

Conformalized Frequency Estimation from Sketched Data

Matteo Sesia

University of Southern California, Los Angeles, USA

SESIA@MARSHALL.USC.EDU

Stefano Favaro

University of Torino and Collegio Carlo Alberto, Torino, Italy

STEFANO.FAVARO@UNITO.IT

Abstract

A flexible conformal inference method is developed to construct confidence intervals for the frequencies of queried objects in a very large data set, based on the information contained in a much smaller sketch of those data. The approach is completely data-adaptive and makes no use of any knowledge of the population distribution or of the inner workings of the sketching algorithm; instead, it constructs provably valid frequentist confidence intervals under the sole assumption of data exchangeability. Although the proposed solution is much more broadly applicable, this paper explicitly demonstrates its use in combination with the famous count-min sketch algorithm and a non-linear variation thereof to facilitate the exposition. The performance is compared to that of existing frequentist and Bayesian alternatives through several experiments with synthetic data as well as with real data sets consisting of SARS-CoV-2 DNA sequences and classic English literature.

Keywords: Sketching, conformal inference, memory constraints, privacy, frequency estimation.

1. Introduction

1.1. Frequency queries from sketched data

Efficiently finding the frequency of an object from a sketch of a very large data set is an important problem in computer science (Misra and Gries, 1982; Charikar et al., 2002), which has applications in many scientific endeavors including machine learning (Shi et al., 2009), cybersecurity (Schechter et al., 2010), natural language processing (Goyal et al., 2012), genetics (Zhang et al., 2014), and data privacy (Cormode et al., 2018). The count-min sketch (CMS) of Cormode and Muthukrishnan (2005) is a renowned algorithm for compressing a data set of m objects $Z_1, \dots, Z_m \in \mathcal{Z}$, belonging to a discrete (and possibly infinite) space \mathcal{Z} , into a representation with reduced memory footprint, while allowing simple approximate queries about the observed frequency of any possible $z \in \mathcal{Z}$. At the heart of the CMS lie $d \geq 1$ different w -wide hash functions $h_j : \mathcal{Z} \rightarrow [w] := \{1, \dots, w\}$, for all $j \in [d] := \{1, \dots, d\}$ and some integer $w \geq 1$. Each hash function maps the elements of \mathcal{Z} into one of w possible buckets, and it is designed to ensure that distinct values of z tend to populate the buckets uniformly. Hash functions are typically chosen at random from a *pairwise independent* family \mathcal{H} , which ensures that the probability (over the randomness in the choice of hash functions) that two distinct objects $z_1, z_2 \in \mathcal{Z}$ are mapped by two different hash functions into the same bucket is equal to $1/w^2$. The data set Z_1, \dots, Z_m is then compressed into a sketch matrix $C \in \mathbb{N}^{d \times w}$ with row sums equal to m . The element in the j -th row and k -th column of C counts the number of data points mapped by the j -th hash function into the k -th bucket:

$$C_{j,k} = \sum_{i=1}^m \mathbb{1}[h_j(Z_i) = k], \quad j \in [d]. \quad (1)$$

Because d and w are such that $d \cdot w \ll m$, the matrix C loses information compared to the full data set, but it has the advantage of requiring much less space to store.

Given a sketch C from (1), one may want to answer queries about the original data. A fundamental instance of this problem consists in estimating the empirical frequency of an object $z \in \mathcal{Z}$:

$$f_m(z) := \sum_{i=1}^m \mathbb{1}[Z_i = z]. \quad (2)$$

One solution is to return the smallest count among the d buckets into which z is mapped:

$$\hat{f}_{\text{up}}^{\text{CMS}}(z) = \min_{j \in [d]} \{C_{j, h_j(z)}\} \geq f_m(z). \quad (3)$$

This is a valid deterministic upper bound for the true frequency $f_m(z)$ (Cormode and Muthukrishnan, 2005). Of course, $\hat{f}_{\text{up}}^{\text{CMS}}(z)$ generally tends to overestimate $f_m(z)$ because different objects are sometimes hashed into the same bucket. However, the independence of the hash functions ensures that collisions are not too common, leading to a probabilistic lower bound for the true $f_m(z)$.

Theorem 1 (Cormode and Muthukrishnan (2005)) *Fix any $\delta, \epsilon \in (0, 1)$. Suppose $d = \lceil -\log \delta \rceil$ and $w = \lceil e/\epsilon \rceil$. Then, $\mathbb{P}_{\mathcal{H}}[f_m(z) \geq \hat{f}_{\text{up}}^{\text{CMS}}(z) - \epsilon m] \geq 1 - \delta$ for any fixed $z \in \mathcal{Z}$, where $\hat{f}_{\text{up}}^{\text{CMS}}(z)$ is as in (1) and depends on d, w .*

For example, if $d = 3$, a 95% lower bound on $f_m(z)$ is $\hat{f}_{\text{up}}^{\text{CMS}}(z) - \lceil e \cdot m/w \rceil$. The subscript \mathcal{H} in Theorem 1 clarifies that the randomness is with respect to the hash functions, while Z_1, \dots, Z_m and z are fixed. Although this bound is useful, for example to choose d and w prior to sketching, it is not fully satisfactory. First, it is often very conservative if the data are randomly sampled from some distribution as opposed to adversarially fixed. Second, it is not flexible: δ is fixed by d (as opposed to being determined by the user), while ϵ is fixed only by the hash width. To address these issues, a bootstrap method was developed by Ting (2018) to compute tighter and data-adaptive lower bounds for $f_m(z)$, under the additional assumptions that the data and the query are independent random samples from some underlying distribution. This approach is nearly exact and optimal for the CMS, up to a possible finite-sample discrepancy between the bootstrap and population distributions, but it relies on the specific *linear* structure of the CMS (the sketch matrix C in (3) can be written as a linear combination of the true frequencies of all objects in the data set), and the idea is therefore not easily extendable to other sketching algorithms that may often be preferable in practice.

One limitation of the CMS is that it does not mitigate as well as possible the adverse effects of random hash collisions, which can result in an overly conservative deterministic upper bound with some data sets. Fortunately, a simple modification provides meaningful improvements: the CMS with *conservative updates* (CMS-CU). Whenever a new object Z_i is sketched, the CMS-CU only updates the row of C with the smallest value of $C_{j, h_j(z)}$, leaving the other counters unaltered. Then, the same CMS deterministic upper bound in (3) remains valid. The procedure is outlined in Algorithm 1, Appendix A. The CMS-CU can greatly improve query accuracy compared to the vanilla CMS (Estan and Varghese, 2002), but this practical advantage comes at some cost because the algorithm is non-linear. In particular, the theoretical analysis of the CMS-CU beyond a deterministic upper bound is more challenging, and it appears to be a relatively less explored topic.

1.2. Contributions

This paper will develop a novel method for constructing tight and provably valid frequentist confidence intervals for frequency queries based on sketched data, at any desired level. This method will work seamlessly with the CMS-CU as well as with any other possible sketching algorithm, regardless of whether it is linear or very complex. Our solution is a non-parametric statistical one, in the sense that it makes minimal exchangeability assumptions about the data generating process and then leverages the sampling randomization instead of the specific structure of the sketch or the randomness in its hash functions. Although this will require higher computational power and memory usage compared to traditional sketching, the difference in these requirements will be negligible compared to the size of the data sets involved. Given that our solution will make use of conformal inference, a brief review of the relevant ideas from that literature is provided in the next section.

1.3. Related work

Other types of lower bounds have been developed for the CMS (Cormode and Yi, 2020); however, most are similar to that of Theorem 1 in the sense that they make no assumptions about the data and only make use of the hashing randomness. A different Bayesian perspective is taken by Cai et al. (2018) and Dolera et al. (2021): they model the data distribution with a prior and then compute the posterior of the queried frequencies given the sketch. Our approach is closer to that of Ting (2018) in that we seek similar frequentist guarantees by leveraging the randomness in the data. However, the theoretical frameworks and methodologies are different: Ting (2018) exploits the specific linear structure of the CMS, while we view the sketch as a black box and rely on conformal inference.

Conformal inference was pioneered by Vovk and collaborators (Vovk et al., 2005) and was brought to the statistics spotlight by Lei et al. (2018). Although primarily conceived for supervised prediction (Vovk et al., 2009; Vovk, 2015; Lei and Wasserman, 2014; Romano et al., 2019; Izbicki et al., 2019), conformal inference has found many other applications including outlier detection (Bates et al., 2021), causal inference (Lei and Candès), and survival analysis (Candès et al., 2021). However, to the best of our knowledge its potential for sketching remained untapped.

There exist many algorithms for using a sketch of a large data set to compute approximate frequency queries; some are relatively similar to the CMS (Fan et al., 2000; Goyal and Daumé, 2011; Pitel and Fouquier, 2015; Cormode and Yi, 2020), while others are decidedly more complicated and may involve complex learning algorithms (Hsu et al., 2019; Jiang et al., 2019). Our methodology is applicable to all of them, since nothing that we do is specifically tied to the CMS. However, the CMS is a classic algorithm and it provides a familiar example that will facilitate the exposition.

2. Preliminaries on conformal prediction

As this paper will develop a novel approach for frequency estimation from sketched data that re-purposes and extends statistical techniques for conformal prediction, it will be helpful to begin with a brief review of the relevant concepts from that field. Conformal prediction is typically discussed in a *supervised learning* context, where the data consist of pairs of observations (X_i, Y_i) , with X_i indicating a vector of *features* for the i -th observation and Y_i denoting the corresponding continuous (or discrete) *label*. In supervised learning, the goal is to use the information contained in $(X_1, Y_1), \dots, (X_m, Y_m)$ to learn a mapping between features and labels that can allow one to predict as accurately as possible the unseen label Y_{m+1} of a new sample with features X_{m+1} . In

particular, conformal prediction assumes that $(X_1, Y_1), \dots, (X_{m+1}, Y_{m+1})$ are exchangeable random samples from some unknown joint distribution over (X, Y) and then constructs a prediction interval $[\hat{L}_{m,\alpha}(X_{m+1}), \hat{U}_{m,\alpha}(X_{m+1})]$ with guaranteed *marginal coverage*:

$$\mathbb{P}[\hat{L}_{m,\alpha}(X_{m+1}) \leq Y_{n+1} \leq \hat{U}_{m,\alpha}(X_{m+1})] \geq 1 - \alpha, \quad (4)$$

for any fixed $\alpha \in (0, 1)$. Conformal prediction is extremely flexible, as it can leverage any machine learning algorithm to approximate from the data the true relation between X and Y , thus yielding relatively short intervals satisfying (4). Conformal inference can also be applied to construct more general prediction sets (Romano et al., 2020; Angelopoulos et al., 2020), which may be more useful than intervals for some applications, but it is appropriate to focus on intervals in this paper.

The most straightforward version of conformal prediction begins by randomly splitting the available observations into two disjoint subsets; these are assumed for simplicity to have equal size $n = m/2$, although this is unnecessary. The first n samples are employed to fit a completely arbitrary black-box machine learning model for predicting Y given X ; think of a deep neural network or a random forest, for example. The remaining data are utilized to assess the out-of-sample predictive accuracy of this model, which is quantified in terms of individual *conformity scores* for all n hold-out samples. Combined with the model learnt from the first half of the data, the empirical distribution of these conformity scores is then translated into an explicit recipe for constructing prediction intervals for future test points as a function of X_{m+1} ; these intervals are guaranteed to cover Y_{n+1} with probability at least $1 - \alpha$, treating all data as random. The details of this procedure will be clarified shortly. Importantly, the conformal coverage guarantee will hold exactly in finite samples, regardless of the accuracy of the machine learning model, under the sole assumption that X_{m+1} is exchangeable with the n hold-out data points. It is worth emphasizing that it is unnecessary for the exchangeability to involve also the data used to train the model, which in fact may even be considered as fixed, although it is sometimes easier to state that all data points are exchangeable.

The implementation of conformal inference depends on the choice of conformity scores. While several different options exist, an intuitive and general explanation is the following. Imagine that associated with the fitted machine learning model there exists a *nested sequence* (Gupta et al., 2019) of possible prediction intervals $[\hat{L}_{m,\alpha}(x; t), \hat{U}_{m,\alpha}(x; t)]$ for each possible x . This sequence is indexed by the index $t \in \mathcal{T} \subseteq \mathbb{R}$, which may be either discrete or continuous, and it is increasing, in the sense that $\hat{L}_{m,\alpha}(x; t_2) \leq \hat{L}_{m,\alpha}(x; t_1)$ and $\hat{U}_{m,\alpha}(x; t_2) \geq \hat{U}_{m,\alpha}(x; t_1)$ for all $t_2 \geq t_1$. Further, assume this sequence to be such that there exists $t_\infty \in \mathcal{T}$ such that $\hat{L}_{m,\alpha}(x; t_\infty) \leq Y \leq \hat{U}_{m,\alpha}(x; t_\infty)$ almost surely for all x . For example, one may consider a sequence of intervals in the form of $\hat{\psi}_m(x) \pm t$, where $\hat{\psi}_m$ is a regression function for a bounded continuous label Y given X learnt by the black-box machine learning model and t plays the role of a predictive standard error. Then, the conformity score for a data point with features $X = x$ and label $Y = y$ can be defined as the smallest index t such that y is contained in the sequence of prediction intervals corresponding to x :

$$E(X, Y) := \min \left\{ t \in \mathcal{T} : Y \in [\hat{L}_{m,\alpha}(x; t), \hat{U}_{m,\alpha}(x; t)] \right\}. \quad (5)$$

With this notation in place, the conformal prediction rule is then easily stated. Let $\mathcal{I}^{\text{calib}} \subset \{1, \dots, m\}$ indicate the subset of hold-out data points, which we have assumed without loss of generality to have size $n = m/2$. Let $\hat{Q}_{n, 1-\alpha}$ denote the $\lceil (1 - \alpha)(n + 1) \rceil$ smallest value among the n conformity scores $E(X_i, Y_i)$ evaluated for all $i \in \mathcal{I}^{\text{calib}}$. Then, the output conformal prediction

interval for the unseen label of a new data point with features X_{m+1} is:

$$\left[\hat{L}_{m,\alpha}(X_{m+1}; \hat{Q}_{n,1-\alpha}), \hat{U}_{m,\alpha}(X_{m+1}; \hat{Q}_{n,1-\alpha}) \right]. \quad (6)$$

Intuitively, this satisfies marginal coverage (4) because Y_{m+1} does not belong to the interval in (6) if and only if the imaginary score $E(X_{m+1}, Y_{m+1})$ is larger than $\hat{Q}_{n,1-\alpha}$. The rest is a simple exchangeability argument; see Romano et al. (2019) for details, or the proof of Theorem 2 below.

3. Conformalized sketching

3.1. Problem statement

Assume the m data points to be sketched, $Z_1, \dots, Z_m \in \mathcal{Z}$, are exchangeable random samples from some distribution P_Z on \mathcal{Z} . Although P_Z may be arbitrary and unknown, the statistical framework assumed here is more restrictive compared to the classical deterministic setting reviewed in Section 1.1, which treats the observations as fixed and therefore can in principle also handle non-stationary data streams as well as completely adversarial data sets. Yet, there are interesting applications in which the assumption of exchangeable random samples can be justified; this will be discussed in more detail in Sections 4 and 5. Consider an arbitrary *sketching* function $\phi : [\mathcal{Z}]^m \rightarrow \mathcal{C}$, where \mathcal{C} is some space with lower dimensions compared to $[\mathcal{Z}]^m$. For example, ϕ may represent the sketching performed by the CMS or CMS-CU, in which case $\mathcal{C} = \mathbb{N}^{d \times w}$, for some d, w such that $d \cdot w \ll m$. In general, our framework will treat ϕ as a *black box* and allow it to be anything, regardless of whether it involves random hashing or any other specific data compression technique. The goal is to leverage the exchangeability and the information in $\phi(Z_1, \dots, Z_m)$ to estimate the unobserved true empirical frequency $f_m(z)$ of some object $z \in \mathcal{Z}$, defined as in (2), while accounting for uncertainty. More precisely, although still informally speaking, we would like to construct a *confidence interval* in the form of $[\hat{L}_{m,\alpha}(z), \hat{U}_{m,\alpha}(z)]$ that is guaranteed to contain $f_m(z)$ “with probability at least $1 - \alpha$ ”, where the randomness here is with respect to the random data sampling.

One way to address the above problem is to assume that the query point z is also drawn exchangeably from the same distribution as Z_1, \dots, Z_m . Taking this viewpoint, we will refer to z as Z_{m+1} and consider the goal of computing a confidence interval $[\hat{L}_{m,\alpha}(Z_{m+1}), \hat{U}_{m,\alpha}(Z_{m+1})]$ depending on $\phi(Z_1, \dots, Z_m)$ that is as short as possible and such that

$$\mathbb{P} \left[\hat{L}_{m,\alpha}(Z_{m+1}) \leq f_m(Z_{m+1}) \leq \hat{U}_{m,\alpha}(Z_{m+1}) \right] \geq 1 - \alpha, \quad (7)$$

where the probability is with respect to Z_1, \dots, Z_{m+1} . The interpretation of (7) is as follows: the interval will cover the true frequency $f_m(z)$ for at least fraction $1 - \alpha$ of all future test points on average, if the queries are drawn exchangeably from the same distribution as the sketched data. Of course, this guarantee is not fully satisfactory because some random queries may be more likely to be observed than others, and therefore not all of our inferences will be equally reliable. In particular, the confidence intervals for rare queries may have lower coverage than expected, as illustrated by the following thought experiment. Imagine for example that P_Z has support on $\mathcal{Z} = \{0, 1, 2, \dots, 10^{100}\}$, with $\mathbb{P}[Z_i = 0] = 0.95$ and $\mathbb{P}[Z_i = z] = 0.05/(|\mathcal{Z}| - 1)$ for all $z \in \mathcal{Z} \setminus \{0\}$ and $i \geq 1$. In that case, a 95% confidence interval such as that in (7) may be completely unreliable for all but one specific query about $f_m(0)$. Therefore, we will focus instead on constructing confidence intervals that are simultaneously valid for both rare and random queries, adapting related ideas from

the conformalized multi-class classification (Vovk et al., 2005; Sadinle et al., 2019; Romano et al., 2020). Precisely, fix any partition $\mathcal{B} = (B_1, \dots, B_L)$ of $\{1, \dots, m\}$ into L bins, such that each B_l is a sub-interval of $\{1, \dots, m\}$, and construct a confidence interval $[\hat{L}_{m,\alpha}(Z_{m+1}), \hat{U}_{m,\alpha}(Z_{m+1})]$ depending on $\phi(Z_1, \dots, Z_m)$ and \mathcal{B} that is as short as possible and such that:

$$\mathbb{P} \left[\hat{L}_{m,\alpha}(Z_{m+1}) \leq f_m(Z_{m+1}) \leq \hat{U}_{m,\alpha}(Z_{m+1}) \mid f_m(Z_{m+1}) \in B \right] \geq 1 - \alpha, \quad \forall B \in \mathcal{B}. \quad (8)$$

The choice of \mathcal{B} will involve some trade-offs: finer partitions yield more reliable theoretical guarantees but possibly wider confidence intervals; in particular, intervals associated with a bin B may tend to be wider if $\mathbb{P}[Z \in B]$ is small. In practice, we will adopt $|\mathcal{B}| = 5$ later in this paper, but much larger $|\mathcal{B}|$ can be utilized when working with very big data, as it will become clear below.

3.2. Conformalization

In the attempt of adapting conformal inference to address the sketching problem, the first difficulty is that the latter is a data recovery problem, not a supervised prediction task. We propose to overcome this challenge by storing in memory the true frequencies for all objects in the first m_0 observations, with $m_0 \ll m$ but sufficiently large subject to memory constraints. Without loss of generality, assume $m_0 \ll m$; otherwise, the problem would be trivial. Let $n \leq m_0$ indicate the number of distinct objects among the first m_0 observations. The memory required to store these frequencies is $O(n)$, which will typically be negligible as long as m_0 is also small compared to the size of the sketch, $|\mathcal{C}|$. This approach turns our problem into a supervised prediction one, as detailed below.

During an initial *warm-up* phase, the frequencies of the n distinct objects among the first m_0 observations from the data stream, Z_1, \dots, Z_{m_0} , are stored exactly into:

$$f_{m_0}^{\text{wu}}(z) := \sum_{i=0}^{m_0} \mathbb{1}[Z_i = z]. \quad (9)$$

Next, the remaining $m - m_0$ data points are streamed and sketched, storing also the true frequencies for all instances of objects already seen during the warm-up phase. In other words, the following counters are computed and stored along with the data sketch $\phi(Z_{m_0+1}, \dots, Z_m)$:

$$f_{m-m_0}^{\text{sv}}(z) := \begin{cases} \sum_{i=m_0+1}^m \mathbb{1}[Z_i = z], & \text{if } f_{m_0}^{\text{wu}}(z) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Again, this requires only $O(n)$ memory. Now, let us define the variable Y_i for all $i \in \{1, \dots, m_0\} \cup \{m+1\}$ as the true frequency of Z_i among Z_{m_0+1}, \dots, Z_m :

$$Y_i := \sum_{i'=m_0+1}^m \mathbb{1}[Z_{i'} = Z_i]. \quad (11)$$

Note that Y_i is observable for $i \in \{1, \dots, m_0\}$, in which case $Y_i = f_{m-m_0}^{\text{sv}}(Z_i)$. For a new query, Z_{m+1} is the target of inference—in truth, the target is $f_m(Z_i) = Y_i + f_{m_0}^{\text{wu}}(Z_i)$, but the second term is already known exactly. To complete the connection between sketching and supervised conformal prediction, one still needs to define meaningful features X , and this is where the sketch

$\phi(Z_{m_0+1}, \dots, Z_m)$ comes into play. For each $i \in \{1, \dots, m_0\} \cup \{m+1, \dots\}$, define X_i as the vector containing the object of the query as well as all the information in the sketch:

$$X_i := (Z_i, \phi(Z_{m_0+1}, \dots, Z_m)). \quad (12)$$

The following result establishes that the pairs $(X_1, Y_1), \dots, (X_{m_0}, Y_{m_0}), (X_{m+1}, Y_{m+1})$ are exchangeable with one another. This anticipates that conformal prediction techniques can be applied to the supervised observations $(X_1, Y_1), \dots, (X_{m_0}, Y_{m_0})$ to predict Y_{m+1} given X_{m+1} , guaranteeing the marginal coverage property in (7). All mathematical proofs are in Appendix B.

Proposition 1 *If the data Z_1, \dots, Z_{m+1} are exchangeable, then the pairs of random variables $(X_1, Y_1), \dots, (X_{m_0}, Y_{m_0}), (X_{m+1}, Y_{m+1})$ defined in (11)–(12) are exchangeable with one another.*

Further, confidence sets satisfying the stronger conditional coverage property defined in (8) can be obtained by modifying the standard conformal inference procedure as follows. First, each data point (X_i, Y_i) in the calibration set $\mathcal{I}^{\text{calib}} = \{1, \dots, m_0\}$ is assigned to the appropriate frequency bin based on Y_i . Define n_l as the number of calibration data points assigned to bin B_l , for all $l \in \{1, \dots, L\}$. Then, $\hat{Q}_{n_l, 1-\alpha}(B_l)$ is defined as the $\lceil (1-\alpha)(n_l + 1) \rceil$ smallest value among the n_l conformity scores assigned to bin B_l . Finally, the confidence interval for a new random query X_{m+1} is given by

$$[\hat{L}_{m,\alpha}(X_{m+1}; \hat{Q}_{n, 1-\alpha}^*), \hat{U}_{m,\alpha}(X_{m+1}; \hat{Q}_{n, 1-\alpha}^*)],$$

where

$$\hat{Q}_{n, 1-\alpha}^* := \max_{l \in \{1, \dots, L\}} \hat{Q}_{n_l, 1-\alpha}(B_l),$$

and $[\hat{L}_{m,\alpha}(\cdot; t), \hat{U}_{m,\alpha}(\cdot; t)]$ is a rule for computing a nested sequence of intervals depending on Z_{m+1} and $\phi(Z_{m_0+1}, \dots, Z_m)$. Examples of such rules are in the next section. Importantly, these rules may involve parameters to be fitted on a subset of m_0^{train} supervised data points (X_i, Y_i) for $i \in \{1, \dots, m_0^{\text{train}}\}$, as long as the scores are only evaluated on the remaining $m_0 - m_0^{\text{train}}$ points. This procedure is outlined in Algorithm 2, Appendix A. The following result states that the confidence interval output by Algorithm 2 has the desired frequency-range conditional coverage.

Theorem 2 *If the data Z_1, \dots, Z_{m+1} are exchangeable, the confidence interval output by Algorithm 2 satisfies the frequency-range conditional coverage property defined in (8).*

3.3. Conformity scores

Algorithm 2 can accommodate virtually any data-adaptive rule, depending on Z_{m+1} and on the sketch $\phi := \phi(Z_{m_0+1}, \dots, Z_m)$, for computing nested confidence intervals. Two concrete options are presented here. For simplicity, we focus on one-sided confidence intervals; that is, we seek a $1 - \alpha$ lower bound for $f_m(X_{m+1})$. This is of particular interest when a deterministic upper bound $\hat{f}_{\text{up}}(Z_{m+1}) \geq f_m(Z_{m+1})$ is already available, as it is the case with the CMS or CMS-CU. Then, one can simply set $\hat{U}_{m,\alpha}((z, \phi); t) := \hat{f}_{\text{up}}(z)$ and focus on computing $\hat{L}_{m,\alpha}(\cdot; t)$.

The first option for $\hat{L}_{m,\alpha}(\cdot; t)$ is to use the following fixed rule:

$$\hat{L}_{m,\alpha}^{\text{fixed}}((z, \phi); t) := \max\{0, \hat{f}_{\text{up}}(Z_{m+1}) - t\}, \quad t \in \{0, 1, \dots, m\}. \quad (13)$$

In words, the lower bound for $f_m(Z_{m+1})$ in (14) is defined by shifting the deterministic upper bound downward by t . The appropriate value of t guaranteeing the desired coverage for future random queries is calculated by the conformal inference procedure outlined in Algorithm 2. This approach has the advantage of being simple and intuitive, and it is very similar to the optimal solution of Ting (2018) for the special case of the CMS. Further, it does not need training data, so all m_0 observations with tracked frequencies can be utilized for computing conformity scores.

The second option involves training but has the advantage of being more flexible; this is inspired by the approaches of Chernozhukov et al. (2021) and Sesia and Romano (2021) for conformalized regression. Concretely, consider a machine learning model that takes as input the known upper bound $\hat{f}_{\text{up}}(Z_i)$ and estimates the conditional distribution of $\hat{f}_{\text{up}}(Z_i) - f_m(Z_i)$ given $\hat{f}_{\text{up}}(Z_i)$. For example, think of a multiple quantile neural network (Taylor, 2000) or a quantile random forest (Meinshausen, 2006). After fitting this model on the m_0^{train} supervised data points (X_i, Y_i) allocated for training, let \hat{q}_t be the estimated α_t lower quantile of the of $\hat{f}_{\text{up}}(Z_i) - f_m(Z_i) \mid \hat{f}_{\text{up}}(Z_i)$, for all $t \in [1, \dots, T]$ and some fixed sequence $1 = \alpha_1 < \dots < \alpha_T = 0$. Without loss of generality, assume the machine learning model is that $\hat{q}_0 = 0$ and $\hat{q}_T = m$. Then, define $\hat{L}_{m,\alpha}(\circ; t)$ as:

$$\hat{L}_{m,\alpha}^{\text{adaptive}}((z, \phi); t) := \max \left\{ 0, \hat{f}_{\text{up}}(X_{m+1}) - \hat{q}_t \left(\hat{f}_{\text{up}}(X_{m+1}) \right) \right\}, \quad t \in \{0, 1, \dots, m\}. \quad (14)$$

The advantage of this second approach is that it can lead to an adaptive lower bound whose distance from the upper bound is data adaptive. For example, it may be the case that the sketching algorithm introduces higher uncertainty about the frequencies of more common items compared to rarer ones, or vice versa, and this pattern can be learnt given a sufficient number of supervised samples.

Of course, the two above examples of $\hat{L}_{m,\alpha}(\cdot; t)$ are not the only possible ones. Algorithm 2 can be applied in combination with any rule for computing nested sequences of lower bounds, and it could potentially leverage all high-dimensional information contained in $\phi(Z_{m_0+1}, \dots, Z_m)$, not just the deterministic upper bound for the CMS and CMS-CU. However, the applications in this paper focus on the CMS and CMS-CU, so other families of lower bounds are not discussed here.

4. Applications

4.1. Experiments with synthetic data sets

Conformalized sketching is applied with the two types of conformity scores described in Section 3.3. The adaptive scores are based on an isotonic distributional regression model (Henzi et al., 2021). The goal is to compute lower frequency bounds for random queries based on simulated data compressed by the CMS-CU, implemented with $d = 3$ hash functions of width $w = 1000$. In particular, $m = 100,000$ observations are sampled independently and identically distributed, from some distribution that will be specified below. The first $m_0 = 5000$ observations are stored without loss during the warm-up phase, as outlined in Algorithm 2, while the remaining 95,000 are compressed through the CMS-CU. The conformity scores are evaluated separately within $L = 5$ frequency bins, seeking the frequency-range conditional coverage property defined in (8). The bins are determined in a data-driven fashion so that each contains approximately the same probability mass; in practice, this is achieved by partitioning the range of frequencies for the objects tracked exactly by Algorithm 2 according to the observed empirical quantiles. Lower bounds for new frequency queries are computed for 10,000 new query points sampled exchangeably from the same distribution, adopting the significance level $1 - \alpha$. The quality of these lower bounds is quantified with two metrics: the

mean *length* of the resulting confidence intervals and the coverage the proportion of queried objects for which the true frequency is correctly covered, or empirical *coverage*. The average performance over 10 independent realizations of each experiment is reported.

Experiments are performed using synthetic data sampled from two families of distributions. First, we consider a Zipf distribution with probability mass $\mathbb{P}[Z_i = z] = z^{-a}/\zeta(a)$ for all $z \in \{1, 2, \dots\}$, where ζ is the Riemann Zeta function and $a > 1$ is a parameter that controls the power-law behavior of the tails and will be varied. Second, synthetic data will be generated from a random probability measure distributed as the Pitman-Yor prior (Pitman and Yor, 1997) with a standard Gaussian base distribution and parameters $\lambda > 0$ and $\sigma \in [0, 1)$, as explained in Appendix A.3. The parameter λ is set to $\lambda = 5000$, while σ is varied. For $\sigma = 0$ the Pitman-Yor prior reduces to the Dirichlet prior (Ferguson, 1973), while $\sigma > 0$ results in heavier tails.

Three alternative methods are considered as benchmarks. The first benchmark is the *classical* 95% lower bound reviewed in Theorem 1. The second benchmark is the *Bayesian* method of Cai et al. (2018), which assumes a non-parametric Dirichlet process prior for the distribution of the data stream, estimates its scaling parameter by maximizing the marginal likelihood of the observed sketch, and then computes the posterior distribution of the queried frequencies conditional on the observed sketch. The performance of the lower 5% posterior quantile is compared to the lower bounds obtained with the other methods according to the frequentist metrics defined above. The third benchmark is the bootstrap method of Ting (2018), which is nearly exact and optimal for the vanilla CMS (up to some possible finite-sample discrepancy between the bootstrap and population distributions) but is not theoretically valid for other sketching techniques.

Figure 1 compares the average performance of the proposed conformal method to those of the three benchmarks on the Zipf data. All methods achieve the desired marginal coverage (4), with the exception of the Bayesian approach which is based on a misspecified prior family in this case. The length of the confidence intervals indicates the classical bound is extremely conservative, while the bootstrap and conformal methods can provide relative informative bounds, particularly when a is larger and hash collisions become rarer. However, the conformal confidence intervals can be shorter, especially when implemented with the adaptive conformity scores. This result should not be surprising because the bootstrap approach is not guaranteed to be optimal for the non-linear compression induced by the CMS-CU. Indeed, as shown in Figure 2, the machine learning model deployed by our adaptive conformity scores is able to detect and take advantage of the fact that this non-linear sketch allows increased precision in the queries for more common objects.

Supplementary results reported in Appendix C demonstrate empirically that the CMS-CU leads to more precise queries with all methods compared to the vanilla CMS; see Figure A1. Figure A2 shows that, as expected, the conformal lower bounds no longer have a clear advantage over the bootstrap ones when the data are sketched with the vanilla CMS instead of the CMS-CU. In fact, although the conformal intervals obtained with the adaptive conformity scores can still be little shorter than the bootstrap ones even for the vanilla CMS, the latter have the advantage of (approximately, in the limit of large samples) satisfying an even stronger frequency-conditional coverage property equivalent to (8) with bins of size 1. Analogous results for the experiments with Pitman-Yor process data are also in Appendix C; see Figures A3–A6.

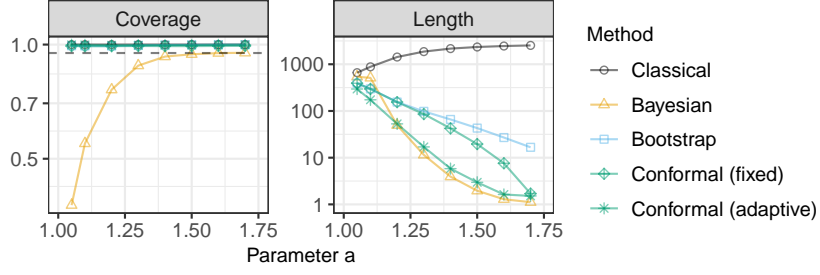


Figure 1: Coverage and length of 95% confidence intervals with data from a Zipf distribution, sketched with CMS-CU. The results are shown as a function of the Zipf parameter a .

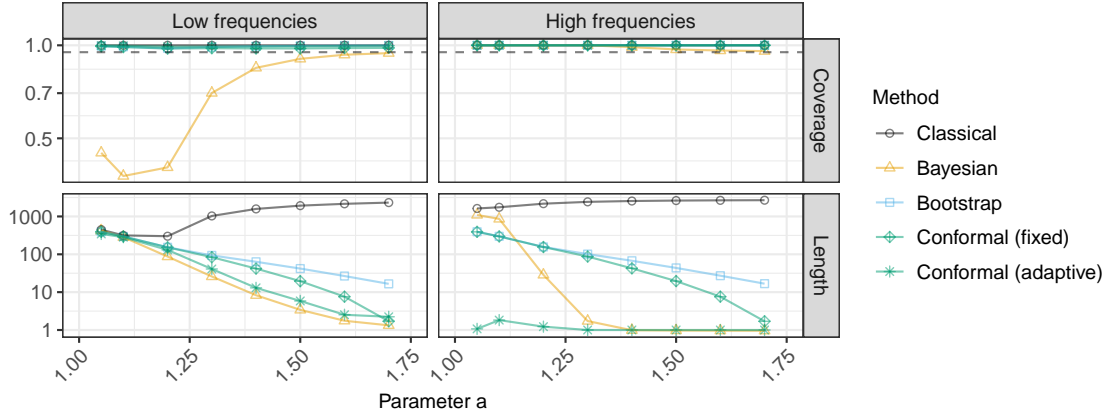


Figure 2: Performance of confidence intervals stratified by the true query frequency. Left: frequency below median; right: frequency above median. Other details are as in Figure 1.

4.2. Analysis of 16-mers in SARS-CoV-2 DNA sequences

This example application involves a data set of nucleotide sequences from SARS-CoV-2 viruses made publicly available by the National Center for Biotechnology Information ([Hatcher et al., 2017](#)). These data include 43,196 sequences, each consisting of approximately 30,000 nucleotides. The goal is to estimate the empirical frequency of each possible *16-mer*, a distinct sequence of 16 DNA bases found in contiguous nucleotides. Given that each nucleotide has one of 4 bases, there are $4^{16} \approx 4.3$ billion possible 16-mers. This means that exact tracking of all 16-mers is not unfeasible with a powerful computer, which allows us to validate the sketch-based queries. Note that sequences containing missing values are removed during pre-processing, for simplicity.

The experiments are carried out as in Section 4.1, but (1) a larger sample of 1,000,000 observations is sketched, and (2) the width w of the hash functions is varied as a control parameter. Figure 3 shows the performance of all methods as a function of the hash width, measured in terms of marginal

coverage and mean confidence interval width. All methods appear to have marginal coverage above the nominal level, except for the Bayesian approach which yields slightly lower coverage if w is large. For small hash widths, all methods construct confidence intervals of similar width, because the distribution of SARS-CoV-2 16-mers frequencies is quite concentrated with relatively narrow support (see Figure A7 in Appendix C), which makes it especially difficult to compress the data without much loss. By contrast, the proposed conformal methods yield noticeably shorter confidence intervals if w is large. Figure A8 reports the same results stratified by the frequency of the queried objects, while Figure A9 confirms the advantage of sketching with the CMS-CU as opposed to the vanilla CMS. Table A1 visualizes a list of 10 common and 10 rare queries along with their corresponding deterministic upper bounds for $w = 50,000$, comparing the lower bounds obtained with each method. Table A2 shows analogous results obtained with $w = 5,000$.

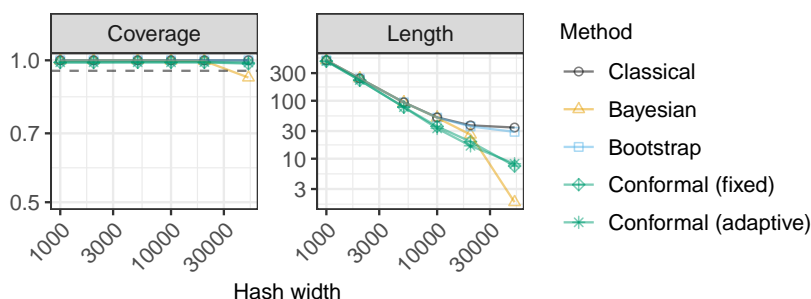


Figure 3: Performance of confidence intervals based on sketched SARS-CoV-2 sequence data. The results are shown as a function of the hash width. Other details are as in Figure 1.

Figure A10 compares the performance of different frequency *point-estimates*, measured in terms of mean absolute deviation from the true frequency. With the classical method we take for simplicity the midpoint of the 95% confidence interval as a point estimate, although there exist of course also other approaches (Cormode and Yi, 2020). For the other methods, the point estimate is defined as the lower confidence bound at level $\alpha = 0.5$; in the Bayesian case, this is the posterior median. Although a conformal lower bound at level $\alpha = 0.5$ is not always a reliable estimator of conditional medians (Medarametla and Candès, 2021), because the conformal coverage guarantees treat the query as random, this approach consistently outperformed the benchmarks in all of our experiments.

4.3. Analysis of 2-grams in English literature

The second application considered in this paper is based on a data set consisting of the full texts of 18 open-domain classic pieces of English literature downloaded using the NLTK Python package (Bird et al., 2009) from the Gutenberg Corpus (Project Gutenberg). The goal is to count the frequencies of all possible 2-grams—consecutive pairs of English words—across this corpus. After some basic preprocessing to remove punctuation and unusual words (only those contained in a dictionary of size 25,487 common English words are retained), the total number of 2-grams left in this data set is approximately 1,700,000 (although the total number of all *possible* 2-grams within this dictionary is

approximately 650,000,000). The same experiments are then carried out as in Section 4.2, sketching 1,000,000 randomly sampled 2-grams and querying 10,000 independent 2-grams.

Figure 4 shows the conformal intervals with adaptive scores achieve the desired coverage and tend to have the shortest width, while the Bayesian intervals are not valid unless the hashes are very wide. The conformal approach has a larger advantage here because these data can be compressed more efficiently compared to the one in the previous section because the frequency distribution of English 2-grams has power-law tails; see Figure A7. Additional results along the lines of those in the previous section are in Appendix C; see Figures A11–A13 and Tables A1–A2.

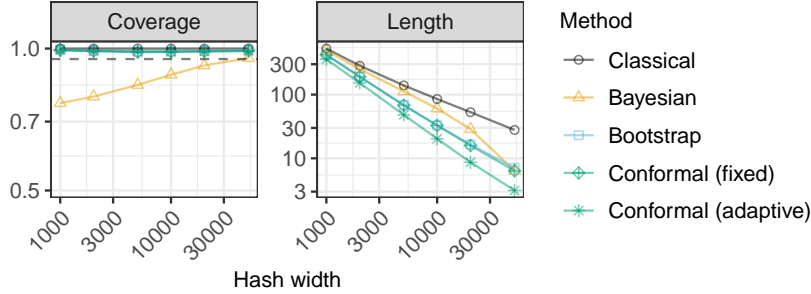


Figure 4: Performance of confidence intervals for random frequency queries, for a sketched data set of English 2-grams in classic English literature. Other details are as in Figure 3.

5. Discussion

Conformalized sketching is a non-parametric and data-adaptive statistical method for quantifying uncertainty in problems involving frequency estimation from sketched data. This paper has revolved around the CMS because that is a prominent technique for which several benchmarks are available. However, the key strength of conformalized sketching is that it operates under the sole assumption that the data are exchangeable with one another as well as with the future queries, without requiring anything from the sketching algorithm, which may be completely arbitrary and even unknown. Therefore, this methodology could be applicable quite broadly when one needs to retrieve information from large data sets that must be sketched due to memory or privacy constraints. Of course, exchangeability is not always an appropriate assumption, and therefore sometimes one has to rely on more conservative bounds based on hashing randomness. Yet there are interesting situations in which the sketched data can be seen as exchangeable random samples from a population. For example, in natural language processing one may wish to count the frequencies of non-contiguous tuples of words co-occurring in the same sentence within a large corpus, expanding on the example of Section 4.3. In that case, it would be unfeasible to keep track of all frequencies exactly, but sketching random samples would be relatively easy. One limitation of conformalized sketching is that it does not provide guarantees about the expected proportion of correct *unique* queries. In fact, if the queries are randomly sampled exchangeably with the sketched data points, some of them may be redundant. Even though the expected proportion of incorrect unique queries is sometimes below

α (Figures A14–A15), this is not always the case (Figure A16). It may be possible to modify the procedure to achieve this additional guarantee, but we leave this intriguing problem for future work.

Software availability

Code and data needed to reproduce these experiments are available online at <https://github.com/msesia/conformalized-sketching>.

References

- Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.
- Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *preprint at arXiv:2104.08279*, 2021.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- Diana Cai, Michael Mitzenmacher, and Ryan P Adams. A Bayesian nonparametric view on count-min sketch. In *Advances in Neural Information Processing Systems 31*, pages 8782–8791, 2018.
- Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.
- Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.
- Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021.
- Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.
- Graham Cormode and Ke Yi. *Small summaries for big data*. Cambridge University Press, 2020.
- Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1655–1658, 2018.
- Emanuele Dolera, Stefano Favaro, and Stefano Peluchetti. A Bayesian nonparametric approach to count-min sketch under power-law data streams. In *International Conference on Artificial Intelligence and Statistics*, pages 226–234. PMLR, 2021.
- Cristian Estan and George Varghese. New directions in traffic measurement and accounting. In *Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 323–336, 2002.

- Li Fan, Pei Cao, Jussara Almeida, and Andrei Z Broder. Summary cache: a scalable wide-area web cache sharing protocol. *IEEE/ACM transactions on networking*, 8(3):281–293, 2000.
- S. Thomas Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.
- Amit Goyal and Hal Daumé. Lossy conservative update (lcu) sketch: Succinct approximate count storage. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 878–883, 2011.
- Amit Goyal, Hal Daumé III, and Graham Cormode. Sketch algorithms for estimating point queries in nlp. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1093–1103, 2012.
- Chirag Gupta, Arun K Kuchibhotla, and Aaditya K Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *preprint at arXiv:1910.10562*, 2019.
- Eneida L Hatcher, Sergey A Zhdanov, Yiming Bao, Olga Blinkova, Eric P Nawrocki, Yuri Ostapchuck, Alejandro A Schäffer, and J Rodney Brister. Virus variation resource—improved response to emergent viral outbreaks. *Nucleic acids research*, 45(D1):D482–D490, 2017.
- Alexander Henzi, Johanna F. Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):963–993, 2021.
- Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. In *International Conference on Learning Representations*, 2019.
- Rafael Izbicki, Gilson T Shimizu, and Rafael B Stern. Flexible distribution-free conditional predictive bands using density estimators. *preprint at arXiv:1910.05575*, 2019.
- Tanqiu Jiang, Yi Li, Honghao Lin, Yisong Ruan, and David P Woodruff. Learning-augmented data stream algorithms. In *International Conference on Learning Representations*, 2019.
- Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- Lihua Lei and Emmanuel Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Dhruv Medarametla and Emmanuel Candès. Distribution-free conditional median inference. *Electronic Journal of Statistics*, 15(2):4625–4658, 2021.
- Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.
- Jayadev Misra and David Gries. Finding repeated elements. *Science of computer programming*, 2(2):143–152, 1982.

- Guillaume Pitel and Geoffroy Fouquier. Count-min-log sketch: Approximately counting with approximate counters. *arXiv preprint arXiv:1502.04885*, 2015.
- Jim Pitman and Marc Yor. The two parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25:855–900, 1997.
- Project Gutenberg. Project Gutenberg. www.gutenberg.org. Accessed: 2022-02-05.
- Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3538–3548, 2019.
- Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.
- Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.
- Stuart Schechter, Cormac Herley, and Michael Mitzenmacher. Popularity is everything: A new approach to protecting passwords from {Statistical-Guessing} attacks. In *5th USENIX Workshop on Hot Topics in Security (HotSec 10)*, 2010.
- Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34, 2021.
- Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alex Smola, and SVN Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10(11), 2009.
- James W. Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.
- Daniel Ting. Count-min: Optimal estimation and tight error bounds using empirical error distributions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2319–2328, 2018.
- Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.
- Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.
- Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *Annals of Statistics*, 37(3):1566–1590, 2009.
- Qingpeng Zhang, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, and C Titus Brown. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PloS one*, 9(7):e101271, 2014.

Appendix A. Algorithms

A.1. CMS-CU

Algorithm 1 CMS-CU

Input: Data set Z_1, \dots, Z_m . Sketch dimensions d, w . Hash functions h_1, \dots, h_d . Query z .

Initialize: $C_{j,k} = 0$ for all $j \in [d], k \in [w]$.

for $i = 1, \dots, m$ **do**

Compute $j^* = \arg \min_{j \in [d]} C_{j, h_j(Z_i)}$.

Increment $C_{j^*, h_{j^*}(Z_i)} \leftarrow C_{j^*, h_{j^*}(Z_i)} + 1$

end for

Compute $\hat{f}_{\text{up}}^{\text{CMS-CU}}(z) = \min_{j \in [d]} \{C_{j, h_j(z)}\}$.

Output: deterministic upper-bound for the frequency of z in the data set: $\hat{f}_{\text{up}}^{\text{CMS-CU}}(z)$.

A.2. Conformalized sketching

Algorithm 2 Conformalized sketching

Input: Data set Z_1, \dots, Z_m . Sketching function ϕ . Warm-up duration $m_0 \ll m$.
 A (trainable) rule for computing nested intervals $[\hat{L}_{m,\alpha}(\cdot; t), \hat{U}_{m,\alpha}(\cdot; t)]$, $t \in \mathcal{T}$.
 Number of data points $m_0^{\text{train}} < m_0$ used for training $[\hat{L}_{m,\alpha}(\cdot; t), \hat{U}_{m,\alpha}(\cdot; t)]$.
 A partition $\mathcal{B} = (B_1, \dots, B_L)$ of $\{0, \dots, m\}$ into L intervals.
 Random query Z_{m+1} . Desired coverage level $1 - \alpha \in (0, 1)$.
Initialize a sparse dictionary $f_{m_0}^{\text{wu}}(z) = 0, \forall z \in \mathcal{Z}$.
for $i = 1, \dots, m_0$ **do**
 Increment $f_{m_0}^{\text{wu}}(Z_i) \leftarrow f_{m_0}^{\text{wu}}(Z_i) + 1$.
end for
Initialize a sparse dictionary $f_{m-m_0}^{\text{sv}}(z) = 0, \forall z \in \mathcal{Z}$.
Initialize an empty sketch $\phi(\emptyset)$.
for $i = m_0 + 1, \dots, m$ **do**
 Update the sketch ϕ with the new observation Z_i .
 if $f_{m_0}^{\text{wu}}(Z_i) > 0$ **then**
 Increment $f_{m-m_0}^{\text{sv}}(Z_i) \leftarrow f_{m-m_0}^{\text{sv}} + 1$.
 end if
end for
Train $[\hat{L}_{m,\alpha}(\cdot; t), \hat{U}_{m,\alpha}(\cdot; t)]$ using the data in $\{(X_i, Y_i)\}_{i=1}^{m_0^{\text{train}}}$.
for $i = m_0^{\text{train}} + 1, \dots, m_0$ **do**
 Set $X_i = (Z_i, \phi(Z_{m_0+1}, \dots, Z_m))$ as in (12).
 Set $Y_i = f_{m-m_0}^{\text{sv}}(Z_i)$.
 Compute the conformity score $E(X_i, Y_i)$ with (5), using $[\hat{L}_{m,\alpha}(\cdot; t), \hat{U}_{m,\alpha}(\cdot; t)]$.
 Assign each score $E(X_i, Y_i)$ to an appropriate frequency bin $B \in \mathcal{B}$ based on Y_i .
end for
for $l = 1, \dots, L$ **do**
 Compute $\hat{Q}_{n_l, 1-\alpha}(B_l)$ as the $\lceil (1 - \alpha)(n_l + 1) \rceil$ smallest value among the n_l scores in bin B_l .
end for
Set $\hat{Q}_{n, 1-\alpha}^* = \max_l \hat{Q}_{n_l, 1-\alpha}(B_l)$.
Set $X_{m+1} = (Z_{m+1}, \phi(Z_{m_0+1}, \dots, Z_m))$ as in (12).
Output: a $(1 - \alpha)$ -level confidence interval

$$\left[f_{m_0}^{\text{wu}}(Z_{m+1}) + \hat{L}_{m,\alpha}(X_{m+1}; \hat{Q}_{n, 1-\alpha}^*), f_{m_0}^{\text{wu}}(Z_{m+1}) + \hat{U}_{m,\alpha}(X_{m+1}; \hat{Q}_{n, 1-\alpha}^*) \right]$$

for the unobserved frequency $f_m(Z_{m+1})$ of Z_{m+1} defined in (2).

A.3. Sampling from a Pitman-Yor predictive distribution

The data points are sampled sequentially from the following predictive distribution, which has parameters $\lambda > 0$ and $\sigma \in [0, 1)$. After sampling Z_1 from a standard normal distribution, $\mathcal{N}(0, 1)$, fix any $i \geq 1$ and let Z_1, \dots, Z_i indicate the data stream observed up to that point. Denote by k_i the number of distinct elements within it, and by $V_i = (V_{i,1}, \dots, V_{i,k_i})$ the set of such distinct values. Further, let $c_{i,l}$ indicate the number of times that object $V_{i,l}$ has been observed in Z_1, \dots, Z_i , for $l \in \{1, \dots, k_i\}$. Then, Z_{i+1} is generated as follows:

$$Z_{i+1} \mid Z_1, \dots, Z_i = \begin{cases} V_{i,l}, & \text{with probability } \frac{c_{i,l} - \sigma}{\lambda + i}, \text{ for } l \in \{1, \dots, k_i\}, \\ \mathcal{N}(0, 1), & \text{with probability } \frac{\lambda + k_i \sigma}{\lambda + i}. \end{cases}$$

Above, the second case which occurs with probability $(\lambda + k_i \sigma)/(\lambda + i)$ corresponds to sampling a new unique value from the standard normal distribution.

Appendix B. Mathematical proofs

B.1. Proof of Proposition 1

Proof Consider $((X_{\pi(1)}, Y_{\pi(1)}), \dots, (X_{\pi(m_0)}, Y_{\pi(m_0)}), (X_{\pi(m+1)}, Y_{\pi(m+1)}))$ for any permutation π of $\{1, \dots, m_0, m+1\}$. This is equal to $((X'_1, Y'_1), \dots, (X'_{m_0}, Y'_{m_0}), (X'_{m+1}, Y'_{m+1}))$, defined by applying the functions in (11)–(12) to a shuffled data set $Z_{\tilde{\pi}(1)}, \dots, Z_{\tilde{\pi}(m+1)}$, where $\tilde{\pi}$ indicates a permutation of $\{1, \dots, m+1\}$ that agrees with π on $\{1, \dots, m_0, m+1\}$ and leaves $\{m_0+1, \dots, m\}$ unchanged. Therefore,

$$\begin{aligned} & ((X_{\pi(1)}, Y_{\pi(1)}), \dots, (X_{\pi(m_0)}, Y_{\pi(m_0)}), (X_{\pi(m+1)}, Y_{\pi(m+1)})) \\ &= ((X'_1, Y'_1), \dots, (X'_{m_0}, Y'_{m_0}), (X'_{m+1}, Y'_{m+1})) \\ &\stackrel{d}{=} ((X_1, Y_1), \dots, (X_{m_0}, Y_{m_0}), (X_{m+1}, Y_{m+1})), \end{aligned}$$

where the last equality in distribution follows directly from the assumption that Z_1, \dots, Z_{m+1} are exchangeable. \blacksquare

B.2. Proof of Theorem 2

Proof The following notation will be helpful: let $B(Y_i) \in \mathcal{B}$ indicate the frequency bin into which Y_i belongs, for $i \in \{1, \dots, m_0, m+1\}$. We begin by proving the result for the simpler case in which Algorithm 2 is applied using conformity scores that do not require training, in which case $m_0^{\text{train}} = 0$. For $i \in \{1, \dots, m_0, m+1\}$, define the random variables Y_i and X_i as in (11)–(12), respectively. We already know from Proposition 1 that $(X_1, Y_1), \dots, (X_{m_0}, Y_{m_0}), (X_{m+1}, Y_{m+1})$ are exchangeable. This implies that the conformity scores $E(X_i, Y_i)$ are exchangeable with one another, for $i \in \{1, \dots, m_0, m+1\}$, because each of them only depends on X_i, Y_i and on the separate data points in the sketch $\phi(Z_{m_0+1}, \dots, Z_m)$. Therefore, E_{m+1} is also exchangeable with the subset of conformity scores with indices in $\{i \in \{1, \dots, m_0\} : B(Y_i) = B(Y_{m+1})\}$. Now, fix any bin $B^* \in \mathcal{B}$ and assume $B(Y_{m+1}) = B^*$. Now, note that the interval output by Algorithm 2 does not cover the true frequency $f_m(Z_{m+1})$ if and only if $E_{m+1} > \hat{Q}_{n, 1-\alpha} \geq \hat{Q}_{n_l, 1-\alpha}(B^*)$. However, a standard exchangeability argument for the conformity scores in $\{i \in \{1, \dots, m_0\} : B(Y_i) = B^*\}$ shows that $\mathbb{P}[E_{m+1} > \hat{Q}_{n_l, 1-\alpha}(B^*) \mid B(Y_{m+1}) = B^*] \leq 1 - \alpha$; for example, see Lemma 1 of Romano et al. (2019). This completes the first part of the proof. The second part with $m_0^{\text{train}} > 0$ follows very similarly: Proposition 1 implies that $(X_{m_0^{\text{train}}+1}, Y_{m_0^{\text{train}}+1}), \dots, (X_{m_0}, Y_{m_0}), (X_{m+1}, Y_{m+1})$ are exchangeable, and so must be the conformity scores E_i for $i \in \{m_0^{\text{train}}+1, \dots, m_0, m+1\}$ because each of them only depends on the corresponding X_i, Y_i and on the separate set of observations indexed by $\{1, \dots, m_0^{\text{train}}\}$, as well as on the sketch $\phi(Z_{m_0+1}, \dots, Z_m)$. The rest of the proof is exactly the same as in the first part because the empirical quantiles $\hat{Q}_{n_l, 1-\alpha}(B)$ are only computed on subsets of the data indexed by $\{m_0^{\text{train}}+1, \dots, m_0\}$. \blacksquare

Appendix C. Supplementary figures and tables

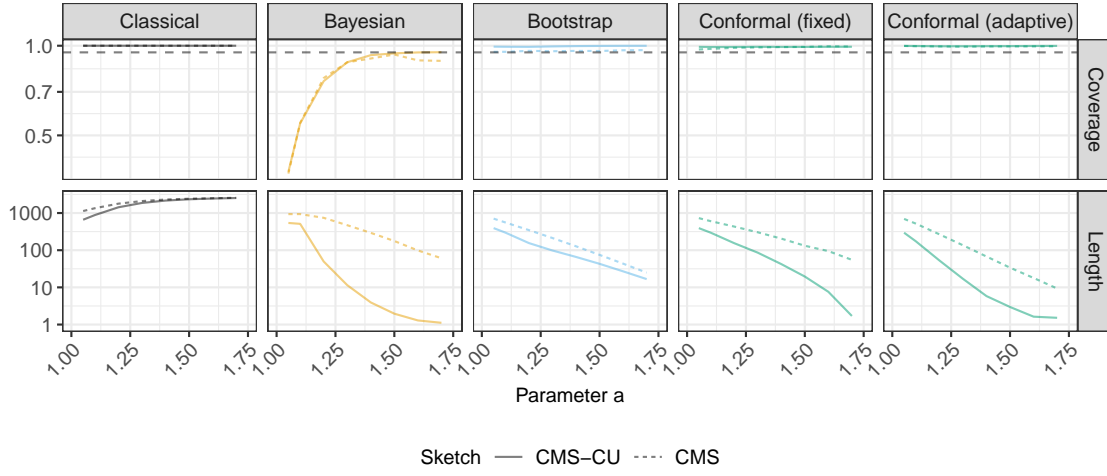


Figure A1: Performance of confidence intervals for random frequency queries, based on data sketched with either the vanilla CMS or the CMS-CU. Other details are as in Figure 1.

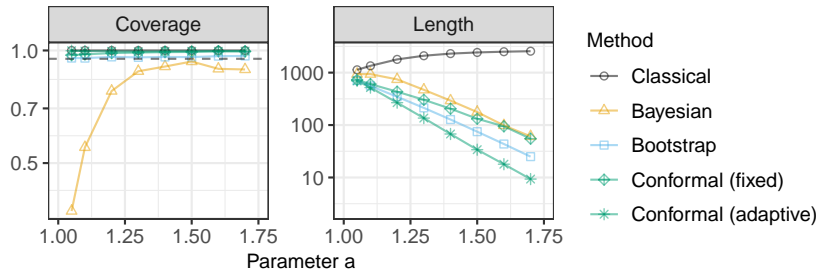


Figure A2: Performance of confidence intervals for random frequency queries, based on data sketched with the vanilla CMS. Other details are as in Figure 1.

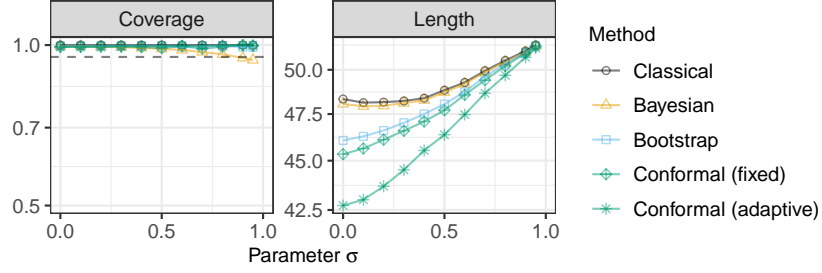


Figure A3: Empirical coverage and length of 95% confidence intervals for random frequency queries on a data set sampled from the predictive distribution of a Pitman-Yor process. The results are shown as a function of the Pitman-Yor process parameter σ . Other details are as in Figure 1.

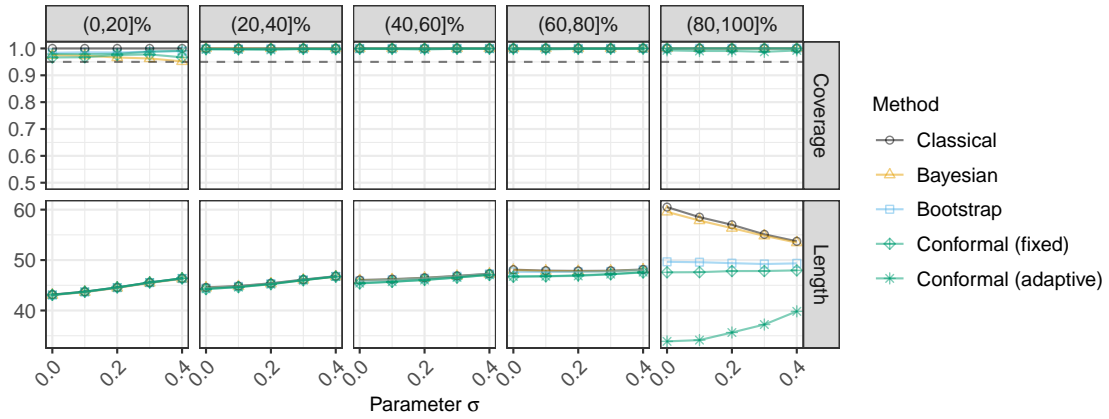


Figure A4: Performance of confidence intervals for random frequency queries, stratified by the quintile of the true query frequency. Other details are as in Figure A3.

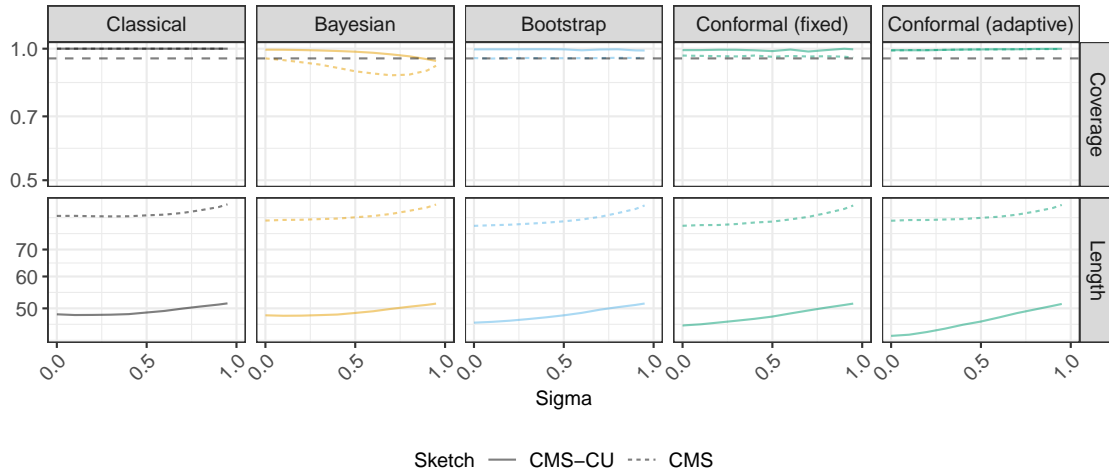


Figure A5: Performance of confidence intervals for random frequency queries, based on data sketched with either the vanilla CMS or the CMS-CU. Other details are as in Figure A3.

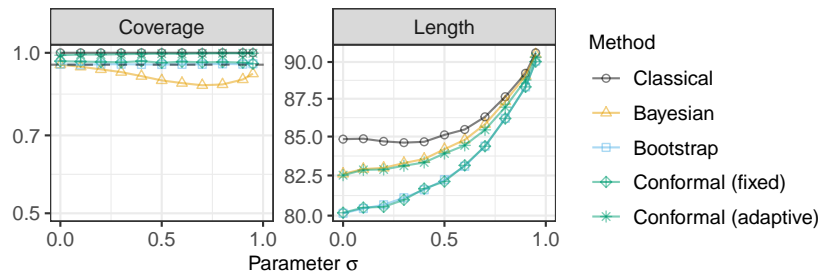


Figure A6: Performance of confidence intervals for random frequency queries, based on data sketched with the vanilla CMS. Other details are as in Figure A3.

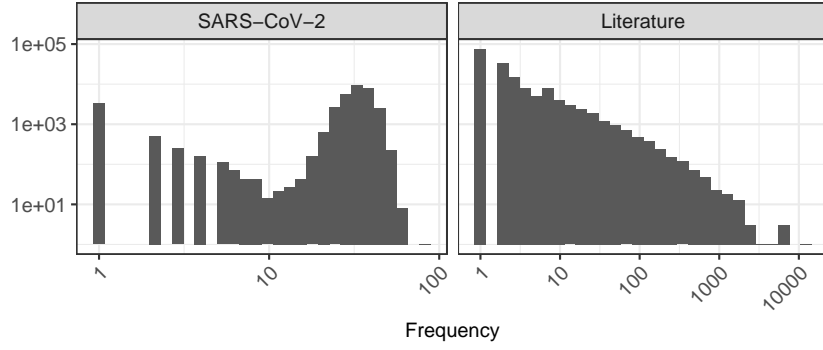


Figure A7: True frequency distribution in two sketched data sets. Left: sequenced SARS-CoV-2 DNA 16-mers. Right: English 2-grams in a corpus of classic English literature.

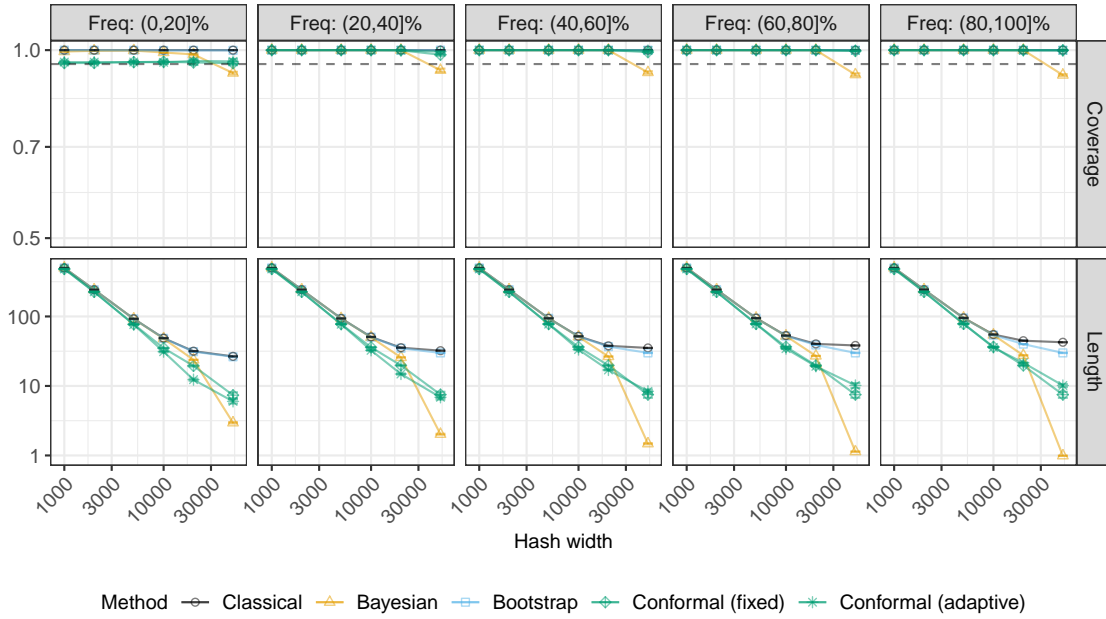


Figure A8: Performance of confidence intervals for random frequency queries, stratified by the quintile of the true query frequency. Other details are as in Figure 3.

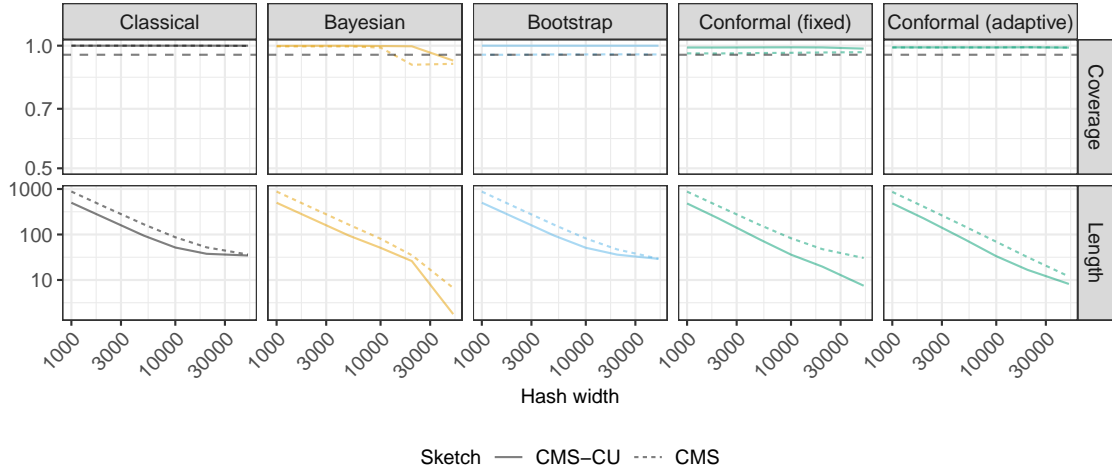


Figure A9: Performance of confidence intervals for random frequency queries, based on data sketched with either the vanilla CMS or the CMS-CU. Other details are as in Figure 3.

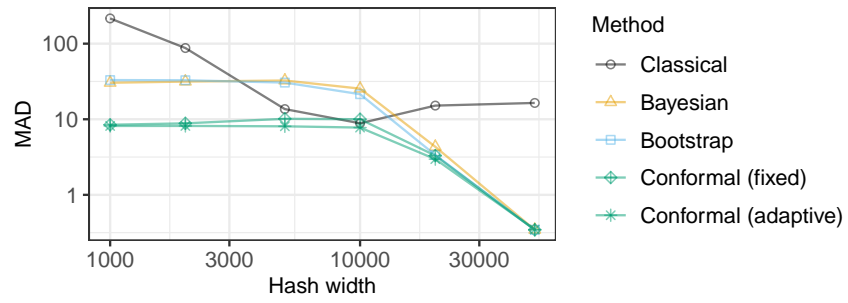


Figure A10: Median absolute deviation of point estimates for random frequency queries. Other details are as in Figure 3.

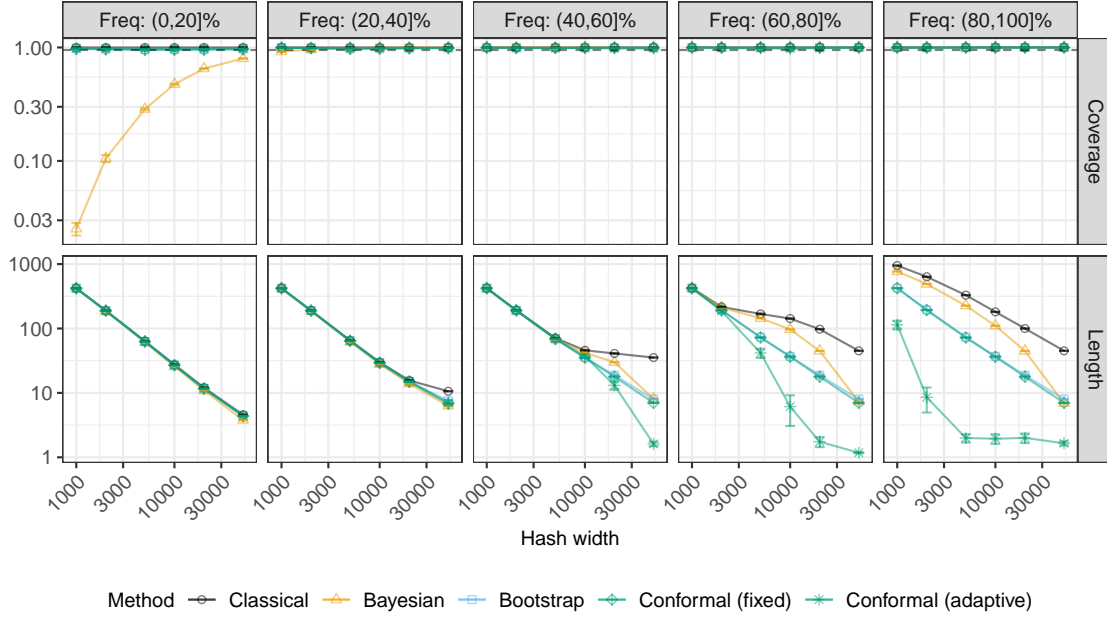


Figure A11: Performance of confidence intervals for random frequency queries, stratified by the quintile of the true query frequency. Other details are as in Figure 4.

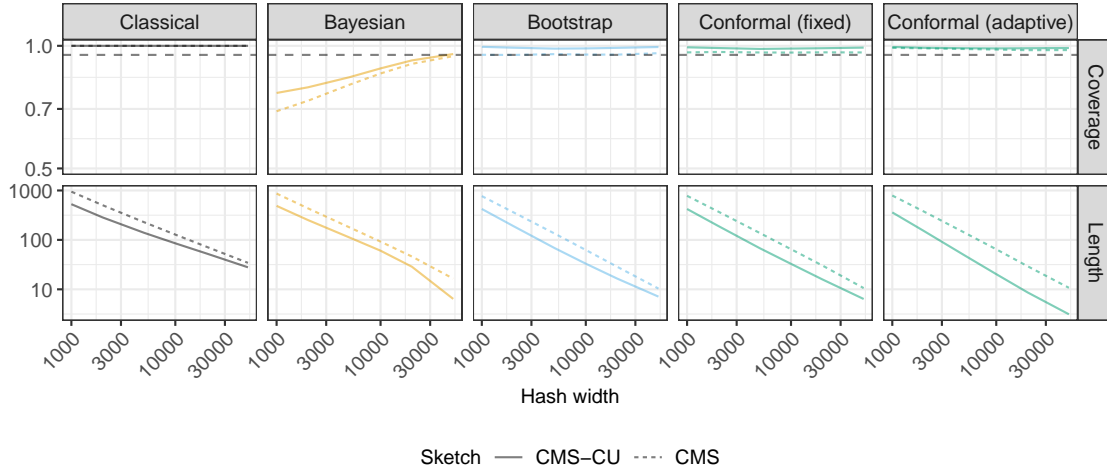


Figure A12: Performance of confidence intervals for random frequency queries, based on data sketched with either the vanilla CMS or the CMS-CU. Other details are as in Figure 4.

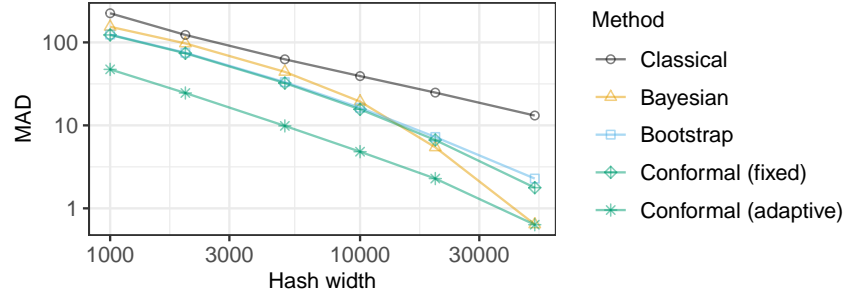


Figure A13: Median absolute deviation of point estimates for random frequency queries. Other details are as in Figure 4.

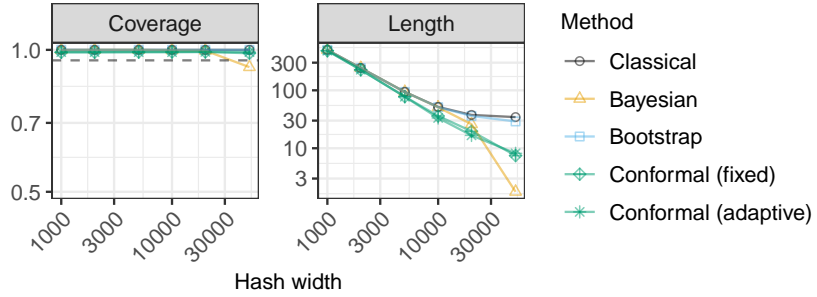


Figure A14: Performance of confidence intervals for random frequency queries, keeping only unique queries. The coverage is defined as the empirical proportion of unique queries whose frequency is correctly covered by the output confidence intervals. Other details are as in Figure 3.

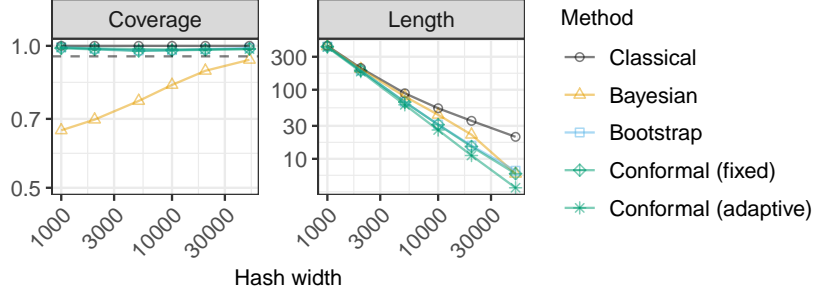


Figure A15: Performance of confidence intervals for random frequency queries, keeping only unique queries. The coverage is defined as the empirical proportion of unique queries whose frequency is correctly covered by the output confidence intervals. Other details are as in Figure 4.

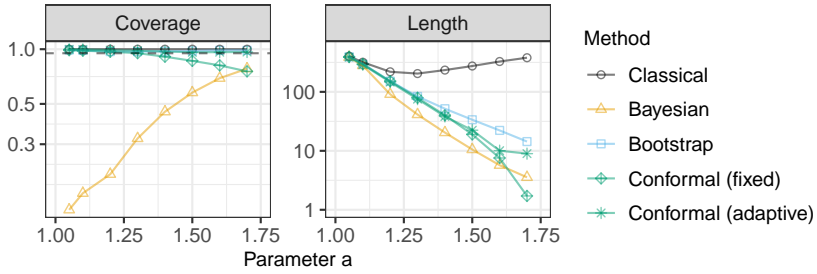


Figure A16: Performance of confidence intervals for random frequency queries, keeping only unique queries. The coverage is defined as the empirical proportion of unique queries whose frequency is correctly covered by the output confidence intervals. Other details are as in Figure 1.

Table A1: True frequencies, deterministic upper bounds, and 95% lower bounds for some a few random queries in two sketched data sets. Sketching with CMS-CU with $w = 50,000$. Lower bounds written in green are below the true frequency; those in red are above. For each query, the highest lowest bound below the true frequency is highlighted in bold.

			95% Lower bound				
Data	Frequency	Upper bound	Classical	Bayesian	Bootstrap	Conformal	
						Fixed	Adaptive
SARS-CoV-2							
AATTATTATAAGAAAG	81	81	26	81	52	50	36
TCAGACAAC TACTATT	76	76	21	55	47	45	32
AAAGTTGATGGTGTTG	73	73	18	59	44	42	31
CAATTATTATAAGAAA	63	63	8	48	34	32	26
ATCAGACAAC TACTAT	60	60	5	44	31	29	26
ACCTTTGACAATCTTA	55	55	0	52	26	24	27
ATTTGAAGTCACCTAA	55	55	0	55	26	24	27
CATGCAAATTACATAT	54	54	0	54	25	23	26
GAATTTACAGTATTC	54	54	0	54	25	23	27
TTTGTAGAAAACCCAG	53	53	0	53	24	22	27
AGTTGCAGAGTGGTTT	24	24	0	13	0	0	20
TCTTCACAATTGGAAC	24	24	0	12	0	1	20
TTCTGCTCGCATAGTG	24	24	0	12	0	0	20
CTACTTTAGATTCGAA	23	23	0	11	0	0	19
GCTGGTGTCTCTATCT	23	23	0	23	0	1	19
TTCTAAGAAGCCTCGG	23	24	0	14	0	0	20
GGGCTGTTGTTCTTGT	22	24	0	12	0	0	20
ACGTTTCGTGTTGTTTT	20	20	0	20	0	0	16
GAAGTCTTTGAATGTG	20	20	0	20	0	0	16
CAAACCTGGTAATTTT	3	3	0	3	0	0	0
Literature							
of the	12565	12568	12513	12544	12557	12556	12562
in the	6188	6190	6135	6169	6179	6179	6180
and the	6173	6175	6120	6151	6164	6164	6165
the of	6015	6017	5962	5990	6006	6006	6007
the lord	4186	4195	4140	4165	4184	4184	4184
to the	3465	3467	3412	3445	3456	3456	3463
the and	2250	2251	2196	2227	2240	2240	2248
all the	2226	2230	2175	2207	2219	2219	2224
and he	2169	2173	2118	2153	2162	2162	2167
to be	2062	2064	2009	2043	2053	2053	2060
man on	22	29	0	10	18	18	18
their hand	22	24	0	9	13	13	0
no need	20	28	0	9	17	17	16
and brother	12	14	0	2	3	3	0
miss would	10	13	0	3	2	2	0
i please	8	12	0	3	1	1	1
also how	3	13	0	2	2	2	0
in under	3	9	0	2	0	0	0
ten old	3	6	0	1	0	0	0
fault he	1	9	0	1	0	0	0

Table A2: True frequencies, upper and lower bounds for a few random queries in two sketched data sets. Hash width $w = 50,000$. Other details are as in Table A1.

			95% Lower bound				
Data	Frequency	Upper bound	Classical	Bayesian	Bootstrap	Conformal	
						Fixed	Adaptive
SARS-CoV-2							
AATTATTATAAGAAAG	81	209	0	4	0	0	18
TCAGACAAC TACTATT	76	213	0	8	0	0	18
AAAGTTGATGGTGTG	73	130	0	2	0	1	18
CAATTATTATAAGAAA	63	233	0	4	11	6	19
ATCAGACAAC TACTAT	60	179	0	2	0	0	18
ACCTTTGACAATCTTA	55	292	0	15	70	67	22
ATTTGAAGTCACCTAA	55	258	0	11	36	31	20
CATGCAAATTACATAT	54	204	0	3	0	0	18
GAATTTACAGTATTC	54	260	0	12	38	35	22
TTTGTAGAAAACCCAG	53	246	0	7	24	18	20
ATGCTGCAATCGTGCT	24	139	0	2	0	0	17
ATTTCTTAATATTACA	24	92	0	1	0	0	17
CTCTATCATTATTGGT	24	121	0	1	0	0	17
TGTTTTATTCTCTACA	24	199	0	3	0	1	19
CAGTACATCGATATCG	23	119	0	2	0	0	17
TAATGGTGACTTTTTG	23	92	0	1	0	0	17
CAACCATAAAACCAGT	22	105	0	1	0	0	17
AGTTATTTGACTCCTG	21	97	0	1	0	1	18
ATAAAGGAGTTGCACC	19	218	0	5	0	0	18
Literature							
of the	12565	12630	12086	12325	12463	12454	12563
in the	6188	6242	5698	5906	6075	6067	6096
and the	6173	6314	5770	5972	6147	6139	6169
the of	6015	6162	5618	5834	5995	5985	6014
the lord	4186	4289	3745	3975	4122	4114	4185
to the	3465	3558	3014	3217	3391	3380	3464
the and	2250	2413	1869	2081	2246	2237	2249
all the	2226	2346	1802	1993	2179	2170	2225
and he	2169	2293	1749	1937	2126	2117	2168
to be	2062	2121	1577	1770	1954	1945	2061
very for	15	59	0	2	0	0	0
and faithful	14	94	0	3	0	0	0
but found	9	74	0	2	0	0	0
my speech	6	98	0	3	0	0	0
of eight	5	66	0	2	0	0	0
and soul	4	140	0	6	0	0	0
her prow	3	79	0	2	0	0	0
usual as	2	56	0	2	0	0	0
a invitation	1	80	0	2	0	0	0
angular log	0	146	0	5	0	0	0