A. Appendix to: "The Generalization Error of Dictionary Learning with Moreau Envelopes"

A.1. Proof of Lemma 1

Proof. For a proof of (9) and (12) see Corollary 1 in (Georgogiannis, 2016). The continuity of e_h follows from Theorem 1.25 in (Rockafellar & Wets, 2009). From the calculus rules of the generalized subgradients (Theorem 9.13 and Corollary 10.9 in (Rockafellar & Wets, 2009)):

$$\partial e_h(t) \subseteq t - P_h(t),$$

and since $t - P_h(t) \geq 0$ (by assumption) for every $t \geq 0$, we conclude that e_h is non-decreasing on $[0, +\infty)$. □

A.2. Proof of Theorem 1

The proof technique outlined in Section 3 is heavily motivated from the proof of Theorem 2 in (Gribonval et al., 2015b). Since it is quite lengthy, we split it into two parts. In the first part, we prove the Lipschitz continuity of the map $F : \mathfrak{D} \mapsto \mathscr{F}_{\mathfrak{D}}$ and in the second the UCEM property for $\mathscr{F}_{\mathfrak{D}}$.

A.2.1. Lipschitz continuity of map $F : \mathfrak{D} \mapsto \mathscr{F}_{\mathfrak{D}}$

The key in this approach is the following lemma taken from the theory of general metric spaces.

Lemma 3. *Let (M, ρ) and (M_1, ρ_1) be metric spaces, $K \subset M$, and define the map $F : K \mapsto M_1$. If F satisfies*

$$\rho_1(F(x), F(y)) \leq L\rho(x, y), \text{ for any } x, y \in K,$$

for some $L > 0$, i.e., F is a Lipschitz continuous map from K to M_1 with constant L, then

$$\mathcal{N}(L\varepsilon, F(K), \rho_1) \leq \mathcal{N}(\varepsilon, K, \rho), \tag{47}$$

for every $\varepsilon > 0$. Here, $\mathcal{N}(L\varepsilon, F(K), \rho_1)$ and $\mathcal{N}(\varepsilon, K, \rho)$ denote the covering numbers of the sets $F(K)$ and K, under the metrics ρ_1 and ρ, respectively.

Proof. The proof is quite straightforward; given an ε-cover of K with size N, say $\{x_1, \dots, x_N\}$, and any $y \in K$, there exists a x_i in in the ε-cover of K such that $\rho(y, x_i) \leq \varepsilon$. Thus

$$\rho_1(F(y), F(x_i)) \leq L\rho(y, x_i) \leq L\varepsilon.$$

In words, for any point $F(y)$, there is a point $F(x_i)$ such that $F(y)$ and $F(x_i)$ are $L\varepsilon$ close; the set $\{F(x_1), \dots, F(x_N)\}$ constitutes an ε-cover. Since $\mathcal{N}(L\varepsilon, F(K), \rho_1) \leq |\{F(x_1), \dots, F(x_N)\}|$, where $|\{F(x_1), \dots, F(x_N)\}|$ denotes the cardinality of this set, the claim follows. □

The first step is to define the metrics used on the (metric) spaces spaces $\mathfrak{D}$ and $\mathscr{F}_{\mathfrak{D}}$.

Definition 6. *Let $p, q \geq 1$. Then, a matrix $A \in \mathbb{R}^{m\times d}$ can be seen as an operator $A : (\mathbb{R}^d, ||\cdot||_p) \mapsto (\mathbb{R}^m, ||\cdot||_q)$. The $l_{p,q}$-induced norm of A is*

$$||A||_{p,q} := \sup_{\substack{x\in\mathbb{R}^d \\ ||x||_p=1}} ||Ax||_q.$$

□

In the sequel, $||\cdot||_{p,q}$ is fixed to $||\cdot||_{1,2}$ which is equivalent to

$$||A||_{1,2} = \max_{1\leq j\leq d} ||A_{\cdot,j}||_2; \tag{48}$$

$A_{\cdot,j}$ is the j-th column of $A \in \mathbb{R}^{m\times d}$ (see also Lemma 17 in (Vainsencher et al., 2011)). The metric on $\mathscr{F}_{\mathfrak{D}}$ is the supremum norm on the ball $\mathbb{B}_{\mathbb{R}^m}(T)$:

$$||f||_\infty := \sup_{x\in\mathbb{B}_{\mathbb{R}^m}(T)} |f(x)|. \tag{49}$$

If $f_D \in \mathscr{F}_{\mathfrak{D}}$ and $g(0) = 0$, then $||f_D||_\infty \leq e_h(||x - D0||_2) + g(0) = e_h(T)$ for all f_D in $\mathscr{F}_{\mathfrak{D}}$. Next, define the map $F : \mathfrak{D} \mapsto \mathscr{F}_{\mathfrak{D}}$ between $(\mathfrak{D}, ||\cdot||_{1,2})$ and $(\mathscr{F}_{\mathfrak{D}}, ||\cdot||_\infty)$. Our aim is to show that F is uniformly Lipschitz continuous or else, there is a constant $L > 0$ such that

$$||F(D) - F(D')||_\infty \leq L||D - D'||_{1,2} \text{ for any } D, D' \text{ in } \mathfrak{D}. \tag{50}$$

For this purpose, the technical Lemmas 4, 5, and 6 below are needed.

Lemma 4 states that the infimum over $a \in \mathbb{R}^d$ in the definition of f_D is achieved. Fix $x \in \mathbb{B}_{\mathbb{R}^m}(T)$ and $D \in \mathbb{R}^{m\times d}$, and consider the function $\mathcal{L}_x(D, \cdot) : \mathbb{R}^d \to \mathbb{R}$ defined as

$$\mathcal{L}_x(D, a) := e_h(||x - Da||_2) + g(a). \tag{51}$$

Associate with $\mathcal{L}_x(D, \cdot)$ the set

$$\mathfrak{U}_0^D(x) := \left\{a \in \mathbb{R}^d : \mathcal{L}_x(D, a) \leq f_D(x)\right\}. \tag{52}$$

Set $\mathfrak{U}_0^D(x)$ contains the values of a that achieve the minimum value of $\mathcal{L}_x(D, a)$ for a fixed D (not necessarily in $\mathfrak{D}$), i.e., $\mathfrak{U}_0^D(x) = \operatorname{argmin}_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$. Note that $f_D(x) = \inf_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$ is still well defined but not necessarily in $\mathscr{F}_{\mathfrak{D}}$, unless D has unit-norms columns. As stated below, $\mathfrak{U}_0^D(x)$ is non-empty and compact for any $x \in \mathbb{B}_{\mathbb{R}^m}(T)$ and D.

Lemma 4. *Let $D \in \mathbb{R}^{m\times d}$ and consider any $x \in \mathbb{B}_{\mathbb{R}^m}(T)$. The value $f_D(x) := \inf_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$ is finite and $\operatorname{argmin}_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$ is non-empty and compact.*

Proof. Let $f_D(x) = \inf_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$; because $\mathcal{L}_x(D, \cdot)$ is proper (as a sum of proper functions), $f_D(x) < +\infty$.[4] For some $\beta \in (f_D(x), +\infty)$, the level set $\text{lev}_{\leq\beta}\mathcal{L}_x(D, \cdot) := \{a \in \mathbb{R}^d : \mathcal{L}_x(D, a) \leq \beta\}$ is non-empty; it is closed because $\mathcal{L}_x(D, \cdot)$ is lsc (in fact it is continuous) as the sum of two continuous functions and bounded because both e_h and g are non-decreasing (and not constant) as $||a||_2 \to +\infty$. The sets $\text{lev}_{\leq\beta}\mathcal{L}_x(D, \cdot)$ for $\beta \in (f_D(x), +\infty)$ are therefore compact and nested: $\text{lev}_{\leq\beta}\mathcal{L}_x(D, \cdot) \subset \text{lev}_{\leq\beta'}\mathcal{L}_x(D, \cdot)$ when $\beta < \beta'$. The intersection of this family of sets, which is $\text{lev}_{\leq f_D(x)}\mathcal{L}_x(D, \cdot)$, is therefore non-empty and compact. Since $\mathcal{L}_x(D, \cdot)$ does not take the value $-\infty$ nowhere, we conclude that $f_D(x)$ is finite. Under the previous assumptions, $\inf_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$ can be written as $\min_{a\in\mathbb{R}^d} \mathcal{L}_x(D, a)$ and the claim follows. □

Remark 3. *In most cases of interest, the functions e_h and g are bounded from above, i.e., $\sup_{x\in\mathbb{B}_{\mathbb{R}^m}(T)} e_h(x)$ and $\sup_{x\in\mathbb{B}_{\mathbb{R}^m}(T)} g(x)$ are finite. So, the value of β in Lemma 4 could be the minimum of the latter two supremas. Also, it is easily gleaned from the proof of Lemma 4 that its conclusions still hold for any non-decreasing lsc function g.* □

Next, a bound for the absolute value of the difference between $|e_h(x) - e_h(x')|$ when h satisfies the assumptions in Lemma 1 is given; these assumptions on h remain valid throughout the article.

Lemma 5. *Let $e_h(x) := \inf_{z\in\mathbb{R}^m} \frac{1}{2}||x - z||_2^2 + h(z)$, where $h : \mathbb{R}^m \to \mathbb{R}$ is lsc and proper. Then*

$$|e_h(x) - e_h(x')| \leq \frac{1}{2}||x - x'||_2^2, \tag{53}$$

for any x, x' in $\mathbb{R}^m$.

[4] We call f proper if $f(x) < \infty$ for at least one $x \in \mathbb{R}^n$, and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$; in words, if the domain of f is a nonempty set on which f is finite, see page 5 in (Rockafellar & Wets, 2009).

Proof. Let x and x' in $\mathbb{R}^m$. Then:

$$\begin{aligned} e_h(x) - e_h(x') &= \inf_{z\in\mathbb{R}^d}\left\{\frac{1}{2}||x-z||_2^2 + g(z)\right\} - \inf_{o\in\mathbb{R}^d}\left\{\frac{1}{2}||x'-o||_2^2 + g(o)\right\} \\ &\leq \inf_{z\in\mathbb{R}^d}\left\{\frac{1}{2}||x-z||_2^2 + g(z)\right\} + \sup_{o\in\mathbb{R}^d}\left\{-\frac{1}{2}||x'-o||_2^2 - g(o)\right\} \\ &= \sup_{o\in\mathbb{R}^d}\left\{\inf_{z\in\mathbb{R}^d}\left\{\frac{1}{2}||x-z||_2^2 + g(z)\right\} - \frac{1}{2}||x'-o||_2^2 - g(o)\right\} \\ &\leq \sup_{o\in\mathbb{R}^d}\left\{\frac{1}{2}||x-o||_2^2 + g(o) - \frac{1}{2}||x'-o||_2^2 - g(o)\right\} \\ &\leq \sup_{o\in\mathbb{R}^d}\left\{\frac{1}{2}||x-o||_2^2 - \frac{1}{2}||x'-o||_2^2\right\} \\ &\leq \frac{1}{2}||x-x'||_2^2. \end{aligned} \tag{54}$$

Interchanging the roles of x and x', we conclude the result. □

Lemma 6. *Fix $x \in \mathbb{B}_{\mathbb{R}^m}(T)$ and let $D, D' \in \mathbb{R}^{m\times d}$. If there exist non-negative constants $C_{D,x}$ and $C_{D',x}$, such that $\sup_{a\in\mathfrak{U}_0^D(x)} ||a||_1 \leq C_{D,x}$ and $\sup_{a\in\mathfrak{U}_0^{D'}(x)} ||a||_1 \leq C_{D',x}$, then*

$$\frac{|F(D) - F(D')|}{||D'-D||_{1,2}} \leq \frac{1}{2}||D'-D||_{1,2} \max\{C_{D,x}, C_{D',x}\}^2, \tag{55}$$

Proof. Fix x in $\mathbb{B}_{\mathbb{R}^m}(T)$ and let $D, D' \in \mathbb{R}^{m\times d}$. The function $f_{D'}(x)$ is upper bounded as

$$\begin{aligned} F(D') = f_{D'}(x) &:= \inf_{a\in\mathbb{R}^d}\{e_h(||x-D'a||_2) + g(a)\} \\ &= \inf_{a\in\mathbb{R}^d}\{e_h(||x-D'a||_2) + g(a) - e_h(||x-Da||_2) + e_h(||x-Da||_2)\} \\ &\leq \inf_{a\in\mathbb{R}^d}\{e_h(||x-Da||_2) + g(a) + \frac{1}{2}||D'a-Da||_2^2\} \quad \text{(from Lemma 5)} \\ &\leq \sup_{a\in\mathfrak{U}_0^D(x)}\{e_h(||x-Da||_2) + g(a) + \frac{1}{2}||D'a-Da||_2^2\} \quad (\text{since } \mathfrak{U}_0^D(x) \text{ is non-empty}) \\ &\leq \sup_{a\in\mathfrak{U}_0^D(x)}\{e_h(||x-Da||_2) + g(a)\} + \sup_{a\in\mathfrak{U}_0^D(x)} \frac{1}{2}||D'a-Da||_2^2 \\ &= f_D(x) + \sup_{a\in\mathfrak{U}_0^D(x)} \frac{1}{2}||D'a-Da||_2^2 \quad (\text{by definition of } \mathfrak{U}_0^D(x)) \\ &\leq f_D(x) + \frac{1}{2}||D'-D||_{1,2}^2 \sup_{a\in\mathfrak{U}_0^D(x)} ||a||_1^2 \quad (\text{by definition of the } ||\cdot||_{1,2}\text{-norm}) \\ &= F(D) + \frac{1}{2}||D'-D||_{1,2}^2 \sup_{a\in\mathfrak{U}_0^D(x)} ||a||_1^2, \end{aligned} \tag{56}$$

or else

$$\frac{F(D') - F(D)}{||D'-D||_{1,2}} \leq \frac{1}{2}||D'-D||_{1,2} \sup_{a\in\mathfrak{U}_0^D(x)} ||a||_1^2. \tag{57}$$

Interchanging the roles of D and D' in (56),

$$\frac{F(D) - F(D')}{||D'-D||_{1,2}} \leq \frac{1}{2}||D'-D||_{1,2} \sup_{a\in\mathfrak{U}_0^{D'}(x)} ||a||_1^2, \tag{58}$$

and thus the inequalities

$$-\frac{1}{2}||D'-D||_{1,2} \sup_{a\in\mathfrak{U}_0^D(x)} ||a||_1^2 \leq \frac{F(D) - F(D')}{||D'-D||_{1,2}} \leq \frac{1}{2}||D'-D||_{1,2} \sup_{a\in\mathfrak{U}_0^{D'}(x)} ||a||_1^2, \tag{59}$$

hold true.

From (59),

$$-\frac{1}{2}||D'-D||_{1,2}\max\{C_{D,x},C_{D',x}\}^2 \leq \frac{F(D)-F(D')}{||D'-D||_{1,2}} \leq \frac{1}{2}||D'-D||_{1,2}\max\{C_{D,x},C_{D',x}\}^2 \tag{60}$$

and the result follows. □

Without loss of generality assume that $C_{D',x} \leq C_{D,x}$.

Proposition 6. *Fix $x \in \mathbb{B}_{\mathbb{R}^m}(T)$ and let $C_{D,x} > 0$ be a finite upper bound for $\sup_{a\in\mathfrak{U}_0^D(x)} ||a||_1$. Then, for any $D, D' \in \mathbb{R}^{m\times d}$, we have*

$$|F(D)-F(D')| \leq \frac{1}{2}C_{D,x}||D-D'||_{1,2}. \tag{61}$$

Proof. The following proof is an adaption of the proof of Theorem 2 in (Gribonval et al., 2015b). Fix $\varepsilon > 0$. From inequality (55),

$$|F(D)-F(D')| \leq \frac{(1+\varepsilon)}{2}C_{D,x}||D-D'||_{1,2},$$

whenever $\delta := ||D-D'||_{1,2} \leq \frac{1+\varepsilon}{C_{D,x}}$. If $\delta > \frac{1+\varepsilon}{C_{D,x}}$, then choose an integer $k > 0$ such that $\frac{\delta}{k} \leq \frac{1+\varepsilon}{C_{D,x}}$ and construct the sequence

$$D_i = D + \frac{i}{k}(D-D'), \quad \text{with } i = 0, \cdots, k-1. \tag{62}$$

For this sequence of D_i's,

$$||D_{i+1}-D_i||_{1,2} = \frac{||D-D'||_{1,2}}{k} \leq \frac{1+\varepsilon}{C_{D,x}}, \tag{63}$$

and

$$\begin{aligned}|F(D_{i+1})-F(D_i)| &\leq \frac{1}{2}C_{D,x}^2||D_{i+1}-D_i||_{1,2}^2 \\ &\leq \frac{1+\varepsilon}{2}C_{D,x}||D_{i+1}-D_i||_{1,2} \ \text{(from (63))}.\end{aligned}$$

Also,

$$\begin{aligned}|F(D)-F(D')| &\leq \sum_{i=0}^{k-1}|F(D_{i+1})-F(D_i)| \\ &\leq \frac{(1+\varepsilon)}{2}C_{D,x}\sum_{i=0}^{k-1}||D_{i+1}-D_i||_{1,2} \\ &= \frac{(1+\varepsilon)}{2}C_{D,x}\sum_{i=0}^{k-1}\frac{||D-D'||_{1,2}}{k} \\ &= \frac{(1+\varepsilon)}{2}C_{D,x}||D-D'||_{1,2}.\end{aligned}$$

Since ε was arbitrary we conclude the result. □

And now the final step in the proof for the Lipschitz continuity of $F : \mathfrak{D} \mapsto \mathscr{F}_{\mathfrak{D}}$; what remains is to find an upper bound for $C_{D,x}$ when $D \in \mathfrak{D}$, $x \in \mathbb{B}_{\mathbb{R}^m}(T)$, and $a \in \mathfrak{U}_0^D(x)$. Note that

$$\begin{aligned}\mathcal{L}_x(D,a) &= e_h(||x-Da||_2) + g(a) \\ &\leq f_D(x) = \inf_{a\in\mathbb{R}^d}\{e_h(||x-Da||_2)+g(a)\} \quad (\text{since } a \in \mathfrak{U}_0^D(x)) \\ &\leq e_h(||x||_2) \quad (\text{for } a = 0)\end{aligned} \tag{64}$$

and consequently

$$g(a) \leq e_h(||x||_2). \tag{65}$$

In the sequel, an upper bound for the l_1-norm of $a \in \mathfrak{U}_0^D(x)$ is inferred from (65); this bound depends on the l_2-norm of x. For example, when $g(a) = ||a||_p$, for $1 \leq p < \infty$, Hölder's inequality yields

$$\begin{aligned} ||a||_1 = \sum_{i=1}^{d} |a_i| \leq \left(\sum_{i=1}^{d} |a_i|^p\right)^{1/p} \left(\sum_{i=1}^{d} 1^{1/(1-1/p)}\right)^{1-1/p} \\ = d^{1-1/p}||a||_p. \end{aligned} \tag{66}$$

From (65) and (66) the following implications hold true

$$a \in \mathfrak{U}_0^D(x) \ \Rightarrow \ ||a||_p \leq e_h(||x||_2) \ \Rightarrow \ ||a||_1 \leq d^{1-1/p} e_h(||x||_2). \tag{67}$$

For the non-convex case of the l_p-norm, $0 < p < 1$, for any D and x (see Section III in (Gribonval et al., 2015b))

$$a \in \mathfrak{U}_0^D(x) \ \Rightarrow \ ||a||_p \leq e_h(||x||_2) \ \Rightarrow \ ||a||_1 \leq d^{\max\{0,1-1/p\}} e_h(||x||_2). \tag{68}$$

In the general case where $g(a) = \sum_{i=1}^d \hat{g}(a_i)$ and $\hat{g}$ is continuous, even, and strictly increasing on $[0, +\infty)$, such as the log-penalty function $g_{\log}(\cdot) : \mathbb{R}^d \to \mathbb{R}_+$ below,

$$g_{\log}(a; \gamma) = \sum_{i=1}^{d} \underbrace{\frac{1}{\gamma+1} \log(\gamma |a_i| + 1)}_{\hat{g}_{\log}(\cdot;\gamma):\mathbb{R}\to\mathbb{R}_+}, \ \gamma > 0,$$

then Lemma 7 gives a rough upper bound for $\sup_{a \in \mathfrak{U}_0^D(x)} ||a||_1$ that depends on x as follows.

Lemma 7. *Let $D \in \mathbb{R}^{m \times d}$ and assume that $\hat{g} : \mathbb{R}^d \to \mathbb{R}_+$ is an i) even, ii) continuous, and iii) strictly increasing function on $[0, +\infty)$. Also, let $g(a) := \sum_{i=1}^d \hat{g}(a_i)$. Then*

$$a \in \mathfrak{U}_0^D(x) \ \Rightarrow \ g(a) \leq e_h(||x||_2) \ \Rightarrow \ ||a||_1 \leq C_{D,x}, \tag{69}$$

where $C_{D,x} := d\hat{g}^{-1}(e_h(||x||_2))$ and $\hat{g}^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$ is the inverse function of $\hat{g}$ on $[0, \infty)$.

Proof. It holds true that

$$\begin{aligned} g(a) &= \sum_{i=1}^{d} \hat{g}(a_i) \leq e_h(||x||_2) \quad \text{(from inequality (65))} \\ &\Rightarrow \max_{1\leq i\leq d} \hat{g}(a_i) \leq e_h(||x||_2) \\ &\Rightarrow |a_i| \leq \hat{g}^{-1}(e_h(||x||_2)) \ \text{ (since } \hat{g} \text{ is continuous and increasing)} \\ &\Rightarrow ||a||_1 \leq d\hat{g}^{-1}(e_h(||x||_2)). \end{aligned} \tag{70}$$

□

As a corollary of $e_h(||x||_2) \leq e_h(T)$, $x \in \mathbb{B}_{\mathbb{R}^m}(T)$ it holds true that

$$a \in \mathfrak{U}_0^D(x) \ \Rightarrow \ ||a||_1 \leq d\hat{g}^{-1}(e_h(||x||_2)) \leq d\hat{g}^{-1}(e_h(T)) \tag{71}$$

for any $D \in \mathbb{R}^{m \times d}$; thus $d\hat{g}^{-1}(e_h(T))$ is an upper bound for $C_{D,x}$. Next proposition states that $F : \mathfrak{D} \mapsto \mathscr{F}_{\mathfrak{D}}$ is Lipschitz continuous. Its proof is a combination of Proposition 6, expression (71), and the monotonicity of $\hat{g}^{-1}$ on $[0, +\infty)$.

Proposition 7. *For any $x \in \mathbb{B}_{\mathbb{R}^m}(T)$ and any $D \in \mathfrak{D}$ it holds*

$$\sup_{a \in \mathfrak{U}_0^D(x)} ||a||_1 \leq d\hat{g}^{-1}(e_h(T)). \tag{72}$$

Thus, the map $F : (\mathfrak{D}, ||\cdot||_{1,2}) \mapsto (\mathscr{F}_{\mathfrak{D}}, ||\cdot||_\infty)$ is Lipschitz continuous, i.e.,

$$||F(D) - F(D')||_\infty \leq C_{\mathfrak{D}} ||D - D'||_{1,2}, \tag{73}$$

where

$$C_{\mathfrak{D}} := \frac{d\hat{g}^{-1}(e_h(T))}{2} \tag{74}$$

is the Lipschitz constant.

A.2.2. Proof of Theorem 1: The UCEM property for $\mathscr{F}_{\mathfrak{D}}$

The deeper meaning of Proposition 7 is that it allows us to invoke Lemma 3 and upper bound $\mathcal{N}_\infty(\varepsilon, \mathscr{F}_{\mathfrak{D}})$ in terms of the covering number $\mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2})$.

Lemma 8. *The following inequality between the covering numbers of the spaces $\mathscr{F}_{\mathfrak{D}}$ and $\mathfrak{D}$ is valid,*

$$\mathcal{N}_\infty(\varepsilon, \mathscr{F}_{\mathfrak{D}}) \leq \mathcal{N}\left(\frac{\varepsilon}{C_{\mathfrak{D}}}, \mathfrak{D}, ||\cdot||_{1,2}\right). \tag{75}$$

Proof. The proof is a direct application of Lemma 3, Proposition 6, and Proposition 7. □

A well known result for the covering number of the dictionary space $\mathfrak{D}$ is the following.

Lemma 9 (Lemma 15 in (Gribonval et al., 2015b))**.** *The covering number of the space $\mathfrak{D} \subset \mathbb{R}^{m \times d}$ is upper bounded as*

$$\mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2}) \leq \left(\frac{3}{\varepsilon}\right)^{md}. \tag{76}$$

The next result is a combination of Lemma 8 and Lemma 9. It states that the size of every minimal ε-cover of $\mathscr{F}_{\mathfrak{D}}$, say $\mathscr{F}_{\mathfrak{D},\varepsilon}$, is upper bounded by $\left(\frac{3C_{\mathfrak{D}}}{\varepsilon}\right)^{md}$.

Corollary 1. *The covering number of the function class $\mathscr{F}_{\mathfrak{D}}$ is upper bounded as*

$$\mathcal{N}_\infty(\varepsilon, \mathscr{F}_{\mathfrak{D}}) \leq \left(\frac{3C_{\mathfrak{D}}}{\varepsilon}\right)^{md}. \tag{77}$$

We are ready to state the proof of Theorem 1 about the UCEM property for $\mathscr{F}_{\mathfrak{D}}$.

Proof. Let $\mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}$ be an $\frac{\varepsilon}{3}$ proper cover of $\mathscr{F}_{\mathfrak{D}}$ w.r.t $||\cdot||_\infty := \sup_{x \in \mathbb{B}_{\mathbb{R}^m}(T)} |f_D(x)|$ of minimal cardinality. Fix $f_D \in \mathscr{F}_{\mathfrak{D}}$. Then, there exists $f \in \mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}$ such that, $||f_D - f||_\infty < \frac{\varepsilon}{3}$, thus

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n f_D(x_i) - \int f_D d\mu\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n f_D(x_i) - \frac{1}{n}\sum_{i=1}^n f(x_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f d\mu\right| \\ &\quad + \left|\int f d\mu - \int f_D d\mu\right| \\ &\leq ||f - f_D||_\infty + \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f d\mu\right| + ||f - f_D||_\infty \\ &< \frac{2}{3}\varepsilon + \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f d\mu\right|.\end{aligned} \tag{78}$$

Thus

$$\mathbb{P}\Big\{\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}\Big|\frac{1}{n}\sum_{i=1}^n f_D(X_i)-\int f_D d\mu\Big|>\varepsilon\Big\}$$

$$\leq \mathbb{P}\Big\{\sup_{f\in\mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-\int f d\mu\Big|>\frac{\varepsilon}{3}\Big\} \tag{79}$$

$$\leq \bigcup_{f\in\mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}}\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-\int f d\mu\Big|>\frac{\varepsilon}{3}\Big\}.$$

From Hoeffding's inequality and $\sup_{x\in\mathbb{B}_{\mathbb{R}^m}(T)}|f(x)|\leq e_h(T)$, for all $f\in\mathscr{F}_{\mathfrak{D},\varepsilon}$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\int f d\mu\right|>\frac{\varepsilon}{3}\right\}\leq 2e^{-\frac{2n\varepsilon^2}{9e_h(T)^2}}. \tag{80}$$

Applying the union bound in (79) together with the fact that $\mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}$ has finite size $\left(\frac{9C_{\mathfrak{D}}}{\varepsilon}\right)^{md}$,

$$\bigcup_{f\in\mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}}\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-\int f d\mu\Big|>\frac{\varepsilon}{3}\Big\}$$

$$\leq \sum_{f\in\mathscr{F}_{\mathfrak{D},\frac{\varepsilon}{3}}}\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-\int f d\mu\Big|>\frac{\varepsilon}{3}\Big\} \tag{81}$$

$$\overset{(80)}{\leq} 2\left(\frac{9C_{\mathfrak{D}}}{\varepsilon}\right)^{md} e^{-\frac{2n\varepsilon^2}{9e_h(T)^2}}.$$

The proof of (22) is finished. To prove

$$\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}\left|\frac{1}{n}\sum_{i=1}^n f_D(X_i)-\int f_D d\mu\right|\to 0 \quad \text{almost surely (as } n\to\infty), \tag{82}$$

note that the inequality

$$\sum_{n=1}^{\infty}\left(\frac{9C_{\mathfrak{D}}}{\varepsilon}\right)^{md} e^{-\frac{2n\varepsilon^2}{9e_h(T)^2}}<\infty \tag{83}$$

clearly holds for any (fixed) $\varepsilon>0$. This implies

$$\sum_{n=1}^{\infty}\mathbb{P}\left\{\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}\left|\frac{1}{n}\sum_{i=1}^n f_D(X_i)-\int f_D d\mu\right|>\frac{1}{k}\right\}<\infty,$$

for each $k\in\mathbb{N}_+$. The Borel-Cantelli Lemma states that, if the sum of the probabilities of the events

$$E_n=\left\{\omega:\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}\left|\frac{1}{n}\sum_{i=1}^n f_D(X_i(\omega))-\int f_D d\mu\right|>\frac{1}{k}\right\}$$

is finite, i.e.,

$$\sum_{n=1}^{\infty}\mathbb{P}\Big\{E_n\Big\}=\sum_{n=1}^{\infty}\mathbb{P}\left\{\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}\left|\frac{1}{n}\sum_{i=1}^n f_D(X_i(\omega))-\int f_D d\mu\right|>\frac{1}{k}\right\}<\infty,$$

then the probability that infinitely many of them occur is 0, i.e., $\mathbb{P}\{\limsup_{n\to\infty}E_n\Big\}=0$, or else,

$$\limsup_{n\to\infty}\sup_{f_D\in\mathscr{F}_{\mathfrak{D}}}\left|\frac{1}{n}\sum_{i=1}^n f_D(X_i)-\int f_D d\mu\right|\leq\frac{1}{k} \quad \text{almost surely,} \tag{84}$$

for each $k \in \mathbb{N}_+$. Hence with probability one

$$\limsup_{n\to\infty} \sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left| \frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu \right| \leq \frac{1}{k} \quad \text{for all } k \in \mathbb{N}_+,$$

which implies (82). Since the above hold for any $\mu \in \bar{\mathcal{P}}$, the class $\mathscr{F}_{\mathfrak{D}}$ has the UCEM property with respect to $\bar{\mathcal{P}}$. □

A.3. Proof of Proposition 1

Proof. Let $\mathscr{F}_{\mathfrak{D},\varepsilon} = \{f_1, \dots, f_N\}$ be an ε-covering of $\mathscr{F}_{\mathfrak{D}}$ with minimal cardinality $N = \mathcal{N}_\infty(\varepsilon, \mathscr{F}_{\mathfrak{D}})$. We have

$$\begin{aligned}
\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left| \frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu \right| &\leq \sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left| \frac{1}{n}\sum_{i=1}^n f_D(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i) + \frac{1}{n}\sum_{i=1}^n f(X_i) \right. \\
&\qquad \left. - \int f d\mu + \int f d\mu - \int f_D d\mu \right| \\
&\leq ||f_D - f||_\infty + \sup_{f \in \mathscr{F}_{\mathfrak{D},\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f d\mu \right| \\
&\qquad + ||f_D - f||_\infty \\
&= 2||f_D - f||_\infty + \sup_{f \in \mathscr{F}_{\mathfrak{D},\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f d\mu \right| \\
&< 2\varepsilon + \sup_{f \in \mathscr{F}_{\mathfrak{D},\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f d\mu \right|.
\end{aligned} \tag{85}$$

Using Hoeffding's inequality and the union bound

$$\mathbb{P}\left\{ \sup_{f \in \mathscr{F}_{\mathfrak{D},\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f d\mu \right| \geq \xi \right\} \leq 2N \exp\left(-\frac{2n\xi^2}{e_h(T)^2} \right). \tag{86}$$

Set

$$\delta := 2N \exp\left(-\frac{2n\xi^2}{e_h(T)^2} \right) = 2\left(\frac{3C_{\mathfrak{D}}}{\varepsilon} \right) \exp\left(-\frac{2n\xi^2}{e_h(T)^2} \right), \tag{87}$$

and note, that after some calculations,

$$\xi = e_h(T)\sqrt{\frac{\log(\frac{2}{\delta}) + md\log(\frac{3C_{\mathfrak{D}}}{\varepsilon})}{2n}}. \tag{88}$$

Thus, with probability at least $1-\delta$,

$$\sup_{f \in \mathscr{F}_{\mathfrak{D},\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f d\mu \right| < e_h(T)\sqrt{\frac{\log(\frac{2}{\delta}) + md\log(\frac{3C_{\mathfrak{D}}}{\varepsilon})}{2n}}. \tag{89}$$

From (85) and (89)

$$\begin{aligned}
\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left| \frac{1}{n}\sum_{i=1}^n f_D(X_i) - \int f_D d\mu \right| &\leq 2\varepsilon + \sup_{f \in \mathscr{F}_{\mathfrak{D},\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f d\mu \right| \\
&\leq \frac{2}{n} + e_h(T)\sqrt{\frac{\log(\frac{2}{\delta}) + md\log(\frac{3C_{\mathfrak{D}}}{\varepsilon})}{2n}} \quad (\text{for } \varepsilon = 1/n) \\
&\leq \frac{2}{n} + e_h(T)\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} + \sqrt{\frac{md\log(3nC_{\mathfrak{D}})}{2n}},
\end{aligned} \tag{90}$$

with probability at least $1-\delta$. The proof is now finished. □

A.4. Proof of Lemma 2

Proof. We follow the same arguments as in the proof of Lemma 10 in (Gribonval et al., 2015b). Without loss of generality assume that the support of the probability measure μ is within the unit ball, i.e., for any $X \sim \mu$ it holds $||X|| \leq 1$. First, we show that for $c = \frac{1}{\sqrt{8}}$,

$$\Gamma_n(c\tau) := \sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu \right| \geq c\tau \right\} \leq 2e^{-n\tau^2}. \tag{91}$$

For fixed $D \in \mathfrak{D}$ the random variables $f_D(X_i)$, $i \in \{1, \ldots, n\}$ are independent. When samples are drawn according to μ,

$$f_D(X_i) \leq e_h(||X_i||_2) \leq \frac{1}{2}||X_i|| \leq \frac{1}{2}.$$

Using Hoeffding's inequality

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu \right| \geq c\tau \right\} \leq 2e^{-8c^2 n\tau^2},$$

which implies that

$$\Gamma_n(c\tau) \leq 2e^{-n\tau^2}, \tag{92}$$

for $c = \frac{1}{\sqrt{8}}$. Now assume that (91) is true. Let $\mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2})$ be an ε-cover of $\mathfrak{D}$ and $L > \frac{d\hat{g}^{-1}(1)}{2}$; recall that $\frac{d\hat{g}^{-1}(e_h(T))}{2}$ is the Lipschitz constant of the map $F : (\mathfrak{D}, ||\cdot||_{1,2}) \mapsto (\mathscr{F}_{\mathfrak{D}}, ||\cdot||_\infty)$. For a fixed dictionary $D \in \mathfrak{D}$ there exists an index $j(D)$ such that $||D - D_j||_{1,2} \leq \varepsilon$. Then

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu \right| &\leq \left| \frac{1}{n} \sum_{i=1}^{n} f_D(X_i) - \frac{1}{n} \sum_{i=1}^{n} f_{D_j}(X_i) + \frac{1}{n} \sum_{i=1}^{n} f_{D_j}(X_i) \right. \\
&\qquad \left. - \int f_{D_j} d\mu + \int f_{D_j} d\mu - \int f_D d\mu \right| \\
&\leq ||f_D - f_{D_j}||_\infty + \sup_{j \in \{1, \ldots, \mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2})\}} \left| \frac{1}{n} \sum_{i=1}^{n} f_{D_j}(X_i) - \int f_{D_j} d\mu \right| \\
&\qquad + ||f_D - f_{D_j}||_\infty \\
&= 2||f_D - f_{D_j}||_\infty + \sup_{j \in \{1, \ldots, \mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2})\}} \left| \frac{1}{n} \sum_{i=1}^{n} f_{D_j}(X_i) - \int f_{D_j} d\mu \right| \\
&\leq 2L\varepsilon + \sup_{j \in \{1, \ldots, \mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2})\}} \left| \frac{1}{n} \sum_{i=1}^{n} f_{D_j}(X_i) - \int f_{D_j} d\mu \right| \\
&\leq 2L\varepsilon + c\tau
\end{aligned} \tag{93}$$

which holds with probability at least $1 - \mathcal{N}(\varepsilon, \mathfrak{D}, ||\cdot||_{1,2}) \cdot \Gamma_n(c\tau)$. Since this is true for any $\varepsilon, \tau > 0$, set

$$\varepsilon = \frac{c\sqrt{\beta}}{2L} \sqrt{\frac{\log(n)}{n}} \quad \text{and } \tau = \sqrt{\frac{md \log\left(\frac{3}{\varepsilon}\right) + t}{n}} = \sqrt{md \log\left(\frac{6L}{c\sqrt{\beta}}\right) + \frac{md}{2} \log\left(\frac{n}{\log(n)}\right) + t} \cdot \frac{1}{\sqrt{n}}. \tag{94}$$

The assumption $\frac{n}{\log n} \geq \max\left\{8, \left(\frac{c}{2L}\right)^2 \beta\right\}$, $c = \frac{1}{\sqrt{8}}$, implies that

$$\begin{aligned}
\frac{c\sqrt{\beta}}{2L} \sqrt{\frac{\log(n)}{n}} &\leq \frac{c\sqrt{\beta}}{2L} \frac{1}{\sqrt{\max\left\{8, \frac{c^2}{4L^2}\beta\right\}}} \\
&= \sqrt{\frac{\frac{c^2\beta}{4L^2}}{\max\left\{8, \frac{c^2\beta}{4L^2}\right\}}} \\
&\leq 1.
\end{aligned} \tag{95}$$

This shows that $0 < \varepsilon \leq 1$. Since $\beta \geq 1$ and $\log(n) \geq 1$, we have

$$\begin{aligned} 2L\varepsilon + c\tau &= 2Lc\sqrt{\beta}\frac{1}{2L}\sqrt{\frac{\log(n)}{n}} + c\sqrt{md\log\left(\frac{6L}{c\sqrt{\beta}}\right) + \frac{md}{2}\log\left(\frac{n}{\log(n)}\right) + t}\cdot\frac{1}{\sqrt{n}} \\ &\leq c\sqrt{\frac{\beta\log(n)}{n}} + c\sqrt{md\log\left(\frac{6L}{c\sqrt{\beta}}\right) + \frac{md}{2}\log(n) + t}\cdot\frac{1}{\sqrt{n}} \quad (\text{since } \log(n) \geq 1) \\ &\leq c\sqrt{\frac{\beta\log(n)}{n}} + c\sqrt{md\log\left(\frac{6L}{c}\right) + \frac{md}{2}\log(n) + t}\cdot\frac{1}{\sqrt{n}} \quad (\text{since } \beta \geq 1) \\ &\leq c\sqrt{\frac{\beta\log(n)}{n}} + c\sqrt{\beta + \frac{\beta}{2}\log(n) + t}\cdot\frac{1}{\sqrt{n}} \quad \left(\text{since } \beta := md\max\left\{\log\left(\frac{6L}{c}\right), 1\right\}\right) \\ &\leq c\sqrt{\frac{\beta\log(n)}{n}} + c\sqrt{\frac{\beta\log(n)}{2n}} + c\sqrt{\frac{\beta + t}{n}} \\ &\leq 2c\sqrt{\frac{\beta\log(n)}{n}} + c\sqrt{\frac{\beta + t}{n}}. \end{aligned} \tag{96}$$

Hence for $c = \sqrt{\frac{1}{8}}$,

$$\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left|\frac{1}{n}\sum_{i=1}^{n} f_D(X_i) - \int f_D d\mu\right| \leq \frac{2}{\sqrt{8}}\sqrt{\frac{\beta\log(n)}{n}} + \frac{1}{\sqrt{8}}\sqrt{\frac{\beta + t}{n}} \tag{97}$$

with probability at least $1 - 2e^{-t}$. The proof is now finished. □

A.5. Discussion: The case of a separable, continuous, even, and bounded g

As already mentioned, Theorem 1 covers a wide range of separable functions $g : \mathbb{R}^d \to \mathbb{R}_+$ but it does not cover popular (separable) penalty functions, like the SCAD or MCP,

$$\hat{g}_{scad}(t; \lambda, \gamma) = \begin{cases} \lambda t, & t \leq \lambda \\ \frac{\lambda\gamma t - \frac{1}{2}(t^2 + \lambda^2)}{\gamma - 1}, & \lambda < t \leq \gamma\lambda \\ \frac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)}, & t > \lambda\gamma. \end{cases} \quad \text{and} \quad \hat{g}_{mcp}(t; \lambda, \gamma) = \begin{cases} \lambda t - \frac{t^2}{2\gamma}, & t \leq \lambda \\ \frac{1}{2}\gamma\lambda^2, & t > \gamma\lambda. \end{cases} \tag{98}$$

The functions $\hat{g}_{scad}$ and $\hat{g}_{mcp}$ are bounded from above and so they fail to satisfy the "strictly increasing" assumption of Section 3. A closer look at the proof of Lemma 7 reveals that if

$$e_h(T) < \sup_{t \in \mathbb{R}} \hat{g}_{mcp}(t; \lambda, \gamma), \tag{99}$$

then the following set of implications are true:

$$\begin{aligned} a \in \mathfrak{U}_0^D(x) \quad &\Rightarrow \quad g_{mcp}(a; \lambda, \gamma) \leq e_h(||x||_2) \ \ (\text{from inequalities (64) and (65)}) \\ &\Rightarrow \sum_{i=1}^{d} \hat{g}_{mcp}(a_i; \lambda, \gamma) \leq e_h(||x||_2) \\ &\Rightarrow \max_{1 \leq i \leq d} \hat{g}_{mcp}(a_i; \lambda, \gamma) \leq e_h(||x||_2) \\ &\Rightarrow |a_i| \leq \hat{g}_{mcp}^{-1}(e_h(||x||_2); \lambda, \gamma) \ \ (\text{since } \hat{g}_{mcp} \text{ is invertible in } [0, e_h(T)] \text{ due to (99)}) \\ &\Rightarrow ||a||_1 \leq d\hat{g}_{mcp}^{-1}(e_h(||x||_2); \lambda, \gamma). \end{aligned} \tag{100}$$

The above reasoning also applies to g_{scad}. Summarizing, the following lemma is proved.

Lemma 10. *Let a separable function $g : \mathbb{R}^d \to \mathbb{R}_+$ of the form $g(a) = \sum_{i=1}^{d} \hat{g}(a_i)$, where $\hat{g} : \mathbb{R} \to \mathbb{R}_+$ is continuous, even, strictly increasing up to some point on $[0, +\infty)$ and then constant. If*

$$e_h(T) < \sup_{t \in \mathbb{R}} \hat{g}(t), \tag{101}$$

then the following set of implications hold true

$$a \in \mathfrak{U}_0^D(x) \;\Rightarrow\; g(a) \le e_h(||x||_2) \;\Rightarrow\; ||a||_1 \le C_{D,x}, \tag{102}$$

where $C_{D,x} := d\hat{g}^{-1}(e_h(||x||_2))$ *and* $\hat{g}^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$ *denotes the inverse function of* $\hat{g}$ *restricted on the domain where* $\hat{g}$ *is strictly increasing.* □

The next Proposition is an analog of Proposition 7 and an immediate consequence of expression (102).

Proposition 8. *For any* $x \in \mathbb{B}_{\mathbb{R}^m}(T)$*,* $D \in \mathfrak{D}$*, and* $g : \mathbb{R}^d \to \mathbb{R}_+$ *satisfying assumptions of Lemma 10,*

$$\sup_{a \in \mathfrak{U}_0^{\mathfrak{D}}(x)} ||a||_1 \le d\hat{g}^{-1}(e_h(T)). \tag{103}$$

Thus the map $F : (\mathfrak{D}, ||\cdot||_{1,2}) \mapsto (\mathscr{F}_{\mathfrak{D}}, ||\cdot||_\infty)$ *is Lipschitz continuous,*

$$||F(D) - F(D')||_\infty \le C_{\mathfrak{D}} ||D - D'||_{1,2}, \tag{104}$$

with Lipschitz constant $C_{\mathfrak{D}} := \frac{d\hat{g}^{-1}(e_h(T))}{2}$. □

The rest of results in Section 3 still remain valid for any function g under consideration; even Lemma 4 as mentioned in Remark 3. So, the family of functions $\mathscr{F}_{\mathfrak{D}}$ retains the UCEM property for $\bar{\mathcal{P}}$ for any bounded separable penalty function g satisfying assumptions of Section 4.

A.6. Proof of Proposition 2

Proof. The proof follows the lines of the proof of Theorem 20 in (Vainsencher et al., 2011); we only depart in the details. First note that the set $\Sigma_k := \{a \in \mathbb{R}^d : |\{i : a_i \neq 0\}| = k\}$ of all k-sparse vectors in $\mathbb{R}^d$ is the union of $\binom{d}{k}$ sets Σ_k^l,

$$\Sigma_k^l := \left\{a \in \mathbb{R}^d : a_i \neq 0, \forall i \in I_l \text{ and } a_i = 0, \forall i \notin I_l\right\},\; l = 1, \ldots, \binom{d}{k}, \tag{105}$$

where I_l is one of the $\binom{d}{k}$ possible k-tuples in $\{1, \ldots, d\}$; in other words,

$$\Sigma_k = \bigcup_{l=1}^{\binom{d}{k}} \Sigma_k^l. \tag{106}$$

The proof is constructive; first it is shown that for any $D \in \mathfrak{D}$ there exist $\gamma > 0$ and $q \in \mathbb{S}^{m-1}$ such that $f_D(q) > \gamma$. Let μ be the uniform measure on the unit sphere $\mathbb{S}^{m-1} := \{x \in \mathbb{R}^m : ||x||_2 = 1\}$. Denote as A_c the probability assigned by μ to the set "within $e_h(c)$" of a k-dimensional subspace, $c > 0$. For example, when $m = 3$ and $k = 1$, the probability A_c can be defined as

$$A_c = \mu\Big\{\{x \in \mathbb{S}^2 : \exists t \in \mathbb{R} \text{ and } z = te_1 \text{ such that } e_h(||x - z||_2) \le e_h(c)\}\Big\},$$

where $e_1 = (1, 0, 0)^\top$. As $c \searrow 0$, A_c also tends to zero. Then there exist $c > 0$ such that $\binom{d}{k}A_c < 1$; for that c and any $D \in \mathfrak{D}$ there exists a set of positive measure, say $\tilde{A}_c$, on which $f_D(x) > e_h(c) = \gamma$. Indeed, for every $l \in \{1, \ldots, \binom{d}{k}\}$, the following inclusion is valid

$$\left\{x \in \mathbb{S}^{m-1} : \min_{a \in \Sigma_k^l} e_h(||x - Da||_2) \le e_h(c)\right\} \subseteq A_c$$

by definition of A_c. Along with assumption $\binom{d}{k}A_c < 1$, the assertion that for any $D \in \mathfrak{D}$ there exists a q such that $f_D(q) > 0$ holds true.

Let q be a sample point in $\tilde{A}_c$ and assume without loss of generality that $\sum_{j=1}^{k-1} q_j > 0$. Next construct two dictionaries D, D' such that $f_D(q) > 0$ while $f_{D'}(q) = 0$. First construct dictionary D; its first $k-1$ columns are the standard basis vectors in $\mathbb{R}^m$, $\{e_1, \ldots, e_{k-1}\}$, its k-th column is

$$D_{\cdot,k} = \frac{1}{\sqrt{k-1}} \sum_{j=1}^{k-1} \sqrt{1 - \varepsilon^2/4}e_j + \varepsilon e_k/2,$$

and the remaining columns are arbitrary unit-norm vectors. Now D' is constructed; it is identical to D with the only difference being in the k-th column. Specifically,

$$D'_{\cdot,k} = \frac{1}{\sqrt{k-1}} \sum_{j=1}^{k-1} \sqrt{1-\varepsilon^2/4} e_j + lq,$$

for some $l \in \mathbb{R}$ such that $||D'_{\cdot,k}||_2 = 1$, or else,

$$\begin{aligned} ||D'_{\cdot,k}||_2 = 1 \Leftrightarrow ||D'_{\cdot,k}||_2^2 = 1 \\ &\Leftrightarrow ||lq||_2^2 + 2l\sqrt{\frac{1-\varepsilon^2/4}{k-1}} \sum_{j=1}^{k-1} q_j - \varepsilon^2/4 = 0 \\ &\Leftrightarrow l^2 + 2l\sqrt{\frac{1-\varepsilon^2/4}{k-1}} \sum_{j=1}^{k-1} q_j - \varepsilon^2/4 = 0 \quad \text{(since } ||q||_2^2 = 1). \end{aligned}$$

The roots of the previous quadratic equation (with respect to l) are

$$l = \begin{cases} -\frac{2\sqrt{\frac{1-\varepsilon^2/4}{k-1}} \sum_{j=1}^{k-1} q_j}{2} - \sqrt{\left(\frac{2\sqrt{\frac{1-\varepsilon^2/4}{k-1}} \sum_{j=1}^{k-1} q_j}{2}\right)^2 + \frac{\varepsilon^2}{4}} \\ \sqrt{\left(\frac{2\sqrt{\frac{1-\varepsilon^2/4}{k-1}} \sum_{j=1}^{k-1} q_j}{2}\right)^2 + \frac{\varepsilon^2}{4}} - \frac{2\sqrt{\frac{1-\varepsilon^2/4}{k-1}} \sum_{j=1}^{k-1} q_j}{2}, \end{cases}$$

which after some simple algebraic manipulations implies that $l \le \varepsilon/2$ (to see this recall that $\sum_{j=1}^{k-1} q_j > 0$ and use the inequality $b^r - a^r \le (b-a)^r$ for any $0 < r < 1$ and $0 < a \le b$). For this q there exist $t_j \in \mathbb{R}$, $j \in \{1, \ldots, k\}$, such that $q = \sum_{j=1}^{k} t_j D'_{\cdot,j}$ and thus $f_{D'}(q) = 0$, proving the second part of the theorem. On the other hand

$$||D - D'||_2 = ||\varepsilon e_k/2 - lq||_2 \le ||\varepsilon e_k/2||_2 + ||lq||_2 = \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

and the proof is now completed. □

A.7. Proof of Proposition 3

Proof. Fix $D \in \mathfrak{D}$ and define the (possibly multivalued) map $\hat{a}_D : \mathbb{R}^m \rightrightarrows \mathbb{R}^d$ as

$$\hat{a}_D(x) := \arg\min_{a \in \Sigma_k} e_h(||x - Da||_2); \tag{107}$$

this map maps any vector $x \in \mathbb{R}^m$ to an optimal solution $\hat{a}_D(x) \in \mathbb{R}^d$. For that D the corresponding subgraph in the family of subgraphs

$$\mathscr{F}_{\mathfrak{D}}^+ := \left\{ \left\{ (x,t) \in \mathbb{R}^{m+1} : \ f_D(x) \ge t \right\}; f_D \in \mathscr{F}_{\mathfrak{D}} \right\} \tag{108}$$

is described by the set of points (x, t), $t \ge 0$, for which

$$e_h(||x - D\hat{a}_D(x)||_2) \ge t. \tag{109}$$

Due to monotonicity of e_h, point (x, t) with $t \ge 0$ satisfies (109) if and only if,

$$||x - D\hat{a}_D(x)||_2 \ge c(t), \tag{110}$$

where $c(t)$ is the smallest value of c for which $e_h(c) \ge t$. In view of (110), if the set $\{(x_i, t_i)\}_{i=1}^n$ is shattered by $\mathscr{F}_{\mathfrak{D}}^+$, then there exist matrix D_0 in $\mathfrak{D}$ and vectors $\{\hat{a}_{D_0}(x_i)\}_{i=1}^n$ in $\mathbb{R}^d$ such that, for those shattered points, it holds true

$$||x_i - D_0\hat{a}_{D_0}(x_i)||_2^2 \ge c(t_i)^2. \tag{111}$$

We claim that the shatter coefficient of $\mathscr{F}_{\mathfrak{D}}^+$ is upper bounded by the shatter coefficient of the collection of all subgraphs generated by functions which belong in

$$\mathcal{G} := \left\{ g_A(y,s) = ||Ay||_2^2 + \beta s \ ; \ A \in \mathbb{R}^{m \times (m+d)}, \ \beta \in \mathbb{R} \right\} \tag{112}$$

with $(y, s) \in \mathbb{R}^{m+d} \times [0, +\infty)$. Indeed, if some points in $\{(x_i, t_i)\}_{i=1}^n$ are shattered by $\mathscr{F}_{\mathfrak{D}}^+$, then for those shattered points it holds true that

$$||x_i - D_0 \hat{a}_{D_0}(x_i)||_2^2 \geq c(t_i)^2, \tag{113}$$

or equivalently,

$$\left\| \begin{bmatrix} I & -D_0 \end{bmatrix} \begin{bmatrix} x_i \\ \hat{a}_{D_0}(x_i) \end{bmatrix} \right\|_2^2 \geq c(t_i)^2. \tag{114}$$

Equivalence between the last two inequalities implies the existence of matrix $A_0 = [I \ -D_0] \in \mathbb{R}^{m \times (m+d)}$, scalar $\beta_0 = 1$, and vectors $\{(y_i, s_i)\}_{i=1}^n \subset \mathbb{R}^{m+d} \times [0, +\infty)$ with

$$(y_i, \ s_i) := \left(\left[\begin{smallmatrix} x_i \\ \hat{a}_{D_0}(x_i) \end{smallmatrix} \right], \ c(t_i) \right) \in \mathbb{R}^{m+d} \times [0, +\infty), \quad i = 1, \dots, n, \tag{115}$$

such that the graph of $||A_0 y||_2^2 - \beta_0 s$,

$$\left\{ (y, s) : ||A_0 y||_2^2 - \beta_0 s \geq 0 \right\}, \tag{116}$$

picks outs only those (y_i, s_i) satisfying inequality (114); note that (116) is the subgraph of some function, sat g_{A_0}, in $\mathcal{G}$. Every set in (116) is the sum of an ellipsoid and a linear function of s. By Lemma 18 in (Pollard, 1984), the sets $\{g_A \geq t\}$, for $g_A \in \mathcal{G}$, pick out only a polynomial number of subsets from $\{(y_i, s_i)\}_{i=1}^n$; those corresponding to functions in $\mathcal{G}$ with $A = [I \quad D]$ pick out even fewer points from $\{(y_i, s_i)\}_{i=1}^n$. The VC dimension of $\mathcal{G}$ is at most $((m+d)^2 + 3(m+d))/2+1)$, see (Akama & Irie, 2011) for an improved bound on the VC dimension of ellipsoids. Consequently, monotonicity of e_h and Theorems 13.5, 13.9 in (Devroye et al., 1997) conclude the result

$$s(\mathscr{F}_{\mathfrak{D}}^+, n) \leq \left(\underbrace{\frac{en}{((m+d)^2 + 3(m+d))/2 + 1)}}_{:= \alpha(m,d)} \right)^{((m+d)^2 + 3(m+d))/2+1)}. \tag{117}$$

□

A.8. Proof of Theorem 2

Proof. Proposition 3 and Corollary 29.1 in (Devroye et al., 1997) imply

$$\mathbb{P}\left\{ \sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left| \frac{1}{n} \sum_{i=1}^n f_D(X_i) - \int f_D d\mu \right| > \varepsilon \right\} \leq 8 s(\mathscr{F}_{\mathfrak{D}}^+, n) e^{-\frac{n\varepsilon^2}{32 e_h(T)^2}}. \tag{118}$$

Since

$$\sum_{n=1}^{\infty} s(\mathscr{F}_{\mathfrak{D}}^+, n) e^{-\frac{2n\varepsilon^2}{32 e_h(T)^2}} < \infty, \tag{119}$$

for all $\epsilon > 0$, by the Borel-Cantelli lemma (and arguments similar to the relevant part in the proof of Theorem 1)

$$\sup_{f_D \in \mathscr{F}_{\mathfrak{D}}} \left| \frac{1}{n} \sum_{i=1}^n f_D(X_i) - \int f_D d\mu \right| \to 0 \quad \text{almost surely (as } n \to \infty), \tag{120}$$

for all probability measures $\mu \in \bar{\mathcal{P}}$. Thus $\mathscr{F}_{\mathfrak{D}}$ has the UCEM property for all $\mu \in \bar{\mathcal{P}}$. □

A.9. Proof of Proposition 4

Proof. Set the right hand side of inequality (35) equal to δ and solve with respect to ε. The result immediately follows. □

A.10. Proof of Theorem 3

Proof. First we bound the shatter coefficient of family

$$\mathscr{F}_{\mathfrak{D}} := \left\{ f_D(x) = \inf_{a \in \mathbb{R}^d} \{ e_h(||x - Da||_2) + g(a);\ D \in \mathfrak{D} \} \right\} \tag{121}$$

when $g : \mathbb{R}^d \to [0, +\infty)$ is a bounded lsc function, i.e., $g(a) \leq M$ for some $M > 0$ and all $a \in \mathbb{R}^d$. To this end, we follow the proof of Proposition 3 and only depart in details.

Fix $D \in \mathfrak{D}$ and define

$$\hat{a}_D(x) := \arg\min_{a \in \mathbb{R}^d} e_h(||x - Da||_2) + g(a); \tag{122}$$

this (possibly multivalued) map maps any vector $x \in \mathbb{R}^m$ to an optimal solution $\hat{a}_D(x) \in \mathbb{R}^d$ for the minimization problem

$$\inf_{a \in \mathbb{R}^d} e_h(||x - Da||_2) + g(a). \tag{123}$$

For fixed $D \in \mathfrak{D}$ and function $f_D \in \mathscr{F}_{\mathfrak{D}}$, the corresponding subgraph in the collection of sets

$$\mathscr{F}^+_{\mathfrak{D}} := \left\{ \left\{ (x, t) \in \mathbb{R}^{m+1} :\ f_D(x) \geq t \right\} ; f_D \in \mathscr{F}_{\mathfrak{D}} \right\} \tag{124}$$

contains the points $(x, t) \in \mathbb{R}^{m+1},\ t \geq 0$, for which

$$f_D(x) \geq t \ \Leftrightarrow\ e_h(||x - D\hat{a}_D(x)||_2) + g(\hat{a}_D(x)) \geq t \ \Leftrightarrow\ e_h(||x - D\hat{a}_D(x)||_2) \geq t - g(\hat{a}_D(x)). \tag{125}$$

Due to boundedness of g and monotonicity of e_h a point (x, t) with $t > 0$ satisfies (125) if and only if

$$||x - D\hat{a}_D(x)||_2 \geq c(t - M), \tag{126}$$

where $c(t - M)$ is the smallest value of c for which $e_h(c) \geq t - M$. In view of (126), if the set $\{(x_i, t_i)\}_{i=1}^n$ is shattered by $\mathscr{F}^+_{\mathfrak{D}}$, then there exist some matrix D_0 in $\mathfrak{D}$ and set of points $\{\hat{a}_{D_0}(x_i)\}_{i=1}^n$ in $\mathbb{R}^d$ such that

$$||x_i - D_0\hat{a}_{D_0}(x_i)||_2^2 \geq c(t_i - M)^2, \tag{127}$$

for every shattered point (x_i, t_i) in $\{(x_i, t_i)\}_{i=1}^n$. The claim that the shatter coefficient of $\mathscr{F}^+_{\mathfrak{D}}$ is upper bounded by the shatter coefficient of the family of subgraphs of the function class $\mathcal{G}$ below

$$\mathcal{G} := \left\{ g_A(y, s) = ||Ay||_2^2 + \beta s\ ;\ A \in \mathbb{R}^{m \times (m+d)},\ \beta \in \mathbb{R} \right\}, \tag{128}$$

is proven in the same way as the relevant part in the proof of Proposition 3. Hence we have proven that

$$s(\mathscr{F}^+_{\mathfrak{D}}, n) \leq \left(\underbrace{\frac{en}{((m+d)^2 + 3(m+d))/2 + 1)}}_{:= \alpha(m,d)} \right)^{((m+d)^2 + 3(m+d))/2+1)} . \tag{129}$$

Proof of inequality (40) follows the proof of Theorem 2 and the proof of (41) is the same as the proof of Proposition 1. □

A.11. Discussion: A note on the approximation error when $m \gg d$

This is a discussion concerning the upper bound for the approximation error of Proposition 5 in the main text. Recall that for sufficiently large values of n, it holds true that $\mathcal{R}(\hat{D}_n) = o(1) + \varepsilon_{\text{app}}$, or else $\mathcal{R}(\hat{D}_n) \simeq \varepsilon_{\text{app}} = \mathcal{R}(D^*)$. What follows, is that we regard $\mathcal{R}(D^*)$ as a function of d ($\ll m$) and describe its rate of decrease as $d \to m$. The function class under study is

$$\mathscr{F}_{\mathfrak{D}} := \{ f_D;\ D \in \mathfrak{D} \}, \tag{130}$$

and each $f_D : \mathbb{R}^m \to [0, +\infty)$ has the form

$$f_D(x) := \inf_{a \in \mathbb{R}^d} \{e_h(||x - Da||_2) + 1_{\mathcal{K}}(a)\}. \tag{131}$$

Assumptions (H1)-(H6) are the key to the proof of Proposition 5 in the text; they implicitly restrict the shape of the Moreau envelope e_h. It can be shown that

$$\partial e_h(t) \subseteq t - P_h(t); \tag{132}$$

see Theorem 10.13 in (Rockafellar & Wets, 2009) and Proposition 7 in (Yu et al., 2015). The continuity of e_h and the differential inclusion in (132) give a description of the epigraph of e_h in $[0, +\infty)$. If the proximal map of h satisfies (H1)-(H6), then in the interval $[0, \tau]$ the Moreau envelope behaves like the quadratic function t^2, i.e., $e_h \sim t^2$ as $t \to 0_+$.[5] Indeed, under assumption (H6), for $t \in [0, \tau]$ it holds true that $\partial e_h(t) \subseteq t - P_h(t) = t$ which implies the previous assertion. Since the proximal map P_h is monotone and non-decreasing on $[0, +\infty)$ with $0 \leq P_h(t) \leq t$, it is always true that $\partial e_h(t) \leq t - P_h(t) \leq t$ and consequently $e_h(t) \leq ct^2$, for some $c > 0$ and any $t \geq 0$ (recall that $0 \leq e_h(0) \leq h(0) = 0$ by assumption). Assumptions (H1)-(H6) are valid for the proximal maps of the l_p-norm, $0 \leq p < \infty$, the SCAD, and the MCP univariate functions and many other pairs of (h, P_h), see also (Antoniadis, 2007).

In order to upper bound the approximation error $\mathcal{R}(D^*)$, we use the quantization error (or else distortion error) for e_h which is defined as

$$E_{d,e_h} = \inf_{\{c_1,\ldots,c_d\} \subset \mathbb{R}^m} \int \min_{j=1,\ldots,d} e_h(||x - c_j||_2) d\mu, \tag{133}$$

and $\{c_1, \ldots, c_d\}$ is a subset of $\mathbb{R}^m$ with d vectors. The rate of convergence of the quantization error E_{d,e_h} as d tends to $+\infty$ is ruled by the following theorem.

Theorem 4 (Delattre, Sylvain, et al (Delattre et al., 2004))**.** *Assume that $V : \mathbb{R}_+ \to [0, +\infty)$ is a non-decreasing function satisfying $V(0) = 0$ and $V(t) \sim t^r$ as $t \to 0_+$, $r > 0$. Assume also that there exists a non-decreasing function $W : \mathbb{R}_+ \to [0, +\infty)$, with $W(0) \geq 1$, such that $V(t) \leq t^r W(t)$ for every $t \in [0, +\infty)$. If the random variable X satisfies $\int ||X||_2^r d\mu < \infty$ and $\int W(||X||_2) d\mu < \infty$, then*

$$E_{d,V} = \inf_{\{c_1,\ldots,c_d\} \subset \mathbb{R}^m} \int \min_{j=1,\ldots,d} V(||x - c_j||_2) d\mu \ \leq \ \mathcal{O}(d^{-r/m}), \ \ \text{as } d \to +\infty; \tag{134}$$

here, m is the dimension of X and d is the number of vectors in $\{c_1, \ldots, c_d\}$.

The assumptions in Theorem 4 are valid for the Moreau envelope of any function h whose proximal map P_h satisfies (H1)-(H6); as already mentioned, under assumptions (H1)-(H6), we have $V(t) = e_h(t) \sim t^2$ locally around zero and $V(t) = e_h(t) \leq ct^2$ for some $c > 0$ and any $t \geq 0$. The approximation error $\mathcal{R}(D^*)$ and the quantization error in (133) are related as follows:

$$\mathcal{R}(D^*) = \inf_{D \in \mathfrak{D}} \int f_D d\mu = \inf_{D \in \mathfrak{D}} \int \inf_{a \in \mathcal{K}} e_h(||x - Da||_2) d\mu \leq \inf_{D \in \mathfrak{D}} \int \min_{c' \in \{D_{\cdot,j}\}_{j=1}^d} e_h(||x - c'||_2) d\mu, \tag{135}$$

since the basis of the positive orthant belongs to $\mathcal{K}$. Therefore,

$$\inf_{D \in \mathfrak{D}} \int \min_{c' \in \{D_{\cdot,j}\}_{j=1}^d} e_h(||x - c'||_2) d\mu \ \geq \ \inf_{\{c_1,\ldots,c_d\} \subset \mathbb{R}^m} \int \min_{j=1,\ldots,d} e_h(||x - c_j||_2) d\mu = E_{d,e_h},$$

and we can carefully choose $\{D_{\cdot,1}, \ldots, D_{\cdot,d}\}$ such that equality is achieved, i.e.,

$$\inf_{D \in \mathfrak{D}} \int \min_{c' \in \{D_{\cdot,j}\}_{j=1}^d} e_h(||x - c'||_2) d\mu = \inf_{\{c_1,\ldots,c_d\} \subset \mathbb{R}^m} \int \min_{j=1,\ldots,d} e_h(||x - c_j||_2) d\mu = E_{d,e_h} \tag{136}$$

(equality always is achieved if the data points x are rescaled to lie inside the Euclidean ball $\mathbb{B}_{\mathbb{R}^m}(1)$). In order to use Theorem 4, assume that i) m is sufficiently large and ii) d approaches m. Under these assumptions, Proposition 5 is an immediate corollary of inequality (135) and Theorem 4.

[5]The notation $a_n \sim b_n$ means $a_n = b_n + o(b_n)$.