2 \wedge \frac{B^2+Y^2 d}{n} \wedge \frac{YB}{\sqrt{n}}\right).$$

4 Proofs

This section is dedicated for proving Theorem 1. We split the proof into two parts. In the first part, we show how to simultaneously exploit properties of the ℓ_2 loss and empirical risk minimization improving upon the approach taken in Section 2 and deriving an unsymmetrized version of the offset Rademacher process. In the second part, we show how standard symmetrization and contraction arguments can be used to prove Theorem 1

4.1 Exploiting Properties of ERM and Square Loss

When conditioned on data, we will view functions as n-dimensional vectors, that is, $h = (h(x_1), \ldots, h(x_n))^\mathsf{T}$ and $y = (y_1, \ldots, y_n)$. Further, we denote $\|h\|_n^2 = \frac{1}{n}\sum_{i=1}^n h(x_i)^2$.

Lemma 2. *Let $\hat{h}$ be an empirical risk minimizer of a closed and convex class $\mathcal{H}$ with respect to the square loss ℓ. Then, for any $h \in \mathcal{H}$, we have*

$$P_n\ell_h - P_n\ell\hat{h} \geq \|h - \hat{h}\|_n^2.$$

Proof. Note that conditionally on the data, $\hat{h}$ is a projection of y on a closed and convex set $\mathcal{H}$ with respect to the ℓ_2 distance:

$$\hat{y} = \operatorname*{argmin}_{h\in\mathcal{H}} \|h - y\|_2^2.$$

By first order optimality conditions the following holds for any $h \in \mathcal{H}$:

$$\left\langle h - \hat{h}, \nabla_{\hat{h}} \left\|\hat{h} - y\right\|_2^2 \right\rangle \geq 0$$

which implies that

$$\begin{aligned}
P_n\ell_h - P_n\ell\hat{h} - \|h - \hat{h}\|_n^2 &= \|h - y\|_n^2 - \|\hat{h} - y\|_n^2 - \|h - \hat{h}\|_n^2 \\
&= \frac{1}{n}\left(\|h - \hat{h} + \hat{h} - y\|_2^2 - \|\hat{h} - y\|_2^2 - \|h - \hat{h}\|_2^2\right) \\
&= \frac{1}{n}\left(2\left\langle h - \hat{h}, \hat{h} - y\right\rangle\right) \\
&= \frac{1}{n}\left\langle h - \hat{h}, \nabla_{\hat{h}} \left\|\hat{h} - y\right\|_2^2 \right\rangle \\
&\geq 0
\end{aligned}$$

which completes our proof. □

In contrast to the approach taken in Equation (1) for the excess risk decomposition, instead of simply removing the term $P_n(\ell_{\hat{h}} - \ell_{h^*})$ we can subtract a non-negative term $\|\hat{h} - h^\star\|_n^2 = P_n(\hat{h} - h^\star)^2$. This observation gives rise to the unsymmetrized version of the offset Rademacher complexity as shown in the next lemma.

Lemma 3. *The following deterministic upper-bound holds for the excess risk of an empirical risk minimization algorithm:*

$$P\ell_{\hat{h}} - P\ell_{h^\star} \leq 2(P - P_n)(\hat{h} - h^\star)(h^\star - Y) + P(\hat{h} - h^\star)^2 - 2P_n(\hat{h} - h^\star)^2.$$

Proof. Applying Lemma 2 to the middle term in the decomposition given in Equation (1) we obtain

$$\begin{aligned} P\ell_{\hat{h}} - P\ell_{h^\star} &= (P\ell_{\hat{h}} - P_n\ell_{\hat{h}}) + (P_n\ell_{\hat{h}} - P_n\ell_{h^\star}) + (P_n\ell_{h^\star} - P\ell_{h^\star}) \\ &\leq (P - P_n)(\hat{h}(X) - Y)^2 - (P - P_n)(h^\star - Y)^2 - P_n(\hat{h} - h^\star)^2 \\ &= (P - P_n)((\hat{h} - h^\star)^2 + 2(\hat{h} - h^\star)(h^\star - Y)) - P_n(\hat{h} - h^\star)^2 \\ &= 2(P - P_n)(\hat{h} - h^\star)(h^\star - Y) + P(\hat{h} - h^\star)^2 - 2P_n(\hat{h} - h^\star)^2. \end{aligned}$$

□

4.2 Proof of Theorem 1

Finishing the proof relies on classical symmetrization and contraction techniques applied to the result of Lemma 3. While the classical symmetrization and contraction results do not directly apply to the offset process given in Lemma 3, the proofs of symmetrization and contraction "go through" when applied to the offset process.

By Lemma 3 we have

$$\begin{aligned} \mathbb{E}\big[P\ell_{\hat{h}} - P\ell_{h^\star}\big] &\leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{2(P - P_n)(\hat{h} - h^\star)(h^\star - Y) + P(\hat{h} - h^\star)^2 - 2P_n(\hat{h} - h^\star)^2\right\}\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2(P - P_n)f(h^\star - Y) + Pf^2 - 2P_nf^2\right\}\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2(P - P_n)f(h^\star - Y) - \frac{1}{2}P_nf^2 + Pf^2 - \frac{3}{2}P_nf^2\right\}\right] \\ &\leq \underbrace{\mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2(P - P_n)f(h^\star - Y) - \frac{1}{4}Pf^2 - \frac{1}{4}P_nf^2\right\}\right]}_{T_1} \\ &\quad + \underbrace{\mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\frac{5}{4}Pf^2 - \frac{7}{4}P_nf^2\right\}\right]}_{T_2}. \end{aligned}$$

We will now deal with the terms T_1 and T_2 separately. Combining the obtained upper-bounds yields the desired result.

4.2.1 Bounding T_1

We introduce an independent copy (x_i', y_i') of the data and let $P_n' = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i', y_i')}$. Let ε_i be i.i.d. Rademacher random variables independent of everything else. Finally, we will denote

$h^\star(x_i) - y_i = \xi_i$ and $h^\star(x_i) - y_i = \xi_i'$ We can then symmetrize T_1 as follows

$$
\begin{aligned}
T_1 &= \mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\mathop{\mathbb{E}}_{(x',y')}\left[2(P_n'-P_n)f(h^\star-Y) - \frac{1}{4}P'f^2 - \frac{1}{4}P_nf^2\right]\right\}\right] \\
&\le \mathop{\mathbb{E}}_{(x_i,y_i),(x_i',y_i')}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2(P_n'-P_n)f(h^\star-Y) - \frac{1}{4}P_n'f^2 - \frac{1}{4}P_nf^2\right\}\right] \\
&= \mathop{\mathbb{E}}_{(x_i,y_i),(x_i',y_i'),\varepsilon}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(x_i')\xi_i' - f(x_i)\xi_i) - \frac{1}{4}P_n'f^2 - \frac{1}{4}P_nf^2\right\}\right] \\
&\le 2\mathop{\mathbb{E}}_{(x_i,y_i),\varepsilon}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\frac{1}{n}\sum_{i=1}^n 2\varepsilon_i f(x_i)\xi_i - \frac{1}{4}P_nf^2\right\}\right]
\end{aligned}
$$

Recall that $|\xi_i| = |h^\star(x_i) - y_i| \le B + Y$. We can now use Talagrand's contraction argument to get rid of the bounded multipliers ξ_i. To do so, we compute the expected supremum with respect to ε_1 conditionally data and $\varepsilon_2, \ldots, \varepsilon_n$ as follows:

$$
\begin{aligned}
&2\mathop{\mathbb{E}}_{\varepsilon_1}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\frac{1}{n}\sum_{i=1}^n 2\varepsilon_i f(x_i)\xi_i - \frac{1}{4}P_nf^2\right\}\right] \\
&= \sup_{f\in\mathcal{H}-h^\star}\left\{2f(x_1)\xi_1 + \frac{1}{n}\sum_{i=2}^n 2\varepsilon_i f(x_i)\xi_i - \frac{1}{4}P_nf^2\right\} \\
&\qquad + \sup_{g\in\mathcal{H}-h^\star}\left\{-2g(x_1)\xi_1 + \frac{1}{n}\sum_{i=2}^n 2\varepsilon_i g(x_i)\xi_i - \frac{1}{4}P_ng^2\right\} \\
&= \sup_{f,g\in\mathcal{H}-h^\star}\left\{2\xi_1(f(x_1)-g(x_1)) + \frac{1}{n}\sum_{i=2}^n 2\varepsilon_i(f(x_i)+g(x_i))\xi_i - \frac{1}{4}P_n(f^2+g^2)\right\} \\
&\le \sup_{f,g\in\mathcal{H}-h^\star}\left\{2(B+Y)|f(x_1)-g(x_1)| + \frac{1}{n}\sum_{i=2}^n 2\varepsilon_i(f(x_i)+g(x_i))\xi_i - \frac{1}{4}P_n(f^2+g^2)\right\} \\
&= \sup_{f,g\in\mathcal{H}-h^\star}\left\{2(B+Y)(f(x_1)-g(x_1)) + \frac{1}{n}\sum_{i=2}^n 2\varepsilon_i(f(x_i)+g(x_i))\xi_i - \frac{1}{4}P_n(f^2+g^2)\right\} \\
&= 2\mathop{\mathbb{E}}_{\varepsilon_1}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2(B+Y)\varepsilon_1 f(x_1) + \frac{1}{n}\sum_{i=2}^n 2\varepsilon_i f(x_i)\xi_i - \frac{1}{4}P_nf^2\right\}\right].
\end{aligned}
$$

Repeating the same argument for $\varepsilon_2, \ldots, \varepsilon_n$, we obtain

$$
T_1 \le 2(B+Y)\mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{2R_nf - \frac{1}{4(B+Y)}P_nf^2\right\}\right]. \tag{3}
$$

4.2.2 Bounding T_2

We begin by symmetrization similarly as before

$$\begin{aligned} T_2 &= \mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\frac{5}{4}Pf^2 - \frac{7}{4}P_nf^2\right\}\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\frac{6}{4}(P-P_n)f^2 - \frac{1}{4}Pf^2 - \frac{1}{4}P_nf^2\right\}\right] \\ &\le 2\mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{\frac{6}{4}R_nf^2 - \frac{1}{4}P_nf^2\right\}\right]. \end{aligned}$$

Further, noting that for any $f, g \in \mathcal{H} - h^\star$ and any x we have

$$\left|f(x)^2 - g(x)^2\right| \le |(f(x) - g(x)| \cdot 4B$$

and following the same contraction argument as before, we obtain

$$\begin{aligned} T_2 &\le 2\mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{3B\cdot 2R_nf - \frac{1}{4}P_nf^2\right\}\right] \\ &= 6B\mathbb{E}\left[\sup_{f\in\mathcal{H}-h^\star}\left\{3B\cdot 2R_nf - \frac{1}{12B}P_nf^2\right\}\right] \end{aligned} \tag{4}$$

4.3 Completing the Proof

We are done immediately by Equations (3) and (4).

References

P. L. Bartlett, O. Bousquet, S. Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

O. Shamir. The sample complexity of learning linear predictors with the squared loss. *The Journal of Machine Learning Research*, 16(1):3475–3486, 2015.