Machine learning meets super-resolution

H. N. Mhaskar
Claremont Graduate University, Claremont.

Inverse Problems and Machine Learning
February 10, 2018

Goals

The problem of super-resolution is dual of the problem of machine learning, viewed as function approximation.

- How to measure the accuracy
- How to ensure lower bounds
- Common tools

Will illustrate on the (hyper-)sphere $\mathbb{S}^q$ of $\mathbb{R}^{q+1}$.

1. Machine learning

Machine learning on $\mathbb{S}^q$

Given data (training data) of the form $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^M$, where $\mathbf{x}_j \in \mathbb{S}^q$, $y_j \in \mathbb{R}$,

find a function $\mathbf{x} \mapsto \sum_{k=1}^N a_k G(\mathbf{x} \cdot \mathbf{z}_k)$

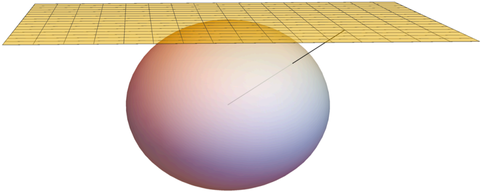
- that models the data well;
- in particular, $\sum_{k=1}^N a_k G(\mathbf{x}_j \cdot \mathbf{z}_k) \approx y_j$.

Tacit assumption: There exists an underlying function f such that $y_j = f(\mathbf{x}_j) +$ noise.

ReLU networks

An ReLU network is a function of form

$$\mathbf{x} \mapsto \sum_{k=1}^{N} a_k |\mathbf{w}_k \cdot \mathbf{x} + b_k|.$$



$$\mathbf{w}_k \cdot \mathbf{x} + b_k \rightsquigarrow \frac{(\mathbf{w}_k, b) \cdot (\mathbf{x}, 1)}{\sqrt{(|\mathbf{w}_k|^2 + 1)(|\mathbf{x}|^2 + 1)}}$$

Approximation on Euclidean space $\rightsquigarrow$ approximation on sphere

Notation on the sphere

$\mathbb{S}^q := \{\mathbf{x} = (x_1, \ldots, x_{q+1}) \ : \ \sum_{k=1}^{q+1} x_k^2 = 1\}$,
$\omega_q =$ Riemannian volume of $\mathbb{S}^q$
$\rho(\mathbf{x}, \mathbf{y}) =$ geodesic distance between $\mathbf{x}$ and $\mathbf{y}$.
$\Pi_n^q =$ class of all spherical polynomials of degree at most n.
$\mathbb{H}_\ell^q =$ class of all homogeneous harmonic polynomials of degree ℓ,
$d_\ell^q =$ the dimension of $\mathbb{H}_\ell^q$,
$\{Y_{\ell,k}\} =$ orthonormal basis for $\mathbb{H}_\ell^q$.
$\Delta =$ Negative Laplace-Beltrami operator.
$\Delta Y_{\ell,k} = \ell(\ell + q - 1) Y_{\ell,k} = \lambda_\ell^2 Y_{\ell,k}$.

Notation on the sphere

With $p_\ell = p_\ell^{(q/2-1,q/2-1)}$ (Jacobi polynomial),

$$\sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\mathbf{x})Y_{\ell,k}(\mathbf{y}) = \omega_{q-1}^{-1}p_\ell(1)p_\ell(\mathbf{x}\cdot\mathbf{y}).$$

If $G : [-1,1] \to \mathbb{R}$,

$$G(\mathbf{x}\cdot\mathbf{y}) = \sum_{\ell=0}^{\infty} \hat{G}(\ell) \sum_{k=1}^{d_\ell^q} Y_{\ell,k}(\mathbf{x})\overline{Y_{\ell,k}(\mathbf{y})}.$$

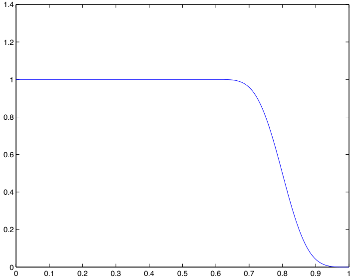
For a measure μ on $\mathbb{S}^q$,

$$\hat{\mu}(\ell,k) = \int_{\mathbb{S}^q} \overline{Y_{\ell,k}(\mathbf{y})}d\mu(\mathbf{y}).$$

Notation on the sphere

$$\Phi_n(t) = \omega_{q-1}^{-1} \sum_{\ell=0}^{n} h\left(\frac{\lambda_\ell}{n}\right) p_\ell(1) p_\ell(t).$$

$$\sigma_n(\mu)(\mathbf{x}) = \int_{\mathbb{S}^q} \Phi_n(\mathbf{x}\cdot\mathbf{y})d\mu(\mathbf{y}) = \sum_{\ell=0}^{n} h\left(\frac{\lambda_\ell}{n}\right) \sum_{k=1}^{d_\ell^q} \hat{\mu}(\ell,k) Y_{\ell,k}(\mathbf{x}).$$



Notation on the sphere

Localization

(Mh. 2004) If $S > q$ and h is sufficiently smooth,

$$|\Phi_n(\mathbf{x}\cdot\mathbf{y})| \le c(h,s)\frac{n^q}{\max(1,(n\rho(\mathbf{x}\cdot\mathbf{y}))^S)}$$

Polynomial approximation

(Mh. 2004)

$$E_n(f) = \min_{P \in \Pi_n^q} \|f - P\|_\infty.$$

$$W_r = \{f \in C(\mathbb{S}^q) : E_n(f) = \mathcal{O}(n^{-r})\}.$$

Theorem TFAE

1. $f \in W_r$
2. $\|f - \sigma_n(f)\| = \mathcal{O}(n^{-r})$
3. $\|\sigma_{2^n}(f) - \sigma_{2^{n-1}}(f)\| = \mathcal{O}(2^{-nr})$ (Littlewood-Paley type expansion)

Data-based approximation

For $\mathcal{C} = \{\mathbf{x}_j\} \subset \mathbb{S}^q$, $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^M$,

1. Find N and $w_j \in \mathbb{R}$ such that

$$\sum_{j=1}^M w_j P(\mathbf{x}_j) = \int_{\mathbb{S}^q} P(\mathbf{x}) d\mathbf{x}, \quad P \in \Pi_{2N}^q$$

   and

$$\sum_{j=1}^M |w_j P(\mathbf{x}_j)| \le c \int_{\mathbb{S}^q} |P(\mathbf{x})| d\mathbf{x}, \quad P \in \Pi_{2N}^q.$$

   Done by least squares or least residual solutions, to ensure a good condition number.

2.

$$S_N(\mathcal{D})(\mathbf{x}) = \sum_{j=1}^M w_j y_j \Phi_N(\mathbf{x} \cdot \mathbf{x}_j)$$

Data-based approximation

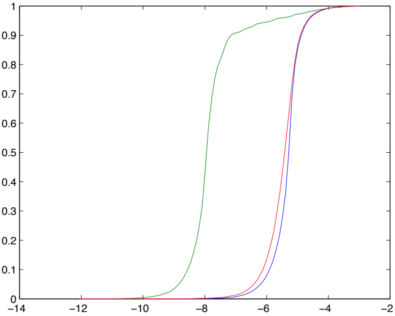
(Le Gia, Mh., 2008)
If $\{\mathbf{x}_j\}_{j=1}^M$ are chosen uniformly from μ_q, and $f \in W_r$, then with high probability,

$$\|f - S_N(\mathcal{D})\|_\infty \lesssim M^{-r/(2r+q)}.$$

If f is locally in W_r, then the results holds locally as well; i.e., accuracy in approximation adapts itself according to local smoothness.

Examples

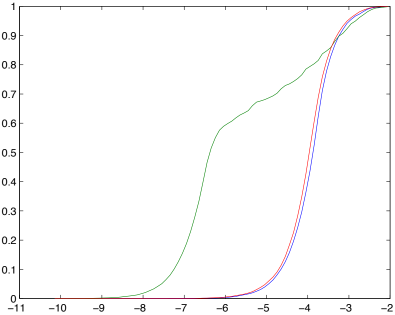
$$f(x, y, z) = [0.01 - (x^2 + y^2 + (z-1)^2)]_+ + \exp(x + y + z)$$



Percentages of error less than 10^x Least square, $\sigma_{63}(h_1)$, $\sigma_{63}(h_5)$.

Examples

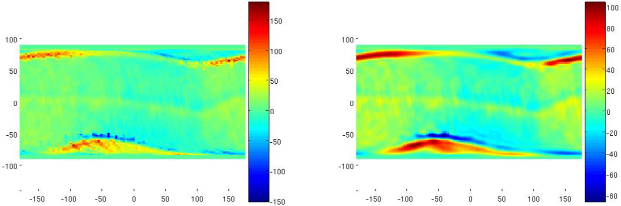
$$f(x,y,z) = (x-0.9)_+^{3/4} + (z-0.9)_+^{3/4}$$



Percentages of error less than 10^x Least square, $\sigma_{63}(h_1)$, $\sigma_{63}(h_5)$.

Examples

East–west component of earth's magnetic field



Original data on left (Courtesy Dr. Thorsten Maier),
reconstruction with $\sigma_{46}(h_7)$ on right

ZF networks

Let $\hat{G}(\ell) \backsim \ell^{-\beta}$, $\beta > q$, $\mathcal{C}_m$ a nested sequence of points with

$$\delta(\mathcal{C}_m) = \max_{\mathbf{x}\in\mathbb{S}^q} \min_{\mathbf{z}\in\mathcal{C}_m} \rho(\mathbf{x},\mathbf{z}) \sim \eta(\mathcal{C}_m) = \min_{\mathbf{z}_1\neq\mathbf{z}_2\in\mathcal{C}_m} \rho(\mathbf{z}_1,\mathbf{z}_2) \geq 1/m.$$

$\mathcal{G}(\mathcal{C}_m) = \mathrm{span}\{G(\circ\cdot\mathbf{z}) : \mathbf{z}\in\mathcal{C}_m\}$.

ZF networks

(Mh. 2010)

Theorem Let $0 < r < \beta - q$, then $f \in W_r$ if and only if

$$\mathrm{dist}(f, \mathcal{G}(\mathcal{C}_m)) = \mathcal{O}(m^{-r}),$$

Remark. The theorem gives lower limits for individual functions.

One problem

$\mathbf{x}_j$'s may not be distributed according to μ_q; their distribution is unknown.

Drusen classification

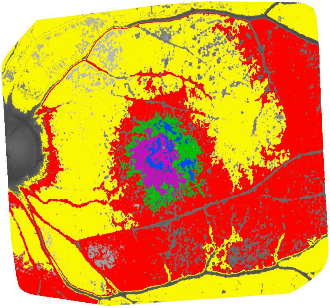
- AMD (Age related Macular Degeneration) is the most common cause of blindness among the elderly in the western world.
- AMD $\leftrightsquigarrow$ RPE (Retinal Pigment Epithelium) $\leftrightsquigarrow$ Drusen accumulation of different kinds

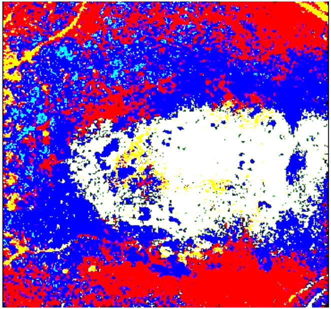
Problem: Automated quantitative prediction of disease progression, based on drusen classification.

Drusen classification

(Ehler, Filbir, Mh., 2012)
We used 24 images (400×400 pixels each) on each patient, at different frequencies. By preprocessing these images at each pixel, we obtained a data set consisting of 160,000 points on a sphere in a 5 dimensional Euclidean space. We used about 1600 of these as training set, and classified the drusen in 4 classes.
While the current practice is based on spatial appearance, our method is based on multi–spectral information.

Drusen classification





2. Super-resolution

Problem statement

Given observations of the form

$$\sum_{m=1}^{L} a_m \exp(-ijx_m) + \text{noise}, \qquad |j| \le N,$$

determine L, a_m's and x_m's.
Hidden periodicities (Lanczos)
Direction finding (Krim, Pillai, $\cdots$)
Singularity detection (Eckhoff, Gelb, Tadmor, Tanner, Mh., Prestin, Batenkov, $\cdots$)
Parameter estimation (Potts, Tasche, Filbir, Mh., Prestin, $\cdots$)
Blind source signal separation (Flandrin, Daubeschies, Wu, Chui, Mh., $\cdots$)

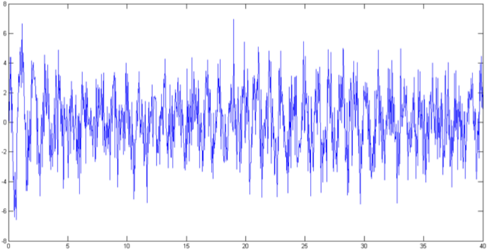
A simple observation

If Φ_N is a highly localized kernel (Mh.-Prestin, 1998), then $\sum_{m=1}^{L} a_m \Phi_N(x - x_m) \approx \sum_{m=1}^{L} a_m \delta_{x_m}$.

A simple observation

Original signal:

$$f(t) = \cos(2\pi t) + \cos(2\pi(0.96)t) + \cos(2\pi(0.92)t) + \cos(2\pi(0.9)t) + \text{noise}$$

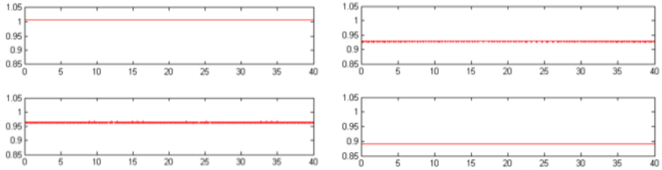


A simple observation

Original signal:

$$f(t) = \cos(2\pi t) + \cos(2\pi(0.96)t) + \cos(2\pi(0.92)t) + \cos(2\pi(0.9)t) + \text{noise}$$

Frequencies obtained by our method (Chui, Mh., van der Walt, 2015): .



Super-resolution

Question How large should N be?
Answer With $\eta = \min_{j \neq k} |x_j - x_k|$, $N \geq c\eta^{-1}$.
Super-resolution (Donoho, Candés, Fernandez-Granda) How can we do this problem with $N \ll \eta^{-1}$?

Spherical variant

Given

$$\sum_{m=1}^{L} a_m Y_{\ell,k}(\mathbf{x}_m) + noise, \quad k = 1, \cdots, d_\ell^q, \ 0 \leq \ell \leq N,$$

determine L, a_m, $\mathbf{x}_m$.

Observation

With $\mu^* = \sum_{m=1}^{L} a_m \delta_{\mathbf{x}_m}$,

$$\hat{\mu^*}(\ell, k) = \sum_{m=1}^{L} a_m Y_{\ell,k}(\mathbf{x}_m).$$

Super-duper-resolution

Given

$$\hat{\mu^*}(\ell, k) + \text{noise}, \quad k = 1, \cdots, d_\ell^q,\ \ell \leq N,$$

determine μ^*.

Remark The minimal separation is 0. Any solution based on finite amount of information is beyond super-resolution.

Duality

$$d\mu_N(\mathbf{x}) = \sigma_N(\mu^*)(\mathbf{x})d\mathbf{x} = \int_{\mathbb{S}^q} \Phi_N(\mathbf{x}\cdot\mathbf{y})d\mu^*(\mathbf{y})d\mathbf{x}.$$

For $f \in C(\mathbb{S}^q)$,

$$\int_{\mathbb{S}^q} f(\mathbf{x})d\mu_N(\mathbf{x}) = \int_{\mathbb{S}^q} \sigma_N(f)(\mathbf{x})d\mu^*(\mathbf{x}).$$

So,

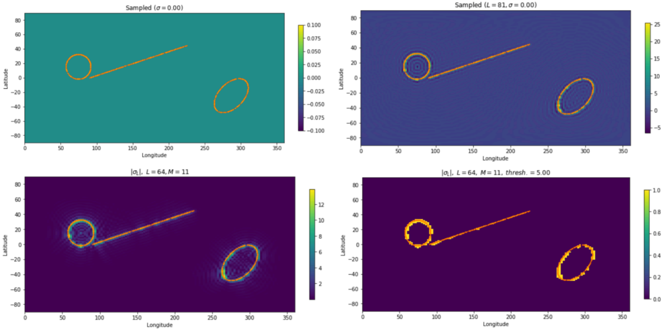
$$\left|\int_{\mathbb{S}^q} f(\mathbf{x})d(\mu_N - \mu^*)(\mathbf{x})\right| \le |\mu^*|_{TV} E_{N/2}(f).$$

Thus, $\mu_N \to \mu^*$ (weak-*). Also,

$$\int_{\mathbb{S}^q} P(\mathbf{x})d\mu_N(\mathbf{x}) = \int_{\mathbb{S}^q} P(\mathbf{x})d\mu^*(\mathbf{x}), \quad P \in \Pi^q_{N/2}.$$

Examples

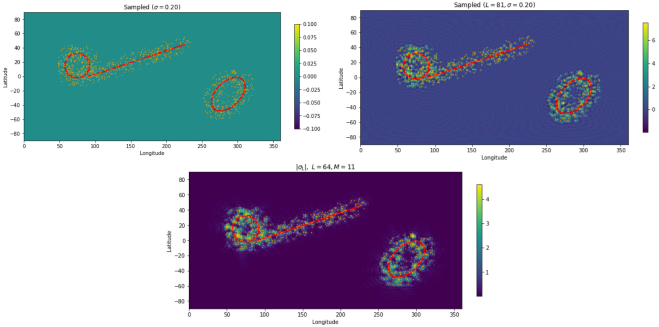
(Courtesy: D. Batenkov)



Original measure (left), Fourier projection (middle), σ_{64} (below left), thresholded $|\sigma_{64}|$ (below right).

Examples

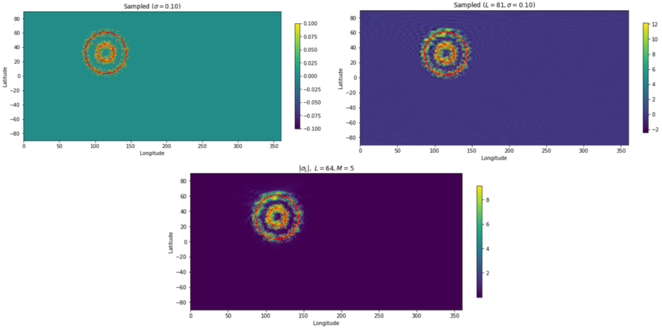
(Courtesy: D. Batenkov)



Original measure (left), Fourier projection (middle), σ_{64} (below).

Examples

(Courtesy: D. Batenkov)



Original measure (left), Fourier projection (middle), σ_{64} (below).

3. Distance between measures

Erdös-Turán discrepancy

Erdös, Turán, 1940 If ν is a signed measure on T,

$$(*) \qquad D[\nu] = \sup_{[a,b]\subset \mathrm{T}} |\nu([a,b])|.$$

Analogues of (*) hard for manifolds, even sphere.
Equivalently, if

$$G(x) = \sum_{k\in \mathbf{Z}\setminus\{0\}} \frac{e^{ikx}}{ik}$$

$$(**) \qquad D[\nu] = \sup_{x\in \mathrm{T}} \left| \int_{\mathrm{T}} G(x-y)d\nu(y) \right|$$

Generalization to multivariate case: Dick, Pillisheimer, 2010.

Wasserstein metric

$$\sup_f \left\{ \left| \int_{\mathbb{S}^q} f d\nu \right| : \max_{\mathbf{x},\mathbf{y}\in\mathbb{S}^q} |f(\mathbf{x}) - f(\mathbf{y})| \le 1 \right\}.$$

Replace $\max_{\mathbf{x},\mathbf{y}\in\mathbb{S}^q} |f(\mathbf{x}) - f(\mathbf{y})| \le 1$ by $\|\Delta(f)\| \le 1$.
Equivalent metric:

$$\left\| \int_{\mathbb{S}^q} G(\circ \cdot \mathbf{y}) d\nu(\mathbf{y}) \right\|_1,$$

where G is Green kernel for Δ.

Measuring weak-* convergence

Let $G : [-1, 1] \to \mathbb{R}$, $\hat{G}(\ell) > 0$ for all ℓ, $\hat{G}(\ell) \backsim \ell^{-\beta}$, $\beta > q$.

$$D_G[\nu] = \left\| \int_{\mathbb{S}^q} G(\circ \cdot \mathbf{y}) d\nu(\mathbf{y}) \right\|_1 .$$

Theorem

$$D_G[\mu_N - \mu^*] \le cN^{-\beta} |\mu^*|_{TV}.$$

Remark The approximating measure is constructed from $\mathcal{O}(N^q)$ pieces of information $\hat{\mu^*}(\ell, k)$. In terms of the amount of information, M, the rate is $\mathcal{O}(M^{-\beta/q})$.

Widths

Let $\mathcal{M}$= set of all Borel measures on $\mathbb{S}^q$ having bounded variation,

$$\mathcal{K} = \{\nu \in \mathcal{M} : |\nu|_{TV} \le 1\}.$$

$$\mathcal{S} = \{S : \mathcal{K} \to \mathbb{R}^M, \text{weak-* continuous}\},$$

For $A : \mathbb{R}^M \to \mathcal{M}$, $S \in \mathcal{S}$,

$$\mathrm{Err}_M(A, S) = \sup_{\mu \in \mathcal{K}} D_G[A(S(\mu)) - \mu].$$

(width)

$$d_M(\mathcal{K}) = \inf_{A,S} \mathrm{Err}_M(A, S) \ge cM^{-\beta/q}.$$

Under the hood

(Mh. 2010)

$$\left\| G(\circ, \mathbf{y}) - \int_{\mathbb{S}^q} G(\mathbf{z}, \mathbf{y}) \Phi_N(\circ \cdot \mathbf{z}) d\mathbf{z} \right\|_1 \le cN^{-\beta}.$$

For function approximaton:

$$\sigma_N(f) \rightsquigarrow \text{Estimate on } \mathrm{dist}(f, \mathcal{G}(\mathcal{C}_m)).$$

For super-duper-resolution: Estimate on $D_G[\mu_N - \mu^*]$.

Under the hood

(Mh. 2010) If $F(\mathbf{x}) = \sum_{k=1}^{L} a_k G(\mathbf{x}\cdot\mathbf{z}_k)$,

$$\eta = \min_{1\le k\neq j\le L} \rho(z_k, z_\ell),$$

then

$$\sum_{k=1}^{L} |a_k| \le c\eta^{-\beta}\|F\|_1.$$

For function approximation: Converse theorem for ZF approximation.
For super-duper-resolution: Estimate on the widths.

Thank you.