SPARSE RECOVERY IN LINEAR SPANS AND CONVEX HULLS OF INFINITE DICTIONARIES

VLADIMIR KOLTCHINSKII

joint work with

STAS MINSKER

School of Mathematics

Georgia Institute of Technology

Marne-la-Vallée, 2010

Regression Problem

(X, Y) a random couple in $S \times T$, $T \subset \mathbb{R}$

P distribution of (X, Y)

Π distribution of X **(design distribution)**

$f_* := \mathrm{argmin}_{f:S\mapsto\mathbb{R}} \mathbb{E}(Y - f(X))^2$

$f_*(X) = \mathbb{E}(Y|X)$ **regression function**

Dictionary

$\mathcal{H}$ a class of functions $h : S \mapsto [-1, 1]$ equipped with a σ-algebra $\mathcal{B}_{\mathcal{H}}$ and with a measure μ

For $\lambda \in L_1(\mu)$,

$$f_\lambda(\cdot) := \int_{\mathcal{H}} \lambda(h) h(\cdot) \mu(dh).$$

L_1-penalization

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. copies of (X, Y)

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \Bigg[n^{-1} \sum_{j=1}^{n} (Y_j - f_\lambda(X_j))^2 + \varepsilon \|\lambda\|_{L_1(\mu)} \Bigg]$$

$\mathbb{D} \subset L_1(\mu)$ is a bounded convex set

$\varepsilon > 0$ regularization parameter

L_2-Error Bounds and Sparsity

Regression problem is called **"sparse"** with respect to $\mathcal{H}$ if there exists a **"sparse"** function $\lambda \in \mathbb{D}$ (i.e., λ is supported in a "small" subset of $\mathcal{H}$) such that the L_2 approximation error $\|f_\lambda - f_*\|^2_{L_2(\Pi)}$ is "small".

Basic Question. Suppose the regression problem is **"sparse"**. Does it imply that $\hat{\lambda}^\varepsilon$ is **"approximately sparse"** and $\|f_{\hat{\lambda}^\varepsilon} - f_*\|^2_{L_2(\Pi)}$ **is "small"?**

Finite Dictionaries

$\mathcal{H} := \{h_1, \dots, h_N\}$

$\mu(\{h_j\}) = 1,\ j = 1, \dots, N$

$\lambda = (\lambda_1, \dots, \lambda_N)$

$\|\lambda\|_{L_1(\mu)} = \|\lambda\|_{\ell_1}$

LASSO (**Tibshirani (1996), Chen, Donoho and Saunders (1996), ...**):

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - f_\lambda(X_j))^2 + \varepsilon \|\lambda\|_{\ell_1} \right],\ \mathbb{D} \subset \mathbb{R}^N.$$

Sparse Recovery: LASSO and related methods

Connections to High Dimensional Geometry: Donoho (2004–), Donoho and Tanner (2005–), Candes and Tao (2006–), ...

Methods of Asymptotic Geometric Analysis: Rudelson and Versynin (2005–), Mendelson, Pajor and Tomczak-Jaegermann (2007–), ...

Sparsity Oracle Inequalities: Bunea, Tsybakov and Wegkamp (2007–), van de Geer (2008–), Koltchinskii (2008–)

Finite Dictionaries: Geometric Characteristics

Gram matrix $K := \left(\langle h_i, h_j\rangle_{L_2(\Pi)}\right)_{i,j=1}^{N}$

For $w \in \mathbb{R}^N$,

$$C_w := \left\{u \in \mathbb{R}^N : \sum_{j \notin \mathrm{supp}(w)} |u_j| \le 4\langle w, u\rangle_{\ell_2}\right\}.$$

Define the **alignment coefficient** of w as

$$a(w) := \sup_{\|f_u\|_{L_2(\Pi)} \le 1, u \in C_w} \langle w, u\rangle_{\ell_2}.$$

Bounds on $a(w)$

- $a(w) \leq \|K^{-1/2}w\|_{\ell_2}, \; w \in \mathrm{Im}(K^{1/2})$

Restricted Isometry Constant δ_d : the smallest $\delta \in (0,1)$ such that for all $J \subset \{1, \dots, N\}$ with $\mathrm{card}(J) = d$, the spectrum of the Gram matrix $\Big(\langle h_i, h_j \rangle_{L_2(\Pi)}\Big)_{i,j \in J}$ belongs to the interval $[1-\delta, 1+\delta]$.

- For $d = \mathrm{card}(\mathrm{supp}(w))$,

$$a(w) \leq \frac{C\|w\|_{\ell_2}}{1 - C(\|w\|_{\ell_\infty} \vee 1)\delta_{3d}} \leq \frac{C\|w\|_{\ell_\infty}\sqrt{d}}{1 - C(\|w\|_{\ell_\infty} \vee 1)\delta_{3d}}.$$

Theorem 1 **Oracle Inequality**. *There exist constants* $C, D > 0$ *such that, for all* $\lambda \in \mathbb{D}$ *with* $\mathrm{card}(\mathrm{supp}\lambda) = d$, *for all* $\varepsilon \geq D\sqrt{\frac{d+\log N}{n}}$, *for all* $t > 0$ *and* $t_{n,N} := t + \log N + 4 \log \log_2 n + 2 \log 2$, *with probability at least* $1 - e^{-t}$

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_*\|^2_{L_2(\Pi)} \leq \left[2\|f_\lambda - f_*\|^2_{L_2(\Pi)} + \xi_n^2\right],$$

where

$$\xi_n^2 := C\left[a^2(\mathrm{sign}(\lambda))\varepsilon^2 \bigvee \frac{d + \log N + t_{n,N}}{n}\right].$$

Let $L \subset L_2(\Pi)$, $d = \dim(L)$,

$$U_L(x) := \sup_{h \in L, \|h\|_{L_2(\Pi)} \leq 1} |h(x)| \text{ and } U(L) := \|U_L\|_{L_\infty}$$

Note that

(a) $\|U_L\|_{L_2(\Pi)} = \sqrt{d}$;

(b) If there exists an orthonormal basis of $L \subset L_2(\Pi)$ consisting of uniformly bounded functions, then $U(L) \asymp \sqrt{d}$.

Theorem 2 *There exist constants* $C, D > 0$ *such that, for all* $\lambda \in \mathbb{D}$, *for all* $L \subset L_2(\Pi)$ *with* $d := \dim(L)$, *for all* $t > 0$ *and* $t_n := t + 4\log\log_2 n + 2\log 2$, *for all* $\varepsilon \geq D\sqrt{\frac{\log N}{n}}$, *the following bounds hold with probability at least* $1 - e^{-t}$:

$$\|f_{\hat{\lambda}^\varepsilon} - f_*\|^2_{L_2(\Pi)} \leq \Big[2\|f_\lambda - f_*\|^2_{L_2(\Pi)} + \xi_n^2\Big],$$

where

$$\xi_n^2 := C\Bigg[a^2(\mathrm{sign}(\lambda))\varepsilon^2 \bigvee \frac{d + t_n}{n} \bigvee$$

$$\max_{j\in\mathrm{supp}(\lambda)} \|P_{L^\perp} h_j\|_{L_2(\Pi)} \sqrt{\frac{\log N}{n}} \bigvee \frac{U(L)\log N}{n}\Bigg].$$

Under Restricted Isometry Condition:

Suppose, for a small enough $c > 0$, $\delta_{3d} \leq c$. Then, taking $d := \mathrm{card}(\mathrm{supp}(\lambda)),\ L := \mathrm{l.s.}(h_j : j \in \mathrm{supp}(\lambda))$, we get

$$a(\mathrm{sign}(\lambda)) \leq C\sqrt{d} \text{ and } U(L) \leq C\sqrt{d},$$

implying

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_*\|^2_{L_2(\Pi)} \leq \left[2\|f_\lambda - f_*\|^2_{L_2(\Pi)} + C\frac{d\log N + t_n}{n}\right]$$

and, if $f_* = f_{\lambda_*},\ \mathrm{card}(\mathrm{supp}(\lambda_*)) = d$,

$$\|\hat{\lambda}^{\varepsilon} - \lambda_*\|^2_{\ell_2} \leq C\left[\|\lambda - \lambda_*\|^2_{\ell_2} + \frac{d\log N + t_n}{n}\right].$$

General Dictionaries: Approximation Error Bounds

$$\lambda^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \left[\|f_{\lambda} - f_{*}\|^2_{L_2(\Pi)} + \varepsilon \|\lambda\|_{L_1(\mu)} \right].$$

Approximation Error: $\|f_{\lambda^{\varepsilon}} - f_{*}\|^2_{L_2(\Pi)}$

Gram Operators and Alignment Coefficients

Let $K : L_2(\mu) \mapsto L_2(\mu)$ denote the **Gram operator** of the dictionary $\mathcal{H}$:

$$(Ku)(h) = \int_{\mathcal{H}} \langle h, g \rangle_{L_2(\Pi)} u(g) \mu(dg).$$

For $w \in L_2(\mu)$, let

$$C_w := \Big\{ u : \mathcal{H} \mapsto \mathbb{R} : \int_{\mathcal{H} \setminus \mathrm{supp}(w)} |u| d\mu \leq 4 \langle w, u \rangle_{L_2(\mu)} \Big\}.$$

Define the **alignment coefficient** of w as

$$a(w) := a_{\mathcal{H}}(w) := \sup_{\|f_u\|_{L_2(\Pi)} \leq 1, u \in C_w} \langle w, u \rangle_{L_2(\mu)}.$$

For all $w \in \mathrm{Im}(K^{1/2})$,

$$a(w) \leq \|K^{-1/2} w\|_{L_2(\mu)}.$$

For $\lambda \in \mathbb{D}$, denote

$$\partial|\lambda| := \left\{ w : \mathcal{H} \mapsto [-1, 1] : w(h) = \mathrm{sign}(\lambda(h)), h \in \mathrm{supp}(\lambda) \right\}.$$

Theorem 3 *There exists a constant* $C > 0$ *such that for all* $\varepsilon > 0$, *for all* $\lambda \in \mathbb{D}$ *and for all* $w \in \partial|\lambda|$,

$$\|f_{\lambda^\varepsilon} - f_*\|^2_{L_2(\Pi)} + \varepsilon \int_{\mathcal{H}\setminus \mathrm{supp}(w)} |\lambda^\varepsilon| d\mu \leq$$

$$C\Big[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + a^2(w)\varepsilon^2\Big].$$

Moreover,

$$\|f_{\lambda^\varepsilon} - f_*\|^2_{L_2(\Pi)} \leq \inf_{\lambda\in\mathbb{D}, w\in\partial|\lambda|} \Big[\|f_\lambda - f_*\|^2_{L_2(\Pi)} +$$

$$Ca(w)\varepsilon\|f_\lambda - f_*\|_{L_2(\Pi)} + C^2 a^2(w)\varepsilon^2\Big].$$

Sobolev Norms and Sparsity

$\mathcal{H} := \left\{ h(t, \cdot) : t \in G \right\}, \; G \subset \mathbb{R}^d$

Suppose

$$a(w) \leq C \|w\|_{\mathbb{W}^{2,\alpha}(G)}$$

Sparse Spikes

Suppose $\lambda \in \mathbb{D}$, $\lambda = \sum_{j=1}^{d} \lambda_j$, where $\lambda_j \in L_1(\mu)$, $\mathrm{supp}(\lambda_j) \subset U_j \subset G$, where $U_j, j = 1, \ldots, d$ are disjoint balls.

Let $w = \sum_{j=1}^{d} w_j \in \partial|\lambda|$, where $w_j \in \mathbb{W}^{2,\alpha}(G)$ and $\mathrm{supp}(w_j), j = 1, \ldots, d$ are disjoint. Then

$$a(w) \leq C \left(\sum_{j=1}^{d} \|w_j\|^2_{\mathbb{W}^{2,\alpha}} \right)^{1/2},$$

implying $a(w) \leq \mathrm{const}\ \sqrt{d}$.

Example: Fourier Dictionary

$S := \mathbb{R}^d$

$\mathcal{H} := \left\{ \cos\langle t, \cdot \rangle : t \in G \right\}$

$G \subset \mathbb{R}^d$ bounded open subset, $G = -G$

μ, Π absolutely continuous measures with densities m, p, $m(t) = m(-t)$

If

$$p(x) \geq L(1 + |x|^2)^{-\alpha},$$

then

$$a(w) \leq C\|w\|_{\mathbb{W}^{2,\alpha}(\mathbb{R}^d)}.$$

Example: Location Dictionary

$S := \mathbb{T}^d$

$\mathcal{H} := \left\{ h(\cdot - t) : t \in \mathbb{T}^d \right\}$

Π probability measure with density p, p bounded away from 0

μ Haar measure in $\mathbb{T}^d$

If

$$|\tilde{h}_n| \geq L(1 + |n|^2)^{-\alpha/2},\ n \in \mathbb{Z}^d,$$

then

$$a(w) \leq C\|w\|_{\mathbb{W}^{2,\alpha}(\mathbb{T}^d)}.$$

Example: Decision Stumps

$S := [0,1]$

$\mathcal{H} := \left\{ I_{[0,t]} - I_{(t,1]} : t \in [0,1] \right\}$

Π absolutely continuous measure in $[0,1]$ with density p that is bounded away from 0

$$a(w) \leq C\|w\|_{\mathbb{W}^{2,1}[0,1]}.$$

Weakly Correlated Partitions and Sparsity

$\{\mathcal{H}_j,\ j=1,\dots,N\}$ a measurable partition of $\mathcal{H}$

$\mathcal{L}_j := \text{c.l.s.}(\mathcal{H}_j)$

$$\sigma_\Pi(g) := \text{cov}_\Pi(g,g), \quad \rho_\Pi(h,g) := \frac{\text{cov}_\Pi(h,g)}{\sigma_\Pi(h)\sigma_\Pi(g)}.$$

Restricted Isometry Constant δ_d : the smallest $\delta \in (0,1)$ such that for all $J \subset \{1,\dots,N\}$ with $\text{card}(J)=d$ and all $h_j \in \mathcal{L}_j,\ j\in J$, the spectrum of the correlation matrix $\Big(\rho_\Pi(h_i,h_j)\Big)_{i,j\in J}$ belongs to the interval $[1-\delta, 1+\delta]$.

Let $K_j : L_2(\mathcal{H}_j, \mu) \mapsto L_2(\mathcal{H}_j, \mu)$,

$$(K_j u)(h) = \int_{\mathcal{H}_j} \mathrm{cov}_{\Pi}(h, g) u(g) \mu(dg),\ \ h \in \mathcal{H}_j.$$

Proposition 1 *For all* $J \subset \{1, \dots, N\}$ *with* $d := \mathrm{card}(J)$ *and all* $w = \sum_{j \in J} w_j$ *with* $w_j \in \mathrm{Im}(K_j^{1/2})$ *and*

$$B := \max_{j \in J} \|K_j^{-1/2} w_j\|_{L_2(\mathcal{H}_j, \mu)}$$

the following bound holds with some numerical constant $C > 0$:

$$a(w) \le \frac{CB\sqrt{d}}{1 - CB\delta_{3d}}.$$

Random Error Bounds and Oracle Inequalities

Under some **complexity assumptions** on the dictionary $\mathcal{H}$, we will provide upper bounds on $\|f_{\hat{\lambda}^{\varepsilon}} - f_{\lambda}\|^2_{L_2(\Pi)}$ for an arbitrary function $\lambda \in \mathbb{D}$ and

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \Big[n^{-1} \sum_{j=1}^{n} (Y_j - f_{\lambda}(X_j))^2 + \varepsilon \|\lambda\|_{L_1(\mu)} \Big].$$

Complexity Assumptions on the Dictionary

Suppose there exists a function $H(u) \geq 0, u > 0$, $H(u) \to \infty$ as $u \to 0$, H regularly varying of exponent $\alpha \in [0, 2)$ and such that the following condition on the **random covering numbers** holds

$$\log N(\mathcal{H}; L_2(\Pi_n); u) \leq H(u), \ \ u > 0 \text{ a.s.},$$

or the following condition on the **bracketing numbers** holds

$$\log N_{[\ ]}(\mathcal{H}; L_2(\Pi); u) \leq H(u), \ \ u > 0.$$

Approximation by Finite Dimensional Subspaces

$L \subset L_2(\Pi)$ a linear subspace, $\dim(L) < +\infty$

For $\mathcal{H}' \subset \mathcal{H}$,

$$\rho(\mathcal{H}'; L) := \sup_{h \in \mathcal{H}'} \|P_{L^{\perp}} h\|_{L_2(\Pi)}$$

Theorem 4 *There exist constants* $C, D > 0$ *such that for all* $\lambda \in \mathbb{D}$, $w \in \partial|\lambda|$, $L \subset L_2(\Pi)$ *with* $d := \dim(L)$ *and* $\rho := \rho(\mathrm{supp}(w); L)$, *for all* $t > 0$ *and* $t_n := t + 4\log\log_2 n + 2\log 2$, *for all* $\varepsilon \geq D\sqrt{\frac{H(1/\sqrt{d})}{n}}$, *the following bounds hold with probability at least* $1 - e^{-t}$:

$$\|f_{\hat{\lambda}_\varepsilon} - f_\lambda\|^2_{L_2(\Pi)} + \varepsilon \int_{\mathcal{H}\setminus \mathrm{supp}(w)} |\hat{\lambda}^\varepsilon| d\mu \leq$$

$$C\Big[\|f_\lambda - f_*\|^2_{L_2(\Pi)} \bigvee a^2(w)\varepsilon^2 \bigvee$$

$$\frac{d + t_n}{n} \bigvee \rho\sqrt{\frac{H(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)H(\rho/\sqrt{d})}{n}\Big]$$

Moreover, with the same probability, the following **sparsity oracle inequality** holds:

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_*\|_{L_2(\Pi)}^2 \leq \Big[\|f_{\lambda} - f_*\|_{L_2(\Pi)}^2 + \|f_{\lambda} - f_*\|_{L_2(\Pi)}\xi_n + \xi_n^2\Big],$$

where

$$\xi_n^2 := C\Bigg[a^2(w)\varepsilon^2 \bigvee \frac{d+t_n}{n} \bigvee$$

$$\rho\sqrt{\frac{H(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)H(\rho/\sqrt{d})}{n}\Bigg].$$

Regularized Boosting

The problem

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - f_\lambda(X_j))^2 + \varepsilon \|\lambda\|_{L_1(\mu)} \right]$$

can be viewed as **a regularized boosting.**

Blanchard, Lugosi and Vayatis (2003) obtained oracle inequalities for regularized boosting (for more general losses than quadratic) with error rate $n^{-\frac{1}{2}\frac{V+2}{V+1}}$, where V is the VC-dimension of the base class $\mathcal{H}$. In the case of N-dimensional decision stumps, $V = [2\log_2(2N)]$.

An Additive Model: High-Dimensional Decision Stumps

$S := [0,1]^N$

$\mathcal{H}_j := \left\{ h_t^{(j)} : t \in [0,1] \right\}$

$h_t^{(j)}(x) := I_{[0,t]}(x_j) - I_{(t,1]}(x_j), x = (x_1, \ldots, x_N) \in S$

$\mathcal{H} := \bigcup_{j=1}^{N} \mathcal{H}_j$

μ "Lebesgue measure"

$\lambda := \sum_{j=1}^{N} \lambda_j, \ \mathrm{supp}(\lambda_j) \subset \mathcal{H}_j$

$f_\lambda(x) = \sum_{j=1}^{N} f_{\lambda_j}(x_j)$

$\|\lambda_j\|_{L_1(\mathcal{H}_j,\mu)} = \frac{1}{2}\|f_{\lambda_j}\|_{TV}$

L_1-penalization is equivalent to

$$(\hat{f}_1^\varepsilon, \ldots, \hat{f}_N^\varepsilon) :=$$

$$\mathrm{argmin}\Big[n^{-1}\sum_{j=1}^{N}(Y_j - (f_1 + \cdots + f_N)(X_j))^2 + \frac{\varepsilon}{2}\sum_{j=1}^{N}\|f_j\|_{TV}\Big].$$

Sparsity in Additive Models

Let $J \subset \{1, \dots, N\}$, $d := \mathrm{card}(J)$ and let Λ_s be the set of **sparse functions** λ such that

(a) $\lambda = \sum_{j \in J} \lambda_j, \ \mathrm{supp}(\lambda_j) \subset \mathcal{H}_j$

(b) there exist $w_j \in \partial|\lambda_j|, \ j \in J$, $\|w_j\|_{\mathbb{W}^{2,1}[0,1]} \leq L$, L is a constant.

Suppose also that the spaces $\mathcal{L}_j = \mathrm{c.l.s.}(\mathcal{H}_j)$ are **"weakly correlated"** (e.g., δ_{3d} is bounded by a small constant).

Let $\varepsilon := D\sqrt{\frac{\log N}{n}}$

Then, for all $\lambda \in \Lambda_s$, with probability at least $1 - e^{-t}$,

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_*\|^2_{L_2(\Pi)} \leq \left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + \xi_n \|f_\lambda - f_*\|_{L_2(\Pi)} + \xi_n^2\right],$$

where

$$\xi_n^2 := C\left[\frac{(d\log(nd))^{1/3}}{n^{2/3}} + \frac{d\log N + t + 4\log(2^{1/2}\log_2 n)}{n}\right]$$

Sparse Recovery in Convex Hulls: Entropy Penalization

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. copies of (X, Y)

$$\hat{\lambda}^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - f_{\lambda}(X_j))^2 + \varepsilon \int_{\mathcal{H}} \lambda \log \lambda d\mu \right]$$

$$\lambda^{\varepsilon} := \mathrm{argmin}_{\lambda \in \mathbb{D}} \left[\|f_{\lambda} - f_*\|^2_{L_2(\mu)} + \varepsilon \int_{\mathcal{H}} \lambda \log \lambda d\mu \right]$$

$\mathbb{D}$ is a convex set of probability densities with respect to μ

$\varepsilon > 0$ regularization parameter

Symmetrized Kullback-Leibler Distance

$$K(\lambda_1|\lambda_2) := \int_{\mathcal{H}} \lambda_1 \log\left(\frac{\lambda_1}{\lambda_2}\right) d\mu$$

$$K(\lambda_1, \lambda_2) := K(\lambda_1|\lambda_2) + K(\lambda_2|\lambda_1)$$

For $\lambda \in \mathbb{D}$,

$$\Lambda(A) := \int_A \lambda d\mu, \ A \subset \mathcal{H}.$$

Theorem 5 **Approximation Error**. *There exists a constant* $C > 0$ *such that for all* $\varepsilon > 0$ *and for all* $\lambda \in \mathbb{D}$

$$\|f_{\lambda^\varepsilon} - f_*\|^2_{L_2(\Pi)} + \varepsilon K(\lambda^\varepsilon, \lambda) \leq$$

$$C\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + a^2(\log\lambda)\varepsilon^2\right].$$

Moreover,

$$\|f_{\lambda^\varepsilon} - f_*\|^2_{L_2(\Pi)} \leq \inf_{\lambda\in\mathbb{D}}\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + \right.$$

$$\left. Ca(\log\lambda)\varepsilon\|f_\lambda - f_*\|_{L_2(\Pi)} + C^2a^2(\log\lambda)\varepsilon^2\right]$$

In addition, for all $\mathcal{H}' \subset \mathcal{H}$

$$\Lambda^{\varepsilon}(\mathcal{H}\setminus\mathcal{H}') \le 2\Lambda(\mathcal{H}\setminus\mathcal{H}') + \frac{C}{\varepsilon}\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + a^2(\log\lambda)\varepsilon^2\right]$$

and

$$\Lambda(\mathcal{H}\setminus\mathcal{H}') \le 2\Lambda^{\varepsilon}(\mathcal{H}\setminus\mathcal{H}') + \frac{C}{\varepsilon}\left[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + a^2(\log\lambda)\varepsilon^2\right].$$

Theorem 6 **Random Error**. *There exist constants* $C, D > 0$ *such that for all* $\mathcal{H}' \subset \mathcal{H}$, $L \subset L_2(\Pi)$ *with* $d := \dim(L)$ *and* $\rho := \rho(\mathcal{H}'; L)$, *for all* $t > 0$ *and* $t_n := t + 4\log\log_2 n + 2\log 2$, *for all* $\varepsilon \geq D\sqrt{\frac{H(1/\sqrt{d})}{n}}$, *the following bounds hold with probability at least* $1 - e^{-t}$:

$$\|f_{\hat{\lambda}^\varepsilon} - f_{\lambda^\varepsilon}\|^2_{L_2(\Pi)} + \varepsilon K(\hat{\lambda}^\varepsilon, \lambda^\varepsilon) \leq C\Bigg[\frac{d + t_n}{n} \bigvee$$

$$\rho\sqrt{\frac{H(\frac{\rho}{\sqrt{d}})}{n}} \bigvee \Lambda^\varepsilon(\mathcal{H} \setminus \mathcal{H}')\sqrt{\frac{H(\frac{1}{\sqrt{d}})}{n}} \bigvee \frac{U(L)H(\frac{\rho}{\sqrt{d}})}{n}\Bigg]$$

In addition,

$$\hat{\Lambda}^{\varepsilon}(\mathcal{H} \setminus \mathcal{H}') \leq C\Bigg[\Lambda^{\varepsilon}(\mathcal{H} \setminus \mathcal{H}') \bigvee \frac{d + t_n}{n\varepsilon} \bigvee$$

$$\frac{\rho}{\varepsilon}\sqrt{\frac{H(\frac{\rho}{\sqrt{d}})}{n} \bigvee \frac{U(L)H(\frac{\rho}{\sqrt{d}})}{n\varepsilon}}\Bigg]$$

and

$$\Lambda^{\varepsilon}(\mathcal{H} \setminus \mathcal{H}') \leq C\Bigg[\hat{\Lambda}^{\varepsilon}(\mathcal{H} \setminus \mathcal{H}') \bigvee \frac{d + t_n}{n\varepsilon} \bigvee$$

$$\frac{\rho}{\varepsilon}\sqrt{\frac{H(\frac{\rho}{\sqrt{d}})}{n} \bigvee \frac{U(L)H(\frac{\rho}{\sqrt{d}})}{n\varepsilon}}\Bigg].$$

Moreover, for all $\lambda \in \mathbb{D}$, *with the same probability, the following* **sparsity oracle inequality** *holds*

$$\|f_{\hat{\lambda}^{\varepsilon}} - f_*\|^2_{L_2(\Pi)} \leq \Big[\|f_\lambda - f_*\|^2_{L_2(\Pi)} + \|f_\lambda - f_*\|_{L_2(\Pi)}\xi_n + \xi_n^2\Big],$$

where

$$\xi_n^2 := C\Big[a^2(\log\lambda)\varepsilon^2 \bigvee \frac{d+t_n}{n} \bigvee$$

$$\rho\sqrt{\frac{H(\rho/\sqrt{d})}{n}} \bigvee \Lambda(\mathcal{H}\backslash\mathcal{H}')\sqrt{\frac{H(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)H(\rho/\sqrt{d})}{n}\Big].$$