Computationally feasible greedy algorithms for neural nets

Andrew R. Barron

Yale University
Department of Statistics

Presentation, December 12, 2015
NIPS Workshop on Non-convex Optimization, Montreal
Joint work with Jason Klusowski
Outline

Flexible high-dimensional function estimation with sigmoidal, sinusoidal and polynomial activation functions

Approximation and estimation bounds

Greedy term selection

Computational strategies
  - Exhaustive search in discretized directions
  - Adaptive Annealing
  - Nonlinear power methods
Data Setting

- **Data**: $(X_i, Y_i), i = 1, 2, \ldots, n$
- **Inputs**: explanatory variable vectors

\[ X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d}) \]

- **Domain**: Either a unit cube in $R^d$ or all of $R^d$
- **Random design**: independent $X_i \sim P$
- **Output**: response variable $Y_i$ in $R$
  - Moment conditions, with Bernstein constant $c$
- **Relationship**: $E[Y_i|X_i] = f(X_i)$ as in:
  - Perfect observation: $Y_i = f(X_i)$
  - Noisy observation: $Y_i = f(X_i) + \epsilon_i$ with $\epsilon_i$ indep $N(0, \sigma^2)$
  - Classification: $Y \in \{0, 1\}$ with $f(X) = P[Y = 1|X]$
- **Function**: $f(x)$ unknown
Univariate 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 $z^\ell$, Hermite $H_\ell(z)$

Products or ridge form builds multivariate activation functions
Flexible multivariate function approximation: $d > 1$

By internally parameterized models & nonlinear least squares

- Fit functions $f_m(x) = \sum_{j=1}^{m} c_k \phi(x, a_k)$ in the span of a parameterized dictionary $\Phi = \{ \phi(\cdot, a) : a \in R^d \}$

- Product bases:
  using continuous powers, frequencies or thresholds

  $\phi(x, a) = \phi_1(x_1, a_1) \phi_1(x_2, a_2) \cdots \phi_1(x_d, a_d)$

- Ridge bases: as in projection pursuit regression models, sinusoidal models, and single-hidden-layer neural nets:

  $\phi(x, a) = \phi(a^T x) = \phi_1(a_1 x_1 + a_2 x_2 + \ldots + a_d x_d)$

- Internal parameter vector $a$ of dimension $d$.
- Activation function built from univariate function $\phi_1(z)$
Notation

• Response vector: \( Y = (Y_i)_{i=1}^n \) in \( R^n \)

• Dictionary vectors: \( \Phi(n) = \{(\phi(X_i, \theta))_{i=1}^n : \theta \in \Theta\} \subset R^n \)

• Sample squared norm: \( \|f\|_2^2(n) = \frac{1}{n} \sum_{i=1}^n f^2(X_i) \)

• Population squared norm: \( \|f\|^2 = \int f^2(x)P(dx) \)

• Normalized dictionary condition: \( \|\phi\| \leq 1 \) for \( \phi \in \Phi \)
Impractical one-shot optimization

- Sample version

\[ \hat{f}_m \text{ achieves } \min_{(\theta_j, c_j)]_{j=1}^m} \| Y - \sum_{j=1}^m c_j \phi_{\theta_j} \|_2^n \]

- Population version

\[ f_m \text{ achieves } \min_{(\theta_j, c_j)]_{j=1}^m} \| f - \sum_{j=1}^m c_j \phi_{\theta_j} \|_2^2 \]

- Optimization of \( (\theta_j, c_j)]_{j=1}^m \) in \( \mathbb{R}^{(d+1)m} \).
GREEDY OPTIMIZATIONS

Step 1: Choose $c_1, \theta_1$ to achieve $\min \| Y - c\phi_{\theta} \|^2_{(n)}$ or
- sample version: $\max_\theta (1/n) \sum_{i=1}^n nY_i \phi(X_i, \theta)$
- population version: $\max_\theta E[f(X)\phi(X, \theta)]$

Step $m > 1$: Arrange

$$\hat{f}_m = \alpha \hat{f}_{m-1} + c \phi(x, \theta_m)$$

with $\alpha_m, c_m, \theta_m$ chosen to achieve

$$\min_{\alpha, c, \theta} \| Y - \alpha \hat{f}_{m-1} - c \phi_{\theta} \|^2_{(n)}.$$

Also acceptable, with $R_i = Y_i - \hat{f}_{m-1}(X_i)$,
- Choose $\theta_m$ to achieve $\max_\theta \sum_{i=1}^n R_i \phi(X_i, \theta)$
- Population version: $\max_\theta E[R(X)\phi(X, \theta)]$

Forward stepwise selection of $S_m = \{ \phi_{\theta_1}, \ldots, \phi_{\theta_m} \}$. Given $S_{m-1}$, choose $\theta_m$ to $\min_\theta d(Y, \text{span}\{ \phi_{\theta_1}, \ldots, \phi_{\theta_m} \})$
For either one-shot or greedy approximation
(B. IT 1993, Lee et al IT 1995)
- Population version:

\[ \| f - f_m \| \leq \frac{\| f \|_\Phi}{\sqrt{m}} \]

and moreover

\[ \| f - f_m \|^2 \leq \inf_g \left\{ \| f - g \|^2 + \frac{2\| g \|^2_\Phi}{m} \right\} \]

- Sample version:

\[ \| Y - \hat{f}_m \|^2_{(n)} \leq \| Y - f \|^2_{(n)} + \frac{2\| f \|^2_\Phi}{m} \]

where \( \| f \|_\Phi \) is the variation of \( f \) with respect to \( \Phi \)
(as will be defined on the next slide).
\( \ell_1 \) norm on coefficients in representation of \( f \)

- Consider the range of a neural net, expressed via the bound,

\[
| \sum_j c_j \text{sgn}(\theta_{0,j} + \theta_{1,j}x_1 + \ldots + \theta_{d,j}x_d) | \leq \sum_j |c_j|
\]

equality if \( x \) is in polygon where \( \text{sgn}(\theta_j \cdot x) = \text{sgn}(c_j) \) for all \( j \)

- Motivates the norm

\[
\|f\|_\Phi = \lim_{\epsilon \to 0} \inf \left\{ \sum_j |c_j| : \| \sum_j c_j \phi_{\theta_j} - f \| \leq \epsilon \right\}
\]

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

\[
\|f\|_\Phi = V_\Phi(f) = \inf \{ V : f/V \in \text{closure}(\text{conv}(\pm \Phi)) \}
\]

- It appears in the bound \( \|f - f_m\| \leq \frac{\|f\|_\Phi}{\sqrt{m}} \)
\( \ell_1 \) norm on coefficients in representation of \( f \)

- Finite sum representations, \( f(x) = \sum_j c_j \phi(x, \theta_j) \). Variation 
  \( \|f\|_\Phi = \sum_j |c_j| \), which is the \( \ell_1 \) norm of the coefficients in representative of \( f \) in the span of \( \Phi \)

- Infinite integral representation 
  \( f(x) = \int e^{i \theta \cdot x} \tilde{f}(\theta) \, d\theta \) 
  (Fourier representation), for \( x \) in a unit cube. The variation 
  \( \|f\|_\Phi \) is bounded by an \( L_1 \) spectral norm:

  \[
  \|f\|_{\text{cos}} = \int_{R^d} |\tilde{\Phi}(\theta)| \, d\theta \\
  \|f\|_{\text{step}} \leq \int |\tilde{\Phi}(\theta)| \|\theta\|_1 \, d\theta \\
  \|f\|_{\text{ramp}} \leq \int |\tilde{\Phi}(\theta)| \|\theta\|_1^2 \, d\theta
  \]

- As we said, this \( \|f\|_\Phi \) appears in the numerator of the approximation bound.
The population accuracy of function estimated from sample

Statistical risk $E\|\hat{f}_m - f\|^2 = E(\hat{f}_m(X) - f(X))^2$

Expected squared generalization error on new $X \sim P$

Minimax optimal risk bound, via information theory

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

Here $\log N(\Phi, \delta)$ is the metric entropy of $\Phi$ at $\delta = 1/m$; it is of order $d \log(1/\delta)$ and, with $\ell_1$ constrained internal parameters, it is of order $(1/\delta) \log d$

$$E\|\hat{f}_m - f\|^2 \leq \frac{\|f\|_\Phi^2}{m} + \frac{c}{n} \min\{md \log(n/d), m^2 \log d\}$$

Bound is $2\|f\|_\Phi [\frac{cd}{n} \log n/d]^{1/2}$ or $3\|f\|_\Phi^{4/3} [\frac{c}{n} \log d]^{1/3}$, whichever is smallest
Adapt network size $m$ and choice of internal parameters

Minimum Description Length Principle leads to Complexity penalized least squares criterion. Let $\hat{m}$ achieve

$$\min_m \left\{ \| Y - \hat{f}_m \|_2^2 (n) + 2c \frac{m}{n} \log N(\Phi, \delta) \right\}$$

Information-theoretic risk bound

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

Performs as well as if the best $m^*$ were known in advance.

$\| f \|^2_\Phi / m$ replaces $\| f_m - f \|^2$ in the greedy case.

$l_1$ penalized least squares

Achieves the same risk bound

Retains the MDL interpretation (B, Huang,Li,Luo,2008)
Confronting the computational challenge

- **Greedy search**
  - Reduces dimensionality of optimization from $md$ to just $d$
  - Obtain a current $\theta_m$ achieving within a constant factor of the maximum of
    \[
    J_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} R_i \phi(X_i, \theta).
    \]
  - This surface can still have many maxima.
    - We might get stuck at a spurious local maximum.

- **New computational strategies:**
  1. Third order tensor methods (pros and cons)
  2. Nonlinear power methods
  3. Adaptive annealing
Tensor and nonlinear power methods

- Know design distribution $p(X)$

- Target $f(x) = \sum_{k=1}^{m_0} g_k(a_k^T x)$ 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 $\ell$ 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:

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

- Ridge combination target functions:
  \[ f(X) = \sum_{k=1}^{m_0} g_k(a_k^T X) \]

- Expected Hessian of \( f(X) \)
  \[ M = \sum_{k=1}^{m_0} a_k a_k^T E[g''_k(a_k^T X)] = E \left[f(X) S^2(X)\right]. \]

  Principle eigenvector:
  \[ \max_a \left\{ a^T M a \right\} \]

  Linear power method finds \( a_k \) if orthogonal (they’re not).

- Third order array (Anandkumar et al):
  \[ \sum_{k=1}^{m_0} a_{j_1,k} a_{j_2,k} a_{j_3,k} E[g''_k(a_k^T X)] = E \left[f(X) S_{j_1,j_2,j_3}(X)\right] \]

  can be whitened and a quadratic power method finds \( a_k \).
Scoring a Ridge Function

- 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.
Scoring a Ridge Function

- 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 case, simplifying when $||a|| = 1$:
  $$\phi(a^T X) = [(a^T X)^2 - 1]g(a^T X) + [2a^T!X]g'(a^T X) + g''(a^T X)$$
  $$\phi(z) = (z^2 - 1)g(z) + 2z g'(z) + g''(z)$$

- Scoring a ridge function permits finding the component of $\phi(a^T X)$ in the target function.
Matrix scored ridge function:

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

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

\[ M_a = E[f(X)M_{a,X}] = E[(\nabla \nabla^T f(X))g(a^T X)] = \sum_{k=1}^{m_0} a_k a_k^T G_k(a_k, a) \]

and

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

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

\[
M_a = E[f(X)M_{a,x}] = E[(\nabla\nabla^T f(X))g(a^TX)] = \sum_{k=1}^{m_0} a_k a_k^T G_k(a_k, a)
\]

and

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

Here $G_k(a_k, a) = E[g''_k(a_k^TX)g(a^TX)]$ measures the strength of the match of $a$ to the direction $a_k$.

\[\cos(z), \sin(z)\] case, with $X$ standard multivariate Normal:

$g_k(a_k^TX) = c_k e^{i a_k^TX}$ and $g(a^TX) = e^{-i a^TX}$

expected product $G_k(a_k, a) = c_k e^{-(1/2)\|a_k - a\|^2}$
Scoring a Ridge Function

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

\[ M_a = E[f(X)M_{a,X}] = E[(\nabla \nabla^T f(X))g(a^T X)] = \sum_{k=1}^{m_0} a_k a_k^T G_k(a_k, a) \]

and

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

- 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 polynomial** case, with $X \sim \text{Normal}(0, I)$:

  $H_\ell(a^T X)$ and $H_{\ell'}(a_k^T X)$ orthogonal for $\ell' \neq \ell$, and

\[ G_k(a_k, a) = c_{k,\ell} (a_k^T a)^\ell \]
Ideal: maximize $E[f(X)\phi(a, X)] = a^T M_a a$ 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 $a_t = V(a_{t-1})$. 

Andrew Barron

Computationally feasible greedy algorithms for neural nets
Suppose $m_o \leq d$ (# components $\leq$ dimension)

Let $R = \sum_k a_k a_k^T \beta_k$ be a reference matrix, for instance $R = M_\theta$ has $\beta_k = G_k(a_k, \theta)$, and let $QDQ^T$ be its eigen-decomposition.

Let $W = QD^{-1/2}$ be the whitening matrix:

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

with orthonormal directions

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

Also represent

$$a = W u \sqrt{\beta}$$

or

$$a = Wu/\|Wu\|$$

for unit vectors $u$. 
Analysis of the Nonlinear Power Method

Criterion

\[ E[f(X)φ(a, X)] = a^T M_a a = u^T \tilde{M}_u u \]

where

\[ \tilde{M}_u = \sum_k \alpha_k \alpha_k^T \tilde{G}_k(\alpha_k, u) \beta_k \]

and \( \tilde{G}_k \) is \( G_k \) with the \( a_k \) and \( a \) expressed via \( \alpha_k \) and \( u \).

The power mapping \( a_t = M_{a_{t-1}} a_{t-1} / \| \cdot \| \) corresponds to \( a_t \) proportional to \( u_t \) with

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

Provably rapidly convergent, when \( \tilde{G}_k \) increasing in the inner product \( \alpha_k^T u \).

Limit of \( u_t \) is \( u^* = \alpha_k \) with the largest initial \( \tilde{G}_k(\alpha_k, u_0) / \beta_k \).

Corresponding limit of \( a_t \) is \( a^* \) proportional to \( Wu^* \).

Direction \( a_k \) is revealed by \( W^{-T} \alpha_k / \sqrt{\beta_k} \).
Optimization path for bounded ridge bases

More general approach to seek approximation optimization of

\[ J(\theta) = \sum_{i=1}^{n} r_i \phi(\theta^T X_i) \]

Adaptive Annealing:

- recent & current work with Luo, Chatterjee, Klusowski
- Sample \( \theta_t \) from the evolving density

\[ p_t(\theta) = e^{tJ(\theta) - ct} p_0(\theta) \]

along a sequence of values of \( t \) from 0 to \( t_{final} \)

- use \( t_{final} \) of order \((d \log d)/n\)
- Initialize with \( \theta_0 \) drawn from a product prior \( p_0(\theta) \), such as normal(0, I) or a product of standard Cauchy
- Starting from the random \( \theta_0 \) define the optimization path \( \theta_t \) such that its distribution tracks the target density \( p_t \)
Optimization path

- **Adaptive Annealing**: Arrange $\theta_t$ from the evolving density
  \[ p_t(\theta) = e^{tJ(\theta)-c_t}p_0(\theta) \]
  with $\theta_0$ drawn from $p_0(\theta)$

- **State evolution** with vector-valued change function $G_t(\theta)$:
  \[ \theta_{t+h} = \theta_t - hG_t(\theta_t) \]
  or better: $\theta_{t+h}$ is the solution to
  \[ \theta_t = \theta_{t+h} + hG_t(\theta_{t+h}), \]
  with small step-size $h$, such that $\theta + hG_t(\theta)$ is invertible with a positive definite Jacobian, and solves equations for the evolution of $p_t(\theta)$.

- As we will see there are many such change functions $G_t(\theta)$, though not all are nice.
Solve for the change $G_t$ to track the density $p_t$

- **Density evolution:** by the Jacobian rule
  \[
p_{t+h}(\theta) = p_t(\theta + hG_t(\theta)) \det(I + h\nabla G_t^T(\theta))
  \]

  Up to terms of order $h$
  \[
p_{t+h}(\theta) = p_t(\theta) + h \left[ (G_t(\theta))^T \nabla p_t(\theta) + p_t(\theta) \nabla^T G_t(\theta) \right]
  \]

- In agreement for small $h$ with the partial diff equation
  \[
  \frac{\partial}{\partial t} p_t(\theta) = \nabla^T [G_t(\theta)p_t(\theta)]
  \]

- The right side is $G_t^T(\theta) \nabla p_t(\theta) + p_t(\theta) \nabla^T G_t(\theta)$. Dividing by $p_t(\theta)$ it is expressed in the log density form
  \[
  \frac{\partial}{\partial t} \log p_t(\theta) = \nabla^T G_t(\theta) + G_t^T(\theta) \nabla \log p_t(\theta)
  \]
Four candidate solutions

Four solutions to the partial differential equation at time $t$

$$\frac{\partial}{\partial t} p_t(\theta) = \nabla^T[G(\theta)p_t(\theta)]$$

1. Solution of smallest $L_2$ norm of $G(\theta)p(\theta)$ in which $G(\theta)p(\theta)$ is a gradient
2. Solution in which pairs of coordinates of $G(\theta)p(\theta)$ are 2–dim gradients
3. Solution of smallest $L_2$ norm of $G(\theta)$ in which $G$ is a gradient
4. Approximate solutions expressed in terms of $u_i = X_i^T \theta$. 
Solution of smallest $L_2$ norm of $G_t(\theta)p_t(\theta)$ at a specific $t$.

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $\text{green}_d(\theta)$ proportional to $1/\|\theta\|^{d-2}$, harmonic for $\theta \neq 0$.

The partial diff equation becomes the Poisson equation:

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

- Solution

$$b(\theta) = (f \star \text{green})(\theta)$$
Candidate solution 1.

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

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $\text{green}_d(\theta)$ proportional to $1/||\theta||^{d-2}$, harmonic for $\theta \neq 0$.

The partial diff equation becomes the Poisson equation:

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

- Solution, using $\nabla \text{green}_d(\theta) = c_d \theta/||\theta||^d$

$$\nabla b(\theta) = (f * \nabla \text{green}_d)(\theta)$$
Candidate solution 1.

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

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

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

- Solution, using $\nabla \text{green}_d(\theta) = c_d \frac{\theta}{\|\theta\|^d}$

$$G_t(\theta)p_t(\theta) = (f \ast \nabla \text{green}_d)(\theta)$$
Candidate solution 1.

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

- Let $G_t(\theta)p_t(\theta) = \nabla b(\theta)$, gradient of a function $b(\theta)$
- Let $f(\theta) = \frac{\partial}{\partial t} p_t(\theta)$
- Set $\text{green}_d(\theta)$ proportional to $1/\|\theta\|^{d-2}$, harmonic for $\theta \neq 0$.
- The partial diff equation becomes the Poisson equation:
  \[ \nabla^T [G_t(\theta)p_t(\theta)] = f(\theta) \]
- Solution, using $\nabla \text{green}_d(\theta) = c_d \theta/\|\theta\|^d$
  \[ G_t(\theta) = \frac{(f * \nabla \text{green}_d)(\theta)}{p_t(\theta)} \]
- Not nice. Convolution is a high-dimensional integral.
Candidate solution 2.

Solution using 2–dimensional convolutions

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

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

• Pair consecutive terms to achieve a portion of the solution

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

• Solution, for each consecutive pair of coordinates,

$$\begin{bmatrix} G_{t,j}(\theta) \\ G_{t,j+1}(\theta) \end{bmatrix} = \frac{2}{d} \left( f \ast \nabla \text{green}_2 \right)(\theta) \frac{\nabla}{p_t(\theta)}$$

The 2–dim Green’s function gradient acts on $(\theta_j, \theta_{j+1})$.

• Solution computed numerically. Stable for particular objective functions $J$ and initial distributions $p_0$. 

Andrew Barron

Computationally feasible greedy algorithms for neural nets
Candidate solution 2.

Solution using 2–dimensional convolutions

- Solution, for each consecutive pair of coordinates,

\[
\begin{bmatrix}
G_{t,j}(\theta) \\
G_{t,j+1}(\theta)
\end{bmatrix} = \frac{2 (f \ast \nabla \text{green}_2)(\theta)}{d} p_t(\theta)
\]

- Stable for particular objective functions \( J \)?
- For \( p_0 \) we use a product of 2–dimensional circularly symmetric Cauchy distributions
- Stable if \( J(\theta) \) can exhibit only small change by changing two consecutive coordinates
- True for sigmoids with coeff squashing and variable replication. Terms \( \phi(a^T X) \) represented using small \( \eta \) as

\[
\phi \left( \eta \sum \phi(\theta_{j,r})X_{j,r} \right)
\]

The internal \( \phi \) is an increasing sigmoid squashing real \( \theta_{j,r} \) into \((-1, 1)\). For each \( X_j \) the aggregate coefficient is

\[
a_j = \eta \sum_{r=1}^{\text{rep}} \phi(\theta_{j,r})
\]
Candidate solution 3.

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

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

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

then becomes an elliptic pde in $b_t(\theta)$ for fixed $t$.

- With $\nabla \log p_t(\theta)$ and $\frac{\partial}{\partial t} \log p_t(\theta)$ 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 3.

Ideal solution of smallest $L_2$ norm of $G_t(\theta)$

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

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

then becomes an elliptic pde in $b_t(\theta)$ for fixed $t$.

- With $\nabla \log p_t(\theta)$ and $\frac{\partial}{\partial t} \log p_t(\theta)$ 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)
- To achieve explicit solution give up $G_t(\theta)$ being a gradient
- For ridge bases, we decompose into a system of first order differential equations and integrate
Candidate solution 4 by decomposition of ridge sum

- Optimize $J(\theta) = \sum_{i=1}^{n} r_i \phi(X_i^T \theta)$
- Target density $p_t(\theta) = e^{tJ(\theta) - c_t} p_0(\theta)$ with $c'_t = E_{p_t}[J]$
- The time score is $\frac{\partial}{\partial t} \log p_t(\theta) = J(\theta) - E_{p_t}[J]$
- Specialize the pde in log density form
  \[ \nabla^T G_t(\theta) + G_t^T(\theta) \nabla \log p_t(\theta) = J(\theta) - E_{p_t}[J] \]
- The right side takes the form of a sum
  \[ \sum r_i [\phi(X_i^T \theta) - a_i]. \]
- Likewise $\nabla \log p_t(\theta) = t \nabla J(\theta) + \nabla \log p_0(\theta)$ is the sum
  \[ \sum X_i \left[ t r_i \phi'(X_i^T \theta) - (1/n)(X_i^T \theta) \right] \]
- from the Gaussian initial distribution with log $p_0(\theta)$ equal to
  \[ -(1/2n) \sum \theta^T X_i X_i^T \theta \]
Approximate solution for ridge sums

- Seek approximate solution of the form

\[ G_t(\theta) = \sum x_i \|x_i\|^2 g_i(u) \]

with \( u = (u_1, \ldots, u_n) \) evaluated at \( u_i = X_i^T \theta \), for which

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

- 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 \( g_i(u) \)
- Arrange \( g_i(u) \) to solve the differential equations

\[ \frac{\partial}{\partial u_i} g_i(u) + g_i(u)\left[t r_i \phi'(u_i) - u_i/n + \text{rest}_i\right] = r_i\left[\phi(u_i) - a_i\right] \]

where \( \text{rest}_i = \sum_{j \neq i}[t r_j \phi'(u_j) - u_j/n]x_j^T x_i/\|x_i\|^2 \).
Integral form of solution

- Differential equation for \( g_i(u_i) \), suppressing dependence on the coordinates other than \( i \)
  \[
  \frac{\partial}{\partial u_i} g_i(u_i) + g_i(u_i) \left[ t \cdot r_i \phi'(u_i) - \frac{u_i}{n} + \text{rest}_i \right] = r_i \left[ \phi(u_i) - a_i \right]
  \]

- Define the density factor
  \[
  m_i(u_i) = e^{t \cdot r_i \phi(u_i) - \frac{u_i^2}{2n} + u_i \cdot \text{rest}_i}
  \]

- Allows the above diff equation to be put back in the form
  \[
  \frac{\partial}{\partial u_i} [g_i(u_i) \cdot m_i(u_i)] = r_i \left[ \phi(u_i) - a_i \right] m_i(u_i)
  \]

- An explicit solution, evaluated at \( u_i = x_i^T \theta \), is
  \[
  g_i(u_i) = r_i \int_{c_i}^{u_i} m_i(\tilde{u}_i) \left[ \phi(\tilde{u}_i) - a_i \right] d\tilde{u}_i
  \]
  \[
  m_i(u_i)
  \]

  where \( c_i \) is such that \( \phi(c_i) = a_i \).
The derived change function $G_t$ for evolution of $\theta_t$

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

$$r_i \int_{c_i}^{u_i} e^t r_i (\phi(\tilde{u}_i) - \phi(u_i)) - (\tilde{u}_i^2 - u_i^2) / 2n + t(\tilde{u}_i - u_i)\text{rest}_i(u) \left[ \phi(\tilde{u}_i) - a_i \right] d\tilde{u}_i$$

- Evaluating at $u = X\theta$ we have the change function

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

for which $\theta_t$ evolves according to

$$\theta_{t+h} = \theta_t + h G_t(\theta_t)$$

- For showing $g_{i,t}$, $G_t$ and $\nabla G_t$ are nice, assume the activation function $\phi$ and its derivative is bounded (e.g. a logistic sigmoid or a sinusoid).

- Run several optimization paths in parallel, starting from independent choices of $\theta_0$. Allows access to empirical computation of $a_{i,t} = E_p[\phi(x_i^T \theta_t)]$.
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 $nd \log d$ 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 $\theta_t$ whose distribution closely tracks the target

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

such that, with high probability, the solutions paths have instances of $J(\theta_t)$ 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 $\Phi = \{\phi(\theta^T x) : \theta \in \mathbb{R}^d\}$.
Flexible approximation models
- Subset selection
- Nonlinearly parameterized bases as with neural nets
- $\ell_1$ 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 with a moderate sample size

Computational challenges are being addressed by
- Nonlinear power methods
- Adaptive annealing