On The Negative Inter-Dependencies of the Multivariate Hypergeometric Distribution

Szymon Snoeck

March 2025

1 Introduction

The multivariate hypergeometric distribution naturally arises in algorithm analysis and discrete math. Often, one is trying to prove large-deviation bounds or bounds on the expectation of a multivariate hypergeometric random variable, $(m_1, \ldots, m_k)$. This is naturally complicated by the inter-dependencies between $m_1, \ldots, m_k$. Towards a resolution, this note shows that the expectation of a multivariate hypergeometric distribution can be upper bounded by suitably chosen independent binomials. In particular, this result can be extended to a bound on the moment generating function and in turn provide large deviation bounds.

2 Setup and Main Result

Suppose one has an urn containing balls, each labeled with a number from 1 to $k \in \mathbb{N}$, and $m \in \mathbb{N}$ balls are selected uniformly, without replacement. Let V of size $n \in \mathbb{N}$ denote the set of all balls in urn. For $i \in [k]$, define S_i as the set of balls with label $i \in [k]$. If S of size m is the set of balls picked then $|S \cap S_1|, \ldots, |S \cap S_k|$—the number of balls picked of each label—form a multivariate hypergeometric distribution. Now it is shown that $|S \cap S_1|, \ldots, |S \cap S_k|$ have a strong negative inter-dependence in that their expectation with respect to a wide range of functions is upper bounded by the same expectation over a multinomial or independent binomial distribution.

Theorem 1. *Fix a partition $S_1, \ldots, S_k$ of V and any $m \in \mathbb{N}$. Let $b_1, \ldots, b_k$ be independent binomial random variables such that $b_i \sim bin(m, |S_i|/n)$. Then for any $g : \mathbb{R} \to \mathbb{R}_{>0}$ that is log convex and increasing, the following holds:*

$$\mathop{\mathbb{E}}_{S \sim \text{Unif}\{S' \subseteq V : |S'| = m\}} [g(|S \cap S_1|) \cdots g(|S \cap S_k|)] \leq \mathbb{E}[g(b_1)] \cdots \mathbb{E}[g(b_k)].$$

Corollary 1. *Fix a partition $S_1, \ldots, S_k$ of V and any $m \in \mathbb{N}$. Let $b_1, \ldots, b_k$ be a multinomial random variable with m trials such that $b_i \sim bin(m, |S_i|/n)$. Then* *for any $f : \mathbb{R}^k \to \mathbb{R}$ that is convex, the following holds:*

$$\mathop{\mathbb{E}}_{S \sim \mathrm{Unif}\{S' \subseteq V : |S'| = m\}}[f(|S \cap S_1|, \ldots, |S \cap S_k|)] \leq \mathbb{E}[f(b_1, \ldots, b_k)].$$

Proof of Theorem 1. This argument was inspired by techniques introduced by Luh and Pippenger [2014]. Consider creating two sequences $C, \tilde{C}$ through the following random process: starting with $C = \tilde{C} = \varnothing$, sample v uniformly from V and append v to $\tilde{C}$. If v is not already in C then add it to the end of C. Repeat until all $v \in V$ have been sampled (this process terminates with probability 1). After the process is done the sequences are: $C = (c_1, c_2, \ldots, c_n)$ and $\tilde{C} = (\tilde{c}_1, \tilde{c}_2, \ldots)$.

At this point it will be useful to define $\sigma : [|\tilde{C}|] \to [n]$ such that $\sigma(j)$ is the smallest index $i \in [n]$ such that $c_i = \tilde{c}_j$ which exists by construction. Moreover, it is easy to see that $\sigma(j) \leq j$ for all $j \in [|\tilde{C}|]$.

For all $i \in [k], j \in [n]$, define the following random variables: $x_{ij} \equiv \mathbb{1}[c_j \in S_i]$ and $\tilde{x}_{ij} \equiv \mathbb{1}[\tilde{c}_j \in S_i]$. Thus, $S \equiv \{c_1, \ldots, c_m\}$ is a uniformly sampled m-sized subset of V, and for all $i \in [m]$, $|S \cap S_i| = \sum_{j \in [m]} x_{ij}$. Furthermore, for $i \in [k]$, $b_i \equiv \sum_{j \in [m]} \tilde{x}_{ij}$ is distributed like a binomial with m trials and probability of success $|S_i|/m$.

To finish the proof, it is shown that for all $i \in [m]$, $\mathbb{E}[b_i |\ |S \cap S_1|, \ldots, |S \cap S_k|] = |S \cap S_i|$. Since C is a uniformly, randomly chosen ordering of the $v \in V$, for any $j, j' \in [m]$, we have that:

$$E[x_{ij} |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k] = E[x_{ij'} |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k].$$

Thus, for any $j \in [m]$, $\mathbb{E}[x_{ij} |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k] = s_i/m$ since $s_i = \sum_{j \in [m]} \mathbb{E}[x_{ij} |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k]$. This gives us that:

$$\begin{aligned}
&\mathbb{E}[b_i |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k] \\
&\qquad = \sum_{j \in [m]} \mathbb{E}[\tilde{x}_{ij} |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k] \\
&\qquad = \sum_{j \in [m]} \mathbb{E}[x_{i\sigma(j)} |\ |S \cap S_1| = s_1, \ldots, |S \cap S_k| = s_k] \\
&\qquad = \sum_{j \in [m]} s_i/m = s_i.
\end{aligned}$$

Hence $\mathbb{E}[b_i |\ |S \cap S_1|, \ldots, |S \cap S_k|] = |S \cap S_i|$. Since g is log convex, the function $f : \mathbb{R}^k \to \mathbb{R}_{>0}$ defined as $f(a_1, \ldots, a_k) = g(a_1) \cdots g(a_k) = \exp\left(\sum_{i \in [k]} \log(g(a_i))\right)$ is convex as well. By multivariate Jensen's inequality, we get:

$$\begin{aligned}
&\mathbb{E}[f(|S \cap S_1|, \ldots, |S \cap S_k|)] \\
&\qquad = \mathbb{E}[f(\mathbb{E}[b_i |\ |S \cap S_1|, \ldots, |S \cap S_k|], \ldots, \mathbb{E}[b_i |\ |S \cap S_1|, \ldots, |S \cap S_k|])] \\
&\qquad \leq \mathbb{E}[\mathbb{E}[f(b_1, \ldots, b_k) |\ |S \cap S_1|, \ldots, |S \cap S_k|]] \\
&\qquad = \mathbb{E}[f(b_1, \ldots, b_k)] = \mathbb{E}[g(b_1) \cdots g(b_k)].
\end{aligned}$$

Note that $b_1, \dots, b_k$ form a multinomial distribution so the above proves Corollary 1. To finish the proof we apply a result by Dubhashi and Ranjan [1996] which proved that for g increasing the following holds:

$$\mathbb{E}[g(b_1) \cdots g(b_k)] \leq \mathbb{E}[g(b_1)] \cdots \mathbb{E}[g(b_k)].$$

$\square$

References

D. P. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), Jan. 1996. doi: 10.7146/brics.v3i25.20006. URL `https://tidsskrift.dk/brics/article/view/20006`.

K. Luh and N. Pippenger. Large-deviation bounds for sampling without replacement. *The American Mathematical Monthly*, 121(5):449–454, 2014. doi: 10.4169/amer.math.monthly.121.05.449. URL `https://www.tandfonline.com/doi/abs/10.4169/amer.math.monthly.121.05.449`.