Approximation Theory

Ben Recht
Center for the Mathematics of Information
Caltech

April 7, 2008

References

- The majority of this material is adapted from F. Girosi's 9.520 lecture from 2003.
  - Available on OCW
  - Very readable with an extensive bibliography

- Random Features
  - Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines." NIPS 2007
  - Ali Rahimi and Benjamin Recht. "On the power of randomized shallow belief networks." In preparation, 2008.

Outline

- Decomposition of the generalization error
- Approximation and rates of convergence
- “Dimension Independent” convergence rates
- Maurey-Barron-Jones approximations
- Random Features

Notation

$$R[f] = \int_{X \times Y} V(f(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy$$

$$R_{\mathrm{emp}}[f] = \frac{1}{L} \sum_{i=1}^{L} V(f(\mathbf{x}_i), y_i)$$

$$f_0 = \arg\min_{f \in \mathcal{T}} R[f]$$

$R[f_0] =$ how well we can do

$$f_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} R[f]$$

$R[f_{\mathcal{H}}] =$ how well we can do in $\mathcal{H}$

$$\hat{f}_{\mathcal{H},L} = \arg\min_{f \in \mathcal{H}} R_{\mathrm{emp}}[f]$$

$R[\hat{f}_{\mathcal{H},\mathcal{L}}] =$ how well we can do in $\mathcal{H}$ with our L observations

$$R[\hat{f}_{\mathcal{H},\mathcal{L}}] - R[f_0] = R[\hat{f}_{\mathcal{H},\mathcal{L}}] - R[f_{\mathcal{H}}] + R[f_{\mathcal{H}}] - R[f_0]$$

Generalization Error Estimation Error Approximation Error

For least squares cost $V(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$

$$R[f] = \|f - f_0\|_2^2 + R[f_0]$$

$$R[\hat{f}_{\mathcal{H},\mathcal{L}}] - R[f_0] = R[\hat{f}_{\mathcal{H},\mathcal{L}}] - R[f_{\mathcal{H}}] + \|f_{\mathcal{H}} - f_0\|_2^2$$

Estimation Error Approximation Error

Independent of *target space* (statistics)

Independent of *examples* (analysis)

Judiciously select $\mathcal{H}$ to balance the tradeoff

- Nested hypothesis spaces

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \cdots \subset \mathcal{H}_n \subset \cdots$$

- Error

$$\epsilon_n = \inf_{f \in \mathcal{H}_n} \|f - f_0\|_2$$

For most families of hypothesis spaces we encounter

$$\lim_{n \to \infty} \epsilon_n = 0$$

- How fast does this error go to zero? We are interested in bounds of the form

$$\epsilon_n \leq c n^{-\alpha}$$

Example Hypothesis Spaces

- Polynomials on [0,1]. $\mathcal{H}_n$ is the set of all polynomials with degree at most *n*

$$\mathcal{H}_n = \operatorname{span}\{1, x, x^2, x^3, \ldots, x^n\}$$

  We can approximate any smooth function with a polynomial (Taylor series).

- Sines and cosines on [-π,π].

$$\mathcal{H}_n = \operatorname{span}\{1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots, \cos(nx), \sin(nx)\}$$

  We can approximate any square integrable function with a Fourier series.

Calculating approximation rates

$$C_2[-\pi,\pi] = C_0[-\pi,\pi] \bigcap L_2[-\pi,\pi]$$

- Functions in this class can be represented by

$$f(x) = \sum_{k=0}^{\infty} c_k e^{ikx} \qquad\qquad c_k \propto \int_{-\pi}^{\pi} f(x) e^{-ikx} dx$$

- Parseval:

$$\|f(x)\|_2^2 = \sum_{k=0}^{\infty} |c_k|^2$$

Target Space

- Sobolev space of smooth functions

$$W_{s,2} \equiv \left\{ f \in C_2[-\pi,\pi] \;\middle|\; \left\| \frac{d^s f}{dx^s} \right\|_2 < \infty \right\}$$

- Using parseval:

$$\|f\|_s^2 \equiv \left\| \frac{d^s f}{dx^s} \right\|_2^2 = \sum_{k=1}^{\infty} k^{2s} c_k^2$$

Hypothesis Space

- $\mathcal{H}_n$ is the set of trig functions of degree n

$$p(x) = \sum_{k=1}^{n} a_k e^{ikx}$$

- If f is of the form

$$f(x) = \sum_{k=1}^{\infty} c_k e^{ikx}$$

Best approximation in L_2 norm by $\mathcal{H}_n$ is given by

$$f_n(x) = \sum_{k=1}^{n} c_k e^{ikx}$$

Approximation Rate

- Note that $\mathcal{H}_n$ has n parameters. How fast does ϵ_n go to zero?

$$\epsilon_n[f]^2 \equiv \|f - f_n\|_2^2 = \sum_{k=n+1}^{\infty} |c_k|^2 = \sum_{k=n+1}^{\infty} \frac{k^{2s}}{k^{2s}} |c_k|^2$$

$$< \frac{1}{n^{2s}} \sum_{k=n+1}^{\infty} k^{2s} |c_k|^2 < \frac{1}{n^{2s}} \sum_{k=1}^{\infty} k^{2s} |c_k|^2 = \frac{\|f\|_s^2}{n^{2s}}$$

- More smoothness, faster convergence

$$\epsilon_n[f] < c[f] n^{-s}$$

- What happens in higher dimension?

$$C_2[-\pi,\pi]^d = C_0[-\pi,\pi]^d \bigcap L_2[-\pi,\pi]^d$$

- Functions can be written

$$f(\mathbf{x}) = \sum_{\mathbf{w}\in\mathbb{Z}_+^d} c_{\mathbf{w}} e^{i\mathbf{w}^*\mathbf{x}}$$

- Target space

$$W_{s,2} \equiv \left\{ f \in C_2[-\pi,\pi]^d \;\middle|\; \|f\|_s < \infty \right\}$$

- Again by Parseval

$$\|f\|_s^2 \equiv \left\| \frac{d^s f}{dx_1^s} \right\|_2^2 + \cdots + \left\| \frac{d^s f}{dx_d^s} \right\|_2^2 = \sum_{\mathbf{w}\in\mathbb{Z}_+^d} \left(\sum_{a=1}^{d} w_a^{2s} \right) |c_{\mathbf{w}}|^2$$

- Hypothesis Space. $\mathcal{H}_t$

$$p(\mathbf{x}) = \sum_{\substack{\mathbf{w}\in\mathbb{Z}_+^d \\ 0\le w_a\le t}} a_{\mathbf{w}} e^{i\mathbf{w}^*\mathbf{x}}$$

- Number of parameters in $\mathcal{H}_t$ is $n = t^d$. Best approximation to f is given by

$$f_t(\mathbf{x}) = \sum_{\substack{\mathbf{w}\in\mathbb{Z}_+^d \\ 0\le w_a\le t}} c_{\mathbf{w}} e^{i\mathbf{w}^*\mathbf{x}}$$

- How fast does ϵ_t go to zero? We do the calculation for d=2:

$$\epsilon_t[f]^2 \equiv \|f - f_t\|_2^2 = \sum_{k,\ell=t+1}^{\infty} |c_{k\ell}|^2 + \sum_{k=1}^{t} \sum_{\ell=t+1}^{\infty} |c_{k\ell}|^2 + \sum_{k=t+1}^{\infty} \sum_{\ell=1}^{t} |c_{k\ell}|^2$$

$$= \sum_{(k,\ell)\in\mathcal{I}}^{\infty} \frac{k^{2s}+\ell^{2s}}{k^{2s}+\ell^{2s}} |c_{k\ell}|^2$$

$$< \frac{1}{t^{2s}} \sum_{k,\ell=t+1}^{\infty} (k^{2s}+\ell^{2s}) |c_{k\ell}|^2$$

$$< \frac{1}{t^{2s}} \sum_{k,\ell=1}^{\infty} (k^{2s}+\ell^{2s}) |c_{k\ell}|^2 = \frac{\|f\|_s^2}{t^{2s}}$$

- Now the approximation scales as (as a function of *n*):

$$\boxed{\epsilon_n[f] < c[f] n^{-\frac{s}{d}}}$$

Curse of dimensionality

Blessing of smoothness

$$\epsilon_n[f] < c[f]n^{-\frac{s}{d}}$$

Curse of dimensionality

- Provides an estimate for the number of parameters

$$n \propto \left(\frac{1}{\epsilon}\right)^{\frac{d}{s}}$$

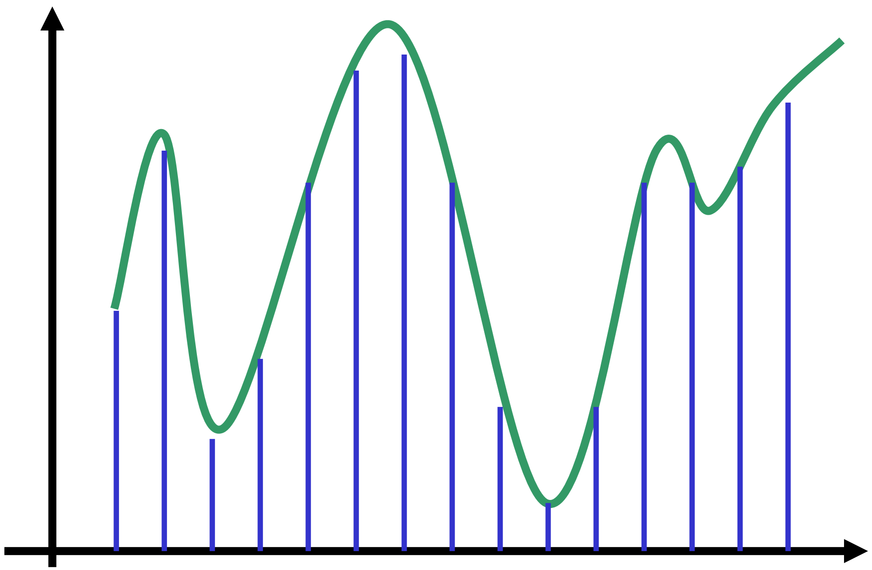
- Is this upper bound very loose?

Hard Limits

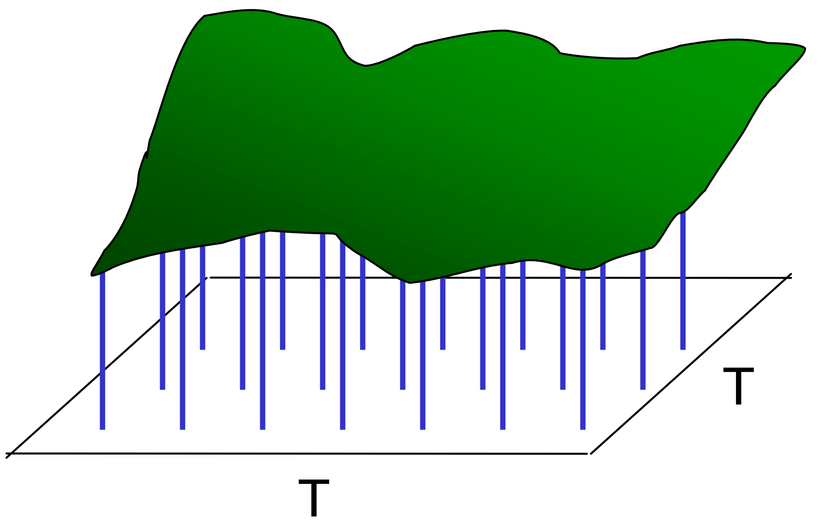
- Tommi Poggio: just remember Nyquist....

Sample rate = 2 x max freq

Num samples = 2 x T x max freq



In dimension d: Num samples = $(2 \times T \times \text{max freq})^d$



N-widths

- Let $\mathcal{X}$ be a normed space of functions. Let $\mathcal{A}$ be a subset of $\mathcal{X}$. We want to approximate $\mathcal{A}$ with a linear combination of a finite set of "basis functions" $\mathcal{X}$.
- Kolmogorov N-widths let us quantify how well we could do over all choices of finite sets of basis functions.

$$d_n(\mathcal{A}, \mathcal{X}) = \inf_{\phi_1, \ldots, \phi_n \in \mathcal{X}} \quad \sup_{f \in \mathcal{A}} \quad \inf_{c_1, \ldots, c_n} \left\| f - \sum_{k=1}^{n} c_k \phi_k \right\|_{\mathcal{X}}$$

The *n-width* of $\mathcal{A}$ in $\mathcal{X}$

Multivariate Example

$$\mathcal{X} = L_2([0,1]^d)$$

$$W_{s,2} = \{f \ : \ \|f\|_s \leq \infty\}$$

s times differentiable
*s*th derivative in L_2

$$\mathcal{A} = \{f \in W_{s,2} \ : \ \|f\|_s \leq 1\}$$

- Theorem (from Pinkus 1980):

$$d_n(\mathcal{A}, \mathcal{X}) \approx \left(\frac{1}{n}\right)^{\frac{s}{d}}$$

This rate is achieved by splines

“Dimension Free” convergence

- Consider networks of the form

$$f_n(\mathbf{x}) = \sum_{k=1}^{n} c_k \phi_k(\mathbf{x}; \omega_k)$$

- “Shallow” networks with parametric basis functions

$$\phi_k(\mathbf{x}; \omega)$$

- Characterize when we can get good approximations

$$\inf_{\omega_1,\ldots,\omega_k} \inf_{c_1,\ldots,c_n} \left\| f - \sum_{k=1}^{n} c_k \phi_k(\cdot; \omega_k) \right\|$$

Maurey-Barron-Jones Lemma

- **Theorem:** If f is in the convex hull of a set G in a Hilbert Space with $||g||_2 \leq b$ for all $g \in G$, then for every $n \geq 1$ and every $c' > b^2 - ||f||_2^2$, there is an f_n in the convex hull of n points in G such that

$$\|f - f_n\|_2^2 \leq \frac{c'}{n}$$

- Also known as Maurey's "empirical method"
- Many uses in computing covering numbers (see, e.g., generalization bounds, random matrices, compressive sampling, etc.)

Maurey-type Approximation Schemes

Define $\tilde{f}(\omega) = \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\omega^* \mathbf{x}} dx$

- Jones (1992)

$$\tilde{f} \in L_1(\mathbb{R}^d) \quad \int_{\mathbb{R}^d} |\tilde{f}(\omega)| d\omega < \infty \qquad f_n = \sum_{k=1}^{n} c_k \cos(\mathbf{w}_k^* \mathbf{x} + b_k)$$

- Barron (1993)

$$\nabla \tilde{f} \in L_1(\mathbb{R}^d) \quad \int_{\mathbb{R}^d} \|\omega\| |\tilde{f}(\omega)| d\omega < \infty \qquad f_n = \sum_{k=1}^{n} c_k \sigma(\mathbf{w}_k^* \mathbf{x} + b_k)$$

- Girosi & Anzellotti (1995)

$$\tilde{f} \in W_{2,s}(\mathbb{R}^d) \quad \text{with } 2s > d \qquad f_n = \sum_{k=1}^{n} c_k \exp(-\|\mathbf{x} - \mathbf{x}_k\|^2)$$

- Using nearly identical analysis, all of these schemes achieve

$$\epsilon_n = O\left(\frac{1}{\sqrt{n}}\right)$$

Hidden Smoothness

- Barron hides the smoothness via the functional

$$\int_{\mathbb{R}^d} \|\omega\| |\tilde{f}(\omega)| d\omega < \infty$$

- Girosi and Anzellotti show that this means

$$f = \frac{1}{\|\mathbf{x}\|^{d-1}} * g \qquad \text{for some} \qquad g \in L_1$$

- Note: functions get smoother as d increases

Algorithmic difficulty

- Training these networks is hard

$$\text{minimize}_{\theta_k, c_k} \left\| f - \sum_{k=1}^{n} c_k \phi(\cdot; \theta_k) \right\|$$

- But for fixed θ_k, the following is almost always trivial:

$$\text{minimize}_{c_k} \left\| f - \sum_{k=1}^{n} c_k \phi(\cdot; \theta_k) \right\|$$

- How to avoid optimizing the θ_k?

Random Features

- What happens if we pick θ_k at random and then optimize the weights?

$$\text{minimize}_{c_k} \left\| f - \sum_{k=1}^{n} c_k \phi(\cdot; \theta_k) \right\|$$

- It turns out, with some *a priori* information about the frequency content of *f*, we can do just as well as the classical approximation results of Maurey and co.

- Fix parameterized basis functions $\phi(\cdot;\omega)$
- Fix a probability distribution $p(\omega)$
- Our target space will be:

$$\mathcal{F}_p \equiv \left\{ f = \int \alpha(\omega)\phi(\cdot;\omega)d\omega \;\middle|\; \sup_\omega \left|\frac{\alpha(\omega)}{p(\omega)}\right| < \infty \right\}$$

- With the convention that

$$\left|\frac{\alpha(\omega)}{0}\right| = \begin{cases} 0 & \alpha(\omega) = 0 \\ \infty & \text{otherwise} \end{cases}$$

Random Features: Example

- Fourier basis functions: $\phi(\mathbf{x}; \omega, b) = \cos(\omega^* \mathbf{x} + b)$

- Gaussian parameters $\omega \sim \mathcal{N}(0, \sigma^2 I)$ $b \sim \mathrm{unif}([0, 2\pi])$

- If $\tilde{f}(\omega) = \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\omega^* \mathbf{x}} dx$, then $\sup_\omega \left| \frac{\tilde{f}(\omega)}{p(\omega)} \right| < \infty$ means

that the frequency distribution of *f* has subgaussian tails.

$$\mathcal{F}_p \equiv \left\{ f = \int \alpha(\omega)\phi(\cdot;\omega)d\omega \;\middle|\; \sup_\omega \left|\frac{\alpha(\omega)}{p(\omega)}\right| \leq \infty \right\}$$

- **Theorem:** Let f be in $\mathcal{F}_p$ with

$$\sup_\omega \left|\frac{\alpha(\omega)}{p(\omega)}\right| \leq C$$

Let $\omega_1,\ldots,\omega_n$ be sampled iid from *p*. Then

$$\min_{c_k} \left\| f - \sum_{k=1}^{n} c_k \varphi(\mathbf{x};\omega_k) \right\|_2 \leq \left(1 + \tfrac{1}{2}\log(\tfrac{1}{\delta})\right) \frac{\sqrt{2}C}{\sqrt{n}}$$

with probability at least $1 - \delta$.

Generalization Error

$$R[\hat{f}_{\mathcal{H},\mathcal{L}}] - R[f_0] = \underbrace{R[\hat{f}_{\mathcal{H},\mathcal{L}}] - R[f_{\mathcal{H}}]}_{\text{Estimation Error}} + \underbrace{\|f_{\mathcal{H}} - f_0\|_2^2}_{\text{Approximation Error}}$$

$$< \frac{c_1 n}{L} + \frac{c_2}{n}$$

- It's a finite sized basis set!
- Choosing $n = O(\sqrt{L})$ gives overall convergence of $O(\frac{1}{\sqrt{L}})$

Kernels

$$k(\mathbf{x}, \mathbf{y}) = \int p(\omega)\phi(\mathbf{x}; \omega)\phi(\mathbf{y}; \omega) d\omega$$

- Note that under the mapping

$$\mathbf{x} \mapsto \xi(\mathbf{x}) \equiv \left[\frac{1}{\sqrt{D}}\phi(\mathbf{x}; \omega_k)\right]_{1 \leq k \leq D}$$

  we have

$$\langle \xi(\mathbf{x}), \xi(\mathbf{y}) \rangle \approx k(\mathbf{x}, \mathbf{y})$$

- *Ridge regression with random features approximates Tikhonov regularized least-squares on an RKHS*

Random Features for Classification

Dataset	Fourier+LS	Binning+LS	CVM	Exact SVM
CPU regression 6500 instances 21 dims	3.6% 20 secs $D = 300$	5.3% 3 mins $P = 350$	5.5% 51 secs	11% 31 secs ASVM
Census regression 18,000 instances 119 dims	5% 36 secs $D = 500$	7.5% 19 mins $P = 30$	8.8% 7.5 mins	9% 13 mins SVMTorch
Adult classification 32,000 instances 123 dims	14.9% 9 secs $D = 500$	15.3% 1.5 mins $P = 30$	14.8% 73 mins	15.1% 7 mins $\text{SVM}^{\text{light}}$
Forest Cover classification 522,000 instances 54 dims	11.6% 71 mins $D = 5000$	2.2% 25 mins $P = 50$	2.3% 7.5 hrs	2.2% 44 hrs libSVM
KDDCUP99 (see footnote) classification 4,900,000 instances 127 dims	7.3% 1.5 min $D = 50$	7.3% 35 mins $P = 10$	6.2% (18%) 1.4 secs (20 secs)	8.3% < 1 s SVM+sampling

Gaussian RKHS vs Random Features

- **Random Features are good:** when L is sufficiently large and the function is sufficiently smooth

- **TR on RKHS is good:** when L is small or the function is not so smooth

```
% Approximates Gaussian Process regression
%   with Gaussian kernel of variance gamma
% lambda: regularization parameter
% dataset: X is dxN, y is 1xN
% test: xtest is dx1
% D: dimensionality of random feature

% training
  w = randn(D, size(X,1));
  b = 2*pi*rand(D,1);
  Z = cos(sqrt(gamma)*w*X + repmat(b,1,size(X,2)));
% Equivalent to
%      alpha = (lambda*eye(size(X,2)+Z*Z')\(Z*y);
  alpha = symmlq(@(v)(lambda*v(:) + *(Z'*v(:))),…
             Z*y(:),1e-6,2000);


% testing
  ztest = alpha(:)’*cos( sqrt(gamma)*w*xtest(:) + …
             + repmat(b,1,size(X,2)) );
```