On Convex Hulls
and the (Im)Possibility of Overparametrization

Sara van de Geer

March 6 2024

2024 International Zurich Seminar on Information and Communication

Outline

This talk is about deriving bounds for the entropy of convex hulls.
If the entropy of a model is small, it means there is no overparametrization.

Motivation for entropy bounds: least squares regression

Least squares estimator:

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \underbrace{\|Y - f\|_2^2}_{:=\sum_{i=1}^n (Y_i - f_i)^2}$$

where

∘ $Y \in \mathbb{R}^n$ is a random vector of observations

and

∘ $\mathcal{F} \subset (\mathbb{R}^n, \|\cdot\|_2)$ is a given class of regression functions.

Estimation error depends on the entropy $\mathcal{H}(\cdot, \mathcal{F})$ of $\mathcal{F}$.

More precisely, define

$$f^0 := \mathbb{E}Y$$

and suppose (say) that

$$\xi := Y - \mathbb{E}Y \sim \mathcal{N}(0, I)/\sqrt{n}$$

If $f^0 \in \mathcal{F}$ (no model misspecification) then

$$\|\hat{f} - f^0\|_2 \overset{\mathbb{P}}{\asymp} \epsilon_n$$

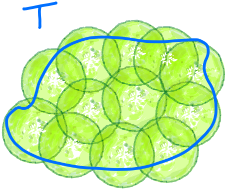
where

$$\epsilon_n^2 \asymp \frac{\mathcal{H}(\epsilon_n, \mathcal{F})}{n}$$

(assuming $\mathcal{H}(\epsilon, \mathcal{F}) \lesssim \epsilon^{-2}$).

Definition

Let (T, d) be a subset of a metric space.
For $\epsilon > 0$ the ϵ-covering number $N(\epsilon, T)$ of T is defined as the minimum number of balls with radius ϵ, necessary to cover the space T.
The entropy of T is $\mathcal{H}(\cdot, T) := \log N(\cdot, T)$.



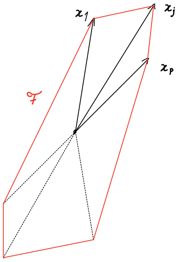
∘ $X \in \mathbb{R}^{n \times p}$ given input matrix: $X = (x_1, \ldots, x_p)$

Definition *The convex hull of X is*

$$\begin{aligned}\mathcal{F} &:= \operatorname{conv}(X) \\ &:= \{f = X\beta : \ \beta \in [0, \infty)^p, \sum_{j=1}^{p} \beta_j \le 1\}.\end{aligned}$$

Definition *The absolute convex hull of X is*

$$\begin{aligned}\mathcal{F} &:= \operatorname{absconv}(X) \\ &:= \{f = X\beta : \ \beta \in \mathbb{R}^p, \ \underbrace{\|\beta\|_1}_{=: \sum_{j=1}^p |\beta_j|} \le 1\}.\end{aligned}$$



We study bounds for $\mathcal{H}(\cdot, \mathcal{F})$.

Example Mixture model U observed, V unobserved

$$x_{i,j} = \underbrace{\mathbb{P}(U = i | V = j)}_{\text{known}}, \ \underbrace{\beta_j = \mathbb{P}(V = j)}_{\text{unknown}}.$$

Then

$$\mathbb{P}(U = i) = \sum_{j=1}^{p} \mathbb{P}(U = i | V = j)\mathbb{P}(V = j) = (X\beta)_i.$$

Example Discrete version of

- $f : \mathbb{R}^d \to \mathbb{R}$
- $\beta = Df$
- $Df(u) = \prod_{k=1}^{d} \partial f(u)/(\partial u_k), \ u = (u_1, \ldots, u_d)$
- $\|Df\|_1 = \int |df|$ the Vitali total variation of $f : \mathbb{R}^d \to \mathbb{R}$
- $\mathcal{F} := \{f : \ \|Df\|_1 \leq 1\}$

We write

$$X \in \mathbb{R}^{n\times p},\ X = (x_1, \dots, x_p),$$

$$x_j \in \mathbb{R}^n,\ j \in [1:p]$$

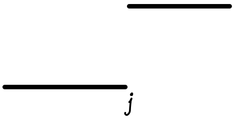
Typically X will be the extreme points of $\mathcal{F}$.

Normalization We assume $\|x_j\|_2 \leq 1$, $j \in [1:p]$

Polynomial covering numbers
We examine the case where for some $\mathbf{V} > 0$

$$N(\epsilon, X) \asymp \epsilon^{-\mathbf{V}},\ \epsilon > 0$$

Example Heaviside functions



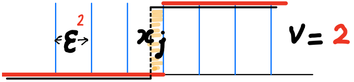
$x_j = \psi_j(\cdot)$
$= 1\{\cdot \geq j\}$

$$x_{i,j} := \psi_j(i),\ (i,j) \in [1:n]^2$$
$$\psi_j(i) := 1(i \geq j)$$

$$\rightsquigarrow \mathcal{F} = \{f :\ [1:n] \to \mathbb{R},\ \mathrm{TV}(f) \leq 1\}$$
$$TV(f) := \sum_{i=2}^{n} |f(i) - f(i-1)| \leq 1 \text{ total variation}$$

V $= 2$:

$$N(\epsilon, X) \asymp \epsilon^{-2}, \epsilon > 0$$



Entropy bounds based on covering numbers

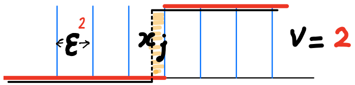
Theorem [Ball and Pajor, 1992] *For $\mathcal{F} = \mathrm{absconv}(X)$*

$$N(\epsilon, X) \lesssim \epsilon^{-\mathbf{V}} \;\Rightarrow\; \mathcal{H}(\epsilon, \mathcal{F}) \le \epsilon^{-\frac{2\mathbf{V}}{2+\mathbf{V}}}.$$

$\square$

Note that $\frac{2\mathbf{V}}{2+\mathbf{V}} < 2$: Dudley's entropy integral exists.
Example Heaviside functions

$$\begin{aligned} \mathbf{V} &= 2: \\ N(\epsilon, X) &\asymp \epsilon^{-2} \\ \frac{2\mathbf{V}}{2+\mathbf{V}} &= 1 \\ \Rightarrow \mathcal{H}(\epsilon, \mathcal{F}) &\le \epsilon^{-1} \end{aligned}$$



This entropy bound is tight.

Example One hidden layer neural networks

Let $Z = (z_1, \ldots, z_n) \in \mathbb{R}^{d\times n}$ be a given input matrix. We define

$$x_{w,c}(i) := (\langle z_i, w\rangle - c)_+, \; w \in \mathcal{S}^{d-1}, c \in \mathbb{R}.$$

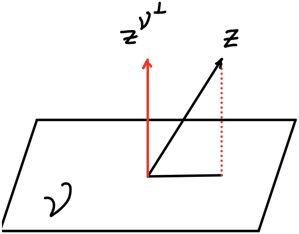
$$\begin{aligned} N(\epsilon, X) &\lesssim \epsilon^{-d} \\ \Rightarrow \mathcal{H}(\epsilon, \mathcal{F}) &\lesssim \epsilon^{-\frac{2d}{2+d}} \end{aligned} \qquad\qquad \begin{aligned} \mathbf{V} &\leq d \\ \frac{2\mathbf{V}}{2+\mathbf{V}} &\leq \frac{2d}{2+d} \end{aligned}$$

The entropy bound based on covering numbers is *not* tight in general.

Entropy bounds based on approximation numbers

Notation Let $\mathcal{V} \subset \mathbb{R}^n$ linear and $z \in \mathbb{R}^n$.

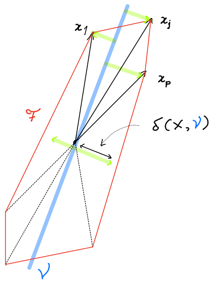
$$
\begin{aligned}
z^{\mathcal{V}^\perp} &:= \arg\min_{f \in \mathcal{V}} \|z - f\|_2 \\
\delta(X, \mathcal{V}) &:= \max_{j \in [1:p]} \|x_j^{\mathcal{V}^\perp}\|_2
\end{aligned}
$$

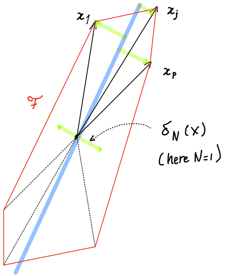


Definition *Let $N \in \mathbb{N}$. Then*

$$\delta_N(X) := \min\left\{\delta(X, \mathcal{V}) : \dim(\mathcal{V}) = N\right\}.$$

is the N-approximation number of $X = (x_1, \ldots, x_p)$.





Definition *We call*

$$M(\epsilon, X) := \min\{N : \ \delta_N(X) \leq \epsilon\}$$

the ϵ-approximation of X.

Lemma

$$M(\epsilon, X) \leq N(\epsilon, X) \; \forall \, \epsilon > 0 \; \square$$

Approximation numbers lead to better entropy bounds

Theorem I *For* $\mathcal{F} := \mathrm{absconv}(X)$

$$\begin{aligned} M(\epsilon, X) &\lesssim \epsilon^{-W} \log^{w}(1/\epsilon) \\ \Rightarrow \mathcal{H}(\epsilon, \mathcal{F}) &\lesssim \epsilon^{-\frac{2W}{2+W}} \log^{\frac{2w}{2+w}}(1/\epsilon) \log^{\frac{W}{2+W}}(1/\epsilon) \quad \square \end{aligned}$$

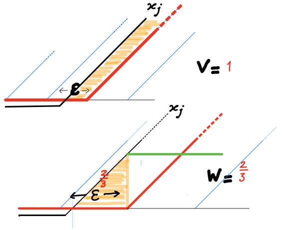
Example Hinge functions.

$$
\begin{aligned}
x_{i,j} &:= \frac{1}{n}\psi_j(i) \\
\psi_j(i) &:= (i-j)_+ = (i-j)\mathbb{1}\{i \geq j\}
\end{aligned}
$$

$$
\rightsquigarrow \ \mathcal{F} = \left\{ f : n \sum_i |f(i) - 2f(i-1) + f(i-2)| \leq 1 \right\}
$$

$$
\begin{aligned}
N(\epsilon, X) &\asymp \epsilon^{-1} \\
M(\epsilon, X) &\asymp \epsilon^{-\frac{2}{3}} \\
\Rightarrow \mathcal{H}(\epsilon, \mathcal{F}) &\lesssim \epsilon^{-\frac{1}{2}} \log^{\frac{1}{4}}(1/\epsilon)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V} &= 1 \\
W &= \frac{2}{3} \\
\frac{2W}{2+W} &= \frac{1}{2}
\end{aligned}
$$



$$
3\int_0^{\epsilon^{2/3}} x^2 dx = x^3\Big|_0^{\epsilon^{2/3}} = \epsilon^2
$$

Example

Truncated power basis

$$\begin{aligned} q &\in \mathbb{N} \text{ given} \\ x_{i,j} &:= \frac{1}{n^{q-1}} \psi_j(i) \\ \psi_j(i) &:= (i-j)_+^{q-1} \end{aligned}$$

$$\begin{aligned} N(\epsilon, X) &\asymp \epsilon^{-1} \ (q > 1) \\ M(\epsilon, X) &\asymp \epsilon^{-\frac{2}{2q-1}} \\ \Rightarrow \mathcal{H}(\epsilon, \mathcal{F}) &\lesssim \epsilon^{-\frac{1}{q}} \log^{\frac{1}{2q}}(1/\epsilon) \end{aligned}$$

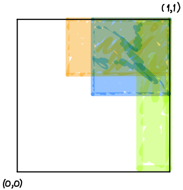
$$\begin{aligned} \mathbf{V} &= 1 \\ W &= \frac{2}{2q-1} \\ \frac{2W}{2+W} &= \frac{1}{q} \end{aligned}$$

Example

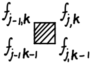
Higher dimensional total variation. Let

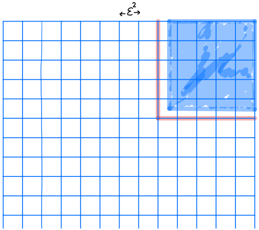
$$\begin{aligned} \Psi &:= \{1\{i \geq j\} : (i,j) \in [1:m]^2\} \in \mathbb{R}^{m\times m}, \\ X &:= \Psi \otimes \Psi, \; n = m^2 \end{aligned}$$

Kronecker product of heaviside functions= half-intervals in $\mathbb{R}^2$:

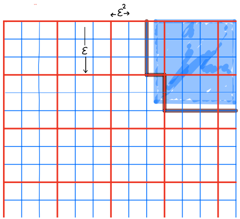


$$\begin{aligned} \rightsquigarrow \; \mathcal{F} &= \mathrm{absconv}(\Psi \otimes \Psi) \\ &= \left\{ f \in \mathbb{R}^{m\times m} : \sum_{j,k} |f_{j,k} - f_{j-1,k} - f_{j,k-1} + f_{j-1,k-1}| \leq 1 \right\}. \end{aligned}$$





$\mathbf{V} = 4$



$W \leq 3$

$$
\begin{aligned}
N(\epsilon, X) &\sim \epsilon^{-4} \\
M(\epsilon, X) &\lesssim \epsilon^{-3} \\
\Rightarrow \mathcal{H}(\epsilon, \mathcal{F}) &\lesssim \epsilon^{-\frac{6}{5}} \log^{\frac{3}{5}}(1/\epsilon)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V} &= 4 \\
W &\leq 3 \\
\frac{2W}{2+W} &\leq \frac{6}{5}
\end{aligned}
$$

Tightness of the bound based on approximation numbers?

$$
\begin{array}{rcl}
M(\epsilon, X) & \asymp & \epsilon^{-W} \log^{\cdots}(1/\epsilon) \\
& \overset{?}{\Leftrightarrow} & \\
\mathcal{H}(\epsilon, \mathcal{F}) & \asymp & \epsilon^{-\frac{2W}{2+W}} \log^{***}(1/\epsilon)
\end{array}
$$

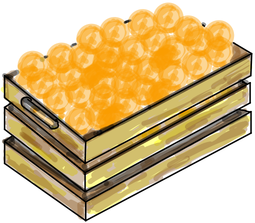
Literature

- R.F. Bass, *Probability inequalities for multiparameter Brownian processes* (1988)
- G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry* (1989)
- J. Kuelbs, W.V. Li, *Metric entropy and the small ball problem for Gaussian measures* (1992)
- T. Dunker, W. Linde, T. Kühn, M.A. Lifshits, *Metric entropy of integration operators and small ball probabilities for the Brownian sheet* (1999)
- W.V. Li, W. Linde, *Approximation, metric entropy and small ball estimates for Gaussian measures* (1999)
- S. Artstein, V. Milman, S. Szarek, N. Tomczak-Jaegermann, *On convexified packing and entropy duality* (2004)
- R. Blei, F. Gao, W.V. Li, *Metric entropy of high dimensional distributions* (2007)

A too brief to be serious look at duality, packing and volume-metric arguments, small balls, Gaussian measures

Ingredients

- ∘ mappings between Banach and Hilbert spaces
- ∘ duality theorem
- ∘ packing numbers
- ∘ volume-metric arguments
- ∘ Gaussian measures
- ∘ small ball estimates



Let $u: \mathbf{H} \to \mathbf{B}$ be a mapping from a Hilbert space $\mathbf{H}$ to a Banach space $\mathbf{B}$.

In our case $\mathbf{H} = \mathbb{R}^n$, $\mathbf{B} = (\mathbb{R}^p, \|\cdot\|_\infty)$ and $u: y \mapsto X^T y$.

Duality theorem [Artstein et al. (2004)] *in our case. Let $\mathcal{B}$ be the unit ball in $\mathbb{R}^n$. Equip $\mathcal{B}$ with the metric induced by the norm $\|y\|_X := \|X^T y\|_\infty$, $y \in \mathbb{R}^n$. Then*

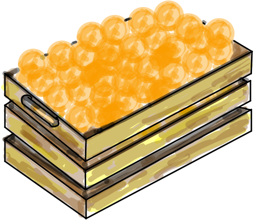
$$\mathcal{H}(\epsilon, \mathcal{F}) \asymp \mathcal{H}(\epsilon, (\mathcal{B}, \|\cdot\|_X)) \quad \square$$

Packing and covering

Definition *Let (T, d) be a subset of a metric space. The ϵ-packing number of T is the maximal number of balls with radius ϵ that T can contain.*

Then

$$N(\epsilon, T) \leq D(\epsilon, T) \leq N(\epsilon/2, T)$$



Volume-metric argument:

Write the balls in the ϵ-packing set as $B_1, \ldots, B_D$. Then obviously

$$D(\epsilon, T) \leq \frac{\mathrm{vol}(T)}{\min_k \mathrm{vol}(B_k)}.$$

Small balls

-Restricted to our setting-

Let $\xi \sim \mathcal{N}(0, I)$ be a standard Gaussian random vector in $\mathbb{R}^n$.
Define the small ball behaviour

$$\phi(\epsilon, X) = -\log \mathbb{P}(\|\xi\|_X \leq \epsilon)$$

where (recall) $\|\xi\|_X = \|X^T\xi\|_\infty$. This measures the size of small balls in the space $(\mathcal{B}, \|\cdot\|_X)$.

Theorem [Kuelbs and Li (1993)]

$$\phi(\epsilon, X) \asymp \epsilon^{-W} \;\Leftrightarrow\; \mathcal{H}(\epsilon, (\mathcal{B}, \|\cdot\|_X)) \asymp \epsilon^{-\frac{2W}{2+W}}.$$

$\square$

Small ball for Brownian sheet

Theorem [Bass, 1988]. *Let* $\{\mathcal{W}(t) : t \in [0,1]^2\}$ *be the 2-dimensional Brownian sheet. Then*

$$\phi(\epsilon) := -\log \mathbb{P}(\sup_t \mathcal{W}(t) \leq \epsilon) \lesssim \epsilon^{-2} \log^3(1/\epsilon).$$

(In d dimensions this becomes $\epsilon^{-2} \log^{3(d-1)}(1/\epsilon)$*.)* □

Conclusion: the entropy bound based on approximation numbers is tight up to log-terms

Literature $\rightsquigarrow$

$$
\begin{array}{rcl}
M(\epsilon, X) & \asymp & \epsilon^{-W} \log^{\cdots}(1/\epsilon) \\
& \overset{!}{\Leftrightarrow} & \\
\mathcal{H}(\epsilon, \mathcal{F}) & \asymp & \epsilon^{-\frac{2W}{2+W}} \log^{***}(1/\epsilon)
\end{array}
$$

We can extend the situation to

$$
\begin{array}{rcl}
X & := & \{x_v : \ v \in \Theta\} \subset L_2(Q) \\
\mathcal{F} & := & \mathrm{absconv}(X) \subset L_2(Q)
\end{array}
$$

Theorem [Blei, Gao, Li, 2007]
∘ μ *Lebesgue measure on* $[0,1]$
∘ $\mathcal{F} \subset L_2(\mu \times \cdots \times \mu)$ *all distribution functions on* $[0,1]^d$.
Then

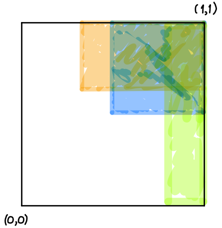
$$\mathcal{H}(\epsilon, \mathcal{F}) \lesssim \epsilon^{-1} \log^{d-1/2}(1/\epsilon)$$

so that $\frac{2W}{2+W} = 1$ *for all dimensions* d.
Proof. Duality theorem + small ball estimates, etc. □

Corollary (reverse thinking)
For Ψ *the heaviside functions*

$$M(\epsilon, \underbrace{\Psi \otimes \cdots \otimes \Psi}_{d \text{ times}}) \lesssim \frac{1}{\epsilon^2} \log^{\cdots}(1/\epsilon),$$

so that $W = 2$ *for all dimensions* d.



The class $\Psi \otimes \Psi$

Example Higher dimensional total variation
Direct proof that $W = 2$ for all dimensions d.
Let

$$\Psi := \{\psi_j(i) := 1\{i \geq j\} : (i,j) \in [1:m]^2\} \in \mathbb{R}^{m\times m}$$

be the heaviside functions. and

$$X := \Psi \otimes \Psi,$$

be half-intervals in $\mathbb{R}^2$.
Recall for $\mathcal{F} = \mathrm{absconv}(X)$

$$\mathcal{F} = \left\{ f \in \mathbb{R}^{m\times m} : \sum_{j,k} |f_{j,k} - f_{j-1,k} - f_{j,k-1} + f_{j-1,k-1}| \leq 1 \right\}.$$

- Before for $d = 2$

$$\begin{aligned} X &:= \Psi \otimes \Psi \\ M(\epsilon, X) &\lesssim \epsilon^{-3} \\ (W &\leq 3) \end{aligned}$$

- We now show for all d

$$\begin{aligned} X &:= \underbrace{\Psi \otimes \cdots \otimes \Psi}_{d \text{ times}} \\ M(\epsilon, X) &\asymp \epsilon^{-2} \log^{2(d-1)}(1/\epsilon) \\ W = 2 & \quad \text{for all dimensions } d \end{aligned}$$

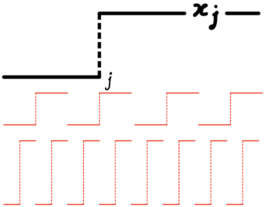
Indeed, $M(\epsilon, X) \lesssim \epsilon^{-2} \log^{2(d-1)}(1/\epsilon)$ can be shown using the Haar basis. Let for $i \in [1:m]$, $m^2 = n$

$$\begin{aligned} h_{j,k}(i) &:= 2^{(k-1)/2} \mathbb{1}\left\{ i/n \in \left[\frac{2j-2}{2^k}, \frac{2j-1}{2^k} \right) \right\} \\ &- 2^{(k-1)/2} \mathbb{1}\left\{ i/n \in \left[\frac{2j-1}{2^k}, \frac{2j}{2^k} \right) \right\}. \end{aligned}$$

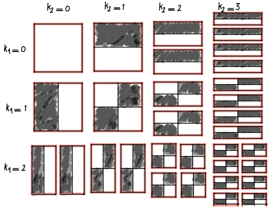
For $\mathbf{k} := (k_1, k_2)$, $\mathbf{j} = (j_1, j_2)$

$$\mathbf{h}_{\mathbf{j},\mathbf{k}}(i_1, i_2) = h_{j_1,k_1}(i_1) h_{j_2,k_2}(i_2)$$

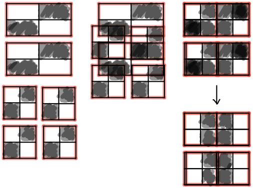
We only keep the basis functions until resolution level $\mathbf{k}$ with $k_1 + k_2 \leq K$ where we choose K such that $K2^{-K} \asymp \epsilon^2$. The number of such $\mathbf{k}$ is $\asymp K2^K$. So $N \asymp 2^K K^2 \asymp \epsilon^{-2} \log^2(1/\epsilon)$.



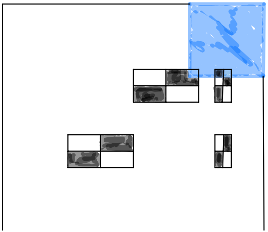
Haar basis in dimension $d = 1$



Haar basis in dimension $d = 2$



Orthogonalty of the basis functions

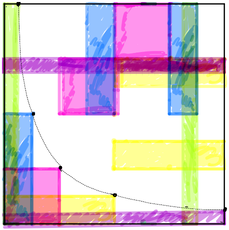


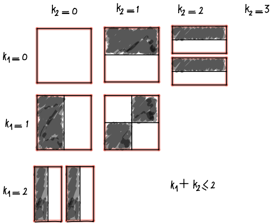
A halfinterval only correlates with basis functions containing its corner

- Haar basis $L_2(\mu)$: $P = \{p_1, p_2, \ldots\}$
- Haar basis $L_2(\mu \times \cdots \times \mu)$: $P \otimes \cdots \otimes P$
- $P_k \subset P$: Haar basis at resolution level k
- $\mathcal{V} := \mathrm{span}\left(\left\{P_{k_1} \otimes \cdots \otimes P_{k_d} \;:\; k_1 + \cdots k_d \leq K\right\}\right)$

= all basis functions with volume (area) below a certain threshold

=: linear space to project on.





End of proof: $W = 2$ for all dimensions d. $\square$

Corollary *We recover the result of [Blei, Gao, Li, 2007]*

$\circ$ $\mu :=$ *uniform measure on* $[0, 1]$ *or on* $[1 : m]$

$\circ$ $\Psi \subset L_2(\mu)$ *heaviside functions*

$\circ$ $\mathcal{F} := \mathrm{absconv}(\Psi \otimes \cdots \otimes \Psi) \subset L_2(\mu \times \cdots \times \mu)$ *Then*

$$\mathcal{H}(\epsilon, \mathcal{F}) \lesssim \frac{1}{\epsilon} \log^{d-1/2}(1/\epsilon).$$

Extension to general tensors

- $\Psi := \{\psi_v : v \in \Theta\} \subset$ unit ball $\in L_2(\mu)$
- $X := \underbrace{\Psi \otimes \cdots \otimes \Psi}_{d \text{ times}}$
- $\mathcal{F} := \text{absconv}(X) \subset L_2(\mu \times \cdots \times \mu)$.

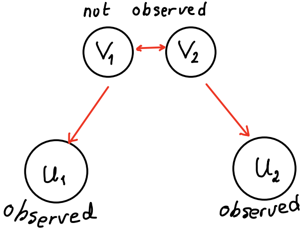
Example Mixtures

Let $U = (U_1, U_2)$ be observed and $V = (V_1, V_2)$ a latent variable.

Assume that given $V = (v_1, v_2)$, U_1 and U_2 are independent with densities ψ_{v_1} resp. ψ_{v_2}.

Let G be the unknown distribution of V.

Then the density of U is



$$
\begin{aligned}
f(u_1, u_2) &= \int \psi_{v_1}(u_1)\psi_{v_2}(u_2)dG(v_1, v_2) \\
&\in \text{conv}(\Psi \otimes \Psi)
\end{aligned}
$$

Definition *Let $q \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $W > 0$. We say that $\{\mathcal{V}_k\}_{k \in \mathbb{N}_0}$ is a nested sequence of approximations for Ψ, with parameter (q, γ, W) if*

- $\mathcal{V}_k \supset \mathcal{V}_{k-1}$, $k \in \mathbb{N}$
- $\dim(\mathcal{V}_k) = q2^k$, $k \in \mathbb{N}_0$
- $\delta(\Psi, \mathcal{V}_k) \leq \gamma 2^{-k/W}$.

Theorem II *Suppose that $\{\mathcal{V}_k\}_{k\in\mathbb{N}_0}$ is a nested sequence of approximations for Ψ, with parameter (q,γ,W). Then*

$$M(\epsilon,\underbrace{\Psi\otimes\cdots\otimes\Psi}_{d\ \text{times}})\lesssim\epsilon^{-W}\log^{\frac{(d-1)(2+W)}{2}}(1/\epsilon).$$

$\square$

Corollary *By Theorem I, for*

$$\mathcal{F}=\mathrm{absconv}(\underbrace{\Psi\otimes\cdots\otimes\Psi}_{d\ \text{times}})$$

we have

$$\mathcal{H}(\epsilon,\mathcal{F})\lesssim\epsilon^{-\frac{2W}{2+W}}\log^{d-1}(1/\epsilon)\log^{\frac{W}{2+W}}(1/\epsilon).$$

Example: higher order Vitali total variation

[Friedman, 1991, multiplicative adaptive regression splines]

- $\mu :=$ Lebesgue measure on $[0, 1]$
- $Df(u_1, \ldots, u_d) := \prod_{j=1}^{d} \frac{\partial^q f(u_1, \ldots, u_d)}{\partial^q u_j}$,
- $\|Df\|_1 := \int |Df| d(\mu \times \cdots \times \mu)$ q-th order Vitali total variation.
- $\Psi := \{\Psi_v = (\cdot - v)_+^{q-1} : \; v \in [0, 1]\}$ truncated power basis
- $\mathcal{F} = \mathrm{absconv}(\Psi \otimes \cdots \otimes \Psi)$
- $\mathcal{N} := \{f : \; Df = 0\}$
- $\mathcal{F} := \{f : \; \|Df\|_1 \leq 1, \; f \perp \mathcal{N}\}$

Lemma $\exists \{\mathcal{V}_k\}_{k \in \mathbb{N}_0}$ *with parameter* (q, γ, W) *where* $\gamma = \sqrt{1/(2q-1)}$ *and* $W = 2/(2q-1)$. □

Corollary

$$\overset{\text{Theorems I\&II}}{\Longrightarrow}$$

$$\mathcal{H}(\epsilon, \mathcal{F}) \lesssim \epsilon^{-\frac{1}{q}} \log^{d-1}(1/\epsilon) \log^{\frac{1}{2q}}(1/\epsilon)$$

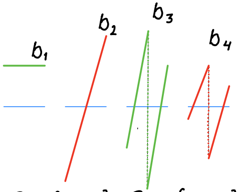
Multi-resolution analysis (a byproduct)

The proof of Theorem II uses an orthonormal basis
$P = (p_1, p_2, \ldots) \subset L_2(\mu)$ such that

$$\underbrace{(\underbrace{p_1, \ldots p_q}_{\text{basis } \mathcal{V}_0}, p_{q+1}, \ldots p_{q2^k}}_{\text{basis } \mathcal{V}_k}, \ldots)$$

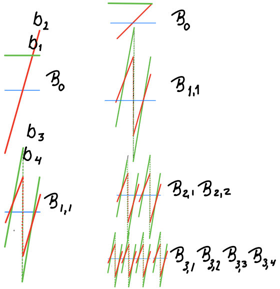
For the case of the truncated power basis, we can construct a multi-resolution basis P
consisting of piecewise polynomials of degree $q-1$

Multi-resolution basis consisting of piecewise linear functions (q=2)



$\mathcal{B}_0 = \{b_1, b_2\}$ $\mathcal{B}_{1,1} = \{b_3, b_4\}$

$\mathcal{B}_0^T \mathcal{B}_{1,1} = 0$



all blocks are orthogonal to each other

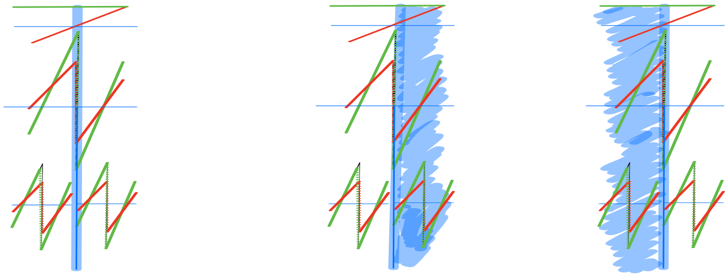


illustration of the orthogonality of the blocks

Statistical application

Let

$$Y = f^0 + \xi,\ \xi \sim \mathcal{N}(0,1)/\sqrt{n}$$

Regularized least squares estimator

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|_2^2 + 2\lambda \|\beta\|_1 \right\},$$

with $\lambda > 0$ a regularization parameter. Let $\hat{f} := X\hat{\beta}$.

In the next theorem we neglect all log-terms for transparency.

Theorem [SvdG, P. Hinz (2019)] *Suppose* $M(\epsilon, X) \lesssim \epsilon^{-W}$.
For all $f^* = X\beta^*$

$$\begin{aligned} \|\hat{f} - f^0\|_2^2 + \lambda\|\hat{\beta}\|_1 &\underset{}{\leq} \|f^* - f^0\|_2^2 + \text{Rem}, \\ \text{Rem} &\overset{\mathbb{P}}{\lesssim} \lambda\|\beta^*\|_1 + n^{-\frac{2+W}{2}}\lambda^{-W} \square \end{aligned}$$

Optimal tradeoff:

$$\lambda \asymp n^{-\frac{2+W}{2(1+W)}} \|\beta^*\|^{-\frac{1}{1+W}} \rightsquigarrow \text{Rem} \overset{\mathbb{P}}{\lesssim} n^{-\frac{2+W}{2(1+W)}} \|\beta^*\|_1^{\frac{W}{1+W}}$$

Special case

- $X := \Psi \otimes \cdots \otimes \Psi$
- $\Psi :=$ truncated power basis order q
- $f^* = f^0 = X\beta^0$
- $\|\beta^0\|_1 \asymp 1$

Then, for $\lambda \asymp n^{-\frac{2q}{2q+1}}$,

$$\|\hat{f} - f^0\|_2^2 + \lambda\|\hat{\beta}\|_1 \overset{\mathbb{P}}{\lesssim} n^{-\frac{2q}{2q+1}}$$

Is overparametrization (im)possible?

We have

$$\mathrm{Rem} = o_{\mathbb{P}}(1) \text{ for } \|\beta^*\|_1 \ll n^{\frac{2+W}{2W}}.$$

Note

$$\frac{2+W}{2W} < \frac{1}{2}$$

Dudley's entropy integral

Let $\mathbf{Q}_N$ be the projection on $\mathrm{span}(\mathbf{P}_N \cup f_0)$. It holds that

$$\mathbb{E} \sup_{f\in\mathcal{F}:\ \|f-f_0\|_2\leq\epsilon} |\xi^T(f-f_0)| \leq \mathbb{E} \sup_{f\in\mathcal{F}:\ \|f-f_0\|_2\leq\epsilon} |\xi^T \mathbf{Q}_N(f-f_0)| + \mathbb{E} \sup_{f\in\mathcal{F}} |(I-\mathbf{P}_N)f|$$

$$\lesssim \sqrt{N}\epsilon + \delta_N(X)$$

$$\lesssim \sqrt{N}\epsilon + N^{-1/W}.$$

Now choose $N \asymp \epsilon^{-\frac{2W}{2+W}}$ We get

$$\mathbb{E} \sup_{f\in\mathcal{F}:\ \|f-f_0\|_2\leq\epsilon} |\xi^T(f-f_0)| \lesssim \epsilon^{\frac{2}{2+W}}.$$

This coincides with Dudley's entropy integral

$$\int_0^\epsilon \sqrt{\mathcal{H}(u,\mathcal{F})}du \asymp \int_0^\epsilon u^{-\frac{W}{2+W}} du \asymp \epsilon^{\frac{2}{2+W}}$$

Summary

◦ we (re)derived entropy bounds for the (absolute) convex hull of a "small" set
◦ "small" in terms of covering numbers being polynomial ...
◦ and "smaller" in terms of approximation numbers (approximation by finite dimensional spaces)
◦ the bounds are tight up to log-terms
◦ main example: classes of functions with bounded higher order Vitali total variation
◦ a by-product is systems of multi-resolution basis functions

◦ statistical applications in e.g regression and density estimation (multiplicative adaptive regression splines, mixture models)
(application to one-hidden layer neural networks & Barron spaces)

Thank you!