K-Randomization

Maxim Zhilyaev   David Zeber

March 14, 2016

1   Outline of the procedure

-

2   Theoretical setup

In the following we work with data in the form of bit vectors. A **bit vector** is a vector $v \in \{0,1\}^L$.

First we define the randomization procedure we will be applying.

Definition. The randomization procedure R with **lie probability** $0 < q < 1/2$ flips a bit with probability q, and leaves it as-is with probability $1 - q$. In other words, for a bit $b \in \{0,1\}$,

$$R(b) = R(b; X) = (1 - b) \cdot X + b \cdot (1 - X) \quad \text{where } X \sim Ber(q).$$

When applied to a vector, each bit is randomized independently:

$$R(v) = R\big(v; (X_1, \dots, X_L)\big) = \big(R(v_1; X_1), \dots, R(v_L; X_L)\big) \quad \text{where } X \overset{\text{iid}}{\sim} Ber(q).$$

Remark. The randomization R reports the original bit value with probability $1 - q > q$, and lies with probability q. This is equivalent to the randomized response procedure where the value is reported as-is with probability $1 - f$, and with probability f the reported value is the outcome of the toss of a fair coin. In this case, $q = f/2$.

Remark. If $q = 1/2$, then $R(0) \overset{d}{=} R(1)$, and the reported value is "completely" randomly generated, i.e., independently of the original value.

Distribution of $R(v)$.

For a bit b, the randomization lies iff $R(b) \neq b$:

$$P[R(b) = s] = q^{\mathbf{1}_{\{b \neq s\}}} (1 - q)^{\mathbf{1}_{\{b = s\}}}$$

Hence, for a bit vector v,

$$P[R(v) = s] = q^{\sum \mathbf{1}_{\{b_i \neq s_i\}}} (1 - q)^{\sum \mathbf{1}_{\{b_i = s_i\}}} = q^{L - m(v,s)} (1 - q)^{m(v,s)},$$

where $m(v, s) = |\{i : v_i = s_i\}|$. Note that this probability is maximized when $m(v, s) = L$ (the reported vector s is identical to the original vector v), and minimized when $m(v, s) = 0$. In other words, the most likely outcome of randomizing a bit vector is obtaining an identical vector.

For a collection T,

$$P[s \in R(T)] = 1 - P[s \notin R(T)] = 1 - \prod_{v \in T} P[R(v) \neq s] = 1 - \prod_{v \in T} \left[1 - q^{L - m(v,s)} (1 - q)^{m(v,s)} \right].$$

3   Differential Privacy

The typical setting for differential privacy is the following. We consider a **database** as a collection of records. The records are elements of some space D, and a database $\boldsymbol{x}$ is a vector of n records: $\boldsymbol{x} \in D^n$.

We wish to release information based on the database by applying a **query** to it. This is a function A mapping the database into another space: $A : D^n \to \boldsymbol{S}$. If the function A is random, i.e., $A(\boldsymbol{x}) = A(\boldsymbol{x}, X)$ for a random element X, then the output $A(\boldsymbol{x})$ is a random element of $\boldsymbol{S}$.

In considering the differential privacy of A, we compare the result of applying A to two very similar databases $\boldsymbol{x}, \boldsymbol{x}' \in D^n$. We say the databases **differ in one row** if $\sum_{i=1}^{n} \mathbf{1}_{\{x_i \neq x_i'\}} = 1$. The random query A is said to be ϵ-**differentially private** if, for any two databases $\boldsymbol{x}, \boldsymbol{x}' \in D^n$ differing in one row,

$$P[A(\boldsymbol{x}) \in S] \leq \epsilon \cdot P[A(\boldsymbol{x}') \in S]$$

for all $S \subset \boldsymbol{S}$ (measurable). An alternative notion of differing in one row that is sometimes used is that $\boldsymbol{x} \in D^n$, $\boldsymbol{x}' \in D^{n+1}$, and $x_i = x_i'$ for $i = 1, \ldots, n$. In other words, $\boldsymbol{x}'$ includes an additional record that is not in $\boldsymbol{x}$.

If $\boldsymbol{S}$ is countable, then we can write

$$P[A(\boldsymbol{x}) \in S] = \sum_{s \in S} P[A(\boldsymbol{x}) = s].$$

Hence,

$$\frac{P[A(\boldsymbol{x}) \in S]}{P[A(\boldsymbol{x}') \in S]} = \frac{\sum_{s \in S} P[A(\boldsymbol{x}) = s]}{\sum_{s \in S} P[A(\boldsymbol{x}') = s]} \leq \max_{s \in S} \frac{P[A(\boldsymbol{x}) = s]}{P[A(\boldsymbol{x}') = s]}$$

by the Lemma (need reference).

Furthermore, if A randomizes each record in the database independently, i.e., $A(\boldsymbol{x}) = A(\boldsymbol{x}, \boldsymbol{X}) := \big(A_0(x_1, X_1), \ldots, A_0(x_n, X_n)\big)$ where X_i are independent, then $\boldsymbol{S} = \boldsymbol{S}_0^n$ and $s = (s_1, \ldots, s_n)$ with $s_i \in \boldsymbol{S}_0$. In this case $P[A(\boldsymbol{x}) = s] = P[A_0(x_1) = s_1, \ldots, A_0(x_n) = s_n] = \prod P[A_0(x_i) = s_i]$. If $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in one row (wlog $x_1 \neq x_1'$ and $x_i = x_i'$ for $i = 2, \ldots, n$), then

$$\frac{P[A(\boldsymbol{x}) = s]}{P[A(\boldsymbol{x}') = s]} = \frac{P[A_0(x_1) = s_1]}{P[A_0(x_1') = s_1]}.$$

Therefore, in this case, the query A will satisfy differential privacy if

$$P[A_0(x) = s] \leq \epsilon \cdot P[A_0(x') = s]$$

for all $x, x' \in D$ and $s \in \boldsymbol{S}_0$. This is the formulation used in the RAPPOR paper that applies to differences between individual records rather than collections differing on a single element.

—————————————————————————————-

Consider a collection T of bit vectors, and write $T_v = T \backslash \{v\}$. The randomization procedure R is ϵ-differentially private if

$$\log \left(\frac{P[R(T) \in S]}{P[R(T_v) \in S]} \right) \leq \epsilon$$

for any set of bit vectors S.

Anonymity:

$$A_p = \min_{v \in T, s \in \{0,1\}^L} \frac{P[s \in R(T_v)]}{P[s = R(v)]}$$

4   Single bit case

4.1   Estimating number of single bits

Suppose there are T set bits in the original collection of N single bit records. After randomization is performed the number of observed synthetic bits S is a random variable which we express as:

$$S = p \cdot T + q \cdot (N - T)$$

From here we can express an estimate for T, computed from observed value of S:

$$\bar{T} = \frac{S - qN}{p - q} \tag{4.1}$$

The expectation, variance and deviation of $\bar{T}$ random variable are given by:

$$E(\bar{T}) = T \tag{4.2}$$

$$VAR(\bar{T}) = \frac{qpN}{(p-q)^2} \tag{4.3}$$

$$\sigma(\bar{T}) = \sqrt{\frac{qpN}{(p-q)^2}} \tag{4.4}$$

4.2   Local Differential Privacy

We now study how differential privacy ratio changes depending on the configuration of underlying database D. Assuming that D consists of N single bit records, we are interested in deriving

the expression of differential probability ratio as a function of observed number of set bits after randomization is performed.

4.2.1 Choice of D

We are seeking collection D that maximizes differential privacy ratio for any number of observed bits in the randomized collection S. Since we initially consider D to consists of single bits only, the modified record switches the original bit to an opposite value. Without loss of generality, assume that the original record was 1 and it was modified to 0. Hence the original collection D contains at least one set bit, and the modified collection D_m contains one less set bits. Both collections generate synthetic collection S. Call the number of set bits in the synthetic collection a random variable s. Then, the differential privacy ratio when s is equal a particular number i of set bits is given by:

$$R_i = \frac{P(s = i | D_m)}{P(s = i | D)}$$

Theorem 4.1. R_i *is maximized when D contains N set bits*

Proof. Suppose there are m set bits in the original collection D_m. Consider generating function for the number of set bits s in S.

$$G_m(x) = (q + px)^m (p + qx)^{N-m} = \sum_{i=0}^{N} a_i^m x^i$$

Note that coefficients a_i^m in the expansion of the generating function G_m represent probabilities of $P(s = i | D)$. We prove that for any i, the differential privacy ratio grows with m:

$$\frac{a_i^m}{a_i^{m+1}} > \frac{a_i^{m-1}}{a_i^m}$$

which holds when

$$(a_i^m)^2 > a_i^{m-1} a_i^{m+1} \tag{4.5}$$
$$(a_i^m)^2 - a_i^{m-1} a_i^{m+1} > 0 \tag{4.6}$$

Consider generating functions for $m + 1$, m and $m - 1$ respectively:

$$G_{m+1}(x) = (p + qx)^{m+1}(p + qx)^{N-m-1} \tag{4.7}$$

$$G_m(x) = (p + qx)^m(p + qx)^{N-m} \tag{4.8}$$

$$G_{m-1}(x) = (p + qx)^{m-1}(p + qx)^{N-m+1} \tag{4.9}$$

Define $Q(x)$ as:

$$Q(x) = (p + qx)^{m-1}(p + qx)^{N-m-1} = \sum_{i=0}^{N-2} b_i x^i$$

Then generating functions above are expressed as:

$$G_{m+1}(x) = Q(x)(q + px)^2 = \sum_{i=0}^{N}[b_i q^2 + 2qpb_{i-1} + b_{i-2}p^2]x^i \tag{4.10}$$

$$G_m(x) = Q(x)(q + px)(p + qx) = \sum_{i=0}^{N}[b_i qp + (q^2 + p^2)b_{i-1} + b_{i-2}qp]x^i \tag{4.11}$$

$$G_{m-1}(x) = Q(x)(p + qx)^2 = \sum_{i=0}^{N}[b_i p^2 + 2qpb_{i-1} + b_{i-2}q^2]x^i \tag{4.12}$$

From here we can express coefficients of each generating function through coefficients of $Q(x)$

$$a_i^{m+1} = b_i q^2 + 2qpb_{i-1} + b_{i-2}p^2 \tag{4.13}$$

$$a_i^m = b_i qp + (q^2 + p^2)b_{i-1} + b_{i-2}qp \tag{4.14}$$

$$a_i^{m-1} = b_i p^2 + 2qpb_{i-1} + b_{i-2}q^2 \tag{4.15}$$

Now, replace the coefficients a_i in the 4.6 with their expressions through b_i.

$$(a_i^m)^2 - a_i^{m-1}a_i^{m+1} = (b_i qp + (q^2 + p^2)b_{i-1} + b_{i-2}qp)^2 - (b_i q^2 + 2qpb_{i-1} + b_{i-2}p^2) \cdot (b_i p^2 + 2qpb_{i-1} + b_{i-2}q^2)$$

After trivial algebraic transformations the above expression simplifies to:

$$(a_i^m)^2 - a_i^{m-1}a_i^{m+1} = (p^2 - q^2)^2 \cdot (b_i^2 - b_{i+1}b_{i-1}) \geq 0$$

Note that the first term of the product is always greater than 0, and we will show that the second term is greater or equal to zero as well.

Lemma 1 If a polynomial has the from bellow

$$Q(x) = \sum_{i=0}^{n} a_i x^i = a_n \prod_{i} (r_i + x), \; where \; r_i \geq 0$$

Then

$$(a_i^2 - a_{i+1} a_{i-1}) \geq 0$$

Proof. Assume polynomial is monic (e.g. $a_n = 1$), and prove lemma by induction.

For n=2:

$$(r_1 + x)(r_2 + x) = r_1 * r_2 + (r_1 + r_2)x + x^2 \qquad (4.16)$$
$$a_1^2 - a_0 a_2 = (r_1 + r_2)^2 - r_1 * r_2 = r_1^2 + r_2^2 + r_1 r_2 > 0 \;, since \; r_1 > 0 \; and \; r_2 > 0 \qquad (4.17)$$

Assume that for n, the statement holds for all i, then for $n+1$ we can express the polynomial as:

$$Q^{n+1}(x) = \sum_{i=0}^{n+1} a_i x^i = \prod_{i} (r_i + x) = Q^n(x) \cdot (r_{n+1} + x) = \left(\sum_{i=0}^{n} b_i x^i \right) \cdot (r_{n+1} + x) \qquad (4.18)$$

$$\sum_{i=0}^{n+1} a_i x^i = \sum_{i=0}^{n+1} [b_i r_{n+1} + b_{i-1}] x^i \qquad (4.19)$$

$$a_i = b_i r_{n+1} + b_{i-1} \qquad (4.20)$$

The index of r_{n+1} is irrelevant for the proof, hence we drop it. We now express $(a_i^2 - a_{i+1} a_{i-1})$ through coefficients of $Q^n(x)$ and preform algebraic simplifications:

$$a_i^2 - a_{i+1} a_{i-1} = [b_i r + b_{i-1}]^2 - [b_{i+1} r + b_i] \cdot [b_{i-1} r + b_{i-2}] \qquad (4.21)$$
$$a_i^2 - a_{i+1} a_{i-1} = r^2 (b_i^2 - b_{i+1} b_{i-1}) + r(b_i b_{i-1} - b_{i+1} b_{i-1}) + (b_{i-1}^2 - b_i b_{i-2}) \qquad (4.22)$$
$$b_i^2 - b_{i+1} b_{i-1} \geq 0 \; by \; induction \; hypothesis \qquad (4.23)$$
$$b_{i-1}^2 - b_i b_{i-2} \geq 0 \; by \; induction \; hypothesis \qquad (4.24)$$
$$\qquad (4.25)$$
$$b_i b_{i-1} - b_{i+1} b_{i-1} \geq 0 \; because \; all \; b_i \; are \; positive \; and \qquad (4.26)$$
$$b_i^2 \geq b_{i+1} b_{i-1} \; and \; b_{i-1}^2 > b_i b_{i-2} \qquad (4.27)$$
$$b_i^2 \cdot b_{i-1}^2 \geq b_{i+1} b_{i-1} b_i b_{i-2} \qquad (4.28)$$
$$b_i b_{i-1} \geq b_{i+1} b_{i-1} \qquad (4.29)$$

This completes the proof of **Lemma 1** for monic polynomials. Same result is true for non-monic polynomials because if $(a_i^2 - a_{i+1}a_{i-1}) \geq 0$, then multiplying each coefficient by constant factor does not change the inequality.

We now ready to finish the proof of **Theorem 4.1**. Consider the generating function $G_m(x)$ again:

$$G_m(x) = (q + px)^m (p + qx)^{N-m} = p^m q^{N-m} \left(\frac{q}{p} + x\right)^m \left(\frac{p}{q} + x\right)^{N-m}$$

Note that since p and q are probabilities, the expressions in parenthesis are of the from necessary for **Lemma 1** to hold. Which proves that:

$$\frac{a_i^m}{a_i^{m+1}} > \frac{a_i^{m-1}}{a_i^m}$$

Which in turn proves that the differential privacy ratio maximizes when $m = N$ $\qquad\square$

$\square$

4.3   Maximum and Local differential privacy

Since we established a notion of a differential privacy ratio R_i to be a function of the observed number of set bits in the synthetic output, it's instructive to see how this ratio changes with i. Since D consists of set bits, we have for any i

$$P(s = i | D) = \binom{N}{i} p^i q^{N-i} \tag{4.30}$$

$$P(s = i | D_m) = \binom{N-1}{i} p^{i+1} q^{N-i} + \binom{N-1}{i-1} p^{i-1} q^{N-i+1} \tag{4.31}$$

$$R_i = \frac{P(s = i | D_m)}{P(s = i | D)} = \frac{N-i}{N} \frac{p}{q} + \frac{i}{N} \frac{q}{p} \tag{4.32}$$

When all $i = 0$ - all synthetic bits are 0, the ratio reaches its maximum:

$$R_0 = \frac{p}{q}$$

When $i = N$ - the synthetic output consists of set bits entirely, the privacy ratio reaches minimum:

$$R_N = \frac{q}{p}$$

The ratio reduces as i increases, and becomes 1 when number of synthetic bits is equal to expected number of set synthetic bits after randomization:

$$R_{pN} = \frac{N - pN}{N}\frac{p}{q} + \frac{pN}{N}\frac{q}{p} = (1-p)\frac{p}{q} + p\frac{q}{p} = p + q = 1$$

This observation raises a question of reducing absolute theoretical bound of classical differential privacy by considering realistic values of i, rather then all possible outcomes of randomization. Indeed, the probability of all N bits of D generating N zeros is very low. For example, assuming $p = 0.7$, $q = 0.3$ and $N = 100$, the probability of seeing no synthetic ones is $q^{100} = 5e^{-53}$, which is improbable for any realistic scenario. Instead, we should consider values of i that are realistic. In statistical sense, we should only consider values of i that fall within certain number of σ away from the expected mean.

This brings about a notion of a **local differential privacy**, whereby the probabilistic ratio is considered only for values of i that have realistic chance of being observed. Consider the expression for R_i again.

$$R_i = \frac{P(s = i | D_m)}{P(s = i | D)} = \frac{N - i}{N}\frac{p}{q} + \frac{i}{N}\frac{q}{p}$$

The expected number of observed synthetic bits is pN, while the deviation of S random variable is $\sigma = \sqrt{pqN}$. Consider the interval $[pN - 3\sigma, pN + 3\sigma]$. Since the probabilistic ratio grows as i decreased, the maximum ratio will be attained when $i = pN - 3\sigma$. Hence, the local differential privacy reaches maximum at $i = pN - 3\sigma$, and we want to express analytically the relationship between the probabilistic privacy ratio λ, number of records N, and RRT parameters p and q:

$$i = pN - 3\sigma = pN - 3\sqrt{pqN} \tag{4.33}$$

$$R_i = \frac{P(s = i | D_m)}{P(s = i | D)} = \frac{N - i}{N}\frac{p}{q} + \frac{i}{N}\frac{q}{p} \le \lambda \tag{4.34}$$

$$Max(R_i) = \frac{N - pN + 3\sqrt{pqN}}{N} \cdot \frac{p}{q} + \frac{pN + 3\sqrt{pqN}}{N} \cdot \frac{q}{p} \le \lambda \tag{4.35}$$

From here:

$$\frac{N - pN + 3\sqrt{pqN}}{N} \cdot \frac{p}{q} + \frac{pN - 3\sqrt{pqN}}{N} \cdot \frac{q}{p} \le \lambda \tag{4.36}$$

$$p + q + 3\sqrt{\frac{pq}{N}}\left(\frac{p}{q} - \frac{q}{p}\right) \le \lambda \tag{4.37}$$

$$1 + 3\sqrt{\frac{pq}{N}}\frac{p^2 - q^2}{pq} \le \lambda \tag{4.38}$$

$$1 + 3\sqrt{\frac{1}{N}} \cdot \frac{p - q}{\sqrt{pq}} \le \lambda \tag{4.39}$$

$$\frac{pqN}{(p - q)^2} \ge \frac{9}{(\lambda - 1)^2} \tag{4.40}$$

This is an interesting result. Note that left side of inequality is the variance of estimate $\bar{T}$. The local differential privacy grantee simply places a lower bound on the variance of RRT estimates:

$$VAR(\bar{T}) = \frac{pqN}{(p-q)^2} \geq \frac{9}{(\lambda-1)^2} \tag{4.41}$$

For a randomization algorithm applied independently to N bits to be ϵ-differentially private in local sense, means that estimate deviation is lower-bounded by:

$$\sigma(\bar{T}) \geq \frac{3}{\lambda-1} = \frac{3}{e^\epsilon - 1} \tag{4.42}$$

From here, we can express RRT noise parameter q through N and λ:

$$\frac{pqN}{(p-q)^2} \geq \frac{9}{(\lambda-1)^2} \tag{4.43}$$

$$\frac{(1-q)q}{(1-2q)^2} \geq \frac{9}{(\lambda-1)^2 N} \tag{4.44}$$

$$q \geq \frac{1}{2}\left(1 - \frac{1}{\sqrt{1 + 4\frac{9}{(\lambda-1)^2 N}}}\right) \tag{4.45}$$

Suppose $\lambda = 2$ and there are 1000 single bits records in D. The required noise is:

$$q = 0.009$$

Compare that to the level of noise that absolute differential privacy bound would require for $\epsilon = ln(2)$.

$$\frac{p}{q} \leq 2 \tag{4.46}$$

$$q \geq \frac{1}{3} = 0.333 \tag{4.47}$$

The notion of local privacy allowed us to reduce RRT noise 37 times and enabled drastic improvement in estimation accuracy. In the classical case, the estimation deviation is $\sigma = 44.7$, while for the local privacy the deviation is $\sigma = 3$, meaning that precision of RRT estimates had grown 10 fold. It's worth reflecting on what's exactly going on and why such a drastic performance increase is achievable.

Consider confidence intervals for both an original collection D and modified collection D_m. D contains 1000 set bits and D_m contains 999 set bits. Corresponding means and deviation for sum

9

of observed synthetic bits in each case is given below:

$$E(S) = p \cdot 1000 \tag{4.48}$$

$$\sigma(S) = \sqrt{pq \cdot 1000} \tag{4.49}$$

$$E(S_m) = p \cdot 999 + q \tag{4.50}$$

$$\sigma(S_m) = \sqrt{pq \cdot 999 + pq} \tag{4.51}$$

Consider the confidence intervals for both S and S_m for RRT under classical and local differential privacy constrains. If $q = 0.333$ the confidence interval for S and S_m are:

$$S- > [621.98, 711.42] \tag{4.52}$$

$$S_m- > [621.65, 711.09] \tag{4.53}$$

Under local differential privacy, the noise level $q = 0.009$, and the confidence intervals become:

$$S- > [982.04, 999.96] \tag{4.54}$$

$$S_m- > [981.06, 998.98] \tag{4.55}$$

The intervals are nearly identical in either case. Which illustrates the point - we do not need the full power of the absolute differential privacy bound: the local privacy bound will guarantee privacy ratio for 99.98% of possible synthetic outcomes. Effectively, we exploit the noise of large collection to reduce the RRT noise required to randomize each individual record. Rephrasing this important idiom - hiding a record among other records needs less noise than obfuscating a single record.

5   K-randomization for a single bit case

We know consider an important technique for further increasing the estimation precision while providing same local privacy guarantees. Recall from previous example, that if collection D consists of $N = 1000$ records, the corresponding RRT noise at $\lambda = 2$ is $q = 0.009$. We saw that deviation in this case is $\sigma = 3$. Hence our estimation error will be roughly 9 in either direction. We can increase the estimate precision by repeating randomization k times, hence the name **k-ranomization**.

It will be shown that repeating randomization k times achieves increase in precision proportional to $\sqrt{k}$, it also causes slight increase in RRT noise necessary to maintain same differential privacy guarantee. However, the RRT noise increase is usually insignificant compared to the precision gain, which gives a nice dimension to the usual privacy vs. precision tradeoff. K-randomization enables precision increase at the same privacy level for the expense of increasing synthetic record volume k times. Instead of trading privacy for precision, k-randomization allows to trade infrastructure cost for precision while keeping privacy the same. This is especially apparent for long multivariate records, but we will lay mathematical grounds starting from a single bit case.

5.1 Estimating number of single bits under k-randomization

Suppose there are T set bits in the original collection of N single bit records. Each record is randomized k-times. The number of observed synthetic bits S is a random variable expressed as:

$$S = p \cdot kT + q \cdot (kN - kT)$$

The estimate for T, computed from observed value of S is:

$$\bar{T} = \frac{S - qkN}{k(p - q)} \tag{5.1}$$

The aggregator simply divides the estimate computed from kN records by k. The expectation, variance and deviation of $\bar{T}$ random variable are given by:

$$E(\bar{T}) = T \tag{5.2}$$

$$VAR(\bar{T}) = \frac{qpkN}{k^2 \cdot (p - q)^2} = \frac{qpN}{k \cdot (p - q)^2} \tag{5.3}$$

$$\sigma(\bar{T}) = \sqrt{\frac{qpN}{k \cdot (p - q)^2}} \tag{5.4}$$

Note that deviation of the estimate is reduced by $\sqrt{k}$ compared to a single randomization case.

5.2 Choice of D

We now prove that D consisting of only set bits maximizes local differential privacy ratio for any number of observed bits in the randomized collection S. Recall that for a time randomization, the generating function for S given that D contains m set bits is:

$$G_m(x) = (q + px)^{km}(p + qx)^{k(N-m)} = [(q + px)^m(p + qx)^{N-m}]^k = \left[\sum_{i=0}^{N} a_i^m x^i \right]^k = \sum_{j=0}^{kN} b_i^m x^i$$

The generating function for S given that D_m contains $m - 1$ set bits is:

$$G_{m-1}(x) = [(q + px)^{(m-1)}(p + qx)^{(N-m+1)}]^k = \left[\sum_{i=0}^{kN} a_i^{m-1} x^i \right]^k == \sum_{j=0}^{kN} b_i^{m-1} x^i$$

Note that each coefficients b_j^m and b_j^{m-1} are products of a_i^m and a_i^{m-1} with exact same indexes of i. Hence, by Theorem 4.1:

$$\frac{b_j^m}{b_j^{m+1}} = \frac{\prod a_i^m}{\prod a_i^{m+1}} > \frac{\prod a_i^{m-1}}{\prod a_i^m} = \frac{b_j^{m-1}}{b_j^m} \tag{5.5}$$

5.3 Local differential privacy under k-randomization

Consider probabilities of seeing s set bits in the synthetic output for D and D_m respectively:

$$P(S = s|D) = \binom{kN}{s} p^s q^{kN-s} \tag{5.6}$$

$$P(S = s|D_m) = \sum_{i=0}^{k} \binom{k(N-1)}{s-i} p^{s-i} q^{k(N-1)-s+i} \cdot \binom{k}{i} p^{k-i} q^i \tag{5.7}$$

$$P(S = s|D_m) = \sum_{i=0}^{k} \binom{k(N-1)}{s-i} \binom{k}{i} p^{s+k-2i} q^{kN-s-(k-2i)} \tag{5.8}$$

Expressing the privacy ratio at given s, we have:

$$R_s = \sum_{i=0}^{k} \frac{\binom{k(N-1)}{s-i} \cdot \binom{k}{i}}{\binom{kN}{s}} \cdot \frac{p^{k-2i}}{q^{k-2i}} \tag{5.9}$$

Consider the binomial ratio in the sum:

$$\frac{\binom{k(N-1)}{s-i}}{\binom{kN}{s}} = \frac{(kN-k)!}{(kN)!} \cdot \frac{s!}{(s-i)!} \cdot \frac{(kN-s)!}{(kN-s-(k-i))!} = \frac{\prod_{j=0}^{i-1}(S-j) \cdot \prod_{j=0}^{k-i-1}(kN-S-j)}{\prod_{j=0}^{k-1}(kN-j)} \tag{5.10}$$

For positive B, A and e such that $A < B$ the following holds:

$$\frac{A-e}{B-e} < \frac{A}{B} \tag{5.11}$$

Hence the expression in 5.10 is upper bounded by:

$$\frac{\prod_{j=0}^{i-1}(s-j) \cdot \prod_{j=0}^{k-i-1}(kN-s-j)}{\prod_{j=0}^{k-1}(kN-j)} < \frac{\prod_{j=0}^{i-1} s \cdot \prod_{j=0}^{k-i-1}(kN-s)}{\prod_{j=0}^{k-1} kN} = \frac{s^i \cdot (kN-s)^{k-i}}{(kN)^k} \tag{5.12}$$

Dividing each numerator term by kN we arrive to an upper bound of the privacy ratio:

$$R_s < \sum_{i=0}^{k} \left(\frac{s}{kN}\right)^i \left(1 - \frac{s}{kN}\right)^{k-i} \cdot \binom{k}{i} \cdot \frac{p^{k-2i}}{q^{k-2i}} \tag{5.13}$$

Again, under local privacy constrains we compute privacy ratio for s located 3σ bellow the mean:

$$s = pkN - 3\sqrt{pqkN}$$

Replacing s in formula 5.13, we get:

$$\sum_{i=0}^{k} \left(\frac{pkN - 3\sqrt{pqkN}}{kN}\right)^i \left(1 - \frac{pkN - 3\sqrt{pqkN}}{kN}\right)^{k-i} \cdot \binom{k}{i} \cdot \frac{p^{k-2i}}{q^{k-2i}} = \tag{5.14}$$

$$\sum_{i=0}^{k} \left(p - 3\sqrt{\frac{pq}{kN}}\right)^i \left(1 - p + 3\sqrt{\frac{pq}{kN}}\right)^{k-i} \cdot \binom{k}{i} \cdot \frac{p^{k-2i}}{q^{k-2i}} = \tag{5.15}$$

$$\sum_{i=0}^{k} \frac{q^i}{p^i} \left(p - 3\sqrt{\frac{pq}{kN}}\right)^i \cdot \frac{p^{k-i}}{q^{k-i}} \left(q + 3\sqrt{\frac{pq}{kN}}\right)^{k-i} \cdot \binom{k}{i} = \tag{5.16}$$

$$\sum_{i=0}^{k} \binom{k}{i} \left(q - 3q\sqrt{\frac{q}{pkN}}\right)^i \cdot \left(p + 3p\sqrt{\frac{p}{qkN}}\right)^{k-i} = \tag{5.17}$$

$$\left(q - 3q\sqrt{\frac{q}{pkN}} + p + 3p\sqrt{\frac{p}{qkN}}\right)^k = \tag{5.18}$$

$$\left(q + p + 3p\sqrt{\frac{p}{qkN}} - 3q\sqrt{\frac{q}{pkN}}\right)^k = \tag{5.19}$$

$$\left(1 + \frac{3(p - q)}{\sqrt{qpkN}}\right)^k \tag{5.20}$$

Should the differential privacy ratio limit be λ we have the lower bound below:

$$R_s < \left(1 + \frac{3(p - q)}{\sqrt{qpkN}}\right)^k \leq \lambda \tag{5.21}$$

From here we have:

$$\left(1 + \frac{3(p - q)}{\sqrt{qpkN}}\right)^k \leq \lambda \tag{5.22}$$

$$\frac{qpkN}{(p - q)^2} \geq \frac{9}{(\sqrt[k]{\lambda} - 1)^2} \tag{5.23}$$

From here we express required RRT noise through λ, N and k.

$$q \geq \frac{1}{2}\left(1 - \frac{1}{\sqrt{1 + 4\frac{9}{(\sqrt[k]{\lambda}-1)^2 kN}}}\right) \tag{5.24}$$

13

6 multivariate vectors

6.1 Differential privacy ratio

6.1.1 Sufficient Statistics proof

The original collection D consists of N unit vectors of length L. One of the unit-vectors 1 is modified into a zero-vector 0. There are 2^L possible distinct synthetic vectors. Denote v_i a distinct synthetic vector. Denote a observed synthetic configuration S as $s_1, s2, \ldots, s_{2^L}$, whereby s_i represents a count of original vectors that mapped into specific synthetic vector v_i after randomization. Denote D' as a collection of $(N-1)$ unit vectors. Then the probability of generating S from collection D' and a single vector y is given by:

$$P(S|D' + y) = P(s_1 - 1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|y) + \cdots + P(s_1, s_2 - 1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|y) \tag{6.1}$$

Re-writing the ratio

$$\frac{P(S|D' + 0)}{P(S|D' + 1)} = \frac{P(s_1 - 1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|0) + \cdots + P(s_1, s_2 - 1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|0)}{P(s_1 - 1, s_2, \ldots, s_{2^L}|D') \cdot P(v_1|1) + \cdots + P(s_1, s_2 - 1, \ldots, s_{2^L}|D') \cdot P(v_{2^L}|1)} \tag{6.2}$$

$$\frac{P(S|D' + 0)}{P(S|D' + 1)} = \frac{P(v_1|0) + \sum_{i=2}^{2^L} \frac{P(s_1, s_2, \ldots, s_i - 1, \ldots, s_{2^L}|D')}{P(s_1 - 1, s_2, \ldots, s_{2^L}|D')} p(v_i|0)}{P(v_1|1) + \sum_{i=2}^{2^L} \frac{P(s_1, s_2, \ldots, s_i - 1, \ldots, s_{2^L}|D')}{P(s_1 - 1, s_2, \ldots, s_{2^L}|D')} p(v_i|1)} \tag{6.3}$$

Note that distribution of randomized vectors generated by D' is multinomial, since the probability of generating a particular v_i from a unit vector remains constant over all N trials.

$$\frac{P(s_1, s_2, \ldots, s_i - 1, \ldots, s_{2^L}|D')}{P(s_1 - 1, s_2, \ldots, s_i, \ldots, s_{2^L}|D')} = \frac{\frac{(2^L)!}{s_1! \cdot s_2! \ldots (s_i - 1)! \ldots} p(v_1|1)^{s_1} \ldots p(v_i|1)^{s_i - 1} \ldots}{\frac{(2^L)!}{(s_1 - 1)! \cdot s_2! \ldots s_i! \ldots} p(v_1|1)^{s_1 - 1} \ldots p(v_i|1)^{s_i} \ldots} = \tag{6.4}$$

$$\frac{(s_1 - 1)! s_i!}{s_1! (s_i - 1)!} \cdot \frac{p(v_1|1)^{s_1} p(v_i|1)^{s_i - 1}}{p(v_1|1)^{s_1 - 1} p(v_i|1)^{s_i}} = \frac{s_i}{s_1} \cdot \frac{p(v_1|1)}{p(v_i|1)} = \frac{s_i}{s_1} \cdot \frac{q^L}{p(v_i|1)} \tag{6.5}$$

Using that result in the ratio expression we have:

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{p^L + \sum_{i=2}^{2^L} \frac{s_i}{s_1} \cdot \frac{q^L}{p(v_i|1)} p(v_i|0)}{q^L + \sum_{i=2}^{2^L} \frac{s_i}{s_1} \cdot \frac{q^L}{p(v_i|1)} p(v_i|1)} = \frac{s_1 \left(\frac{p}{q}\right)^L + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|1)}}{s_1 + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|1)}{p(v_i|1)}} = \quad (6.6)$$

$$\frac{s_1 \left(\frac{p}{q}\right)^L + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|1)}}{s_1 + \sum_{i=2}^{2^L} s_i} = \frac{s_1 \left(\frac{p}{q}\right)^L + \sum_{i=2}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|1)}}{N} = \frac{1}{N} \cdot \sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|1)} \quad (6.7)$$

Note that if v_i and v_j have same number of set bits, the ratio inside the sum is the same:

$$\text{if } v_i \text{ has same number of set bits as } v_j, \text{ and this number is } l \text{ ,then} \quad (6.8)$$

$$\frac{p(v_i|0)}{p(v_i|1)} = \frac{p(v_j|0)}{p(v_j|1)} = \frac{p^{L-l}q^l}{p^l q^{L-l}} = \left(\frac{p}{q}\right)^{L-2l} \quad (6.9)$$

This allows us to express privacy ratio through counts of synthetic vectors that have same number of set bits l:

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{1}{N} \cdot \sum_{i=1}^{2^L} s_i \cdot \frac{p(v_i|0)}{p(v_i|1)} = \frac{1}{N} \cdot \sum_{l=0}^{L} s_l \cdot \left(\frac{p}{q}\right)^{L-2l} \quad (6.10)$$

Hence, an observer does not gain any more privacy insight by looking at individual vectors than by looking at aggregated counts in a histogram buckets each collecting synthetic vectors with same bit count.

6.1.2 Local differential privacy

As mentioned above, we can equivalently represent collection S by set-bits-histogram counts. For vectors of length L, there are $L+1$ histogram buckets raging from $l=0$ to $l=L$. Let's consider the privacy ratio when the synthetic collection is in the expected state S_e and assume bucket l is sufficiently filled, that is $s_0 \geq 1$. We should represent state S as:

$$S = [s_0, s_1, \ldots, s_L]$$

For the expected synthetic state S_e we choose the state generated from modified collection since this increases privacy ratio. Recall that modified collection $D'+0$ consists of $(N-1)$ unit vectors and one zero valued vector. Hence the distribution S is a some of N independent random vectors of size L consisting of probabilities of finding 1 in a bucket l:

$$S = (N-1) \begin{bmatrix} p(l=0|1) \\ p(l=1|1) \\ \cdots \\ p(l=L|1) \end{bmatrix} + \begin{bmatrix} p(l=0|0) \\ p(l=1|0) \\ \cdots \\ p(l=|0) \end{bmatrix} = N \begin{bmatrix} p(l=0|1) \\ p(l=1|1) \\ \cdots \\ p(l=L|1) \end{bmatrix} + \begin{bmatrix} p(l=0|0) \\ p(l=1|0) \\ \cdots \\ p(l=L|0) \end{bmatrix} - \begin{bmatrix} p(l=0|1) \\ p(l=1|1) \\ \cdots \\ p(l=L|1) \end{bmatrix}$$
(6.11)

Note that probability of generating a synthetic vector containing l set bits from either unit or zero original is given by:

$$p(l|1) = \binom{L}{l} p^l q^{L-l} \tag{6.12}$$

$$p(l|0) = \binom{L}{l} q^l p^{L-l} \tag{6.13}$$

For a single unit vector, the expected value of the privacy ratio is given by:

$$\sum_{l=0}^{L} p(l|1) \cdot \left(\frac{p}{q}\right)^{L-2l} = \sum_{l=0}^{L} \cdot \binom{L}{l} p^l q^{L-l} \left(\frac{p}{q}\right)^{L-2l} = \sum_{l=0}^{L} \binom{L}{l} \cdot p^{L-l} q^l = (p+q)^L = 1 \tag{6.14}$$

For a single zero-valued vector, the expected value of the privacy ratio is given by:

$$\sum_{l=0}^{L} p(l|0) \cdot \left(\frac{p}{q}\right)^{L-2l} = \sum_{l=0}^{L} \cdot \binom{L}{l} p^{L-l} q^l \left(\frac{p}{q}\right)^{L-2l} = \sum_{l=0}^{L} \binom{L}{l} \frac{p^{2L-3l}}{q^{L-3l}} = \tag{6.15}$$

$$\sum_{l=0}^{L} \binom{L}{l} \frac{p^{3L-3l} q^{3l}}{(pq)^L} = \frac{1}{(pq)^L} \sum_{l=0}^{L} \binom{L}{l} (p^3)^{L-l} (q^3)^l = \left(\frac{p^3+q^3}{pq}\right)^L \tag{6.16}$$

Using these expressions we have for the expected state S_e

$$\frac{P(S_e|D'+0)}{P(S_e|D'+1)} = \frac{N-1}{N} + \frac{1}{N}\left(\frac{p^3+q^3}{pq}\right)^L \tag{6.17}$$

The local differential privacy requires that state S should not be too far away from S_e. Which we express as a requirement that the count in every bucket can't be more than 3σ away from expected value. Since, the number of synthetic vectors in each bucket has to sum to N, then the total number

of deviations in each bucket must sum to 0. Which immediately gives us the upper bound for the privacy ratio:

$$\frac{P(S_e|D' + 0)}{P(S_e|D' + 1)} < \frac{1}{N} \cdot \sum_{l=0}^{L} (E(s_l) + 3\sqrt{VAR(s_l)}) \cdot \left(\frac{p}{q}\right)^{L-2l} \tag{6.18}$$

Note that variance in each l-bucket is identical for either unit or zero original vector and is:

$$VAR(s_l) = \frac{1}{N} \cdot \sum_{l=0}^{L} (E(s_l) + 3\sqrt{VAR(s_l)}) \cdot \left(\frac{p}{q}\right)^{L-2l} \tag{6.19}$$

7 ===== IGNORE BELOW THIS POINT =============∎

Suppose there are two vectors of length L in D. One is modified to all-zeros vector. Assume synthetic vectors are nl_1 and l_2, and original collection is vector z and x

$$\frac{P(S|D'+0)}{P(S|D'+1)} = \frac{1}{N} \cdot \sum_{l=0}^{L} N\binom{L}{l}p^l q^{L-l} \cdot \left(\frac{p}{q}\right)^{L-2l} = \sum_{l=0}^{L}\binom{L}{l}p^{L-l}q^l = 1 \tag{7.1}$$

$$\frac{p(l_1|0)p(l_2|x) + p(l_1|x)p(l_2|0)}{p(l_1|z)p(l_2|x) + p(l_1|x)p(l_2|z)} = \frac{\frac{p(l_1|0)}{p(l_1|z)} + \frac{p(l_1|x)p(l_2|0)}{p(l_1|z)p(l_2|x)}}{1 + \frac{p(l_1|x)p(l_2|z)}{p(l_1|z)p(l_2|x)}} \tag{7.2}$$

$$\frac{p(l_1|0)p(l_2|x) + p(l_1|x)p(l_2|0)}{p(l_1|1)p(l_2|x) + p(l_1|x)p(l_2|1)} = \frac{\frac{p(l_1|0)}{p(l_1|1)} + \frac{p(l_1|x)p(l_2|0)}{p(l_1|1)p(l_2|x)}}{1 + \frac{p(l_1|x)p(l_2|1)}{p(l_1|1)p(l_2|x)}} \tag{7.3}$$

$$p(l_1|0) = p^{L-l_1}q^{l_1} \tag{7.4}$$

$$p(l_1|1) = p^{l_1}q^{L-l_1} \tag{7.5}$$

$$\frac{\left(\frac{p}{q}\right)^{L-2\cdot l_1} + \frac{p(l_2|0)}{p(l_2|1)} \cdot \frac{p(l_1|x)p(l_2|1)}{p(l_1|1)p(l_2|x)}}{1 + \frac{p(l_1|x)p(l_2|1)}{p(l_1|1)p(l_2|x)}} \tag{7.6}$$

$$\frac{\left(\frac{p}{q}\right)^{L-2\cdot l_1} + \left(\frac{p}{q}\right)^{L-2\cdot l_2} \cdot p^{(l_2-l_1)}q^{L-(l_2-l_1)} \cdot \frac{p(l_1|x)}{p(l_2|x)}}{1 + p^{(l_2-l_1)}q^{L-(l_2-l_1)} \cdot \frac{p(l_1|x)}{p(l_2|x)}} \tag{7.7}$$

$$\left(\frac{p}{q}\right)^{L-2\cdot l_1} \cdot \frac{1 + \left(\frac{p}{q}\right)^{2(l_1-l_2)} \cdot p^{(l_2-l_1)}q^{L-(l_2-l_1)} \cdot \frac{p(l_1|x)}{p(l_2|x)}}{1 + p^{(l_2-l_1)}q^{L-(l_2-l_1)} \cdot \frac{p(l_1|x)}{p(l_2|x)}} \tag{7.8}$$

$$\left(\frac{p}{q}\right)^{L-2\cdot l_1} \cdot \frac{1 + q^L\left(\frac{p}{q}\right)^{(l_1-l_2)} \cdot \frac{p(l_1|x)}{p(l_2|x)}}{1 + q^L\left(\frac{q}{p}\right)^{(l_1-l_2)} \cdot \frac{p(l_1|x)}{p(l_2|x)}} \tag{7.9}$$

$$\tag{7.10}$$

$$p(S|D) \tag{7.11}$$

$$p(S|D_m) = \sum_{i=1}^{N-1} p(s_i|0)p(S-s_i|D_{N-1}) \tag{7.12}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{\sum_{i=1}^{N} p(s_i|0)p(S-s_i|D_{N-1})}{\sum_{i=1}^{N} p(s_i|1)p(S-s_i|D_{N-1})} \tag{7.13}$$

18

$$p(S|D) = N! \prod_{i=1}^{N} p^{s_i} q^{L-s_i} \tag{7.14}$$

$$p(S|D_m) = \sum_{i=1}^{N} p(s_i|0) \cdot (N-1)! \prod_{j=1, j\neq i}^{N} p^{s_j} q^{L-s_j} \tag{7.15}$$

$$\frac{p(S|D_m)}{p(S|D)} = \sum_{i=1}^{N} p(s_i|0) \cdot \frac{(N-1)! \prod_{j=1, j\neq i}^{N} p^{s_j} q^{L-s_j}}{N! \prod_{i=1}^{N} p^{s_i} q^{L-s_i}} \tag{7.16}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(s_i|0)}{p(s_i|1)} \tag{7.17}$$

$$\frac{p(S|D_m)}{p(S|D)} = K/N \sum_{i=1}^{K} \frac{p(0|0)}{p(0|1)} + M/N \sum_{i=1}^{M} \frac{p(1|0)}{p(1|1)} \tag{7.18}$$

$$p(S|D_m) = \sum_{i=1}^{N} p(s_i|0) \cdot (N-1)! \prod_{j=1, j\neq i}^{N} p^{s_j} q^{L-s_j} \tag{7.19}$$

$$p(S|D) = \sum_{i=1}^{N} p(s_i|x) \cdot (N-1)! \prod_{j=1, j\neq i}^{N} p^{s_j} q^{L-s_j} \tag{7.20}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{\sum_{i=1}^{N} p(s_i|0) \cdot (N-1)! \prod_{j=1, j\neq i}^{N} p^{s_j} q^{L-s_j}}{\sum_{i=1}^{N} p(s_i|x) \cdot (N-1)! \prod_{j=1, j\neq i}^{N} p^{s_j} q^{L-s_j}} \tag{7.21}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{p(s_1|0)/p(s_1|1)+}{1} \tag{7.22}$$

8 WILD WEST

Suppose S consists of 0 and identical s vectors If D is N unit vectors:

$$p(S|D) = Nq^L p(s|1)^{N-1} \tag{8.1}$$

$$p(S|D_m) = p(0|0)p(s|1)^{N-1} + (N-1)p(0|1)p(s|0)p(s|1)^{N-2} \tag{8.2}$$

$$p(S|D_m) = p^L p(s|1)^{N-1} + (N-1)q^L p(s|0)p(s|1)^{N-2} \tag{8.3}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{p^L p(s|1)^{N-1} + (N-1)q^L p(s|0)p(s|1)^{N-2}}{Nq^L p(s|1)^{N-1}} \tag{8.4}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{p^L p(s|1)^{N-1} + (N-1)q^L p(s|0)p(s|1)^{N-2}}{Nq^L p(s|1)^{N-1}} \tag{8.5}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{\left(\frac{p}{q}\right)^L + (N-1)\frac{p(s|0)}{p(s|1)}}{N} \tag{8.6}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{\left(\frac{p}{q}\right)^L + (N-1)\left(\frac{p}{q}\right)^{L-2s}}{N} \tag{8.7}$$

$$\frac{p(S|D_m)}{p(S|D)} = \frac{1}{N}\left(\frac{p}{q}\right)^L + \frac{N-1}{N}\left(\frac{p}{q}\right)^{L-2s} \tag{8.8}$$

$$\tag{8.9}$$

9   MSTACK QUESTION

Bounding ratio of probabilities of Poisson-Binomial Distribution.

There are N Bernoulli trials, where m trails have probability of success p and $N-m$ trails have probability of success $q = 1-p$. Assume $p > q$. The number of successes in these N Bernoulli random variables is a random variable S expressed as a sum of two binomial random variables:

$$S = Bin(p, m) + Bin(q, N-m) \tag{9.1}$$

The distribution of S is known to be a Poisson-Binomial Distribution. I am studying the behavior of the ratio between P(S=k) and P(S=k-1) with respect to m and k. For a given m and number of successes k, denote such ratio as $R(k, m)$:

$$R(k, m) = \frac{P(S = k \mid m)}{P(S = k - 1 \mid m)} \tag{9.2}$$

I am particularly interested in the behavior of this ratio when for small quantiles, and ran numerical simulation for $R(k, m)$ when $k << mean$. It appears that for values of k that are equally distant from the mean, the following holds:

Let $\mu_m = m \cdot (p - q) + N \cdot q$ be the mean of the corresponding distribution and α be the distance away from the mean. Then for $k = \mu_m - \alpha$:

$$R(\mu_0 - \alpha, 0) > R(\mu_1 - \alpha, 1) > \cdots > R(\mu_{m-1} - \alpha, m - 1) > R(\mu_m - \alpha, m) > \cdots > R(\mu_N - \alpha, N) \tag{9.3}$$

The ratio seems to bounded by two binomial distribution for $m = 0$ and $m = N$ respectively. It is easy to see why $R(\mu_0 - \alpha, 0) > R(\mu_N - \alpha, N)$.

$$R(k, 0) = \frac{\binom{N}{k} q^k p^{N-k}}{\binom{N}{k-1} q^{k-1} p^{N-k+1}} = \frac{N - k + 1}{k} \cdot \frac{q}{p} \tag{9.4}$$

$$R(k, N) = \frac{\binom{N}{k} q^k p^{N-k}}{\binom{N}{k-1} q^{k-1} p^{N-k+1}} = \frac{N - k + 1}{k} \cdot \frac{p}{q} \tag{9.5}$$

Setting $k = \mu - \alpha$ for each case, we have

$$R(k, 0) = \frac{N - qN + \alpha + 1}{qN - \alpha} \cdot \frac{q}{p} \approx \frac{qp + \frac{\alpha q}{N}}{qp - \frac{\alpha p}{N}} \tag{9.6}$$

$$R(k, N) = \frac{N - pN + \alpha + 1}{pN - \alpha} \cdot \frac{p}{q} \approx \frac{qp + \frac{\alpha p}{N}}{qp - \frac{\alpha q}{N}} \tag{9.7}$$

$$\frac{qp + \frac{\alpha q}{N}}{qp - \frac{\alpha p}{N}} > \frac{qp + \frac{\alpha p}{N}}{qp - \frac{\alpha q}{N}} \tag{9.8}$$

$$(qp + \frac{\alpha q}{N})(qp - \frac{\alpha q}{N}) > (qp + \frac{\alpha p}{N})(qp - \frac{\alpha p}{N}) \tag{9.9}$$

$$(qp)^2 - \left(\frac{\alpha q}{N}\right)^2 > (qp)^2 - \left(\frac{\alpha p}{N}\right)^2 \tag{9.10}$$

Since $p > q$, the above inequality holds.

But I am unable to express $R(k, m)$ analytically, no prove that it is bounded between by ratios corresponding to the limiting binomial distribution. Any help with proving this and/or pointers to related papers will be much appreciated