The Generalization Error of Dictionary Learning with Moreau Envelopes

Alexandros Georgogiannis [1]

Abstract

This is a theoretical study on the sample complexity of dictionary learning with general type of reconstruction losses. The goal is to estimate a $m \times d$ matrix D of unit-norm columns when the only available information is a set of training samples. Points x in $\mathbb{R}^m$ are subsequently approximated by the linear combination Da after solving the problem $\min_{a\in\mathbb{R}^d} \Phi(x - Da) + g(a)$ with function $g : \mathbb{R}^d \to [0, +\infty)$ being either an indicator function or a sparsity promoting regularizer. Here is considered the case where

$$\Phi(x) = \inf_{z\in\mathbb{R}^m} ||x - z||_2^2 + h(||z||_2)$$

and h is an even and univariate function on the real line. Connections are drawn between Φ and the Moreau envelope of h. A new sample complexity result concerning the k-sparse dictionary problem removes the spurious condition regarding the coherence of D appearing in previous works. Finally comments are made on the approximation error of certain families of losses. The derived generalization bounds are of order $\mathcal{O}(\sqrt{\log n/n})$.

1. Introduction

The dictionary learning problem, also known as sparse coding, was initially studied in the context of Neuroscience (Olshausen & Field, 1997); the relevant literature has grown enormously since; see (Zhang et al., 2015) and references therein. The problem is described as follows: given set $\{X_i\}_{i=1}^n \subset \mathbb{R}^m$ with n points sampled from an unknown fixed probability measure μ, a dictionary matrix $D \in \mathbb{R}^{m\times d}$ is to be constructed so that any unseen sample from μ can be approximated well by linear combinations of columns of D. The quality of approximation, for a given dictionary D, is measured by function $f_D : \mathbb{R}^m \to [0, +\infty)$ while D usually belongs to some predefined family of matrices. From the statistical learning theory perspective, the aim is to minimize the population risk $\mathcal{R} : \mathbb{R}^{m\times d} \to [0, +\infty)$,

$$\mathcal{R}(D) := \int f_D(X)d\mu = \int f_D d\mu, \tag{1}$$

when the only accessible information is a set of n training samples, say $\{X_i\}_{i=1}^n$, usually independent and identically distributed. Notation X is used for random vectors sampled from μ and notation x for real vectors in $\mathbb{R}^m$.

The empirical risk minimization principle (ERM) is a natural approach in search of the best dictionary (Vapnik, 1998). It suggests that since the only availiable information is the set of training samples, one should search for the matrix $\hat{D}_n$ that minimizes the empirical risk $\mathcal{R}_n : \mathbb{R}^{m\times d} \to [0, +\infty)$,

$$\mathcal{R}_n(D) := \frac{1}{n}\sum_{i=1}^n f_D(X_i). \tag{2}$$

The empirical estimate $\hat{D}_n$ is not of much use unless $|\mathcal{R}_n(\hat{D}_n) - \mathcal{R}(\hat{D}_n)|$ decreases as the number of samples n increases. Subsuming all computational difficulties on computing the global minimizing argument of (2), the problem addressed here is a "generalization problem". Given the family $\mathfrak{D}$ of all $m \times d$ matrices with unit-norm columns, we design a loss function f_D that measures the quality of approximation $x \simeq Da$ and ask: Does the difference

$$|\mathcal{R}(\hat{D}_n) - \inf_{D\in\mathfrak{D}} \mathcal{R}(D)| = \left| \int f_{\hat{D}_n} d\mu - \inf_{D\in\mathfrak{D}} \int f_D d\mu \right| \tag{3}$$

decrease as the number of samples n increases, and if so, at what rate? Or even further, if $\mathcal{R}(\hat{D}_n)$ is close to $\inf_{D\in\mathfrak{D}} \mathcal{R}(D)$, is $\hat{D}_n$ close to the global minimizing argument of $\mathcal{R}(D)$? Intuitively, the decrement of the absolute difference in expression (3) guarantees that by increasing the amount of data the probability that the population risk is within a very small distance of the optimal achievable gets arbitrarily close to one. The answers to the previous questions of course depend on the number of samples n, the predefined family of dictionaries $\mathfrak{D}$ and the form of f_D.

The proposed loss functions in the literature of dictionary learning vary according to the application but it would not be an exaggeration to say that almost all of them may be

[1]School of Electrical and Computer Engineering, Technical University of Crete, Greece. Correspondence to: Alexandros Georgogiannis <alexandrosgeorgogiannis@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

described by a function of the form:

$$f_D(x) := \inf_{a\in\mathbb{R}^d} \Phi(x - Da) + g(a), \tag{4}$$

with $\Phi : \mathbb{R}^m \to [0, +\infty)$ and $g : \mathbb{R}^d \to [0, +\infty)$. This article focuses on the generalization properties of dictionary learning when Φ has the form:

$$\Phi(x - Da) := \inf_{z\in\mathbb{R}^m} \frac{1}{2}||x - Da - z||_2^2 + h(||z||_2). \tag{5}$$

Functions h and g take values on $[0, +\infty)$ and are described in further detail later on. Definition (5) is not novel and has been used in many applications of sparse coding or dictionary learning (Adler et al., 2015; Amini et al., 2014; Forero et al., 2015; 2017; Jiang et al., 2015; Liu et al., 2015; Zhao & Tan, 2017). Although there is no formal robustness analysis yet to justify the superiority of (5) over the common square Euclidean loss $||x - Da||_2^2$, experimental evaluations in the previous applications suggest that this modification is a computationally "cheap" alternative, achieving better reconstruction error in some cases.

As can be seen from (4), if g is a sparsity promoting penalty on $\mathbb{R}^d$ then approximations that are linear combinations of a few columns of D are favored. The rationale behind the choice of h in (5) is not so obvious but if h satisfies a set of assumptions, then the following simplification holds true:

$$\Phi(x - Da) := e_h(||x - Da||_2). \tag{6}$$

Here, $e_h : \mathbb{R} \to \mathbb{R}$ is a univariate continuous function with special name and properties, the so called Moreau envelope of h (Rockafellar & Wets, 2009). Interestingly enough, the epigraphical form of e_h is completely determined by the generating function h. Roughly speaking, with a suitably chosen h, we can design loss functions f_D able to ignore the influence of points x, the distance of which from their approximation Da is above a predefined threshold. The consistency results of this study should be regarded as complementary extensions−and in some cases refinements−of the generalization bounds in (Gribonval et al., 2015b) and (Vainsencher et al., 2011). Contrary to previous works, all bounds presented here are valid for the whole of space of dictionaries with unit-norm columns.

In Section 3 is considered the case where g is a separable function from $\mathbb{R}^d$ to $[0, +\infty)$, that is, g is of the form $g(a) = \sum_{i=1}^d \hat{g}(a_i)$, and $\hat{g} : \mathbb{R} \to [0, +\infty)$ is univariate, continuous, even, and strictly increasing on $[0, +\infty)$, with minimum value $\hat{g}(0) = 0$. These assumptions are valid for many coordinate-separable regularizers, e.g., the l_p-norms on $\mathbb{R}^d$ and variants of the logarithmic function. Let us point out here that if $h = 0$, using the results of Section 3 we revert to previously known bounds for the penalized squared Euclidean loss $f_D(x) = \inf_{a\in\mathbb{R}^d} (1/2)\|x - Da\|_2^2 + g(a)$. Section 4 is an attempt to cover, beyond the class of strictly increasing penalties $\hat{g}$ of Section 3, continuous and bounded penalties from robust statistics, such as the MCP or SCAD. This type of penalty functions have achieved widespread use, and to the best of our knowledge, the bounds presented here are among the first that consider them.

However, the extended bounds of Section 4 turn out to be of limited applicability and do not work when g is the indicator function of all k-sparse vectors. To overcome this difficulty, in Section 5, we remove the continuity assumption from g and rely on combinatorial tools from Vapnik-Chervonenkis (VC) theory in order to present bounds valid for any bounded, lower semicontinuous function g. Whenever possible, the sample complexity bounds presented here are compared to similar ones in literature. Next follows a brief overview of the relevant literature.

1.1. Related Work and Contribution

The authors in (Gribonval et al., 2015b; Vainsencher et al., 2011) derive sample complexity bounds for the rate of convergence towards 0 of the absolute difference in (3) when $\Phi(x) = ||x||_2^2$, $\mathfrak{D}$ is a general constraint set, and $g(a)$ ranges from the l_p-norms and characteristic functions of compact sets to the indicator function of non-negative vectors or k-sparse vectors. The results in (Maurer & Pontil, 2010) are independent of dimension m, as well as some results in (Vainsencher et al., 2011).

A closer look on results of (Gribonval et al., 2015b) and (Vainsencher et al., 2011) concerning the finite case for dimension d, reveals that those are valid under joint assumptions on g and D. For instance, if g is the indicator function of k-sparse vectors in $\mathbb{R}^d$, then the generalization bounds in (Vainsencher et al., 2011) are valid under an incoherence assumption on D while in (Gribonval et al., 2015b) under a "restricted isometry"-like property. General non-asymptotic results can be extracted from the previous analyses, as the case $\Phi(x) = \omega(||x||)$, for any convex function $\omega : \mathbb{R} \to [0, +\infty)$ and any norm $||\cdot||$ on $\mathbb{R}^m$. In (Liu & Tao, 2016) authors focus on the l_1-non-negative matrix factorization problem where $\Phi(x) = ||x||_1$ and g is the indicator function of the non-negative orthant in $\mathbb{R}^d$.

The main contribution of our work is the addition of generalization bounds concerning loss functions that are combinations of Moreau envelopes with bounded and lower semicontinuous regularizers. Some results are refinements of previously known ones, meaning that a spurious assumption on dictionary D has been removed.

2. Preliminaries and some Technical Remarks

This is mainly a technical section where we take a closer look at the loss function f_D and describe the statistical

framework for the analysis. The value of f_D at point $x \in \mathbb{R}^m$, in light of equations (4) and (5), is expressed through the solution of the minimization problem:

$$\underbrace{\inf_{a\in\mathbb{R}^d}\left\{\overbrace{\inf_{z\in\mathbb{R}^m}\left\{\frac{1}{2}||x-Da-z||_2^2+h(||z||_2)\right\}}^{:=\Phi(x-Da)}+g(a)\right\}}_{f_D(x)}. \tag{7}$$

The close connection between Φ and h is captured in Lemma 1 that, among others, gives a description of the set of points $z \in \mathbb{R}^m$ that achieve the minimum in (5).

Lemma 1. *Let $h : \mathbb{R} \to [0, +\infty)$ be a lower semicontinuous (lsc) and even function with its restriction on $[0, +\infty)$ non-decreasing and $h(0) = 0$. Assume that the multivalued map $P_h : \mathbb{R} \rightrightarrows \mathbb{R}$, defined as*

$$P_h(t) := \operatorname{argmin}_{u\in\mathbb{R}} \frac{1}{2}(t-u)^2 + h(u), \tag{8}$$

*(**H1**) is odd, i.e., $P_h(-t) = -P_h(t)$, (**H2**) compact-valued, (**H3**) non-decreasing, (**H4**) has a closed graph and (**H5**) satisfies $P_h(t) \leq t$ for all $t \in \mathbb{R}$. Then function Φ in (5) becomes*

$$\Phi(x - Da) = e_h(||x - Da||_2), \tag{9}$$

where $e_h : \mathbb{R} \to [0, +\infty)$ is defined as

$$e_h(t) := \inf_{u\in\mathbb{R}} \frac{1}{2}(t-u)^2 + h(u),\ t \in \mathbb{R} \tag{10}$$

and is continuous with its restriction on $[0, +\infty)$ being non-decreasing. Furthermore, map $P^h : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$,

$$P^h(x-Da) := \operatorname{argmin}_{z\in\mathbb{R}^m} \frac{1}{2}||x-Da-z||_2^2 + h(||z||_2), \tag{11}$$

is equivalently represented as

$$P^h(x - Da) = \frac{x - Da}{||x - Da||_2} P_h(||x - Da||_2). \tag{12}$$

According to Lemma 1, if h satisfies a certain set of assumptions, then Φ is equal to the composition of the Moreau envelope of h with the Euclidean norm. Although the restrictions surrounding h and its proximal map seem to be strict, the lemma is valid for a large number of h and P_h pairs; see Section 3.1 in (Antoniadis, 2007) for various examples. Hereafter, any univariate function h in this article satisfies assumptions (**H1**) through (**H5**).

Example 1. *The case of the l_0-norm on $\mathbb{R}$ is an example that clearly describes the influence of h on the boundedness properties of Φ. Let $h : \mathbb{R} \to [0, \infty)$ be the l_0-(pseudo)norm on the real line defined as $l_0(t; \lambda) = \frac{\lambda^2}{2} 1_{\{t\neq 0\}}$ for some $\lambda > 0$. The l_0-norm satisfies all assumptions of Lemma 1: it is even, non-decreasing and lower semicontinuous while its proximal map P_{l_0} equals $P_{l_0}(t) = \operatorname{argmin}_{u\in\mathbb{R}} \frac{1}{2}(t-u)^2 + l_0(u; \lambda)$ and is defined as*

$$P_{l_0}(t) = \begin{cases} 0, & |t| < \lambda, \\ \{0, t\}, & |t| = \lambda, \\ t, & |t| > \lambda. \end{cases} \tag{13}$$

Now function $f_D : \mathbb{R}^m \to [0, +\infty)$ reads as

$$f_D(x) = \inf_{a\in\mathbb{R}^d}\left\{\underbrace{\tfrac{1}{2}\min\{||x-Da||_2^2, \lambda^2\}}_{:=e_{l_0}(||x-Da||_2)} + g(a)\right\}. \tag{14}$$

Boundedness of e_{l_0} implies that whenever the distance $||x - Da^*_D(x)||_2$ between a point x and its best linear approximation $Da^*_D(x)$ is greater than the predefined value λ, then $e_{l_0}(||x - Da^*_D(x)||_2) = \lambda^2/2$; here $a^*_D(x)$ is the (possibly multivalued) map

$$a^*_D(x) := \operatorname*{argmin}_{a\in\mathbb{R}^d} \left\{e_{l_0}(||x - Da||_2^2) + g(a)\right\}. \tag{15}$$

As long as g is globally upper bounded by some $M > 0$, if $||x - Da^*_D(x)||_2 > \lambda$ and $a^*_D(x)$ is sufficiently large, then $f_D(x) = \lambda^2/2 + M$. Since the empirical optimal dictionary $\hat{D}_n$ is defined through the minimization of the empirical risk $\mathcal{R}_n(D)$ in (2), and $\mathcal{R}_n$ is solely a function of D, points x for which $f(x) = \lambda^2/2 + M$ have no influence on the estimation of $\hat{D}_n$, and in that sense are "outliers".

The previous example is merely used to build some intuition behind the popularity of fidelity term (9) in the presence of "outliers". As "outliers" are considered points, the distance of which from their approximation $Da^*_D(x)$ is larger than a predefined threshold, say $\gamma > 0$. Note that any function h with proximal map satisfying $P_h(t) = t$ when $|t| > \gamma$ behaves like the l_0-norm in Example 1.

Remark 1. *The simple example described above may serve to anchor intuition, but it should be kept in mind that although we use the term "outlier", this is rather a study that focuses on the generalization error of dictionary learning. We do not provide robustness analysis of dictionary learning, since this would require a detailed mathematical definition of the notion "outlier". Robustness analysis results for Moreau envelope losses using notions from robust statistics, as the breakdown value, are provided in (Georgogiannis, 2016) for the generalized k-means problem; k-means is an unstructured dictionary learning problem−as $D^{m\times d}$ does not have unit-norm columns−with $m \ll d$, $h(t) = 0$, and $g(\cdot)$ the indicator of the basis vectors in $\mathbb{R}^d$.*

A robustness analysis different from the previous one has already been developed in (Gribonval et al., 2015a); the authors show that under coherence-based assumptions on D,

it is highly probable that the empirical risk $\frac{1}{n}\sum_{i=1}^{n} f_D(X_i)$, *when* $f_D(x) = \inf_{a\in\mathbb{R}^d} \frac{1}{2}||x - Da||_2^2 + g(a)$, *has a guaranteed empirical local minimum around the neighborhood of a population global minimum dictionary. A study motivated by the above references is of great interest and would fill the gap between theoretical and actual performance of dictionary learning algorithms using Moreau envelopes.*

Next is introduced the statistical learning framework. Denote as X, X_1, $X_2, \ldots$, independent and identically distributed random vectors with values in a closed ball in $\mathbb{R}^m$, say $\mathbb{B}_{\mathbb{R}^m}(T)$ with radius T centered at the origin, and denote as $\bar{\mathcal{P}}$ the set of all probability distributions μ on the Borel σ-algebra $\mathcal{B}(\mathbb{B}_{\mathbb{R}^m}(T))$ generated by this ball.[1] The aim is to show that the family of functions

$$\mathscr{F}_{\mathfrak{D}} = \left\{ f_D(x) : \mathbb{R}^d \to \mathbb{R};\ D \in \mathfrak{D} \right\} \tag{16}$$

has the *uniform convergence of empirical means* property on the measure space $(\mathbb{B}_{\mathbb{R}^m}(T), \mathcal{B}(\mathbb{B}_{\mathbb{R}^m}(T)), \mu)$, $\mu \in \bar{\mathcal{P}}$. Here $\mathfrak{D}$ is the set of all $m \times d$ real matrices with unit Euclidean-norm columns and f_D is of the form (4). The collection of functions $\mathscr{F}_{\mathfrak{D}}$ has the *uniform convergence of empirical means* (UCEM) property if the following convergence

$$\mathbb{P}\left\{ \underbrace{\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}}_{\sup_{D\in\mathfrak{D}}} \left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_D(X_i)}_{\mathcal{R}_n(D)} - \underbrace{\int f_D d\mu}_{\mathcal{R}(D)} \right| > \varepsilon \right\} \overset{n\to\infty}{\longrightarrow} 0 \tag{17}$$

is valid for every positive number ε and probability measure $\mu \in \bar{\mathcal{P}}$ on $\mathbb{B}_{\mathbb{R}^m}(T)$ (Vidyasagar, 2002).[2] This asymptotic result immediately answers the question raised in the introduction: if (17) holds true, then an application of inequality

$$\mathcal{R}(\hat{D}_n) - \inf_{D\in\mathfrak{D}} \mathcal{R}(D) \lesssim \sup_{D\in\mathfrak{D}} |\mathcal{R}_n(D) - \mathcal{R}(D)|$$

assures that $\mathcal{R}(\hat{D}_n)$ tends to the optimal value $\inf_{D\in\mathfrak{D}} \mathcal{R}(D)$ as the number of samples increases.

In most of our proofs, standard arguments from empirical processes theory are followed. In Sections 3 and 4 an appropriate form for h and g is chosen and then are used techniques based on either deterministic (Kolmogorov & Širjaev, 1993) or random ε-covers of the function class $\mathscr{F}_{\mathfrak{D}}$ (Györfi et al., 2006); let us recall their definitions.

Definition 1 (ε-cover)**.** *Let* $\varepsilon > 0$ *and let* $\mathscr{F}$ *be a class of functions from* $A \subseteq \mathbb{R}^m$ *to* $\mathbb{R}$. *Every finite collection of functions* $\tilde{f}_1, \ldots, \tilde{f}_N : \mathbb{R}^m \to \mathbb{R}$ *, for which for each* $f \in \mathscr{F}$ *there is a* $j(f) \in \{1, \ldots, N\}$ *such that*

$$||f - \tilde{f}_j||_\infty := \sup_{x\in A} |f(x) - \tilde{f}_j(x)| < \varepsilon, \tag{18}$$

is called ε*-cover of* $\mathscr{F}$ *under the supremum norm.*

Let $\mathscr{F}_{\mathfrak{D},\varepsilon} = \{f_1, \ldots, f_N\}$ be a ε-cover of $\mathscr{F}_{\mathfrak{D}}$ with respect to $||\cdot||_\infty$. As intuitively expected, the fewer the balls needed to cover $\mathscr{F}_{\mathfrak{D}}$, the smaller the $\mathscr{F}_{\mathfrak{D}}$.

Definition 2 (ε-covering number)**.** *Let* $\varepsilon > 0$ *and let* $\mathscr{F}$ *be a class of functions from a set* $A \subseteq \mathbb{R}^m$ *to* $\mathbb{R}$. *Let* $\mathcal{N}(\varepsilon, \mathscr{F}, ||\cdot||_\infty)$ *be the size of the smallest* ε*-cover of* $\mathscr{F}$ *under the supremum norm in (18). If no finite* ε*-cover exists, take* $\mathcal{N}(\varepsilon, \mathscr{F}, ||\cdot||_\infty) = \infty$. *Then* $\mathcal{N}(\varepsilon, \mathscr{F}, ||\cdot||_\infty)$ *is named the* ε*-covering number of* $\mathscr{F}$*, abbreviated to* $\mathcal{N}_\infty(\varepsilon, \mathscr{F})$.

The method of proof used in Sections 3 and 4 to establish the UCEM property for $\mathscr{F}_{\mathfrak{D}}$ when g is continuous is based on deterministic ε-covers, basic exponential inequalities and the Borel-Cantelli lemma. Unfortunately, this approach does not work when g is the indicator function of all k-sparse vectors; see Section 5. To overcome this difficulty, we rely on tools from VC theory, such as the shatter coefficient of the family of subgraphs of a function class.

Definition 3 (subgraphs of a function class)**.** *Consider a function class* $\mathscr{F}$ *with functions* $f : \mathbb{R}^m \to \mathbb{R}_+$. *The set*

$$\mathscr{F}^+ := \left\{ \{(x,t) \in \mathbb{R}^{m+1} : f(x) \geq t\};\ f \in \mathscr{F} \right\} \tag{19}$$

is the collection of all subgraphs of functions f *in* $\mathscr{F}$.

A family of subgraphs is a family of sets for which the shatter coefficient and VC dimension are defined as follows.

Definition 4 (shatter coefficient)**.** *Let* $\mathcal{A}$ *be a family of sets. For* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^m$, *let* $N_{\mathcal{A}}(x_1, \ldots, x_n)$ *be the number of different sets in* $\{\{x_1, \ldots, x_n\} \cap A;\ A \in \mathcal{A}\}$. *The* n*-th shatter coefficient* $s(\mathcal{A}, n)$ *of* $\mathcal{A}$ *is*

$$s(\mathcal{A}, n) := \max_{x_1,\ldots,x_n} N_{\mathcal{A}}(x_1, \ldots, x_n).$$

The shatter coefficient is the maximal number of different subsets of n *points that can be picked out by sets of* $\mathcal{A}$.

Definition 5 (VC dimension)**.** *Let* $\mathcal{A}$ *be a collection of sets with* $|\mathcal{A}| \geq 2$. *The largest integer* $k \geq 1$ *for which* $s(\mathcal{A}, k) = 2^k$ *is denoted by* $V_{\mathcal{A}}$ *and is called the VC dimension of the class* $\mathcal{A}$.

If for some hypothetical function class $\mathscr{F}$ the corresponding shatter coefficient $s(\mathscr{F}^+, n)$ is a polynomial of degree b with respect to n, i.e., $s(\mathscr{F}^+, n) = \mathcal{O}(n^b)$, then the popular Vapnik-Chervonenkis's inequality (Theorem 12.5 in (Devroye et al., 1997)) implies UCEM for $\mathscr{F}$. Later on, in Section 5, we show that this is the case for $s(\mathscr{F}_{\mathfrak{D}}^+, n)$ as well, where $\mathscr{F}_{\mathfrak{D}}^+$ denotes the collection of all subgraphs of functions in $\mathscr{F}_{\mathfrak{D}}$ with g the indicator of k-sparse vectors in $\mathbb{R}^d$—recall the definitions of f_D and $\mathscr{F}_{\mathfrak{D}}$ in (4) and (16).

[1] The Borel σ-algebra $\mathcal{B}(Y)$ of a subset Y of a metric space S is the one generated by $\mathcal{B}(Y) = \{Y \cap E : E \in \mathcal{B}(S)\}$. Thus the Borel σ-algebra $\mathcal{B}(\mathbb{B}_{\mathbb{R}^m}(T))$ is precisely the class of all subsets of $\mathbb{B}_{\mathbb{R}^m}(T)$ which are Borel sets in $\mathbb{R}^m$ (Folland, 2013).

[2] Symbol $\mathbb{P}$ in (17) denotes the product measure $\mu_{\times_1^\infty}$ on the product σ-algebra $\bigotimes_1^\infty \mathcal{B}(\mathbb{B}_{\mathbb{R}^m}(T))$ (Folland, 2013).

3. The case of a separable, continuous, even, and strictly increasing $g : \mathbb{R}^d \to [0, +\infty)$

In this section we prove the UCEM property for the function class $\mathscr{F}_{\mathfrak{D}}$ in (16) when f_D is defined as

$$f_D(x) := \inf_{a \in \mathbb{R}^d} \{e_h(||x - Da||_2) + g(a)\} \tag{20}$$

and g has the following form:

$$g(a) = \sum_{i=1}^{d} \hat{g}(a_i). \tag{21}$$

Here is assumed that $\hat{g} : \mathbb{R} \to [0, +\infty)$ is a univariate, continuous, even, and strictly increasing function on $[0, +\infty)$ with minimum value $\hat{g}(0) = 0$. The aforementioned assumptions on g are valid for many coordinate-separable regularizers, e.g., the l_p norms on $\mathbb{R}^d$, $g(a) = \lambda||a||_p$, $0 < p < +\infty$ for some $\lambda > 0$, and the log penalty function $g(a) = \sum_{i=1}^{d} \frac{\lambda}{\log(\gamma+1)} \log(\gamma|a_i| + 1)$, $\gamma > 0$. From now on, a separable function of the previous form is called (strictly) increasing if for all i, $\hat{g}(a_i)$ is (strictly) increasing as $|a_i| \to +\infty$. The main result is the following theorem.

Theorem 1. *Let $\varepsilon > 0$ and consider the function class $\mathscr{F}_{\mathfrak{D}}$ in (16) with $f_D : \mathbb{B}_{\mathbb{R}^m}(T) \to [0, e_h(T)]$ and $g : \mathbb{R}^d \to [0, +\infty)$ defined as in (20) and (21) respectively. Then*

$$\mathbb{P}\left\{\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu\right| > \varepsilon\right\} \leq 2\left(\frac{9d\hat{g}^{-1}(e_h(T))}{2\varepsilon}\right)^{md} e^{-\frac{2n\varepsilon^2}{9e_h(T)^2}}. \tag{22}$$

Furthermore,

$$\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu\right| \to 0 \quad (n \to \infty) \tag{23}$$

almost surely, for any $\mu \in \bar{\mathcal{P}}$. Hence, the function class $\mathscr{F}_{\mathfrak{D}}$ has the UCEM property with respect to $\bar{\mathcal{P}}$.

An outline of Theorem's 1 proof is the following:

1. We define map F that maps any $m \times p$ matrix to some function of the form (20). Using appropriate metrics, F is shown to be globally Lipschitz.
2. The Lipschitz continuity of F and the covering number of $\mathfrak{D}$ generate an upper bound for $\mathcal{N}_\infty(\varepsilon, \mathscr{F}_{\mathfrak{D}})$.
3. Standard theorems from the empirical process theory imply the concentration result in (22) and finally prove the UCEM property for the function class $\mathscr{F}_{\mathfrak{D}}$.

The above outline makes clear that the main difficulty in proving Theorem 1 is the verification of the Lipschitz continuity of map F. Let us mention that factor $\frac{d\hat{g}^{-1}(e_h(T))}{2}$ appearing on the right hand side (rhs) of (22) is an upper bound for the Lipschitz constant of the aforementioned map.

There exist other approaches that do not require any form of continuity on F to prove the UCEM property for $\mathscr{F}_{\mathfrak{D}}$. However, theoretical questions regarding the existence of the optimal dictionary are answered quite easily if we manage to construct such a map. For example, as well known, a continuous map maps compact sets to compact sets. If F is continuous, the compactness of $\mathfrak{D}$ implies the compactness of $\mathscr{F}_{\mathfrak{D}}$. This in turn implies the existence of the optimal solution f_D^* of minimization problem $\inf_{f_D \in \mathscr{F}_{\mathfrak{D}}} \int f_D d\mu$; indeed, the integral is a linear operator and $\mathscr{F}_{\mathfrak{D}}$ is compact.

Remark 2. *Another theoretical question, of great importance for the measure theory enthusiasts, concerns the measurability of the supremum appearing on the left hand side (lhs) of (22). This is a random variable of which the measurability stems from total boundedness of $\mathscr{F}_{\mathfrak{D}}$ with respect to the supremum norm $||f||_\infty := \sup_{\{x:||x||_2 \leq T\}} |f(x)|$.*

Proposition 1. *Assume a set up as the one in Theorem 1. Then for any $\delta > 0$,*

$$\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu\right| \leq \mathcal{O}\left(\sqrt{\frac{\log(nd)}{n}}\right) \tag{24}$$

with probability at least $1 - \delta$.

The term $\log(d)$ in (24), responsible for the sub-optimality of the bound in case of convex Moreau envelopes, results from our proof method; similar bounds in the literature are of order $\mathcal{O}(\sqrt{\log n/n})$ (Gribonval et al., 2015b; Vainsencher et al., 2011). This term is eliminated in Lemma 2 below to end up with a same order upper bound. The latter is in alignment with the sample complexity results presented in (Gribonval et al., 2015b) and (Liu & Tao, 2016) for the cases where $f_D(x)$ equals $\inf_{a \in \mathbb{R}^d} \frac{1}{2}||x - Da||_2^2 + g(a)$ and $\inf_{a \in \mathbb{R}^d} \frac{1}{2}||x - Da||_1 + g(a)$ respectively.

Lemma 2. *Let $L > \frac{d\hat{g}^{-1}(e_h(T))}{2}$ and define $\beta > 0$ as $\beta := md \max\{\log(6L\sqrt{8}), 1\}$. Assume that n satisfies condition*

$$\frac{n}{\log(n)} \geq \max\left\{8, \left(\frac{1}{2\sqrt{8}L}\right)^2 \beta\right\} \tag{25}$$

and consider the same set up as in Theorem 1. Then,

$$\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu\right| \leq \frac{2}{\sqrt{8}}\sqrt{\frac{\beta \log n}{n}} + \frac{1}{\sqrt{8}}\sqrt{\frac{\beta + t}{n}}, \tag{26}$$

with probability at least $1 - 2e^{-t}$.

The rationale behind this lemma is to find conditions, under which for large values of the sample size n, an exponential tail for the error kicks in but without the term $\log(d)$ of inequality (24). Although the analysis seems finer, the result is valid only if the sample size satisfies the quite strict and complex inequality (25).

4. The case of a separable, continuous, even, and bounded $g : \mathbb{R}^d \to [0, +\infty)$

Analysis of Section 3 covers a broad range of regularizers g but it does not cover popular penalty functions from robust statistics, like SCAD, $g_{scad}(a) = \sum_{i=1}^d \hat{g}_{scad}(a_i)$ or MCP, $g_{mcp}(a) = \sum_{i=1}^d \hat{g}_{mcp}(a_i)$ (Mazumder et al., 2012):

$$\hat{g}_{scad}(t;\lambda,\gamma) = \begin{cases} \lambda t, & t \le \lambda \\ \frac{\lambda\gamma t - \frac{1}{2}(t^2+\lambda^2)}{\gamma-1}, & \lambda < t \le \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & t > \lambda\gamma, \end{cases} \tag{27}$$

and

$$\hat{g}_{mcp}(t;\lambda,\gamma) = \begin{cases} \lambda t - \frac{t^2}{2\gamma}, & t \le \lambda \\ \frac{1}{2}\gamma\lambda^2, & t > \gamma\lambda. \end{cases} \tag{28}$$

Although the previous univariate functions are continuous, even, and satisfy the assumptions of Lemma 1, they fail to satisfy the assumptions of Theorem 1 because they are bounded above and thus not strictly increasing.

This section is an attempt to extend the results of Section 3 and handle a very special case of coordinate-separable regularizers: those $g(a) = \sum_{i=1}^d \hat{g}(a_i)$, where $\hat{g} : \mathbb{R} \to [0, +\infty)$ is not only continuous and symmetric around zero, but also strictly increasing up to some point in $[0, +\infty)$ and then constant. For this purpose, we require that $\hat{g}$ satisfies the additional (strict) inequality

$$e_h(T) < \sup_{t\in\mathbb{R}} \hat{g}(t). \tag{29}$$

Under assumption (29), all results presented in Section 3 remain valid; see the relevant discussion in Appendix A.5. Example 2 describes the impact of this assumption on penalty function $\hat{g}_{mcp}$ while the same applies to $\hat{g}_{scad}$.

Example 2. *Let h be the l_0-norm on the real line and $\hat{g}(a) = \hat{g}_{mcp}(a;\gamma,\lambda_2)$, $\lambda_2 > 0$; recall the definition of the l_0-norm on the real line: $l_0(t;\lambda_1) = \frac{\lambda_1}{2} 1_{\{t\neq 0\}}$, $\lambda_1 > 0$. In this case, the Moreau envelope is*

$$e_{l_0}(t;\lambda_1) = \frac{1}{2}\min\{t^2, \lambda_1^2\}$$

and $g_{mcp}(a) = \sum_{i=1}^d \hat{g}_{mcp}(a_i;\gamma,\lambda_2)$. Now assumption (29) reads as

$$\sup_{t\in\mathbb{R}} \hat{g}_{mcp}(t;\lambda_2,\gamma) > \frac{1}{2}\min\{T^2, \lambda_1^2\} \tag{30}$$

or after some simple algebraic calculations,

$$\frac{1}{2}\lambda_2^2\gamma > \frac{1}{2}\min\{T^2,\lambda_1^2\} \;\Leftrightarrow\; \lambda_2 > \sqrt{\frac{1}{\gamma}\min\{T^2,\lambda_1^2\}}. \tag{31}$$

Thus, function class $\mathscr{F}_{\mathfrak{D}}$ in (16) with $f_D(x)$ defined as

$$f_D(x) = \inf_{a\in\mathbb{R}^d}\left\{ e_{l_0}(||x - Da||_2) + \sum_{i=1}^d \hat{g}_{mcp}(a_i;\gamma,\lambda_2)\right\}$$

has the UCEM property only for pairs of values (λ_1,λ_2) with $\lambda_2 > \sqrt{\frac{1}{\gamma}\min\{T^2,\lambda_1^2\}}$.

Example 2 reveals that the ease with which we extend the results of Section 3 has great impact on the diversity of functions $\hat{g}$ that we could handle. In order to use the upper bounds in Proposition 1 or Lemma 2, our focus needs to be restricted on families $\mathscr{F}_{\mathfrak{D}}$ where the rightmost inequality in (31) holds. This artificial restriction on the available pair of values (λ_1,λ_2) makes this extension quite useless; in many applications, when setting up λ_1 and λ_2, we search on a wider grid of values.

In the next section, we remove the continuity assumption from g and derive generalization bounds valid for any bounded lsc function, such as SCAD, MCP or the indicator function of all k-sparse vectors in $\mathbb{R}^d$.

5. The case of the indicator function of all k-sparse vectors in $\mathbb{R}^d$ and its extension

Denote as $\Sigma_k = \{a \in \mathbb{R}^d : |\{i : a_i \neq 0\}| = k\}$ the set of all k-sparse vectors in $\mathbb{R}^d$. The approach followed in Sections 3 and 4 to prove the UCEM property for $\mathscr{F}_{\mathfrak{D}}$ heavily relies on the assumption that g is continuous. Consequently, it does not work for the function

$$g(a) = \begin{cases} 0, & \text{if } a \in \Sigma_k \\ \infty, & \text{otherwise,} \end{cases} \tag{32}$$

the non-separable and lsc indicator function of all k-sparse vectors in $\mathbb{R}^d$. Using combinatorial tools from VC theory, we remove the spurious condition on the coherence of $D \in \mathfrak{D}$ appearing in previous works (Gribonval et al., 2015b; Vainsencher et al., 2011) and prove the UCEM property when g is bounded and lsc. Starting the analysis with function (32), the results are then extended to cover any bounded lsc function on $\mathbb{R}^d$ with values in $[0, +\infty)$.

Next is presented Proposition 2, a modification of Theorem 20 in (Vainsencher et al., 2011): it states that map F from metric space $(\mathfrak{D}, ||\cdot||_{1,2})$ to metric space $(\mathscr{F}_{\mathfrak{D}}, ||\cdot||_\infty)$,

$$\mathscr{F}_{\mathfrak{D}} := \left\{ \min_{a\in\Sigma_k} e_h(||x - Da||_2);\ D \in \mathfrak{D}\right\}, \tag{33}$$

is not uniformly Lipschitz for any Lipschitz constant.[3] This is the main reason we resign (ourselves) from previous proof techniques. Without an explicit upper bound for the Lipschitz constant of map F, we cannot infer a bound for the covering number of $\mathscr{F}_{\mathfrak{D}}$ in terms of the one of $\mathfrak{D}$.

Proposition 2. *Consider the family of functions $\mathscr{F}_{\mathfrak{D}}$ in (33). Then, there exist $\gamma > 0$ and $q \in \mathbb{B}_{\mathbb{R}^m}(T)$ such that for every $\varepsilon > 0$, there exist $D, D' \in \mathfrak{D}$ such that*

$$\max_{1\leq j\leq d} ||D_{\cdot,j} - D'_{\cdot,j}||_2 \leq \varepsilon \quad \textit{but} \quad |f_D(q) - f_{D'}(q)| > \gamma.$$

In other words, map F from $\mathfrak{D}$ to $\mathscr{F}_{\mathfrak{D}}$ with $D \in \mathfrak{D} \mapsto F(D) \in \mathscr{F}_{\mathfrak{D}}$ is not globally Lipschitz.

Proposition 2 suggests that there are two ways to overcome the limitations when dealing with k-sparse vectors: either more restrictions shall be imposed on the class of dictionaries $\mathfrak{D}$ or a different proof method has to be followed. The former approach was adopted by (Vainsencher et al., 2011) and (Gribonval et al., 2015b), who both use deterministic ε-net arguments under an incoherence assumption on D and a lower RIP-property, respectively. In such a way, the authors restrict their analysis on a subspace of original space of all unit-norm column dictionaries.

Here the latter approach is adopted: without additional assumptions on the dictionaries, standard tools from VC theory verify the UCEM property of $\mathscr{F}_{\mathfrak{D}}$. The main result is Proposition 3 which delivers an upper bound for $s(\mathscr{F}^+_{\mathfrak{D}}, n)$, the shatter coefficient of

$$\mathscr{F}^+_{\mathfrak{D}} := \left\{\{(x,t) \in \mathbb{R}^{m+1} : f_D(x) \geq t\};\ f_D \in \mathscr{F}_{\mathfrak{D}}\right\}; \tag{34}$$

the previous set collection is the family of all subgraphs of functions f_D which belong to $\mathscr{F}_{\mathfrak{D}}$ (as defined in (33)).

Proposition 3. *The shatter coefficient $s(\mathscr{F}^+_{\mathfrak{D}}, n)$ of the collection of sets $\mathscr{F}^+_{\mathfrak{D}}$, as defined in (34), is bounded above as $s(\mathscr{F}^+_{\mathfrak{D}}, n) \leq \left(\frac{en}{\alpha(m,d)}\right)^{\alpha(m,d)}$ with $\alpha(m,d)$ independent of n and $\alpha(m,d) = ((m+d)^2 + 3(m+d))/2 + 1$.*

A direct use of Proposition's 3 bound in the popular Vapnik-Chervonenkis's theorem (Theorem 12.5, (Devroye et al., 1997)) generates Theorem 2 and its byproduct Proposition 4. The latter characterizes the rate of convergence to zero of the difference of the sample average from the true mean of $f_D(X)$. All random variables appearing in Theorems 2, 3 and Proposition 4 below are assumed measurable.

Theorem 2. *Let $f_D : \mathbb{B}_{\mathbb{R}^m}(T) \to [0, e_h(T)]$ for each f_D in the function class $\mathscr{F}_{\mathfrak{D}}$ in (33) and let $\varepsilon > 0$. Then*

$$\mathbb{P}\left\{\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu\right| > \epsilon\right\} \leq 8s(\mathscr{F}^+_{\mathfrak{D}}, n)e^{-\frac{n\varepsilon^2}{32e_h(T)^2}}. \tag{35}$$

Furthermore,

$$\sum_{n=1}^{\infty}\left(\frac{en}{\alpha(m,d)}\right)^{\alpha(m,d)} e^{-\frac{2n\varepsilon^2}{32e_h(T)^2}} < \infty \tag{36}$$

for all $\varepsilon > 0$, and by the Borel-Cantelli lemma,

$$\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu\right| \to 0 \quad \textit{(almost surely)}. \tag{37}$$

Hence, function class $\mathscr{F}_{\mathfrak{D}}$ has the UCEM property.

Proposition 4. *Assume the same setup as in Theorem 2 and let $\delta > 0$. With probability at least $1 - \delta$, holds true that*

$$\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu\right| \leq \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right). \tag{38}$$

When $f_D(x) = e_h(||x||_2)$ and $e_h(t) = t^2$, the bounds for the absolute difference in the rhs of (38) in (Gribonval et al., 2015b) and (Vainsencher et al., 2011) are of order $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$ and $\mathcal{O}\left(\sqrt{\frac{\log(\sqrt{n})}{n}}\right)$ respectively.

Although Proposition 4 is suboptimal compared to the latter, let us recall that Proposition 4 is valid for all dictionaries with unit-norm columns, in contrast to the last referenced bounds that do not cover the whole of space $\mathfrak{D}$. With slight modifications, Theorem 2 extends to Theorem 3 which covers any bounded lsc function g, including MCP or SCAD.

Theorem 3. *Let $g : \mathbb{R}^d \to [0, +\infty)$ bounded and lsc, and $\mathscr{F}_{\mathfrak{D}}$ the function class with functions $f_D : \mathbb{R}^m \to [0, +\infty)$,*

$$f_D(x) := \inf_{a\in\mathbb{R}^d} e_h(||x - Da||_2) + g(a). \tag{39}$$

Let $\varepsilon > 0$. Then

$$\mathbb{P}\left\{\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu\right| > \epsilon\right\} \leq 8s(\mathscr{F}^+_{\mathfrak{D}}, n)e^{-\frac{n\varepsilon^2}{32e_h(T)^2}}, \tag{40}$$

where $s(\mathscr{F}^+_{\mathfrak{D}}, n) \leq \left(\frac{en}{\alpha(m,d)}\right)^{\alpha(m,d)}$ and $\alpha(m,d) := ((m+d)^2 + 3(m+d))/2 + 1$. Furthermore, with probability at least $1 - \delta$, holds true that

$$\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu\right| \leq \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right). \tag{41}$$

[3] Although Proposition 2 has the same formulation as Theorem 20 of (Vainsencher et al., 2011), the latter cannot apply directly in our case except for $k = 2$. Proposition 2 clarifies through minor modifications what happens when $k > 2$.

6. On the approximation error when $m \gg d$

As already mentioned, our aim is to analyze the expected reconstruction error of the learned bases $\hat{D}_n$, $\mathcal{R}(\hat{D}_n) := \int f_{\hat{D}_n} d\mu$, when $\hat{D}_n$ is the (ERM)-estimator $\hat{D}_n := \operatorname{argmin}_{D\in\mathfrak{D}} \mathcal{R}_n(D)$. This reconstruction error decomposes into the estimation error ϵ_{est} and the approximation error ϵ_{app} as follows:

$$\mathcal{R}(\hat{D}_n) = \underbrace{\mathcal{R}(\hat{D}_n) - \mathcal{R}(D^*)}_{:=\,\epsilon_{\text{est}}} + \underbrace{\mathcal{R}(D^*)}_{:=\,\epsilon_{\text{app}}}, \tag{42}$$

where $D^* := \operatorname{argmin}_{D\in\mathfrak{D}} \mathcal{R}(D)$ is the optimal dictionary, the global minimizer of the population risk. The estimation error exists because $\hat{D}_n$ is just an estimate for D^*. The approximation error measures the risk of restricting ourselves to $\mathfrak{D}$ rather than to a larger family of matrices. The optimal choice for $\hat{D}_n$ guarantees that both ϵ_{est} and ϵ_{app} are the smallest possible. The estimation error is bounded as

$$\epsilon_{\text{est}} := \mathcal{R}(\hat{D}_n) - \mathcal{R}(D^*) \leq 2 \sup_{D\in\mathfrak{D}} |\mathcal{R}_n(D) - \mathcal{R}(D)|. \tag{43}$$

In previous sections was proven that the rhs of (43) approaches zero as $n \to +\infty$ and that, in view of (42), the reconstruction error $\mathcal{R}(\hat{D}_n)$ is asymptotically equal to ϵ_{app}. The approximation error does not depend on the sample size n; it is determined by the family of losses under study and the probability distribution of the data. In the k-sparse case, ϵ_{app} is rarely zero, even for well behaved probability measures μ. The authors in (Vovk, 2016), Section 24, show that the two objectives, of good data approximation and of sparsity of the combination vector a, are incompatible if the data distribution puts its mass far from any low dimensional subspace and in such cases $\epsilon_{\text{app}} \neq 0$.

In this section, assuming $m \gg d$, ϵ_{app} is considered a function of d. An upper bound for ϵ_{app} as $d \to m$, valid for any probability measure $\mu \in \bar{\mathcal{P}}$, gives insights to the problem of approximating points in $\mathbb{R}^m$ with combinations of points lying on subspaces of dimension d. Following the approach in (Liu & Tao, 2016), we relate the optimal population risk $\mathcal{R}(D^*)$ to the quantization error of probability measure μ.

Next proposition is meaningful only in the case where g is the indicator function of special compact subsets of $\mathbb{R}^d$, i.e., $g(a) = 1_{\mathcal{K}}(a)$ with $\mathcal{K} \subset \mathbb{R}^d$. Specifically, $\mathcal{K}$ is assumed to contain the basis vectors of the positive orthant. Assumptions **(H1)** to **(H5)** in Lemma 1 regarding h and its proximal map P_h remain valid, but is also required that

$$(\mathbf{H6}) \quad P_h(t) = 0, \ \text{ when } t \in [-\tau, \tau], \tag{44}$$

for some predefined value $\tau > 0$. These assumptions simplify the proof of Proposition 5: if they are true, then the Moreau envelope behaves like the quadratic function t^2 in a neighborhood around zero. Although assumptions **(H1)** through **(H6)** may seem strict, they are valid for many univariate penalty functions and compact sets, such as the closed unit-norm balls in $\mathbb{R}^d$.

Proposition 5. *Assume $m \gg d$. Let the family of losses $\mathscr{F}_{\mathfrak{D}}$ be defined as $\mathscr{F}_{\mathfrak{D}} := \{f_D(x);\ D \in \mathfrak{D}\}$ with*

$$f_D(x) := \inf_{a\in\mathbb{R}^d} e_h(||x - Da||_2) + 1_{\mathcal{K}}(a), \tag{45}$$

where $1_{\mathcal{K}}$ is the indicator function of some compact set $\mathcal{K} \subset \mathbb{R}^d$ that contains all basis vectors of the positive orthant, i.e., $\{e_j\}_1^d \in \mathcal{K}$ and e_j is the j-th column of the identity matrix.

If $h : \mathbb{R} \to [0, +\infty)$ satisfies the assumptions of Lemma 1 while its proximal map P_h satisfies assumptions **(H1)-(H6)***, then for the approximation error it holds true that*

$$\mathcal{R}(D^*) := \inf_{D\in\mathfrak{D}} \int f_D(x) d\mu \leq \mathcal{O}(d^{-2/m}). \tag{46}$$

The bound in (46) depends on m and d. Despite being "weak", as $m^{-2/m} \to 1$, this upper bound provides an insight to the problem: when m is fixed, but sufficiently large, and $d \to m$, the approximation error decreases as d increases, at rate $\mathcal{O}(d^{-2/m})$. Let us note here that the UCEM property for the family of risk functions $\mathscr{F}_{\mathfrak{D}}$ as defined in Proposition 5 can be proved using elements from the proof of Theorem 1 in Section 3.

7. Conclusions

This article is a theoretical analysis on the sample complexity of dictionary learning when the loss function to be minimized is the sum of the Moreau envelope of some univariate lsc function h on the real line and a regularization function g. We derive generalization bounds for a wide range of g, including the case of the indicator function of all k-sparse vectors. As a byproduct of this analysis is provided some intuition behind the popularity of loss functions under study in the context of "gross outliers", that is, samples with arbitrary "large" values. Finally, we comment on the approximation error of an ideal family of losses when the dimension $m \gg d$, where d is the size of the dictionary. In the future, it would be interesting to characterize the differentiability properties of the losses under study. Such an analysis would have direct practical applications on the design of numerical optimization algorithms.

Acknowledgements

The author thanks Athanasios P. Liavas and the anonymous reviewers for helpful comments and suggestions that improved the quality of the article.

References

Adler, A., Elad, M., Hel-Or, Y., and Rivlin, E. Sparse coding with anomaly detection. *Journal of Signal Processing Systems*, 79(2):179–188, 2015.

Akama, Y. and Irie, K. VC dimension of ellipsoids. *arXiv preprint arXiv:1109.4347*, 2011.

Amini, S., Sadeghi, M., Joneidi, M., Babaie-Zadeh, M., and Jutten, C. Outlier-aware dictionary learning for sparse representation. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pp. 1–6. IEEE, 2014.

Antoniadis, A. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.

Delattre, S., Graf, S., Luschgy, H., and Pages, G. Quantization of probability distributions under norm-based distortion measures. *Statistics & Decisions*, 22(4/2004): 261–282, 2004.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer New York, 1997. ISBN 9780387946184. URL https://books.google.gr/books?id=uDgXoRkyWqQC.

Folland, G. B. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

Forero, P. A., Shafer, S., and Harguess, J. Structured outlier models for robust dictionary learning. In *Information Sciences and Systems (CISS), 2015 49th Annual Conference on*, pp. 1–6. IEEE, 2015.

Forero, P. A., Shafer, S., and Harguess, J. D. Sparsity-driven laplacian-regularized outlier identification for dictionary learning. *IEEE Transactions on Signal Processing*, 65 (14):3803–3817, 2017.

Georgogiannis, A. Robust k-means: a theoretical revisit. In *Advances in Neural Information Processing Systems*, pp. 2883–2891, 2016.

Gribonval, R., Jenatton, R., and Bach, F. Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory*, 61(11):6298–6319, 2015a.

Gribonval, R., Jenatton, R., Bach, F., Kleinsteuber, M., and Seibert, M. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015b.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Jiang, W., Nie, F., and Huang, H. Robust dictionary learning with capped l1-norm. In *IJCAI*, pp. 3590–3596, 2015.

Kolmogorov, A. N. and Širjaev, A. N. *Selected works of AN Kolmogorov. Vol. 3, Information theory and the theory of algorithms*. Kluwer, 1993.

Liu, H., Qin, J., Cheng, H., and Sun, F. Robust kernel dictionary learning using a whole sequence convergent algorithm. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 3678–3684. AAAI Press, 2015.

Liu, T. and Tao, D. On the performance of manhattan nonnegative matrix factorization. *IEEE transactions on neural networks and learning systems*, 27(9):1851–1863, 2016.

Maurer, A. and Pontil, M. k-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

Mazumder, R., Friedman, J. H., and Hastie, T. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 2012.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

Pollard, D. *Convergence of stochastic processes*. Springer Science & Business Media, 1984.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Vainsencher, D., Mannor, S., and Bruckstein, A. M. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(Nov):3259–3281, 2011.

Vapnik, V. N. Statistical learning theory, 1998.

Vidyasagar, M. *A theory of learning and generalization*. Springer-Verlag New York, Inc., 2002.

Vovk, V. *Measures of Complexity: Festschrift for Alexey Chervonenkis*. Springer Publishing Company, Incorporated, 2016. ISBN 3319357786, 9783319357782.

Yu, Y., Zheng, X., Marchetti-Bowick, M., and Xing, E. P. Minimizing nonconvex non-separable functions. In *AISTATS*, 2015.

Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. A survey of sparse representation: algorithms and applications. *IEEE access*, 3:490–530, 2015.

Zhao, R. and Tan, V. Y. F. Online nonnegative matrix factorization with outliers. *IEEE Transactions on Signal Processing*, 65(3):555–570, Feb 2017. ISSN 1053-587X. doi: 10.1109/TSP.2016.2620967.