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

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

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

Univariate activation functions

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

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

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 $H_\ell(z)$

Products or ridge form builds multivariate activation functions

Flexible multivariate function approximation: $d > 1$

By internally parameterized models & nonlinear least squares

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

  using continuous powers, frequencies or thresholds

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

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

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

- Internal parameter vector $\underline{a}$ 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: $\Phi_{(n)} = \left\{ (\phi(\underline{X}_i, \underline{\theta}))_{i=1}^n \ : \ \underline{\theta} \in \Theta \right\} \subset R^n$
- 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 one-shot optimization

- Sample version

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

- Population version

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

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

GREEDY OPTIMIZATIONS

- Step 1: Choose c_1, $\underline{\theta}_1$ to achieve $\min \| Y - c\phi_{\underline{\theta}} \|^2_{(n)}$ or
  - sample version: $max_{\theta}(1/n)\sum_{i=1} nY_i\phi(\underline{X}_i, \underline{\theta})$
  - population version: $\max_{\theta} E[f(X)\phi(\underline{X}, \underline{\theta})]$
- Step $m > 1$: Arrange

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

with α_m, c_m, $\underline{\theta}_m$ chosen to achieve

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

- Also acceptable, with $R_i = Y_i - \hat{f}_{m-1}(\underline{X}_i)$,
  - Choose $\underline{\theta}_m$ to achieve $\max_{\underline{\theta}} \sum_{i=1}^{n} R_i\, \phi(\underline{X}_i, \underline{\theta})$
  - Population version: $max_{\theta}\, E[R(X)\phi(\underline{X}, \underline{\theta})]$
- Foward stepwise selection of $S_m = \{\phi_{\underline{\theta}_1}, \ldots, \phi_{\underline{\theta}_m}\}$. Given S_{m-1}, choose θ_m to $\min_{\theta} d(Y, span\{\phi_{\underline{\theta}_1} \ldots \phi_{\underline{\theta}}, \phi_{\underline{\theta}_m}\})$

Basic $m-$term approximation and computation bounds

For either one-shot or greedy approximation
(B. *IT* 1993, Lee et al *IT* 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:

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

  where $\|f\|_\Phi$ is the variation of f with respect to Φ (as will be defined on the next slide).

ℓ_1 norm on coefficients in representation of f

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

$$\Big|\sum_j c_j \,\mathrm{sgn}(\theta_{0,j} + \theta_{1,j}x_1 + \ldots + \theta_{d,j}x_d)\Big| \leq \sum_j |c_j|$$

equality if $\underline{x}$ is in polygon where $\mathrm{sgn}(\underline{\theta}_j \cdot \underline{x}) = \mathrm{sgn}(c_j)$ for all j

- Motivates the norm

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

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

ℓ_1 norm on coefficients in representation of f

- Finite sum representations, $f(\underline{x}) = \sum_j c_j \phi(\underline{x}, \underline{\theta}_j)$. Variation $\|f\|_\Phi = \sum_j |c_j|$, which is the ℓ_1 norm of the coefficients in representation of f in the span of Φ
- Infinite integral representation $f(\underline{x}) = \int e^{i\,\underline{\theta}\cdot\underline{x}}\,\tilde{f}(\underline{\theta})\,d\theta$ (Fourier representation), for $\underline{x}$ in a unit cube. The variation $\|f\|_\Phi$ is bounded by an L_1 spectral norm:

$$\|f\|_{cos} = \int_{R^d} |\tilde{f}(\underline{\theta})|\,d\underline{\theta}$$

$$\|f\|_{step} \leq \int |\tilde{f}(\underline{\theta})|\;\|\underline{\theta}\|_1\,d\underline{\theta}$$

$$\|f\|_{ramp} \leq \int |\tilde{f}(\underline{\theta})|\;\|\underline{\theta}\|_1^2\,d\underline{\theta}$$

- As we said, this $\|f\|_\Phi$ appears in the numerator of the approximation bound.

Statistical Risk

- The population accuracy of function estimated from sample
- 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

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

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

Adaptation

- Adapt network size m and choice of internal parameters
- Minimum Description Length Principle leads to
  Complexity penalized least squares criterion.
  Let $\hat{m}$ achieve

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

- Information-theoretic risk bound

$$E \| \hat{f}_{\hat{m}} - f \|^2 \; \le \; \min_{m} \left\{ \| f_m - f \|^2 + 2c \frac{m}{n} \log N(\Phi, \delta) \right\}$$

  - Performs as well as if the best m^* were known in advance.
  - $\|f\|_{\Phi}^2/m$ 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 $\underline{\theta}_m$ achieving within a constant factor of the maximum of

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

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

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

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*):

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

can be whitened and a quadratic power method finds a_k.

Scoring a Ridge Function

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

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

- 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.

Scoring a Ridge 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

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

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*

Scoring a Ridge Function

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

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

  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.
- $\cos(z)$, $\sin(z)$ 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 $G_k(a_k, a) = c_k e^{-(1/2)\|a_k - a\|^2}$

Scoring a Ridge Function

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

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

  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.
- Hermite polynomial case, with $X \sim$ Normal(0, I): $H_\ell(a^T X)$ and $H_{\ell'}(a_k^T X)$ orthogonal for $\ell' \neq \ell$, and

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

Nonlinear Power Method

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

Analysis via Whitening

- 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

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

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 the a_k and a expressed via α_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

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

Adaptive Annealing:

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

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

  along a sequence of values of t from 0 to t_{final}
- use t_{final} of order $(d \log d)/n$
- Initialize with θ_0 drawn from a product prior $p_0(\underline{\theta})$, such as normal$(0, I)$ or a product of standard Cauchy
- Starting from the random θ_0 define the optimization path θ_t such that its distribution tracks the target density p_t

Optimization path

- Adaptive Annealing: Arrange θ_t from the evolving density

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

  with θ_0 drawn from $p_0(\theta)$

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

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

  or better: θ_{t+h} is the solution to

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

  with small step-size h, such that $\underline{\theta} + h \, G_t(\underline{\theta})$ 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

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

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

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

- 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

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

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 $u_i = X_i^T \theta$.

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 $green_d(\theta)$ 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

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

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 $green_d(\theta)$ 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 $\nabla green_d(\theta) = c_d\, \theta/\|\theta\|^d$

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

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 $green_d(\theta)$ 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 $\nabla green_d(\theta) = c_d\, \theta/\|\theta\|^d$

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

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 $green_d(\theta)$ 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 $\nabla green_d(\theta) = c_d \, \theta / \|\theta\|^d$

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

- Not nice. Convolution is a high-dimensional integral.

Candidate solution 2.

Solution using 2−dimensional convolutions

- Write the pde $\nabla^T[G_t(\theta)p_t(\theta)] = f(\theta)$ 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,

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

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

- Solution computed numerically. Stable for particular objective functions J and initial distributions p_0?

Candidate solution 2.

Solution using 2−dimensional convolutions

- Solution, for each consecutive pair of coordinates,

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

- 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 η as

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

The internal ϕ is is an increasing sigmoid squashing real $\theta_{j,r}$ into $(-1, 1)$. For each X_j the aggregate coefficient is $a_j = \eta \sum_{r=1}^{rep} \phi(\theta_{j,r})$

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

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

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

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

- 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

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

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

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

- 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

$$\frac{\partial}{\partial u_i}[g_i(u_i)\, m_i(u_i)] \;=\; r_i\big[\phi(u_i) - a_i\big]\, m_i(u_i)$$

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

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

  where c_i is such that $\phi(c_i) = a_i$.

The derived change function G_t for evolution of θ_t

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

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

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

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

  for which θ_t evolves according to

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

- 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 θ_0. Allows access to empirical computation of $a_{i,t} = E_{p_t}[\phi(x_i^T \theta_t)]$

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 θ_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 $\Phi = \{\phi(\theta^T x) : \theta \in R^d\}$

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 with a moderate sample size
- Computational challenges are being addressed by
  - Nonlinear power methods
  - Adaptive annealing