Learning Sparse Polynomials

over product measures

Kiran Vodrahalli
knv2109@columbia.edu

Columbia University

December 11, 2017

The Problem

"Learning Sparse Polynomial Functions" [Andoni, Panigrahy, Valiant, Zhang '14]

Consider learning a polynomial $f : \mathbb{R}^n \to \mathbb{R}$ of degree d of k monomials. Key features of setting:

- real-valued (in contrast to many works considering $f : \{-1, 1\}^n \to \{-1, 1\}$)
- "sparse" (only k monomials)
- distribution over data x: Gaussian or uniform
  - only consider product measures
- realizable setting: assume we try to exactly recover the polynomial

Why this setting?

- notion of "low-dimension" in sparsity
- Boolean settings are hard (parity functions)

We outline the results of Andoni et. al. '14 in this talk.

Background and Motivation

computation and sample complexities

Goal: Learn the polynomial in time and samples $< o(n^d)$.

- many approaches for learning take sample/computation time $\mathcal{O}(n^d)$
- polynomial kernel regression in $\binom{n}{d}$-sized basis
  - sample complexity: same as linear regression (depends linearly on dimension, in this case n^d)
  - computation complexity: worse than n^d
- compressed sensing in $\binom{n}{d}$
  - $f(x) := \langle v, x^{\otimes d} \rangle$ where v is k-sparse, x is data
  - sub-linear complexity results only hold for particular settings of data (RIP, incoherence, nullspace property)
  - unclear if these hold for $X^{\otimes d}$ (probably not)
- dimension reduction + regression (ex: principal components regression) — note this is *improper learning*

The Results

sub-$\mathcal{O}(n^d)$ samples and computation

Two key results: oracle setting and learning from samples.

Definition

Inner product $\langle h_1, h_2 \rangle$ is defined with respect to a distribution D over the data X as $\mathbb{E}_D [h_1(x)h_2(x)]$. We also have $\|h\|^2 = \langle h, h \rangle$.

Definition

A correlation oracle pair calculates $\langle f^*, f \rangle$ and $\langle (f^*)^2, f \rangle$ where f^* is the true polynomial.

- in the oracle setting, can exactly learn polynomial f^* in $\mathcal{O}(k \cdot nd)$ oracle calls
- if learning from samples $(x, f^*(x))$, learn $\hat{f}$ s.t. $\|\hat{f} - f\| \leq \epsilon$:
  - sample complexity: $\mathcal{O}(\text{poly}(n, k, 1/\epsilon, m))$
  - $m = 2^d$ if D uniform, $m = 2^{d \log d}$ if D Gaussian
  - computation complexity: (# samples)$\cdot\mathcal{O}(nd)$
  - $(x, f^*(x) + g), g \sim \mathcal{N}(0, \sigma^2)$: same bounds $\times$ poly$(1 + \sigma)$

Methodology

overview of `Growing-Basis`

Key idea: Greedily build a polynomial in an orthonormal basis, one basis function at a time. Identify first the existence of variable x_i using correlation, and then find its degree in the basis function. This strategy will work for the following reasons:

- We can work in an orthonormal basis and pay a factor 2^d increase in the sparsity of the representation.
- We can identify the degree of a variable in a particular basis function by examining the correlation of several basis functions with $(f^*)^2$ in an iterative fashion. This search procedure takes time $\mathcal{O}(nd)$.

Methodology

orthogonal polynomial bases over distributions

Definition

Consider inner product space $\langle \cdot, \cdot \rangle_D$ for distribution D, where $D = \mu^{\otimes n}$ is a product measure over $\mathbb{R}^n$. For any coordinate, we can find an orthogonal basis of polynomials depending on distribution D by Gram-Schmidt. Let $H_t(x_i)$ be the degree t basis function for variable x_i. Then for $T = (t_1, \cdots, t_n)$ such that $\sum_i t_i = d$, $H_T(x) = \prod_i H_{t_i}(x_i)$ defines the orthogonal basis function parametrized by T in the product basis.

Thus we can write

$$f^*(x) := \sum_T \alpha_T H_T(x)$$

for any polynomial f^*. There are at most $k2^d$ terms in the sum.

Methodology

algorithm

Algorithm 1 `Growing-Basis`

procedure GROWING-BASIS(degree $d, \langle \cdot, f^* \rangle, \langle \cdot, (f^*)^2 \rangle$)
  $\hat{f} := 0$
  while $\langle 1, (f^* - \hat{f})^2 \rangle > 0$ **do**
    $H := 1, B := 1$
    for $r = 1, \cdots, n$ **do**
      for $t = d, \cdots, 0$ **do**
        if $\langle H \cdot H_{2t}(x_r), (f^* - \hat{f})^2 \rangle > 0$ **then**
          $H := H \cdot H_{2t}(x_r)$, $B := B \cdot H_t(x_r)$
          break out of double loop.
        end if
      end for
    end for
    $\hat{f} := \hat{f} + \langle B, f^* \rangle \cdot B$
  end while
  return $\hat{f}$

Methodology

sparsity in orthogonal basis

We give a lemma which allows us to work in an orthogonal basis without blowing up the sparsity too much.

Lemma

Suppose f^* is k-sparse in product basis H_1. Then it is $k2^d$ sparse in product basis H_2.

Proof.

Write each term $H^{(1)}_{t_i}(x_i)$ of f^* in basis H_1 in basis H_2: each will have t_i terms. Since each monomial term in H_1 is a product of such $H_{t_i}(x_i)$, there will be $\prod_i (t_i + 1) \leq 2^{\sum_i t_i} \leq 2^d$ terms for each monomial. Since there are k monomials, there are at most $k2^d$ terms when expressed in H_2. $\square$

Methodology

detecting degrees (1)

We now give a lemma which suggests the correctness of the search procedure used in `Growing-Basis`.

Lemma

Let d_1 denote the maximum degree of variable x_1 in f^*. Then, $\langle H_{2t}(x_1), (f^*)^2(x) \rangle > 0$ iff $t \leq d_1$.

Proof.

We have

$$(f^*)^2(x) = \sum_T \alpha_T^2 \prod_{i=1}^{n} H_{t_i}(x_i)^2 + \sum_{T \neq U} \alpha_T \alpha_U \prod_{i=1}^{n} H_{t_i}(x_i) H_{u_i}(x_i)$$

Note that if $t > t_1$, $H_{t_1}(x_1)^2$ will only be supported on basis functions $H_0, \cdots, H_{2t_1}$. This set does not include H_{2t} since $2t > 2t_1$, so $\langle H_{2t}(x_1), H_{t_1}(x_1)^2 \rangle = 0$. Likewise for second term if $t > u_1$, thus, if $t > d_1$, correlation is zero. If $t = d_1$, the correlation is nonzero for the first term, but zero for the second term. $\square$

Methodology

detecting degrees (2)

Let's get some intuition.

$$(f^*)^2(x) = \sum_T \alpha_T^2 \prod_{i=1}^n H_{t_i}(x_i)^2 + \sum_{T \neq U} \alpha_T \alpha_U \prod_{i=1}^n H_{t_i}(x_i) H_{u_i}(x_i)$$

Let's look at

$$\left\langle H_{2t}(x_1), \prod_{i=1}^n H_{t_i}(x_i)^2 \right\rangle = \left\langle H_{2t}(x_1), \prod_{i=1}^n \left(1 + \sum_{j=1}^{2t_i} c_{t,j} H_j(x_i)\right) \right\rangle$$

Since $t_i = t$ (for T such that $t_1 = d_1$), the coefficient of the term $H_{2t}(x_1) \prod_{i=2}^n H_0(x_i)$ is the only thing that remains since everything else will get zeroed out. Then just sum over T such that $t_1 = d_1$. The second term does not contribute since either $i \neq 1$ or $t_i + u_i < 2t$ since $u_i \neq t_i$.

$$\left\langle H_{2t}(x_1), \prod_{i=1}^n H_{t_i}(x_i) H_{u_i}(x_i) \right\rangle = 0$$

Methodology

detecting degrees (3)

Thus, it makes sense that if we proceed from the largest degree possible, we will be able to detect the degree of x_1 in one of the basis functions in the representation of f^*. With some more analysis of a similar flavor, we extend this to finding a complete product basis representation.

- Key idea: lexicographic order
  - example: $1544300 \succeq 1544000$ since $0 < 3$.
  - we will use to compare degree lists T and U, which correspond to basis functions H_T, H_U.
- We can essentially proceed inductively.
- Recap: Suppose f^* contains basis functions $H_{t_1}(x_1), \cdots, H_{t_r}(x_r)$. Then, check $\langle H_{2t_1,\cdots,2t_r,t,0,\cdots,0}(x), f^*(x)^2 \rangle > 0$ for $t = d \to 0$. Assign $t_{r+1} := t^*$ such that t^* is the first value making the correlation > 0.

Methodology

sampling version

In the sampling situation, we only get data points $\{(z_i, f^*(z_i)\}_{i=1}^m$ and no oracle. We will run the same algorithm, replacing the oracles with an emulated version.

- Have to emulate correlation oracle: $\hat{C}(f) = \frac{1}{m}\sum_{i=1}^m f(z_i)f^*(z_i)^2$.
- Chebyshev inequality suffices to bound $m = \mathcal{O}\left(\frac{1}{\epsilon^2}\mathbb{E}\left[f^2(f^*)^4\right]\right) < \mathcal{O}\left(\frac{\max_f \mathbb{E}\left[f^2(f^*)^4\right]}{\epsilon^2}\right)$ to get a constant probability bound.
- Can repeat $\log(1/\delta)$ times and take the median to boost the probability of success to $1-\delta$.
- For the noisy case, compute correlation up to 4^{th} moments instead and apply standard concentration inequalities (subgaussian noise is very standard).

Methodology

getting 2^d sample complexity

To actually get a bound for sample complexity, we bound $\max_f \mathbb{E}\left[f^2(f^*)^4\right]$ assuming a uniform distribution $[-1,1]^n$.

- Legendre orthogonal polynomials for this distribution
- Fact: $|H_{d_i}(x_i)| \leq \sqrt{2d_i+1}$.
- Thus: $|H_S(x)| = \prod_i |H_{S_i}(x_i)| \leq \prod_i \sqrt{2S_i+1} \leq \prod_i 2^{S_i} \leq 2^d$.
- Thus: $|f^*(x)| = |\sum_S \alpha_S H_S(x)| \leq 2^d \sum_S |\alpha_S|$.
- By Parseval (Pythagorean thm. for inner product spaces), $\sum_S \alpha_S^2 = 1$. Since f^* is k-sparse, $\sum_S |\alpha_S| \leq \sqrt{k}$.
- Thus $|f^*(x)| \leq 2^d\sqrt{k}$.
- Thus $f(x)^2 f^*(x)^4 \leq 2^{6d}k^2$ if f^* is degree d and f is represented in a degree $2d$ basis.

Key Takeaways

proof methodology

The key methodology in the proof has the following properties:

- relies heavily on orthogonal properties of polynomials
- is "term-by-term": we examine and find each basis function one at a time.
- achieves 2^d dependence because
  - transforming to an orthogonal basis only causes 2^d blow-up in sparsity
  - fact about Legendre polynomials (for uniform distribution)
- weakness: relies heavily on product distribution assumption in order to construct orthogonal polynomial bases over n variables.

Thank you for your attention!