Lecture 18

Negative Correlation and Applications

We have seen Chernoff-Hoeffding bounds for sums of independent random variables. There are several situations where we have dependent random variables and we need to reason about them. In some situations we can get concentration even with dependence.

<u>Example</u>: Suppose we throw n balls into n bins. Let X_i be the indicator for bin i to be empty. We see that

$$\Pr[X_i = 1] = \left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e}.$$

Let $X = \sum_{i=1}^{n} X_i$ be the number of empty bins. $E[X] \approx \frac{n}{e}$.

Note that $X_1, X_2, \ldots, X_n$ are not independent. So we cannot use the Chernoff-bound.

However it turns out that Chernoff bound holds for the upper tail.

<u>Defn:</u> A collection $\{X_1, \ldots, X_n\}$ of random variables is negatively correlated if

$$\forall\, S \subseteq [n] \quad E\left[\prod_{i \in S} X_i\right] \leq \prod_{i \in S} E[X_i]$$

Claim: In the example we saw with balls and bins, $X_1, \dots, X_n$ are negatively correlated.

Proof: $E[X_i] = \left(1 - \frac{1}{n}\right)^n$

$E\left[\prod_{i \in S} X_i\right]$ is $\Pr$[all bins in S are empty]

Let $|S| = k$ then

$$E\left[\prod_{i \in S} X_i\right] = \left(1 - \frac{k}{n}\right)^n.$$

Check that $\left(1 - \frac{1}{n}\right)^{nk} \geq \left(1 - \frac{k}{n}\right)^n$ $\square$.

Theorem: Suppose $X_1, X_2, \dots, X_n$ are binary random variables and are negatively correlated. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$.

Then

$$\Pr\left[X > (1+\delta)\mu\right] \leq \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}.$$

Note that the bound is exactly the same as for the standard upper tail in the multiplicative Chernoff bound. The reason for that is that in a certain formal sense the whole moment generating function based proof goes through. We sketch the proof.

Let $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ be <u>independent</u> bin random variables where $E[\tilde{X}_i] = E[X_i]$.

Let $\tilde{X} = \sum_{i=1}^{n} \tilde{X}_i$. $E[\tilde{X}] = E[X] = \mu$

The mgf proof for Chernoff bound proceeds as follows

$$\Pr[\tilde{X} > (1+\delta)\mu] = \Pr[e^{t\tilde{X}} > e^{t(1+\delta)\mu}]$$
$$\leq \frac{E[e^{t\tilde{X}}]}{e^{t(1+\delta)\mu}}.$$

The key step where independence is used is in expanding $E[e^{t\tilde{X}}]$ as $\prod_{i=1}^{n} E[e^{t\tilde{X}_i}]$.

After this step we only work with bounds on $E[\tilde{X}_i]$ etc.

Now consider $X = \sum_{i=1}^{n} X_i$ where X_i are negatively correlated and $E[X_i] = E[\tilde{X}_i]$ $\forall i \in [n]$.

If we can argue that

$$\mathbb{E}[e^{tX}] \leq \mathbb{E}[e^{t\tilde{X}}].$$

then we are done. For this we expand $\mathbb{E}[e^{tX}]$ as $\mathbb{E}[\prod_i e^{tX_i}]$ and use Taylor series expansion for each term with expectation outside.

$$\prod_i \left(1 + tX_i + \frac{t^2 X_i^2}{2!} + \cdots\right)$$

and $\prod_i (1 + t\tilde{X}_i + \cdots)$.

In the product we have terms which are polynomials in t and in the variables. Consider a term

$X_1 X_2^2 X_3^5 X_5^3$ and the

corresponding term $\tilde{X}_1 \tilde{X}_2^2 \tilde{X}_3^5 \tilde{X}_5^3$

Since variables are binary we can drop the exponent so we have

$$X_1 X_2 X_3 X_5. \quad \text{and} \quad \tilde{X}_1 \tilde{X}_2 \tilde{X}_3 \tilde{X}_5$$

Now by negative correlation assumption

$$E[X_1 X_2 X_3 X_5] \le E[\tilde{X}_1 \tilde{X}_2 \tilde{X}_3 \tilde{X}_5].$$

Thus term by term we have $\le$.

$$\Rightarrow E[e^{tX}] \le E[e^{t\tilde{X}}]$$

and we can proceed with the rest of the proof with $\tilde{X}$ and $\tilde{X}_1, \dots, \tilde{X}_n$.

$\square$.

Sometimes people define binary random variables $X_1, \ldots, X_n$ to be negatively correlated if $\forall S \subseteq [n]$

(i) $E\left[\prod_{i \in S} X_i\right] \le \prod_{i \in S} E[X_i]$ and

(ii) $E\left[\prod_{i \in S} (1 - X_i)\right] \le \prod_{i \in S} (1 - E[X_i])$.

If both conditions are satisfied then we also get the lower tail bound for $X = \sum X_i$

$$\Pr[X < (1-\delta)\mu] \le e^{-\frac{\mu\delta^2}{2}}.$$

Application

Consider Max-Coverage problem which is a problem related to Set Cover.

Given a $\mathcal{U}$ of n elements and m sets $S_1, S_2, \ldots, S_m \subseteq \mathcal{U}$ and an inter k, pick k of the given sets to maximize the size of their union. In other words pick k sets to cover as many elements as possible.

A simple Greedy algorithm gives a $(1-\frac{1}{e})$ approximation. However it does not give the same ratio for a slightly more general constraint

So we will instead consider an LP relaxation based approach.

Variable x_i for set S_i chosen or not

Variable z_j for whether ~~its~~ element j is covered.

$$\max \sum_{j=1}^{n} z_j$$

$$\sum_{i=1}^{m} x_i = k$$

$$\sum_{j \in S_i} x_i \geq z_j \quad \forall j \in [n]$$

$$z_j \leq 1 \quad \forall j \in [n]$$

$$x_i \geq 0 \quad \forall i \in [m].$$

Suppose we solve above LP relaxation. Let OPT_{LP} be the value of the opt fractional solution (x^*, z^*).
How can we round it?

Randomized Rounding

A simple strategy is to pick each set S_i independently with prob $\bar{x}_i^*$.
First let us evaluate the expected # of elements covered.

Let $Y_j = 1$ if element j is covered.

$$\Pr[Y_j = 1] = 1 - \prod_{i: j \in S_i} (1 - \bar{x}_i^*)$$

$$\geq 1 - \prod_{i : j \in S_i} e^{-x_i^*}$$

$$\geq 1 - e^{-z_j^*}$$

$$\geq \left(1 - \frac{1}{e}\right) z_j^*$$

Thus, by linearity of expectation, the expected # of elements covered is $\sum_j \left(1 - \frac{1}{e}\right) z_j^* \geq \left(1 - \frac{1}{e}\right) OPT_{LP}$.

The problem with the preceding rounding is that we may not satisfy the constraint that we pick at most k sets.

How can we ensure that we satisfy the constraint and still get a good approximation for covering elements?

We will discuss a rounding strategy called "pipage" rounding. This is a <u>dependent</u> rounding strategy.

1. Solve LP to obtain fractional solution $(\bar{x}, \bar{z})$.
2. While ($\bar{x}$ has fractional variables)
   - Let x_i, x_j be s.t. $x_i, x_j \in (0, 1)$
   - Let $\varepsilon = \min(x_i, x_j, 1-x_i, 1-x_j)$

- Toss a coin. If heads

$$x_i \leftarrow x_i + \varepsilon$$

$$x_j \leftarrow x_j - \varepsilon$$

Else

$$x_i \leftarrow x_i - \varepsilon$$

$$x_j \leftarrow x_j + \varepsilon$$

3. Output sets with $x_i = 1$.

<u>Claim</u>: After the while loop terminates $\bar{x}$ is integral and $\sum_{i=1}^{m} x_i = k$.

<u>Lemma</u>: Let X_i be the value of S_i at end of alg. Then $E[X_i] = \bar{x}_i$.

Proof: In each step it is easy to see that $E[X_i]$ does not change. By induction on steps.

D.

Lemma: The algorithm terminates in T steps where $E[T] = \text{poly}(m)$.

Proof: In each iteration with prob $\frac{1}{2}$ at least one variable becomes 0 or 1. If a variable is 0 or 1 it is not touched again. Initially at most m fractional variables. Implies, in expectation $T \leq 2m$. Can also prove high probability bound using Chernoff bounds.

$\square$.

The main technical lemma that we will not prove is the following.

<u>Lemma</u>: $X_1, X_2, \ldots, X_m$ are negatively correlated.

The proof relies on the fact that expectations are preserved and only two variables are modified at ~~at~~ each step. It is not difficult but we omit details.

Thus the rounding ensures that

$$\sum_{i=1}^{m} X_i = k \quad \text{deterministically.}$$

and $X_1, X_2, \ldots, X_m$ are negatively correlated.

Now consider an element j.

What is $\Pr[j \text{ is covered}]$

$$= 1 - \prod_{i: j \in S_i} (1 - X_i)$$

By negative correlation

$$\prod_{i: j \in S_i} (1 - X_i) \leq \prod_{i: j \in S_i} (1 - E[X_i])$$

$$= \prod_{i: j \in S_i} (1 - x_i)$$

$\Rightarrow \Pr[j \text{ is covered}]$

$$\geq 1 - \prod_{i : j \in S_i} (1 - x_i)$$

and hence we can use the same analysis as before.

$\Rightarrow$ expected # of elements covered is $\geq \left(1 - \frac{1}{e}\right) OPT_{LP}$.

Thus we maintain the constraint and obtain a $\left(1 - \frac{1}{e}\right)$ approx.

The above approach generalizes quite a bit to submodular function maximization subject to an arbitrary matroid constraint. We will not go into details but consider the following extension of Max K-Coverage.

As before we have $\mathcal{U}$ and sets $S_1, S_2, \ldots, S_m$. Now the sets are colored. In other words we partition the sets into l groups $G_1, G_2, \ldots, G_l$.

Each group h has a bound k_h and this implies that at most k_h sets can be chosen from G_h.

We can write a natural LP with this more complicated constraint.

$$\max \sum_{j=1}^{n} z_j$$

$$\sum_{i \in G_h} x_i \leq k_h \quad h \in [\ell]$$

$$\sum_{j \in S_i} x_i \geq z_j \quad \forall j \in [n]$$

$$z_j \leq 1 \quad \forall j \in [n]$$

$$x_i \geq 0 \quad \forall i \in [m].$$

Now as before we can see that if we randomly round by picking each set S_i independently with probability x_i, we get expected coverage $\geq (1-\frac{1}{e})\, OPT_{LP}$.

It is not hard to generalize pipage rounding to this slightly more complex constraint.

This yields a $(1-\frac{1}{e})$ approx.

The natural greedy algorithm yields only a $\frac{1}{2}$ approximation for this generalization.