Large Deviation Bounds for Decision Trees and Sampling Lower Bounds for $AC^0$ Circuits

Chris Beck, Russell Impagliazzo, Shachar Lovett
History of Sampling Problems

• Earliest uses of randomness in algorithms were for sampling, not decision problems
  – Complexity Theory of Randomized Sampling [Jerrum Valiant Vazirani ‘86]
  – Markov Chain Monte Carlo method [Jerrum Sinclair ‘89]
History of Sampling Problems

• Many researchers observed that just because a function $f$ is hard doesn’t mean that $\{(x, f(x))\}_{x \in \{0,1\}^n}$ is hard to sample uniformly.
  
  – Even when $f =$ integer factorization, we can sample this in polynomial time [Bach’85, Kalai]
  – Even when $f =$ parity, we can still sample this distribution in $AC^0$.  
    [Babai’87, Boppana Lagarias ‘87]
History of Sampling Problems

• The task of explicitly exhibiting a distribution which is hard to sample for concrete models was raised more recently:
Prior Work

• In a CCC ‘2011 paper, Lovett and Viola gave the first result for $AC^0$:

  No distribution sampled by depth $d$ $AC^0$ circuits of size $2^{n^{O(1/d)}}$ with any amount of randomness has statistical distance better than $1 - n^{-\Omega(1)}$ to the uniform distribution over any good code.

This work: Improve to $1 - \exp(-n^{\Omega(1/d)})$. 
Applications to Data Structures

• Lovett and Viola observed that *sampling lower bounds* imply lower bounds for *succinct data structures*. Plugging in our improved bound:

\[
\text{Suppose codewords of an } (n, k, d)\text{-code of constant rate and distance can be stored using } k + r \text{ bits, and recovered by an } \mathcal{AC}^0 \text{ circuit. Then } r \geq n^\varepsilon.
\]
Second Main Result

• We were led to a new concentration bound:

Suppose we have $n$ boolean decision trees of height $n^\varepsilon$, such that no variable expects to be queried by $n^\delta$ trees, on a random input.

  – Almost surely, no variable is queried by $n^{\varepsilon'}$ trees.
  – Almost surely, the number of ones computed is not more than a constant times expectation $+n^{\varepsilon'}$

• We will discuss this in detail shortly.
Sketch of Proof

- Following the outline of Lovett and Viola, we argue based on the noise sensitivity of $AC^0$.
- Lovett and Viola showed that for any function $F$ which approximately samples a code, it holds that for random $x$ and sparse noise $\delta$, $F(x)$ and $F(x + \delta)$ are distinct codewords noticeably often.

That is, $F$ often has a large response to noise.
Sketch of Proof

• On the other hand, we know by [LMN’93] that for size $S$ circuits noised at a rate of $p$, each output bit flips with probability $p \cdot \text{polylog}(S)$.

• By a markov argument, the probability that $\Omega(1)$ fraction of outputs flip is $p \cdot \text{polylog}(S)$.

• Naively, to improve this, we would like a “tail-bound LMN” which states the output bits flip approximately independently of each other.
Sketch of Proof

• Unfortunately, tail-bound LMN is false for $AC^0$
• However, the following seems more plausible: For some models, we get “regularity lemmas”:
  – Circuits in class C have a small number of “high influence” inputs, which, if restricted, leave a “pseudorandom” circuit.
  – LTFs: O’Donnell Servedio, PTFs: Harsha et al., more

• Perhaps “tail-bound LMN” holds for the restricted circuit.
Detour: Regularity in Ckt. Complexity

- Given a circuit, get a small decision tree.
- Each circuit at a leaf is the restriction of original circuit, according to path to that leaf.
- All or “most” leaves are “balanced” / “regular” / “pseudorandom”.

\[ \approx \varepsilon \]
Actual Sketch of Proof

• We don’t prove a regularity lemma for $AC^0$.
• Instead, we give a reduction, showing that if $AC^0$ approximately samples a code, then a collection of decision trees does so as well.
• Then, we give a regularity lemma for collections of decision trees. This allows us to further reduce to “balanced” decision trees.
• Finally, we prove “tail-bound LMN” for balanced collections of decision trees. □
Balanced Decision Forests

• We introduce the following variation on the idea of \textit{influence} of a variable.

<table>
<thead>
<tr>
<th>Let $F$ be a decision forest. The \textit{significance} of a variable $x_i$ is the expected fraction of trees which read $x_i$ on a random input.</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Significance is an upper bound on influence.</td>
</tr>
<tr>
<td>– Significance is a “whitebox” definition; it may be different for two forests which compute the same function.</td>
</tr>
</tbody>
</table>
Balanced Decision Forests

• We say that a decision forest is $\beta$-balanced if no variable has significance exceeding $\beta$.

• Balanced decision forests make up the pseudorandom part of our regularity lemma, which we won’t state here explicitly.

• Our main lemma gives a large deviations bound for the hamming weight of the output of a balanced decision forest.
Balanced Decision Forests

• Second Main Result:

Let $F$ be a $\beta$-balanced forest of height $h$, let $W$ be the fractional hamming weight of $F(x)$.

$$\Pr_x \left[ W \geq O(E[W] + h\sqrt{\beta \log(h^4/\varepsilon)}) \right] \leq \varepsilon.$$

• Let $n$ be the number of outputs. In a typical application, $\beta \sim n^{-1+\delta}$, $h \sim n^\varepsilon$, and theorem still gives exponential concentration for $W$. 
Contrast with Other Results

• The most similar result is Kim-Vu polynomial concentration, since decision trees can also be written as polynomials. But, their result deteriorates rapidly in the degree, becoming useless for height $\log n$.

• Gavinsky, Lovett, Saks and Srinivasan recently gave a bound for “read k” families of functions which recovers the Chernoff bound, except they lose a factor $k$ in the exponent, which is optimal. However, this is not adaptive read $k$. 
Handy Corollary of Main Result

• Fix a forest $F$, let $\text{sig}_i(x)$ be the fraction of trees which read $x_i$ on input $x$. So $E_x[\text{sig}_i(x)]$ is what we have termed the significance of $x_i$.

• Then for any $\beta$-balanced height $h$ forest $F$, $\Pr_x \left[ \max_i \text{sig}_i(x) \geq h\sqrt{\beta \log(h^5/\varepsilon)} \right] \leq \varepsilon$.

• If we think of $\text{sig}_i(x)$ as a Lipschitz constant, Thm: “Lipschitz On Average” $\rightarrow$ “Lipschitz a.e.”
Conclusion: “Tail-Bound LMN”

• This corollary also implies that balanced decision forests satisfy “tail-bound LMN”, that is, the chance of a large response to sparse noise is exponentially small.

• Why? For almost all inputs, no bit has high significance (Corollary). Almost surely, only a few queried bits are noised (Chernoff). \( \rightarrow \) Noise response is \( o(n) \) almost surely.

• Thus, balanced forests can’t sample codes.
Open Questions

• Our bound on statistical distance seems to be the best you could do using the switching lemma, and “beating the switching lemma” is a notorious frontier in circuit complexity.
• Other applications for main lemma?
• Improve Viola’s extractor for $AC^0$ sources?
• Is the main lemma true for influence instead of significance?
Thanks!
Second Main Result

• In proving the strengthened sampling lower bound, we were led to discover a new concentration bound for sums of random variables computed by decision trees.

• This concentration bound extends a long line of work aimed at obtaining Chernoff-like concentration despite limited independence. We believe it will have other applications.
Decision Forests
Balanced Decision Trees

• Main Lemma / Second Main Result:
Let $F$ be a $\beta$-balanced forest of height $h$, let $W$ be the fractional hamming weight of $F(x)$.

$$\Pr[W \geq O(E[W] + h\sqrt{\beta \log(h^4/\varepsilon)})] \leq \varepsilon.$$  

• Let $n$ be the number of outputs. This lemma is effective, and we use it, even when $h$ is as large as $n^\varepsilon$, giving exponential concentration.
Applications of [LV’11]

• Lovett and Viola also gave several interesting applications for their work:
  – Lower Bounds for Succinct Data Structures
Let $C$ be any $(n, k, d)$-code, let $F$ be an $AC^0$ circuit of size $S$, depth $t$, with $m$ inputs. Then

$$sd(U_C, F(U_m)) \geq 1 - O \left(\frac{n}{dk} \cdot \log^{t-1} S\right)^{1/3}$$
Prior Work

• In a CCC ‘2011 paper, Lovett and Viola gave the first result for $AC^0$:

Even exponentially large $AC^0$ circuits with any amount of randomness cannot sample the uniform distribution over a good code with statistical distance better than $1 - n^{-O(1)}$. 