CAS CS 599 B: Mathematical Methods for TCS

Lecturer: Mark Bun Spring 2022

Lecture Notes 9:

Concentration Inequalities, Small-Bias Distributions

Reading.

- Vadhan, Pseudorandomness, §3.5

1 Derandomizing Concentration Inequalities

Chernoff-Hoeffding bounds tell us that sums of fully independent random variables are tightly concentrated around their means. It turns out that these concentration bounds hold even when there is only limited independence between the summands.

Proposition 1. *Let $X_1, \ldots, X_t$ be pairwise independent random variables in $[0,1]$ and let $\bar{X} = \frac{1}{t}\sum_{i=1}^{t} X_i$. Then for every $\varepsilon > 0$,*

$$\Pr[|\overline{X} - \mathbb{E}[\overline{X}]| \geq \varepsilon] \leq \frac{1}{t\varepsilon^2}.$$

Proof. We'll end up applying Chebyshev's inequality to $\overline{X}$. To do this, we'll bound its variance. Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \mathbb{E}[\overline{X}]$. Then

$$\begin{aligned}
\mathrm{Var}\left[\overline{X}\right] &= \mathbb{E}[(\overline{X} - \mu)^2] \\
&= \frac{1}{t^2}\,\mathbb{E}\left[\left(\sum_{i=1}^{t} X_i - \mu_i\right)^2\right] \\
&= \frac{1}{t^2}\sum_{i,j}\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \\
&= \frac{1}{t^2}\sum_{i=1}^{t}\mathbb{E}[(X_i - \mu_i)^2] \\
&= \frac{1}{t^2}\sum_{i=1}^{t}\mathrm{Var}\left[X_i\right] \\
&\leq \frac{1}{t}.
\end{aligned}$$

Applying Chebyshev's inequality:

$$\Pr[|\overline{X} - \mu| \geq \varepsilon] \leq \frac{1}{\varepsilon^2} \cdot \mathrm{Var}\left[\overline{X}\right]$$

gives the result. □

For k-wise independent random variables, we can obtain concentration exponential in k for some settings of parameters:

Proposition 2. *Let $X_1, \dots, X_t$ be k-wise independent random variables in $[0,1]$ and let $\bar{X} = \frac{1}{t}\sum_{i=1}^t X_i$. Then for every $\varepsilon > 0$,*

$$\Pr[|\overline{X} - \mathbb{E}[\overline{X}]| \geq \varepsilon] \leq \left(\frac{k^2}{4t\varepsilon^2}\right)^{k/2}.$$

(This was assigned as an exercise.)

These concentration results are useful for reducing the randomness needed to amplify the error of randomized algorithms. Instead of repeating a randomized algorithm with independent coin tosses, one can repeat it with k-wise independent coin tosses (using much less randomness) and still get efficient error reduction.

To illustrate, suppose A is a randomized algorithm that computes a function f to error at most $1/3$ using m bits of randomness:

$$\forall x \in \{-1,1\}^n \qquad \Pr_{r \sim \{-1,1\}^m}[A(x;r) \neq f(x)] \leq 1/3.$$

Consider the algorithm $\widetilde{A}$ which takes the majority vote of $A(x;r^1), \dots, A(x;r^t)$ for $r^1, \dots, r^t \sim \{-1,1\}^m$ i.i.d. For each $i = 1, \dots, t$, let X_i be the 0-1 indicator for whether $A(x;r^i) \neq f(x)$. Then

$$\begin{aligned}\Pr[\widetilde{A}(x;r^1,\dots,r^t) \neq f(x)] &= \Pr\left[\frac{1}{t}(X_1 + \cdots + X_t) \geq \frac{1}{2}\right] \\ &\leq \Pr\left[\frac{1}{t}(X_1 + \cdots + X_t) - \mathbb{E}\left[\frac{1}{t}(X_1 + \cdots + X_t)\right] \geq \frac{1}{2} - \frac{1}{3}\right] \\ &\leq \exp(-2(1/6)^2 \cdot t)\end{aligned}$$

by Hoeffding. If we want to drive this error probability down to δ, it suffices to take $t = O(\log(1/\delta))$ repetitions. This takes a total of $O(m\log(1/\delta))$ uniform bits.

Meanwhile, if we use Proposition 2 instead, it suffices to take $t = O(k^2\delta^{-2/k})$, for a total of $k \cdot \max\{m, \log t\} = O(km + \log(1/\delta) + k\log k)$ uniform bits. Thus, k-wise independence lets us trade an increase in runtime for a reduction of the multiplicative dependence $m\log(1/\delta)$ to an additive dependence $m + \log(1/\delta)$.

2 Pseudorandom Generators

Let's attempt to capture what we mean when we say that k-wise independent (and other) distributions are pseudorandom. A distribution $\mathcal{D}$ over $\{-1,1\}^n$ is the uniform distribution $\mathcal{U}_n$ iff it looks like the uniform distribution to every statistical test you can run on it, i.e., $\mathbb{E}_{x\sim\mathcal{D}}[f(x)] = \mathbb{E}_{x\sim\mathcal{U}_n}[f(x)]$ for every function $f : \{-1,1\}^n \to \{-1,1\}$. We can also state a robust version of this result:

Definition 3. The *total variation* or *statistical* distance between two distributions $\mathcal{D}, \mathcal{D}'$ over $\{-1,1\}^n$

$$TV(\mathcal{D},\mathcal{D}') = \frac{1}{2}\sum_{x\in\{-1,1\}^n} |\mathcal{D}[x] - \mathcal{D}'[x]| = \max_{E\subseteq\{-1,1\}^n} |\Pr_{x\sim\mathcal{D}}[x \in E] - \Pr_{x\sim\mathcal{D}'}[x \in E]|.$$

The second characterization shows that $\mathcal{D}$ is close to the uniform distribution if and only if it looks close to uniform to ("fools") every statistical test f:

Theorem 4. *For a distribution $\mathcal{D}$ over $\{-1,1\}^n$,*

$$TV(\mathcal{D},\mathcal{U}_n) = \frac{1}{2} \max_{f:\{-1,1\}^n \to \{-1,1\}} \left| \mathop{\mathbb{E}}_{x\sim D}[f(x)] - \mathop{\mathbb{E}}_{x\sim \mathcal{U}_n}[f(x)] \right|.$$

Unfortunately, even for constant ε, fooling every statistical test requires a sample space of size $\Omega(2^n)$. To have any hope of constructing pseudorandom distributions, we restrict the class of tests we care to fool to those of low complexity.

Let $\mathcal{C} \subseteq \{f : \{-1,1\}^n \to \{-1,1\}\}$ be a class of *test functions*.

Definition 5. A distribution $\mathcal{D}$ over $\{-1,1\}^n$ is ε-pseudorandom against a class of tests $\mathcal{C}$ if it ε-fools every $f \in \mathcal{C}$, i.e., for every $f \in \mathcal{C}$,

$$| \mathop{\mathbb{E}}_{x\sim \mathcal{D}}[f(x)] - \mathop{\mathbb{E}}_{x\sim \mathcal{U}}[f(x)] \,| \leq \varepsilon.$$

Example 6. If $\mathcal{D}$ is a k-wise independent distribution, then $\mathcal{D}$ is 0-pseudorandom against the class $\{f \mid \deg(f) \leq k\}$. Question to ponder: Do k-wise independent distributions approximately fool spectrally concentrated functions?

A *pseudorandom generator* is an (efficient) algorithm for sampling from a pseudorandom distribution using only a small number of uniform bits. That is, a PRG is a deterministic algorithm that takes as input a small random seed (of "seed length" ℓ) and stretches it to a longer string of length n that is pseudorandom against a class of tests.

Definition 7. Let $G : \{-1,1\}^\ell \to \{-1,1\}^n$ be a deterministic function. We say that G is an ε-pseudorandom generator against $\mathcal{C}$ if the distribution $G(\mathcal{U}_\ell)$ is ε-pseudorandom against $\mathcal{C}$, i.e., for every $f \in \mathcal{C}$,

$$|\mathbb{E}[f(G(\mathcal{U}_\ell))] - \mathbb{E}[f(\mathcal{U}_n)]| \leq \varepsilon.$$

Example 8. Our construction of a k-wise independent distribution is a 0-PRG for degree-k polynomials with seed length $\ell = O(k \log n)$.

3 Small-Bias Distributions

Besides k-wise independent distributions, "small-bias" distributions are another important primitive class of pseudorandom distributions.

Definition 9. A distribution $\mathcal{D}$ is ε-biased if it is ε-pseudorandom against the class of (nonempty) parity functions, i.e.,

$$\left| \mathop{\mathbb{E}}_{x\sim D}[\chi_S(x)] \right| \leq \varepsilon$$

for all $S \neq \emptyset$.

To describe a small-bias generator (due to Alon, Goldreich, Høastad, and Peralta), we introduce a map $L : \mathbb{F}_{2^t} \times \mathbb{F}_{2^t} \to \{-1,1\}$ that helps us bridge arithmetic in finite fields with the Fourier representation. Define

$$L(x,y) = (-1)^{\langle x,y \rangle}$$

where the inner product uses a representation of $\mathbb{F}_{2^t}$ as a vector space over $\mathbb{F}_2$. The map has the following useful properties:

- $L(x,y) \cdot L(z,y) = (-1)^{\langle x,y\rangle + \langle z,y\rangle} = L(x+z,y)$
- $L(0,y) = 1$ for all y
- For all $x \neq 0$, we have $\mathbb{E}_{y\sim\mathbb{F}_{2^t}}[L(x,y)] = 0$.

Our generator $G : \mathbb{F}_{2^t} \times \mathbb{F}_{2^t} \to \{-1,1\}^n$ is defined by

$$G(x,y) = (L(1,y), L(x,y), \ldots, L(x^{n-1},y)).$$

Our goal is to estimate

$$\begin{aligned}\mathop{\mathbb{E}}_{r\sim G(\mathcal{U}_\ell)}[\chi_S(r)] &= \mathop{\mathbb{E}}_{x,y}\left[\chi_S(L(1,y), L(x,y), \ldots, L(x^{n-1},y))\right] \\ &= \mathop{\mathbb{E}}_{x,y}\left[L\left(\left(\sum_{i\in S} x^i\right) y\right)\right] \\ &=: \mathop{\mathbb{E}}_{x,y}\left[L(p(x),y)\right].\end{aligned}$$

If $p(x) = 0$, the value of L is fixed to 1; otherwise, L is uniformly random. This suggests breaking the expectation up as

$$\begin{aligned}\mathop{\mathbb{E}}_{x,y}[L(p(x),y)] &= \mathop{\mathbb{E}}_{x,y}[L(p(x),y) \mid p(x) = 0] \cdot \Pr_x[p(x) = 0] + \mathop{\mathbb{E}}_{x,y}[L(p(x),y) \mid p(x) \neq 0] \cdot \Pr_x[p(x) \neq 0] \\ &= \Pr_x[p(x) = 0] \\ &\in \left[0, \frac{n-1}{2^t}\right].\end{aligned}$$

Thus, it suffices to take $(n-1)/2^t \leq \varepsilon$, or $t \geq \log((n-1)/\varepsilon)$. This gives a PRG with seed length $2\lceil \log((n-1)/\varepsilon)\rceil$ which is within a constant factor of optimal. For relatively recent progress on improving the leading constant, see Ta-Shma "Explicit, almost optimal, epsilon-balanced codes" and the references therein.

4 Almost k-Wise Independence

At first glance, there may seem to be no relationship between small-bias distributions and k-wise independent distributions. After all, the pairwise independent distribution $(r_S = \chi_S(x))_{S\neq\emptyset}$ we constructed does not fool the parity function, which is constant on this distribution. Meanwhile, the ε-biased distribution we constructed above has nonzero bias on degree-2 parities. Nevertheless, it turns out that ε-biased distributions are automatically *almost* k-wise independent.

Definition 10. Random variables $r_1, \ldots, r_n$ are δ-almost k-wise independent if for every $S \subseteq [n]$ with $|S| = k$ and every $t \in \{-1,1\}^k$,

$$\frac{1}{2}\sum_{t\in\{-1,1\}^k} |\Pr[(r_i)_{i\in S} = t] - 2^{-k}| \leq \delta.$$

That is, for every set of coordinates S, the distribution of the projection $(r_i)_{i\in S}$ is at most δ-far from the uniform distribution on k bits in total variation distance. You may see some authors define almost k-wise independence in terms of the L_∞ distance from each projection to uniform, rather than the L_1 distance we've used here. The two notions are equivalent up to a 2^k factor in δ, which generally translates to an additive k in seed length.

Theorem 11. *If $\mathcal{D}$ is an ε-biased distribution, it is also $(2^{k/2}\varepsilon)$-almost k-wise independent for every k.*

Proof. It suffices to show that for every function $f : \{-1,1\}^n \to \{-1,1\}$ that depends on only k variables, that

$$|\mathbb{E}[f(\mathcal{D})] - \mathbb{E}[f(\mathcal{U}_n)]| \leq 2^{k/2}\varepsilon.$$

Examining the LHS:

$$\begin{aligned}
|\mathbb{E}[f(\mathcal{D})] - \mathbb{E}[f(\mathcal{U}_n)]| &= \left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x) \right] - \hat{f}(\emptyset) \right| \\
&= \left| \sum_{S\subseteq[n]} \hat{f}(S) \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [\chi_S(x)] - \hat{f}(\emptyset) \right| \\
&= \left| \sum_{S\neq\emptyset} \hat{f}(S) \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [\chi_S(x)] \right| \\
&\leq \sum_{S\neq\emptyset} |\hat{f}(S)| \left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [\chi_S(x)] \right| \\
&\leq \varepsilon \sum_{S\neq\emptyset} |\hat{f}(S)|.
\end{aligned}$$

Now we use the fact that f only depends on k variables. By relabeling variables if necessary, we can assume $f(x_1,\ldots,x_n) = g(x_1,\ldots,x_k)$. Then by Cauchy-Schwarz,

$$\sum_{S\neq\emptyset} |\hat{f}(S)| \leq \sum_{S\subseteq[k]} |\hat{g}(S)| \cdot 1 \leq \sqrt{\sum_{S\subseteq[k]} (\hat{g}(S))^2 \cdot \sum_{S\subseteq[k]} 1} = 2^{k/2}.$$

□

Setting $\varepsilon = 2^{-k/2}\delta$, a δ-almost k-wise independent distribution can be sampled using seed length $O(\log(n/\varepsilon)) = O(k+\log n+\log(1/\delta))$. This should be compared to the (optimal) seed length of $O(k\cdot\log n)$ needed to sample a k-wise independent distribution.