High-Dimensional Neural Networks: Statistical and Computational Properties

Andrew R. Barron

YALE UNIVERSITY
DEPARTMENT OF STATISTICS

Presentation, 27 September, 2016

International Conference on Operational Research

KOI 2016, Osijek, Croatia

Joint work with Jason Klusowski

Outline

- Flexible high-dimensional function estimation using sinusoid, sigmoid, ramp or polynomial activation functions
- Approximation and estimation bounds to reveal the effect of dimension, model size, and sample size
- Computation with greedy term selection
  - Adaptive Annealing Method (for general designs) to guide parameter search
  - Nonlinear Power Method (for specific designs) to improve upon tensor methods for parameter estimation

Example papers for some of what is to follow

Papers illustrating my background addressing these questions of high dimensionality (available from www.stat.yale.edu)

- A. R. Barron, R. L. Barron (1988). Statistical learning networks: a unifying view. *Computing Science & Statistics: Proc. 20th Symp on the Interface*, ASA, p.192-203.
- A. R. Barron (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, Vol.39, p.930-944.
- A. R. Barron, A. Cohen, W. Dahmen, R. DeVore (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*, Vol.36, p.64-94.
- J. M. Klusowski, A. R. Barron (2016) Risk bounds for high-dimensional ridge function combinations including neural networks, Submitted to the *Annals of Statistics*. arXiv:1607.01434v1

Data Setting

- Data: $(\underline{X}_i, Y_i)$, $i = 1, 2, \ldots, n$
- Inputs: explanatory variable vectors (arbitrary dependence)

$$\underline{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})$$

- Domain: Cube in R^d
- Random design: independent $\underline{X}_i \sim P$
- Output: response variable Y_i in R
  - Bounded or subgaussian
- Relationship: $E[Y_i|\underline{X}_i] = f(\underline{X}_i)$ as in:
  - Perfect observation: $Y_i = f(\underline{X}_i)$
  - Noisy observation: $Y_i = f(\underline{X}_i) + \epsilon_i$ with ϵ_i indep
- Function: $f(\underline{x})$ unknown

Activation functions

Activation functions denoted $\phi(z)$ or $g(z)$

Piecewise constant: $1_{\{z-b\geq 0\}}$ or $\text{sgn}(z-b)$

Sigmoid: $(e^z - e^{-z})/(e^z + e^{-z})$

Linear spline, ramp: $(z-b)_+$

Sinusoidal: $\cos(2\pi f z)$, $\sin(2\pi f z)$

Polynomial: standard z^ℓ, Hermite $H_\ell(z)$

Multivariate Activation functions

- built from products or from ridge forms: $\phi(\underline{a}^T \underline{x})$
- constructed using univariate function ϕ
- internal parameter vector $\underline{a}$ of dimension d.

Product and Ridge Bases

- Product bases: for polynomials, sinusoids, splines
  using continuous powers, frequencies or thresholds

$$\phi(\underline{x}, \underline{a}) = \phi(x_1, a_1)\, \phi(x_2, a_2) \cdots \phi(x_d, a_d)$$

- Ridge bases: for projection pursuit regression, sinusoids, neural networks and polynomial networks:

$$\phi(\underline{x}, \underline{a}) = \phi(\underline{a}^T \underline{x}) = \phi(a_1 x_1 + a_2 x_2 + \ldots + a_d x_d)$$

Flexible multivariate function modeling

Internally parameterized models & nonlinear least squares

- Functions $f_m(\underline{x}) = \sum_{j=1}^{m} c_k \phi(\underline{x}, \underline{a}_k)$ in the span of a parameterized dictionary $\Phi = \{\phi_{\underline{a}}(\cdot) = \phi(\cdot, \underline{a}) : \underline{a} \in \mathcal{A}\}$ with parameter set $\mathcal{A} \subset R^d$
- Flexible function approximation
- Statistically challenging
- Computationally challenging

Notation

- Response vector: $Y = (Y_i)_{i=1}^n$ in R^n
- Dictionary vectors: $\Phi_{(n)} = \{(\phi(\underline{X}_i, \underline{a}))_{i=1}^n \; : \; \underline{a} \in \mathcal{A}\} \subset R^n$
- Sample squared norm: $\|f\|_{(n)}^2 = \frac{1}{n}\sum_{i=1}^n f^2(\underline{X}_i)$
- Population squared norm: $\|f\|^2 = \int f^2(\underline{x})P(d\underline{x})$
- Normalized dictionary condition: $\|\phi\| \leq 1$ for $\phi \in \Phi$

Flexible $m-$term nonlinear optimization

Impractical complete nonlinear least squares optimization

- Sample version

$$\hat{f}_m \quad \text{achieves} \quad \min_{(\underline{a}_j, c_j)_{j=1}^m} \| Y - \sum_{j=1}^m c_j \, \phi_{\underline{a}_j} \|_{(n)}^2$$

- Population version

$$f_m \quad \text{achieves} \quad \min_{(\underline{a}_j, c_j)_{j=1}^m} \| f - \sum_{j=1}^m c_j \, \phi_{\underline{a}_j} \|^2$$

- Optimization of $(\underline{a}_j, c_j)_{j=1}^m$ in $R^{(d+1)m}$.

GREEDY OPTIMIZATIONS

Optimize one term at a time

- Step 1: Choose $\underline{a}_1$, c_1 to produce a single best term:
  - sample version: $\min_{\underline{a},c} \|Y - c\phi_{\underline{a}}\|^2_{(n)}$
  - population version: $\min_{\underline{a},c} \|f - c\phi_{\underline{a}}\|^2$
- Step $m > 1$: Arrange

$$\hat{f}_m = \alpha\, \hat{f}_{m-1} + c\, \phi_{\underline{a}}$$

with $\underline{a}_m$, c_m, α_m providing the term with best improvement:

  - sample version: $\min_{\underline{a},c,\alpha} \|Y - \alpha\hat{f}_{m-1} - c\phi_{\underline{a}}\|^2_{(n)}$
  - population version: $\min_{\underline{a},c,\alpha} \|f - \alpha f_{m-1} - c\phi_{\underline{a}}\|^2$

Acceptable variants of greedy optimization

- Step $m > 1$: Arrange

$$\hat{f}_m = \alpha \, \hat{f}_{m-1} + c \, \phi_{\underline{a}}$$

with $\underline{a}_m$, c_m, α_m providing the term with best improvement:

  - least squares: $\min_{\underline{a},c,\alpha} \| Y - \alpha \hat{f}_{m-1} - c\phi_{\underline{a}} \|^2_{(n)}$

- Acceptable variants

  - Inner-product maximization: With $Res_i = Y_i - \hat{f}_{m-1}(\underline{X}_i)$ choose $\underline{a}_m$ to achieve $\max_{\underline{a}} \sum_{i=1}^{n} Res_i \, \phi(\underline{X}_i, \underline{a})$ to within a constant factor

  - Foward stepwise selection: Given $\underline{a}_1, \ldots, \underline{a}_{m-1}$, choose $\underline{a}_m$ to obtain $\min_{\underline{a}} d(Y, span\{\phi_{\underline{\theta}_1} \ldots \phi_{\underline{a}_{m-1}}, \phi_{\underline{a}}\})$

  - Orthogonal matching pursuit: Project onto span after choosing inner-product maximizer.

  - ℓ_1 Penalization: Update $v_m = \alpha_m v_{m-1} + c_m$. Choose α_m, c_m to achieve $\min_{\alpha,c} \| Y - \alpha \hat{f}_{m-1} - c\phi_{\underline{a}} \|^2_{(n)} + \lambda[\alpha v_{m-1} + c]$

Basic $m-$term approximation and computation bounds

For complete or greedy approximation (B. 1993, Lee et al 1995)

- Population version:

$$\|f - f_m\| \le \frac{\|f\|_\Phi}{\sqrt{m}}$$

  and moreover

$$\|f - f_m\|^2 \le \inf_g \left\{ \|f - g\|^2 + \frac{2\|g\|_\Phi^2}{m} \right\}$$

- Sample version:

$$\|Y - \hat{f}_m\|_{(n)}^2 \le \|Y - f\|_{(n)}^2 + \frac{2\|f\|_\Phi^2}{m}$$

- Optimization with ℓ_1 penalization:

$$\|Y - \hat{f}_m\|_{(n)}^2 + \lambda\|f_m\|_\Phi \le \inf_g \left\{ \|Y - f\|_{(n)}^2 + \lambda\|g\|_\Phi + \frac{2\|g\|_\Phi^2}{m} \right\}$$

- Here $\|g\|_\Phi$ is the minimal ℓ_1 norm of coefficients of g in span of Φ

The norm of f with respect to the dictionary Φ

- Minimal ℓ_1 norm on coefficients in approximation of f

$$\|f\|_\Phi = \lim_{\epsilon \to 0} \inf \Big\{ \sum_j |c_j| \;:\; \| \sum_j c_j \phi_{\underline{a}_j} - f \| \le \epsilon \Big\}$$

called the variation of f with respect to Φ (B. 1991)

$$\|f\|_\Phi = V_\Phi(f) = \inf\{ V \;:\; f/V \in closure(conv(\pm\Phi)) \}$$

- It appears in the bound $\|f - f_m\| \le \frac{\|f\|_\Phi}{\sqrt{m}}$
- Later also called the atomic norm of f with respect to Φ
- In the case of the signum activation function it matches the minimal range of neural nets arbitrarily well approximating f

Greedy proof of the approximation bound:

- Consider the case $\|f\|_\Phi = 1$
- Take Φ to be closed under sign changes.
- The $\min_\phi$ is not more than ave_ϕ
- Take average with respect to the weights representing f

$$\begin{aligned}\|f - f_m\|^2 &\le \min_\phi \|f - (1-\lambda) f_{m-1} - \lambda\phi\|^2 \\ &\le \mathrm{ave}_\phi \|f - (1-\lambda) f_{m-1} - \lambda\phi\|^2 \\ &= (1-\lambda)^2 \|f - f_{m-1}\|^2 + \lambda^2\end{aligned}$$

- Bound follows by induction with $\lambda = 1/m$

$$\|f - f_m\|^2 \le \frac{1}{m}$$

- Jones (AS 1992), B. (IT 1993)
- extensions: Lee et al (IT 1995), DeVore et al (AS 2008)

Relating the variation to a spectral norm of f

- Neural net approximation using the Fourier representation
$$f(\underline{x}) = \int e^{i\,\underline{\omega}\cdot\underline{x}}\,\tilde{f}(\underline{\omega})\,d\underline{\omega}.$$
- L_1 spectral norm: $\|f\|_{spectrum,s} = \int_{R^d} |\tilde{f}(\underline{\omega})|\ \|\underline{\omega}\|_1^s\,d\underline{\omega}$
- Let Φ be the dictionary for sinusoids, sigmoids or ramps. With unit cube domain for $\underline{x}$, the variation $\|f\|_\Phi$ satisfies
$$\begin{aligned} \|f\|_{sinusoid} &= \|f\|_{spectrum,0} \\ \|f\|_{sigmoid} &\le \|f\|_{spectrum,1} \\ \|f\|_{ramp} &\le \|f\|_{spectrum,2} \end{aligned}$$
- For sinusoid, sigmoid cases the parameter set is $\mathcal{A} = R^d$
- For the ramp case $\mathcal{A} = \{\underline{a} : \|\underline{a}\|_1 \le 1\}$.
- As we said, this $\|f\|_\Phi$ appears in the approximation bound
$$\|f - f_m\| \le \|f\|_\Phi/\sqrt{m}.$$

Statistical Risk

- Statistical risk $E\|\hat{f}_m - f\|^2 = E(\hat{f}_m(\underline{X}) - f(\underline{X}))^2$
- Expected squared generalization error on new $\underline{X} \sim P$
- Minimax optimal risk bound, via information theory

$$E\|\hat{f}_m - f\|^2 \;\leq\; \|f_m - f\|^2 + c\frac{m}{n}\log N(\Phi, \delta)$$

  where $\log N(\Phi, \delta)$ is the metric entropy of Φ at $\delta_m = 1/m$
- With greedy optimization using ℓ_1 penalty or suitable $\hat{m}$

$$E\|\hat{f} - f\|^2 \;\leq\; \min_{g.m} \left\{ \|g - f\|^2 + \frac{\|g\|_\Phi^2}{m} + c\frac{m}{n}\log N(\Phi, \delta_m) \right\}$$

  achieves ideal approximation, complexity tradeoff.

Statistical Risk with MDL Selection

- Again, general risk bound, using metric entropy $\log N(\Phi, \delta)$

$$E\|\hat{f} - f\|^2 \;\leq\; \min_{g.m} \left\{ \|g - f\|^2 + \frac{\|g\|_\Phi^2}{m} + c\frac{m}{n} \log N(\Phi, \delta_m) \right\}.$$

- For greedy algorithm with Minimum Description Length $\hat{m}$

$$\min_m \left\{ \|Y - \hat{f}_m\|_{(n)}^2 + 2c\frac{m}{n} \log N(\Phi, \delta) \right\}$$

- Performs as well as if the best m^* were known in advance.
- ℓ_1 penalty retains MDL interpretation and risk (B.,Huang,Li, Luo,2008)
- Risk bound specializes when $\|f\|_\Phi$ is finite

$$E\|\hat{f} - f\|^2 \;\leq\; \min_m \left\{ \frac{\|f\|_\Phi^2}{m} + c\frac{m}{n} \log N(\Phi, \delta_m) \right\}$$

$$\leq\; c\,\|f\|_\Phi \left(\frac{1}{n} \log N(\Phi, \delta_{m^*}) \right)^{1/2}$$

Statistical risk for neural nets

- Specialize the metric entropy $\log N(\Phi, \delta)$ (Klusowski, Barron 2016).
- It is not more than order $d \log(1/\delta)$ for Lipshitz activation functions such as sigmoids.
- With ℓ_1 constrained internal parameters, as in ramp case with finite $\|f\|_{spectrum,2}$, also not more than order $(1/\delta) \log d$
- Risk bound is $\|f\|_\Phi [\frac{d}{n} \log(n/d)]^{1/2}$ or $\|f\|_\Phi^{4/3} [\frac{1}{n} \log d]^{1/3}$, whichever is smallest.
- The $[(\log d)/n]^{1/3}$ is for the no-noise case. For the case with noise that is sub-exponential and sub-Gaussian, may replace it by $[(\log d)/n]^{1/4}$, to within $\log n$ factors.
- Implication: Can allow $d >> n$ provided n is large enough to accomodate the worse exponent of $1/3$ in place of $1/2$.

Confronting the computational challenge

- Greedy search
  - Reduces dimensionality of optimization from *md* to just *d*
  - Obtain a current $\underline{a}_m$ achieving within a constant factor of the maximum of

$$J_n(\underline{a}) \;=\; \frac{1}{n}\sum_{i=1}^{n} R_i\,\phi(\underline{X}_i, \underline{a}).$$

- This surface can still have many maxima.
  - We might get stuck at a spurious local maximum.
- New computational strategies identify approximate maxima with high probability
  1. Adaptive Annealing
  2. Third-order Tensor Methods (pros and cons)
  3. Nonlinear Power Methods
- These are stochastically initialized search methods

Optimization path for bounded ridge bases

Adaptive Annealing:

- A more general approach to seek approx optimization of

$$J(\underline{a}) = \sum_{i=1}^{n} r_i \, \phi(\underline{a}^T \underline{X}_i)$$

- recent & current work with Luo, Chatterjee, Klusowski
- Sample $\underline{a}_t$ from the evolving density

$$p_t(\underline{a}) \;=\; e^{\,t\,J(\underline{a})-c_t}\, p_0(\underline{a})$$

  along a sequence of values of t from 0 to t_{final}
- use t_{final} of order $(d\log d)/n$
- Initialize with a_0 drawn from a product prior $p_0(\underline{a})$:
  - Uniform$[-1, 1]$ for each coefficient in bounded a case
  - Normal(0,I) or product of Cauchy in unbounded a case
- Starting from the random a_0 define the optimization path a_t such that its distribution tracks the target density p_t.

Optimization path

- Adaptive Annealing: Arrange a_t from the evolving density

$$p_t(a) = e^{tJ(a)-c_t} p_0(a)$$

with a_0 drawn from $p_0(a)$

- State evolution with vector-valued change function $G_t(a)$:

$$a_{t+h} = a_t \; - \; h\, G_t(a_t)$$

- Better state evolution: $a_{t+h} = a^*$ is the solution to

$$a_t = a^* \; + \; h\, G_t(a^*),$$

with small step-size h, such that $a + h\,G_t(a)$ is invertible with a positive definite Jacobian, and solves equations for the evolution of $p_t(a)$.

- As we will see there are many such change functions $G_t(a)$, though not all are nice.

Boundary requirements

- Suppose a is restricted to a bounded domain $\mathcal{A}$ with smooth boundary $\partial\mathcal{A}$
- For $a \in \partial\mathcal{A}$, let v_a be outward normal to $\partial\mathcal{A}$ at a.
  - Either $G_t(a) = 0$ (vanishing at the boundary)
  - Or $G_t(a)^T v_a \geq 0$ (to move inward, not move outside)
- Likewise, for a near $\partial\mathcal{A}$, if $G_t(a)$ approaches 0, it should do so at order not larger than the distance of a from $\partial\mathcal{A}$.

Solve for the change G_t to track the density p_t

- Density evolution: by the Jacobian rule

$$p_{t+h}(a) = p_t\big(a + h\,G_t(a)\big)\,det\big(I + h\nabla G_t^T(a)\big)$$

Up to terms of order h

$$p_{t+h}(a) = p_t(a) + h\left[(G_t(a))^T\,\nabla p_t(a) + p_t(a)\,\nabla^T G_t(a)\right]$$

- In agreement for small h with the partial diff equation

$$\frac{\partial}{\partial t}\,p_t(a) \;=\; \nabla^T\left[G_t(a)p_t(a)\right]$$

- The right side is $G_t^T(a)\nabla p_t(a) + p_t(a)\nabla^T G_t(a)$. Dividing by $p_t(a)$ it is expressed in the log density form

$$\frac{\partial}{\partial t}\log p_t(a) \;=\; \nabla^T G_t(a) \;+\; G_t^T(a)\,\nabla\log p_t(a)$$

Five candidate solutions

Five solutions to the partial differential equation at time t

$$\nabla^T \left[G(a) p_t(a) \right] = \partial_t \, p_t(a)$$

1. Solution in which $G(a)p(a)$ is a gradient
2. Solution using pairs of coefficients
3. Solution with j random, $\partial_{a_j}[G_j(a)p(a)]$ provides $\partial_t p_t(a)$
4. Solution in which $G(a)$ is a gradient
5. Approximate solutions expressed in terms of $u_i = X_i^T a$.

Candidate solution 1.

Solution of smallest L_2 norm of $G_t(a)p_t(a)$ at a specific t.

- Let $G_t(a)p_t(a) = \nabla b(a)$, gradient of a function $b(a)$
- Let $f(a) = \frac{\partial}{\partial t} p_t(a)$
- Set $green_d(a)$ proportional to $1/\|a\|^{d-2}$, harmonic $a \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T \nabla b(a) = f(a)$$

- Solution

$$b(a) \quad = \quad (f * green)(a)$$

Candidate solution 1.

Solution of smallest L_2 norm of $G_t(a)p_t(a)$ at a specific t

- Let $G_t(a)p_t(a) = \nabla b(a)$, gradient of a function $b(a)$
- Let $f(a) = \frac{\partial}{\partial t} p_t(a)$
- Set $green_d(a)$ proportional to $1/\|a\|^{d-2}$, harmonic $a \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T \nabla b(a) = f(a)$$

- Solution, using $\nabla green_d(a) = c_d \, a/\|a\|^d$

$$\nabla b(a) \;=\; (f * \nabla green_d)(a)$$

Candidate solution 1.

Solution of smallest L_2 norm of $G_t(a)p_t(a)$ at a specific t

- Let $G_t(a)p_t(a) = \nabla b(a)$, gradient of a function $b(a)$
- Let $f(a) = \frac{\partial}{\partial t} p_t(a)$
- $green_d(a)$ proportional to $1/\|a\|^{d-2}$, harmonic for $a \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T [G_t(a)p_t(a)] = f(a)$$

- Solution, using $\nabla green_d(a) = c_d \, a/\|a\|^d$

$$G_t(a)p_t(a) = \ (f * \nabla green_d)(a)$$

- If $\mathcal{A} \subset R^d$, set $green_{\mathcal{A}}(a, \tilde{a})$ to be the associated Green's function, replacing $green_d(a - \tilde{a})$ in the convolution

Candidate solution 1.

Solution of smallest L_2 norm of $G_t(a)p_t(a)$ at a specific t

- Let $G_t(a)p_t(a) = \nabla b(a)$, gradient of a function $b(a)$
- Let $f(a) = \frac{\partial}{\partial t} p_t(a)$
- $green_d(a)$ proportional to $1/\|a\|^{d-2}$, harmonic for $a \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T [G_t(a)p_t(a)] = f(a)$$

- Solution, using $\nabla green_d(a) = c_d \, a/\|a\|^d$

$$G_t(a)p_t(a) = \ (f * \nabla green_d)(a)$$

- If $\mathcal{A} \subset R^d$, let $green_{\mathcal{A}}(a, \tilde{a})$ replace $green_d(a - \tilde{a})$.
- Then $G_t(a)p_t(a)$ tends to 0 as a tends to $\partial\mathcal{A}$.

Candidate solution 1.

Solution of smallest L_2 norm of $G_t(a)p_t(a)$ at a specific t

- Let $G_t(a)p_t(a) = \nabla b(a)$, gradient of a function $b(a)$
- Let $f(a) = \frac{\partial}{\partial t} p_t(a)$
- $green_d(a)$ proportional to $1/\|a\|^{d-2}$, harmonic for $a \neq 0$.
- The partial diff equation becomes the Poisson equation:

$$\nabla^T [G_t(a)p_t(a)] = f(a)$$

- Solution, using $\nabla green_d(a) = c_d \, a/\|a\|^d$

$$G_t(a) \;=\; \frac{(f * \nabla green_d)(a)}{p_t(a)}$$

- Comp. challenge: high-dimensional convolution integral.

Candidate solution 2.

Solution using 2−dimensional Green integrals

- Write the pde $\nabla^T[G_t(a)p_t(a)] = f(a)$ in the coordinates $G_{t,j}$

$$\sum_{j=1}^{d} \frac{\partial}{\partial a_j}[G_{t,j}(a)p_t(a)] = f(a)$$

- Pair consecutive terms to achieve a portion of the solution

$$\sum_{i\in\{j,j+1\}} \frac{\partial}{\partial a_i}[G_{t,i}(a)p_t(a)] = \frac{2}{d} f(a)$$

- Solution, for each consecutive pair of coordinates,

$$\begin{bmatrix} G_{t,j}(a) \\ G_{t,j+1}(a) \end{bmatrix} = \frac{2}{d}\frac{(f * \nabla green_2)(a)}{p_t(a)}$$

The 2−dim Green's function gradient acts on (a_j, a_{j+1}).

- Yields a numerical solution. Stable for particular J and p_0? Do we lose the desirable boundary behavior?

Candidate solution 2.

Solution using 2−dimensional Green integrals

- Solution, for each consecutive pair of coordinates,

$$\begin{bmatrix} G_{t,j}(a) \\ G_{t,j+1}(a) \end{bmatrix} = \frac{2}{d} \frac{(f * \nabla green_2)(a)}{p_t(a)}$$

- Stable for particular objective functions J?
- For p_0 we use a product of 2−dimensional uniform densities on unit disks
- Stable if $J(a)$ can exhibit only small change by changing two consecutive coordinates
- True for Lipshitz sigmoids with variable replication. Terms $\phi(a^T X)$ represented using small η as $\phi\left(\eta \sum_{j,r} a_{j,r} X_j\right)$. For each X_j the aggregate coefficient is $a_j = \eta \sum_{r=1}^{rep} a_{j,r}$.
- Challenge is that $G(a)$ is not necessarily zero at boundary.

Candidate solution 3. Solving 1-dim diff equations

Consider the equation with each a_j in an interval $[-1, 1]$,

$$\partial_{a_j}[G_j(a)p(a)] = \partial_t p_t(a)$$

- Call the right side $f(a)$. It is an ordinary diff equation in a_j if other coordinates a_{-j} held fixed

$$\frac{\partial}{\partial a_j}\left[G_j(a_j, a_{-j})p(a_j, a_{-j})\right] = f(a_j, a_{-j})$$

- Solutions take the form

$$G_j(a) = \frac{1}{p(a)}\left[\int_{-1}^{a_j} f(\tilde{a}_j, a_{-j})d\tilde{a}_j - C(a_{-j})\right]$$

- Natural choices for the "constant" $C(a_{-j})$ are 0 or

$$S(a_{-j}) = \int_{-1}^{1} f(\tilde{a}_j, a_{-j})d\tilde{a}_j$$

- $G_j(a)p(a)$ is either $\int_{-1}^{a_j} f(\tilde{a}_j, a_{-j})d\tilde{a}_j$ or $-\int_{a_j}^{1} f(\tilde{a}_j, a_{-j})d\tilde{a}_j$

Candidate solution 3. Solving 1-dim diff equations

$G_j(a)p(a)$ is either $\int_{-1}^{a_j} f(\tilde{a}_j, a_{-j})d\tilde{a}_j$ or $-\int_{a_j}^{1} f(\tilde{a}_j, a_{-j})d\tilde{a}_j$

- These choices make $G_j(a)$ be zero as a_j approaches one or the other of the end points, but not both.
- Thus led to define the solution $G_j(a)$ given by

$$\frac{1}{p(a)}\left[\int_{-1}^{a_j} f(\tilde{a}_j, a_{-j})d\tilde{a}_j\, 1_{\{S(a_{-j})\geq 0\}} - \int_{a_j}^{1} f(\tilde{a}_j, a_{-j})d\tilde{a}_j\, 1_{\{S(a_{-j})<0\}}\right]$$

- This makes move be inward near the non-zero edge.
- If we pick j at random in $1,\ldots,d$, get similar density update.
- The rule solves differential equation for the order h term.
- It satisfies some desired boundary properties.
- Challenge: Boundary slightly moved at the non-zero edge.

Candidate solution 4.

Perhaps the ideal solution is one of smallest L_2 norm of $G_t(a)$

- It has $G_t(a) = \nabla b_t(a)$ equal to the gradient of a function
- The pde in log density form

$$\nabla^T G_t(a) + G_t^T(a)\,\nabla \log p_t(a) = \frac{\partial}{\partial t} \log p_t(a)$$

  then becomes an elliptic pde in $b_t(a)$ for fixed t.
- With $\nabla \log p_t(a)$ and $\frac{\partial}{\partial t} \log p_t(a)$ arranged to be bounded, the solution may exist and be nice.
- But explicit solution to this elliptic pde is not available (except perhaps numerically in low dim cases).

Candidate solution 4.

Ideal solution of smallest L_2 norm of $G_t(a)$

- It has $G_t(a) = \nabla b_t(a)$ equal to the gradient of a function
- The pde in log density form

$$\nabla^T G_t(a) + G_t^T(a) \nabla \log p_t(a) = \frac{\partial}{\partial t} \log p_t(a)$$

  then becomes an elliptic pde in $b_t(a)$ for fixed t.
- With $\nabla \log p_t(a)$ and $\frac{\partial}{\partial t} \log p_t(a)$ arranged to be bounded, the solution may exist and be nice.
- But explicit solution to this elliptic pde is not available (except perhaps numerically in low dim cases).
- Next seek approximate solution.
- For ridge bases, decompose into a system of first order differential equations and integrate.

Candidate solution 5 by decomposition of ridge sum

- Optimize $J(a) = \sum_{i=1}^{n} r_i \, \phi(X_i^T a)$
- Target density $p_t(a) = e^{tJ(a)-c_t} \, p_0(a)$ with $c_t' = E_{p_t}[J]$
- The time score is $\frac{\partial}{\partial t} \log p_t(a) = J(a) - E_{p_t}[J]$
- Specialize the pde in log density form

$$\nabla^T G_t(a) \; + \; G_t^T(a) \, \nabla \log p_t(a) \; = \; J(a) - E_{p_t}[J]$$

- The right side, setting $b_{i,t} = E_{p_t}[\phi(x_i^T a)]$, takes the form of a sum

$$\sum r_i \, [\phi(X_i^T a) - b_{i,t}].$$

- Likewise $\nabla \log p_t(a) = t \, \nabla J(a) + \nabla \log p_0(a)$ is the sum

$$\sum X_i \, \left[t \, r_i \phi'(X_i^T a) - (1/n)(X_i^T a) \right]$$

- Use a Gaussian initial distribution with $\log p_0(a)$ equal to

$$-(1/2n) \sum a^T X_i X_i^T a.$$

Account for prior by appending d extra input vectors as columns

Approximate solution for ridge sums

- Seek approximate solution of the form

$$G_t(a) = \sum \frac{x_i}{\|x_i\|^2} g_i(\underline{u})$$

with $\underline{u} = (u_1, \ldots, u_n)$ evaluated at $u_i = X_i^T a$, for which

$$\nabla^T G_t(a) = \sum_i \frac{\partial}{\partial u_i} g_i(\underline{u}) + \sum_{i,j:i \neq j} \frac{x_i^T x_j}{\|x_i\|^2} \frac{\partial}{\partial u_j} g_i(\underline{u})$$

- Can we ignore the coupling in the derivative terms?
- $x_j^T x_i / \|x_i\|^2$ are small for uncorrelated designs, large d.
- Match the remaining terms in the sums to solve for $g_i(\underline{u})$
- Arrange $g_i(\underline{u})$ to solve the differential equations

$$\frac{\partial}{\partial u_i} g_i(\underline{u}) \,+\, g_i(\underline{u}) \big[t\, r_i \phi'(u_i) - u_i/n + rest_i\big] \,=\, r_i \big[\phi(u_i) - b_{i,t}\big]$$

where $rest_i = \sum_{j \neq i} [t\, r_j\, \phi'(u_j) - u_j/n] x_j^T x_i / \|x_i\|^2$.

Integral form of solution

- Differential equation for $g_i(u_i)$, suppressing dependence on the coordinates other than i

$$\frac{\partial}{\partial u_i} g_i(u_i) + g_i(u_i)\big[t\, r_i \phi'(u_i) - u_i/n + rest_i\big] = r_i\big[\phi(u_i) - b_{i,t}\big]$$

- Define the density factor

$$m_i(u_i) = e^{t\, r_i\, \phi(u_i) - u_i^2/2n + u_i\, rest_i}$$

- Allows the above diff equation to be put back in the form

$$\frac{\partial}{\partial u_i}[g_i(u_i)\, m_i(u_i)] = r_i\big[\phi(u_i) - b_{i,t}\big]\, m_i(u_i)$$

- An explicit solution, evaluated at $u_i = x_i^T a$, is

$$g_i(u_i) = r_i \frac{\int_{c_i}^{u_i} m_i(\tilde{u}_i)\big[\phi(\tilde{u}_i) - b_{i,t}\big]\, d\tilde{u}_i}{m_i(u_i)}$$

where $c_i = c_{i,t}$ is such that $\phi(c_{i,t}) = b_{i,t}$.

The derived change function G_t for evolution of a_t

- Include the u_j for $j \neq i$ upon which $rest_i$ depends. Our solution for $g_{i,t}(\underline{u})$ is

$$r_i \int_{c_i}^{u_i} e^{t\, r_i\, (\phi(\tilde{u}_i) - \phi(u_i)) - (\tilde{u}_i^2 - u_i^2)/2n + t(\tilde{u}_i - u_i) rest_i(\underline{u})} \left[\phi(\tilde{u}_i) - b_{i,t}\right] d\tilde{u}_i$$

- Evaluating at $\underline{u} = Xa$ we have the change function

$$G_t(a) = \sum \frac{x_i}{\|x_i\|^2} g_{i,t}(Xa)$$

for which a_t evolves according to

$$a_{t+h} = a_t - h\, G_t(a_t)$$

- For showing $g_{i,t}$, G_t and ∇G_t are nice, assume the activation function ϕ and its derivative is bounded (e.g. a logistic sigmoid or a sinusoid).
- Run several optimization paths in parallel, starting from independent choices of a_0. Allows access to empirical computation of $b_{i,t} = E_{p_t}[\phi(x_i^T a_t)]$

Conjectured conclusion

Derived the desired optimization procedure and the following.

Conjecture: With step size h of order $1/n^2$ and a number of steps of order $n\, d \log d$ and $X_1, X_2, \ldots, X_n$ i.i.d. Normal$(0, I)$. *With high probability on the design X, the above procedure produces optimization paths a_t whose distribution closely tracks the target*

$$p_t(a) = e^{t\,J(a) - c_t}\, p_0(a)$$

such that, with high probability, the solutions paths have instances of $J(a_t)$ which are at least $1/2$ the maximum.

Consequently, the relaxed greedy procedure is computationally feasible and achieves the indicated bounds for sparse linear combinations from the dictionary $\Phi = \{\phi(a^T x) : a \in R^d\}$.

Summary

- Flexible approximation models
  - Subset selection
  - Nonlinearly parameterized bases as with neural nets
  - ℓ_1 control on coefficients of combination
- Accurate approximation with moderate number of terms
  - Proof analogous to random coding
- Information theoretic risk bounds
  - Based on the minimum description length principle
  - Shows accurate estimation, even for very large dimension
- Computational challenges are being addressed by
  - Adaptive annealing
  - Nonlinear power methods

Tensor and nonlinear power methods (overview)

- Know design distribution $p(X)$
- Target $f(x) = \sum_{k=1}^{m_o} g_k(a_k^T x)$ is a combination of ridge functions with distinct linearly independent directions a_k
- Ideal: maximize $E[f(X)\phi(a^T X)]$ or $(1/n)\sum_i Y_i \phi(a^T X_i)$
- Score functions operating on $f(X)$ and $f(X)\, g(a^T X)$ yield population and sample versions of tensors

$$E\left[\frac{\partial^3}{\partial X_{j_1} \partial X_{j_2} \partial X_{j_3}} f(X)\right]$$

and nonlinearly parameterized matrixes

$$E\left[(\nabla\nabla^T f(X)) g(a^T X)\right]$$

- Spectral decompositions then identify the directions a_k

Score method for representing expected derivatives

- Score function (tensor) $S^\ell(X)$ of order ℓ from known $p(X)$

$$S_{j_1,\ldots j_\ell}(X)\, p(X) \;=\; (-1)^\ell \frac{\partial^\ell}{\partial X_{j_1} \cdot \partial X_{j_\ell}} p(X)$$

Gaussian score: $S^1(X) = X$,

$$S^2(X) = XX^T - I,$$

$$S^3_{j_1,j_2,j_3}(X) = X_{j_1} X_{j_2} X_{j_3} - X_{j_1} 1_{j_2,j_3} - X_{j_2} 1_{j_1,j_3} - X_{j_3} 1_{j_1,j_2}.$$

- Expected derivative:

$$E\left[\frac{\partial^\ell}{\partial X_{j_1} \cdot \partial X_{j_\ell}} f(X)\right] = E\left[f(X) S_{j_1,\ldots j_\ell}(X)\right]$$

- Repeated integration by parts

Expected derivatives of ridge combinations

- Ridge combination target functions:

$$f(X) = \sum_{k=1}^{m_o} g_k(a_k^T X)$$

- Expected Hessian of $f(X)$

$$M = \sum_{k=1}^{m_o} a_k a_k^T \, E[g_k''(a_k^T X)] = E\left[f(X) S^2(X)\right].$$

Principle eigenvector:

$$\max_a \left\{a^T M \, a\right\}$$

Linear power method finds a_k if othogonal (the're not).

- Third order array (Anandkumar *et al* 2015, draft):

$$\sum_{k=1}^{m_o} a_{j_1,k} a_{j_2,k} a_{j_3,k} E[g_k'''(a_k^T X)] = E\left[f(X) S_{j_1,j_2,j_3}(X)\right]$$

can be whitened and a quadratic power method finds a_k.

Scoring a Ridge Function

- A suitable activation function $\phi(a, X)$ for optimization of

$$E[f(X)\phi(a, X)]$$

- Matrix scoring of a ridge function $g(a^T X)$:

$$M_{a,X} = S^2 g(a^T X) + [S^1 a^T + a(S^1)^T] g'(a^T X) + [aa^T] g''(a^T X)$$

- Activation function formed by scoring a ridge function

$$\phi(a, X) = a^T [M_{a,X}] a$$
$$= (a^T S^2 a) g(a^T X) + 2(a^T S^1)(a^T a) g'(a^T X) + (a^T a)^2 g''(a^T X)$$

- Scoring a ridge function permits finding the component of $\phi(a, X)$ in the target function using

$$E[f(X)\phi(a, X)] = a^T E[f(X) M_{a,X}] a = a^T E[(\nabla \nabla^T f(X)) g(a^T X)] a$$

- Twice itegrating by parts

Scoring a Ridge Function (Gaussian design case)

- Matrix scoring of a ridge function $g(a^T X)$:

$$M_{a,X} = S^2 g(a^T X) + [S^1 a^T + a(S^1)^T] g'(a^T X) + [aa^T] g''(a^T X)$$

- Activation function formed by scoring a ridge function

$$\phi(a, X) = a^T [M_{a,X}] a$$

$$= (a^T S^2 a) g(a^T X) + 2(a^T S^1)(a^T a) g'(a^T X) + (a^T a)^2 g''(a^T X)$$

- Gaussian design case, simplifying when $\|a\| = 1$:

$$\phi(a^T X) = [(a^T X)^2 - 1] g(a^T X) + [2a^T X] g'(a^T X) + g''(a^T X)$$

$$\phi(z) = (z^2 - 1) g(z) + 2z\, g'(z) + g''(z)$$

- Hermite poly: If $g(z) = H_{\ell-2}(z)$ then $\phi(z) = H_\ell(z)$ for $\ell \geq 2$.

Scored Ridge Function Decomposes $E[f(X)\phi(a,X)]$

- Matrix scored ridge function, providing $\phi(a,X) = a^T M_{a,X} a$,

$$M_{a,X} = S^2 g(a^T X) + [Sa^T + aS^T] g'(a^T X) + [aa^T] g''(a^T X)$$

- The amount of $\phi(a,X)$ in $f(X)$ via the matrix decomposition

$$M_a = E[f(X) M_{a,X}] = E[(\nabla\nabla^T f(X)) g(a^T X)] = \textstyle\sum_{k=1}^{m_o} a_k a_k^T G_k(a_k, a)$$

is quantified by

$$E[f(X)\phi(a,X)] = a^T [M_a] a = \sum_{k=1}^{m_0} (a_k^T a)^2 G_k(a_k, a)$$

- Here $G_k(a_k, a) = E[g_k''(a_k^T X) g(a^T X)]$ measures the strength of the match of a to the direction a_k.
- It replaces $E[g_k''(a_k^T X) S^T] a = (a_k^T a) E[g_k'''(a_k^T X)]$ in the tensor method of Anandkumar *et al*

Using Sinusoids or Sigmoids

- The amount of $\phi(a, X)$ in $f(X)$ via the matrix decomposition

$$M_a = E[f(X)M_{a,X}] = \sum_{k=1}^{m_o} a_k a_k^T G_k(a_k, a)$$

quantified by

$$E[f(X)\phi(a, X)] = a^T[M_a]a = \sum_{k=1}^{m_0} (a_k^T a)^2 G_k(a_k, a)$$

- Here $G_k(a_k, a) = E[g_k''(a_k^T X) g(a^T X)]$ measures the strength of the match of a to the direction a_k.
- $\cos(z)$, $\sin(z)$ case, with X standard multivariate Normal:

$g_k(a_k^T X) = -c_k e^{i\, a_k^T X}$ and $g(a^T X) = e^{-i\, a^T X}$

expected product $G_k(a_k, a) = c_k e^{-(1/2)\|a_k - a\|^2}$
- Step sigmoid case $\phi(z) = 1_{\{z>0\}}$: The $G_k(a_k, a)$ is determined by the angle between a_k and a.

Using Hermite polynomials

- The amount of $\phi(a, X)$ in $f(X)$ via the matrix decomposition

$$M_a = E[f(X)M_{a,X}] = \sum_{k=1}^{m_o} a_k a_k^T G_k(a_k, a)$$

is given by

$$E[f(X)\phi(a, X)] = a^T[M_a]a = \sum_{k=1}^{m_0} (a_k^T a)^2 G_k(a_k, a)$$

- Here $G_k(a_k, a) = E[g_k''(a_k^T X)g(a^T X)]$ measures the strength of the match of a to the direction a_k.
- Hermite case: $g(z) = H_{\ell-2}(z)$, with $X \sim \text{Normal}(0, I)$. $H_\ell(a^T X)$ and $H_{\ell'}(a_k^T X)$ orthonormal for $\ell' \neq \ell$.

$$G_k(a_k, a) = c_{k,\ell}\,(a_k^T a)^\ell$$

with $c_{k,\ell} = E[g_k(Z)H_\ell(Z)]$ in $g_k(z) = \sum_{\ell'} c_{k,\ell'} H_{\ell'}(z)$

Nonlinear Power Method

- Maximize $J(a) = E[f(X)\phi(a, X)] = a^T M_a\, a,$ s.t. $\|a\| = 1$
- Cauchy-Schwartz inequality:

$$a^T M_a\, a \leq \|a\|\, \|M_a a\|$$

with equality iff a is proportional to $M_a a$.

- Motivates the mapping of the nonlinear power method

$$V(a) = \frac{M_a\, a}{\|M_a a\|}$$

- Seek fixed points $a^* = V(a^*)$ via iterations $a_t = V(a_{t-1})$.
- Construct a whitened version.
- Verify that $J(a_t)$ is increasing.
- The nonlinear power method provides maximizers of

$$J(a) = E[f(X)\phi(a, X)]$$

Analysis with Whitening

- Suppose $m_o \leq d$ (# components $\leq$ dimension)
- Let $Ref = \sum_k a_k a_k^T \beta_k$ be a reference matrix, for instance $Ref = M_{a_{ref}}$ has $\beta_k = G_k(a_k, a_{ref})$, and let $Q\,D\,Q^T$ be its eigen-decomposition.
- Let $W = QD^{-1/2}$ be the whitening matrix:

$$I = W^T Ref\, W = \sum_k (W^T a_k)(a_k^T W)\beta_k = \sum_k \alpha_k\, \alpha_k^T$$

with orthonormal directions

$$\alpha_k = W^T a_k \sqrt{\beta_k}$$

- Represent $a = W\,u / \|Wu\| = W\,u\,\sqrt{\beta}$ for unit vectors u.
- Then $\quad a^T a_k = u^T \alpha_k\, (\beta/\beta_k)^{1/2}$
- Let u_{ref} be the unit vector prop to $W^{-1} a_{ref} = D^{1/2} Q^T a_{ref}$

Analysis of the Nonlinear Power Method

- Criterion $E[f(X)\phi(a,X)] = a^T M_a\, a = u^T \tilde{M}_u\, u$ where

$$\tilde{M}_u = \sum_k \alpha_k \alpha_k^T\; \tilde{G}_k(\alpha_k, u)\, \beta/\beta_k$$

and $\tilde{G}_k$ is G_k with a_k, a expressed via α_k, u. Example

$$\tilde{G}_k(\alpha_k^T u) = c_{k,\ell}\,(\alpha_k^T u)^\ell\,(\beta/\beta_k)^{\ell/2}$$

$$\tilde{M}_u = \sum_k \alpha_k \alpha_k^T\; (\alpha_k^T u/\alpha_k^T u_{ref})^\ell$$

- The power mapping $a_t = M_{a_{t-1}} a_{t-1}/\|\cdot\|$ corresponds to

$$u_t = \tilde{M}_{u_{t-1}} u_{t-1}/\|\cdot\|$$

- Provably rapidly convergent, when $\tilde{G}_k$ is increasing in $\alpha_k^T u$.
- Limit of u_t is $u^* = \pm\alpha_k$ with largest initial $(\alpha_k^T u_0/\alpha_k^T u_{ref})^\ell$.
- Each $+\alpha_k$ or $-\alpha_k$ is a local maximizer.
- Global maximizer corresponds to largest $1/|\alpha_k^T u_{ref}|$
- Corresponding maximizer of $a^T M_a\, a$ is a^* prop to Wu^*.

Analysis of Nonlinear Power Method, Polynomial Case

- Let $c_k(t) = \alpha_k^T u_t$ be coefficient of u_t in the direction α_k
- Let $c_{k,ref} = \alpha_k^T u_{ref}$ be coefficient of u_{ref} in direction α_k

$$\tilde{M}_{u_t} = \sum_k \alpha_k \, \alpha_k^T \, (\alpha_k^T u_t / \alpha_k^T u_{ref})^\ell$$

- So that

$$\tilde{M}_{u_t} \, u_t = \sum_k \alpha_k (\alpha_k^T u_t) \, (\alpha_k^T u_t / \alpha_k^T u_{ref})^\ell$$

Thus the coefficienct for u_{t+1} satisfies the recursion:

$$c_k(t+1) = \frac{\left[c_k(t)/c_{k,ref}\right]^{\ell+1} c_{k,ref}}{[\sum_k (\quad)^2]^{1/2}}$$

- By induction

$$c_k(t) = \frac{\left[c_k(0)/c_{k,ref}\right]^{(\ell+1)^t} c_{k,ref}}{[\sum_k (\quad)^2]^{1/2}}$$

- It rapidly concentrates on the index k with the largest

$$\frac{c_k(0)}{c_{k,ref}} = \frac{\alpha_k^T u_0}{\alpha_k^T u_{ref}}$$

Analysis of Nonlinear Power Method, Polynomial Case

- Suppose $k = 1$ has the largest

$$\frac{c_k(0)}{c_{k,ref}} = \frac{\alpha_k^T u_0}{\alpha_k^T u_{ref}}$$

  with the others less by the factor $1 - \Delta$. Then

$$\|u_t - \alpha_1\|^2 \leq 2(1 - \Delta)^{2(\ell+1)^t}$$

- Moreover $J(a_t) = E[f(X)\phi(a_t, X)] = u_t^T \tilde{M}_{u_t} u_t$ equals

$$\sum_k \left[c_k(t)/c_{k,ref}\right]^{\ell+2} c_{k,ref}^2$$

  which is strictly increasing in t, proven by applications of Holder's inequality
- Factor of increase quantified by the exponential of a relative entropy.
- The increase each step is large unless $c_k^2(t)$ is close to concentrated on the maximizers of $\alpha_k^T u_0 / \alpha_k^T u_{ref}$.