The Small Ball Method Applied to Regression

By

Jochem Mullink

University of Twente
August 2023

Supervisors
prof. dr. Johannes Schmidt-Hieber
dr. Sophie Langer

Summary

Many problems in applied mathematics amount to estimating an object (i.e. some parameter, function or measure) from measurements. Examples encompass *regression* and *missing value imputation* in statistics and *computational tomography* (ct) and *medical resonance imaging* (mri) in imaging. So the object one needs to recover is unknown and one only has access to a set of (noisy) measurements. The goal of the small ball method is to formulate sufficient conditions under which it is possible to (approximately) recover the unknown object.

In order to recover the unknown object one wants to construct an estimator using only the measurements. Of course one wants this estimator to be close to the unknown object in a pre-specified sense. A large family of estimators can be written as the minimizer of an empirical risk functional. The empirical risk functional only depends on the measurements. The problem is that in general these estimators depend in a very complicated way on the measurements. Only in a limited number of examples it is possible to write down a closed form expression for the estimator. The small ball method can be used to formulate recovery guarantees for empirical risk minimizers.

In this report we first of all we describe the small ball method. The main difference between the approach taken here and previous work is that we introduce a delocalized small ball assumption (DSBA). This is a weaker variant of the small ball assumption. In some situations the DSBA holds, but the classical small ball assumption fails to hold. Examples are spaces of Sobolev and Hölder continuous functions. Also uniformly bounded function spaces satisfy the DSBA. We also look at some applications where we partially extend the small ball method beyond the regression setup.

Dankwoord

Allereerst wil ik graag Johannes en Sophie bedanken voor hun geduld, begrip en steun gedurende de gehele afstudeerperiode. Verder wil ik graag Gertjan, Ivo, Lilian, Marcel, Allyce en alle andere mensen waar ik gedurende mijn tijd in Enschede mee heb mogen samenwerken bedanken. Last but not least wil ik mijn familie bedanken voor hun support gedurende mijn gehele studententijd.

Contents

1 Introduction

We first introduce some notation. Let Ω be a probability space. A real-valued random variable is any measurable function $f : \Omega \to \mathbb{R}$. For a real-valued random variable f, we define the expected value of f as $\mathbb{E}f = \int f(x)d\mu(x)$ and given a sample $X_1, ..., X_N$ of i.i.d. random elements in Ω, we let $P_N f = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ be the empirical mean of f. For $p \in [1, \infty)$, let $\|f\|_{L^p(\mu)} = (\mathbb{E}f(X)^p)^{1/p}$ be the L^p-norm of any measurable function f on Ω. In this report we will ignore any measurability issues and assume that any minimizer exists.

1.1 Problem setting

The learning problem in regression is defined by a pair of random elements (X, Y) on a probability space Ω. We say that (X, Y) is distributed according to a law $\mathcal{P}$. We let μ be the law of X. The reandom element X takes values in a space of covariates $\mathcal{X}$ and Y takes values in $\mathcal{Y}$. In this report $\mathcal{Y} = \mathbb{R}$. The object we want to infer is the conditional expectation of Y given that $X = x$. So we aim to reconstruct $\mathbb{E}[Y|X = x]$ as well as possible. The idea is that X and Y are not independent, such that X contains information about the value of Y.

In regression, the random element X encodes covariates that can be used to predict Y. So in regression the goal is to construct an estimator that predicts Y as well as possible given X. In inverse problems such as computational tomography the goal is to recover the true underlying object.

The perspective common is that the measurements are independent copies of the random variables (X, Y). We observe a sample $D_N = ((X_1, Y_1), ..., (X_N, Y_N))$, where each pair (X_i, Y_i) is sampled independently and identically (i.i.d.) according to the same distribution as (X, Y). An estimator is any function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that depends on the sample D_N.

Since one only has access to a sample D_N and does not know the underlying distribution P, one can only hope to recover $\mathbb{E}[Y|X = x]$ approximately. Therefore we need to quantify the quality of an estimator.

A cost function is a function that measures the precision of a prediction. Let us say we have constructed an estimator $\hat{f}$. We observe X and we predict $\hat{f}(X)$. Conditionally on the value of X, we sample Y and compare $\hat{f}(X)$ with Y. A real-valued function $c(\hat{f}(X), Y)$ of these two variables is called a cost function. Recall that we consider the situation where Y is a real-valued random variable. We will almost always work with the square cost function $c(y_1, y_2) = (y_1 - y_2)^2$, although it turns out this is not strictly necessary. Related to this cost function we have a loss function $l_f(X, Y) = c(f(X), Y)$. The loss function corresponding to squared cost is called squared loss. We will work with the squared loss throughout this report and only mention how the arguments need to be changed extended to other cost functions. In particular in the rest of this introduction we will specialize to working with the squared cost function.

The pair (X, Y) and the sample D_N are independent. So we do not want to

evaluate the estimator on a pair (X, Y) in the sample, but we want to know how it performs on a fresh pair (X, Y) that was not used to construct the estimator $\hat{f}$. This is the right framework for many prediction tasks.

One is not really interested in the performance of our estimator on a single new measurement, but rather wants to know how well an estimator performs on average. The expected loss is the expected value of the loss function under repeated sampling of X, Y according to the underlying model. The expected loss is defined as

$$\mathbb{E}l_f = \mathbb{E}\Big[(f(X) - Y)^2\Big], \tag{1}$$

where the expectation is taken with respect to (X, Y).

The expected value of Y given $X = x$ minimizes the expected loss. Therefore it is natural to measure the performance of an estimator relative to the minimizer of this criterion. We denote $\bar{f}(x) = \mathbb{E}[Y|X = x]$. Denote the excess risk of f relative to g as $\mathcal{L}(f, g) = l_f - l_g$ and the excess expected risk of f relative to g is defined as $\mathbb{E}\mathcal{L}(f, g)$.

1.2 Estimation-approximation trade-off

We will see later that it is necessary to restrict attention to a function class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$. When $\mathcal{F}$ is too large, it becomes impossible to construct an efficient estimator in a sense that we will make precise later.

On the other hand, when the size of $\mathcal{F}$ is restricted, it is not necessarily true that the function $\bar{f}(x)$ is in $\mathcal{F}$. So we do not only incur an error due to not having access to P. We also incur an error due to the fact that the function $\bar{f}(x)$ is not in $\mathcal{F}$. It is quite easy to see that

$$\mathbb{E}\mathcal{L}(f, \bar{f}) = \mathbb{E}\mathcal{L}(f, f^*) + \mathbb{E}\mathcal{L}(f^*, \bar{f}), \tag{2}$$

where f^* is the minimizer of $\mathbb{E}l_f$ over the function class $\mathcal{F}$. $\mathbb{E}\mathcal{L}(f, f^*)$ is called the excess expected risk over $\mathcal{F}$. $\mathbb{E}\mathcal{L}(f^*, \bar{f})$ is the approximation error.

Thus the choice of $\mathcal{F}$ implies a trade-off. Making $\mathcal{F}$ larger decreases the approximation error, but makes it more difficult to estimate $\hat{f}$. Making $\mathcal{F}$ smaller decreases the excess expected risk, but might make the approximation error larger. For example, it is easier to estimate a second degree polynomial than a third degree polynomial, but if the data is generated by a third degree polynomial then one can necessarily not recover $\bar{f}$ when only fitting a second degree polynomial.

Rather than working with $\mathbb{E}\mathcal{L}(f, f^*)$, we can introduce a metric called the estimation error. Recall that μ is the law of X. We define the $L^2(\mu)$ norm of a function to be $\|f\|_{L^2(\mu)} = (\mathbb{E}[f(X)^2])^{1/2} = (\int f(x)^2 d\mu(x))^{1/2}$ and the space $L^2(\mu)$ to be the set of all measurerable functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_{L^2(\mu)}$ is finite.

The estimation error is

$$\|\hat{f} - f^*\|_{L^2(\mu)}. \tag{3}$$

The excess expected risk and the estimation error are related. These two similarity measures are both zero if $\hat{f} - f^*$ is supported on a set of measure 0. Thus only the contribution of $\hat{f} - f^*$ on the support of the measure μ matters. Furthermore when $\mathcal{F}$ has certain structural properties, then these two quantities are related.

Lemma 1.1. *Let $\mathcal{F} \subset L^2(\mu)$.*

1. *If $\mathcal{F}$ is a closed subvectorspace of $L^2(\mu)$, then $\mathcal{L}(f, f^*) = \|f - f^*\|_{L^2(\mu)}$ for any $f \in \mathcal{F}$.*
2. *If $\mathcal{F}$ is a closed convex subset of $L^2(\mu)$, then $\mathcal{L}(f, f^*) \geq \|f - f^*\|_{L^2(\mu)}$ for any $f \in \mathcal{F}$.*

So if $\mathcal{F}$ is a subvectorspace bounding the excess expected risk and the estimation error are equivalent.

The proof of this lemma is provided in the Appendix. In this report we assume that $\mathcal{F} \subset L^2(\mu)$ is a closed convex set.

1.3 Types of estimators

Remember that we want to approximate the minimizer of

$$\mathbb{E}l_f = \mathbb{E}\Big[(f(X) - Y)^2\Big] \tag{4}$$

over the function class $\mathcal{F}$ without knowing the distribution of (X, Y). We only have access to the sample D_N. Thus it is natural to replace the risk in the previous equation by the empirical risk given by

$$P_N l_f = \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - Y_i)^2. \tag{5}$$

This approach is called empirical risk minimization. We select the function $\hat{f}$ that minimizes the empirical risk. This is the most straightforward approach, but sometimes there is good reason to amend this estimator.

The first reason is when some prior information about the function $\bar{f}$ is known. For example the following types of prior information might be reasonable in practice.

1. $\bar{f}$ has a certain degree of smoothness. Then it is sensible to approximate $\bar{f}$ by a smooth function as opposed to a very rough function.
2. maybe some structural properties of $\bar{f}$ are known and one wants to select an estimator that satisfies such a structural property.

It can also happen that one has no prior information on $\bar{f}$, but it is still necessary to make concessions on the size of $\mathcal{F}$. For example when one has insufficient data so that the empirical risk minimizer is non-unique. In that case, it is useful as a tie-breaker, to select a "low-complexity" estimator.

So now we want to amend Equation5 in order to include prior information. A popular way to do this is through adding a complexity penalty to Equation5. Then the estimator $\hat{f}$ minimizes

$$P_N l_f^\lambda := \frac{1}{N}\sum_{i=1}^{N}(f(X_i) - Y_i)^2 + \lambda C(f). \tag{6}$$

where $C(f)$ is a complexity function and λ is a trade-off parameter. λ determines by how much the complexity of f is penalized. This approach is called regularization and any minimizer of Equation 6 is called a penalized empirical risk minimizer.

The properties of this estimator depend heavily on the properties of the complexity function $C(f)$.

Finally we would like to mention the following estimator, which is the limiting estimator of the estimator defined in Equation6 as $\lambda \to 0$. This estimator is the minimal complexity interpolating estimator which is defined (if it exists) as the minimizer of

$$C(f)$$

over all functions $f \in \mathcal{F}$ such that $f(X_i) = Y_i$ for all (X_i, Y_i) in the sample D_N.

1.4 Examples

In this report we assume that $C(f)$ is a norm [1] on a vectorspace.

There are a lot of different models that promote smoothness of the estimator.

1. The most simple example of an estimator of the type of Equation6 is called smooting spline. Here a function $f : \mathbb{R} \to \mathbb{R}$ is estimated. The complexity function
$$C(f) = \int_{\mathbb{R}} \left(f^{(p)}(x)\right)^2 dx,$$
where $f^{(p)}$ is the p-th (weak) derivative of f.

2. A very general construction is the following. For a (say) locally integrable function $f : \mathbb{R}^d \to \mathbb{R}$, we define the Fourier transform $\mathcal{F}(f) : \mathbb{R}^d \to \mathbb{R}$ by

[1] We state our results when $C(f)$ *is* a norm on a vectorspace. A lot of the examples in this chapter do not directly fit into this framework, but the results can be amended to also apply to all of these examples.

$$\mathcal{F}(f)(\omega) = \int_{\mathbb{R}^d} \exp(-2\pi i \omega \cdot x) f(x) dx$$

.

The smoothness of f is related to the rate of decay of $\mathcal{F}(f)$ at infinity. Thus we can choose $C(f)$ to be the L^q-norm of the function

$$\omega \mapsto \|\omega\|_p^k \mathcal{F}(f)(\omega)$$

for some $(p, q, k) \in [1, \infty] \times [1, \infty] \times \mathbb{R}_+$. When $q = 2$, this corresponds to an inner product on $L^2(\mu)$. When $(p, q, k) = (1, 1, 2)$, then this type of regularization has a strong connection with neural network models with a single hidden layer Ma et al. [2022].

3. A general construction that can be used to promote smoothness of the estimator $\hat{f}$ is through Reproducing kernel Hilbert spaces. Moreover, this approach leads to an efficient algorithm (computing the estimator requires solving a system of N linear equations in N variables).

The second class of functions are linear functions. So we let $\mathcal{X} = \mathcal{F} = \mathbb{R}^d$ and $f(x) = x \cdot f$. This setup is called linear regression.

1. (Ridge regression) The first example that we consider is Ridge regression. Here $C(f) = \|f\|_2^2$ is the Euclidean norm squared. This is the most important example of a linear regression model. It is the "must unstructured" linear regression model in the sense that the norm is rotation invariant. Compared to the empirical risk minimizer, this estimator perturbs all coefficients in the direction of the origin.

2. (LASSO) Consider now the same setting, but now suppose that $\bar{f}$ has few non-zero coefficients. One would be inclined to choose $C(f) = \|f\|_0 = |\{i : f_i \neq 0\}|$. But it turns out that this optimization problem is difficult to solve. Moreover, when for example $\bar{f}$ only *approximately* has few non-zero coeffients, then this also turns out to be sub-optimal.

   In the LASSO one chooses $C(f) = \|f\|_1 = \sum_i |f_i|$.

When we consider matrix completion or matrix recovery, we let $\mathcal{X} = \mathcal{F} = \mathbb{R}^{d \times d}$. $\mathcal{F}$ acts on $\mathcal{X}$ by

$$f(x) = \mathrm{Tr}(fx),$$

where fx is matrix multiplication of f and x. This is a special case of the linear regression setup by embedding $\mathbb{R}^{d \times d}$ into $\mathbb{R}^{d^2}$, but it leads to some interesting choices of complexity function $C(f)$.

Recall that any $d \times d$-matrix f can be decomposed as

$$f = U^T \Sigma V,$$

where U and V are orthogonal matrices and Σ is a diagonal matrix with diagonal $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d \geq 0$. The σ_i are invariant under conjugation of the matrix X. If X has rank r, then $\sigma_i = 0$ for all $i \geq r+1$.

1. Analogously with the LASSO, we can for any matrix f choose $C(f) = \mathrm{Tr}(f) = \|\mathrm{diag}(\Sigma)\|_1$, where $\mathrm{diag}(\Sigma)$ is the diagonal of the matrix Σ. This choice of complexity function promotes low-rank minimizers of f in the same way as the LASSO promotes sparsity.

2. Likewise, we can choose $C(f) = \|\mathrm{diag}(\Sigma)\|_2$ in analogy with Ridge regression.

3. In some situations it is sensible that multiple type of sparsity simultaneously occur in a problem. A prime example of this is the case where the variables are grouped, and simultaneously few variable groups are active and within groups few individual variables are active. A second type of structural assumption is in matrix completion. Here a matrix can simultaneously be low-ranked and have few non-zero entries (see Gui et al. [2016]).

1.5 A result when $\mathcal{F}$ is uniformly bounded

Bounds on respectively the expected excess risk and the estimation error have quite a long history. Here we will not give a comprehensive history, but we want to provide enough background to motivate the small ball method. One of the main results that were obtained were a two-sided bound in the bounded setting. In the bounded setting $\mathcal{F}$ is a space of measurable functions $f : \Omega \to [0, 1]$.

In the result below, we need a first complexity measure of the function space $\mathcal{F}$. In this report we will always let $\epsilon_i \in \{-1, 1\}$ be i.i.d. Rademacher random variables each taking the value ± 1 independently with probability 0.5. We let Rad $\mathcal{F}$ be the Rademacher complexity of the set $\mathcal{F}$ defined by

$$\text{Rad } \mathcal{F} = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i).$$

In a later chapter we show why the Rademacher complexity is a reasonable complexity measure in this context. Let $\psi(r)$ be a function satisfying

$$\text{Rad } \{f \in F | \mathbb{E}f \leq r\} \leq \psi(r),$$

and certain technical conditions that we skip here. Let $r^* \geq 0$ be any real number that satisfies the fixed point equation $\psi(r) \leq r$.

The claim in this context is that there exists a constant $C > 0$ such that for any $K > 1$, with probability at least $1 - 2\exp(-x)$, every $f \in \mathcal{F}$ satisfies

$$\max \left\{ P_N f - \frac{K+1}{K} \mathbb{E}f, \mathbb{E}f - \frac{K}{K-1} P_N f \right\} \leq P_N f + KCr^* + \frac{x(C+CK)}{N}. \quad (7)$$

This result is Corollary 3.5 in Bartlett et al. [2005]. This is called a two-sided bound, because both $P_N f - \frac{K+1}{K}\mathbb{E}f$ and $\mathbb{E}f - \frac{K}{K-1}P_N f$ are upper bounded (so the empirical process $f \mapsto P_N f$ is both upper and lower bounded). An upper bound on $\mathbb{E}f - \frac{K}{K-1}P_N f$ is called an one-sided bound. The main observation due to Mendelson [2015] that led to the development of the small ball method was that often only one-sided bounds are needed to bound the estimation error. Moreover, an one-sided bound can hold, while the corresponding two-sided bound does not. Thus these one-sided bounds exist under much less stringent conditions than two-sided bounds. The small ball method provides a general method to establish one-sided bounds. Notice that an one-sided bound is a lower bound on $P_N f$. So sometimes we will call such a bound a lower bound on an empirical process.

Finally we would like to say something about how this result can be used. Typically it is not applied directly to the space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$. It is typically applied to a loss class, for example $\mathcal{G} = \{x \mapsto (f(x) - f^*(x))^2 : f \in \mathcal{F}\}$. Thus a one-sided bound in this example leads to an upper bound on $(\mathbb{E} - \frac{K}{K-1}P_N)(f(x) - f^*(x))^2$, which is similar to the result that one can obtain from the small ball method. Note however that this two-sided bound does not apply to the function class $\mathcal{G}$, unless one poses stringent uniform boundedness conditions on $\mathcal{F}$.

1.6 The small ball method

The small ball method was developed to prove one-sided bounds. The original motivation for the small ball method was lower bounding the smallest singular value of a random $m \times n$-matrix with general random i.i.d. rows Koltchinskii and Mendelson [2015], but it is applicable to a much wider array of problems.

A large number of papers has appeared applying the small ball method to various estimation problems. A sequence of papers by Mendelson and co-authors applied the small ball method to regression problems. In Lecué and Mendelson [2013] it was applied to empirical risk minimizers. In Lecué and Mendelson [2018, 2017a] it was used to prove bounds on the estimation error for penalized empirical risk minimizers. The same types of arguments can be used to derive recovery guarantees for inverse problems and sparse recovery Tropp [2015], Lecué and Mendelson [2017b]. Finally we would like to mention Chinot et al. [2022], where an adversarial noise model is considered.

1.7 Noise model

In this subsection we specifically focus on the small ball method applied in the regression setup. First we recall the setup. We have random variables (X, Y) on $\mathcal{X} \times \mathbb{R}$ distributed according to $\mathcal{P}$. Let μ be the law of X. Let $\mathcal{F} \subset L^2(\mu)$ be closed and convex. We let $f^* \in \mathcal{F}$ be the minimizer of $\mathbb{E}(f(X) - Y)^2$.

We have the following equivalent characterization of f^*. The function f^* is the projection of Y onto $\mathcal{F}$ with respect to the $L^2(\mu)$ inner product. This means

that for all $f \in \mathcal{F}$,

$$\mathbb{E}(f - f^*)(Y - f^*) \geq 0.$$

When $\mathcal{F}$ is a closed subspace of $L^2(\mu)$, then for all $f \in \mathcal{F}$

$$\mathbb{E}(f - f^*)(Y - f^*) = 0.$$

This is an orthogonality relation. We define $\xi = Y - f^*(X)$. So ξ and $f - f^*$ are orthogonal. We say that the model $\mathcal{P}$ is well-specified if $\mathbb{E}[\xi | X = x] = 0$.

In Chinot et al. [2022] an adversarial noise model is assumed. Here the following alternative data generating setup is used. Choose $f^* \in \mathcal{F}$. Sample $X_1, ..., X_N \in \mathcal{X}$ according to the law μ. Now an adversary can choose $\xi_1, ..., \xi_N$ subject to a constraint on $\|(\xi_i)\|_2^2$. Conditional on the ξ_i the measurements $Y_i = f^*(X_i) + \xi_i$ are revealed. In this context the following obstruction to learnability exists.

Lemma 1.2. *(Chinot et al. [2022]) Let $0 < \epsilon < 1$. Let $\mathcal{F}$ be a function space with $f, g \in \mathcal{F}$ such that $\|f - g\|_{L^2(\mu)} = \epsilon^2/8$. Then*

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{F} : \|(\xi_i)\|_2^2 \leq N\epsilon^2} P(\|\hat{f} - f^*\|_{L^2(\mu)} \geq \epsilon^2/16) \geq 3/8$$

This is a minimax lower bound for estimation over the class $\mathcal{F}$. The infimum in this bound is taken with respect to all possible estimators. This lower bound says that for any estimator $\hat{f}$ there exists a function f^* and a choice of noise vector such that the estimation error is larger than a constant with constant probability.

Note that in the i.i.d. setting $\|(\xi_i)\|_2^2 = O(\epsilon^2 N)$, so that the magnitude of the noise is similar. But Lemma 1.2 implies that the estimation error is bounded from below by the noise level with constant probability, under a weak condition on the function space $\mathcal{F}$. Thus this lemma shows that we need to make assumptions beyond a boundedness assumption on the noise in order to show that the estimation error $\|\hat{f} - f^*\|_{L^2(\mu)} \to_{N \to \infty} 0$.

In our results we need to make additional assumptions on the noise generating procedure. In addition to the orthogonality relations defined above we need an empirical counterpart to these relations.

Assumption 1.1. *D_N is a sample such that there exists a constant c_2 sufficiently small such that for every $f \in F$ with $\|f - f^*\|_{L^2(\mu)} \geq r$, then*

$$\left| \frac{1}{N} \sum_{i=1}^{N} \xi_i (f - f^*)(X_i) - \mathbb{E}\xi(f - f^*) \right| \leq c_2 \|f - f^*\|_{L^2(\mu)}^2, \tag{8}$$

where $\xi_i = Y_i - f^*(X_i)$ and $\xi = Y - f^*(X)$ and the expectation on the LHS is taken with respect to (X, Y).

This will be the first ingredient that we need in order to demonstrate the small ball method. Later on we will elaborate more on this type of condition and present it in a more general framework. This inequality says that outside a ball of radius r centered at f^*, the quantity

$$(P_N - \mathbb{E})\xi(f - f^*)$$

can be bounded. We have formulated it here in a way so that it can be directly applied to the empirical risk minimizer.

1.8 Approximate isometries

We have already explained that by using the small ball method makes it possible to lower bound empirical processes. A possible outcome of the small ball method is the following type of lower bound.

Assumption 1.2. *Assume we have a sample D_N such that there exists a constant c_1 sufficiently large such that whenever $f \in \mathcal{F}$ and $\|f - f^*\|_{L^2(\mu)} \geq r$, then*

$$\frac{1}{N}\sum_{i=1}^{N}(f - f^*)^2(X_i) \geq c_1\|f - f^*\|^2_{L^2(\mu)} \tag{9}$$

This is a lower bound on the empirical process corresponding to the function class $\mathcal{F} = \{x \mapsto (f - f^*)^2(x) : f \in \mathcal{F}\}$. It says that outside of a ball of radius r centered at f^*, the empirical process corresponding to $\mathcal{F}$ can be lower bounded by the mean of this process. Details on these types of estimates are provided in the next chapter.

The radius r in Assumption 1.1 and 1.2 depends both on the function class $\mathcal{F}$ and the distribution $\mathcal{P}$. The value of r in both assumptions can be computed using a fixed point equation similar to the definition of r^* in Equation 7. The radius r is a problem dependent "critical radius". The point behind these two asssumptions is that whenever $\|f - f^*\|_{L^2(\mu)} \geq r$, then it can be shown that f cannot be an empirical risk minimizer. That will also be the proof method that we use to prove the following theorem, which is essentially due to Mendelson [2015].

Theorem 1.1. *Assume that $\mathcal{F} \subset L^2(\mu)$ is closed and convex. Assume that we have a sample D_N such that Assumption 1.1 and 1.2 hold with critical radius r. Then for any empirical risk minimizer $\hat{f}$,*

$$\|f - f^*\|_{L^2(\mu)} \leq r.$$

In a later chapter we will show that under the i.i.d. assumption and under certain conditions, it is possible to prove that the sample D_N satisfies Assumption 1.1 and 1.2 with high probability.

Proof. We need to show that whenever $f \in \mathcal{F}$ and $\|f - f^*\|_{L^2(\mu)} \geq r$, then f cannot be an empirical risk minimizer. Recall that by definition, an empirical risk minimzer is any function $f \in \mathcal{F}$ that minimizes

$$P_N(f(X) - Y)^2.$$

Then, since $f^* \in \mathcal{F}$, for any empirical risk minimizer f,

$$P_N(f(X) - Y)^2 - P_N(f^*(X) - Y)^2 \leq 0.$$

In order to show that f is not an empirical risk minimizer, we need to lower bound $P_N\mathcal{L}_f = P_N(f(X) - Y)^2 - P_N(f^*(X) - Y)^2$.

So assume that $f \in \mathcal{F}$ and $\|f - f^*\|_{L^2(\mu)} \geq r$. Then by explicit computations it follows that

$$P_N(f(X)-Y)^2-P_N(f^*(X)-Y)^2 = P_N(f(X)-f^*(X))^2+2P_N(f(X)-f^*(X))\xi.$$

We first lower bound the second term. By the orthogonality relation it follows that $\mathbb{E}\xi(f(X) - Y) \geq 0$. So adding and subtracting $\mathbb{E}\xi(f(X) - Y)$ shows that

$$P_N\xi(f(X)-f^*(X)) \geq (P_N-\mathbb{E})\xi(f(X)-f^*(X))+\mathbb{E}\xi(f(X)-f^*(X)) \geq (P_N-\mathbb{E})\xi(f(X)-f^*(X)),$$

where in the second step we used this orthogonality relation. Now we use Assumption 1.1, which shows that

$$(P_N - \mathbb{E})\xi(f(X) - f^*(X)) \geq -|(P_N - \mathbb{E})\xi(f(X) - f^*(X))| \geq -c_2\|f - f^*\|^2_{L^2(\mu)}.$$

This lower bounds the second term. The first term can be lower bounded directly using Assumption 1.2, since

$$\frac{1}{N}\sum_{i=1}^{N}(f - f^*)^2(X_i) \geq c_1\|f - f^*\|^2_{L^2(\mu)}.$$

So whenever $f \in \mathcal{F}$ and $\|f - f^*\|_{L^2(\mu)} \geq r$, then

$$P_N\mathcal{L}_f \geq (c_1 - 2c_2)\|f - f^*\|^2_{L^2(\mu)}$$

which shows that $P_N\mathcal{L}_f \geq 0$ whenver $c_1 - 2c_2 \geq 0$. This concludes the proof. □

This proof shows the strategy behind many applications of the small ball method. We want to show that a function or class of functions shares a certain property. We show that any function that does not satisfy this property cannot be a member of the class of functions it is presumed to live in.

To formulate this theorem a little differently, we need to show that we have the following inclusion of sets;

$$\{f \in \mathcal{F} : P_N\mathcal{L}_f \leq 0\} \subset \{f \in \mathcal{F} : \|f - f^*\|_{L^2(\mu)} \leq r\}.$$

This type of argument not only works for empirical risk minimizers, but can also be applied to families of other estimators.

1.9 Overview of the report

As mentioned we are going to focus on penalized empirical risk minimizers where the complexity functional $C(f)$ is a norm $\Psi(f)$. For a given choice of norm $\Psi(f)$ we basically want to understand how the estimation error depends on the trade-off parameter λ. The starting point for this problem is the paper Lecué and Mendelson [2018], where a optimal range of values $[\lambda_0, \lambda_1]$ of λ is identified on which it is possible to upper bound the estimation error. But it is impossible to directly use their argument to proof bounds on the estimation error outside of this range.

The first result that we prove is an extension of their result to the range where $\lambda \in [\lambda_0, \infty)$. The proof of this result is substantially more straightforward and extends to all sufficiently large values of λ. This partially answers what happens when λ is outside the range $[\lambda_0, \lambda_1]$.

When $\lambda < \lambda_0$ we say that the penalized empirical risk minimizer is under-regularized and now an interesting phenomenon happens. Classically speaking, one would expect using the bias-variance decomposition, that whenever λ is large, then the bias is large and this would make the estimation error large. This also always happens because when $\lambda \to \infty$, then the empirical risk minimizer will shrink to zero. Conversely, when λ is too small, one would expect that the variance becomes large and this would imply that the estimation error is large. In fact, when $\lambda \to 0$, then one interpolates the data if this is possible.

Sometimes the estimation error improves as $\lambda \to 0$. This phenomenon is called benign overfitting Bartlett et al. [2020]. An example is shown in Figure 1. To understand this phenomenon many papers have been published in the last few years. It is interesting to understand this phenomenon in the context of the small ball method.

We want to (partially) address the following questions in this report.

(i) What is the influence of the norm Ψ on the estimation error of the penalized empirical risk minimizer?

(ii) What is the influence of the choice of trade-off parameter λ?

(iii) What happens when the trade-off parameter $\lambda \to 0$?

In some cases, it is possible to represent the empirical risk minimizer and the penalized empirical risk minimizer in a closed form and then some of these questions can be answered directly based on this representation. But in general it is not possible to find such a representation.

One of the earliest approaches to this problem is based on localized Rademacher complexity Bartlett et al. [2002], see also the two-sided bound above. The main drawback of this approach is that it only works in the bounded setting (both the

function class and measurements are bounded by B a.s.). In Mendelson [2015] it was shown that even in the bounded setting the rates obtained in Bartlett et al. [2002] can be sub-optimal.

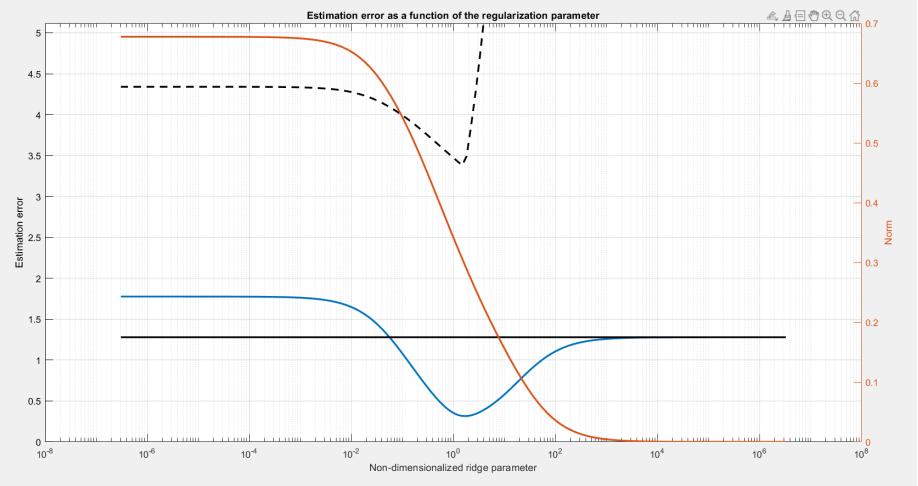
The paper Lecué and Mendelson [2018] extends Mendelson [2015]. It uses the small ball method to obtain estimation bounds for penalized empirical risk minimizers when λ is in a specific range.

Based on these previous works already question (i) and (ii) can be partially answered. In Lecué and Mendelson [2018] it is shown that when $C(f) = \Psi(f)$ is a norm, *sparsity* of f^* relative to Ψ improves the estimation properties of the penalized empirical risk minimizer.

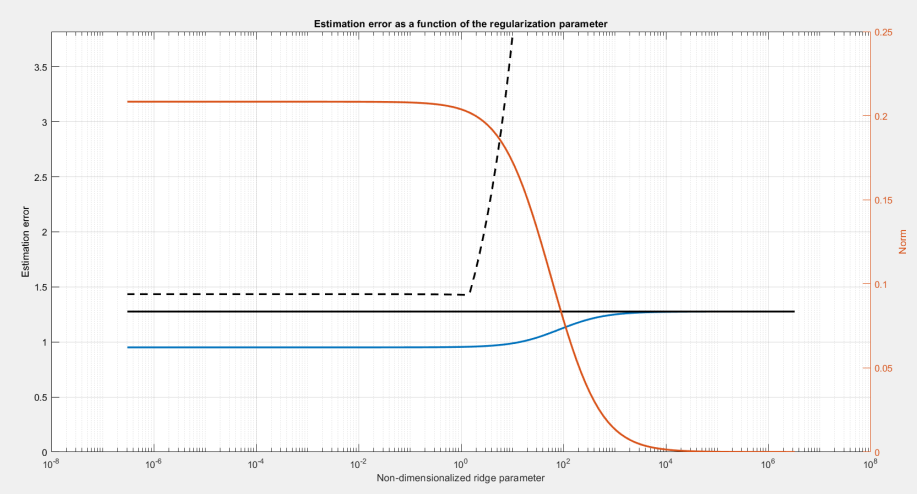
Regarding the second question in Lecué and Mendelson [2018] a specific range of λ is identified for which the resulting penalized empirical risk estimator defined in Equation6 attains the optimal estimation rate in a certain sense.

The third question is the recent and concerns so-called *norm minimizing interpolating estimators*. These are estimators minimizing $C(f)$ subject to $f(X_i) = Y_i$. In the last few years many results have been obtained on that. In particular for *well-specified* models with Gaussian noise, general results have been obtained for example in Koehler et al. [2021]. The main advantage of the approach taken here is that it allows for more general noise generating models and in the *misspecified* setting. The approach taken here can be considered as an application of the methods obtained in Lecué and Mendelson [2013], Mendelson [2015], Lecué and Mendelson [2018] to extend results obtained in Chinot and Lerasle [2020], Chinot et al. [2022] to different noise models in the interpolating and under-regularized regime.

The report is ordered in the following way. In chapter two we review the small ball method and prove the estimates that will be used to prove our results. In the third chapter we will prove the main results in this report. We will try to answer the questions posed in this introduction.



(a) $\sigma_i = \exp(-i)$



(b) $\sigma_i = i^{-1}\log(i+1)^{-1}$

Figure 1: In both pictures the estimation error as a function of the regularization parameter is plotted. The regularization parameter is plotted on a log-scale and normalized. The blue line is the estimation error. The orange line is the norm of the estimator. The black line is $\mathbb{E}[Y^2]$ as a reference. In both pictures ridge regression is applied with Gaussian design with covariance matrices Σ with diagonal σ_i. Figure 1a: This is the classical situation where the estimation error has a clear optimal range and respectively the bias or the variance is large whenever λ becomes to large or too small. Figure 1b: Benign overfitting occurs. The estimation error does not increase as $\lambda \to 0$.

2 The small ball method

2.1 A global result.

Let μ be a probability measure on Ω and let $\mathcal{F} \subset L^2(\mu)$. Recall the definition of the Rademacher complexity

$$\text{Rad } \mathcal{F} = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i),$$

where $X_1, ..., X_i \in \Omega$ are i.i.d. random variables with law μ. The ϵ_i random variables taking the value ± 1 independently with probability 0.5.

The main result of this section is the following. The proof of a similar statement was first presented in Mendelson [2015].

Theorem 2.1. *Let $\mathcal{F} \subset L^2(\mu)$. Let $T : \mathcal{F} \mapsto \mathbb{R}^+$ be any function. Assume that $(\epsilon, \kappa, r) \in (0, 1] \times \mathbb{R}^+ \times \mathbb{R}^+$ is a triple that satisfies the following two properties.*

1. *((delocalized) Small ball assumption) For all $f \in \mathcal{F}$ such that $T(f) \geq r$,*

$$P(f(X) \geq \kappa T(f)) \geq \epsilon.$$

2. *(Rademacher bound) The Rademacher complexity of $\mathcal{F}$ is bounded,*

$$Rad\ \mathcal{F} \leq \frac{\kappa \epsilon}{32} r.$$

Then, there exists an event $\mathcal{A}$ with probability $\geq 1 - 2\exp(-\epsilon^2 N/16)$, such that on $\mathcal{A}$ and for all $f \in \mathcal{F}$ such that $T(f) \geq r$,

$$\# \left\{ i \in [N] : f(X_i) \geq \frac{\kappa T(f)}{2} \right\} \geq \frac{N\epsilon}{2}.$$

Before we show the proof of this statement, we first would like to say something about the assumptions in this theorem and the implications of this theorem.

The function $T(f)$ is usually chosen to be $T(f) = \|f\|_{L^2(\mu)}$, $T(f) = \|f\|^2_{L^2(\mu)}$ or $T(f) = \mathbb{E}[l_f - l_{f^*}]$, for a loss function l_f. So in most applications $T(f)$ will be chosen to be the estimation error or the excess expected risk. Depending on the choice of T different types of estimates can be proven using this method. In Mendelson [2015] a similar theorem was proven for $T(f) = \|f\|_{L^2(\mu)}$.

As mentioned in the introduction, the conclusion of this theorem can be used to lower bound various empirical processes related to statistical learning problems. Choosing $T(f) = \|f\|_{L^2(\mu)}$, it can be shown that on the event $\mathcal{A}$ Assumption 1.2 holds with $c_1 = \frac{\kappa^2 \epsilon}{8}$.

The theorem states that whenever $T(f)$ exceeds some tolerance, then with high probability at a certain fraction of instances X_i that is uniform over the function class, $|f(X_i)|$ exceeds some tolerance level. Then this implies a high probability lower bound on the empirical risk associated to the function class.

Now we want to discuss the (delocalized) small ball assumption (DSBA).

Definition 2.1. *(small ball condition) Any set of functions $\mathcal{F}$ that satisfies condition 1 in Theorem 2.1 is said to satisfy a delocalized small ball condition. When $\mathcal{F}$ satisfies that condition with $r = 0$, then $\mathcal{F}$ is said to satisfy a small ball condition.*

Compared to the small ball assumption in Mendelson [2015], this assumption is strictly weaker, because of the observation that this assumption only needs to hold for all f with $T(f) \geq r$. We will see later that in some applications for which the DSBA holds, the small ball assumption does not necessarily hold.

The small ball assumption is a quantitative identifiability assumption. The function space $\mathcal{F}$ is identifiable if $P(|f(X) - g(X)| > 0) > 0$ whenever $f \neq g$. One could argue that the conclusion of Theorem 2.1 says that f, g are identifiable using only the values of $f - g$ on the X_i, whenever $T(f - g)$ is sufficiently large.

The small ball assumption is scale invariant under dilation by $\mathcal{F} \to c\mathcal{F}$ with $c > 0$. So in particular this assumption can hold for unbounded function classes.

The following important condition implies the DSBA.

Assumption 2.1. *($L^p - L^q$-norm equivalence) Assume that for some $1 \leq q < p \leq \infty$ there exists a constant $B > 0$, such that for all $f \in \mathcal{F}$ with $\|f\|_{L^q(\mu)} \geq r$,*

$$\|f\|_{L^p(\mu)} \leq B\|f\|_{L^q(\mu)}.$$

When $\mathcal{F}$ satisfies Assumption 2.1 for some (p, q, r) with $p > q$, then by the Paley-Zygmund inequality, $\mathcal{F}$ satisfies the (ϵ, κ, r) DSBA with $\kappa \in [0, 1]$ and $T(f) = \|f\|_{L^q(\mu)}$ and with

$$\epsilon = \left[\frac{1 - \kappa^q}{B^q}\right]^{\frac{2p}{2p-2q}}.$$

This result is especially useful when $q = 1$ or $q = 2$. A proof of this fact is provided in the Appendix.

Now we give two situations where DSBA holds.

Example 2.1. *(Gaussian random variables) Suppose that for any $f \in \mathcal{F}$, $f(X)$ is distributed according to a Gaussian random variable with mean zero and variance σ_f^2 (the variance can depend on f). For any mean zero Gaussian random variable f,*

$$\mathbb{E}|f| = \sqrt{\frac{2}{\pi}}\sigma, \qquad \mathbb{E}f^2 = \sigma^2, \qquad \mathbb{E}f^4 = 3\sigma^4.$$

So when $q = 2$ and $p = 4$, then it follows that we can choose

$$\epsilon = \left(\frac{1-\kappa^2}{\sqrt{3}}\right)^2.$$

When $q = 1$ and $p = 2$, then we can choose

$$\epsilon = \frac{\left(1-\kappa^2\right)^2}{\frac{\pi}{2}}.$$

This example is particularly useful when $\Omega = \mathbb{R}^d$ is equipped with a centered Gaussian measure and $\mathcal{F}$ consists of all linear functionals on $\mathbb{R}^d$.

Example 2.2. *(bounded random variables) Suppose that there exists a constant $b > 0$ such that $|f|$ is bounded by b almost surely for all $f \in \mathcal{F}$. Let $p = \infty$. Since for all $f \in \mathcal{F}$, $\|f\|_{L^\infty} \leq b$ it follows that Assumption 2.1 holds for all $q \in [1, \infty)$ and all $r > 0$ with $B = b/r$.*

So $\mathcal{F}$ satisfies the DSBA with $\kappa \in (0, 1)$, $q \in (1, \infty)$ and with

$$\epsilon = \frac{r^q(1-\kappa^q)}{b^q}.$$

Example 2.2 is interesting because in the bounded setting the classical SBA (with $r = 0$) does not hold in general. Thus this example shows that the DSBA is strictly more general than the classical SBA.

Before we present the proof of Theorem 2.1, we want to finish with one remark regarding condition 2 in Theorem 2.1. We need to choose r such that

$$\text{Rad } \mathcal{F} \leq \frac{\kappa\epsilon}{32} r.$$

But Rad $\mathcal{F} = \mathbb{E}\left[\sup_{f\in\mathcal{F}} R_N f(X)\right] = O(N^{-1/2})$ for quite general function classes (see Bartlett et al. [2005]). To obtain lower bounds on empirical processes of order smaller than $N^{-1/2}$, we need a better approach. This is the goal of the upcoming chapters.

Now we will present the proof of Theorem 2.1.

Proof. We need to show that with high probability and a fraction of the X_i the function value $h(X_i)$ is larger than some critical value. It is standard to first bound the expected value of this quantity, and then bound the deviation between this quantity and its expectation under the imposed assumption that the X_i are sampled i.i.d. to X with law μ. The fact that the result holds uniformly over all f such that $T(f)$ is sufficiently large will be critical in the applications that we consider. Before we present the proof we need some notation.

Define the function

$$\phi(t) = \begin{cases} 0 & \text{when } t < 1, \\ t - 1 & \text{when } 1 \leq t \leq 2 \\ 1 & \text{when } t > 2 \end{cases}.$$

Let 1 be the indicator function. The function ϕ is 1−Lipschitz and $0 \leq \phi(t) \leq 1$. Furthermore $1(x \geq 1) \geq \phi(x) \geq 1(x \geq 2)$. Let $\mathcal{F}_{\geq r} = \{f \in \mathcal{F} : T(f) \geq r\}$ and let $f \in \mathcal{F}_{\geq r}$. Let $\eta = \eta(f) = \kappa T(f)/2$.

Now we show how we decompose the quantity of interest in terms of its expectation and its deviation around the mean. By the properties of the function ϕ, and by adding and substracting the expectation of $\phi(\frac{f(X)}{2\eta})$ it follows that

$$\begin{aligned} &P_N 1(f(X) \geq \eta) \geq P_N \phi\left(\frac{f(X)}{\eta}\right) \geq \\ &\mathbb{E}\phi\left(\frac{f(X)}{\eta}\right) - \left|(P_N - \mathbb{E})\phi\left(\frac{f(X)}{\eta}\right)\right| \geq \\ &\mathbb{P}\left(\frac{f(X)}{2\eta} \geq 1\right) - \left|(P_N - \mathbb{E})\phi\left(\frac{f(X)}{\eta}\right)\right| = (i) + (ii), \end{aligned} \tag{10}$$

were the properties of ϕ and the fact that for $a, b \geq 0$, $|a + b| \geq |a| - |b|$.

Now we are going to lower bound both terms uniformly over $\mathcal{F}_{\geq r}$. We start with the first term.

(i) We show that $(i) \geq \epsilon$. This fact follows directly from the small ball assumption and the fact that $\frac{2\eta}{T(f)} = \kappa$,

$$\inf_{f \in \mathcal{F}_{\geq r}} \mathbb{P}(\frac{f(X)}{2\eta} \geq 1) = \inf_{f \in \mathcal{F}_{\geq r}} \mathbb{P}(f(X) \geq \frac{2\eta}{T(f)} T(f)) \geq \epsilon.$$

(ii) In order to bound the second term, we first apply the bounded differences concentration inequality $Z = \sup_{f \in \mathcal{F}_{\geq r}} |(P_N - \mathbb{E})\phi(\frac{f(X)}{\eta})|$, see the appendix. Recall that $\phi(x) \in [0, 1]$. Then with probability $\geq 1 - t$,

$$\sup_{f \in \mathcal{F}_{\geq r}} |(P_N - \mathbb{E})\phi(\frac{f(X)}{\eta})| \leq \mathbb{E} \sup_{f \in \mathcal{F}_{\geq r}} |(P_N - \mathbb{E})\phi(\frac{f(X)}{\eta})| + \sqrt{\log(2/t)/N}.$$

By the symmetrization inequality for Rademacher processes,

$$\mathbb{E} \sup_{f \in \mathcal{F}_{\geq r}} |(P_N - \mathbb{E})\phi(\frac{f(X)}{\eta})| \leq 2\mathbb{E} \sup_{f \in \mathcal{F}_{\geq r}} R_N \phi(\frac{f(X)}{\eta}).$$

Recall that ϕ is a 1−Lipschitz function. Thus we can apply the contraction inequality for Rademacher processes. This implies that

$$2\mathbb{E} \sup_{f \in \mathcal{F}_{\geq r}} R_N \phi(\frac{f(X)}{\eta}) \leq 4\mathbb{E} \sup_{h \in \mathcal{F}_{\geq r}} R_N \frac{f(X)}{\eta}.$$

The constant η was defined as $\eta = \kappa T(f)/2$ and for all $f \in \mathcal{F}_{\geq r}$, $T(f) \geq r$. Further, recall that since $\mathcal{F}_{\geq r} \subset \mathcal{F}$, that r satisfies $\text{Rad } \mathcal{F}_{\geq r} \leq \text{Rad } \mathcal{F} \leq \frac{\kappa\epsilon}{32} r$. So

$$4\mathbb{E} \sup_{f \in \mathcal{F}_{\geq r}} R_N \frac{f(X)}{\eta} \leq \frac{8}{\kappa r} \mathbb{E} \sup_{f \in \mathcal{F}_{\geq r}} R_N f(X) \leq \frac{\epsilon}{4}.$$

Setting $t = 2\exp(-\epsilon^2 N/16)$ implies that $\sqrt{\log(2/t)/N} = \epsilon/4$. Putting everything together shows that with probability $\geq 1 - 2\exp(-\epsilon^2 N/16)$,

$$\inf_{f \in \mathcal{F}_{\geq r}} \frac{1}{N} \sum_{i=1}^{N} 1(f(X_i) \geq \eta) \geq \frac{\epsilon}{2}.$$

This proofs the theorem. □

2.2 Localization: The role of convexity.

In what ways Theorem 2.1 can be adapted to obtain better lower bounds on the associated empirical process? In general it will not be possible to do this for any set $\mathcal{F}$ and any function $T : \mathcal{F} \to \mathbb{R}$.

Definition 2.2. *(star-shaped setting) We will say that $(\mathcal{F}, T)$ is in the star-shaped setting if*

1. *$\mathcal{F}$ is star-shaped around 0, which means that for any $\alpha \in [0,1]$ and for any $f \in \mathcal{F}$, $\alpha f \in \mathcal{F}$.*

2. *for some symmetric convex subset $S \subset L^2(\mu)$, T is of the form $T(f) = \|f\|_S = \inf\{t \in \mathbb{R}^+ : f/t \in S\}$.*

The first condition can always be satisfied by replacing any function space $\mathcal{F}$ by the star-hull $\text{star}(\mathcal{F}, 0) = \{\alpha f : f \in \mathcal{F}, \alpha \in [0,1]\}$. Moreover it can be shown (see Lemma 3.9 in Mendelson [2003]) that the Rademacher complexity of the star-hull is not significantly larger than the Rademacher complexity of the original class. Thus in many situations only the second condition is important.

Important examples of T that satisfy condition 2 are the $\|\cdot\|_{L^2(\mu)}$ and $\|\cdot\|_{L^1(\mu)}$ norms. Any norm $\|\cdot\|$ satisfies condition 2 with S the unit ball $\{f : \|f\| \leq 1\}$.

We will see that it is enough for Theorem 2.1 to hold on the set $\{f \in \mathcal{F} : T(f) = r\}$ for some appropriate radius r then the conclusion of Theorem 2.1 will also hold for all $f \in \mathcal{F}_{\geq r}$. This will motivate the idea that we will exploit. This method is called localization.

We are going to apply Theorem 2.1 to the set $\mathcal{F}_{\leq r} = \{\{f \in \mathcal{F} : T(f) \leq r\}\}$. Now condition 2 in Theorem 2.1 will hold for $\mathcal{F}_{\leq r}$ if

$$\text{Rad } \mathcal{F}_{\leq r} \leq \frac{\kappa\epsilon}{32} r.$$

The following lemma implies that for any problem of star-shaped type, r^* which is the infimum over all r that satisfy this condition has the property that for all $r \geq r^*$ this condition also holds.

Lemma 2.1. *In the star-shaped setting, for any $r \geq r^*$,*

$$\textit{Rad } \mathcal{F}_{\leq r} \leq \frac{\kappa\epsilon}{32} r.$$

Proof. Clearly by the star-shaped property and the definition of T, $\mathcal{F}_{\leq r} \subset \frac{r}{r^*}\mathcal{F}_{\leq r^*}$. So by the scaling property of Rademacher complexity

$$\text{Rad } \mathcal{F}_{\leq r} \leq \text{Rad } \frac{r}{r^*}\mathcal{F}_{\leq r^*} = \frac{r}{r^*}\text{Rad } \mathcal{F}_{\leq r^*} \leq \frac{r}{r^*}\frac{\kappa\epsilon}{32}r^* = \frac{\kappa\epsilon}{32}r.$$

This concludes the proof. □

Now we will present the main result of this section. This is essentially Corollary 5.3 in Mendelson [2015].

Theorem 2.2. *In the star-shaped setting, let r^* be the infimum of all $r > 0$ such that the (κ, ϵ, r) DSBA holds and that*

$$\textit{Rad } \mathcal{F}_{\leq r} \leq \frac{\kappa\epsilon}{32}r.$$

Assume that r^* is finite. Then with probability $\geq 1 - 2\exp(-\epsilon^2 N/16)$ for all $f \in \mathcal{F}$ with $T(f) \geq r$,

$$\#\{i \in [N] : |f(X_i)| \geq \frac{\kappa T(f)}{2}\} \geq \frac{N\epsilon}{2}.$$

Proof. Applying Theorem 2.1 to the set $\mathcal{F}_{\leq r^*}$ shows that with probability greater than or equal $\geq 1 - 2\exp(-\epsilon^2 N/16)$, for all $f \in \mathcal{F}$ such that $T(f) = r^*$,

$$\#\{i \in [N] : |f(X_i)| \geq \frac{\kappa T(f)}{2}\} \geq \frac{N\epsilon}{2}.$$

But by the star-shaped condition any $f \in \mathcal{F}$ such that $T(f) \geq r^*$ can be written as $f = cg$ for some $c \geq 1$ and for $g \in \mathcal{F}$ with $T(g) = r^*$. The implication follows because the condition

$$|f(X_i)| \geq \frac{\kappa T(f)}{2}$$

is invariant under scaling. □

Observe that from the proof of this statement it follows that when $\mathcal{F}$ is a subspace of $L^2(\mu)$, then once there exists a finite r^* that satisfies the conditions from the theorem, then the conclusion holds with that same probability over all of $\mathcal{F}$.

Now we want to show how the conclusion of Theorem 2.1 and Theorem 2.2 can be exploited.

1. We say that $\mathcal{F}$ satisfies the (r, θ) empirical small ball assumption if Assumption 1.2 holds with constant $c_1 = 2\theta$. That is given a sample $X_1, ..., X_N \in \Omega$, for all $f \in \mathcal{F}$ such that $\|f\|_{L^2(\mu)} \geq r$,

$$\frac{1}{N}\sum_{i=1}^{N} f^2(X_i) \geq 2\theta\|f\|^2_{L^2(\mu)}.$$

Lemma 2.2. *Under the conditions of Theorem 2.2, with $T(f) = \|f\|_{L^2(\mu)}$, $\mathcal{F}$ satisfies the (r^*, θ) empirical small ball assumption for $\theta = \epsilon\kappa^2/8$.*

Proof. (see the proof of Theorem 3.1 in Mendelson [2015].) From the conclusion it follows that whenever $\|f\|_{L^2(\mu)} \geq r^*$,

$$P_N f^2 \geq \frac{\kappa^2 \epsilon}{8} \|f\|^2_{L^2(\mu)}.$$

This implies the lemma. □

This type of result is useful when deriving bounds on the estimation error, because it can be applied to $f - f^*$ and $\|f - f^*\|_{L^2(\mu)}$ is the estimation error.

2. We can also apply the small ball method directly to loss classes. We let Y be any random variable on Ω and we choose $\mathcal{F} = \{(f - Y)^2 : f \in \mathcal{G}\}$ for some underlying set of random variables $\mathcal{G}$. Choose for $g \in \mathcal{F}$, $T(g) = \mathbb{E}[g] = \mathbb{E}\left[(f - Y)^2\right]$. Suppose that the set $\mathcal{F}$ satisfies some small ball property. Let r^* be the modulus of continuity of $\mathcal{F}$ around 0. That is r^* is the infimum over all $r > 0$ such that

$$\text{Rad}\,\{g \in \text{conv}(\mathcal{F}, 0) : \mathbb{E}[g] \leq r\} \leq \frac{\kappa\epsilon}{32} r.$$

The conclusion of Theorem 2.2 implies that whenever $\mathbb{E}[g] \geq r^*$, then

$$P_N g \geq \frac{\kappa\epsilon}{4} \mathbb{E}[g]$$

Now we provide a few examples showing how to compute the localized Rademacher complexities. To show this systematically, we need to introduce convering numbers of a class. Consider the set $\mathcal{F}$ equipped with the $L^2(\mu)$-norm. This norm induces a distance function on $\mathcal{F}$, given by $d(f, g) = \|f - g\|_{L^2(\mu)}$. We can measure the size of $\mathcal{F}$ in the following way. First of all $f_1, ..., f_m \in L^2(\mu)$ is a $\epsilon-$covering of $\mathcal{F}$ if for all $f \in \mathcal{F}$ there exists a f_i such that $d(f, f_i) \leq \epsilon$. We let $N(\mathcal{F}, d, \epsilon)$ be the smallest $m > 0$ such that there exist a $m-$covering of $\mathcal{F}$. Often it is convenient to work with the entropy numbers $\log N(\mathcal{F}, d, \epsilon)$. For many function classes the dependence of the entropy numbers on ϵ is one of the following two types. 1. For parametric classes $\log N(\mathcal{F}, d, \epsilon) \sim d \log \frac{\text{diam}\,\mathcal{F}}{\epsilon}$ 2. For nonparametric classes we have that $\log N(\mathcal{F}, d, \epsilon) \sim \epsilon^{-\alpha}$. Examples of classes of the first type are, intuitively speaking classes that can be smoothly parametrized by at most d parameters. Nonparametric classes are classes that cannot be parametrized by a finite dimensional space. Examples are spaces of Lipschitz functions, spaces of Sobolev functions or Besov functions etc. See for example Wainwright [2019].

By Dudley's entropy integral (see Wainwright [2019], for any $\epsilon \in \mathbb{R}^+$,

$$\text{Rad}\,\mathcal{F} \lesssim \epsilon + \int_\epsilon^{\text{diam}\,\mathcal{F}} \sqrt{\frac{\log N(\mathcal{F}, d, u)}{N}} du.$$

1. For parametric classes,

$$\text{Rad } \mathcal{F}_{\leq r} \lesssim r\sqrt{\frac{d}{N}}.$$

So when we localize in $\|\cdot\|_{L^2(\mu)}$, then it follows that $r^* = 0$ if $d \lesssim N$ and there exists no r satifying the fixed point equation otherwise. This makes sense, since to recover a vector $v \in \mathbb{R}^d$ one can perfectly recover v when one has at least d measurements and it is impossible to recover v otherwise.

2. A second important example of parametric classes are when $\Omega = \mathbb{R}^d$ and when X is a mean zero random variable on Ω with covariance matrix $\Sigma = \mathbb{E}\left[XX^T\right]$. Let $\mathcal{F} = \{f \in \mathbb{R}^d : \|f\|_2 \leq \sigma\}$. Let $\lambda_1 \geq ... \geq \lambda_d \geq 0$ be the eigenvalues corresponding to Σ. From Lemma 4 in Chinot and Lerasle [2020] it follows that

$$\text{Rad } \mathcal{F}_{\leq r} \leq \sqrt{\frac{2}{N}}\left(\sum_{i=1}^{d} \lambda_i \sigma^2 \wedge r^2\right)^{1/2}.$$

The dependence of r^* on N depends on the rate of decay of the eigenvalues of Σ.

3. For nonparametric classes we need to make a distinction between the cases (a) $0 < \alpha < 2$, (b) $\alpha = 2$ and (c) $\alpha > 2$.

   (a) When $0 < \alpha < 2$, then we can set $\epsilon = 0$ and

$$\text{Rad } \mathcal{F}_{\leq r} \lesssim \frac{r^{1-\alpha/2}}{\sqrt{N}(\alpha - 2)}.$$

   Solving the fixed point equation shows that we can choose $r^* \sim_\alpha N^{-1/\alpha}$.

   (b) When $\alpha = 2$,

$$\text{Rad } \mathcal{F}_{\leq r} \lesssim \epsilon + \frac{1}{\sqrt{N}} \log \frac{r}{\epsilon}.$$

   Choosing $\epsilon = \frac{1}{\sqrt{N}}$, r^* can in principle be computed.

   (c) When $\alpha > 2$, then

$$\text{Rad } \mathcal{F}_{\leq r} \lesssim \frac{\epsilon^{1-\alpha/2}}{\sqrt{N}(2 - \alpha)} + \epsilon,$$

   which is independent of r. Choosing $\epsilon = \left(\frac{\alpha - 2}{2\sqrt{N}}\right)^{2/\alpha}$, shows that

$$\text{Rad } \mathcal{F}_{\leq r} \lesssim_\alpha N^{-1/\alpha}.$$

   So we can choose $r^* \sim N^{-1/\alpha}$.

From these examples it is clear that the dependence of r^* on N is determined by the dependence of $N(\mathcal{F}, d, \epsilon)$ on ϵ.

2.3 Weighted empirical processes: non-convex results.

In the last chapter we saw that in the star-shaped setting it was possible to replace the global Rademacher complexity with a local variant called the modulus of continuity. But for function classes that do not belong to this family it is in general not possible to extend this analysis in the same way. Therefore we need to use a family of related empirical processes. We will see that in some cases it is possible to bound these weighted processes even when the underlying function class is not star-shaped. We will also see that these types of processes are of interest in the star-shaped setting, because using them it is possible to derive results that are distinct from the results obtained in the previous sections. A good reference for such processes is Bousquet [2002]. Right here we use these processes from the point of view of the small ball method.

Given a set $\mathcal{F} \subset L^2(\mu)$, we can define some new function spaces that are related to $\mathcal{F}$. To do so, we introduce a weight function $w : \mathcal{F} \mapsto \mathbb{R}$ on $\mathcal{F}$ and define

$$\mathcal{F}_w = \{f/w(f) : f \in \mathcal{F}\}.$$

First of all, observe that the DSBA still holds for an appropriate value of r if the original function space $\mathcal{F}$ satisfies the DSBA. We can use both the global result Theorem 2.1 and in the star-shaped setting we can use Theorem 2.2 in order to lower bound the weighted empirical process associated to $\mathcal{F}_w$.

So we need to relate the reweighted empirical process to the empirical process $\mathcal{F}$ for appropriate weight functions w. Observe that

$$\mathbb{E}f = cP_N f + w(f)\frac{\mathbb{E}f - cP_N f}{w(f)} \leq cP_N f + w(f)\sup_{f\in\mathcal{F}}\frac{\mathbb{E}f - cP_N f}{w(f)}. \tag{11}$$

In order to make this inequality tight, we need to choose $w(f)$ such that $\mathbb{E}f - cP_N f$ is of the same order as $w(f)$. In particular when these two quantities are equal, we have equality in this bound.

The first weight function we consider is $w(f) = \sqrt{\mathbb{E}f}$. This choice of weight can be applied to obtain bounds for interpolating estimators.

Proposition 2.1. *Assume that $\mathcal{F}$ is a set of non-negative random variables. Let $w(f) = \sqrt{\mathbb{E}f}$ and let $c > 0$. Define*

$$V = \sup_{f\in\mathcal{F}}\frac{\mathbb{E}f - cP_N f}{\sqrt{\mathbb{E}f}}.$$

Then

$$\mathbb{E}f \leq cP_N f + V\sqrt{cP_N f} + V^2.$$

Proof. This follows from Equation 11 and the fact that $x \leq B\sqrt{x} + A$, with $A = \sqrt{cP_N f}$ and $B = V$ implies $x \leq A + B\sqrt{A} + B^2$ (see Bousquet [2002]). □

We do not go further into this topic, because it is not strictly necessary for the final chapter. The premise is that outside a $T(f) = \mathbb{E}f$ ball it is possible to use Theorem 2.1 to lower bound the empirical process associated to the weighted

function space. Then it is possible to exploit the Proposition above to lower bound the unweighted empirical process. Examples of similar results for different weight functions are presented in Bousquet [2002].

2.4 Further examples.

In this subsection we want to explain the relevance of the small ball method in the context of function estimation.

Example 2.3 (Mendelson [2017]). *Assume there exists a sequence f_n of $\{0,1\}$-valued functions such that $\rho_n = \mu(supp\ f_n) \to 0$. $P(f_n \geq \kappa\rho_n) = \rho_n$ for sufficiently large n and $\|f_n\|_{L^2(\mu)} = \rho_n$. So $\mathcal{F}$ does not satisfy the small ball assumption.*

By this example it follows that any function space $\mathcal{F}$ that has the property that it can approximate (in $L^2(\mu)$) functions of this type does not satisfy the classical small ball assumption.

Sometimes this is a desirable property, because this example shows that f_n is 0 on the sample with probability $(1-\rho_n)^N$. Thus when ρ_n is very small, then with probability larger than some constant it is not possible to distinguish between 0 and f_n. This type of unidentifiability forms an obstruction towards learnability. So in a sense it is good to exclude these type of functions from the analysis.

On the other hand, it might as well be possible to sometimes include these type of functions. Namely if $\|f_n\|_{L^2(\mu)}$ is small, then it does not matter whether the values of f_n and 0 coincide on the sample. Only functions need to be identifiable if they are far enough from each other. This is the main motivation for introducing the DSBA.

Now we want to show that a very large space satisfies DSBA. The idea is that Ω is equipped with a metric d and that every $f \in \mathcal{F}$ has Lipschitz constant $Lip(f) \leq L$ for some fixed constant L. Recall that the (delocalized) small ball property holds relative to a probability measure on Ω. We need a compatibility condition between μ and d.

Definition 2.3 (Doubling property). *(Ω, μ, d) satisfies a doubling property if $\exists C \in (0,1)$ such that*

$$\mu(B(x,r)) \geq C\mu(B(x,2r))$$

for all $x \in \Omega$ and $r > 0$.

We have that $\mu(\Omega) = 1$, since μ is a probability measure. We will also assume that (Ω, d) has a finite diameter. Without loss of generality we can assume that d is chosen such that the diameter of (Ω, d), $\text{diam}(d) = \sup_{x,y\in\Omega} d(x,y) = 1$.

Lemma 2.3. *Assume (Ω, μ, d) satisfies a doubling property with constant C and let $d = -\log_2(C)$ be the intrinsic dimensionality of (Ω, d). Assume $diam(d) = 1$. Let $\mathcal{F}$ be a class of $L-$Lipschitz continuous functions. Let $\kappa \in (0,1)$. Then $\mathcal{F}$ satisfies the (r, κ, ϵ)-delocalized small ball assumption with $\epsilon = \left(\frac{r(1-\kappa)}{L}\right)^{-d}$.*

Proof. Let $\kappa \in (0,1)$. By assumption, $f \in \mathcal{F}$ is $L-$Lipschitz. Assume that $\|f\|_{L^2(\mu)} \geq r$. Then since μ is a probability measure, there exists a point $x \in \Omega$ such that $f(x) \geq r$.

By the Lipschitz property it follows that for all y with $d(x,y) \leq \frac{r(1-\kappa)}{L}$, $f(y) \geq \kappa r$.

By assumption $\text{diam}(d) = 1$ and $\mu(\Omega) = 1$. So by iterating the doubling property it follows that

$$\mu(B(x, 2^{-i})) \geq C^i \mu(B(x,1)) = C^i. \tag{12}$$

And choosing $i = \log_2(\frac{r(1-\kappa)}{L})$ implies that the DSBA holds with $\epsilon = C^{\log_2(r(1-\kappa)/L)}$, which implies the conclusion. □

The doubling property holds for a variety of common metric probability spaces. In particular $\mathbb{R}^d$ with the Euclidean metric satisfies the doubling property with $C = 2^{-d}$. For further examples see Stein and Murphy [1993].

The DSBA also holds for Sobolev spaces $W^{k,p} = \{f \in L^p(\mu) : \text{for every multindex } \alpha \text{ such that } |\alpha| \leq k, D^\alpha f \in L^p(\mu)\}$. This follows for certain ranges of $k, p, \log C$ from the Sobolev embedding theorem. For example if $\log C < p$, then if $f \in W^{1,p}$, then $f \in C^\alpha$ with $\alpha = 1 - \frac{\log C}{p}$, where C^α is the space of α-Hölder continuous functions. Moreover the C^α norm of f can be controlled in terms of the $W^{1,p}$ norm of f, see Evans [1998], Heinonen et al. [2015].

3 Applications

In the last chapter it was shown how to lower bound empirical processes. In this chapter these results will be used to derive performance guarantees on estimators.

The plan for this chapter it to first provide some more background and results for empirical risk minimizers. After that we will focus on results for penalized empirical risk minimizers. An important result that motivated this report is the main theorem in Lecué and Mendelson [2017a]. In that paper an optimal range for the trade-off parameter λ is identified. The main gain achieved in this report regarding these types of estimators is a simplified proof of this result. Furthermore this simplified argument enables extending this result both to the "over-regularized" (when λ is too large) and the "under-regularized" setting (when λ is too small).

3.1 Setting i: General empirical risk minimization

Let $\mathcal{F} \subset L^2(\mu)$ and consider the problem of minimizing $\mathbb{E}f$ over $\mathcal{F}$. We assume a minimizer f^* exists. We have the following simple result.

Lemma 3.1. *Consider the excess risk class $\mathcal{G} = \mathcal{F} - f^*$. Let $\mathcal{A}$ be any event. Assume that there exists a constants $c, r > 0$ such that on the event $\mathcal{A}$ for all $g \in \mathcal{G}$ with $\mathbb{E}[g] \geq r$ it follows that*

$$\mathbb{E}[g] \leq cP_N g.$$

Then for any empirical risk minimizer $\hat{f}$,

$$\mathbb{E}\left[\hat{f} - f^*\right] \leq r.$$

Proof. Let $\hat{g} = \hat{f} - f^*$. By definition of the empirical risk minimizer, $P_N\hat{g} \leq 0$. For sake of contradiction assume that $\mathbb{E}[\hat{g}] \geq r$. Then by assumption

$$cP_N\hat{g} \geq \mathbb{E}[\hat{g}] \geq r.$$

Thus $P_N\hat{g} \geq r/c$ which is strictly positive, as $c, r > 0$ by assumption. Thus we arrive at a contradiction. And it follows that $\mathbb{E}\left[\hat{f} - f^*\right] \leq r$. □

This type of result is called an exact oracle inequality. Rewrighting it says that $\mathbb{E}\hat{f} \leq \mathbb{E}f^* + r$. This is an exact oracle inequality because there is a constant one before the term $\mathbb{E}f^*$. In a non-exact oracle inequality this constant is strictly greater than one. The term r is the complexity term. A similar type of result is obtained in Koltchinskii [2006], in the case where $\mathcal{F}$ is $[0, 1]$-valued. Now we show two important settings where this lemma can be applied. First we show that we can recover this result in the bounded setting.

Example 3.1 (Uniformly bounded functions)**.** *Assume that $\mathcal{F}$ is a set of functions $f : \Omega \to [0, b]$ and let $\mathcal{G} = \mathcal{F} - f^*$, where f^* again minimizes $\mathbb{E}f$ over*

$\mathcal{F}$. Observe that the range of the function $f - f^* + b$ is $[0, 2b]$. Let $g \in \mathcal{G}$. It follows that

$$P(g(\omega) \geq \kappa \mathbb{E}g) = P(g(\omega) + b \geq \kappa \mathbb{E}g + b)$$

and $\mathbb{E}[g + b] = \mathbb{E}[g] + b$. Since the function $g + b$ has range $[0, 2b]$ it follows from Example 2.2 that $\mathcal{G}$ satisfies the (κ, ϵ, r) DSBA with $\kappa \in (0, 1)$ and $\epsilon = \frac{r(1-\kappa)}{2b}$.

Now we need to choose r satisfying

$$Rad\ \mathcal{F}_r \leq \frac{\kappa\epsilon}{32},$$

but this needs to be done based on a case by case basis.

Example 3.2. *(regression) Let $\Omega = \Omega_0 \times \mathbb{R}$ be a measure space equipped with probability distribution $\mathcal{P}$. Assume (X, Y) is distributed according to P. Let $(X_1, Y_1), ..., (X_N, Y_N)$ be any i.i.d. sample. We let $\mathcal{F}_0$ be any set of real-valued functions on Ω_0 and let $\mathcal{F} = \{(f(X) - Y)^2 : f \in \mathcal{F}_0\}$.*

By Jensens's inequality, if $\mathcal{F}_0 - Y$ satisfies a DSBA, then $\mathcal{F}$ also satisfies a DSBA. More specifically, since the function $f(x) : x \mapsto x^2$ is convex, it follows that $\mathbb{E}[(f - Y)^2] \leq (\mathbb{E}|f - Y|)^2$. Hence

$$\begin{aligned} &P\left((f - Y)^2 \geq \kappa \mathbb{E}(f - Y)^2\right) \geq \\ &P\left((f - Y)^2 \geq \kappa (\mathbb{E}|f - Y|)^2\right) = \\ &P\left(|f - Y| \geq \sqrt{\kappa}\mathbb{E}[|f - Y|]\right). \end{aligned}$$

Thus the loss class $\mathcal{F}$ satisfies DSBA if the function class $\mathcal{F} - Y$ satisfies a DSBA over $L^1(P)$. This can once again be verified for a variety of function classes using the fact that $L^p - L^1$ norm equivalences imply the DSBA over $L^1(P)$.

Secondly we need to compute the critical radius r. The computation of this critical exponent can be reduced to the computation of two simpler radii r_M and r_Q. This will be explained later.

In the exact oracle inequality, we assumed that $\hat{f}$ is an empirical risk minimizer in the sense that $P_N(\hat{f} - f^*) \leq 0$. But from the argument it directly follows that it is possible to extend this result to approximate empirical risk minimizers. In this case an estimator $\hat{f}$ such that $P_N(\hat{f} - f^*) < r/c$. This is an reoccurring observation in this chapter.

3.2 Setting ii: Regression over $L^2(\mu)$ and bounds on the estimation error

Now we return to the problem described in the introduction. The set $\mathcal{F}$ is fixed and the goal is to minimize the expected square loss over $\mathcal{F}$. In the introduction the computation of the critical radius r was not discussed in detail. In chapter two it was already observed that the small ball method can be used to verify Assumption 1.2. This will lead to a radius r_Q. The verification of Assumption 1.1 will lead to a second critical radius r_M depending on the set $\mathcal{F}$ and the

distribution of the pair (X, ξ). The DSBA gives a critical radius r_S. The critical radius related to this problem then will be $r = \max\{r_S, r_Q, r_M\}$. This subsection is based on Mendelson [2015] and it will be explicitly mentioned if something is based on any other source.

Recall that f^* is the minimizer of $\mathbb{E}\left[(f(X) - Y)^2\right]$ over all $f \in \mathcal{F}$. We assume that the centered space $\mathcal{F}_c = \mathcal{F} - f^*$ satisfies the DSBA with parameters (ϵ, κ, r_S) with respect to $T(f) = \|f\|_{L^2(\mu)}$. Let r_Q be equal to the infimum over all $r > 0$ such that

$$\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f^*\|_{L^2(\mu)} \leq r} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i) \leq \gamma_Q r$$

with $\gamma_Q = \kappa\epsilon/32$.

We have the following Lemma showing that Assumption 1.2 holds under the previous assumptions.

Lemma 3.2. *Assume that $\mathcal{F}$ is closed and convex. Under the previous conditions there exists an event $\mathcal{A}$ with probability mass at least $1 - 2\exp(-\epsilon^2 N/16)$ such that $\mathcal{F}_c$ satisfies Assumption 1.2 with parameters $r = \max r_S, r_Q$ and with $c_1 = \kappa^2\epsilon/8$.*

Proof. Choosing $T(f) = \|f\|_{L^2(\mu)}$, by Theorem 2.2 it follows that there exists an event $\mathcal{A}$ with probability mass at least $1 - 2\exp(-\epsilon^2 N/16)$ such that

$$\# \left\{ i \in [N] : |f(X_i)| \geq \frac{\kappa \|f\|_{L^2(\mu)}}{2} \right\} \geq \frac{N\epsilon}{2}.$$

So it immediately follows that

$$\frac{1}{N} \sum_{i=1}^{N} (f - f^*)^2(X_i) \geq \|f\|^2_{L^2(\mu)} \kappa^2 \epsilon/8,$$

which finishes the proof. □

Secondly we need to verify Assumption 1.1. Namely we need to upper bound the supremum of the empirical process $(X, \xi) \mapsto \xi f(X)$ parametrized by the function class $\mathcal{F}_c$ localized around f^*. Thus we want to find a constant $r > 0$ such that

$$P\left(\sup_{f \in \mathcal{F}_c: \|f\|_{L^2(\mu)} \leq r} \frac{1}{N} \sum_{i=1}^{N} \xi_i f(X_i) - \mathbb{E}\xi f(X) \leq \gamma_M r^2 \right) \geq 1 - \delta/4.$$

We denote the infimum of all such $r > 0$ as r_M. In order to compute r_M we relate this empirical process to the Rademacher process of $(X, \xi) \mapsto \xi f(X)$ with $f \in \mathcal{F}_c$.

Lemma 3.3. *(Giné-Zinn Symmetrization theorem) Let $\epsilon_1, ..., \epsilon_N$ be i.i.d. Rademacher random variables. If for some $r > 0$ with probability at least $1 - \delta/4$*

$$\sup_{f \in \mathcal{F}_c: \|f\|_{L^2(\mu)} \leq r} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \xi_i f(X_i) \leq \gamma_M r^2,$$

then also for this same r with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}_c:\|f\|_{L^2(\mu)}\leq r} \frac{1}{N}\sum_{i=1}^N \xi_i f(X_i) - \mathbb{E}\xi f(X) \leq \gamma_M r^2.$$

The Rademacher process $\sup_{f\in\mathcal{F}_c:\|f\|_{L^2(\mu)}\leq r} \frac{1}{N}\sum_{i=1}^N \epsilon_i\xi_i f(X_i)$ can in principle be upper bounded by similar means as previous types of Rademacher complexities. By making assumptions about the noise ξ, the quantity $\sup_f \frac{1}{N}\sum_{i=1}^N \epsilon_i\xi_i f(X_i)$ can be bounded in terms of the Gaussian complexity $G(f) = \mathbb{E}\sup_f \frac{1}{N}\sum_{i=1}^N g_i f(X_i)$, where g_i are i.i.d. standard Gaussian random variables. Examples can be found in Lecué and Mendelson [2018].

Finally we would like to condense the previous two lemmas and the result from the introduction into a single theorem. After that we give an additional interpretation of the critical radius r in this context.

Theorem 3.1. *In the regression setup, let $\mathcal{F}$ be a closed convex space of functions on Ω and let (X,Y) be a pair of random variables on $\Omega\times\mathbb{R}$. Let $r\geq 0$ be any constant such that the following statements hold.*

(i) The space $\mathcal{F}$ satisfies the $\|\cdot\|_{L^2(\mu)}$ DSBA with parameters (ϵ,κ,r).

(ii)

$$\mathbb{E}\sup_{f\in\mathcal{F}:\|f-f^*\|_{L^2(\mu)}\leq r} \frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i) \leq \gamma_Q r$$

with $\gamma_Q = \kappa\epsilon/32$.

(iii) With probability at least $1-\delta/4$

$$\sup_{f\in\mathcal{F}_c:\|f\|_{L^2(\mu)}\leq r} \frac{1}{N}\sum_{i=1}^N \epsilon_i\xi_i f(X_i) \leq \gamma_M r^2,$$

with $\gamma_M = \kappa^2\epsilon/160$.

Then for any empirical risk minimizer $\hat{f}$ it follows that $\|\hat{f}-f^*\|_{L^2(\mu)}\leq r$.

Proof. By the previous lemma, the conditions of Theorem 1.1 hold with $c_1 = \kappa^2\epsilon/16$ and $c_2 = \kappa^2\epsilon/80$. So $c_1 - 2c_2$ is strictly positive and the result follows. □

The critical radius r used here has another interpretation. Consider the family of models M parameterized by $\mathcal{F}$ such that for any $f^*\in\mathcal{F}$ the covariates X are generated according to a fixed distribution and $Y = f^*(X)+e$ where e is standard normal. By a result in Lecué and Mendelson [2013] under weak conditions the critical radius r used here is the minimax rate associated with the class of models M. This means that for any estimator $\hat{f}$,

$$\sup_M \|\hat{f}-f^*\|_{L^2(\mu)} \geq r$$

with constant probability.

Finally in situations where the underlying space $\mathcal{F}$ is not clear we write $r(\mathcal{F})$ for r.

3.3 Norm penalized estimators

We adapt the framework considered in the following way. We let E be a vector space equipped with a norm Ψ and assume that $\mathcal{F}$ is a subset of E. Given $f \in \mathcal{F}$, we denote the corresponding loss function by l_f. Thus we assume that $\mathcal{F}$ parametrizes a family of loss functions on Ω.

Later we will need to make certain assumptions about the loss function l_f, but first we will will show some definitions and results that will be used in the proofs, related to the norm Ψ.

The types of estimators we consider are $\hat{f}$ minimizing

$$P_N l_f + \lambda \Psi(f),$$

over all $f \in \mathcal{F}$. As always we denote by f^* the minimizer of $\mathbb{E} l_f$ over all $f \in \mathcal{F}$. It is a reasonable idea that the larger the norm of f^* is, the more difficult it is to estimate f^*. In general this is indeed true, but in some specific situations it is possible to estimate f^* with a large norm, but that satisfies a certain structural assumption compatible with the norm. Such a compatible structure is called sparsity. We will right now formalize this type of sparsity along the lines of Lecué and Mendelson [2018].

This idea will lead to an adapted complexity measure ρ, that can in general be significantly smaller than $\Psi(f^*)$. We will now introduce this notion of complexity. Note that on the space $\mathcal{F}$, we have a function $T : \mathcal{F} \to \mathbb{R}$ that was used in the small ball method.

Definition 3.1. *Recall that E is a normed linear space. We denote the ball of radius r centered at f as $B_\Psi(r, f) = \{g \in E : \Psi(g - f) \leq r\}$ and we let $S_\Psi(r, f)$ be the sphere of radius r centered at f. Denote the unit ball and unit sphere by $B_\Psi = B_\Psi(1, 0)$ and $S_\Psi = S_\Psi(1, 0)$. We denote by E^* the dual space to E consisting of all continuous linear functionals on E. The dual space E^* is equipped with the dual norm; for a linear functional z^* on E, $\Psi^*(z^*) = \sup_{f \in B_\Psi} z^*(f)$. A functional $z^* \in S_{\Psi^*}$ is norming for $z \in E$ if $z^*(z) = \Psi(z)$.*

Let $\Gamma_{f^*}(\rho) \subset S_\Psi^*$ be the collection of all functionals $z^* \in S_{\Psi^*}$ that are norming for some $f \in B_\Psi(\rho/20, f^*)$. So $\Gamma_{f^*}(\rho) \subset S_\Psi^*$ consists of all linear functionals that are norming for an f close to f^*. We will see that a vector $f \in E$ is sparse if the set of norming functionals of f is large.

The key quantity measuring the degree of sparsity is

$$\Delta(\rho) = \inf_{f \in \mathcal{F} \cap S_\Psi(\rho, f^*) \cap D_T(r(\rho))} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*). \tag{13}$$

In this equation, T is the function on $\mathcal{F}$ coming from the small ball method and $r(\rho)$ is the critical radius corresponding to the set $\mathcal{F}_\rho = \{f \in \mathcal{F} : \Psi(f - f^*) \leq \rho\}$. We denote $D_T(r(\rho)) = \{f \in \mathcal{F} : T(f) \leq r(\rho)\}$, where we abbreviated $r(\rho) = r(\mathcal{F}_\rho)$.

It is not easy to see why $\Delta(\rho)$ is the right quantity to consider. Later on we will provide an example to show that this indeed leads to a notion of sparsity. Further it becomes clear that $\Delta(\rho)$ occurs naturally in the proofs.

In all of our results we need that we choose ρ (which is the complexity of f^*) such that $\Delta(\rho)$ is non-zero.

Assumption 3.1. *Let $\rho > 0$ be some fixed constant such that $\Delta(\rho) > 0$. We write $f^* = u + v$ with $\Psi(u) \leq \rho/20$ and with z^* norming for v, such that $\Delta(\rho) = z^*(v)$.*

Now we come to the main motivating example for the theory just introduced. We consider the linear regression setup with $\Psi(f) = \|f\|_1 = \sum |f_i|$. So μ is a measure on $\mathbb{R}^d$ and $\mathcal{F} = \mathbb{R}^d$.

In this case f^* is sparse whenever the number of non-zero coefficients $s = |\text{supp } f^*| = \#\{f_i^* : f_i^* \neq 0\}$ of f^* is small. So if $s \ll d$, we can construct norming functionals for f^* in the following way. Recall that the dual norm to $\|\cdot\|_1$ is $\|\cdot\|_\infty$. We let $z_i^* = \pm 1$ on the support of f^*, where we choose the sign of z_i^* depending on the sign of f^*, so that $z^*(f^*) = \Psi(f^*)$. All other coefficients of z^* can be chosen arbitrarily. So if f^* is $s-$sparse, then the set of norming funcionals at f^* consists of a subspace of codimension s. By the same construction we have that whenever f_1, f_2 have disjoint support, then there exists a z^* that is both norming for f_1 and f_2. This idea leads to the following result.

Lemma 3.4. *Lecué and Mendelson [2018] Choose $T(f) = \|f\|_2$. If $f^* = u + v$ and $u \in (\rho/20)B_1$ and $100|supp\ v| \leq (\rho/r(\rho))^2$, then $\Delta(\rho) \geq 4\rho/5$.*

Proof. We need to show that for any $f \in \mathcal{F} \cap S_\Psi(\rho, t^*) \cap D_T(r(\rho))$ there exists some $z^* \in \Gamma_{t^*}(\rho)$ such that $z^*(f - t^*) \geq 4\rho/5$.

So let $f \in \mathcal{F} \cap S_\Psi(\rho, t^*) \cap D_T(r(\rho))$. Because $\Psi(v - t^*) \leq \rho/20$, there exists some $z^* \in \Gamma_{t^*}(\rho)$ that is norming for v.

Let $I = \text{supp } (v)$. For $w \in \mathbb{R}^d$, we denote by $P_I w$ the vector obtained by setting all coefficients of w outside I to zero. So $P_I v = v$ and it follows that v and $P_{I^c} w$ have disjoint support, where I^c is the complement of I. So we can choose z^* such that z^* is also norming for $P_{I^c} w$. Thus

$$z^*(w) = z^*(P_I w) + z^*(P_{I^c} w) \geq \|P_{I^c} w\|_1 - \|P_I w\|_1 \geq \|w\|_1 - 2\|P_I w\|_1.$$

By assumption $\|w\|_2 \leq r(\rho)$, so $\|P_I w\|_1 \leq \sqrt{s}\|P_I w\|_2 \leq \sqrt{s} r(\rho)$. So finally

$$z^*(w) \geq \rho - 2\sqrt{s} r(\rho) \geq 4\rho/5$$

precisely whenever $100s \leq (\rho/r(\rho))^2$. □

In general a norm does not need to have a relevant notion of sparsity, but it is always possible to find a ρ such that $\Delta(\rho) = \rho$. Namely if we choose $\rho = 20\Psi(f^*)$, then $0 \in B_\Psi(\rho/20, f^*)$ and any $z^* \in S_{\Psi^*}$ is norming for 0 and thus $\Delta(\rho) = \rho$. More examples can be found in Lecué and Mendelson [2018].

Now we are going to explain how the quantity $\Delta(\rho)$ can be used in our results.

The following two results are used in the analysis of penalized empirical risk minimizers for the square loss. So we want to upper bound the estimation error and we choose $T(f) = \|f - f^*\|_{L^2(\mu)}$. From the proof of this statement it will become clear it is possible to generalize this for other choices of $T(f)$. We only use the homegeneity property of $T(f)$. This first result is used in the analysis of over-regularized estimators.

Lemma 3.5 (Along the lines of Lecué and Mendelson [2018])**.** *Assume that $f \in \mathcal{F}$ with $\rho\|f - f^*\|_{L^2(\mu)} \leq \Psi(f - f^*)r(\rho)$. Under Assumption 3.1,*

$$\Psi(f) - \Psi(f^*) \geq \frac{\Psi(f - f^*)}{\rho}\Delta(\rho) - \rho/10.$$

Proof. Let $f \in \mathcal{F}$ and assume that $\rho\|f - f^*\|_{L^2(\mu)} \leq \Psi(f - f^*)r(\rho)$. Let $f^* = u + v$ with z^* such that $\Delta(\rho) = z^*(v)$ and $\Psi(u) \leq \rho/20$. First note that by duality for any z^* with $\Psi^*(z^*) \leq 1$, it follows that $\Psi(f) \geq z^*(f)$ for all $f \in E$. Then by the triangle inequality

$$\Psi(f) - \Psi(f^*) = \Psi(f) - \Psi(u + v) \geq \Psi(f) - \Psi(v) - \Psi(u).$$

By definition of $\Delta(\rho)$ and since $\Psi(f) \geq z^*(f)$ one concludes that $\Psi(f) - \Psi(v) \geq z^*(f - v)$. Applying duality again and using that $f^* = u + v$ shows that $z^*(f - v) - \Psi(u) \geq z^*(f - f^*) - 2\Psi(u)$. We can now multiply $z^*(f - f^*)$ by $\frac{\Psi(f-f^*)}{\rho}\frac{\rho}{\Psi(f-f^*)} = 1$. We can moreover use that $g = \frac{\rho}{\Psi(f-f^*)}(f - f^*)$ has norm ρ and by the condition on $f - f^*$ it follows that $\|g\|_{L^2(\mu)} \leq r(\rho)$. So now using the definition of $\Delta(\rho)$ it follows that $z^*(f - f^*) \geq \frac{\Psi(f-f^*)}{\rho}\Delta(\rho)$. Finally exploiting that $\Psi(u) \leq \rho/20$ proves the Lemma. □

The following result is used in the analysis of underregularized penalized empirical risk minimizers for the square loss.

Lemma 3.6. *Whenever $\rho\|f-f^*\|_{L^2(\mu)} \leq \Psi(f-f^*)r(\rho)$, then under Assumption 3.1,*

$$\Psi(f - f^*) \leq \rho\frac{\Psi(f) - \Psi(f^*) + \rho/10}{\Delta(\rho)}.$$

Proof. This Lemma follows easily from 3.5 by rearranging the conclusion. □

Notice that the only used property of the function $f - f^* \mapsto \|f - f^*\|_{L^2(\mu)}$ was homogeneity with respect to scaling. So this proof also works for functions $T(f)$ that satisfy a similar homogeneity assumption. These types of functions $T(f)$ are going to be considered later, but we do not explicitly look at the underregularized case in that situation.

3.4 Setting ii revisited: Regression over $L^2(\mu)$ and bounds on the estimation error

Now we consider again the regression setting with squared loss, but with penalized empirical risk minimizers instead of empirical risk minimizers. To be precise we consider any minimizer of

$$P_N(f(X) - Y)^2 + \lambda\Psi(f)$$

with Ψ being any norm. This falls within the framework sketched in the previous section because the loss $l_f(X, Y) = (f(X) - Y)^2$ is parametirized by a family of functions $\mathcal{F}$ in a vector space. We assume that $\mathcal{F}$ is closed and convex and we once again let f^* be any minimizer of $\mathbb{E}l_f$ over $\mathcal{F}$.

Given an estimator $\hat{f}$, we want to bound the estimation error $\|\hat{f} - f^*\|^2_{L^2(\mu)}$. Recall that Ψ is a norm on $\mathcal{F}$. For penalized empirical risk minimizers it is possible to proof a variety of bounds under different conditions. A shared property of these results is that it is useful to distinguish between the case when $\|\hat{f} - f^*\|_{L^2(\mu)} \leq \frac{r(\rho)}{\rho}\Psi(\hat{f} - f^*)$ and the case when $\|\hat{f} - f^*\|_{L^2(\mu)} \geq \frac{r(\rho)}{\rho}\Psi(\hat{f} - f^*)$. We call these two cases respectively *norm-dominated* and *loss-dominated* results. The idea is that we can combine a norm-dominated result and a loss-dominated result in order to give a bound that holds for the estimator $\hat{f}$. Depending on the problem, it is possible to choose the combination of norm-dominated and loss-dominated bounds that is most suited for the problem or estimator at hand. This is reminiscent of the classical bias-variance decomposition, and we will see that these two cases also correspond to a bias-like term and a variance-like term.

The main advantage of this approach is that it is possible to more flexibly apply these results. We will see that it is sometimes possible to either recover or improve upon known results in the literature.

To state the results more compactly, we introduce the following notation. We let $\mathcal{A}$ be an event such that

1. $$\sup_{f \in \mathcal{F}_c : \|f\|_{L^2(\mu)}, \Psi(f) \leq \rho \leq r(\rho)} \frac{1}{N}\sum_{i=1}^{N} \epsilon_i \xi_i f(X_i) \leq \gamma_M r^2,$$

   with $\gamma_M = \kappa^2\epsilon/160$.

2. For all $f \in \mathcal{F}$ such that $\|f - f^*\|_{L^2(\mu)} \geq r(\rho)$,

$$P_N(f - f^*)^2 \geq \frac{\kappa^2\epsilon}{8}\mathbb{E}(f - f^*)^2.$$

By the Lemma 3.2 and Lemma 3.3 the event $\mathcal{A}$ can be chosen to have probability mass at least $1 - 2\exp(\epsilon^2 N/16) - \delta$. So in the rest of this chapter we will say that the event $\mathcal{A}$ holds if the these two statements hold.

Also recall the definition of $\mathcal{L}_f = l_f - l_{f^*}$ and $\mathcal{L}_f^\lambda = \mathcal{L}_f + \lambda(\Psi(f) - \Psi(f^*))$ which will be used in this chapter.

3.4.1 Results for (over-)regularized estimators

The following result is a generalization of the main result in Lecué and Mendelson [2018]. The main difference between their approach and the approach taken here is that the proof is structured in a more efficient way. This makes the proof of this result more comprehensible. The following result holds for all λ that are sufficiently large.

Theorem 3.2. *Let $\hat{f}$ be the $\lambda-$penalized empirical risk minimizer. Let $\rho > 0$ and recall the definition of the sparsity parameter $\Delta(\rho)$. Let λ_0 be defined by the equation $\lambda = \lambda_0 \frac{\kappa^2 \epsilon}{80\rho} r^2(\rho)$. Then on $\mathcal{A}$,*

$$\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho) \max\{1, \frac{\rho/10}{\lambda_0 \Delta(\rho) - \rho}, \lambda_0/9\}, \tag{14}$$

whenever $\lambda_0 \Delta(\rho) - \rho > 0$.

First of all this theorem has the following interpretation. Recall that $r(\rho)$ is more or less the rate that one would expect to be attained whenever one performs empirical risk minimization over the set $\mathcal{F}_\rho = \{f \in \mathcal{F} : \Psi(f - f^*) \leq \rho\}$. Thus whenever ρ is sufficiently large, and given a proper choice of λ it follows that it is possible to recover f^* just as accurately as if the set $\mathcal{F}_\rho$ is known beforehand. This phenomenon is called adaptivity of the estimator $\hat{f}$ to the choice of ρ.

The proof incorporates quite a few of the steps of the original proof of a related statement in Lecué and Mendelson [2018]. The main difference is that the proof is simplified and it is easier to generalize this proof to different settings.

A direction in which this result can be generalized is to estimators that are only approximate minimizers to the penalized empirical risk functional. Recall that a penalized empirical risk minimizer $\hat{f}$ also minimizes $P_N \mathcal{L}_f^\lambda$ and that $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$. An estimator is called an $\alpha-$approximate penalized empirical risk minimizer if $P_N \mathcal{L}_{\hat{f}}^\lambda \leq \alpha^2$. These types of results are a straightforward extension of the proofs presented here. But first we want to discuss what this result states about the choice of optimal trade-off parameter λ.

The rescaled trade-off parameter λ_0 is used, because it is "dimensionless".Indeed, when we rescale the norm $\Psi \to c\Psi$, then in order for the estimator to remain the same one needs to rescale λ also appropriately. But under this rescaling λ_0 remains constant. Also one can rescale $(f, Y) \to (cf, cY)$ and λ_0 also does not change whenever one rescales λ to compensate for this second transformation. From a physics point of view these two different types of transformations correspond to different units and λ_0 is dimensionless in this sense. Thus it becomes interesting to understand the dependence of the proportionality constant $\frac{\kappa^2 \epsilon}{8\rho} r^2(\rho)$ on ρ, because this says something about how the optimal parameter λ needs to be chosen. In particular when this proportionality constant is independent of ρ then it is possible to use the same value of λ. Sometimes $r^2(\rho)/\rho$ is not constant. For example, as we will see, when $r_Q(\rho) > r_M(\rho)$, then $r^2(\rho)$ is a quadratic function of ρ. So in that case $r^2(\rho)/\rho$ is not constant. In Lecué and Mendelson [2018] a variety of examples are presented were this

proportionality constant is independent of ρ. For example this happens for the LASSO, slope and trace-norm regression.

Choice of λ

From the previous result it becomes clear that choosing $\lambda_0 \sim 1$ is optimal. Thus the optimal choice of λ, which is the parameter that is important in practice should be chosen in the order of magnitude of $r^2(\rho)/\rho$. Depending on the application, sometimes the size of $r(\rho)$ can be estimated. For example when one has control over the measurements X_i and one has an idea about the characteristics of the noise, it is possible to estimate $r(\rho)$ for fixed ρ. Alternatively $r(\rho)$ can be computed conditionally on the X_i when one has an idea about the noise level. But in general the value of ρ is not known (because it depends on f^* which is unknown). Therefore it is of interest to understand under what conditions $r^2(\rho)/\rho$ is independent of the value of ρ.

When $\mathcal{F}$ is a vector space, it is is possible to determine the dependence of $r^2(\rho)$ on ρ.

Lemma 3.7. *Assume that $\mathcal{F}$ is a vector space. Then $r_Q(\rho) = \rho r_Q(1)$.*

Proof. Let $s = \rho r_Q(1)$ and notice that

$$\frac{1}{\rho}\mathbb{E}\sup_{f:\Psi(f)\leq\rho,\|f\|_{L^2(\mu)}\leq s}\frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i) = \mathbb{E}\sup_{f:\Psi(f)\leq 1,\|f\|_{L^2(\mu)}\leq s/\rho}\frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i) \leq \gamma_Q s/\rho.$$

Letting $\tilde{s} = s/\rho$ and substituting this in the equation above shows that $r_Q(1)$ is equal to the infimum over all $\tilde{s}$ satisfying the equation above. Thus $r_Q(\rho) = \rho r_Q(1)$. □

Thus when $r_Q(\rho)$ dominates, $r^2(\rho) \sim \rho^2$ and by dimentionality arguments it is reasonable to expect that $\Psi(f)^2$ is a reasonable penalty function in this case. These different types of penalty functions are not discussed further.

Relationship to the main result in Lecué and Mendelson [2018]

The main result in Lecué and Mendelson [2018] can be recovered, since assuming that we choose ρ such that $\Delta(\rho) \geq 4\rho/5$, then in order for the RHS of Equation 17 to be equal to $r(\rho)$, we can choose λ_0 satisfying $\frac{11}{8} \leq \lambda_0 \leq 9$. Thus we can prove the following corollary.

Corollary 3.1. *Lecué and Mendelson [2018] Let $\hat{f}$ be a penalized empirical risk minimizer. Let $\rho > 0$ with $\Delta(\rho) \geq \frac{4\rho}{5}$. Whenever*

$$\frac{11\kappa^2\epsilon}{640\rho}r^2(\rho) \leq \lambda \leq \frac{9\kappa^2\epsilon}{80\rho}r^2(\rho), \tag{15}$$

then $\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho)$ and $\Psi(\hat{f} - f^*) \leq \rho$.

Approximate minimizers

Until now, mostly estimators were considered that minimize $P_N l_f + \lambda\Psi(f)$. For any such minimizer f we have that $P_N L_f^\lambda \leq 0$. Right here we consider approximate minimizers $\hat{f}_\alpha$ which are any estimator such that $P_N L_f^\lambda \leq \alpha^2$. Due to the nature of the proofs it is straightforward to extend these results.

We only provide an analogue of Theorem 3.2.

Theorem 3.3. *Under the assumptions of Theorem 3.2, we have that*

$$\|\hat{f}_\alpha - f^*\|_{L^2(\mu)} \leq r(\rho) \max\left\{1, \frac{\rho/10 + \rho\frac{80\alpha^2}{\kappa^2\epsilon r^2(\rho)}}{\lambda_0\Delta(\rho) - \rho}, \frac{\lambda_0}{18} + \sqrt{\frac{\lambda_0^2}{324} + \frac{80\alpha^2}{9\kappa^2\epsilon r^2(\rho)}}\right\}, \tag{16}$$

for any α^2-approximate penalized risk minimizer $\hat{f}_\alpha$.

Even though this expression on the RHS looks complicated it reduces to the result in Theorem 3.2 when $\alpha = 0$. It can also be observed that as long as

$$\frac{80\alpha^2}{\kappa^2\epsilon r^2(\rho)} \lesssim 1$$

the rate of convergence (as $N \to \infty$) is of the same order as the rate that would be obtained for an exact (= not approximate) penalized empirical risk minimizer. To obtain a similar rate of convergence an optimization problem only needs to be solved up to an accuracy proportional to $r^2(\rho)$.

3.4.2 Under-regularized estimators

In the previous subsection, risk bounds were proven that hold when $\lambda_0\Delta(\rho) - \rho > 0$. So when λ_0 is too small, this bound is not useful. The following result holds for arbitrary small λ_0.

Theorem 3.4. *Let $\hat{f}$ be the $\lambda-$penalized empirical risk minimizer. Let $\rho > 0$ and recall the definition of the sparsity parameter $\Delta(\rho)$. Let λ_0 be defined by the equation $\lambda = \lambda_0\frac{\kappa^2\epsilon}{80\rho}r^2(\rho)$. Then on $\mathcal{A}$,*

$$\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho)\max\{1, \frac{\Psi(\hat{f}) - \Psi(f^*) + \rho/10}{\Delta(\rho)}, \lambda_0/9\}, \tag{17}$$

whenever $\lambda_0\Delta(\rho) - \rho > 0$.

The proof of this statement follows actually very direct from the sublemmas used in the proof for the over-regularized case (see the proofs section).

These types of under-regularized estimators are quite topical, because from the classical point of view these types of under regularized types of estimators should not be able to perform well. From the classical point of view the parameter λ should be finely tuned in order to balance the bias and variance of the estimator.

Before we compare this result to some other results in the literature, we first illustrate the bounds that can be expected to be derived from this theorem. The quantity on the right hand side in Equation 17 is random, because in general $\Psi(\hat{f})$ is a random quantity. So one needs to bound $\Psi(\hat{f})$. How large $\Psi(\hat{f})$ is, is highly problem dependent. Some examples are provided in Chinot et al. [2022], Koehler et al. [2021]. In general it seems that one can expect that

$$\Psi(\hat{f}) \leq \Psi(f^*) + O(|\xi|).$$

As a consequence, according to this bound, we need $\frac{O(|\xi|)+\rho/10}{\Delta(\rho)} \leq 1$ in order to obtain a bound of similar order as for an optimal choice of λ in Corollary 3.1.

Now we would like to relate this result to some other results in the literature. For this we make a distinction between the situation where the signal to noise ratio is low (when $r_M > r_Q$) and the situation where the signal to noise level is high (when $r_Q > r_M$). Under a high signal to noise level this result is very similar to the main theorem in Chinot et al. [2022]. The difference between this result is that they consider adversarial noise, while here we consider the situation where the noise ξ is sampled i.i.d. according to a fixed distribution. But this difference is immaterial when the signal to noise level is high.

Now we consider the situation with a low signal to noise ratio. Here we need to make a distinction. In the literature there exist many results where bounds on the estimation error are proven under the condition that $Y = f^*(X) + e$ with e a mean zero Gaussian, $f^* \in \mathcal{F}$ and with $\mathcal{F}$ a vector space. But this result holds under much weaker conditions. The only condition related to the distribution of the noise is through the definition of the critical radius r_M. In general it is not possible to recover the aforementioned results under the Gaussian noise assumption using Theorem 3.4.

In Shamir [2022] it is shown that these types of severely under-regularized estimators are biased towards an inconsistent solution in general. This indicates that for general noise models (in general) these under-regularized estimators perform far worse than regularized models with an optimal choice of λ. Only under Gaussian noise assumptions these types of estimators can in general be expected to be consistent. So an interesting problem would be to exploit this Gaussianity assumption in the general context that we consider right here. This problem has already been discussed in Zhou et al. [2021], Koehler et al. [2021] under the additional assumption that not only the noise is Gaussian but also that the vector of covariates is multivariate Gaussian. The main tool that was used in these results is the Convex Gaussian Min-max Theorem Thrampoulidis et al. [2014]. The small ball method and the convex Gaussian Min-max Theorem seem to be interrelated, as for a lot of applications where the small ball method was used, the Convex Gaussian Min-max Theorem is also applicable (see the applications in Thrampoulidis et al. [2014] and Thrampoulidis et al. [2015] and compare to Koltchinskii and Mendelson [2015]).

3.5 Proof of previous results

3.5.1 Norm dominated bounds

The first result is a variant of the argument in Lecué and Mendelson [2018], with the main difference that it holds for a larger range of values of λ. The main result in Lecué and Mendelson [2018] is stated for completeness later this chapter.

Lemma 3.8. *Let $\rho > 0$. Recall the definition of $r(\rho)$. Let $\lambda = \lambda_0 \frac{\kappa^2 \epsilon}{80\rho} r^2(\rho)$. Assume that λ_0 is sufficiently large such that $\lambda_0 \Delta(\rho) > \rho$. Then, on $\mathcal{A}$, for any $\lambda-$penalized empirical risk minimizer $\hat{f}$,*

$$\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho) \frac{\rho/10}{\lambda_0 \Delta(\rho) - \rho} \tag{18}$$

whenever $\rho \|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho) \Psi(\hat{f} - f^*)$.

Proof. Assume that $f \in \mathcal{F}$ and $\rho \|f - f^*\|_{L^2(\mu)} \leq r(\rho)\Psi(f - f^*)$. Define $\mathcal{L}_f^\lambda = \mathcal{L}_f + \lambda(\Psi(f) - \Psi(f^*))$ where $\mathcal{L}_f = l_f - l_{f^*}$. Since $f^* \in \mathcal{F}$ and because $\hat{f}$ is a penalized empirical risk minimizer it follows that $P_N l_{\hat{f}} + \lambda \Psi(\hat{f}) \leq P_N l_f + \lambda \Psi(f)$ for all $f \in \mathcal{F}$. So $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$. We are going to show that $P_N \mathcal{L}_f^\lambda > 0$ whenever $\|f - f^*\|_{L^2(\mu)} \geq r(\rho) \frac{\rho/10}{\lambda_0 \Delta(\rho) - 1}$. Because $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$ this implies the conclusion of the theorem.

First we write out the definition of $P_N \mathcal{L}_f^\lambda$. Recall that $P_N \mathcal{L}_f = P_N(f(X) - Y)^2 - P_N(f^*(X) - Y)^2 = P_N(f(X) - f^*(X))^2 + 2P_N \xi(f(X) - f^*(X))$. So

$$P_N \mathcal{L}_f^\lambda = P_N(f(X) - f^*(X))^2 + 2P_N \xi(f(X) - f^*(X)) + \lambda(\Psi(f) - \Psi(f^*)).$$

We need to lower bound $P_N \mathcal{L}_f^\lambda$ and in order to do so we are going to lower bound each of these three terms.

The first term is non-negative.

The second term can be bounded using a similar method as the method used for empirical risk minimizers. Let $C > 0$ be any constant to be determined later. We want to lower bound $2P_N \xi(f(X) - f^*(X))$. By the convexity of $\mathcal{F}$ it follows that $2\mathbb{E}\xi(f(X) - f^*(X)) \geq 0$ (this is the orthogonality relation). So

$$2P_N \xi(f(X) - f^*(X)) \geq 2(P_N - \mathbb{E})\xi(f(X) - f^*(X)).$$

Now we use a homogeneity argument. Let $\sigma = C\rho/\Psi(f - f^*)$. Define $g = \frac{\rho}{\Psi(f - f^*)}(f - f^*)$. It directly follows that $\Psi(g) \leq \rho$ and by the condition on f it follows that $\|g\|_{L^2(\mu)} \leq r(\rho)$.

But recall that by the definition of $r_M(\rho)$ it follows that with probability at least $1 - \delta$

$$\sup_{f \in \mathcal{F}_c : \Psi(f) \leq \rho, \|f\|_{L^2(\mu)} \leq r(\rho)} (P_N - \mathbb{E})|\xi f| \leq \frac{\kappa^2 \epsilon}{160} r^2(\rho), \tag{19}$$

where $\mathcal{F}_c = \mathcal{F} - f^*$ is $\mathcal{F}$ centered around f^*. Now it follows that

$$2(P_N - \mathbb{E})\xi(f(X) - f^*(X)) \geq \frac{2C}{\sigma}(P_N - \mathbb{E})\xi g \geq \frac{2C}{\sigma}(P_N - \mathbb{E})(-|\xi g|) \geq -\frac{2C}{\sigma}\frac{\kappa^2\epsilon}{160}r^2(\rho).$$

So the second term can be lower bounded by $-\frac{C}{\sigma}\frac{\kappa^2\epsilon}{80}r^2(\rho)$.

Finally we lower bound the last term. By lemma 3.5 it directly follows that

$$\Psi(f) - \Psi(f^*) \geq \frac{\Psi(f - f^*)}{\rho}\Delta(\rho) - \rho/10.$$

Multiplying by C/C and recalling the definition of ρ shows that $\frac{\Psi(f-f^*)}{\rho}\Delta(\rho) - \rho/10 = \frac{C}{\sigma}\Delta(\rho) - \rho/10$.

Putting everything together shows that

$$P_N\mathcal{L}_f^\lambda \geq -\frac{C}{\sigma}\frac{\kappa^2\epsilon}{80}r^2(\rho) + \frac{\lambda C}{\sigma}\Delta(\rho) - \lambda\rho/10.$$

So $P_N\mathcal{L}_f^\lambda > 0$ if

$$\lambda\left(\frac{C\Delta(\rho)}{\sigma} - \rho/10\right) > \frac{C}{\sigma}\frac{\kappa^2\epsilon}{80}r^2(\rho).$$

Using the fact that $\lambda = \lambda_0\frac{\kappa^2\epsilon}{80\rho}r^2(\rho)$, it follows that $P_N\mathcal{L}_f^\lambda > 0$ if $\lambda_0\left(\frac{C\Delta(\rho)}{\rho} - \sigma/10\right) > C$. Rearranging, this condition is equivalent to the condition that $C\left(\lambda_0\frac{\Delta(\rho)}{\rho} - 1\right) > \sigma/10$.

Now we can make some assumptions. Right here we assume that $\sigma \leq 1$. This implies that under the condition that

$$C > \frac{1/10}{\lambda_0\frac{\Delta(\rho)}{\rho} - 1},$$

it follows that $P_N\mathcal{L}_f^\lambda > 0$.

To recap we have shown that for the choice of the constant C large enough it follows that, whenever $\sigma \leq 1$, f cannot be an empirical risk minimizer. So for any empirical risk minimizer $\hat{f}$ it follows that $\sigma > 1$. Using the definition of σ shows that

$$C\rho/\Psi(\hat{f} - f^*) > 1$$

which implies an upper bound on $\Psi(\hat{f} - f^*)$. Finally using the norm dominating condition implies the Lemma. □

Lemma 3.9. *In the regression setting, when $\rho\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho)\Psi(\hat{f} - f^*)$, then*

$$\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho)\frac{\Psi(\hat{f}) - \Psi(f^*) + \rho/10}{\Delta(\rho)}.$$

Proof. This result follows directly from Lemma 3.6 in combination with the norm dominated condition.

□

3.5.2 Loss dominated bounds

The following theorem is a variant of the main result in Lecué and Mendelson [2018], in the loss dominated setting, with the main difference that it holds for all λ.

Lemma 3.10. *Recall the definition of $r(\rho)$. Let $\lambda = \lambda_0 \frac{\kappa^2 \epsilon}{80\rho} r^2(\rho)$. Then on the event $\mathcal{A}$, for any $\lambda-$penalized empirical risk minimizer $\hat{f}$,*

$$\|\hat{f} - f^*\|_{L^2(\mu)} \leq r(\rho) \max\{1, \lambda_0/9\} \tag{20}$$

whenever $\rho\|\hat{f} - f^*\|_{L^2(\mu)} \geq r(\rho)\Psi(\hat{f} - f^*)$.

Proof. Assume that $f \in \mathcal{F}$ and that $\rho\|f - f^*\|_{L^2(\mu)} \geq r(\rho)\Psi(f - f^*)$. We are going to show that whenever $\|f - f^*\|_{L^2(\mu)} > r(\rho)\max\{1, \lambda_0\rho/4\}$, then $P_N\mathcal{L}_f^\lambda > 0$. By Proposition 5.2

$$P_N\mathcal{L}_f^\lambda \geq P_N(f(X) - f^*(X))^2 + 2P_N\xi(f(X) - f^*(X)) + \lambda(\Psi(f) - \Psi(f^*)) \tag{21}$$

Now we first lower bound the first two terms of the right hand side by a scaling argument. Let $g = \frac{r(\rho)}{\|f-f^*\|_{L^2(\mu)}}(f - f^*)$. By the definition of g it directly follows that $\|g\|_{L^2(\mu)} = r(\rho)$. Additionally by the loss dominated condition it follows that $\Psi(g) \leq \rho$. Denote by $\sigma = \frac{r(\rho)}{\|f-f^*\|_{L^2(\mu)}}$. First multiplying by $\sigma^2/\sigma^2 = 1$ and noting that $\sigma\xi(f(X) - f^*(X)) = \xi g$ it follows that

$$\begin{aligned}
&P_N(f - f^*)^2 + 2P_N\xi(f(X) - f^*(X)) \\
=&\frac{1}{\sigma^2}\left[P_N g^2 + 2\sigma P_N \xi g\right] \\
\geq&\frac{1}{\sigma^2}\left[P_N g^2 - 2\sigma P_N |\xi g|\right].
\end{aligned}$$

Now we are going to again lower bound both terms in this expression. The first term can be lower bounded by observing that $\mathbb{E}g^2 = r^2(\rho) \geq r_Q(\rho)$. So by the small ball method it follows that $P_N g^2 \geq \frac{\kappa^2\epsilon}{8}\mathbb{E}g^2 = \frac{\kappa^2\epsilon}{8}r^2(\rho)$. The second term can be lower bounded in exactly the same way as in Equation 19. This shows that $P_N\xi g \geq -P_N|\xi g| \geq -\frac{\kappa^2\epsilon}{80}r^2(\rho)$.

Now we assume that $\sigma \leq 1$. It directly follows that

$$\frac{1}{\sigma^2}\left[P_N g^2 - 2\sigma P_N|\xi g|\right] \geq \frac{1}{\sigma^2}\left[P_N g^2 - 2P_N|\xi g|\right] \geq \frac{1}{\sigma^2}\left[\frac{9\kappa^2\epsilon}{160}r^2(\rho)\right].$$

Finally by definition of σ it follows that $\frac{1}{\sigma^2}\left[\frac{9\kappa^2\epsilon}{80}r^2(\rho)\right] = \frac{9\kappa^2\epsilon}{80}\|f - f^*\|^2_{L^2(\mu)}$, which lower bounds the first two terms.

Bounding the last term of Equation21 is easier. By the triangle inequality and the loss-dominating condition

$$\lambda(\Psi(f) - \Psi(f^*)) \geq -\lambda\Psi(f - f^*) \geq -\frac{\lambda\rho}{r(\rho)}\|f - f^*\|_{L^2(\mu)}.$$

Thus $P_N \mathcal{L}_f^\lambda > 0$ if

$$\frac{9\kappa^2\epsilon}{80}\|f - f^*\|_{L^2(\mu)}^2 > \frac{\lambda\rho}{r(\rho)}\|f - f^*\|_{L^2(\mu)}.$$

Dividing both sides by $\|f - f^*\|_{L^2(\mu)}$ shows that this happens when $\|f - f^*\|_{L^2(\mu)} > \frac{80\lambda\rho}{9\kappa^2\epsilon r(\rho)} = r(\rho)\frac{\lambda_0}{9}$, where the equality follows from the definition of λ in terms of λ_0.

Thus we have proven that $P_N \mathcal{L}_f^\lambda > 0$ if $\|f - f^*\|_{L^2(\mu)} > r(\rho)\frac{\lambda_0}{9}$ and $r(\rho) < \|f - f^*\|_{L^2(\mu)}$. This finishes the proof. □

Now we can finish the proofs of Theorem 3.2 and 3.4. Theorem 3.2 directly follows from Lemma 3.8 and Lemma 3.10. Theorem 3.4 directly follows from Lemma 3.9 and Lemma 3.10.

3.5.3 Approximate minimizers

As mentioned, the proofs for approximate minimizers follow directly from the proofs of the previous lemmas. We will proof both a norm dominated and a loss dominated analogue that will jointly imply Theorem 3.3. Because the statements and proof are so similar to the proofs of Lemma 3.10 and 3.8, we will only point out the differences between both proofs. First the norm dominated statement. By the proof of the norm dominated statement, we have that

$$P_N \mathcal{L}_f^\lambda \geq \lambda\left(\frac{C\Delta(\rho)}{\sigma} + \rho/10\right) - \frac{C}{\sigma}\frac{\kappa^2\epsilon}{80}r^2(\rho).$$

Thus $P_N \mathcal{L}_f^\lambda > \alpha^2$ if

$$\lambda\left(\frac{C\Delta(\rho)}{\sigma}\rho/10\right) > \frac{C}{\sigma}\frac{\kappa^2\epsilon}{80}r^2(\rho) + \alpha^2.$$

Performing the same steps as in the proof of Lemma 3.8 and assuming that $\sigma \leq 1$ shows that $P_N \mathcal{L}_f^\lambda > \alpha^2$ when

$$C > \frac{1/10 + \frac{80\alpha^2}{\kappa^2\epsilon r^2(\rho)}}{\lambda_0\Delta(\rho)/\rho - 1},$$

which finishes the norm dominated part.

The loss dominated result follows in a similar way. Namely we have in that case that $P_N \mathcal{L}_f^\lambda > \alpha^2$ if

$$\frac{9\kappa^2\epsilon}{80}\|f - f^*\|_{L^2(\mu)}^2 > \frac{\lambda\rho}{r(\rho)}\|f - f^*\|_{L^2(\mu)} + \alpha^2.$$

Multiplying this equation by $\frac{80}{9\kappa^2\epsilon}$ and applying the quadratic formula shows that $P_N\mathcal{L}_{\hat{f}}^{\lambda} > \alpha^2$ if

$$\|f - f^*\|_{L^2(\mu)} > r(\rho)\frac{\lambda_0}{18} + \sqrt{r^2(\rho)\frac{\lambda_0^2}{324} + \frac{80\alpha^2}{9\kappa^2\epsilon}},$$

which implies a loss dominated statement. These two results combined imply Theorem 3.3.

3.6 Setting i revisited: General penalized empirical risk minimization

Now we revisit the problem of empirical risk minimization for general loss functions, in the context of penalized estimators. Given a linear space E equipped with norm Ψ, let $\mathcal{F}$ be a subset of E. For every $f \in \mathcal{F}$, let l_f be a loss function on a probability space Ω. So $\mathcal{F}$ parameterizes a family of loss functions on Ω. Given an i.i.d. sample $X_1, .., X_N \in \Omega$, we consider any estimator $l_{\hat{f}}$ that minimizes

$$P_N l_f + \lambda\Psi(f)$$

over all $f \in \mathcal{F}$. The goal is to minimize $\mathbb{E}l_f$ over $\mathcal{F}$. Let f^* be any minimizer of $\mathbb{E}l_f$ in $\mathcal{F}$ (which we assume exists). We measure how good an estimator is in terms of the excess risk $\mathbb{E}\left[l_{\hat{f}} - l_{f^*}\right]$. Right here we are going to assume that $l_f - l_{f^*}$ is of a certain form.

Recall that in the context of penalized estimators in regression, homogeneity played an important role in the proofs. We will make a similar assumption. First we will write $l_f - l_{f^*} = B(f - f^*, f^*)$. So $B(f - f^*, f^*)$ is a loss function on Ω parameterized by $f - f^*$ and f^*. We will assume that B is homogeneous of order $\alpha \geq 1$ under scaling of $f - f^*$, which means that for any constant $c \geq 0$

$$B(c(f - f^*), f^*) = c^{\alpha}B(f - f^*, f^*).$$

Important examples of loss functions of this type are linear and quadratic forms in $f - f^*$. An example of a loss function that is homogeneous of order 1 is shown in Vu et al. [2013], where sparse principal component analysis is considered with a $\|\cdot\|_1$ norm penalty. Notice that for example the regression setting does not fall within this framework, because here $l_f - l_{f^*}$ is the sum of two homogeneous loss functions. The result below can be generalized to sums of homogeneous loss functions. To keep the proofs short and simple we will state and proof the results only for a single homogeneous loss function.

As mentioned, regression with respect to the squared loss falls within this framework. Another motivation is provided by computing a truncated Taylor expansion of $l_f - l_{f^*}$ around f^* which also leads to a representation of this form. Especially this second example shows that a wide range of loss functions $l_f - l_{f^*}$ admits such a representation.

Now we present a result when B is homogeneous of order α which is an analogue to Theorem 3.2. We present it in a fashion similar to Corollary 3.1.

Theorem 3.5. *Assume that $l_f - l_{f^*} = B(f - f^*, f^*)$ is homogeneous of order $\alpha \geq 1$. Let $T(f) = \mathbb{E}[l_f - l_{f^*}]$. Recall the definition of $\Delta(\rho)$ for a general function $T(f)$ on $\mathcal{F}$ and choose $\rho > 0$ such that $\Delta(\rho) \geq \frac{4\rho}{5}$. Let r, C_1, C_2 be constants such that*

1. *(Rademacher bound)*

$$\mathbb{E} \sup_{f \in \mathcal{F}: \Psi(f-f^*) \leq \rho, T(f) \leq r} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i (l_f - l_{f^*})(X_i) \leq C_1 r$$

2. *(Small ball assumption) For all $f \in \mathcal{F}$ with $\Psi(f - f^*) \leq \rho$ and $T(f) \geq r$, we have that*

$$P_N(l_f - l_{f^*}) \geq C_1 T(f).$$

Define the normalized regularization parameter λ_0 by $\lambda = \lambda_0 \frac{20r}{\rho}$. Assume that λ_0 satisfies $C_2 < \lambda_0 < 20C_1$. Then any penalized empirical risk minimizer $\hat{f}$ satisfies

$$T(\hat{f}) \leq r.$$

First we proof a lemma that is an exact analogue of Lemma 3.5.

Lemma 3.11. *Recall the definition of $\Delta(\rho)$ for a general function T. Then for any $f \in \mathcal{F}$ with $\rho^\alpha T(f) \leq r(\rho)\Psi(f - f^*)^\alpha$,*

$$\Psi(f) - \Psi(f^*) \geq \left[\frac{\rho}{\Psi(f - f^*)}\right]^\alpha T(f) - \rho/10.$$

Proof. By the proof of Lemma 3.5 we have that

$$\Psi(f) - \Psi(f^*) \geq z^*(f - f^*) - \rho/10.$$

Once again a scaling argument is applied. Let $g = \frac{\rho}{\Psi(f-f^*)}(f - f^*)$. Observe that $\Psi(g) = \rho$ and by homogeneity,

$$T(g) = \mathbb{E}B(g, f^*) = \left[\frac{\rho}{\Psi(f - f^*)}\right]^\alpha \mathbb{E}B(f - f^*, f^*) = \left[\frac{\rho}{\Psi(f - f^*)}\right]^\alpha T(f).$$

Finally using the norm dominated condition shows that $T(g) \leq r(\rho)$. It now follows from the definition of $\Delta(\rho)$, that $z^*(f - f^*) = \frac{\Psi(f-f^*)}{\rho} z^*(g) = \frac{\Psi(f-f^*)}{\rho}\Delta(\rho)$. □

Lemma 3.12. *Assume that $\lambda > \frac{20C_2 r}{\rho}$. For any $f \in \mathcal{F}$ such that $\rho^\alpha T(f) \leq r\Psi(f - f^*)^\alpha$ and $\Psi(f - f^*) \geq \rho$ and $T(f) \geq r$,*

$$P_N(l_f - l_{f^*}) + \lambda(\Psi(f) - \Psi(f^*)) > 0.$$

A direct consequence of this result is that in the norm dominated setting whenever λ is large enough it follows that $\Psi(\hat{f} - f^*) < \rho$ or $T(\hat{f}) < r$. But the norm dominated condition implies that when $\Psi(\hat{f} - f^*) < \rho$, then $T(\hat{f}) < r$.

Proof. Let $f \in \mathcal{F}$ such that $\rho^\alpha T(f) \leq r\Psi(f-f^*)^\alpha$ and $\Psi(f-f^*) \geq \rho$ and $T(f) \geq r$. Let $g = \frac{\rho}{\Psi(f-f^*)}(f-f^*)$. By the proof of Lemma 3.11, $\Psi(g) = \rho$ and $T(g) \leq r$. Also $B(f-f^*, f^*) = \left[\frac{\Psi(f-f^*)}{\rho}\right]^\alpha B(g, f^*)$. But $\frac{\Psi(f-f^*)}{\rho} \geq 1$, so $B(f-f^*, f^*) \geq B(g, f^*)$. We are going to lower bound $P_N(l_f - l_{f^*})$. Using the fact that f^* is an empirical risk minimizer shows that

$$P_N(l_f - l_{f^*}) = \mathbb{E}(l_f - l_{f^*}) + (P_N - \mathbb{E})(l_f - l_{f^*}) \geq (P_N - \mathbb{E})(l_f - l_{f^*}) \geq -2C_2 r(\rho),$$

where the last inequality follows from the symmetrization theorem and the Rademacher condition.

By Lemma 3.11 and the norm dominating condition, it follows that

$$\lambda(\Psi(f) - \Psi(f^*)) \geq \lambda\left[\left(\frac{T(f)}{r}\right)^\alpha \Delta(\rho) - \rho/10\right].$$

Using that $T(f) \geq r$ and the fact that $\Delta(\rho) \geq 4\rho/5$ shows that

$$P_N(l_f - l_{f^*}) + \lambda(\Psi(f) - \Psi(f^*)) > 0$$

if $\lambda\rho/10 > 2C_2 r$, which proves the lemma. □

Finally we are going to proof a loss dominated result.

Lemma 3.13. *Let $f \in \mathcal{F}$ such that $\rho^\alpha T(f) \geq r\Psi(f-f^*)^\alpha$. Then*

$$P_N(l_f - l_{f^*}) + \lambda(\Psi(f) - \Psi(f^*)) > 0$$

if $C_1 T(f) > \lambda\rho\left[\frac{T(f)}{r}\right]^{1/\alpha}$.

Proof. Let $f \in \mathcal{F}$ such that $\rho^\alpha T(f) \geq r\Psi(f-f^*)^\alpha$. Once again we apply a scaling argument. This time we let $g = \left[\frac{r}{T(f)}\right]^{1/\alpha}(f-f^*)$. An easy computation using the homogeneity of B shows that $T(g) = r$. By the loss dominated condition, it follows that $\Psi(g) \leq \rho$. Now we use the small ball assumption. By the definition of g, we have that

$$P_N(l_f - l_{f^*}) = P_N B(f-f^*, f^*) = \frac{T(f)}{r} P_N B(g, f^*) \geq \frac{T(f)}{r} C_1 \mathbb{E}B(g, f^*) = C_1 T(f).$$

Now we lower bound $\lambda(\Psi(f) - \Psi(f^*))$ exactly in the same way as in the proof of Lemma 3.10. By the loss dominating condition it follows that

$$\lambda(\Psi(f) - \Psi(f^*)) \geq -\lambda\rho\left[\frac{T(f)}{r}\right]^{1/\alpha}.$$

So putting both bounds together shows that

$$P_N(l_f - l_{f^*}) + \lambda(\Psi(f) - \Psi(f^*)) > 0$$

if $C_1 T(f) > \lambda\rho \left[\frac{T(f)}{r}\right]^{1/\alpha}$.

Recall that $\lambda = \lambda_0 \frac{20r}{\rho}$. The condition that $C_1 T(f) > \lambda\rho \left[\frac{T(f)}{r}\right]^{1/\alpha}$ is equivalent to $20 C_1 T(f)^{1-1/\alpha} > \lambda_0 r^{1-1/\alpha}$. So if $T(f) > r$, it is sufficient to choose $C_1 > \lambda_0/20$. □

4 Discussion and conclusion

In this report we developed a version of the small ball assumption that is useful for a wider variety of function classes. This extension is called the delocalized small ball assumption (DSBA), because it only needs to hold outside of a ball of radius r. It was shown that the DSBA can be applied in the bounded setting and to functions in Sobolev spaces that lie above the critical line for the Sobolev embedding theorem.

The application of the small ball method to regression problems was explained. A simplified proof of a statement due to Lecue and Mendelson Lecué and Mendelson [2018] was given, which afforded generalizing this result. In Lecué and Mendelson [2018] various applications were discussed using the standard small ball assumption, but further applications using the DSBA were not given. So it would be interesting to see what results could be obtained by applying the DSBA to this setting.

Our simplified proof provides in a reasonably direct way new bounds for under regularized penalized empirical risk minimizers. It is clear that this new bound is not optimal for models having Gaussian noise. In general it is unclear whether this result can be improved, e.g. under Gaussian noise assumptions.

The DSBA was applied to generic empirical risk minimizers. Here it was possible to recover a result from Bartlett et al. [2005] up to constants. Regarding generic penalized empirical risk minimizers, a result was established analogous to Theorem 3.2. A certain homogeneity assumption were presupposed. Also in this setting it would be interesting to see what results can be obtained in specific applications.

It is further interesting to analyze to which settings the DSBA can be applied, e.g. in hypothesis testing. Furthermore contraction rates for Bayesian estimators can be established under the condition that certain tests exist (see Theorem 8.9 in Ghosal and Van der Vaart [2017] for example). It would be of interest to understand how the small ball method can be applied in those settings.

5 Appendix

In this section some background results are stated and proven.

5.1 Basic results

First we discuss some properties of the functional $\mathcal{L}_f^\lambda$ that are used in this report.

Proposition 5.1. *Whenever $\hat{f}$ is a minimizer of the penalized empirical risk functional $f \mapsto P_N \mathcal{L}_f^\lambda$ and f^* minimizes the population risk, then*

1. $P_N \mathcal{L}_{f^*}^\lambda = 0$
2. $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$

Proof. The first statement in this proposition immediately follows from the definition of $\mathcal{L}_f^\lambda$ and the second part follows from the fact that $\hat{f}$ minimizes $P_N \mathcal{L}_{\hat{f}}^\lambda$ over $\mathcal{F}$ and the fact that $f^* \in \mathcal{F}$. □

The following lower bound on the empirical penalized excess risk is often used.

Proposition 5.2. *Recall that $\xi = f^*(X) - Y$. Assume that f^* minimizes $f \mapsto \mathbb{E}[(Y - f(X))^2]$ in $\mathcal{F}$. Then*

$$P_N \mathcal{L}_f \geq P_N (f - f^*)^2(X) + 2(P_N \xi (f - f^*)(X) - \mathbb{E}\xi(f - f^*)(X)). \tag{22}$$

Proof. By definition,

$$\begin{aligned} P_N \mathcal{L}_f =& P_N l_f - P_N l_{f^*} = P_N (Y - f(X))^2 - P_N (Y - f^*(X))^2 \\ =& P_N (f - f^*)^2(X) + 2(P_N \xi (f - f^*)(X). \end{aligned}$$

It remains to show that $\mathbb{E}\xi(f - f^*)(X) \geq 0$. This follows because f^* minimizes $\mathbb{E}(Y - f(X))^2$. Thus for any $f \in \mathcal{F}$ we have that $0 \geq \mathbb{E}(Y - f^*(X))^2 - \mathbb{E}(Y - f(X))^2 = \mathbb{E}(f - f^*)^2(X) - 2\mathbb{E}\xi(f - f^*)(X) \geq -2\mathbb{E}\xi(f - f^*)(X)$. This final claim follows because $\mathbb{E}(f - f^*)^2(X)$ is non-negative. So $0 \leq 2\mathbb{E}\xi(f - f^*)(X)$. □

Lemma 5.1. *Let $\mathcal{F} \subset L^2(\mu)$.*

1. *If $\mathcal{F}$ is a closed subvectorspace of $L^2(\mu)$, then $\mathcal{L}(f, f^*) = \|f - f^*\|_{L^2(\mu)}$ for any $f \in \mathcal{F}$.*
2. *If $\mathcal{F}$ is a closed convex subset of $L^2(\mu)$, then $\mathcal{L}(f, f^*) \geq \|f - f^*\|_{L^2(\mu)}$ for any $f \in \mathcal{F}$.*

Proof. By definition $\mathcal{L}(f, f^*) = \mathbb{E}\mathcal{L}_f$. Using the definition of $\mathcal{L}_f$ it follows that

$$\mathcal{L}_f = (f - f^*)^2(X) + 2\xi(f - f^*).$$

So the proposition follows if $\mathbb{E}\xi(f - f^*)(X)$ is zero or non-negative respectively when $\mathcal{F}$ is a subspace or when $\mathcal{F}$ is convex. In the convex case this follows from the previous proposition and when $\mathcal{F}$ is a subspace this follows from the characterization of the nearest point map. □

5.2 Concentration inequalities

Lemma 5.2. *(Bounded differences inequality) Let $f : \Omega^N \mapsto [0,1]$ be a measurable function. Let X be a random variable on Ω and let $X_1, ..., X_N$ be i.i.d. copies of X. Let $Z = f(X_1, ..., X_N)$. Then*

$$P\{Z - \mathbf{E}[Z] > t\} \leq e^{-8t^2/N}. \tag{23}$$

Proof. See Theorem 6.2 on page 166 in Boucheron et al. [2013]. □

5.3 Contraction and symmetrization theorems

Lemma 5.3. *Let F be a class of functions on a probability space (Ω, μ). Suppose that $\mathcal{L}(X) = \mu$ and let $X_1, ..., X_N \sim X$ be i.i.d. Let P_N be the associated empirical measure. Let $\epsilon \in \{-1, 1\}^N$ be uniformly distributed. Then*

(i)

$$\mathbb{E}\Big[\sup_{f\in F} P_N f(X) - \mathbb{E}f(X)\Big] \leq 2\mathbb{E}\Big[\sup_{f\in F} \frac{1}{N}\sum_{i=1}^{N} \epsilon_i f(X_i)\Big] \tag{24}$$

(ii) If $\phi : \mathbb{R} \mapsto \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $Lip(\phi)$, then

$$\mathbb{E}\Big[\sup_{f\in F} \frac{1}{N}\sum_{i=1}^{N} \phi(\epsilon_i f(X_i))\Big] \leq 2Lip(\phi)\mathbb{E}\Big[\sup_{f\in F} \frac{1}{N}\sum_{i=1}^{N} \epsilon_i f(X_i)\Big] \tag{25}$$

Proof. Observe that $f(X) - \mathbb{E}f(X)$ is centered. Therefore (i) follows from lemma 11.4 on page 322 from Boucheron et al. [2013]. Statement (ii) follows from Theorem 4.12 in Ledoux and Talagrand [1991]. □

Finally we present the proof of the fact that the small ball assumption holds for non-negative functions f that satisfy an $L^p - L^q$ norm equivalence. First we state a result Mendelson [2015] obtained using the Paley-Zygmund inequality De la Pena and Giné [2012].

Proposition 5.3. *Let $r > 2 \geq 1$ and let g be a function on Ω such that $\|g\|_{L^r} \leq C\|g\|_{L^2}$. Then for any $u \in (0,1)$, $P(|g| \geq u\|g\|_{L^2}) \geq \left(\frac{1-u^2}{C^2}\right)^{\frac{r}{r-2}}$.*

Now let $p > q \geq 1$ and assume that f is such that

$$\|f\|_{L^p} \leq B\|f\|_{L^q}.$$

Observe that $\|f\|_{L^q} = \||f|^{q/2}\|_{L^2}^{2/q}$ and that $\|f\|_{L^p} = \||f|^{q/2}\|_{L^{2p/q}}^{2/q}$. So f also satisfies

$$\||f|^{q/2}\|_{L^{2p/q}} \leq B^{q/2}\||f|^{q/2}\|_{L^2}.$$

Now we apply Proposition 5.3 with $C = B^{q/2}$, $g = |f|^{q/2}$, and $r = 2p/q$. Notice that $r > 2$ as long as $p > q$. This implies for any $u \in (0,1)$ that

$$P(|f|^{q/2} \geq u \||f|^{q/2}\|_{L^2} \geq \left(\frac{1-u^2}{C^2}\right)^{\frac{r}{r-2}} = \left(\frac{1-u^2}{B^q}\right)^{\frac{2p}{2p-2q}}.$$

Setting $u = \kappa^{q/2}$ proves that

$$P(|f| \geq \kappa \|f\|_{L^q}) \geq \left[\frac{1-\kappa^q}{B^q}\right]^{\frac{2p}{2p-2q}},$$

which finishes the proof.

References

Chao Ma, Lei Wu, et al. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Jie Gui, Zhenan Sun, Shuiwang Ji, Dacheng Tao, and Tieniu Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE transactions on neural networks and learning systems*, 28(7):1490–1507, 2016.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.

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

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015:12991–13008, 2015. ISSN 1073-7928. doi:10.1093/imrn/rnv096.

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

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method ii: complexity dependent error rates. *The Journal of Machine Learning Research*, 18(1):5356–5403, 2017a.

Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, pages 67–101, 2015.

Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, 19(3):881–904, 2017b.

Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4):2306–2333, 2022.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings 15*, pages 44–58. Springer, 2002.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Geoffrey Chinot and Matthieu Lerasle. On the robustness of the minimum l_2 interpolator. *arXiv preprint arXiv:2003.05838*, 2020.

Shahar Mendelson. A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia, February 11–22, 2002 Revised Lectures*, pages 1–40. Springer, 2003.

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

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

Shahar Mendelson. Extending the scope of the small-ball method. *arXiv preprint arXiv:1709.00843*, 2017.

Elias M Stein and Timothy S Murphy. *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, volume 3. Princeton University Press, 1993.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 1998.

Juha Heinonen, Pekka Koskela, Nageswari Shanmugalingam, and Jeremy T Tyson. *Sobolev spaces on metric measure spaces*. Number 27. cambridge university press, 2015.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. 2006.

Ohad Shamir. The implicit bias of benign overfitting. In *Conference on Learning Theory*, pages 448–478. PMLR, 2022.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *arXiv preprint arXiv:2112.04470*, 2021.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. *Advances in neural information processing systems*, 26, 2013.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

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

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.