Learning with Square Loss: Localization through Offset Rademacher Complexity

Tengyuan Liang
Department of Statistics, The Wharton School, University of Pennsylvania

Alexander Rakhlin
Department of Statistics, The Wharton School, University of Pennsylvania

Karthik Sridharan
Department of Computer Science, Cornell University

Abstract

We consider regression with square loss and general classes of functions without the boundedness assumption. We introduce a notion of offset Rademacher complexity that provides a transparent way to study localization both in expectation and in high probability. For any (possibly non-convex) class, the excess loss of a two-step estimator is shown to be upper bounded by this offset complexity through a novel geometric inequality. In the convex case, the estimator reduces to an empirical risk minimizer. The method recovers the results of (Rakhlin et al., 2015) for the bounded case while also providing guarantees without the boundedness assumption.

1. Introduction

Determining the finite-sample behavior of risk in the problem of regression is arguably one of the most basic problems of Learning Theory and Statistics. This behavior can be studied in substantial generality with the tools of empirical process theory. When functions in a given convex class are uniformly bounded, one may verify the so-called "Bernstein condition." The condition—which relates the variance of the increments of the empirical process to their expectation—implies a certain localization phenomenon around the optimum and forms the basis of the analysis via *local Rademacher complexities*. The technique has been developed in (Koltchinskii and Panchenko, 2000; Koltchinskii, 2011; Bousquet et al., 2002; Bartlett et al., 2005; Bousquet, 2002), among others, based on Talagrand's celebrated concentration inequality for the supremum of an empirical process.

In a recent pathbreaking paper, Mendelson (2014a) showed that a large part of this heavy machinery is not necessary for obtaining tight upper bounds on excess loss, even—and especially—if functions are unbounded. Mendelson observed that only one-sided control of the tail is required in the deviation inequality, and, thankfully, it is the tail that can be controlled under very mild assumptions.

In a parallel line of work, the search within the online learning setting for an analogue of "localization" has led to a notion of an "offset" Rademacher process (Rakhlin and Sridharan, 2014), yielding—in a rather clean manner—optimal rates for minimax regret in online supervised learning. It was also shown that the supremum of the offset process is a lower bound on the minimax value, thus establishing its intrinsic nature. The present paper blends the ideas of Mendelson (2014a) and Rakhlin and Sridharan (2014). We introduce the notion of an offset Rademacher process for i.i.d.

data and show that the supremum of this process upper bounds (both in expectation and in high probability) the excess risk of an empirical risk minimizer (for convex classes) and a two-step Star estimator of Audibert (2007) (for arbitrary classes). The statement holds under a weak assumption even if functions are not uniformly bounded.

The offset Rademacher complexity provides an intuitive alternative to the machinery of local Rademacher averages. Let us recall that the Rademacher process indexed by a function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as a stochastic process $g \mapsto \frac{1}{n} \sum_{t=1}^{n} \epsilon_t g(x_t)$ where $x_1, \ldots, x_n \in \mathcal{X}$ are held fixed and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables. We define the offset Rademacher process as a stochastic process

$$g \mapsto \frac{1}{n} \sum_{t=1}^{n} \epsilon_t g(x_t) - cg(x_t)^2$$

for some $c \geq 0$. The process itself captures the notion of localization: when g is large in magnitude, the negative quadratic term acts as a compensator and "extinguishes" the fluctuations of the term involving Rademacher variables. The supremum of the process will be termed *offset Rademacher complexity*, and one may expect that this complexity is of a smaller order than the classical Rademacher averages (which, without localization, cannot be better than the rate of $n^{-1/2}$).

The self-modulating property of the offset complexity can be illustrated on the canonical example of a linear class $\mathcal{G} = \{x \mapsto w^{\mathsf{T}} x : w \in \mathbb{R}^p\}$, in which case the offset Rademacher complexity becomes

$$\frac{1}{n} \sup_{w \in \mathbb{R}^p} \left\{ w^{\mathsf{T}} \left(\sum_{t=1}^{n} \epsilon_t x_t \right) - c\|w\|_{\Sigma}^2 \right\} = \frac{1}{4cn} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_{\Sigma^{-1}}^2$$

where $\Sigma = \sum_{t=1}^{n} x_t x_t^{\mathsf{T}}$. Under mild conditions, the above expression is of the order $\mathcal{O}(p/n)$ in expectation and in high probability — a familiar rate achieved by the ordinary least squares, at least in the case of a well-specified model. We refer to Section 6 for the precise statement for both well-specified and misspecified case.

Our contributions can be summarized as follows. First, we show that offset Rademacher complexity is an upper bound on excess loss of the proposed estimator, both in expectation and in deviation. We then extend the chaining technique to quantify the behavior of the supremum of the offset process in terms of covering numbers. By doing so, we recover the rates of aggregation established in (Rakhlin et al., 2015) and, unlike the latter paper, the present method does not require boundedness (of the noise and functions). We provide a lower bound on minimax excess loss in terms of offset Rademacher complexity, indicating its intrinsic nature for the problems of regression. While our in-expectation results for bounded functions do not require any assumptions, the high probability statements rest on a lower isometry assumption that holds, for instance, for subgaussian classes. We show that offset Rademacher complexity can be further upper bounded by the fixed-point complexities defined by Mendelson Mendelson (2014a). We conclude with the analysis of ordinary least squares.

2. Problem Description and the Estimator

Let $\mathcal{F}$ be a class of functions on a probability space $(\mathcal{X}, P_X)$. The response is given by an unknown random variable Y, distributed jointly with X according to $P = P_X \times P_{Y|X}$. We observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ distributed i.i.d. according to P and aim to construct an estimator $\widehat{f}$ with

small excess loss $\mathcal{E}(\widehat{f})$, where

$$\mathcal{E}(g) \triangleq \mathbb{E}(g - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2 \tag{1}$$

and $\mathbb{E}(f - Y)^2 = \mathbb{E}(f(X) - Y)^2$ is the expectation with respect to (X, Y). Let $\widehat{\mathbb{E}}$ denote the empirical expectation operator and define the following two-step procedure:

$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathbb{E}}(f(X) - Y)^2, \quad \widehat{f} = \operatorname*{argmin}_{f \in \operatorname{star}(\mathcal{F}, \widehat{g})} \widehat{\mathbb{E}}(f(X) - Y)^2 \tag{2}$$

where $\operatorname{star}(\mathcal{F}, g) = \{\lambda g + (1 - \lambda)f : f \in \mathcal{F}, \lambda \in [0, 1]\}$ is the star hull of $\mathcal{F}$ around g. (we abbreviate $\operatorname{star}(\mathcal{F}, 0)$ as $\operatorname{star}(\mathcal{F})$.) This two-step estimator was introduced (to the best of our knowledge) by Audibert (2007) for a finite class $\mathcal{F}$. We will refer to the procedure as the Star estimator. Audibert showed that this method is deviation-optimal for finite aggregation — the first such result, followed by other estimators with similar properties (Lecué and Mendelson, 2009; Dai et al., 2012) for the finite case. We present analysis that quantifies the behavior of this method for arbitrary classes of functions. The method has several nice features. First, it provides an alternative to the 3-stage discretization method of Rakhlin et al. (2015), does not require the prior knowledge of the entropy of the class, and goes beyond the bounded case. Second, it enjoys an upper bound of offset Rademacher complexity via relatively routine arguments under rather weak assumptions. Third, it naturally reduces to empirical risk minimization for convex classes (indeed, this happens whenever $\operatorname{star}(\mathcal{F}, \widehat{g}) = \mathcal{F}$).

Let f^* denote the minimizer

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2,$$

and let ξ denote the "noise"

$$\xi = Y - f^*.$$

We say that the model is misspecified if the regression function $\mathbb{E}[Y | X = x] \notin \mathcal{F}$, which means ξ is not zero-mean. Otherwise, we say that the model is well-specified.
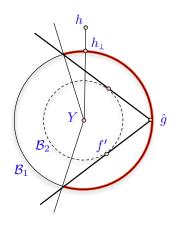
3. A Geometric Inequality

We start by proving a geometric inequality for the Star estimator. This deterministic inequality holds conditionally on $X_1, \ldots, X_n$, and therefore reduces to a problem in $\mathbb{R}^n$.

Lemma 1 (Geometric Inequality) *The two-step estimator $\widehat{f}$ in (2) satisfies*

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 \geq c \cdot \widehat{\mathbb{E}}(\widehat{f} - h)^2 \tag{3}$$

for any $h \in \mathcal{F}$ and $c = 1/18$. If $\mathcal{F}$ is convex, (3) holds with $c = 1$. Moreover, if $\mathcal{F}$ is a linear subspace, (3) holds with equality and $c = 1$ by the Pythagorean theorem.

Remark 2 *In the absence of convexity of $\mathcal{F}$, the two-step estimator $\widehat{f}$ mimics the key Pythagorean identity, though with a constant $1/18$. We have not focused on optimizing c but rather on presenting a clean geometric argument.*

Proof [Proof of Lemma 1]

Define the empirical ℓ_2 distance to be, for any f, g, $\|f\|_n := [\widehat{\mathbb{E}} f^2]^{1/2}$ and empirical product to be $\langle f, g \rangle_n := \widehat{\mathbb{E}}[fg]$. We will slightly abuse the notation by identifying every function with its finite-dimensional projection on $(X_1, \ldots, X_n)$.

Denote the ball (and sphere) centered at Y and with radius $\|\widehat{g} - Y\|_n$ to be $\mathcal{B}_1 := \mathcal{B}(Y, \|\widehat{g} - Y\|_n)$ (and $\mathcal{S}_1$, correspondingly). In a similar manner, define $\mathcal{B}_2 := \mathcal{B}(Y, \|\widehat{f} - Y\|_n)$ and $\mathcal{S}_2$. By the definition of the Star algorithm, we have $\mathcal{B}_2 \subseteq \mathcal{B}_1$. The statement holds with $c = 1$ if $\widehat{f} = \widehat{g}$, and so we may assume $\mathcal{B}_2 \subset \mathcal{B}_1$. Denote by $\mathcal{C}$ the conic hull of $\mathcal{B}_2$ with origin at $\widehat{g}$. Define the spherical cap outside the cone $\mathcal{C}$ to be $\mathcal{S} = \mathcal{S}_1 \setminus \mathcal{C}$ (drawn in red in Figure 3).

First, by the optimality of $\widehat{g}$, for any $h \in \mathcal{F}$, we have $\|h - Y\|_n^2 \geq \|\widehat{g} - Y\|_n^2$, i.e. any $h \in \mathcal{F}$ is not in the interior of $\mathcal{B}_1$. Furthermore, h is not in the interior of the cone $\mathcal{C}$, as otherwise there would be a point inside $\mathcal{B}_2$ strictly better than $\widehat{f}$. Thus $h \in (\text{int}\mathcal{C})^c \cap (\text{int}\mathcal{B}_1)^c$.

Second, $\widehat{f} \in \mathcal{B}_2$ and it is a contact point of $\mathcal{C}$ and $\mathcal{S}_2$. Indeed, $\widehat{f}$ is necessarily on a line segment between $\hat{g}$ and a point outside $\mathcal{B}_1$ that does not pass through the interior of $\mathcal{B}_2$ by optimality of $\widehat{f}$. Let K be the set of all contact points – potential locations of $\widehat{f}$.

Now we fix $h \in \mathcal{F}$ and consider the two dimensional plane $\mathcal{L}$ that passes through three points $(\hat{g}, Y, h)$, depicted in Figure 3. Observe that the left-hand-side of the desired inequality (3) is constant as $\widehat{f}$ ranges over K. To prove the inequality it therefore suffices to choose a value $f' \in K$ that maximizes the right-hand-side. The maximization of $\|h - f'\|^2$ over $f' \in K$ is achieved by $f' \in K \cap \mathcal{L}$. This can be argued simply by symmetry: the two-dimensional plane $\mathcal{L}$ intersects $\text{span}(K)$ in a line and the distance between h and K is maximized at the extreme point of this intersection. Hence, to prove the desired inequality, we can restrict our attention to the plane $\mathcal{L}$ and f' instead of $\widehat{f}$.

For any $h \in \mathcal{F}$, define the projection of h onto the shell $\mathcal{L} \cap \mathcal{S}$ to be $h_\perp \in \mathcal{S}$. We first prove (3) for $h_\perp$ and then extend the statement to h. By the geometry of the cone,

$$\|f' - \widehat{g}\|_n \geq \frac{1}{2} \|\widehat{g} - h_\perp\|_n.$$

By triangle inequality,

$$\|f' - \widehat{g}\|_n \geq \frac{1}{2} \|\widehat{g} - h_\perp\|_n \geq \frac{1}{2} \left(\|f' - h_\perp\|_n - \|f' - \widehat{g}\|_n \right).$$

Rearranging,

$$\|f' - \widehat{g}\|_n^2 \geq \frac{1}{9}\|f' - h_\perp\|_n^2.$$

By the Pythagorean theorem,

$$\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2 = \|\widehat{g} - Y\|_n^2 - \|f' - Y\|_n^2 = \|f' - \widehat{g}\|_n^2 \geq \frac{1}{9}\|f' - h_\perp\|_n^2,$$

thus proving the claim for $h_\perp$ for constant $c = 1/9$.

We can now extend the claim to h. Indeed, due to the fact that $h \in (\text{int}\mathcal{C})^c \cap (\text{int}\mathcal{B}_1)^c$ and the geometry of the projection $h \to h_\perp$, we have $\langle h_\perp - Y, h_\perp - h \rangle_n \leq 0$. Thus

$$\begin{aligned}
\|h - Y\|_n^2 - \|f' - Y\|_n^2 &= \|h_\perp - h\|_n^2 + \|h_\perp - Y\|_n^2 - 2\langle h_\perp - Y, h_\perp - h \rangle_n - \|f' - Y\|_n^2 \\
&\geq \|h_\perp - h\|_n^2 + (\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2) \\
&\geq \|h_\perp - h\|_n^2 + \frac{1}{9}\|f' - h_\perp\|_n^2 \geq \frac{1}{18}(\|h_\perp - h\|_n + \|f' - h_\perp\|_n)^2 \\
&\geq \frac{1}{18}\|f' - h\|_n^2.
\end{aligned}$$

This proves the claim for h with constant $1/18$. ∎

An upper bound on excess loss follows immediately from Lemma 1.

Corollary 3 *Conditioned on the data $\{X_n, Y_n\}$, we have a deterministic upper bound for the Star algorithm:*

$$\mathcal{E}(\widehat{f}) \leq (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - \widehat{f})] + \mathbb{E}(f^* - \widehat{f})^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f^* - \widehat{f})^2, \qquad (4)$$

with the value of constant c given in Lemma 1.

Proof

$$\begin{aligned}
\mathcal{E}(\widehat{f}) &= \mathbb{E}(\widehat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 \\
&\leq \mathbb{E}(\widehat{f} - Y)^2 - \mathbb{E}(f^* - Y)^2 + \left[\widehat{\mathbb{E}}(f^* - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 - c \cdot \widehat{\mathbb{E}}(\widehat{f} - f^*)^2\right] \\
&= (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - \widehat{f})] + \mathbb{E}(f^* - \widehat{f})^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f^* - \widehat{f})^2.
\end{aligned}$$

∎

An attentive reader will notice that the multiplier on the negative empirical quadratic term in (4) is slightly larger than the one on the expected quadratic term. This is the starting point of the analysis that follows.

4. Symmetrization

We will now show that the discrepancy in the multiplier constant in (4) leads to offset Rademacher complexity through rather elementary symmetrization inequalities. We perform this analysis both in expectation (for the case of bounded functions) and in high probability (for the general unbounded case). While the former result follows from the latter, the in-expectation statement for bounded functions requires no assumptions, in contrast to control of the tails.

Theorem 4 *Define the set $\mathcal{H} := \mathcal{F} - f^* + star(\mathcal{F} - \mathcal{F})$. The following expectation bound on excess loss of the Star estimator holds:*

$$\mathbb{E}\mathcal{E}(\widehat{f}) \leq (2M + K(2+c)/2) \cdot \mathbb{E} \sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i h(X_i) - c' h(X_i)^2 \right\}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables, $c' = \min\{\frac{c}{4M}, \frac{c}{4K(2+c)}\}$, $K = \sup_f |f|_\infty$, and $M = \sup_f |Y - f|_\infty$ almost surely.

The proof of the theorem involves an introduction of independent Rademacher random variables and two contraction-style arguments to remove the multipliers $(Y_i - f^*(X_i))$. These algebraic manipulations are postponed to the appendix.

The term in the curly brackets will be called an offset Rademacher process, and the expected supremum — an offset Rademacher complexity. While Theorem 4 only applies to bounded functions and bounded noise, the upper bound already captures the localization phenomenon, even for non-convex function classes (and thus goes well beyond the classical local Rademacher analysis).

As argued in (Mendelson, 2014a), it is the contraction step that requires boundedness of the functions when analyzing square loss. Mendelson uses a small ball assumption (a weak condition on the distribution, stated below) to split the analysis into the study of the multiplier and quadratic terms. This assumption allows one to compare the expected square of any function to its empirical version, to within a multiplicative constant that depends on the small ball property. In contrast, we need a somewhat stronger assumption that will allow us to take this constant to be at least $1 - c/4$. We phrase this condition—the lower isometry bound—as follows. [1]

Definition 5 (Lower Isometry Bound) *We say that a function class $\mathcal{F}$ satisfies the lower isometry bound with some parameters $0 < \eta < 1$ and $0 < \delta < 1$ if*

$$\mathbb{P}\left(\inf_{f \in \mathcal{F} \setminus \{0\}} \frac{1}{n} \sum_{i=1}^{n} \frac{f^2(X_i)}{\mathbb{E}f^2} \geq 1 - \eta \right) \geq 1 - \delta \tag{5}$$

for all $n \geq n_0(\mathcal{F}, \delta, \eta)$, where $n_0(\mathcal{F}, \delta, \eta)$ depends on the complexity of the class.

In general this is a mild assumption that requires good tail behavior of functions in $\mathcal{F}$, yet it is stronger than the small ball property. Mendelson Mendelson (2015) shows that this condition holds for heavy-tailed classes assuming the small ball condition plus a norm-comparison property $\|f\|_{\ell_q} \leq L\|f\|_{\ell_2}, \forall f \in \mathcal{F}$. We also remark that Assumption 5 holds for sub-gaussian classes $\mathcal{F}$ using concentration tools, as already shown in Lecué and Mendelson (2013). For completeness, let us also state the small ball property:

---

1. We thank Shahar Mendelson for pointing out that the small ball condition in the initial version of this paper was too weak for our purposes.

Definition 6 (Small Ball Property Mendelson **(2014a,b))** *The class of functions $\mathcal{F}$ satisfies the small-ball condition if there exist constants $\kappa > 0$ and $0 < \epsilon < 1$ for every $f \in \mathcal{F}$,*

$$\mathbb{P}\big(|f(X)| \geq \kappa(\mathbb{E}f^2)^{1/2}\big) \geq \epsilon.$$

Armed with the lower isometry bound, we now prove that the tail behavior of the deterministic upper bound in (4) can be controlled via the tail behavior of offset Rademacher complexity.

Theorem 7 *Define the set $\mathcal{H} := \mathcal{F} - f^* + \text{star}(\mathcal{F} - \mathcal{F})$. Assume the lower isometry bound in Definition 5 holds with $\eta = c/4$ and some $\delta < 1$, where c is the constant in (3). Let $\xi_i = Y_i - f^*(X_i)$. Define*

$$A := \sup_{h \in \mathcal{H}} \frac{\mathbb{E}h^4}{(\mathbb{E}h^2)^2} \quad and \quad B := \sup_{X,Y} \mathbb{E}\xi^4.$$

Then there exist two absolute constants $c', \tilde{c} > 0$ (only depends on c), such that

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) > 4u\right) \leq 4\delta + 4\mathbb{P}\left(\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - \tilde{c} \cdot h(X_i)^2 > u\right)$$

for any

$$u > \frac{32\sqrt{AB}}{c'} \cdot \frac{1}{n},$$

as long as $n > \frac{16(1-c')^2 A}{c'^2} \vee n_0(\mathcal{H}, \delta, c/4)$.

Theorem 7 states that excess loss is stochastically dominated by offset Rademacher complexity. We remark that the requirement in A, B holds under the mild moment conditions.

Remark 8 *In certain cases, Definition 5 can be shown to hold for $f \in \mathcal{F} \setminus r^*\mathcal{B}$ (rather than all $f \in \mathcal{F}$), for some critical radius r^*, as soon as $n \geq n_0(\mathcal{F}, \delta, \eta, r^*)$ (see Mendelson (2015)). In this case, the bound on the offset complexity is only affected additively by $(r^*)^2$.*

We postpone the proof of the Theorem to the appendix. In a nutshell, it extends the classical probabilistic symmetrization technique (Giné and Zinn, 1984; Mendelson, 2003) to the non-zero-mean offset process under the investigation.

5. Offset Rademacher Process: Chaining and Critical Radius

Let us summarize the development so far. We have shown that excess loss of the Star estimator is upper bounded by the (data-dependent) offset Rademacher complexity, both in expectation and in high probability, under the appropriate assumptions. We claim that the necessary properties of the estimator are now captured by the offset complexity, and we are now squarely in the realm of empirical process theory. In particular, we may want to quantify rates of convergence under complexity assumptions on $\mathcal{F}$, such as covering numbers. In contrast to local Rademacher analyses where one would need to estimate the data-dependent fixed point of the critical radius in some way, the task is much easier for the offset complexity. To this end, we study the offset process with the tools of empirical process theory.

5.1. Chaining Bounds

The first lemma describes the behavior of offset Rademacher process for a finite class.

Lemma 9 *Let $V \subset \mathbb{R}^n$ be a finite set of vectors of cardinality N. Then for any $C > 0$,*

$$\mathbb{E}_\epsilon \max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] \leq \frac{1}{2C} \frac{\log N}{n}.$$

Furthermore, for any $\delta > 0$,

$$\mathbb{P} \left(\max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] \geq \frac{1}{2C} \frac{\log N + \log 1/\delta}{n} \right) \leq \delta.$$

When the noise ξ is unbounded,

$$\mathbb{E}_\epsilon \max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i \xi_i v_i - C v_i^2 \right] \leq M \cdot \frac{\log N}{n},$$

$$\mathbb{P}_\epsilon \left(\max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i \xi_i v_i - C v_i^2 \right] \geq M \cdot \frac{\log N + \log 1/\delta}{n} \right) \leq \delta,$$

where

$$M := \sup_{v \in V \setminus \{0\}} \frac{\sum_{i=1}^n v_i^2 \xi_i^2}{2C \sum_{i=1}^n v_i^2}. \tag{6}$$

Armed with the lemma for a finite collection, we upper bound the offset Rademacher complexity of a general class through the chaining technique. We perform the analysis in expectation and in probability. Recall that a δ-cover of a subset S in a metric space (T, d) is a collection of elements such that the union of the δ-balls with centers at the elements contains S. A covering number at scale δ is the size of the minimal δ-cover.

One of the main objectives of symmetrization is to arrive at a stochastic process that can be studied conditionally on data, so that all the relevant complexities can be made sample-based (or, empirical). Since the functions only enter offset Rademacher complexity through their values on the sample $X_1, \ldots, X_n$, we are left with a finite-dimensional object. Throughout the paper, we work with the empirical ℓ_2 distance

$$d_n(f, g) = \left(\frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2 \right)^{1/2}.$$

The covering number of $\mathcal{G}$ at scale δ with respect to d_n will be denoted by $\mathcal{N}_2(\mathcal{G}, \delta)$.

Lemma 10 *Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then for any $z_1, \ldots, z_n \in \mathcal{Z}$*

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i g(z_i) - C g(z_i)^2 \right] \leq \inf_{\gamma \geq 0, \alpha \in [0, \gamma]} \left\{ \frac{(2/C) \log \mathcal{N}_2(\mathcal{G}, \gamma)}{n} \right.$$

$$\left. + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\}$$

where $\mathcal{N}_2(\mathcal{G}, \gamma)$ is an ℓ_2-cover of $\mathcal{G}$ on $(z_1, \ldots, z_n)$ at scale γ (assumed to contain $\mathbf{0}$).

Instead of assuming that $\mathbf{0}$ is contained in the cover, we may simply increase the size of the cover by 1, which can be absorbed by a small change of a constant.

Let us discuss the upper bound of Lemma 10. First, we may take $\alpha = 0$, unless the integral diverges (which happens for very large classes with entropy growth of $\log \mathcal{N}_2(\mathcal{G}, \delta) \sim \delta^{-p}$, $p \geq 2$). Next, observe that first term is precisely the rate of aggregation with a finite collection of size $\mathcal{N}_2(\mathcal{G}, \gamma)$. Hence, the upper bound is an optimal balance of the following procedure: cover the set at scale γ and pay the rate of aggregation for this finite collection, plus pay the rate of convergence of ERM within a γ-ball. The optimal balance is given by some γ (and can be easily computed under assumptions on covering number behavior — see (Rakhlin and Sridharan, 2014)). The optimal γ quantifies the localization radius that arises from the curvature of the loss function. One may also view the optimal balance as the well-known equation

$$\frac{\log \mathcal{N}(\mathcal{G}, \gamma)}{n} \asymp \gamma^2,$$

studied in statistics (Yang and Barron, 1999) for well-specified models. The present paper, as well as (Rakhlin et al., 2015), extend the analysis of this balance to the misspecified case and non-convex classes of functions.

Now we provide a high probability analogue of Lemma 10.

Lemma 11 *Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then for any $z_1, \ldots, z_n \in \mathcal{Z}$ and any $u > 0$,*

$$\mathbb{P}_\epsilon \left(\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i g(z_i) - C g(z_i)^2 \right] > u \cdot \inf_{\alpha \in [0, \gamma]} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\} + \frac{2}{C} \frac{\log \mathcal{N}_2(\mathcal{G}, \gamma) + u}{n} \right)$$
$$\leq \frac{2}{1 - e^{-2}} \exp(-cu^2) + \exp(-u)$$

where $\mathcal{N}_2(\mathcal{G}, \gamma)$ is an ℓ_2-cover of $\mathcal{G}$ on $(z_1, \ldots, z_n)$ at scale γ (assumed to contain $\mathbf{0}$) and $C, c > 0$ are universal constants.

The above lemmas study the behavior of offset Rademacher complexity for abstract classes $\mathcal{G}$. Observe that the upper bounds in previous sections are in terms of the class $\mathcal{F} - f^* + \text{star}(\mathcal{F} - \mathcal{F})$. This class, however, is not more complex that the original class $\mathcal{F}$ (with the exception of a finite class $\mathcal{F}$). More precisely, the covering numbers of $\mathcal{F} + \mathcal{F}' := \{f + g : f \in \mathcal{F}, g \in \mathcal{F}'\}$ and $\mathcal{F} - \mathcal{F}' := \{f - g : f \in \mathcal{F}, g \in \mathcal{F}'\}$ are bounded as

$$\log \mathcal{N}_2(\mathcal{F} + \mathcal{F}', 2\epsilon), \ \log \mathcal{N}_2(\mathcal{F} - \mathcal{F}', 2\epsilon) \leq \log \mathcal{N}_2(\mathcal{F}, \epsilon) + \log \mathcal{N}_2(\mathcal{F}', \epsilon)$$

for any $\mathcal{F}, \mathcal{F}'$. The following lemma shows that the complexity of the star hull $\text{star}(\mathcal{F})$ is also not significantly larger than that of $\mathcal{F}$.

Lemma 12 (Mendelson (2002), Lemma 4.5) *For any scale $\epsilon > 0$, the covering number of $\mathcal{F} \subset \mathcal{B}_2$ and that of $\text{star}(\mathcal{F})$ are bounded in the sense*

$$\log \mathcal{N}_2(\mathcal{F}, 2\epsilon) \leq \log \mathcal{N}_2(star(\mathcal{F}), 2\epsilon) \leq \log \frac{2}{\epsilon} + \log \mathcal{N}_2(\mathcal{F}, \epsilon).$$

5.2. Critical Radius

Now let us study the critical radius of offset Rademacher processes. Let $\xi = f^* - Y$ and define

$$
\alpha_n(\mathcal{H}, \kappa, \delta) \triangleq \inf \left\{ r > 0 : \mathbb{P} \left(\sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} \leq \kappa r^2 \right) \geq 1 - \delta \right\}.
$$
(7)

Theorem 13 *Assume $\mathcal{H}$ is star-shaped around 0 and the lower isometry bound holds for δ, ϵ. Define the critical radius*

$$
r = \alpha_n(\mathcal{H}, c'(1 - \epsilon), \delta).
$$

Then we have with probability at least $1 - 2\delta$,

$$
\sup_{h \in \mathcal{H}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} = \sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\},
$$

which further implies

$$
\sup_{h \in \mathcal{H}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} \leq r^2.
$$

The first statement of Theorem 13 shows the self-modulating behavior of the offset process: there is a critical radius, beyond which the fluctuations of the offset process are controlled by those within the radius. To understand the second statement, we observe that the complexity α_n is upper bounded by the corresponding complexity in (Mendelson, 2014a), which is defined without the quadratic term subtracted off. Hence, offset Rademacher complexity is no larger (under our Assumption 5) than the upper bounds obtained by Mendelson (2014a) in terms of the critical radius.

6. Examples

In this section, we briefly describe several applications. The first is concerned with parametric regression.

Lemma 14 *Consider the parametric regression $Y_i = X_i^T \beta^* + \xi_i, 1 \leq i \leq n$, where ξ_i need not be centered. The offset Rademacher complexity is bounded as*

$$
\mathbb{E}_\epsilon \sup_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i X_i^T \beta - C\beta^T X_i X_i^T \beta \right\} = \frac{\text{tr} \left(G^{-1} H \right)}{Cn}
$$

and

$$
\mathbb{P}_\epsilon \left(\sup_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i X_i^T \beta - C\beta^T X_i X_i^T \beta \right\} \geq \frac{\text{tr} \left(G^{-1} H \right)}{Cn} + \frac{\sqrt{\text{tr} \left([G^{-1} H]^2 \right)}}{n} \left(4\sqrt{2 \log \frac{1}{\delta}} + 64 \log \frac{1}{\delta} \right) \right) \leq \delta
$$

where $G := \sum_{i=1}^n X_i X_i^T$ is the Gram matrix and $H = \sum_{i=1}^n \xi_i^2 X_i X_i^T$. In the well-specified case (that is, ξ_i are zero-mean), assuming that conditional variance is σ^2, then conditionally on the design matrix, $\mathbb{E} G^{-1} H = \sigma^2 I_p$ and excess loss is upper bounded by order $\frac{\sigma^2 p}{n}$.

10

Proof The offset Rademacher can be interpreted as the Fenchel-Legendre transform, where

$$\sup_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i X_i^T \beta - C\beta^T X_i X_i^T \beta \right\} = \frac{\sum_{i,j=1}^n \epsilon_i \epsilon_j \xi_i \xi_j X_i^T G^{-1} X_j}{Cn}. \tag{8}$$

Thus we have in expectation

$$\mathbb{E}_\epsilon \frac{1}{n} \sup_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i X_i^T \beta - C\beta^T X_i X_i^T \beta \right\} = \frac{\sum_{i=1}^n \xi_i^2 X_i^T G^{-1} X_i}{Cn} = \frac{\mathsf{tr}[G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)]}{Cn}. \tag{9}$$

For high probability bound, note the expression in Equation (8) is Rademacher chaos of order two. Define symmetric matrix $M \in \mathbb{R}^{n \times n}$ with entries

$$M_{ij} = \xi_i \xi_j X_i^T G^{-1} X_j$$

and define

$$Z = \sum_{i,j=1}^n \epsilon_i \epsilon_j \xi_i \xi_j X_i^T G^{-1} X_j = \sum_{i,j=1}^n \epsilon_i \epsilon_j M_{ij}.$$

Then

$$\mathbb{E}Z = \mathsf{tr}[G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)],$$

and

$$\mathbb{E} \sum_{i=1}^n (\sum_{j=1}^n \epsilon_j M_{ij})^2 = \|M\|_F^2 = \mathsf{tr}[G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)].$$

Furthermore,

$$\|M\| \leq \|M\|_F = \sqrt{\mathsf{tr}[G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)]}$$

We apply the concentration result in (Boucheron et al., 2013, Exercise 6.9),

$$\mathbb{P}\left(Z - \mathbb{E}Z \geq 4\sqrt{2}\|M\|_F \sqrt{t} + 64\|M\|t \right) \leq e^{-t}. \tag{10}$$

∎

For the finite dictionary aggregation problem, the following lemma shows control of offset Rademacher complexity.

Lemma 15 *Assume $\mathcal{F} \in \mathcal{B}_2$ is a finite class of cardinality N. Define $\mathcal{H} = \mathcal{F} - f^* + star(\mathcal{F} - \mathcal{F})$ which contains the Star estimator $\widehat{f} - f^*$ defined in Equation (2). The offset Rademacher complexity for $\mathcal{H}$ is bounded as*

$$\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - Ch(X_i)^2 \right\} \leq \tilde{C} \cdot \frac{\log(N \vee n)}{n}$$

and

$$\mathbb{P}_\epsilon \left(\sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - Ch(X_i)^2 \right\} \leq \tilde{C} \cdot \frac{\log(N \vee n) + \log \frac{1}{\delta}}{n} \right) \leq \delta.$$

where $\tilde{C}$ is a constant depends on $K := 2(\sqrt{\sum_{i=1}^n \xi_i^2/n} + 2C)$ and

$$M := \sup_{h \in \mathcal{H} \setminus \{0\}} \frac{\sum_{i=1}^n h(X_i)^2 \xi_i^2}{2C \sum_{i=1}^n h(X_i)^2}.$$

We observe that the bound of Lemma 15 is worse than the optimal bound of (Audibert, 2007) by an additive $\frac{\log n}{n}$ term. This is due to the fact that the analysis for finite case passes through the offset Rademacher complexity of the star hull, and for this case the star hull is more rich than the finite class. For this case, a direct analysis of the Star estimator is provided in (Audibert, 2007).

While the offset complexity of the star hull is crude for the finite case, the offset Rademacher complexity *does* capture the correct rates for regression with larger classes, initially derived in (Rakhlin et al., 2015). We briefly mention the result. The proof is identical to the one in (Rakhlin and Sridharan, 2014), with the only difference that offset Rademacher is defined in that paper as a sequential complexity in the context of online learning.

Corollary 16 *Consider the problem of nonparametric regression, as quantified by the growth*

$$\log \mathcal{N}_2(\mathcal{F}, \epsilon) \leq \epsilon^{-p}.$$

In the regime $p \in (0, 2)$, the upper bound of Lemma 11 scales as $n^{-\frac{2}{2+p}}$. In the regime $p \geq 2$, the bound scales as $n^{-1/p}$, with an extra logarithmic factor at $p = 2$.

For the parametric case of $p = 0$, one may also readily estimate the offset complexity. Results for VC classes, sparse combinations of dictionary elements, and other parametric cases follow easily by plugging in the estimate for the covering number or directly upper bounding the offset complexity (see Rakhlin et al. (2015); Rakhlin and Sridharan (2014)).

7. Lower bound on Minimax Regret via Offset Rademacher Complexity

We conclude this paper with a lower bound on minimax regret in terms of offset Rademacher complexity.

Theorem 17 (Minimax Lower Bound on Regret) *Define the offset Rademacher complexity over $\mathcal{X}^{\otimes n}$ as*

$$\mathfrak{R}^\circ(n, \mathcal{F}) = \sup_{\{x_i\}_{i=1}^n \in \mathcal{X}^{\otimes n}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i f(x_i) - f(x_i)^2 \right\}$$

then the following minimax lower bound on regret holds:

$$\inf_{\hat{g} \in \mathcal{G}} \sup_P \left\{ \mathbb{E}(\hat{g} - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2 \right\} \geq \mathfrak{R}^\circ((1+c)n, \mathcal{F}) - \frac{c}{1+c} \mathfrak{R}^\circ(cn, \mathcal{G}),$$

for any $c > 0$.

For the purposes of matching the performance of the Star procedure, we can take $\mathcal{G} = \mathcal{F} + \text{star}(\mathcal{F} - \mathcal{F})$.

Acknowledgments

References

Audibert, J. (2007). Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20(2).

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

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Bousquet, O. (2002). *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Ecole Polytechnique.

Bousquet, O., Koltchinskii, V., and Panchenko, D. (2002). Some local measures of complexity of convex hulls and generalization bounds. In Springer, editor, *Computational Learning Theory*, pages 164–171, Sydney, Australia.

Dai, D., Rigollet, P., and Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Annals of Statistics*.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Annals of Probability*, 12(4):929–989.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. *High Dimensional Probability*, II:443–459.

Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probability Theory and Related Fields*, 145(3):591–613.

Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*.

Mendelson, S. (2002). Improving the sample complexity using global data. *Information Theory, IEEE Transactions on*, 48(7):1977–1991.

Mendelson, S. (2003). A few notes on statistical learning theory. In Mendelson, S. and Smola, A. J., editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer.

Mendelson, S. (2014a). Learning without Concentration. In *Conference on Learning Theory*.

Mendelson, S. (2014b). Learning without Concentration for General Loss Functions. *ArXiv e-prints*.

Mendelson, S. (2015). On aggregation for heavy-tailed classes. Preprint.

Rakhlin, A. and Sridharan, K. (2014). Online non-parametric regression. In *Conference on Learning Theory*.

Rakhlin, A., Sridharan, K., and Tsybakov, A. (2015). Empirical entropy, minimax regret and minimax risk. *Bernoulli*. Forthcoming.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599.

Appendix A. Proofs

Proof [Proof of Theorem 4] Since $\widehat{f}$ is in the star hull around $\widehat{g}$, $\widehat{f}$ must lie in the set $\mathcal{H} := \mathcal{F} + \text{star}(\mathcal{F} - \mathcal{F})$. Hence, in view of (4), excess loss $\mathcal{E}(\widehat{f})$ is upper bounded by

$$\sup_{f \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] + \mathbb{E}(f^* - f)^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f^* - f)^2 \right\} \tag{11}$$

$$\leq \sup_{f \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] + (1 + c/4)\mathbb{E}(f^* - f)^2 - (1 + 3c/4) \cdot \widehat{\mathbb{E}}(f^* - f)^2 \right.$$
$$\left. - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \mathbb{E}(f^* - f)^2 \right) \right\}$$

$$\leq \sup_{f \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \mathbb{E}(f^* - f)^2 \right) \right\} \tag{12}$$

$$+ \sup_{f \in \mathcal{H}} \left\{ (1 + c/4)\mathbb{E}(f^* - f)^2 - (1 + 3c/4) \cdot \widehat{\mathbb{E}}(f^* - f)^2 \right\} \tag{13}$$

We invoke the supporting Lemma 18 (stated and proved below) for the term (13):

$$\mathbb{E} \sup_{f \in \mathcal{H}} \left\{ (1 + c/4)\mathbb{E}(f^* - f)^2 - (1 + 3c/4) \cdot \widehat{\mathbb{E}}(f^* - f)^2 \right\} \tag{14}$$

$$\leq \frac{K(2 + c)}{2} \cdot \mathbb{E} \sup_{f \in \mathcal{H}} \frac{1}{n} \left\{ \sum_{i=1}^{n} 2\epsilon_i(f(X_i) - f^*(X_i)) - \frac{c}{4K(2 + c)} \cdot \sum_{i=1}^{n}(f(X_i) - f^*(X_i))^2 \right\}. \tag{15}$$

Let $\widehat{\mathbb{E}}'$ stand for empirical expectation with respect to an independent copy $(X_1', \ldots, X_n')$. For the term (12), Jensen's inequality yields

$$\mathbb{E} \sup_{f \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \mathbb{E}(f^* - f)^2 \right) \right\}$$

$$\leq \mathbb{E} \sup_{f \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[2(f^* - Y)(f^* - f)] - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \widehat{\mathbb{E}}'(f^* - f)^2 \right) \right\}.$$

When introducing i.i.d. Rademacher random variables, we observe that the quadratic term remains unchanged by renaming X_i and X_i', and thus the preceding expression is upper bounded by

$$2\mathbb{E} \sup_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i(f^*(X_i) - Y_i)(f^*(X_i) - f(X_i)) - (c/4)(f^*(X_i) - f(X_i))^2 \right\}.$$

Using a contraction technique as in the proof of Lemma 18, we obtain an upper bound of

$$2M \cdot \mathbb{E} \sup_{f \in \mathcal{H}} \frac{1}{n} \left\{ \sum_{i=1}^{n} 2\epsilon_i(f^*(X_i) - f(X_i)) - \frac{c}{4M} \cdot \sum_{i=1}^{n}(f^*(X_i) - f(X_i))^2 \right\} \tag{16}$$

Combining the bounds yields the statement of the theorem. ∎

Lemma 18 *For any class $\mathcal{F}$ of uniformly bounded functions with $K = \sup_{f \in \mathcal{F}} |f|_\infty$, for any $f^* \in \mathcal{F}$, and for any $c > 0$, it holds that*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}(f - f^*)^2 - (1 + 2c)\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

$$\leq c \cdot \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \left\{ \frac{4K(1+c)}{c} \sum_{i=1}^n \epsilon_i (f(X_i) - f^*(X_i)) - \sum_{i=1}^n (f(X_i) - f^*(X_i))^2 \right\}.$$

Proof [Proof of Lemma 18] We write

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}(f - f^*)^2 - (1 + 2c)\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

$$= \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ (1+c)\mathbb{E}(f - f^*)^2 - (1+c)\widehat{\mathbb{E}}(f - f^*)^2 - c\mathbb{E}(f - f^*)^2 - c\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

which, by Jensen's inequality, is upper bounded by

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ (1+c)(\widehat{\mathbb{E}}'(f - f^*)^2 - \widehat{\mathbb{E}}(f - f^*)^2) - c\widehat{\mathbb{E}}'(f - f^*)^2 - c\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

We recall that $\widehat{\mathbb{E}}'$ is an empirical mean operator with respect to an independent copy $(X_1', \ldots, X_n')$. Writing out the empirical expectations in the above expression, the above is equal to

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=1}^n \epsilon_i \left((f(X_i') - f^*(X_i'))^2 - (f(X_i) - f^*(X_i))^2 \right) - c\widehat{\mathbb{E}}'(f - f^*)^2 - c\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

$$\leq 2 \cdot \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=1}^n \epsilon_i (f(X_i) - f^*(X_i))^2 - c\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

with the last expectation taken over ϵ_i and data X_i, $1 \leq i \leq n$.

We proceed with a contraction-style proof. Condition on $X_1, \ldots, X_n$ and $\epsilon_2, \ldots, \epsilon_n$, and write out the expectation with respect to ϵ_1:

$$\frac{1}{2} \sup_{f \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=2}^n \epsilon_i (f(X_i) - f^*(X_i))^2 - c\widehat{\mathbb{E}}(f - f^*)^2 + \frac{1+c}{n}(f(X_1) - f^*(X_1))^2 \right\}$$

$$+ \frac{1}{2} \sup_{g \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=2}^n \epsilon_i (g(X_i) - f^*(X_i))^2 - c\widehat{\mathbb{E}}(g - f^*)^2 - \frac{1+c}{n}(g(X_1) - f^*(X_1))^2 \right\}$$

$$\leq \frac{1}{2} \sup_{f,g \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=2}^n \epsilon_t (f(X_i) - f^*(X_i))^2 - c\widehat{\mathbb{E}}(f - f^*)^2 + \frac{1+c}{n} \sum_{i=2}^n \epsilon_t (g(X_i) - f^*(X_i))^2 \right.$$

$$\left. - c\widehat{\mathbb{E}}(g - f^*)^2 + \frac{4K(1+c)}{n}|f(X_1) - g(X_1)| \right\}$$

The absolute value can be dropped since the expression is symmetric in f, g. We obtain an upper bound of

$$\frac{1}{2} \sup_{f,g \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=2}^{n} \epsilon_t (f(X_i) - f^*(X_i))^2 - c\widehat{\mathbb{E}}(f - f^*)^2 + \frac{1+c}{n} \sum_{i=2}^{n} \epsilon_t (g(X_i) - f^*(X_i))^2 \right.$$
$$\left. -c\widehat{\mathbb{E}}(g - f^*)^2 + \frac{4K(1+c)}{n}(f(X_1) - g(X_1)) \right\}$$

$$= \mathbb{E}_{\epsilon_1} \sup_{f \in \mathcal{F}} \left\{ \frac{1+c}{n} \sum_{i=2}^{n} \epsilon_i (f(X_i) - f^*(X_i))^2 - c\widehat{\mathbb{E}}(f - f^*)^2 + \frac{4K(1+c)}{n}\epsilon_1 f(X_1) \right\}$$

Proceeding in this fashion for ϵ_2 until ϵ_n, we conclude

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}(f - f^*)^2 - (1 + 2c)\widehat{\mathbb{E}}(f - f^*)^2 \right\}$$

$$\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \frac{4K(1+c)}{n} \sum_{i=1}^{n} \epsilon_t (f(X_i) - f^*(X_i)) - \frac{c}{n} \sum_{i=1}^{n} (f(X_i) - f^*(X_i))^2 \right\}$$

where we added f^* back in for free since random signs are zero-mean. ∎

Proof [Proof of Theorem 7] We start with the deterministic upper bound (11) on excess loss (see the proof of Theorem 4):

$$\sup_{h \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] + \mathbb{E}h^2 - (1 + c) \cdot \widehat{\mathbb{E}}h^2 \right\} \tag{17}$$

where $h = f - f^* \in \mathcal{H}$. Define

$$U_{X_i, Y_i}(h) = 2\xi_i h(X_i) - \mathbb{E}[2\xi h] + \mathbb{E}h^2 - (1 + c) \cdot h(X_i)^2,$$
$$V_{X_i, Y_i}(h) = 2\xi_i h(X_i) - \mathbb{E}[2\xi h] - \mathbb{E}h^2 + (1 - c') \cdot h(X_i)^2.$$

where c' will be specified later. We now prove a version of probabilistic symmetrization lemma Giné and Zinn (1984); Mendelson (2003) for

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i, Y_i}(h) > x \right). \tag{18}$$

Note that unlike the usual applications of the technique in the literature, we perform symmetrization with the quadratic terms. Define

$$\mathcal{B} = \left\{ \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i, Y_i}(h) > x \right\}, \quad \beta = \inf_{h \in \mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i, Y_i}(h) < \frac{x}{2} \right). \tag{19}$$

Clearly for $\{X_i, Y_i\}_{i=1}^{n} \in \mathcal{B}$, there exists a $h \in \mathcal{H}$ satisfies condition in $\mathcal{B}$. If in addition h satisfies

$$\sum_{i=1}^{n} V_{X_i', Y_i'}(h) < \frac{x}{2}$$

17

then

$$\sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2}$$

and therefore

$$\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2}.$$

The latter can be written as

$$\sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{n} 2\xi_i h(X_i) - 2\xi_i' h(X_i') + 2\mathbb{E}h^2 - (1+c) \cdot h(X_i)^2 - (1-c') \cdot h(X_i')^2 \right\} > \frac{x}{2}.$$

Then for this particular h,

$$\beta = \inf_{g \in \mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i',Y_i'}(g) < \frac{x}{2} \right) \leq \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i',Y_i'}(h) < \frac{x}{2} \right)$$

$$\leq \mathbb{P}\left(\sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2} \right) \leq \mathbb{P}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2} \right).$$

Note that the right-hand-side does not depend on h. We integrate over $\{X_i, Y_i\}_{i=1}^{n} \in \mathcal{B}$ to obtain

$$\beta \cdot \mathbb{P}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) > x \right)$$

$$\leq \mathbb{P}\left(\sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] + 2\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \right\} > \frac{x}{2} \right) \quad (20)$$

Next, we apply Assumption 5 with $\epsilon = c/4 = 1/72$ to terms in (20) to construct an offset Rademacher process. Note

$$\frac{2}{1-\epsilon} < 2(1+2\epsilon) = 2 + c.$$

We can now choose $\tilde{c}, c' > 0$ in that satisfy

$$\frac{2}{1-\epsilon} \leq 2 + c - c' - 2\tilde{c} \quad \Longleftrightarrow \quad 1 - (1 - c' - \tilde{c})(1-\epsilon) \leq (1 + c - \tilde{c})(1-\epsilon) - 1. \quad (21)$$

Choose b now such that

$$1 - (1 - c' - \tilde{c})(1-\epsilon) \leq b \leq (1 + c - \tilde{c})(1-\epsilon) - 1. \quad (22)$$

Then we have on the set $\mathcal{H}$, applying lower isometry bound and Eq. (22), with probability at least $1 - 2\delta$,

$$\widehat{\mathbb{E}}(f - f^*)^2 \geq (1-\epsilon) \cdot \mathbb{E}(f - f^*)^2 \implies (1+b)\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 \leq -\tilde{c} \cdot \widehat{\mathbb{E}}h^2,$$

$$\widehat{\mathbb{E}}'(f - f^*)^2 \geq (1-\epsilon) \cdot \mathbb{E}(f - f^*)^2 \implies (1-b)\mathbb{E}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \leq -\tilde{c} \cdot \widehat{\mathbb{E}}'h^2.$$

Thus we can continue bounding the expression in (20) as

$$\sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] + 2\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \right\}$$

$$= \sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] + (1+b)\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 + (1-b)\mathbb{E}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \right\}$$

$$\leq \sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] - \tilde{c} \cdot \widehat{\mathbb{E}}h^2 - \tilde{c} \cdot \widehat{\mathbb{E}}'h^2 \right\}$$

For the probability of deviation, we obtain

$$\beta \cdot \mathbb{P} \left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i, Y_i}(h) > x \right)$$

$$\leq \mathbb{P} \left(\sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] - \tilde{c} \cdot \widehat{\mathbb{E}}h^2 - \tilde{c} \cdot \widehat{\mathbb{E}}'h^2 \right\} > \frac{x}{2} \right) + 2\delta$$

$$= \mathbb{P} \left(\sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\epsilon \xi h] - \tilde{c} \cdot \widehat{\mathbb{E}}h^2 - \tilde{c} \cdot \widehat{\mathbb{E}}'h^2 \right\} > \frac{x}{2} \right) + 2\delta$$

$$\leq 2\mathbb{P} \left(\sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^{n} h(X_i)^2 \right\} > \frac{x}{4} \right) + 2\delta.$$

To estimate β, write

$$\beta = \inf_{h \in \mathcal{H}} \mathbb{P} \left(\sum_{i=1}^{n} V_{X_i, Y_i}(h) < \frac{x}{2} \right) \tag{23}$$

$$= 1 - \sup_{h \in \mathcal{H}} \mathbb{P} \left(\sum_{i=1}^{n} 2\xi_i h(X_i) - \mathbb{E}[2\xi h] - \mathbb{E}h^2 + (1-c') \cdot h(X_i)^2 \geq \frac{x}{2} \right). \tag{24}$$

Let's bound the last term in above equation, for any $h \in \mathcal{H}$

$$\mathbb{P} \left((\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] + (1-c')\widehat{\mathbb{E}}h^2 - \mathbb{E}h^2 > \frac{x}{2n} \right) \tag{25}$$

$$\leq \mathbb{P} \left((\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] > \frac{x}{2n} + \frac{c'}{2}\mathbb{E}h^2 \right) + \mathbb{P} \left((\widehat{\mathbb{E}} - \mathbb{E})[h^2] > \frac{c'}{2(1-c')}\mathbb{E}h^2 \right). \tag{26}$$

$$\tag{27}$$

Define

$$A := \sup_{h \in \mathcal{H}} \frac{\mathbb{E}h^4}{(\mathbb{E}h^2)^2} \quad \text{and} \quad B := \sup_{X,Y} \mathbb{E}\xi^4.$$

Then for the second term in Eq (26), using Chebyshev's inequality

$$\mathbb{P} \left((\widehat{\mathbb{E}} - \mathbb{E})[h^2] > \frac{c'}{2(1-c')}\mathbb{E}h^2 \right) \leq \frac{4(1-c')^2 A}{c'^2 n} \leq 1/4$$

if

$$n \geq \frac{16(1 - c')^2 A}{c'^2}.$$

For the first term in Eq (26), note

$$\mathsf{Var}[2\xi h] \leq 4\mathbb{E}[\xi^2 h^2] \leq 4\sqrt{AB} \cdot \mathbb{E}h^2$$

and thus through Chebyshev inequality

$$\mathbb{P}\left((\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] > \frac{x}{2n} + \frac{c'}{2}\mathbb{E}h^2\right) \leq \frac{4\sqrt{AB} \cdot \mathbb{E}h^2}{n\left(\frac{x}{2n} + \frac{c'}{2}\mathbb{E}h^2\right)^2}$$

$$\leq \frac{4\sqrt{AB} \cdot \mathbb{E}h^2}{n \cdot 4\frac{x}{2n} \cdot \frac{c'}{2}\mathbb{E}h^2} \leq \frac{1}{4}$$

if

$$x \geq \frac{16\sqrt{AB}}{c'}.$$

Assemble above bounds, for any $h \in \mathcal{H}$

$$\sup_{h \in \mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} 2\xi_i h(X_i) - \mathbb{E}[2\xi h] - \mathbb{E}h^2 + (1 - c') \cdot h(X_i)^2 \geq \frac{x}{2}\right) \leq \frac{1}{2}$$

which further implies $\beta \geq 1/2$ for any $x > \frac{16\sqrt{AB}}{c'}$ and whenever

$$n > \frac{16(1 - c')^2 A}{c'^2}.$$

Under the above regime,

$$\frac{1}{2}\mathbb{P}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i, Y_i}(h) > x\right) \leq 2\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left\{\sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^{n} h(X_i)^2\right\} > \frac{x}{4}\right) + 2\delta$$

and so

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i, Y_i}(h) > 4t\right)$$

$$\leq 4\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left\{\sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^{n} h(X_i)^2\right\} > t\right) + 4\delta.$$

We conclude by writing

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} (\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] + \mathbb{E}h^2 - (1 + c) \cdot \widehat{\mathbb{E}}h^2 > 4t\right)$$

$$\leq 4\mathbb{P}\left(\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^{n} h(X_i)^2 > t\right) + 4\delta.$$

Proof [**Proof of Lemma 9**] Using a standard argument,

$$\mathbb{E}_\epsilon \max_{v \in V} \left[\sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] \le \frac{1}{\lambda} \log \sum_{v \in V} \mathbb{E}_\epsilon \exp \left\{ \sum_{i=1}^n \lambda \epsilon_i v_i - \lambda C v_i^2 \right\}.$$

For any $v \in V$,

$$\mathbb{E}_\epsilon \exp \left\{ \sum_{i=1}^n \lambda \epsilon_i v_i - \lambda C v_i^2 \right\} \le \exp \left\{ \sum_{i=1}^n \lambda^2 v_i^2 / 2 - \lambda C v_i^2 \right\} \le 1$$

by setting $\lambda = 2C$. The first claim follows. For the second claim,

$$\mathbb{P}\left(\max_{v \in V} \left[\sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] \ge \frac{1}{2C} \log(N/\delta) \right)$$

$$\le \mathbb{E} \exp \left\{ \lambda \max_{v \in V} \left[\sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] - \lambda \frac{1}{2C} \log(N/\delta) \right\}$$

$$\le \sum_{v \in V} \mathbb{E} \exp \left\{ \lambda \left[\sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] - \lambda \frac{1}{2C} \log(N/\delta) \right\}$$

$$\le \sum_{v \in V} \exp \left\{ - \log(N/\delta) \right\} = \delta.$$

Now let's move to the case where ξ, the noise is unbounded.

$$\mathbb{E}_\epsilon \frac{1}{n} \max_{v \in V} \left\{ \sum_{i=1}^n \epsilon_i \xi_i v_i - C v_i^2 \right\} \le \frac{1}{n\lambda} \log \mathbb{E}_\epsilon \sum_{v \in V} \exp \left(\lambda \sum_{i=1}^n \epsilon_i \xi_i v_i - \lambda C v_i^2 \right)$$

$$\le \frac{1}{n\lambda} \log \sum_{v \in V} \exp \left(\sum_{i=1}^n \frac{\lambda^2}{2} \xi_i^2 v_i^2 - \lambda C v_i^2 \right) \le \max_{v \in V \setminus \{0\}} \frac{\sum_{i=1}^n v_i^2 \xi_i^2}{2C \sum_{i=1}^n v_i^2} \cdot \frac{\log N}{n}$$

if we take $\lambda = \min_{v \in V \setminus \{0\}} \frac{2C \sum_{i=1}^n v_i^2}{\sum_{i=1}^n v_i^2 \xi_i^2}$. The high probability statement follows also use this particular choice of λ.

Proof [**Proof of Lemma 10**] The proof proceeds as in (Rakhlin and Sridharan, 2014). Fix $\gamma \in [0, 1]$. By definition of a cover, there exists a set $V \subset \mathbb{R}^n$ vectors of size $N = \mathcal{N}_2(\mathcal{G}, \gamma)$ with the following property: for any $g \in \mathcal{G}$, there exists a $v = v[g] \in V$ such that

$$\frac{1}{n} \sum_{i=1}^n (g(z_i) - v_i)^2 \le \gamma^2.$$

Then we may write,

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^n \epsilon_i g(z_i) - Cg(z_i)^2 \right] \tag{28}$$

$$\leq \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i(g(z_i) - v[g]_i) \right] + \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^n (C/4)v[g]_i^2 - Cg(z_i)^2 \right] \tag{29}$$

$$+ \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^n \epsilon_i v[g]_i - (C/4)v[g]_i^2 \right] \tag{30}$$

We now argue that the second term is nonpositive. More precisely, we claim that for any $g \in \mathcal{G}$,

$$\frac{1}{4} \sum_{t=1}^n v[g]_i^2 \leq \sum_{t=1}^n g(z_i)^2 \tag{31}$$

for some element $v[g] \in V \cup \{\mathbf{0}\}$. First, consider the case $\sum_{t=1}^n g(z_i)^2 \leq \gamma^2$. Then $v[g] = \mathbf{0}$ is an element γ-close to values of g on the sample, and (31) is trivially satisfied. Next, consider the case $\sum_{t=1}^n g(z_i)^2 > \gamma^2$ and write $u = (g(z_1), \ldots, g(z_n))$. The triangle inequality for the Euclidean norm yields

$$\|v[g]\| \leq \|v[g] - u\| + \|u\| \leq \gamma + \|u\| \leq 2\|u\|,$$

establishing non-positivity of the second term in (28). The third term in (28) is upper bounded with the help of Lemma 9 as

$$\mathbb{E}_\epsilon \max_{g \in \mathcal{G}} \left[\sum_{t=1}^n \epsilon_i v[g]_i - (C/4)v[g]_i^2 \right] \leq \frac{2}{C} \log \mathcal{N}_2(\mathcal{G}, \gamma)$$

Finally, the first term in (28) is upper bounded using the standard chaining technique, keeping in mind that the ℓ_2-diameter of the indexing set is at most γ. ∎

Proof [Proof of Lemma 11] The proof is similar to the proof of Lemma 10. We proceed with the following decomposition:

$$\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i g(z_i) - Cg(z_i)^2 \right] \leq \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i(g(z_i) - v[g]_i) \right] + \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i v[g]_i - \frac{C}{4}v[g]_i^2 \right].$$

For the first term, we can employ the traditional high probability chaining bound. For some $c > 0$, the following holds,

$$\mathbb{P}_\epsilon \left(\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i(g(z_i) - v[g]_i) \right] > u \cdot \inf_{\alpha \in [0,\gamma]} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\} \right)$$

$$\leq \frac{2}{1 - e^{-2}} \exp(-cu^2).$$

For the second term,

$$\mathbb{P}_\epsilon \left(\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i v[g]_i - (C/4)v[g]_i^2 \right] > \frac{2}{C} \frac{\log \mathcal{N}_2(\mathcal{G}, \gamma) + u}{n} \right) \leq \exp(-u).$$

Combining the above two bounds, we have

$$\mathbb{P}_\epsilon \left(\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i g(z_i) - C g(z_i)^2 \right] > u \cdot \inf_{\alpha \in [0,\gamma]} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\} \right.$$
$$\left. + \frac{2}{C} \frac{\log \mathcal{N}_2(\mathcal{G}, \gamma) + u}{n} \right)$$

$$\leq \mathbb{P}_\epsilon \left(\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i (g(z_i) - v[g]_i) \right] > u \cdot \inf_{\alpha \in [0,\gamma]} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\} \right)$$

$$+ \mathbb{P}_\epsilon \left(\sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i v[g]_i - (C/4) v[g]_i^2 \right] > \frac{2}{C} \frac{\log \mathcal{N}_2(\mathcal{G}, \gamma) + u}{n} \right)$$

$$\leq \frac{2}{1 - e^{-2}} \exp(-cu^2) + \exp(-u).$$

∎

Proof [Proof of Theorem 13] Denote by $\mathcal{B}$ the unit ball with respect to ℓ_2 distance, $\mathcal{B} = \{h : (\mathbb{E}h^2)^{1/2} \leq 1\}$, and let $\mathcal{S}$ denote the unit sphere. Choosing any $h \in \mathcal{H} \backslash r\mathcal{B}$, we have $\|h\|_{\ell_2} > r \triangleq \alpha_n(\mathcal{H}, \kappa', \delta)$ with k' to be chosen later. Under the assumption that $\mathcal{H}$ is star-shaped, we know $h_r := r/\|h\|_{\ell_2} \cdot h \in \mathcal{H}$, thus

$$\frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i)$$

$$= \frac{\|h\|_{\ell_2}}{r} \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h_r(X_i) - \left(\frac{\|h\|_{\ell_2}}{r} \right)^2 c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i)$$

$$= \frac{\|h\|_{\ell_2}}{r} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h_r(X_i) - c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} - \frac{\|h\|_{\ell_2}}{r} \left(\frac{\|h\|_{\ell_2}}{r} - 1 \right) c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i).$$

Comparing the supremum of the offset Rademacher process outside the ball $r\mathcal{B}$ with the one inside the ball $r\mathcal{B}$, we have

$$\sup_{h \in \mathcal{H} \backslash r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} - \sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\}$$

$$\leq \sup_{h \in \mathcal{H} \backslash r\mathcal{B}} \left\{ \left(\frac{\|h\|_{\ell_2}}{r} - 1 \right) \sup_{h_r \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h_r(X_i) - c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} \right.$$

$$\left. - \frac{\|h\|_{\ell_2}}{r} \left(\frac{\|h\|_{\ell_2}}{r} - 1 \right) \inf_{h_r \in \mathcal{H} \cap r\mathcal{S}} \left\{ c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} \right\}$$

$$\leq \sup_{h \in \mathcal{H} \backslash r\mathcal{B}} \left\{ \left(\frac{\|h\|_{\ell_2}}{r} - 1 \right) \left\{ \sup_{h_r \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h_r(X_i) - c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} \right. \right.$$

$$\left. \left. - \inf_{h_r \in \mathcal{H}_r \cap r\mathcal{S}} \left\{ c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} \right\} \right\}. \tag{32}$$

If

$$\kappa' r^2 \le c'(1-\epsilon)r^2,$$

we can apply the lower isometry bound 5 and conclude

$$\sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} \le k'r^2 \le c'(1-\epsilon)r^2 \le \inf_{h_r \in \mathcal{H} \cap r\mathcal{S}} \left\{ c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\}$$

with probability at least $1 - 2\delta$.

Under this event, the difference of terms in (32) is smaller than 0, and we conclude

$$\sup_{h \in \mathcal{H} \backslash r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} - \sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\}$$

$$\le \sup_{h \in \mathcal{H} \backslash r\mathcal{B}} \left\{ \left(\frac{\|h\|_{\ell_2}}{r} - 1 \right) \left\{ \sup_{h_r \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h_r(X_i) - c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} \right. \right.$$

$$\left. \left. - \inf_{h_r \in \mathcal{H}_r \cap r\mathcal{S}} \left\{ c' \frac{1}{n} \sum_{i=1}^n h_r^2(X_i) \right\} \right\} \right\}$$

$$\le \sup_{h \in \mathcal{H} \backslash r\mathcal{B}} \left\{ \left(\frac{\|h\|_{\ell_2}}{r} - 1 \right) \left(\kappa' r^2 - c'(1-\epsilon)r^2 \right) \right\} \le 0$$

Thus the excess loss is upper bounded by the offset Rademacher process, and the latter is further bounded by the process restricted within the critical radius:

$$\sup_{h \in \mathcal{H}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} \le \sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\}$$

$$\le \alpha_n^2(\mathcal{H}, c'(1-\epsilon), \delta)$$

with probability at least $1 - 2\delta$.

$\blacksquare$

Proof [Proof of Theorem 17] Denote $\mathcal{F} \subset \mathcal{G} = \mathcal{F} + \text{star}(\mathcal{F} - \mathcal{F})$. The minimax excess loss can be written as

$$\inf_{\hat{g} \in \mathcal{G}} \sup_P \left\{ \mathbb{E}(\hat{g} - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2 \right\}$$

$$= \inf_{\hat{g} \in \mathcal{G}} \sup_P \left\{ \left\{ -\mathbb{E}2Y\hat{g} + \mathbb{E}\hat{g}^2 \right\} + \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}2Yf - \mathbb{E}f^2 \right\} \right\}.$$

Now let's construct a particular distribution P in the following way: take any $x_1, x_2, ..., x_{(1+c)n} \in \mathcal{X}$ and let P_X be the uniform distribution on these $(1 + c)n$ points. For any $\epsilon = (\epsilon_1, \ldots, \epsilon_{(1+c)n}) \in \{\pm 1\}^{(1+c)n}$, denote the distribution P_ϵ of (X, Y) indexed by ϵ to be: X is sampled from P_X, and $Y_{|X=x_i} = \epsilon_i$, $\forall 1 \le i \le (1+c)n$. Note here $\hat{g} : (X, Y)^{\otimes n} \to \mathcal{F} + \text{star}(\mathcal{F} - \mathcal{F})$. Now we proceed

with this particular distribution

$$\inf_{\hat{g}\in\mathcal{G}} \sup_{P} \left\{ \left\{ -\mathbb{E}2Y\hat{g} + \mathbb{E}\hat{g}^2 \right\} + \sup_{f\in\mathcal{F}} \left\{ \mathbb{E}2Yf - \mathbb{E}f^2 \right\} \right\}$$

$$\geq \inf_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \left\{ \left\{ -\mathbb{E}2Y\hat{g} + \mathbb{E}\hat{g}^2 \right\} + \sup_{f\in\mathcal{F}} \left\{ \mathbb{E}2Yf - \mathbb{E}f^2 \right\} \right\}$$

$$\geq \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \left\{ \sup_{f\in\mathcal{F}} \frac{1}{(1+c)n} \left\{ \sum_{i=1}^{(1+c)n} 2\epsilon_i f(x_i) - f(x_i)^2 \right\} \right\}$$

$$- \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \left\{ 2\mathbb{E}Y\hat{g} - \mathbb{E}\hat{g}^2 \right\}.$$

Note that the first term is exactly $\mathfrak{R}^{\mathrm{o}}((1+c)n, \mathcal{F})$. Let us upper bound the second term. Denote the indices of a uniform n samples from $(1+c)n$ samples $\{x_i\}_{i=1}^{(1+c)n}$ with replacement as $i_1, i_2, \ldots, i_n$, and I be the set of unique indices $|I| \leq n$. Observe that $\hat{g}$ is a function of (x_I, Y_I) only, independent of $\epsilon_j, j \notin I$.

$$\sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \left\{ 2\mathbb{E}Y\hat{g} - \mathbb{E}\hat{g}^2 \right\}$$

$$\leq \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \mathbb{E}_{i_1,\ldots,i_n} \left\{ \frac{1}{(1+c)n} \sum_{i=1}^{(1+c)n} \left\{ 2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2 \right\} \right\}$$

$$= \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_{i_1,\ldots,i_n} \mathbb{E}_\epsilon \left\{ \frac{1}{(1+c)n} \sum_{i=1}^{(1+c)n} \left\{ 2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2 \right\} \right\} \tag{33}$$

Conditionally on $i_1, i_2, \ldots, i_n$,

$$\frac{1}{(1+c)n} \sum_{i\notin I} \left\{ 2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2 \right\} = 0 - \frac{1}{(1+c)n} \sum_{i\notin I} \hat{g}(x_i)^2 < 0.$$

Expression in (33) is upper bounded by

$$\sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_{i_1,\ldots,i_n} \mathbb{E}_\epsilon \left\{ \frac{1}{(1+c)n} \sum_{i\in I} \left\{ 2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2 \right\} \right\}$$

$$\leq \sup_{\hat{g}\in\mathcal{G}} \mathbb{E}_{i_1,\ldots,i_n} \sup_{\{x_i\}_{i=1}^{|I|}\in\mathcal{X}^{\otimes|I|}} \mathbb{E}_\epsilon \left\{ \frac{1}{(1+c)n} \sum_{i\in I} \left\{ 2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2 \right\} \right\}$$

$$\leq \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{n}\in\mathcal{X}^{\otimes n}} \mathbb{E}_\epsilon \left\{ \frac{1}{(1+c)n} \sum_{i=1}^{cn} \left\{ 2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2 \right\} \right\}$$

$$\leq \sup_{\{x_i\}_{i=1}^{n}\in\mathcal{X}^{\otimes n}} \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left\{ \frac{1}{(1+c)n} \sum_{i=1}^{cn} \left\{ 2\epsilon_i g(x_i) - g(x_i)^2 \right\} \right\}$$

$$= \frac{c}{1+c} \mathfrak{R}^{\mathrm{o}}(cn, \mathcal{G}).$$

Thus the claim holds.

∎

Proof [Proof of Lemma 15] From Lemma 12, we know for $\mathcal{H} = \mathcal{F} - f^* + \text{star}(\mathcal{F} - \mathcal{F})$,

$$\log \mathcal{N}_2(\mathcal{H}, 8\epsilon) \leq \log \mathcal{N}_2(\mathcal{F} - f^*, 4\epsilon) + \log \mathcal{N}_2(\text{star}(\mathcal{F} - \mathcal{F}), 4\epsilon) \leq \log \frac{2}{\epsilon} + 3\log \mathcal{N}_2(\mathcal{F}, \epsilon).$$

Consider the δ-covering net of $\mathcal{H}$, where for any $h \in \mathcal{H}$, $v[h]$ is the closest point on the net.

$$\frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i h(X_i) - Ch(X_i)^2 \right\}$$

$$\leq \frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i (h(X_i) - v[h]) - C(h(X_i)^2 - v[h]^2) \right\} + \frac{1}{n} \sup_{v \in \mathcal{N}_2(\mathcal{H}, \delta)} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i v - Cv^2 \right\}$$

$$\leq 2\left(\sqrt{\sum_{i=1}^{n} \xi_i^2/n} + 2C\right) \cdot \delta + \frac{1}{n} \sup_{v \in \mathcal{N}_2(\mathcal{H}, \delta)} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i v - Cv^2 \right\}.$$

The second term is the offset Rademacher for a finite set of cardinality at most $\log(16/\delta) + 3\log N$, thus applying Lemma 9,

$$\mathbb{E}_\epsilon \frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i h(X_i) - Ch(X_i)^2 \right\} \leq \inf_{\delta > 0} \left\{ K \cdot \delta + M \cdot \frac{3\log N + \log(16/\delta)}{n} \right\}$$

$$\leq \tilde{C} \cdot \frac{\log(N \vee n)}{n}$$

where $K := 2\left(\sqrt{\sum_{i=1}^{n} \xi_i^2/n} + 2C\right)$ and M is defined in Equation (6). We also have the high probability bound via Lemma 9:

$$\mathbb{P}_\epsilon \left(\frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{n} 2\epsilon_i \xi_i h(X_i) - Ch(X_i)^2 \right\} \leq \tilde{C} \cdot \frac{\log(N \vee n) + u}{n} \right) \leq e^{-u}.$$

∎