CSE 291: Analysis of Polynomial Threshold Functions Spring 2015

Lecture 8: Pseudorandom Generators for PTFs; Noise Sensitivity

Instructor: Danie Kane **Date:** Apr. 28

1 Recap

Last time we have covered regularity lemma.

Lemma 1 (Regularity Lemma). *Any polynomial p that takes input from $\{-1,+1\}^n$ can be written as a decision tree of depth $O(\frac{1}{\tau}(\ln\frac{1}{\tau\epsilon})^{O(d)})$, with leaves as polynomials p_ρ, such that the following holds. With probability $1-\epsilon$ over a random leaf, the associated polynomial p_ρ is either (1) τ-regular, or (2) $var(p_\rho) \leq \epsilon\|p_\rho\|_2^2$.*

Remark. In the latter case, p_ρ can be thought of as a large constant plus small variations, where its sign stays constant (-1 or +1).

We have also defined pseudorandom generators(PRGs) of PTFs, which will be the main focus of this lecture. Formally, our goal is to explicitly construct a simple (low entropy) random variable Y, so that for any degree-d PTF over n variables,

$$|\mathbb{E}f(Y) - \mathbb{E}_{B\in_U\{\pm1\}^n}f(B)| \leq \epsilon \tag{1}$$

As we have seen, there is an implicit construction with seedlength as small as $O(d\ln n + \ln\frac{1}{\epsilon})$, but this requires a computationally inefficient random sampling.

2 The Construction of Meka-Zuckerman PRG

The main tool we will be using is k-wise independence.

Definition 1. *A sequence of random variables $(W_1,\ldots,W_n)$ are k-wise independent if any k of them are independent.*

Note that d-wise independent random variables fools in expectation all degree-d polynomials. But fooling degree-d PTFs appears to be a harder task. We begin with a standard fact.

Fact 1. *We can generate a set of k-wise independent random variables $(W_1,\ldots,W_n)$, with $W_i \in_U \{1,2,\ldots,m\}$, from a seed of length $O(k\ln(nm))$.*

The construction is as follows. We first use Fact 1 to build $F:[n]\to[t]$ where $(F(1),\ldots,F(n))$ are 2-wise independent, requiring a seedlength of $O(2\ln(nt))$. Next, we repeatedly use Fact 1 t times, to build independent random vector $Z_1,\ldots,Z_t \in \{\pm1\}^n$, where each Z_i is a random vector whose n coordinates are k-wise independent. This step requires a seed length of $t\cdot O(k\ln(2n))$. We put each Z_i as a row vector and concatenate them vertically to an array:

$$\begin{bmatrix} Z_{1,1}, & Z_{1,2}, & \ldots, & Z_{1,n} \\ Z_{2,1}, & Z_{2,2}, & \ldots, & Z_{2,n} \\ Z_{t,1}, & Z_{t,2}, & \ldots, & Z_{t,n} \end{bmatrix}$$

Our pseudorandom number generated is defined as:

$$(Y_1,\ldots,Y_n) = (Z_{F(1),1}, Z_{F(2),2}, \ldots, Z_{F(n),n})$$

In the next section, we claim that with decent size of k and t, Equation (1) can be established.

3 Idea of Analysis

Step 1: Replacement Method Conditioned on F. First we fix F. For any degree-d polynomial p, without loss of generality, assume $\|p\|_2 = 1$, since scaling does not change the sign. $p(Y_1, \ldots, Y_n)$ can be written as a degree-d polynomial p_F of $Z_1, \ldots Z_t$ (therefore there are $nt - n$ dummy variables). Consider $f(\cdot) = \text{sign}(p_F(\cdot))$. The high level idea is to use replacement method, as we have seen in the invariance principle. In particular, we are going to show

$$\mathbb{E}f(Z_1, \ldots, Z_t) \approx \mathbb{E}f(G_1, \ldots, G_t) \tag{2}$$

$$\mathbb{E}f(B_1, \ldots, B_t) \approx \mathbb{E}f(G_1, \ldots, G_t) \tag{3}$$

where $(G_1, \ldots, G_t)$ are independent standard Gaussians, and $(B_1, \ldots, B_t)$ are independent uniform Bernoullis. We start with Equation (2).

For notational simplicity, Let $Z = (Z_1, \ldots, Z_{i-1})$, $B = (B_1, \ldots, B_{i-1})$ and $G = (G_{i+1}, \ldots, G_t)$. With foresight, we find smooth functions $\rho_+, \rho_- : \mathbb{R} \to [-1, +1]$ such that $\rho_+(x) = \rho_-(x) = \text{sign}(x)$ except for a small interval of length $O((\epsilon/d)^d)$ and $\rho_-(x) \leq \text{sign}(x) \leq \rho_+(x)$, with $\|\rho^{(4)}\|_\infty \leq O((\epsilon/d)^{-4d})$. Consider $\rho \in \{\rho_+, \rho_-\}$. Our goal now comes down to bounding

$$\mathbb{E}[\rho(p(Z, G, Z_i)) - \rho(p(Z, G, G_i))]$$

Now consider $W = Z_i$ or $W = G_i$. Let $p_0(Z, G, W) = \mathbb{E}_W p(Z, G, W)$. Since p is multilinear, the conditional expectation is the same in either case and W simply become a dummy variable in p_0. A Taylor expansion of ρ yields

$$\mathbb{E}\rho(p(Z, G, W)) = \mathbb{E}[\text{ polynomial of degree 3 in } p(Z, G, W)\ ] + O(\|\rho^{(4)}\|_\infty \mathbb{E}(p(Z, G, W) - p_0(Z, G, W))^4)$$

If we set $k = 4d$, then and the variables among the set (Z, G, W) are $4d$-wise independent. Consequently the first terms are the same in both cases, since the variables among the set (Z, G, W) are $3d$-wise independent. Now consider the second term. Since $(p(Z, G, W) - p_0(Z, G, W))^4$ is a polynoimal of degree $4d$, Z_i's can be safely replaced with B_i's when computing the expectation.

In either case, by hypercontractivity (over a hybrid of Gaussians and Bernoullis),

$$\mathbb{E}(p(B, G, W) - p_0(B, G, W))^4 \leq 2^{O(d)}(\mathbb{E}(p(B, G, W) - p_0(B, G, W))^2)^2$$

We expand p in Fourier domain:

$$p(x) = \sum_{S \subseteq [n]} \hat{p}(S) x^S$$

Note that F determines a partition of $[n]$; for example, if we replace Z_i with G_i, only a subset of arguments of p are affected. We call The set of variable *the ith bucket* with respect to F, abbreviated as $B(i)$. Using this notation, it can be seen that

$$p_0(x) = \sum_{S \subseteq [n] : S \cap B(i) = \emptyset} \hat{p}(S) x^S$$

Therefore,

$$\begin{aligned}
& \mathbb{E}|p(X) - p_0(X)|^2 \\
= & \sum_{S \subseteq [n] : S \cap B(i) \neq \emptyset} \hat{p}(S)^2 \\
\leq & \sum_{S \subseteq [n]} \sum_{j \in S \cap B(i)} \hat{p}(S)^2 \\
= & \sum_{j \in B(i)} \sum_{S \in [n] : j \in S} \hat{p}(S)^2 \\
= & \sum_{j \in B(i)} \text{Inf}_j(p)
\end{aligned}$$

As a result, the total sum of each individual term in (**??**) can be bounded as

$$\mathbb{E}[f(Z_1,\dots,Z_t)-f(G_1,\dots,G_t)]|\le O(\epsilon)+O((\epsilon/d)^{-4d})\sum_{i=1}^{t}2^{O(d)}(\sum_{j\in B(i)}\mathrm{Inf}_j(p))^2$$

where the first term comes from replacing $\rho_+(\cdot)$ ($\rho_-(\cdot)$) with $\mathrm{sign}(\cdot)$ using Carbery-Wright, the second term comes from the bound of $\mathbb{E}[\rho(p(Z,G,Z_i))-\rho(p(Z,G,G_i))]$, as we have just shown above.
Similarly,

$$\mathbb{E}[f(B_1,\dots,B_t)-f(G_1,\dots,G_t)]|\le O(\epsilon)+O((\epsilon/d)^{-4d})\sum_{i=1}^{t}2^{O(d)}(\sum_{j\in B(i)}\mathrm{Inf}_j(p))^2$$

Therefore:

$$\mathbb{E}[f(B_1,\dots,B_t)-f(Z_1,\dots,Z_t)]|\le O(\epsilon)+O((\epsilon/d)^{-4d})\sum_{i=1}^{t}2^{O(d)}(\sum_{j\in B(i)}\mathrm{Inf}_j(p))^2 \tag{4}$$

Step 2: Averaging Over F. Now, taking the expectation over the random choice of F on Equation (4), the second term can be bounded as follows:

$$\begin{aligned}
& O\Big((\epsilon/d)^{-4d}\cdot\mathbb{E}_F[\sum_{i=1}^{t}2^{O(d)}(\sum_{j\in B(i)}\mathrm{Inf}_j(p))^2]\Big)\\
\le\ & O\Big((\epsilon/d)^{-4d}2^{O(d)}\cdot\mathbb{E}_F[\sum_{j,k\in[n]:F(j)=F(k)}\mathrm{Inf}_j(p)\mathrm{Inf}_k(p)]\Big)\\
\le\ & O\Big((\epsilon/d)^{-4d}2^{O(d)}\cdot\mathbb{E}_F[\sum_{j=1}^{n}\mathrm{Inf}_j(p)^2+\frac{1}{t}\sum_{j,k=1}^{n}\mathrm{Inf}_j(p)\mathrm{Inf}_k(p)]\Big)
\end{aligned}$$

where the first inequality follows from the definition of $B(\cdot)$, the second inequality is by the 2-wise independence in each of F's coordinates. To summarize, the error is bounded by

$$O(\epsilon)+O\Big((\epsilon/d)^{-4d}2^{O(d)}\cdot\mathbb{E}_F[\sum_{j=1}^{n}\mathrm{Inf}_j(p)^2+\frac{1}{t}\sum_{j,k=1}^{n}\mathrm{Inf}_j(p)\mathrm{Inf}_k(p)]\Big) \tag{5}$$

Let $\tau=(\max_j\mathrm{Inf}_j(p))/(\sum_j\mathrm{Inf}_j(p))$ be the regularity parameter of p. Then

$$\sum_{j=1}^{n}\mathrm{Inf}_j(p)^2\le\tau\sum_{j=1}^{n}\mathrm{Inf}_j(p)\le\tau d\cdot\mathrm{var}(p)\le\tau d$$

In the meantime,

$$\frac{1}{t}\sum_{j,k=1}^{n}\mathrm{Inf}_j(p)\mathrm{Inf}_k(p)\le\frac{1}{t}(\sum_{j=1}^{n}\mathrm{Inf}_j(p))^2\le\frac{(d\cdot\mathrm{var}(p))^2}{t}\le\frac{d^2}{t}$$

Step 3: Applying the Regularity Lemma. At this point, it may be tempting to set $\tau=\tau_0=O((\frac{\epsilon}{d})^{4d+1})$ and $t=O(\frac{1}{\epsilon}(\frac{\epsilon}{d})^{-4d})$, which lets us conclude the total expected error (5) is bounded by $O(\epsilon)$. But in general this bound on τ may not be true.

Fortunately there is a quick fix: apply regularity lemma on p. Essentially, p can be written as a decision tree of depth $D=\tilde{O}(\tau_0^{-1})$ such that with probability $1-\epsilon$, a random leaf is either (1) τ_0-regular, or (2) a constant plus small variation term. Now we modify our setting of k from $4d$ to $D+4d$, ensuring even after D levels of variable conditioning along the path of the tree, the remaining variables are still $4d$-wise independent. In case (1), the previous result now can be applied to ensure the expected error on each leaf is at most ϵ. In case (2), the sign of p over this leaf is a constant, thus has constant sign (-1 or +1). Applying the previous result to every leaf and averaging let us conclude the result.

Seedlength. In summary, $t = O(\frac{1}{\epsilon}(\frac{\epsilon}{d})^{-4d})$ and $k = \tilde{O}((\frac{\epsilon}{d})^{-(4d+1)})$. Therefore the total number of seedlength is

$$O(tk\ln(2n)) + O(2\ln(nt)) = O\Big(\frac{1}{\epsilon}(\frac{d}{\epsilon})^{8d}(\ln\frac{1}{\epsilon})^{O(d)}\ln n\Big) = \tilde{O}\Big((\frac{d}{\epsilon})^{O(d)}\ln n\Big)$$

We emphasize that this is still a nontrivial PRG, since its seedlength is $O(\ln n)$.

Remark. The state of the art right now is $O_d(\epsilon^{-12}\ln n)$ for PTFs, although for LTFs, a construction with seedlengh $O(\ln\frac{n}{\epsilon}\ln\ln\frac{n}{\epsilon})$ has been shown. As of PRG for Gaussians, we can do a lot better: the best results so far are $O_{d,c}(\epsilon^{-c}\ln n)$ for arbitrary $c>0$ and $O(2^{O(d)}\epsilon^{-5}\ln n)$. A recent result of seedlength polylog$(\frac{n}{\epsilon})$ for $d=2$ has been shown.

4 Noise Sensitivity

Bernoulli and Gaussian Noise Sensitivity. Consider a boolean function $f:\mathbb{R}^n \to \{\pm 1\}$. The noise sensitivity of f measures how likely small changes with input to f leads to small changes of output. This is opposite to the notion of stability we have seen.

Definition 2. *$NS_\epsilon(f)$, the noise sensitivity of f, is defined as*

$$NS_\epsilon(f) = \Pr_{(X,Y)}(f(X) \neq f(Y))$$

where (X,Y) are $(1-\epsilon)$-correlated Bernoullis.
$GNS_\epsilon(f)$, the Gaussian noise sensitivity of f, is defined as

$$GNS_\epsilon(f) = \Pr_{(X,Y)}(f(X) \neq f(Y))$$

where (X,Y) are $(1-\epsilon)$-correlated Gaussians.

It is instructive to look at $\mathrm{NS}_\epsilon(f)$ in Fourier domain. In particular,

$$\begin{aligned} NS_\epsilon(f) &= \frac{1-\mathbb{E}f(X)f(Y)}{2} \\ &= \frac{1-\mathrm{Stab}_{1-\epsilon}(f)}{2} \\ &= \frac{1-\mathbb{E}f(X)(T_{1-\epsilon}f)(X)}{2} \\ &= \frac{1-\sum_{S\subseteq[n]}\hat{f}(S)^2(1-\epsilon)^{|S|}}{2} \\ &= \sum_{S\subseteq[n]}\hat{f}(S)^2\frac{1-(1-\epsilon)^{|S|}}{2} \end{aligned}$$

Roughly, if f has large high degree Fourier coefficients, $\mathrm{NS}_\epsilon(f)$ is likely to be high. Similar to the operator T_ρ we have seen in Bernoulli case, we can define operator U_ρ in Gaussian case. Formally, for $0 \le \rho \le 1$, $U_\rho f$ is the function from $\mathbb{R}$ to $\mathbb{R}$ such that

$$(U_\rho f)(x) = \mathbb{E}[f(Y)|X = x]$$

where Y is ρ-correlated with X, that is, $Y = \rho X + \sqrt{1-\rho^2}Z$ where Z is a standard Gaussian independent of X. What does this operator do in the Fourier domain?

Lemma 2. *For $0 < \rho \le 1$, function f that has Fourier expansion $f = \sum_{\boldsymbol{a}} c_{\boldsymbol{a}} h_{\boldsymbol{a}}$, we have*

$$U_\rho f = \sum_{\boldsymbol{a}} \rho^{\|\boldsymbol{a}\|_1} c_{\boldsymbol{a}} h_{\boldsymbol{a}}$$

Specifically,

$$U_\rho h_{\boldsymbol{a}} = \rho^{\|\boldsymbol{a}\|_1} h_{\boldsymbol{a}}$$

Proof. First we show that $(\{U_{e^{-s}} : s \ge 0\}, \circ)$ is a semigroup. To check associativity, it suffices to show

$$U_{e^{-s}} U_{e^{-t}} = U_{e^{-(s+t)}}$$

This follows from straightforward calculations:

$$\begin{aligned}
& U_{e^{-s}}[(U_{e^{-t}} f)(x)] \\
= \; & U_{e^{-s}}[\mathbb{E}_{A\sim N(0,I)} f(e^{-t}x + \sqrt{1-e^{-2t}}A)] \\
= \; & \mathbb{E}_{A\sim N(0,I), B\sim N(0,I)} f(e^{-s-t}x + e^{-s}\sqrt{1-e^{-2t}}A + \sqrt{1-e^{-2s}}B) \\
= \; & \mathbb{E}_{N\sim N(0,I)} f(e^{-(s+t)}x + \sqrt{1-e^{-2(s+t)}}N) \\
= \; & (U_{e^{-(s+t)}} f)(x)
\end{aligned}$$

where in the first equality we introduce a standard Gaussian A, in the second inequality we introduce a standard Gaussian B independent of A.

Now consider $f = \sum_{\mathbf{a}} c_{\mathbf{a}} h_{\mathbf{a}}$. We would like to find the representation of $g_t = U_{e^{-t}} f = \sum_{\mathbf{a}} c_{\mathbf{a}}(t) h_{\mathbf{a}}$. Note that $U_{e^{-0}} f = f$, thus, in this notation, $c_{\mathbf{a}}(0) = c_{\mathbf{a}}$. We take derivative of g_t with respect to t. First note that

$$\frac{\mathrm{d}}{\mathrm{d}t} g_t = \frac{\mathrm{d}}{\mathrm{d}t} U_{e^{-t}} f = \frac{\mathrm{d}}{\mathrm{d}s} U_{e^{-(s+t)}} f\Big|_{s=0} = \frac{\mathrm{d}}{\mathrm{d}s} U_{e^{-s}} (U_{e^{-t}} f)\Big|_{s=0} = \frac{\mathrm{d}}{\mathrm{d}s} U_{e^{-s}} g_t\Big|_{s=0}$$

Then

$$\begin{aligned}
(U_{e^{-s}} g_t)(x) &= \mathbb{E}_Y g(e^{-s}x + \sqrt{1-e^{-2s}}Y) \\
&= \mathbb{E}_Y g_t((1 - s + O(s^2))x + \sqrt{2s + O(s^2)}Y) \\
&= \mathbb{E}_Y g_t(x - sx + \sqrt{2s}Y + O(s^{3/2})) \\
&= \mathbb{E}_Y g_t(x) + \nabla g_t(x) \cdot (-sx + \sqrt{2s}Y) + \frac{1}{2} \sum_{i,j} \frac{\partial^2 g}{\partial x_i \partial x_j} 2s Y_i Y_j + O(s^{3/2}) \\
&= g_t(x) + \nabla g_t(x) \cdot (-sx) + s\nabla^2 g_t + o(s^{3/2})
\end{aligned}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}s} U_{e^{-s}} g_t\Big|_{s=0} = -\nabla g_t + \nabla^2 g_t = L g_t$$

where L is the differential operator we have seen in the alternative definition of Hermite polynomials. Recall that $Lh_{\mathbf{a}} = -\|a\|_1 h_a$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t} g_t = L g_t = -\sum_{\mathbf{a}} c_{\mathbf{a}}(t) \|\mathbf{a}\|_1 h_{\mathbf{a}}$$

On the other hand,

$$\frac{\mathrm{d}}{\mathrm{d}t} g_t = L g_t = \sum_{\mathbf{a}} c'_{\mathbf{a}}(t) h_{\mathbf{a}}$$

By uniqueness of Fourier expansion, $c'_{\mathbf{a}}(t) = -\|\mathbf{a}\|_1 c_{\mathbf{a}}(t)$. In conjunction with the initial condition $c_{\mathbf{a}}(0) = c_{\mathbf{a}}$, we get $c_{\mathbf{a}}(t) = c_{\mathbf{a}} e^{-\|\mathbf{a}\| t}$. Thus for all $t \geq 0$,

$$U_{e^{-t}} f = \sum_{\mathbf{a}} c_{\mathbf{a}} e^{-t\|\mathbf{a}\|_1} h_{\mathbf{a}}$$

That is,

$$U_\rho f = \sum_{\mathbf{a}} c_{\mathbf{a}} \rho^{\|\mathbf{a}\|_1} h_{\mathbf{a}}$$

□

Using the above lemma, we see that exactly analogous to the Bernoulli case,

$$\mathrm{GNS}_\epsilon(f) = \frac{1 - \mathbb{E} f(X) U_{(1-\epsilon)} f(X)}{2} = \sum_{\mathbf{a}} \hat{f}(\mathbf{a}) \frac{1 - (1-\epsilon)^{\|\mathbf{a}\|_1}}{2}$$

Average Sensitivity. The notion of average sensitivity measures the total influence of coordinates. In particular,

$$\mathrm{AS}(f) = \sum_{i=1}^{n} \mathrm{Inf}_i(f) = n \Pr(f(X) \neq f(X'))$$

where X is a uniform Bernoulli random variable, X' differs from X on one single randomly chosen coordinate.

Gaussian Surface Area. Consider $f : \mathbb{R}^n \to \{\pm 1\}$. $S = \{x \in \mathbb{R}^n : f(x) = +1\}$. The Gaussian surface area, $\Gamma(f)$ is defined as:

$$\Gamma(f) = \lim_{\varepsilon \to 0} \frac{\Pr(X : X \text{ is within } \varepsilon \text{ Euclidean distance of } \partial S)}{2\varepsilon}$$

We expect "nice" surfaces of ∂S(e.g. f is a PTF). In this case, a equivalent definition is through integral over the surface area:

$$\Gamma(f) = \int_{\partial S} \phi(x) \mathrm{d}\sigma$$

where $\phi(x)$ is the Gaussian pdf.