Lecture 1: Generalized discrepancy and composed functions

Troy Lee

Oct. 3, 2012

1 Discrepancy

My favorite way to introduce the discrepancy bound is through a game with no communication, an XOR game. In an XOR game $G = (f, \mu)$ there is a function $f : \{0,1\}^{|X|} \times \{0,1\}^{|Y|} \to \{-1,+1\}$ and a distribution μ over $|X| \times |Y|$ known to the players Alice and Bob. I will often think without warning of f and μ as $|X|$-by-$|Y|$ matrices.

In an XOR game the verifier chooses (x, y) according to μ, and sends x to Alice and y to Bob. Alice then answers with a bit $a_x \in \{-1, 1\}$ and Bob with $b_y \in \{-1, +1\}$. The verifier checks if $a_x b_y = f(x, y)$. Alice and Bob want to maximize their winning probability

$$\omega_G = \max_{\substack{a \in \{-1,+1\}^{|X|} \\ b \in \{-1,+1\}^{|Y|}}} \Pr_{(x,y) \leftarrow \mu} [f(x, y) = a_x b_y].$$

What winning probability can Alice and Bob always guarantee? They can always win with probability 1/2 by agreeing to output the most popular outcome under μ. For this reason it is more interesting, and mathematically nicer, to study the *correlation* of a strategy with f under μ.

$$\begin{aligned} \beta_G &= \max_{\substack{a \in \{-1,+1\}^{|X|} \\ b \in \{-1,+1\}^{|Y|}}} \sum_{x,y} \mu(x, y) f(x, y) a_x b_y \\ &= \max_{\substack{a \in \{-1,+1\}^{|X|} \\ b \in \{-1,+1\}^{|Y|}}} \Pr_{(x,y) \leftarrow \mu} [f(x, y) = a_x b_y] - \Pr_{(x,y) \leftarrow \mu} [f(x, y) \neq a_x b_y]. \end{aligned}$$

We will call β_G the maximal bias. You can check that $\omega_G = 1/2 + \beta_G/2$. The maximal bias can be written very neatly in matrix notation:

$$\beta_G = \max_{\substack{a \in \{-1,+1\}^{|X|} \\ b \in \{-1,+1\}^{|Y|}}} a^t (f \circ \mu) b.$$

Here the $\circ$ indicates entrywise product $(f \circ \mu)(x, y) = f(x, y)\mu(x, y)$. We will use this matrix viewpoint of the bias much more later on.

Shared randomness Alice and Bob can use shared randomness but it is easy to see by a convexity argument that it does not increase their winning probability. For upper bounds we will often assume they have shared randomness and for lower bounds we will assume they don't.

Here is the first relationship between the bias in an XOR game and communication protocols.

Theorem 1. *If f has a deterministic c-bit communication protocol, then $\beta_G \geq 2^{-c}$ in the game $G = (f, \mu)$ for any distribution μ.*

Proof. We assume Alice and Bob share a random string r of length c. Alice and Bob interpret r as a communication transcript. On input x Alice simulates her part of the protocol, assuming that the communication from Bob is given as in r. If ever r disagrees with what she would say, then she outputs a random bit. Otherwise, she outputs the last bit of the protocol (which we can assume is the answer). Bob does something similar: if ever r disagrees with what he would say, he answers a random bit. Otherwise, he outputs 1.

With probability at least 2^{-c}, the random string r actually will be a valid transcript of the protocol on x, y and thus the last bit of r actually will be $f(x, y)$. Otherwise, Alice and Bob output a random bit and are correct as often as they are wrong. Thus the bias is at least 2^{-c} for any distribution μ. □

Now we leverage this proof to randomized communication complexity. For this we will use Yao's minimax principle:

$$R_\epsilon(f) = \max_\mu D_{\mu,\epsilon}(f)$$

Here $D_{\mu,\epsilon}(f)$ is the minimum communication complexity of a deterministic protocol that correctly answers f on at least a $1 - \epsilon$ fraction of inputs (x, y) under the distribution μ.

Theorem 2 (discrepancy method)**.**

$$2^{R_\epsilon(f)} \geq \frac{1 - 2\epsilon}{\beta_{(f,\mu)}}$$

for any distribution μ.

Proof. Fix a distribution μ. Say that $R_\epsilon(f) = c$. By Yao's minimax principle, $D_{\mu,\epsilon}(f) \leq c$. Thus there is a c-bit deterministic protocol that succeeds with probability $1 - \epsilon$ when x, y are chosen according to μ. The XOR protocol given above will also succeed on these "good" (x, y) pairs thus will achieve bias $(1 - 2\epsilon)2^{-c}$. □

This is the standard discrepancy method.

Disjointness example The disjointness function is the most important function in communication complexity. As far as applications go, it is almost always some form of disjointness that comes up. However, the discrepancy method gives terrible bounds for disjointness (left as an exercise!). We will next see that the generalized discrepancy method that can show a $\Omega(\sqrt{n})$ bound for disjointness.

Generalized Discrepancy Now we look at an extension of the discrepancy method called the generalized discrepancy method. This method can show a $\Omega(\sqrt{n})$ lower bound for disjointness as we will see later in the lecture.

The idea of the generalized discrepancy method is the following. If there is a distribution μ such that f has non-negligible correlation with the function g under μ, and no strategy achieves good bias in the XOR game (g, μ), then f must have large communication complexity. The reason is that if f had small communication complexity this would give us a good strategy in the XOR game (f, μ). But because f correlates with g under μ this strategy would also do quite well in the game (g, μ), which contradicts the fact that there is no good strategy.

Now we see the formal proof.

Theorem 3.

$$2^{R_\epsilon(f)} \geq \max_{g,\mu} \frac{\langle f, g \circ \mu \rangle - 2\epsilon}{\beta_{(g,\mu)}}$$

Proof. Fix a distribution μ. Say that $R_\epsilon(f) = c$. This means that there is a c-bit deterministic protocol that is correct for f on a $1 - \epsilon$ fraction of inputs under μ. We call the inputs where this protocol is correct "good" and where it is not "bad". We use the XOR strategy from this communication protocol for f in the game (g, μ). Let's see what happens.

$$\begin{aligned}
\beta_{(g,\mu)} &\geq 2^{-c} \left(\sum_{x,y:\text{good}} \mu(x,y) f(x,y) g(x,y) - \sum_{x,y:\text{bad}} \mu(x,y) f(x,y) g(x,y) \right) \\
&= 2^{-c} \left(\sum_{x,y} \mu(x,y) f(x,y) g(x,y) - 2 \sum_{x,y:\text{bad}} \mu(x,y) f(x,y) g(x,y) \right) \\
&\geq 2^{-c} \left(\langle f, g \circ \mu \rangle - 2\epsilon \right).
\end{aligned}$$

□

Spectral Discrepancy Now we have our bound. Let's think about how to go about evaluating it. Let's look again at what $\beta_{(f,\mu)}$ is:

$$\beta_G = \max_{\substack{a \in \{-1,+1\}^{|X|} \\ b \in \{-1,+1\}^{|Y|}}} a^t (f \circ \mu) b.$$

This is a hard optimization problem. NP-hard in fact. That's not very good for us.

Let's look at an easier quantity to compute, known as the spectral discrepancy. Recall that the spectral norm of a matrix A is defined as

$$\|A\| = \max_{a,b} \frac{a^t A b}{\|a\| \|b\|}.$$

We can give an upper bound on the discrepancy by the spectral discrepancy as this optimization problem is less constrained. Notice that $a \in \{-1,+1\}^{|X|}$ has ℓ_2 norm equal to $\sqrt{|X|}$. Thus

$$\beta_{(g,\mu)} \leq \|f \circ \mu\| \sqrt{|X||Y|}.$$

We can now plug this into Theorem 3 to get the spectral discrepancy bound.

$$2^{R_\epsilon(f)} \geq \frac{1}{\sqrt{|X||Y|}} \max_{g,\mu} \left(\frac{\langle f, g \circ \mu \rangle - 2\epsilon}{\|g \circ \mu\|} \right) \tag{1}$$

This is a quantity that is easier to compute. Now the problem is how to choose g and μ? We will refer to these as witnesses to the hardness of f.

Note that g and μ always appear together as $g \circ \mu$ in Equation (1). Mathematically, it makes more sense to consider these as a single object $\psi = g \circ \mu$. What is the property of ψ? Well, g is a sign matrix, and μ is a probability distribution. So ψ can be arbitrary subject to the constraint $\ell_1(\psi) = 1$. Let's summarize the properties that we need in a good witness ψ:

1. ψ correlates with f. We need $\langle f, \psi \rangle > 2\epsilon$
2. $\ell_1(\psi) = 1$
3. ψ has small spectral norm.

2 Composed functions

From now on we will focus on *composed* communication functions. Let $f : \{-1,+1\}^n \to \{-1,+1\}$ and $g : X \times Y \to \{-1,+1\}$. We will be interested in the function

$$F(x,y) = f(g(x^1,y^1),\ldots,g(x^n,y^n))$$

where $x = (x^1,\ldots,x^n)$ and similarly for y.

Many interesting functions in communication complexity are of this form, for example Set Intersection (the negation of Disjointness) is $\mathrm{OR}_n(x_1 \wedge y^1,\ldots,x_n \wedge y_n)$.

What should we expect the communication complexity of a composed function to be?

Exercise: Show that the deterministic communication complexity of F is at most the deterministic *query* complexity of f times the communication complexity of g.

One goal is to understand for what kind of inner functions g this upper bound is tight. We will now see what is known in this direction.

3 Choosing a witness

We left off simplifying our generalized discrepancy bound but still with the problem of how to choose a witness to the hardness of f. We will now see a great idea due to Sherstov [She09] and Shi-Zhu [SZ09] for how to do this in the case of composed functions.

The bound will be in terms of the approximate polynomial degree of f, denoted $\deg_\epsilon(f)$. This is the minimal degree of a polynomial p such that $|p(x)-f(x)| \leq \epsilon$ for all $x \in \{-1,+1\}^n$. First, we need to review a few concepts from Fourier analysis over the boolean cube.

Fourier analysis: For $T \subseteq [n]$ we define a function $\chi_T : \{-1,+1\}^n \to \{-1,+1\}$ as $\chi_T(x) = \prod_{i \in T} x_i$ that takes the parity of the bits in T. Note that these function χ_T have an orthogonality property $\langle \chi_S, \chi_T \rangle = 2^n \delta_{S,T}$. We can express f in the basis of the χ_T. Let $\hat{f}_T = (1/2^n)\langle f, \chi_T \rangle$. Note that $|\hat{f}_T| \leq 1$. We can write

$$f(x) = \sum_{T \subseteq [n]} \hat{f}_T \chi_T(x).$$

Finding the best degree d polynomial which approximates f (in terms of ℓ_∞ approximation on the Boolean cube) is an optimization problem that can be written as a linear program. Looking at the dual of this linear program, it follows that if no degree d polynomial can approximate f with error ϵ over the Boolean cube then there is a witness to this fact ψ satisfying the following properties.

1. $\langle \psi, f \rangle \geq \epsilon$
2. $\ell_1(\psi) = 1$
3. $\langle \chi_T, \psi \rangle = 0$ for all characters $|T| \leq d$

These properties turn out to mesh extremely well with the properties that we need for a good witness. From this ψ we construct witness for F as

$$\Psi(x,y) = \left(\frac{2}{|X||Y|}\right)^n \psi(g(x^1,y^1), \cdots, g(x^n,y^n)). \tag{2}$$

All we need is to figure out the properties of the inner function g needed to transfer the good properties of ψ to Ψ. For Sherstov's proof the key property seems to be that g is *strongly balanced*, meaning that every row and column of g sums to zero.

Theorem 4 (Lee-Zhang [LZ10]). *Let $g : X \times Y \to \{-1,+1\}$ be a strongly balanced function. Then*

$$R_\epsilon(f \circ g^n) \geq \deg_{\epsilon_0}(f) \log_2\left(\frac{\sqrt{|X||Y|}}{\|g\|}\right) - O(1).$$

for any $\epsilon > 0$ and $\epsilon_0 > 2\epsilon$.

Using he first property of ψ, we will prove that F correlates well with Ψ, the second property gives $\ell_1(\Psi) = 1$, and in the most difficult step the third property will be used to upper bound the spectral norm of Ψ. Combining these three items, Equation (1) will give the lower bound.

Item One

$$\begin{aligned}\langle F, \Psi \rangle &= \sum_{x,y} F(x,y)\Psi(x,y) \\ &= \left(\frac{2}{|X||Y|}\right)^n \sum_{z \in \{-1,+1\}^n} f(z)\psi(z) \prod_i \left(\sum_{\substack{x^i, y^i \\ g(x^i,y^i)=z_i}} 1 \right) \\ &= \epsilon\end{aligned}$$

Item Two Works in the same way, $\ell_1(\Psi) = 1$

Item Three This is the most difficult step of the proof, and also where the proofs of Sherstov and Shi-Zhu differ. We can use the Fourier decomposition of ψ to express Ψ as $\Psi(x,y) = \sum_T \hat{\psi}_T \chi_T(g(x^1,y^1), \dots, g(x^n,y^n))$.

Let's look at the matrix Z_T where $Z_T(x,y)\chi_T(g(x^1,y^1), \dots, g(x^n,y^n))$. This can be written as a tensor product as $Z_T = \otimes_{i=1}^n g^{(i \in T)}$ where $g^0 = J$ (the all ones matrix) and $g^1 = g$. This means that

$$\Psi = \left(\frac{2}{|X||Y|}\right)^n \sum_{T, |T| > d} \hat{\psi}_T \bigotimes_{i=1}^n g^{(i \in T)}.$$

Shi-Zhu upper bound this using the triangle inequality. Sherstov uses instead

Fact 5. *If $AB^t = 0$ and $A^t B = 0$ then $\|A + B\| = \max\{\|A\|, \|B\|\}$.*

That g is strongly balanced implies $gJ = 0$ and $g^T J = 0$. This gives the key property to the proof: whenever $T \neq S$ we have $Z_T Z_S^t = 0$ and $Z_T^t Z_S = 0$. This allows us to use the fact to very simply upper bound the spectral norm of Ψ.

We have $\hat{\psi}_T = \frac{1}{2^n}\langle \psi, \chi_T \rangle$, and so $|\hat{\psi}_T| \leq \frac{1}{2^n}$ as $\ell_1(\psi) \leq 1$. Thus

$$\begin{aligned}
\|\Psi\| &= \left(\frac{2}{|X||Y|}\right)^n \max_{T,|T|\geq d} \|\hat{\psi}_T Z_T\| \\
&\leq \left(\frac{1}{|X||Y|}\right)^n \max_{T,|T|\geq d} \|Z_T\| \\
&= \max_{T,|T|\geq d} \prod_i \frac{\|g^{(i\in T)}\|}{|X||Y|} \\
&= \frac{1}{(|X||Y|)^{n/2}} \max_{T,|T|\geq d} \frac{\|g\|^{|T|}}{(|X||Y|)^{|T|/2}}
\end{aligned}$$

This is largest when $|T|$ is smallest, i.e. equal to d. Putting these three items into Equation (1) gives the bound.

4 Application to Disjointness

Set Intersection (the complement of Disjointness) can be expressed as composition of OR and AND functions.

$$\mathrm{SI}_n(x, y) = \mathrm{OR}_n(\mathrm{AND}(x_1, y_1), \cdots, \mathrm{AND}(x_n, y_n)) \tag{3}$$

The lower bound cannot directly work here as AND is not strongly balanced. So we consider different composition. The outer function is now OR on $\frac{n}{4}$ bits instead of n bits, and the inner g function is $\mathrm{OR}_4(\mathrm{AND}(x_1, y_1), \cdots, \mathrm{AND}(x_4, y_4))$ (Fig. 1).

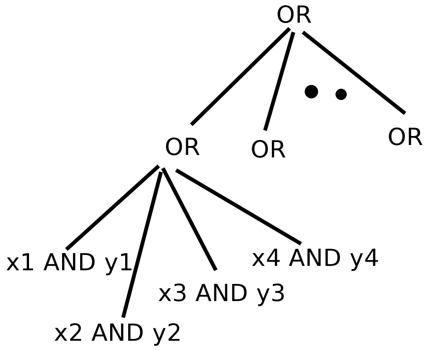

Figure 1: Equivalent way to view set intersection

The OR ∘ AND inner function is still not strongly balanced, but does have a strongly balanced submatrix. So we can use the above framework to get a lower bound on this subfunction which will then imply the bound for disjointness. Let's write the submatrix explicitly.

$$\begin{array}{ccccc} & 0001 & 0010 & 1000 & 0100 \\ 0011 & -1 & -1 & 1 & 1 \\ 0101 & -1 & 1 & 1 & -1 \\ 1100 & 1 & 1 & -1 & -1 \\ 1010 & 1 & -1 & -1 & 1 \end{array}$$

As we see that this sub matrix is strongly balanced and also it can be written as

$$\begin{pmatrix} -H & 0 \\ 0 & -H \end{pmatrix} + \begin{pmatrix} 0 & H \\ H & 0 \end{pmatrix}$$

So the norm of g is at most $2\sqrt{2}$. This proves that

$$R_{1/4}(\mathrm{DISJ}_n) = \Omega(\deg_{1/4}(OR_n)) = \Omega(\sqrt{n}) \tag{4}$$

Generalized discrepancy is also a lower bound on the quantum communication complexity, so this bound also holds in that setting, where it is tight.

References

[LZ10] T. Lee and S. Zhang. Composition theorems in communication complexity. In *Proceedings of the 37th International Colloquium On Automata, Languages and Programming*, pages 475–489. Springer-Verlag, 2010.

[She09] A. Sherstov. The pattern matrix method. *SIAM Journal on Computing*, 2009.

[SZ09] Y. Shi and Y. Zhu. Quantum communication complexity of block-composed functions. *Quantum information and computation*, 9(5,6):444–460, 2009.