Journal of Machine Learning Research 25 (2024) 1-43 Submitted 1/23; Revised 6/24; Published 11/24

Exponential Tail Local Rademacher Complexity Risk Bounds Without the Bernstein Condition

Varun Kanade *Department of Computer Science, University of Oxford*

Patrick Rebeschini *Department of Statistics, University of Oxford*

Tomas Vaškevičius[*]

Editor: Ambuj Tewari

Abstract

The local Rademacher complexity framework is one of the most successful general-purpose toolboxes for establishing sharp excess risk bounds for statistical estimators based on empirical risk minimization. However, applying this toolbox typically requires using the Bernstein condition, which often restricts the applicability domain to convex and proper settings. Recent years have witnessed several examples of problems where optimal statistical performance is only achievable via non-convex and improper estimators originating from aggregation theory, including the fundamental problem of model selection. These examples are currently outside the reach of the classical local Rademacher complexity theory.

In this work, we build upon the recent approach to localization via offset Rademacher complexities, for which a general high-probability theory has yet to be established. Our main result is an exponential-tail offset Rademacher complexity excess risk upper bound that yields results at least as sharp as those obtainable via the classical theory. However, our bound applies under an estimator-dependent geometric condition (the "offset condition") instead of the estimator-independent (but, in general, distribution-dependent) Bernstein condition on which the classical theory relies. Our results apply to improper prediction regimes not directly covered by the classical theory, such as optimal model selection aggregation for arbitrary classes (including infinite and non-convex classes), and early-stopping/iterative regularization; the Bernstein condition does not hold in both examples.

Keywords: excess risk bounds, local Rademacher complexity, improper learning, Bernstein condition, concentration inequalities, aggregation

1. Introduction

We study the problem of obtaining statistical performance estimates for a general class of prediction procedures. Let $S_n = (X_i, Y_i)_{i=1}^n$ denote an i.i.d. sample of input-output pairs $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ distributed according to some *unknown* distribution P. A function mapping $\mathcal{X}$ to $\mathcal{Y}$ is called a *predictor*. A *statistical estimator* is a procedure mapping the observed random sample S_n to some predictor $\widehat{f} = \widehat{f}(S_n) \in \mathcal{F}$, where the class $\mathcal{F}$ is called the *range* of the estimator $\widehat{f}$. In order to measure the quality of an estimator $\widehat{f}$, we introduce a *loss*

∗. This work was completed while the author was affiliated with the Department of Statistics, University of Oxford, and the Institute of Mathematics, EPFL.

function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ and define the performance measure called *risk* as follows:

$$R(\widehat{f}) = \mathbf{E}_{(X,Y)\sim P}[\ell(\widehat{f}(X), Y)|S_n].$$

The above performance measure is absolute, and its scale depends on the properties of the loss function ℓ as well as the distribution P. In order to obtain a performance measure whose value can approach zero as the sample size n increases, it is customary to introduce a class of *reference predictors* $\mathcal{G}$. The risk incurred by the estimator $\widehat{f}$, relative to the smallest risk achievable via predictors in class $\mathcal{G}$, is called *excess risk* and it is defined as

$$\mathcal{E}_P(\widehat{f}, \mathcal{G}) = R(\widehat{f}) - \inf_{g\in\mathcal{G}} R(g).$$

Observe that we have not imposed any restrictions on the distribution P, other than constraining it to be supported on $\mathcal{X} \times \mathcal{Y}$. Such a setting is sometimes called *agnostic*, *distribution-free*, *model-free* or *misspecified*, and it has been a central object of study in Statistical Learning Theory since the early works of Vapnik and Chervonenkis (1968, 1971, 1974). This setup should be contrasted with the *well-specified* setting, where the reference class of functions $\mathcal{G}$ is taken to be $\mathcal{F}$, the range of the estimator $\widehat{f}$, and the observations are assumed to follow the distribution $Y_i = f(X_i) + \xi_i$ for some $f \in \mathcal{F}$ and zero-mean noise ξ_i. The present paper focuses on obtaining excess risk bounds that hold for *any* distribution P supported on $\mathcal{X} \times \mathcal{Y}$, henceforth referred to as the *agnostic* setting.[1]

1.1 Expected and High-Probability Bounds

Upper bounds on the excess risk $\mathcal{E}_P(\widehat{f}, \mathcal{G})$ can be obtained either in *expectation* or in *deviation*. The former type of bounds aims to find the smallest remainder term $\Delta_{\mathbf{E}}(n, \mathcal{G})$ that depends on properties of the estimator $\widehat{f}$ such as its range $\mathcal{F}$ so that for some universal constant $c > 0$ the following holds:

$$\mathbf{E}_{S_n}\big[\mathcal{E}_P(\widehat{f}, \mathcal{G})\big] \le c\Delta_{\mathbf{E}}(n, \mathcal{G}).$$

Similarly, bounds in deviation aim to find the smallest remainder term $\Delta_{\mathbf{Pr}}$ that depends on properties of the estimator $\widehat{f}$ so that the following holds for any $\delta \in (0, 1]$:

$$\mathbf{P}_{S_n}\big(\mathcal{E}_P(\widehat{f}, \mathcal{G}) > c'\Delta_{\mathbf{Pr}}(n, \mathcal{G}, \delta)\big) \le \delta,$$

where $c' > 0$ is some universal constant. Observe that bounds of the above type can be transformed to in-expectation bounds via tail integration arguments; hence, obtaining sharp excess risk bounds that hold with high probability is typically a more challenging problem than obtaining in-expectation guarantees. If the remainder term $\Delta_{\mathbf{Pr}}(n, \mathcal{G}, \delta)$ is of order $\log(1/\delta)$ as a function of δ, we call such guarantees *exponential tail* bounds.

1. While our setup is sometimes called distribution-free, notice that we are not entirely free of assumptions on the distribution P. In particular, we constrain its support to the set $\mathcal{X}\times\mathcal{Y}$, which can be considered a distributional assumption. In what follows, we shall use the term "agnostic" to highlight that beyond the support set $\mathcal{X} \times \mathcal{Y}$ the observations are otherwise free of any modelling assumptions. We will typically take $\mathcal{Y}$ to be a compact subset of $\mathbb{R}$, however, in some settings, we may allow taking $\mathcal{X} = \mathbb{R}^d$, thus being completely free of any assumptions on the distribution of the covariates (see Section 1.3.2).

Several frameworks have been developed for obtaining both types of statistical performance guarantees. One of the simplest ways to obtain sharp in-expectation guarantees without imposing strong distributional assumptions is via *average stability* (or *leave-one-out*) arguments (Rogers and Wagner, 1978; Devroye and Wagner, 1979; Haussler, Littlestone, and Warmuth, 1994). Among other approaches are in-expectation guarantees obtainable via stochastic approximation arguments (e.g., (Robbins and Monro, 1951; Walk and Zsidó, 1989; Nemirovski, Juditsky, Lan, and Shapiro, 2009; Dieuleveut and Bach, 2016)), or by transporting regret bounds from the framework of prediction of individual sequences (Cesa-Bianchi and Lugosi, 2006) to the stochastic setting via an online-to-batch conversion (e.g., (Cesa-Bianchi, Conconi, and Gentile, 2004; Audibert, 2009)).

Recently, there has been a growing interest in obtaining sharp excess risk bounds that hold with high probability. One challenge in converting in-expectation guarantees to in-deviation counterparts is that, typically, simply applying concentration tools results in extra deviation terms of order $n^{-1/2}$. Consequently, stochastic conversions of "fast rate" in-expectation guarantees of order n^{-1} are converted to in-deviation guarantees with the "slow rate" $n^{-1/2}$. To preserve optimal rates, stochastic conversions need to be performed via probabilistic tools capable of taking some notion of variance into account (e.g., Bernstein's inequality) while, *at the same time*, extinguishing the resulting variance terms by exploiting curvature of the loss function, or imposed "niceness" (e.g., low noise) assumptions on the underlying data-generating distribution. While this conversion has been carried out successfully in a few important cases of interest, as we are going to describe below, the wide applicability of this machinery is limited as typically either the variance terms are too large or because properly bounding them comes at the price of introducing restrictive assumptions.

For the class of *uniformly stable* algorithms (which is a more restrictive notion than average stability; see the work by Bousquet and Elisseeff (2002)), "fast rate" excess risk bounds that hold with high-probability were recently obtained by Klochkov and Zhivotovskiy (2021), while for online-to-batch conversions see the work by Kakade and Tewari (2009) and the references therein. In terms of probabilistic tools, the former work builds on the notion of (weakly) self-bounding functions (Boucheron, Lugosi, and Massart, 2000; Maurer, 2006), while the latter relies on the tail bound for martingales due to Freedman (1975). However, both works cited above impose strong assumptions on the loss function – assumptions that we will not use in the theory we are going to develop in this paper. These assumptions are typically not satisfied in classical settings of interest, such as in the case of regression with the squared loss. Specifically, these works assume that the loss function is strongly convex when the domain of the loss function is taken to be the parameter space of the predictors. For example, in the setting of linear regression with quadratic loss, such an assumption would amount to restrictions on the smallest eigenvalue of the empirical covariance matrix.

One of the most successful general-purpose tool for obtaining sharp excess risk upper bounds is the *local Rademacher complexity* (Bartlett, Bousquet, and Mendelson, 2005; Koltchinskii, 2006, 2011). This approach automatically comes with exponential-tail in-deviation guarantees due to the underlying mathematical machinery resting on a powerful concentration bound for controlling the supremum of empirical processes due to Talagrand (1994, 1996). At the same time, (localized) Rademacher averages are relatively simple to

upper bound, with many settings of interest covered in the existing literature; for some examples, see the textbook by Wainwright (2019, Chapters 13 and 14).

However, due to technical reasons related to the so-called Bernstein condition (see Section 2.1 for a detailed discussion), local Rademacher complexity bounds are primarily suitable when two conditions hold: the reference class $\mathcal{G}$ is convex and the estimator's range $\mathcal{F}$ is equal to the reference class $\mathcal{G}$. A setup when $\mathcal{F} = \mathcal{G}$ is called *proper*.

Our goal is to extend the classical local Rademacher complexity bounds beyond the analysis of proper procedures. In particular, we aim to obtain exponential-tail excess risk bounds when the reference class $\mathcal{G}$ is possibly non-convex and when the statistical learning procedure is allowed to be improper. Before summarizing our contributions in Section 1.4, in Sections 1.2 and 1.3, we explain the boundedness assumptions that enter our analysis and provide some example problems not covered via the classical local Rademacher complexity theory.

1.2 The Bounded Setting

In this paper, we consider the setting where the response variable Y and all the predictors are bounded, both in the reference class of functions $\mathcal{G}$ and in the range $\mathcal{F}$ of the estimator. However, the space of the covariates $\mathcal{X}$ does not necessarily need to be bounded; see Section 1.3 for an example. In the rest of the paper, we will refer to such a setup as the *bounded setting*.

The bounded setting is classical and has been extensively studied since the early days of Statistical Learning Theory, primarily in the classification setting with the zero-one loss. When learning with the zero-one loss, the empirical risk minimization (ERM) estimator satisfies "slow-rate" excess risk bounds, that is, bounds of order $1/\sqrt{n}$, where n is the sample size. Such bounds can be obtained via classical symmetrization arguments, and they are unimprovable[2] in the agnostic setting (Vapnik, 1998).

Turning to the regression setting, the main one investigated in this paper, there exist settings where the ERM estimator satisfies excess risk guarantees that decay faster than $1/\sqrt{n}$ as a function of the sample size. For example, when the loss function is sufficiently curved and the reference class of functions is convex, the ERM algorithm automatically favours choosing low variance elements. In other words, the combination of convexity of the reference class and curvature of the loss function results in a *localization* effect, effectively reducing the complexity of the problem. This situation is more formally captured via the Bernstein condition; see Section 2 for a more detailed discussion.

The localization phenomenon was successfully exploited to obtain "fast-rate" bounds for empirical risk minimization algorithms via empirical processes theory arguments (van de Geer, 2000; Massart, 2000b), yielding sharp covering number bounds. Subsequently, local Rademacher complexity bounds (Bartlett, Bousquet, and Mendelson, 2005; Koltchinskii, 2006) extended the previous arguments to data-dependent complexity measures that are easier to compute and always result in bounds at least as sharp as those obtainable via

2. When learning with the zero-one loss, it is possible to obtain excess risk bounds that decay faster than $1/\sqrt{n}$ if we impose additional assumptions on the data-generating distribution P. See, for example, the works by Mammen and Tsybakov (1999); Tsybakov (2004); Massart and Nédélec (2006). In the classification context, the Bernstein condition is closely related to these works; see the discussion following the statement of Lemma 13 for further details.

direct covering number analysis. Concerning the *analysis of ERM estimator* in the *bounded and convex setting*, local Rademacher complexities machinery remains the state-of-the-art tool. However, we note that even for bounded and convex problems, ERM is not necessarily optimal and obtaining optimal statistical performance is only possible via improper procedures (Vaškevičius and Zhivotovskiy, 2023), further motivating the need to extend the classical local Rademacher complexity framework beyond the convex and proper setting (i.e., when the Bernstein condition fails).

Let us now discuss two research directions that received a lot of attention following the development of local Rademacher complexities. The first direction continued to build our understanding of the bounded framework through the analysis of model selection aggregation procedures, where the classical local Rademacher complexity theory does not apply due to the non-convexity of the reference class (see Section 1.2.1). The second direction concerns the analysis of learning algorithms in the setting where the data and the prediction functions are allowed to be unbounded and heavy-tailed. This setup significantly departs from the bounded setting considered in this work, and neither setup includes the other as a sub-problem; we briefly discuss the latter research direction in Section 1.2.2.

1.2.1 Model Selection Aggregation

It was noticed in the discussion paper by Tsybakov (2006) that a very natural problem called *model selection aggregation* (Nemirovski, 2000; Tsybakov, 2003) falls outside the scope of the classical local Rademacher complexity theory developed by Bartlett, Bousquet, and Mendelson (2005); Koltchinskii (2006). In this problem, the reference class of functions $\mathcal{G}$ is taken to be a finite set of bounded functions; particularly, it is a non-convex set, and local Rademacher complexity theory does not apply directly. Understanding how to optimally aggregate statistical models constructed from i.i.d. data, e.g., models arising from different tuning parameters, or different statistical estimators, is a fundamental problem in statistics. At the same time, deviation-optimal model selection aggregation procedures have been used to construct computable procedures (not necessarily computationally efficient) to demonstrate the achievability of some statistical minimax lower bounds; see, e.g., (Rakhlin, Sridharan, and Tsybakov, 2017; Mendelson, 2019; Mourtada, Vaškevičius, and Zhivotovskiy, 2022).

One challenge concerning the analysis of optimal model selection aggregation estimators is that only *improper procedures*, i.e., ones for which the range $\mathcal{F}$ is strictly larger than the reference class $\mathcal{G}$, can obtain optimal performance (that is, improperness is *necessary*). Regarding in-expectation bounds, optimal performance is achievable via exponential weights (or progressive mixture) algorithms, with different proofs available in the literature; see, e.g., the works by Catoni (1997); Yang (2000); Vovk (2001); Juditsky, Rigollet, and Tsybakov (2008). However, none of the proofs for the in-expectation optimality of exponential weights algorithm follow traditional strategies based on empirical processes theory, such as those based on local Rademacher complexities; see Section 3.2.2 in the work by Audibert (2010). As it turns out, a successful application of such strategies would be impossible because they would lead to optimal exponential-tail deviation bounds which were shown not to hold by Audibert (2008). Audibert (2008) also proposed a deviation-optimal method for model selection aggregation, called the *star algorithm*. One of the key takeaways from Audibert's

analysis is that the excess risk random variable $\mathcal{E}(\widehat{f}, \mathcal{G})$ can take *negative values* for improper estimators $\widehat{f}$. It follows that, in general, in-expectation guarantees for improper methods cannot be used to derive high-probability bounds because Markov's inequality does not apply. For example, Mourtada, Vaškevičius, and Zhivotovskiy (2022, Theorems 1 and 2) exhibit two different statistical estimators for the problem of linear regression, both of which satisfy expectation-optimal excess risk bounds obtainable via average stability arguments, and both of which incur excess risk lower bounded by an absolute constant, with a constant probability.

The phenomenon concerning deviation-optimality of model selection aggregation estimators has generated a lot of attention; for example, see the works by Lecué and Mendelson (2009); Rigollet (2012); Dai, Rigollet, and Zhang (2012); Lecué and Rigollet (2014); Wintenberger (2017); Bellec (2017) for analysis of different model selection aggregation procedures. More broadly, the analysis of improper statistical estimators is receiving increased attention, as such procedures were shown to be necessary for optimal statistical performance in logistic regression, see (Hazan, Koren, and Levy, 2014; Foster, Kale, Luo, Mohri, and Sridharan, 2018; Mourtada and Gaïffas, 2022), and linear regression, see (Vaškevičius and Zhivotovskiy, 2023; Mourtada, Vaškevičius, and Zhivotovskiy, 2022).

In the bounded setting, the only known local Rademacher complexity analysis of improper procedures, particularly of Audibert's star aggregation algorithm, was shown to be possible by Liang, Rakhlin, and Sridharan (2015), who introduced the notion of *offset Rademacher complexity*. In contrast to earlier results in the literature, offset Rademacher complexity analysis allows us to analyze the star estimator directly applied to *infinite* classes of reference functions without any additional discretization steps used to approximate the infinite class via a finite class. However, the high probability guarantees obtained by Liang, Rakhlin, and Sridharan (2015) hold under a certain lower isometry assumption, leaving open the question of obtaining high probability guarantees in the bounded setting, particularly relevant for model selection aggregation. As a corollary of our main results, we close this gap by providing an exponential-tail offset Rademacher complexity excess risk bound for the star estimator (see Appendix A.2) that holds for infinite reference classes of functions.

1.2.2 Comparison With the Heavy-Tailed/Moment-Equivalence Setting

Within the Statistical Learning Theory framework, learning in the presence of heavy-tailed data has attracted a lot of attention over the past decade, particularly concerning the mean-estimation problem in the direction initiated by Catoni (2012), and regression problems, primarily when learning with the quadratic loss; see, e.g., (Audibert and Catoni, 2011; Mendelson, 2015; Oliveira, 2016; Lugosi and Mendelson, 2019b). The interested reader will find a detailed account of the progress made on both problems in the recent survey by Lugosi and Mendelson (2019a).

While the classical localization theory heavily relies on Talagrand's concentration inequality, a two-sided concentration result, the key insight highlighted by Mendelson (2015) and Oliveira (2016) is that in some cases, *one-sided* concentration arguments suffice to provide learning guarantees. The validity of these *one-sided* concentration inequalities can be ensured by imposing certain *moment-equivalence* conditions on the unknown data generating distribution. Within this setting, localized complexity bounds were first obtained

by Mendelson (2015, 2018) using the *small-ball* method and later by Liang, Rakhlin, and Sridharan (2015) using the slightly stronger *lower-isometry* condition.

The bounded framework considered in this work differs from the moment-equivalence framework described above. Indeed, it is well-known that moment-equivalence type assumptions do not allow for agnostic treatment of the bounded setting. Let us provide a simple example in the case of L_q–L_2 moment equivalence for some $q > 2$, a frequently used assumption in the above-cited papers and follow-up work. A family of random variables satisfies the L_q–L_2 moment-equivalence condition if there exists some absolute constant $c > 0$ such that any random variable X in this family satisfies $\mathbf{E}[|X|^q]^{1/q} \leq c\mathbf{E}[X^2]^{1/2}$. However, it is now easy to see that such an assumption does not include all random variables bounded by one. For example, a Bernoulli random variable with a small enough parameter p will violate the above assumption. For further examples and more extensive discussions highlighting the differences between the two frameworks we refer to the works by Lecué and Mendelson (2013); Oliveira (2016); Saumard (2018); Vaškevičius and Zhivotovskiy (2023).

Finally, we emphasize that in this paper our aim is to extend the scope of classical localization beyond convex setups, addressing the limitation that arises through the Bernstein condition on which the classical arguments rely. Such a limitation was recently addressed in the moment-equivalence framework by Mendelson (2019). The main ingredient that allows us to handle the bounded setting is a moment generating function bound on an offset multiplier process that we obtain in Proposition 7. Consequently, we are able to bypass the steps used in the classical localization arguments that rely on the Bernstein condition to handle variance terms arising through an application of Talagrand's concentration inequality; see Section 2.1 for further details.

1.3 Motivating Examples

To make our discussions more concrete, we will now discuss three example problems where the need to analyze improper estimators arises naturally. These examples provide additional motivation for extending the scope of the classical local Rademacher complexity theory beyond the Bernstein condition. We revisit each of the three problems in Appendix A.

1.3.1 Model Selection Aggregation

As already discussed in Section 1.2.1, in a model selection aggregation problem the reference class of functions is finite, and so, non-convex. Therefore, the Bernstein condition needed to apply the classical local Rademacher complexity bounds fails[3], and in fact, the ERM estimator is a suboptimal procedure for this problem, as well as any other proper procedure (this happens already when the reference class contains only two functions). Moreover, it is worth mentioning that using an ERM estimator over the convex hull of the finite reference class also fails – it results in a "slow-rate" bound of order $1/\sqrt{n}$ as shown by Lecué and Mendelson (2009).

3. We shall remark that the local Rademacher complexity of a finite class coincides with the minimax-optimal rate; however, the issue is that the bound involving the local Rademacher complexity is not applicable to the ERM estimator due to the violation of the Bernstein condition.

1.3.2 Learning Non-Convex Classes

A natural generalization of the model selection aggregation problem is to consider possibly infinite non-convex reference classes of functions. Because the model selection aggregation problem falls outside the scope of the classical localization theory, so does this more general class of problems.

In the bounded framework, general treatment of non-convex reference classes via covering number bounds was recently considered by Rakhlin, Sridharan, and Tsybakov (2017). They construct a rather involved three-stage procedure based on sample splitting and discretization of the original class and show that this procedure obtains minimax-optimal excess risk bounds in terms of empirical entropy of the reference class. In the moment-equivalence framework, localized complexity bounds for non-convex reference classes were obtained by Mendelson (2019). In the bounded framework, the only local Rademacher complexity bound applicable for non-convex reference classes is the offset Rademacher complexity bound due to Liang, Rakhlin, and Sridharan (2015); however, their bound is only valid in expectation.

The need to handle non-convex problems may arise even when the reference class of predictors is convex. For example, let us consider a special case of the problem analyzed by Mourtada, Vaškevičius, and Zhivotovskiy (2022). Let the reference class of functions contain all linear functions $\mathcal{G} = \{\langle w, \cdot\rangle : w \in \mathbb{R}^d\}$ and let P be any distribution supported on $\mathbb{R}^d \times [-1, 1]$. While for arbitrary covariate vectors in $\mathbb{R}^d$ the class of linear functions $\mathcal{G}$ is unbounded, the condition that response variable is supported on the bounded interval $[-1, 1]$ allows us to replace $\mathcal{G}$ with a class of truncated linear functions $\mathcal{G}_{\mathrm{trunc}} = \{\mathrm{trunc}(\langle w, \cdot\rangle) : w \in \mathbb{R}^d\}$, where for any $x \in \mathbb{R}^d$ we define $\mathrm{trunc}(\langle w, x\rangle) = \min\{1, \max\{-1, \langle w, x\rangle\}\}$. The reference class $\mathcal{G}_{\mathrm{trunc}}$ is a class of bounded functions and hence, it falls within the scope of the bounded framework investigated in this paper. Since $\mathcal{G}_{\mathrm{trunc}}$ is non-convex, the classical local Rademacher complexity theory does not apply. However, with the results obtained in our paper we can directly apply the star estimator on the non-convex class $\mathcal{G}_{\mathrm{trunc}}$, without relying on any additional discretization/sample-splitting steps. See Appendix A.2.

We remark that in the example problem described above, the marginal distribution of the covariates P_X is arbitrary. In particular, heavy-tailed distributions, even those that violate moment-equivalence conditions (cf. Section 1.2.2) are allowed within our framework as long as the response variable is bounded. Moreover, because P_X is arbitrary and the covariates are unbounded, any proper procedure (i.e., any procedure that outputs a linear function) can incur arbitrarily large excess risk, as shown by Shamir (2015).

1.3.3 Iterative Regularization

Taking its roots in stochastic approximation (Robbins and Monro, 1951) and the theory of inverse problems (Landweber, 1951), an iterative regularization procedure generates a sequence of predictors $(f_t)_{t\geq 0}$ and outputs an estimator $f_{t^\star}$ based on some *stopping rule* that selects the stopping-time $t^\star$, for example, by running a model-selection procedure on held-out data. Compared to the classical way of regularizing via penalized empirical risk minimization, iterative regularization provides an appealing strategy for carefully balancing statistical/computational cost by considering procedures tailored to generate successive iterates at a low computational cost; see, e.g., (Bühlmann and Yu, 2003; Zhang and Yu, 2005;

Yao, Rosasco, and Caponnetto, 2007; Raskutti, Wainwright, and Yu, 2014; Wei, Yang, and Wainwright, 2019; Kanade, Rebeschini, and Vaškevičius, 2023).

One of the simplest iterative regularization procedures is obtained by running gradient descent on an unregularized empirical risk function. For example, let $\mathcal{G} = \{\langle w, \cdot\rangle : w \in \mathbb{R}^d, \|w\|_2 \leq 1\}$ be a bounded reference class of linear predictors. Let $R_n(w)$ denote the empirical risk of the linear predictor $\langle w, \cdot\rangle$, let $w_0 = 0$, and suppose that the sequence $(w_t)_{t\geq 0}$ is obtained by recursively applying the update-rule $w_{t+1} = w_t - \eta\nabla R_n(w_t)$, where $\eta > 0$ is the step size. The difficulty in applying the classical localized complexity tools to analyze the statistical performance of the suitably stopped iterate $w_{t^\star}$ comes from the fact that the linear function $\langle w_{t^\star}, \cdot\rangle$ does not necessarily belong to the reference class $\mathcal{G}$. Thus, the early-stopped iterate $w_{t^\star}$ can be seen as an improper estimator. On the other hand, it was recently shown by Kanade, Rebeschini, and Vaškevičius (2023) that a large class of iterative procedures can be analyzed via the offset Rademacher complexity theory (see Appendix A.3). Thus, the results obtained in this paper automatically extend the in-expectation offset Rademacher complexity guarantees obtained in that work to their exponential-tail counterparts.

1.4 Paper Outline and Summary of Main Results

In this paper, we obtain *exponential-tail* excess risk upper bounds that hold for a *general class* of estimators satisfying a certain geometric condition that we call the *offset condition* (see Definition 4). This geometric condition can serve as a design principle for statistical estimators that satisfy sharp excess risk guarantees with high probability. In particular, arguments based on convex geometry can be used to establish that such a condition holds for a broad class of known estimators (see the examples in Appendix A). The class of estimators satisfying the geometric condition includes improper learning settings that are not covered by the classical theory of local Rademacher complexities. In the classical setting of empirical risk minimization performed over a convex class under boundedness assumptions, our complexity measure yields results *at least as sharp* as those obtainable by the classical theory of local Rademacher complexities (this is made more precise in Section 3.4). The starting point of our analysis is the work of Liang, Rakhlin, and Sridharan (2015), who were the first to provide an *in-expectation* analysis of the star aggregation algorithm based on *offset Rademacher complexity*, a modified notion of classical localization that arises from the analysis of *offset empirical processes*.

The main contribution of the current paper is obtaining results analogous to the ones achievable via the classical local Rademacher complexity theory, yet applicable under a different set of assumptions. In particular, the main element of the classical theory is an *estimator-independent* Bernstein condition (see Section 2.1 for details) that ensures a linear relationship between the variance and expectation of the excess loss class. In contrast, our results build on an *estimator-dependent* geometric condition, called the offset condition. The theory developed in this paper shows that the offset condition is sufficient to ensure sharp excess risk guarantees for improper estimators. For example, as discussed in Appendix A, any estimator that satisfies the offset condition while outputting a sparse combination of a given finite dictionary of functions attains deviation-optimal excess risk rate for the problem of model selection aggregation, where improperness is necessary for optimality.

The rest of the paper is organized as follows.

- In Section 1.5, we summarize the notation used in this paper.
- In Section 2, we provide background on local Rademacher complexity measures. Section 2.1 contains a sketch of how the classical theory of localization, through its foundation built on Talagrand's concentration inequality, is applicable in regimes where the variance of the excess loss class is controlled by a linear function of its expectation (which results in the use of the Bernstein condition for Lipschitz losses). In Theorem 3, we formulate an excess risk bound guaranteed via the classical theory for empirical risk minimization algorithms under the Bernstein condition. This result serves as a benchmark for our paper, which we aim to match without invoking the Bernstein condition. We achieve this (to the extent quantified in Section 3) by establishing a general machinery of localization via offset Rademacher complexities, the background on which is provided in Section 2.2.
- The main results are presented in Section 3.
  - Section 3.1 contains the definition of the geometric condition (called the offset condition) that serves as our replacement of the Bernstein condition and the definition of offset Rademacher complexity, which is slightly modified from the one appearing in prior work by Liang, Rakhlin, and Sridharan (2015). Specifically, we include additional negative terms, which play an important role in our concentration arguments and in proving that our notion of complexity is never worse than the classical notion of local Rademacher complexities (cf. Lemma 11).
  - Section 3.2 contains a moment generating function bound for offset multiplier empirical processes (Proposition 7), which is the main technical contribution of the present paper. This result serves as our replacement for Talagrand's concentration inequality, on which the classical theory of localization is built. The key feature of our concentration result is the fact that the variance of the supremum of offset multiplier processes is automatically controlled by a linear function of their expectations due to the presence of the negative quadratic terms inside the supremum. In contrast, the classical theory of localization needs to *assume* that a certain variance-expectation relationship holds, as elaborated in Section 2.1. We prove Proposition 7 via an application of an exponential Efron-Stein inequality as discussed in greater detail in Section 5.
  - In Section 3.3, we present our main theorem – an exponential-tail excess risk bound stated in terms of the offset Rademacher complexity (cf. Theorem 8). The key difference from the usual theory of localization is that the estimator-independent Bernstein condition appearing in Theorem 3 is replaced via the estimator-dependent offset condition. We prove Theorem 8 by bounding the Laplace transform of the offset empirical processes (arising through the geometric condition imposed on an estimator) in terms of the Laplace transform of a related offset multiplier empirical process. We then complete the proof via an application of Proposition 7, which provides tight control on the Laplace transform of the obtained offset multiplier process.

  - Further connections between the classical theory and the theory developed in this paper are discussed in Section 3.4. In Lemma 11, we show that the offset Rademacher complexity is at most as large as the classical local Rademacher complexity. Thus, the bounds obtained in our paper, when they apply, are at least as sharp as those obtainable via the classical theory (cf. Corollary 12). Finally, we discuss the sense in which the offset condition can be interpreted as an analogue of the Bernstein condition, when the roles of empirical and population quantities are interchanged. (cf. Lemma 13).

- Appendix A contains several applications of the theory developed in this paper. In Lemma 16, we bound the offset Rademacher complexity of sparse linear classes; in Corollary 17, we show how this bound can be applied for non-linear classes via a change-of-basis argument. As a direct consequence, we show how our theory can yield deviation-optimal bounds for two different model selection aggregation procedures, both of which output a sparse combination of dictionary elements and satisfy the offset condition. Such applications are outside the scope of the classical theory of localization, due to the necessary improperness of optimal estimators, as discussed in the introduction. Finally, we discuss how the analysis of iterative regularization schemes fits within the theory developed in this paper.

- Sections 4, 5 and Appendix B contain the proofs.

1.5 Notation

We denote by P the unknown distribution from which an i.i.d. sample $S_n = (X_i, Y_i)_{i=1}^n$ is drawn, where $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. We denote the marginal distribution on $\mathcal{X}$ by P_X and for the sample $S_n = (X_i, Y_i)_{i=1}^n$, let $S_n^X = (X_i)_{i=1}^n$. An estimator $\widehat{f}$ is a mapping between datasets and some class of predictors $\mathcal{F}$, called the range of the estimator $\widehat{f}$. The loss function is denoted by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. For any function $f : \mathcal{X} \to \mathbb{R}$, denote $\ell_f(X, Y) = \ell(f(X), Y)$. The population risk functional is defined by $R(f) = \mathbf{E}\ell_f(X, Y)$, where the expectation is computed with respect to $(X, Y) \sim P$ and f is always assumed to be measurable. We say that the loss function ℓ is L-Lipschitz in its first argument if for any $y, y_1, y_2 \in \mathcal{Y}$ we have $|\ell(y_1, y) - \ell(y_2, y)| \leq L|y_1 - y_2|$. As a function of the sample S_n, define the empirical risk functional R_n by $R_n(f) = n^{-1} \sum_{i=1}^n \ell_f(X_i, Y_i)$. The function class $\mathcal{F}$ always denotes the range of some estimator, while $\mathcal{G}$ denotes a set of reference functions. We let $g^\star \in \operatorname{argmin}_{g \in \mathcal{G}} R(g)$, assuming without loss of generality that such a function exists; otherwise $g^\star$ could be replaced by some function that is arbitrarily close to attaining $\inf_{g \in \mathcal{G}} R(g)$. For any function class $\mathcal{H}$, denote its star-hull by $\operatorname{star}(\mathcal{H}) = \{\lambda h : h \in \mathcal{H}, \lambda \in [0, 1]\}$, where $(\lambda h)(x) = \lambda h(x)$. We say that a function class $\mathcal{H}$ is star-shaped (around the origin) if $\operatorname{star}(\mathcal{H}) = \mathcal{H}$. For any $\mathcal{F}$ and g, the class $\mathcal{F} - g$ denotes $\{f - g : f \in \mathcal{F}\}$. Finally, we denote by $a \lesssim b$ the existence of some universal constant $c > 0$ such that $a \leq cb$.

2. Background on Local Complexity Measures

This section provides background on local complexity measures. In Section 2.1, we recall the classical notion of local Rademacher averages, developed in the series of works by

Koltchinskii and Panchenko (2000); Koltchinskii (2001); Bartlett, Boucheron, and Lugosi (2002); Lugosi and Wegkamp (2004); Bartlett, Bousquet, and Mendelson (2005); Koltchinskii (2006), among others. In particular, we explain why this theory is primarily applicable in the proper learning setup, and explain how convexity assumptions enter this theory through the so-called Bernstein condition. This paper aims to replace such assumptions and establish a methodology that applies to improper and non-convex problems of interest, such as model selection aggregation. In Section 2.2, we discuss a more recent approach of localization via offset Rademacher complexities, introduced in the statistical context with the quadratic loss by Liang, Rakhlin, and Sridharan (2015); see also (Rakhlin and Sridharan, 2014). The offset Rademacher complexity approach replaces the Bernstein condition with an estimator-dependent offset condition, and thus paves the way to achieve the goals set out in this paper – obtaining sharp *exponential-tail* excess risk guarantees that hold for improper estimators.

2.1 Local Rademacher Complexity

Let $\mathcal{F}$ be the range of some estimator $\widehat{f}$, $\mathcal{G}$ be a reference class of functions, and let $g^\star$ denote any population risk minimizer over the class $\mathcal{G}$, i.e., $g^\star \in \operatorname{argmin}_{f\in\mathcal{G}} R(f)$. The first step in the classical local Rademacher complexity analysis proceeds by noting that

$$\begin{aligned}\mathcal{E}(\widehat{f},\mathcal{G}) &= (R(\widehat{f}) - R(g^\star)) - (R_n(\widehat{f}) - R_n(g^\star)) + (R_n(\widehat{f}) - R_n(g^\star)) \\ &\le \sup_{f\in\mathcal{F}} \{(R(f) - R(g^\star)) - (R_n(f) - R_n(g^\star))\} + (R_n(\widehat{f}) - R_n(g^\star))\end{aligned}$$

The term $R_n(\widehat{f}) - R_n(g^\star)$ is typically controlled by assuming that it is at most 0 almost surely. This is true, for example, if $\widehat{f}$ is an empirical risk minimizer over $\mathcal{F}$ and $\mathcal{G} \subseteq \mathcal{F}$.

The supremum term is controlled via Talagrand's concentration inequality[4] for empirical processes (Talagrand, 1994), a functional Bernstein-type concentration inequality with variance proxy

$$\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left\{ \operatorname{Var}_{(X,Y)\sim P} \left[\ell_f(X,Y) - \ell_{g^\star}(X,Y) \right] \right\}.$$

In particular, denoting $Z = \sup_{f\in\mathcal{F}} \{(R(f) - R(g^\star)) - (R_n(f) - R_n(g^\star))\}$ and letting $c > 0$ be some universal constant, for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have

$$Z \le 2\mathbf{E}Z + c\sqrt{\frac{\sigma^2(\mathcal{F})\log(1/\delta)}{n}} + c\frac{C_\ell \log(1/\delta)}{n}, \tag{1}$$

where C_ℓ is a boundedness constant such that the for any $f \in \mathcal{F}$ and any $(X,Y) \in \mathcal{X}\times Y$ we have $|\ell_f(X,Y) - \ell_{g^\star}(X,Y)| \le C_\ell$.

Let us now informally discuss how the above concentration bound leads to a localization theory via Rademacher complexities. Let $\psi(f, g^\star) \ge 0$ be some measure of distance between the functions f and $g^\star$ (for the sake of this high-level presentation, we ignore the properties that ψ needs to satisfy). The idea of localization is to replace $\mathcal{F}$ in (1) by a localized subset

4. We state a version with absolute constants. Of independent interest, various extensions and refinements of Talagrand's concentration bound are available in the literature; we refer the interested reader to (Ledoux, 1997; Massart, 2000a; Bousquet, 2002; Klein and Rio, 2005; Mendelson, 2010; Lederer and van de Geer, 2014).

$\mathcal{F}(r) = \{f \in \mathcal{F} : \psi(f, g^\star) \leq r\}$ for some radius $r > 0$. The theory of local Rademacher complexities then aims to compute the smallest value of $r > 0$ such that the supremum of the empirical process computed over the localized class $\mathcal{F}(r)$ yields an upper bound on the excess risk of an estimator of interest (typically the empirical risk minimization estimator).

To allow for an explicit control of the variance proxy $\sigma^2(\mathcal{F}(r))$, it is further assumed that for any $f \in \mathcal{F}$, we have $\mathrm{Var}(\ell_f - \ell_{g^\star}) \leq \psi(f, g^\star)$. There are two consequences of the above assumed relation between the variance and the distance function. First, it holds that $\sigma^2(\mathcal{F}(r)) \leq r$. Second, it is possible to obtain a uniform Bernstein-type concentration bound on the excess risk over the full class $\mathcal{F}$, such that for each $f \in \mathcal{F}$, the variance-proxy is proportional to $\sqrt{\psi(f, g^\star)/n}$. For more details and a precise quantification of the above statements we refer to (Wainwright, 2019, Theorem 14.20, Equation 14.51).

When the obtained uniform Bernstein-type concentration bound is applied to the estimator $\widehat{f}$ of interest, we obtain an upper bound on its excess risk in terms of the supremum over a localized class $\mathcal{F}(r)$ (for some radius $r > 0$), and the "slow rate" variance term $\sqrt{\psi(\widehat{f}, g^\star)/n}$. To compensate for this variance term and to obtain a "fast rate" excess risk bound, it is further assumed that for some constant $B > 0$ the following inequality holds for any $f \in \mathcal{F}$: $\psi(f, g^\star) \leq B\mathbf{E}[\ell_f - \ell_{g^\star}]$. Since the left hand side of the above equation is a non-negative distance, the right hand side also needs to be non-negative. This, in turn, constrains us to the settings where $\mathcal{F}$, the range of the estimator of interest, cannot be larger than the reference class $\mathcal{G}$, for otherwise there would exist a data generating distribution P and a function $f \in \mathcal{F}$ such that $\mathbf{E}[\ell_f - \ell_{g^\star}] < 0$.

Summing up the above, the theory of local Rademacher complexities is rooted in the following variance-expectation assumption – a widely used condition in the empirical processes analysis of M-estimators (see, e.g., the works by van de Geer (2000); Massart (2000b)):

$$\mathrm{Var}(\ell_f - \ell_{g^\star}) \leq \psi(f, g^\star) \leq B\mathbf{E}[\ell_f - \ell_{g^\star}] \quad \text{for any } f \in \mathcal{F}. \tag{2}$$

In applications in learning theory, a natural choice for the distance function ψ is a suitably rescaled squared $L_2(P_X)$ norm. Indeed, if the loss function ℓ is C_b-Lipschitz in its first argument, then $\mathrm{Var}(\ell_f - \ell_{g^\star}) \leq C_b^2\mathbf{E}(f(X) - g^\star(X))^2$. Thus, the remaining question is what is the smallest allowed value $r > 0$ such that Talagrand's concentration inequality (1) applied to $\mathcal{F}(r)$ yields an upper bound on the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$. Using a peeling argument applied to a reweighted excess loss class (cf. Bartlett, Bousquet, and Mendelson (2005, Section 3)), this value can be shown to equal a solution to a certain fixed-point equation, leading to the following definition.

Definition 1 (Local Rademacher Complexity) *Let P_X denote any distribution supported on $\mathcal{X}$ and let $\mathcal{H}$ denote any class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. For $r > 0$, let $\mathcal{H}(r) = \{h \in \mathcal{H} : \mathbf{E}_{X \sim P_X}[h(X)^2] \leq r\}$. Let $\sigma = (\sigma_i)_{i=1}^n$ be a sequence of i.i.d. Rademacher (i.e., symmetric and $\{\pm 1\}$-valued) random variables and let $S_n^X = (X_i)_{i=1}^n$ denote n independent random variables distributed according to P_X. Then, for any $\gamma > 0$, the local Rademacher complexity of the class $\mathcal{H}$ is defined by*

$$\mathfrak{R}_n^{loc}(P_X, \mathcal{H}, \gamma) = \inf\left\{ r > 0 : \mathbf{E}_{S_n^X, \sigma}\left[\sup_{h \in \mathcal{H}(\gamma^{-1}r)} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) \right\} \right] \leq r \right\}.$$

It now remains to discuss when the second inequality of (2) holds in an *agnostic*[5] sense (as opposed to, e.g., imposing low-noise assumptions on the underlying distribution, as is frequently done in the classification setting). The primary application domain where this is true is when a function class $\mathcal{F}$ is convex, $g^\star \in \mathcal{G}$ denotes a population risk minimizer over all functions in $\mathcal{F}$ (thus, $\mathcal{F} \subseteq \mathcal{G}$, constraining to study the proper learning setting), and the loss function ℓ is strongly convex in its first argument (cf. Bartlett, Bousquet, and Mendelson (2005, Section 5.2)). The second inequality in (2), when ψ is taken to be the squared $L_2(P_X)$ norm, is often called the *Bernstein condition* (cf. Bartlett and Mendelson (2006)), which we state below.

Definition 2 (Bernstein Condition) *Let P be a distribution supported on $\mathcal{X} \times \mathcal{Y}$ and let ℓ be a loss function with domain $\mathcal{Y} \times \mathcal{Y}$. The tuple $(P, \ell, \mathcal{F}, g^\star)$ satisfies the Bernstein condition with parameter $\gamma > 0$ if the following holds for any $f \in \mathcal{F}$:*

$$\mathbf{E}_{X \sim P_X}(f(X) - g^\star(X))^2 \leq \frac{1}{\gamma}\mathbf{E}_{(X,Y)\sim P}\left[\ell_f(X,Y) - \ell_{g^\star}(X,Y)\right].$$

Summing up all of the above, let us now state a result obtained by Bartlett, Bousquet, and Mendelson (2005). In our notation, it reads as follows.

Theorem 3 (Corollary 5.3 in (Bartlett et al., 2005)) *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $[-b, b]$ for some $b > 0$. Let P be a distribution supported on $\mathcal{X} \times [-b, b]$ and let $g^\star \in \operatorname{argmin}_{g\in\mathcal{G}} R(g)$, where $\mathcal{G}$ is some reference class of functions. Suppose that the following three conditions hold:*

1. *The loss function $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ is C_b-Lipschitz in its first argument;*
2. *The tuple $(P, \ell, \mathcal{F}, g^\star)$ satisfies the Bernstein condition with parameter $\gamma > 0$;*
3. *The function class $\mathcal{F} - g^\star = \{f - g^\star : f \in \mathcal{F}\}$ is star-shaped around 0 (cf. Section 1.5).*

Let $\widehat{f}$ be an estimator such that $R_n(\widehat{f}) - R_n(g^\star) \leq 0$ almost surely. Then, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$, we have

$$\mathcal{E}(\widehat{f}, \mathcal{G}) \leq c_1 C_b \mathfrak{R}_n^{loc}(P_X, \mathcal{F} - g^\star, C_b^{-1}\gamma) + c_2 \frac{(C_b b + C_b^2 \gamma^{-1}) \log(1/\delta)}{n},$$

where $c_1, c_2 > 0$ are universal constants.

Limitations. We conclude this section by briefly summarizing two limitations of the above framework.

The first limitation is its reliance on the Bernstein condition. As already discussed, a natural application domain where this condition holds, together with the condition that $R_n(\widehat{f}) - R_n(g^\star) \leq 0$ almost surely, is when $\mathcal{F} = \mathcal{G}$ and $\mathcal{F}$ is a convex class. Since improper learning settings do not satisfy the Bernstein condition uniformly for all data generating distributions P, Theorem 3 does not easily lend itself to non-convex and improper application

5. Recall that, as discussed in the introduction, the present paper aims to obtain excess risk bounds that hold for *any* distribution P supported on $\mathcal{X} \times \mathcal{Y}$.

domains that arise, for instance, in model selection aggregation or iterative regularization applications (cf. Appendix A). The present paper addresses these limitations (see, in particular, Theorem 8 and example applications in Appendix A).

The second limitation concerns the boundedness assumptions, also present in our work. Such assumptions prevent us from capturing unbounded, and in particular, heavy-tailed problems that have recently received a lot of attention. On the other hand, heavy-tailed problems are typically analyzed under assumptions that do not cover the bounded framework. Thus, the two setups are different and complementary to each other; see Section 1.2.2 for a comparison between the bounded and heavy-tailed settings.

2.2 Offset Rademacher Complexity

We now describe the offset Rademacher complexity approach due to Liang, Rakhlin, and Sridharan (2015), an empirical processes theory-based technique shown to yield agnostic *in-expectation* guarantees for Audibert's star algorithm in the bounded setting[6]. Let us preface the rest of this section by noting that the analysis in the above-cited paper is constrained to the case when ℓ is the quadratic loss, i.e., for any y, y' we have $\ell(y, y') = (y - y')^2$.

Let $\mathcal{G} = \{g_1, \dots, g_m\}$ denote a dictionary of m functions mapping $\mathcal{X} \to [-b, b]$. Then, as discussed in the introduction, any estimator whose range $\mathcal{F}$ is equal to $\mathcal{G}$ (i.e., any proper estimator) can only yield slow excess risk rates of order $n^{-1/2}$ instead of the optimal rate $b^2 \log(m)/n$. Hence, due to the necessary improperness of optimal estimators, the model selection aggregation problem does not easily fit into the classical theory of localization discussed in the previous section. The optimal in-expectation and in-deviation performance is attained by the star estimator $\widehat{f}^{(\mathrm{star})}$ due to Audibert (2008), defined as follows:

$$\widehat{f}^{(\mathrm{star})} = \mathrm{argmin}_{f \in \mathcal{G}, \lambda \in [0,1]} R_n(\lambda \widehat{f}^{(\mathrm{ERM})} + (1-\lambda) f), \text{ where } \widehat{f}^{(\mathrm{ERM})} = \mathrm{argmin}_{f \in \mathcal{G}} R_n(f).$$

Recall that in the above expressions R_n denotes the empirical risk functional.

The key observation of Liang, Rakhlin, and Sridharan (2015, Lemma 1) is that the star estimator satisfies a *deterministic* condition that we state below. For any observed sample $S_n = (X_i, Y_i)_{i=1}^n$, the following holds with a constant $\gamma = 1/18$:

$$R_n(\widehat{f}^{(\mathrm{star})}) - R_n(g^\star) \le -\frac{\gamma}{n} \sum_{i=1}^n (\widehat{f}^{(\mathrm{star})}(X_i) - g^\star(X_i))^2. \tag{3}$$

The above condition can be interpreted as an analogue of the Bernstein condition (cf. Definition 2), with population quantities replaced by its empirical counterparts (see Section 3.4 and Lemma 13); however, the above inequality does not require the estimator $\widehat{f}^{(\mathrm{star})}$ to be proper (in fact, it is improper), nor does it require its range $\mathcal{F}$ to be convex. We defer an extended discussion to Section 3.4.

6. Liang, Rakhlin, and Sridharan (2015) also develop high-probability bounds for heavy-tailed, unbounded classes, as long as a certain lower isometry condition holds.

The condition (3) can be used to upper bound the excess risk as follows:

$$\begin{aligned}
&\mathcal{E}(\widehat{f}^{(\mathrm{star})}, \mathcal{G}) \\
&= (R(\widehat{f}^{(\mathrm{star})}) - R(g^\star)) - (R_n(\widehat{f}^{(\mathrm{star})}) - R_n(g^\star)) + (R_n(\widehat{f}^{(\mathrm{star})}) - R_n(g^\star)) \\
&\le (R(\widehat{f}^{(\mathrm{star})}) - R(g^\star)) - (R_n(\widehat{f}^{(\mathrm{star})}) - R_n(g^\star)) - \gamma \frac{1}{n}\sum_{i=1}^n (\widehat{f}^{(\mathrm{star})}(X_i) - g^\star(X_i))^2 \\
&\le \sup_{f \in \mathcal{F}} \big\{ (R(f) - R(g^\star)) - (R_n(f) - R_n(g^\star)) - \gamma \frac{1}{n}\sum_{i=1}^n (f(X_i) - g^\star(X_i))^2 \big\}.
\end{aligned}$$

Taking expectations on both sides and applying classical symmetrization and contraction arguments, Liang, Rakhlin, and Sridharan (2015, Theorem 3) show that the following holds for some absolute constants $c_1, c_2 > 0$:

$$\mathbf{E}_{S_n} \mathcal{E}(\widehat{f}^{(\mathrm{star})}, \mathcal{G}) \le c_1 b \mathbf{E}_{S_n, \sigma} \left[\sup_{h \in \mathcal{F} - g^\star} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) - \frac{\gamma}{b} h(X_i)^2 \right\} \right], \tag{4}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ denotes a sequence of i.i.d. Rademacher random variables. The right-hand side of the above equation is called the *offset Rademacher complexity* of the class $\mathcal{F} - g^\star$; the negative quadratic terms produce a localization phenomenon similar to that of Definition 1. As we shall see in Corollary 12, a modified notion of the above complexity measure yields guarantees at least as sharp as those obtainable via local Rademacher complexities introduced in the previous section.

Limitations. We now discuss the limitations of the existing results based on the above approach. First, the bound (4) holds only *in-expectation*. However, the star estimator was introduced to address the *in-deviation* optimality for model selection aggregation, and thus, obtaining in-deviation guarantees for this estimator is of particular interest (Audibert, 2008). As discussed in the introduction, transforming in-expectation guarantees to in-deviation guarantees for *improper* statistical estimators presents several technical difficulties. High probability alternatives to the bound (4) have not been obtained before our work since there is no replacement for Talagrand's concentration inequality on which the classical theory of localization resides. We develop such a (one-sided) concentration result in Proposition 7, using which we obtain an exponential-tail offset Rademacher complexity deviation bound in Theorem 8.

While high probability bounds featuring offset Rademacher complexity term have not been previously developed, let us now discuss some deviation bounds that have been obtained using the framework described above. The main high probability result obtained in (Liang, Rakhlin, and Sridharan, 2015, Theorem 4) holds under a certain lower isometry condition, which differs from the bounded setting considered in this work, as discussed in the introduction. Setting aside the difference in assumptions, there is an important qualitative difference between (Liang, Rakhlin, and Sridharan, 2015, Theorem 4) and Theorem 8 proved in this paper. The former result upper bounds excess risk by another random variable, while our theorem features a deterministic quantity (offset Rademacher complexity). The further control on the random variable in the upper bound of (Liang, Rakhlin, and Sridharan, 2015, Theorem 4) typically results in looser bounds (e.g., suffering from excess logarithmic terms for finite classes; see (Liang, Rakhlin, and Sridharan, 2015, Lemma 11)).

The recent work of Vijaykumar (2021) extends the geometric inequality (3) to general loss functions. However, the high probability bounds obtained therein are expressed in terms of empirical covering numbers where the covering is performed with the *worst-case* metric. In contrast, local Rademacher complexity (cf. Definition 1) can be upper bounded using covering number arguments where the covering is performed with the $L_2(P_X)$ metric, leading to minimax optimal bounds in many cases (see Wainwright (2019, Chapters 13 and 14) for some examples). Crucially, in general the notion of complexity based on empirical covering numbers using worst-case metric used by Vijaykumar (2021) does not capture statistical minimax optimality and results in suboptimal bounds even for the star estimator applied to a problem with a *finite* reference class $\mathcal{G}$. In contrast, we show in Appendix A how the geometric inequality obtained by Vijaykumar (2021), when used with offset Rademacher complexity bounds developed in this paper, results in minimax optimal bounds for the star aggregation algorithm.

3. Main Results

The main results of this paper are presented in this section. In Section 3.1, we introduce the geometric condition (called the offset condition) used to replace the Bernstein condition; further, we define the offset Rademacher complexity (slightly modified from the one appearing in prior works) used to replace the classical notion of local Rademacher complexity. Section 3.2 contains a moment generating function bound for shifted multiplier empirical processes. This result serves as our replacement for Talagrand's concentration inequality, the foundation of the classical theory of localization. Section 3.3 contains a high probability excess risk bound in terms of the offset Rademacher complexity; this result applies in settings where the Bernstein condition does not hold. Finally, in Section 3.4, we provide a comparison between the offset and Bernstein conditions and demonstrate that the theory presented in this paper can recover the classical agnostic learning setup bounds overviewed in Section 2.1.

3.1 Definitions

We begin with the definition of the *offset condition*. Observe that this condition is *estimator-dependent*, as opposed to the Bernstein condition (cf. Definition 2). For corresponding notions in the context of improper statistical estimators see, e.g., Liang, Rakhlin, and Sridharan (2015, Lemma 1), Vijaykumar (2021, Section 3), and the analysis of the star estimator by Audibert (2008).

Definition 4 (Offset Condition) *Let $\mathcal{G}$ be a class of functions mapping $\mathcal{X}$ to $[-b, b]$ for some $b > 0$. Fix a loss function $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ and recall that R_n denotes the induced empirical risk functional. Let $\varepsilon : [0, 1] \to \mathbb{R}$ be some function and let $\gamma > 0$ be some positive real number. Let P be a distribution supported on $\mathcal{X} \times \mathcal{Y}$. An estimator $\widehat{f}$ satisfies the* offset condition *with respect to $(\mathcal{G}, \ell, \varepsilon, \gamma)$ for the distribution P, if for any any $\delta \in [0, 1]$ the following holds:*

$$\mathbf{P}_{S_n}\left(R_n(\widehat{f}) - R_n(g^\star) \leq -\frac{\gamma}{n}\sum_{i=1}^{n}(\widehat{f}(X_i) - g^\star(X_i))^2 + \varepsilon(\delta)\right) \geq 1 - \delta,$$

where $S_n = (X_i, Y_i)_{i=1}^n$ is an i.i.d. sample drawn from the distribution P and $g^\star = g^\star(\mathcal{G}, P, \ell)$ denotes any population risk minimizer in the class $\mathcal{G}$.

Whenever the following deterministic inequality holds for any sample $S_n = (X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$:

$$R_n(\widehat{f}) - R_n(g^\star) \leq -\frac{\gamma}{n}\sum_{i=1}^n (\widehat{f}(X_i) - g^\star(X_i))^2 + \varepsilon,$$

we say that the estimator $\widehat{f} = \widehat{f}(S_n)$ satisfies the deterministic offset condition *with respect to $(\mathcal{G}, \ell, \varepsilon, \gamma)$.*

In the above definition the function $\varepsilon(\cdot)$ allows for the offset condition to fail with probability δ, while incurring a penalty $\varepsilon(\delta)$. As we shall see in Appendix A, such a condition naturally enters the analysis of some improper estimators. Also, we will discuss some example estimators that satisfy the deterministic offset condition.

In Section 2.1, we described how the Bernstein condition implies local Rademacher complexity excess risk bounds for empirical risk minimization estimators. Likewise, we shall see that offset condition implies excess risk bounds expressed in terms of the offset Rademacher complexity defined below.

Definition 5 (Offset Rademacher Complexity) *Let P_X be any distribution supported on $\mathcal{X}$ and let $\mathcal{H}$ be any class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Let $\sigma = (\sigma_i)_{i=1}^n$ denote a sequence of i.i.d. Rademacher (i.e., symmetric and $\{\pm 1\}$-valued) random variables and let $S_n^X = (X_i)_{i=1}^n$ denote n independent random variables distributed according to P_X. Then, for any $\gamma > 0$, the offset Rademacher complexity of the class $\mathcal{H}$ is defined by*

$$\mathfrak{R}_n^{off}(P_X, \mathcal{H}, \gamma) = \mathbf{E}_{S_n^X, \sigma}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) - \gamma h(X_i)^2 - \gamma \mathbf{E}_{X\sim P_X}[h(X)^2]\right\}\right].$$

Let us remark that our definition above differs from the one presented in Section 2.2 since we include extra negative terms $-\gamma\mathbf{E}_{X\sim P_X}[h(X)^2]$ inside the above supremum. This refinement is necessary for our concentration argument to work, since we establish moment bounds for shifted multiplier processes that contain negative population terms (cf. Section 3.2). At the same time, the inclusion of the negative quadratic population terms allows us to show that the above notion of complexity is at least as sharp as the classical one introduced in Definition 1 (see Lemma 11 in Section 3.4 for details).

Remark 6 *The introduction of the negative term $-\gamma\mathbf{E}_{X\sim P_X}[h(X)^2]$ is a necessary component for our proof of Proposition 7 and for the proof of Lemma 11, where we show that the offset Rademacher complexity is never larger than the classical local Rademacher complexity.*

On the other hand, when computing specific application-dependent upper bounds on the offset Rademacher complexity, either the term $-\gamma\mathbf{E}_{X\sim P_X}[h(X)^2]$ or $-\gamma\frac{1}{n}\sum_{i=1}^n h(X_i)^2$ may be dropped. Keeping the term $-\gamma\mathbf{E}_{X\sim P_X}[h(X)^2]$ may be interpreted as performing localizing with respect to population norms $\|\cdot\|_{L_2(P_X)}$ while keeping the empirical term $-\gamma\frac{1}{n}\sum_{i=1}^n h(X_i)^2$ may be seen as localization with respect to the empirical L_2 norms. In the bounded setting, the two forms of localization are known to be equivalent; see (Wainwright, 2019, Section 14.5) for details.

3.2 Concentration of Shifted Multiplier Processes

The primary technical tool in this paper is the following proposition, which proves a Bernstein-type one-sided concentration bound for the supremum of shifted multiplier processes (defined below in Equation (5)). This proposition plays a crucial role in establishing our main result, Theorem 8 presented in the next section. In particular, provided that an estimator satisfies the offset condition, we will show that the moment generating function of its excess risk can be controlled by the moment generating function of a certain shifted multiplier process. We defer the proof of the below proposition to Section 5.

Proposition 7 *Let $\mathcal{H}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Let $P_{(X,\zeta)}$ be a joint distribution on $\mathcal{X} \times \mathbb{R}$ with marginal distributions P_X and P_ζ, and let $S_n = (X_i, \zeta_i)_{i=1}^n$ be a set of n i.i.d. samples from $P_{(X,\zeta)}$. Fix any positive constant $\gamma > 0$ and define a random variable $U = U(S_n)$ to be the supremum of the offset multiplier process as follows:*

$$U = \sup_{h \in \operatorname{star}(\mathcal{H})} \left\{ \sum_{i=1}^{n} \zeta_i h(X_i) - \mathbf{E}_{(X,\zeta)\sim P_{(X,\zeta)}}[\zeta h(X)] - \gamma h(X_i)^2 - \gamma \mathbf{E}_{X \sim P_X}[h(X)^2] \right\}. \quad (5)$$

Suppose that there exist positive constants κ and σ such that $\sup_{h\in\mathcal{H}} \|h\|_{L_\infty(P_X)} \leq \kappa$ and $\|\zeta\|_{L_\infty(P_\zeta)} \leq \sigma$. Then, for $\eta = 8(\sigma^2\gamma^{-1} + \gamma\kappa^2)$ and any $\lambda \in (0, 1/\eta)$ the following holds:

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \leq \frac{\lambda^2 \eta \mathbf{E}U}{2(1 - \eta\lambda)}.$$

Before turning to the offset Rademacher complexity upper bounds, let us remark that in the above moment bound, the variance proxy/variance factor (in the sense of (Boucheron et al., 2013, Section 2.4)) is equal to $\eta\mathbf{E}U$; thus the variance of the random variable U is automatically controlled by its expectation. In particular, the above bound can be transformed into deviation bounds of the form $U \leq 2\mathbf{E}[U] + c\eta \log(1/\delta)$, where $\delta > 0$ is the confidence parameter. To prove the above bound, we leverage a type of self-boundedness property (through an application of the Exponential Efron-Stein inequality) directly satisfied by the above supremum process (see the proof for more details). In contrast, recall that the variance proxy in Talagrand's concentration inequality (1) is not controlled by the expectation of the corresponding empirical process. In the classical localized complexity bounds, Talagrand's inequality is combined with the Bernstein condition and an intricate peeling argument to induce a similar self-bounding effect. On the other hand, using the above concentration result, our theory allows us to obtain high probability bounds in terms of the offset Rademacher complexity without relying on the Bernstein condition as we show in the following section.

3.3 Exponential-Tail Offset Rademacher Complexity Bound

We now present the main result of this paper, the proof of which can be found in Section 4. The following theorem provides an alternative to Theorem 3, but with Bernstein condition replaced via the offset condition. As a consequence, the offset condition can serve as a design principle for estimators in the regimes where the Bernstein condition fails to hold; some examples are given in Appendix A.

Theorem 8 *Let $\widehat{f}$ be an estimator with range $\mathcal{F}$, where $\mathcal{F}$ denotes a class of functions mapping $\mathcal{X}$ to $[-b,b]$ for some $b>0$. Let P be any distribution supported on $\mathcal{X}\times[-b,b]$ and denote $g^\star \in \operatorname{argmin}_{g\in\mathcal{G}} R(g)$, where $\mathcal{G}$ is some reference class of functions mapping $\mathcal{X}$ to $[-b,b]$. Suppose that the following two conditions hold:*

1. *The loss function $\ell:[-b,b]\times[-b,b]\to[0,\infty)$ is C_b-Lipschitz in its first argument;*
2. *The estimator $\widehat{f}$ satisfies the offset condition with respect to $(\mathcal{G},\ell,\varepsilon,\gamma)$ for the distribution P, where ε is some function mapping $[0,1]$ to $\mathbb{R}$ and $\gamma>0$ is some positive real number.*

Then, for any $\delta_1,\delta_2\in(0,1)$ with probability at least $1-\delta_1-\delta_2$, we have

$$\mathcal{E}(\widehat{f},\mathcal{G}) \le c_1 C_b' \mathfrak{R}_n^{off}(P_X, \operatorname{star}(\mathcal{F}-g^\star), (C_b')^{-1}\gamma) + c_2 \frac{\gamma^{-1}(C_b')^2 \log(1/\delta_1)}{n} + \varepsilon(\delta_2),$$

where $c_1,c_2>0$ are some universal constants and $C_b' = C_b + \gamma b$.

Observe that in the above theorem, to upper bound the excess risk of the estimator $\widehat{f}$, we pay for the complexity of its range $\mathcal{F}$ as opposed to the complexity of the reference class $\mathcal{G}$. Let us comment on why such behaviour is expected in our setup.

First, as discussed in Section 1.3, there are natural problems of interest where any estimator whose range equals $\mathcal{G}$ (i.e., any proper estimator) is bound to incur sub-optimal excess risk rate. Thus, we are led to consider improper estimators, i.e., estimators such that $\mathcal{G}\subset\mathcal{F}$. In the above theorem, we further restrict our attention to those estimators that satisfy the offset condition. For this family of estimators, their range $\mathcal{F}$ necessarily serve as a natural proxy of their complexity.

For example, let $\mathcal{G}$ be a finite reference class of functions. Then, Audibert's star estimator (see Appendix A.1) satisfies the offset condition while having a relatively small range $\{\lambda g_1 + (1-\lambda)g_2 : g_1,g_2\in\mathcal{G}, \lambda\in[0,1]\}$. This range is small enough to preserve minimax-optimal statistical rates for the model selection aggregation problem. On the other hand, performing empirical risk minimization over the convex hull of $\mathcal{G}$ is also an estimator that satisfies the offset condition. However, the convex hull of $\mathcal{G}$ is a much larger set than the range of the star estimator, and its local Rademacher complexity yields the "slow-rate" excess risk bound of order $1/\sqrt{n}$. Indeed, Lecué and Mendelson (2009) prove a matching lower bound for performing ERM over the convex hull of a finite reference class of functions.

Remark 9 *As explained above, when the only property of an estimator exploited in the analysis is the offset condition, then paying for the complexity of the estimator's range (as opposed to the reference class) is unavoidable. However, as we discuss in Appendix A, there exist several improper estimators of interest that satisfy the offset condition with range $\mathcal{F}$ whose complexity is of the same order of magnitude as that of the reference class $\mathcal{G}$, yielding minimax-optimal rates (e.g., for model selection aggregation of arbitrary classes). Moreover, when applied to the ERM estimator over a convex class, the bound of Theorem 8 reduces to the classical local Rademacher complexity guarantee discussed in Section 2.1. From this point of view, we may view Theorem 8 as an extension of the classical localization theory to a broader class of estimators.*

On the other hand, there exist statistical estimators whose excess risk is not governed solely by the complexity of their range. For example, the Q-Aggregation estimator (Lecué and Rigollet, 2014) outputs a function from the convex hull of a given finite class, yet it satisfies a deviation-optimal "fast-rate" excess risk guarantee for the model selection aggregation problem. This is possible because the Q-Aggregation estimator satisfies a more restrictive condition than the offset condition. More specifically, it satisfies an offset-type condition with a different negative term than the one stated in Definition 4. See the PhD thesis Vaškevičius (2021, Section 3.7) for more details on this example and for suggestions for future work.

Remark 10 *In comparison with Theorem 3, the above result replaces C_b with a worse constant $C_b' = C_b + \gamma b$. However, the primary application domain where the above theorems hold is the setting where for any $y \in [-b, b]$, the function $\ell(\cdot, y)$ is C_b-Lipschitz and γ-strongly convex in the fist argument (see Appendix A for examples). In such a setting it can be shown that $\gamma b \leq C_b$ and hence $C_b' \leq 2C_b$.*

3.4 Recovering Local Rademacher Complexity Results Without The Bernstein Condition

In this section, we discuss how Theorem 8 yields excess risk bounds that are no worse than the ones stated in Theorem 3. We begin by stating the following lemma, which is proved in Appendix B.1.

Lemma 11 *Let P_X be any distribution supported on $\mathcal{X}$ and let $\mathcal{H}$ be any star-shaped class of functions (i.e., $\mathcal{H} = \operatorname{star}(\mathcal{H})$) mapping $\mathcal{X}$ to $\mathbb{R}$. Then, for any $\gamma > 0$ we have*

$$\mathfrak{R}_n^{off}(P_X, \mathcal{H}, \gamma) \leq \mathfrak{R}_n^{loc}(P_X, \mathcal{H}, \gamma).$$

An immediate consequence of the above lemma is the following corollary, which shows that the classical local Rademacher complexity bounds hold when the Bernstein condition is replaced via the *estimator-dependent* offset condition.

Corollary 12 *Consider the setting of Theorem 8. For any $\delta_1, \delta_2 \in (0,1)$ with probability at least $1 - \delta_1 - \delta_2$, we have*

$$\mathcal{E}(\widehat{f}, \mathcal{G}) \leq c_1 C_b' \mathfrak{R}_n^{loc}(P_X, \operatorname{star}(\mathcal{F} - g^\star), (C_b')^{-1}\gamma) + c_2 \frac{\gamma^{-1}(C_b')^2 \log(1/\delta_1)}{n} + \varepsilon(\delta_2),$$

where $c_1, c_2 > 0$ are some universal constants and $C_b' = C_b + \gamma b$.

It remains to discuss the relationship between the offset and Bernstein conditions. A typical example where the Bernstein condition holds for *any* distribution P is when $\mathcal{F} = \mathcal{G}$ is a convex class, and the loss function is strongly convex. In such regimes, any empirical risk minimizer over $\mathcal{F}$ satisfies the offset condition. Thus, when applied to an empirical risk minimization estimator, the offset condition can be seen as an analogue of the Bernstein condition, where the roles played by empirical and population quantities are interchanged. We formalize this observation in the lemma below.

Lemma 13 *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a loss function and let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ be the set of all distributions P supported on $\mathcal{X} \times \mathcal{Y}$. Let $f^\star = f^\star(\mathcal{F}, P, \ell)$ be any population risk minimizer over $\mathcal{F}$. Let $\widehat{f}^{(ERM)}$ be an estimator that returns any empirical risk minimizer in the class $\mathcal{F}$. If for any $P \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ the tuple $(P, \ell, \mathcal{F}, f^\star)$ satisfies the Bernstein condition with parameter γ, then the estimator $\widehat{f}^{(ERM)}$ satisfies the deterministic offset condition with respect to $(\mathcal{F}, \ell, 0, \gamma)$.*

Proof Given an i.i.d. sample $S_n = (X_i, Y_i)_{i=1}^n$ from some distribution $P \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, let P_n denote a distribution on $\mathcal{X} \times \mathcal{Y}$ assigning equal mass to each (X_i, Y_i). Since $P_n \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, by the assumption of this lemma $(P_n, \ell, \mathcal{F}, \widehat{f}^{(ERM)}(S_n))$ satisfies the Bernstein condition with parameter γ. This is equivalent to saying that $\widehat{f}^{(\mathrm{ERM})}$ satisfies the deterministic offset condition with respect to $(\mathcal{F}, \ell, 0, \gamma)$. ■

Let us conclude this section by highlighting one difference between the offset and Bernstein conditions. In some settings, the Berstein condition is used as a *distributional* assumption, which imposes constraints on the data distribution itself – as opposed to *agnostic* learning results, required to hold for any data-generating distribution subject to constrained support. For example, in the classification setting with zero-one loss, the Bernstein condition corresponds to bounded noise assumptions (see the discussions in (Boucheron, Bousquet, and Lugosi, 2005)), under which empirical risk minimization estimator can achieve fast rates of convergence of the excess risk. For sharp treatment of the classification setting under the bounded noise assumptions via ideas related to offset Rademacher averages, see (Zhivotovskiy and Hanneke, 2018). At the same time, let us remark that the offset condition can be exploited to design statistical estimators that achieve fast rates in the classification setting in an agnostic sense (i.e., without bounded noise assumptions), provided an option to abstain from prediction exists; for an extended discussion see (Bousquet and Zhivotovskiy, 2021).

4. Proof of Theorem 8

Recall that P denotes the underlying distribution of (X, Y) and let P_n denote its empirical counterpart supported on the sample S_n so that

$$P\ell = \mathbf{E}_{(X,Y)\sim P}[\ell(X, Y)] \quad \text{and} \quad P_n\ell = \frac{1}{n}\sum_{i=1}^n \ell(X_i, Y_i) \quad \text{for any function } \ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R};$$

$$Ph = \mathbf{E}_{X\sim P_X}[h(X)] \quad \text{and} \quad P_nh = \frac{1}{n}\sum_{i=1}^n h(X_i) \quad \text{for any function } h : \mathcal{X} \to \mathbb{R}.$$

With the above notation we have $R(f) = P\ell_f$ and $R_n(f) = P_n\ell_f$. Denote the event

$$E_{\delta_2} = \{P_n\ell_{\widehat{f}} - P_n\ell_{g^\star} \le -\gamma P_n(\widehat{f} - g^\star)^2 + \varepsilon(\delta_2)\}$$

Since $\widehat{f}$ satisfies the $(\mathcal{G}, \ell, \varepsilon, \gamma)$-offset condition we have $\mathbf{P}(E_{\delta_2}) \geq 1 - \delta_2$; on E_{δ_2} we have

$$\begin{aligned} P\ell_{\widehat{f}} - P\ell_{g^\star} &= (P - P_n)(\ell_{\widehat{f}} - \ell_{g^\star}) + P_n(\ell_{\widehat{f}} - \ell_{g^\star}) \\ &\leq (P - P_n)(\ell_{\widehat{f}} - \ell_{g^\star}) - \gamma P_n(\widehat{f} - g^\star)^2 + \varepsilon(\delta_2) \\ &\leq \underbrace{\sup_{f \in \mathcal{F}} \left\{ (P - P_n)(\ell_f - \ell_{g^\star}) - \gamma P_n(f - g^\star)^2 \right\}}_{:= Z} + \varepsilon(\delta_2). \end{aligned}$$

The rest of the proof is structured as follows:

1. We first symmetrize a suitably rearranged Laplace transform of the empirical offset process Z. Since for $\lambda \geq 0$ the map $x \mapsto e^{\lambda x}$ is convex and non-decreasing, this step of the proof follows via standard arguments.

2. Next, we apply Talagrand's Contraction Lemma to the symmetrized offset empirical process. This step turns our process into a multiplier-type process of Proposition 7.

3. We conclude the proof via an application of Proposition 7, which yields a Bernstein-type upper bound on the moment generating function of $Z - \mathfrak{R}_n^{\text{off}}(\text{star}(\mathcal{H}), \gamma')$, for a suitably defined constant $\gamma' > 0$. The desired tail bound then follows via Markov's inequality.

Remark 14 *Our proof strategy is inspired by the work of Lecué and Rigollet (2014), where symmetrization and contraction arguments are also performed on the Laplace transform of the empirical process of interest. The contraction step is needed there to make the corresponding complexity measure linear in the model parameters so that the supremum over a convex hull is attained at a vertex. In contrast, we need to apply the contraction step to put us in the setting of Proposition 7.*

Symmetrization step. We begin by rewriting the random variable Z as follows:

$$\begin{aligned} Z &= \sup_{f \in \mathcal{F}} \left\{ (P - P_n)\left(\ell_f - \ell_{g^\star}\right) - \gamma P_n(f - g^\star)^2 \right\} \\ &= \sup_{f \in \mathcal{F}} \left\{ (P - P_n)\left(\ell_f - \ell_{g^\star} + \frac{3\gamma}{4}(f - g^\star)^2 \right) - \frac{\gamma}{4} P_n(f - g^\star)^2 - \frac{3\gamma}{4} P(f - g^\star)^2 \right\}, \quad (6) \end{aligned}$$

where in the last equation above we have added and subtracted $(3\gamma/4)P(f - g^\star)^2$. For any function $f \in \mathcal{F}$ introduce a shorthand notation

$$\phi_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \text{ such that } \phi_f(X, Y) = \ell_f(X, Y) - \ell_{g^\star}(X, Y) + \frac{3\gamma}{4}(f(X) - g^\star(X))^2.$$

Let $S_n' = (X_i', Y_i')_{i=1}^n$ denote an independent copy of $S_n = (X_i, Y_i)_{i=1}^n$ and denote $\mathbf{E}'$ as a shorthand notation for expectation computed with respect to S_n' only, conditionally on all other random variables. Let P_n' denote a counterpart to P_n with the sample S_n replaced by

S_n'. Carrying on from equation (6) we can rewrite Z as follows:

$$\begin{aligned}
Z &= \sup_{f\in\mathcal{F}}\left\{(P-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{3\gamma}{4}P(f-g^\star)^2\right\}\\
&= \sup_{f\in\mathcal{F}}\left\{(P-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{\gamma}{4}P(f-g^\star)^2 - \frac{2\gamma}{4}P(f-g^\star)^2\right\}\\
&= \sup_{f\in\mathcal{F}}\left\{(\mathbf{E}'P_n'-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{\gamma}{4}\mathbf{E}'P_n'(f-g^\star)^2 - \frac{2\gamma}{4}P(f-g^\star)^2\right\}. \quad (7)
\end{aligned}$$

Observe that in the above equation we have left the term $(2\gamma/4)P(f-g^\star)$ unchanged. This is needed to put us in the setting of Proposition 7, as we shall see below.

Let us now introduce a sequence of n independent Rademacher (symmetric and $\{\pm1\}$ valued) random variables σ_i and let $\mathbf{E}_\sigma$ denote expectation with $\sigma_1,\dots,\sigma_n$ only, conditionally on all other random variables. Let P_n^σ denote the symmetrized empirical measure so that for any function $\ell:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ and any function $h:\mathcal{X}\to\mathbb{R}$ we have

$$P_n^\sigma\ell = \frac{1}{n}\sum_{i=1}^n \sigma_i\ell(X_i,Y_i) \quad \text{and} \quad P_n^\sigma h = \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i).$$

For $\lambda>0$ the map $x\mapsto e^{\lambda x}$ is convex and non-decreasing; hence, for any $\lambda>0$, using the identity (7), we can proceed to symmetrize the Laplace transform of Z as follows:

$$\begin{aligned}
\mathbf{E}\exp(\lambda Z) &\le \mathbf{E}\mathbf{E}'\exp\left(\lambda\sup_{f\in\mathcal{F}}\left\{(P_n'-P_n)\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2\right.\right.\\
&\qquad\left.\left. - \frac{\gamma}{4}P_n'(f-g^\star)^2 - \frac{2\gamma}{4}P(f-g^\star)^2\right\}\right)\\
&\le \mathbf{E}\mathbf{E}_\sigma\exp\left(2\lambda\sup_{f\in\mathcal{F}}\left\{P_n^\sigma\phi_f - \frac{\gamma}{4}P_n(f-g^\star)^2 - \frac{\gamma}{4}P(f-g^\star)^2\right\}\right). \quad (8)
\end{aligned}$$

Notice that the above moment generating function is almost of the form that can be bounded via Proposition 7. It remains to replace the term $P_n^\sigma\phi_f$ with a term $\rho P_n^\sigma(f-g^\star)$, for some constant ρ. This is the aim of the contraction step of this proof, which follows below.

Contraction step. Recall that by the assumptions of this theorem, there exists some constant C_b such that for any $f,f'\in\mathcal{F}, x\in\mathcal{X}, y\in\mathcal{Y}$ we have

$$|\ell_f(x,y)-\ell_{f'}(x,y)| \le C_b|f(x)-f'(x)|.$$

In particular, for any $f,f'\in\mathcal{F}$ and any $x\in\mathcal{X}, y\in\mathcal{Y}$ we have

$$\begin{aligned}
|\phi_f(x,y)-\phi_{f'}(x,y)| &= \left|\ell_f(x,y) + \frac{3\gamma}{4}(f(x)-g^\star(x))^2 - \ell_{f'}(x,y) - \frac{3\gamma}{4}(f'(x)-g^\star(x))^2\right|\\
&\le C_b|f(x)-f'(x)| + \frac{3\gamma}{4}|(f(x)-f'(x))(f(x)+f'(x)-2g^\star(x))|\\
&\le (C_b+3\gamma b)|f(x)-f'(x)|\\
&= (C_b+3\gamma b)|(f(x)-g^\star(x))-(f'(x)-g^\star(x))|.
\end{aligned}$$

Hence, applying Talagrand's contraction inequality (Ledoux and Talagrand, 2013, Theorem 4.12) (conditionally on the sample S_n) with the set T_{S_n} and contraction mappings $\phi^{(i)}_{S_n}$:

$$T_{S_n} = \{((f-g^\star)(X_1),\dots,(f-g^\star)(X_n))^\mathsf{T} : f \in \mathcal{H}\},$$
$$\phi^{(i)}_{S_n}(t_i) = (2C_b+6\gamma b)^{-1}\cdot 2\left(\ell(t_i+g^\star(X_i),Y_i) - \ell_{g^\star}(X_i,Y_i) - \frac{3\gamma}{4}t_i^2\right),$$

we may proceed upper bounding (8) as follows (cf. (Lecué and Rigollet, 2014, Eq. (3.11))):

$$\begin{aligned}
&\mathbf{E}\exp(\lambda Z)\\
&\le \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda \sup_{f\in\mathcal{F}}\left\{P_n^\sigma 2\phi_f - \frac{\gamma}{2}P_n(f-g^\star)^2 - \frac{\gamma}{2}P(f-g^\star)^2\right\}\right)\\
&\le \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda \sup_{f\in\mathcal{F}}\left\{(2C_b+6\gamma b)P_n^\sigma(f-g^\star) - \frac{\gamma}{2}P_n(f-g^\star)^2 - \frac{\gamma}{2}P(f-g^\star)^2\right\}\right)\\
&= \mathbf{E}\mathbf{E}_\sigma \exp\left(\lambda \sup_{h\in\mathcal{H}}\left\{(2C_b+6\gamma b)P_n^\sigma h - \frac{\gamma}{2}P_n h^2 - \frac{\gamma}{2}P h^2\right\}\right)\\
&\le \mathbf{E}\mathbf{E}_\sigma \exp\Bigg(\frac{\lambda}{n}\cdot \underbrace{n \sup_{h\in\mathrm{star}(\mathcal{H})}\left\{(2C_b+6\gamma b)P_n^\sigma h - \frac{\gamma}{2}P_n h^2 - \frac{\gamma}{2}P h^2\right\}}_{:=U}\Bigg),
\end{aligned}$$

where in the penultimate line we introduced $\mathcal{H} = \{f-g^\star : f\in\mathcal{F}\}$, and in the last step the inequality comes from replacing $\mathcal{H}$ by $\mathrm{star}(\mathcal{H}) = \{\lambda h : h\in\mathcal{H}, \lambda\in[0,1]\}$.

We will now show that the random variable U is a supremum of an offset multiplier process satisfying the conditions of Proposition 7. Let $\zeta_i = (2C_b+6\gamma b)\sigma_i$ and denote the distribution of ζ by P_ζ. Then, for any $h\in\mathcal{H}$ and for (X,ζ) distributed according to the product distribution $P_X\otimes P_\zeta$, we have $\mathbf{E}[\zeta h(X)] = 0$. Therefore,

$$\begin{aligned}
U &= n\cdot \sup_{h\in\mathrm{star}(\mathcal{H})}\left\{(2C_b+6\gamma b)P_n^\sigma h - \frac{\gamma}{2}P_n h^2 - \frac{\gamma}{2}P h^2\right\}\\
&= \sup_{h\in\mathrm{star}(\mathcal{H})}\left\{\sum_{i=1}^n \zeta_i h(X_i) - \mathbf{E}_{(X,\zeta)\sim P_X\otimes P_\zeta}[\zeta h(X)] - \frac{\gamma}{2}h(X_i)^2 - \frac{\gamma}{2}\mathbf{E}_{X\sim P_X}h(X)^2\right\}.
\end{aligned}$$

Hence, the moment generating function of the random variable U can be bounded via Proposition 7, taking $P_{(X,\zeta)} = P_X\otimes P_\zeta$.

Concluding the proof. Let $c_3>0$ be some universal constant such that

$$\eta = 8((2C_b+6\gamma b)^2(\gamma/2)^{-1} + (\gamma/2)4b^2) \le c_3(\gamma^{-1}C_b^2 + bC_b + \gamma b^2).$$

Relabelling λ/n by λ and applying Proposition 7 to the random variable U, the following holds for any $\lambda\in(0,1/\eta)$:

$$\log\mathbf{E}\exp(\lambda((nZ)-\mathbf{E}\mathbf{E}_\sigma U)) \le \log\mathbf{E}\mathbf{E}_\sigma\exp(\lambda(U-\mathbf{E}\mathbf{E}_\sigma U)) \le \frac{\lambda^2\eta\mathbf{E}\mathbf{E}_\sigma U}{2(1-\eta\lambda)}. \tag{9}$$

The desired tail bound now follows via standard arguments that we sketch below. By (Boucheron, Lugosi, and Massart, 2013, Section 2.4), the upper bound (9) shows that the

random variable $nZ - \mathbf{E}\mathbf{E}_\sigma U$ is sub-gamma on the right-tail with variance proxy $\eta \mathbf{E}\mathbf{E}_\sigma U$ and scale parameter η. Hence, via Markov's inequality, for any $\delta_1 \in (0,1]$ we have

$$\mathbf{P}\left(nZ - \mathbf{E}\mathbf{E}_\sigma[U] \geq \sqrt{2\eta \mathbf{E}\mathbf{E}_\sigma[U]\log(\delta^{-1})} + \eta\log(\delta_1^{-1})\right) \leq \delta_1.$$

Subtracting $\mathbf{E}\mathbf{E}_\sigma[U]$ from both sides of the inequality defining the event inside $\mathbf{P}(\cdot)$ and optimizing the quadratic function in $\sqrt{\mathbf{E}\mathbf{E}_\sigma[U]}$, we deduce that

$$\begin{aligned}
\delta_1 &\geq \mathbf{P}\left(nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \geq \sqrt{2\eta \mathbf{E}\mathbf{E}_\sigma[U]\log(\delta_1^{-1})} - \mathbf{E}\mathbf{E}_\sigma[U] + \eta\log(\delta_1^{-1})\right) \\
&\geq \mathbf{P}\left(nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \geq \sup_{x\in\mathbb{R}}\left\{x\sqrt{2\eta\log(\delta^{-1})} - x^2\right\} + \eta\log(\delta_1^{-1})\right) \\
&= \mathbf{P}\left(nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \geq (3/2)\eta\log(\delta_1^{-1})\right).
\end{aligned}$$

Thus, denoting the event

$$E_{\delta_1} = \{nZ - 2\mathbf{E}\mathbf{E}_\sigma[U] \leq (3/2)\eta\log(\delta_1^{-1})\}$$

we have $\mathbf{P}(E_{\delta_1}) \geq 1 - \delta_1$. Finally, observe that

$$\begin{aligned}
\mathbf{E}_{S_n}\mathbf{E}_\sigma U &= n(2C_b + 6\gamma b)\mathfrak{R}_n^{\mathrm{off}}\left(P_X, \mathrm{star}(\mathcal{H}), \frac{\gamma}{2}\cdot(2C_b+6\gamma b)^{-1}\right) \\
&\leq 74\cdot n(C_b+\gamma b)\mathfrak{R}_n^{\mathrm{off}}\left(P_X, \mathrm{star}(\mathcal{H}), \gamma\cdot(C_b+\gamma b)^{-1}\right).
\end{aligned}$$

The desired result follows by the union bound on the events E_{δ_1} and E_{δ_2}. ■

5. Proof of Proposition 7

Let us first discuss the key insight into our proof. Without loss of generality, assume that the supremum in the definition of the random variable U (cf. (5)) is always attained by some function, and denote this (random) function by $\tilde{h} = \tilde{h}(S_n)$. The following lemma shows that the empirical and population L_2 norms of $\tilde{h}$ are upper bounded by $c^{-1}U$. Thus, intuitively the supremum over $\mathrm{star}(\mathcal{H})$ in the multiplier process is computed over a "self-localized" (in a random/data-dependent way) subset of $\mathrm{star}(\mathcal{H})$. In contrast, we remark that the classical theory of localization via fixed-point equations proceeds by localizing the function class $\mathrm{star}(\mathcal{H})$ by constraining it to an *explicitly* chosen subset of functions with small L_2 population or empirical norms.

Lemma 15 *Consider the setting of Proposition 7 and let $\tilde{h} = \tilde{h}(S_n)$ denote a random function that attains the supremum of the offset multiplier process U (cf. (5)) given the sample $S_n = (X_i, \zeta_i)_{i=1}^n$. That is, $\tilde{h}$ satisfies*

$$\sum_{i=1}^n\left(\zeta_i\tilde{h}(X_i) - \mathbf{E}[\zeta\tilde{h}(X)|S_n] - \gamma\tilde{h}(X_i)^2 - \gamma\mathbf{E}[\tilde{h}(X)^2|S_n]\right) = U(S_n).$$

Then, the following deterministic inequality holds for any realization of S_n:

$$\sum_{i=1}^n\left(\mathbf{E}[\tilde{h}(X)^2|S_n] + \tilde{h}(X_i)^2\right) \leq \frac{1}{\gamma}U(S_n).$$

Proof Fix any realization $S_n = (X_i, \zeta_i)_{i=1}^n$ and in the rest of this proof we work conditionally on S_n. For any $h \in \mathrm{star}(\mathcal{H})$, define $A(h)$ and $B(h)$ as follows:

$$A(h) = \sum_{i=1}^{n} \left(\zeta_i h(X_i) - \mathbf{E}[\zeta h(X)|S_n]\right), \quad B(h) = \gamma \sum_{i=1} \left(\mathbf{E}[h(X)^2|S_n] + h(X_i)^2]\right).$$

Thus, since $\tilde{h} = \tilde{h}(S_n)$ denotes a maximizer of the offset multiplier process, we have

$$A(\tilde{h}) - B(\tilde{h}) = \sup_{h \in \mathrm{star}(\mathcal{H})} (A(h) - B(h)) = U(S_n). \tag{10}$$

For any $\lambda \in [0, 1)$, let $\lambda h : x \mapsto \lambda h(x)$. Observe that for any h and λ, the term $A(\lambda h)$ scales *linearly* as a function of λ (i.e., $A(\lambda h) = \lambda A(h)$), while the term $B(\lambda h)$ scales *quadratically* (i.e., $B(\lambda h) = \lambda^2 B(h)$) as a function of λ. Fix any $\lambda \in [0, 1)$ and note that by the definition of star-hulls, the function $\lambda\tilde{h}$ is in the set $\mathrm{star}(\mathcal{H})$. Therefore, the identity (10) implies that

$$\lambda A(\tilde{h}) - \lambda^2(B(\tilde{h})) = A(\lambda\tilde{h}) - B(\lambda\tilde{h}) \leq \sup_{h \in \mathrm{star}(\mathcal{H})} (A(h) - B(h)) = U(S_n). \tag{11}$$

Rearranging the identity (10) we also have $A(\tilde{h}) = U(S_n) + B(\tilde{h})$, which plugged into the left hand side of (11) yields

$$\lambda(1 - \lambda)B(\tilde{h}) \leq (1 - \lambda)U(S_n).$$

Dividing both sides by $(1-\lambda) > 0$ shows that $\lambda B(\tilde{h}) \leq U(S_n)$. Since the last equation holds for any $\lambda \in [0, 1)$ it follows that $B(\tilde{h}) \leq U(S_n)$ which completes the proof of this lemma. ■

With the above lemma in place, we are ready to prove Proposition 7. In the below proof, we follow the standard approach for obtaining Bernstein-type concentration bounds for the supremum of empirical processes (see (Boucheron, Lugosi, and Massart, 2013, Section 12.2)). In particular, such bounds often build on the entropy method, which in our case appears through an application of the exponential Efron-Stein inequality. For a survey of tail bounds on the supremum of empirical processes, see the bibliographic remarks in (Boucheron, Lugosi, and Massart, 2013, Chapter 12). We now introduce some additional notation.

1. Let $S_n^{(i)}$ be equal to the sample S_n with the i-th element (X_i, ζ_i) replaced by an independent copy $(X_i', \zeta_i') \sim P_{(X,\zeta)}$.

2. For $i = 1, \ldots, n$, let $U_i' = U(S_n^{(i)}))$. Thus U_i' is the supremum of the offset multiplier process computed on the sample $S_n^{(i)}$, which differs from S_n by the i-th sample only.

3. Let $\mathbf{E}'[\cdot] = \mathbf{E}[\cdot|S_n]$ denote the expectation computed with respect to the random variables (X_i', ζ_i') only. In particular, we have $\mathbf{E}'[U] = U$.

The exponential Efron-Stein inequality (Boucheron, Lugosi, and Massart, 2013, Theorem 6.16) asserts that for $\theta > 0$ and any $\lambda \in (0, 1/\theta)$ we have

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbf{E} e^{\lambda V^+/\theta}, \quad \text{where} \quad V^+ = \sum_{i=1}^{n} \mathbf{E}'[(U - U_i')_+^2]. \tag{12}$$

To complete the proof of Proposition 7, it remains to upper bound the random variable V^+. This will be achieved via a combination of Lemma 15 and boundedness assumptions on the function class $\mathcal{H}$ and the multipliers ζ. Indeed, let $\tilde{h} = \tilde{h}(S_n)$ be a function that attains the supremum in the definition of U (cf. Lemma 15) Then, evaluating the multiplier process defined on the sample $S_n^{(i)}$ with the function $\tilde{h}$ yields a lower bound on U_i. Therefore, for $i = 1, \dots, n$ we have

$$U - U_i' \le \zeta_i \tilde{h}(X_i) - \gamma \tilde{h}(X_i)^2 - \zeta_i' \tilde{h}(X_i') + \gamma \tilde{h}(X_i')^2$$

and hence,

$$(U - U_i')_+^2 \le \left(\zeta_i \tilde{h}(X_i) - \gamma \tilde{h}(X_i)^2 - \zeta_i' \tilde{h}(X_i') + \gamma \tilde{h}(X_i')^2\right)^2.$$

Noting that for any $a, b, c, d \in \mathbb{R}$ we have $(a+b+c+d)^2 \le 4a^2 + 4b^2 + 4c^2 + 4d^2$ (for example, by the Cauchy-Schwarz inequality) it follows that

$$\begin{aligned}
\mathbf{E}'[(U - U_i')_+^2] &\le 4\mathbf{E}'[\zeta_i^2 \tilde{h}(X_i)^2 + \gamma^2 \tilde{h}(X_i)^4 + \zeta_i'^2 \tilde{h}(X_i')^2 + \gamma^2 \tilde{h}(X_i')^4] \\
&\le 4\mathbf{E}'[(\sigma^2 + \gamma^2\kappa^2)(\tilde{h}(X_i)^2 + \tilde{h}(X_i')^2)] \\
&\le 4(\sigma^2 + \gamma^2\kappa^2)(\tilde{h}(X_i)^2 + \mathbf{E}[\tilde{h}(X)^2 | S_n]),
\end{aligned}$$

where the second line follows by the boundedness assumptions and the last line follows by noting that $\tilde{h}(X_i)$ depends on S_n only and renaming X_i' to X. Hence, we can now obtain an upper bound on V^+ defined in (12) via Lemma 15 as follows:

$$0 \le V^+ \le 4(\sigma^2 + \gamma^2\kappa^2) \sum_{i=1}^n \left(\tilde{h}(X_i)^2 + \mathbf{E}[\tilde{h}(X)^2 | S_n]\right) \le 4(\sigma^2\gamma^{-1} + \gamma\kappa^2) U$$

Plugging the above upper bound on V^+ into the exponential Efron-Stein inequality (12) with the choice $\theta = 4(\sigma^2\gamma^{-1} + \gamma\kappa^2)$ yields, for any $\lambda \in (0, 1/\theta)$:

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \le \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbf{E} e^{\lambda U} = \frac{\lambda\theta}{1 - \lambda\theta} \left(\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} + \lambda \mathbf{E} U\right).$$

Rearranging the above inequality, we obtain

$$\frac{1 - 2\lambda\theta}{1 - \lambda\theta} \log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \le \frac{\lambda^2 \theta \mathbf{E} U}{1 - \lambda\theta}.$$

For any $\lambda \in (0, 1/(2\theta))$ we have $(1 - 2\lambda\theta)/(1 - \lambda\theta) > 0$, thus for $\lambda \in (0, 1/(2\theta))$ we have

$$\log \mathbf{E} e^{\lambda(U - \mathbf{E}U)} \le \frac{\lambda^2 \theta \mathbf{E}[U]}{1 - 2\lambda\theta} = \frac{\lambda^2 (\eta \mathbf{E} U)}{2(1 - \eta\lambda)},$$

where $\eta = 2\theta$. This finishes our proof. ■

Acknowledgments

Tomas Vaškevičius would like to thank Jaouad Mourtada and Nikita Zhivotovskiy for many discussions related to high probability excess risk bounds.

Varun Kanade and Patrick Rebeschini were supported in part by the Alan Turing Institute under the EPSRC grant EP/N510129/1. Tomas Vaškevičius was supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). Patrick Rebeschini was also partially funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number EP/Y028333/1).

References

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

Jean-Yves Audibert. *PAC-Bayesian aggregation and multi-armed bandits*. Hdr thesis, Université Paris-Est, 2010.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher Complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Pierre C Bellec. Optimal exponential bounds for aggregation of density estimators. *Bernoulli*, 23(1):219–248, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16(3):277–292, 2000.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet and Nikita Zhivotovskiy. Fast classification rates without standard margin assumptions. *Information and Inference: A Journal of the IMA*, 2021.

Peter Bühlmann and Bin Yu. Boosting with the l2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Olivier Catoni. A mixture approach to universal model selection. Technical report, École Normale Supérieure, 1997.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, volume 75, pages 167–208, 2018.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209, 2014.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. The statistical complexity of early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12(4): 3010–3041, 2023.

Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 34, 2021.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Vladimir Koltchinskii. Local Rademacher Complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

Louis Landweber. An iteration formula for Fredholm integral equations of the first kind. *American journal of mathematics*, 73(3):615–624, 1951.

Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

Johannes Lederer and Sara van de Geer. New concentration inequalities for suprema of empirical processes. *Bernoulli*, 20(4):2020–2038, 2014.

Michel Ledoux. On talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, 1:63–87, 1997.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through Offset Rademacher Complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *The Journal of Machine Learning Research*, 17(1):2718–2755, 2016.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019a.

Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019b.

Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000a.

Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000b.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Shahar Mendelson. Empirical processes with a bounded ψ_1 diameter. *Geometric and Functional Analysis*, 20(4):988–1027, 2010.

Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3): 1–25, 2015.

Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1):459–502, 2018.

Shahar Mendelson. An unrestricted learning procedure. *Journal of the ACM (JACM)*, 66 (6):1–42, 2019.

Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23(31):1–49, 2022.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 2022.

Arkadi Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Arkadiĭ Nemirovsky and David Yudin. *Problem complexity and method efficiency in optimization.* Wiley, New York, 1983.

Roberto Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4):1175–1194, 2016.

Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3806–3832. PMLR, 15–19 Aug 2021.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Philippe Rigollet. Kullback-leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, pages 639–665, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Adrien Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018.

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

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

A. B. Tsybakov. Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2681 – 2687, 2006. doi: 10.1214/009053606000001064.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Alexandre B Tsybakov. Optimal rates of aggregation. *Conference on Learning Theory*, pages 303–313, 2003.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

Vladimir Vapnik and Alexey Chervonenkis. Uniform convergence of frequencies of occurence of events to their probabilities. In *Dokl. Akad. Nauk SSSR*, volume 181, pages 781–783, 1968.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

Tomas Vaškevičius and Nikita Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *Bernoulli*, 29(1):473–495, 2023.

Tomas Vaškevičius. *Risk bounds for improper prediction procedures*. PhD thesis, University of Oxford, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Suhas Vijaykumar. Localization, convexity, and star aggregation. *Advances in Neural Information Processing Systems*, 34, 2021.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

H Walk and L Zsidó. Convergence of the robbins-monro method for linear problems in a banach space. *Journal of Mathematical Analysis and Applications*, 139(1):152–177, 1989.

Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 65(10):6685–6703, 2019.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Yuhong Yang. Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74(1):135–161, 2000.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

Nikita Zhivotovskiy and Steve Hanneke. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

Appendix A. Example Applications

In this section, we discuss some applications of our theory to problems where the Bernstein condition does not hold, yet there exist estimators that satisfy the offset condition. As a result, sharp deviation-optimal excess risk rates can be obtained for such estimators via the theory developed in this paper.

For any function class $\mathcal{H}$ mapping $\mathcal{X}$ to $\mathbb{R}$ and any sample $S_n^X = (X_i)_{i=1}^n$, where $X_i \in \mathcal{X}$, define

$$\mathfrak{R}^{\text{off}}(S_n^X, \mathcal{H}, \gamma) = \mathbf{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i) - \gamma h(X_i)^2 \right\} \middle| S_n^X \right],$$

where $\sigma = (\sigma_1, \dots, \sigma_n)$ denotes a sequence of i.i.d. Rademacher random variables. Observe, in particular, that for any distribution P_X supported on $\mathcal{X}$, we have

$$\mathfrak{R}_n^{\text{off}}(P_X, \mathcal{H}, \gamma) \le \mathbf{E}_{S_n^X} \left[\mathfrak{R}^{\text{off}}(S_n^X, \mathcal{H}, \gamma) \right]. \tag{13}$$

Thus, upper bounds on $\mathfrak{R}_n^{\text{off}}(S_n^X, \mathcal{H}, \gamma)$ imply corresponding upper bounds on the offset Rademacher complexity. Let us now state a bound on $\mathfrak{R}_n^{\text{off}}(S_n^X, \mathcal{H}, \gamma)$ for sparse linear classes, which will be used to yield sharp bounds for the examples considered in this section.

Lemma 16 *For any $w \in \mathbb{R}^d$ let $\|w\|_0$ denote the number of non-zero coordinates of w. Denote a class of k-sparse linear predictors by*

$$\mathcal{H}_{lin}^{d,k} = \{\langle w, \cdot \rangle : w \in \mathbb{R}^d, \|w\|_0 \le k\}.$$

Let $S_n^\Phi = (\Phi_i)_{i=1}^n$, where $\Phi_i \in \mathbb{R}^d$ are arbitrary. Then, for any $\gamma > 0$ we have

$$\mathfrak{R}^{off}(S_n^\Phi, \mathcal{H}_{lin}^{d,k}, \gamma) \lesssim \frac{1}{\gamma} \log\left(\frac{ed}{k}\right) \frac{k}{n}.$$

The above lemma is proved in Section B.2 via a direct argument involving comparison inequalities for Rademacher and Gaussian chaos. As an immediate consequence, let us state the following corollary that will simplify the exposition of the applications to follow.

Corollary 17 *Let $\mathcal{G} = \{g_1, \dots, g_m\}$ denote a finite class of arbitrary functions mapping $\mathcal{X}$ to $\mathbb{R}$. For any positive integer $k \in \{1, \dots, m\}$ define the function class containing k-sparse linear combinations of elements of $\mathcal{G}$ by*

$$\mathcal{G}_{lin}^k = \left\{ g_w(\cdot) = \sum_{i=1}^m w_i g_i(\cdot) : w \in \mathbb{R}^d \text{ and } \|w\|_0 \le k \right\}.$$

Let $k_1, k_2 \in \{1, \dots, m\}$, $\mathcal{F} = \mathcal{G}_{lin}^{k_1}$, and fix any $g^\star \in \mathcal{G}_{lin}^{k_2}$. Then, for any distribution P_X supported on $\mathcal{X}$ and for any $\gamma > 0$ we have

$$\mathfrak{R}_n^{off}(P_X, \mathrm{star}(\mathcal{F} - g^\star), \gamma) \lesssim \frac{1}{\gamma} \log\left(\frac{em}{(k_1+k_2)}\right) \frac{(k_1+k_2)}{n}.$$

Proof Let $k = k_1 + k_2$ and note that $\mathrm{star}(\mathcal{F} - g^\star) \subseteq \mathcal{G}_{\mathrm{lin}}^k$. Hence, the bound (13) yields

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathrm{star}(\mathcal{F} - g^\star), \gamma) \le \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{G}_{\mathrm{lin}}^k, \gamma) \le \mathbf{E}_{S_n^X}\left[\mathfrak{R}^{\mathrm{off}}(S_n^X, \mathcal{G}_{\mathrm{lin}}^k, \gamma)\right]. \tag{14}$$

For any sample S_n^X and any $i = 1, \dots, n$ define $\Phi_i^X \in \mathbb{R}^m$ by $(\Phi_i^X)_j = g_j(X_i)$. Then, for any $w \in \mathbb{R}^d$ and $g_w = \sum_{i=1}^m w_i g_i$ we have $g_w(X_i) = \sum_{j=1}^m w_j g_j(X_i) = \langle w, \Phi_i^X \rangle$. Hence, letting $S_n^\Phi(S_n^X) = (\Phi_i^X)_{i=1}^n$ and applying Lemma 16 yields

$$\mathfrak{R}^{\mathrm{off}}(S_n^X, \mathcal{G}_{\mathrm{lin}}^k, \gamma) = \mathfrak{R}^{\mathrm{off}}(S_n^\Phi(S_n^X), \mathcal{F}_{\mathrm{lin}}^{m,k}, \gamma) \lesssim \frac{1}{\gamma} \log\left(\frac{em}{k}\right) \frac{k}{n}.$$

Plugging in the above inequality into (14) completes the proof. ■

We now turn to the example applications.

A.1 Model Selection Aggregation

In a model selection aggregation problem, we are given a finite dictionary $\mathcal{G} = \{g_1, \dots, g_m\}$ of functions mapping $\mathcal{X}$ to $[-b, b]$. Given a sample $S_n = (X_i, Y_i)_{i=1}^n$, a statistical estimator $\widehat{f}$ aims to construct a new function such that the excess risk $\mathcal{E}(\widehat{f}, \mathcal{G})$ is small with high probability.

In what follows, we consider loss functions $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ that are C_b-Lipschitz and γ-strongly convex in the first coordinate. More precisely, we assume that for any $y, y_1, y_2 \in [-b, b]$ we have $|\ell(y_1, y) - \ell(y_2, y)| \le C_b |y_1 - y_2|$ and for any $\lambda \in [0, 1]$ we have $\ell(\lambda y_1 + (1-\lambda) y_2, y) \le \lambda \ell(y_1, y) + (1-\lambda)\ell(y_2, y) - \frac{\gamma}{2}\lambda(1-\lambda)(y_1 - y_2)^2$.

An identical setup to the one described above was recently treated by Lecué and Rigollet (2014); Wintenberger (2017). Optimal model selection aggregation rates $\gamma^{-1} C_b^2 \log(m/\delta)/n$ were obtained therein for the Q-aggregation and online Bernstein aggregation procedures. Below, we show how the offset Rademacher complexity analysis yields the same rates for two other estimators: Audibert's star algorithm and the midpoint estimator.

Audibert's Star Algorithm. The star algorithm due to (Audibert, 2008) is defined by

$$\widehat{f}^{(\mathrm{star})} = \mathrm{argmin}_{f \in \mathcal{G}, \lambda \in [0,1]} R_n(\lambda \widehat{f}^{(\mathrm{ERM})} + (1-\lambda) f), \text{ where } \widehat{f}^{(\mathrm{ERM})} = \mathrm{argmin}_{f \in \mathcal{G}} R_n(f).$$

It was shown by Liang, Rakhlin, and Sridharan (2015, Lemma 1), that this estimator satisfies the deterministic offset condition. In more recent work, Vijaykumar (2021, Proposition 5) shows that $\widehat{f}^{(\mathrm{star})}$ satisfies the $(\mathcal{G}, \ell, 0, \gamma/9)$-deterministic offset condition.

Remark 18 *While in this section we only consider finite reference classes $\mathcal{G}$, the star estimator $\widehat{f}^{(star)}$ satisfies the deterministic offset condition for* arbitrary *classes $\mathcal{G}$; in particular, $\mathcal{G}$ is allowed to be an infinite class. We will return to this point in Appendix A.2 to*

formulate a deviation-optimal local Rademacher complexity excess risk bound for the star estimator for arbitrary reference classes.

In the view of Corollary 17, the range of the star estimator $\widehat{f}^{(\mathrm{star})}$ is equal to $\{\lambda g + (1-\lambda)g' : g, g' \in \mathcal{G}, \lambda \in [0,1]\} \subseteq \mathcal{G}^2_{\mathrm{lin}}$. Thus, combining Theorem 8 (see also Remark 10) and Corollary 17 yields, for any $\delta \in (0,1)$ with probability at least $1-\delta$,

$$\mathcal{E}(f^{(\mathrm{star})}, \mathcal{G}) \lesssim \gamma^{-1} C_b^2 \frac{\log(m/\delta)}{n}.$$

Midpoint Estimator. Let $c_1 > 0$ be some sufficiently large universal constant (as elaborated in the proof of Lemma 19). For any $\delta \in (0,1)$, the midpoint estimator is defined by

$$\widehat{f}^{(\mathrm{mid})}_{\delta} = \mathrm{argmin}_{f \in \mathcal{G}_{\delta,c_1}(S_n)} R_n\left(\frac{\widehat{f}^{(\mathrm{ERM})} + f}{2}\right),$$

where $\widehat{f}^{(\mathrm{ERM})} = \widehat{f}^{(\mathrm{ERM})}(S_n)$ is any function in $\mathcal{G}$ that minimizes the empirical risk $R_n(\cdot)$ (induced by the sample S_n) and the set $\mathcal{G}_{\delta,c_1}(S_n)$ is a random (data-dependent) set of *almost empirical risk minimizers* defined by

$$\mathcal{G}_{\delta,c_1}(S_n) = \{g \in \mathcal{G} : R_n(g) \leq R_n(f^{(\mathrm{ERM})}) + c_1 C_b d_{\delta,n}(\widehat{f}^{(\mathrm{ERM})}, g)\}$$

with the empirical distance function $d_{\delta,n}$ given by, for any functions g, g':

$$d_{\delta,n}(g,g') = \sqrt{\frac{n^{-1}\sum_{i=1}^n (g(X_i) - g'(X_i))^2 \cdot \log(2m/\delta)}{n}} + \frac{b\log(2m/\delta)}{n}.$$

In the context of model selection aggregation, the idea of applying empirical risk minimization over some set preselected set of almost minimizers goes back to Lecué and Mendelson (2009). For the recent use of midpoint procedures in statistical literature, see, for example, (Mendelson, 2019; Bousquet and Zhivotovskiy, 2021; Mourtada, Vaškevičius, and Zhivotovskiy, 2022).

Since $\widehat{f}^{(\mathrm{mid})}$ outputs 2-sparse convex combinations of elements of the dictionary $\mathcal{G}$, similarly to the above analysis of Audibert's star algorithm, it is enough to establish that $\widehat{f}^{(\mathrm{mid})}$ satisfies the offset condition. For the midpoint estimator, this fact is already implicit in the proofs of Puchkin and Zhivotovskiy (2021) in the context of active learning. While, admittedly, the direct analysis of the midpoint estimator is no more difficult than establishing the below lemma, for exposition purposes, let us demonstrate that $\widehat{f}^{(\mathrm{mid})}$ does indeed satisfy the offset condition.

Lemma 19 *Fix any $\delta \in (0,1)$ and any distribution P supported on $\mathcal{X} \times [-b,b]$. In the setup described above, the estimator $\widehat{f}^{(mid)}_{\delta}$ satisfies the $(\mathcal{G}, \ell, \varepsilon, (64)^{-1}\gamma)$-offset condition for the distribution P, with $\varepsilon(\delta) \lesssim C_b^2 \gamma^{-1} \log(2m/\delta)/n$.*

The proof is deferred to Appendix B.3. An immediate consequence of the above lemma, via an application of Theorem 8 (with $\delta_1 = \delta_2 = \delta/2$) and Corollary 17 is that for any $\delta \in (0,1)$ with probability at least $1-\delta$ the following holds:

$$\mathcal{E}(\widehat{f}^{(\mathrm{mid})}_{\delta}, \mathcal{G}) \lesssim \gamma^{-1} C_b^2 \frac{\log(4m/\delta)}{n}.$$

A.2 Learning Non-Convex Classes

Beyond the model selection aggregation setting, let $\mathcal{G}$ be an arbitrary and possibly infinite class of functions with range $[-b, b]$. For convex classes $\mathcal{G}$, we may use the classical theory of localization to obtain sharp bounds for the empirical risk minimization algorithm. However, when $\mathcal{G}$ is allowed to be non-convex, the Bernstein condition no longer holds and alternative procedures need to be considered.

When learning arbitrary classes of bounded functions, Rakhlin, Sridharan, and Tsybakov (2017) obtain sharp exponential-tail entropy number bounds for a complicated three-stage procedure that involves the star algorithm (or any other deviation optimal procedure) as a sub-algorithm. Upon the introduction of offset Rademacher complexities, Liang, Rakhlin, and Sridharan (2015) notice that offset Rademacher complexity bounds (which are at least as sharp as the entropy number bounds obtained in the above-cited paper) can be obtained for the star estimator as a direct consequence of the offset condition. This provides a simple alternative to the three-stage procedure of Rakhlin, Sridharan, and Tsybakov (2017); however, in the bounded setting considered in this paper, the results of Liang, Rakhlin, and Sridharan (2015) only yield *expected* excess risk bounds. As a corollary of Theorem 8, we can obtain the desired *exponential-tail* bound for the star algorithm that holds for arbitrary classes of reference functions.

Corollary 20 *Let $\mathcal{G}$ be an arbitrary class of reference functions with domain $\mathcal{X}$ and range $[-b, b]$ and let $\ell : [-b, b] \times [-b, b] \to [0, \infty)$ be a loss function that is C_b-Lipschitz and γ-strongly convex in its first argument. Let P be any distribution supported on $\mathcal{X} \times [-b, b]$ and fix any $g^\star \in \operatorname{argmin}_{g \in \mathcal{G}} R(g)$. Then, for any $\delta \in (0, 1)$, the star estimator $\widehat{f}^{(star)}$ defined in Appendix A.1 satisfies*

$$\mathcal{E}(\widehat{f}^{(star)}, \mathcal{G}) \le c_1 C_b \mathfrak{R}_n^{off}(P_X, \operatorname{star}(\mathcal{F} - g^\star), C_b^{-1}\gamma) + c_2 \frac{\gamma^{-1} C_b^2 \log(1/\delta)}{n},$$

where $c_1, c_2 > 0$ are some universal constants and the range of the star estimator is equal to $\mathcal{F} = \{\lambda g_1 + (1 - \lambda) g_2 : \lambda \in [0, 1], g_1, g_2 \in \mathcal{G}\}$.

Proof As discussed in Appendix A.1, the star estimator satisfies the deterministic offset condition with parameters $(\mathcal{G}, \ell, 0, \gamma/9)$. The result follows from Theorem 8 combined with Remark 10. ■

Finally, observe that when the class $\mathcal{G}$ is convex, the estimator $\widehat{f}^{(\mathrm{star})}$ becomes the ERM estimator over the class $\mathcal{G}$. At the same time, the convexity of the set $\mathcal{G}$ implies that the range of the star estimator is equal to $\mathcal{G}$, and thus, the above result reduces to the classical localized complexity bound for ERM applied to learning a convex class.

A.3 Iterative Regularization

The idea of iterative regularization is to apply some optimization procedure to the *unregularized* empirical risk function $R_n(\cdot)$ and induce a regularizing effect by early stopping. Thus, the number of iterations performed acts as a regularization parameter, in a similar way that the size of penalty acts as a regularization parameter for penalized procedures based

on empirical risk minimization. Iterative regularization schemes are actively studied since they have a built-in warm-restart feature: obtaining a new model only costs one iteration of the optimization algorithm, usually amounting to a gradient descent or stochastic gradient descent update. In contrast, for explicitly penalized procedures, obtaining new models (corresponding to different regularization parameters) amount to solving a new optimization problem. Let us demonstrate an example of how a general family of such algorithms fit into the framework of offset Rademacher complexity.

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. In this section, we fix the set of reference functions to be $\mathcal{G} = \{f_w(\cdot) = \langle w, \cdot\rangle : w \in G \subset \mathbb{R}^d\}$, where the set G is arbitrary. Denote any population risk minimizer in $\mathcal{G}$ by $g^\star = f_{w^\star}$, where $w^\star \in G$. Further, for any $w \in \mathbb{R}^d$, let $R(w) = R(f_w)$ and $R_n(f_w) = R_n(w)$.

We consider a family of mirror descent algorithms (Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003) that admit the more frequently studied gradient descent procedure as a special case. Let $\mathcal{D} \subseteq \mathbb{R}^d$ be an open and convex set. Let $\psi : \mathcal{D} \to \mathbb{R}^d$ denote a continuously differentiable strictly convex function whose gradient diverges at the boundary of $\mathcal{D}$. We call such a function a *mirror map*. The associated *Bregman divergence* $D_\psi : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is defined by $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$; note that for any $x, y \in \mathcal{D}$ we have $D_\psi(x, y) \geq 0$ due to the convexity of ψ. In *continuous-time*, the mirror descent algorithm is defined by the following differential equation, where $t \geq 0$ is the time parameter:

$$\frac{d}{dt} w_t = -\left(\nabla^2\psi(w_t)\right)^{-1} \nabla R_n(w_t). \tag{15}$$

We now present an argument due to Kanade, Rebeschini, and Vaškevičius (2023), where it was shown that early-stopped mirror descent algorithms satisfy the offset condition.

Lemma 21 *As defined above, let $\mathcal{G}$ be any reference class of linear functions and denote $g^\star = f_{w^\star}$. Let ℓ be a differentiable and γ-strongly convex loss function in its first argument (cf. Appendix A.1). Fix an arbitrary initialization point $w_0 \in \mathbb{R}^d$ and let $(w_t)_{t>0}$ be generated by the mirror descent flow* (15)*. Then, for any $\varepsilon > 0$ there exists a (random) stopping time $t^\star = t^\star(S_n, w^\star, w_0)$ such that the following three deterministic conditions hold:*

1. *The stopping time satisfies the deterministic bound $t^\star \leq 2D_\psi(w^\star, w_0)/\varepsilon$;*
2. *The early-stopped iterate $w_{t^\star}$ satisfies $w_{t^\star} \in \{w \in \mathbb{R}^d : D_\psi(w^\star, w) \leq D_\psi(w^\star, w_0)\}$;*
3. *The estimator $\widehat{f} = f_{w_{t^\star}}$ satisfies the $(\mathcal{G}, \ell, \varepsilon, \frac{\gamma}{2})$-deterministic offset condition.*

Proof For any $t \geq 0$, let $\delta(t) = R_n(w_t) - R_n(w^\star) + \frac{\gamma}{2}\sum_{i=1}^n (f_{w_t}(X_i) - f_{w^\star}(X_i))^2$. Let $t^\star := \inf\{t \geq 0 : \delta(t) \leq \varepsilon\}$ A direct computation shows the following well-known identity: $-\frac{d}{dt} D_\psi(w^\star, w_t) = \langle -R_n(w_t), w^\star - w_t\rangle$. By the γ-strong convexity assumption, it hence follows that for any $t \geq 0$ we have $-\frac{d}{dt} D_\psi(w^\star, w_t) \geq \delta(t)$. Integrating both sides, it follows that the following infimum is well defined and it satisfies all the conditions of this theorem: $t^\star = \inf\{0 \leq t \leq 2D_\psi(w^\star, w_0)/\varepsilon : \delta(t) \leq \varepsilon\}$. ∎

Remark 22 *Notice that the stopping time $t^\star$ depends on the unknown reference point $w^\star$. Thus, the above lemma establishes the* existence *of a point that satisfies the offset condition*

along the mirror descent flow. To obtain a procedure that recovers the excess risk guarantees satisfied by this optimally stopped point, we could adopt a sample-splitting and model selection approach, for instance, selecting the stopping time by running the star algorithm on held out data.

At the same time, we remark that tuning the stopping time plays an analogous role to tuning regularization parameters in explicitly penalized procedures such as ridge regression or the lasso, where the optimal regularization parameter also depends on unknown properties of the problem. For more detailed discussions, we refer to the below-cited references.

Observe that the above argument only involves the tools from convex optimization, yet Theorem 8 readily implies probabilistic performance bounds for the estimator considered above. Condition 1 in the above lemma establishes a statistical-computational trade-off. Condition 2 determines the range of the early-stopped estimator. Condition 3 shows that the early-stopped mirror descent estimator can be analyzed via offset Rademacher complexities; indeed, this is the only known approach for obtaining sharp guarantees for this general class of iterative regularization schemes (see (Kanade, Rebeschini, and Vaškevičius, 2023) for further discussion and for discrete-time results). For more examples and further background on iterative regularization, see, for example, (Bühlmann and Yu, 2003; Yao, Rosasco, and Caponnetto, 2007; Raskutti, Wainwright, and Yu, 2014; Lin, Rosasco, and Zhou, 2016; Wei, Yang, and Wainwright, 2019).

Appendix B. Deferred Proofs

B.1 Proof of Lemma 11

Fix any $\varepsilon > 0$ and let $\lambda = (1+\varepsilon)^{-1} \in (0,1)$. Let $\lambda\mathcal{H} = \{\lambda h : h \in \mathcal{H}\}$ and observe that by the star-shapedness assumption we have $\lambda\mathcal{H} \subseteq \mathcal{H}$. It follows that

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) = \lambda^{-1}\mathfrak{R}_n^{\mathrm{off}}(P_X, \lambda\mathcal{H}, \lambda^{-1}\gamma) \le \lambda^{-1}\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \lambda^{-1}\gamma). \tag{16}$$

We now proceed via a peeling argument. For any $r_1 \ge 0, r_2 > 0$ denote $\mathcal{H}(r_1, r_2) = \{h \in \mathcal{H} : \mathbf{E}_{X\sim P_X}[h(X)^2] \in [r_1, r_2]\}$. Denote $\mathfrak{R}_n^{\mathrm{loc}} = \mathfrak{R}_n^{\mathrm{loc}}(P_X, \mathcal{H}, \gamma)$. Let $\mathcal{H}_0 = \mathcal{H}(0, \gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}})$ and for $k = 1, 2, \ldots$, let $\mathcal{H}_k = \mathcal{H}(\lambda^{1-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}, \lambda^{-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}) \cup \{h_0\}$, where h_0 denotes the identically zero function. Since $\mathcal{H} = \cup_{k\ge 0}\mathcal{H}_k$, by (16) we have

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) \le \lambda^{-1}\sum_{k\ge 0}\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, \lambda^{-1}\gamma). \tag{17}$$

Observe that by the definition of $\mathfrak{R}_n^{\mathrm{loc}}$ (cf. Definition 1) we have

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_0, \lambda^{-1}\gamma) \le \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_0, 0) \le \mathfrak{R}_n^{\mathrm{loc}}.$$

At the same time, for any $k \ge 1$ we have $h_0 \in \mathcal{H}_k$ and hence $\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, \lambda^{-1}\gamma) \ge 0$. Also, by (Bartlett, Bousquet, and Mendelson, 2005, Lemmas 3.2 and 3.4) we have $\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}(0, \lambda^{-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}), 0) \le \lambda^{-k}\mathfrak{R}_n^{\mathrm{loc}}$ and consequently

$$\begin{aligned}
0 &\le \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, \lambda^{-1}\gamma) \le \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, 0) - \lambda^{-1}\gamma\cdot\lambda^{1-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}} \\
&= \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}_k, 0) - \lambda^{-k}\mathfrak{R}_n^{\mathrm{loc}} \le \mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}(0, \lambda^{-k}\gamma^{-1}\mathfrak{R}_n^{\mathrm{loc}}), 0) - \lambda^{-k}\mathfrak{R}_n^{\mathrm{loc}} \le 0.
\end{aligned}$$

Hence, using the two display equations above, the inequality (17) simplifies to

$$\mathfrak{R}_n^{\mathrm{off}}(P_X, \mathcal{H}, \gamma) \leq \lambda^{-1}\mathfrak{R}_n^{\mathrm{loc}} = (1+\varepsilon)\mathfrak{R}_n^{\mathrm{loc}}.$$

Since the choice of $\varepsilon > 0$ is arbitrary, our proof is complete. ■

B.2 Proof of Lemma 16

Let $\Phi \in \mathbb{R}^{n\times d}$ denote a matrix such that $\Phi_{i,j} = (\Phi_i)_j$ for any $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,d\}$. To simplify the notation let $\mathcal{F} = \mathcal{F}_{\mathrm{lin}}^{d,k}$. For any $S \subseteq \{1,2,\ldots,d\}$, let $\Phi_S \in \mathbb{R}^{n\times|S|}$ denote the matrix obtained by keeping only the columns of Φ indexed by the set S and let

$$\mathcal{S}^{d,k} = \{S \subseteq \{1,\ldots,d\} : |S| \leq k\}.$$

Observe that for any $\lambda > 0$ by Jensen's inequality, the fact that $x \mapsto e^{\lambda x}$ is increasing, and replacing maximum by a sum, we have

$$\begin{aligned}
&n\mathfrak{R}^{\mathrm{off}}(S_n^\Phi, \mathcal{F}, \gamma)\\
&= \mathbf{E}_\sigma \sup_{\langle w,\cdot\rangle\in\mathcal{F}} \left\{\sum_{i=1}^n \sigma_i \langle w, \Phi_i\rangle - \gamma \langle w, \Phi_i\rangle^2\right\}\\
&= \mathbf{E}_\sigma \sup_{\langle w,\cdot\rangle\in\mathcal{F}} \left\{\langle \Phi w, \sigma\rangle - \gamma w^\top(\Phi^\top\Phi)w\right\}\\
&= \mathbf{E}_\sigma \max_{S\in\mathcal{S}^{d,k}} \sup_{w\in\mathbb{R}^{|S|}} \left\{\langle \Phi_S w, \sigma\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\}\\
&\leq \frac{1}{\lambda}\log \mathbf{E}_\sigma \exp\left(\lambda \max_{S\in\mathcal{S}^{d,k}} \sup_{w\in\mathbb{R}^{|S|}} \left\{\langle \Phi_S w, \sigma\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\}\right)\\
&\leq \frac{1}{\lambda}\log \sum_{S\in\mathcal{S}^{d,k}} \mathbf{E}_\sigma \exp\left(\lambda \sup_{w\in\mathbb{R}^{|S|}} \left\{\langle \Phi_S w, \sigma\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\}\right)\\
&\leq \frac{1}{\lambda}\log \left(\left|\mathcal{S}^{d,k}\right| \max_{S\in\mathcal{S}^{d,k}} \mathbf{E}_\sigma \exp\left(\lambda \sup_{w\in\mathbb{R}^{|S|}} \left\{\langle \Phi_S w, \sigma\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\}\right)\right). \qquad (18)
\end{aligned}$$

We now proceed to upper bound the expectation inside the logarithm. For any matrix A, denote its Moore-Penrose inverse by $A^\dagger$. Fix any $S \in \mathcal{S}^{d,k}$. For any vector $\sigma \in \mathbb{R}^n$, the vector $\Phi_S^\top\sigma$ belongs to the orthogonal complement of the null space of $\Phi_S^\top\Phi_S$. Hence, following (Rockafellar, 1970, Section 12, page 108), the following identity holds:

$$\begin{aligned}
\sup_{w\in\mathbb{R}^{|S|}} \left\{\langle \Phi_S w, \sigma\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\} &= \sup_{w\in\mathbb{R}^{|S|}} \left\{\left\langle w, \Phi_S^\top\sigma\right\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\}\\
&= (4\gamma)^{-1}\sigma^\top\Phi_S(\Phi_S^\top\Phi_S)^\dagger\Phi_S^\top\sigma.
\end{aligned}$$

To simplify the notation, denote by $H = \Phi_S(\Phi_S^\top\Phi_S)^\dagger\Phi_S^\top$ the hat matrix, keeping the dependence on an arbitrary fixed $S \in \mathcal{S}^{d,k}$ implicit. By the above equation, it follows that

$$\mathbf{E}_\sigma \exp\left(\lambda \sup_{w\in\mathbb{R}^{|S|}} \left\{\langle \Phi_S w, \sigma\rangle - \gamma w^\top(\Phi_S^\top\Phi_S)w\right\}\right) = \mathbf{E}_\sigma \exp\left(\frac{\lambda}{4\gamma}\sum_{i,j=1}^n \sigma_i\sigma_j H_{i,j}\right).$$

We will now control the moment generating function of the above Rademacher chaos by decoupling and comparison with Gaussian chaos. Let $\sigma' = (\sigma'_1, \dots, \sigma'_n)^\top$ be an independent copy of σ. Let $g = (g_1, \dots, g_n)^\top \in \mathbb{R}^n$ be a vector of independent standard Normal random variables and let g' be an independent copy of g. Then, for some universal constant $c_1 > 0$ we have

$$\begin{aligned}
&\mathbf{E}_\sigma \exp\left(\frac{\lambda}{4\gamma}\sum_{i,j=1}^n \sigma_i\sigma_j H_{i,j}\right) \\
&\le \mathbf{E}_{\sigma,\sigma'} \exp\left(\frac{\lambda}{\gamma}\sum_{i,j=1}^n \sigma_i\sigma'_j H_{i,j}\right) \qquad \text{(Vershynin, 2018, (Decoupling) Theorem 6.1.1)} \\
&\le \mathbf{E}_{g,g'} \exp\left(\frac{c_1\lambda}{\gamma}\sum_{i,j=1}^n g_i g'_j H_{i,j}\right) \qquad \text{(Vershynin, 2018, (Comparison) Lemma 6.2.3).}
\end{aligned}$$

Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm and let $\|\cdot\|_F$ denote the Frobenius norm. Then, by the Gaussian chaos moment generating function bound (Vershynin, 2018, Lemma 6.2.2), there exist some universal constants $c_2, c_3 > 0$ such that for any $\lambda \in (0, \gamma c_2/\|H\|_{\mathrm{op}}]$ we have

$$\mathbf{E}_{g,g'} \exp\left(\frac{c_1\lambda}{\gamma}\sum_{i,j=1}^n g_i g'_j H_{i,j}\right) \le \exp\left(\frac{c_3\lambda^2}{\gamma^2}\|H\|_F^2\right).$$

We will now plug in the above bound into (18). Notice that the hat matrix H has at most $|S|$ non-zero eigenvalues, all of which are equal to 1; hence,$\|H\|_{op} = 1$ and $\|H\|_F^2 \le |S|$. It follows that for any $\lambda \in (0, \gamma c_2]$ we have

$$\mathbf{E}_\sigma \sup_{w\in\mathbb{R}^d, \|w\|_0\le k}\left\{\langle \Phi w, \sigma\rangle - \gamma w^\top(\Phi^\top\Phi)w\right\} \le \frac{1}{\lambda}\log|\mathcal{S}^{d,k}| + \frac{c_3\lambda k}{\gamma^2}. \tag{19}$$

Recalling the standard bound

$$|\mathcal{S}^{d,k}| = \sum_{i=1}^k \binom{d}{i} \le \left(\frac{ed}{k}\right)^k$$

and plugging in $\lambda = \gamma c_2$ in (19) yields the desired result

$$n\mathfrak{R}^{\mathrm{off}}(S_n^\Phi, \mathcal{F}, \gamma) \le \frac{1}{\gamma}\left(c_2^{-1}k\log\frac{ed}{k} + c_2c_3k\right) \lesssim \frac{1}{\gamma}\log\left(\frac{ed}{k}\right)k.$$

■

B.3 Proof of Lemma 19

For any $g, g' \in \mathcal{G}$ define the event

$$E(g,g') = \left\{R(g) - R(g') \le R_n(g) - R_n(g') + c_1C_b d_{\delta,n}(g,g')\right\}.$$

By the empirical Bernstein inequality (Maurer and Pontil, 2009, Thereom 11) applied to the random variables $(2bC_b)^{-1}(\ell_g(X_i,Y_i) - \ell_{g'}(X_i,Y_i))$ we have $\mathbf{P}(E(g,g')) \geq 1 - \delta/m^2$. Hence, defining the event $E = \cup_{g,g' \in \mathcal{G}} E(g,g')$, by the union bound $\mathbf{P}(E) \geq 1 - \delta$.

We will now show that on the event E, the estimator $\widehat{f}^{(\mathrm{mid})}$ satisfies the offset condition. First observe that on the event $E(\widehat{f}^{(\mathrm{ERM})}, g^\star) \subseteq E$, the population risk minimizer $g^\star$ belongs to the set $\mathcal{G}_{\delta,c_1}(S_n)$ of the empirical almost minimizers. Define the diameter

$$D_n^{\max} = \max_{g,g' \in \mathcal{G}_{\delta,c_1}(S_n)} \|g - g'\|_n^2, \quad \text{where} \quad \|g - g'\|_n^2 = \frac{1}{n}\sum_{i=1}^n (g(X_i) - g'(X_i))^2.$$

We may assume without loss of generality that $D_n^{\max} > 0$ since otherwise the offset condition is trivially satisfied. Since $g^\star \in \mathcal{G}_{\delta,c_1}(S_n)$, it follows that $\|\widehat{f}^{(\mathrm{mid})} - g^\star\|_n^2 \leq D_n^{\max}$. Also, since $D_n^{\max} > 0$, there exists some function $g' \in \mathcal{G}_{\delta,c_1}(S_n)$ such that $\|\widehat{f}^{(\mathrm{ERM})} - g'\| \geq D_n^{\max}/4$. Hence, on the event E it holds that

$$\begin{aligned}
&R_n(\widehat{f}^{(\mathrm{mid})}) - R_n(g^\star) \\
&\leq R_n\left(\frac{\widehat{f}^{(\mathrm{ERM})} + g'}{2}\right) - R_n(g^\star) \\
&\leq \frac{1}{2}(R_n(\widehat{f}^{(\mathrm{ERM})}) - R_n(g^\star)) + \frac{1}{2}(R_n(g') - R_n(g^\star)) - \frac{\gamma}{32} D_n^{\max}, \\
&\leq \left(\frac{1}{2} c_1 C_b \sqrt{\frac{D_n^{\max} \log(2m/\delta)}{n}} - \frac{\gamma}{64} D_n^{\max}\right) + \frac{1}{2} c_1 b C_b \frac{\log(2m/\delta)}{n} - \frac{\gamma}{64} D_n^{\max}, \\
&\leq \left(4c_1^2 C_b^2 \gamma^{-1} + \frac{1}{2} c_1 b C_b\right) \frac{\log(2m/\delta)}{n} - \frac{\gamma}{64} \|\widehat{f}^{(\mathrm{mid})} - g^\star\|_n^2,
\end{aligned}$$

where the third line follows by the strong convexity of the loss function; the fourth line follows by the fact that $g' \in \mathcal{G}_{\delta,c_1}(S_n)$ and $R_n(\widehat{f}^{(\mathrm{ERM})}) - R_n(g^\star) \leq 0$; the fifth line follows by optimizing the quadratic function in $\sqrt{D_n^{\max}}$ in the brackets and replacing $D_n^{\max}$ by $\|\widehat{f}^{(\mathrm{mid})} - g^\star\|_n^2$. By Remark 10, we have $bC_b \leq \gamma^{-1} C_b^2$ and thus our proof is complete. ■