Lecture 18: Negative Correlation and Applications

1 Introduction

We have seen Chernoff-Hoeffding bounds for sums of independent random variables. However, 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.

1.1 Example: Balls and Bins

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

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

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

Note that $X_1, X_2, \ldots, X_n$ are **not independent**, so we cannot use the Chernoff bound directly. However, it turns out that the Chernoff bound holds for the upper tail.

2 Negative Correlation

Definition 1. A collection $X_1, \ldots, X_n$ of random variables is **negatively correlated** if:

$$E\left[\prod_{i \in S} X_i\right] \leq \prod_{i \in S} E[X_i]$$

for all subsets $S \subseteq \{1, 2, \ldots, n\}$.

Claim 1. *In the example we saw with balls and bins, $X_1, \ldots, X_n$ are negatively correlated.*

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

$E\left[\prod_{i \in S} X_i\right]$ is the probability that all bins in S are empty.

Let $|S| = k$. Then:

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

One can check that $\left(1 - \frac{k}{n}\right)^n \leq \left(1 - \frac{1}{n}\right)^{kn} = \left[\left(1 - \frac{1}{n}\right)^n\right]^k$. □

3 Chernoff Bound for Negatively Correlated Variables

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

$$P[X \geq (1+\delta)\mu] \leq e^{-\frac{\delta^2 \mu}{3}}$$

Note: 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.

3.1 Proof Sketch

Let $\bar{X}_1, \ldots, \bar{X}_n$ be independent binary random variables where $E[\bar{X}_i] = E[X_i]$.

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

The MGF proof for Chernoff bound proceeds as follows:

$$P[\bar{X} \geq (1+\delta)\mu] = P[e^{t\bar{X}} \geq e^{t(1+\delta)\mu}] \leq \frac{E[e^{t\bar{X}}]}{e^{t(1+\delta)\mu}}$$

The key step where independence is used is in expanding:

$$e^{t\bar{X}} = \prod_{i=1}^{n} e^{t\bar{X}_i}$$

After this step, we only work with bounds on $E[e^{t\bar{X}_i}]$, etc.

Now consider $X = \sum X_i$ where X_i are negatively correlated and $E[X_i] = E[\bar{X}_i]$ for all i.

If we can show that:

$$E[e^{tX}] \leq E[e^{t\bar{X}}]$$

then we are done. For this, we expand e^{tX} as $\sum_{j=0}^{\infty} \frac{t^j X^j}{j!}$ and use Taylor series expansion to each term with expectation outside:

$$e^{tX} = \prod_{i=1}^{n} (1 + (e^t - 1)X_i)$$

and similarly:

$$e^{t\bar{X}} = \prod_{i=1}^{n} (1 + (e^t - 1)\bar{X}_i)$$

In the product, we have terms which are polynomials in t and in the variables. Consider a term $X_{i_1}^{a_1} X_{i_2}^{a_2} \cdots X_{i_k}^{a_k}$ and the corresponding term $\bar{X}_{i_1}^{a_1} \bar{X}_{i_2}^{a_2} \cdots \bar{X}_{i_k}^{a_k}$.

Since variables are binary, we can drop the exponents, so we have $X_{i_1} X_{i_2} \cdots X_{i_k}$ and $\bar{X}_{i_1} \bar{X}_{i_2} \cdots \bar{X}_{i_k}$.

Now by negative correlation assumption:

$$E[X_{i_1} X_{i_2} \cdots X_{i_k}] \leq E[\bar{X}_{i_1}] E[\bar{X}_{i_2}] \cdots E[\bar{X}_{i_k}] = E[\bar{X}_{i_1} \bar{X}_{i_2} \cdots \bar{X}_{i_k}]$$

Thus, term by term, we have $E[e^{tX}] \leq E[e^{t\bar{X}}]$, and we can proceed with the rest of the proof with $\bar{X}$ and $\bar{X}_i$.

3.2 Lower Tail Bound

Sometimes people define binary random variables $X_1, \ldots, X_n$ to be negatively correlated if for all $S \subseteq \{1, \ldots, n\}$:

$$E\left[\prod_{i \in S} X_i\right] \leq \prod_{i \in S} E[X_i]$$

and:

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

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

$$P[X \leq (1 - \delta)\mu] \leq e^{-\frac{\delta^2 \mu}{2}}$$

4 Application: Max Coverage Problem

Consider the **Max Coverage** problem, which is a problem related to Set Cover.

Problem: Given a universe U of n elements and m sets $S_1, \ldots, S_m \subseteq U$, and an integer 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 $(1 - 1/e)$ 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.

4.1 LP Relaxation

Variables:

- x_i for set S_i (chosen or not)
- z_j for whether element j is covered

$$\begin{aligned}
\max \quad & \sum_{j=1}^{n} z_j \\
\text{s.t.} \quad & \sum_{i=1}^{m} x_i \leq k \\
& \sum_{i : j \in S_i} x_i \geq z_j \quad \forall j \in [n] \\
& z_j \leq 1 \quad \forall j \in [n] \\
& x_i \geq 0 \quad \forall i \in [m]
\end{aligned}$$

Suppose we solve the above LP relaxation. Let OPT_{LP} be the value of the optimal fractional solution (x^*, z^*).

4.2 Randomized Rounding (Naive Approach)

A simple strategy is to pick each set S_i independently with probability x_i^*.

Let us evaluate the expected number of elements covered. Let $Y_j = 1$ if element j is covered.

$$P[Y_j = 1] = 1 - \prod_{i:j \in S_i} (1 - x_i^*)$$

Since $1 - x \geq e^{-x/(1-x)} \geq e^{-2x}$ for $x \leq 1/2$:

$$P[Y_j = 1] \geq 1 - \prod_{i:j \in S_i} e^{-x_i^*} = 1 - e^{-\sum_{i:j \in S_i} x_i^*} \geq 1 - e^{-z_j^*} \geq \frac{z_j^*}{2}$$

(Using $1 - e^{-z} \geq z/2$ for $z \leq 1$.)

Thus, by linearity of expectation, the expected number of elements covered is:

$$\geq \sum_{j=1}^{n} \frac{z_j^*}{2} = \frac{1}{2} \cdot \mathrm{OPT}_{\mathrm{LP}}$$

Problem: We may not satisfy the constraint that we pick at most k sets.

5 Pipage Rounding

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 dependent rounding strategy.

5.1 Algorithm

1. Solve LP to obtain fractional solution $\bar{x} \in [0,1]^m$
2. While $\bar{x}$ has fractional variables:
   (a) Let x_i, x_j be such that $0 < x_i, x_j < 1$
   (b) Let $\epsilon = \min\{x_i, 1 - x_j, 1 - x_i, x_j\}$
   (c) Toss a coin. If heads:
   - $x_i \leftarrow x_i + \epsilon$
   - $x_j \leftarrow x_j - \epsilon$

   Else:
   - $x_i \leftarrow x_i - \epsilon$
   - $x_j \leftarrow x_j + \epsilon$
3. Output sets with $x_i = 1$

Claim 2. *After the while loop terminates, $\bar{x}$ is integral and $\sum_{i=1}^{m} x_i = k$.*

Lemma 1. *Let X_i be the value of x_i at end of algorithm. Then $E[X_i] = x_i^*$.*

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

Lemma 2. *The algorithm terminates in T steps where $E[T] \leq$ poly(m).*

Proof. In each iteration with probability $1/2$, 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. □

5.2 Main Technical Lemma

Lemma 3. *$X_1, \ldots, X_m$ are negatively correlated.*

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

5.3 Analysis

Thus, the rounding ensures that $\sum_{i=1}^{m} X_i = k$ deterministically and $X_1, \ldots, X_m$ are negatively correlated.

Now consider an element j. What is $P[j \text{ is covered}]$?

$$P[j \text{ is covered}] = 1 - \prod_{i:j \in S_i} (1 - X_i)$$

By negative correlation:

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

Therefore:

$$P[j \text{ is covered}] \geq 1 - \prod_{i:j \in S_i} (1 - x_i^*)$$

and hence we can use the same analysis as before: expected number of elements covered is $(1 - 1/e) \cdot \mathrm{OPT}_{\mathrm{LP}}$.

Thus, we maintain the constraint and obtain a $(1 - 1/e)$ approximation.

6 Generalization: Matroid Constraints

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 U and sets $S_1, \ldots, S_m$. Now the sets are colored. In other words, we partition the sets into ℓ groups $A_1, \ldots, A_\ell$.

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

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

$$\begin{aligned}
\max \quad & \sum_{j=1}^{n} z_j \\
\text{s.t.} \quad & \sum_{i \in A_h} x_i \leq k_h \quad \forall h \in [\ell] \\
& \sum_{i:j \in S_i} x_i \geq z_j \quad \forall j \in [n] \\
& z_j \leq 1 \quad \forall j \in [n] \\
& x_i \geq 0 \quad \forall i \in [m]
\end{aligned}$$

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 $(1 - 1/e) \cdot \mathrm{OPT}_{\mathrm{LP}}$.

It is not hard to generalize pipage rounding to this slightly more complex constraint. This yields a $(1 - 1/e)$ approximation.

Note: The natural greedy algorithm yields only a $1/2$ approximation for this generalization.