Bounds for matrices of inclusion probabilities in rejective sampling

Simon Ruetz FIRSTNAME.LASTNAME@UIBK.AC.AT
Karin Schnass
University of Innsbruck
Technikerstraße 13
6020 Innsbruck, Austria

Abstract

Motivated by a desire to model n-sparse signals more realistically, we study matrices of inclusion probabilities for conditional Poisson or rejective sampling in the non-asymptotic regime. In particular, we provide bounds in the semi-definite ordering for matrices that collect first and second order inclusion probabilities as their diagonal and off-diagonal entries respectively, and operator norm bounds of their Hadamard products with other matrices. Finally, we demonstrate the usefulness on these bounds, which only involve first order inclusion probabilities, in a toy application from dictionary learning.

Keywords: conditional Poisson sampling; conditional Bernoulli sampling; Hadamard product, operator norm bounds; semi-definite order bounds

1. Introduction

The goal of this paper is to provide bounds for matrices of inclusion probabilities in rejective sampling and thus make rejective sampling a practicable model for n-sparse signals. Rejective sampling, known also as conditional Poisson/Bernoulli sampling, selects a subset S of size n from $[K] = \{1, \dots, K\}$, by flipping K independent Bernoulli switches δ_k with $\mathbb{P}(\delta_k = 1) = p_k \in (0, 1)$ and keeping the set of active indices $S = \{k : \delta_k = 1\}$ if it has the correct size. First order inclusion probabilities then count how often such a set contains a fixed index i, $\pi_i(n) = \mathbb{P}(\delta_i = 1 | \sum_k \delta_k = n) = \mathbb{P}_n(i \in S)$, and higher order inclusion probabilities how often they contain a fixed set $L = \{i_1, \dots, i_\ell\}$, $\pi_L(n) = \pi_{i_1 \dots i_\ell}(n) = \mathbb{P}_n(L \subseteq S)$. The bounds presented here are of the following form.

Featured Theorem A *Consider rejectively sampling sets of size n with generating probabilities p. Let $\pi = \pi(n)$ be the vector of first order inclusion probabilities, D_π the diagonal matrix with π on the diagonal and Σ_n be the matrix of first and second order inclusion probabilities, ie. with ij-th entry $\pi_{ij}(n)$ and $\pi(n)$ on the diagonal. Further, let $\preceq$ denote the semi-definite order and $\odot$ the Hadamard (entrywise) product. Then we have*

$$\Sigma_n \preceq \pi\pi^t + 2D_\pi \quad \textit{as well as}$$
$$\|A \odot \Sigma_n\| \leq \kappa \cdot \|D_\pi[A - A \odot \mathbb{I}]D_\pi\| + \|A \odot D_\pi\|,$$

for $A \in \mathbb{R}^{K \times K}$ and $\kappa = (1 + \|\pi\|_\infty)(1 - \|\pi\|_\infty)^{-2}$.

Since the bounds above are valid in the non-asymptotic setting, they make rejective sampling usable and useful for people working in sparse signal processing, data science or machine learning, where population size K and sample size n cannot be assumed to be large. Indeed, our main message is, that if you have an equality in uniform n-sampling or Poisson sampling that features n/K or p, you can often get an analogue inequality for rejective sampling, featuring instead π or even p with a benign constant.
To motivate the renewed interest in a classical topic such as inclusion probabilities of rejective sampling, we next provide a motivation for and introduction to rejective sampling aimed at practitioners and interested sampling experts using an application example in the modelling of sparse signals. Section 2 contains the main result in more detail together with some useful auxiliary results. We also provide a short application example in dictionary learning, while a more elaborate example can be found in Appendix A. Finally, Section 3 contains the proofs of the main result.

1.1 Motivation and rejective sampling model

One of the most useful priors in signal processing, e.g. for signal acquisition and restoration such as compressed sensing, denoising or inpainting, [8, 7, 14], is that the signals of interest are sparse in a dictionary or basis. This means that there exists a set of vectors $\phi_i \in \mathbb{R}^d$, called atoms and collected in the dictionary $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d\times K}$, such that we can write each signal y of interest as

$$y = \Phi x = \sum_{k=1}^{K} x(k)\phi_k = \sum_{k\in I} x(k)\phi_k,$$

where the coefficients $x \in \mathbb{R}^K$ are n-sparse, which means that only n entries $x(k)$ are non-zero. The (sorted) set $I = \{k : x(k) \neq 0\}$, which collects all indices k of non-zeros entries $x(k)$, is then called the support of the sparse signal and sometimes, it will be convenient to rewrite the sparse signals as superposition of scaled dictionary atoms in the support in several ways, meaning for $\Phi_I = (\phi_{i_1}, \dots, \phi_{i_n})$ and $x_I = (x_{i_1}, \dots, x_{i_n})$, where we usually assume that $i_j \leq i_{j+1}$, we have

$$y = \Phi_I x_I = \Phi D_x \mathbf{1}_I = \Phi D_I x.$$

Here $\mathbf{1}_I$ denotes the indicator vector of the set $I \subseteq [K] = \{1, \dots, K\}$, meaning $\mathbf{1}_I(j) = 1$ for $j \in I$ and zero else, while D_x for $x \in \mathbb{R}^K$ and D_I for $I \subseteq [K]$ denote the diagonal matrices with x and respectively $\mathbf{1}_I$ on the diagonal.
Ideally, given an algorithm and the sparse model, meaning the dictionary Φ and the sparsity level n, one can quantify whether the algorithm is successful. For instance if the dictionary has small coherence $\mu(\Phi) = \max_{i\neq j} |\langle \phi_i, \phi_j \rangle|$, ℓ_1-minimisation based schemes, [6, 7], or OMP, [16], can recover the sparse support and coefficients x_I from y as long as $2\mu n = 2\mu|I| < 1$, meaning one can compress y to x_I, [9, 20].
Similarly, if one has the design freedom to choose Φ, the task of reconstructing any n-sparse $x \in \mathbb{R}^K$ from $y = \Phi x \in \mathbb{R}^d$ is known as compressed sensing. There one of the key results is that if one chooses Φ randomly, e.g. with i.i.d Bernoulli entries, then as long as $n \cdot \log(K)/d \lesssim 1$ with high probability all n-sparse $x \in \mathbb{R}^K$ can be reconstructed via

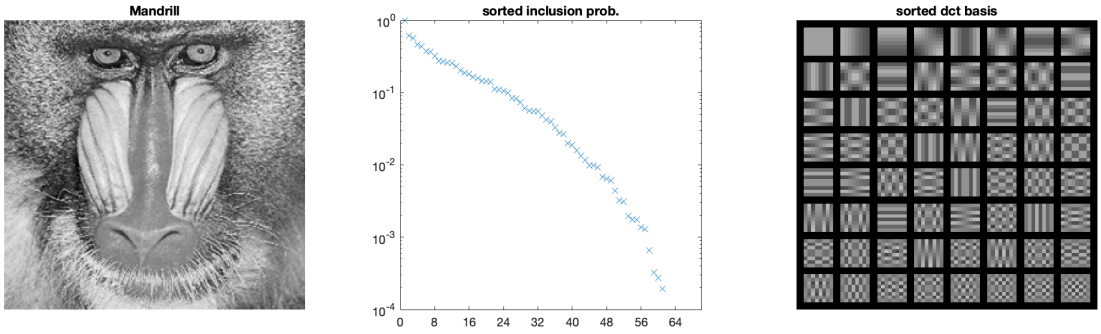

Figure 1: For each 8×8-patch in *Mandrill* (left) we find the best approximation using 8 atoms of the DCT basis (right). We then count how often each atom is used and calculate the empirical inclusion probabilities (middle). The DCT atoms are sorted from most to least frequent row-wise, i.e. most frequent top left, least frequent bottom right.

ℓ_1-minimisation, [4, 5]. Since for $K > d$ we have μ at best of order $1/\sqrt{d}$, choosing a random dictionary, means that the scheme works for much larger n.
However, in many situations the dictionary Φ is predetermined by the application, meaning it cannot be chosen or modelled as random. At the same time all deterministic coherence based results are too weak to be useful or do not reflect the practical performance of an algorithm well.
In these situations one can try to get a better fit between theory and practical performance by adopting a random model on the signal coefficients, meaning the sparse support and the coefficient size, and by studying the average performance. After all, most algorithms are designed to be used for many signals or in the case of machine learning they use as input a batch of *typical* signals.
Going back to our example of sparse approximation, one can show that if the signals are modelled as

$$y = \sum_{k \in S} c(k)\sigma(k)\phi_k = \Phi_S(c_S \odot \sigma_S), \tag{1}$$

where S chooses supports I uniformly at random among all sets of size n, the signs σ are a Rademacher sequence, c is a strictly positive sequence and $\odot$ is the Hadamard (entrywise) product, then with high probability one can recover the support and non-zero coefficients via ℓ_1-minimisation, as long as $n\mu^2 \log(K) \lesssim 1$, [21].
In order to reflect that some atoms usually appear with higher coefficients than others, it is of course easy to add a random model on c, e.g. by letting each entry $c(k)$ follow a distribution bounded away from zero (and bounded) but with different mean. However, it is a little more tricky to incorporate another property that we are likely to encounter in a realistic set of sparse signals, i.e. that some atoms are more likely to be in the n-sparse support that others. For instance, if we look at the best approximation of all 8×8 patches of *Mandrill* with $n = 8$ DCT atoms, and count how often each DCT atom is used, we see in Figure 1 that some atoms clearly appear for frequently than others.

In order to get a still simple but more realistic model, let us forget for a moment that we want only n atoms in the support. In this case we can give each atom a probability to be in the support, meaning we define a series of independent Bernoulli 0-1 variables δ_k which are 1 with probability $p_k \in (0,1)$ and zero with probability $(1-p_k)$. Then for the random vector δ and each $v \in \{0,1\}^K = \Delta$ we have

$$\mathbb{P}(\delta = v) = \prod_{i:v(i)=1} p_i \prod_{j:v(j)=0} (1-p_j) =: p^v \cdot (\mathbf{1}-p)^{1-v}.$$

Now if we identify vectors in Δ with subsets of $[K]$ via $\iota : \Delta \to \mathcal{P}([K]), v \mapsto \{i : v(i) = 1\}$, meaning $\iota^{-1}(I) = \mathbf{1}_I$ we get the set-valued random variable $S = \iota \circ \delta$, which is distributed as

$$\mathbb{P}(S = I) := \prod_{i \in I} p_i \prod_{j \notin I} (1-p_j) =: p^I (\mathbf{1}-p)^{I^c} \quad \text{for} \quad I \subseteq [K].$$

Choosing a support or sample according to the distribution above is called Poisson sampling. Now if our probability weights p_k sum to n, meaning for the vector $p = (p_1, \dots, p_K)$ we have $\|p\|_1 = n$, we know that in expectation the size of the chosen supports is n, since

$$\mathbb{E}(|S|) = \mathbb{E}\Big(\sum_{k=1}^K \mathbf{1}_S(k)\Big) = \mathbb{E}\Big(\sum_{k=1}^K \delta_k\Big) = \sum_{k=1}^K p_k = \|p\|_1 = n,$$

and concentration of measure tells us that most supports that we choose with this method will have support size close to n. Unfortunately, this does not mean that most supports have actual size n. However, we we can force this property by conditioning on it, meaning we define a new distribution of S via

$$\mathbb{P}_n(S = I) := \mathbb{P}(S = I \ \mid \ |S| = n) = \begin{cases} c^{-1} \cdot p^I (\mathbf{1}-p)^{I^c} & \text{if} \quad |I| = n \\ 0 & \text{else} \end{cases},$$

where $c = \mathbb{P}(|S| = n) = \sum_{I:|I|=n} p^I (\mathbf{1}-p)^{I^c}$. Due to its construction, choosing a support or sample according to the distribution above is called conditional Poisson sampling, conditional Bernoulli sampling or rejective sampling, since we reject all samples of the wrong size. Conditional Poisson sampling was popularised and intensively studied by J. Hájek, e.g. [10, 11]. It belongs to a larger class of sampling measures called exponential designs and we refer to [19] for a thorough introduction, [1] for algorithmic aspects and [22, 2] for more recent developments.

Coming back to our goal of modelling n-sparse signals, we can now generalise the model in (1), by allowing the n-sparse supports to be chosen according to rejective sampling with parameter vector p, such that $\|p\|_1 = n$. If we then consider our sparse approximation problem, indeed we get an analog statement, that with high probability one can recover the sparse supports via ℓ_1-minimisation, as long as the hollow Gram matrix of the dictionary, $H = \Phi^t \Phi - \mathbb{I}_K$, weighted by the diagonal matrix Λ with $\Lambda_{ii} = \sqrt{p_i}$, has small operator and column-wise Euklidean norm, that is $\|\Lambda H \Lambda\|_{2,2} \lesssim 1$ and $\|H\Lambda\|_{\infty,2}^2 \log(K) \lesssim 1$, [17].

While rejective sampling clearly allows to model n-sparse signals more accurately, it comes

with certain complications or drawbacks. For instance, the first order inclusion probabilities $\pi_k = \pi_k(n) = \mathbb{P}_n(k \in S)$, that we can observe, differ from the generating Bernoulli probabilities $p_k = \mathbb{P}(k \in S)$, we use in the model, meaning unless $p_k = c$ for all k, we have

$$\pi = \pi(n) = \mathbb{E}_n[\mathbf{1}_S] \neq \mathbb{E}[\mathbf{1}_S] = p.$$

Similarly, the higher order inclusion probabilities are not a product of the first order inclusion probabilities, so for $L \subseteq [K]$ we have

$$\pi_L(n) = \mathbb{P}_n(L \subseteq S) \neq \prod_{\ell \in L} \pi_\ell \quad \text{vs.} \quad \prod_{\ell \in L} p_\ell = \mathbb{P}(L \subseteq S).$$

This also means that the matrix of inclusion probabilities $\Sigma_n = \mathbb{E}_n[\mathbf{1}_S \mathbf{1}_S^t]$ does not have a simple structure like $\Sigma = \mathbb{E}[\mathbf{1}_S \mathbf{1}_S^t] = pp^t + D_p(\mathbb{I} - D_p)$, so weighting a matrix A with Σ_n, is not as straightforward as weighting with Σ, where we simply have

$$A \odot \Sigma = D_p A D_p + A \odot D_p(\mathbb{I} - D_p).$$

However, the results presented in the next section show that many equalities in uniform n-sampling or Poisson sampling that feature n/K or p respectively, have analogue inequalities in rejective sampling, featuring instead of n/K or p either π or even p with a benign constant. As such they provide the key to making rejective sampling an attractive and usable sparse model.

2. Main results

We start with a simple comparison bound between first order inclusion probabilities π and generating probabilities p, which will later allow us to transfer all matrix bounds in terms of π into bounds in terms of p.

Lemma 1 (Comparison bound) *Let $\pi_i(n) = \mathbb{P}_n(i \in S)$ be the first order inclusion probabilities associated to a rejective sampling model with parameter n and weight vector $p = (p_1, \ldots, p_K)$ with $p_i \in (0,1)$ and $\|p\|_1 = n$, then we have*

$$1 - \|p\|_\infty \leq \frac{\pi_i(n)}{p_i} \leq 2.$$

The relation between $\pi_i(n)$ and p_i was first studied by Hájek, [10, 11], who showed that

$$\max_{1 \leq i \leq K} |\pi_i/p_i - 1| \to 0 \quad \text{as} \quad q := \sum_{i=1}^K p_i(1-p_i) \to \infty,$$

and conjectured, as later shown by Milbrodt, [15], that for $\alpha_i := \pi_i(1) = \frac{p_i}{1-p_i}$,

$$\min_i \alpha_i = \min_i \frac{p_i}{1-p_i} \leq \min_i \pi_i \qquad \text{and} \qquad \max_i \pi_i \leq \max_i \frac{p_i}{1-p_i} = \max_i \alpha_i.$$

The result above, which already gives control over the extreme points of the two sequences, was further generalised in [13, 22] to

$$\pi \prec \alpha \quad \text{and} \quad (K^{-1}, \dots, K^{-1}) = \frac{\pi(K)}{K} \prec \cdots \prec \frac{\pi(n)}{n} \prec \cdots \prec \pi(1) = \alpha, \tag{2}$$

where $u \prec v$ for two vectors $u, v \in \mathbb{R}^K$ means, that after sorting them in increasing order, their sorted versions $\vec{u}, \vec{v}$ satisfy

$$\sum_{k=j}^{K} \vec{u}_k \leq \sum_{k=j}^{K} \vec{v}_k, \qquad \forall j = 1, 2, \dots, K.$$

The advantage of the bound here is that, while maybe less sharp, it holds in the non-asymptotic setting, where $q \leq S \ll K$, and that it is valid for all pairs (p_i, π_i), rather than just the extremal ones or partial (sorted) sums.
While the comparison bound above has its asymptotic analogue, bounds in the operator norm or semi-definite order for matrices of inclusion probabilities, such as Σ_n with $\pi_i(n)$ and $\pi_{ij}(n)$ as diagonal and off-diagonal entries, seem to be a unstudied so far. However, Σ_n appears quite naturally in a sparse context.
For instance, the idea behind many dictionary learning algorithms is to construct glimpses of the dictionary Φ using a current guess Ψ. Assuming a sparse signal model as in (1) with constant coefficient sequence c these glimpses can (in first order approximation) be of the form $\Phi D_S + \Phi D_S Z^t \Psi D_S$ for $Z = \Phi - \Psi$, so that aggregating them one arrives at

$$\hat{\Psi} := \Phi \, \mathbb{E}[D_S + D_S Z^t \Psi D_S] = \Phi[D_\pi + (Z^t \Psi) \odot \Sigma_n]. \tag{3}$$

Determining whether a normalised version of $\hat{\Psi}$ is a better approximation to Φ than Ψ, then becomes a question of bounding $(Z^t \Psi) \odot \Sigma_n$ or rather bounding $H \odot \Sigma_n$ where $H = Z^t \Psi - (Z^t \Psi) \odot \mathbb{I}$.

Theorem 2 (Operator norm & semi-definite order bounds) *Let $\mathbb{E}_n$ be the expectation according to the rejective sampling probability with parameter n and weights $p_i \in (0, 1)$. Further let $\pi \in \mathbb{R}^K$ be the vector of first order inclusion probabilities, meaning $\pi = \mathbb{E}_n[\mathbf{1}_S]$, D_π be the $K \times K$ matrix with π on the diagonal and zero else, and Σ_n be the matrix of (up to) second order inclusion probabilities, meaning $\Sigma_n = \mathbb{E}_n[\mathbf{1}_S \mathbf{1}_S^t]$. Then for the Hadamard (entrywise) product with any matrix $A \in \mathbb{R}^{K \times K}$ we have*

$$\|A \odot \Sigma_n\| \leq \frac{1 + \|\pi\|_\infty}{(1 - \|\pi\|_\infty)^2} \cdot \|D_\pi[A - A \odot \mathbb{I}]D_\pi\| + \|A \odot D_\pi\|. \tag{a}$$

Writing $A \preceq B$ for two positive semi-definite (p.s.d) matrices if $A - B$ is p.s.d, we have

$$\Sigma_n \preceq \pi\pi^t + 2D_\pi, \tag{b}$$

and, defining for $L \subseteq [K]$ with $|L| < n$ the set $\mathcal{L} := \{I \subseteq [K] : L \subseteq I\}$, we have

$$\mathbb{E}_n\big[\mathbf{1}_{S \setminus L} \mathbf{1}_{S \setminus L}^t \cdot \mathbb{1}_{\mathcal{L}}(S)\big] \preceq \Sigma_{n-|L|} \cdot \prod_{\ell \in L} \frac{\pi_\ell}{1 - \pi_\ell}. \tag{c}$$

Interestingly the theorem above allows us to bound quantities involving second order inclusion probabilities using only first order inclusion probabilities or in combination with Lemma 1 even generating probabilities, which are much easier to handle. For instance, we can answer the question how close $\hat{\Psi}$ is to a scaled version of Φ.

Example 1 (Toy application in dictionary learning) *Recall that we have a dictionary Φ, a guess $\Psi = \Phi + Z$ and are wondering when the updated guess $\hat{\Psi} = \Phi[D_\pi + (Z^t\Psi) \odot \Sigma_n]$ will be an improvement over Ψ.*
We set $\varepsilon_i = \langle z_i, \psi_i \rangle$ so that $D_\varepsilon = (Z^t\Psi) \odot \mathbb{I}$ and $H = Z^t\Psi - D_\varepsilon$. We can then rewrite $\hat{\Psi} = \Phi[D_\pi(\mathbb{I} + D_\varepsilon) + H \odot \Sigma_n]$ and get abbreviating $C_\pi = (1 + \|\pi\|_\infty)(1 - \|\pi\|_\infty)^{-2}$ and $D_{\sqrt{\pi}} = D_\pi^{1/2}$ that

$$\begin{aligned}
\|\hat{\Psi}(\mathbb{I} + D_\varepsilon)^{-1}D_{\sqrt{\pi}}^{-1} - \Phi D_{\sqrt{\pi}}\| &= \|\Phi D_{\sqrt{\pi}} D_{\sqrt{\pi}}^{-1} (H \odot \Sigma_n) D_{\sqrt{\pi}}^{-1}(\mathbb{I} + D_\varepsilon)^{-1}\| \\
&\leq \|\Phi D_{\sqrt{\pi}}\| \cdot \|(D_{\sqrt{\pi}}^{-1} H D_{\sqrt{\pi}}^{-1}) \odot \Sigma_n\| \cdot \|(\mathbb{I} + D_\varepsilon)^{-1}\| \\
&\leq \|\Phi D_{\sqrt{\pi}}\| \cdot C_\pi \cdot \|D_{\sqrt{\pi}} H D_{\sqrt{\pi}}\| \cdot \|(\mathbb{I} + D_\varepsilon)^{-1}\| \\
&\leq C_\pi \cdot \|\Phi D_{\sqrt{\pi}}\| \cdot \|(\mathbb{I} + D_\varepsilon)^{-1}\| \cdot (\|Z D_{\sqrt{\pi}}\| \cdot \|\Psi D_{\sqrt{\pi}}\| + \|D_\varepsilon D_\pi\|) \\
&\leq C_\pi \cdot \|\Phi D_{\sqrt{\pi}}\| \cdot \|(\mathbb{I} + D_\varepsilon)^{-1}\| \cdot (\|\Psi D_{\sqrt{\pi}}\| + \|D_{\sqrt{\pi}}\|) \cdot \|Z D_{\sqrt{\pi}}\|.
\end{aligned}$$

Since $\|Z D_{\sqrt{\pi}}\| = \|(\Phi - \Psi) D_{\sqrt{\pi}}\|$ is the weighted distance between the current estimate Ψ and the desired dictionary Φ, it is now easy to infer from the bound above under which conditions on Φ and Ψ the normalised version of $\hat{\Psi}$ or equivalently of $\hat{\Psi}(\mathbb{I} + D_\varepsilon)^{-1}D_{\sqrt{\pi}}^{-1}$ will be a better approximation to Φ in this weighted distance.

Moreover, using Theorem 2, it is also possible to estimate terms involving higher order inclusion probabilities of the form $\mathbb{E}_n[D_S H D_S H^t D_S]$, where H is again a matrix with zero diagonal. In dictionary learning such terms appear considering glimpses in second order approximation, and we can bound them using the following corollary of Theorem 2.

Corollary 3 *Under the same conditions as in Theorem 2 we have for any matrix H with zero diagonal that*

$$\left\|\mathbb{E}_n[D_S H D_S H^t D_S]\right\| \leq (1 + \|\pi\|_\infty) \cdot \frac{\|D_\pi H D_\pi H^t D_\pi\|}{(1 - \|\pi\|_\infty)^3} + \frac{\|D_\pi \odot (H D_\pi H^t)\|}{1 - \|\pi\|_\infty}.$$

Note that $D_\pi \odot (H D_\pi H^t)$ is a diagonal matrix, so we can also write $\|D_\pi \odot (H D_\pi H^t)\| = \max_k \pi_k (H D_\pi H^t)_{k,k}$. The interesting fact about the bound above is, that it allows to bound the matrix $\mathbb{E}_n[D_S H D_S H^t D_S]$, which involves up to of third order inclusion probabilities, again using only the first order inclusion probabilities π.
For the interested reader we provide a short explanation how the first and second order glimpses in dictionary learning are constructed and how they can be bounded using Corollary 3 in Appendix A. For even more details and a real application we refer to [18].
Here we next we turn to the proofs of both results.

3. Proof of main results

We start with the proof of Lemma 1.

Proof [of Lemma 1] We want to show that $1 - \|p\|_\infty \leq \pi_i(n)/p_i \leq 2$ whenever $\|p\| = n$. The upper bound follows directly from [17, Lemma 7] so we only need to show the lower bound. We abbreviate $\pi_i = \pi_i(n)$ and $c := 1 - \|p\|_\infty \leq \pi_i/p_i$ as well as $\mathbb{P}(I) = \mathbb{P}(S = I)$ for any set $I \subseteq [K]$. By definition, we have

$$\pi_i = \mathbb{P}(i \in S \mid |S| = n) = \frac{\mathbb{P}(\{i \in S\} \cap \{|S| = n\})}{\mathbb{P}(|S| = n)} = \frac{\sum_{I:|I|=n, i\in I} \mathbb{P}(I)}{\sum_{I:|I|=n} \mathbb{P}(I)} \underbrace{\sum_J \mathbb{P}(J)}_{=1},$$

and $p_i = \sum_{J:i\in J} \mathbb{P}(J)$. So the desired inequality $c \cdot p_i \leq \pi_i$ is equivalent to

$$c \sum_{I:|I|=n} \mathbb{P}(I) \sum_{J:i\in J} \mathbb{P}(J) \leq \sum_{I:|I|=n, i\in I} \mathbb{P}(I) \sum_J \mathbb{P}(J).$$

Splitting the sum over I, J into sums over those containing i and those not containing i we see that the inequality above is implied by

$$c \sum_{I:|I|=n, i\notin I} \mathbb{P}(I) \sum_{J:i\in J} \mathbb{P}(J) \leq \sum_{I:|I|=n, i\in I} \mathbb{P}(I) \sum_{J:i\notin J} \mathbb{P}(J). \tag{4}$$

Note that for any set I not containing the index i we have

$$\frac{p_i}{1-p_i} \cdot \mathbb{P}(I) = \frac{p_i}{1-p_i} \prod_{k\in I} p_k \prod_{k\notin I} (1-p_k) = \prod_{k\in I\cup\{i\}} p_k \prod_{k\notin I\cup\{i\}} (1-p_k) = \mathbb{P}(I \cup \{i\}).$$

Multiplying both sides in (4) with $p_i/(1-p_i)$ we get

$$c \sum_{I:|I|=n+1, i\in I} \mathbb{P}(I) \sum_{J:i\in J} \mathbb{P}(J) \leq \sum_{I:|I|=n, i\in I} \mathbb{P}(I) \sum_{J:i\in J} \mathbb{P}(J),$$

so it suffices to show that

$$c \sum_{I:|I|=n+1, i\in I} \mathbb{P}(I) \leq \sum_{I:|I|=n, i\in I} \mathbb{P}(I).$$

Indeed, we have

$$\begin{aligned}
c \sum_{I:|I|=n+1, i\in I} \mathbb{P}(I) &= c \sum_{I:|I|=n+1, i\in I} \mathbb{P}(I) \sum_{k:k\in I, k\neq i} \frac{1}{n} \\
&= c \sum_{I:|I|=n+1, i\in I} \frac{1}{n} \sum_{k:k\in I, k\neq i} \mathbb{P}(I \setminus \{k\}) \frac{p_k}{1-p_k} \\
&\leq \frac{c}{n(1-\|p\|_\infty)} \sum_{\substack{(I,k):|I|=n+1, i\in I \\ k\in I, k\neq i}} \mathbb{P}(I \setminus \{k\}) \cdot p_k \\
&= \frac{1}{n} \sum_{J:|J|=n, i\in J} \mathbb{P}(J) \sum_{k\notin J} p_k \leq \sum_{J:|J|=n, i\in J} \mathbb{P}(J),
\end{aligned}$$

where we used that $\sum_{k\notin J} p_k \leq \sum_k p_k = n$. ∎

In order to prove Theorem 2 we need two auxiliary results. The first is a corollary of Schur's product theorem, see e.g. [12]. For convenience we include its short proof.

Corollary 4 (of Schur's theorem) *Let $A, B \in \mathbb{R}^{K\times K}$ be two square matrices of the same dimension. If A is positive-semidefinite (p.s.d.), then*

$$\|A \odot B\| \leq \|A \odot \mathbb{I}\| \cdot \|B\|.$$

Proof The matrix

$$\begin{pmatrix} \|B\| \cdot (A \odot \mathbb{I}) & A \odot B \\ (A \odot B)^t & \|B\| \cdot (A \odot \mathbb{I}) \end{pmatrix} = \begin{pmatrix} A & A \\ A & A \end{pmatrix} \odot \begin{pmatrix} \|B\| \cdot \mathbb{I} & B \\ B^t & \|B\| \cdot \mathbb{I} \end{pmatrix}$$

is p.s.d., since the right hand side of the equation is a Hadamard product of two p.s.d. matrices and thus by Schur's product theorem also p.s.d. Further by [12, Theorem 7.7.9] this means that there exists a contraction C, meaning $\|C\| \leq 1$, such that

$$A \odot B = \|B\| \cdot (A \odot \mathbb{I})^{1/2}\, C\, (A \odot \mathbb{I})^{1/2},$$

and hence $\|A \odot B\| \leq \|A \odot \mathbb{I}\| \cdot \|B\|$. ■

The second result is the following lemma, which collects relations between first and second order inclusion probabilities for n resp. $n-1$, corresponding to [19, Eq. 5.21], and bounds between first order inclusion probabilities for n resp. $n-1$.

Lemma 5 *Let $\pi_i(n) = \mathbb{P}_n(i \in I)$ be the first order inclusion probabilities associated to a rejective sampling model with parameter n and weight vector $p = (p_1, \ldots, p_K)$ with $p_i \in (0,1)$, then we have*

$$\pi_{ij}(n) = \pi_i(n) \cdot \frac{\pi_j(n-1) - \pi_{ij}(n-1)}{1 - \pi_i(n-1)} \quad \textit{for all } i \neq j \tag{5}$$

$$\textit{and} \quad \pi_i(n-1) \leq \pi_i(n) \quad \textit{for all } i. \tag{6}$$

A proof of Eq. (5) can be found in [19]. The inequality in (6) follows from [3], using that $\mu = \mathbb{P}$ is strongly Raleigh. Therefore $\mu_{n-1} = \mathbb{P}_{n-1} \preceq \mathbb{P}_n = \mu_n$, which further implies that for increasing functions f, such as $f(S) = \mathbb{1}_S(k)$, $\pi_i(n-1) = \int f \mathrm{d}\mu_{n-1} \leq \int f \mathrm{d}\mu_n = \pi_i(n)$. For convenience, alternative proofs, which use only the language and techniques introduced in this paper, are provided in Appendix B. With the two results above we are ready to prove Theorem 2.

Proof [of Theorem 2] **(a)** We want to show that

$$\|A \odot \Sigma_n\| \leq \frac{1 + \|\pi\|_\infty}{(1 - \|\pi\|_\infty)^2} \cdot \|D_\pi[A - A \odot \mathbb{I}]D_\pi\| + \|D_\pi \odot A\|.$$

First note that since Σ_n has $\pi(n)$ on the diagonal, a simple application of the triangle inequality yields

$$\|A \odot \Sigma_n\| \leq \|A \odot \Sigma_n - A \odot D_{\pi(n)}\| + \|A \odot D_{\pi(n)}\| = \|(A - A \odot \mathbb{I}) \odot \Sigma_n\| + \|A \odot D_{\pi(n)}\|,$$

which already proves the theorem for $n = 1$, where all off-diagonal entries of Σ_n are zero, meaning the first norm term is zero. In case $n \geq 2$ it remains to show that for $H = A - A \odot \mathbb{I}$

we have $\|H \odot \Sigma_n\| \leq c \cdot \|D_{\pi(n)} H D_{\pi(n)}\|$ with constant c as above. Using the abbreviation $\Delta_{\pi(n)} = \mathbb{I} - D_{\pi(n)}$ we know from (5) that

$$\begin{aligned}(H \odot \Sigma_n)_{ij} &= H_{ij} \cdot \pi_{ij}(n) = \frac{\pi_i(n)}{1 - \pi_i(n-1)} \cdot H_{ij} \cdot [\pi_j(n-1) - \pi_{ij}(n-1)] \\ &= \left(\Delta^{-1}_{\pi(n-1)} D_{\pi(n)} H D_{\pi(n-1)} - \Delta^{-1}_{\pi(n-1)} D_{\pi(n)} H \odot \Sigma_{n-1}\right)_{ij}.\end{aligned}$$

Next Lemma 5 tells us that $D_{\pi(n-1)} \preceq D_{\pi(n)}$, which leads to

$$\begin{aligned}\|H \odot \Sigma_n\| &\leq \|\Delta^{-1}_{\pi(n-1)}\| \cdot \|D_{\pi(n)} H D_{\pi(n-1)} - D_{\pi(n)} H \odot \Sigma_{n-1}\| \\ &\leq (1 - \|\pi(n)\|_\infty)^{-1} \cdot \left(\|D_{\pi(n)} H D_{\pi(n)}\| + \|D_{\pi(n)} H \odot \Sigma_{n-1}\|\right). \end{aligned} \tag{7}$$

Applying the inequality above to H^t and using the symmetry of Σ_n we also get

$$\|H \odot \Sigma_n\| \leq (1 - \|\pi(n)\|_\infty)^{-1} \cdot \left(\|D_{\pi(n)} H D_{\pi(n)}\| + \|H D_{\pi(n)} \odot \Sigma_{n-1}\|\right). \tag{8}$$

For $n = 2$, the matrix Σ_{n-1} is again a diagonal matrix, meaning the second norm term vanishes and we are done. For $n > 2$ we simply apply the inequality in (8) to $\bar{H} \odot \Sigma_{n-1}$ with $\bar{H} = D_{\pi(n)} H$, leading to

$$\begin{aligned}\|D_{\pi(n)} H \odot \Sigma_{n-1}\| &= \|\bar{H} \odot \Sigma_{n-1}\| \\ &\leq (1 - \|\pi(n-1)\|_\infty)^{-1} \cdot \left(\|D_{\pi(n-1)} \bar{H} D_{\pi(n-1)}\| + \|\bar{H} D_{\pi(n-1)} \odot \Sigma_{n-2}\|\right) \\ &\leq (1 - \|\pi(n)\|_\infty)^{-1} \cdot \left(\|\pi(n)\|_\infty \|D_{\pi(n)} H D_{\pi(n)}\| + \|D_{\pi(n)} H D_{\pi(n)} \odot \Sigma_{n-2}\|\right).\end{aligned}$$

Substituting the inequality above into (7) yields

$$\|H \odot \Sigma_n\| \leq (1 - \|\pi(n)\|_\infty)^{-2} \cdot \left(\|D_{\pi(n)} H D_{\pi(n)}\| + \|D_{\pi(n)} H D_{\pi(n)} \odot \Sigma_{n-2}\|\right)$$

and since Σ_{n-2} has $\pi(n-2)$ on its diagonal the result follows from Corollary 4 (with $A = \Sigma_{n-2}$) and Lemma 5. **(a)✓**

(b) We want to show that $\mathbb{E}_n[\mathbf{1}_S \mathbf{1}_S^t] = \Sigma_n \preceq \pi\pi^t + 2D_\pi$, meaning

$$\frac{\sum_{I:|I|=n} \mathbf{1}_I \mathbf{1}_I^t \mathbb{P}(I)}{\sum_{I:|I|=n} \mathbb{P}(I)} \preceq \frac{\sum_{(I,J):|I|=|J|=n} \mathbf{1}_I \mathbf{1}_J^t \mathbb{P}(I)\mathbb{P}(J)}{(\sum_{I:|I|=n} \mathbb{P}(I))^2} + \frac{2 \cdot \sum_{I:|I|=n} D_I \mathbb{P}(I)}{\sum_{I:|I|=n} \mathbb{P}(I)}.$$

Multiplying both sides by $(\sum_{I:|I|=n} \mathbb{P}(I))^2$ we therefore have to show that

$$\sum_{(I,J):|I|=|J|=n} \mathbf{1}_I \mathbf{1}_I^t \mathbb{P}(I)\mathbb{P}(J) \preceq \sum_{(I,J):|I|=|J|=n} \mathbf{1}_I \mathbf{1}_J^t \mathbb{P}(I)\mathbb{P}(J) + 2 \sum_{(I,J):|I|=|J|=n} D_I \mathbb{P}(I)\mathbb{P}(J).$$

Now the crucial step is to partition these sums in a special way. For a pair (I, J), by definition of the Poisson sampling model, we can rewrite $\mathbb{P}(I)\mathbb{P}(J)$ as

$$\mathbb{P}(I)\mathbb{P}(J) = \prod_{i \in I} p_i \prod_{j \notin I} (1 - p_j) \prod_{i \in J} p_i \prod_{j \notin J} (1 - p_j) = \prod_{i \in I \cap J} p_i^2 \prod_{i \in I \triangle J} p_i(1 - p_i) \prod_{j \notin I \cup J} (1 - p_j)^2,$$

where $I \triangle J$ denotes the symmetric difference of I, J. This implies that for two pairs (I, J), (I', J') we have

$$\mathbb{P}(I)\mathbb{P}(J) = \mathbb{P}(I')\mathbb{P}(J') \quad \text{whenever} \quad I \cap J = I' \cap J' \quad \text{and} \quad I \triangle J = I' \triangle J'.$$

This allows us to define natural partitions on the set of pairs (I, J) such that the probability $\mathbb{P}(I)\mathbb{P}(J)$ is constant on each partition. Concretely, for any integer $m \in \{1, \dots, n\}$, together with a set $A \subseteq [K]$ with $|A| = n - m$ and a set $B \subseteq [K] \setminus A$ with $|B| = 2m$, we look at the collection of pairs (I, J) with intersection A and symmetric difference B, i.e.,

$$\mathcal{Q}_{A,B} := \{(I, J) : I, J \subseteq [K], |I| = |J| = n, I \cap J = A, I \triangle J = B\}.$$

Note that each pair (I, J) with $|I| = |J| = n$ can be *uniquely* assigned to a collection $\mathcal{Q}_{A,B}$ and that $\mathbb{P}(I)\mathbb{P}(J)$ is constant for all $(I, J) \in \mathcal{Q}_{A,B}$. Using that the sum of positive semi-definite matrices is positive semi-definite, it therefore suffices to show that for all possible choices of A, B we have

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbf{1}_I \mathbf{1}_I^t \preceq \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbf{1}_I \mathbf{1}_J^t + 2 \sum_{(I,J)\in\mathcal{Q}_{A,B}} D_I.$$

For A, B fixed we abbreviate $Q = \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbf{1}_I \mathbf{1}_I^t$ and $\bar{Q} = \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbf{1}_I \mathbf{1}_J^t$. Note that $\sum_{(I,J)\in\mathcal{Q}_{A,B}} D_I = Q \odot \mathbb{I}$, so the inequality above is equivalent to showing that

$$0 \preceq \bar{Q} - Q + 2 \cdot Q \odot \mathbb{I}. \tag{9}$$

For the entries of these matrices we have

$$Q_{ij} = |\{(I, J) \in \mathcal{Q}_{A,B} : i, j \in I\}| \quad \text{resp.} \quad \bar{Q}_{ij} = |\{(I, J) \in \mathcal{Q}_{A,B} : i \in I, j \in J\}|.$$

In case $i, j \in A$ we obviously have $Q_{ij} = \bar{Q}_{ij} = |\mathcal{Q}_{A,B}|$, so

$$Q \odot (\mathbf{1}_A \mathbf{1}_A^t) = \bar{Q} \odot (\mathbf{1}_A \mathbf{1}_A^t).$$

In particular, this means that (9) holds trivially for $m = 0$, where $B = \emptyset$.
In case that $i \in A, j \in B$ we have $Q_{ij} = \bar{Q}_{ij} = \binom{2m-1}{m-1} =: d_m$ and therefore

$$Q \odot (\mathbf{1}_A \mathbf{1}_B^t + \mathbf{1}_B \mathbf{1}_A^t) = \bar{Q} \odot (\mathbf{1}_A \mathbf{1}_B^t + \mathbf{1}_B \mathbf{1}_A^t).$$

It only remains to check what happens for $i, j \in B$. The case $m = 0$ is already settled, thus we assume $m \geq 1$. On the diagonal we get $\bar{Q}_{ii} = 0$ while $Q_{ii} = \binom{2m-1}{m-1} = d_m$. We have $Q_{ij} = \binom{2m-2}{m-2} =: q_m$ and $\bar{Q}_{ij} = \binom{2m-2}{m-1} =: \bar{q}_m$. In summary

$$\bar{Q} - Q + 2 \cdot Q \odot \mathbb{I} = (\bar{q}_m - q_m) \cdot \mathbf{1}_B \mathbf{1}_B^t - (\bar{q}_m - q_m + d_m) \cdot D_B + 2 \cdot Q \odot \mathbb{I}. \tag{10}$$

Since $\bar{q}_m \geq q_m$ the matrix $(\bar{q}_m - q_m) \cdot \mathbf{1}_B \mathbf{1}_B^t$ is positive semi-definite. Finally, as $\binom{2m-2}{m-2} + \binom{2m-2}{m-1} = \binom{2m-1}{m-1}$ we have $\bar{q}_m + q_m = d_m$. Thus for all $m \geq 1$, we have

$$(\bar{q}_m - q_m + d_m) \cdot D_B = 2 \cdot \bar{q}_m \cdot D_B \preceq 2 \cdot d_m \cdot D_B = 2 \cdot Q \odot D_B \preceq 2 \cdot Q \odot \mathbb{I},$$

showing that also the remaining terms in (10) are positive semi-definite, which completes the proof of (b). **(b)✓**

(c) We will prove the statement by induction. Let $\hat{L}$ be a set of size $m \leq n-2$ and $\hat{\mathcal{L}} = \{I \subseteq [K] : \hat{L} \subseteq I\}$. We first show that for $k \notin \hat{L}$ and $L = \hat{L} \cup \{k\}$ we have

$$(1-\pi_k(n)) \cdot \mathbb{E}_n\big[\mathbf{1}_{S\setminus L}\mathbf{1}_{S\setminus L}^t \cdot \mathbb{1}_{\mathcal{L}}(S)\big] \preceq \pi_k(n) \cdot \mathbb{E}_{n-1}\big[\mathbf{1}_{S\setminus \hat{L}}\mathbf{1}_{S\setminus \hat{L}}^t \cdot \mathbb{1}_{\hat{\mathcal{L}}}(S)\big]. \tag{11}$$

We again use that for any set J not containing the index k we have

$$\frac{p_k}{1-p_k} \cdot \mathbb{P}(J) = \mathbb{P}(J \cup \{k\}).$$

Thus expanding the expectation we get

$$\begin{aligned}
\mathbb{E}_n\big[\mathbf{1}_{S\setminus L}\mathbf{1}_{S\setminus L}^t \cdot \mathbb{1}_{\mathcal{L}}(S)\big] &= \frac{\sum_{I:|I|=n,L\subseteq I}\mathbb{P}(I)(\mathbf{1}_{I\setminus L}\mathbf{1}_{I\setminus L}^t)}{\sum_{I:|I|=n}\mathbb{P}(I)} \cdot \frac{\sum_{I:|I|=n,k\in I}\mathbb{P}(I)}{\sum_{I:|I|=n,k\in I}\mathbb{P}(I)} \\
&= \frac{\sum_{I:|I|=n,L\subseteq I}\mathbb{P}(I)(\mathbf{1}_{I\setminus L}\mathbf{1}_{I\setminus L}^t)}{\sum_{I:|I|=n,k\in I}\mathbb{P}(I)} \cdot \frac{\sum_{I:|I|=n,k\in I}\mathbb{P}(I)}{\sum_{I:|I|=n}\mathbb{P}(I)} \\
&= \frac{\sum_{J:|J|=n-1,k\notin J,\hat{L}\subseteq J}\mathbb{P}(J)(\mathbf{1}_{J\setminus \hat{L}}\mathbf{1}_{J\setminus \hat{L}}^t)}{\sum_{J:|J|=n-1,k\notin J}\mathbb{P}(J)} \cdot \pi_k(n) \\
&\preceq \frac{\sum_{J:|J|=n-1,\hat{L}\subseteq J}\mathbb{P}(J)(\mathbf{1}_{J\setminus \hat{L}}\mathbf{1}_{J\setminus \hat{L}}^t)}{\sum_{J:|J|=n-1,k\notin J}\mathbb{P}(J)} \cdot \frac{\sum_{I:|I|=n-1}\mathbb{P}(I)}{\sum_{I:|I|=n-1}\mathbb{P}(I)} \cdot \pi_k(n) \\
&= \frac{\sum_{J:|J|=n-1,\hat{L}\subseteq J}\mathbb{P}(J)(\mathbf{1}_{J\setminus \hat{L}}\mathbf{1}_{J\setminus \hat{L}}^t)}{\sum_{I:|I|=n-1}\mathbb{P}(I)} \cdot \frac{\sum_{I:|I|=n-1}\mathbb{P}(I)}{\sum_{J:|J|=n-1,k\notin J}\mathbb{P}(J)} \cdot \pi_k(n) \\
&= \mathbb{E}_{n-1}\big[\mathbf{1}_{S\setminus \hat{L}}\mathbf{1}_{S\setminus \hat{L}}^t \cdot \mathbb{1}_{\hat{\mathcal{L}}}(S)\big] \cdot \frac{\sum_{I:|I|=n-1}\mathbb{P}(I)}{\sum_{J:|J|=n-1,k\notin J}\mathbb{P}(J)} \cdot \pi_k(n).
\end{aligned}$$

Now all that remains to do in order to prove (11) is to bound the fraction above. Writing out the expression in the denominator we get

$$\begin{aligned}
\frac{\sum_{I:|I|=n-1}\mathbb{P}(I)}{\sum_{J:|J|=n-1,k\notin J}\mathbb{P}(J)} &= \frac{\sum_{I:|I|=n-1}\mathbb{P}(I)}{\sum_{I:|I|=n-1}\mathbb{P}(I) - \sum_{I:|I|=n-1,k\in I}\mathbb{P}(I)} \\
&= \frac{1}{1-\mathbb{P}_{n-1}(k\in S)} \leq \frac{1}{1-\mathbb{P}_n(k\in S)} = \frac{1}{1-\pi_k(n)}.
\end{aligned}$$

By induction and using again that $\pi_k(n-1) \leq \pi_k(n)$, we finally get

$$\mathbb{E}_n\big[\mathbf{1}_{S\setminus L}\mathbf{1}_{S\setminus L}^t \cdot \mathbb{1}_{\mathcal{L}}(S)\big] \prod_{\ell\in L}(1-\pi_\ell(n)) \preceq \mathbb{E}_{n-|L|}[\mathbf{1}_S\mathbf{1}_S^t] \cdot \prod_{\ell\in L}\pi_\ell(n),$$

which completes the proof of (c). **(c)✓**

■

Finally, it only remains to prove Corollary 3.

Proof [of Corollary 3] We want to show that for a matrix H with zero diagonal we have

$$\|\mathbb{E}[D_S H D_S H^t D_S]\| \leq \frac{1+\|\pi\|_\infty}{(1-\|\pi\|_\infty)^3}\|D_\pi H D_\pi H^t D_\pi\| + \frac{1}{1-\|\pi\|_\infty}\|D_\pi \odot (H D_\pi H^t)\|.$$

Let H_k be the k-th column of H and $H_k^t = (H_k)^t$. We rewrite the expression, whose expectation we want to estimate, as

$$D_S H D_S H^t D_S = (H D_S H^t) \odot (\mathbf{1}_S \mathbf{1}_S^t) = \big(\sum_{k\in S} H_k H_k^t\big) \odot (\mathbf{1}_S \mathbf{1}_S^t) = \sum_{k\in S} (H_k H_k^t) \odot (\mathbf{1}_S \mathbf{1}_S^t).$$

Since the k-th entry of H_k and therefore both the k-th row and k-th column of $H_k H_k^t$ are zero, we have $(H_k H_k^t) \odot (\mathbf{1}_S \mathbf{1}_S^t) = (H_k H_k^t) \odot (\mathbf{1}_{S\setminus\{k\}} \mathbf{1}_{S\setminus\{k\}}^t)$, yielding

$$\begin{aligned} D_S H D_S H^t D_S &= \sum_{k\in S} (H_k H_k^t) \odot (\mathbf{1}_{S\setminus\{k\}} \mathbf{1}_{S\setminus\{k\}}^t) \\ &= \sum_k (\mathbf{1}_S(k) \cdot H_k H_k^t) \odot (\mathbf{1}_{S\setminus\{k\}} \mathbf{1}_{S\setminus\{k\}}^t) \\ &= \sum_k (H_k H_k^t) \odot (\mathbf{1}_{S\setminus\{k\}} \mathbf{1}_{S\setminus\{k\}}^t \cdot \mathbf{1}_S(k)). \end{aligned}$$

Note that $\mathbf{1}_S(k) = \mathbb{1}_{\mathcal{L}}(S)$ for $L = \{k\}$ and $\mathcal{L} = \{I \subseteq [K] : L \subseteq I\}$. Thus using the Schur Product Theorem, which says that the Hadamard product of positive semi-definite (p.s.d) matrices is p.s.d., meaning for $A, P, \bar{P}$ p.s.d. with $P \preceq \bar{P}$ we have $A \odot P \preceq A \odot \bar{P}$, together with Theorem 2(c) further leads to

$$\begin{aligned} \mathbb{E}_n\left[D_S H D_S H^t D_S\right] &= \sum_k (H_k H_k^t) \odot \mathbb{E}_n\big[\mathbf{1}_{S\setminus\{k\}} \mathbf{1}_{S\setminus\{k\}}^t \cdot \mathbf{1}_S(k)\big] \\ &\preceq \sum_k (H_k H_k^t) \odot \big(\Sigma_{n-1} \cdot \tfrac{\pi_k}{1-\pi_k}\big) \\ &\preceq \underbrace{(1-\|\pi\|_\infty)^{-1}}_{=:c_\pi} \big(\sum_k H_k \cdot \pi_k \cdot H_k^t\big) \odot \Sigma_{n-1} = c_\pi \cdot (H D_\pi H^t) \odot \Sigma_{n-1}. \end{aligned}$$

This means that for the operator norm of the expression above we have the bound

$$\|\mathbb{E}_n[D_S H D_S H^t D_S]\| \leq c_\pi \cdot \|(H D_\pi H^t) \odot \Sigma_{n-1}]\|,$$

and all that remains to be done is to apply Theorem 2(a) to $A = H D_\pi H^t$ in the bound above. So defining $C_{\pi(n)} := (1+\|\pi(S)\|_\infty)(1-\|\pi(S)\|_\infty)^{-2}$ and observing that due to Lemma 5 we have $\pi_k(n-1) \leq \pi_k(n) = \pi_k$ and thus $C_{\pi(n-1)} \leq C_{\pi(n)} = C_\pi$ we get

$$\begin{aligned} \|\mathbb{E}_n[D_S H D_S H^t D_S]\| &\leq c_\pi \cdot \|A \odot \Sigma_{n-1}]\| \\ &\leq c_\pi \cdot \big(C_{\pi(n-1)} \cdot \|D_{\pi(n-1)}[A - A \odot \mathbb{I}] D_{\pi(n-1)}\| + \|A \odot D_{\pi(n-1)}\|\big) \\ &\leq c_\pi \cdot C_{\pi(n)} \cdot \|D_{\pi(n)} A D_{\pi(n)}\| + c_\pi \cdot \|A \odot D_{\pi(n)}\| \\ &= c_\pi \cdot C_\pi \cdot \|D_\pi H D_\pi H^t D_\pi\| + c_\pi \cdot \|D_\pi \odot (H D_\pi H^t)\|, \end{aligned}$$

where for the last inequality we have also used that $A = H D_\pi H^t$ is positive semidefinite. ■

Appendix A. Motivation and derivation of Relation (3)

For the interested reader we give a short motivation how one constructs the glimpses in dictionary learning, which in first order approximation lead to Relation (3). We then derive a similar relation using second order approximations and show how Corollary 3 comes into play.

The goal of dictionary learning is to identify the dictionary Φ from a set of sparse signals, following a model $y = \Phi_S x_S = \Phi D_S x$. The main idea behind most iterative dictionary learning algorithms is to take a guess Ψ of the same size as Φ, and from each signal realisation $y = \Phi_I x_I$ try to identify I and subsequently x_I, using some sparse approximation scheme. Combinations of y and x_I are then aggregated in a way suitable to approximate Φ. For instance if S follows a rejective sampling model and for simplicity $x = \sigma$, a Rademacher sequence, we have $\mathbb{E}[yx^t] = \mathbb{E}[\Phi D_S x x^t] = \mathbb{E}[\Phi D_S \mathbb{I}] = \Phi D_\pi$, so after renormalisation we can retrieve Φ from $\mathbb{E}[yx^t]$ or rather its approximation by finite averages. Of course, if we only have Ψ instead of Φ we will not get x but only an estimate $\hat{x}$ and the question is if using $\mathbb{E}[y\hat{x}^t]$ will lead to a better approximation of Φ than Ψ. Let us for simplicity assume that the sparse approximation routine always manages to retrieve the correct support realisation I, which is true with high enough probability that the resulting error remains small whenever Ψ is close to Φ, cp. [18]. Then the best approximation to y using Ψ_I is $\hat{y} = \Psi_I \Psi_I^\dagger y$, meaning $\hat{x}_I = \Psi_I^\dagger y = \Psi_I^\dagger \Phi_I x_I$, where $\Psi_I^\dagger$ is the Moore-Penrose pseudo-inverse. If Ψ_I has full rank n, then we have

$$\hat{x}_I = \Psi_I^\dagger \Phi_I x_I = \Psi_I^\dagger \Psi_I x_I + \Psi_I^\dagger (\Phi_I - \Psi_I) x_I = x_I + \Psi_I^\dagger Z_I x_I = x_I + (\Psi_I^t \Psi_I)^{-1} \Psi_I^t Z_I x_I.$$

If $G_{I,I} = \Psi_I^t \Psi_I - \mathbb{I}$ has small operator norm, which for realisations of rejective sampling happens with high probability as long as $G = \Psi^t \Psi - \mathbb{I}$ has small entries and small weighted operator norm, [17], then using a Neumann series we have $(\Psi_I^t \Psi_I)^{-1} = \sum_{n=0}^{\infty} (-G_{I,I})^n$. This means that in first order we can approximate $\Psi_I^\dagger$ by Ψ_I^t and in second order by $\Psi_I^t - G_{I,I}\Psi_I^t$. So for most support realisations I we have

$$\begin{aligned}
\hat{x}_I &\doteq x_I + \Psi_I^t Z_I x_I = D_I x + D_I \Psi^t Z D_I x \quad \text{and} \\
\hat{x}_I &\ddot{=} x_I + \Psi_I^t Z_I x_I - G_{I,I} \Psi_I^t Z_I x_I = D_I x + D_I \Psi Z D_I x - D_I G D_I \Psi^t Z D_I x.
\end{aligned}$$

Assuming again that $x = \sigma$ is a Rademacher sequence, we now arrive at Relation (3),

$$\mathbb{E}[y\hat{x}^t] \doteq \Phi\, \mathbb{E}[D_S x x^t D_S + D_S x x^t D_S Z^t \Psi D_S] = \Phi[D_\pi + (Z^t \Psi) \odot \Sigma_n].$$

Similarly, using the second order approximation and recalling that $Z^t \Psi = D_\varepsilon + H$ we get

$$\begin{aligned}
\mathbb{E}[y\hat{x}^t] &\ddot{=} \Phi \mathbb{E}[D_S + D_S Z^t \Psi D_S - D_I Z^t \Psi D_S G D_S] \\
&= \Phi[D_\pi + (Z^t \Psi - D_\varepsilon G) \odot \Sigma_n] - \Phi \mathbb{E}[D_S H D_S G D_S].
\end{aligned}$$

Theorem 2 already allows us to directly bound the first two terms. To control the last term note that I is a discrete random variable so we can interpret the expectation as a sum of matrices $\sum_\ell A_\ell B_\ell$. Let $A = (A_1, \dots, A_L)$ and $B = (B_1^t, \dots, B_L^t)^t$, then $\| \sum_\ell A_\ell B_\ell \|^2 = \|AB\|^2 \leq \|A\|^2 \cdot \|B\|^2 = \|AA^t\| \cdot \|B^t B\|$. Applying this to the expectation above where $A_\ell \sim \mathbb{P}_n(S = I)^{1/2} D_I H D_I$ and $B_\ell \sim \mathbb{P}_n(S = I)^{1/2} D_I G D_I$ leads to

$$\|\mathbb{E}[D_S H D_S G D_S]\|^2 \leq \|\mathbb{E}[D_S H D_S H^t D_S]\| \cdot \|\mathbb{E}[D_S G^t D_S G D_S]\|,$$

which we can finally bound using Corollary 3. Note, that using tricks similar to the ones above together with Lemma 5, Theorem 2 and Corollary 3 also allows to treat $\mathbb{E}[y\hat{x}^t]$ in full accuracy. This opens up the way to the convergence analysis of practically relevant dictionary learning algorithms, which usually involve additional components such as $\mathbb{E}[\hat{x}\hat{x}^t]$, see e.g [18].

Appendix B. Elementary proof of Lemma 5

For the reader's convenience we here prove Lemma 5 using only the language and techniques introduced in the paper.

Proof [of Lemma 5] We first want to show that for two indices $i \neq j$ we have

$$\pi_{ij}(n) = \pi_i(n) \cdot \frac{\pi_j(n-1) - \pi_{ij}(n-1)}{1 - \pi_i(n-1)}.$$

Recall that for any set J not containing the index i we have

$$\frac{p_i}{1-p_i} \cdot \mathbb{P}(J) = \mathbb{P}(J \cup \{i\}).$$

Using this relation we get

$$\begin{aligned}
\pi_{ij}(n) &= \frac{\sum_{I:|I|=n} \mathbb{1}_I(i)\mathbb{1}_I(j) \cdot \mathbb{P}(I)}{\sum_{I:|I|=n} \mathbb{P}(I)} \cdot \frac{\sum_{I:|I|=n, i\in I} \mathbb{P}(I)}{\sum_{I:|I|=n, i\in I} \mathbb{P}(I)} \\
&= \frac{\sum_{I:|I|=n, i\in I} \mathbb{1}_I(j) \cdot \mathbb{P}(I)}{\sum_{|I|=n, i\in I} \mathbb{P}(I)} \cdot \pi_i(n) \\
&= \frac{p_i}{1-p_i} \cdot \frac{1-p_i}{p_i} \cdot \frac{\sum_{J:|J|=n-1, i\notin J} \mathbb{1}_J(j) \cdot \mathbb{P}(J)}{\sum_{J:|J|=n-1, i\notin J} \mathbb{P}(J)} \cdot \pi_i(n) \cdot \frac{\sum_{J:|J|=n-1} \mathbb{P}(J)}{\sum_{J:|J|=n-1} \mathbb{P}(J)} \\
&= \pi_i(n) \cdot \frac{\sum_{J:|J|=n-1, i\notin J} \mathbb{1}_J(j) \cdot \mathbb{P}(J)}{\sum_{J:|J|=n-1} \mathbb{P}(J)} \cdot \frac{\sum_{J:|J|=n-1} \mathbb{P}(J)}{\sum_{J:|J|=n-1, i\notin J} \mathbb{P}(J)}.
\end{aligned}$$

Further rewriting the fractions in the expression above yields

$$\frac{\sum_{J:|J|=n-1, i\notin J} \mathbb{1}_J(j) \cdot \mathbb{P}(J)}{\sum_{J:|J|=n-1} \mathbb{P}(J)} = \underbrace{\frac{\sum_{J:|J|=n-1} \mathbb{1}_J(j) \cdot \mathbb{P}(I)}{\sum_{J:|J|=n-1} \mathbb{P}(J)}}_{\pi_j(n-1)} - \underbrace{\frac{\sum_{J:|J|=n-1} \mathbb{1}_J(i)\mathbb{1}_J(j) \cdot \mathbb{P}(J)}{\sum_{J:|J|=n-1} \mathbb{P}(J)}}_{\pi_{ij}(n-1)}$$

as well as

$$\frac{\sum_{J:|J|=n-1} \mathbb{P}(J)}{\sum_{J:|J|=n-1, i\notin J} \mathbb{P}(J)} = \frac{\sum_{J:|J|=n-1} \mathbb{P}(J)}{\sum_{J:|J|=n-1} \mathbb{P}(J) - \sum_{J:|J|=n-1, i\in J} \mathbb{P}(J)} = \frac{1}{1-\pi_i(n-1)},$$

which completes the proof of the first equality.

Next we will show more generally that for arbitrary $L \subseteq [K]$, $\pi_L(n-1) \leq \pi_L(n)$. We define $\mathcal{L} = \{I \subseteq [K] : L \subseteq I\}$. Using this together with the definition of $\mathbb{P}_n$ we can rewrite $\pi_L(n-1) \leq \pi_L(n)$ as

$$\frac{\sum_{J:|J|=n-1} \mathbb{1}_{\mathcal{L}}(J) \cdot \mathbb{P}(J)}{\sum_{J:|J|=n-1} \mathbb{P}(J)} \leq \frac{\sum_{I:|I|=n} \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}(I)}{\sum_{I:|I|=n} \mathbb{P}(I)},$$

which is equivalent to

$$\sum_{(I,J):|J|=n-1,|I|=n} \mathbb{1}_{\mathcal{L}}(J)\cdot\mathbb{P}(J)\mathbb{P}(I) \leq \sum_{(I,J):|J|=n-1,|I|=n} \mathbb{1}_{\mathcal{L}}(I)\cdot\mathbb{P}(J)\mathbb{P}(I).$$

Next we use a similar partition as in the proof of Theorem 2(b), that is for $m \in \{1,\dots,n\}$, together with a set $A \subseteq [K]$ with $|A| = n-m$ and a set $B \subseteq [K]\setminus A$ with $|B| = 2m-1$, we look at the collection of pairs (I,J) with intersection A and symmetric difference B, i.e.,

$$\mathcal{Q}_{A,B} := \{(I,J) : I,J \subseteq [K], |I| = n, |J| = n-1, I\cap J = A, I\triangle J = B\}.$$

Since each pair (I,J) with $|I| = n, |J| = n-1$ can be uniquely assigned to a collection $\mathcal{Q}_{A,B}$ and $\mathbb{P}(I)\mathbb{P}(J)$ is constant for all $(I,J) \in \mathcal{Q}_{A,B}$, it is sufficient to show that

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_{\mathcal{L}}(J) \leq \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_{\mathcal{L}}(I)$$

or equivalently that

$$|\{(I,J)\in\mathcal{Q}_{A,B} : L\subseteq J\}| \leq |\{(I,J)\in\mathcal{Q}_{A,B} : L\subseteq I\}. \tag{12}$$

If L is not contained in $A\cup B$, there is no valid pair $(I,J)\in\mathcal{Q}_{A,B}$ and the inequality trivially holds. If $L\subseteq A\cup B$ we abbreviate $L_A = L\cap A$ and $L_B = L\cap B$. Since $L_A\subseteq A$, all pairs in $(I,J)\in\mathcal{Q}_{A,B}$ automatically satisfy $L_A\subseteq I$ and $L_A\subseteq J$ so we can rewrite (12) as

$$|\{(I,J)\in\mathcal{Q}_{A,B} : L_B\subseteq J\}| \leq |\{(I,J)\in\mathcal{Q}_{A,B} : L_B\subseteq I\}|. \tag{13}$$

Since we need to have $A\cup L_B\subseteq J$, in case $|A\cup L_B| = |A| + |L_B| \geq n$ the left hand side in (13) is zero and the inequality holds. Finally, if $k = |L_B| < n - |A| = m$, we can still choose $m-k-1$ out of the $2m-k-1$ resp. $m-k$ out of the $2m-k-1$ remaining elements in B to fill I resp. J and create a valid pair. Since

$$\binom{2m-k-1}{m-k-1} \leq \binom{2m-k-1}{m-k},$$

the inequality in (13) is satisfied, which completes the proof. ■

Finally, let us quickly exploit the partitioning trick to get an elementary proof for another bound, that comes in handy when working with rejective sampling inclusion probabilities,

$$\pi_{L\cup M}(n) \leq \pi_L(n)\cdot\pi_M(n) \quad \text{if} \quad M\cap L = \emptyset. \tag{14}$$

Again this bound can be proved using 3 lines and some heavy machinery from [3], i.e., since $\mu = \mathbb{P}$ is strongly Raleigh, so is $\mu_n = \mathbb{P}_n$, which then implies that $\mathbb{P}_n$ is negatively associated, so for increasing functions $f = \sum_{\ell\in L}\mathbf{1}_S(\ell)$ and $g = \sum_{m\in M}\mathbf{1}_S(m)$ we have $\pi_{L\cup M}(n) = \int fg\mathrm{d}\mu_n \leq \int f\mathrm{d}\mu_n \int g\mathrm{d}\mu_n = \pi_L(n)\cdot\pi_M(n)$.
Alternatively, we can use somewhat more lines but only elementary combinatorics.
We define $\mathcal{L} = \{I\subseteq[K] : L\subseteq I\}$ and $\mathcal{M} = \{I\subseteq[K] : M\subseteq I\}$. Using this together with the definition of $\mathbb{P}_n$ we can rewrite $\pi_{L\cup M}(n) \leq \pi_L(n)\cdot\pi_M(n)$ as

$$\frac{\sum_{I:|I|=n}\mathbb{1}_{\mathcal{M}}(I)\cdot\mathbb{1}_{\mathcal{L}}(I)\cdot\mathbb{P}(I)}{\sum_{I:|I|=n}\mathbb{P}(I)} \leq \frac{\sum_{J:|J|=n}\mathbb{1}_{\mathcal{M}}(J)\cdot\mathbb{P}(J)}{\sum_{J:|J|=n}\mathbb{P}(J)} \cdot \frac{\sum_{I:|I|=n}\mathbb{1}_{\mathcal{L}}(I)\cdot\mathbb{P}(I)}{\sum_{I:|I|=n}\mathbb{P}(I)},$$

which is equivalent to

$$\sum_{(I,J):|I|=|J|=n} \mathbb{1}_{\mathcal{M}}(I) \cdot \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}(I)\mathbb{P}(J) \leq \sum_{(I,J):|I|=|J|=n} \mathbb{1}_{\mathcal{M}}(J) \cdot \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}(I)\mathbb{P}(J).$$

We now use the same decomposition as in the proof of Theorem 2(b), meaning for $m \in \{0, \dots, n\}$, $A \subseteq [K]$ with $|A| = n - m$ and $B \subseteq [K] \setminus A$ with $|B| = 2m$, we define

$$\mathcal{Q}_{A,B} := \{(I, J) : I, J \subseteq [K], |I| = |J| = n, I \cap J = A, I \triangle J = B\}.$$

Using the usual argument we now only need to show that

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_{\mathcal{M}}(I) \cdot \mathbb{1}_{\mathcal{L}}(I) \leq \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_{\mathcal{M}}(J) \cdot \mathbb{1}_{\mathcal{L}}(I)$$

or equivalently that

$$|\{(I, J) \in \mathcal{Q}_{A,B} : M \subseteq I, L \subseteq I\}| \leq |\{(I, J) \in \mathcal{Q}_{A,B} : M \subseteq J, L \subseteq I\}|.$$

If $L \cup M$ is not contained in $A \cup B$ there is no valid pair $(I, J) \in \mathcal{Q}_{A,B}$ and the inequality trivially holds. If $(L \cup M) \subseteq A \cup B$ we abbreviate $L_A = L \cap A$, $L_B = L \cap B$, $M_A = M \cap A$ and $M_B = M \cap B$. Since $(L_A \cup M_A) \subseteq A$, all pairs in $(I, J) \in \mathcal{Q}_{A,B}$ automatically satisfy $(L_A \cup M_A) \subseteq I$, $M_A \subseteq J$ and $L_A \subseteq I$ so we can rewrite the inequality we want to show as

$$|\{(I, J) \in \mathcal{Q}_{A,B} : M_B \subseteq I, L_B \subseteq I\}| \leq |\{(I, J) \in \mathcal{Q}_{A,B} : M_B \subseteq J, L_B \subseteq I\}|. \tag{15}$$

Since we need to have $(A \cup L_B \cup M_B) \subseteq I$, in case $|A \cup L_B \cup M_B| = |A| + |L_B| + |M_B| > n$ the left hand side in (15) is zero and the inequality trivially holds. Finally, if $k = |L_B| + |M_B| \leq n - |A| = m$, we can still choose $m - k$ out of the $2m - k$ resp. $m - |L_B|$ out of the $2m - k$ remaining elements in B to fill I resp. J and create a valid pair. Since $|L_B| \leq k$, we have

$$\binom{2m-k}{m-k} \leq \binom{2m-k}{m-|L_B|},$$

meaning the inequality in (15) is again satisfied, which completes the proof.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under Grant no. Y760.

References

[1] N. Aires. Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodology and Computing in Applied Probability*, 1(4):457–469, 1999.

[2] P. Bertail and S. Clémençon. Bernstein-type exponential inequalities in survey sampling: Conditional Poisson sampling schemes. *Bernoulli*, 25(4B):3527–3554, 2019.

[3] J. Borcea, P. Brändén, and T.M. Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2):521–567, 2009.

[4] E. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[5] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[6] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[7] D. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Aca. Sci.,*, 100(5):2197–2202, March 2003.

[8] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[9] J. J. Fuchs. Extension of the Pisarenko method to sparse linear arrays. *IEEE Transactions on Signal Processing*, 45:2413–2421, October 1997.

[10] J. Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[11] J. Hájek. *Sampling from a Finite Population*. Marcel Dekker Inc., New York, 1981.

[12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 2nd edition, 2013.

[13] S. Kochar and R. Korwar. On random sampling without replacement from a finite population. *Annals of the Institute of Statistical Mathematics*, 53(3):631–646, 2001.

[14] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

[15] H. Milbrodt. Comparing inclusion probabilities and drawing probabilities for rejective sampling and successive sampling. *Statistics and Probability Letters*, 14(3):243–246, 1992.

[16] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal Matching Pursuit: recursive function approximation with application to wavelet decomposition. In *Asilomar Conf. on Signals Systems and Comput.*, 1993.

[17] S. Ruetz and K. Schnass. Submatrices with non-uniformly selected random supports and insights into sparse approximation. *SIAM Journal on Matrix Analysis and Applications*, 42(3):1268–1289, 2021.

[18] S. Ruetz and K. Schnass. Convergence regions of alternating minimisation algorithms for dictionary learning. *SIAM Journal on Optimization*, 36(1):320–349, 2026.

[19] Y. Tillé. *Sampling Algorithms*. Springer Series in Statistics, New York: Springer, 2006.

[20] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[21] J. Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

[22] Y. Yaming. On the inclusion probabilities in some unequal probability sampling plans without replacement. *Bernoulli*, 18(1):279–289, 2012.