

The Mismatch Principle: Statistical Learning Under Large Model Uncertainties

Martin Genzel Gitta Kutyniok

Technische Universität Berlin, Department of Mathematics
Straße des 17. Juni 136, 10623 Berlin, Germany

E-Mail: [\[genzel,kutyniok@math.tu-berlin.de\]](mailto:genzel,kutyniok@math.tu-berlin.de)

Abstract. We study the learning capacity of empirical risk minimization with regard to the squared loss and a convex hypothesis class consisting of linear functions. While these types of estimators were originally designed for noisy linear regression problems, it recently turned out that they are in fact capable of handling considerably more complicated situations, involving highly non-linear distortions. This work intends to provide a comprehensive explanation of this somewhat astonishing phenomenon. At the heart of our analysis stands the mismatch principle, which is a simple, yet generic recipe to establish theoretical error bounds for empirical risk minimization. The scope of our results is fairly general, permitting arbitrary sub-Gaussian input-output pairs, possibly with strongly correlated feature variables. Noteworthy, the mismatch principle also generalizes to a certain extent the classical orthogonality principle for ordinary least squares. This adaptation allows us to investigate problem setups of recent interest, most importantly, high-dimensional parameter regimes and non-linear observation processes. In particular, our theoretical framework is applied to various scenarios of practical relevance, such as single-index models, variable selection, and strongly correlated designs. We thereby demonstrate the key purpose of the mismatch principle, that is, learning (semi-)parametric output rules under large model uncertainties and misspecifications.

Key words. Statistical learning, constrained empirical risk minimization, orthogonality principle, Gaussian mean width, semi-parametric models, single-index models, variable selection.

1 Introduction

One of the key objectives in statistical learning and related fields is to study *sampling processes* that arise from a random pair (x, y) in $\mathbb{R}^p \times \mathbb{R}$ whose joint probability distribution μ is *unknown*. In this context, the random vector $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ is typically regarded as a collection of *input variables* or *features*, whereas $y \in \mathbb{R}$ corresponds to an *output variable* that one would like to predict from x .¹ More precisely, the main goal is to select a function $f^* : \mathbb{R}^p \rightarrow \mathbb{R}$ from a certain *hypothesis class* $\mathcal{F} \subset L^2(\mathbb{R}^p, \mu_x)$ such that the expected risk is minimized:

$$\min_{f \in \mathcal{F}} \mathbb{E}[(y - f(x))^2]. \quad (1.1)$$

A solution $f^* \in \mathcal{F}$ to (1.1) is then called an *expected risk minimizer* and yields the *optimal estimator* (approximation) of y in \mathcal{F} , with respect to the mean squared error.

However, it is not possible to directly solve (1.1) in practice because the underlying probability measure μ of (x, y) is unknown. Instead, one is merely given a finite amount of *observed data*

$$(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$$

where each of these pairs is an independent random sample of (x, y) . This limitation suggests to consider the empirical analog of (1.1), which is well-known as *empirical risk minimization* (ERM):

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2. \quad (1.2)$$

Such types of optimization problems have been extensively studied in statistical learning theory during the last decades, e.g., see [CZ07; SB14; Vap98] for comprehensive overviews. One of the primary concerns in this field of research is to establish (non-asymptotic) bounds on the *estimation error* $\mathbb{E}_x[(\hat{f}(x) - f^*(x))^2]$ with $\hat{f} \in \mathcal{F}$ being a minimizer of (1.2). In particular, it has turned

¹There exist many synonyms for x and y . We collected some of them in Table 1, which summarizes the most important notations and terminology of this work.

out that, in many situations of interest, empirical risk minimization constitutes a *consistent* (or *asymptotically unbiased*) estimator of f^* , in the sense that $\mathbb{E}_x[(\hat{f}(x) - f^*(x))^2] \rightarrow 0$ as $n \rightarrow \infty$.

In this paper, we only focus on those hypothesis classes that contain *linear* functions, i.e.,

$$\mathcal{F} = \{x \mapsto \langle x, \beta \rangle \mid \beta \in K\}$$

for a certain *convex* subset $K \subset \mathbb{R}^p$, which is referred to as the *hypothesis set* or *constraint set*. While this choice of \mathcal{F} is actually one of the simplest examples to think of, the associated empirical risk minimization program has prevailed as a standard approach in modern statistics and signal processing. Indeed, (1.2) just turns into a constrained least squares estimator in the linear case:

$$\min_{\beta \in K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2. \quad (\text{P}_K)$$

The purpose of the *parameter vector* $\beta \in K$ is now to “explain” the observed data by a linear model. At the same time, the hypothesis set K imposes additional structural constraints that restrict the set of admissible models. In this way — as K is appropriately chosen — accurate estimation can even become possible in *high-dimensional* scenarios where $n \ll p$. Perhaps the most popular example is the *Lasso* [Tib96], which was originally designed for *sparse* linear regression and employs a scaled ℓ^1 -ball as constraint set. For that reason, one sometimes also speaks of the *generalized Lasso* or *K-Lasso* when referring to (P_K) , cf. [PV16].

A particular strength of many results in statistical learning is that almost no specific assumptions on the output variable y are made, except for some mild regularity. While these theoretical findings in principle cover a wide class of estimation problems, the actual error bounds are however rather implicit and technical. On the other hand, there is usually more knowledge available about the sampling process in concrete model settings, eventually allowing for more practicable guarantees. This work is an attempt to bridge the gap between these two conceptions: our main results still take the abstract viewpoint of statistical learning, but also provide an easy-to-use toolbox to derive *explicit error bounds* for (P_K) . Before presenting further details of our approach in the next subsection, we wish to highlight two important aspects of the sample pair $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ that will emerge over and over again in the remainder of this article:

- (1) *Correlated features.* In case x represents real-world data, the input variables are typically strongly correlated and affected by noise. But as long as the covariance matrix of x is well-defined (and is of rank d), there at least exists an isotropic decomposition of the form

$$x = As = \sum_{k=1}^d s_k a_k \quad (1.3)$$

where $s = (s_1, \dots, s_d)$ is an *isotropic* random vector¹ in \mathbb{R}^d and $A = [a_1 \dots a_d] \in \mathbb{R}^{p \times d}$ is a *deterministic* matrix (cf. Proposition 2.2). The *mixing matrix* A is often (partially) unknown, so that s is not directly accessible. Therefore, we sometimes also refer to the components of s as *latent variables* or *latent factors*. The statistical fluctuations of x are however completely determined by s , which justifies why all parametric relationships between y and x are stated in terms of s below.

The *linear factor model* of (1.3) leads to the following simple, yet important conclusion on the estimation error of (P_K) : Let $\hat{\beta} \in \mathbb{R}^p$ be a solution to (P_K) and let $\beta^* \in \mathbb{R}^p$ be an expected risk minimizer, i.e., $f^* := \langle \cdot, \beta^* \rangle$ minimizes (1.1). Setting $\hat{f} := \langle \cdot, \hat{\beta} \rangle$ and using the isotropy of s , we observe that²

$$\begin{aligned} \mathbb{E}_x[(\hat{f}(x) - f^*(x))^2] &= \mathbb{E}_x[\langle x, \hat{\beta} - \beta^* \rangle^2] = \mathbb{E}_s[\langle As, \hat{\beta} - \beta^* \rangle^2] \\ &= \mathbb{E}_s[\langle s, A^\top \hat{\beta} - A^\top \beta^* \rangle^2] = \|A^\top \hat{\beta} - A^\top \beta^*\|_2^2. \end{aligned} \quad (1.4)$$

¹A random vector s in \mathbb{R}^d is *isotropic* if $\mathbb{E}[\langle s, z \rangle^2] = \|z\|_2^2$ for all $z \in \mathbb{R}^d$.

²More precisely, we condition on the sample set $\{(x_i, y_i)\}_{i=1}^n$ in (1.4), so that \hat{f} and $\hat{\beta}$ are both non-random at this point.

This identity reflects a well-known issue of least squares estimators: while both $\hat{\beta}$ and β^* can be highly non-unique due to (perfectly) correlated input variables, the error measure of (1.4) does not suffer from this “ambiguity.” In a certain sense, the weighting by A^\top reverts the mixing of (1.3) and therefore allows for a comparison of two parameter vectors. But it is worth mentioning that (1.4) is primarily of theoretical interest and does not resolve the problem of non-unique solutions per se, unless A is roughly known.

- (2) *Semi-parametric models.* It is quite obvious that the linear hypothesis functions used in (P_K) are not capable of learning complicated non-linear output rules. More specifically, we cannot expect that $\hat{f}(x) = \langle x, \hat{\beta} \rangle$ or $f^*(x) = \langle x, \beta^* \rangle$ provide a good approximation of y . On the other hand, the output y often (approximately) follows a semi-parametric model and one aims at estimating some unknown parameters, rather than predicting the value of y . For that reason, there is still hope that the outcome of (P_K) allows us to infer the parametric structure of $(x, y) = (As, y)$. Although our results below are valid in much greater generality, it is very helpful to bear in mind the following two scenarios which will serve as running examples in our framework:

- *Single-index models.* Let $z_0 \in \mathbb{R}^d$ and assume that

$$y = g(\langle s, z_0 \rangle) \quad (1.5)$$

for a scalar function $g: \mathbb{R} \rightarrow \mathbb{R}$ which can be non-linear, unknown, and random (independently of s). The goal is to construct an estimator of the unknown *index vector* z_0 (or at least of its direction $z_0 / \|z_0\|_2$) using (P_K) .

- *Variable selection.* Let $\{k_1, \dots, k_S\} \subset \{1, \dots, d\}$ and assume that

$$y = G(s_{k_1}, \dots, s_{k_S}) \quad (1.6)$$

for a function $G: \mathbb{R}^S \rightarrow \mathbb{R}$ which can be again non-linear, unknown, and random. Can we use empirical risk minimization (P_K) to extract the set of *active variables* $\mathcal{S} := \{k_1, \dots, k_S\}$?

The above concerns give rise to several general issues that we would like to address in this work:

- (Q1) *Estimation.* Can we prove a non-asymptotic upper bound on the estimation error in (1.4) with respect to the sample size n ? When does $\hat{z} := A^\top \hat{\beta} \in \mathbb{R}^d$ provide a consistent estimator of $z^* := A^\top \beta^* \in \mathbb{R}^d$?
- (Q2) *Approximation.* Can we extract the parameters of interest from \hat{z} , or according to (Q1), what information about (x, y) does z^* carry? For example, how close is z^* to $\text{span}\{z_0\}$ when learning a single-index model (1.5), or is z^* supported on \mathcal{S} as y obeys (1.6)?
- (Q3) *Complexity.* What role is played by the hypothesis set K (in high-dimensional problems)? How to exploit low-complexity features of the underlying model, e.g., if $S \ll d$ in variable selection (1.6)?

1.1 The Mismatch Principle

In order to highlight the main ideas of our theoretical approach and to avoid unnecessary technicalities, let us assume in this subsection that (1.3) holds true with $A = I_d$, where $I_d \in \mathbb{R}^{d \times d}$ denotes the identity matrix in \mathbb{R}^d . This simplification of the data model implies that $x = s$ is an *isotropic* random vector in \mathbb{R}^d , and adapting the terminology of the previous paragraphs, we particularly have $\hat{\beta} = \hat{z}$, $\beta^* = z^*$, and $x_i = s_i$ for $i = 1, \dots, n$. For the sake of consistency (with the definitions of Section 2), we agree on writing s , s_i , \hat{z} , z^* instead of x , x_i , $\hat{\beta}$, β^* , respectively. The estimator (P_K) then reads as follows:¹

$$\min_{z \in K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle s_i, z \rangle)^2. \quad (\text{Piso})$$

¹When referring to (P_K) as an *estimator*, we actually mean a *minimizer* of (P_K) .

A key quantity of our framework is the so-called *mismatch covariance* which is defined as

$$\rho(\mathbf{z}^{\natural}) := \|\mathbb{E}[(y - \langle \mathbf{s}, \mathbf{z}^{\natural} \rangle) \mathbf{s}]\|_2, \quad \mathbf{z}^{\natural} \in \mathbb{R}^d.$$

The name ‘mismatch covariance’ is due to the fact that $\rho(\mathbf{z}^{\natural})$ basically measures the covariance between the input variables $\mathbf{s} \in \mathbb{R}^d$ and the *mismatch* $y - \langle \mathbf{s}, \mathbf{z}^{\natural} \rangle$ that arises from approximating the (possibly non-linear) output y by a linear function $f(\mathbf{s}) := \langle \mathbf{s}, \mathbf{z}^{\natural} \rangle$. To better understand the meaning of this statistical parameter, it is very helpful to take a closer look at the following equivalence:

$$\rho(\mathbf{z}^{\natural}) = 0 \quad \Leftrightarrow \quad \forall \mathbf{z} \in \mathbb{R}^d: \mathbb{E}[(y - \langle \mathbf{s}, \mathbf{z}^{\natural} \rangle) \langle \mathbf{s}, \mathbf{z} \rangle] = 0. \quad (1.7)$$

The right-hand side of (1.7) in fact corresponds to the fundamental *orthogonality principle* in linear estimation theory. According to its classical formulation, the orthogonality principle states that the prediction error of the optimal estimator is orthogonal (uncorrelated) to every possible linear estimator $f(\mathbf{s}) = \langle \mathbf{s}, \mathbf{z} \rangle$ with $\mathbf{z} \in \mathbb{R}^d$ (cf. [Kay93, Sec. 12.4]). This argument becomes more plausible when computing the gradient of the expected risk at $\mathbf{z}^{\natural} \in \mathbb{R}^d$,

$$\nabla_{\mathbf{z}=\mathbf{z}^{\natural}} \mathbb{E}[(y - \langle \mathbf{s}, \mathbf{z} \rangle)^2] = -2\mathbb{E}[(y - \langle \mathbf{s}, \mathbf{z}^{\natural} \rangle) \mathbf{s}],$$

which implies that (1.7) is equivalent to $f = \langle \cdot, \mathbf{z}^{\natural} \rangle$ being a critical point of the objective function in (1.1). Consequently, if the mismatch covariance vanishes at $\mathbf{z}^{\natural} \in K$, then \mathbf{z}^{\natural} is an expected risk minimizer in the sense that $\mathbf{z}^{\natural} = \mathbf{z}^*$.

However, we emphasize that (1.7) is only a *sufficient* condition for $\mathbf{z}^{\natural} = \mathbf{z}^*$, since the solution space of (1.1) is restricted to a certain hypothesis set $K \subset \mathbb{R}^n$. There indeed exist several relevant model setups sketched further below where the traditional orthogonality principle fails to work. In this light, the mismatch covariance can be considered as a refined concept that quantifies how far one is from fulfilling the optimality condition of (1.7). Let us make this idea more precise by stating an informal version of our main result (cf. Theorem 2.5 and Corollary 3.1):

Theorem 1.1 (informal) *Let y be a sub-Gaussian variable and let \mathbf{s} be an isotropic, mean-zero sub-Gaussian random vector in \mathbb{R}^d (cf. Subsection 1.3(4)). Assume that $K \subset \mathbb{R}^d$ is a bounded, convex subset and fix a vector $\mathbf{z}^{\natural} \in K$. If $n \geq C \cdot w^2(K)$, then every minimizer $\hat{\mathbf{z}}$ of $(\mathbb{P}_K^{\text{iso}})$ satisfies with high probability*

$$\|\hat{\mathbf{z}} - \mathbf{z}^{\natural}\|_2 \lesssim C' \cdot \left(\frac{w^2(K)}{n} \right)^{1/4} + \rho(\mathbf{z}^{\natural}), \quad (1.8)$$

where $w(K)$ denotes the Gaussian mean width of the hypothesis set K (cf. Definition 2.4), and $C, C' > 0$ are model-dependent constants detailed in Section 2.

At first sight, this result appears somewhat odd, as it states an error bound for every choice of the *target vector* $\mathbf{z}^{\natural} \in K$. But the actual significance of (1.8) clearly depends on the size of the mismatch covariance $\rho(\mathbf{z}^{\natural})$. In particular, if $\rho(\mathbf{z}^{\natural}) > 0$, we do not even obtain a consistent estimate of \mathbf{z}^{\natural} , i.e., the approximation error does not tend to 0 as $n \rightarrow \infty$. Following the above reasoning, it is therefore quite natural to select $\mathbf{z}^{\natural} = \mathbf{z}^*$, which yields the best possible outcome from Theorem 1.1 and reflects the common strategy of statistical learning.

While this viewpoint primarily addresses the sampling-related issues of (Q1) and (Q3), the approximation problem of (Q2) remains untouched. A precise answer to (Q2) in fact strongly depends on the interplay between the structure of the (semi-parametric) output variable y and the corresponding expected risk minimizer \mathbf{z}^* . For example, Plan and Vershynin recently verified in [PV16] that a single-index model (1.5) can be consistently learned via $(\mathbb{P}_K^{\text{iso}})$, as long as the input data is *Gaussian*. Although not stated explicitly, their finding is underpinned by the orthogonality principle of (1.7), in the sense that one can always achieve $\rho(\mu \mathbf{z}_0) = 0$ for an appropriate scalar factor $\mu \in \mathbb{R}$; see Subsection 3.1.1 for more details. Interestingly, the situation is very different for *non-Gaussian* single-index models, where \mathbf{z}^* does not necessarily belong to $\text{span}\{\mathbf{z}_0\}$ anymore; see [PV13a, Rmk. 1.5] and [ALPV14]. But fortunately, the versatility of Theorem 1.1 pays off at this point because it still permits to select a vector $\mathbf{z}^{\natural} \in T := \text{span}\{\mathbf{z}_0\}$. Thus, if there exists $\mathbf{z}^{\natural} \in T \cap K$ such that $\rho(\mathbf{z}^{\natural})$ becomes sufficiently small, (1.8) turns into a meaningful recovery guarantee for single-index models. A similar strategy applies to the problem of variable selection

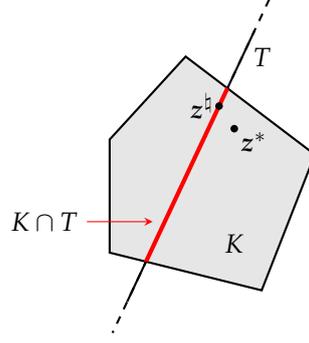


Figure 1: Visualization of the mismatch principle. The mismatch covariance $\rho(\cdot)$ is minimized on the intersection $K \cap T$ (red part). Note that the minimizer $z^h \in \mathbb{R}^d$ is not necessarily equal to the expected risk minimizer $z^* \in \mathbb{R}^d$, and the latter might not even belong to $K \cap T$.

(1.6) by considering $z^h \in T := \{z \in \mathbb{R}^d \mid \text{supp}(z) \subset \mathcal{S}\}$; see Subsection 3.1.3.

These prototypical examples demonstrate that semi-parametric models often come along with a certain *target set* $T \subset \mathbb{R}^d$ containing all those vectors which allow us to extract the parameters of interest. In other words, one would be satisfied as soon as an estimation procedure approximates any vector $z^h \in T$. Transferring this notion to the general setup of Theorem 1.1, we simply assume that there exists such a target set $T \subset \mathbb{R}^d$, which encodes the desired (parametric) information about the data pair (s, y) . The meaning of the term ‘information’ is of course highly application-specific in this context, for instance, it is also reasonable to ask for the support of a (sparse) index vector z_0 in (1.5), instead of its direction. Therefore, we will leave T unspecified in the following and treat it as an abstract component of the underlying observation model. Note that, compared to the hypothesis set K , the target set T is *unknown* in practice, and to some extent, it plays the role of the *ground truth* parameters that one would like to estimate.

Combining the concept of target sets with Theorem 1.1, we can now formulate an informal version of the *mismatch principle*:

Specify a target vector $z^h \in T \cap K$ such that the mismatch covariance $\rho(z^h)$ is minimized. Then invoke Theorem 1.1 to obtain an error estimate for z^h by a minimizer of (P_K^{iso}) .

This simple recipe includes the key aspects of (Q1)–(Q3): By enforcing $z^h \in T \cap K$, we ensure that the target vector is consistent with our (parametric) assumptions on the output variable y and satisfies the model hypothesis imposed by K at the same time (see Figure 1). The error bound of (1.8) then quantifies the compatibility of these desiderata by means of the mismatch covariance $\rho(z^h)$ and the Gaussian mean width $w(K)$. This does not only provide a theoretical guarantee for empirical risk minimization (P_K^{iso}) , but also more insight into its capacity to solve (or not) a specific estimation problem.

Finally, let us point out once again that the mismatch principle can be regarded as an extension of classical learning theory. While the standard learning setting would suggest to choose $T = \mathbb{R}^d$ and $z^h = z^*$, we allow for a restriction of T to a certain family of *interpretable* models that meets the demands of (Q2). In particular, one should not confuse (Q2) with the well-known challenge of approximation in statistical learning, which rather concerns the approximation error caused by (1.1).

1.2 Main Contributions and Overview

The major purpose of this work is to shed more light on the capability of constrained empirical risk minimization to learn *non-linear* output rules. A deeper understanding of these standard estimators is of considerable practical relevance because they established themselves as a benchmark method in many different application fields, such as machine learning, signal- and image processing, econometrics, or bioinformatics. This popularity is primarily due to the fact that — compared to sophisticated non-convex methods — the associated minimization problem (P_K) does not rely on prior knowledge and can be (often) efficiently solved via convex programming.

But although conceptually quite simple, the performance of (P_K) is not fully understood to this day. A particular shortcoming of most theoretical approaches from the literature is that they either focus on very restrictive model settings (e.g., noisy linear regression), or do not make any structural assumptions on y . This paper aims at bridging these two converse viewpoints by employing the novel concept of *target sets*. As already sketched in the previous subsection, this refinement of the traditional learning setting forms the basis of the *mismatch principle*, which eventually enables us to prove out-of-the-box guarantees for empirical risk minimization.

In Section 2, we make the approach of Subsection 1.1 more precise and provide technical definitions of the so-called *mismatch parameters* and the *Gaussian mean width* (Subsection 2.2). Our first main result, Theorem 2.5, formalizes the statement of Theorem 1.1 and allows for arbitrary mixing matrices in (1.3), meaning that the features of $x \in \mathbb{R}^p$ may be (even perfectly) correlated. In this context, we do also continue our discussion on the mismatch principle (see Recipe 2.6) and its implications. Subsection 2.4 then revisits (Q3), refining the global Gaussian mean width $w(K)$ in the first error term of (1.8). Indeed, by using a *localized* complexity parameter, it is possible to achieve the optimal error decay rate of $O(n^{-1/2})$ in terms of n (see Theorem 2.8). This improvement is of particular importance to high-dimensional problems where the dimension of the data p is significantly larger than the actual sample size n . The proofs of our main results are postponed to Section 6. We very loosely follow the frameworks from [Gen17; GK16; Men15], making use of recent advances in uniform bounds for empirical quadratic and multiplier processes.

Section 3 presents a variety of applications of the mismatch principle. In the first part (Subsection 3.1), we return to our initial examples of single-index models and variable selection, but several other popular output rules will be elaborated as well, including generalized linear models, multiple-index models, and superimposed observations. By contrast, Subsection 3.2 does not focus on specific models for y but investigates issues related to *correlated* feature variables, i.e., $A \neq I_d$. In the underdetermined case ($p \geq d$), we will show that estimation via (P_K) is still feasible as long as A is only approximately known. On the other hand, the overdetermined case ($p < d$) is relevant to modeling real-world data that is affected by noise, which does not contribute to the output variable y . This is another important scenario where the classical orthogonality principle does not remain valid (cf. (1.7)), whereas the mismatch covariance naturally incorporates the signal-to-noise ratio of the observation process. Let us emphasize at this point that our focus is clearly on the *theoretical* aspects of empirical risk minimization. While the mismatch principle per se is a tool to analyze the performance of (P_K) , it can also have interesting practical implications if applied appropriately. For instance, one could think of modifying the data (e.g., by standardization) in such a way that the error bound (1.8) predicts a better approximation behavior.

We have outlined above how our approach fits into the field of statistical learning. Section 4 provides a more detailed discussion of the related literature. This does not only concern recent results in learning theory (Subsection 4.1) but also advances in signal processing and compressed sensing (Subsection 4.2) as well as strongly correlated designs (Subsection 4.3). Our concluding remarks and various future directions are presented in Section 5. In fact, the findings of this work are amenable to many possible extensions and generalizations, such as general (non-)convex hypothesis classes, different types of loss functions, or heavy-tailed data.

1.3 Notations and Preliminaries

Before proceeding with the main results, let us fix some notations and conventions that are frequently used in this work:

(1) Vectors and matrices are denoted by lower- and uppercase boldface symbols, respectively. Their entries are indicated by subscript indices and lowercase letters, e.g., $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ for a vector and $M = [m_{k,k'}] \in \mathbb{R}^{d \times d'}$ for a matrix. The (horizontal) *concatenation* of $M \in \mathbb{R}^{d \times d'}$ and $\tilde{M} \in \mathbb{R}^{d' \times d''}$ is denoted by $[M, \tilde{M}] \in \mathbb{R}^{d \times (d'+d'')}$. The *identity matrix* in \mathbb{R}^d is $I_d \in \mathbb{R}^{d \times d}$ and $e_k \in \mathbb{R}^d$ denotes the k -th *unit vector*.

(2) The *support* of a vector $v \in \mathbb{R}^d$ is the index set of its non-zero components, i.e., $\text{supp}(v) := \{k \in \{1, \dots, d\} \mid v_k \neq 0\}$, and we set $\|v\|_0 := |\text{supp}(v)|$. In particular, v is called *s-sparse* if

$\|v\|_0 \leq s$. For $q \geq 1$, the ℓ^q -norm of v is given by

$$\|v\|_q := \begin{cases} (\sum_{k=1}^d |v_k|^q)^{1/q}, & q < \infty, \\ \max_{k=1, \dots, d} |v_k|, & q = \infty. \end{cases}$$

The associated *unit ball* is denoted by $B_q^d := \{v \in \mathbb{R}^d \mid \|v\|_q \leq 1\}$ and the *Euclidean unit sphere* is $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$. The *operator norm* of $M \in \mathbb{R}^{d \times d'}$ is $\|M\|_{\text{op}} := \sup_{v' \in \mathbb{S}^{d'-1}} \|Mv'\|_2$.

(3) Let $L \subset \mathbb{R}^d$ and $v \in \mathbb{R}^d$. The *linear hull* of L is denoted by $\text{span } L$, and its *conic hull* is defined as $\text{cone}(L) := \{\tau v \mid v \in L, \tau \geq 0\}$. Moreover, we write $L \pm v := \{\tilde{v} \pm v \mid \tilde{v} \in L\}$.

If $E \subset \mathbb{R}^d$ is a linear subspace, the associated *orthogonal projection* onto E is denoted by $P_E \in \mathbb{R}^{d \times d}$, and we write $P_E^\perp := I_d - P_E$ for the projection onto the orthogonal complement $E^\perp \subset \mathbb{R}^d$. Moreover, if $E = \text{span}\{v\}$, we use the short notations $P_v := P_E$ and $P_v^\perp := P_E^\perp$.

(4) The *expected value* is denoted by $\mathbb{E}[\cdot]$ and we sometimes use a subscript to indicate that the expectation (integral) is only computed with respect to a certain random variable. Let v be a real-valued random variable. Then v is *sub-Gaussian* if

$$\|v\|_{\psi_2} := \sup_{q \geq 1} q^{-1/2} (\mathbb{E}[|v|^q])^{1/q} < \infty, \quad (1.9)$$

and $\|\cdot\|_{\psi_2}$ is called the *sub-Gaussian norm* (cf. [Ver12]). The *sub-Gaussian norm* of a random vector v in \mathbb{R}^d is then given by $\|v\|_{\psi_2} := \sup_{u \in \mathbb{S}^{d-1}} \|\langle v, u \rangle\|_{\psi_2}$. Moreover, v is called *isotropic* if $\mathbb{E}[vv^\top] = I_d$, or equivalently, if

$$\mathbb{E}[\langle v, u \rangle \langle v, u' \rangle] = \langle u, u' \rangle \quad \text{for all } u, u' \in \mathbb{R}^d. \quad (1.10)$$

If \tilde{v} is another random variable (not necessarily real-valued), we denote the expectation of v conditional on \tilde{v} by $\mathbb{E}[v \mid \tilde{v}]$. Finally, we write $v \sim \mathcal{N}(\mathbf{0}, \Sigma)$ if v is a (mean-zero) *Gaussian random vector* with covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$.

(5) The letter C is reserved for a (generic) constant, whose value could change from time to time. We refer to $C > 0$ as a *numerical constant* if its value does not depend on any other involved parameter. If an (in-)equality holds true up to a numerical constant C , we may simply write $A \lesssim B$ instead of $A \leq C \cdot B$, and if $C_1 \cdot A \leq B \leq C_2 \cdot A$ for numerical constants $C_1, C_2 > 0$, we use the abbreviation $A \asymp B$.

2 Main Results

This part builds upon the key ideas from Subsection 1.1 and derives a formal version of the mismatch principle (Recipe 2.6). Starting with a definition of our model setup in Subsection 2.1 and several technical parameters in Subsection 2.2, our main results are presented in Subsection 2.3 (see Theorem 2.5) and Subsection 2.4 (see Theorem 2.8), respectively. Note that this section rather takes an abstract viewpoint, whereas applications and examples of our framework are elaborated in Section 3.

2.1 Formal Model Setup

Let us begin with a rigorous definition of the random sampling process considered in the introduction. For the sake of convenience, we have summarized the our terminology in Table 1.

Assumption 2.1 (Sampling process) Let (s, y) be a joint random pair in $\mathbb{R}^d \times \mathbb{R}$, where y is a sub-Gaussian random variable and s is an isotropic, mean-zero sub-Gaussian random vector in \mathbb{R}^d with $\|s\|_{\psi_2} \leq \kappa$ for some $\kappa > 0$. Moreover, let $A \in \mathbb{R}^{p \times d}$ be a (deterministic) matrix and define the input random vector in \mathbb{R}^p as $x := As$. The *observed samples* $\{(x_i, y_i)\}_{i=1}^n$ are now generated as follows: Let $\{(s_i, y_i)\}_{i=1}^n$ be independent copies of (s, y) . Then we set $x_i := As_i$ for $i = 1, \dots, n$. In particular, $\{(x_i, y_i)\}_{i=1}^n$ can be also regarded as independent samples of (x, y) .

| Term | Notation | Sampling notation |
|--|---|---|
| Output variable (dependent variable, response variable, label) | $y \in \mathbb{R}$ | y_1, \dots, y_n |
| Latent variables (hidden variables, factors) | $\mathbf{s} \in \mathbb{R}^d$ | $\mathbf{s}_1, \dots, \mathbf{s}_n$ |
| Mixing matrix (pattern matrix, factor loadings) | $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_d] \in \mathbb{R}^{p \times d}$ | |
| Input variables (independent variables, explanatory variables, data variables, features, covariates, predictors) | $\mathbf{x} = \mathbf{A}\mathbf{s} \in \mathbb{R}^p$ | $\mathbf{x}_1 = \mathbf{A}\mathbf{s}_1, \dots, \mathbf{x}_n = \mathbf{A}\mathbf{s}_n$ |
| Observation (data, sample pair) | $(\mathbf{x}, y) = (\mathbf{A}\mathbf{s}, y) \in \mathbb{R}^p \times \mathbb{R}$ | $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ |
| Hypothesis set (constraint set, hypothesis class, models) | $K \subset \mathbb{R}^p$ convex | |
| Target set | $T \subset \mathbb{R}^d$ | |
| (Latent) Parameter vectors | $\boldsymbol{\beta} \in K, \mathbf{z} \in \mathbf{A}^\top K \subset \mathbb{R}^d$ | |

Table 1: A summary of frequently used notations. The terms in parentheses are widely-used synonyms.

As usual in statistical learning, Assumption 2.1 leaves the output variable y largely unspecified and only makes a regularity assumption, namely that the tail of y is sub-Gaussian. But it is still useful to keep in mind our running examples from (1.5) and (1.6), which indicate that y typically depends on the latent factors \mathbf{s} according to a certain (semi-)parametric model. Moreover, Assumption 2.1 builds upon a simple linear factor model for the input variables, i.e., $\mathbf{x} = \mathbf{A}\mathbf{s}$ with \mathbf{s} being isotropic. The following proposition shows that this is actually not a severe restriction.

Proposition 2.2 *Let \mathbf{x} be a mean-zero random vector in \mathbb{R}^p whose covariance matrix is of rank d . Then there exists a deterministic matrix $\mathbf{A} \in \mathbb{R}^{p \times d}$ and an isotropic, mean-zero random vector \mathbf{s} in \mathbb{R}^d such that $\mathbf{x} = \mathbf{A}\mathbf{s}$ almost surely.*

A proof of Proposition 2.2 is given in Subsection 6.3. Nevertheless, this statement of existence is only of limited practical relevance because the resulting latent factors are not necessarily linked to y in a meaningful way. Indeed, while the above isotropic decomposition of \mathbf{x} is highly non-unique, the actual output variable may exhibit a very specific structure, such as in (1.5) or (1.6). We therefore rather take the viewpoint that the factorization $\mathbf{x} = \mathbf{A}\mathbf{s}$ is prespecified, say by a certain data acquisition process or a physical law. In particular, our purpose is *not* to learn the entries of \mathbf{A} , but to consider them as possibly unknown model parameters (cf. Subsection 4.3). Finally, let us point out that the situation of $p < d$ can be also of interest, e.g., when modeling noisy data as in Subsection 3.2.2.

2.2 The Mismatch Parameters and Gaussian Mean Width

We now introduce the technical ingredients of our general error bounds below. Let us begin with a formal definition of the mismatch parameters, which are supposed to quantify the model mismatch that occurs when approximating the true output y by a linear hypothesis function of the form $f(\mathbf{s}) := \langle \mathbf{s}, \mathbf{z}^\natural \rangle$.

Definition 2.3 (Mismatch parameters) Let Assumption 2.1 be satisfied and let $\mathbf{z}^\natural \in \mathbb{R}^d$. Then we define the following two parameters (as a function of \mathbf{z}^\natural):

$$\text{Mismatch covariance:} \quad \rho(\mathbf{z}^\natural) := \rho_{\mathbf{s}, y}(\mathbf{z}^\natural) := \|\mathbb{E}[(y - \langle \mathbf{s}, \mathbf{z}^\natural \rangle)\mathbf{s}]\|_2,$$

$$\text{Mismatch deviation:} \quad \sigma(\mathbf{z}^\natural) := \sigma_{\mathbf{s}, y}(\mathbf{z}^\natural) := \|y - \langle \mathbf{s}, \mathbf{z}^\natural \rangle\|_{\psi_2}.$$

The meaning of the mismatch covariance was already discussed in the course of Subsection 1.1: it is useful to regard $\rho(\mathbf{z}^\natural)$ as a measure of how close \mathbf{z}^\natural is to satisfying the optimality condition of (1.7), which corresponds to the classical orthogonality principle. Compared to that, the mismatch deviation of \mathbf{z}^\natural captures the sub-Gaussian tail behavior of the model mismatch $y - \langle \mathbf{s}, \mathbf{z}^\natural \rangle$. The meaning of this parameter is best illustrated by the case of noisy linear observations, i.e., $y = \langle \mathbf{s}, \mathbf{z}^\natural \rangle + e$, so that $\sigma(\mathbf{z}^\natural) = \|e\|_{\psi_2}$ would basically measure the power of noise. Note that Definition 2.3 does only rely on the latent factors \mathbf{s} , but not on the mixing matrix A . The role of A will become clear in Subsection 2.3 below when combining the concepts of target and hypothesis sets.

Our next definition introduces the (Gaussian) mean width, which is a widely-used notion of complexity for hypothesis sets (cf. (Q3)).

Definition 2.4 (Mean width) Let $L \subset \mathbb{R}^d$ be a subset and let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, I_d)$ be a standard Gaussian random vector. The (global) mean width of L is given by

$$w(L) := \mathbb{E} \left[\sup_{\mathbf{h} \in L} \langle \mathbf{g}, \mathbf{h} \rangle \right].$$

Moreover, we define the conic mean width of L as

$$w_\wedge(L) := w(\text{cone}(L) \cap \mathbb{S}^{d-1}).$$

These parameters originate from geometric functional analysis and convex geometry (e.g., see [GM04; Gor85; Gor88]) and recently turned out to be very useful for the statistical analysis of various signal estimation problems [ALMT14; BM02; CRPW12; GS18; MPT07; PV13b; PVY16; Ver15]. We refer the reader to these works for a more extensive discussion and basic properties. In what follows below, we will treat the (conic) mean width as an abstract quantity that measures the size of the hypothesis set K in (P_K) and thereby determines the approximation rate of our error bounds. Some concrete examples of hypothesis sets (related to sparsity priors) are presented in Subsection 3.1.5. Finally, it is worth mentioning that the conic mean width is essentially equivalent to the notion of *statistical dimension*, which is well known for characterizing the phase transition behavior of many convex programs with Gaussian input data; see [ALMT14] and particularly Proposition 10.2 therein.

2.3 General Error Bounds and the Mismatch Principle

For the sake of convenience, let us restate the constrained empirical risk minimization problem which is in the focus of our theoretical study:

$$\min_{\boldsymbol{\beta} \in K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2. \quad (\text{P}_K)$$

Our first error bound for (P_K) is based on the global mean width and extends Theorem 1.1 with regard to arbitrary mixing matrices. Its proof is contained in Subsection 6.1 and is a consequence of a more general result (Theorem 6.3) that employs the concept of local mean width.

Theorem 2.5 (Estimation via (P_K) – Global version) *Let Assumption 2.1 be satisfied. Let $K \subset \mathbb{R}^p$ be a bounded, convex subset and fix a vector $\mathbf{z}^\natural \in A^\top K \subset \mathbb{R}^d$. Then there exists a numerical constant $C > 0$ such that for every $\delta \in (0, 1]$, the following holds true with probability at least $1 - 5 \exp(-C \cdot \kappa^{-4} \cdot \delta^2 \cdot n)$.¹ If the number of observed samples obeys*

$$n \gtrsim \kappa^4 \cdot \delta^{-4} \cdot w^2(A^\top K), \quad (2.1)$$

then every minimizer $\hat{\boldsymbol{\beta}}$ of (P_K) satisfies

$$\|A^\top \hat{\boldsymbol{\beta}} - \mathbf{z}^\natural\|_2 \lesssim \max\{1, \kappa \cdot \sigma(\mathbf{z}^\natural)\} \cdot \delta + \rho(\mathbf{z}^\natural). \quad (2.2)$$

¹More precisely, the probability refers to the samples $\{(x_i, y_i)\}_{i=1}^n$, and we condition on this random set in the assertion of Theorem 2.5.

The parameter δ can be regarded as an oversampling factor in the sense that it enables a trade-off between the number of required samples in (2.1) and the accuracy of (2.2). Adjusting δ such that (2.1) holds true with equality, we can rephrase (2.2) as a more convenient error bound that explicitly depends on n :

$$\|A^\top \hat{\beta} - z^\natural\|_2 \lesssim \max\{\kappa, \kappa^2 \cdot \sigma(z^\natural)\} \cdot \left(\frac{w^2(A^\top K)}{n}\right)^{1/4} + \rho(z^\natural). \quad (2.3)$$

This expression refines the statement of Theorem 1.1 and reveals that first term in (2.2) is actually of the order $O(n^{-1/4})$. From this, we can particularly conclude that the first error term does not concern the consistency of the estimator, but nevertheless it has a huge impact on the sampling rate. In contrast, the mismatch covariance $\rho(z^\natural)$ does not depend on n , implying that the second error term controls the *bias* of (P_K) when approximating z^\natural . Remarkably, the individual parameters of (2.3) are assigned to very different roles: while the mean width $w(A^\top K)$ only involves the (transformed) hypothesis set $A^\top K$, the mismatch parameters $\rho(z^\natural)$ and $\sigma(z^\natural)$ do exclusively consider the output variable y and the choice of target vector z^\natural . Apart from that, κ can be regarded as a model constant which bounds the sub-Gaussian tail of the input vector s .

Before further discussing the significance of the error bound (2.2) and deriving the mismatch principle (Recipe 2.6), let us briefly investigate how the mixing matrix A affects the assertion of Theorem 2.5. For this purpose, we first make the following important yet simple observation about the solution set of (P_K) :

$$A^\top \cdot \left(\operatorname{argmin}_{\beta \in K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2 \right) = \operatorname{argmin}_{z \in A^\top K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle s_i, z \rangle)^2, \quad (2.4)$$

which follows from the identity $\langle x_i, \beta \rangle = \langle s_i, A^\top \beta \rangle$ and substituting $A^\top \beta$ by z . As already pointed out in the course of (1.4), solving (P_K) may lead to a highly non-unique minimizer, but by a linear transformation according to (2.4), it becomes related to an optimization problem that just takes *isotropic* inputs $s_1, \dots, s_n \in \mathbb{R}^d$. This particularly explains why Theorem 2.5 establishes an error bound on a parameter vector z^\natural in the “latent space” \mathbb{R}^d and why the estimator actually takes the form $\hat{z} := A^\top \hat{\beta}$. Moreover, (2.4) indicates that, in order to better understand the estimation performance of (P_K) , it essentially suffices to study

$$\min_{z \in A^\top K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle s_i, z \rangle)^2. \quad (2.5)$$

The isotropy of data in (2.5) indeed simplifies the statistical analysis, whereas the intricacy of the mixing matrix is now concealed by the transformed hypothesis set $A^\top K$. For that specific reason, the complexity of K is measured by means of $w(A^\top K)$ in (2.1), rather than $w(K)$.

However, the original estimator (P_K) is still of great practical interest because it neither requires knowledge of A nor of the s_i . Even if A would be exactly known, an extraction of the latent factors can be difficult and unstable, implying that potential errors would propagate to the inputs of (2.5). Instead, (2.4) suggests to first solve the “unspoiled” optimization problem (P_K) and then to appropriately reweight the outcome by A^\top as a post-processing step. Our discussion on the mixing matrix is continued in Subsection 3.2 and Subsection 4.3, presenting several implications of Theorem 2.5 and related approaches. In this context, we demonstrate that estimation via (P_K) is still feasible when A is only approximately known, and by carefully adapting the hypothesis set, one can even obtain error bounds for the untransformed minimizer $\hat{\beta}$.

The Mismatch Principle

Similar to the informal statement of Theorem 1.1, a key feature of Theorem 2.5 is that it permits to freely select a target vector $z^\natural \in A^\top K$. This flexibility is an essential advantage over the traditional viewpoint of statistical learning theory, which would simply recommend to set $z^\natural = z^* := A^\top \beta^*$, where $f^* = \langle \cdot, \beta^* \rangle$ is an expected risk minimizer of (1.1), see also (Q1). Indeed, according to (Q2) and Subsection 1.1, learning a semi-parametric model does also require a certain interpretability

of the target vector, in the sense that it encodes the parameters of interest, e.g., an index vector (cf. (1.5)) or the set of active variables (cf. (1.6)). For that purpose, we have introduced the abstract notion of a *target set* $T \subset \mathbb{R}^d$, which contains all those vectors that a user would consider as desirable outcome of an estimation procedure.¹ It is therefore more natural to ask whether (P_K) is capable of estimating any vector in T , rather than $z^* = A^\top \beta^*$. This important concern is precisely addressed by the mismatch principle:

Recipe 2.6 (Mismatch principle) Let Assumption 2.1 be satisfied.

- (1) *Target set.* Select a subset $T \subset \mathbb{R}^d$ containing all admissible target vectors associated with the model (s, y) .
- (2) *Hypothesis set.* Select a subset $K \subset \mathbb{R}^p$ imposing structural constraints on the solution set of (P_K) .
- (3) *Target vector.* Minimize the mismatch covariance on $T \cap A^\top K$, i.e., specify

$$z^\natural = \operatorname{argmin}_{z \in T \cap A^\top K} \rho(z). \quad (2.6)$$

- (4) *Estimation performance.* Invoke (2.2) in Theorem 2.5 to obtain a bound on the estimation error of z^\natural via (P_K) .

While the steps (1) and (2) are related to the underlying model specifications, the actual key step is (3): By restricting the objective set to $T \cap A^\top K$, we ensure that the target vector z^\natural carries the desired (parametric) information and is compatible with the (transformed) hypothesis set $A^\top K$ of the optimization problem (2.5). Under these constraints every minimizer of the mismatch covariance then leads to the smallest possible bias term in (2.2). Meanwhile, the sampling-related term in (2.2) is only slightly affected by the size of the mismatch deviation $\sigma(z^\natural)$. In other words, the mismatch principle simply selects an admissible target vector in a such way that Theorem 2.5 yields the best outcome. Studying the significance of the resulting error bound eventually allows us to draw a conclusion on the ability of (P_K) to solve the examined estimation problem or not. We close this subsection with a useful simplification of Recipe 2.6 that is applicable whenever one has $T \subset \operatorname{cone}(A^\top K)$:

Remark 2.7 (How to apply the mismatch principle?) Invoking Recipe 2.6(3) can be a challenging step because $\rho(\cdot)$ needs to be minimized on the intersection $T \cap A^\top K$. On the other hand, the hypothesis set K can be freely chosen in practice. Thus, as long as $T \subset \operatorname{cone}(A^\top K)$, one can simplify the mismatch principle as follows:

*First minimize the mismatch covariance on the target set T , i.e., set $z^\natural = \operatorname{argmin}_{z \in T} \rho(z)$.
Then specify a scaling factor $\lambda > 0$ for the hypothesis set K such that $z^\natural \in A^\top(\lambda K)$.*

The target vector z^\natural is now admissible for step (3), when using λK as hypothesis set. Consequently, (2.2) states an error bound for the rescaled estimator $(P_{\lambda K})$, while the number of required samples is only enlarged by a quadratic factor of λ^2 . To some extent, the above strategy allows us to decouple the tasks of minimizing the mismatch covariance (on T) and selecting an appropriate hypothesis set. This simplification will turn out to be very beneficial in Subsection 3.1, where we apply the mismatch principle to various types of output models but leave K unspecified.

However, let us emphasize that the choice of K and the scaling parameter λ both remain important practical issues. In view of our approach, it is helpful to consider the hypothesis set as a means to exploit additional structural knowledge about the observation model. A typical example is a *sparse* single-index model, for which the index vector $z_0 \in \mathbb{R}^d$ in (1.5) is unknown but its support is small. In this case, the uncertainty about z_0 would be just captured by $T := \operatorname{span}\{z_0\}$,

¹Note that the target set is part of the “latent space” \mathbb{R}^d , which is due to the fact that we state semi-parameters output models for y in terms of the latent factors $s \in \mathbb{R}^d$ and not in terms of the correlated data $x = As \in \mathbb{R}^p$.

whereas an ℓ^1 -constraint could serve as sparsity prior (cf. Subsection 3.1.5). But nevertheless, finding a suitable value for λ is not straightforward in practice and might require careful tuning of the estimator, since the target vector \mathbf{z}^\natural is still unknown in general. \diamond

2.4 Refined Error Bounds Based on the Conic Mean Width

The above discussion on Theorem 2.5 indicates that the hypothesis set K does only play a minor role for the estimation performance of (\mathbf{P}_K) , see also (Q3). This fallacious impression is especially due to the fact that the complexity parameter $w(\mathbf{A}^\top K)$ in (2.1) does not explicitly depend on \mathbf{z}^\natural . The proposed sample size therefore remains widely unaffected by the choice of the target vector, but on the other hand, the involved oversampling factor of δ^{-4} is clearly suboptimal. Our next main result tackles this issue and in fact achieves the optimal factor of δ^{-2} . The key difference to Theorem 2.5 is that the complexity of the (transformed) hypothesis set $\mathbf{A}^\top K$ is now measured *locally* at the target vector \mathbf{z}^\natural .

Theorem 2.8 (Estimation via (\mathbf{P}_K) – Conic version) *Let Assumption 2.1 be satisfied. Let $K \subset \mathbb{R}^p$ be a convex subset and fix a vector $\mathbf{z}^\natural \in \mathbf{A}^\top K \subset \mathbb{R}^d$. Then there exists a numerical constant $C > 0$ such that for every $\delta \in (0, 1]$, the following holds true with probability at least $1 - 5 \exp(-C \cdot \kappa^{-4} \cdot \delta^2 \cdot n)$: If the number of observed samples obeys*

$$n \gtrsim \kappa^4 \cdot \delta^{-2} \cdot w_\lambda^2(\mathbf{A}^\top K - \mathbf{z}^\natural), \quad (2.7)$$

then every minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) satisfies

$$\|\mathbf{A}^\top \hat{\boldsymbol{\beta}} - \mathbf{z}^\natural\|_2 \lesssim \kappa^{-1} \cdot \sigma(\mathbf{z}^\natural) \cdot \delta + \rho(\mathbf{z}^\natural). \quad (2.8)$$

Analogously to (2.3), we can restate (2.8) as a sample-dependent error bound:

$$\|\mathbf{A}^\top \hat{\boldsymbol{\beta}} - \mathbf{z}^\natural\|_2 \lesssim \kappa \cdot \sigma(\mathbf{z}^\natural) \cdot \left(\frac{w_\lambda^2(\mathbf{A}^\top K - \mathbf{z}^\natural)}{n} \right)^{1/2} + \rho(\mathbf{z}^\natural), \quad (2.9)$$

showing that the decay rate $O(n^{-1/2})$ of the first term is optimal with respect to n .¹ The price to pay for this improvement is the presence of a more complicated complexity parameter. Indeed, using the conic mean width $w_\lambda(\mathbf{A}^\top K - \mathbf{z}^\natural)$ instead of its global counterpart $w(\mathbf{A}^\top K)$ comes along with two difficulties:

- *Computability.* The conic mean width is a quite implicit parameter that is exactly computable only in special cases. This issue is further complicated by the deformation of mixing matrix \mathbf{A} : Our primary goal is to exploit low-complexity features of \mathbf{z}^\natural , whereas the actual hypothesis set K merely restricts the solution space of (\mathbf{P}_K) . Hence, in order to compute or accurately bound $w_\lambda(\mathbf{A}^\top K - \mathbf{z}^\natural)$, the impact of \mathbf{A}^\top needs to be precisely understood.
- *Stability.* The value of $w_\lambda(\mathbf{A}^\top K - \mathbf{z}^\natural)$ is highly sensitive to the position of \mathbf{z}^\natural within $\mathbf{A}^\top K$. For example, if the set $\mathbf{A}^\top K$ is full-dimensional and \mathbf{z}^\natural does not lie exactly on its boundary, one would simply end up with $w_\lambda^2(\mathbf{A}^\top K - \mathbf{z}^\natural) \asymp d$. This leads to an overly pessimistic error bound in (2.9), which does not reflect any benefit of using a “small” hypothesis set.

The latter point particularly implies a challenging tuning problem: compared the approach of Remark 2.7, it does not suffice to just achieve $\mathbf{z}^\natural \in \mathbf{A}^\top(\lambda K)$ for some $\lambda > 0$, but one even requires that $\mathbf{z}^\natural \in \partial(\mathbf{A}^\top(\lambda K))$. Fortunately, such a perfect tuning is usually not necessary in practice and the estimator (\mathbf{P}_K) turns out to be quite stable under these types of model inaccuracies. The above issue of stability is in fact rather an artifact of employing the conic mean width as complexity measure. Based on the refined notion of *local mean width* (see Definition 6.1), we prove a more general version of Theorem 2.8 in Subsection 6.2 which does not suffer from this drawback. This

¹More precisely, we mean that the rate $O(n^{-1/2})$ is optimal in the sense of [PVY16, Sec. 4]: If the output y would be linear in \mathbf{s} with independent Gaussian noise, no estimator can achieve a better rate. Thus, if y obeys a non-linear (and therefore more “complicated”) model, one cannot expect a better estimation performance than in the linear case.

refined error estimate (Theorem 6.3) actually forms the basis for proving Theorem 2.5 as well as Theorem 2.8, and in principle, it allows us to establish more advanced results, e.g., on the stability of (P_K) . However, for the sake of brevity, we have decided to omit a detailed discussion on these extensions and refer the interested reader to [GKM17, Sec. 6.1 + Lem. A.2].

Finally, let us mention that Recipe 2.6(4) is of course also applicable to Theorem 2.8, but as pointed out above, the simplification of Remark 2.7 would only work poorly in this case. For that reason, the examples of the mismatch principle presented in Section 3 are mainly based on an application of Theorem 2.5.

3 Applications and Examples

This section elaborates several applications of the general framework developed in Section 2. In the first part (Subsection 3.1), we still focus on isotropic input data (i.e., $A = I_p$ in Assumption 2.1) and apply the mismatch principle to various types of semi-parametric output models. In particular, we revisit our two prototypical examples from the introduction, that is, single-index models (1.5) and variable selection (1.6). Subsection 3.2 is then devoted to issues and challenges associated with correlated feature variables (i.e., $A \neq I_p$), distinguishing the underdetermined ($p \geq d$) and overdetermined case ($p < d$).

3.1 Semi-Parametric Output Rules

In the following subsections, we investigate several popular examples of semi-parametric models. While the underlying output rules for y actually rely on very different structural assumptions, the common goal is to study the capability of (P_K) to estimate a certain set of “ground truth” parameters. In this context, the mismatch principle (Recipe 2.6) turns out to be very useful, as it allows us to deal with each of the cases in a highly systematic way.

In order to keep the exposition simple and clear, we will only consider the situation of isotropic input vectors, i.e., Assumption 2.1 is fulfilled with $A = I_d$. Consequently, it holds that $x = s$ and $x_i = s_i$ for $i = 1, \dots, n$, and $\hat{\beta} = \hat{z}$. Adapting the notation from Subsection 1.1, we particularly agree on writing s, s_i, z, \hat{z} instead of $x, x_i, \beta, \hat{\beta}$, respectively, so that the estimator (P_K) takes the form

$$\min_{z \in K} \frac{1}{n} \sum_{i=1}^n (y_i - \langle s_i, z \rangle)^2. \quad (\text{P}_K^{\text{iso}})$$

For the sake of convenience, let us also restate Theorem 2.5 in this specific setup (which is in fact the formal version of Theorem 1.1):

Corollary 3.1 (Estimation via (P_K) – Isotropic data) *Let Assumption 2.1 be satisfied with $A = I_d$. Let $K \subset \mathbb{R}^d$ be a bounded, convex subset and fix a vector $z^{\natural} \in K$. Then there exists a numerical constant $C > 0$ such that for every $\delta \in (0, 1]$, the following holds true with probability at least $1 - 5 \exp(-C \cdot \kappa^{-4} \cdot \delta^2 \cdot n)$: If the number of observed samples obeys*

$$n \gtrsim \kappa^4 \cdot \delta^{-4} \cdot w^2(K), \quad (3.1)$$

then every minimizer \hat{z} of (P_K^{iso}) satisfies

$$\|\hat{z} - z^{\natural}\|_2 \lesssim \max\{1, \kappa \cdot \sigma(z^{\natural})\} \cdot \delta + \rho(z^{\natural}). \quad (3.2)$$

Note that the subsequent analysis of output models does only concern step (1) and step (3) of Recipe 2.6. Indeed, according to the simplification of Remark 2.7, it already suffices to analyze the mismatch covariance on the target set, whereas the hypothesis set can be left unspecified. Some examples of K are however discussed in Subsection 3.1.5 below.

3.1.1 Single-Index Models

Let us begin with defining a single-index model, which was already informally introduced in (1.5):

Assumption 3.2 (Single-index models) Let Assumption 2.1 be satisfied with $A = I_d$ and let

$$y = g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \quad (3.3)$$

where $\mathbf{z}_0 \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ is an (unknown) *index vector* and $g: \mathbb{R} \rightarrow \mathbb{R}$ is a measurable scalar *output function* which can be random (independently of \mathbf{s}).¹

Since Assumption 3.2 imposes rather mild restrictions on the output function g , (3.3) covers many model situations of practical interest, e.g., noisy linear regression ($g = \text{Id} + e$ with noise e), 1-bit compressed sensing ($g = \text{sign}(\cdot)$), linear classification, or phase retrieval ($g = |\cdot|$). Apart from these applications, single-index models are also of great relevance in the field of econometrics [Hor09]. More conceptually, g plays the role of a (non-parametric) model uncertainty that is unknown to the estimator ($\mathbf{P}_K^{\text{iso}}$). If g is non-linear, we particularly cannot expect that the mean squared error in ($\mathbf{P}_K^{\text{iso}}$) is small, no matter what parameter vector $\mathbf{z} \in K$ is selected. Fortunately, this issue does not bother us too much because our actual goal is to learn the index vector \mathbf{z}_0 , but not the true output variable y .

For that reason, it appears quite natural to choose $T = \text{span}\{\mathbf{z}_0\}$ as target set. We emphasize that using just a singleton, e.g., $T = \{\mathbf{z}_0\}$, is not necessarily meaningful, since the magnitude of \mathbf{z}_0 might be “absorbed” by g , for instance, if $g = \text{sign}(\cdot)$. Hence, without any further assumptions on g , the best we can expect is to recover the direction (or a scalar multiple) of \mathbf{z}_0 . Invoking step (3) of Recipe 2.6 now leads to the following outcome:

Proposition 3.3 *Let Assumption 3.2 be satisfied and set $T := \text{span}\{\mathbf{z}_0\}$. Then $\mathbf{z}^\natural := \mu \mathbf{z}_0 \in T$ with*

$$\mu := \frac{1}{\|\mathbf{z}_0\|_2^2} \mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \langle \mathbf{s}, \mathbf{z}_0 \rangle]$$

minimizes the mismatch covariance on T and we have

$$\rho(\mathbf{z}^\natural) = \|\mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}]\|_2.$$

In particular, if $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, I_d)$, we have $\rho(\mathbf{z}^\natural) = 0$.

Proof. Let us make the ansatz $\mathbf{z}^\natural = \mu \mathbf{z}_0 \in T$ where $\mu \in \mathbb{R}$ is specified later on. Using the orthogonal decomposition

$$\mathbf{s} = \mathbf{P}_{\mathbf{z}_0} \mathbf{s} + \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s} = \langle \mathbf{s}, \frac{\mathbf{z}_0}{\|\mathbf{z}_0\|_2} \rangle \frac{\mathbf{z}_0}{\|\mathbf{z}_0\|_2} + \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s},$$

the mismatch covariance of \mathbf{z}^\natural simplifies as follows:

$$\begin{aligned} \rho(\mathbf{z}^\natural)^2 &= \|\mathbb{E}[(y - \langle \mathbf{s}, \mu \mathbf{z}_0 \rangle) \mathbf{s}]\|_2^2 = \|\mathbb{E}[(g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) - \langle \mathbf{s}, \mu \mathbf{z}_0 \rangle) (\langle \mathbf{s}, \frac{\mathbf{z}_0}{\|\mathbf{z}_0\|_2} \rangle \frac{\mathbf{z}_0}{\|\mathbf{z}_0\|_2} + \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s})]\|_2^2 \\ &= \left\| \frac{1}{\|\mathbf{z}_0\|_2^2} \mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \langle \mathbf{s}, \mathbf{z}_0 \rangle] \mathbf{z}_0 - \frac{\mu}{\|\mathbf{z}_0\|_2^2} \underbrace{\mathbb{E}[\langle \mathbf{s}, \mathbf{z}_0 \rangle^2]}_{=\|\mathbf{z}_0\|_2^2} \mathbf{z}_0 \right\|_2^2 + \|\mathbb{E}[(g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) - \langle \mathbf{s}, \mu \mathbf{z}_0 \rangle) \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}]\|_2^2 \\ &= \left(\frac{1}{\|\mathbf{z}_0\|_2^2} \mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \langle \mathbf{s}, \mathbf{z}_0 \rangle] - \mu \right)^2 \cdot \|\mathbf{z}_0\|_2^2 + \|\mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}]\|_2^2, \end{aligned} \quad (3.4)$$

where we have used that $\mathbb{E}[\langle \mathbf{s}, \mu \mathbf{z}_0 \rangle \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}] = \mathbf{0}$ due to the isotropy of \mathbf{s} . Since the second summand in (3.4) does not depend on μ , the minimum of $\mu \mapsto \rho(\mu \mathbf{z}_0)$ is indeed attained at $\mu = \mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \langle \mathbf{s}, \mathbf{z}_0 \rangle] / \|\mathbf{z}_0\|_2^2$.

Finally, if $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, I_d)$, the random vector $\mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}$ is independent (not just uncorrelated) from $\langle \mathbf{s}, \mathbf{z}_0 \rangle$, so that $\mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}] = \mathbf{0}$. \blacksquare

The above proof may serve as a template for applying the mismatch principle: First, make a parametric ansatz according to the target set and simplify the mismatch covariance as much as possible. Then select the free parameters such that the simplified expression is minimized (or at least gets small). We will see below that this strategy works out for many other examples as well.

¹Note that we implicitly assume that y is a sub-Gaussian variable.

The Gaussian case of Proposition 3.3 leads to a very desirable situation, as it states that single-index models can be *consistently* learned via (P_K^{iso}) if $s \sim \mathcal{N}(\mathbf{0}, I_d)$. This conclusion precisely corresponds to a recent finding of Plan and Vershynin from [PV16], which relies on a calculation similar to (3.4). But beyond that, Proposition 3.3 even extends the results of [PV16] to sub-Gaussian input variables, where the situation is in fact more complicated: In general, the mismatch covariance does not necessarily vanish at $z^\natural = \mu z_0$, implying that (P_K^{iso}) does not constitute a consistent estimator of any target vector in $T = \text{span}\{z_0\}$. Consequently, according to Corollary 3.1, the actual learning capacity of (P_K^{iso}) is specified by the expression $\rho(z^\natural) = \|\mathbb{E}[g(\langle s, z_0 \rangle) P_{z_0}^\perp s]\|_2$, which in a certain sense measures the compatibility between the non-linear output model and the isotropic data vector. We will return to this issue in the course of variable selection (see Remark 3.11), showing that asymptotically unbiased estimates can be still achieved as long as one is only aiming at the support of z_0 .

Let us conclude our discussion with some interesting model setups for which the outcome of Proposition 3.3 is well interpretable:

Example 3.4 (1) *Rotationally invariant distributions.* This class of probability distributions is a natural generalization of Gaussian random vectors, according to which s takes the form ru , where the radius r is sub-Gaussian on $(0, \infty)$ and u is uniformly distributed on S^{d-1} . In the context of single-index models, this scenario was recently studied by Goldstein et al. in [GMW16], constructing a consistent estimator that is equivalent to (P_K) . A simple calculation confirms their result by means of Proposition 3.3: For every $z \in \mathbb{R}^d$, it holds that

$$\begin{aligned} \mathbb{E}[g(\langle s, z_0 \rangle) \langle P_{z_0}^\perp s, z \rangle] &= \mathbb{E}[\mathbb{E}[g(\langle s, z_0 \rangle) \langle s, P_{z_0}^\perp z \rangle \mid \langle s, z_0 \rangle]] \\ &= \mathbb{E}[g(\langle s, z_0 \rangle) \cdot \underbrace{\mathbb{E}[\langle s, P_{z_0}^\perp z \rangle \mid \langle s, z_0 \rangle]}_{= \frac{1}{\|z_0\|_2^2} \langle P_{z_0}^\perp z, z_0 \rangle \langle s, z_0 \rangle = 0}] = 0, \end{aligned}$$

where we have used the law of total expectation and [GMW16, Cor. 2.1]. Hence, we conclude that $\rho(z^\natural) = \|\mathbb{E}[g(\langle s, z_0 \rangle) P_{z_0}^\perp s]\|_2 = 0$.

(2) *Linear regression.* The most simple class of single-index models are linear observations of the form $y = g(\langle s, z_0 \rangle) := \langle s, z_0 \rangle + e$ with e being independent, mean-zero sub-Gaussian noise. In this case, it is not hard to see that Proposition 3.3 yields $\mu = 1$ and $\rho(z_0) = 0$. Note that, compared to non-linear output functions, μ does not depend on $\|z_0\|_2$ here. Due to $\sigma(z_0) = \|e\|_{\psi_2}$, the error bound (3.2) of Corollary 3.1 is determined by the strength of noise, which is a well-known fact from the literature. In particular, one can achieve exact recovery in the noiseless case $e = 0$.

(3) *Worst case error bounds.* While the latter two examples lead to desirable situations, there exist model configurations where learning the index vector fails. For instance, if the entries of s are Rademacher distributed, recovery of extremely sparse signals from noiseless 1-bit measurements is impossible ($g = \text{sign}(\cdot)$), regardless of the considered estimator [PV13a, Rmk. 1.5]. We encourage the interested reader to verify that the mismatch covariance is indeed very badly behaved in this case.

On the other hand, it turned out in [ALPV14] that these types of adversarial examples can be excluded by worse case error bounds. Although the results of [ALPV14] are actually related to a simple linear estimator, we believe that the applied techniques could be used to derive an upper bound for $\rho(z^\natural)$ as well. The assertion of Proposition 3.3 would then provide a means to determine those index vectors which can(not) be accurately approximated by (P_K^{iso}) . Another interesting approach was recently taken by Dirksen and Mendelson [DM18], demonstrating that consistency in 1-bit compressed sensing can be still achieved by *dithering*, i.e., quantizing at uniformly distributed thresholds.

(4) *Even output functions.* If g is an even function and $s \sim \mathcal{N}(\mathbf{0}, I_d)$, one ends up with $\mu = 0$ in Proposition 3.3. Thus, Corollary 3.1 states that (P_K^{iso}) simply approximates the null vector, which is clearly not what we are aiming at. This negative result nevertheless underpins the versatility of the mismatch principle because it reflects that linear hypothesis functions are not suited for problems like phase retrieval ($g = |\cdot|$). Based on a lifting argument, a recent work by Yang et

al. [YBL17b; YBWL17] shows that learning single-index models is still feasible for even output functions (or more generally, second-order output functions), but their method comes along with the price of a sub-optimal sampling rate in the sparse case. \diamond

3.1.2 Generalized Linear Models

Generalized linear models form a natural extension of linear regression problems and are conceptually very similar to single-index models (cf. [MN89; NW72]). Let us begin with a formal definition:¹

Assumption 3.5 (Generalized linear models) Let Assumption 2.1 be satisfied with $A = I_d$ and let the output variable obey

$$\mathbb{E}[y \mid \mathbf{s}] = g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \quad (3.5)$$

where $\mathbf{z}_0 \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ is an (unknown) *parameter vector* and $g: \mathbb{R} \rightarrow \mathbb{R}$ is a measurable scalar *link function*.

Compared to the single-index model from (3.3), the condition in (3.5) does not concern the output y itself, but it still assumes a single-index structure for the expectation of y conditional on the input data \mathbf{s} . In that way, one may incorporate different (and more general) noise models compared to Assumption 3.2. But despite their strong resemblance, we point out that these concepts are not completely equivalent; see [PVY16, Sec. 6] for a more detailed comparison. On the other hand, the assertion of Proposition 3.3 remains literally true for generalized linear models, so that we can expect similar recovery guarantees for the parameter vector \mathbf{z}_0 :

Proposition 3.6 Let Assumption 3.5 be satisfied and set $T := \text{span}\{\mathbf{z}_0\}$. Then $\mathbf{z}^\natural := \mu \mathbf{z}_0 \in T$ with

$$\mu := \frac{1}{\|\mathbf{z}_0\|_2^2} \mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \langle \mathbf{s}, \mathbf{z}_0 \rangle]$$

minimizes the mismatch covariance on T and we have

$$\rho(\mathbf{z}^\natural) = \|\mathbb{E}[g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) \mathbf{P}_{\mathbf{z}_0}^\perp \mathbf{s}]\|_2.$$

Proof. Using the law of total expectation, we observe that

$$\begin{aligned} \rho(\mathbf{z}^\natural) &= \|\mathbb{E}[(y - \langle \mathbf{s}, \mu \mathbf{z}_0 \rangle) \mathbf{s}]\|_2 = \|\mathbb{E}[\mathbb{E}[(y - \langle \mathbf{s}, \mu \mathbf{z}_0 \rangle) \mathbf{s} \mid \mathbf{s}]]\|_2 \\ &= \|\mathbb{E}[\underbrace{\mathbb{E}[y \mid \mathbf{s}]}_{=g(\langle \mathbf{s}, \mathbf{z}_0 \rangle)} \cdot \mathbf{s}] - \mathbb{E}[\langle \mathbf{s}, \mu \mathbf{z}_0 \rangle \mathbf{s}]\|_2 = \|\mathbb{E}[(g(\langle \mathbf{s}, \mathbf{z}_0 \rangle) - \langle \mathbf{s}, \mu \mathbf{z}_0 \rangle) \mathbf{s}]\|_2. \end{aligned}$$

Now, we may proceed exactly as in the proof of Proposition 3.3. \blacksquare

3.1.3 Multiple-Index Models and Variable Selection

A natural extension of single-index models, which covers a much wider class of output rules, are so-called multiple-index models (e.g., see [Hor09]). They are formally defined as follows:

Assumption 3.7 (Multiple-index models) Let Assumption 2.1 be satisfied with $A = I_d$ and let

$$y = G(\langle \mathbf{s}, \mathbf{z}_1 \rangle, \dots, \langle \mathbf{s}, \mathbf{z}_S \rangle) \quad (3.6)$$

where $\mathbf{z}_1, \dots, \mathbf{z}_S \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ are (unknown) *index vectors* and $G: \mathbb{R}^S \rightarrow \mathbb{R}$ is a measurable *output function* which can be random (independently of \mathbf{s}).

The case of $S = 1$ obviously coincides with the single-index model from Assumption 3.2. In general, the ultimate goal is to recover all index vectors $\mathbf{z}_1, \dots, \mathbf{z}_S$, or at least the spanned subspace

¹This is not exactly how generalized linear models are usually introduced in the literature. We adapt our definition from [PVY16, Subsec. 3.4], disregarding that y also belongs to an exponential family.

$\text{span}\{z_1, \dots, z_S\}$. The feasibility of this problem, however, strongly depends on the output function G . If G is linear, for example, (3.6) would simply degenerate to a linear single-index model. Apart from that, one may have to face certain ambiguities, such as permutation invariance of the indices.

Applying the estimator (P_K^{iso}) to multiple-index observations appears somewhat odd at first sight, as a linear hypothesis function $s \mapsto \langle s, z \rangle$ does certainly not lead to a small mean squared error. Indeed, we cannot expect recovery of all index vectors, but there is at least hope that (P_K^{iso}) approximates *any* vector in the index space $\text{span}\{z_1, \dots, z_S\}$. Regarding the mismatch principle, this suggests to apply Recipe 2.6(3) (and Remark 2.7) with $T = \text{span}\{z_1, \dots, z_S\}$ as target set:

Proposition 3.8 *Let Assumption 3.7 be satisfied. Moreover, assume that the index vectors $z_1, \dots, z_S \in \mathbb{R}^d$ form an orthonormal system and set $T := \text{span}\{z_1, \dots, z_S\}$. Then $z^{\natural} := \sum_{j=1}^S \mu_j z_j \in T$ with*

$$\mu_j := \mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) \langle s, z_j \rangle], \quad j = 1, \dots, S,$$

minimizes the mismatch covariance on T and we have

$$\rho(z^{\natural}) = \|\mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) P_T^{\perp} s]\|_2.$$

In particular, if $s \sim \mathcal{N}(\mathbf{0}, I_d)$, we have $\rho(z^{\natural}) = 0$.

Proof. We proceed analogously to the proof of Proposition 3.3: First, let us make the parametric ansatz $z^{\natural} = \sum_{j=1}^S \mu_j z_j \in T$ with unspecified scalars $\mu_1, \dots, \mu_S \in \mathbb{R}$, and consider the orthogonal decomposition

$$s = P_T s + P_T^{\perp} s = \sum_{j=1}^S \langle s, z_j \rangle z_j + P_T^{\perp} s,$$

where we have used that z_1, \dots, z_S form an orthonormal system. Using the isotropy of s now implies

$$\begin{aligned} \rho(z^{\natural})^2 &= \left\| \mathbb{E} \left[\left(G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) - \sum_{j'=1}^S \mu_{j'} \langle s, z_{j'} \rangle \right) \left(\sum_{j=1}^S \langle s, z_j \rangle z_j + P_T^{\perp} s \right) \right] \right\|_2^2 \\ &= \sum_{j=1}^S \left\| \mathbb{E} \left[\left(G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) - \sum_{j'=1}^S \mu_{j'} \langle s, z_{j'} \rangle \right) \langle s, z_j \rangle z_j \right] \right\|_2^2 \\ &\quad + \left\| \mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) P_T^{\perp} s] - \underbrace{\mathbb{E} \left[\left(\sum_{j'=1}^S \mu_{j'} \langle s, z_{j'} \rangle \right) P_T^{\perp} s \right]}_{=0} \right\|_2^2 \\ &= \sum_{j=1}^S \left(\mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) \langle s, z_j \rangle] - \underbrace{\mu_j \mathbb{E}[\langle s, z_j \rangle^2]}_{=\|z_j\|_2^2=1} \right)^2 \cdot \underbrace{\|z_j\|_2^2}_{=1} \\ &\quad + \left\| \mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) P_T^{\perp} s] \right\|_2^2 \\ &= \sum_{j=1}^S \left(\mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) \langle s, z_j \rangle] - \mu_j \right)^2 + \left\| \mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) P_T^{\perp} s] \right\|_2^2. \end{aligned}$$

Thus, the mismatch covariance is minimized on T if $\mu_j = \mathbb{E}[G(\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle) \langle s, z_j \rangle]$ for all $j = 1, \dots, S$. The claim in the Gaussian case again follows from the fact that $\langle s, z_1 \rangle, \dots, \langle s, z_S \rangle$, and $P_T^{\perp} s$ are independent if $s \sim \mathcal{N}(\mathbf{0}, I_d)$. \blacksquare

We note that the orthogonality assumption on the index vectors is not overly restrictive because one may incorporate an orthogonalizing transformation by redefining G . The actual outcome of Proposition 3.8 strongly resembles the situation of single-index models in Subsection 3.1.1: (P_K^{iso}) approximates a certain target vector z^{\natural} in the index space, supposed that the ‘‘compatibility parameter’’ $\rho(z^{\natural})$ is sufficiently small. In fact, we cannot expect a stronger statement at

this point, since the scalar factors strongly depend on the (still) unknown index vectors and the possibly unknown output function G . Consequently, a recovery strategy for the entire subspace $T = \text{span}\{z_1, \dots, z_S\}$ demands a more sophisticated method, e.g., a lifting approach as recently proposed by [YBL17b; YBWL17].

Remark 3.9 (Shallow neural networks) If $g: \mathbb{R} \rightarrow \mathbb{R}$ is a scalar function and $G = \sum_{j=1}^S \alpha_j g(\langle \cdot, z_j \rangle)$ for certain coefficients $\alpha_1, \dots, \alpha_S \in \mathbb{R}$, the output rule of (3.6) corresponds to a *shallow neural network*. While there is a tremendous interest in these types of models in the past few years, the learning setting of this work fits into the theory of neural networks only to a limited extent. Indeed, the key goal of this field of research is to accurately predict the output variable y , which is clearly not achieved by the least squares estimator ($\mathbf{P}_K^{\text{iso}}$). Instead, we are rather asking for an estimate of the underlying parameter vectors, which is of secondary relevance for prediction problems. An interesting theoretical study of this issue can be found in a recent work by Mondelli and Montanri [MM18]. \diamond

Despite the above mentioned limitations, Proposition 3.8 still allows us to treat an important special case of multiple-index models, namely *variable selection*: If the index vectors are unit vectors, say $z_1 = e_{k_1}, \dots, z_S = e_{k_S}$ for a certain set of *active variables* $\mathcal{S} := \{k_1, \dots, k_S\} \subset \{1, \dots, d\}$, then (3.6) simplifies as follows:

$$y = G(\langle s, e_{k_1} \rangle, \dots, \langle s, e_{k_S} \rangle) = G(s_{k_1}, \dots, s_{k_S}), \quad (3.7)$$

which exactly corresponds to the variable selection model introduced in (1.6). Adapting Proposition 3.8 to this specific situation reveals that the optimal target vector is indeed supported on \mathcal{S} :

Corollary 3.10 (Variable selection) *Let $\mathcal{S} := \{k_1, \dots, k_S\} \subset \{1, \dots, d\}$ be a set of active variables, let Assumption 3.7 be satisfied with $z_j = e_{k_j}$ for $j = 1, \dots, S$ (see (3.7)), and set*

$$T := \text{span}\{e_{k_1}, \dots, e_{k_S}\} = \{z \in \mathbb{R}^d \mid \text{supp}(z) \subset \mathcal{S}\}.$$

Then $z^\natural := \sum_{j=1}^S \mu_j e_{k_j} \in T$ with

$$\mu_j := \mathbb{E}[G(s_{k_1}, \dots, s_{k_S}) s_{k_j}], \quad j = 1, \dots, S, \quad (3.8)$$

minimizes the mismatch covariance on T and we have

$$\rho(z^\natural) = \|\mathbb{E}[G(s_{k_1}, \dots, s_{k_S}) \mathbf{P}_T^\perp s]\|_2.$$

If the feature variables of $s = (s_1, \dots, s_d)$ are independent, we particularly have $\rho(z^\natural) = 0$.

Proof. This is an immediate consequence of Proposition 3.8 and (3.7). The ‘in particular part’ is due to the fact that $\mathbf{P}_T^\perp s$ only depends on the non-active feature variables in $\{1, \dots, d\} \setminus \mathcal{S}$, which are independent from s_{k_1}, \dots, s_{k_S} . \blacksquare

This results shows that variable selection via ($\mathbf{P}_K^{\text{iso}}$) becomes feasible, as long as the scalars μ_1, \dots, μ_S in (3.8) are not too close to 0 and the sample size n is sufficiently large. More precisely, if \hat{z} is a minimizer of ($\mathbf{P}_K^{\text{iso}}$), the set of active variables can be detected by setting all small entries of \hat{z} to zero, so that its support coincides with $\text{supp}(z^\natural) = \mathcal{S}$. A remarkable conclusion of Corollary 3.10 is that — in contrast to single-index models — we even obtain a consistent estimator in the sub-Gaussian case, supposed that the feature variables are independent.

Remark 3.11 (Sparse recovery) The complexity of the variable selection problem is essentially determined by the number of active variables S . In practice, one typically has to face high-dimensional situations where S is significantly smaller than d , implying that the target vector z^\natural in Corollary 3.10 is sparse. As sketched in Subsection 3.1.5 below, such a sparsity assumption can be easily exploited by means of an ℓ^1 -hypothesis set, enabling accurate estimation with $n \ll d$.

Moreover, it is worth emphasizing that a single-index model (3.3) with $\|z_0\|_0 \leq S$ is a special case of (3.7). A comparison of Proposition 3.3 and Corollary 3.10 shows that we can at least expect

a consistent estimate of $\text{supp}(z_0)$, even though recovery of z_0 may fail. Hence, after projecting the data onto $\text{supp}(z_0)$, one could eventually perform a more sophisticated method (operating in a much lower dimensional space) to estimate the weights of z_0 . Such a two-step procedure is in fact often successfully applied in real-world learning tasks. In that spirit, Corollary 3.10 provides more theoretical evidence of why Lasso-type estimators can serve as a powerful feature selector. \diamond

3.1.4 Superimposed Observations

A different way to generalize Assumption 3.2 are superpositions of single-index observations in the following sense:

Assumption 3.12 (Superimposed single-index models) Let s^1, \dots, s^M be independent, isotropic, mean-zero random vectors in \mathbb{R}^d with $\|s^j\|_{\psi_2} \leq \kappa$ for some $\kappa > 0$. The output variable is given by

$$y := \frac{1}{\sqrt{M}} \sum_{j=1}^M g_j(\langle s^j, z_0 \rangle) \quad (3.9)$$

where $z_0 \in \mathbb{R}^d \setminus \{0\}$ is an unknown *index vector* and $g_j: \mathbb{R} \rightarrow \mathbb{R}$ are a measurable scalar *output functions*. Moreover, we define the input vector as $s := \frac{1}{\sqrt{M}} \sum_{j=1}^M s^j$.

Using the isotropy of s^1, \dots, s^M and Hoeffding's inequality [Ver12, Lem. 5.9], it is not hard to see that the random pair (s, y) indeed fulfills Assumption 2.1 with $A = I_d$ (cf. [GJ17b, Prop. 6.5(1)]). The observation scheme of Assumption 3.12 was extensively studied by the first author in [GJ17a; GJ17b], motivated by an application in distributed wireless sensing. The basic idea behind this model is that M autonomous *nodes* transmit their linear measurements $\langle s^j, z_0 \rangle$ simultaneously to a central receiver. Due to non-linear distortions during this transmission procedure, which are modeled by the output functions g_j , the receiver eventually measures a *superposition* of all distorted signals, as described in (3.9).

Since [GJ17b] actually relies on similar proof techniques as this work and considers a more general setup than Assumption 3.12, we omit a detailed discussion at this point. Nevertheless, for the sake of completeness, let us provide an adaption of Proposition 3.3 that verifies the feasibility of the associated recovery task in the Gaussian case:

Proposition 3.13 *Let Assumption 3.12 be satisfied with $s^j \sim \mathcal{N}(0, I_d)$ for $j = 1, \dots, M$ and $\|z_0\|_2 = 1$. Set $T := \text{span}\{z_0\}$ as target set. Then $z^\natural := \bar{\mu}z_0 \in T$ with*

$$\bar{\mu} := \frac{1}{M} \sum_{j=1}^M \mathbb{E}[g_j(\langle s^j, z_0 \rangle) \langle s^j, z_0 \rangle]$$

minimizes the mismatch covariance on T and we have $\rho(z^\natural) = 0$.

Proof. This works analogously to the proof of Proposition 3.3. See [GJ17b, Prop. 6.5(1)] for more details. \blacksquare

3.1.5 High-Dimensional Problems and ℓ^1 -Constraints

So far we did only marginally discuss the impact of the hypothesis set K in the above examples, in particular, the issues raised by (Q3). In fact, the presence of a constraint set leads to fundamentally different estimation results than one would obtain from ordinary (unconstrained) least squares minimization. This concern becomes clearer in the context of Theorem 2.8, where one just ends up with $w_\lambda^2(A^\top K - z^\natural) \asymp d$ in (2.7) if $K = \mathbb{R}^p$ and A is injective.

In the following, we sketch how an ℓ^1 -constraint may significantly improve the sampling rate when the target vector z^\natural is known to be sparse. Regarding the above examples of single-index models (cf. Subsection 3.1.1) and variable selection (cf. Subsection 3.1.3), this assumption would be simply fulfilled if $\|z_0\|_0 \ll d$ in (3.3) or $S \ll d$ in (3.7), respectively. More generally, let us

assume that z^\natural is S -sparse for a certain $S > 0$, i.e., $\|z^\natural\|_0 \leq S$. The Cauchy-Schwarz inequality then implies an upper bound for the ℓ^1 -norm:

$$\|z^\natural\|_1 \leq \sqrt{\|z^\natural\|_0} \cdot \|z^\natural\|_2 \leq \sqrt{S} \cdot \|z^\natural\|_2,$$

which in turn suggests to select $K = \sqrt{S} \cdot \|z^\natural\|_2 \cdot B_1^d$ as hypothesis set. In order to apply Corollary 3.1, we may use that $w(B_1^d) \lesssim \sqrt{\log(2d)}$ (cf. [Ver15, Ex. 3.8]), so that (3.1) yields the following condition:

$$n \gtrsim \kappa^4 \cdot \delta^{-4} \cdot \|z^\natural\|_2^2 \cdot S \cdot \log(2d). \quad (3.10)$$

A similar argument also works for the conic mean width in Theorem 2.8 (with $A = I_d$): Choosing $K = \|z^\natural\|_1 \cdot B_1^d$ as hypothesis set, it holds that $w_\wedge(K - z^\natural) \lesssim \sqrt{S \cdot \log(2d/S)}$ (cf. [CRPW12, Prop. 3.10]). Consequently, the condition of (2.7) is already satisfied if

$$n \gtrsim \kappa^4 \cdot \delta^{-2} \cdot S \cdot \log\left(\frac{2d}{S}\right). \quad (3.11)$$

The above bounds are consistent with the traditional theory of compressed sensing and sparse signal estimation, according to which the number of needed observations essentially scales linearly with the degree of sparsity S , and the ambient dimension d only appears in log-terms (cf. [CRT05; CT05; CT06; Don06]). On the other hand, (3.10) and (3.11) are both only of limited practical relevance, since the involved hypothesis sets require prior knowledge of the sparsity or ℓ^1 -norm of z^\natural , respectively. Fortunately, using the concept of local mean width (see Definition 6.1 and Theorem 6.3), it is possible to derive stable recovery results that allow for different types of model uncertainties. A more detailed discussion of this advanced approach is contained in [GKM17, Sec. 2.4] and [GS18, Sec. 3.1]. For comprehensive overviews of compressed sensing, high-dimensional estimation problems, and the benefit of sparsity, we refer to [BV11; DDEK12; FR13; Ver15].

Finally, let us point that there are many extensions of sparsity that can be incorporated by an appropriate choice of hypothesis set, e.g., *group sparsity* [HZ10], *sparsifying transformations* [EMR07], or *weighted sparsity* [KXAH09]. Analyzing the mean width and the resulting error bounds in each of these situations could then provide a means to adapt certain model parameters; see [OKH12] for an example in weighted ℓ^1 -minimization.

3.2 Correlated Input Variables

In this subsection, we investigate several instances of Assumption 2.1 where the mixing matrix $A \in \mathbb{R}^{p \times d}$ is not the identity. The first part (Subsection 3.2.1) deals with the case of $p \geq d$, meaning that there are more observed than latent variables. In this context, it will turn out that the statement of Theorem 2.5 can be significantly strengthened if A is approximately known and the hypothesis set is carefully adapted. Subsection 3.2.2 is then devoted to the case of $p < d$, which is especially useful to model noisy data. Interestingly, the mismatch principle reveals in this situation that the mismatch covariance basically corresponds to the signal-to-noise ratio of the underlying observation process. Finally, let us emphasize that, compared to the previous subsection, we now take a rather abstract viewpoint and do not further specify to the output variable y .

3.2.1 The Underdetermined Case ($p \geq d$)

We first note that the term ‘underdetermined’ refers to the condition $z^\natural \in A^\top K$ in Theorem 2.5. Indeed, if $p \geq d$, the linear system $z^\natural = A^\top \beta^\natural$ typically has many solutions in K , implying that the associated parameter vector $\beta^\natural \in K$ is not uniquely defined. This ambiguity is in fact a particular reason why all theoretical guarantees from Section 2 do only concern the estimation of a target vector $z^\natural = A^\top \beta^\natural$ but not β^\natural . The purpose of this subsection is therefore to show that, based on approximate knowledge of A , one may adapt the hypothesis set K in a such way that β^\natural becomes well-defined and the mean width in (2.1) does not depend on A anymore.

To simplify the following exposition, let us assume that the mixing matrix $A \in \mathbb{R}^{p \times d}$ has full rank. Hence, A is injective and its pseudo-inverse $A^\dagger \in \mathbb{R}^{d \times p}$ satisfies $A^\dagger A = I_d$. In principle, as long as A is known, the issue of correlated feature variables can be resolved by first computing

$$A^\dagger x_i = A^\dagger A s_i = s_i, \quad i = 1, \dots, n, \quad (3.12)$$

and then proceeding with isotropic data as in Subsection 3.1. Unfortunately, such a pre-processing step is often unstable in practice and A is usually not exactly known. The following corollary of Theorem 2.5 takes a different approach that involves a transformation of the hypothesis set in (P_K) but not of the actual input data. Moreover, it just assumes that an estimated mixing matrix $\tilde{A} \in \mathbb{R}^{p \times d}$ is available.

Corollary 3.14 (Adaptive estimation via (P_K)) *Let Assumption 2.1 be satisfied with $p \geq d$ and assume that A has full rank. Moreover, assume that $\tilde{A} \in \mathbb{R}^{p \times d}$ is a full rank matrix such that $M := \tilde{A}^\dagger A \in \mathbb{R}^{d \times d}$ is invertible. Let $\tilde{K} \subset \mathbb{R}^d$ be a bounded, convex subset and fix a vector $z^\natural \in M^\top \tilde{K}$. Then there exists a numerical constant $C > 0$ such that for every $\delta \in (0, 1]$, the following holds true with probability at least $1 - 5 \exp(-C \cdot \kappa^{-4} \cdot \delta^2 \cdot n)$: If the number of observed samples obeys*

$$n \gtrsim \kappa^4 \cdot \delta^{-4} \cdot \|M\|_{\text{op}}^2 \cdot w^2(\tilde{K}), \quad (3.13)$$

then every minimizer $\hat{\beta}$ of (P_K) with $K := (\tilde{A}^\dagger)^\top \tilde{K}$ satisfies

$$\|\tilde{A}^\top \hat{\beta} - z^\natural\|_2 \lesssim \|M^{-1}\|_{\text{op}} \cdot (\max\{1, \kappa \cdot \sigma(z^\natural)\} \cdot \delta + \rho(z^\natural)) + \|(I_d - M^{-\top})z^\natural\|_2 \quad (3.14)$$

and

$$\|\hat{\beta} - \beta^\natural\|_2 \lesssim \|M^{-1} \tilde{A}^\dagger\|_{\text{op}} \cdot (\max\{1, \kappa \cdot \sigma(z^\natural)\} \cdot \delta + \rho(z^\natural)) \quad (3.15)$$

where $\beta^\natural := (\tilde{A}^\dagger)^\top M^{-\top} z^\natural \in K$.

Proof. We simply apply Theorem 2.5 with $K = (\tilde{A}^\dagger)^\top \tilde{K}$. Let us first observe that

$$z^\natural \in M^\top \tilde{K} = A^\top (\tilde{A}^\dagger)^\top \tilde{K} = A^\top K.$$

Furthermore, Slepian's inequality [FR13, Lem. 8.25] yields

$$w(A^\top K) = w(M^\top \tilde{K}) \leq \|M\|_{\text{op}} \cdot w(\tilde{K}),$$

so that the assumption (3.13) implies (2.1). On the event of Theorem 2.5, we therefore obtain

$$\begin{aligned} \|\tilde{A}^\top \hat{\beta} - z^\natural\|_2 &\leq \|\tilde{A}^\top \hat{\beta} - M^{-\top} z^\natural\|_2 + \|(I_d - M^{-\top})z^\natural\|_2 \\ &= \|M^{-\top}\|_{\text{op}} \cdot \|M^\top \tilde{A}^\top \hat{\beta} - z^\natural\|_2 + \|(I_d - M^{-\top})z^\natural\|_2 \\ &= \|M^{-1}\|_{\text{op}} \cdot \|A^\top \underbrace{(\tilde{A}^\dagger)^\top \tilde{A}^\top \hat{\beta}}_{\stackrel{(*)}{\hat{\beta}}} - z^\natural\|_2 + \|(I_d - M^{-\top})z^\natural\|_2 \\ &\stackrel{(2.2)}{\lesssim} \|M^{-1}\|_{\text{op}} \cdot (\max\{1, \kappa \cdot \sigma(z^\natural)\} \cdot \delta + \rho(z^\natural)) + \|(I_d - M^{-\top})z^\natural\|_2, \end{aligned}$$

where $(*)$ follows from $\hat{\beta} \in \text{ran}((\tilde{A}^\dagger)^\top)$ and the identity $(\tilde{A}^\dagger)^\top \tilde{A}^\top (\tilde{A}^\dagger)^\top = (\tilde{A}^\dagger)^\top$.

In order to establish (3.15), we observe that

$$(\tilde{A}^\dagger)^\top = (\tilde{A}^\dagger)^\top M^{-\top} M^\top = (\tilde{A}^\dagger)^\top M^{-\top} A^\top (\tilde{A}^\dagger)^\top.$$

Since $\hat{\beta} \in \text{ran}((\tilde{A}^\dagger)^\top)$, this implies $\hat{\beta} = (\tilde{A}^\dagger)^\top M^{-\top} A^\top \hat{\beta}$. Using the definition of β^\natural , we therefore end up with

$$\|\hat{\beta} - \beta^\natural\|_2 = \|(\tilde{A}^\dagger)^\top M^{-\top} (A^\top \hat{\beta} - z^\natural)\|_2 \leq \|M^{-1} \tilde{A}^\dagger\|_{\text{op}} \cdot \|A^\top \hat{\beta} - z^\natural\|_2,$$

and (3.15) now follows from (2.2) again. \blacksquare

If $\tilde{A} = A$ (and therefore $M = I_d$), the error bound of (3.14) still coincides with (2.2), whereas (3.15) provides an approximation guarantee for the actual solution vector $\hat{\beta}$ of (P_K) . This achievement of Corollary 3.14 is due to the specific choice of K , which is the image of a lower dimensional subset \tilde{K} that accounts for the effect of the mixing matrix A and thereby resolves the ambiguity pointed out above. The only price to pay is the presence of an additional factor $\|A^\dagger\|_{\text{op}}$ in (3.15). It is also noteworthy that the complexity in (3.13) is now measured in terms of the hypothesis set $\tilde{K} \subset \mathbb{R}^d$, allowing us to exploit structural assumptions on z^\dagger in a direct manner.

A practical feature of Corollary 3.14 is that it even remains valid when the estimators $\hat{\beta}$ and $\hat{z} := \tilde{A}^\top \hat{\beta}$ are based on an approximate mixing matrix \tilde{A} . The quality of this approximation is essentially captured by the transformation matrix $M = \tilde{A}^\top A$, in the sense that $M \approx I_d$ if $\tilde{A} \approx A$. This particularly affects the sample size in (3.13) as well as the error bound in (3.14), which is in fact only significant if \tilde{A} is sufficiently close to A . Remarkably, the output vector $\hat{\beta}$ may still constitute a consistent estimator of β^\dagger , since M just appears in terms of a multiplicative constant in (3.15). However, we point out that (3.15) is actually a weaker statement than (3.14), as the former does not necessarily address our initial concern of learning semi-parametric output rules.

Let us conclude with some possible applications of Corollary 3.14, and in particular (3.15):

Remark 3.15 (1) *Clustered variables.* A situation of practical interest are *clustered features*, where the components of x are divided into (disjoint) groups of strongly correlated variables. In this case, the associated hypothesis set $K := (\tilde{A}^\dagger)^\top \tilde{K}$ would basically turn into a group constraint, which leads to a block-like support of the parameter vector $\beta^\dagger \in K$. A formal study of such a problem scenario is in principle straightforward, but however requires a certain technical effort and therefore goes beyond the scope of this paper.

(2) *Unknown mixing matrix.* The above strategy clearly relies on (partial) knowledge of A . The same problem persists in the general setup of Theorem 2.5, as the mixing matrix is needed to construct the estimator $\hat{z} := A^\top \hat{\beta}$. Thus, if A is unknown, the error bound (2.2) is rather of theoretical interest, merely indicating that — in principle — estimation via (P_K) is feasible. Nevertheless, one may still draw heuristic conclusions from Theorem 2.5. For instance, if A generates clustered features as in (1), the support pattern of the (possibly non-unique) parameter vector β^\dagger would at least indicate what clusters are active. This approach was already discussed in a previous work by the authors [GK16], which is inspired by the example of *mass spectrometry data* (see also Remark 3.18). \diamond

3.2.2 The Overdetermined Case ($p < d$) and Noisy Data

In most real-world applications, the input data is corrupted by noise. With regard to our statistical model setup, this basically means that the input vector $x = As$ is generated from two types of latent factors, namely those affecting the output variable y and those that do not. Let us make this idea more precise:

Assumption 3.16 (Noisy data) Let Assumption 2.1 be satisfied and assume that the latent factors split into two parts, $s = (v, n)$, where $v = (v_1, \dots, v_{d_1}) \in \mathbb{R}^{d_1}$ are referred to as *signal variables* and $n = (n_1, \dots, n_{d_2}) \in \mathbb{R}^{d_2}$ as *noise variables* ($\Rightarrow d = d_1 + d_2$). The output variable does only depend on the signal variables, i.e.,

$$y = G(v) = G(v_1, \dots, v_{d_1}) \quad (3.16)$$

for a measurable *output function* $G: \mathbb{R}^{d_1} \rightarrow \mathbb{R}$ which can be random (independently of v), and we also assume that y and n are independent. According to the above partition, the mixing matrix takes the form $A = [A_v, A_n] \in \mathbb{R}^{p \times d}$ where $A_v \in \mathbb{R}^{p \times d_1}$ is associated with v and $A_n \in \mathbb{R}^{p \times d_2}$ with n . Thus, the input vector can be decomposed as $x = As = A_v v + A_n n$.

The key concern of Assumption 3.16 is that the factors of n do not contribute to the output variable y and are therefore regarded as *noise*. In this context, it could easily happen that $d > p$, so that A is not injective anymore. Retrieving the latent factors s from x then becomes impossible, even if A is exactly known (cf. (3.12)). A typical example case would be component-wise “background” noise within the input vector x , i.e., $A_n \in \mathbb{R}^{p \times p}$ is a diagonal matrix and we have $d = d_1 + d_2 = d_1 + p > p$. Moreover, we emphasize that Assumption 3.16 should not be confused

with the variable selection problem from (3.7). The latter comes along with the task to detect the *unknown* subset of active variables, whereas the above partition into signal and noise variables is known to a certain extent.¹ For that reason, the order of the variables in $\mathbf{s} = (\mathbf{v}, \mathbf{n})$ is rather a matter of convenience and does not restrict the generality.

Let us now apply the mismatch principle (Recipe 2.6) to investigate the learning capacity of (\mathbf{P}_K) under the noise model of Assumption 3.16. According to the output rule of (3.16), it is natural to select a target set of the form $T = T_v \times \mathbb{R}^{d_2} \subset \mathbb{R}^{d_1+d_2}$. Indeed, given any $\mathbf{z}^{\natural} = (\mathbf{z}_v^{\natural}, \mathbf{z}_n^{\natural}) \in T$, we are actually only interested in the signal component $\mathbf{z}_v^{\natural} \in T_v$ — encoding the desired (parametric) information about (3.16) in terms of $T_v \subset \mathbb{R}^{d_1}$ — while the noise component $\mathbf{z}_n^{\natural} \in \mathbb{R}^{d_2}$ is irrelevant for our purposes. This important observation is particularly consistent with the following decomposition of the mismatch covariance:

Proposition 3.17 *Let Assumption 3.16 be satisfied. For every $\mathbf{z}^{\natural} = (\mathbf{z}_v^{\natural}, \mathbf{z}_n^{\natural}) \in \mathbb{R}^{d_1+d_2}$, we have that*

$$\rho_{s,y}(\mathbf{z}^{\natural})^2 = \|\mathbb{E}[(y - \langle \mathbf{v}, \mathbf{z}_v^{\natural} \rangle) \mathbf{v}]\|_2^2 + \|\mathbf{z}_n^{\natural}\|_2^2 = \rho_{v,y}(\mathbf{z}_v^{\natural})^2 + \|\mathbf{z}_n^{\natural}\|_2^2. \quad (3.17)$$

Proof. Using the isotropy of $\mathbf{s} = (\mathbf{v}, \mathbf{n})$ and the independence of y and \mathbf{n} , we obtain

$$\begin{aligned} \rho_{s,y}(\mathbf{z}^{\natural})^2 &= \|\mathbb{E}[(y - \langle \mathbf{s}, \mathbf{z}^{\natural} \rangle) \mathbf{s}]\|_2^2 \\ &= \|\mathbb{E}[(y - \langle \mathbf{v}, \mathbf{z}_v^{\natural} \rangle - \langle \mathbf{n}, \mathbf{z}_n^{\natural} \rangle) \mathbf{v}]\|_2^2 + \|\mathbb{E}[(y - \langle \mathbf{v}, \mathbf{z}_v^{\natural} \rangle - \langle \mathbf{n}, \mathbf{z}_n^{\natural} \rangle) \mathbf{n}]\|_2^2 \\ &= \|\mathbb{E}[(y - \langle \mathbf{v}, \mathbf{z}_v^{\natural} \rangle) \mathbf{v}]\|_2^2 + \|\mathbb{E}[\langle \mathbf{n}, \mathbf{z}_n^{\natural} \rangle \mathbf{n}]\|_2^2 = \|\mathbb{E}[(y - \langle \mathbf{v}, \mathbf{z}_v^{\natural} \rangle) \mathbf{v}]\|_2^2 + \|\mathbf{z}_n^{\natural}\|_2^2. \quad \blacksquare \end{aligned}$$

The identity (3.17) asks us to select a target vector $\mathbf{z}^{\natural} = (\mathbf{z}_v^{\natural}, \mathbf{z}_n^{\natural}) \in T$ such that the “signal-related” mismatch covariance $\rho_{v,y}(\mathbf{z}_v^{\natural})$ is sufficiently small, and at the same time, the magnitude of noise component \mathbf{z}_n^{\natural} must not be too large. This task is unfortunately not straightforward because the transformed hypothesis set $A^T K \subset \mathbb{R}^d$ is not full-dimensional if $d > p$, so that the simplification of Remark 2.7 is not applicable anymore. For example, solely enforcing $\mathbf{z}_n^{\natural} = \mathbf{0}$ could eventually lead to a large value of $\rho_{v,y}(\mathbf{z}_v^{\natural})$, as the condition $\mathbf{z}^{\natural} \in A^T K$ implies an ‘overdetermined’ equation system with $\mathbf{z}_v^{\natural} = A_v^T \boldsymbol{\beta}^{\natural}$ and $\mathbf{z}_n^{\natural} = A_n^T \boldsymbol{\beta}^{\natural}$. More generally speaking, the minimization problem (2.6) in Recipe 2.6(3) is not separable in the signal and noise component, which can cause difficulties.

For that reason, we do not exactly apply Recipe 2.6(3) in the following, but rather construct an admissible target vector $\mathbf{z}^{\natural} = (\mathbf{z}_v^{\natural}, \mathbf{z}_n^{\natural}) \in T \cap A^T K$ that approximately solves (2.6). Motivated by the decomposition of (3.17), our basic idea is to first minimize $\mathbf{z}_v \mapsto \rho_{v,y}(\mathbf{z}_v)$ and then to choose among all *admissible* vectors in $A_n^T K$ the one of minimal magnitude. Such an approach appears in fact quite natural, since our primary goal is to accurately estimate the signal component \mathbf{z}_v^{\natural} by $A_v^T \hat{\boldsymbol{\beta}}$, with $\hat{\boldsymbol{\beta}}$ being a minimizer of (\mathbf{P}_K) . Thus, let us simply define

$$\mathbf{z}_v^{\natural} := \operatorname{argmin}_{\mathbf{z}_v \in T_v \cap A_v^T K} \rho_{v,y}(\mathbf{z}_v).$$

Denoting the preimage of \mathbf{z}_v^{\natural} by $K_v := \{\boldsymbol{\beta} \in K \mid A_v^T \boldsymbol{\beta} = \mathbf{z}_v^{\natural}\} \subset \mathbb{R}^p$, we are now allowed to select any vector $\mathbf{z}_n^{\natural} \in A_n^T K_v \subset \mathbb{R}^{d_2}$. The one of minimal ℓ^2 -norm is therefore

$$\mathbf{z}_n^{\natural} := \operatorname{argmin}_{\mathbf{z}_n \in A_n^T K_v} \|\mathbf{z}_n\|_2 = \operatorname{argmin}_{\boldsymbol{\beta} \in K_v} \|A_n^T \boldsymbol{\beta}\|_2,$$

and we also set $\rho_n(\mathbf{z}_v^{\natural}) := \|\mathbf{z}_n^{\natural}\|_2$, in order to indicate the dependence \mathbf{z}_v^{\natural} .

According to Recipe 2.6(4), this choice of \mathbf{z}^{\natural} yields the following bound:

$$\|A_v^T \hat{\boldsymbol{\beta}} - \mathbf{z}_v^{\natural}\|_2 \leq \|A^T \hat{\boldsymbol{\beta}} - \mathbf{z}^{\natural}\|_2$$

¹This does not necessarily mean that \mathbf{v} and \mathbf{n} are directly accessible, but rather that there exists (partial) knowledge of the mixing matrix $A = [A_v, A_n]$ and the associated noise pattern.

$$\stackrel{(2.3)}{\lesssim} \max\{\kappa, \kappa^2 \cdot \sigma(\mathbf{z}^\natural)\} \cdot \left(\frac{w^2(\mathbf{A}^\top \mathbf{K})}{n}\right)^{1/4} + \left(\rho_{v,y}(\mathbf{z}_v^\natural)^2 + \rho_n(\mathbf{z}_v^\natural)^2\right)^{1/2}. \quad (3.18)$$

The significance of this error estimate clearly depends on the size of the bias terms $\rho_{v,y}(\mathbf{z}_v^\natural)$ and $\rho_n(\mathbf{z}_v^\natural)$. The former parameter corresponds to the mismatch covariance of the noiseless data model (v, y) and can be treated as in the isotropic case of Subsection 3.1. On the other hand, $\rho_n(\mathbf{z}_v^\natural)$ captures the impact of the noise variables \mathbf{n} on the input data \mathbf{x} . As already pointed out above, the presence of such a noise term is inevitable to some extent, since the “ideal” target vector $\mathbf{z}^\natural := (\mathbf{z}_v^\natural, \mathbf{0}) \in \mathbb{R}^{d_1+d_2}$ is usually not contained in $\mathbf{A}^\top \mathbf{K}$.¹ From a more practical perspective, $\rho_n(\mathbf{z}_v^\natural)$ serves as a measure of the *noise power* of the underlying observation model (x, y) and its reciprocal, $1/\rho_n(\mathbf{z}_v^\natural)$, can be interpreted as the signal-to-noise ratio.

Remark 3.18 (1) *Previous approaches.* The above methodology is inspired by [GK16], where the authors investigated noisy data in conjunction with Gaussian single-index models. The key idea of that work is to specify a parameter vector $\boldsymbol{\beta}^\natural \in \mathbb{R}^p$ such that the linear mapping $\mathbf{x} \mapsto \langle \mathbf{x}, \boldsymbol{\beta}^\natural \rangle$ imitates the output model for y as well as possible, coining the notion of *optimal representations*. Interestingly, the mismatch principle essentially leads us to the same results as in [GK16], but in a much more systematic and less technical way. In particular, the error bound of (3.18) is not exclusively restricted to single-index models.

(2) *Real-world data.* A natural way to think of Assumption 3.16 is that the columns of \mathbf{A}_v form building blocks (atoms) of the input vector and the signal variables v_i determine their individual contributions to each sample $x_i, i = 1, \dots, n$. The same interpretation applies to the noise variables n_i , but they do not affect the actual output y_i .

A prototypical example of real-world data that was extensively studied in [GK16] is so-called *mass spectrometry data*. In this case, the columns of \mathbf{A}_v do simply form (discretized) Gaussian-shaped peaks, each one representing a certain type of *molecule* (or a compound). The signal variables are in turn proportional to the molecular concentration of these molecules. On the other hand, the output variable y_i only takes values in $\{-1, +1\}$, indicating the *health status* of a human individual (e.g., healthy or suffering from a specific disease). The key task is now to use empirical risk minimization to identify those molecules (features) that are relevant to the health status, eventually enabling for early diagnostics and a deeper understanding of pathological mechanisms. In the context of statistical learning, this challenge is closely related to the variable selection model considered in (3.7). For more details about this specific problem setup, we refer to [Gen15; GK16], focusing on a theoretical analysis, as well as to [Con+17] for practical aspects and numerical experiments. \diamond

4 Related Literature and Approaches

This part gives a brief overview of recent approaches from the literature that are related to the main ideas and conclusions of this work. The discussion below is more of conceptual nature and intends to point out similarities as well as important differences of our approach to others. In particular, we do only marginally address the specific examples of output models that were already investigated in Subsection 3.1.

4.1 Statistical Learning Theory

The setting and terminology of this work is clearly based on statistical learning theory; see [CZ07; HTF09; SB14; Vap98] for comprehensive overviews. As already pointed out in the introduction, one of the key objectives in this research area is to study empirical risk minimization (1.2) as a means to approximate the inaccessible problem of expected risk minimization (1.1). Despite many different variants, there are basically two common ways to assess the quality of a solution

¹Or equivalently, there does not exist a linear hypothesis function $f = \langle \cdot, \boldsymbol{\beta}^\natural \rangle$ with $\boldsymbol{\beta}^\natural \in K$ and $\mathbf{z}^\natural = \mathbf{A}^\top \boldsymbol{\beta}^\natural$.

$\hat{f} \in \mathcal{F}$ to (1.2):

$$\text{Prediction:} \quad \mathbb{E}_{(x,y)}[(y - \hat{f}(x))^2] - \mathbb{E}_{(x,y)}[(y - f^*(x))^2], \quad (4.1)$$

$$\text{Estimation:} \quad \mathbb{E}_x[(\hat{f}(x) - f^*(x))^2], \quad (4.2)$$

where $f^* \in \mathcal{F}$ is a minimizer of (1.1). The main purpose of (4.1) is to compare the predictive power of the empirical risk minimizer \hat{f} to the best possible in \mathcal{F} , which is f^* by definition. In contrast, the estimation error (4.2) measures how well \hat{f} approximates f^* , while not depending on the true output variable y .

Among a large amount of works on estimation and prediction problems, we think that the most related ones are those by Mendelson [Men15; Men18], which establish a very general learning framework for empirical risk minimization. His results are based on a few complexity parameters that relate the noise level and geometric properties of the hypothesis set to the estimation error and sample size. A crucial feature of Mendelson's approach is that it does not rely on concentration inequalities but rather on a mild *small ball condition*. In that way, the theoretical bounds of [Men15; Men18] remain even valid when the input and output variables are heavy-tailed. This method turned out to be very useful for controlling non-negative empirical processes in general, and therefore found many applications beyond statistical learning.

While the problem setting of Mendelson certainly resembles ours, there are several crucial differences. Firstly, we focus on much more restrictive model assumptions, in particular, sub-Gaussian data and *linear* hypothesis functions. This makes our results less general, but at the same time, more explicit and accessible for concrete observation models. The second important difference comes along with the first one: Instead of simply approximating the expected risk minimizer as in (4.2), we refine the classical estimation problem by the notion of target sets. This allows us to encode the desired (parametric) information about the underlying output rule, which in turn forms the basis of the mismatch principle (Recipe 2.6). For a more technical comparison of our approach to [Men15; Men18] see Remark 6.6.

Finally, it is worth noting that the prediction error (4.1) is only of minor interest in this work. Indeed, since we restrict ourselves to linear hypothesis functions in \mathcal{F} , one cannot expect that $f(x) = \langle x, \beta \rangle$ yields a reliable predictor of y , unless it linearly depends on x .

4.2 Signal Processing and Compressed Sensing

A considerable amount of the recent literature on (high-dimensional) signal processing and compressed sensing deals with *non-linear* distortions of linear sampling schemes, such as quantized, especially 1-bit compressed sensing [BB08; ZBC10]. In this context, single-index models proved very useful (cf. Assumption 3.2), as they permit many different types of perturbations by means of the output function. A remarkable line of research by Plan, Vershynin, and collaborators [ALPV14; PV13a; PV13b; PV14; PV16; PVY16] as well as related works [Gen17; GMW16; TAH15] has made significant progress in the statistical analysis of these model situations. Somewhat surprisingly, it turned out that recovery of the index vector is already achievable by convex programming, even when the output function is unknown. Of particular relevance to our approach is [PV16], where the estimator (P_K) was studied for the first time with Gaussian single-index observations. Although not mentioned explicitly, the key argument of [PV16] is based on an application of the orthogonality principle (cf. (1.7)), eventually leading to the same conclusions as in Subsection 3.1.1. In that light, our results show that the original ideas of [PV16] apply to a much wider class of semi-parametric models.

On the other hand, the discussion of Proposition 3.3 also reveals several shortcomings of learning single-index models via empirical risk minimization, most importantly, the inconsistency of (P_K) in the non-Gaussian case. A recent series of papers by Yang et al. [YBL17a; YBL17b; YBWL17; YWLEZ15; Yan+17] tackles this issue by investigating adaptive, possibly non-convex estimators. While these findings are very interesting and actually include more complicated estimation tasks, such as phase retrieval (cf. Example 3.4(4)) and multiple-index models (cf. Subsection 3.1.3), the proposed algorithms either assume knowledge of the output function or of the density function of the input vector. Such prerequisites are in fact somewhat different from the conception of

this work, but nevertheless, we think that these approaches are a promising direction of future research (see also Section 5).

In the situation of sparse index vectors and ℓ^1 -constraints, our results allow us to reproduce classical recovery guarantees from compressed sensing (cf. Subsection 3.1.5). But despite obvious similarities, the theoretical foundations of compressed sensing rely on quite different concepts, most prominently the *null space property* and *restricted isometry property* [FR13]. This methodology indeed enables the treatment of many practically relevant sensing schemes that are by far not included in Assumption 2.1, e.g., Fourier subsampling or circular matrices.

4.3 Correlated Features and Factor Analysis

An important practical concern of this work is to allow for correlated variables in the input vector x . However, most traditional approaches rely on assumptions that preclude stronger correlations between features, such as restricted isometry, restricted eigenvalue condition, or irrepresentability (cf. [BV11; FR13]). For that reason, various strategies have been suggested in the literature to deal with these delicate situations, for instance, *hierarchical clustering* [BRGZ13] or *OSCAR/OWL* [BR08; FN14; FN16]. The very recent work [LRW18] provides a general methodology to incorporate correlations between covariates by *graph-based regularization*. We do also recommend [LRW18, Sec. 2] for an extensive discussion of related literature in that direction. Although these results are limited to more specific model settings (e.g., noisy linear regression), they bear a resemblance to Corollary 3.14, where the (estimated) mixing matrix \hat{A} is used to adapt the constraint set of (P_K) .

If the correlation structure of $x = As$ is unknown, one rather aims at recovering the mixing matrix A based on the available sample set $\{x_i\}_{i=1}^n$. This task is usually referred to as *factor analysis* or *matrix factorization* (cf. [HTF09, Chap. 14]). Unfortunately, such a factor analysis is unstable in general, as long as the problem is high-dimensional ($n \ll p, d$) and noise is present. It was already succinctly emphasized by Vapnik in [Vap98, p. 12] that a direct approach is preferable in this situation:

“If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve the more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.”

Empirical risk minimization (P_K) precisely reflects this desire, as it only takes the “unspoiled” data $\{(x_i, y_i)\}_{i=1}^n$ as input. Our theoretical findings in Section 2 then show that the output vector $\hat{\beta}$ of (P_K) can be used to compute the actual estimator $\hat{z} = A^T \hat{\beta}$, which involves the (unknown) mixing matrix A only by a simple post-processing step. In that way, one may hope that the entire estimation process is less sensitive to noise and a small sample size. Moreover, as mentioned in Remark 3.15(2), it is often possible to extract the information of interest directly from $\hat{\beta}$ by means of expert knowledge, rather than explicitly computing $\hat{z} = A^T \hat{\beta}$.

5 Conclusion and Outlook

The key results of this work are the mismatch principle (Recipe 2.6) and the associated error bounds (Theorem 2.5 and Theorem 2.8). As demonstrated in Section 3, this simple recipe allows us to derive theoretical guarantees for empirical risk minimization in a very systematic way and thereby to explore the limits of fitting linear models to non-linear output rules. Of particular relevance in that context is the concept of target sets, which encodes (parametrizes) the information in which a user is interested. In a certain sense, this enables us to “bridge the gap” between specific estimation problems, such as learning single-index models, and the abstract setting of statistical learning.

From the application side, it has turned out in Section 3 that the estimator (P_K) is surprisingly robust against various types of model uncertainties and misspecifications. Our findings indicate that the estimated parameter vector often carries the desired information (if correctly interpreted), without performing any complicated pre-processing steps. As a general practical

guideline, we therefore conclude that non-adaptive empirical risk minimization may be already powerful enough to obtain a (rough) initial estimate of the true parameters. Ultimately, the outcome could at least serve as a good initialization for a more sophisticated method that is specifically tailored to the problem under investigation.

Let us close our discussion with several open issues, possible extensions as well as some interesting future research directions:

- *Hypothesis classes.* As long as the hypothesis class $\mathcal{F} \subset L^2(\mathbb{R}^p, \mu_x)$ is a convex set (cf. (1.1) and (1.2)), the arguments of Section 6 essentially remain valid. More precisely, the applied concentration inequalities from Theorem 6.4 and Theorem 6.5 hold true in a much more general setting (see [Men16]). Nevertheless, such an adaption would require a certain technical effort and is actually not the major concern of this paper. For that reason, we stick to case of linear hypothesis functions in order to work out the key ideas.

On the other hand, it is well known that the use of *non-convex* hypothesis classes can substantially increase the learning capacity of empirical risk minimization, e.g., in *deep learning* [GBC16; LBH15]. But this gain may come along with very challenging non-convex optimization tasks. An interesting approach related to single-index models (cf. Assumption 3.2) is taken by Yang et al. in [YWLEZ15], where \mathcal{F} basically corresponds to a subset of $\{x \mapsto g(\langle x, \beta \rangle) \mid \beta \in K\}$. In that way, they obtain a consistent estimator even in the sub-Gaussian case but it only applies under very restrictive assumption on $g: \mathbb{R} \rightarrow \mathbb{R}$. This drawback is mainly due to the fact that the landscape of the empirical risk in (1.2) becomes very complicated for non-linear output functions (see also [MBM16]). However, these results give evidence that the mismatch principle could be extended to non-linear hypothesis functions.

- *Consistent estimators.* The analysis of Subsection 3.1 reveals that (P_K) is typically inconsistent if the input data is non-Gaussian. Interestingly, one can resolve this issue by introducing an adaptive estimator, such as in [YBL17a; YBL17b; YBWL17]. But we rather aim at a generic strategy that does not leverage explicit knowledge of the probability distribution of x . The observation of Remark 3.11 could be useful at this point, as it indicates that (P_K) successfully detects the active variables (or the support of an index vector) under much weaker conditions.
- *Structured input data.* As already pointed out in Subsection 4.2, many problems in signal processing require to deal with structured data vectors, including *non-i.i.d.* samples $\{x_i\}_{i=1}^n$ and *heavy-tailed* feature variables. While the latter concern can be addressed by Mendelson's small ball method [Men15; Men18], the assumption of independent sampling is fundamental in empirical process theory. Some promising progress in this direction has been recently made by Dirksen and Mendelson in [DM18], showing that unbiased robust 1-bit compressed sensing is possible for non-Gaussian measurements by *dithering*.
- *Loss functions.* A natural variation of (P_K) is to consider a different convex *loss function* $\mathcal{L}: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$:

$$\min_{\beta \in K} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, \langle x_i, \beta \rangle). \quad (5.1)$$

Using the concept of *restricted strong convexity* according to [Gen17], one can establish similar error bounds for the estimator (5.1) as in Section 2. However, this asks for a careful adaption of the mismatch covariance (cf. Definition 2.3), which is based on a tighter estimate in (6.9). In that way, the adapted mismatch covariance can even take negative values, thereby improving the estimation performance in certain situations. Although not stated explicitly, this phenomenon is exploited in [GS18], where (5.1) is studied for the so-called *hinge loss function* in the context of 1-bit compressed sensing. This problem setting comes along with a series of technical challenges, since the hinge loss is not coercive and the expected risk minimizer is not necessarily uniquely defined (e.g., if $y = \text{sign}(\langle s, z_0 \rangle)$). Therefore, the geometry of the hypothesis set K plays a much larger role in [GS18], and in fact, the simplification of Remark 2.7 is not applicable anymore. This particularly shows that extending the mismatch principle with regard to (5.1) is not straightforward in general.

- *Regularized estimators.* From an optimization perspective, it is often beneficial to solve the *regularized* analog of (P_K) , that is

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2 + \lambda \|\beta\|_K \quad (5.2)$$

where $\lambda > 0$ is a (tunable) regularization parameter and $\|\cdot\|_K$ the Minkowski functional associated with K . Even though (5.2) is conceptually strongly related to (P_K) , it is far from obvious how one can adapt our results to these types of estimators. The interested reader is referred to [LM16; LM17] for recent advances in that direction.

6 Proofs

The proofs of Theorem 2.5 and Theorem 2.8 in Subsection 6.1 are consequences of a more general error bound (cf. Theorem 6.3) which is based on the so-called local mean width as refined complexity parameter. Theorem 6.3 is then proven in Subsection 6.2. This part is in fact the heart of our analysis and relies on two very recent concentration bounds for empirical stochastic processes (Theorem 6.4 and Theorem 6.5), allowing for *uniformly* controlling the linear and quadratic term of the excess risk functional.

6.1 Proofs of the Main Results (Theorem 2.5 and Theorem 2.8)

Let us begin with defining the local mean width, which, in a certain sense, generalizes the notion of conic mean width from Definition 2.4:

Definition 6.1 (Local mean width) Let $L \subset \mathbb{R}^d$ be a subset. The *local mean width* of L at scale $t > 0$ is defined as

$$w_t(L) := w\left(\frac{1}{t}L \cap \mathbb{S}^{d-1}\right) = \frac{1}{t}w(L \cap t\mathbb{S}^{d-1}).$$

The purpose of this geometric parameter is to measure the complexity of a set L in a “local” neighborhood of $\mathbf{0}$, whose size is determined by the scale t . Thus, intuitively speaking, $w_t(L)$ captures certain features of L at different resolution levels as t varies. The following lemma relates the local mean width to its conic counterpart, showing that the latter essentially corresponds to the limit case $t \rightarrow 0$:

Lemma 6.2 Let $L \subset \mathbb{R}^d$ be convex and $\mathbf{0} \in L$. Then, the mapping $t \mapsto w_t(L)$ is non-increasing, and it holds that $w_t(L) \rightarrow w_\wedge(L)$ as $t \rightarrow 0$. In particular, for all $t > 0$, we have that $w_t(L) \leq w_\wedge(L)$ and $w_t(L) \leq \frac{1}{t}w(L)$.

Proof. This follows from the definition and the basic properties of the (global) mean width as well as the inclusion $\frac{1}{t}L \subset \text{cone}(L)$ for all $t > 0$. \blacksquare

With regard to the problem setting of this work, the most important conclusion from Lemma 6.2 is that the local complexity of a set may be significantly smaller if the considered neighborhood is small but not infinitesimal (cf. [GS18, Sec. 2.1]). Indeed, as we will see next, this relaxation greatly pays off when investigating the performance of the estimator (P_K) . The following result basically shows that, by accepting an approximation error of order t , the required sample size is determined by the value of $w_t(\mathbf{A}^\top K - \mathbf{z}^\natural)$, rather than $w_\wedge(\mathbf{A}^\top K - \mathbf{z}^\natural)$:

Theorem 6.3 (Estimation via (P_K) – Local version) Let Assumption 2.1 be satisfied. Let $K \subset \mathbb{R}^p$ be a convex subset and fix a vector $\mathbf{z}^\natural \in \mathbf{A}^\top K \subset \mathbb{R}^d$. Then there exists numerical constants $C_1, C_2, C_3 > 0$ such that for every $u > 0$ and $t > 0$, the following holds true with probability at least $1 - 2 \exp(-C_1 \cdot u^2) - 2 \exp(-C_1 \cdot n) - \exp(-C_1 \cdot \kappa^{-4} \cdot n)$: If the number of observed samples obeys

$$n \geq C_2 \cdot \kappa^4 \cdot w_t^2(\mathbf{A}^\top K - \mathbf{z}^\natural), \quad (6.1)$$

and

$$t > C_3 \cdot \left(\kappa \cdot \sigma(\mathbf{z}^\natural) \cdot \frac{w_t(\mathbf{A}^\top K - \mathbf{z}^\natural) + u}{\sqrt{n}} + \rho(\mathbf{z}^\natural) \right), \quad (6.2)$$

then every minimizer $\hat{\beta}$ of (P_κ) satisfies $\|A^\top \hat{\beta} - z^\natural\|_2 < t$.

Compared to our main results from Section 2, the above guarantee is quite implicit, as both sides of the condition (6.2) depend on t . A convenient way to read Theorem 6.3 is therefore as follows: First, fix an estimation accuracy t that one is willing to tolerate. Then adjust the sample size n and the target vector z^\natural such that (6.2) is fulfilled (if possible at all). In particular, by allowing for a larger approximation error, the conditions of (6.1) and (6.2) become weaker according to Lemma 6.2. Consequently, the assertion of Theorem 6.3 enables a compromise between the desired accuracy and the budget of available samples, which is controlled by means of the local mean width. The interested reader is referred to [Gen17, Sec. III.D] for a more geometric interpretation of this important trade-off.

The remainder of this part is devoted to an application of Theorem 6.3 in order to derive the statements of Theorem 2.5 and Theorem 2.8. The basic idea is to set t equal to the right-hand side of the (simplified) error bounds in (2.3) and (2.9), respectively, and then to verify that the assumptions (6.1) and (6.2) hold true in either case.

Let us start with proving Theorem 2.8, which in fact requires less technical effort than the proof of Theorem 2.5 below.

Proof of Theorem 2.8. Following the above roadmap, we apply Theorem 6.3 with

$$t = 2C_3 \cdot \left(\kappa \cdot \sigma(z^\natural) \cdot \frac{w_\wedge(A^\top K - z^\natural) + u}{\sqrt{n}} + \rho(z^\natural) \right)$$

where $u > 0$ is specified later on. Let us first assume that $t > 0$. By Lemma 6.2, we have that

$$\begin{aligned} t &> C_3 \cdot \left(\kappa \cdot \sigma(z^\natural) \cdot \frac{w_\wedge(A^\top K - z^\natural) + u}{\sqrt{n}} + \rho(z^\natural) \right) \\ &\geq C_3 \cdot \left(\kappa \cdot \sigma(z^\natural) \cdot \frac{w_t(A^\top K - z^\natural) + u}{\sqrt{n}} + \rho(z^\natural) \right), \end{aligned}$$

implying that (6.2) is satisfied. Moreover, the condition (6.1) follows from (2.7):

$$n \gtrsim \kappa^4 \cdot \underbrace{\delta^{-2}}_{\geq 1} \cdot w_\wedge^2(A^\top K - z^\natural) \geq \kappa^4 \cdot w_t^2(A^\top K - z^\natural).$$

Consequently, Theorem 6.3 states that, with probability at least $1 - 2 \exp(-C_1 \cdot u^2) - 2 \exp(-C_1 \cdot n) - \exp(-C_1 \cdot \kappa^{-4} \cdot n)$, the following error bound holds true:

$$\|A^\top \hat{\beta} - z^\natural\|_2 < t = 2C_3 \cdot \left(\kappa \cdot \sigma(z^\natural) \cdot \frac{w_t(A^\top K - z^\natural) + u}{\sqrt{n}} + \rho(z^\natural) \right).$$

To obtain (2.8), we observe that

$$\kappa \cdot \frac{w_\wedge(A^\top K - z^\natural)}{\sqrt{n}} \stackrel{(2.7)}{\lesssim} \kappa^{-1} \cdot \delta,$$

and set $u = \kappa^{-2} \cdot \delta \cdot \sqrt{n}$. Note that the desired probability of success $1 - 5 \exp(-C \cdot \kappa^{-4} \cdot \delta^2 \cdot n)$ is achieved by adjusting the constant C and using that $\delta \leq 1$ and $\kappa \gtrsim 1$, where the latter is due to

$$\kappa \geq \|s\|_{\psi_2} = \sup_{z \in \mathbb{S}^{d-1}} \|\langle s, z \rangle\|_{\psi_2} \stackrel{(1.9)}{\geq} \sup_{z \in \mathbb{S}^{d-1}} 2^{-1/2} \underbrace{\mathbb{E}[\|\langle s, z \rangle\|^2]^{1/2}}_{\stackrel{(1.10)}{=} 1} = 2^{-1/2}. \quad (6.3)$$

Finally, let us consider the case of $t = 0$ (i.e., exact recovery), which is equivalent to $\sigma(z^\natural) = \rho(z^\natural) = 0$. While we cannot apply Theorem 6.3 directly in this situation, the proof steps from Subsection 6.2 can be still easily adapted: If $0 = \sigma(z^\natural) = \|y - \langle s, z^\natural \rangle\|_{\psi_2}$, the observation model is actually linear and noiseless: $y = \langle s, z^\natural \rangle = \langle x, \beta^\natural \rangle$. This immediately implies that the multiplier

term $\mathcal{M}(\cdot, z^\natural)$ vanishes, so that $\mathcal{E}(\beta, \beta^\natural) = \mathcal{Q}(A^\top \beta, z^\natural)$ for all $\beta \in \mathbb{R}^p$. Repeating the argument of Step 3 in the proof of Theorem 6.3 with $t = 1$ and $L = \text{cone}(A^\top K - z^\natural) \cap \mathbb{S}^{d-1}$, we conclude that, with probability at least $1 - \exp(-C_1 \cdot \kappa^{-4} \cdot n)$, it holds that

$$\frac{1}{n} \sum_{i=1}^n |\langle s_i, \mathbf{h} \rangle|^2 > 0 \quad (6.4)$$

for all $\mathbf{h} \in \text{cone}(A^\top K - z^\natural) \cap \mathbb{S}^{d-1}$. Now, if $\beta' \in K_{z^\natural, >0} = \{\beta \in K \mid A^\top \beta \neq z^\natural\}$, we have $\frac{A^\top \beta' - z^\natural}{\|A^\top \beta' - z^\natural\|_2} \in \text{cone}(A^\top K - z^\natural) \cap \mathbb{S}^{d-1}$ by the convexity of $A^\top K - z^\natural$. Therefore, on the event of (6.4), it follows that

$$\mathcal{E}(\beta', \beta^\natural) = \mathcal{Q}(A^\top \beta', z^\natural) = \|A^\top \beta' - z^\natural\|_2^2 \cdot \frac{1}{n} \sum_{i=1}^n \left| \langle s_i, \frac{A^\top \beta' - z^\natural}{\|A^\top \beta' - z^\natural\|_2} \rangle \right|^2 > 0.$$

Hence, β' cannot be a minimizer of (P_K) , implying that every minimizer $\hat{\beta}$ belongs to $K_{z^\natural, 0}$, i.e., $A^\top \hat{\beta} = z^\natural$. \blacksquare

The proof of Theorem 2.5 is slightly more involved. We loosely follow the argumentation from [GJ17b, Thm. 3.6].

Proof of Theorem 2.5. We apply Theorem 6.3 with

$$t = D \cdot \left[\kappa \cdot \left(\frac{w(A^\top K)}{\sqrt{n}} \right)^{1/2} + \frac{u}{\sqrt{n}} \right] + D' \cdot \rho(z^\natural) \quad (6.5)$$

where the values of $D, D' \gtrsim 1$ are specified later on. Let us first establish an upper bound on the local mean width:

$$\begin{aligned} w_t^2(A^\top K - z^\natural) &\leq \frac{1}{t^2} w^2(A^\top K - z^\natural) = \frac{1}{t^2} w^2(A^\top K) \\ &\stackrel{(6.5)}{\leq} \frac{1}{D^2 \cdot \kappa^2} \cdot \frac{\sqrt{n}}{w(A^\top K)} \cdot w^2(A^\top K) = \frac{1}{D^2 \cdot \kappa^2} \cdot \sqrt{n} \cdot w(A^\top K), \end{aligned} \quad (6.6)$$

where we have also used Lemma 6.2 and the translation invariance of the mean width (cf. [PV13b, Prop. 2.1]). Next, we adjust D such that $D \gtrsim \max\{1, \kappa \cdot \sigma(z^\natural)\}$ and use (6.6) to obtain the following bound on the right-hand side of (6.2):

$$\begin{aligned} &C_3 \cdot \left(\underbrace{\kappa \cdot \sigma(z^\natural)}_{\lesssim D \lesssim D^2} \cdot \frac{w_t(A^\top K - z^\natural) + u}{\sqrt{n}} + \rho(z^\natural) \right) \\ &\lesssim D^2 \cdot \frac{w_t(A^\top K - z^\natural)}{\sqrt{n}} + D \cdot \frac{u}{\sqrt{n}} + D' \cdot \rho(z^\natural) \\ &\stackrel{(6.6)}{\leq} D^2 \cdot \frac{1}{D \cdot \kappa} \cdot \frac{\sqrt{w(A^\top K)} \cdot n^{1/4}}{\sqrt{n}} + D \cdot \frac{u}{\sqrt{n}} + D' \cdot \rho(z^\natural) \\ &= D \cdot \underbrace{\kappa^{-1}}_{\stackrel{(6.3)}{\leq} 2\kappa} \cdot \left(\frac{w(A^\top K)}{\sqrt{n}} \right)^{1/2} + D \cdot \frac{u}{\sqrt{n}} + D' \cdot \rho(z^\natural) \\ &\lesssim D \cdot \left[\kappa \cdot \left(\frac{w(A^\top K)}{\sqrt{n}} \right)^{1/2} + \frac{u}{\sqrt{n}} \right] + D' \cdot \rho(z^\natural) = t. \end{aligned} \quad (6.7)$$

Consequently, if $D' \gtrsim 1$ is large enough and $D = \tilde{C} \cdot \max\{1, \kappa \cdot \sigma(z^\natural)\}$, with $\tilde{C} > 0$ being an appropriate numerical constant, we conclude that (6.7) implies the condition (6.2) of Theorem 6.3.

Moreover, combining (6.6) and (2.1), we obtain

$$w_t^2(\mathbf{A}^\top \mathbf{K} - \mathbf{z}^\natural) \leq \frac{1}{D^2 \cdot \kappa^2} \cdot \sqrt{n} \cdot w(\mathbf{A}^\top \mathbf{K}) \stackrel{(2.1)}{\lesssim} \frac{1}{D^2 \cdot \kappa^4} \cdot \underbrace{\delta^2}_{\leq 1} \cdot n \leq \frac{1}{\tilde{C}^2 \cdot \kappa^4} \cdot n,$$

which implies (6.1), where \tilde{C} might have to be slightly enlarged again.

Consequently, Theorem 6.3 with $u = \delta \cdot \sqrt{n}$ states that, with probability at least $1 - 2 \exp(-C_1 \cdot \delta^2 \cdot n) - 2 \exp(-C_1 \cdot n) - \exp(-C_1 \cdot \kappa^{-4} \cdot n)$, every minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) satisfies the error bound

$$\begin{aligned} \|\mathbf{A}^\top \hat{\boldsymbol{\beta}} - \mathbf{z}^\natural\|_2 &< t \lesssim \max\{1, \kappa \cdot \sigma(\mathbf{z}^\natural)\} \cdot \left[\underbrace{\kappa \cdot \left(\frac{w(\mathbf{A}^\top \mathbf{K})}{\sqrt{n}} \right)^{1/2}}_{\stackrel{(2.1)}{\lesssim} \delta} + \frac{u}{\sqrt{n}} \right] + \rho(\mathbf{z}^\natural) \\ &\lesssim \max\{1, \kappa \cdot \sigma(\mathbf{z}^\natural)\} \cdot \delta + \rho(\mathbf{z}^\natural). \end{aligned}$$

Finally, since $\delta \leq 1$ and $\kappa \gtrsim 1$ by (6.3), we achieve the desired probability of success $1 - 5 \exp(-C \cdot \kappa^{-4} \cdot \delta^2 \cdot n)$ by adjusting C . \blacksquare

6.2 Proof of Theorem 6.3

Throughout this subsection, we assume that Assumption 2.1 is satisfied, and recall the notations from Table 1. For the sake of readability, let us also introduce some additional notation: The objective function of (\mathbf{P}_K) — typically referred to as the *empirical risk* — is denoted by

$$\tilde{\mathcal{L}}(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^n (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2, \quad \boldsymbol{\beta} \in \mathbb{R}^p,$$

and the associated *excess risk* by

$$\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}') := \tilde{\mathcal{L}}(\boldsymbol{\beta}) - \tilde{\mathcal{L}}(\boldsymbol{\beta}'), \quad \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p.$$

Moreover, we define the following subsets of K :

$$\begin{aligned} K_{\mathbf{z}^\natural, t} &:= \{\boldsymbol{\beta} \in K \mid \|\mathbf{A}^\top \boldsymbol{\beta} - \mathbf{z}^\natural\|_2 = t\}, \\ K_{\mathbf{z}^\natural, > t} &:= \{\boldsymbol{\beta} \in K \mid \|\mathbf{A}^\top \boldsymbol{\beta} - \mathbf{z}^\natural\|_2 > t\}, \\ K_{\mathbf{z}^\natural, < t} &:= \{\boldsymbol{\beta} \in K \mid \|\mathbf{A}^\top \boldsymbol{\beta} - \mathbf{z}^\natural\|_2 < t\}. \end{aligned}$$

Finally, by the assumption $\mathbf{z}^\natural \in \mathbf{A}^\top K$ in Theorem 6.3, there exists $\boldsymbol{\beta}^\natural \in K$ such that $\mathbf{z}^\natural = \mathbf{A}^\top \boldsymbol{\beta}^\natural$. Note that this parameter vector could be highly non-unique, but the arguments below apply to every choice of $\boldsymbol{\beta}^\natural$ with $\mathbf{z}^\natural = \mathbf{A}^\top \boldsymbol{\beta}^\natural$.

The proof of Theorem 6.3 is based on a classical localization idea from learning theory (cf. [BBM05; Men02; MPT07]): The claim follows if we can show that $\hat{\boldsymbol{\beta}} \in K_{\mathbf{z}^\natural, < t}$ for every minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) . The key step of our proof below is therefore to verify that, with high probability, we have $\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}^\natural) > 0$ uniformly for all $\boldsymbol{\beta} \in K_{\mathbf{z}^\natural, t}$. This particularly implies that $K_{\mathbf{z}^\natural, t}$ cannot contain a minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) , since $\mathcal{E}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\natural) \leq 0$. Finally, by the convexity of $\tilde{\mathcal{L}}$ and K , it even follows that $\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}^\natural) > 0$ for all $\boldsymbol{\beta} \in K_{\mathbf{z}^\natural, > t}$. In turn, every minimizer of (\mathbf{P}_K) must belong to $K_{\mathbf{z}^\natural, < t}$, which concludes the argument; see Figure 2 for a visualization.

Step 1: Decomposing the Excess Risk

We first decompose the excess risk functional into its linear and quadratic part, corresponding to the first and second order Taylor term, respectively:

$$\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}^\natural) = \tilde{\mathcal{L}}(\boldsymbol{\beta}) - \tilde{\mathcal{L}}(\boldsymbol{\beta}^\natural)$$

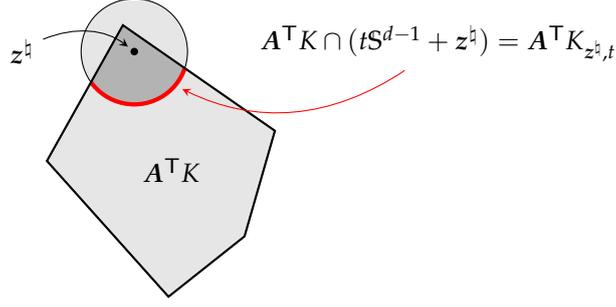


Figure 2: Illustration of the main proof argument. The key difficulty is to show that $\mathcal{E}(\beta, \beta^h) > 0$ for all $\beta \in K_{z^h, t}$ (see red arc). Hence, if $\hat{\beta}$ is a minimizer of (P_K) , we can conclude that $A^T \hat{\beta} \in \mathbb{R}^d$ must belong to the dark gray intersection $A^T K_{z^h, <t}$, which simply means that $\|A^T \hat{\beta} - z^h\|_2 < t$.

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2 - \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta^h \rangle)^2 \\
&= \frac{2}{n} \sum_{i=1}^n (\langle x_i, \beta^h \rangle - y_i) \langle x_i, \beta - \beta^h \rangle + \frac{1}{n} \sum_{i=1}^n |\langle x_i, \beta - \beta^h \rangle|^2 \\
&= \frac{2}{n} \sum_{i=1}^n \underbrace{(\langle A s_i, \beta^h \rangle - y_i)}_{=\langle s_i, z^h \rangle - y_i =: \zeta_i(z^h)} \langle A s_i, \beta - \beta^h \rangle + \frac{1}{n} \sum_{i=1}^n |\langle A s_i, \beta - \beta^h \rangle|^2 \\
&= \frac{2}{n} \sum_{i=1}^n \underbrace{\zeta_i(z^h) \langle s_i, A^T \beta - z^h \rangle}_{=: \mathcal{M}(A^T \beta, z^h)} + \frac{1}{n} \sum_{i=1}^n \underbrace{|\langle s_i, A^T \beta - z^h \rangle|^2}_{=: \mathcal{Q}(A^T \beta, z^h)}. \tag{6.8}
\end{aligned}$$

The factors $\zeta_i(z^h) := \langle s_i, z^h \rangle - y_i$ can be regarded as *multipliers* that do not depend on β , but are *not* independent of s_i . For that reason, $\mathcal{M}(A^T \beta, z^h)$ is referred to as the *multiplier term*, whereas $\mathcal{Q}(A^T \beta, z^h)$ is called the *quadratic term*. Moreover, we emphasize that both terms do actually only depend on $A^T \beta$ and $z^h = A^T \beta^h$, implying that $\mathcal{E}(\tilde{\beta}, \tilde{\beta}^h) = \mathcal{E}(\beta, \beta^h)$ for all $\tilde{\beta}, \tilde{\beta}^h \in \mathbb{R}^p$ with $A^T \tilde{\beta} = A^T \beta$ and $A^T \tilde{\beta}^h = A^T \beta^h$.

The goal of the following two steps is to establish *uniform* lower bounds for both empirical processes, $\beta \mapsto \mathcal{M}(A^T \beta, z^h)$ and $\beta \mapsto \mathcal{Q}(A^T \beta, z^h)$, on the neighborhood set $K_{z^h, t}$. Under the hypothesis of (6.1) and (6.2), this eventually leads to the desired positivity of the excess risk in Step 4.

Step 2: Bounding the Multiplier Term

Let us start with the multiplier term $\mathcal{M}(\cdot, z^h)$. For this purpose, we apply the following recent concentration inequality on empirical multiplier processes due to Mendelson [Men16], which is based on a sophisticated chaining argument:

Theorem 6.4 ([Men16, Thm. 4.4]) *Let Assumption 2.1 be satisfied and let $L \subset tS^{d-1}$ for some $t > 0$. Let ζ be a sub-Gaussian random variable (not necessarily independent of s) and let ζ_1, \dots, ζ_n be independent copies of ζ .¹ Then there exist numerical constants $C_1, C' > 0$ such that for every $u > 0$, the following holds true with probability at least $1 - 2 \exp(-C_1 \cdot u^2) - 2 \exp(-C_1 \cdot n)$:*

$$\sup_{h \in L} \left| \left(\frac{1}{n} \sum_{i=1}^n \zeta_i \langle s_i, h \rangle \right) - \mathbb{E}[\zeta \langle s, h \rangle] \right| \leq C' \cdot \kappa \cdot \|\zeta\|_{\psi_2} \cdot \frac{w(L) + u \cdot t}{\sqrt{n}}.$$

The definition of $\mathcal{M}(\cdot, z^h)$ implies that we can apply Theorem 6.4 with $L = A^T K_{z^h, t} - z^h \subset tS^{d-1}$ and $\zeta = \langle s, z^h \rangle - y$. Accordingly, we have that, with probability at least $1 - 2 \exp(-C_1 \cdot u^2) -$

¹More precisely, the independent copies of ζ are generated *jointly* with s and y , i.e., (s_i, y_i, ζ_i) is an independent copy of (s, y, ζ) .

$2 \exp(-C_1 \cdot n)$, the following bound holds true for every $\beta \in K_{z^{\natural}, t}$ ($\Rightarrow A^T \beta - z^{\natural} \in L$):

$$\begin{aligned}
\frac{1}{2} \cdot \mathcal{M}(A^T \beta, z^{\natural}) &= \frac{1}{n} \sum_{i=1}^n \zeta_i(z^{\natural}) \langle s_i, A^T \beta - z^{\natural} \rangle \\
&\geq \mathbb{E}[\zeta \langle s, A^T \beta - z^{\natural} \rangle] - C' \cdot \kappa \cdot \|\zeta\|_{\psi_2} \cdot \frac{w(L) + u \cdot t}{\sqrt{n}} \\
&= -t \cdot \underbrace{\left(\mathbb{E} \left[\zeta \langle s, \frac{z^{\natural} - A^T \beta}{t} \rangle \right] \right)}_{\stackrel{(*)}{\leq} \rho(z^{\natural})} + C' \cdot \kappa \cdot \underbrace{\|\zeta\|_{\psi_2}}_{=\sigma(z^{\natural})} \cdot \frac{\frac{1}{t} w(A^T K_{z^{\natural}, t} - z^{\natural}) + u}{\sqrt{n}} \\
A^T K_{z^{\natural}, t} &= A^T K \cap (tS^{d-1} + z^{\natural}) \\
&\geq -t \cdot \left(\rho(z^{\natural}) + C' \cdot \kappa \cdot \sigma(z^{\natural}) \cdot \frac{w_t(A^T K - z^{\natural}) + u}{\sqrt{n}} \right) \\
&\geq -t \cdot \max\{1, C'\} \cdot \underbrace{\left(\rho(z^{\natural}) + \kappa \cdot \sigma(z^{\natural}) \cdot \frac{w_t(A^T K - z^{\natural}) + u}{\sqrt{n}} \right)}_{=: t_0},
\end{aligned}$$

where $(*)$ is a consequence of the Cauchy-Schwarz inequality:

$$\begin{aligned}
\mathbb{E} \left[\zeta \langle s, \frac{z^{\natural} - A^T \beta}{t} \rangle \right] &= \mathbb{E} \left[(\langle s, z^{\natural} \rangle - y) \langle s, \frac{z^{\natural} - A^T \beta}{t} \rangle \right] \\
&= \left\langle \mathbb{E}[(\langle s, z^{\natural} \rangle - y) s], \frac{z^{\natural} - A^T \beta}{t} \right\rangle \\
&\leq \|\mathbb{E}[(\langle s, z^{\natural} \rangle - y) s]\|_2 \cdot \underbrace{\left\| \frac{z^{\natural} - A^T \beta}{t} \right\|_2}_{=1} = \rho(z^{\natural}). \tag{6.9}
\end{aligned}$$

Hence, we end up with

$$\mathcal{M}(A^T \beta, z^{\natural}) \geq -2 \cdot \max\{1, C'\} \cdot t \cdot t_0 \quad \text{for all } \beta \in K_{z^{\natural}, t}. \tag{6.10}$$

Step 3: Bounding the Quadratic Term

The quadratic term $\mathcal{Q}(\cdot, z^{\natural})$ can be handled by another chaining-based concentration inequality from random matrix theory:

Theorem 6.5 ([LMPV17, Thm. 1.3]) *Let Assumption 2.1 be satisfied and let $L \subset tS^{d-1}$ for some $t > 0$. Then there exists a numerical constant $C'' > 0$ such that for every $u \geq 0$, the following holds true with probability at least $1 - \exp(-u^2)$:*

$$\sup_{h \in L} \left| \left(\frac{1}{n} \sum_{i=1}^n |\langle s_i, h \rangle|^2 \right)^{1/2} - t \right| \leq C'' \cdot \kappa^2 \cdot \frac{w(L) + u \cdot t}{\sqrt{n}}.$$

Similarly to Step 2, we apply Theorem 6.5 with $L = A^T K_{z^{\natural}, t} - z^{\natural} \subset tS^{d-1}$ and $u = \sqrt{C_1 \cdot \kappa^{-4} \cdot n}$. Hence, with probability at least $1 - \exp(-C_1 \cdot \kappa^{-4} \cdot n)$, the following holds true for every $\beta \in K_{z^{\natural}, t}$:

$$\begin{aligned}
\sqrt{\mathcal{Q}(A^T \beta, z^{\natural})} &= \left(\frac{1}{n} \sum_{i=1}^n |\langle s_i, A^T \beta - z^{\natural} \rangle|^2 \right)^{1/2} \\
&\geq t - C'' \cdot \kappa^2 \cdot \frac{w(L) + u \cdot t}{\sqrt{n}} \\
&= t \cdot \left(1 - C'' \cdot \frac{\kappa^2 \cdot \frac{1}{t} w(A^T K_{z^{\natural}, t} - z^{\natural})}{\sqrt{n}} - C'' \cdot \frac{\kappa^2 \cdot u}{\sqrt{n}} \right)
\end{aligned}$$

$$\begin{aligned}
&= t \cdot \left(1 - C'' \cdot \underbrace{\frac{\kappa^2 \cdot w_t(\mathbf{A}^\top K - \mathbf{z}^\natural)}{\sqrt{n}}}_{\stackrel{(6.1)}{\leq 1/\sqrt{C_2}}} - C'' \sqrt{C_1} \right) \\
&\geq t \cdot \underbrace{\left(1 - C'' / \sqrt{C_2} - C'' \sqrt{C_1} \right)}_{=: C_0}.
\end{aligned}$$

Finally, we adjust the numerical constants C_1 and C_2 such that $C_0 > 0$, implying that

$$\mathcal{Q}(\mathbf{A}^\top \boldsymbol{\beta}, \mathbf{z}^\natural) \geq C_0^2 \cdot t^2 \quad \text{for all } \boldsymbol{\beta} \in K_{\mathbf{z}^\natural, t}. \quad (6.11)$$

Step 4: Bounding the Excess Risk and Conclusion

We now assume that the events of Step 2 and Step 3 have indeed occurred (with probability at least $1 - 2 \exp(-C_1 \cdot u^2) - 2 \exp(-C_1 \cdot n) - \exp(-C_1 \cdot \kappa^{-4} \cdot n)$). Hence, the lower bounds of (6.10) and (6.11) yield

$$\begin{aligned}
\mathcal{E}(\boldsymbol{\beta}, \boldsymbol{\beta}^\natural) &= \mathcal{Q}(\mathbf{A}^\top \boldsymbol{\beta}, \mathbf{z}^\natural) + \mathcal{M}(\mathbf{A}^\top \boldsymbol{\beta}, \mathbf{z}^\natural) \\
&\geq C_0^2 \cdot t^2 - 2 \cdot \max\{1, C'\} \cdot t \cdot t_0 \\
&= t \cdot \underbrace{(C_0^2 \cdot t - 2 \cdot \max\{1, C'\} \cdot t_0)}_{\stackrel{(6.2)}{> 0}} > 0 \quad \text{for all } \boldsymbol{\beta} \in K_{\mathbf{z}^\natural, t}, \quad (6.12)
\end{aligned}$$

where C_3 needs to be appropriately adjusted in (6.2), depending on the numerical constants C_0 and C' .

Since $\mathcal{E}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\natural) \leq 0$ for every minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) , we can conclude that $\hat{\boldsymbol{\beta}}$ cannot belong to $K_{\mathbf{z}^\natural, t}$. Next, on the same event, assume that there is a minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) with $\hat{\boldsymbol{\beta}} \in K_{\mathbf{z}^\natural, >t}$. By the convexity of K , it is not hard to see that the line segment $\lambda \mapsto \text{Seg}(\lambda) := \lambda \hat{\boldsymbol{\beta}} + (1 - \lambda) \boldsymbol{\beta}^\natural$ intersects $K_{\mathbf{z}^\natural, t}$ at a point $\boldsymbol{\beta}' := \lambda' \hat{\boldsymbol{\beta}} + (1 - \lambda') \boldsymbol{\beta}^\natural$ with $\lambda' \in [0, 1]$. Thus,

$$\begin{aligned}
\mathcal{E}(\text{Seg}(0), \boldsymbol{\beta}^\natural) &= \mathcal{E}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\natural) \leq 0, \\
\mathcal{E}(\text{Seg}(\lambda'), \boldsymbol{\beta}^\natural) &= \mathcal{E}(\boldsymbol{\beta}', \boldsymbol{\beta}^\natural) \stackrel{(6.12)}{>} 0, \\
\mathcal{E}(\text{Seg}(1), \boldsymbol{\beta}^\natural) &= \mathcal{E}(\boldsymbol{\beta}^\natural, \boldsymbol{\beta}^\natural) = 0,
\end{aligned}$$

which contradicts the convexity of the mapping $\lambda \mapsto \mathcal{E}(\text{Seg}(\lambda), \boldsymbol{\beta}^\natural)$. In turn, every minimizer $\hat{\boldsymbol{\beta}}$ of (\mathbf{P}_K) must belong to $K_{\mathbf{z}^\natural, <t}$, which implies the desired error bound $\|\mathbf{A}^\top \hat{\boldsymbol{\beta}} - \mathbf{z}^\natural\|_2 < t$. ■

Remark 6.6 (Related approaches) The above proof strategy loosely follows the learning framework of Mendelson from [Men15; Men18]. Based on a similar decomposition of the excess risk as in (6.8), the key idea of Mendelson's approach is to show that the quadratic term dominates the multiplier term except for a small neighborhood of the expected risk minimizer. Thus, according to the convexity argument of Step 4, one can conclude that the empirical risk minimizer belongs to this small neighborhood, which in turn yields an error bound.

While we are able to invoke a concentration result for the quadratic term due to the sub-Gaussianity of s , a major concern of [Men15; Men18] is that one can already establish lower bounds under a so-called small ball condition. This technique is referred to as *Mendelson's small ball method* and allows for proving learning guarantees under much weaker moment assumptions on the input data.

On the other hand, as pointed out in Subsection 4.1, a key difference to [Men15; Men18] is that we do not restrict to the expected risk minimizer as target vector but permit any choice $\mathbf{z}^\natural \in \mathbf{A}^\top K$. This particularly explains why the multiplier term $\mathcal{M}(\cdot, \mathbf{z}^\natural)$ needs to be treated somewhat differently in our setup. Apart from that, it is worth mentioning that the above analysis also improves the related approaches of [Gen17; PV16]. These works on *Gaussian* single-index models handle

the multiplier term by a more naive argument based on Markov's inequality, which eventually leads to a very pessimistic probability of success. \diamond

6.3 Proof of Proposition 2.2

Proof of Proposition 2.2. Since the covariance matrix of x is positive semi-definite and of rank d , there exists $U \in \mathbb{R}^{p \times d}$ with orthonormal columns and a diagonal matrix $D \in \mathbb{R}^{d \times d}$ with positive entries such that $\mathbb{E}[xx^\top] = UDDU^\top$. We set $s := D^{-1}U^\top x$ and $A := UD \in \mathbb{R}^{p \times d}$.

It is not hard to see that s is an isotropic, mean-zero random vector in \mathbb{R}^d . Indeed, we have $\mathbb{E}[s] = D^{-1}U^\top \mathbb{E}[x] = \mathbf{0}$ and

$$\mathbb{E}[ss^\top] = D^{-1}U^\top \mathbb{E}[xx^\top]UD^{-1} = D^{-1}U^\top UDDU^\top UD^{-1} = I_d.$$

Finally, let us compute the covariance matrix of $x - As$:

$$\begin{aligned} \mathbb{E}[(x - As)(x - As)^\top] &= \mathbb{E}[xx^\top] + \mathbb{E}[Ass^\top A^\top] - \mathbb{E}[xs^\top A^\top] - \mathbb{E}[Asx^\top] \\ &= UDDU^\top + UDDU^\top - \mathbb{E}[xx^\top]UD^{-1}DU^\top - UDD^{-1}U^\top \mathbb{E}[xx^\top] \\ &= 2UDDU^\top - UDDU^\top UD^{-1}DU^\top - UDD^{-1}U^\top UDDU^\top \\ &= 2UDDU^\top - 2UDDU^\top = \mathbf{0}. \end{aligned}$$

Hence, we conclude that $x = As$ almost surely. \blacksquare

Acknowledgments

The authors thank Chandrajit Bajaj, Peter Jung, Maximilian März, and Alexander Stollenwerk for fruitful discussions, in particular for pointing out potential applications. M.G. is supported by the European Commission Project DEDALE (contract no. 665044) within the H2020 Framework Program. G.K. acknowledges partial support by the Einstein Foundation Berlin, Einstein Center for Mathematics Berlin (ECMath), DFG CRC/TR 109 "Discretization in Geometry and Dynamics", DFG CRC 1114 "Scaling Cascades in Complex Systems", RTG BIOQIC (RTG 2260), RTG DAEDALUS (RTG 2433), and DFG SPP 1798 "Compressed Sensing in Information Processing".

References

- [ALPV14] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. "One-bit compressed sensing with non-Gaussian measurements". *Linear Algebra Appl.* 441 (2014), 222–239.
- [ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. "Living on the edge: phase transitions in convex programs with random data". *Inf. Inference* 3.3 (2014), 224–294.
- [BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. "Local Rademacher Complexities". *Ann. Statist.* 33.4 (2005), 1497–1537.
- [BM02] P. L. Bartlett and S. Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". *J. Mach. Learn. Res.* 3 (2002), 463–482.
- [BR08] H. D. Bondell and B. J. Reich. "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR". *Biometrics* 64.1 (2008), 115–123.
- [BB08] P. T. Boufounos and R. G. Baraniuk. "1-bit compressive sensing". *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS)*. 2008, 16–21.
- [BRGZ13] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. "Correlated variables in regression: Clustering and sparse estimation". *J. Statist. Plann. Inference* 143.11 (2013), 1835–1858.
- [BV11] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- [CRT05] E. J. Candès, J. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements". *Comm. Pure Appl. Math.* 59.8 (2005), 1207–1223.
- [CT05] E. J. Candès and T. Tao. "Decoding by Linear Programming". *IEEE Trans. Inf. Theory* 51.12 (2005), 4203–4215.
- [CT06] E. J. Candès and T. Tao. "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Trans. Inf. Theory* 52.12 (2006), 5406–5425.

- [CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. "The convex geometry of linear inverse problems". *Found. Comput. Math.* 12.6 (2012), 805–849.
- [Con+17] T. Conrad, M. Genzel, N. Cvetkovic, N. Wulkow, A. Leichtle, J. Vybiral, G. Kutyniok, and C. Schütte. "Sparse Proteomics Analysis – A compressed sensing-based approach for feature selection and classification of high-dimensional proteomics mass spectrometry data". *BMC Bioinform.* 18 (2017), 160.
- [CZ07] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.
- [DDEK12] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. "Compressed Sensing Theory and Applications". Ed. by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. Chap. Introduction to compressed sensing, 1–64.
- [DM18] S. Dirksen and S. Mendelson. "Robust one-bit compressed sensing with non-Gaussian measurements". Preprint arXiv:1805.09409. 2018.
- [Don06] D. L. Donoho. "Compressed sensing". *IEEE Trans. Inf. Theory* 52.4 (2006), 1289–1306.
- [EMR07] M. Elad, P. Milanfar, and R. Rubinfeld. "Analysis versus synthesis in signal priors". *Inverse Probl.* 23.3 (2007), 947–968.
- [FN14] M. A. T. Figueiredo and R. D. Nowak. "Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization". Preprint arXiv:1409.4005. 2014.
- [FN16] M. A. T. Figueiredo and R. D. Nowak. "Ordered Weighted ℓ_1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects". *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2016, 930–938.
- [FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.
- [Gen15] M. Genzel. "Sparse Proteomics Analysis". Master's thesis. TU Berlin, 2015.
- [Gen17] M. Genzel. "High-Dimensional Estimation of Structured Signals From Non-Linear Observations With General Convex Loss Functions". *IEEE Trans. Inf. Theory* 63.3 (2017), 1601–1619.
- [GJ17a] M. Genzel and P. Jung. "Blind Sparse Recovery From Superimposed Non-Linear Sensor Measurements". *Proceedings of the 12th International Conference on Sampling Theory and Applications (SampTA)*. 2017, 106–110.
- [GJ17b] M. Genzel and P. Jung. "Recovering Structured Data From Superimposed Non-Linear Measurements". Preprint arXiv:1708.07451. 2017.
- [GK16] M. Genzel and G. Kutyniok. "A Mathematical Framework for Feature Selection from Real-World Data with Non-Linear Observations". Preprint arXiv:1608.08852. 2016.
- [GKM17] M. Genzel, G. Kutyniok, and M. März. " ℓ^1 -Analysis Minimization and Generalized (Co-)Sparsity: When Does Recovery Succeed?" Preprint arXiv:1710.04952. 2017.
- [GS18] M. Genzel and A. Stollenwerk. "Robust 1-Bit Compressed Sensing via Hinge Loss Minimization". Preprint arXiv:1804.04846. 2018.
- [GM04] A. A. Giannopoulos and V. D. Milman. "Asymptotic Convex Geometry Short Overview". *Different Faces of Geometry*. Ed. by S. Donaldson, Y. Eliashberg, and M. Gromov. Springer, 2004, 87–162.
- [GMW16] L. Goldstein, S. Minsker, and X. Wei. "Structured signal recovery from non-linear and heavy-tailed measurements". Preprint arXiv: 1609.01025. 2016.
- [GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
- [Gor85] Y. Gordon. "Some inequalities for Gaussian processes and applications". *Isr. J. Math.* 50.4 (1985), 265–289.
- [Gor88] Y. Gordon. "On Milman's inequality and random subspaces which escape through a mesh in \mathbb{R}^n ". *Geometric aspects of functional analysis (1986/87)*. Ed. by J. Lindenstrauss and V. D. Milman. Vol. 1317. Lecture Notes in Math. Springer, 1988, 84–106.
- [HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- [Hor09] J. L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics*. Springer, 2009.
- [HZ10] J. Huang and T. Zhang. "The benefit of group sparsity". *Ann. Statist.* 38.4 (2010), 1978–2004.
- [Kay93] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. PTR Prentice-Hall, 1993.
- [KXAH09] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi. "Weighted ℓ^1 minimization for sparse recovery with prior information". *Proceedings of the IEEE International Symposium on Information Theory Proceedings (ISIT)*. 2009, 483–487.
- [LM16] G. Lecué and S. Mendelson. "Regularization and the small-ball method I: Sparse recovery". *Ann. Statist.* 46.2 (2016), 611–641.
- [LM17] G. Lecué and S. Mendelson. "Regularization and the small-ball method II: Complexity dependent error rates". *J. Mach. Learn. Res.* 18.146 (2017), 1–48.
- [LBH15] Y. LeCun, Y. Bengio, and G. Hinton. "Deep Learning". *Nature* 521.7553 (2015), 436.

- [LRW18] Y. Li, G. Raskutti, and R. Willett. “Graph-based regularization for regression problems with highly-correlated designs”. Preprint arXiv:1803.07658. 2018.
- [LMPV17] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. “A simple tool for bounding the deviation of random matrices on geometric sets”. *Geometric aspects of functional analysis*. Ed. by B. Klartag and E. Milman. Vol. 2169. Lecture Notes in Math. Springer, 2017, 277–299.
- [MN89] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.
- [MBM16] S. Mei, Y. Bai, and A. Montanari. “The landscape of empirical risk for non-convex losses”. Preprint arXiv:1607.06534. 2016.
- [Men16] S. Mendelson. “Upper bounds on product and multiplier empirical processes”. *Stoch. Proc. Appl.* 126.12 (2016), 3652–3680.
- [Men02] S. Mendelson. “Improving the sample complexity using global data”. *IEEE Trans. Inf. Theory* 48.7 (2002), 1977–1991.
- [Men15] S. Mendelson. “Learning without concentration”. *J. ACM* 62.3 (2015), Art. 21, 25.
- [Men18] S. Mendelson. “Learning without concentration for general loss functions”. *Probab. Theory Related Fields* 171.1–2 (2018), 459–502.
- [MPT07] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. “Reconstruction and subgaussian operators in asymptotic geometric analysis”. *Geom. Funct. Anal.* 17.4 (2007), 1248–1282.
- [MM18] M. Mondelli and A. Montanari. “On the Connection Between Learning Two-Layers Neural Networks and Tensor Decomposition”. Preprint arXiv:1802.07301. 2018.
- [NW72] J. A. Nelder and R. W. M. Wedderburn. “Generalized Linear Models”. *J. R. Stat. Soc. Ser. A (General)* 135.3 (1972), 370–384.
- [OKH12] S. Oymak, M. A. Khajehnejad, and B. Hassibi. “Recovery threshold for optimal weight ℓ_1 minimization”. *Proceedings of the IEEE International Symposium on Information Theory Proceedings (ISIT)*. 2012, 2032–2036.
- [PV13a] Y. Plan and R. Vershynin. “One-bit compressed sensing by linear programming”. *Comm. Pure Appl. Math.* 66.8 (2013), 1275–1297.
- [PV13b] Y. Plan and R. Vershynin. “Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach”. *IEEE Trans. Inf. Theory* 59.1 (2013), 482–494.
- [PV14] Y. Plan and R. Vershynin. “Dimension reduction by random hyperplane tessellations”. *Discrete Comput. Geom.* 51.2 (2014), 438–461.
- [PV16] Y. Plan and R. Vershynin. “The generalized Lasso with non-linear observations”. *IEEE Trans. Inf. Theory* 62.3 (2016), 1528–1537.
- [PVY16] Y. Plan, R. Vershynin, and E. Yudovina. “High-dimensional estimation with geometric constraints”. *Inf. Inference* 6.1 (2016), 1–40.
- [SB14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.
- [TAH15] C. Thrampoulidis, E. Abbasi, and B. Hassibi. “The LASSO with Non-linear Measurements is Equivalent to One With Linear Measurements”. *Advances in Neural Information Processing Systems 28 (NIPS)*. 2015, 3402–3410.
- [Tib96] R. Tibshirani. “Regression Shrinkage and Selection via the Lasso”. *J. Roy. Statist. Soc. Ser. B* 58.1 (1996), 267–288.
- [Vap98] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
- [Ver12] R. Vershynin. “Compressed Sensing Theory and Applications”. Ed. by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. Chap. Introduction to the non-asymptotic analysis of random matrices, 210–268.
- [Ver15] R. Vershynin. “Estimation in High Dimensions: A Geometric Perspective”. *Sampling Theory, a Renaissance*. Ed. by G. E. Pfander. Applied and Numerical Harmonic Analysis. Birkhäuser, 2015, 3–66.
- [YBL17a] Z. Yang, K. Balasubramanian, and H. Liu. “High-dimensional non-Gaussian single index models via thresholded score function estimation”. *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, 3851–3860.
- [YBL17b] Z. Yang, K. Balasubramanian, and H. Liu. “On Stein’s Identity and Near-Optimal Estimation in High-dimensional Index Models”. Preprint arXiv:1709.08795. 2017.
- [YBWL17] Z. Yang, K. Balasubramanian, Z. Wang, and H. Lui. “Learning Non-Gaussian Multi-Index Model via Second-Order Stein’s Method”. *Advances in Neural Information Processing Systems 30 (NIPS)*. 2017, 6097–6106.
- [YWLEZ15] Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang. “Sparse nonlinear regression: Parameter estimation and asymptotic inference”. Preprint arXiv: 1511.04514. 2015.
- [Yan+17] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov. “Misspecified Nonconvex Statistical Optimization for Phase Retrieval”. Preprint arXiv:1712.06245. 2017.
- [ZBC10] A. Zymnis, S. Boyd, and E. J. Candès. “Compressed sensing with quantized measurements”. *IEEE Signal Process. Lett.* 17.2 (2010), 149–152.