Journal de l'École polytechnique
Mathématiques

Vladimir Koltchinskii & Stanislav Minsker

L_1-penalization in functional linear regression with subgaussian design

Tome 1 (2014), p. 269-330.

<http://jep.cedram.org/item?id=JEP_2014__1__269_0>

Journal de l'École polytechnique
Mathématiques

Tome 1, 2014, p. 269–330 DOI: 10.5802/jep.11

L_1-PENALIZATION IN FUNCTIONAL LINEAR REGRESSION WITH SUBGAUSSIAN DESIGN

BY VLADIMIR KOLTCHINSKII & STANISLAV MINSKER

ABSTRACT. — We study functional regression with random subgaussian design and real-valued response. The focus is on the problems in which the regression function can be well approximated by a functional linear model with the slope function being "sparse" in the sense that it can be represented as a sum of a small number of well separated "spikes". This can be viewed as an extension of now classical sparse estimation problems to the case of infinite dictionaries. We study an estimator of the regression function based on penalized empirical risk minimization with quadratic loss and the complexity penalty defined in terms of L_1-norm (a continuous version of LASSO). The main goal is to introduce several important parameters characterizing sparsity in this class of problems and to prove sharp oracle inequalities showing how the L_2-error of the continuous LASSO estimator depends on the underlying sparsity of the problem.

RÉSUMÉ (Pénalisation L_1 en régression fonctionnelle linéaire avec design sous-gaussien)

Nous étudions la régression fonctionnelle linéaire avec design sous-gaussien et la réponse à valeurs réelles. Nous nous concentrons sur les problèmes où la fonction de régression est bien approchée par un modèle fonctionnel linéaire dont la pente est « sparse » dans le sens où elle peut être représentée comme une somme d'un petit nombre de « pics » séparés. Nous pouvons considérer ce problème comme une extension du problème classique d'estimation « sparse » au cas d'un dictionnaire infini. Nous étudions un estimateur de la fonction de régression basé sur la minimisation du risque empirique pénalisé avec une perte quadratique et avec une pénalité de complexité définie en termes de la norme L_1 (une version continue du LASSO). L'objectif principal est d'introduire certains paramètres importants qui caractérisent la « sparsité » dans cette classe de problèmes et de prouver des inégalités d'oracle « sparses » montrant comment l'erreur L_2 de la version continue du LASSO dépend de la sparsité sous-jacent du problème.

CONTENTS

MATHEMATICAL SUBJECT CLASSIFICATION (2010). — 62J02, 62G05, 62J07.
KEYWORDS. — Functional regression, sparse recovery, LASSO, oracle inequality, infinite dictionaries.

V. Koltchinskii was partially supported by NSF grants DMS-1207808, DMS-0906880, CCF-0808863 and CCF-1415498.
S. Minsker was partially supported by grant R01-ES-017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH), NSF grants DMS-0650413, CCF-0808847 and DOE contract 113054 G002745.

1. Introduction

Let (X, Y) be a random couple defined on a probability space $(\Omega, \Sigma, \mathbb{P})$, where $X = \{X(t) : t \in \mathbb{T}\}$ is a stochastic process with parameter set $\mathbb{T}$ and Y is a real valued response variable. In what follows, it will be assumed that the process X is *subgaussian*. Denote

$$d_X(s,t) := \sqrt{\mathrm{Var}\left(X(s) - X(t)\right)}, \quad s,t \in \mathbb{T}. \tag{1.1}$$

It will be also assumed that the space $\mathbb{T}$ is totally bounded with respect to pseudo-metric d_X and, moreover, it satisfies Talagrand's generic chaining conditions ensuring that there exists a version of the process $X(t)$, $t \in \mathbb{T}$ that is a.s. uniformly bounded and d_X-uniformly continuous. In what follows, we assume that $X(t)$, $t \in \mathbb{T}$ is such a version. Let μ be a finite measure on the Borel σ-algebra $\mathscr{B}_{\mathbb{T}}$ of the pseudometric space $(\mathbb{T}, d_X)$.

Consider the following regression model

$$Y = f_*(X) + \xi,$$

where $f_*(X) = \mathbb{E}(Y|X)$ is the regression function and ξ is a random noise with $\mathbb{E}\xi = 0$ and variance $\mathrm{Var}(\xi) = \sigma_\xi^2$ independent of the design variable X. We will be interested in estimating the regression function $f_*(X)$ under an underlying assumption that $f_*(X)$ can be well approximated by a functional linear model ("oracle model")

$$f_{\lambda,a}(X) = a + \int_{\mathbb{T}} X(t)\lambda(t)\mu(dt),$$

where $\lambda \in L_1(\mu)$ is the "slope" function and $a \in \mathbb{R}$ is the intercept of the model. More precisely, we will focus on the problems in which the oracle models are "sparse" in the sense that the slope function λ is supported in a relatively small subset $\mathrm{supp}(\lambda) := \{t \in \mathbb{T} : \lambda(t) \neq 0\}$ of parameter space $\mathbb{T}$ such that the set of random variables $\{X(t) : t \in \mathrm{supp}(\lambda)\}$ can be well approximated by a linear space of a small dimension. Often, λ will be a sum of several "spikes" with disjoint and well separated supports. Such models might be useful in a variety of applications, in particular, in image processing where, in many cases, only sparsely located regions of the image are correlated with the response variable. In what follows, Π denotes the marginal distribution of X in the space $C_{bu}(\mathbb{T}; d_X)$ of all uniformly bounded and uniformly continuous functions on $(\mathbb{T}; d_X)$, and P denotes the joint distribution of (X, Y) in

$C_{bu}(\mathbb{T}; d_X) \times \mathbb{R}$. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sample consisting of n i.i.d. copies of (X, Y) defined on $(\Omega, \Sigma, \mathbb{P})$. The regression function f_* is to be estimated based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$. Our estimation method can be seen as a direct extension of (a version of) LASSO to the infinite-dimensional case. Namely, let $\mathbb{D}$ be a convex subset of the space $L_1(\mu)$ such that $0 \in \mathbb{D}$. Consider the following penalized empirical risk minimization problem:

$$(\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon) := \operatorname*{argmin}_{\lambda \in \mathbb{D},\, a \in \mathbb{R}} \left[\frac{1}{n} \sum_{j=1}^{n} (Y_j - f_{\lambda,a}(X_j))^2 + \varepsilon \|\lambda\|_1 \right], \tag{1.2}$$

where

$$\|\lambda\|_1 := \|\lambda\|_{L_1(\mu)} = \int_{\mathbb{T}} |\lambda(t)| \mu(dt)$$

and $\varepsilon > 0$ is the regularization parameter. The function $f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon}$ will be used as an estimator of the regression function f_*.

When the parameter set $\mathbb{T}$ is finite, (1.2) defines a standard LASSO-estimator of the vector of parameters of linear regression model (see [40]). This estimator is among the most popular in high-dimensional statistics and it has been intensively studied in the recent years (e.g., see [9], [16], [25], [6], [26], [4], [27]; see also the book by Bühlmann and van de Geer [8] for further references).

We will be more interested in the case of uncountable infinite parameter sets $\mathbb{T}$ (functional linear models). In such problems, standard characteristics of finite dictionaries used in the theory of sparse recovery (restricted isometry constants, restricted eigenvalues, etc) are not directly applicable. Our goal will be to develop proper parameters characterizing sparsity in the case of functional models and to prove oracle inequalities for the $L_2(\Pi)$-error $\|f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon} - f_*\|^2_{L_2(\Pi)}$ of continuous LASSO-estimator in terms of these sparsity parameters. We concentrate on the case of subgaussian random design (that, of course, includes an important example of Gaussian design processes) since, in this case, we can rely on a number of probabilistic tools from the theory of subgaussian and empirical processes. In particular, we extensively use in the proofs recent generic chaining bounds for empirical processes due to Mendelson [33], [32].

It should be emphasized that there is vast literature on functional regression (see, e.g., [35], [36] and references therein). A commonly used general idea in this literature is to estimate the eigenfunctions of the covariance operator and to project the unknown slope function onto the linear span of the "principal components" corresponding to the largest eigenvalues (see [34], [10] and references therein). Under smoothness assumptions on the slope function, a natural approach to its estimation is to use a regularization penalty (see [14] for construction of estimators based on smoothing splines and [42] for a more general reproducing kernel Hilbert space approach).

The problem studied in our paper is much closer to the theory of sparse estimation in high-dimensional statistics and can be viewed as an extension of this theory to the case of functional models and uncountable dictionaries. Our approach is similar in spirit to [24], [26] where such characteristics as "alignment coefficient" (used below

for functional models) were introduced and studied in the case of finite dictionaries, and [28] which extended some of these results to the case of infinite dictionaries. For a review of some modern methods in functional data processing and their connections to various notions of sparsity, we refer the reader to [21]. A recent paper by James, Wang and Zhu [22] is similar to the present work in terms of motivation and approach, however, the theoretical analysis in [22] is performed under the assumptions on the design distribution that might not hold if X has smooth trajectories.

It is important to note that in practice we never observe the whole trajectory of X but rather its densely sampled version. In this case, the natural choice for μ is a uniform measure on the sampling grid, whence (1.2) becomes the usual LASSO once again. However, there is often no reasons to assume that Gram matrix of the design satisfies RIP [13] or restricted eigenvalue type conditions [6, 23] in this case. Although LASSO might not perform well as a variable selection procedure in such a framework, we will provide examples showing that prediction power of an estimator can still benefit from the fact that the underlying model is (approximately) sparse. In particular, oracle inequalities with error rates depending on sparsity can be derived from the general results of our paper. Other interesting approaches to theoretical analysis of LASSO with highly correlated design were proposed in [17], [19]. For instance, in [17] (see, in particular, Corollary 4.2) the authors show that in the case of highly correlated design, it is often possible to choose the regularization parameter to be small $\varepsilon \ll n^{-1/2}$ and achieve reasonable error rates.

It should be also mentioned that in a number of very important applications one has to deal with sparse recovery in infinite dictionaries with random designs that are not subgaussian, or with deterministic designs. For instance, in [12], the authors develop a theory of super-resolution. In this case, the dictionary consists of complex exponentials $e^{i\langle t,\cdot\rangle}$, $t \in \mathbb{T} \subset \mathbb{R}^d$, the design is deterministic and the estimation method is based on minimizing the total variation norm of a signed measure Λ on $\mathbb{T}$ subject to data dependent constraints. Although the results of our paper do not apply immediately to such problems, it is possible to extend our approach in this direction.

We will introduce several assumptions and definitions used throughout the paper.

Definition 1.1. — A closed linear subspace $\mathscr{L} \subset L_2(\mathbb{P})$ will be called a *subgaussian space* if there exists a constant $\Gamma > 0$ such that for all $\eta \in \mathscr{L}$

$$\mathbb{E}e^{s\eta} \leqslant e^{\Gamma s^2 \sigma_\eta^2}, \quad s \in \mathbb{R},$$

where $\sigma_\eta^2 := \mathrm{Var}(\eta)$.

It is well known that $\mathbb{E}\eta = 0, \eta \in \mathscr{L}$ and that ψ_2-and L_2-norms are equivalent on $\mathscr{L}$ (more precisely, they are within a constant $\sim \Gamma$). Also, if $\mathscr{L}$ is a closed linear subspace of $L_2(\mathbb{P})$ such that $\{\eta : \eta \in \mathscr{L}\}$ are jointly normal centered random variables, then $\mathscr{L}$ is a subgaussian space with $\Gamma = 1$. Another example is the closed linear span of independent centered subgaussian random variables $\{\eta_j\}$ such that

$$\mathbb{E}e^{s\eta_j} \leqslant e^{\Gamma \sigma_{\eta_j}^2 s^2}, \quad s \in \mathbb{R},\ j \geqslant 1$$

for some $\Gamma > 0$:

$$\mathscr{L} := \Big\{ \textstyle\sum_{j\geqslant 1} c_j\eta_j : \sum_{j\geqslant 1} \sigma_{\eta_j}^2 c_j^2 < +\infty \Big\}.$$

For instance, one can consider a sequence $\{\eta_j\}$ of i.i.d. Rademacher random variables (that is, η_j takes valued $+1$ and -1 with probability $1/2$). In the case of a single random variable η, its linear span is a subgaussian space if and only if η is subgaussian.

In what follows, a subgaussian space $\mathscr{L}$ and constant Γ will be fixed. All the constants depending only on Γ will be called *absolute*.

ASSUMPTION 1.1. — Suppose that

$$X(t) - \mathbb{E}X(t) \in \mathscr{L} \quad \text{for all } t \in \mathbb{T}. \tag{1.3}$$

Denote by $\mathscr{L}_X$ the closed (in L_2 and, as a consequence, also in the ψ_2-norm) linear span of $\{X(t) - \mathbb{E}X(t) : t \in \mathbb{T}\}$.

This assumption easily implies that the stochastic process $Z(t) := X(t) - \mathbb{E}X(t)$, $t \in \mathbb{T}$, is subgaussian, meaning the for all $t, s \in \mathbb{T}$, $Z(t) - Z(s)$ is a subgaussian random variable with parameter $\Gamma d_X^2(t, s)$.

Next, we recall the notion of Talagrand's generic chaining complexity (see [39] for a comprehensive introduction). Given a pseudo-metric space $(\mathbb{T}, d_X)$, let $\{\Delta_n\}$ be a nested sequence of partitions such that $\operatorname{Card}\Delta_0 = 1$ and $\operatorname{Card}\Delta_n \leqslant 2^{2^n}$. For $s \in \mathbb{T}$, let $\Delta_n(s)$ be the unique subset of Δ_n containing s. The generic chaining complexity $\gamma_2(\mathbb{T}; d_X)$ is defined as

$$\gamma_2(\mathbb{T}; d_X) := \inf_{\{\Delta_n\}} \sup_{s\in\mathbb{T}} \sum_{n\geqslant 0} 2^{\frac{n}{2}} D(\Delta_n(s)),$$

where $D(A)$ stands for the diameter of a set A. Let

$$\gamma_2(\delta) := \gamma_2(\mathbb{T}; d_X; \delta) = \inf_{\{\Delta_n\}} \sup_{t\in\mathbb{T}} \sum_{n\geqslant 0} 2^{n/2} \left(D(\Delta_n(t)) \wedge \delta\right).$$

If d_Y is another metric on $\mathbb{T}$ such that $d_Y(t, s) \leqslant d_X(t, s)$ for all $t, s \in \mathbb{T}$, and $\sup_{t,s\in\mathbb{T}} d_Y(t, s) \leqslant \delta$, then clearly

$$\gamma_2(\mathbb{T}; d_Y) \leqslant \gamma_2(\delta). \tag{1.4}$$

This bound will be often used below. Our main complexity assumptions on the design distribution are the following:

ASSUMPTION 1.2. — Pseudometric space $(\mathbb{T}, d_X)$ is such that $\gamma_2(\mathbb{T}; d_X) < \infty$ and, moreover,

$$\gamma_2(\mathbb{T}; d_X; \delta) \longrightarrow 0 \quad \text{as } \delta \longrightarrow 0.$$

Under these assumptions, the process $Z = X - \mathbb{E}X$ has a version that is uniformly bounded and d_X-uniformly continuous a.s. Moreover, $\big\|\|X - \mathbb{E}X\|_\infty\big\|_{\psi_2} < \infty$ (in particular, all the moments of $\|X - \mathbb{E}X\|_\infty$ are finite). It what follows, we will denote

$$S(\mathbb{T}) := S(\mathbb{T}, d_X) = \inf_{t\in\mathbb{T}} \sqrt{\operatorname{Var}(X(t))} + L\gamma_2(\mathbb{T}; d_X).$$

Note that Theorem A.2 implies that there exists a numerical constant $L > 0$ such that

$$\mathbb{E}\sup_{t\in\mathbb{T}}|X(t)-\mathbb{E}X(t)| \leqslant S(\mathbb{T}). \tag{1.5}$$

We will also need the following assumptions on the regression function f_* and the noise ξ :

Assumption 1.3. — Suppose that $f_*(X)-\mathbb{E}f_*(X)\in\mathscr{L}$ and $\xi\in\mathscr{L}$.

Since $\mathbb{E}f_*(X)=\mathbb{E}Y$, this assumption also implies that $Y-\mathbb{E}Y\in\mathscr{L}$. Note that if $\{X(t),\ t\in\mathbb{T}\}\cup\{Y\}$ is a family of centered Gaussian random variables and $\mathscr{L}$ is its closed linear span, then $\mathscr{L}$ is a subgaussian space and $f_*(X)$ is the orthogonal projection of Y onto the subspace $\mathscr{L}_X$. Thus, $f_*(X)\in\mathscr{L}_X\subset\mathscr{L}$.

Acknowledgements. — We want to thank the anonymous Referees for carefully reading the paper and for providing constructive feedback that helped us to improve the quality of results and presentation.

The authors are very thankful to Mikhail Lifshits and Mauro Maggioni for insightful discussions and their valuable input.

2. Approximation error bounds, alignment coefficient and Sobolev norms

Recall that P is the joint distribution of (X,Y) and let P_n be the empirical distribution based on the sample $(X_1,Y_1),\dots,(X_n,Y_n)$. The integrals with respect to P and P_n are denoted by

$$Pg:=\mathbb{E}g(X,Y),\quad P_ng:=\frac{1}{n}\sum_{i=1}^n g(X_i,Y_i).$$

In what follows, it will be convenient to denote $\ell(y,u):=(y-u)^2,\ y,u\in\mathbb{R}$ and

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

We also use the notation $\ell'(y,u)$ for the derivative of quadratic loss $\ell(y,u)$ with respect to u : $\ell'(y,u)=2(u-y)$. Throughout the paper, $\langle\cdot,\cdot\rangle$ denotes the bilinear form

$$\langle f,g\rangle:=\int_{\mathbb{T}} f(t)g(t)\mu(dt).$$

Let

$$F_n(\lambda,a):=P_n(\ell\bullet f_{\lambda,a})+\varepsilon\|\lambda\|_1,\quad F(\lambda,a):=P(\ell\bullet f_{\lambda,a})+\varepsilon\|\lambda\|_1.$$

Denote also

$$\overline{Y}_n:=n^{-1}\sum_{j=1}^n Y_j,\quad \overline{X}_n(t):=n^{-1}\sum_{j=1}^n X_j(t),\quad t\in\mathbb{T}.$$

Note that

$$\begin{aligned}\widehat{a}(\lambda)&:=\operatorname*{argmin}_{a\in\mathbb{R}}F_n(\lambda,a)=\overline{Y}_n-\langle\lambda,\overline{X}_n\rangle,\\ a(\lambda)&:=\operatorname*{argmin}_{a\in\mathbb{R}}F(\lambda,a)=\mathbb{E}Y-\langle\lambda,\mathbb{E}X\rangle.\end{aligned} \tag{2.1}$$

The following penalized empirical risk minimization problem

$$(\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon) := \operatorname*{argmin}_{\lambda\in\mathbb{D}, a\in\mathbb{R}} F_n(\lambda; a) \tag{2.2}$$

is exactly problem (1.2) written in a more concise form. Note that (2.2) is the empirical version of

$$(\lambda_\varepsilon, a_\varepsilon) := \operatorname*{argmin}_{\lambda\in\mathbb{D}, a\in\mathbb{R}} F(\lambda, a). \tag{2.3}$$

Due to convexity of the loss, both (2.3) and (2.2) are convex optimization problems. It will be shown (Theorem A.1 in the appendix) that, under certain assumptions, they admit (not necessarily unique) solutions λ_ε, $\widehat{\lambda}_\varepsilon$.

ASSUMPTION 2.1. — It is assumed throughout the paper that the solutions $(\lambda_\varepsilon, a_\varepsilon)$ of (2.3) and $(\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon)$ of (2.2) exist.

It might be also possible to study the problem under an assumption that $(\lambda_\varepsilon, a_\varepsilon)$ and $(\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon)$ are approximate solutions of the corresponding optimization problems, but we are not pursuing this to avoid further technicalities.

The goal of this section is to determine the parameters responsible for the size of the $L_2(\Pi)$ risk of $f_{\lambda_\varepsilon, a_\varepsilon}$, where $(\lambda_\varepsilon, a_\varepsilon)$ is the (distribution-dependent) solution of the problem (2.3), and to find upper bounds on these parameters in terms of classical Sobolev type norms. Later on, it will be shown that the same parameters affect the error rate of empirical solution $f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon}$.

Recall that $\mathbb{D} \subset L_1(\mu)$ is a convex subset that contains zero. It immediately follows from (2.3) that we can take $a_\varepsilon = a(\lambda_\varepsilon)$ and also that

$$\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)} \leqslant q(\varepsilon) := \inf_{\lambda\in\mathbb{D}, a\in\mathbb{R}} \left[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + \varepsilon\|\lambda\|_1\right]. \tag{2.4}$$

Clearly, q is a non-decreasing concave function (concavity follows from the fact that it is an infimum of linear functions). Therefore, $q(\varepsilon)/\varepsilon$ is a non-increasing function. Note also that $q(\varepsilon) \leqslant \|f_* - \Pi f_*\|^2_{L_2(\Pi)}$ (take $\lambda = 0$, $a = \mathbb{E}Y = \mathbb{E}f_*(X)$ in the expression under the infimum) and

$$q(\varepsilon) \leqslant \varepsilon\|\lambda_*\|_1$$

provided that $f_* = f_{\lambda_*, a_*}$, where $\lambda_* \in \mathbb{D}$, $a_* \in \mathbb{R}$ (take $\lambda = \lambda_*, a = a_*$). The infimum in the definition of $q(\varepsilon)$ is attained at $(\lambda_\varepsilon, a_\varepsilon)$ (a solution of problem (2.3) that is assumed to exist). Then, in addition to the bound $\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)} \leqslant q(\varepsilon)$, (2.3) also implies

$$\|\lambda_\varepsilon\|_1 \leqslant \frac{q(\varepsilon)}{\varepsilon}.$$

We will be interested, however, in other bounds on $\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)}$, in which the "regularization error" is proportional to ε^2 rather than to ε (as it is the case in the bounds for $q(\varepsilon)$). To this end, we have to introduce some new characteristics of the oracles $\lambda \in \mathbb{D}$.

Let $k(s,t) := \mathrm{Cov}(X(s), X(t))$, $s,t \in \mathbb{T}$ be the covariance function of the stochastic process X. Clearly, under Assumption 1.2, $\iint k^2(s,t)\mu(ds)\mu(dt) < \infty$ and the *covariance operator* $K : L_2(\mathbb{T},\mu) \mapsto L_2(\mathbb{T},\mu)$ defined by

$$(Kv)(s) := \int\limits_{\mathbb{T}} k(s,t)v(t)\mu(dt).$$

is Hilbert–Schmidt. For $u \in L_2(\mathbb{T})$, define

$$\|u\|_K := \sup_{\langle Kv,v\rangle \leqslant 1} \langle u, v\rangle . \tag{2.5}$$

Remark 2.1. — In the case when $\mathbb{T}$ is finite, operator K is represented by the Gram matrix of a finite dictionary and standard "restricted isometry" and "restricted eigenvalue" type constants are defined in terms of K and are involved in oracle inequalities for LASSO and other related estimators.

Note that $\langle Kv, v\rangle = \mathrm{Var}(f_v(X))$, where $f_v(X) := \int_{\mathbb{T}} v(t)X(t)\mu(dt)$. The set

$$\mathbb{H}(K) := \{u \in L_2(\mathbb{T}) : \ \|u\|_K < \infty\}$$

is a reproducing kernel Hilbert space of the covariance kernel k.

We will need the following description of the subdifferential of the convex function $\|\cdot\|_1$:

$$\partial\|\lambda\|_1 = \{w : \mathbb{T} \mapsto [-1,1] : \ \mu - \text{a.s. } w(t) = \mathrm{sign}(\lambda(t)) \text{ whenever } \lambda(t) \neq 0\}. \tag{2.6}$$

It follows from the general description of the subdifferential of a norm $\|\cdot\|$ in a Banach space $\mathfrak{X}$:

$$\partial\|x\| = \begin{cases} \{x^* \in \mathfrak{X}^* : \ \|x^*\| = 1, \ x^*(x) = \|x\|\}, & x \neq 0, \\ \{x^* \in \mathfrak{X}^* : \ \|x^*\| \leqslant 1\}, & x = 0, \end{cases}$$

where $\mathfrak{X}^*$ is the dual space. For details on our specific example, see [20, §4.5.1].

Note that, in standard examples (such as $\mathbb{T} \subset \mathbb{R}^d$), the "canonical" version of the subgradient of $\|\lambda\|_1$, $w(t) = \mathrm{sign}(\lambda(t))$, $t \in \mathbb{T}$, lacks smoothness and RKHS-norms are often large or infinite for such a choice of w. It will be seen below that existence of smoother versions of the subgradient is important in such cases. Given a measurable $w : \mathbb{T} \mapsto [-1,1]$, let $\mathbb{T}_w = \{t \in \mathbb{T} : \ |w(t)| \geqslant 1/2\}$. For smooth w, $\mathbb{T}_w$ will play a role of support of λ. Given $b \in [0,\infty]$, define the cone $C_w^{(b)}$ by

$$C_w^{(b)} := \left\{ u \in L_1(\mu) : \int\limits_{\mathbb{T}\setminus\mathbb{T}_w} |u|d\mu \leqslant b\,\langle w, u\rangle \right\}. \tag{2.7}$$

Note that, for $w \in \partial\|\lambda\|_1$, we have $|w(t)| \leqslant 1$, $t \in \mathbb{T}$. Therefore, $u \in C_w^{(b)}$ implies that

$$\int\limits_{\mathbb{T}\setminus\mathbb{T}_w} |u|d\mu \leqslant b \int\limits_{\mathbb{T}_w} |u|d\mu.$$

Roughly, this means that, for functions $u \in C_w^{(b)}$, $\mathbb{T}_w$ is a "dominant set". Let

$$\mathfrak{a}^{(b)}(w) := \sup\left\{ \langle w, u\rangle : \ u \in C_w^{(b)}, \ \|f_u\|_{L_2(\Pi)} = 1 \right\}. \tag{2.8}$$

Such quantities were introduced in the framework of sparse recovery in [24], [26] and its size is closely related to the RIP and restricted eigenvalue-type conditions. In some sense, $\mathfrak{a}^{(b)}(w)$ characterizes the way in which the vector (function) w is "aligned" with eigenspaces of the covariance operator of the process X and, following [24], it will be called the *alignment coefficient*. Clearly, we always have the bound $\mathfrak{a}^{(b)}(w) \leqslant \|w\|_K$, however, it can be improved in several important cases, see Section 4.4. Note that $\mathfrak{a}^{(b)}(w)$ is a nondecreasing function of b. For $b = \infty$, we have $C_w^{(\infty)} = L_1(\mu)$. In this case, $\mathfrak{a}^{(\infty)}(w) = \|w\|_K$, so the alignment coefficient coincides with the RKHS-norm associated to the covariance function k. For $b = 0$, we have

$$C_w^{(0)} = \{u \in L_1(\mu) : u = 0 \text{ a.s. on } \mathbb{T} \setminus \mathbb{T}_w\},$$

so the cone $C_w^{(0)}$ coincides with the subspace of functions supported in $\mathbb{T}_w$. In this case, $\mathfrak{a}^{(0)}(w)$ is the RKHS-norm associated with restriction of the kernel k to $\mathbb{T}_w$.

In what follows, it will be convenient to take $b = 16$ and denote $\mathfrak{a}(w) = \mathfrak{a}^{(16)}(w)$ (although in the statement of Theorem 2.1 below a smaller value $b = 2$ could be used).

We will be interested in those oracles λ for which there exists a subgradient $w \in \partial\|\lambda\|_1$ such that $\mathfrak{a}(w)$ is not too large and $\mathbb{T}_w$ is a "small" subset of $\mathbb{T}$. Such functions provide a natural analogue of sparse vectors in finite-dimensional problems.

THEOREM 2.1. — *The following inequality holds:*

$$\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)} \leqslant \inf_{\substack{\lambda \in \mathbb{D}, w \in \partial\|\lambda\|_1 \\ a \in \mathbb{R}}} \left[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + \frac{1}{4}\varepsilon^2 \mathfrak{a}^2(w)\right]. \tag{2.9}$$

REMARK 2.2. — It easily follows from the proof of this theorem that for all $\lambda \in \mathbb{D}$, $w \in \partial\|\lambda\|_1$, $a \in \mathbb{R}$,

$$\int\limits_{\mathbb{T}\setminus\mathbb{T}_w} |\lambda_\varepsilon| d\mu \leqslant \frac{4}{\varepsilon}\left[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + \frac{1}{4}\varepsilon^2 \mathfrak{a}^2(w)\right].$$

The intuition behind these results is the following: if there exists an oracle (λ, w, a) with a small approximation error $\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)}$ (say, of the order $o(\varepsilon)$) and not very large alignment coefficient $\mathfrak{a}(w)$, then the risk $\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)}$ is also small and λ_ε is "almost" concentrated on the set $\mathbb{T}_w$.

As we show below, in some cases $\|\cdot\|_K$ and $\mathfrak{a}(\cdot)$ can be bounded in terms of Sobolev-type norms.

Since self-adjoint integral operator K with kernel k is Hilbert–Schmidt, it is compact, and the orthogonal complement to its kernel possesses an orthonormal system of eigenfunctions $\{f_j\}_{j=1}^\infty \subset L_2(\mathbb{T}, \mu)$ corresponding to positive eigenvalues ν_j. It is well-known that

$$\mathbb{H}(K) = \left\{w(\cdot) = \textstyle\sum_{j=1}^\infty w_j f_j(\cdot) : \ \|w\|_K^2 = \sum_{j=1}^\infty w_j^2/\nu_j < \infty\right\}. \tag{2.10}$$

However, one might want to find a more direct characterization of $\mathbb{H}(K)$. One way to proceed is to use the so-called *Factorization theorem*:

THEOREM 2.2 ([30], Theorem 4 in Section 9). — *Assume that there exists a Hilbert space $\mathbb{V}$ and an injective linear operator $L : \mathbb{V} \mapsto \ell_\infty(\mathbb{T})$ such that $K = LL^*$, where L^* is the adjoint of L. Then $\mathbb{H}(K) = L(\mathbb{V})$, and $\langle Lu_1, Lu_2\rangle_{\mathbb{H}(K)} = \langle u_1, u_2\rangle_{\mathbb{V}}$.*

The most obvious choice is $\mathbb{V} = \ker(K)^\perp$ and $L = K^{1/2}$, whence $\|w\|_K = \|K^{-1/2}w\|_{L_2(\mu)}$ which again gives (2.10). Other choices often lead to more insightful description. For example, if X is the standard Brownian motion on $[0,1]$, then one can check [30] that $\mathbb{V} = L_2[0,1]$ with the standard Lebesgue measure and $(Lx)(t) := \int_0^t x(s)ds$ satisfy the requirements. It immediately implies

COROLLARY 2.1. — *The reproducing kernel Hilbert space associated with the Brownian motion is defined by*

$$\mathbb{H}(K) = \left\{h \in L_2[0,1],\, h(0) = 0,\, \|h\|_K^2 := \int_0^1 \left(h'(s)\right)^2 ds < \infty\right\} \subset \mathbb{W}^{2,1}[0,1], \tag{2.11}$$

where

$$\mathbb{W}^{2,1}[0,1] = \Big\{h \in L_2[0,1],\ h \text{ is abs. continuous,} \\ \|h\|_{\mathbb{W}^{2,1}}^2 := \int_0^1 \left[h^2(s) + \left(h'(s)\right)^2\right] ds < \infty\Big\}$$

is the Sobolev space.

In particular, it means that $\mathfrak{a}(w) \leqslant \|w\|_{\mathbb{W}^{2,1}}$. Suppose now that $\mathbb{T} \subset \mathbb{R}^m$ is a bounded open subset and, for some $C > 0$ and $\beta > 0$,

$$\mathfrak{a}^2(w) \leqslant C\|w\|_{\mathbb{W}^{2,\beta}}^2. \tag{2.12}$$

Let $\lambda \in L_1(\mathbb{T}, \mu)$ be a "sparse" oracle such that $\mathrm{supp}(\lambda) := \bigcup_{j=1}^d \mathbb{T}_j$, where $\mathbb{T}_j$, $j = 1, \dots, d$, are disjoint sets. Moreover, assume that the distance between $\mathbb{T}_j$ and $\mathbb{T}_k$ is positive for all $j \neq k$. In other words, λ has d components with well separated supports and it is zero in between. In this case, one can find $w \in \partial\|\lambda\|_1$ such that $w = \sum_{j=1}^d w_j$ and w_j, $j = 1, \dots, d$, are smooth functions (from the space $\mathbb{W}^{2,\beta}$ to be specific) with disjoint supports. For any such function w, we have

$$\mathfrak{a}^2(w) \leqslant C\|w\|_{\mathbb{W}^{2,\beta}}^2 \leqslant C_1 \sum_{j=1}^d \|w_j\|_{\mathbb{W}^{2,\beta}}^2 \leqslant C_1 d \max_{1\leqslant j\leqslant d} \|w_j\|_{\mathbb{W}^{2,\beta}}^2$$

and the bound of Theorem 2.1 implies that

$$\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|_{L_2(\Pi)}^2 \leqslant \left\|f_{\lambda, a(\lambda)} - f_*\right\|_{L_2(\Pi)}^2 + \frac{C}{4}\, d \max_{1\leqslant j\leqslant d} \|w_j\|_{\mathbb{W}^{2,\beta}}^2 \varepsilon^2. \tag{2.13}$$

Thus, the size of the error explicitly depends on the number d of components of "sparse" oracles λ approximating the target.

In Section 4, we will show that bound (2.12) holds for a number of stochastic processes X and, moreover, there are other ways to take advantage of sparsity in the cases when the domain $\mathbb{T}$ of X can be partitioned in a number of regions $\mathbb{T}_j$, $j = 1, \dots, N$, such that the processes $\{X(t),\, t \in \mathbb{T}_j\}$ are "weakly correlated".

3. Basic oracle inequalities

In this section, we present general oracle inequalities for the L_2-risk of estimator $f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon}$. The main goal is to show that if there exists an oracle (λ, w, a), $\lambda \in \mathbb{D}$, $w \in \partial\|\lambda\|_1$, $a \in \mathbb{R}$ such that the approximation error $\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)}$ is small, the alignment coefficient $\mathfrak{a}(w)$ is not large and λ is "sparse" in the sense that the set of random variables $\{X(t) : t \in \mathbb{T}_w\}$ can be well approximated by a linear space $L \subset \mathscr{L}_X$ of small dimension, then the L_2-error $\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\|^2_{L_2(\Pi)}$ of the estimator $f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon}$ can be controlled in terms of the dimension of L and the alignment coefficient $\mathfrak{a}(w)$. To state the result precisely, we have to introduce one more parameter, an "approximate dimension", providing an optimal choice of approximating space L. Thus, the degree of "sparsity" of the oracle will be characterized by the alignment coefficient that already appeared in approximation error bounds of Section 2 and also by "approximate dimension" $d(w, \lambda)$ introduced below.

We start, however, with a "slow-rate" oracle inequality that does not depend on "sparsity". The inequalities of this type are well known in the literature on sparse recovery, in particular, for LASSO estimator in the case of finite dictionaries, see [4], [31].

Recall that $\mathbb{D} \subseteq L_1(\mu)$ is a convex set and $0 \in \mathbb{D}$. Recall also the definition of $q(\varepsilon)$ (see 2.4) and its properties. Note that

$$\sigma_Y^2 = \mathrm{Var}(f_*(X)) + \sigma_\xi^2 = \|f_* - \Pi f_*\|^2_{L_2(\Pi)} + \sigma_\xi^2. \tag{3.1}$$

Theorem 3.1. — *There exist absolute constants $\mathfrak{C}, \mathfrak{c}$ and D such that the following holds. For any $s \geqslant 1$ with $\overline{s} := s + 3\log(\log_2 n + 2) + 3 \leqslant \mathfrak{c}\sqrt{n}/\log n$ and for all ε satisfying*

$$\varepsilon \geqslant D\,\frac{\sigma_Y S(\mathbb{T})}{\sqrt{n}}, \tag{3.2}$$

with probability at least $1 - e^{-s}$

$$\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\|^2_{L_2(\Pi)} + \frac{3}{4}\varepsilon\|\widehat{\lambda}_\varepsilon\|_1 \leqslant \inf_{\lambda\in\mathbb{D},a\in\mathbb{R}}\left[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + \frac{3}{2}\,\varepsilon\|\lambda\|_1\right] + \mathfrak{C}\,\frac{\sigma_Y^2\,\overline{s}}{n}. \tag{3.3}$$

As was mentioned earlier, our main goal is to obtain sharper bounds which would demonstrate connections between the risk of $f_{\widehat{\lambda}_\varepsilon,\widehat{a}}$ and the degree of sparsity of an underlying model. Our next result is a step in this direction. We will need the notion of Kolmogorov's d-width of the set of random variables $C \subset \mathscr{L}_X$ defined as follows:

$$\rho_d(C) := \inf_{\substack{L\subset\mathscr{L}_X\\ \dim(L)\leqslant d}} \sup_{\eta\in C} \|P_{L^\perp}\eta\|_{L_2(\mathbb{P})}.$$

It characterizes the optimal accuracy of approximation of the set C by d-dimensional linear subspaces of $\mathscr{L}_X$. Given $\mathbb{T}' \subset \mathbb{T}$, let

$$X_{\mathbb{T}'} := \{X(t) - \mathbb{E}X(t) : t \in \mathbb{T}'\}.$$

Recall that $\mathbb{T}_w := \{t \in \mathbb{T} : |w(t)| \geqslant 1/2\}$. Given an oracle $\lambda \in \mathbb{D}$ and $w \in \partial\|\lambda\|_1$, let

$$\rho_d(w) := \rho_d\left(X_{\mathbb{T}_w}\right).$$

The following number will play a role of approximate dimension of the set of random variables $X_{\mathbb{T}_w}$:

$$d(w,\lambda) := \min\Big\{ d \geqslant 0 : \frac{d\sigma_Y^2}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\rho_d(w)\right)}{\sqrt{n}} \Big\}. \tag{3.4}$$

Theorem 3.2. — *There exist absolute constants $\mathfrak{C}, \mathfrak{c}$ and D such that the following holds. For any $s \geqslant 1$ with $\overline{s} := s + 3\log(\log_2 n + 2) + 3 \leqslant \mathfrak{c}\sqrt{n}/\log n$ and for all ε satisfying*

$$\varepsilon \geqslant D\,\frac{\sigma_Y S(\mathbb{T})\sqrt{s}}{\sqrt{n}}, \tag{3.5}$$

with probability at least $1 - e^{-s}$

$$\begin{aligned} &\big\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \\ &\leqslant \inf_{\substack{\lambda\in\mathbb{D}\\ w\in\partial\|\lambda\|_1\\ a\in\mathbb{R}}} \Big[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + 2\varepsilon^2\mathfrak{a}^2(w) + \mathfrak{C}\frac{\sigma_Y^2\, d(w,\lambda)}{n} + \mathfrak{C}\frac{\|\lambda\|_1^2 S^2(\mathbb{T})}{n}\Big] + \mathfrak{C}\frac{\sigma_Y^2\,\overline{s}}{n}. \end{aligned} \tag{3.6}$$

Under an additional assumption that $\|\lambda\|_1$ is not too large, it is possible to prove the following modified version of Theorem 3.2 without the term $\mathfrak{C}\|\lambda\|_1^2 S^2(\mathbb{T})/n$ in the oracle inequality.

Theorem 3.3. — *Assume that conditions of Theorem* 3.2 *hold. If $\mathbb{D}$ is such that*

$$\mathbb{D} \subset \Big\{ \lambda \in L_1(\mu) :\ \|\lambda\|_1 \leqslant \frac{\mathfrak{c}\sigma_Y\sqrt{n}}{S(\mathbb{T})} \Big\}$$

for some absolute constant $\mathfrak{c} > 0$, then with probability $\geqslant 1 - e^{-s}$

$$\begin{aligned} &\big\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \\ &\qquad\leqslant \inf_{\substack{\lambda\in\mathbb{D}\\ w\in\partial\|\lambda\|_1\\ a\in\mathbb{R}}} \Big[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + 2\varepsilon^2\mathfrak{a}^2(w) + \mathfrak{C}\frac{\sigma_Y^2\, d(w,\lambda)}{n}\Big] + \mathfrak{C}\frac{\sigma_Y^2\,\overline{s}}{n}. \end{aligned} \tag{3.7}$$

The proof of this result follows from the proof of Theorem 3.2, see remark 7.1 for more details.

Remark 3.1. — Note that the oracle inequality of Theorem 3.2 is *sharp*, meaning that the constant in front of $\|f_{\lambda,a} - f_*\|_{L_2(\Pi)}$ (the leading constant) is 1. It is possible to derive an oracle inequality with the leading constant larger than 1 which might yield faster rates when the variance of the noise σ_ξ^2 is small. Define the following version of the "approximate dimension" (compare to (3.4)):

$$d_{\sigma_\xi}(w,\lambda) := \min\Big\{ d \geqslant 0 : \frac{d\sigma_\xi^2}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\rho_d(w)\right)}{\sqrt{n}} \Big\}.$$

Then, under the assumptions of Theorem 3.2, the following inequality holds with probability $\geqslant 1 - e^{-s}$:

$$\begin{aligned}&\left\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\right\|^2_{L_2(\Pi)}\\ &\leqslant \inf_{\substack{\lambda\in\mathbb{D}\\ w\in\partial\|\lambda\|_1\\ a\in\mathbb{R}}}\left[2\left\|f_{\lambda,a} - f_*\right\|^2_{L_2(\Pi)} + 2\varepsilon^2\mathfrak{a}^2(w) + \mathfrak{C}\frac{\sigma_\xi^2 d_{\sigma_\xi}(w,\lambda)}{n} + \mathfrak{C}\frac{\|\lambda\|_1^2 S^2(\mathbb{T})}{n}\right] + \mathfrak{C}\frac{\sigma_Y^2\overline{s}}{n}.\end{aligned}$$

The proof of this result uses arguments similar to the proof of Theorem 3.2, so we omit the details.

Inequality (3.6) above depends on rather abstract parameters (such as the alignment coefficient and the approximate dimension) that have to be further bounded before one can get a meaningful bound in concrete examples. This will be discussed in some detail in the following sections.

4. Bounding the alignment coefficient

First, we discuss the bounds on the alignment coefficient in terms of Sobolev-type norms in some detail. After this, we turn to the problem of bounding the alignment coefficient in the cases when there exists a weakly correlated partition for the design process X.

4.1. Sacks-Ylvisaker conditions. — In the univariate case $\mathbb{T} = [0,1]$, it is possible to determine whether (a certain subspace of) the Sobolev space can be continuously embedded into $\mathbb{H}(K)$ based on the smoothness of the covariance function $k(\cdot\,,\cdot)$. Existence of such an embedding is given by the so-called *Sacks-Ylvisaker* conditions [38]. This provides a way to bound the RKHS norm $\|\cdot\|_K$ generated by the covariance function of X (and, thus, also the alignment coefficient) in terms of a Sobolev norm. Definitions and statements below are taken from [37], Section 3.

Set $\Omega_+ := \big\{(s,t)\in(0,1)^2 :\ s > t\big\}$ and $\Omega_- := \big\{(s,t)\in(0,1)^2 :\ s < t\big\}$. Let G be a continuous function on $\Omega_+\cup\Omega_-$ such that the restrictions $G|_{\Omega_j}$ are continuously extendable to the closures $\mathrm{cl}(\Omega_j)$, $j\in\{+,-\}$. G_j will stand for the extension of G to $[0,1]^2$ which is continuous on $\mathrm{cl}(\Omega_j)$ and on $[0,1]^2\setminus\mathrm{cl}(\Omega_j)$. Set $R^{(k,l)}(s,t) = \frac{\partial^{k+l}}{\partial s^k\partial t^l}R(s,t)$. Then, the covariance kernel $k(\cdot\,,\cdot)$ defined on $[0,1]^2$ satisfies the Sacks-Ylvisaker conditions of order $r\in\mathbb{N}$ if the following holds true:

(A) $k\in C^{r,r}([0,1]^2)$, the partial derivatives of $G = k^{(r,r)}$ up to order 2 are continuous on $\Omega_+\cup\Omega_-$ and are continuously extendable to $\mathrm{cl}(\Omega_+)$ and to $\mathrm{cl}(\Omega_-)$.

(B) $\min_{0\leqslant t\leqslant 1}\big(G_-^{(1,0)}(t,t) - G_+^{(1,0)}(t,t)\big) > 0$.

(C) $G_+^{(2,0)}(t,\cdot)$ belongs to the RKHS with reproducing kernel G and

$$\sup_{t\in[0,1]}\|G_+^{(2,0)}(t,\cdot)\|_G < \infty.$$

(D) In the case $r\geqslant 1$, $k^{(0,j)}(\cdot,0) = 0$ for $0\leqslant j\leqslant r-1$.

Let $\mathbb{W}_0^{2,r+1}$ be the subspace of $\mathbb{W}^{2,r+1}$ defined by

$$\mathbb{W}_0^{2,r+1} = \big\{ f \in \mathbb{W}^{2,r+1} : \ f^{(j)}(0) = f^{(j)}(1) = 0 \text{ for } 0 \leqslant j \leqslant r \big\}.$$

THEOREM 4.1 (Corollary 1 in [37]). — *Assume $k(\cdot\,,\cdot)$ satisfies the Sacks-Ylvisaker conditions of order r. Then $\mathbb{W}_0^{2,r+1} \subset \mathbb{H}(K)$ and the embedding $\mathbb{W}_0^{2,r+1} \hookrightarrow \mathbb{H}(K)$ is continuous.*

As a result, we have the bound $\|w\|_K \leqslant C\|w\|_{\mathbb{W}^{2,r+1}}$ that holds for all w with some constant $C > 0$.

It is well-known that the covariance function $k_1(s,t) = s \wedge t$ of the Brownian motion and $k_2(s,t) = e^{-|s-t|}$ of the Ornstein-Uhlenbeck process satisfy Sacks-Ylvisaker conditions of order $r = 0$.

COROLLARY 4.1. — *Let $X(t)$, $t \in [0,1]$ be the Ornstein-Uhlenbeck process and let $\mathbb{H}(K)$ be the associated reproducing kernel Hilbert space. If $w \in \mathbb{W}^{2,1}[0,1]$ is such that $w(0) = w(1) = 0$, then $w \in \mathbb{H}(K)$ and*

$$\|w\|_K \leqslant C\|w\|_{\mathbb{W}^{2,1}[0,1]}.$$

This should be compared to the exact description of $\mathbb{H}(K)$, the kernel of the Ornstein-Uhlenbeck process, which is known to be

$$\mathbb{H}(K) = \Big\{ w \in L_2[0,1] : \ \|w\|_K^2 = \frac{w^2(0) + w^2(1)}{2} + \frac{1}{4}\int_0^1 w^2(t)dt + \int_0^1 \big(w'(t)\big)^2\, dt < \infty \Big\}.$$

4.2. DISCRETE SOBOLEV NORMS AND THE BROWNIAN MOTION. — In this example, we look back at the case when the design process is a Brownian motion (it was already discussed in Section 2). However, this time we make the more realistic assumption that the design processes are observed only at discrete points.

Assume that $\{X(t),\ t \in [0,1]\}$ is a standard Brownian motion released at zero, that is, $X(t) = Z + W(t)$, where Z is a standard normal random variable independent of W. Suppose that we observe n iid copies of X, $X_1, \dots, X_n$ on a grid $\mathbb{T} = \mathscr{G}_N = \{0 \leqslant t_1 < \cdots < t_N \leqslant 1\}$. Let μ be a counting measure on $\mathbb{T}$. If, for example, the grid is uniform with mesh size $1/N$ for some large N, with high probability the adjacent columns of the design matrix $\big(X_i(t_j)\big)_{i\leqslant n, j\leqslant N}$ will be almost collinear. To the best of our knowledge, a direct analysis based on the restricted eigenvalue type conditions [6] provides unsatisfactory bounds in such cases. On the other hand, results that hold true without any assumptions on the design (e.g., [27], first statement of Theorem 1) only guarantee "slow" rates of convergence (of order $n^{-1/2}$, where n is the size of a training data set).

The covariance function $k(\cdot\,,\cdot)$ of X satisfies $k(t_i,t_j) = 1 + t_i \wedge t_j$. Let $K = \big(1 + t_i \wedge t_j\big)_{i,j=1}^N$ be the associated Gram matrix and let $K = LL^T$ be its Cholesky

factorization. Note that

$$L = \begin{pmatrix} \sqrt{1+t_1} & 0 & \dots & 0 \\ \sqrt{1+t_1} & \sqrt{t_2-t_1} & 0 & \vdots \\ \vdots & & \ddots & 0 \\ \sqrt{1+t_1} & \sqrt{t_2-t_1} & \dots & \sqrt{t_N-t_{N-1}} \end{pmatrix}$$

By the Factorization theorem (or a straightforward argument), for any $w \in \mathbb{R}^N$, $\|w\|_K = \left\|L^{-1}w\right\|_2$. If $w = Lv$, then a direct computation shows

$$\|v\|_2^2 = \|w\|_K^2 = \frac{w_1^2}{1+t_1} + \sum_{j=2}^{N} \frac{(w_j - w_{j-1})^2}{t_j - t_{j-1}}. \tag{4.1}$$

The latter expression can be seen as a discrete analogue of the Sobolev norm. For example, let the grid $\mathscr{G}_N$ be uniform, that is, $t_j = (j-1)/N$, $j = 1 \dots N$, and let $\lambda : \mathscr{G}_N \mapsto \mathbb{R}$ be sparse in the following sense: $\operatorname{supp}(\lambda) = \{t_{i_1} < t_{i_2} < \dots < t_{i_s}\}$ so that $|\operatorname{supp}(\lambda)| = s$ and

$$\min_{2 \leqslant k \leqslant s} |t_{i_k} - t_{i_{k-1}}| = \sigma \gg \frac{1}{N}.$$

It is clear from (4.1) that $\inf_{w \in \partial \|\lambda\|_1} \|w\|_K \leqslant C\sqrt{s/\sigma}$ for some absolute constant $C > 0$ (e.g., take a vector whose entries linearly interpolate the sign pattern of λ) while the trivial choice $w(t_j) = \operatorname{sign}(\lambda(t_j))$ leads to $\|w\|_K \geqslant c\sqrt{Ns}$.

Note also that if $w_j := w(t_j)$, $j = 1, \dots, N$, for a smooth function $w \in \mathbb{W}^{2,1}([0,1])$ (with a slight abuse of notation, we write w both for the vector in $\mathbb{R}^N$ and for the function), then, by Cauchy-Schwarz inequality,

$$\frac{(w_j - w_{j-1})^2}{t_j - t_{j-1}} = \frac{(w(t_j) - w(t_{j-1}))^2}{t_j - t_{j-1}} = \frac{\left(\int_{t_{j-1}}^{t_j} w'(s)ds\right)^2}{t_j - t_{j-1}} \leqslant \int_{t_{j-1}}^{t_j} |w'(s)|^2 ds.$$

It immediately implies that $\|w\|_K^2 \leqslant |w(0)|^2 + \int_0^1 |w'(s)|^2 ds$, so the discrete Sobolev norm needed to control the alignment coefficient is bounded from above by its continuous counterpart. As a matter of fact, we have that

$$\|w\|_K^2 \leqslant \inf_{\widetilde{w}} \left[|\widetilde{w}(0)|^2 + \int_0^1 |\widetilde{w}'(s)|^2 ds \right],$$

where the infimum is taken over all functions $\widetilde{w} \in \mathbb{W}^{2,1}([0,1])$ such that $\widetilde{w}(t_j) = w_j$, $j = 1, \dots, N$.

These observations allow one to characterize the prediction performance of the LASSO estimator in terms of s and σ, in particular, rates faster than $n^{-1/2}$ can be deduced from Theorem 3.2.

4.3. Stationary processes. — In this subsection, we derive Sobolev norm bounds on the alignment coefficient in the case when X is a stationary process (or a stationary random field).

Let $\mathbb{T} \subset \mathbb{R}^d$ be a bounded open set and let μ be the Lebesgue measure. Consider a stationary random field $\{X(t),\ t \in \mathbb{R}^d\}$ with continuos covariance function k :

$$k(t-s) = \mathrm{Cov}(X(t), X(s)), \quad t, s \in \mathbb{R}^d.$$

By Bochner's theorem, there exists a finite Borel measure ν such that

$$k(t) = \int_{\mathbb{R}^d} e^{i\langle t,u\rangle} \nu(du), \quad t \in \mathbb{R}^d \tag{4.2}$$

called the *spectral measure* of X. In what follows, we assume that ν is absolutely continuous with spectral density $v : \mathbb{R}^d \mapsto \mathbb{R}_+$.

Proposition 4.1. — *Suppose that*

$$v(t) \geqslant \frac{c}{(1+|t|^2)^p}, \quad t \in \mathbb{R}^d \tag{4.3}$$

for some $p > d/2$ and $c > 0$. For w defined on $\mathbb{T}$, let

$$\Omega(w) := \left\{\widetilde{w} : \mathbb{R}^d \mapsto \mathbb{R} :\ \widetilde{w}(t) = w(t),\ t \in \mathbb{T}\right\}.$$

Then

$$\|w\|_K \leqslant C \inf_{\widetilde{w} \in \Omega(w)} \|\widetilde{w}\|_{\mathbb{W}^{2,p}(\mathbb{R}^d)}.$$

Note that condition (4.3) could not hold for $p \leqslant d/2$ since this would contradict integrability of the spectral density v.

Proof. — Given $u \in L_1(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$, let $\widehat{u}$ be its Fourier transform. Observe that

$$\mathrm{Var}(f_u(X)) = \iint k(t-s)u(t)u(s)dtds = \int |\widehat{u}(z)|^2 v(z)dz.$$

For u supported in $\mathbb{T}$ and $\widetilde{w} \in \Omega(w)$, this gives

$$\begin{aligned}\langle w, u\rangle_{L_2(\mathbb{T},\mu)} = \langle \widetilde{w}, u\rangle_{L_2(\mathbb{R}^d)} = \left\langle \widehat{\widetilde{w}}, \widehat{u}\right\rangle_{L_2(\mathbb{R}^d)} &= \left\langle \frac{\widehat{\widetilde{w}}}{\sqrt{v}}, \widehat{u}\sqrt{v}\right\rangle \\ &\leqslant C\left\|(1+|x|^2)^{p/2}\widehat{\widetilde{w}}\right\|_{L_2(\mathbb{R}^d)} \mathrm{Var}(f_u(X)),\end{aligned}$$

hence $\|w\|_K \leqslant C\left\|(1+|x|^2)^{p/2}\widehat{\widetilde{w}}\right\|_{L_2(\mathbb{R}^d)}$. It remains to note that by the properties of Fourier transform

$$\left\|(1+|x|^2)^{p/2}\widehat{\widetilde{w}}\right\|_{L_2(\mathbb{R}^d)} \leqslant C\|\widetilde{w}\|_{\mathbb{W}^{2,p}(\mathbb{R}^d)}. \qquad \square$$

We now turn to the case of stationary processes observed at discrete points. Let $\{X(t),\ t \in \mathbb{R}^d\}$ be a (weakly) stationary random field, and let $X_1, \dots, X_n$ be i.i.d. copies of X observed on the grid $\mathbb{T} = \mathscr{G}_N = \{t_j = 2\pi j/N,\ j \in \{1, \dots, N\}^d\}$ for some even N. In this case, functions on $\mathbb{T}$ can be identified with vectors in $\mathbb{R}^{N^d}$. We also assume that μ is the counting measure on $\mathbb{T}$.

Assumption 4.1. — Suppose the following condition on the spectral density v of the process X holds:

$$c_1\Big(\frac{1}{1+|t|^2}\Big)^p \leqslant v(t) \leqslant c_2\Big(\frac{1}{1+|t|^2}\Big)^p \quad \text{for some } p > \frac{d}{2}, \tag{4.4}$$

where $0 < c_1 \leqslant c_2 < \infty$.

Proposition 4.2. — *Given* $\vec{w} = (w_1, \dots, w_{N^d})^T \in \partial\|\lambda\|_1$, *let*

$$\Omega_N(\vec{w}) = \big\{w \in \mathbb{W}^{2,p}(\mathbb{R}^d) : \ w(2\pi j/N) = w_j, \ j \in \mathbb{Z}^d\big\},$$

where w_j *are defined arbitrarily for* $j \notin \{1, \dots, N\}^d$. *Under the above-stated assumptions,*

$$\|\vec{w}\|_K \leqslant C \inf_{w \in \Omega_N(\vec{w})} \|w\|_{\mathbb{W}^{2,p}(\mathbb{R}^d)}.$$

The proof is outlined in section 7.5. Implications of this result for the risk of $\widehat{\lambda}_\varepsilon$ are presented in Theorem 6.3 below. In particular, we show that rates faster than $n^{-1/2}$ are often possible.

4.4. Sparse multiple linear models and weakly correlated partitions. — In this section, we assume that

$$Y = a + \sum_{j=1}^{N} \int_{\mathbb{T}_j} X^{(j)}(t_j) d\Lambda_j(t_j) + \xi, \tag{4.5}$$

where $a \in \mathbb{R}$, $\mathbb{T}_1, \dots, \mathbb{T}_N$ are measurable spaces equipped with σ-algebras $\mathscr{B}_1, \dots, \mathscr{B}_N$ and finite measures $\mu_1, \dots, \mu_N$, $X^{(1)}, \dots, X^{(N)}$ are subgaussian stochastic processes on $\mathbb{T}_1, \dots, \mathbb{T}_N$, $\Lambda_1, \dots, \Lambda_N$ are signed measures on spaces $\mathbb{T}_1, \dots, \mathbb{T}_N$ with bounded total variations, and ξ is a zero-mean random variable independent of $X^{(1)}, \dots, X^{(N)}$. Suppose $\mathscr{B}_j = \mathscr{B}_{\mathbb{T}_j}$ (Borel σ-algebra in the semimetric space $(\mathbb{T}_j, d_{X^{(j)}})$). Without loss of generality, we can assume that the sets $\mathbb{T}_1, \dots, \mathbb{T}_N$ form a partition of the space $\mathbb{T} := \bigcup_{j=1}^N \mathbb{T}_j$ equipped with a σ-algebra $\mathscr{B}$ and a measure μ such that the measures μ_j are restrictions of μ on $\mathbb{T}_j$. Similarly, signed measures Λ_j become restrictions on $\mathbb{T}_j$ of a signed measure Λ_* on $(T, \mathscr{B})$. We will set $X(t) := X^{(j)}(t)$, $t \in \mathbb{T}_j$, $j = 1, \dots, N$, and, finally, we can assume that $\mathscr{B} = \mathscr{B}_{\mathbb{T}}$ is the Borel σ-algebra in the semimetric space $(\mathbb{T}, d_X)$.

We are interested in the situation when the processes $\big\{X^{(j)}(t), t \in \mathbb{T}_j\big\}$, $j = 1, \dots, N$, are *weakly correlated* (in particular, they can be independent). The number of predictors N can be very large, but Y might depend only on $X^{(j)}(t)$, $t \in \mathbb{T}_j$, $j \in J \subset \{1, \dots, N\}$, where $\mathrm{Card}(J) \ll N$, whence $\mathrm{Card}(J)$ naturally represents the degree of sparsity of the problem. Another interpretation of the model is to assume that the domain $\mathbb{T}$ of the stochastic process X can be partitioned in disjoint sets $\mathbb{T}_j$ so that $\{X(t) : t \in \mathbb{T}_j\}$, $j = 1, \dots, N$, are "weakly correlated", but only few of the elements of partition are correlated with the response variable Y. It is important to emphasize that the results of the following sections concerning the estimator (1.2) are *adaptive* with respect to the partitions, in particular, we do not need to know the

"weakly correlated" parts in advance, but the estimator adapts to such a structure (given that it exists).

Let K_j be the covariance operator of $X^{(j)}$ and k_j its kernel (the covariance function of $X^{(j)}$). Our next goal is to understand how to control the alignment coefficient $\mathfrak{a}(\cdot)$ associated with the process X in terms of the RKHS-norms $\|\cdot\|_{K_j},\ j=1,\dots,N$.

Without loss of generality, assume that $X^{(j)},\ j=1,\dots,N$, are centered. Given $u\in L_1(\mathbb{T},\mu)$, it can be represented as $u=\sum_{j=1}^N u_j$ with $\operatorname{supp}(u_j)\subseteq\mathbb{T}_j$. Given $\gamma>0$, define

$$C_{\gamma,J}:=\Big\{u\in L_1(\mathbb{T},\mu):\textstyle\sum_{j\notin J}\|f_{u_j}\|_{L_2(\Pi)}\leqslant\gamma\sum_{j\in J}\|f_{u_j}\|_{L_2(\Pi)}\Big\}$$

and
$$\beta_2^{(\gamma)}(J):=\inf\Big\{\beta>0:\textstyle\sum_{j\in J}\|f_{u_j}\|^2_{L_2(\Pi)}\leqslant\beta^2\|f_u\|^2_{L_2(\Pi)},\ u\in C_{\gamma,J}\Big\}.$$

Clearly, if $X^{(j)}(t),\ t\in\mathbb{T},\ j=1\dots N$ are uncorrelated, then $\beta_2^{(\gamma)}(J)=1$ for any nonempty $J\subseteq\{1,\dots,N\}$. More generally, we have the following result:

Proposition 4.3. — *For all $J\subset\{1,\dots,N\}$ and all $w=\sum_{j\in J}w_j$ such that $\operatorname{supp}(w_j)\subseteq\mathbb{T}_j$ and $\|w_j\|_{K_j}<\infty$, we have*

$$\mathfrak{a}^{(b)}(w)\leqslant\beta_2^{(\gamma)}(J)\Big(\sum_{j\in J}\|w_j\|^2_{K_j}\Big)^{1/2},\tag{4.6}$$

where

$$\gamma=b\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\max_{j\in J}\|w_j\|_{K_j}.$$

Proof. — Note that since $w=\sum_{j\in J}w_j$ with $\operatorname{supp}(w_j)\subseteq\mathbb{T}_j$,

$$\mathbb{T}_w\subset\bigcup_{j\in J}\mathbb{T}_j\quad\text{and}\quad\mathbb{T}\setminus\mathbb{T}_w\supset\bigcup_{j\notin J}\mathbb{T}_j.$$

For all $u\in C_w^{(b)}$ (defined in (2.7)), we have

$$\begin{aligned}\sum_{j\notin J}\|f_{u_j}\|_{L_2(\Pi)}&\leqslant\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\sum_{j\notin J}\|u_j\|_1\\&\leqslant\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\int\limits_{\mathbb{T}\setminus\mathbb{T}_w}|u|d\mu\leqslant b\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\langle w,u\rangle.\end{aligned}$$

Since

$$\langle w,u\rangle=\sum_{j\in J}\langle w_j,u_j\rangle\leqslant\sum_{j\in J}\|w_j\|_{K_j}\big\|f_{u_j}\big\|_{L_2(\Pi)}\leqslant\max_{j\in J}\|w_j\|_{K_j}\sum_{j\in J}\big\|f_{u_j}\big\|_{L_2(\Pi)},$$

we can conclude that

$$\sum_{j\notin J}\big\|f_{u_j}\big\|_{L_2(\Pi)}\leqslant b\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\max_{j\in J}\|w_j\|_{K_j}\sum_{j\in J}\big\|f_{u_j}\big\|_{L_2(\Pi)}.$$

We proved that $C_w^{(b)} \subseteq C_{\gamma,J}$ for $\gamma := b\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\max_{j\in J}\|w_j\|_{K_j}$. For all $u \in C_w^{(b)} \subseteq C_{\gamma,J}$, we have

$$\langle w,u\rangle = \sum_{j\in J}\langle w_j,u_j\rangle \leqslant \Big(\sum_{j\in J}\|w_j\|_{K_j}^2\Big)^{1/2}\Big(\sum_{j\in J}\|f_{u_j}\|_{L_2(\Pi)}\Big)^{1/2}$$
$$\leqslant \beta_2^{(\gamma)}(J)\Big(\sum_{j\in J}\|w_j\|_{K_j}^2\Big)^{1/2}\|f_u\|_{L_2(\Pi)},$$

implying that

$$\mathfrak{a}^{(b)}(w) \leqslant \beta_2^{(\gamma)}(J)\Big(\sum_{j\in J}\|w_j\|_{K_j}^2\Big)^{1/2}$$

with $\gamma := b\max_{1\leqslant j\leqslant N}\|k_j\|_\infty^{1/2}\max_{j\in J}\|w_j\|_{K_j}$. □

Next, we will relate $\beta_2^{(\gamma)}(J)$ to the size of *restricted isometry* [13] constants associated with partition $\mathbb{T}_1,\dots,\mathbb{T}_N$. Given an integer $d \geqslant 1$, we define the restricted isometry constant δ_d as the smallest $\delta > 0$ with the following property: for any $J \subset \{1,\dots,N\}$ with $\mathrm{Card}(J) = d$, any u_j, $j \in J$ such that $\mathrm{supp}(u_j) \subseteq \mathbb{T}_j$ and $\mathrm{Var}(f_{u_j}(X)) = 1$, the spectrum of the $d\times d$ matrix $\big(\mathrm{Cov}(f_{u_i}(X),f_{u_j}(X))\big)_{i,j\in J}$ belongs to $[1-\delta,1+\delta]$.

Proposition 4.4. — *The following inequality holds for all $J \subset \{1,\dots,N\}$ with $\mathrm{Card}(J) \leqslant d$:*

$$\beta_2^{(\gamma)}(J) \leqslant \frac{1+\delta_{2d}}{(1-\delta_{2d})^2-\gamma\delta_{3d}}.$$

In particular, it means that $\beta_2^{(\gamma)}$ can be bounded by a constant as soon as $\delta_{3d} < 1/(2+\gamma)$.

Proof. — The argument is similar to Lemma 7.2 in [26], the details are included in Appendix A.6 for the reader's convenience. □

5. Oracle inequalities and weakly correlated partitions

First, we will state a corollary of Theorem 3.2 concerning the model of weakly correlated partitions discussed in Section 4. Let $\Delta := \{\mathbb{T}_1,\dots,\mathbb{T}_N\}$ be a partition of the parameter space $\mathbb{T}$ into $N \geqslant 1$ measurable disjoint sets. Let $\mathscr{T}$ be the set of all such partitions. Let $X^{(j)}$ denote the restriction of stochastic process X to the set $\mathbb{T}_j$ and let K_j be the covariance operator of the process X_j and k_j be its covariance function. Consider an oracle $\lambda \in L_1(\mu)$ and denote

$$J_\lambda := \{j = 1,\dots,N : \mathbb{T}_j \cap \mathrm{supp}(\lambda) \neq \varnothing\}.$$

Also, denote $N(\lambda) := \mathrm{Card}(J_\lambda)$. Usually, we assume that N is very large and $N(\lambda)$ is much smaller than N, so, $N(\lambda)$ plays the role of "sparsity parameter" in this framework. Let $w = \sum_{j\in J_\lambda} w_j \in \partial\|\lambda\|_1$ be a subgradient such that $\mathrm{supp}(w_j) \subset \mathbb{T}_j$,

$j \in J_\lambda$. In what follows, denote $\mathscr{W}_{\lambda,\Delta}$ the set of all such subgradients w. Recall the definition of the quantity $\beta_2^{(\gamma)}(J)$ (Section 4) and denote

$$\beta(w,\lambda) := \beta_2^{(\gamma)}(J_\lambda), \quad \gamma := 16 \max_{1\leqslant j\leqslant N} \|k_j\|_\infty^{1/2} \max_{j\in J_\lambda} \|w_j\|_{K_j}.$$

Proposition 4.3 implies that

$$\mathfrak{a}(w) \leqslant \beta(w,\lambda)\Big(\sum_{j\in J_\lambda} \|w_j\|_{K_j}^2\Big)^{1/2}. \tag{5.1}$$

We will also need the following quantities that would play the role of "approximate dimensions" of the sets of random variables $X_{\mathbb{T}_{w_j}}$, $j \in J$ (local versions of $d(w,\lambda)$):

$$\mathfrak{d}_j(w,\lambda) := \min\Big\{m \geqslant 0 : \frac{m\sigma_Y^2}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\rho_m(w_j)\right)}{\sqrt{n}}\Big\}. \tag{5.2}$$

Proposition 5.1. — *Under the above notations, the following bound holds:*

$$d(w,\lambda) \leqslant \sum_{j\in J_\lambda} \mathfrak{d}_j(w,\lambda).$$

Proof. — Denote $m_j := \mathfrak{d}_j(w,\lambda)$. Then,

$$\frac{m_j\sigma_Y^2}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\rho_{m_j}(w_j)\right)}{\sqrt{n}}, \quad j \in J_\lambda$$

and, for all $j \in J_\lambda$ and $\delta > 0$, there exist $L_j \subset \mathscr{L}_X$ such that $\dim(L_j) \leqslant m_j$ and

$$\sup_{t\in\mathbb{T}_{w_j}} \|P_{L_j}^\perp(X(t) - \mathbb{E}X(t))\|_{L_2(\Pi)} \leqslant \rho_{m_j}(w_j) + \delta.$$

Denote $L := \text{l.s.}\big(\bigcup_{j\in J_\lambda} L_j\big)$. Then,

$$\begin{aligned}\sup_{t\in\mathbb{T}_w} \|P_L^\perp(X(t) - \mathbb{E}X(t))\|_{L_2(\Pi)} &\leqslant \max_{j\in J_\lambda} \sup_{t\in\mathbb{T}_{w_j}} \|P_{L_j}^\perp(X(t) - \mathbb{E}X(t))\|_{L_2(\Pi)} \\ &\leqslant \max_{j\in J_\lambda} \rho_{m_j}(w_j) + \delta\end{aligned}$$

and

$$\frac{\sigma_Y^2 \sum_{j\in J_\lambda} m_j}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\max_{j\in J_\lambda} \rho_{m_j}(w_j)\right)}{\sqrt{n}}.$$

Since $\dim(L) \leqslant \sum_{j\in J_\lambda} m_j := m$, we have

$$\rho_m(w) \leqslant \sup_{t\in\mathbb{T}_w} \|P_L^\perp(X(t) - \mathbb{E}X(t))\|_{L_2(\Pi)} \leqslant \max_{j\in J_\lambda} \rho_{m_j}(w_j) + \delta.$$

It follows that

$$\frac{\sigma_Y^2 m}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\rho_m(w) - \delta\right)}{\sqrt{n}}.$$

Since $\delta > 0$ is arbitrary, this yields

$$\frac{\sigma_Y^2 m}{n} \geqslant \|\lambda\|_1 \frac{\gamma_2\left(\rho_m(w)\right)}{\sqrt{n}},$$

and the result follows. □

As a very simple example, let $I_1, \dots, I_N$ be disjoint finite subsets of the set $\mathbb{N}$ of natural numbers and let

$$X(t) = X^{(j)}(t) = \sum_{k \in I_j} \eta_k^{(j)} \phi_k^{(j)}(t), \quad t \in \mathbb{T}_j,\ j = 1, \dots, N,$$

where $\phi_k^{(j)}$, $k \in I_j$ are bounded measurable functions on $\mathbb{T}_j$, $j = 1, \dots, N$, and $\{\eta_k^{(j)} : k \in I_j,\, j = 1, \dots, N\}$ are centered jointly normal random variables. Denote $m_j := \mathrm{Card}(I_j)$, $j = 1, \dots, N$. Let $\lambda \in \mathbb{D}$ and $w \in \mathscr{W}_{\lambda,\Delta}$. Obviously,

$$\mathfrak{d}_j(w, \lambda) \leqslant m_j,\ j \in J_\lambda,$$

so, we have a simple bound

$$d(w, \lambda) \leqslant \sum_{j \in J_\lambda} m_j.$$

The next statement immediately follows from Theorem 3.2, Proposition 5.1 and bound (5.1).

COROLLARY 5.1. — *Suppose that assumptions and notations of Theorem* 3.2 *hold. There exists an absolute constant* $\mathfrak{C} > 0$ *such that with probability at least* $1 - e^{-s}$

$$\begin{aligned} (5.3)\quad \big\|f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \leqslant \inf_{\substack{\Delta \in \mathscr{T}, \lambda \in \mathbb{D}, \\ w \in \mathscr{W}_{\lambda,\Delta}, a \in \mathbb{R}}} \Bigg[& \|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} \\ & + 2\varepsilon^2 \beta^2(w, \lambda) \textstyle\sum_{j \in J_\lambda} \|w_j\|^2_{K_j} + \mathfrak{C}\, \dfrac{\sigma_Y^2 \sum_{j \in J_\lambda} \mathfrak{d}_j(w, \lambda)}{n} + \mathfrak{C}\, \dfrac{\|\lambda\|_1^2 S^2(\mathbb{T})}{n} \Bigg] + \mathfrak{C}\, \frac{\sigma_Y^2 \overline{s}}{n}. \end{aligned}$$

The term $\|\lambda\|_1^2 S^2(\mathbb{T})/n$ that depends on $\|\lambda\|_1^2$ can be dropped if $\|\lambda\|_1$ is not too large (see Theorem 3.3). In general, this term can be controlled in terms of sparsity parameter $N(\lambda)$ and $\|\lambda\|_{L_2(\mu)}$. To this end, note that, by Cauchy-Schwarz inequality,

$$\begin{aligned} \|\lambda\|_1 &= \sum_{j \in J_\lambda} \int_{\mathbb{T}_j} |\lambda| d\mu \leqslant \sum_{j \in J_\lambda} \left(\int_{\mathbb{T}_j} |\lambda|^2 d\mu \right)^{1/2} \mu^{1/2}(\mathbb{T}_j) \\ &\leqslant \left(\sum_{j \in J_\lambda} \int_{\mathbb{T}_j} |\lambda|^2 d\mu \right)^{1/2} \left(\sum_{j \in J_\lambda} \mu(\mathbb{T}_j) \right)^{1/2} \leqslant \|\lambda\|_{L_2(\mu)} \max_{j \in J_\lambda} \mu^{1/2}(\mathbb{T}_j) \sqrt{N(\lambda)}. \end{aligned}$$

For an arbitrary oracle $\lambda \in \mathbb{T}$, arbitrary partition $\Delta \in \mathscr{T}$, arbitrary subgradient $w \in \mathscr{W}_{\lambda,\mathscr{T}}$ and for

$$\varepsilon = D\, \frac{\sigma_Y S(\mathbb{T}) \sqrt{s}}{\sqrt{n}},$$

we have the following inequality that holds with probability at least $1 - e^{-s}$:

$$(5.4)\qquad \big\|f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \leqslant \big\|f_{\lambda, a(\lambda)} - f_*\big\|^2_{L_2(\Pi)} + \mathfrak{C} \left[\mathfrak{Q}(w, \lambda, \Delta) \frac{N(\lambda)}{n} + \frac{\sigma_Y^2 \overline{s}}{n} \right],$$

where

$$\mathfrak{Q}(w,\lambda,\Delta) := \sigma_Y^2 S(\mathbb{T})^2 \beta^2(w,\lambda) \max_{j\in J_\lambda} \|w_j\|_{K_j}^2 s + \sigma_Y^2 \max_{j\in J_\lambda} \mathfrak{d}_j(w,\lambda) + S^2(\mathbb{T})\|\lambda\|_{L_2(\mu)}^2 \max_{j\in J_\lambda} \mu(\mathbb{T}_j).$$

Thus, if there is an oracle $\lambda \in \mathbb{D}$ for which the approximation error $\big\|f_{\lambda,a(\lambda)} - f_*\big\|_{L_2(\Pi)}^2$ is small and the quantity $\mathfrak{Q}(w,\lambda,\Delta)$ is of a moderate size, then the error of the estimator $(\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon)$ is essentially controlled by the quantity $N(\lambda)/n$ (up to log factors). Since $N(\lambda)$ can be viewed as a degree of sparsity of the oracle λ, this explains the connection of the oracle inequality of Corollary 5.1 and now classical bounds for LASSO in the case of large finite dictionaries. Once again, it is important to emphasize that the estimation method (1.2) does not require any knowledge of a "weakly correlated partition" Δ. The method is adaptive in the sense that, if there exists a partition Δ such that $\beta(w,\lambda)$ and other quantities involved in the definition of $\mathfrak{Q}(w,\lambda,\Delta)$ are not large, then the size of the error depends on the degree of sparsity $N(\lambda)$ with respect to the partition of oracles λ that provide good approximation of the target.

In the simplest example, $\mathbb{T} := \{1,\dots,N\}$ and the partition $\Delta := \big\{\{1\},\dots,\{N\}\big\}$ (so, $\mathbb{T}$ is partitioned in one point sets). Let μ be the counting measure. Thus, X is an N-dimensional subgaussian vector and we are in the framework of standard high-dimensional multiple regression model. For simplicity, assume that X is scaled in such a way that $\mathbb{E}X(t) = 0$, $\mathbb{E}X^2(t) = 1$. The estimator (1.2) becomes a version of usual LASSO-estimator. Then, it is easy to check that $S(\mathbb{T}) \leqslant C\sqrt{\log N}$. Also, in this case RKHS-spaces $\mathbb{H}(K_j)$, $j = 1,\dots,N$, are one-dimensional and we have $\|w_j\|_{K_j} = |w(j)|$, $j = 1,\dots,N$. For an oracle $\lambda \in \mathbb{D}$,

$$N(\lambda) = \mathrm{Card}(J_\lambda),\ J_\lambda = \mathrm{supp}(\lambda) = \{1 \leqslant j \leqslant N : \lambda_j \neq 0\}.$$

In this case, we can set $w(j) = \mathrm{sign}(\lambda(j))$, $j = 1,\dots,N$. Also, we obviously have $\mathfrak{d}_j(w,\lambda) = 1$. Finally, in this case the quantity $\beta_2^{(\gamma)}(J)$ coincides with standard "cone constrained" characteristics frequently used in the literature on sparse recovery (see, e.g., [26], Section 7.2.2). We will use $\beta(\lambda) = \beta(w,\lambda) = \beta_2^{(16)}(J_\lambda)$. Then, Corollary 5.1 takes the following form.

Corollary 5.2. — *There exist absolute constants $\mathfrak{C}, \mathfrak{c}$ and D such that the following holds. For any $s \geqslant 1$ with $\overline{s} := s + 3\log(\log_2 n + 2) + 3 \leqslant \mathfrak{c}\sqrt{n}/\log n$ and for all ε satisfying*

$$\varepsilon \geqslant D\,\frac{\sigma_Y\sqrt{s\log N}}{\sqrt{n}}, \tag{5.5}$$

with probability at least $1 - e^{-s}$

$$\big\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\big\|_{L_2(\Pi)}^2 \leqslant \inf_{\lambda\in\mathbb{D},a\in\mathbb{R}}\Big[\|f_{\lambda,a} - f_*\|_{L_2(\Pi)}^2 + 2\beta^2(\lambda)N(\lambda)\varepsilon^2 + \mathfrak{C}\frac{\sigma_Y^2 N(\lambda)}{n} + \mathfrak{C}\frac{\|\lambda\|_1^2\log N}{n}\Big] + \mathfrak{C}\frac{\sigma_Y^2\overline{s}}{n}. \tag{5.6}$$

More generally, assume that $\mathbb{T}$ is a finite set with a counting measure μ and consider an arbitrary partition $\Delta = \{\mathbb{T}_1, \dots, \mathbb{T}_N\}$ of $\mathbb{T}$. Denote

$$m_j := \mu(\mathbb{T}_j) = \mathrm{Card}(\mathbb{T}_j), \quad j = 1, \dots, N$$

and $m := \mu(\mathbb{T}) = \mathrm{Card}(\mathbb{T})$. As before, X is subgaussian and $\mathbb{E}X(t) = 0$, $\mathbb{E}X^2(t) = 1$. Then, we have $S(\mathbb{T}) \leqslant C\sqrt{\log m}$. In this case, covariance operators K_j are acting in m_j-dimensional Euclidean spaces and we have

$$\|w_j\|_{K_j} = \|K_j^{-1/2} w_j\|_2, \quad j = 1, \dots, N.$$

Clearly, we also have $\mathfrak{d}_j(w, \lambda) \leqslant m_j$. Thus, the oracle inequality of Corollary 5.1 implies that

$$\begin{aligned} (5.7) \quad \big\|f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \leqslant \inf_{\substack{\Delta \in \mathscr{T}, \lambda \in \mathbb{D}, \\ w \in \mathscr{W}_{\lambda,\Delta}, a \in \mathbb{R}}} \Bigg[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} \\ + 2\beta^2(w,\lambda) \textstyle\sum_{j \in J_\lambda} \|K_j^{-1/2} w_j\|_2^2 \varepsilon^2 + \mathfrak{C}\, \dfrac{\sigma_Y^2 \sum_{j \in J_\lambda} m_j}{n} + \mathfrak{C}\, \dfrac{\|\lambda\|_1^2 \log m}{n} \Bigg] + \mathfrak{C}\, \dfrac{\sigma_Y^2 \overline{s}}{n}. \end{aligned}$$

This holds with probability at least $1 - e^{-s}$ for all ε satisfying $\varepsilon \geqslant D\sigma_Y \sqrt{s \log m}/\sqrt{n}$.

6. Stationary and piecewise stationary processes

Suppose $\mathbb{T}$ is a bounded subset of $\mathbb{R}^d$ with Lebesgue measure μ and let $\Delta = \{\mathbb{T}_1, \dots, \mathbb{T}_N\}$ be a measurable partition of $\mathbb{T}$.

Assumption 6.1. — Suppose that each set $\mathbb{T}_j$ is contained in a ball of radius r. In what follows, we assume that $r \geqslant N^{-1/d}$. It is easy to see that there exists a constant $\kappa \geqslant 2$ depending only on d such that the ε-covering numbers of $\mathbb{T}$ with respect to the standard Euclidean distance satisfy the condition

$$N(\mathbb{T}; \varepsilon) \leqslant \left(\frac{R}{\varepsilon}\right)^d \bigvee N, \quad \varepsilon \in (0, R), \tag{6.1}$$

where $R = \kappa N^{1/d} r$.

Let $X^{(j)}$, $j = 1, \dots, N$, be centered stationary subgaussian processes on $\mathbb{R}^d$ and let

$$X(t) := \sum_{j=1}^N X^{(j)}(t) I_{\mathbb{T}_j}(t), \quad t \in \mathbb{T}.$$

Thus, we can view the process X as "piecewise stationary". Let K_j denote the covariance operator and v_j denote the spectral density of $X^{(j)}$, $j = 1, \dots, N$ (we assume that the spectral densities exist).

Assumption 6.2. — Suppose that, for some constant $B > 0$ and some $p > d/2$,

$$\frac{1}{B} \frac{1}{(1 + |t|^2)^p} \leqslant v_j(t) \leqslant B \frac{1}{(1 + |t|^2)^p}, \quad t \in \mathbb{R}^d,\ j = 1, \dots, N. \tag{6.2}$$

We use the notations J_λ, $N(\lambda) = \mathrm{Card}(J_\lambda)$ and $\beta(\lambda) = \beta(w, \lambda)$ introduced in Section 5. Let $\lambda \in \mathbb{D}$ be an oracle such that, for each $j \in J_\lambda$ we either have that $\lambda(t) \geqslant 0$ for all $t \in \mathbb{T}_j$, or $\lambda(t) \leqslant 0$ for all $t \in \mathbb{T}_j$. Thus, λ does not change its sign inside the elements of the partition. Denote $\mathbb{D}_\Delta$ the set of all such oracles in $\mathbb{D}$.

Finally, denote $R(\lambda) = \kappa(N(\lambda))^{1/d} r$. Clearly, $r \leqslant R(\lambda) \leqslant R$ (we assume that $N(\lambda) \geqslant 1$) and condition (6.1) holds for the covering numbers of the set $\bigcup_{j \in J_\lambda} \mathbb{T}_j$ with $R(\lambda)$ in place of R.

Theorem 6.1. — *There exist constants $\mathfrak{C}, \mathfrak{c}$ and D depending only on B, p, d such that the following holds. For any $s \geqslant 1$ with $\overline{s} := s + 3\log(\log_2 n + 2) + 3 \leqslant \mathfrak{c}\sqrt{n}/\log n$, for all ε satisfying*

$$\varepsilon \geqslant D\,\frac{\sigma_Y \sqrt{s(\log N \vee \log r)}}{\sqrt{n}},$$

with probability at least $1 - e^{-s}$

$$\begin{aligned}(6.3)\quad \big\|f_{\widehat{\lambda}_\varepsilon, \widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \leqslant \inf_{\lambda \in \mathbb{D}_\Delta, a \in \mathbb{R}} \Bigg[&\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} \\ &+ \mathfrak{C}\left(\sigma_Y^2 r^d\right)^{\frac{2p-d}{2p+d}} L^{d/(2p+d)}\, \frac{\|\lambda\|_1^{2d/(2p+d)} N(\lambda)^{(2p-d)/(2p+d)}}{n^{2p/(2p+d)}} + \mathfrak{C}\,\frac{\sigma_Y^2 N(\lambda)}{n} \\ &+ \mathfrak{C} r^d (1 + r^{-p})^2 \beta^2(\lambda) N(\lambda) \varepsilon^2 + \mathfrak{C}\,\frac{\|\lambda\|_1^2 (\log N \vee |\log r|)}{n} \Bigg] + \mathfrak{C}\,\frac{\sigma_Y^2 \overline{s}}{n},\end{aligned}$$

where $L := \log N \vee \log n \vee |\log \sigma_Y| \vee |\log r|$.

We will now consider a stationary subgaussian random field $X(t)$, $t \in \mathbb{R}^d$, observed in a ball $\mathbb{T} = \{t : |t| \leqslant R\}$ of radius $R \geqslant 2$.

Assumption 6.3. — Suppose that X has a spectral density $v(t)$, $t \in \mathbb{R}^d$ and, for some constant $B > 0$ and some $p > d/2$,

$$\frac{1}{B}\frac{1}{(1 + |t|^2)^p} \leqslant v(t) \leqslant B\frac{1}{(1 + |t|^2)^p}, \quad t \in \mathbb{R}^d. \tag{6.4}$$

Let $\lambda \in \mathbb{D}$ be an oracle such that $\mathrm{supp}(\lambda)$ can be covered by a union of $N(\lambda)$ disjoint balls $B(t_1; r), \dots, B(t_{N(\lambda)}; r)$ of radius $r \leqslant R/2$. Moreover, let us assume that the balls in this covering are well separated in the sense that the distance between any two distinct balls is at least $2r$. In addition to this, assume that λ does not change sign on each of the sets $B(t_j; r) \cap \mathrm{supp}(\lambda)$, $j = 1, \dots, N(\lambda)$. Let $\mathbb{D}_r$ denote the set of all such oracles $\lambda \in \mathbb{D}$.

Then, the following theorem holds.

Theorem 6.2. — *There exist constants $\mathfrak{C}, \mathfrak{c}$ and D depending only on B, p, d such that the following holds. For any $s \geqslant 1$ with $\overline{s} := s + 3\log(\log_2 n + 2) + 3 \leqslant \mathfrak{c}\sqrt{n}/\log n$, for all ε satisfying*

$$\varepsilon \geqslant D\,\frac{\sigma_Y \sqrt{s \log R}}{\sqrt{n}},$$

with probability at least $1 - e^{-s}$

$$\begin{aligned}(6.5)\quad \big\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \leqslant \inf_{\lambda\in\mathbb{D}_r, a\in\mathbb{R}}\Bigg[&\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)}\\ &+ \mathfrak{C}\left(\sigma_Y^2 r^d\right)^{\frac{2p-d}{2p+d}} L^{d/(2p+d)}\,\frac{\|\lambda\|_1^{2d/(2p+d)} N(\lambda)^{(2p-d)/(2p+d)}}{n^{2p/(2p+d)}}\\ &+ \mathfrak{C}\,\frac{\sigma_Y^2 N(\lambda)}{n} + \mathfrak{C} r^d (1+r^{-p})^2 N(\lambda)\varepsilon^2 + \mathfrak{C}\,\frac{\|\lambda\|_1^2 \log R}{n}\Bigg] + \mathfrak{C}\,\frac{\sigma_Y^2 \overline{s}}{n},\end{aligned}$$

where $L := \log n \vee |\log \sigma_Y| \vee \log R \vee |\log r|$.

Note that in Theorems 6.1 and 6.2 the error rate depends on "sparsity parameter" $N(\lambda)$ (its meaning is somewhat different in these two cases). Moreover, the error rate involves a "nonparametric term" $O(n^{-2p/(2p+d)})$. Thus, $p > d/2$ plays a role of smoothness parameter in this problem.

Often, it is natural to assume that the target $f_*(X)$ can be approximated by $f_\Lambda(X)$, where Λ is a discrete signed measure supported on a "well-separated" subset of $\mathbb{T}$, so that $f_\Lambda(X) = \sum_{j=1}^{N(\Lambda)} \lambda_j X(t_j)$, where

$$\min_{1\leqslant i<j\leqslant N(\Lambda)} |t_i - t_j| \geqslant 3\delta(\Lambda) > 0$$

and $\delta(\Lambda)$ is large enough. Such a discrete oracle Λ can be further approximated by a linear combination of continuous "spikes" supported in well separated disjoint balls of radius $r > 0$. This can be done for an arbitrary $r < \delta(\Lambda)$ and optimizing the bound of Theorem 6.2 with respect to r would lead to a bound with a faster error rate. We will implement this in a special (and practically important) case when the design processes X_j, $j = 1, \dots, n$ are observed on a discrete grid in $\mathbb{R}^d$. Specifically, assume that $\mathbb{T} = \mathscr{G}_N = \big\{t_j = 2\pi j/N,\ j \in \{1, \dots, N\}^d\big\}$ and it is equipped with the counting measure μ, see section 4.3 for more details. Note that in this case we are in the framework of a standard high-dimensional linear regression with highly correlated design. Functions λ on $\mathbb{T}$ can be identified with vectors in $\mathbb{R}^{N^d}$ and we will assume that $\mathbb{D} := R^{N^d}$. Suppose that Assumption 6.3 holds and let λ be an oracle such that $J(\lambda) = \mathrm{supp}(\lambda) \subset \{1, \dots, N\}^d$, $N(\lambda) := \mathrm{Card}(J(\lambda))$, and

$$\min_{i,j\in J(\lambda), i\neq j} \frac{|i-j|}{N} =: 2\delta(\lambda) \geqslant \frac{1}{N},$$

where $|i-j|$ stands for the usual Euclidean distance in $\mathbb{R}^d$. We are mainly interested in the oracles λ with "well-separated" non-zero elements, meaning that $\delta(\lambda) \gg 1/N$. In this setting, the following result holds.

THEOREM 6.3. — *There exist constants* $\mathfrak{C}, \mathfrak{c}$ *and* D *depending only on* B, p, d *such that the following holds. For any* $s \geqslant 1$ *with* $\overline{s} := s + 3\log(\log_2 n + 2) + 3 \leqslant \mathfrak{c}\sqrt{n}/\log n$, *let*

$$\varepsilon = D\,\frac{\sigma_Y \sqrt{s}}{\sqrt{n}}.$$

Then with probability at least $1 - e^{-s}$

$$\big\|f_{\widehat{\lambda}_\varepsilon,\widehat{a}_\varepsilon} - f_*\big\|^2_{L_2(\Pi)} \leqslant \inf_{\lambda\in\mathbb{R}^{N^d},a\in\mathbb{R}}\Bigg[\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + \mathfrak{C}\delta(\lambda)^{d-2p}\sigma_Y^2\frac{N(\lambda)s}{n} + \mathfrak{C}\sigma_Y^{2p/(p+d)}\big(sL\|\lambda\|_1^2\big)^{d/(2p+2d)}\frac{N(\lambda)^{p/(p+d)}}{n^{(2p+d)/(2p+2d)}} + \mathfrak{C}\,\frac{\|\lambda\|_1^2}{n}\Bigg] + \mathfrak{C}\,\frac{\sigma_Y^2\overline{s}}{n}.$$

where $L = \log n \vee \log N \vee |\log\sigma_Y|$.

7. Proofs of the main results

7.1. Preliminaries. — Recall that

$$F_n(\lambda,a) := P_n(\ell\bullet f_{\lambda,a}) + \varepsilon\|\lambda\|_1, \quad F(\lambda,a) := P(\ell\bullet f_{\lambda,a}) + \varepsilon\|\lambda\|_1.$$

In the proofs of the main results, we will use necessary conditions for the minima in problems (1.2), (2.3) that will be stated now. Given a convex functional $H : L_1(\mu)\times\mathbb{R}\mapsto\mathbb{R}$, define its directional derivative at a point $(\lambda,a)\in L_1(\mu)\times\mathbb{R}$ in the direction $u = (u_1,u_2)\in L_1(\mu)\times\mathbb{R}$ as

$$DH(\lambda,a)(u) := \lim_{t\downarrow 0}\frac{H((\lambda,a)+tu) - H(\lambda,a)}{t}.$$

Proposition 7.1. — *For any* $\lambda_1,\lambda_2\in\mathbb{D}$ *and* $a_1,a_2\in\mathbb{R}$,

$$DF_n(\lambda_1,a_1)(\lambda_2-\lambda_1,a_2-a_1) = P_n(\ell'\bullet f_{\lambda_1,a_1})(f_{\lambda_2,a_2} - f_{\lambda_1,a_1}) + \varepsilon\,\langle w_1,\lambda_2-\lambda_1\rangle$$

for some $w_1\in\partial\|\lambda_1\|_1$ *that depends on* λ_2. *Similarly,*

$$DF(\lambda_1,a_1)(\lambda_2-\lambda_1,a_2-a_1) = P(\ell'\bullet f_{\lambda_1,a_1})(f_{\lambda_2,a_2} - f_{\lambda_1,a_1}) + \varepsilon\,\langle w_1,\lambda_2-\lambda_1\rangle\,.$$

Proof. — The treatment of the terms $P_n(\ell\bullet f_{\lambda,a}), P(\ell\bullet f_{\lambda,a})$ is straightforward, so it only remains to examine the L_1-penalty term. Let $v := \lambda_2 - \lambda_1$. Since the function $(0,1)\ni s\mapsto|\lambda_1(t) + sv(t)|$ is convex, we have that

$$(0,1)\ni s\longmapsto\frac{|\lambda_1(t)+sv(t)| - |\lambda_1(t)|}{s}$$

is nondecreasing. Given a decreasing sequence $\{s_n\}_{n\geqslant 0}\subset(0,1)$ such that $s_n\to 0$, the sequence of functions

$$g_n(t) := \frac{|\lambda_1(t)+s_nv(t)| - |\lambda_1(t)|}{s_n}$$

monotonically converges to

$$g(t) = \begin{cases}\operatorname{sign}(\lambda_1(t))v(t), & \lambda_1(t)\neq 0\\ \operatorname{sign}(v(t))v(t), & \text{else.}\end{cases}$$

Moreover, $g_n(t)$ are integrable, and the monotone convergence theorem implies that

$$\lim_{n\to\infty}\int_{\mathbb{T}} g_n d\mu = \int_{\mathbb{T}} g d\mu = \int_{\mathbb{T}} w_1 u d\mu,$$

where $|w_1(t)|\leqslant 1$, $t\in\mathbb{T}$ and $w_1(t) = \operatorname{sign}(\lambda_1(t))$, $\lambda_1(t)\neq 0$. In particular, $w_1\in\partial\|\lambda_1\|_1$. □

When $\lambda_1 = \widehat{\lambda}_\varepsilon$ (which minimizes F_n), the corresponding directional derivatives must be nonnegative for any $\lambda_2 \in \mathbb{D}$.

7.2. Proof of Theorem 2.1. — Let $(\overline{\lambda}, \overline{w}, \overline{a})$ be a triple that minimizes the right-hand side of (2.9). If the infimum is not attained, one can consider the triple for which the right-hand side is arbitrarily close to the infimum and follow the argument below.

Since $(\lambda_\varepsilon, a_\varepsilon)$ minimizes $F(\lambda, a)$ over $\mathbb{D} \times \mathbb{R}$, the directional derivative

$$DF(\lambda_\varepsilon, a_\varepsilon)(\overline{\lambda} - \lambda_\varepsilon, \overline{a} - a_\varepsilon)$$

is nonnegative for any $\overline{\lambda} \in \mathbb{D}$, $\overline{a} \in \mathbb{R}$. By Proposition 7.1, this is equivalent to the following: there exists $w_\varepsilon \in \partial\|\lambda_\varepsilon\|_1$ such that

$$P(\ell' \bullet f_{\lambda_\varepsilon, a_\varepsilon})(f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}}) + \varepsilon \left\langle w_\varepsilon, \lambda_\varepsilon - \overline{\lambda} \right\rangle \leqslant 0. \tag{7.1}$$

Let $\overline{w} \in \partial\|\overline{\lambda}\|_1$. Since

$$(\ell' \bullet f_{\lambda_\varepsilon, a_\varepsilon})(x, y) = 2(f_{\lambda_\varepsilon, a_\varepsilon}(x) - y)$$

and also $Y = f_*(X) + \xi$, where $\mathbb{E}(\xi | X) = 0$, we have

$$\begin{aligned} P(\ell' \bullet f_{\lambda_\varepsilon, a_\varepsilon})(f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}}) &= 2\mathbb{E}(f_{\lambda_\varepsilon, a_\varepsilon}(X) - Y)(f_{\lambda_\varepsilon, a_\varepsilon}(X) - f_{\overline{\lambda}, \overline{a}}(X)) \\ &= 2\langle f_{\lambda_\varepsilon, a_\varepsilon} - f_*, f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}} \rangle_{L_2(\Pi)}. \end{aligned}$$

Thus, (7.1) can be rewritten as

$$2\langle f_{\lambda_\varepsilon, a_\varepsilon} - f_*, f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}} \rangle_{L_2(\Pi)} + \varepsilon \left\langle w_\varepsilon - \overline{w}, \lambda_\varepsilon - \overline{\lambda} \right\rangle \leqslant \varepsilon \left\langle \overline{w}, \lambda_\varepsilon - \overline{\lambda} \right\rangle. \tag{7.2}$$

Note that

$$\begin{aligned} 2\langle f_{\lambda_\varepsilon, a_\varepsilon} - f_*, f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}} \rangle_{L_2(\Pi)} = \|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)} + \|f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}}\|^2_{L_2(\Pi)} \\ - \|f_{\overline{\lambda}, \overline{a}} - f_*\|^2_{L_2(\Pi)} \end{aligned}$$

and

$$\left\langle w_\varepsilon - \overline{w}, \lambda_\varepsilon - \overline{\lambda} \right\rangle \geqslant \frac{1}{2} \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |\lambda_\varepsilon| d\mu.$$

Hence

$$\begin{aligned} \|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)} + \|f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda}, \overline{a}}\|^2_{L_2(\Pi)} + \frac{\varepsilon}{2} \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |\lambda_\varepsilon| d\mu \\ \leqslant \|f_{\overline{\lambda}, \overline{a}} - f_*\|^2_{L_2(\Pi)} + \varepsilon \left\langle \overline{w}, \lambda_\varepsilon - \overline{\lambda} \right\rangle. \end{aligned} \tag{7.3}$$

Consider two cases: first, if

$$\|f_{\lambda_\varepsilon, a_\varepsilon} - f_*\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4} \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |\lambda_\varepsilon| d\mu \leqslant \|f_{\overline{\lambda}, \overline{a}} - f_*\|^2_{L_2(\Pi)},$$

then inequality (2.9) clearly holds. Otherwise, (7.3) implies that

$$\int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |\lambda_\varepsilon| d\mu = \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |\lambda_\varepsilon - \overline{\lambda}| d\mu \leqslant 4 \left\langle \overline{w}, \lambda_\varepsilon - \overline{\lambda} \right\rangle.$$

Hence, $\lambda_\varepsilon - \overline{\lambda} \in C^{(4)}_{\overline{w}}$ and

$$\varepsilon \langle \overline{w}, \lambda_\varepsilon - \overline{\lambda} \rangle \leqslant \varepsilon \big\| f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda},\overline{a}} \big\|_{L_2(\Pi)} \mathfrak{a}(\overline{w}) \leqslant \frac{1}{4}\varepsilon^2 \mathfrak{a}^2(\overline{w}) + \big\| f_{\lambda_\varepsilon, a_\varepsilon} - f_{\overline{\lambda},\overline{a}} \big\|^2_{L_2(\Pi)},$$

where we used the definition of $\mathfrak{a}(\overline{w})$ and a simple inequality $ab \leqslant \frac{1}{4}a^2 + b^2$. Substituting this bound into (7.3) gives the result.

7.3. Proof of Theorem 3.2. — Throughout the proof, C, C_1, c, c_1, etc. denote absolute constants whose values may change from line to line.

Step 1. Reduction to empirical processes. — Let $(\overline{\lambda}, \overline{w}, \overline{a})$ be a triple that minimizes the right-hand side of bound (3.6). Clearly, $\overline{a} = a(\overline{\lambda})$, see (2.1). If the infimum is not attained, it is easy to modify the argument by considering a triple for which the right-hand side is arbitrarily close to the infimum. Since $0 \in \mathbb{D}$ and, for $\lambda = 0$, one can also take $w = 0$ and $a = \mathbb{E}Y = \mathbb{E}f_*(X)$, we have that

$$\|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} \leqslant \|f_* - \Pi f_*\|^2_{L_2(\Pi)} = \mathrm{Var}(f_*(X)) \tag{7.4}$$

and

$$\|\overline{\lambda}\|_1^2 \leqslant \frac{\|f_* - \Pi f_*\|^2_{L_2(\Pi)} n}{CS^2(\mathbb{T})}. \tag{7.5}$$

We will write in what follows $\widehat{\lambda} = \widehat{\lambda}_\varepsilon$ and set $\widehat{a} := \widehat{a}(\widehat{\lambda})$. Since $(\widehat{\lambda}, \widehat{a})$ minimizes $F_n(\lambda, a)$ over $\mathbb{D} \times \mathbb{R}$, the directional derivative $DF_n(\widehat{\lambda}, \widehat{a})(\overline{\lambda} - \widehat{\lambda}, \overline{a} - \widehat{a})$ is nonnegative. Here and in what follows, we use the "optimal" value $\overline{a} = a(\overline{\lambda})$, see (2.1). By Proposition 7.1, this is equivalent to the following: there exists $\widehat{w} \in \partial\|\widehat{\lambda}\|_1$ such that

$$P_n(\ell' \bullet f_{\widehat{\lambda},\widehat{a}})(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + \varepsilon\langle \widehat{w}, \widehat{\lambda} - \overline{\lambda} \rangle \leqslant 0. \tag{7.6}$$

Since $\overline{w} \in \partial\|\overline{\lambda}\|_1$, (7.6) can be rewritten as

$$\begin{aligned} P(\ell' \bullet f_{\widehat{\lambda},\widehat{a}})(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + \varepsilon\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda} \rangle \\ \leqslant \varepsilon\langle \overline{w}, \overline{\lambda} - \widehat{\lambda} \rangle + (P - P_n)(\ell' \bullet f_{\widehat{\lambda},\widehat{a}})\big(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big). \end{aligned} \tag{7.7}$$

Denote $\eta(x, y) := y - f_{\overline{\lambda},\overline{a}}(x)$. Observe that

$$\big(\ell' \bullet f_{\widehat{\lambda},\widehat{a}}\big)(x, y) = -2\big(y - f_{\widehat{\lambda},\widehat{a}}(x)\big) = -2\eta(x, y) + 2\big(f_{\widehat{\lambda},\widehat{a}}(x) - f_{\overline{\lambda},\overline{a}}(x)\big)$$

and, since $Y = f_*(X) + \xi$, $\mathbb{E}(\xi|X) = 0$,

$$\begin{aligned} &- P\big[\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})\big] = -\mathbb{E}\eta(X, Y)\big(f_{\widehat{\lambda},\widehat{a}}(X) - f_{\overline{\lambda},\overline{a}}(X)\big) \\ &\quad = -\mathbb{E}\big(\xi + f_*(X) - f_{\overline{\lambda},\overline{a}}(X)\big)\big(f_{\widehat{\lambda},\widehat{a}}(X) - f_{\overline{\lambda},\overline{a}}(X)\big) = \big\langle f_{\overline{\lambda},\overline{a}} - f_*, f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}} \big\rangle_{L_2(\Pi)}. \end{aligned}$$

Therefore, we get the following bound:

$$\begin{aligned} 2\big\langle f_{\overline{\lambda},\overline{a}} - f_*, f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}} \big\rangle_{L_2(\Pi)} + 2\big\| f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}} \big\|^2_{L_2(\Pi)} + \varepsilon\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda} \rangle \\ \leqslant \varepsilon\langle \overline{w}, \overline{\lambda} - \widehat{\lambda} \rangle + 2(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + 2(\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2. \end{aligned}$$

Using the fact that

$$2\langle f_{\overline{\lambda},\overline{a}} - f_*, f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\rangle_{L_2(\Pi)} = \|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} - \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} - \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)},$$

it can be rewritten as

$$\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \varepsilon\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + \varepsilon\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle + 2(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + 2(\Pi - \Pi_n)\big(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big)^2. \tag{7.8}$$

The main part of the proof deals with bounding the empirical processes in the right-hand side of (7.8). In what follows, L denotes a subspace of the subgaussian space $\mathscr{L}_X \subset L_2(\mathbb{P})$ (the closed linear span of $\{X(t) - \mathbb{E}X(t) : t \in \mathbb{T}\}$). Let $d := \dim(L)$ and let $P_L, P_{L^\perp}$ be the orthogonal projections onto the subspace L and its orthogonal complement $L^\perp \subset \mathscr{L}_X$, and

$$\rho := \rho(L) := \sup_{t\in\mathbb{T}_{\overline{w}}} \|P_{L^\perp}(X(t) - \mathbb{E}X(t))\|_{L_2(\mathbb{P})}\,. \tag{7.9}$$

Step 2. Bounds for $(P_n - P)\big[\eta\big(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big)\big]$. — Let $f^0_\lambda(\cdot) := f_{\lambda,a}(\cdot) - \Pi f_{\lambda,a}$, which clearly satisfies $\Pi f^0_\lambda = 0$. Observe that the following decomposition holds:

$$f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}} = f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}} + \overline{Y}_n - \mathbb{E}Y + \big\langle \widehat{\lambda} - \overline{\lambda}, \mathbb{E}X - \overline{X}_n\big\rangle + \big\langle \overline{\lambda}, \mathbb{E}X - \overline{X}_n\big\rangle\,. \tag{7.10}$$

This implies

$$(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) = (P_n - P)\eta(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}) + (P_n - P)\eta(\overline{Y}_n - \mathbb{E}Y) + (P_n - P)\eta\cdot(\Pi - \Pi_n)(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}) + (P_n - P)\eta\,\big\langle \overline{\lambda}, \mathbb{E}X - \overline{X}_n\big\rangle\,.$$

Denote

$$\Lambda(\delta,\Delta,R) := \Big\{\lambda\in\mathbb{D} : \big\|f^0_\lambda - f^0_{\overline{\lambda}}\big\|_{L_2(\Pi)} \leqslant \delta,\ \int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\lambda| d\mu \leqslant \Delta, \|\lambda\|_1 \leqslant R\Big\},$$

$$\alpha_n(\delta;\Delta;R) := \sup_{\lambda\in\Lambda(\delta,\Delta,R)} \big|(P_n - P)\eta(f^0_\lambda - f^0_{\overline{\lambda}})\big|\,,$$

$$\tau_n(\delta;\Delta;R) := \sup_{\lambda\in\Lambda(\delta,\Delta,R)} \big|(\Pi_n - \Pi)(f^0_\lambda - f^0_{\overline{\lambda}})\big|\,.$$

Then

$$\Big|(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})\Big| \leqslant \alpha_n\Big(\|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\|_{L_2(\Pi)}, \int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}| d\mu, \|\widehat{\lambda}\|_1\Big) + \big|\overline{Y}_n - \mathbb{E}Y\big|\cdot|(P_n - P)\eta| + |(P_n - P)\eta|\cdot\tau_n\Big(\|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\|_{L_2(\Pi)}, \int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}| d\mu, \|\widehat{\lambda}\|_1\Big) + |(P_n - P)\eta|\cdot\big|\big\langle \overline{\lambda}, \mathbb{E}X - \overline{X}_n\big\rangle\big|\,. \tag{7.11}$$

To provide upper bounds on each of the terms in the right-hand side of (7.11) we need several lemmas.

LEMMA 7.1. — *Let $\{Y(t), t \in \mathbb{T}\}$ be a centered subgaussian process such that*

$$\mathbb{E}Y(t)Y(s) = \mathrm{Cov}(X(t), X(s)), \quad t, s \in \mathbb{T}.$$

There exists a constant $C > 0$ such that

$$\mathbb{E} \sup_{\lambda \in \Lambda(\delta, \Delta, R)} \big|\langle Y, \lambda - \overline{\lambda}\rangle\big| \leqslant C\Big[\delta\sqrt{d} \vee (R + \|\overline{\lambda}\|_1)\gamma_2(\rho) \vee \Delta S(\mathbb{T})\Big].$$

Proof. — Denote $\mathscr{L}_Y$ the closed linear span of $\{Y(t), t \in \mathbb{T}\}$, the subgaussian space of the process Y. Clearly, the mapping $X(t) - \mathbb{E}X(t) \mapsto Y(t)$, $t \in \mathbb{T}$ can be extended to an $L_2(\mathbb{P})$-isometry of the spaces $\mathscr{L}_X, \mathscr{L}_Y \subset L_2(\mathbb{P})$. Let $\widetilde{L}$ be the image of the subspace L under this isometry. For all $\lambda \in \Lambda(\delta, \Delta, R)$ and for $u := \lambda - \overline{\lambda}$,

$$\langle Y, u\rangle = P_{\widetilde{L}}\langle Y, u\rangle + \int\limits_{\mathbb{T}_{\overline{w}}} P_{\widetilde{L}^\perp} Y(t)u(t)\mu(dt) + \int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} P_{\widetilde{L}^\perp} Y(t)u(t)\mu(dt). \tag{7.12}$$

We will use this representation and bound separately the supremum of each term to control

$$\sup\{\langle Y, \lambda - \overline{\lambda}\rangle : \lambda \in \Lambda(\delta, \Delta, R)\}.$$

For the first term, let $\xi_1, \dots, \xi_d$ be an orthonormal basis of $\widetilde{L}$. Note that for $u = \lambda - \overline{\lambda}$, $\lambda \in \Lambda(\delta, \Delta, R)$, we have $\mathbb{E}\langle Y, u\rangle^2 = \|f_u^0\|^2_{L_2(\Pi)} \leqslant \delta^2$. Therefore,

$$\begin{aligned}\mathbb{E}\sup\big\{|P_{\widetilde{L}}\langle Y, \lambda - \overline{\lambda}\rangle| : \lambda \in \Lambda(\delta, \Delta, R)\big\} &\leqslant \mathbb{E}\sup\big\{P_{\widetilde{L}}\langle Y, u\rangle : \mathbb{E}\langle Y, u\rangle^2 \leqslant \delta^2\big\} \\ &\leqslant \mathbb{E}\sup\Big\{\Big|\textstyle\sum_{k=1}^d \alpha_k\xi_k\Big| : \sum_{k=1}^d \alpha_k^2 \leqslant \delta^2\Big\} = \delta\mathbb{E}\left(\sum_{k=1}^d \xi_k^2\right)^{1/2} \leqslant \delta\sqrt{d}.\end{aligned}$$

For the second term, observe that for $u = \lambda - \overline{\lambda}, \lambda \in \Lambda(\delta, \Delta, R)$

$$\left|\int\limits_{\mathbb{T}_{\overline{w}}} P_{\widetilde{L}^\perp} Y(t)u(t)\mu(dt)\right| \leqslant \sup_{t \in \mathbb{T}_{\overline{w}}} |P_{\widetilde{L}^\perp} Y(t)|\|u\|_1 \leqslant (R + \|\overline{\lambda}\|_1) \sup_{t \in \mathbb{T}_{\overline{w}}} |P_{\widetilde{L}^\perp} Y(t)|.$$

Denote $U(t) := P_{\widetilde{L}^\perp} Y(t)$, $t \in \mathbb{T}$. Clearly, U is a centered subgaussian process such that

$$\mathbb{E}(U(t) - U(s))^2 \leqslant \mathbb{E}(Y(t) - Y(s))^2, \quad t, s \in \mathbb{T}$$

and, as a consequence, $\|U(t) - U(s)\|_{\psi_2} \leqslant c\|Y(t) - Y(s)\|_{\psi_2}$ with an absolute constant $c > 0$. Moreover, since the spaces $\mathscr{L}_Y, \widetilde{L}$ are isometric images of the spaces $\mathscr{L}_X, L$, we also have that

$$\sup_{t \in \mathbb{T}_{\overline{w}}} \mathbb{E}U^2(t) = \sup_{t \in \mathbb{T}_{\overline{w}}} \mathbb{E}|P_{\widetilde{L}^\perp} Y(t)|^2 = \sup_{t \in \mathbb{T}_{\overline{w}}} \mathbb{E}|P_{L^\perp}(X - \mathbb{E}X)(t)|^2 = \rho^2,$$

which implies that $\sup_{t \in \mathbb{T}_{\overline{w}}} \|U(t)\|_{\psi_2} \leqslant c\rho$. Then, it follows from the upper bound on sup-norms of subgaussian processes in terms of generic chaining complexities (in particular, (1.4)) that

$$\mathbb{E} \sup_{t \in \mathbb{T}_{\overline{w}}} |P_{\widetilde{L}^\perp} Y(t)| \leqslant C\gamma_2(\rho)$$

and

$$\mathbb{E} \sup_{\lambda \in \Lambda(\delta, \Delta, R)} \left|\int\limits_{\mathbb{T}_{\overline{w}}} P_{\widetilde{L}^\perp} Y(t)(\lambda - \overline{\lambda})(t)\mu(dt)\right| \leqslant C(R + \|\overline{\lambda}\|_1)\gamma_2(\rho).$$

with an absolute constant $C > 0$. Finally, for the third term, note that for $u = \lambda - \overline{\lambda}$, $\lambda \in \Lambda(\delta, \Delta, R)$

$$\left| \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} P_{\widetilde{L}^\perp} Y(t) u(t) \mu(dt) \right| \leqslant \sup_{t \in \mathbb{T}} |P_{\widetilde{L}^\perp} Y(t)| \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |u| d\mu \leqslant \Delta \sup_{t \in \mathbb{T}} |P_{\widetilde{L}^\perp} Y(t)|,$$

where we used the fact that $\overline{\lambda}(t) = 0$, $u(t) = \lambda(t)$ for $t \in \mathbb{T} \setminus \mathbb{T}_{\overline{w}}$. By an argument similar to the one used for the second term of (7.12), we get

$$\mathbb{E} \sup_{\lambda \in \Lambda(\delta, \Delta, R)} \left| \int_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} P_{\widetilde{L}^\perp} Y(t) (\lambda - \overline{\lambda})(t) \mu(dt) \right| \leqslant C \Delta S(\mathbb{T}),$$

which implies the bound of the lemma. □

LEMMA 7.2. — *There exists a constant $C > 0$ such that, for all $\delta > 0, \Delta > 0, R > 0$,*

$$\mathbb{E}\alpha_n(\delta, \Delta, R) \leqslant C \left(\|f_{\overline{\lambda}, \overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2} \right) \left[\delta \sqrt{\frac{d}{n}} \bigvee (R \vee \|\overline{\lambda}\|_1) \frac{\gamma_2(\rho)}{\sqrt{n}} \bigvee \Delta \frac{S(\mathbb{T})}{\sqrt{n}} \right] \bigvee C \left[\delta \sqrt{\frac{d}{n}} \bigvee (R \vee \|\overline{\lambda}\|_1) \frac{\gamma_2(\rho)}{\sqrt{n}} \bigvee \Delta \frac{S(\mathbb{T})}{\sqrt{n}} \right]^2.$$

Proof. — Let $\mathscr{F} := \mathscr{F}(\delta, \Delta, R) := \{f^0_\lambda - f^0_{\overline{\lambda}} : \lambda \in \Lambda(\delta, \Delta, R)\}$. We use a recent result by S. Mendelson (see Theorem A.4, statement (i)) which implies that, for all δ, Δ, R,

$$\mathbb{E}\alpha_n(\delta, \Delta, R) \leqslant C \left[\frac{\|\eta\|_{\psi_2} \gamma_2(\mathscr{F}; \psi_2)}{\sqrt{n}} \bigvee \frac{\gamma_2^2(\mathscr{F}; \psi_2)}{n} \right], \tag{7.13}$$

for an absolute constant $C > 0$. Since $\{f(X) : f \in \mathscr{F}(\delta, \Delta, R)\} \subset \mathscr{L}$ and $\mathscr{L}$ is a subgaussian space, we have that $\|f\|_{\psi_2} \leqslant c_1 \|f\|_{L_2(\Pi)}, f \in \mathscr{F}(\delta, \Delta, R)$ for some constant c_1. Therefore,

$$\gamma_2 (\mathscr{F}; \|\cdot\|_{\psi_2}) \leqslant c_1 \gamma_2 (\mathscr{F}; L_2(\Pi)).$$

Let $G(t)$, $t \in \mathbb{T}$ be a centered Gaussian process with the same covariance as the process $\{X(t), t \in \mathbb{T}\}$. Then the stochastic processes $u \mapsto \langle G, u \rangle$ has the same covariance as $u \mapsto \langle X - \mathbb{E}X, u \rangle = f^{(0)}_u(X)$, that is, $\mathbb{E} \langle G, u_1 \rangle \langle G, u_2 \rangle = \langle f^0_{u_1}, f^0_{u_2} \rangle_{L_2(\Pi)}$. By Talagrand's generic chaining theorem for Gaussian processes (Theorem 2.1.1 in [39]), this implies

$$\gamma_2 (\mathscr{F}, L_2(\Pi)) \leqslant c_2 \, \mathbb{E} \sup \{ |\langle G, \lambda - \overline{\lambda} \rangle| : \lambda \in \Lambda(\delta, \Delta, R) \}.$$

Using Lemma 7.1, we get the following bound on $\gamma_2 (\mathscr{F}, \|\cdot\|_{\psi_2})$:

$$\gamma_2 (\mathscr{F}, \|\cdot\|_{\psi_2}) \leqslant C \Big[\delta \sqrt{d} \vee (R + \|\overline{\lambda}\|_1) \gamma_2(\rho) \vee \Delta S(\mathbb{T}) \Big]. \tag{7.14}$$

Next, note that

$$\eta(X, Y) = Y - f_{\overline{\lambda}, \overline{a}}(X) = f_*(X) + \xi - f_{\overline{\lambda}, \overline{a}}(X).$$

We also have $\mathbb{E}f_*(X) = \mathbb{E}Y$ and

$$\mathbb{E}f_{\overline{\lambda}, \overline{a}}(X) = \mathbb{E}Y - \langle \overline{\lambda}, \mathbb{E}X \rangle + \mathbb{E}\langle \overline{\lambda}, X \rangle = \mathbb{E}Y$$

which implies that $\mathbb{E}\eta(X,Y)=0$. The random variable $f_*(X)-f_{\overline{\lambda},\overline{a}}(X)$ belongs to the subgaussian space $\mathscr{L}$, implying that

$$\|\eta\|_{\psi_2}=\|f_*(X)-f_{\overline{\lambda},\overline{a}}(X)+\xi\|_{\psi_2}\leqslant\|f_*(X)-f_{\overline{\lambda},\overline{a}}(X)\|_{\psi_2}+\|\xi\|_{\psi_2}\leqslant c\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}+\|\xi\|_{\psi_2} \tag{7.15}$$

with an absolute constant $c>0$. In view of (7.13), (7.14) and (7.15) easily imply the bound of the lemma. □

Our next goal is to derive an upper bound on $\alpha_n(\delta,\Delta,R)$ that holds uniformly in $\delta\in[\delta_-,\delta_+]$, $\Delta\in[\Delta_-,\Delta_+]$, $R\in[R_-,R_+]$ for some $\delta_-<\delta_+,\Delta_-<\Delta_+,R_-<R_+$ to be determined later. Let

$$J_1:=[\log_2(\delta_+/\delta_-)]+1,\quad J_2:=[\log_2(\Delta_+/\Delta_-)]+1,\quad J_3:=[\log_2(R_+/R_-)]+1$$

and, given $s>0$, let

$$\overline{s}:=s+\log\bigl((J_1+1)(J_2+1)(J_3+1)\bigr).$$

Finally, denote

$$\nu_n(\delta,\Delta,R):=\inf_{L\subset\mathscr{L}}\left[\delta\sqrt{\frac{\dim(L)}{n}}\bigvee(R\vee\|\overline{\lambda}\|_1)\frac{\gamma_2(\rho(L))}{\sqrt{n}}\bigvee\Delta\frac{S(\mathbb{T})}{\sqrt{n}}\right], \tag{7.16}$$

where the infimum is taken over all finite dimensional subspaces $L\subset\mathscr{L}$ and $\rho(L)$ is defined in (7.9).

Lemma 7.3. — *There exists a constant $C>0$ with the following property. With probability at least $1-e^{-s}$, the following inequality holds uniformly for all $\delta\in[\delta_-,\delta_+]$, $\Delta\in[\Delta_-,\Delta_+]$, $R\in[R_-,R_+]$:*

$$\alpha_n(\delta,\Delta,R)\leqslant C\left(\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}\vee\|\xi\|_{\psi_2}\right)\left[\delta\sqrt{\frac{\overline{s}}{n}}\bigvee\nu_n(\delta,\Delta,R)\right]\bigvee C\nu_n^2(\delta,\Delta,R).$$

Proof. — First, we use Adamczak's version of Talagrand's inequality (A.7) to deduce an exponential bound on $\alpha_n(\delta,\Delta,R)$ from the bound on $\mathbb{E}\alpha_n(\delta,\Delta,R)$ (for fixed $\delta,\Delta,R>0$). To this end, observe that, by the properties of Orlicz norms and subgaussian spaces,

$$\bigl\|\eta(f_\lambda^0-f_{\overline{\lambda}}^0)\bigr\|_{L_2(P)}\leqslant\|\eta\|_{L_4(P)}\bigl\|f_\lambda^0-f_{\overline{\lambda}}^0\bigr\|_{L_4(\Pi)}\leqslant c_1\|\eta\|_{\psi_2}\bigl\|f_\lambda^0-f_{\overline{\lambda}}^0\bigr\|_{L_4(\Pi)}\leqslant c_2\left(\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}+\|\xi\|_{\psi_2}\right)\bigl\|f_\lambda^0-f_{\overline{\lambda}}^0\bigr\|_{L_2(\Pi)},$$

where we used (7.15) to bound $\|\eta\|_{\psi_2}$. For all $\lambda\in\Lambda(\delta,\Delta,R)$, this implies

$$\|\eta(f_\lambda^0-f_{\overline{\lambda}}^0)\|_{L_2(P)}\leqslant c\delta\left(\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}+\|\xi\|_{\psi_2}\right).$$

Using (A.3), we will also estimate the envelope of the class $\mathscr{F}(\delta,\Delta,R)$ as follows:

$$\Bigl\|\sup_{\lambda\in\Lambda(\delta,\Delta,R)}\bigl|\eta(X,Y)(f_\lambda^0(X)-f_{\overline{\lambda}}^0(X))\bigr|\Bigr\|_{\psi_1}\leqslant c\|\eta(X,Y)\|_{\psi_2}\Bigl\|\sup_{\lambda\in\Lambda(\delta,\Delta,R)}\bigl|f_\lambda^0(X)-f_{\overline{\lambda}}^0(X)\bigr|\Bigr\|_{\psi_2}.$$

Recall that L is a subspace of the subgaussian space $\mathscr{L}$ with $\dim(L) = d$ and

$$\rho = \sup_{t\in\mathbb{T}_{\overline{w}}} \|P_{L^\perp}(X(t) - \mathbb{E}X(t))\|_{L_2(\mathbb{P})}.$$

Let $\zeta_1, \dots, \zeta_d$ be an orthonormal basis of $L \subset L_2(\mathbb{P})$. For $u = \lambda - \overline{\lambda}$, the following decomposition holds:

$$f^0_\lambda(X) - f^0_{\overline{\lambda}}(X) = \langle u, X - \mathbb{E}X\rangle = P_L\langle u, X - \mathbb{E}X\rangle + \int\limits_{\mathbb{T}_{\overline{w}}} P_{L^\perp}(X - \mathbb{E}X)(t)u(t)\mu(dt) + \int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} P_{L^\perp}(X - \mathbb{E}X)(t)u(t)\mu(dt).$$

We have

$$\begin{aligned}\Big\| \sup_{\lambda\in\Lambda(\delta,\Delta,R)} \Big|P_L\langle \lambda - \overline{\lambda}, X - \mathbb{E}X\rangle\Big|\Big\|_{\psi_2} &\leqslant \Big\|\sup\Big\{|\textstyle\sum_{k=1}^d \alpha_k\zeta_k| : \sum_{k=1}^d \alpha_k^2 \leqslant \delta^2\Big\}\Big\|_{\psi_2}\\ &\leqslant \delta\sqrt{d}\Big\|\Big(\frac{1}{d}\sum_{k=1}^d \zeta_k^2\Big)^{1/2}\Big\|_{\psi_2} \leqslant \delta\sqrt{d}\Big\|\frac{1}{d}\sum_{k=1}^d \zeta_k^2\Big\|_{\psi_1}^{1/2}\\ &\leqslant \delta\sqrt{d}\max_{1\leqslant k\leqslant d}\|\zeta_k^2\|_{\psi_1}^{1/2} \leqslant \delta\sqrt{d}\max_{1\leqslant k\leqslant d}\|\zeta_k\|_{\psi_2} \leqslant C\delta\sqrt{d}.\end{aligned}$$

We also easily get

$$\begin{aligned}&\Big\| \sup_{\lambda\in\Lambda(\delta,\Delta,R)} \int\limits_{\mathbb{T}_{\overline{w}}} P_{L^\perp}(X - \mathbb{E}X)(t)(\lambda - \overline{\lambda})(t)\mu(dt)\Big\|_{\psi_2}\\ &\qquad\leqslant (R + \|\overline{\lambda}\|_1)\Big\|\sup_{t\in\mathbb{T}_{\overline{w}}} P_{L^\perp}(X - \mathbb{E}X)(t)\Big\|_{\psi_2} \leqslant C(R + \|\overline{\lambda}\|_1)\gamma_2(\rho)\end{aligned}$$

and

$$\begin{aligned}\Big\| \sup_{\lambda\in\Lambda(\delta,\Delta,R)} \int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} P_{L^\perp}(X - \mathbb{E}X)(t)(\lambda - \overline{\lambda})(t)\mu(dt)\Big\|_{\psi_2} &\leqslant \Delta\Big\|\sup_{t\in\mathbb{T}} P_{L^\perp}(X - \mathbb{E}X)(t)\Big\|_{\psi_2}\\ &\leqslant C\cdot\Delta S(\mathbb{T}).\end{aligned}$$

It implies that

$$\Big\| \sup_{\lambda\in\Lambda(\delta,\Delta,R)} |f^0_\lambda(X) - f^0_{\overline{\lambda}}(X)|\Big\|_{\psi_2} \leqslant C\Big[\delta\sqrt{d} + (R + \|\overline{\lambda}\|_1)\gamma_2(\rho) + \Delta S(\mathbb{T})\Big]. \tag{7.17}$$

Thus,

$$\begin{aligned}&\Big\| \sup_{\lambda\in\Lambda(\delta,\Delta,R)} |\eta(X,Y)(f^0_\lambda(X) - f^0_{\overline{\lambda}}(X))|\Big\|_{\psi_1}\\ &\qquad\leqslant C\left(\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} + \|\xi\|_{\psi_2}\right)\Big[\delta\sqrt{d} + (R + \|\overline{\lambda}\|_1)\gamma_2(\rho) + \Delta S(\mathbb{T})\Big].\end{aligned}$$

It follows from Adamczak's bound (A.7) and the second statement of Proposition A.1 that, with probability at least $1-e^{-s}$,

$$\alpha_n(\delta,\Delta,R) \leqslant C\Bigg[\mathbb{E}\alpha_n(\delta,\Delta,R) + (\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} + \|\xi\|_{\psi_2})\delta\sqrt{\frac{s}{n}} + (\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} + \|\xi\|_{\psi_2})\Big[\delta\sqrt{d} + (R + \|\overline{\lambda}\|_1)\gamma_2(\rho) + \Delta S(\mathbb{T})\Big]\frac{s\log n}{n}\Bigg].$$

Combining this with the bound of Lemma 7.2, taking the infimum of the right-hand side with respect to $L \subset \mathscr{L}$ and recalling that, according to our assumptions, $s\log n/\sqrt{n}$ is bounded by an absolute constant, we derive the following inequality:

$$\alpha_n(\delta,\Delta,R) \leqslant \beta_n(\delta,\Delta,R;s) := C(\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2})\Big[\delta\sqrt{\frac{s}{n}} \bigvee \nu_n(\delta,\Delta,R)\Big] \bigvee C\nu_n^2(\delta,\Delta,R) \tag{7.18}$$

that holds with probability at least $1-e^{-s}$.

We still need to make the last bound uniform in $\delta \in [\delta_-,\delta_+]$, $\Delta \in [\Delta_-,\Delta_+]$, $R \in [R_-,R_+]$. To this end, define $\delta_{j_1} := \delta_+ 2^{-j_1}$, $\Delta_{j_2} := \Delta_+ 2^{-j_2}$ and $R_{j_3} := R_+ 2^{-j_3}$ for $j_1 = 0,1,\dots,J_1$, $j_2 = 0,1,\dots,J_2$, and $j_3 = 0,1,\dots,J_3$. Using bound (7.18) for each $\delta_{j_1},\Delta_{j_2},R_{j_3}$ with s replaced by $\overline{s} := s + \log((J_1+1)(J_2+1)(J_3+1))$ and applying then the union bound, we get that with probability at least $1-e^{-s}$ $\alpha_n(\delta_{j_1},\Delta_{j_2},R_{j_3}) \leqslant \beta_n(\delta_{j_1},\Delta_{j_2},R_{j_3};\overline{s})$ for all $j_k = 0,\dots,J_k$, $k=1,2,3$. By monotonicity of the functions α_n,β_n in their variables this easily implies that with the same probability

$$\alpha_n(\delta,\Delta,R) \leqslant C\left(\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2}\right)\left[\delta\sqrt{\frac{\overline{s}}{n}} \bigvee \nu(\delta,\Delta,R)\right] \bigvee C\nu_n^2(\delta,\Delta,R).$$

for all $\delta \in [\delta_-,\delta_+]$, $\Delta \in [\Delta_-,\Delta_+]$, $R \in [R_-,R_+]$ and for a large enough constant $C>0$. □

Bounding the last three terms in the right-hand side of (7.11) is easier. Since $\eta(X,Y)$ is a subgaussian random variable (its mean is equal to zero and its ψ_2-norm is finite) and (7.15) holds, we have the following tail bound:

$$|(P_n - P)\eta| = \left|n^{-1}\sum_{j=1}^n \eta(X_j,Y_j)\right| \leqslant C(\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2})\sqrt{\frac{s}{n}} \tag{7.19}$$

with probability at least $1-e^{-s}$ and with some constant $C>0$. Moreover, using the representation $Y - \mathbb{E}Y = f_*(X) - \mathbb{E}f_*(X) + \xi$ and the assumption that $f_*(X) - \mathbb{E}f_*(X) \in \mathscr{L}$, we get

$$\begin{aligned}\|Y - \mathbb{E}Y\|_{\psi_2} &\leqslant \|f_*(X) - \mathbb{E}f_*(X)\|_{\psi_2} + \|\xi\|_{\psi_2}\\ &\leqslant c\|f_*(X) - \mathbb{E}f_*(X)\|_{L_2(\Pi)} + \|\xi\|_{\psi_2}\\ &= c\|f_* - \Pi f_*\|_{L_2(\Pi)} + \|\xi\|_{\psi_2}.\end{aligned}$$

Since $Y - \mathbb{E}Y$ is subgaussian, it is easy to deduce that

$$|\overline{Y}_n - \mathbb{E}Y| = \left| n^{-1} \sum_{j=1}^{n} (Y_j - \mathbb{E}Y) \right| \leqslant C(\|f_* - \Pi f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2}) \sqrt{\frac{s}{n}} \tag{7.20}$$

with probability at least $1 - e^{-s}$ and with some constant $C > 0$. Therefore,

$$\begin{aligned} &\left|\overline{Y}_n - \mathbb{E}Y\right| \cdot |(P_n - P)\eta| \\ &\qquad \leqslant C \left(\|f_* - \Pi f_*\|^2_{L_2(\Pi)} \vee \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} \vee \|\xi\|^2_{\psi_2} \right) \frac{s}{n} \end{aligned} \tag{7.21}$$

with probability at least $1 - 2e^{-s}$. Since $\langle \overline{\lambda}, X_j - \mathbb{E}X_j \rangle$ are i.i.d. subgaussian random variables, their average $\langle \overline{\lambda}, \overline{X}_n - \mathbb{E}X \rangle$ is also subgaussian. This easily yields the bound

$$\begin{aligned} &\left|\langle \overline{\lambda}, \overline{X}_n - \mathbb{E}X \rangle\right| \\ &\qquad \leqslant C \|f^0_{\overline{\lambda}}\|_{L_2(\Pi)} \sqrt{\frac{s}{n}} \leqslant C \left(\|f_* - \Pi f_*\|_{L_2(\Pi)} \vee \|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \right) \sqrt{\frac{s}{n}} \end{aligned} \tag{7.22}$$

that holds with probability at least $1 - e^{-s}$ and with some $C > 0$. Therefore, with probability at least $1 - 2e^{-s}$

$$\begin{aligned} &\left|\langle \overline{\lambda}, \overline{X}_n - \mathbb{E}X \rangle\right| |(P_n - P)\eta| \\ &\qquad \leqslant C \left(\|f_* - \Pi f_*\|^2_{L_2(\Pi)} \vee \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} \vee \|\xi\|^2_{\psi_2} \right) \frac{s}{n}. \end{aligned} \tag{7.23}$$

The proof of the next lemma is a simplified version of the proofs of Lemmas 7.2, 7.3. Together with (7.19) it will be used to control the term

$$|(P_n - P)\eta| \cdot \tau_n \left(\|f^0_\lambda - f^0_{\overline{\lambda}}\|_{L_2(\Pi)}, \int\limits_{\mathbb{T} \setminus \mathbb{T}_{\overline{w}}} |\widehat{\lambda}| d\mu, \|\widehat{\lambda}\|_1 \right)$$

in the right-hand side of (7.11).

Lemma 7.4. — *There exists a constant $C > 0$ such that the following holds. Under the notations of Lemma 7.3, with probability at least $1 - e^{-s}$ and with*

$$\overline{s} := s + \log\big((J_1 + 1)(J_2 + 1)(J_3 + 1)\big)$$

satisfying the condition $\overline{s}\sqrt{\log n} \leqslant \sqrt{n}$,

$$\tau_n(\delta, \Delta, R) = \sup\left\{ \left|(\Pi_n - \Pi)(f^0_\lambda - f^0_{\overline{\lambda}})\right| : \lambda \in \Lambda(\delta, \Delta, R) \right\} \leqslant C \left[\delta \sqrt{\frac{\overline{s}}{n}} \bigvee \nu_n(\delta, \Delta, R) \right]$$

uniformly for all $\delta \in [\delta_-, \delta_+]$, $\Delta \in [\Delta_-, \Delta_+]$, $R \in [R_-, R_+]$.

Step 3. Bounds for $(\Pi - \Pi_n)\big(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big)^2$. — We will need the following representation (that is a consequence of (7.10)):

$$\begin{aligned} &(\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2 \\ &\quad = (\Pi - \Pi_n)(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}})^2 + 2(\Pi - \Pi_n)(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}})(\overline{Y}_n - \mathbb{E}Y) \\ &\qquad + 2(\Pi - \Pi_n)(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}) \langle \overline{\lambda}, \mathbb{E}X - \overline{X}_n \rangle + 2\Big[(\Pi - \Pi_n)(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}})\Big]^2 \\ &\qquad := (\Pi - \Pi_n)(f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}})^2 + \zeta_n(\widehat{\lambda}). \end{aligned} \tag{7.24}$$

Using bounds (7.20), (7.22) and Lemma 7.4, it yields that with probability at least $1-3e^{-s}$ for the same δ, Δ, R

$$\sup\Big\{|\zeta_n(\lambda)| : \lambda \in \Lambda(\delta,\Delta,R)\Big\} \leqslant C\left[\delta\sqrt{\frac{\overline{s}}{n}} \bigvee \nu_n(\delta,\Delta,R)\right]^2 \tag{7.25}$$
$$\bigvee C\left(\|f_* - \Pi f_*\|_{L_2(\Pi)} \vee \|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2}\right)\sqrt{\frac{s}{n}}\left[\delta\sqrt{\frac{\overline{s}}{n}} \bigvee \nu_n(\delta,\Delta,R)\right].$$

Next, we have to estimate

$$\psi_n(\delta,\Delta,R) := \sup_{\lambda\in\Lambda(\delta,\Delta,R)}\Big|(\Pi_n - \Pi)(f_\lambda^0 - f_{\overline{\lambda}}^0)^2\Big|.$$

Lemma 7.5. — *There exists a constant $C>0$ such that the following holds. Under the notations of Lemma 7.3, with probability at least $1-e^{-s}$*

$$\psi_n(\delta,\Delta,R) \leqslant C\delta\left[\delta\sqrt{\frac{\overline{s}}{n}} \vee \nu_n(\delta,\Delta,R)\right] \bigvee C\nu_n^2(\delta,\Delta,R)$$

uniformly for all $\delta\in[\delta_-,\delta_+]$, $\Delta\in[\Delta_-,\Delta_+]$, $R\in[R_-,R_+]$.

Proof. — The proof is based on the inequality due to S. Dirksen and W. Bednorz (see Theorem A.5 in the appendix). To this end, we need to estimate several quantities appearing in that bound. First, note that, since $\big\{f(X):\ f\in\mathscr{F}(\delta,\Delta,R)\big\}$ is a subset of a subgaussian space,

$$\sup_{f\in\mathscr{F}(\delta,\Delta,R)}\|f\|_{\psi_2} \leqslant c\sup_{f\in\mathscr{F}(\delta,\Delta,R)}\|f\|_{L_2(\Pi)} \leqslant c\delta.$$

Together with the bound (7.14) on $\gamma_2(\mathscr{F};\psi_2)$, Theorem A.5 implies that with probability $\geqslant 1-e^{-s}$,

$$\psi_n(\delta,\Delta,R) \leqslant C\delta\left[\delta\sqrt{\frac{d}{n}}\bigvee(R\vee\|\overline{\lambda}\|_1)\frac{\gamma_2(\rho)}{\sqrt{n}}\bigvee\Delta\frac{S(\mathbb{T})}{\sqrt{n}}\right]$$
$$\bigvee C\left[\delta\sqrt{\frac{d}{n}}\bigvee(R\vee\|\overline{\lambda}\|_1)\frac{\gamma_2(\rho)}{\sqrt{n}}\bigvee\Delta\frac{S(\mathbb{T})}{\sqrt{n}}\right]^2 \bigvee C\delta^2\left[\sqrt{\frac{s}{n}}\vee\frac{s}{n}\right].$$

It remains to combine the discretization argument as in the proof of Lemma 7.3 with an application of the union bound to get an estimate for $\psi_n(\delta,\Delta,R)$ that holds uniformly in δ,Δ,R with a high probability. As a result, we get that

$$\psi_n(\delta,\Delta,R) \leqslant C\delta\left[\delta\sqrt{\frac{\overline{s}}{n}} \vee \nu_n(\delta,\Delta,R)\right] \bigvee C\nu_n^2(\delta,\Delta,R)$$

with probability at least $1-e^{-s}$ for all $\delta\in[\delta_-,\delta_+]$, $\Delta\in[\Delta_-,\Delta_+]$, $R\in[R_-,R_+]$ uniformly, and for a large enough constant $C>0$. □

Step 4. Upper bound on $\|\widehat{\lambda}\|_1$.

Lemma 7.6. — *There exist constants* $C, D > 0$ *such that the following holds. For all* $s \geqslant 1$ *and* ε *satisfying the assumptions* $s \log n \leqslant \sqrt{n}$ *and* $\varepsilon \geqslant D\|\xi\|_{\psi_2} S(\mathbb{T})\sqrt{s/n}$*, with probability at least* $1 - 5e^{-s}$*,*

$$\|\widehat{\lambda}\|_1 \leqslant C\left(\frac{q(\varepsilon)}{\varepsilon} + \frac{\sigma_Y^2 s}{n\varepsilon}\right).$$

Proof. — By the definition of $\widehat{\lambda}$, for all $\lambda \in \mathbb{D}, a \in \mathbb{R}$

$$P_n(\ell \bullet f_{\widehat{\lambda},\widehat{a}}) + \varepsilon\|\widehat{\lambda}\|_1 \leqslant P_n(\ell \bullet f_{\lambda,a}) + \varepsilon\|\lambda\|_1. \tag{7.26}$$

We will take $a = a(\lambda) = \mathbb{E}Y - \langle \lambda, \mathbb{E}X\rangle$ everywhere below. Let $\xi(x,y) = y - f_*(x)$ (then, $\xi_j = \xi(X_j, Y_j)$). Since

$$\ell \bullet f_{\widehat{\lambda},\widehat{a}} - \ell \bullet f_{\lambda,a} = (f_{\widehat{\lambda},\widehat{a}} + f_{\lambda,a} - 2f_* - 2\xi)(f_{\widehat{\lambda},\widehat{a}} - f_{\lambda,a}),$$

it is easy to conclude that

$$P_n(\ell \bullet f_{\widehat{\lambda},\widehat{a}}) - P_n(\ell \bullet f_{\lambda,a}) = \|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi_n)} - \|f_{\lambda,a} - f_*\|^2_{L_2(\Pi_n)} - 2P_n\xi(f_{\widehat{\lambda},\widehat{a}} - f_{\lambda,a}).$$

Thus, (7.26) implies that

$$\begin{aligned}\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi_n)} &+ \varepsilon\|\widehat{\lambda}\|_1 \\ &\leqslant \|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)} + (\Pi_n - \Pi)(f_{\lambda,a} - f_*)^2 + 2P_n\left[\xi(f_{\widehat{\lambda},\widehat{a}} - f_{\lambda,a})\right] + \varepsilon\|\lambda\|_1.\end{aligned} \tag{7.27}$$

Using Bernstein's inequality for the random variable with finite $\|\cdot\|_{\psi_1}$-norm (see [26], section A.2) we get that with probability at least $1 - e^{-s}$, for $s \leqslant n$

$$\begin{aligned}|(\Pi_n - \Pi)(f_{\lambda,a} - f_*)^2| &\leqslant C_1\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)}\left[\sqrt{\frac{s}{n}} \bigvee \frac{s}{n}\right] \\ &\leqslant C_1\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)}\sqrt{\frac{s}{n}},\end{aligned} \tag{7.28}$$

where we also used the fact that

$$\|(f_{\lambda,a} - f_*)^2\|_{\psi_1} = \|f_{\lambda,a} - f_*\|^2_{\psi_2} \leqslant c\|f_{\lambda,a} - f_*\|^2_{L_2(\Pi)}.$$

Next, we apply representation (7.10) to term $f_{\widehat{\lambda},\widehat{a}} - f_{\lambda,a}$ in $P_n\xi(f_{\widehat{\lambda},\widehat{a}} - f_{\lambda,a})$ to get the following bound:

$$\begin{aligned}\left|P_n\xi(f_{\widehat{\lambda},\widehat{a}} - f_{\lambda,a})\right| &\leqslant \|\widehat{\lambda} - \lambda\|_1\left\|\frac{1}{n}\sum_{j=1}^n \xi_j(X_j - \mathbb{E}X)\right\|_\infty \\ &\quad + \left|\frac{1}{n}\sum_{j=1}^n \xi_j\right|\left|\overline{Y}_n - \mathbb{E}Y + \langle\widehat{\lambda}, \mathbb{E}X - \overline{X}_n\rangle\right|.\end{aligned} \tag{7.29}$$

To bound the first term in the right-hand side of (7.29), we use a general multiplier inequality (see [41], Lemma 2.9.1):

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{j=1}^n \xi_j(X_j - \mathbb{E}X)\right\|_\infty \leqslant 2\sqrt{2}\|\xi\|_{2,1}\max_{1\leqslant k\leqslant n}\mathbb{E}\left\|\frac{1}{\sqrt{k}}\sum_{j=1}^k \varepsilon_j(X_j - \mathbb{E}X)\right\|_\infty,$$

where $\|\xi\|_{2,1} := \int_0^\infty \sqrt{\mathbb{P}\{\xi \geqslant u\}}\, du$. Note that the process

$$t \longmapsto \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \varepsilon_j (X_j(t) - \mathbb{E}X(t))$$

is subgaussian for every k with respect to the distance d_X. Therefore,

$$\mathbb{E}\left\| \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \varepsilon_j (X_j - \mathbb{E}X) \right\|_\infty \leqslant C_1 S(\mathbb{T}),$$

which yields

$$\mathbb{E}\left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j (X_j - \mathbb{E}X) \right\|_\infty \leqslant C_2 \|\xi\|_{\psi_2} \frac{S(\mathbb{T})}{\sqrt{n}}, \tag{7.30}$$

where we also used the bound $\|\xi\|_{2,1} \leqslant c\|\xi\|_{\psi_2}$.

Adamczak's inequality (A.7) implies that with probability $\geqslant 1 - e^{-s}$

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j (X_j - \mathbb{E}X) \right\|_\infty \\ &\qquad \leqslant C \left[\|\xi\|_{\psi_2} \frac{S(\mathbb{T})}{\sqrt{n}} + \sigma_\xi \sup_{t \in \mathbb{T}} \sqrt{\operatorname{Var}(X(t))} \sqrt{\frac{s}{n}} + \|\xi\|_{\psi_2} S(\mathbb{T}) \frac{s \log n}{n} \right] \\ &\qquad \leqslant C' \|\xi\|_{\psi_2} \left[S(\mathbb{T}) \sqrt{\frac{s}{n}} \bigvee S(\mathbb{T}) \frac{s \log n}{n} \right] \leqslant C \|\xi\|_{\psi_2} S(\mathbb{T}) \sqrt{\frac{s}{n}}, \end{aligned} \tag{7.31}$$

where we also used the bound

$$\sup_{t \in \mathbb{T}} \sqrt{\operatorname{Var}(X(t))} \leqslant \mathbb{E}^{1/2} \sup_{t \in \mathbb{T}} |X(t) - \mathbb{E}X(t)|^2 \leqslant C S(\mathbb{T}). \tag{7.32}$$

To estimate the second term in (7.29), we use inequality (7.20) and also the following tail bounds: with probability at least $1 - e^{-s}$,

$$\left| \frac{1}{n} \sum_{j=1}^{n} \xi_j \right| \leqslant C \|\xi\|_{\psi_2} \sqrt{\frac{s}{n}} \tag{7.33}$$

and, with the same probability,

$$\left\| \overline{X}_n - \mathbb{E}X \right\|_\infty \leqslant C S(\mathbb{T}) \sqrt{\frac{s}{n}}. \tag{7.34}$$

Together with (7.20), these bounds imply that, for some $C > 0$, with probability at least $1 - 3e^{-s}$

$$\begin{aligned} &\left| \frac{1}{n} \sum_{j=1}^{n} \xi_j \right| \left| \overline{Y}_n - \mathbb{E}Y + \langle \widehat{\lambda}, \mathbb{E}X - \overline{X}_n \rangle \right| \\ &\qquad \leqslant C \|\xi\|_{\psi_2} \left[S(\mathbb{T}) \frac{s}{n} \|\widehat{\lambda}\|_1 + \frac{(\|f_* - \Pi f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2}) s}{n} \right]. \end{aligned} \tag{7.35}$$

It follows from bounds (7.27), (7.28), (7.29), (7.31) and (7.35) that with probability at least $1-5e^{-s}$

$$\varepsilon\|\widehat{\lambda}\|_1 \leqslant \|f_{\lambda,a}-f_*\|^2_{L_2(\Pi)} + C_1\|f_{\lambda,a}-f_*\|^2_{L_2(\Pi)}\sqrt{\frac{s}{n}} + \varepsilon\|\lambda\|_1$$
$$+ C'\|\xi\|_{\psi_2}S(\mathbb{T})\sqrt{\frac{s}{n}}\|\widehat{\lambda}-\lambda\|_1 + C\|\xi\|_{\psi_2}S(\mathbb{T})\frac{s}{n}\|\widehat{\lambda}\|_1$$
$$+ C\,\frac{(\|f_*-\Pi f_*\|^2_{L_2(\Pi)}\vee\|\xi\|^2_{\psi_2})s}{n}.$$

If constant D in the assumption on ε is large enough and $s\leqslant n$, it implies that with some $C>0$

$$\frac{\varepsilon}{2}\|\widehat{\lambda}\|_1 \leqslant C\|f_{\lambda,a}-f_*\|^2_{L_2(\Pi)} + 2\varepsilon\|\lambda\|_1 + C\,\frac{(\|f_*-\Pi f_*\|^2_{L_2(\Pi)}\vee\|\xi\|^2_{\psi_2})s}{n},$$

and the result immediately follows. □

Step 5. Putting all the bounds together. — We have all the necessary estimates to complete the proof. Let E denote the event on which the bounds of Lemma 7.3, Lemma 7.4, Lemma 7.5 and also bounds (7.19), (7.20), (7.22), (7.28), (7.31), (7.33) and (7.34) hold. The probability of this event is at least $1-10e^{-s}$. In what follows, we assume that event E occurs. Note that in this case the bound of Lemma 7.6 also holds. Denote

$$\widehat{\delta} := \|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\|_{L_2(\Pi)},\ \widehat{\Delta} := \int\limits_{\mathbb{T}\setminus\mathbb{T}_w} |\widehat{\lambda}|d\mu,\ \widehat{R} := \|\widehat{\lambda}\|_1.$$

Suppose that

$$\widehat{\delta}\in[\delta_-,\delta_+],\quad \widehat{\Delta}\in[\Delta_-,\Delta_+],\quad \widehat{R}\in[R_-,R_+]. \tag{7.36}$$

It follows from bound (7.11), Lemma 7.3 and bounds (7.21) – (7.23) that

$$\frac{1}{C}(P_n-P)\left[\eta\left(f_{\widehat{\lambda},\widehat{a}}-f_{\overline{\lambda},\overline{a}}\right)\right] \tag{7.37}$$
$$\leqslant (\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}\vee\|\xi\|_{\psi_2})\left[\widehat{\delta}\sqrt{\frac{\overline{s}}{n}}\bigvee\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R})\right]$$
$$\bigvee\nu_n^2(\widehat{\delta},\widehat{\Delta},\widehat{R})\bigvee\left(\|f_*-\Pi f_*\|^2_{L_2(\Pi)}\vee\|f_{\overline{\lambda},\overline{a}}-f_*\|^2_{L_2(\Pi)}\vee\|\xi\|^2_{\psi_2}\right)\frac{s}{n}$$
$$\bigvee\left(\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}\vee\|\xi\|_{\psi_2}\right)\sqrt{\frac{s}{n}}\left[\widehat{\delta}\sqrt{\frac{\overline{s}}{n}}\bigvee\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R})\right]$$

for some absolute constant $C>0$. Similarly, Lemma 7.5 and bound (7.25) imply that

$$\frac{1}{C}(\Pi-\Pi_n)(f_{\widehat{\lambda},\widehat{a}}-f_{\overline{\lambda},\overline{a}})^2 \leqslant \widehat{\delta}\left[\widehat{\delta}\sqrt{\frac{\overline{s}}{n}}\bigvee\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R})\right]\bigvee\nu_n^2(\widehat{\delta},\widehat{\Delta},\widehat{R}) \tag{7.38}$$
$$\bigvee\left(\|f_*-\Pi f_*\|_{L_2(\Pi)}\vee\|f_{\overline{\lambda},\overline{a}}-f_*\|_{L_2(\Pi)}\vee\|\xi\|_{\psi_2}\right)\sqrt{\frac{s}{n}}\left[\widehat{\delta}\sqrt{\frac{\overline{s}}{n}}\bigvee\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R})\right]$$
$$\bigvee\left[\widehat{\delta}\sqrt{\frac{\overline{s}}{n}}\bigvee\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R})\right]^2.$$

The last two inequalities will be replaced by simplified upper bounds. To this end, we use elementary inequalities such as $ab \leqslant (a^2/2c) + (cb^2/2)$, for instance:

$$C(\|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2})\widehat{\delta}\sqrt{\frac{\overline{s}}{n}} \leqslant \frac{2C^2}{2}(\|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} \vee \|\xi\|^2_{\psi_2})\frac{\overline{s}}{n} + \frac{1}{8}\widehat{\delta}^2.$$

Also recall that by (7.4), (3.1) and the assumption that $\xi \in \mathscr{L}$,

$$\|f_* - \Pi f_*\|_{L_2(\Pi)} \vee \|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)} \vee \|\xi\|_{\psi_2} \leqslant \sigma_Y.$$

Whenever it is more convenient, we can replace the maximum $\bigvee$ by the sum, or vice versa (with a proper change of constant C), we can drop repetitive terms in the maximum, etc. With such simple transformations, it is easy to get the following bound (with some constant $C > 0$ and under the assumption that $\overline{s} \leqslant n$):

$$(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + (\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2 \leqslant \frac{1}{8}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}} + C\sigma_Y\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) + C\nu_n^2(\widehat{\delta},\widehat{\Delta},\widehat{R}) + C\,\frac{\sigma_Y^2\overline{s}}{n}. \tag{7.39}$$

Note that

$$\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) = \inf_{L\subset\mathscr{L}}\left[\widehat{\delta}\sqrt{\frac{\dim(L)}{n}} \bigvee (\widehat{R} \vee \|\overline{\lambda}\|_1)\frac{\gamma_2(\rho(L))}{\sqrt{n}} \bigvee \widehat{\Delta}\frac{S(\mathbb{T})}{\sqrt{n}}\right] \leqslant (\widehat{R} \vee \|\overline{\lambda}\|_1)\frac{S(\mathbb{T})}{\sqrt{n}},$$

where we used the bounds $\widehat{\Delta} \leqslant \widehat{R}$, $\gamma_2(\rho) \leqslant S(\mathbb{T})$ and computed the expression in the right-hand side of the definition of ν_n for a trivial subspace of zero dimension. Using Lemma 7.6, we get the following bound:

$$\widehat{R} \vee \|\overline{\lambda}\|_1 \leqslant c\left(\frac{q(\varepsilon)}{\varepsilon} \bigvee \frac{\sigma_Y^2 s}{n\varepsilon}\right) \bigvee \|\overline{\lambda}\|_1,$$

which holds with probability at least $1 - e^{-s}$. Therefore,

$$\begin{aligned}\nu_n^2(\widehat{\delta},\widehat{\Delta},\widehat{R}) &\leqslant c\left(\frac{q(\varepsilon)}{\varepsilon} \bigvee \frac{\sigma_Y^2 s}{n\varepsilon}\right)\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) + \|\overline{\lambda}\|_1\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) \\ &\leqslant c\left(\frac{q(\varepsilon)}{\varepsilon} \bigvee \frac{\sigma_Y^2 s}{n\varepsilon}\right)\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) + \frac{1}{2}\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + \frac{1}{2}\nu_n^2(\widehat{\delta},\widehat{\Delta},\widehat{R}),\end{aligned}$$

and inequality (7.39) easily yields

$$\begin{aligned}(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + (\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2 &\leqslant \frac{1}{8}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}} \\ &+ C\sigma_Y\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) + C\left(\frac{q(\varepsilon)}{\varepsilon} \bigvee \frac{\sigma_Y^2 s}{n\varepsilon}\right)\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) \\ &+ C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.\end{aligned} \tag{7.40}$$

Remark 7.1. — Note that under an additional assumption that

$$\mathbb{D} \subset \left\{\lambda \in L_1(\mu) : \ \|\lambda\|_1 \leqslant \frac{C\sigma_Y\sqrt{n}}{S(\mathbb{T})}\right\}$$

(in particular, $\|\overline{\lambda}\|_1 \leqslant C\sigma_Y\sqrt{n}/S(\mathbb{T})$), we have

$$\nu_n^2(\widehat{\delta},\widehat{\Delta},\widehat{R}) \leqslant c\Big(\frac{q(\varepsilon)}{\varepsilon} \bigvee \frac{\sigma_Y^2 s}{n\varepsilon}\Big)\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) + \|\overline{\lambda}\|_1\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R})$$
$$\leqslant c\Big(\frac{q(\varepsilon)}{\varepsilon} \bigvee \frac{\sigma_Y^2 s}{n\varepsilon}\Big)\frac{S(\mathbb{T})}{\sqrt{n}}\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}) + C\sigma_Y\nu_n(\widehat{\delta},\widehat{\Delta},\widehat{R}),$$

so that the term $\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})/n$ disappears from (7.40). In this case, the remainder of the proof yields Theorem 3.3.

Under the assumption (3.5) on ε and the inequality $s \leqslant cn$ (which easily follows from the the main conditions of the theorem), we have

$$\frac{s\sigma_Y}{n\varepsilon}\,\frac{S(\mathbb{T})}{\sqrt{n}} \leqslant c_1$$

with some constant $c_1 > 0$. Also, condition (3.5) and the inequality

$$q(\varepsilon) \leqslant \|f_* - \Pi f_*\|_{L_2(\Pi)}^2 \leqslant \sigma_Y^2$$

imply that $\dfrac{q(\varepsilon)}{\varepsilon}\,\dfrac{S(\mathbb{T})}{\sqrt{n}} \leqslant \mathfrak{c}\sigma_Y$. Hence, with some constant $C > 0$ and for any subspace $L \subset \mathscr{L}_X$ with $\dim(L) = d$ and $\rho(L) = \rho$,

$$(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + (\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2 \tag{7.41}$$
$$\leqslant \frac{1}{8}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}} + C\sigma_Y\bigg[\widehat{\delta}\sqrt{\frac{d}{n}}\bigvee(\widehat{R}\vee\|\overline{\lambda}\|_1)\frac{\gamma_2(\rho)}{\sqrt{n}}\bigvee\widehat{\Delta}\frac{S(\mathbb{T})}{\sqrt{n}}\bigg]$$
$$+ C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.$$

We will now substitute (7.41) in the right-hand side of bound (7.8). Recall that $\mathbb{T}_{\overline{w}} = \{t : \overline{w}(t) \geqslant 1/2\}$. Since, by monotonicity of subdifferentials, we have $(\widehat{w}(t) - w(t))(\widehat{\lambda}(t) - \lambda(t)) \geqslant 0$ for all $t \in \mathbb{T}$, and $\overline{w}, \widehat{w}$ take their values in $[-1,1]$ by definition, we also have that

$$\big\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\big\rangle \geqslant \frac{1}{2}\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}|d\mu. \tag{7.42}$$

Taking this into account, we get

$$\|f_{\widehat{\lambda},\widehat{a}} - f_*\|_{L_2(\Pi)}^2 + \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{2}\big\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\big\rangle + \frac{\varepsilon}{4}\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}|d\mu \tag{7.43}$$
$$\leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|_{L_2(\Pi)}^2 + \varepsilon\big\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\big\rangle + \frac{1}{8}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}}$$
$$+ C\sigma_Y\bigg[\widehat{\delta}\sqrt{\frac{d}{n}}\bigvee(\widehat{R}\vee\|\overline{\lambda}\|_1)\frac{\gamma_2(\rho)}{\sqrt{n}}\bigvee\widehat{\Delta}\frac{S(\mathbb{T})}{\sqrt{n}}\bigg]$$
$$+ C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.$$

Note also that

$$\|\widehat{\lambda}\|_1 \leqslant \|\overline{\lambda}\|_1 + \langle \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle + \langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle,$$

which will be used to control $\widehat{R} \vee \|\overline{\lambda}\|_1 = \|\widehat{\lambda}\|_1 \vee \|\overline{\lambda}\|_1$. Then, bound (7.43) implies the following (with a different value of C):

$$\begin{aligned}(7.44)\quad \|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \frac{\varepsilon}{2}\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle + \frac{\varepsilon}{4}\widehat{\Delta}\\ \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + \varepsilon\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle + \frac{1}{4}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}}\\ + C\,\frac{\sigma_Y^2 d}{n} + C\sigma_Y\|\overline{\lambda}\|_1\frac{\gamma_2(\rho)}{\sqrt{n}} + \widehat{\Delta}C\sigma_Y\frac{S(\mathbb{T})}{\sqrt{n}}\\ + C\sigma_Y\frac{\gamma_2(\rho)}{\sqrt{n}}\big(\langle \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle \vee 0\big) + C\sigma_Y\frac{\gamma_2(\rho)}{\sqrt{n}}\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle\\ + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.\end{aligned}$$

If constant D in the condition on ε is large enough, we have

$$C\sigma_Y\,\frac{\gamma_2(\rho)}{\sqrt{n}} \leqslant C\sigma_Y\,\frac{S(\mathbb{T})}{\sqrt{n}} \leqslant \varepsilon/8,$$

which implies

$$\begin{aligned}(7.45)\quad \|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4}\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle + \frac{\varepsilon}{8}\widehat{\Delta}\\ \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + \frac{9}{8}\varepsilon\big(\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle \vee 0\big) + \frac{1}{4}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}}\\ + C\,\frac{\sigma_Y^2 d}{n} + C\sigma_Y\|\overline{\lambda}\|_1\frac{\gamma_2(\rho)}{\sqrt{n}} + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.\end{aligned}$$

Finally, note that $\widehat{\delta} = \|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\|_{L_2(\Pi)} \leqslant \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|_{L_2(\Pi)}$. Because of this, under the assumption that $C, \overline{s}$ and n are such that $C\sqrt{\frac{\overline{s}}{n}} \leqslant 1/4$, we get from (7.45)

$$\begin{aligned}(7.46)\quad \|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \frac{1}{2}\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4}\langle \widehat{w} - \overline{w}, \widehat{\lambda} - \overline{\lambda}\rangle + \frac{\varepsilon}{8}\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}|d\mu\\ \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + \frac{9}{8}\varepsilon\left(\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle \vee 0\right) + C\,\frac{\sigma_Y^2 d}{n}\\ + C\sigma_Y\|\overline{\lambda}\|_1\frac{\gamma_2(\rho)}{\sqrt{n}} + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.\end{aligned}$$

First, assume that

$$(7.47)\quad \frac{7}{8}\varepsilon\big(\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle \vee 0\big) \geqslant C\,\frac{\sigma_Y^2 d}{n} + C\sigma_Y\|\overline{\lambda}\|_1\frac{\gamma_2(\rho)}{\sqrt{n}} + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.$$

In this case, bound (7.46) implies that

$$\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \frac{1}{2}\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \frac{\varepsilon}{8}\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}|d\mu \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + 2\varepsilon\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle. \tag{7.48}$$

If $\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)}$, the inequality of the theorem trivially holds. Otherwise, (7.48) implies that

$$\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda} - \overline{\lambda}|d\mu \leqslant 16\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle,$$

which means that $\overline{\lambda} - \widehat{\lambda} \in C^{(16)}_{\overline{w}}$ and

$$\langle \overline{w}, \overline{\lambda} - \widehat{\lambda}\rangle \leqslant \mathfrak{a}(\overline{w})\big\|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\big\|_{L_2(\Pi)} \leqslant \mathfrak{a}(\overline{w})\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|_{L_2(\Pi)}$$

by the definition of $\mathfrak{a}(\cdot) = \mathfrak{a}^{(16)}(\cdot)$. Therefore, we have

$$\begin{aligned} &\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \frac{1}{2}\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \frac{\varepsilon}{8}\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}|d\mu \\ &\qquad \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + 2\varepsilon\mathfrak{a}(\overline{w})\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|_{L_2(\Pi)} \\ &\qquad \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + 2\mathfrak{a}^2(\overline{w})\varepsilon^2 + \frac{1}{2}\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)}, \end{aligned} \tag{7.49}$$

which again implies the bound of the theorem.

If condition (7.47) does not hold, then bound (7.46) implies that with some constant $C > 0$

$$\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \frac{\varepsilon}{8}\int\limits_{\mathbb{T}\setminus\mathbb{T}_{\overline{w}}} |\widehat{\lambda}|d\mu \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + C\,\frac{\sigma_Y^2 d}{n} + C\sigma_Y\|\overline{\lambda}\|_1\frac{\gamma_2(\rho)}{\sqrt{n}} + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}, \tag{7.50}$$

which gives the bound of the theorem in this case. To complete the proof, it remains to choose the values of quantities $\delta_-, \delta_+, \Delta_-, \Delta_+$ and R_-, R_+ and to explain how to establish the bound of the theorem in the case when conditions (7.36) do not hold. We will choose the values

$$\delta_+ := C_1\sigma_Y\sqrt{n}, \qquad \delta_- := \frac{C_1\sigma_Y}{\sqrt{n}},$$

$$R_+ = \Delta_+ = \frac{C_1\sigma_Y\sqrt{n}}{S(\mathbb{T})}, \qquad R_- = \Delta_- = \frac{C_1\sigma_Y}{S(\mathbb{T})\sqrt{n}},$$

where C_1 is a large enough constant. Recall that $\overline{s} = s + \log((J_1+1)(J_2+1)(J_3+1))$ and, for our choice of $\delta_-, \delta_+, \Delta_-, \Delta_+, R_-, R_+$ we have $J_1 = J_2 = J_3 = \lfloor \log_2 n\rfloor + 1$. Therefore, $\overline{s} = s + 3\log(\lfloor \log_2 n\rfloor + 2)$. Since $q(\varepsilon) \leqslant \|f_* - \Pi f_*\|^2_{L_2(\Pi)}$, it easily follows from Lemma 7.6 that

$$\|\widehat{\lambda}\|_1 \leqslant C\left[\frac{q(\varepsilon)}{\varepsilon} + \frac{\sigma_Y^2 s}{n\varepsilon}\right] \leqslant 2C_2\frac{\sigma_Y^2}{\varepsilon} \leqslant C_3\frac{\sigma_Y}{S(\mathbb{T})}\sqrt{n} \leqslant R_+,$$

provided that constant C_1 is large enough. It is also easy to see from (7.5) that $\|\overline{\lambda}\|_1 \leqslant R_+$. Thus, $\widehat{R} \vee \|\overline{\lambda}\|_1 \leqslant R_+$, and also $\widehat{\Delta} \leqslant \widehat{R} \leqslant R_+ = \Delta_+$. In addition,

$$\widehat{\delta} = \|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\|_{L_2(\Pi)} = \mathbb{E}^{1/2}\big\langle \widehat{\lambda} - \overline{\lambda}, X - \mathbb{E}X\big\rangle^2 \leqslant \|\widehat{\lambda} - \overline{\lambda}\|_1 \sup_{t\in\mathbb{T}} \sqrt{\mathrm{Var}(X(t))}$$
$$\leqslant C_2(\widehat{R} \vee \|\overline{\lambda}\|_1)S(\mathbb{T}) \leqslant C_3 \frac{\sigma_Y}{S(\mathbb{T})}\sqrt{n}S(\mathbb{T}) \leqslant \delta_+,$$

again, provided that constant C_1 is large enough. Here, we also used the bound (7.32) to estimate $\sup_{t\in\mathbb{T}} \sqrt{\mathrm{Var}(X(t))}$.

Thus, conditions $\widehat{\delta} \leqslant \delta_+$, $\widehat{\Delta} \leqslant \Delta_+$, $\widehat{R} \leqslant R_+$ hold on the event E. If some of the conditions $\widehat{\delta} \geqslant \delta_-$, $\widehat{\Delta} \geqslant \Delta_-$, $\widehat{R} \geqslant R_-$ are violated, we can still use bound (7.44) with quantities $\widehat{\delta}, \widehat{\Delta}, \widehat{R}$ that fall outside the intervals being replaced in its right-hand side by the corresponding upper bound δ_-, Δ_-, R_-. It is easy to check that the inequality of the theorem still holds in this case with a proper constant C.

It now remains to replace s by $s + 3$ (so that $\mathbb{P}(E) \geqslant 1 - 10e^{-s-3} \geqslant 1 - e^{-s}$) to get that the bound of the theorem holds with probability at least $1 - e^{-s}$.

7.4. Proof of Theorem 3.1. — Most of the necessary ingredients have been already developed in the proof of Theorem 3.2. Let $(\overline{\lambda}, \overline{a})$ be a couple that minimizes the right-hand side of bound (3.3). As before, if the infimum is not attained, the proof can be easily modified. We also have that (plugging $(0, \Pi f_*)$ in the right-hand side of (3.3))

$$\|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} \leqslant \|f_* - \Pi f_*\|^2_{L_2(\Pi)}, \qquad \|\overline{\lambda}\|_1 \leqslant \frac{2}{3}\frac{\|f_* - \Pi f_*\|^2_{L_2(\Pi)}}{\varepsilon}.$$

The following inequality is equivalent to (7.8):

$$\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \varepsilon\big\langle \widehat{w}, \widehat{\lambda} - \overline{\lambda}\big\rangle \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)}$$
$$+ 2(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + 2(\Pi - \Pi_n)\big(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big)^2. \tag{7.51}$$

Note that

$$\varepsilon\big\langle \widehat{w}, \widehat{\lambda} - \overline{\lambda}\big\rangle \geqslant \varepsilon\big(\|\widehat{\lambda}\|_1 - \|\overline{\lambda}\|_1\big). \tag{7.52}$$

To bound the empirical processes on the right-hand side of (7.51), we use inequalities (7.37) and (7.38) which imply that (see (7.41) above for details) with some constant $C > 0$ and for any subspace $L \subset \mathscr{L}$ with $\dim(L) = d$ and $\rho(L) = \rho$,

$$(P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + (\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2$$
$$\leqslant \frac{1}{8}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}} + C\sigma_Y\left[\widehat{\delta}\sqrt{\frac{d}{n}}\bigvee(\widehat{R} \vee \|\overline{\lambda}\|_1)\frac{\gamma_2(\rho)}{\sqrt{n}} \bigvee \widehat{\Delta}\frac{S(\mathbb{T})}{\sqrt{n}}\right]$$
$$+ C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n} \tag{7.53}$$

holds on the event E (defined in the proof of Theorem 3.2) of probability at least $\geqslant 1 - 10e^{-s}$, where

$$\widehat{\delta} := \big\|f^0_{\widehat{\lambda}} - f^0_{\overline{\lambda}}\big\|_{L_2(\Pi)} \leqslant \big\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big\|_{L_2(\Pi)}, \quad \widehat{\Delta} := \int\limits_{\mathbb{T}\setminus\mathbb{T}_w} |\widehat{\lambda}|d\mu, \quad \widehat{R} := \|\widehat{\lambda}\|_1$$

and we assume that bounds (7.36) hold. Using the inequalities $\gamma_2(\rho) \leqslant S(\mathbb{T})$, $\widehat{\Delta} \leqslant \widehat{R}$ and choosing L to be the trivial subspace of dimension 0, we get

$$\begin{aligned}(7.54)\quad (P_n - P)\eta(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}) + (\Pi - \Pi_n)(f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}})^2 \\ \leqslant \frac{1}{8}\widehat{\delta}^2 + C\widehat{\delta}^2\sqrt{\frac{\overline{s}}{n}} + C\sigma_Y(\|\overline{\lambda}\|_1 \vee \|\widehat{\lambda}_\varepsilon\|_1)\frac{S(\mathbb{T})}{\sqrt{n}} \\ + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n},\end{aligned}$$

Substituting (7.54) and (7.52) back in (7.51), we get that with some $C > 0$

$$\begin{aligned}(7.55)\quad \|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\|^2_{L_2(\Pi)} + \varepsilon\|\widehat{\lambda}_\varepsilon\|_1 \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} \\ + \varepsilon\|\overline{\lambda}\|_1 + \frac{1}{4}\big\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big\|^2_{L_2(\Pi)} + C\big\|f_{\widehat{\lambda},\widehat{a}} - f_{\overline{\lambda},\overline{a}}\big\|^2_{L_2(\Pi)}\sqrt{\frac{\overline{s}}{n}} \\ + C\sigma_Y(\|\overline{\lambda}\|_1 + \|\widehat{\lambda}_\varepsilon\|_1)\frac{S(\mathbb{T})}{\sqrt{n}} + C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} + C\,\frac{\sigma_Y^2\overline{s}}{n}.\end{aligned}$$

If the constant D in condition (3.2) is large enough, we have $C\sigma_Y S(\mathbb{T})/\sqrt{n} \leqslant \varepsilon/4$ and, since $\|\overline{\lambda}\|_1 \leqslant 2\|f_* - \Pi f_*\|^2_{L_2(\Pi)}/3\varepsilon \leqslant 2\sigma_Y^2/3\varepsilon$,

$$C\,\frac{\|\overline{\lambda}\|_1^2 S^2(\mathbb{T})}{n} \leqslant C\varepsilon\|\overline{\lambda}\|_1\frac{2\|f_* - \Pi f_*\|^2_{L_2(\Pi)}S^2(\mathbb{T})}{3\varepsilon^2 n} \leqslant \frac{\varepsilon}{4}\|\overline{\lambda}\|_1.$$

Moreover, if $C\sqrt{\overline{s}/n} \leqslant 3/4$, (7.55) yields

$$\|f_{\widehat{\lambda},\widehat{a}} - f_*\|^2_{L_2(\Pi)} + \frac{3}{4}\varepsilon\|\widehat{\lambda}_\varepsilon\|_1 \leqslant \|f_{\overline{\lambda},\overline{a}} - f_*\|^2_{L_2(\Pi)} + \frac{3}{2}\varepsilon\|\overline{\lambda}\|_1 + C\,\frac{\sigma_Y^2\overline{s}}{n}.$$

The case when (7.36) does not hold can be handled exactly as at the end of the proof of Theorem 3.2.

7.5. Proof of Proposition 4.2. — For simplicity, we consider the case $d = 1$. Extension to arbitrary dimension follows the same proof pattern.

Note that by (4.2),

$$(7.56)\qquad \mathrm{Var}\left(\textstyle\sum_{j=1}^N u_j X(t_j)\right) = \sum_{1\leqslant j,l\leqslant N} k(t_j - t_l)u_j u_l = \int_{\mathbb{R}}\left|\sum_{j=1}^N e^{it_j z}u_j\right|^2 v(z)dz.$$

Clearly, the function $q(z) = \left|\sum_{j=1}^N e^{it_j z}u_j\right|^2$ is periodic with period $2\pi N$; let $I := \int_{-\pi N}^{\pi N} q(z)v(z)dz$ and $0 \neq m \in \mathbb{Z}$. Together with (4.4), this gives

$$\int\limits_{2\pi mN-\pi N}^{2\pi mN+\pi N} q(z)v(z)dz = \int\limits_{2\pi mN-\pi N}^{2\pi mN+\pi N} q(z)v(z-2\pi mN)\,\frac{v(z)}{v(z-2\pi mN)}dz$$
$$\leqslant \sup_{|y-2\pi mN|\leqslant \pi N}\frac{v(y)}{v(y-2\pi mN)}\int\limits_{-\pi N}^{\pi N} q(z)v(z)dz \leqslant \frac{C}{(|m|-1/2)^2}\cdot I.$$

Hence

$$\int\limits_{-\pi N}^{\pi N}\left|\sum_{j=1}^N e^{it_j z}u_j\right|^2 v(z)dz \leqslant \int_{\mathbb{R}}\left|\sum_{j=1}^N e^{it_j z}u_j\right|^2 v(z)dz$$
$$\leqslant \underbrace{C\sum_{m\in\mathbb{Z}}\frac{1}{(|m|-1/2)^2}}_{C_2}\int\limits_{-\pi N}^{\pi N}\left|\sum_{j=1}^N e^{it_j z}u_j\right|^2 v(z)dz.$$

Recall that our goal is to bound $\|\vec{w}\|_K$ for $\vec{w}\in\partial\|\lambda\|_1$ where $\lambda\in\mathbb{R}^N$. It will be convenient to represent $\vec{w} = (w(t_1),\dots,w(t_N))^T$ as a restriction of a smooth, compactly supported function $w(t),\ t\in\mathbb{R}$ on a grid $\mathscr{G}_N$. Clearly, $w(t)$ is not unique, and we will be interested in the interpolation of "minimal energy", as explained below.

Note that the map

$$\ell_2(\mathbb{Z}) \ni x \longmapsto \widehat{x}_N \in L_2([-\pi N,\pi N],dy),\quad \widehat{x}_N(y) := \frac{1}{\sqrt{2\pi N}}\sum_{j\in\mathbb{Z}} x_j e^{ijy/N}$$

is an isometry. With the convention $u_j = 0,\ j\notin\{1,\dots,N\}$, this implies

$$\langle\vec{w},\vec{u}\rangle_2 = \sum_{j\in\mathbb{Z}} w\Big(\frac{2\pi j}{N}\Big)u_j = \langle\widehat{w}_N,\widehat{u}_N\rangle_{L_2([-\pi N,\pi N],dy)}$$
$$= \Big\langle\frac{\widehat{w}_N}{\sqrt{Nv}},\widehat{u}_N\sqrt{Nv}\Big\rangle_{L_2([-\pi N,\pi N],dy)}$$
$$\leqslant \frac{1}{\sqrt{N}}\Big(\int\limits_{-\pi N}^{\pi N}\frac{|\widehat{w}_N(y)|^2}{v(y)}dy\int\limits_{-\pi N}^{\pi N} N\,|\widehat{u}_N(y)|^2\,v(y)dy\Big)^{1/2}$$
$$\leqslant \frac{c}{\sqrt{N}}\Big(\int\limits_{-\pi N}^{\pi N}(1+y^2)^p\,|\widehat{w}_N(y)|^2\,dy\int\limits_{-\infty}^{\infty} N\,|\widehat{u}_N(y)|^2\,v(y)dy\Big)^{1/2},$$

hence by (7.56)

$$\|\vec{w}\|_K^2 \leqslant \frac{C}{N}\int\limits_{-\pi N}^{\pi N}(1+y^2)^p\,|\widehat{w}_N(y)|^2\,dy. \tag{7.57}$$

Next, define $w_N(y) := \frac{1}{\sqrt{2\pi N}} \int_{-\pi N}^{\pi N} e^{-it\cdot y} \widehat{w}_N(t)dt$. A simple direct computation gives

$$w_N(y) = \sum_{j\in\mathbb{Z}} w\Big(\frac{2\pi j}{N}\Big) \operatorname{sinc}(\pi N(y - j/N)),$$

where $\operatorname{sinc}(x) = (\sin x)/x$. In other words, $w_N(y)$ is the *spectral approximation* of $w(y)$. Define

$$w_N^{(p)}(y) := \frac{e^{-i\pi p/2}}{\sqrt{2\pi N}} \int\limits_{-\pi N}^{\pi N} e^{-it\cdot y}\, t^p \widehat{w}_N(t)dt$$

(note that for $p \in \mathbb{N}$ this is just the p'th derivative of $w_N(y)$). By the isometric property of Fourier transform, this gives

$$\frac{1}{N} \int\limits_{-\pi N}^{\pi N} t^{2p} |\widehat{w}_N(t)|^2 dt = C \int_{\mathbb{R}} |w_N^{(p)}(t)|^2 dt,$$

hence (7.57), together with the triangle inequality, implies

$$\|\vec{w}\|_K^2 \leqslant C_1 \|w_N\|^2_{\mathbb{W}^{2,p}(\mathbb{R})} \leqslant 2C_1 \big(\|w\|^2_{\mathbb{W}^{2,p}(\mathbb{R})} + \|w - w_N\|^2_{\mathbb{W}^{2,p}(\mathbb{R})}\big). \tag{7.58}$$

We will need the following important fact (it will be used for $m = p$):

THEOREM 7.1 ([3], Theorem 5.4). — *Assume that $w \in \mathbb{W}^{2,p}(\mathbb{R})$ and $m \leqslant p$. Then*

$$\big\|w_N^{(m)} - w^{(m)}\big\|_{L_2(\mathbb{R})} \leqslant C(p,m) N^{-(p-m)} \|w\|_{\mathbb{W}^{2,p}(\mathbb{R})},$$

where $C(p,m)$ is independent of w and N.

Together with (7.58) this implies the claim of the proposition.

7.6. PROOFS OF THEOREMS 6.1, 6.2 AND 6.3. — Recall that for every $\lambda \in \mathbb{D}_\Delta$, for each $j \in J_\lambda$ we either have that $\lambda(t) \geqslant 0$ for all $t \in \mathbb{T}_j$ (in this case, set $\sigma_j = +1$), or $\lambda(t) \leqslant 0$ for all $t \in \mathbb{T}_j$ (set $\sigma_j = -1$). Clearly, the function

$$w := \sum_{j\in J_\lambda} \sigma_j I_{\mathbb{T}_j}$$

satisfies the conditions $|w(t)| \leqslant 1$, $t \in \mathbb{T}$ and $w(t) := \operatorname{sign}(\lambda(t))$ if $\lambda(t) \neq 0$. Therefore, $w \in \partial\|\lambda\|_1$. In what follows, we will use such w as a version of subgradient of $\lambda \in \mathbb{D}_\Delta$.

We will start by providing upper bounds on RKHS-norms of w_j, $j \in J_\lambda$.

LEMMA 7.7. — *Suppose that, for each $j = 1, \dots, N$, the set $\mathbb{T}_j$ is contained in a ball $B(t_j; r)$ with some center $t_j \in \mathbb{R}^d$ and of radius r. Suppose also that*

$$v_j(t) \geqslant \frac{c}{(1 + |t|^2)^p}, \quad t \in \mathbb{R}^d,\ j = 1, \dots, N. \tag{7.59}$$

Then

$$\|w_j\|_{K_j} \leqslant C r^{d/2}(1 + r^{-p}), \quad j \in J_\lambda.$$

Proof. — Note that for arbitrary functions w_j defined on $\mathbb{T}_j$,

$$\|w_j\|_{K_j} \leqslant C \inf_{\widetilde{w}_j \in \Omega(w_j)} \|\widetilde{w}_j\|_{\mathbb{W}^{2,p}(\mathbb{R}^d)}, \tag{7.60}$$

where $\Omega(w_j)$ is the set of all extensions of w_j onto $\mathbb{R}^d$ (see Proposition 4.1). To control the RKHS-norms of w_j, consider an arbitrary nonnegative C^∞-function ϕ supported in the unit ball $\{t : |t| \leqslant 1\}$ such that $\int_{\mathbb{R}^d} \phi(t)dt = 1$. Denote $\phi_r(t) := r^{-d}\phi(t/r)$ and let

$$\widetilde{w}_j(t) := \sigma_j \int_{\mathbb{R}^d} \phi_r(t-s) I_{B(t_j,2r)}(s)ds = \sigma_j(\phi_r * I_{B(t_j,2r)})(t), \quad t \in \mathbb{R}^d.$$

It is immediate that for $t \in \mathbb{T}_j$, $\widetilde{w}_j(t) = w_j(t)$, so, $\widetilde{w}_j \in \Omega(w_j)$. Thus, we have

$$\|w_j\|_{K_j} \leqslant C\|\widetilde{w}_j\|_{\mathbb{W}^{p,2}(\mathbb{R}^d)} = C\|\phi_r * I_{B(t_j,2r)}\|_{\mathbb{W}^{p,2}(\mathbb{R}^d)} \leqslant C'\left\|(1+|t|^2)^{p/2}\widehat{\phi_r}\widehat{I_{B(t_j,2r)}}\right\|_{L_2(\mathbb{R}^d)}.$$

Since, by an easy computation,

$$\left\|\widehat{I_{B(t_j,2r)}}\right\|_{L_\infty} \leqslant \mu(B(t_j,2r)) \leqslant c' r^d$$

and

$$\left\|(1+|t|^2)^{p/2}\widehat{\phi_r}\right\|_{L_2(\mathbb{R}^d)} \leqslant C_1 r^{-d/2}(1+r^{-p}),$$

we conclude that $\|w_j\|_{K_j} \leqslant Cr^{d/2}(1+r^{-p})$, $j \in J_\lambda$ for some constant C. □

The next lemma provides bounds on $S(\mathbb{T}, d_X)$ and $\gamma_2(\delta; d_X)$.

Lemma 7.8. — *Let $\mathbb{T}$ be a bounded measurable subset of $\mathbb{R}^d$ and let $X(t)$, $t \in \mathbb{R}^d$ be a centered subgaussian stationary random field with spectral measure ν and spectral density v. Suppose that bound* (6.1) *holds for some $R \geqslant 1$. Suppose also that*

$$v(t) \leqslant \frac{B}{(1+|t|^2)^p} \tag{7.61}$$

for some $p > d/2, B > 0$. Then, there exists a constant $C > 0$ depending on d, p, B such that

$$S(\mathbb{T}; d_X) \leqslant C\sqrt{\log N \vee |\log r|}$$

and

$$\gamma_2(\mathbb{T}; d_X; \delta) \leqslant C\delta\sqrt{\log \frac{CR^{(p-d/2)\wedge 1}}{\delta} \bigvee \log N}.$$

Proof. — By the spectral representation of covariance, for all $t_1, t_2 \in \mathbb{R}^d$ and for $A > 0$,

$$\begin{aligned} d_X^2(t_1,t_2) = \mathrm{Var}(X(t_1)-X(t_2)) &= \int_{\mathbb{R}^d} \left|e^{i\langle t_1-t_2,s\rangle}-1\right|^2 v(s)ds \\ &\leqslant B|t_1-t_2|^2 \int_{|s|\leqslant A} \frac{|s|^2}{(1+|s|^2)^p}ds + B\int_{|s|>A} \frac{1}{(1+|s|^2)^p}ds. \end{aligned} \tag{7.62}$$

If $2p > d+2$, we take $A = \infty$ and get $d_X^2(t_1,t_2) \leqslant C'|t_1-t_2|^2$ for some $C' > 0$ that depends on p and d. If $2p = d+2$, a simple computation of the integrals in the right-hand side of (7.62) and minimizing the resulting bound with respect to A yields

$$d_X^2(t_1,t_2) \leqslant C'|t_1-t_2|^2\Big(\log\Big(\frac{1}{|t_1-t_2|}\Big) \bigvee 1\Big).$$

Finally, if $2p < d+2$, then a similar argument yields the bound

$$d_X^2(t_1,t_2) \leqslant C'|t_1-t_2|^{2p-d}.$$

Using bound (6.1), it is easy to show that in each of these three cases we have

$$\log N(\mathbb{T}; d_X; \varepsilon) \leqslant C\Big(\log \frac{CR^{(p-d/2)\wedge 1}}{\varepsilon} \vee \log N\Big), \quad \varepsilon \in (0, CR^{(p-d/2)\wedge 1}).$$

The bound on $\gamma_2(\mathbb{T}; d_X; \delta)$ now follows by controlling the generic chaining complexity in terms of Dudley's entropy integral. We also have that, under condition (7.61), the diameter $D(\mathbb{T}; d_X)$ admits the following estimate:

$$D^2(\mathbb{T}; d_X) \leqslant 2\sup_{t\in\mathbb{T}} \mathrm{Var}(X(t)) = 2\int_{\mathbb{R}^d} v(s)ds \leqslant C'',$$

where C'' is a constant depending on d, p, B. The bound on $S(\mathbb{T}; d_X)$ now follows from the bound on $\gamma_2(\mathbb{T}; d_X; \delta)$ by substituting $\delta = \sqrt{C''}$. □

We will also need a bound on Kolmogorov's width of the set of random variables $X_{\mathbb{T}}$ given in the following lemma.

LEMMA 7.9. — *Let $\mathbb{T}$ be a bounded measurable subset of $\mathbb{R}^d$ satisfying condition* (6.1) *and let $X(t)$, $t \in \mathbb{R}^d$ be a centered subgaussian stationary random field with spectral measure ν and spectral density v. Suppose that*

$$v(t) \leqslant \frac{B}{(1+|t|^2)^p} \tag{7.63}$$

for some $p > d/2$, $B > 0$. Then, there exists a constant $C > 0$ depending only on d, p and B such that for all $m \geqslant CN$

$$\rho_m(X_{\mathbb{T}}) \leqslant C\Big(\frac{R}{m^{1/d}}\Big)^{p-d/2}. \tag{7.64}$$

REMARK 7.2. — Note that bound (7.64) is sharp (up to a constant). A matching lower bound can be proved via an argument based on replacing the spectral density v by a smaller density $\overline{v}$ that is constant in a cube of a proper size and zero outside of the cube. For such a smaller density, it is possible to find a grid of points of sufficiently large cardinality m such that the values of the stationary random field with spectral density $\overline{v}$ at the points of the grid are uncorrelated. Bounding the corresponding Kolmogorov's width from below can be now reduced to bounding Kolmogorov numbers of the embedding of ℓ_1^m into ℓ_2^m, see Gluskin [18] for the solution of the last problem. The authors are very thankful to M. Lifshits for pointing out this beautiful argument.

Proof. — We will construct an approximation of the set of random variables $X_{\mathbb{T}} = \{X(t) : t \in \mathbb{T}\}$ by a finite dimensional subspace of subgaussian random variables $L \subset \mathscr{L}_X$. Since X is a stationary random field, the following spectral representation holds

$$X(t) = \int_{\mathbb{R}^d} e^{i\langle t,s\rangle} Z(ds),$$

where Z is an orthogonal random measure such that

$$\mathbb{E}Z(A)\overline{Z(B)} = \nu(A \cap B), A, B \in \mathscr{B}_{\mathbb{R}^d}.$$

By a standard isometry argument, to approximate the random variable $X(t)$ in the space $L_2(\mathbb{P})$, it is enough to approximate the function $e^{2\pi i\langle t,\cdot\rangle}$ in the space $L_2(\mathbb{R}^d, \nu)$. For $\delta \leqslant r$, consider a δ-net of the set $\mathbb{T}$ that consists of $N' \leqslant (R/\delta)^d$ points $\tau_1, \dots, \tau_{N'}$. To construct an approximation of the exponential function, we will use Taylor expansion of order l in a δ-neighborhood of each of the points τ_k. We use the the following standard bound on the remainder of Taylor expansion:

$$\left|e^{i\langle h,s\rangle} - Q_l(h;s)\right| \leqslant \frac{|h|^l|s|^l}{l!}, \quad Q_l(h;s) := \sum_{j=0}^{l-1} \frac{i^j\langle h,s\rangle^j}{j!}. \tag{7.65}$$

For $t \in B(\tau_k;\delta)$,

$$e^{i\langle t,s\rangle} = e^{i\langle \tau_k,s\rangle}e^{i\langle t-\tau_k,s\rangle} = e^{i\langle \tau_k,s\rangle}Q_l(t-\tau_k;s) + e^{i\langle \tau_k,s\rangle}(e^{i\langle t-\tau_k,s\rangle} - Q_l(t-\tau_k;s)).$$

Denote (for some $A > 0$ to be chosen later)

$$\zeta_l^{(k)}(h) := \mathrm{Re}\left(\int_{\mathbb{R}^d} e^{i\langle \tau_k,s\rangle}Q_l(h;s)I(|s| \leqslant A\delta^{-1})Z(ds)\right).$$

By spectral isometry (using the fact that X is real valued), we get that for all $k = 1, \dots, N'$ and all $t \in B(\tau_k;\delta)$ (thus, for all $t \in \mathbb{T}$)

$$\begin{aligned}\mathbb{E}\left|X(t) - \zeta_l^{(k)}(t-\tau_k)\right|^2 &\leqslant \mathbb{E}\left|X(t) - \int_{\mathbb{R}^d} e^{i\langle \tau_k,s\rangle}Q_l(h;s)I(|s| \leqslant A\delta^{-1})Z(ds)\right|^2 \\ &\leqslant \int\limits_{|s|\leqslant A\delta^{-1}} \left|e^{i\langle t,s\rangle} - e^{i\langle \tau_k,s\rangle}Q_l(t-\tau_k;s)\right|^2 v(s)ds + \int\limits_{|s|>A\delta^{-1}} v(s)ds.\end{aligned} \tag{7.66}$$

Under condition (7.63) and the assumption $p > d/2$, using (7.65), we get that with some constant $C > 0$ depending only on B, d and for all $k = 1, \dots, N'$ and $l \geqslant (2p-d) \vee 1$

$$\begin{aligned}\int\limits_{|s|\leqslant A\delta^{-1}} \left|e^{i\langle t,s\rangle} - e^{i\langle \tau_k,s\rangle}Q_l(t-\tau_k;s)\right|^2 v(s)ds &\leqslant \frac{\delta^{2l}}{(l!)^2}\int\limits_{|s|\leqslant A\delta^{-1}} |s|^{2l}v(s)ds \\ &\leqslant B\frac{\delta^{2l}}{(l!)^2}\int\limits_{|s|\leqslant A\delta^{-1}} \frac{|s|^{2l}}{(1+|s|^2)^p}ds \leqslant C\,\frac{\delta^{2p-d}}{A^{2p-d-2l}(2l-2p+d)(l!)^2}.\end{aligned}$$

We also have

$$\int\limits_{|s|>A\delta^{-1}} v(s)ds \leqslant B \int\limits_{|s|>A\delta^{-1}} \frac{1}{(1+|s|^2)^p} ds \leqslant C\,\frac{\delta^{2p-d}}{(2p-d)A^{2p-d}}.$$

We will now set

$$A := A_l := (2l)^{1/(2l)}(l!)^{1/l}.$$

Then, (7.66) easily implies that with some constant C depending only on p and d

$$\mathbb{E}\big|X(t) - \zeta_l^{(k)}(t-\tau_k)\big|^2 \leqslant C\Big(\frac{\delta}{A_l}\Big)^{2p-d}.$$

Using Stirling's approximation, it is easy to see that $A_l \geqslant l/e$, implying that

$$\mathbb{E}\big|X(t) - \zeta_l^{(k)}(t-\tau_k)\big|^2 \leqslant C\Big(\frac{\delta}{l}\Big)^{2p-d}. \tag{7.67}$$

Note that $Q_l(h;\cdot)$ is polynomial of degree $l-1$ of d variables, hence, the family of functions

$$\Big\{e^{i\langle\tau_k,\cdot\rangle} Q_l(h;\cdot) I(|\cdot| \leqslant A\delta^{-1}) : h \in \mathbb{R}^d\Big\}$$

belongs to a (complex) linear space of dimension $\binom{l-1+d}{d} \leqslant (l+d-1)^d$. This immediately implies that the family of random variables $\{\zeta_l^{(k)}(h) : h \in \mathbb{R}^d\}$ belongs to a linear subspace of $\mathscr{L}_X$ whose dimension is at most $2(l+d-1)^d$. Therefore, $\big\{\zeta_l^{(k)}(t-\tau_k) : t \in B(\tau_k;\delta), k = 1,\dots,N'\big\}$ belongs to a subspace of $\mathscr{L}$ of dimension $\leqslant 2(l+d-1)^d N' \leqslant 2(l+d-1)^d (R/\delta)^d$. Let $m \geqslant 2(l+d-1)^d$ and let

$$\delta = 2^{1/d}(l+d-1)\frac{R}{m^{1/d}}.$$

Assuming that $m \geqslant C_1 N$, where $C_1 := 2(l+d-1)^d\kappa^d$, we have $\delta \leqslant r$. Then $2(l+d-1)^d(R/\delta)^d = m$ and it follows from (7.67) that

$$\rho_m(X_{\mathbb{T}}) \leqslant C\Big(\frac{l+d-1}{l}\Big)^{p-d/2} \frac{R^{p-d/2}}{m^{p/d-1/2}},$$

with some constant C depending on B, d, p. The claim of the lemma follows by substituting the smallest $l \geqslant (2p-d) \vee d$. □

We will now provide an upper bound on the "approximate dimension" $d(w;\lambda)$ needed to complete the proof of the theorem. To this end, recall that we assume that for all $j = 1,\dots,N$ the set $\mathbb{T}_j$ belongs to a ball of radius $r \geqslant N^{-1/d}$ and $R = \kappa N^{1/d} r$, $R \geqslant 2$. Also, for an oracle λ, $R(\lambda) = \kappa(N(\lambda))^{1/d} r$, so, we have $r \leqslant R(\lambda) \leqslant R$. In what follows, C, C', etc are constants depending on B, d, p. First, let us upper bound $\gamma_2(\rho_m(w)) = \gamma_2(\rho_m(X_{T_w}))$. Using Lemmas 7.8 and 7.9, we get that for all $m \geqslant C_1 N(\lambda)$

$$\gamma_2(\rho_m(w)) \leqslant C\,\frac{(R(\lambda))^{p-d/2}}{m^{p/d-1/2}} \sqrt{\log\Big(\frac{CR^{p-d/2}m^{p/d-1/2}}{(R(\lambda))^{p-d/2}}\Big) \bigvee \log N}. \tag{7.68}$$

Since $R/R(\lambda) = \kappa N^{1/d} r/\kappa (N(\lambda))^{1/2} r \leqslant N^{1/d}$, it is easy to conclude that

$$\gamma_2(\rho_m(w)) \leqslant C \frac{(R(\lambda))^{p-d/2}}{m^{p/d-1/2}} \sqrt{\log m} \bigvee C \frac{(R(\lambda))^{p-d/2}}{m^{p/d-1/2}} \sqrt{\log N}.$$

To provide an upper bound on $d(w,\lambda)$, we first find the smallest m satisfying the inequality

$$\frac{\sigma_Y^2 m}{n} \geqslant C \frac{\|\lambda\|_1}{\sqrt{n}} \frac{(R(\lambda))^{p-d/2}}{m^{p/d-1/2}} \sqrt{\log m} \bigvee C \frac{\|\lambda\|_1}{\sqrt{n}} \frac{(R(\lambda))^{p-d/2}}{m^{p/d-1/2}} \sqrt{\log N}.$$

This is equivalent to the bound

$$m \geqslant C \frac{n^{d/(2p+d)} \|\lambda\|_1^{2d/(2p+d)} R(\lambda)^{d(2p-d)/(2p+d)}}{\sigma_Y^{4d/(2p+d)}} \cdot \Bigl((\log m)^{d/(2p+d)} \bigvee (\log N)^{d/(2p+d)} \Bigr) \tag{7.69}$$

Note that in the oracle inequality of Theorem 3.2, it is enough to restrict oracles λ to the ball

$$\|\lambda\|_1 \leqslant C' \|f_* - \Pi f_*\|_{L_2(\Pi)} n^{1/2}$$

for some constant $C' > 0$ (see bound (7.5) in the proof of this theorem). Recall that also

$$N^{-1/d} \leqslant r \leqslant R(\lambda) \leqslant R = \kappa N^{1/d} r.$$

Therefore, bound (7.69) easily implies that

$$m \geqslant C \frac{(n\|\lambda\|_1^2)^{d/(2p+d)} R(\lambda)^{d(2p-d)/(2p+d)}}{\sigma_Y^{4d/(2p+d)}} \cdot \Bigl(\log N \vee \log n \vee |\log \sigma_Y| \vee |\log r| \Bigr)^{d/(2p+d)}. \tag{7.70}$$

It easily follows from the definition of $d(w,\lambda)$ that either we have $d(w,\lambda) \leqslant C_1 N(\lambda)$, or $d(w,\lambda) \leqslant m$ for any m satisfying (7.70). Therefore, with some constant $C > 0$

$$d(w;\lambda) \leqslant CN(\lambda) \bigvee C \frac{(n\|\lambda\|_1^2)^{d/(2p+d)} R(\lambda)^{d(2p-d)/(2p+d)}}{\sigma_Y^{4d/(2p+d)}} \Bigl(\log N \vee \log n \vee |\log \sigma_Y| \vee |\log r| \Bigr)^{d/(2p+d)}. \tag{7.71}$$

To complete the proof, it is enough to substitute this bound on $d(w;\lambda)$ in the oracle inequality of Theorem 3.2. Bounds (5.1) and Lemma 7.7 should be used to control the alignment coefficient $\mathfrak{a}(w)$.

As to the proof of Theorem 6.2, the main difference is in the bounds on the alignment coefficient $\mathfrak{a}(w)$. For a given oracle $\lambda \in \mathbb{D}_r$ and a covering $B(t_1;r), \dots, B(t_{N(\lambda)};r)$ of $\operatorname{supp}(\lambda)$, let σ_j be the sign of λ on $B(t_j;r) \cap \operatorname{supp}(\lambda)$ and $\widetilde{w}_j := \sigma_j (\phi_r * I_{B(t_j;2r)})$, $j = 1, \dots, N(\lambda)$ (see the notations of the proof of Lemma 7.7). It is easy to see that $\sum_{j \in J_\lambda} \widetilde{w}_j$ is an extension of a subgradient $w \in \partial \|\lambda\|_1$. Thus, by

Proposition 4.1,

$$\mathfrak{a}^2(w) \leqslant \|w\|_K^2 \leqslant \Big\| \sum_{j=1}^{N(\lambda)} \widetilde{w}_j \Big\|_{\mathbb{W}^{2,p}}^2 .$$

Since functions $\widetilde{w}_j$ have disjoint support, we can further bound this using Proposition A.2 and of Lemma 7.7 as

$$\mathfrak{a}^2(w) \leqslant C \Big[\sum_{j=1}^{N(\lambda)} \|\widetilde{w}_j\|_{\mathbb{W}^{2,p}}^2 + \frac{1}{r^{2\alpha}} \sum_{j=1}^{N(\lambda)} \|\widetilde{w}_j\|_{\mathbb{W}^{2,\lfloor p \rfloor}}^2 \Big] \leqslant C \left(r^d + r^{d-2p} \right) N(\lambda).$$

Finally, to prove the result of Theorem 6.3, we need to bound the alignment coefficient $\mathfrak{a}(w)$ as follows. Let ϕ be an arbitrary nonnegative C^∞-function supported in the unit ball $\{t : |t| \leqslant 1\}$ such that for all $t \in \mathbb{R}^d$, $\phi(t) \leqslant \phi(0) = 1$. Given $\lambda \in \overline{\mathbb{D}}$ and $r \leqslant \delta(\lambda)$, let

$$\widetilde{w}_j = \operatorname{sign}(\lambda_j) \phi\Big(\frac{t - t_j}{r} \Big), \quad j \in J(\lambda).$$

Clearly, restriction of $w = \sum_{j \in J(\lambda)} \widetilde{w}_j$ to the grid $\mathscr{G}_N$ is an element of $\partial \|\lambda\|_1$. By Proposition 4.2, we have

$$\mathfrak{a}^2(w) \leqslant \|w\|_K^2 \leqslant \Big\| \sum_{j=1}^{N(\lambda)} \widetilde{w}_j \Big\|_{\mathbb{W}^{2,p}}^2 .$$

By Proposition A.2 and a simple computation is spirit of Lemma 7.7

$$\mathfrak{a}^2(w) \leqslant C \left(r^d + r^{d-2p} \right) N(\lambda).$$

It is easy to see that $\gamma_2(\rho_m(w))$ and $d(w;\lambda)$ can be bounded above by their "continuous" counterparts for $\mathbb{T} = [0, 2\pi]^d$, in particular, inequalities (7.68) and (7.71) hold. To complete the proof, it is enough to substitute bounds on $d(w;\lambda)$ and $\mathfrak{a}^2(w)$ in the oracle inequality of Theorem 3.2 and optimize the resulting expression with respect to r. Choose $r_*(\lambda)$ as $r_*(\lambda) = \min(\widetilde{r}, \delta(\lambda))$ with $\widetilde{r}$ defined as

$$\widetilde{r}^{2p-d} = \Big(\frac{N(\lambda)^2}{n} \Big)^{d/(2p+2d)} \frac{\sigma_Y^{2d/(p+d)} s^{(2p+d)/(2p+2d)}}{(L\|\lambda\|_1^2)^{d/(2p+2d)}},$$

where $L = \log n \vee \log N \vee |\log \sigma_Y|$. The claim now follows from simple algebra.

Appendix A. Technical background and remaining proofs

A.1. Existence of solutions of optimization problems. — We provide below sufficient conditions for existence of solutions to the problems (2.3) and (1.2).

Theorem A.1. — *Let $\mathbb{D}$ be a convex, weakly compact subset of $L_1(\mu)$. Then*

(1) *$F(\lambda, a)$, $F_n(\lambda, a)$ are weakly lower semicontinuous;*
(2) *Solutions to problems (2.3) and (1.2), denoted by λ_ε and $\widehat{\lambda}_\varepsilon$, exist.*

Proof. — We prove the statement for $F(\lambda)$, and the result for $F_n(\lambda)$ follows similarly. The functional $\lambda \mapsto \|\lambda\|_1$ is continuous. Assume $\|\lambda_k - \lambda_0\|_1 \to 0$. Using Hölder's inequality, we get

$$
\begin{aligned}
&P(\ell_\bullet f_{\lambda_k,a(\lambda_k)}) - P(\ell_\bullet f_{\lambda_0,a(\lambda_0)}) = \mathbb{E}(Y - f_{\lambda_k,a(\lambda_k)}(X))^2 - \mathbb{E}(Y - f_{\lambda_0,a(\lambda_0)}(X))^2 \\
&= \mathbb{E}\Big[\big(2Y - f_{\lambda_k,a(\lambda_k)}(X) - f_{\lambda_0,a(\lambda_0)}(X)\big) \int\limits_{\mathbb{T}} (\lambda_0 - \lambda_k)(X - \mathbb{E}X) d\mu \Big] \\
&\leqslant \mathbb{E}^{1/2}\bigg(\int\limits_{\mathbb{T}} (\lambda_0 - \lambda_k)(X - \mathbb{E}X) d\mu \bigg)^2 \mathbb{E}^{1/2} \big(2Y - (f_{\lambda_k,a(\lambda_k)} + f_{\lambda_0,a(\lambda_0)})(X)\big)^2 \\
&\leqslant \|\lambda_k - \lambda_0\|_1 \mathbb{E}^{1/2} \|X - \mathbb{E}X\|_\infty^2 \Big(2\sqrt{\mathrm{Var}(Y)} + \|\lambda_k + \lambda_0\|_1 \mathbb{E}^{1/2} \|X - \mathbb{E}X\|_\infty^2 \Big) \longrightarrow 0,
\end{aligned}
$$

where in the last step we used the fact that

$$
\bigg| \int\limits_{\mathbb{T}} \lambda(t)(X(t) - \mathbb{E}X(t)) \mu(dt) \bigg| \leqslant \|\lambda\|_1 \sup_{t \in \mathbb{T}} |X(t) - \mathbb{E}X(t)|.
$$

Thus, $F(\lambda)$ is continuous, hence it is lower semi-continuos. In turn, this is equivalent to the fact that the level sets $\mathscr{L}_t = \{\lambda : F(\lambda) \leqslant t\}$ are closed. Moreover, they are convex since F is. Mazur's theorem (see [29], Theorem 2.1) implies that they are also closed in weak topology, so F is weakly lower semi-continuous.

Now it is easy to show existence of solutions. Given a minimizing sequence $\{\lambda_k\} \subset \mathbb{D}$, we can extract a weakly convergent subsequence

$$
\lambda_{k_l} \xrightarrow{\sigma} \lambda_\infty.
$$

It remains to note that by weak compactness and lower semi-continuity $\lambda_\infty \in \mathbb{D}$ and $-\infty < F(\lambda_\infty) \leqslant \liminf_{l \to \infty} F(\lambda_{k_l})$, which means that λ_∞ is the solution. □

When $\mathbb{T}$ is finite, then one can clearly take $\mathbb{D} = L_1(\mathbb{T}, \mu) \subseteq \mathbb{R}^{|T|}$, and Theorem A.1 is not needed to prove existence of $\widehat{\lambda}_\varepsilon$. However, in general the unit ball in $L_1(\mathbb{T}, \mu)$ is not weakly compact, so one way to proceed is to choose $\mathbb{D}$ to be uniformly integrable (which implies weak compactness, see Theorem 4.7.18 in [7]). A possible choice is

$$
\mathbb{D} = \bigg\{ \lambda : \ \bigg| \int\limits_{\mathbb{T}} \max\left(|\lambda(t)| \log |\lambda(t)|, 0\right) d\mu(t) \bigg| \leqslant L \bigg\} \text{ for some } L > 0.
$$

A.2. Orlicz norms. — Let $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing function with $\psi(0) = 0$.

Definition A.1. — The Orlicz norm of a random variable η on a probability space $(\Omega, \Sigma, \mathbb{P})$ is defined via

$$
\|\eta\|_\psi := \inf \{C > 0 : \ \mathbb{E}\psi\left(|\eta|/C\right) \leqslant 1\}
$$

By $\|\cdot\|_{\psi_1},\ \|\cdot\|_{\psi_2}$ we denote the Orlicz norms for $\psi_1(x) := e^x - 1$ and $\psi_2(x) := e^{x^2} - 1$, respectively; the following inequalities are elementary:

$$\|\eta\|_{\psi_1} \leqslant \sqrt{\log 2}\|\eta\|_{\psi_2}, \tag{A.1}$$

$$\|\eta^2\|_{\psi_1} = \|\eta\|^2_{\psi_2}, \tag{A.2}$$

$$\|\xi\eta\|_{\psi_1} \leqslant \|\xi\|_{\psi_2}\|\eta\|_{\psi_2}. \tag{A.3}$$

It is easy to check from the definition that every subgaussian random variable η (meaning that $\mathbb{E}e^{s\eta} \leqslant e^{\Gamma\sigma_\eta^2 s^2},\ s \in \mathbb{R}$) satisfies the following property:

$$\|\eta\|^2_{\psi_2} \leqslant 8\Gamma\sigma_\eta^2. \tag{A.4}$$

In what follows, we use the same notations for Orlicz norms on other probability spaces (for instance, $C_{bu}(\mathbb{T}; d_X)$ with its Borel σ-algebra and probability measure Π).

A.3. Bounds for subgaussian processes and Talagrand's generic chaining complexities

THEOREM A.2. — *Let $\{Z(t),\ t \in \mathbb{T}\}$ be a centered subgaussian process. Then, for all $u \geqslant 0,\ t_0 \in \mathbb{T}$,*

(1) $\mathbb{P}\left(\sup_t (Z(t) - Z(t_0)) \geqslant 2u \cdot \gamma_2(\mathbb{T}, d_Z)\right) \leqslant Ce^{-u^2/4}$,

(2) $\mathbb{E}\|Z\|_\infty \leqslant \mathbb{E}|Z(t_0)| + L\gamma_2(\mathbb{T}, d_Z)$,

where $d_Z(t,s) = \sqrt{\mathrm{Var}(Z(t) - Z(s))}$.

Proof. — See Chapter 1.2 in [39]. □

A simple corollary is the following inequality:

$$\mathbb{P}\Big(\|Z\|_\infty \geqslant C\sqrt{t}\big(\gamma_2(\mathbb{T}, d_Z) + \inf_{t\in\mathbb{T}}\sqrt{\mathrm{Var}(Z(t))}\,\big)\Big) \leqslant e^{-t}. \tag{A.5}$$

We mention another result which is useful in our investigation:

PROPOSITION A.1. — *Let Z be a centered subgaussian stochastic process such that*

$$\gamma_2(\mathbb{T}, d_Z) < \infty$$

and let $Z_1, \dots, Z_n$ be iid copies of Z. Then for any $t_0 \in \mathbb{T}$

(1) $(\log 2)^{-1/2}\big\|\|Z\|_\infty\big\|_{\psi_1} \leqslant \big\|\|Z\|_\infty\big\|_{\psi_2} \leqslant \|Z(t_0)\|_{\psi_2} + L\gamma_2(\mathbb{T}, d_Z)$,

(2) $\big\|\max_{j=1\dots n}\|Z_j\|_\infty\big\|_{\psi_1} \leqslant C\log n\big\|\|Z\|_\infty\big\|_{\psi_1}$.

Proof. — First statement is a straightforward corollary of Talagrand's result and integration-by-parts formula. For the proof of the second claim, see [41], Lemma 2.2.2. □

In the case when $Z(t)$, $t \in \mathbb{T}$ is a centered Gaussian process, a famous result of Talagrand (see Theorem 2.1.1 in [39]) states that

$$\frac{1}{K}\gamma_2(\mathbb{T}; d_Z) \leqslant \mathbb{E}\sup_{t\in\mathbb{T}} Z(t) \leqslant K\gamma_2(\mathbb{T}; d_Z) \tag{A.6}$$

for some universal constant K. Moreover, the upper bound also holds for the centered subgaussian process Z.

In practice, a useful way to estimate the generic chaining complexity $\gamma_2(\mathbb{T}; d_Z)$ and its "local version" $\gamma_2(\delta)$ is to evaluate Dudley's entropy integral:

Theorem A.3. — *The following inequality holds for all $\delta \leqslant \sup_{t,s\in\mathbb{T}} d_X(t,s)$:*

$$\gamma_2(\delta) \leqslant (2\sqrt{2}-1)^{-1} \int_0^\delta \sqrt{\log N(\mathbb{T}, d_Z, \varepsilon/4)}\, d\varepsilon,$$

where $N(\mathbb{T}, d_Z, \varepsilon)$ is the minimal number of balls of radius ε required to cover $\mathbb{T}$.

Proof. — This well-known bound can be obtained by repeating the argument of Proposition 1.2.1 in [39]. □

The following immediate corollary covers two important examples.

Corollary A.1

(1) *If* $\mathrm{Card}(\mathbb{T}) = N$, *then*

$$\gamma_2(\delta) \leqslant C\delta\sqrt{\log N};$$

(2) *If the covering numbers grow polynomially, i.e.* $N(\mathbb{T}, d_X, \varepsilon) \leqslant C_1 (A/\varepsilon)^V$, *then*

$$\gamma_2(\delta) \leqslant C_2\delta\sqrt{V \log(A/\delta)}\,.$$

A.4. Empirical processes. — We state a version of generic chaining bounds for empirical processes due to S. Mendelson, S. Dirksen and W. Bednorz which are used in our proofs. Let $\mathscr{F}$ be a class of functions defined on a measurable space $(S, \mathscr{A})$. Suppose $\mathscr{F}$ is symmetric, that is, $f \in \mathscr{F}$ implies $-f \in \mathscr{F}$ (in applications, we often deal with the classes that do not satisfy this assumption and then replace $\mathscr{F}$ by $\mathscr{F} \cup -\mathscr{F}$). Let $(X,\xi), (X_1,\xi), \dots, (X_n,\xi)$ be i.i.d. random variables with values in $S \times \mathbb{R}$ such that $\mathbb{E}f(X) = 0$, $f \in \mathscr{F}$ and ξ is a subgaussian random variable. Let Π be the marginal distribution of X. It will be used as a measure on $(S, \mathscr{A})$.

Theorem A.4. — *There exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\sup_{f\in\mathscr{F}}\left|n^{-1}\sum_{j=1}^n \xi_j f(X_j) - \mathbb{E}\xi f(X)\right| \leqslant C\left[\|\xi\|_{\psi_2}\frac{\gamma_2(\mathscr{F};\psi_2)}{\sqrt{n}} \bigvee \frac{\gamma_2^2(\mathscr{F};\psi_2)}{n}\right].$$

This inequality follows from Corollary 3.9 in [32]. We will often combine it with a version of Talagrand's concentration inequality for unbounded function classes due to Adamczak [1] (stated in a convenient form for our purposes). Let $\mathscr{F}$ be a class of functions defined on a measurable space $(S, \mathscr{A})$ and let $X, X_1, \dots, X_n$ be i.i.d. random variables sampled from distribution P on $(S, \mathscr{A})$. Let F be a measurable

envelope for $\mathscr{F}$, that is F is a measurable function on S such that $|f(x)| \leqslant F(x)$, $x \in S$, $f \in \mathscr{F}$. Then, there exists a universal constant $K > 0$ such that

$$\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{j=1}^{n} f(X_j) - \mathbb{E}f(X)\right| \leqslant K\left[\mathbb{E}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{j=1}^{n} f(X_j) - \mathbb{E}f(X)\right| + \sup_{f\in\mathscr{F}}\sqrt{\mathrm{Var}(f(X))}\sqrt{\frac{s}{n}} + \Big\|\max_{1\leqslant j\leqslant n}|F(X_j)|\Big\|_{\psi_1}\frac{s}{n}\right] \tag{A.7}$$

with probability $\geqslant 1 - e^{-s}$.

Finally, we state a recent sharp bound for the empirical processes due to S. Dirksen [15] and W. Bednorz [5] (earlier versions of exponential generic chaining bounds for similar empirical processes are due to Mendelson [33], [32]). Assume that $\{f(X),\, f \in \mathscr{F}\}$ is a subset of the subgaussian space $\mathscr{L}$.

THEOREM A.5. — *There exists an absolute constant $C > 0$ such that*

$$\sup_{f\in\mathscr{F}}\left|n^{-1}\sum_{j=1}^{n} f^2(X_j) - \mathbb{E}f^2(X)\right| \leqslant C\left[\sup_{f\in\mathscr{F}}\|f\|_{\psi_2}\frac{\gamma_2(\mathscr{F};\psi_2)}{\sqrt{n}} + \frac{\gamma_2^2(\mathscr{F};\psi_2)}{n} + \sup_{f\in\mathscr{F}}\|f\|_{\psi_2}^2\left(\sqrt{\frac{s}{n}}\vee\frac{s}{n}\right)\right]$$

with probability $\geqslant 1 - e^{-s}$.

For a proof and discussion of this bound, see Theorem 5.5 in [15]. Note that in (ii), the generic chaining complexity $\gamma_2(\mathscr{F};\psi_2)$ in the right-hand side is for the class $\mathscr{F}$ itself rather than $\mathscr{F}^2$.

A.5. SOBOLEV NORMS. — For any $p \in \mathbb{R}_+$, define the Sobolev space $\mathbb{W}^{2,p}(\mathbb{R}^d)$ as

$$\mathbb{W}^{2,p}(\mathbb{R}^d) = \left\{f \in L^2(\mathbb{R}^d) :\ \|f\|^2_{\mathbb{W}^{2,p}(\mathbb{R}^d)} := \int_{\mathbb{R}^d}(1+|t|^2)^p|\widehat{f}(t)|^2dt < \infty\right\},$$

where $\widehat{f}$ is the Fourier transform of f. It is well known that for $p \in \mathbb{Z}_+$, this coincides with another definition of Sobolev spaces (in terms of partial derivatives)

Assume that $f \in \mathbb{W}^{2,p}(\mathbb{R}^d)$ for $p \in \mathbb{Z}_+$ is such that $f = \sum_{j=1}^k f_j$, where f_j, $j = 1\dots k$ have disjoint supports. Clearly, in this case we have $\|f\|^2_{\mathbb{W}^{2,p}(\mathbb{R}^d)} = \sum_{j=1}^k \|f_j\|^2_{\mathbb{W}^{2,p}(\mathbb{R}^d)}$. When p is not an integer, we will use the following proposition.

PROPOSITION A.2. — *Assume that $p \in \mathbb{R}_+$, $\alpha := p - \lfloor p\rfloor > 0$ and $f \in \mathbb{W}^{2,p}(\mathbb{R}^d)$ is such that $f = \sum_{j=1}^k f_j$, where f_j, $j = 1\dots k$, have disjoint supports and*

$$\min_{1\leqslant i<j\leqslant k}\mathrm{dist}(\mathrm{supp}(f_i),\mathrm{supp}(f_j)) \geqslant r > 0,$$

where dist *is the Euclidean distance. Then*

$$\|f\|^2_{\mathbb{W}^{2,p}(\mathbb{R}^d)} \leqslant C(d,p)\left[\sum_{j=1}^k\|f_j\|^2_{\mathbb{W}^{2,p}(\mathbb{R}^d)} + \frac{1}{r^{2\alpha}}\sum_{j=1}^k\|f_j\|^2_{\mathbb{W}^{2,\lfloor p\rfloor}(\mathbb{R}^d)}\right].$$

Proof. — We will need to use equivalence of certain norms defined on Sobolev spaces $\mathbb{W}^{2,p}(\mathbb{R}^d)$. Let

$$'\|f\|^2_{\mathbb{W}^{2,p}(\mathbb{R}^d)} := \|f\|^2_{\mathbb{W}^{2,\lfloor p\rfloor}} + \max_{|m|=\lfloor p\rfloor} \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \frac{\left(\partial^m f(x) - \partial^m f(y)\right)^2}{|x-y|^{2\alpha+d}}dxdy,$$

where we use the usual multi-index notation. It is known that $\|\cdot\|_{\mathbb{W}^{2,p}}$ and $'\|\cdot\|_{\mathbb{W}^{2,p}}$ are equivalent (e.g., see [2] p. 219). Let $f = \sum_{j=1}^k f_i$ satisfy the conditions of the proposition. Making the change of variables $x = t+u$, $y = u$ in the expression of $'\|\cdot\|_{\mathbb{W}^{2,p}(\mathbb{R}^d)}$, we have

$$\begin{aligned}
'\|f\|^2_{\mathbb{W}^{2,p}} &= \|f\|^2_{\mathbb{W}^{2,\lfloor p\rfloor}} + \max_{|m|=\lfloor p\rfloor} \iint\limits_{\mathbb{R}^d\times\mathbb{R}^d} \left(\partial^m f(t+u) - \partial^m f(u)\right)^2 du\frac{dt}{|t|^{2\alpha+d}} \\
&= \sum_{j=1}^k \|f_j\|^2_{\mathbb{W}^{2,\lfloor p\rfloor}} + \max_{|m|=\lfloor p\rfloor} \Bigg[\iint\limits_{\mathbb{R}^d\times B(0,r)} \left(\partial^m f(t+u) - \partial^m f(u)\right)^2 du\frac{dt}{|t|^{2\alpha+d}} \\
&\qquad + \iint\limits_{\mathbb{R}^d\times \overline{B}(0,r)} \left(\partial^m f(t+u) - \partial^m f(u)\right)^2 du\frac{dt}{|t|^{2\alpha+d}}\Bigg].
\end{aligned}$$

It remains to notice that

$$\begin{aligned}
&\iint\limits_{\mathbb{R}^d\times B(0,r)} \left(\partial^m f(t+u) - \partial^m f(u)\right)^2 du\frac{dt}{|t|^{2\alpha+d}} \\
&\qquad = \sum_{j=1}^k \iint\limits_{\mathbb{R}^d\times B(0,r)} \left(\partial^m f_j(t+u) - \partial^m f_j(u)\right)^2 du\frac{dt}{|t|^{2\alpha+d}} \\
&\qquad \leqslant \sum_{j=1}^k \iint\limits_{\mathbb{R}^d\times \mathbb{R}^d} \frac{\left(\partial^m f_j(x) - \partial^m f_j(y)\right)^2}{|x-y|^{2\alpha+d}}dxdy
\end{aligned}$$

and, since $(f(t+u)-f(u))^2 \leqslant 2f^2(t+u) + 2f^2(u)$,

$$\begin{aligned}
\iint\limits_{\mathbb{R}^d\times \overline{B}(0,r)} \left(\partial^m f(t+u) - \partial^m f(u)\right)^2 du\frac{dt}{|t|^{2\alpha+d}} &\leqslant 2\|\partial^m f\|^2_{L_2(\mathbb{R}^d)} \int_{|t|\geqslant r} \frac{dt}{|t|^{d+2\alpha}} \\
&= C_1(d,\alpha)\frac{\|\partial^m f\|^2_{L_2(\mathbb{R}^d)}}{r^{2\alpha}},
\end{aligned}$$

where $C_1(d,\alpha) = 2\pi^{d/2}(d+2\alpha)/\alpha\Gamma(d/2)$, and the claim easily follows. □

A.6. Proof of Proposition 4.4. — Let J_1, J_2 be two disjoint subsets of $\{1,\dots,N\}$, and define

$$r(J_1;J_2) := \sup_{u,v} \left| \frac{\left\langle \sum_{j\in J_1} f_{u_j}, \sum_{j\in J_2} f_{v_j}\right\rangle_{L_2(\Pi)}}{\sqrt{\sum_{j\in J_1}\|f_{u_j}\|^2_{L_2(\Pi)} \sum_{j\in J_2}\|f_{v_j}\|^2_{L_2(\Pi)}}}\right|$$

where the supremum is taken over all

$$u = \sum_{j \in J_1} u_j, \quad v = \sum_{j \in J_2} v_j$$

such that

$$\operatorname{supp}(u_j) \subseteq \mathbb{T}_j, \quad \operatorname{supp}(v_j) \subseteq \mathbb{T}_j \quad \text{and} \quad \sum_{j \in J_1} \|f_{u_j}\|^2_{L_2(\Pi)} \neq 0, \quad \sum_{j \in J_2} \|f_{v_j}\|^2_{L_2(\Pi)} \neq 0.$$

Next, let

$$\rho_d := \max \{r(J_1; J_2) : \ J_1 \cap J_2 = \varnothing, \ \operatorname{Card}(J_1) + \operatorname{Card}(J_2) \leqslant 3d\}.$$

In what follows, we set $\lambda(u_j) := \big\|f_{u_j}\big\|_{L_2(\Pi)}$ and $h(u_j) := f_{u_j}/\|f_{u_j}\|_{L_2(\Pi)}$.

LEMMA A.1. — *The following inequality holds: $\rho_d \leqslant \delta_{3d}$.*

Proof. — See Lemma 2.1 in [11]. □

Set $J_0 := J$, $\lambda^{(0)} := \{\lambda(u_j), \ j \in J_0\}$, and let $(\lambda(u_{\pi(1)}), \dots, \lambda(u_{\pi_{N-d}}))$ be the vector $(\lambda(u_j), \ j \in J_0^c)$ sorted in the decreasing order, so that π is some permutation. We further define $J_1 := (\pi(1), \dots, \pi(d))$, $J_2 := (\pi(d+1), \dots, \pi(2d))$, etc., and $\lambda^{(k)} = (\lambda(u_j), \ j \in J_k)$. Everywhere below, $\|\cdot\|_1$, $\|\cdot\|_2$ denote the usual vector p-norms.

First, we will show that

$$\sum_{k \geqslant 2} \|\lambda^{(k)}\|_2 \leqslant b\|\lambda^{(0)}\|_2 := b\sqrt{\textstyle\sum_{j \in J_0} \lambda^2(u_j)}\,. \tag{A.8}$$

Indeed, for all $j \in J_k$, $k \geqslant 2$ we have $|\lambda(u_j)| \leqslant \frac{1}{d} \sum_{i \in J_{k-1}} |\lambda(u_i)|$, implying that $\|\lambda^{(k)}\|_2 \leqslant \frac{1}{\sqrt{d}} \|\lambda^{(k-1)}\|_1$. Summing up, we get

$$\sum_{k \geqslant 2} \|\lambda^{(k)}\|_2 \leqslant \frac{1}{\sqrt{d}} \sum_{j \notin J_0} |\lambda(u_j)| \leqslant \frac{b}{\sqrt{d}} \sum_{j \in J_0} |\lambda(u_j)|,$$

where the last inequality follows from the definition of the cone $C_{b,J}$. Inequality (A.8) follows since $\frac{b}{\sqrt{d}} \sum_{j \in J_0} |\lambda(u_j)| \leqslant b\sqrt{\sum_{j \in J_0} \lambda^2(u_j)}$.

Let P_J be the $L_2(\Pi)$-orthogonal projection onto L_J, the linear span of $\{h(u_j),\, j\in J\}$. The following sequence of inequalities establishes the claim of Proposition 4.4:

$$\begin{aligned}
\Big\|\sum_{j=1}^N \lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)} &\geqslant \Big\|P_{J_0\cup J_1}\sum_{j=1}^N \lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)}\\
&\geqslant \Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)} - \sum_{k\geqslant 2}\|P_{J_0\cup J_1}\sum_{j\in J_k}\lambda(u_j)h(u_j)\|_{L_2(\Pi)}\\
&\geqslant \Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)} - \rho_d\sum_{k\geqslant 2}\|\lambda^{(k)}\|_2 \underbrace{\sup_{v\in L_{J_0\cup J_1},\|v\|_{L_2(\Pi)}=1}\|v\|_2}_{\leqslant 1/(1-\delta_{2d})}\\
&\geqslant \Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)} - \frac{\rho_d}{1-\delta_{2d}}\sum_{k\geqslant 2}\|\lambda^{(k)}\|_2\\
&\geqslant \Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)} - \frac{\delta_{3d}b}{1-\delta_{2d}}\sqrt{\textstyle\sum_{j\in J_0\cup J_1}\lambda^2(u_j)}\\
&\geqslant \Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)} - \frac{\delta_{3d}b}{(1-\delta_{2d})^2}\Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)}\\
&= \Big(1-\delta_{3d}\frac{b}{(1-\delta_{2d})^2}\Big)\Big\|\sum_{j\in J_0\cup J_1}\lambda(u_j)h(u_j)\Big\|_{L_2(\Pi)}\\
&\geqslant \Big(1-\delta_{3d}\frac{b}{(1-\delta_{2d})^2}\Big)(1-\delta_{2d})\sqrt{\textstyle\sum_{j\in J_0}\lambda^2(u_j)},
\end{aligned}$$

hence $\beta_2^{(b)}(J)\leqslant (1-\delta_{2d})/\big[(1-\delta_{2d})^2-b\delta_{3d}\big]$.

It remains to show that $\delta_{3d}<1/(2+b)$ is sufficient for $\beta_2^{(b)}<\infty$. Since $\delta_{2d}\leqslant\delta_{3d}$, it is enough to show that $\delta_{3d}<1/(2+b)$ implies $(1-\delta_{3d})^2-b\delta_{3d}>0$. The latter is satisfied whenever $\delta_{3d}<\frac{2+b}{2}\big(1-\sqrt{1-4/(2+b)^2}\big)$. The elementary inequality $\sqrt{1-x}\leqslant 1-x/2$, $x\in[0,1]$, gives $\frac{2+b}{2}\big(1-\sqrt{1-4/(2+b)^2}\big)\geqslant 1/(2+b)$, and the result follows.

INDEX

References

[1] R. Adamczak – "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains", *Electron. J. Probab.* **13** (2008), p. 1000–1034.
[2] R. Adams – *Sobolev spaces*, Academic Press, New York, 1975.
[3] G. Bal – "Numerical methods for PDEs", Lecture notes available at `http://www.columbia.edu/~gb2030/COURSES/E6302/NumAnal.pdf`, 2009.
[4] P. L. Bartlett, S. Mendelson & J. Neeman – "ℓ_1-regularized linear regression: persistence and oracle inequalities", *Probab. Theory Relat. Fields* **154** (2012), p. 193–224.
[5] W. Bednorz – "Concentration via chaining method and its applications", `arXiv:1405.0676v2`, 2014.
[6] P. J. Bickel, Y. Ritov & A. B. Tsybakov – "Simultaneous analysis of Lasso and Dantzig selector", *Ann. Statist.* **37** (2009), no. 4, p. 1705–1732.
[7] V. I. Bogachev – *Measure theory. Vol. I, II*, Springer-Verlag, Berlin, 2007.
[8] P. Bühlmann & S. A. van de Geer – *Statistics for high-dimensional data*, Springer- Verlag, Berlin-Heidelberg, 2011.
[9] F. Bunea, A. B. Tsybakov & M. Wegkamp – "Sparsity oracle inequalities for the Lasso", *Electron. J. Statist.* **1** (2007), p. 169–194.
[10] T. T. Cai & P. Hall – "Prediction in functional linear regression", *Ann. Statist.* **34** (2006), no. 5, p. 2159–2179.
[11] E. Candès – "The restricted isometry property and its implications for compressed sensing", *Comptes Rendus Mathématique* **346** (2008), no. 9, p. 589–592.
[12] E. Candès & C. Fernandez-Granda – "Towards a mathematical theory of super-resolution", *Comm. Pure Appl. Math.* **67** (2014), no. 6, p. 906–956.
[13] E. J. Candès, J. K. Romberg & T. Tao – "Stable signal recovery from incomplete and inaccurate measurements", *Comm. Pure Appl. Math.* **59** (2006), no. 8, p. 1207–1223.
[14] C. Crambes, A. Kneip & P. Sarda – "Smoothing splines estimators for functional linear regression", *Ann. Statist.* **37** (2009), no. 1, p. 35–72.
[15] S. Dirksen – "Tail bounds via generic chaining", `arXiv:1309.3522`, 2013.
[16] S. A. van de Geer – "High-dimensional generalized linear models and the Lasso", *Ann. Statist.* **36** (2008), no. 2, p. 614–645.
[17] S. A. van de Geer & J. Lederer – "The Lasso, correlated design, and improved oracle inequalities", in *A Festschrift in Honor of Jon Wellner*, IMS Collections, Institute of Mathematical Statistics, 2012, p. 3468–3497.
[18] E. D. Gluskin – "Norms of random matrices and widths of finite-dimensional sets", *Mat. Sb.* **120(162)** (1983), no. 2, p. 180–189.
[19] M. Hebiri & J. Lederer – "How correlations influence Lasso prediction", *IEEE Trans. Information Theory* **59** (2013), no. 3, p. 1846–1854.
[20] A. D. Ioffe & V. M. Tikhomirov – *Theory of extremal problems*, Nauka, Moscow, 1974.
[21] G. James – "Sparseness and functional data analysis", in *The Oxford handbook of functional data analysis*, Oxford University Press, New York, 2011, p. 298–323.
[22] G. M. James, J. Wang & J. Zhu – "Functional linear regression that's interpretable", *Ann. Statist.* **37** (2009), no. 5A, p. 2083–2108.
[23] V. Koltchinskii – "The Dantzig selector and sparsity oracle inequalities", *Bernoulli* **15** (2009), no. 3, p. 799–828.
[24] ———, "Sparse recovery in convex hulls via entropy penalization", *Ann. Statist.* **37** (2009), no. 3, p. 1332–1359.

[25] ———, "Sparsity in penalized empirical risk minimization", *Ann. Inst. H. Poincaré Probab. Statist.* **45** (2009), no. 1, p. 7–57.
[26] ———, "Oracle inequalities in empirical risk minimization and sparse recovery problems", in *38th Probability Summer School (Saint-Flour, 2008)*, Springer, 2011.
[27] V. Koltchinskii, K. Lounici & A. B. Tsybakov – "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion", *Ann. Statist.* **39** (2011), no. 5, p. 2302–2329.
[28] V. Koltchinskii & S. Minsker – "Sparse recovery in convex hulls of infinite dictionaries", in *COLT 2010, 23rd Conference on Learning Theory*, 2010, p. 420–432.
[29] S. Lang – *Real and functional analysis*, 3rd ed., Graduate Texts in Math., vol. 142, Springer, 1993.
[30] M. A. Lifshits – *Gaussian random functions*, Mathematics and its Applications, vol. 322, Kluwer Academic Publishers, Dordrecht, 1995.
[31] P. Massart & C. Meynet – "The Lasso as an ℓ_1-ball model selection procedure", *Electron. J. Statist.* **5** (2011), p. 669–687.
[32] S. Mendelson – "Oracle inequalities and the isomorphic method", Preprint, 2012. Available at `http://maths-people.anu.edu.au/~mendelso/papers/subgaussian-12-01-2012.pdf`.
[33] ———, "Empirical processes with a bounded ψ_1 diameter", *Geom. Funct. Anal.* **20** (2010), no. 4, p. 988–1027.
[34] H. G. Müller & U. Stadtmüller – "Generalized functional linear models", *Ann. Statist.* **33** (2005), no. 2, p. 774–805.
[35] J. O. Ramsay – *Functional data analysis*, Wiley Online Library, 2006.
[36] J. O. Ramsay & B. W. Silverman – *Applied functional data analysis: methods and case studies*, Springer Series in Statistics, vol. 77, Springer, New York, 2002.
[37] K. Ritter, G. W. Wasilkowski & H. Woźniakowski – "Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions", *Ann. Appl. Probab.* (1995), p. 518–540.
[38] J. Sacks & D. Ylvisaker – "Designs for regression problems with correlated errors", *Ann. Statist.* **37** (1966), no. 1, p. 66–89.
[39] M. Talagrand – *The generic chaining*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2005.
[40] R. Tibshirani – "Regression shrinkage and selection via the Lasso", *J. R. Stat. Soc. Ser. B Stat. Methodol.* (1996), p. 267–288.
[41] A. W. van der Vaart & J. A. Wellner – *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996.
[42] M. Yuan & T. T. Cai – "A reproducing kernel Hilbert space approach to functional linear regression", *Ann. Statist.* **38** (2010), no. 6, p. 3412–3444.

Manuscript received January 13, 2014
accepted September 4, 2014

Vladimir Koltchinskii, School of Mathematics, Georgia Institute of Technology
686 Cherry Street, Atlanta, GA 30332-0160 USA
E-mail : `vlad@math.gatech.edu`
Url : `http://www.math.gatech.edu/users/vlad`

Stanislav Minsker, Department of Mathematics, Duke University
Box 90320, Durham NC 27708-0320,
E-mail : `stas.minsker@gmail.com`
Url : `https://sminsker.wordpress.com`