On Missing Mass Variance

Maciej Skorski
University of Luxembourg

Abstract—The missing mass refers to the probability of elements not observed in a sample, and since the work of Good and Turing during WWII, has been studied extensively in many areas including ecology, linguistic, networks and information theory. This work determines what is the maximal variance of the missing mass, for any sample and alphabet sizes. The result helps in understanding the missing mass concentration properties.

Index Terms—Missing Mass, Sample Coverage, Second Moment, Measure Concentration

I. INTRODUCTION

A. Background
Let $X_1, \ldots, X_n \sim_{\text{IID}} p$ be a sample from a distribution $(p_s)_{s \in S}$ on a countable alphabet $S$, and $f_s = \#\{i : X_i = s\}$ be the empirical (observed) frequencies. The missing mass
\[ M_0 = \sum_s p_s I(f_s = 0), \]
which quantifies how much of the population does not appear in the sample, is of interest to statistics [1] and applied disciplines, such as: ecology [2]–[5], linguistic [6]–[9], archaeology [10], network design [11], [12], information theory [13], [14], microbiology [15] and bio-molecular modeling [16], [17]. The problem was studied first by Good and Turing [18].

This work studies the maximal variance of the missing mass
\[ \max_{(p_s) \in \mathcal{P}(S)} \text{Var}[M_0] = ? \]
given the alphabet size (equivalently, the constrained support):
\[ \#S = m, \]
where we allow for $m = +\infty$ and denote by $\mathcal{P}(S)$ the set of probability measures on $S$. We are interested in deriving a formula valid for any sample size $n$ and any alphabet size $m$, with the relative error of at most $o(1)$ when $n$ grows.

The motivation is two-fold. First, our work continues the line of research on extreme missing mass properties [19]–[22]. Second, more importantly, understanding the variance extremes determines no-go regimes for concentration inequalities, which have been actively studied [23]–[29]; the need for such no-go results was highlighted in prior works [19], [20].

B. Related Work
No prior work has addressed the problem of finding sharp variance bounds for the missing mass, even under no alphabet constraints. Below we review weaker results that are available.

The maximal variance is $\Theta(n^{-1})$ under no constraints ($m = +\infty$). This has been noticed in the line of work on concentration properties [23]. Furthermore, this asymptotic holds up to a constant for many distributions with certain regularity properties [26]. The asymptotic $\Theta(n^{-1})$, when $m = +\infty$, can be also deduced from works that studied the Good-Turing estimator [19], [20]; unfortunately the estimator variance differs from our variance already by $\Omega(n^{-1})$, and so it is insufficient to give the sharp asymptotic.

As we will see, in the context of our problem techniques from prior works are not sufficient. While certain tricks [19], [20] can used to handle the smaller order term, determining the leading asymptotic term involves subtle non-linear programming, mainly due to the support constraint. Analogues programs from prior works on extreme missing mass properties luckily had simpler objectives and boundaries. For example, the problem [21] is straightforward due to Schur-convexity [30]; in [20] the reduced program involves only one equality constraint; in [19] the constant was found up to a small gap, but with elementary methods avoiding optimization.

The reader interested in missing mass beyond the scope of our problem may also refer to other works [31]–[38].

C. Summary of Contribution
The contribution of this work can be summarized as follows:

- The formula accurately approximating missing mass variance approximation, with the poissonized version which is better suited for analysis and numerical evaluation.
- The characterization of the worst-case variance value, along with the maximizing distribution, for any sample and alphabet size. The maximal variance behaves like $\Theta(\min(\frac{1}{m}, \frac{1}{n}))$, with the explicit leading constant. The maximizer is a mixture of a uniform distribution and a point mass, with the explicitly given proportions.
- An application of the extreme variance values to benchmarking results on missing mass concentration.

II. RESULTS

A. Accurate Variance Estimation
We present the accurate bound on the missing mass variance, valid for any discrete distribution. The derivation is similar as for the formula for the Good-Turing estimator, obtained in [19], although quantitatively quite different.

Theorem 1 (Missing Mass Variance). We have:
\[ \text{Var}[M_0] = -n \left( \sum_s p_s^2 (1 - p_s)^n \right)^2 + n \sum_s p_s^3 (1 - p_s)^n + O(n^{-2}). \]
We can further approximate the variance, similarly as for moments of occupancy numbers [39], [40], with "poissonized" terms. This makes the formula easier to analyze and also more stable numerically (via the "log-sum-exp" trick [41]).

**Corollary 1 (Missing Mass Variance, Poissonized).** We have:

\[
\text{Var}[M_0] = -n \left( \sum_s p_s^2 e^{-np_s} \right)^2 + n \sum_s p_s^3 e^{-np_s} + O(n^{-2}). \tag{5}
\]

**B. Extreme Variance Behavior**

Building on Corollary 1, we use non-linear optimization to characterize the maximal variance. The result is stated below:

**Theorem 2.** Denote \( m = \#S, b = \frac{m}{w} \). Consider the program:

\[
\begin{align*}
\max_{w,c} \quad & \alpha(w, c) = -w^2 e^{-2c} + wc^2 e^{-c} \\
\text{s.t.} \quad & 0 \leq w \leq 1 \\
& \leq w = bc,
\end{align*}
\]

and let \( \alpha \) be its optimal value achieved at \( w, c \). Then:

\[
\max_{(p_s) \in \mathcal{P}(S)} \text{Var}[M_0] = \frac{\alpha}{n} + O(n^{-2}), \tag{7}
\]

with the maximum achieved at \( (p_s) \) being the mixture of a uniform distribution on \( \min\{\lfloor wn/c \rfloor, m-1 \} \) elements and a Dirac mass, with the proportion \( w \) and \( 1-w \) respectively.

For \( m = +\infty \), the constraints reduce to \( 0 \leq w \leq 1, 0 \leq c \).

The shape of the worst-case distribution, a Uniform-Dirac mixture, appears often for extreme problems in nonparametric statistics [42] and could be expected; the proof is however quite subtle. The optimization is illustrated in Figure 1.

![Figure 1: Optimization landscape of the problem in Theorem 2.](image)

Finally, we give a reduction to a one-dimensional problem, which can be solved numerically for any given \( m, n \).

**Corollary 2.** Let \( c^* = 2.26281 \ldots \) be the root of \( 2 - 2e^c + c(-2 + e^c) = 0 \). Then the optimal solution for Theorem 2 is:

\[
\alpha = \begin{cases} 
\max_{0 \leq c \leq \frac{1}{2}} -c^2 e^{-2c} + c^2 e^{-c} = 0.477 \ldots & \frac{m}{n} \geq \frac{1}{c^*} \\
\max_{0 < c < \frac{1}{2}} -c^2 e^{-2c} + bc^3 e^{-c} = \frac{m}{n} < \frac{1}{c^*},
\end{cases}
\tag{8}
\]

and \( w \) equals 1 and respectively \( bc \), where \( c \) is the maximizer.

We see that the constant in the maximal variance asymptotic depends on the ratio of the sample size and alphabet size. Furthermore, for sufficiently large \( n \) the Dirac part vanishes (phase transition), as illustrated in Figure 2.

![Figure 2: The constant \( \alpha \) in the asymptotic \( \Theta(n^{-1}) \) for the maximal missing mass variance (2), depending on the alphabet size \( m = \#S \) (Corollary 2). The phase transition (the maximizer becomes uniform) occurs at \( m = \frac{1}{c^*} \).](image)

C. Application: Gap in Missing Mass Concentration

We show how our result can be used to benchmark concentration bounds. This solves the problem posed in [20].

The state-of-the-art bounds for the concentration of \( M_0 \) [26] give the following: \( M_0 \) is sub-gamma with variance-factor \( v = \Theta(n^{-1}) \) and scale \( c = O(n^{-1}) \), and thus satisfies the concentration bound \( \Pr[|M_0 - \mathbb{E}[M_0]| > \epsilon] \leq 2 \exp \left( \frac{-\epsilon^2}{2v^2 + 2\epsilon c} \right) \) (in particular \( 2e^{-\Omega(n^{-2})} \) as we can assume \( \epsilon < 1 \)). However we know that \( \epsilon^2 \geq \text{Var}[M_0] \) for the sub-gamma bound, and this is sharp [43]. Using our results, we observe that:

- The biggest variance is \( \text{Var}[M_0] \approx 0.477 \). In turn, by the numeric formula [43], we have \( v = \sum_p p_s^2(1-p_s)n + n^{-1} \sum_s p_s^3(1-p_s)^n \) which can be as big as \( v \approx 0.839 \). This leaves the gap of \( \frac{0.862}{\epsilon} \) w.r.t. the "ideal" case.
- Missing mass is a sum of negatively-dependent variables, and the all prior works on missing mass concentration handle this by the IID majorization. However ignoring correlation terms gives the variance-term of at least \( v \approx \sum_s p_s^2(1-p_s)^n - (1-p_s)^{2n} \), which can be as far as \( \Omega(n^{-1}) \) from the "ideal" value of \( \text{Var}[M_0] \).

III. PRELIMINARIES

We will need the mean-value theorem to obtain asymptotic bounds, as well as to local and count roots of equations.

**Lemma 1 (Mean-Value Theorem [44]).** If a function \( f(u) \) is continuous on the closed interval \([a, b]\) and differentiable on the open interval \((a, b)\), then there exists a point \( c \in (a, b) \) such that \( \frac{f(b) - f(a)}{b-a} = f'(c) \).

We will also need the facts below in our estimations:
**Lemma 2** (Generalized-Bernoulli Inequality). Let $1 \leq k \leq n$ be integers. Then for any $-1 \leq u \leq 0$ it holds that: $1 + nu \leq (1 + u)^n \leq 1 + nu + O(nu^2)$.

**Lemma 3** (Taylor’s Expansion with Remainder). Let $f$ be a twice differentiable real function, then $f(x) = f(x_0) + \frac{df}{dx}(x_0) \cdot (x - x_0) + \frac{d^2 f}{dx^2}(z)(x - x_0)^2$ for some $z \in [x, x_0]$.

Another fact handles the commonly occurring expressions:

**Lemma 4** (Occuancy-like Expressions). Let $k, n \geq 1$. Then $p^k(1-p)^n$ for $p \in [0,1]$ is maximized for $p = \frac{k}{k+n}$, and $p^k e^{-np}$ for $p \in [0,1]$ is maximized for $p = \frac{k}{n}$. In particular if $(p_s)$ is a probability distribution, then $\sum p_s^k(1-p_s)^n = O(n^{1-k})$ and $\sum p_s^k e^{-np_s} = O(n^{1-k})$ when $k = O(1)$.

Finally, we will need some theory of non-linear programming. For a detailed discussion we refer to books [45]–[47].

**Lemma 5** (KKT Conditions). Consider the program

\[
\begin{align*}
\text{max} & \quad f(x) \\
\text{s.t.} & \quad h_i(x) = 0, \quad i \in I \\
& \quad g_j(x) \leq 0, \quad i \in J
\end{align*}
\]

(9)

with differentiable real functions $f, (h_i)_{i \in I}, (g_j)_{j \in J}$ in variables $x = x_1, \ldots, x_d$. If the maximum occurs at $x$, then:

\[
\frac{\partial}{\partial x} f(x) = \sum_i \lambda_i \frac{\partial}{\partial x} h_i(x) + \sum_j \mu_j \frac{\partial}{\partial x} g_j(x),
\]

where $\frac{\partial}{\partial x} = (\frac{\partial}{\partial x_1} \ldots \frac{\partial}{\partial x_d})$, for $\lambda_i \in \mathbb{R}$, $\mu_j \geq 0$ such that

\[
\lambda_i \in \mathbb{R}, \quad \mu_j \geq 0, \quad \mu_j g_j(x) = 0,
\]

(10)

provided that regularity conditions hold at $x$.

We briefly remind the optimization terminology. The function $f$ is called objective, and any $x \in \mathbb{R}^d$ satisfying the constraints is called feasible. The optimal value is also called the program value. The constraint $h_i$ respectively $g_j$ is called active at $x$, when $h_i(x) = 0$, respectively $g_j(x) = 0$.

**Remark 1** (Constraints Qualification). The KKT conditions hold for optimal $x$ when the active constraints gradients are linearly independent (LICQ) or for affine constraints (LQC).

**Remark 2** (Extreme Values [48]). A continuous function on a compact subset of $\mathbb{R}^d$ achieves its maximum and minimum.

**IV. PROOFS**

**A. Proof of Theorem 1**

Let $\xi_s$ be the indicator that $s$ does not occur in the sample. By **Lemma 2**:

\[
\text{Var}[\xi_s] = (1-p_s)^n - (1-p_s)^2
\]

\[
= (1-p_s)^n (1 - (1-p_s)^n)
\]

\[
= np_s (1-p_s)^n(1 + O(np_s)).
\]

In turn, for $s \neq s'$ by **Lemma 1**:

\[
\text{Cov}[\xi_s, \xi_{s'}] = (1-p_s-p_{s'})^n - (1-p_s)(1-p_{s'})^n
\]

\[
= -np_s p_{s'} (1-p_s)(1-p_{s'})^n + O(n^2 p_s^2 p_{s'}^2 (1-p_s)^n - (1-p_{s'})^n).
\]

(13)

These bounds imply that:

\[
\sum_s p_s^2 \text{Var}[\xi_s] = n \sum_s p_s^3 (1-p_s)^n + O(n^{-2}),
\]

(14)

where we used **Lemma 4** to estimate the $O(\cdot)$ term, and

\[
\sum_{s \neq s'} p_s p_{s'} \text{Cov}[\xi_s, \xi_{s'}] = -n \sum_{s \neq s'} p_s^2 p_{s'}^2 (1-p_s)^n - (1-p_{s'})^n + O(n^{-2}),
\]

(15)

where we estimated the $O(\cdot)$ term using $n^2 \sum_{s \neq s'} p_s^2 p_{s'}^2 (1-p_s)^n (1-p_{s'})^n < (n \sum_s p_s^2 (1-p_s)^n - 2) = O(n^{-2})$ due to **Lemma 4**; since $n \sum_s p_s^2 (1-p_s)^n = O(n^{-2})$ by **Lemma 4**, we obtain the following simpler expression:

\[
\sum_{s \neq s'} p_s p_{s'} \text{Cov}[\xi_s, \xi_{s'}] = \text{Var}[M] = -n (\sum_s p_s^2 (1-p_s)^n)^2 + O(n^{-2}).
\]

(16)

Since $\sum_s p_s^2 (1-p_s)^n = \sum_s p_s^3 (1-p_s)^n + O(n^{-2})$, as seen by $p_s^2 (1-p_s) = p_s^2 (1-p_s)^n (1-2p_s + O(p_s^2))$ and **Lemma 4**, we can replace the exponent $n-2$ with $n$.

By the identity $\text{Var}[M] = \sum_s p_s \text{Var}[\xi_s] + \sum_{s \neq s'} p_s p_{s'} \text{Cov}[\xi_s, \xi_{s'}]$, and the bounds we derived above for the two parts of the summation:

\[
\text{Var}[M_0] = -n \left(\sum_s p_s^2 (1-p_s)^n\right)^2 + n \sum_s p_s^3 (1-p_s)^n + O(n^{-2}).
\]

(17)

At the final step we observe $e^{-np_0} - (1-p)^n \leq O(np^2 e^{-np})$ by **Lemma 1** applied to $f(u) = u^n$. By **Lemma 4**, we conclude that in the above identity we can replace $1-p$ by $e^{-p}$ making the error of $O(n \sum_s p_s^3 e^{-np}) = O(n^{-2})$ in the first sum and $O(n \sum_s p_s^5 e^{-np}) = O(n^{-3})$ in the third sum. Both sums are $O(n^{-1})$, thus the total additive error is $O(n^{-2})$.

**B. Proof of Theorem 2**

1) Non-linear Programming Formulation: In view of **Corollary 1**, we need to solve the optimization program:

\[
\begin{align*}
\text{max} & \quad -n \left(\sum_s p_s^2 e^{-np_s}\right)^2 + n \sum_s p_s^3 e^{-np_s} \\
\text{s.t.} & \quad \sum_s p_s = 1, \quad p_s \geq 0.
\end{align*}
\]

(18)

Observe that $(p_s)$ may be not finite-dimensional, but rather countable, so it is not clear whether the maximum is achieved on a feasible point. However, the objective is bounded, so the supremum may be achieved on a sequence of feasible points.
2) Maximizer is 4-Mixture: We will show that every feasible \((p_s)\) with at least 4 distinct non-zero values can be improved, that is replaced by a feasible \((p_s)\) with the strictly bigger objective value. To this end we fix the values of \(p_s\) on all indices except a finite subset \(S'\) and consider the program

\[
\max -n \left( \sum_{s \in S'} p_s^2 e^{-np_s} + B \right)^2 + n \sum_{s \in S'} p_s^3 e^{-np_s} + C \\
\text{s.t.} \quad \sum_{s} p_s = A, \ p_s \geq 0.
\]

for some \(A, B, C > 0\). It suffices to prove that regardless of values of \(A, B, C\), the maximum is achieved for \((p_s)_{s \in S'}\) which takes at most 3 distinct non-zero values. By Remark 2 the maximum is indeed achieved at some feasible point. By Lemma 5 the non-zero components of the optimal point \((p_s)_{s}\) are solutions of the following equation in \(u\):

\[
-2n \cdot D \cdot \frac{\partial g}{\partial u}(u) + n \cdot \frac{\partial^2 g}{\partial u^2}(u) = \lambda,
\]

where \(D = 2(\sum_{s \in S'} p_s^2 e^{-np_s} + B)\), for some constant \(\lambda\). The equation is of the form \(P(u) e^{-nu} = \lambda\) where \(P(u)\) is a polynomial of degree 3. We claim that such an equation has at most 4 solutions; to this end it suffices to show that the derivative of \(P(u)e^{-nu}\) has at most 3 solutions; indeed by Lemma 1 between two points with equal function value there is a root of its derivative (a.k.a. Rolle’s Theorem). But the derivative is \(P'(u)e^{-nu}\) for a degree-3 polynomial \(P'\), and has at most 3 roots by the Fundamental Theorem of Algebra.

This claim shows that, the equivalent program is

\[
\max_{(p_s), (k_s)} -n \left( \sum_{s=1}^{4} k_s p_s^2 e^{-np_s} \right)^2 + n \sum_{s=1}^{4} k_s p_s^3 e^{-np_s} \\
\text{s.t.} \quad \sum_{s=1}^{4} k_s p_s = 1, \ p_s > 0, \ k_s \geq 0, \ k_s \in \mathbb{Z},
\]

whether the maximum in (18) is achieved at a concrete point, or in the limit on a sequence of points. We can state it equivalently: when seeking the maximizer we can restrict the feasible set to mixtures of (at most) 4 uniform distributions.

3) Relaxation: We prove, as before, that (21) holds when \(S\) is bounded. Let \(m = \#S\), consider the relaxed program:

\[
\max_{(p_s), (k_s)} -n \left( \sum_{s=1}^{4} k_s p_s^2 e^{-np_s} \right)^2 + n \sum_{s=1}^{4} k_s p_s^3 e^{-np_s} \\
\text{s.t.} \quad \sum_{s=1}^{4} k_s p_s = 1, \ p_s > 0, \ k_s \leq 1, \ k_s \leq m
\]

We prove that the global maximum is achieved. Consider the supremum of the objective under the constraints; it can be approached in the limit of the objective values on the sequence of feasible points \((k_j^s), (p_j^s)\) for \(j = 1, 2, \ldots\). By the diagonal argument (passing to a subsequence) we can assume that for every \(s\) the sequences \(k_j^s, p_j^s\) are monotone. Observe now that \(p_j^s \to \infty\) for \(j \to \infty\), then the objective in the limit and the feasibility does not change when we replace \(p_j^s = 0\); this is because \(p^2 e^{-np} \to 0\) for \(p \to \infty\). We can thus assume that \((p_j^s)\) are bounded. Suppose now that \(k_j^s \to \infty\) for some \(s\), \(j \to \infty\). Then it must be that \(p_j^s = 0\) because \(k_j^s p_j^s \leq 1\); then the objective in the limit and the feasibility do not change when we replace \(k_j^s = 0\) for all \(j\). We can thus assume that \(k_j^s\) are also bounded. Since \(p_j^s\) are bounded and \(k_j^s\) are bounded, they converge to finite values \(p_s^*, k_s^*\). By the continuity, the sequence of objective values converges to the objective on \(p_s^*, k_s^*\). Furthermore, these values are feasible, as the constraint set is closed. This shows that the supremum is achieved.

4) Solving Relaxation: Note that the constraints are bilinear; thus we can apply the KKT conditions for with respect to \(p_s\) with optimal \(k_s\), and with respect to \(k_s\) with fixed optimal \(p_s\). At the optimal solution \((p_s), (k_s)\) for \(s\) such that \(p_s k_s > 0\) we have that \(u = p_s\) solves the system of equations

\[
\frac{\partial g}{\partial u}(u) = \lambda u + \mu \quad \frac{\partial g}{\partial u}(u) = \lambda
\]

where we introduced the auxiliary function

\[
g(u) = (u^3 - A u^2) e^{-nu}
\]

with \(A = 2 \sum_{s} k_s p_s^2 e^{-np_s}\). The first equation in (23) is the KKT condition for \((k_s)\) divided by \(n\), and the second equation in (23) is the KKT condition for \((p_s)\) divided by \(n\). Geometrically, (23) means that the line \(u \to \lambda u + \mu\) is tangent to the plot of \(g\) at \(u\). We will now argue that this does not happen for distinct values of \(u\), so the system has one solution.

This is best seen on Figure 3, but the formal argument is rather subtle. Suppose that (23) has two solutions \(u_1 < u_2\). The derivative \(\frac{d^2 g}{d u^2}(u) = -e^{-nu} u (2 - nu) + u (-3 + nu)\) has 3 roots which split the domain into intervals \(I_1, I_2, I_3\) such that \(g\) is decreasing on \(I_1\), increasing on \(I_2\) and decreasing on \(I_3\); the endpoints are the roots 0, \(\frac{3 + A n - \sqrt{9 - 4 A n + 4 A^2 n^2}}{2n}\), \(\frac{3 + A n + \sqrt{9 - 4 A n + 4 A^2 n^2}}{2n}\). We then claim that on every interval \(I_j\) the derivative takes every value at most twice. Otherwise, one of the intervals contains two roots of the second derivative \(\frac{d^2 g}{d u^2}\), by Lemma 1 (applied to the repeating values); since every interval contains at least one root of the second derivative, by Lemma 1 applied to the endpoints (for \(I_3\), we use \(+\infty\) as the right endpoint), in total this gives 4 roots. On the other hand the roots of \(\frac{d^2 g}{d u^2}\) satisfy the polynomial equation of order 3 with at most 3 roots, a contradiction.

Thus, for the optimal solution of (22) \(k_s p_s \neq 0\) for only one \(s\). In other words, the program is equivalent to:

\[
\max_{p, k} -n k^2 p^4 e^{-2np} + n k p^3 e^{-np} \\
\text{s.t.} \quad \left\{ \begin{array}{l}
kp \leq 1 \\
0 \leq k \leq m \\
0 \leq p
\end{array} \right.
\]
5) \textbf{Relaxation Gap:} Consider the optimal solution \((p, k)\) for (25). We will slightly modify it so that it is feasible for (21), but the relaxed objective does not change much. Define:

\[
k_1 = \min\{[k], m - 1\}, \quad p_1 = p
\]
\[
k_2 = 1, \quad p_2 = 1 - k_1 p_1
\]
\[
k_3, k_4, p_3, p_4 = 0.
\]

Consider the following quantities:

\[
\begin{align*}
P' &= -n \left( \sum_s k_s p_s^2 e^{-np_s} \right)^2 + n \sum_s k_s p_s^3 e^{-np_s} .
\end{align*}
\]
\[
P = -n(k p e^{-np})^2 + n k p^3 e^{-np}.
\]

We then have:

\[
\Delta_1 \triangleq n \left( \sum_s k_s p_s^2 e^{-np_s} \right)^2 - n(kp^2 e^{-np})^2
\leq n(\sum_s k_s p_s^2 e^{-np_s} - kp^2 e^{-np}) \cdot O(n^{-1})
= (k_1 - k) p^2 e^{-np} + O(n^{-2}) \cdot O(1)
\leq O(n^{-2}).
\]

Where in the second line we use the identity \(x^2 - y^2 = (x - y)(x + y)\) with \(x = \sum_s k_s p_s^2 e^{-np_s}, y = kp^2 e^{-np}\) and the fact that \(x + y = O(n^{-1})\) by Lemma 4, in the third line we use again Lemma 4 and \(|k_1 - k| \leq O(1)\) by the definition of \(k_1\). Similarly we obtain:

\[
\Delta_2 \triangleq n k p^3 e^{-np} - n \sum_s k_s p_s^3 e^{-np_s}
= n(k - k_1) p^3 e^{-np} - O(n^{-2})
\leq O(n^{-2}),
\]

where we used Lemma 4 (twice) and \(|k_1 - k| \leq O(1)\).

Therefore \(P - P' = \Delta_1 + \Delta_2 \leq O(n^{-2})\), since also \(P' \leq P\) (by relaxation) we finally obtain:

\[
|P - P'| \leq O(n^{-2}).
\]

6) \textbf{Concluding Result:} By the change of variables \(w = kp, c = pn\) and \(b \equiv \frac{m}{n}\) we can rewrite (25) as follows:

\[
\begin{align*}
\max_{w,c} & \quad \frac{-w^2 c^2 e^{-2c} + w^2 e^{-c}}{n} \\
\text{s.t.} & \quad \begin{cases}
0 \leq w \leq 1 \\
0 \leq c. 
\end{cases}
\end{align*}
\]

This proves the first part of Theorem 2 (the maximal value); note that the relaxation gap of \(O(n^{-2})\) with respect to the program (18) and the difference of \(O(n^{-2})\) between this program and the variance, give the total error of \(O(n^{-2})\).

For the second part (the maximizing distribution) we use the construction (26), which gives the stated characterization.

\textbf{C. Proof of Corollary 2}

Consider the program (31). The maximum is achieved, as inherited from (22) (it follows also directly: the objective is non-negative and goes to zero when \(c \to \infty\), uniformly in \(w\)); in fact is is achieved for \(w, c > 0\) (then the objective, under the constraints, is positive; for \(w c = 0\) it equals zero).

We first note that the optimal solution must be on the boundary. Otherwise if \(0 < w < 1\) and \(0 < w < b c\) the first order conditions give \(w = \frac{c^2}{2} e^{-c} c^2 e^{-c} w (2(e - 1) w - (e - 2) c^2) = 0\), which leads to \(w = \frac{c^2}{2} e^{-c}\) and \(2(c - 1) w - (e - 2) c^2 = 0\), which in turn leads to a contradiction that \(c - 1 = c - 2\).

Thus our optimal value equals \(\frac{c^2}{2}\) times the maximum of the following two cases which determine the constant \(\alpha\): either

\[
c > \frac{1}{e},
\]

or

\[
\max_{0 < c \leq \frac{1}{e}} -c^2 e^{-2c} + c^2 e^{-c}.
\]

For numerical optimization below we used the software [49].

Suppose that \(b \leq \frac{1}{e}\). Let \(g(c) = -c^2 e^{-2c} + c^2 e^{-c}\); this function has the unique maximizer \(c^* \approx 2.6281\), is increasing for \(0 < c < c^*\) and decreasing for \(c > c^*\). Then (32) equals \(g(1/b)\) which matches the objective of (33) at \(c = \frac{1}{b}\); thus the optimal value of (32) is smaller or equal to that of (33).

Suppose that \(b > \frac{1}{e}\). Then (32) equals \(g(c^*)\). We will prove that (33) is smaller or equal to \(g(c^*)\). This is true when \(\frac{c^*}{2} < 1\), because then (33) is at most \(c^2 e^{-c}\) maximized over \(c \leq \log 2\), which is 0.24... (we used \(bc \leq 1\)). We can assume that \(1 < \frac{c^*}{2}\); note that the objective of (33) increases in \(b\) when \(b < \frac{1}{c^*}\), and under the constraint we have \(0 < b \leq \frac{1}{c^*}\). Thus, (33) can be upper-bounded by setting \(b = \frac{1}{c^*}\) in the objective and removing the constraint on \(c\). This gives the upper bound of \(\max_c g(c) = g(c^*)\), and finishes the proof.

\textbf{V. Conclusion}

This work shows the exact behavior of the missing mass variance in the extreme case. The result can be used to benchmark concentration results and to understand no-go regimes. The optimization techniques can be possibly generalized and used to obtain similar characterization for other occupancy numbers, such as small counts recently studied [50].
