Deterministic Approximate Counting for juntas of degree-2 PTFs

Anindya De
University of California, Berkeley

Ilias Diakonikolas
U. Edinburgh

Rocco Servedio
Columbia U.

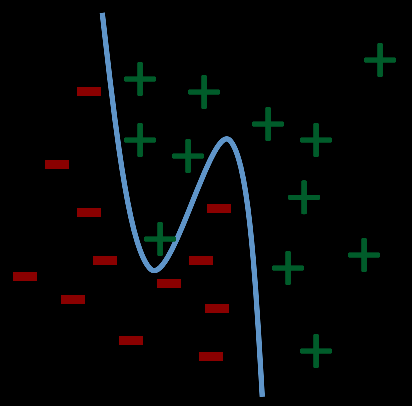
PTFs and LTFs

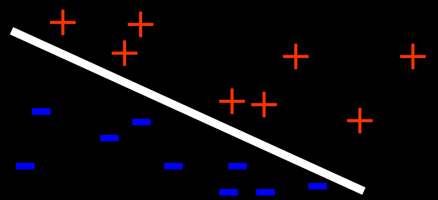
Degree-d polynomial threshold function (PTF): sign of a degree-d polynomial

$$f : \{-1,1\}^n \to \{-1,1\}$$

$$f = sign(p(x_1,\ldots,x_n))$$

where $deg(p) = d$





Juntas of degree-2 PTFs

Input : k degree-2 PTFs $f_1, \ldots, f_k$

where $f_1, \ldots, f_k : \{\pm 1\}^n \to \{0,1\}$

and $g : \{0,1\}^k \to \{0,1\}$.

Task : Deterministically approximate (up to error ϵ) the quantity :

$$\Pr_{x \in \{-1,1\}^n} [g(f_1(x), \ldots, f_k(x)) = 1]$$

The Challenge ..

Deterministically approximate the quantity

$$\Pr_{x \in \{-1,1\}^n} [g(f_1(x), \ldots, f_k(x)) = 1]$$

in time $poly(n) \cdot h(k, \epsilon)$.

Motivation

- Previous talk ☺
- Counting versions of all *self-respecting* decision problems are #P-hard.
- This motivates study of approximate counting.

Motivation

- If the problem is really *self-respecting*: Deciding if the number of satisfying assignments is non-zero is itself NP-hard.

- This rules out efficient multiplicative algorithms.

- Of course, there is a trivial random sampling algorithm for additive approximation.

Motivation

- As in the previous talk, we would like to get efficient deterministic algorithms for additive approximation.

- Circuit lower bounds => every efficient randomized algorithm can be derandomized.

- While proving lower bounds isn't in reach, we should *at least* try to prove its consequences.

Approximate counting of PTFs

- For LTFs : [SVV, GKM] - $poly(n, 1/\epsilon)$ time deterministic counting with multiplicative error.

- For PTFs of degree 2: Last talk –
  $poly(n) \cdot 2^{poly(1/\epsilon)}$ for degree 2 for additive error ϵ

- What about richer classes of functions ?

Approximate counting of juntas of LTFs

- Gopalan, O'Donnell, Wu, Zuckerman : Deterministic approximate counting for k-juntas of halfspaces - $n^{O(k+\log(k/\epsilon))}$.

- For $\epsilon = \log^{-o(1)} n$, the running time is $2^{k^{O(1)}} \cdot poly(n)$.

- What about functions of PTFs?

Approximate counting of functions of degree-2-PTFs

- Diakonikolas, Kane, Nelson – Deterministic approximate counting k-juntas of degree-2 PTFs in time $n^{O(k \cdot poly(1/\epsilon))}$ over $\mathcal{N}^n(0,1)$.

- Slightly worse dependence on k for the Boolean hypercube.

- Thus, for any $k = \omega(1)$ or $\epsilon = o(1)$, the running time of the algorithm is super-polynomial in n .

Main result

Theorem: There is an algorithm running in deterministic time $poly(n) \cdot h(k, \epsilon)$ which given a function $g : \{0,1\}^k \to \{0,1\}$ and k degree-2 PTFs $f_1, \ldots, f_k : \{-1,1\}^n \to \{0,1\}$, outputs a number ν such that

$$\left| \nu - \Pr_{x \in \{\pm 1\}^n} [g(f_1(x), \ldots, f_k(x)) = 1] \right| \leq \epsilon$$

Main technical result

Theorem: There is an algorithm running in deterministic time $poly(n) \cdot h(k, \epsilon)$ which given a function $g : \{0,1\}^k \to \{0,1\}$ and k degree-2 PTFs $f_1, \ldots, f_k : \mathbb{R}^n \to \{0,1\}$, outputs a number ν such that

$$\left| \nu - \Pr_{x \sim \mathcal{N}^n(0,1)} [g(f_1(x), \ldots, f_k(x)) = 1] \right|$$

Technique for proving main result

- Prove the result over the distribution $\mathcal{N}^n(0,1)$.

- Following the previous talk : Multi-dimensional Invariance principle (Mossel) shows that the same result holds over $\{-1,1\}^n$ for k regular degree-2 polynomials.

Technique for proving main result

- We next prove a *new* regularity lemma: Given k degree-2 PTFs, we show that we can construct a decision tree of depth $c(k, \epsilon)$ such that w.h.p. over the leaves of the decision tree : If all the variables appearing on the path from the root to the leaf are restricted, then the resulting k degree-2 PTFs are all regular.

Regularity lemma

- For k=1, results due to DSTW, MZ, HKM implied this.

- For LTFs (with k>1), GOWZ provided such a regularity lemma.

- The new regularity lemma follows arguments similar to DSTW.

Thus, it boils down to ...

Theorem: There is an algorithm running in deterministic time $poly(n) \cdot h(k, \epsilon)$ which given a function $g : \{0,1\}^k \to \{0,1\}$ and k degree-2 PTFs $f_1, \ldots, f_k : \mathbb{R}^n \to \{0,1\}$, outputs a number ν such that

$$\left| \nu - \underset{x \in \mathcal{N}^n(0,1)}{\mathbf{Pr}} [g(f_1(x), \ldots, f_k(x))] \right| \leq \epsilon$$

Roadmap

- State a new CLT.
- Show how the CLT is useful for approximate counting when some *nice* conditions are met.
- Show how the general case can be decomposed to a combination of CLT + brute force.

Proof of the main technical result

- We prove a new CLT for the joint distribution of k degree-2 polynomials which have ``small eigenvalues".
- Recall that for any degree-2 polynomial
  $p(x) = x^T Ax + \langle B, x\rangle + C$
  we define $\lambda_{\max}(p) = \sigma_{\max}(A)$.

New Central Limit Theorem

Theorem : Let $p_1, \ldots, p_k : \mathcal{N}^n(0,1) \to \mathbb{R}$ be k degree 2 polynomials such that :

- $\forall i \in [k] \quad \lambda_{\max}(p_i) \leq \epsilon$,
- $\forall i \in [k] \quad Var(p_i) \leq 1$,
- $\exists i \in [k] \quad Var(p_i) \geq c$.

Let $Z \sim (p_1, \ldots, p_k)$ and $Z' \sim \mathcal{N}(\mu, \Sigma)$ where

$$\mu = \mathbf{E}[Z] \quad ; \quad \Sigma = \mathbf{Cov}[Z].$$

Then,

$$d_K(Z, Z') \leq \frac{k^{2/3} \cdot \epsilon^{1/6}}{c^{1/6}}.$$

Remarks about the CLT

- The k-dimensional Kolmogorov distance between Z and Z' is defined to be:

$$\sup_{\theta_1,\ldots,\theta_k\in\mathbb{R}} \left|\Pr[\forall i \in [k],\ Z_i \leq \theta_i] - \Pr[\forall i \in [k],\ Z'_i \leq \theta_i]\right|$$

- The case k=1 follows by Berry-Esseen theorem.

- Why is this CLT useful?

Applying the CLT

Let $f_1, \ldots, f_k : \mathbb{R}^n \to \{0,1\}$ be k degree-2 PTFs where $f_i = sign(p_i)$ satisfying the following:

- $Var(p_i) = 1$ and $\lambda_{\max}(p_i) \leq \epsilon$.

If $g = \mathrm{AND}$, then we need to compute

$$\Pr_{x \in \mathcal{N}^n(0,1)} [f_1(x) \wedge \ldots \wedge f_k(x)]$$

$$= \Pr_{x \sim \mathcal{N}^n(0,1)} [p_1(x) \geq 0 \wedge \ldots \wedge p_k(x) \geq 0]$$

Applying the CLT

- However,

$$\Pr[p_1(x) \geq 0 \wedge \ldots \wedge p_k(x) \geq 0] \approx \Pr[Z_1 \geq 0 \wedge \ldots \wedge Z_k \geq 0]$$

where $(Z_1, \ldots, Z_k)$ are jointly normal with with the same mean and covariance as the distribution of $(p_1, \ldots, p_k)$.

Applying the CLT

However, $\Pr[Z_1 \geq 0 \wedge \ldots \wedge Z_k \geq 0]$ can be computed to good accuracy in time $k^{O(k)}$.

Thus, if the eigenvalues of all the polynomials are small enough ($\leq \epsilon^6/k^4$), then we're done …

Decomposition

- So, what happens if some of the polynomials have large eigenvalues …

- To understand the idea behind the strategy, consider a *toy* case where the polynomials $p_1, \; p_2, \; \ldots, \; p_k$ are diagonalizable in the same basis.

Toy case : Diagonalization

- In other words,

$$p_1 = \sum_{j=1}^{n} \alpha_{1j} L_j(x)^2 + \sum_{j=1}^{n} \beta_{1j} L_j(x) + C_1$$

$$\vdots$$

$$p_k = \sum_{j=1}^{n} \alpha_{kj} L_j(x)^2 + \sum_{j=1}^{n} \beta_{kj} L_j(x) + C_k$$

Renaming linear forms

Here $L_1(x), \ldots, L_n(x)$ forms an orthonormal basis. Since, Gaussians are invariant under orthogonal transformations, we can rewrite

$$p_1 = \sum_{j=1}^{n} \alpha_{1j} y_j^2 + \sum_{j=1}^{n} \beta_{1j} y_j + C_1$$

$$\vdots$$

$$p_k = \sum_{j=1}^{n} \alpha_{kj} y_j^2 + \sum_{j=1}^{n} \beta_{kj} y_j + C_k$$

Applying GOWZ

- If $\max_i \lambda_{\max}(p_i) \leq \epsilon$, then it translates to saying that $\max_{i \in [k]} \max_{j \in [n]} |\alpha_{ij}| \leq \epsilon$.

- If this condition is not satisfied, then following the analysis of GOWZ, it can be shown that there is a small set L $(|L| \leq k/\epsilon^2)$ such that for any p_i at least one of the following is true :

The two cases

- With high probability, over the restriction of the variables in L, $sign(p_i)$ is close to being constant.

- After the restriction of the variables in L, $\lambda_{\max}(p_i)/Var(p_i) \leq \epsilon$.

Win-win analysis

- Win-win analysis : First, restrict all the variables in L. For each $i \in [k]$, we end up with one of the following:

(i) Either $sign(p_i)$ is close to a constant.

(ii) $\lambda_{\max}(p_i)$ is small compared to its variance implying that we can apply the CLT.

All this can clearly be done in time $poly(n) \cdot h(k, \epsilon)$

Decomposition

- However, we're in a more complicated situation i.e. all of $p_1, \ldots, p_k$ may not be diagonalizable in the same basis …

- What's the way out ??

Decomposition

- The key concept used is that of *renaming linear forms.* In other words, consider a function $F(x_1, \ldots, x_n)$. Given any linear form $L_1(x)$ such that $\|L_1(x)\|_2 = 1$, consider an orthonormal completion $\{L_1, \ldots, L_n\}$.

  Then, $F(x_1, \ldots, x_n)$ can be re-expressed as $G(L_1(x), \ldots, L_n(x))$ where the distribution of $L_1(x), \ldots, L_n(x)$ is $\mathcal{N}^n(0, 1)$.

Steps in the decomposition

- Either the conditions of the CLT is met or without loss of generality, we can assume $\lambda_{\max}(p_1) \geq \epsilon$.

- This means that there is a linear form $L_1(x)$ such that if $p_1 = \alpha_1 L_1(x)^2 + \beta_1 \cdot L_1(x) \cdot r_1 + q_1$ where q_1 and r_1 are independent of $L_1(x)$ and $Var(q_1) \leq 1 - \epsilon$.

Restricting a linear form

- Using the concept of renaming a linear form, we can consider *restricting* on all possible values of $L_1(x)$.

- We continue recursively until all the q_i satisfy:

(i) Either $\lambda_{\max}(q_i)/Var(q_i) \leq \epsilon$,

(ii) Or $Var(q_i) \leq \epsilon^2$.

- This can go on for at most $\tilde{O}(k/\epsilon^2)$ steps.

Our Central Limit Theorem

Theorem : Let $p_1, \ldots, p_k : \mathcal{N}^n(0,1) \to \mathbb{R}$ be k degree 2 polynomials such that :

- $\forall i \in [k] \quad \lambda_{\max}(p_i) \le \epsilon$,
- $\forall i \in [k] \quad Var(p_i) \le 1$,
- $\exists i \in [k] \quad Var(p_i) \ge c$.

Let $Z \sim (p_1, \ldots, p_k)$ and $Z' \sim \mathcal{N}(\mu, \Sigma)$ where

$$\mu = \mathbf{E}[Z] \quad ; \quad \Sigma = \mathbf{Cov}[Z].$$

Then,

$$d_K(Z, Z') \le \frac{k^{2/3} \cdot \epsilon^{1/6}}{c^{1/6}}.$$

Proof sketch

- Key word (i): Stein's method
- Key word (ii) : Malliavin calculus

Stein's method

- Easy to show that for every absolutely continuous f with bounded f', if Z denotes the standard normal, then

$$\mathbf{E}[Z \cdot f(Z)] = \mathbf{E}[f'(Z)]$$

Proof : Integration by parts

Stein's method

Converse : If for a random variable Z it holds that for every absolutely continuous f with bounded f', $\mathbf{E}[Z \cdot f(Z)] = \mathbf{E}[f'(Z)]$, then Z is the standard normal.

Proof : Some basic ODE (not difficult).

Stein's method

- Is this characterization robust?

Lemma : For any random variable W,

$$d_{TV}(W, \mathcal{N}(0,1)) \leq \sup_{f \in \mathcal{F}} |\mathbf{E}[f'(W) - W \cdot f(W)]|$$

where $\mathcal{F} = \{f : \|f\| \leq \sqrt{\pi/2},\ \|f'\| \leq 2\}$.

Stein's method

- Is this characterization robust?

Lemma : For any random variable W,

$$d_W(W, \mathcal{N}(0,1)) \leq \sup_{f \in \mathcal{F}} |\mathbf{E}[f'(W) - W \cdot f(W)]|$$

where $\mathcal{F} = \{f : \|f\|,\ \|f'\|,\ \|f''\| \leq 2\}$.

Stein's method

- Similar characterization available for closeness to multivariate normal.

- To explain the gist of the idea, we will just focus on the univariate case.

Stein's method

- Assume $W = p(x_1, \ldots, x_n)$ where $x_1, \ldots, x_n \sim \mathcal{N}(0,1)$.

- Suppose, we want to show that $d_{TV}(W, \mathcal{N}(0,1))$ is small.

- All we need to do is to bound $\sup_{f \in \mathcal{F}} |\mathbf{E}[f'(W) - W \cdot f(W)]|$.

Enter Malliavin Calculus …

- In a nutshell, it allows us to take derivatives of functions of stochastic processes.

- Informally, if the *chance parameter* is ω, we are taking a derivative w.r.t. ω .

Malliavin calculus

Let $F : \mathbb{R}^n \to \mathbb{R}$ where the domain is equipped with the $\mathcal{N}^n(0,1)$ measure. The Malliavin derivative operator D maps F to a $\mathbb{R}^n$ valued random variable where $DF_i = \dfrac{\partial F}{\partial x_i}$.

To see why it is the derivative, we need to consider functions of Brownian motion

Malliavin calculus

Malliavin derivatives satisfy some nice properties:

For every $h \in \mathbb{R}^n$, let $W(h) = \sum_{i=1}^{n} h_i x_i$.

Then, $\mathbf{E}[F \cdot W(h)] = \mathbf{E}[\langle DF, h \rangle]$.
(Integration by parts)

Stein meets Malliavin

(Nualart and Peccati): The fundamental relation between Stein's method and Malliavin derivatives:

$$\mathbf{E}[f'(W) - W \cdot f(W)] = \mathbf{E}[f'(W)(1 - \langle DW, -DL^{-1}W \rangle)]$$

Here L^{-1} is an operator which attenuates the q^{th} level of the Hermite expansion by $(-1/q)$.

Stein meets Malliavin

Recall

$$\mathbf{E}[f'(W) - W \cdot f(W)] = \mathbf{E}[f'(W)(1 - \langle DW, -DL^{-1}W\rangle)]$$

It is easy to show that
$$Var(W) = 1 \implies \mathbf{E}[\langle DW, -DL^{-1}W\rangle] = 1$$

Since the f appearing in Stein's method always satisfies $\|f'\| \leq 2$, hence by Cauchy-Schwartz,

$$|\mathbf{E}[f'(W) - W \cdot f(W)]| \leq \sqrt{Var(\langle DW, -DL^{-1}W\rangle)} \cdot$$

Stein meets Malliavin

Thus, it all boils down to controlling the variance of the quantity $\langle DW, DL^{-1}W \rangle$.

Stein meets Malliavin

For closeness to multivariate normal, things are slightly more complicated.

Let us define $\mathcal{H} = \{h : \mathbb{R}^k \to \mathbb{R} : \|h''\| < 1\}$. Let $(Z_1, \ldots, Z_k)$ be a Gaussian distribution with the same mean and covariance as $(W_1, \ldots, W_k)$.

Stein meets Malliavin

Key result (Nourdin and Peccati)

$$|\mathbf{E}[h(Z_1,\ldots,Z_k)] - \mathbf{E}[h(W_1,\ldots,W_k)]| = O(k^2\epsilon)$$

where $\sup_{i,j} Var(\langle DW_i, -DL^{-1}W_j\rangle) \leq \epsilon$.

Our result

We show that if $W_i = F_i(X_1, \ldots, X_n)$ where F_i are degree-2 polynomials with $Var(F_i) = 1$ and $\lambda_{\max}(F_i) \leq \epsilon$, then

$$\sup_{i,j} Var(\langle DW_i, -DL^{-1}W_j \rangle) \leq \epsilon .$$

Proof : calculation + Matrix analysis

Our result

This proves closeness of $(Z_1, \ldots, Z_k)$ and $(W_1, \ldots, W_k)$ w.r.t. class of test functions $\mathcal{H}$.

To prove closeness in Kolmogorov distance, we need closeness w.r.t. the class

$$\mathcal{H}_K = \{(x_1 \leq \theta_1) \wedge \ldots \wedge (x_k \leq \theta_k) : \theta_1, \ldots, \theta_k \in \mathbb{R}\}$$

Mollification

To go from closeness in class $\mathcal{H}$ to closeness in class $\mathcal{H}_K$, we do the following steps:

- ✓ Show that closeness in $\mathcal{H}$ implies closeness in class $\widetilde{\mathcal{H}_K}$ where $\widetilde{\mathcal{H}_K}$ is a smoothened version of $\mathcal{H}_K$ (uses mollification machinery)
- ✓ Carbery-Wright shows that closeness in $\widetilde{\mathcal{H}_K}$ implies closeness in $\mathcal{H}_K$.

Recap …

Theorem : Let $p_1, \ldots, p_k : \mathcal{N}^n(0,1) \to \mathbb{R}$ be k degree 2 polynomials such that :

- $\forall i \in [k] \quad \lambda_{\max}(p_i) \leq \epsilon$,
- $\forall i \in [k] \quad Var(p_i) \leq 1$,
- $\exists i \in [k] \quad Var(p_i) \geq c$.

Let $Z \sim (p_1, \ldots, p_k)$ and $Z' \sim \mathcal{N}(\mu, \Sigma)$ where

$$\mu = \mathbf{E}[Z] \quad ; \quad \Sigma = \mathbf{Cov}[Z].$$

Then,

$$d_K(Z, Z') \leq \frac{k^{2/3} \cdot \epsilon^{1/6}}{c^{1/6}}.$$

Recap …

- The CLT allows us to do approximate counting as long as all the $\lambda_{\max}(p_i)$ are small.

- If some of the $\lambda_{\max}(p_i)$ are large, then we can apply the decomposition method to reduce the counting to CLT + brute force.

- Apply the regularity lemma to move from $\mathcal{N}^n(0,1)$ to the Boolean hypercube.

THANKS