How Many Subpopulations Is Too Many?: Exponential Lower Bounds for Inferring Population Histories

How Many Subpopulations is Too Many? Exponential Lower Bounds for Inferring Population Histories[*]

Younhun Kim[†] Frederic Koehler[‡] Ankur Moitra[§] Elchanan Mossel[¶]
Govind Ramnarayan[‖]

Abstract

Reconstruction of population histories is a central problem in population genetics. Existing coalescent-based methods, like the seminal work of Li and Durbin (*Nature*, 2011), attempt to solve this problem using sequence data but have no rigorous guarantees. Determining the amount of data needed to *correctly reconstruct* population histories is a major challenge. Using a variety of tools from information theory, the theory of extremal polynomials, and approximation theory, we prove new sharp information-theoretic lower bounds on the problem of reconstructing *population structure* — the history of multiple subpopulations that merge, split and change sizes over time. Our lower bounds are *exponential* in the number of subpopulations, even when reconstructing recent histories. We demonstrate the sharpness of our lower bounds by providing algorithms for distinguishing and learning population histories with matching dependence on the number of subpopulations. Along the way and of independent interest, we essentially determine the optimal number of samples needed to learn an exponential mixture distribution information-theoretically, proving the upper bound by analyzing natural (and efficient) algorithms for this problem.

[*] *Keywords*: Population History, Coalescent Theory, Hyperexponential/Exponential Mixture Distributions, Parameter Estimation, Sample Complexity, Data Requirement.

[†] Massachusetts Institute of Technology. Department of Mathematics. Email: younhun@mit.edu.

[‡] Massachusetts Institute of Technology. Department of Mathematics. Email: fkoehler@mit.edu. Research is partially supported by NSF Large CCF-1565235 and Ankur Moitra's David and Lucile Packard Fellowship.

[§] Massachusetts Institute of Technology. Department of Mathematics and CSAIL. Email: moitra@mit.edu. This work was supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship, an Alfred P. Sloan Fellowship, and an ONR Young Investigator Award.

[¶] Massachusetts Institute of Technology. Department of Mathematics and IDSS. Email: elmos@mit.edu. Partially supported by awards ONR N00014-16-1-2227, NSF CCF1665252 and DMS-1737944.

[‖] Massachusetts Institute of Technology. CSAIL. Email: govind@mit.edu. Partially supported by awards NSF CCF1665252 and DMS-1737944.

1 Introduction

1.1 Background: inference of population size history

A central task in population genetics is to reconstruct a species' *effective population size* over time. Coalescent theory [22] provides a mathematical framework for understanding the relationship between effective population size and genetic variability. In this framework, observations of present-day genetic variability — captured by DNA sequences of individuals — can be used to make inferences about changes in population size over time.

There are many existing methods for estimating the size history of a *single population* from *sequence data*. Some rely on Maximum Likelihood methods [15, 25, 26, 27] and others utilize Bayesian inference [5, 11, 21] along with a variety of simplifying assumptions. A well-known work of Li and Durbin [15] is based on using sequence data from just a single human (a single pair of haplotypes) and revolves around the assumption that coalescent trees of alleles along the genome satisfy a certain conditional independence property [17]. By and large, methods such as these do not have any associated provable guarantees. For example, Expectation-Maximization (EM) is a popular heuristic for maximizing the likelihood but can get stuck in a local maximum. Similarly, Markov Chain Monte Carlo (MCMC) methods are able to sample from complex posterior distributions if they are run for a long enough time, but it is rare to have reasonable bounds on the mixing time. In the absence of provable guarantees, simulations are often used to give some sort of evidence of correctness.

Under what sorts of conditions is it possible to infer a single population history? Kim, Mossel, Rácz and Ross [14] gave a strong lower bound on the number of samples needed even when one is given exact coalesence data. In particular, they showed that the number of samples must be at least exponential in the number of generations. Thus there are serious limitations to what kind of information we can hope to glean from (say) sequence data from a single human individual. In a sense, their work provides a quantitative answer to the question: *How far back into the past can we hope to reliably infer population size, using the data we currently have?* We emphasize that although they work in a highly idealized setting, this only makes their problem easier (e.g. assuming independent inheritance of loci along the genome and assuming that there are no phasing errors) and thus their lower bounds more worrisome.

1.2 Our setting: inference of multiple subpopulation histories

A more interesting and challenging task is the reconstruction of *population structure*, which refers to the sub-division of a single population into several subpopulations that merge, split, and change sizes over time. There are two well-known works that attack this problem using coalescent-based approaches. Both use sequence data to infer population histories where present-day subpopulations were formed via divergence events of a single, ancestral population in the distant past. The first is Schiffels and Durbin [25], who used their method to infer the population structure of nine human subpopulations up to about 200,000 years into the past. More recently, Terhorst, Kamm and Song [27] inferred population structures of up to three human subpopulations. Just as in the single population case, these methods do not come with provable guarantees of correctness due to the simplifying assumptions they invoke and the heuristics they employ.

As for theoretical work, the lower bounds proven for a single population trivially carry over to the setting of inferring population structure. However, the lower bound in [14] only applies when we are trying to reconstruct events in the distant past, leading us to a natural question: can we infer *recent* population structure, but, when there are multiple subpopulations?

In this paper, we establish strong limitations to inferring the population sizes of multiple subpopulation histories using pairwise coalescent trees. We prove sample complexity lower bounds that are *exponential in the number of subpopulations*, even for reconstructing recent histories. Our results provide a quantitative answer to the question, *Up to what granularity of dividing a population into multiple subpopulations, can we hope to reliably infer population structure?*

Our methods incorporate tools from information theory, approximation theory, and analysis (from [29]). To complement our lower bounds, we also give an algorithm for hypothesis testing based on the celebrated Nazarov-Turán lemma [20]. Our upper and lower bounds match up to constant factors and establish sharp bounds for the number of samples needed to distinguish between two known population structures as a function of the number of subpopulations. Finally, for the more general problem of learning the population structure (as opposed to testing which of two given population structures is more accurate) we give an algorithm with provable guarantees based on the Matrix Pencil Method [12] from signal processing. We elaborate on our results in Section 1.4.

1.3 Modeling Assumptions

Our results will apply under the following assumptions: (1) individuals are haploids[1], (2) the genome can be divided into *known* allelic blocks that are inherited independently and (3) for each pair of blocks, we are given the exact coalescence time. Indeed, in practice, one must start with sequenced genomes — and in the context of recovering events in human history, (potentially unphased) genotypes of diploid individuals. The problem of recovering coalescence times from sequences provides a major challenge and often requires one to either know the population history beforehand, or leverage simultaneous recovery of history and coalescence times using various joint models that enable probabilistic inference.

But since the main message of our paper is a *lower bound* on the number of exact pairwise coalescent samples needed to recover population history, in practice it would only be harder. Even in our idealized setting, handling 7 or 8 subpopulations already requires more data than one could reasonably be assumed to possess. Thus, our work provides a rather direct challenge to empirical work in the area: Either results with 7 or 8 subpopulations are not to be trusted or there must be some biological reason why the types of population histories that arise in our lower bounds, that are information-theoretically impossible to distinguish from each other using too few samples, can be ruled out.

1.3.1 The Multiple-Subpopulation Coalescent Model

Consider a panmictic haploid[2] population, such that each subpopulation evolves according to the standard Wright-Fisher dynamics[3] — we direct the reader to [3] for an overview. For simplicity, we assume no admixture between distinct subpopulations as long as they are separated in the model (i.e. they have not merged into a single population in the time period under consideration).

As a reminder, if one assumes that a single population has size N which is large and constant throughout time, then the time to the most recent common ancestor (TMRCA) of two randomly sampled individuals closely follows the Kingman coalescent [3] with exponential rate N:

$$\Pr(T > t) = \exp(-t/N). \tag{1}$$

[1] Alternatively, diploids whose phasing is provided.

[2] In a diploid population, the exponents are scaled by a constant factor 2. This can be handled easily via scaling and therefore makes little difference in the analysis.

[3] The distinction between the Wright-Fisher and Moran models is of no consequence in this work, as the latter yields the same exponential model in the diffusion limit [3].

where T, the coalescence time for two randomly chosen individuals, is measured in **generations**. Henceforth, we will assume that this is the distribution of T in the single-component case.

If instead we have a population which is partitioned into a collection of distinct subpopulations with non-constant sizes, let $\mathbf{N}(t)$ be the function that describes the sub-population sizes over time. As in [14], we will assume that the function $\mathbf{N}(t)$ is piecewise constant with respect to some unknown collection of intervals $I_1, I_2, \ldots$ partitioning the real line. In particular, for each $t \in I_k$, there is an associated vector of effective subpopulation sizes $\mathbf{N}(t) = (N_1^{(k)}, \ldots, N_{D_k}^{(k)})$, indexed by the D_k subpopulations present at time t. The indexing need not be consistent across different intervals, as their semantic meaning will change as subpopulations merge and split. For example, $N_1^{(k)}$ and $N_1^{(k+1)}$ need not always represent the sizes of the same subpopulation.

Consider the case where $\mathbf{N}(t)$ is constant for all $t \in I = [a, b]$, where $0 < a < b$, with no admixture and no migration in-between subpopulations in the time interval I. In this case, the coalescence time follows the law of a convex combination of exponential functions:

$$\Pr(T > a + t \mid T > a) = \sum_{\ell=0}^{D} p_\ell e^{-\lambda_\ell t} \tag{2}$$

where $p_0 + p_1 + \cdots + p_D = 1$, $\lambda_0 = 0$ and the other λ_i are $\frac{1}{N_i}$ (refer to Appendix A for a more careful treatment).

The population structure is assumed to undergo changes over time, where the positive direction points towards the past. The three possible changes are:

1. (Split) One subpopulation at time t^- becomes two subpopulations at time t (i.e. $D_k = D_{k-1} + 1$).

2. (Merge) Two subpopulations at time t^- join to form one subpopulation at time t (i.e. $D_k = D_{k-1} - 1$).

3. (Change Size) An arbitrary number of subpopulations change size at time t.

Figure 1 provides an illustrative example. If an individual at time t^- is from a subpopulation of size M which splits into two subpopulations of sizes M_1, M_2 at time t, then its ancestral subpopulation is random: for $i \in \{1, 2\}$, subpopulation i is chosen with probability M_i/M. In our model, we only allow at most one of these events at any particular time point. For us, a "split" looking backward in time refers to a *convergence* event of two subpopulations going *forward* in time, while a "merge" refers to a divergence event. This convention is chosen because we think of reconstruction as proceeding backwards in time from the present.

1.4 Our Results

The main theoretical contribution of this work is an essentially tight bound on the sample complexity of learning population history in the multiple-subpopulation model. In particular, we show sample complexity lower bounds which are *exponential in the number of subpopulations k*. Here is an organized summary of our results:

- First, we show a two-way relationship between the problem of learning a population history (in our simplified model) and the problem of learning a mixture of exponentials. Recall that when the effective subpopulation sizes are all constant, the distribution of coalescence times follows Eq. (2) and thus is equivalent to learning the parameters p_t and λ_t in a mixture of exponentials. Conversely, we show how to use an algorithm for learning mixtures of

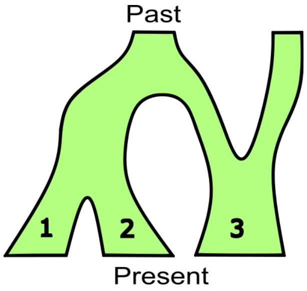

Figure 1: An example of population structure history, illustrating merges and splits starting with three present-day subpopulations.

exponentials to reconstruct the entire population history by locating the intervals where there are no genetic events and then learning the associated parameters in each, separately. (Section 2.1 with details in Appendix B and Appendix F.)

- *(Main Result)* Using this equivalence, we show an information-theoretic lower bound on the sample complexity that applies regardless of what algorithm is being used. In particular, we construct a pair of population histories that have different parameters but which require $\Omega((1/\Delta)^{4k})$ samples in order to tell apart. This lower bound is *exponential in the number of subpopulations* k. Here, $\Delta \leq 1/k$ is the smallest gap between any pair of the λ_t's. The proof of this result combines tools spanning information theory, extremal polynomials, and approximation theory. (Section 2.3 with details in Appendix D.)

- In the *hypothesis testing* setting where we are given a pair of population histories that we would like to use coalescence statistics to distinguish between, we give an algorithm that succeeds with only $O((1/\Delta)^{4k})$ samples. The key to this result is a powerful tool from analysis, the Nazarov-Turán Lemma [20] which lower bounds the maximum absolute value of a sum of exponentials on a given interval in terms of various parameters. This result matches our lower bounds, thus resolving the sample complexity of hypothesis testing up to constant factors. (Section 2.4 with details in Appendix E)

- In the *parameter learning* setting when we want to directly estimate population history from coalescence times, we give an efficient algorithm which provably learns the parameters of a (possibly truncated) mixture of exponentials given only $O((1/\Delta)^{6k})$ samples. We accomplish this by analyzing the *Matrix Pencil Method* [12], a classical tool from signal processing, in the real-exponent setting. (Section 2.2 with details in Appendix C.)

- Finally, we demonstrate using simulated data that our sample complexity lower bounds really do place serious limitations on what can be done in practice. From our plots it is easy to see that the sample complexity grows exponentially in the number of subpopulations even in the optimistic case where the separation $\Delta = 1/k$ which minimizes our lower bounds. In particular, the number of samples we would need very quickly exceeds the number of functionally relevant genes (on the order of 10^4) and even the number of SNPs available in the human genome (on the order of 10^7). In fact, through a direct numerical analysis of our chosen instances, we can give even stronger sample-complexity lower bounds (Section 3, with details in Appendix G.2).

Discussion of Results: In summary, this work highlights some of the fundamental difficulties of reconstructing population histories from pairwise coalescence data. Even for recent histories, the lower bounds grow exponentially in the number of subpopulations. Empirically, and in the absence of provable guarantees, and even with much noisier data than we are assuming, many works suggest that it is possible to reconstruct population histories with as many as nine subpopulations. While testing out heuristics on real data and assessing the biological plausibility of what they find is important, so too is delineating sharp theoretical limitations. Thus we believe that our work is an important contribution to the discussion on reconstructing population histories. It points to the need for the methods that are applied in practice to be able to justify why their findings ought to be believed. Moreover they need to somehow preclude the types of population histories that arise in our lower bounds and are genuinely impossible to distinguish between given the finite amount of data we have access to.

1.5 Related works

As mentioned in Section 1.1, existing methods that attempt to empirically estimate the population history of a *single population* from *sequence data* generally fall into one of two categories: Many are based on (approximately) maximizing the likelihood [15, 25, 26, 27] and others perform Bayesian inference [5, 11, 21]. Generally, they are designed to recover a *piecewise constant* function $N(t)$ that describes the size of a population, with the goal of accurately summarizing divergence events, bottleneck events and growth rates throughout time.

Many notable methods that fall into the first category rely on Hidden Markov Models (HMMs), which implicitly make a Markovian assumption on the coalescent trees of alleles across the genome. One notable work is Li and Durbin [15], which gave an HMM-based method (PSMC) that reconstructs the population history of a single population using the genome of a single diploid individual. Later related works gave alternative HMMs that incorporate more than two haplotypes (diCal [26] and MSMC [25]) and improve robustness under phasing errors (SMC++ [27]).

Methods in the second category operate under an assumption about the probability distribution of coalescence events and the observed data. For instance, Drummond [5] prescribes a prior for the distribution of coalescence trees and population sizes, under which MCMC techniques are used to compute both an output and a corresponding 95% credibility interval. However, given the highly idealized nature of their models and the limitations of their methodology (for example, there is no guarantee their MCMC method has actually mixed), it is unclear whether the ground truth actually lies in those credibility intervals.

In the *multiple subpopulations* case, there are two major coalescent-based methods. The first is Schiffels and Durbin [25], which introduced the MSMC model as an improvement over PSMC. These authors used their method to infer the population history of *nine human subpopulations up to about 200,000 years into the past.* Terhorst, Kamm and Song [27] introduced a variant (SMC++) that was directly designed to work on genotypes with missing phase information. In particular, they demonstrate the potential dangers of relying on phase information, by showing that MSMC is sensitive to such errors. In an experiment, SMC++ was used to perform inference of population histories of various combinations of *up to three human subpopulations.* In these experiments, individuals are purposefully chosen from specific subpopulations. We emphasize that in our model, due to the presence of population merges *and* splits, one does not always know what subpopulation an ancestral individual is from.

As a side remark, there are approaches that attempt to infer a (single-component) population history using different types of information. We briefly touch upon some of these known works. One alternative strategy is to use the site frequency spectrum (SFS), e.g. [2, 7]. The earliest theoretical result regarding SFS-based reconstruction is due to Myers, Fefferman and Patterson [19], who proved that *generic* 1-component population histories suffer from unidentifiability issues. Their lower bound constructions have a caveat: They are pathological examples of oscillating functions which are unlikely to be observed in a biological context. Later works [1, 28] prove both identifiability and lower bounds for reconstructing *piecewise constant* population histories using information from the SFS. (In contrast, as our algorithms show, reconstruction from coalescence data does not suffer from the same lack of identifiability issues.)

Most recently, Joseph and Pe'er [13] developed a Bayesian time-series model that incorporates data from *ancient DNA* to recover the history for multiple subpopulations only under size changes, without considering merges or splits. While our analysis does not directly account for such data, the necessity of considering such models is consistent with our assertion: extra information about the ground truth, such as directly observable information about the past (e.g. ancestral DNA), is probably *required* in order for the problem to even be information-theoretically feasible. In addition, [13] does not solve for subpopulation *sizes*, but rather subpopulation *proportions*, which contains less information than what we are after.

2 Theoretical Discussion

2.1 Reductions between mixtures of exponentials and population history

In the rest of our theoretical analysis, we will focus on the mixture of exponentials viewpoint of population history. To justify this, note that if we can learn truncated mixtures of exponentials, then we can easily learn population history. Details are given in Appendix F including a concrete algorithm based on our analysis of the Matrix Pencil Method. Conversely, we observe that an arbitrary mixture of exponentials can be embedded as a sub-mixture of a simple population history with two time periods, so that recovering the population history requires in particular learning the mixture of exponentials. The following theorem makes this precise; its proof is delegated to Appendix B.

Theorem 2.1. *Let P with $P(T > t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t}$ be a the distribution of an arbitrary mixture of k exponentials (over random variable T) with all $\lambda_i > 0$. Then for any $t_0 > 0$, there exists a two-period population history with k populations which induces a distribution Q on coalescence times such that*

$$Q(T > t + t_0 | T \neq \infty, T > t_0) = P(T > t).$$

Remark 1. By choosing a small value for t_0, we ensure that very few coalesence times occur in the more recent period, so that the reconstruction algorithm must rely on the information from the second (less recent) period with our planted mixture of exponentials.

Additionally, we provide a more sophisticated version of this reduction which maps two mixtures of exponentials to different population histories simultaneously, while preserving statistical indistinguishability.

Theorem 2.2. *Let P with $P(T > t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t}$ and Q with $Q(T > t) = \sum_{j=1}^{\ell} q_j e^{-\mu_j t}$ be arbitrary mixtures of exponentials with all $\lambda_i, \mu_j > 0$. Then for all $t_0 > 0$ sufficiently small there exist two distinct 2-period population histories R with $k + 2$ subpopulations and S with $\ell + 2$ subpopulations such that:*

1. *$R[T > t + t_0 | T \neq \infty, T > t_0] = P(T > t)$ and $S[T > t + t_0 | T \neq \infty, T > t_0] = Q(T > t)$*
2. *$R[T = t_0] = S[T = t_0]$ and $R[T = \infty] = S[T = \infty]$.*

Again, if we take t_0 small enough, we ensure that any distinguishing algorithm must rely on information from the second (less recent) period, and hence because the probability of all other events match, must distinguish between the mixtures of exponentials Q and R.

2.2 Guaranteed recovery of exponential mixtures via the Matrix Pencil Method

Given samples from a hyperexponential distribution

$$\Pr(T > t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t}, \tag{3}$$

can we learn the parameters $p_1, \ldots, p_k$, $\lambda_1, \ldots, \lambda_k$? In Section 2.1, we established the equivalence between solving this problem and learning population history. Suppose for now that we are given access to the exact values of probabilities $v_t := \Pr(T \geq t)$ for $t \in \mathbb{R}_{\geq 0}$, i.e. $v_t = \sum_{j=1}^{k} p_i \alpha_i^t$, where $\alpha_i = e^{-\lambda_i}$. The *Matrix Pencil Method* is the following linear-algebraic method, originating in the signal processing literature [12], which solves for the parameters $\{p_i, \lambda_i\}_{i=1}^{k}$:

1. Let A, B be $k \times k$ matrices where $A_{ij} = v_{i+j-1}$ and $B_{ij} = v_{i+j-2}$.
2. Solve the generalized eigenvalue equation $\det(A - \gamma B) = 0$ for the pair (A, B). The γ which solve $\det(A - \gamma B) = 0$ are the α's.
3. Finish by solving for the p's in a linear system of equations $\vec{v} = V\vec{p}$, where $\vec{v} = (v_0, \ldots, v_{k-1})$, V is the $k \times k$ Vandermonde matrix generated by $\alpha_1, \ldots, \alpha_k$ and $\vec{p}$ is the vector of unknowns $(p_1, \ldots, p_k)$.

To understand why the algorithm works in the noiseless setting, consider the decomposition $A = VD_pD_\alpha V^T$ and $B = VD_pV^T$ where $V = V_k(\alpha_1, \ldots, \alpha_k)$ is the $k \times k$ Vandermonde matrix whose (i, j) entry is α_j^{i-1}, $D_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_k)$ and $D_p = \mathrm{diag}(p_1, \ldots, p_k)$. Then it's clear that the α_i are indeed the generalized eigenvalues of the pair (A, B). However, in our setting, we do not have access to the exact measurements v_t, but instead have *noisy* empirical measurements $\tilde{v}_t$; in practice, the output of the MPM can be very sensitive to noise.

The Matrix Pencil Method is a close cousin of *Prony's Method* [23]. Prior to this work, Feldmann et al. [8] considered the strategy of using Prony's Method to *fit* exponential mixtures to general long-tail distributions. In the upcoming section, we provide a finite-sample guarantee of the MPM in the context of *learning* exponential mixture distributions. As it turns out (Remark 3 and Remark 4), this algorithm is nearly optimal in terms of the number of samples required.

2.2.1 Analysis of MPM under noise

We now describe our analysis of the MPM in the more realistic setup where the CDF is estimated from sample data. First note that the model (Equation 3) is statistically unidentifiable if there exist two identical λ's. Indeed, the mixture $\frac{1}{2}e^{-\lambda t} + \frac{1}{2}e^{-\lambda t}$ is exactly same as the single-component model $e^{-\lambda t}$, as is any other re-weighting of the coefficients into arbitrarily many components with exponent λ. Therefore it is natural to introduce a *gap parameter* $\Delta := \min_{i \neq j} |\lambda_i - \lambda_j|$ which is required to be nonzero, as in the work on super-resolution (e.g. [4, 18]).

Without loss of generality, we also assume that: (1) the components are sorted in decreasing order of exponents, so that $\lambda_1 > \cdots > \lambda_k > 0$, and (2) time has been re-scaled[4] by a constant factor, so that $\lambda_i \in (0, 1)$ for each i. Now we can state our guarantee for the MPM under noise:

[4] In practice, even if this scaling is unknown, this is easily handled by e.g. trying powers of 2 and picking the best result in CDF distance, for instance $\|F - G\|_\infty = \sup_t |F(t) - G(t)|$.

Theorem 2.3. *Let $\Delta = \min_{i \neq j} |\lambda_i - \lambda_j|$ and let $p_{\min} = \min_i p_i$. For all $\delta > 0$, there exists $N_0 = O\left(\frac{k^{10}}{p_{\min}^4}\left(\frac{2e}{\Delta}\right)^{6k}\log\frac{1}{\delta}\right)$ such that, with probability $1-\delta$, using empirical estimates $\widetilde{v}_0, \ldots, \widetilde{v}_{2k-1}$ from $N \geq N_0$ samples, the matrix pencil method outputs $\{(\widetilde{\lambda}_j, \widetilde{p}_j)\}_{j=1}^k$ satisfying*

$$|\widetilde{\lambda}_j - \lambda_j| = O\left(\frac{k^{3.5}}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k}\sqrt{\frac{1}{N}\log\frac{1}{\delta}}\right) \quad \textit{and} \quad |\widetilde{p}_j - p_j| = O\left(\frac{k^5}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{3k}\sqrt{\frac{1}{N}\log\frac{1}{\delta}}\right)$$

for all j.

Remark 2. Letting α_i denote $e^{-\lambda_i}$, we note that we can equivalently focus on learning the α_i's, and that guarantees for recovering λ_i and α_i are equivalent up to constants: $e^{-1}|\alpha_i - \widetilde{\alpha}_i| \leq |\lambda_i - \widetilde{\lambda}_i| \leq |\alpha_i - \widetilde{\alpha}_i|$. since e^{-x} is monotone decreasing on $[0,1]$ with derivative lying in $[-1, -1/e]$.

The full proof of Theorem 2.3 is given in Appendix C. As in previous work analyzing the MPM in the *super-resolution* setting with imaginary exponents [18], we see that the stability of MPM ultimately comes down to analyzing the condition number of the corresponding Vandermonde matrix, which in our case is very well-understood [9].

2.3 Strong information-theoretic lower bounds

In this section we describe our main results, strong information theoretic lower bounds establishing the difficulty of learning mixtures of exponentials (and hence, by our reductions, population histories). The full proofs of all results found in this section are given in Appendix D. First, we state a lower bound on learning the exponents λ_j, which is an informal restatement of Corollary D.5.

Theorem 2.4. *For any $k > 1$, there exists an infinite family of parameters $a_1, \ldots, a_k, \lambda_1, \ldots, \lambda_k$ and $b_1, \ldots, b_k, \mu_1, \ldots, \mu_k$ parametrized by integers $m > 2(k-1)$ and $\alpha \in (0, \frac{1}{2})$ such that:*

1. *Each λ_i and μ_j is in $(0,1]$, $\lambda_1 = \mu_1$, and the elements of $\{\lambda_i\}_{i=2}^k \cup \{\mu_i\}_{i=2}^k$ are all distinct and separated by at least $\Delta = 1/(m+2k)$. Furthermore $\lambda_2, \mu_2 > \alpha/k$.*

2. *Let H_1 and H_2 be hypotheses, under which the random variable T respectively follows the distributions*

$$\Pr_{H_1}[T \geq t] = \sum_{i=1}^k a_i e^{-\lambda_i t} \quad \textit{and} \quad \Pr_{H_2}[T \geq t] = \sum_{i=1}^k b_i e^{-\mu_i t}.$$

If N samples are observed from either H_1 or H_2, each with prior probability $1/2$, then the Bayes error rate for any classifier that distinguishes H_1 from H_2 is at least $\frac{1-\delta}{2}$, where

$$\delta = \frac{\alpha\sqrt{2N}}{2k-3}[\Delta(2k-3)]^{2k-4}.$$

Remark 3. From the square-root dependence of N in Theorem 2.3, the required number of samples N_0 has rate $4k$ in the exponent of $\frac{2e}{\Delta}$ if one just wants to learn the λ's, and Theorem 2.4 confirms that the exponent $4k$ is tight for learning the λ's.

Next we state an additional information-theoretic lower bound showing that the information-theoretic (minimax) rate is necessarily of the form $\frac{1}{\sqrt{N}}\Delta^{-O(k)}$ up to lower order terms, *even if* all of the λ_i are already known and we are only asked to reconstruct the mixing weights p_j.

Theorem 2.5. *Let m, k be positive integers such that $m > k > 3$ and let $\Delta = 1/(m+k-1)$. There exists a fixed choice of $\lambda_1, \dots, \lambda_k$ which are Δ-separated such that*

$$\inf_{\hat{p}} \max_{p} \mathbb{E}_p \|p - \hat{p}\|_1 \geq \frac{1}{4} \min\left(1, \frac{k-3}{\sqrt{2N}} \left(\frac{1}{\Delta(k-3)}\right)^{k-4}\right) \tag{4}$$

where the max is taken over feasible choices of p, and the infimum is taken over possible estimators $\hat{p}$ from N samples of the mixture of exponentials with CDF $F(t) = 1 - \sum_j p_j e^{-\lambda_j t}$.

Remark 4. Recall that in Theorem 2.3, the number of samples needed was exponential in $4k$ when learning just the λ's and in $6k$ for learning both the λ's and the p's. The exponent of $2k$ in Theorem 2.5 suggests that the discrepancy of $2k$ for MPM in Theorem 2.3 is tight.

As expected, our lower bounds show that the learning problem becomes harder as Δ approaches 0. The "easiest" case, then, ought to be when Δ is as large as possible, so that the λ_i are equally spaced apart in the unit interval. This raises the following question: as Δ grows, does the sample complexity remains exponential in k, or is there a phase transition (as in super-resolution [18]) where the problem becomes easier? In the Appendix, we completely resolve this question: the sample complexity still grows exponentially in $4k$ (Theorem D.7) when Δ is maximally large.

2.4 A tight upper bound: Nazarov-Turán-based hypothesis testing

As an alternative to the learning problem that the Matrix Pencil Method solves, we also consider the hypothesis testing scenario in which we want to test if the sampled data matches a hypothesized mixture distribution. In this case, we can give guarantees from weaker assumptions and requiring smaller numbers of samples. To state our guarantee, we need the following additional notation: for P a mixture of exponentials, let $p_\lambda(P)$ denote the coefficient of $e^{-\lambda t}$, which is 0 if this component is not present in the mixture. We study the following simple-versus-composite hypothesis testing problem using N samples:

Problem 1. Fix $k_0, k_1, \delta, \Delta > 0$ and let P be a known mixture of k_0 exponentials.

- H_0: The sampled data is drawn from P.
- H_1: The sampled data is drawn from a different, unknown mixture of at most k_1 exponentials Q. Let $\nu_1 := \max\{\lambda : p_\lambda(P) > p_\lambda(Q)\}$ and $\nu_2 := \max\{\lambda : p_\lambda(Q) > p_\lambda(P)\}$. We assume that $\min\{|p_{\nu_1}(P) - p_{\nu_1}(Q)|, |p_{\nu_2}(P) - p_{\nu_2}(Q)|\} \geq \delta$ and $|\nu_1 - \nu_2| \geq \Delta$.

Henceforth, we will refer to H_0 as the *null hypothesis* and H_1 as the *alternative hypothesis* (note that H_1 is a composite hypothesis). To solve this hypothesis testing problem, we propose a finite-sample variant of the Kolmogorov-Smirnov test:

1. Let $\alpha > 0$ be the significance level.
2. Let F_N be the empirical CDF and let F be the CDF under the null hypothesis H_0.
3. Reject H_0 if $\sup_t |F_n(t) - F(t)| > \sqrt{\log(2/\alpha)/2N}$.

We show that this test comes with a provable finite-sample guarantee.

Theorem 2.6. *Consider the problem setup as in Problem 1 and fix a significance level $\alpha > 0$. Let $k := \frac{k_0+k_1}{2}$ and $c_\Delta = 8e^2/\min(1/\Delta, 2k-1)$. Then:*

1. *(Type I Error) Under the null hypothesis, the above test rejects H_0 with probability at most α.*
2. *(Type II Error) There exists $N_0(\alpha) = O((c_\Delta/\Delta)^{4k-2}\log(2/\alpha)/\delta^2)$ such that if $N \geq N_0$, then the power of the test at significance level α is at least:*

$$\Pr_{Q \in H_1}[\textit{Reject } H_0] \geq 1 - 2\exp\left(-N\delta^2(\Delta/c_\Delta)^{4k-2}/8\right). \tag{5}$$

The full proof of Theorem 2.6 is given in Appendix E. The key step in the proof is a careful application of the celebrated Nazarov-Turán Lemma [20].

Remark 5. This improves upon the Matrix Pencil Method upper bound (Theorem 2.3), in terms of the exponent found above Δ (Δ^{-6k} versus Δ^{-4k}) and above the mixing weights ($p_{\min}^4$ versus δ^2). Even when the alternative Q is fixed and known, we see from Theorem 2.4 that $\Omega((1/\delta^2)(1/\Delta)^{4k})$ many samples are information-theoretically required, which matches Theorem 2.6.

3 Simulations and Indistinguishability in Simple Examples

Our theoretical analysis rigorously establishes the worst-case dependence on the number of samples needed in order to learn the parameters of a single period of population history under our model – recall the construction of Theorem 2.4 of two hard-to-distinguish mixtures of exponentials and the result Theorem 2.2 converting these to population histories.

In our simulations, we will analyze both the performance and information-theoretic difficulty of learning not a specially constructed worst-case instance, but instead an *extremely simple population history* with k populations. More precisely we consider the following instance:

Simulation Instance(k):

> **Population history description:** We consider reconstructing a single period model with k populations in which the ratio of the population sizes is $1 : 2 : \cdots : k$ and the relative probability of tracing back to each of these populations (i.e. $\Pr(\mathcal{E}_{i,i}|T > t_0)$ from Appendix A) are all equal to $1/k^2$. This can easily be realized as a one period of a 2-period population history model, in which in the second (more recent) era all populations are the same size[5]
>
> **Mixture of exponentials description:** We consider the following mixture of exponentials:
>
> $$\Pr(T > t) = (1 - 1/k) + \sum_{i=1}^{k}(1/k^2)e^{-t/k}.$$
>
> The constant term represents atomic mass at ∞ and corresponds to no coalescence. When $k = 1$ this is a standard exponential distribution, otherwise it is a mixture of $k+1$ exponentials, counting the degenerate constant term.

We do not believe that this is an unusually difficult instance of a mixture of exponentials on k components. If anything, the situation is likely the opposite: our worst-case analysis (Theorems 2.5, 2.3) suggests that this is comparatively *easy* as the gap parameter Δ is maximally large.

[5] As in Remark 1, we can optionally make the more recent era short so that almost all samples will be from the earlier period.

In order to evaluate the error in parameter space from the result of the learning algorithm, we adopted a natural metric, the well-known *Earthmover's distance*. Informally, this measures the minimum distance (weighted by p_i and recovered $\tilde{p}_i$) that the recovered exponents must be moved to agree with the ground truth; we give the precise definition in Appendix G.2.

For a point of comparison to MPM, we also tested a natural convex programming formulation which essentially minimizes $\| \int e^{-\lambda t} d\mu(\lambda) - (1 - \tilde{F}(t))\|_\infty$ over probability measures μ on $\mathbb{R}_{\geq 0}$, where $\tilde{F}$ is the empirical CDF – refer to Appendix G.1 for details. The results of running both the convex program and the MPM are shown in Figure 3 (blue and green lines) plotted on a log (base 10) scale; details of the setup are provided in Appendix G.2. As expected based on our theoretical analysis, the number of samples needed scaled exponentially in k, the number of populations in our instance. Details of the setup are provided in Appendix G.2; due to limitations of machine precision, the convex program could not reliably reconstruct at 5 components with any noise level and so this point is omitted.

Besides showing the performance of the algorithms, we were able to deduce *rigorous, unconditional* lower bounds on the information-theoretic difficulty of these particular instances. Each point on the red line corresponds to the existence of a different mixture of exponentials (found by examining the output of the convex program), with a comparable number of mixture components[6], which is far in parameter space[7] from the ground truth and yet the distribution of N samples from this model (where $N = 10^y$ and y is the y-coordinate in the plot) has total-variation (TV) distance at most 0.5 from the distribution of N samples from the true distribution. By the Neyman-Pearson Lemma, this implies that if the prior distribution is $\left(\frac{1}{2}, \frac{1}{2}\right)$ between these two distributions, then we cannot successfully distinguish them with greater than 75% probability. We describe the mathematical derivation of the TV bound in Appendix G.2, and illustrate such a hard-to-distinguish pair in Example 1. Recall that by Theorem 2.2, such a hard to distinguish pair of mixtures can automatically be converted into a pair of hard-to-distinguish population histories.

Notably, the lower bound shows that reliably learning the underlying parameters in this simple model with 5 components necessarily requires at least 10 trillion samples from the true coalescence distribution. In reality, since we do not truly have access to clean i.i.d. samples from the distribution, this is likely a significant underestimate.

Example 1. Consider the mixtures of exponentials with CDFs $F(t)$ and $G(t)$, where $1 - F(t) = 0.5 + 0.25e^{-0.5t} + 0.25e^{-t}$ and

$$1 - G(t) = 0.49975946 + 0.15359557e^{-0.45t} + 0.30642727e^{-0.81t} + 0.0402177e^{-1.55t}.$$

Despite being very different in parameter space, their H^2 distance is $7.9727 \cdot 10^{-6}$ so any learning algorithm requires at least 15660 samples to distinguish them with better than 75% success rate.

As a remark, we point out that the CDFs F and G in this example have exponents that are interlaced. Observe that this is a characteristic also shared by the information-theoretic obstructions referenced in Section 2.3 and Appendix D. This likely illustrates a major source of difficulty of most reasonable-looking instances: "averaging" adjacent exponents of an exponential mixture may produce a different mixture with a similar distribution whose components interlace with the original.

[6] The alternative hypothesis had no more than a few additional mixture components. A byproduct of this analysis is that even estimating the number of populations is in these examples requires a very large number of samples.

[7] More precisely, with Earthmover's distance in parameter space greater than 0.01. For comparison, an estimator which only gets the (easy) constant component correct already has Earthmover distance at most $1/k$.

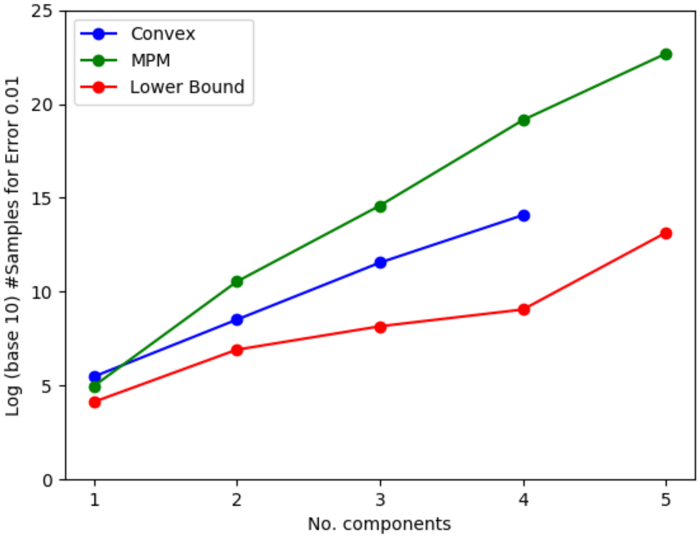

Figure 2: Plot of #components versus log (base 10) number of samples needed for accurate reconstruction (parameters within Earthmover's distance 0.01). Below the red line, it is mathematically impossible for *any method* to distinguish with greater than 75% success between the ground truth and a fixed alternative instance which has significantly different parameters.

References

[1] Anand Bhaskar and Yun S Song. Descartes rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of statistics*, 42(6):2469, 2014.

[2] Anand Bhaskar, YX Rachel Wang, and Yun S Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome research*, pages gr–178756, 2015.

[3] Richard A Blythe and Alan J McKane. Stochastic models of evolution in genetics, ecology and linguistics. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(07):P07018, 2007.

[4] Emmanuel J Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

[5] AJ Drummond, A Rambaut, B Shapiro, and OG Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.*, 22(5):1185–1192, 2005.

[6] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[7] Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10):e1003905, 2013.

[8] Anja Feldmann and Ward Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance evaluation*, 31(3-4):245–279, 1998.

[9] Walter Gautschi. On inverses of vandermonde and confluent vandermonde matrices. *Numerische Mathematik*, 4(1):117–123, 1962.

[10] Walter Gautschi. How (un) stable are vandermonde systems. 1990.

[11] Joseph Heled and Alexei Drummond. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8(1):289, 2008.

[12] Yingbo Hua and Tapan K Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824, 1990.

[13] Tyler A Joseph and Itsik Peer. Inference of population structure from ancient dna. In *International Conference on Research in Computational Molecular Biology*, pages 90–104. Springer, 2018.

[14] Junhyong Kim, Elchanan Mossel, Miklós Z Rácz, and Nathan Ross. Can one hear the shape of a population history? *Theoretical population biology*, 100:26–38, 2015.

[15] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493, 2011.

[16] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

[17] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1459):1387–1393, 2005.

[18] Ankur Moitra. Super-resolution, extremal functions and the condition number of vandermonde matrices. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 821–830, New York, NY, USA, 2015. ACM.

[19] Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theoretical population biology*, 73(3):342–348, 2008.

[20] Fedor L'vovich Nazarov. Local estimates for exponential polynomials and their applications to inequalities of the uncertainty principle type. *Algebra i analiz*, 5(4):3–66, 1993.

[21] Rasmus Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, 2000.

[22] Magnus Nordborg. Coalescent theory. *Handbook of statistical genetics*, 2:843–877, 2001.

[23] R Prony. Essai xperimental et analytique: sur les lois de la dilatabilit de uides lastique et sur celles de la force expansive de la vapeur de l'alkool, direntes tempratures. *Journal de lEcole Polytechnique*, 2, 1795.

[24] EJ Remez. Sur une propriété des polynômes de tchebycheff, comm. linst. *Sci., Kharkow*, 13:93–95, 1936.

[25] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919, 2014.

[26] Sara Sheehan, Kelley Harris, and Yuns S. Song. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics*, 194:647–662, 2013.

[27] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2):303, 2017.

[28] Jonathan Terhorst and Yun S Song. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, 112(25):7677–7682, 2015.

[29] Paul Turán. *On a new method of analysis and its applications.* Wiley New York, 1984.

A Derivation of the multiple-subpopulation coalescent model

For $b > a > 0$, let $I = [a, b]$ be an interval such that the population structure $\mathbf{N}$ is constant over I. Then Eq. (1), together with the Markov property of Kingman's coalescent model tells us that the coalescence time T of two randomly sampled individuals in the ith sub-population is given, for any $t \in [0, b - a]$, by

$$\Pr\left(T > a + t \;\middle|\; \mathcal{E}_{i,i} \wedge \{T > a\}\right) = \exp\left(-\frac{1}{N_i}t\right). \tag{6}$$

Here, $\mathcal{E}_{i,j}$ represents the event where the ancestry of one of the individuals traces back to subpopulation i, and the other traces back to j.

Let D be the number of subpopulations restricted to the interval I. By the law of total probability, the random variable T satisfies, again for any $t \in [0, b - a]$,

$$\Pr(T > a + t \mid T > a) = \sum_{i<j} \Pr(\mathcal{E}_{i,j} \mid T > a) + \sum_{i=1}^{D} \Pr(\mathcal{E}_{i,i} \mid T > a) \Pr(T > a + t \mid \mathcal{E}_{i,i} \wedge \{T > a\})$$

The first summation over $i < j$ uses the fact that $\Pr(T > a + t \mid \mathcal{E}_{i,j} \wedge \{T > a\}) = 1$, via the "no admixture" assumption; whenever the two individuals' lineages at time a lie in distinct subpopulations, they do not coalesce anywhere in I. Via Eq. (6), the right hand side can be re-written as seen in Eq. (2), i.e.

$$\Pr(T > a + t \mid T > a) = \sum_{\ell=0}^{D} p_\ell e^{-\lambda_\ell t}.$$

B Reduction from Mixtures of Exponentials to Population History

B.1 Proof of Theorem 2.1

Proof. We consider the following population history:

- In the (more recent) period $[0, t_0]$ there are k populations and population i has size $\sqrt{q_i}$, where q_i is the (unique) nonnegative solution to

  $$p_i = q_i e^{-t_0/\sqrt{q_i}}.$$

  To see that the solution exists and is unique, observe that the rhs of this equation is a strictly increasing function in q_i which maps $(0, \infty)$ to $(0, \infty)$.

- In the (less recent) period $[t_0, \infty)$ each of the k populations changes to size $1/\lambda_i$.

By construction, the probability that two independently sampled individuals being in the same population is proportional to q_i, and conditioned on no coalescence before time t_0 this probability is proportional to p_i. Therefore the distribution Q of coalescence times satisfies

$$Q(T > t + t_0 | T > t_0) = q_0 + \frac{1}{1 - q_0} \sum_{i=1}^{k} p_i e^{-\lambda_i t} \tag{7}$$

where q_0 is the probability that the two individuals sampled were in different populations. □

B.2 Hardness for Distinguishing Population Histories

Theorem 2.1 shows that any arbitrary mixture of exponentials can be embedded as a sub-mixture of a simple population history with two time periods. We can leverage this equivalence to reduce distinguishing two mixtures of exponentials to distinguishing two population histories, and hence conclude from Theorem 2.4 that distinguishing two population histories is exponentially hard in the number of subpopulations.

The high-level idea is to take the reduction from Theorem 2.1 and apply it to two arbitrary mixtures of exponentials, and argue that the problem of distinguishing the resulting two population histories is at least as hard as distinguishing the two mixtures they came from. The problem that arises is that the constant term (q_0 in Equation 7) is exactly the probability of no coalescence, which is fixed by the subpopulation sizes, and therefore fixed by the desired mixture. Therefore, if we are not careful, the probability of no coalescence will be significantly different between our two population histories, making them easily distinguishable.

Proof of Theorem 2.2. Consider the following population history for R:

- The first k populations of R are constructed according to the proof of Theorem 2.1 from the mixture P. Let N_i denote the size if the i^{th} population in the more recent period $[0, t_0)$.
- The $(k+1)^{th}$ and $(k+2)^{th}$ populations of R will have sizes N_{k+1}, N_{k+2} in the period $[0, t_0)$, such that $N_{k+1}^2 + N_{k+2}^2 = 64 \max(k^2, \ell^2) \sum_{i=1}^{k} N_i^2$. The exact sizes of N_{k+1} and N_{k+2} will be set later.
- Every pair of samples in populations $k+1$ and $k+2$ coalesce in the closed interval $[0, t_0]$. This can be achieved by having populations $k+1$ and $k+2$ shrink to size "zero" at time t_0, so any pair of individuals in these populations who have not coalesced in the interval $[0, t_0)$ will coalesce at time t_0.

Similarly, let the population history S be constructed from the mixture Q in the same way (substituting Q for P and ℓ for k in what appears above), and let M_j denote the size of j^{th} population of the population history S in the (most recent) period $[0, t_0)$ (where j ranges from 1 to $\ell + 2$). Recall that we want to establish the following four properties:

1. R has $k+2$ subpopulations and S has $\ell + 2$ subpopulations.
2. $R[T > t + t_0 | T \neq \infty, T > t_0] = P(T > t)$ and $S[T > t + t_0 | T \neq \infty, T > t_0] = Q(T > t)$
3. $R[T = t_0] = S[T = t_0]$.
4. $R[T = \infty] = S[T = \infty]$.

Note that the coalescence probability for the population history R at exactly t_0 is

$$R[T = t_0] = \frac{1}{(\sum_{i=1}^{k+2} N_i)^2} \left(N_{k+1}^2 e^{-t_0/N_{k+1}} + N_{k+2}^2 e^{-t_0/N_{k+2}} \right).$$

First, we analyze what happens in the limit $t_0 \to 0$, disregarding the term that comes from the probability of coalescing before time t_0, so the equations simplify to

$$R[T = t_0] = \frac{N_{k+1}^2 + N_{k+2}^2}{(\sum_{i=1}^{k+2} N_i)^2} \tag{8}$$

and similarly

$$S[T=t_0] = \frac{M_{k+1}^2 + M_{k+2}^2}{(\sum_{i=1}^{\ell+2} M_i)^2} \tag{9}$$

We do this mainly for simplicity, and at the end show how to handle the general case, supposing t_0 is sufficiently small.

It is clear that R and S satisfy Property (1) of Theorem 2.2 by construction, as they have the correct numbers of subpopulations. Property (2) is similarly satisfied by construction and by an application to Theorem 2.1. Indeed, when we sample from the history R conditioned on coalescence before time ∞ and no coalescence at or before time t_0, we can see that the two individuals sampled must be from one of subpopulations 1 to k (since any pair of individuals sampled from subpopulations $k+1$ or $k+2$ coalesce in the interval $[0, t_0]$). Repeating this argument for S establishes property (2).

It remains to show that Properties (3) and (4) hold, and for this we will need to set $N_{k+1}, N_{k+2}, M_{\ell+1}$, and $M_{\ell+2}$ appropriately. In order to do this, we first rewrite Properties (3) and (4) using Equations 8 and 9, to see what conditions we need to satisfy with $N_{k+1}, N_{k+2}, M_{\ell+1}$, and $M_{\ell+2}$.

Noting that the probability of coalescence at *exactly* time t_0 from population history R (resp. S) is exactly the probability of sampling a pair of individuals from N_{k+1} or N_{k+2} (resp. $M_{\ell+1}$ or $M_{\ell+2}$), we get that R and S will satisfy Property (3) if and only if the population sizes satisfy the following (desired) equation:

$$\frac{N_{k+1}^2 + N_{k+2}^2}{\left(\sum\limits_{i=1}^{k+2} N_i\right)^2} = \frac{M_{\ell+1}^2 + M_{\ell+2}^2}{\left(\sum\limits_{j=1}^{\ell+2} M_j\right)^2} \tag{10}$$

Similarly, R and S will satisfy Property (4) if and only if the population sizes satisfy the following (desired) equation:

$$\frac{\sum\limits_{j=1}^{\ell+2} M_j^2}{\left(\sum\limits_{j=1}^{\ell+2} M_j\right)^2} = \frac{\sum\limits_{i=1}^{k+2} N_i^2}{\left(\sum\limits_{i=1}^{k+2} N_i\right)^2} \tag{11}$$

So it suffices to describe how to set $N_{k+1}, N_{k+2}, M_{\ell+1}$, and $M_{\ell+2}$ such that Equations 10 and 11 are satisfied. First, we rearrange Equations 10 and 11 to get the equivalent (desired) set of equalities

$$\frac{\left(\sum\limits_{i=1}^{k+2} N_i\right)^2}{\left(\sum\limits_{j=1}^{\ell+2} M_j\right)^2} = \frac{N_{k+1}^2 + N_{k+2}^2}{M_{\ell+1}^2 + M_{\ell+2}^2} = \frac{\sum\limits_{i=1}^{k+2} N_i^2}{\sum\limits_{j=1}^{\ell+2} M_j^2} \tag{12}$$

Note that the second equality is trivially satisfied by our stipulations that $N_{k+1}^2 + N_{k+2}^2 = 64\max(k^2, \ell^2)\sum_{i=1}^{k} N_i^2$ and $M_{\ell+1}^2 + M_{\ell+2}^2 = 64\max(k^2, \ell^2)\sum_{i=1}^{\ell} M_j^2$.

Now we move on to describing how to set population sizes to satisfy the first equality. Note that we can write

$$\left(\sum_{j=1}^{\ell+2} M_j\right)^2 = ((M_{\ell+1} + M_{\ell+2}) + \sum_{j=1}^{\ell} M_j)^2 \tag{13}$$

Note that the RHS of Equation 13 is a continuous and strictly increasing function of $M_{\ell+1} + M_{\ell+2}$, and furthermore that we can set $M_{\ell+1} + M_{\ell+2}$ to be anywhere in the range $\left[\sqrt{M_{\ell+1}^2 + M_{\ell+2}^2}, \sqrt{2} \cdot \sqrt{M_{\ell+1}^2 + M_{\ell+2}^2}\right]$ while keeping $M_{\ell+1}^2 + M_{\ell+2}^2$ constant. Furthermore, note that

$$\begin{aligned}\left(\sum_{j=1}^{\ell+2} M_j\right)^2 &= ((M_{\ell+1} + N_{\ell+2}) + \sum_{j=1}^{\ell} M_j)^2 \\ &\leq ((9/8)(M_{\ell+1} + M_{\ell+2}))^2 \\ &= \frac{81}{64}(M_{\ell+1} + M_{\ell+2})^2 \end{aligned} \tag{14}$$

where the second line uses the fact that we set $M_{\ell+1}^2 + M_{\ell+2}^2$ sufficiently large such that

$$M_{\ell+1} + M_{\ell+2} \geq \sqrt{64\ell^2 \sum_{j=1}^{\ell} M_j^2} \geq \sqrt{64 \left(\sum_{j=1}^{\ell} M_j\right)^2} \geq 8 \sum_{j=1}^{\ell} M_j$$

We can do the same bound for $\left(\sum_{i=1}^{k+2} N_i\right)^2$, and hence we get that the LHS of Equation 12 can be bounded on both sides as follows

$$\frac{64}{81} \frac{(N_{k+1} + N_{k+2})^2}{(M_{\ell+1} + M_{\ell+2})^2} \leq \frac{\left(\sum\limits_{i=1}^{k+2} N_i\right)^2}{\left(\sum\limits_{j=1}^{\ell+2} M_j\right)^2} \leq \frac{81}{64} \frac{(N_{k+1} + N_{k+2})^2}{(M_{\ell+1} + M_{\ell+2})^2} \tag{15}$$

Now suppose that we initially set $N_{k+2} = M_{\ell+2} = 0$, and suppose that this gives us that

$$\frac{N_{k+1}^2 + N_{k+2}^2}{M_{\ell+1}^2 + M_{\ell+2}^2} = \alpha \cdot \frac{\left(\sum\limits_{i=1}^{k+2} N_i\right)^2}{\left(\sum\limits_{j=1}^{\ell+2} M_j\right)^2} \tag{16}$$

for some $\alpha < 1$. We know by $\alpha \geq \frac{64}{81}$ by noting that $N_{k+2} = M_{\ell+2} = 0$ and applying the upper bound of Equation 15.

We can continuously increase $M_{\ell+1} + M_{\ell+2}$ while keeping $M_{\ell+1}^2 + M_{\ell+2}^2$ constant by moving $M_{\ell+1}$ and $M_{\ell+2}$ relatively closer together, until we satisfy Equation 16 with $\alpha = 1$, satisfying the first equality in Equation 12. This is because we can increase $(M_{\ell+1} + M_{\ell+2})^2$ by a factor of up to 2 overall in this manner, and using the fact that

$$(M_{\ell+1} + M_{\ell+2})^2 \leq \left(\sum_{j=1}^{\ell+2} M_j\right)^2 \leq \frac{81}{64}(M_{\ell+1} + M_{\ell+2})^2$$

we conclude that we can increase $\left(\sum_{j=1}^{\ell+2} M_j\right)^2$ by at least a multiplicative factor of $2 \cdot (64/81) \geq (81/64)$ in this fashion. Hence, there exists a setting of $M_{\ell+1}$ and $M_{\ell+2}$ that satisfies both equalities in Equation 12.

To handle the case where $1 < \alpha \le 81/64$, we instead continuously increase $N_{k+1} + N_{k+2}$ while keeping $N_{k+1}^2 + N_{k+2}^2$ fixed, and the argument goes through mutatis mutandis.

Finally, we describe how to handle the general case where $t_0 > 0$ and t_0 is sufficiently small. We can ensure that property (2) by using the same rescaling method as in the proof of Theorem 2.1. It remains to show that properties (3) and (4) hold by modifying $N_{k+1}, N_{k+2}, M_{\ell+1}, M_{\ell+2}$ appropriately. Recall that in this case, $R[T = t_0] = \frac{1}{(\sum_{i=1}^{k+2} N_i)^2}\left(N_{k+1}^2 e^{-t_0/N_{k+1}} + N_{k+2}^2 e^{-t_0/N_{k+2}}\right)$ and similarly for S. As before, property (3) holds by fixing the ratio of $N_{k+1}^2 e^{-t_0/N_{k+1}} + N_{k+2}^2 e^{-t_0/N_{k+2}}$ and $M_{k+1}^2 e^{-t_0/M_{k+1}} + M_{k+2}^2 e^{-t_0/M_{k+2}}$. As before, this leaves a degree of freedom in the value of $M_{k+1} + M_{k+2}$ which we use to guarantee property (4) holds assuming t_0 is sufficiently small. □

C Analysis of the Matrix Pencil Method

In this section, $\|A\|$ and $\|A\|_F$ respectively denotes the operator norm and the Frobenius norm of matrices. For a vector x, $\|x\|$ is its Euclidean norm, and more generally $\|x\|_p$ denotes its ℓ_p-norm.

C.1 The condition number of Vandermonde matrices

In 1962, Gautschi [9] observed an exact formula for the $\ell_\infty \to \ell_\infty$ condition number of a real Vandermonde matrix, which we now recall:

Definition 1. The $\ell_\infty \to \ell_\infty$ norm of a matrix is defined by[8]

$$\|A\|_{\infty\to\infty} := \sup_{x:\|x\|_\infty=1} \|Ax\|_\infty = \max_i \sum_j |A_{ij}|.$$

Theorem C.1 ([9],[10]). *Suppose $V = V_n(\alpha_1, \dots, \alpha_n)$ such that $\alpha_1, \dots, \alpha_n$ are all distinct. Then*

$$\max_i \prod_{j:j\neq i} \frac{\max(1, |\alpha_j|)}{|\alpha_i - \alpha_j|} \le \|V^{-1}\|_{\infty\to\infty} \le \max_i \prod_{j:j\neq i} \frac{1 + |\alpha_j|}{|\alpha_j - \alpha_i|}$$

and furthermore equality is attained in the upper bound whenever $\alpha_j = |\alpha_j| e^{i\theta}$ for θ independent of j.

Since we are interested in real-valued Vandermonde matrices with positive entries, the above expression is an exact formula. Furthermore, we can relate $\|V^{-1}\|_{\infty\to\infty}$ to the bottom singular value of V, i.e. top singular value of V^{-1}, because for any matrix A we have

$$\sigma_{\max}(A) = \sup_{x:\|x\|_2=1} \|Ax\|_2 \le \sqrt{n} \sup_{x:\|x\|_2=1} \|Ax\|_\infty \le \sqrt{n} \sup_{x:\|x\|_\infty=1} \|Ax\|_\infty = \sqrt{n}\|A\|_{\infty\to\infty}$$

for an upper bound, and for a lower bound we have

$$\sigma_{\max}(A) = \sup_{x:x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \ge \sup_{x:x\neq 0} \frac{\|Ax\|_2}{\sqrt{n}\|x\|_\infty} \ge \sup_{x:x\neq 0} \frac{\|Ax\|_\infty}{\sqrt{n}\|x\|_\infty} = \frac{1}{\sqrt{n}}\|A\|_{\infty\to\infty}$$

Hence, by applying these bounds to the matrix V^{-1} and from Theorem C.1, we see that

$$\sigma_{min}(V)^{-1} = \sigma_{max}(V^{-1}) \in \left(\frac{1}{\sqrt{n}} \max_i \prod_{j:j\neq i} \frac{1 + |\alpha_j|}{|\alpha_j - \alpha_i|}, \sqrt{n} \max_i \prod_{j:j\neq i} \frac{1 + |\alpha_j|}{|\alpha_j - \alpha_i|}\right)$$

[8] The second equality follows from Holder's inequality (and its equality case, where one takes x to be the appropriate sign vector).

In particular, these bounds show that the condition number of a square Vandermonde is exponentially bad in the dimension. (Note that these bounds do not contradict Moitra's results [18], which instead shows that in the Fourier setting, rectangular V can sometimes be well-conditioned – this is because Theorem C.1 looks specifically at the conditioning of square Vandermonde matrices.)

Next, we translate this result to a bound in terms of the parameters described in Section 2.2.

Lemma C.2. *Let $\alpha_i = e^{-\lambda_i}$ and $1 > \lambda_1 > \cdots > \lambda_k > 0$. Define $V = V_k(\alpha_1, \ldots, \alpha_k)$, and let $\Delta = \min_{i \neq j} |\lambda_i - \lambda_j|$. Then $\frac{1}{\sigma_{min}(V)} \leq \sqrt{k}\left(\frac{2e}{\Delta}\right)^k$ and $\kappa(V) \leq k^{3/2}\left(\frac{2e}{\Delta}\right)^k$.*

Proof. Observe that

$$\prod_{j \neq i} \frac{1 + |\alpha_j|}{|\alpha_j - \alpha_i|} \leq \left(\frac{2e}{\Delta}\right)^k$$

because $|\alpha_i| \leq 1$ and

$$\alpha_i - \alpha_j = e^{-\lambda_i} - e^{-\lambda_j} = \int_{-\lambda_j}^{-\lambda_i} e^x dx \geq \Delta e^{-1}$$

for $i > j$. Thus $\|V^{-1}\|_{\infty \to \infty} \leq \left(\frac{2e}{\Delta}\right)^k$, which implies $\frac{1}{\sigma_{\min}(V)} = \sigma_{\max}(V^{-1}) \leq \sqrt{k}\left(\frac{2e}{\Delta}\right)^k$. We also know that $\sigma_{\max}(V) \leq \|V\|_F \leq k$, which gives the bound on the condition number. □

C.2 Matrix perturbation bounds

In this section, we will establish key lemmas that allow us to prove bounds on how close $\widetilde{\alpha}_j$ are to α_j.

Lemma C.3. *Let (A, B) be a pair of $n \times n$ matrices with generalized eigenvalues $\{\mu_j\}$, such that B is nonsingular. Take $A = VD_AV^T$ and $B = VD_BV^T$ where D_A and D_B are diagonal matrices and V is an arbitrary, nonsingular $n \times n$ matrix. Consider the perturbed system $\widetilde{A} = A + E$ and $\widetilde{B} = B + F$ where E and F are symmetric matrices, and let $\{\widetilde{\mu}_j\}$ be its generalized eigenvalues. Assume further that $\|F\| < \sigma_{min}(D_B)\sigma_{min}(V)^2$. Then for all j,*

$$|\mu_j - \widetilde{\mu}_j| \leq \frac{2k^{3/2}}{\sigma_{\min}(D_B)\sigma_{\min}(V)^2 - \|F\|}\left(\frac{\sigma_{\max}(D_A)}{\sigma_{\min}(D_B)}\|F\| + \|E\|\right) \tag{17}$$

Proof. Observe that the generalized eigenvalue problem $Ax = \mu Bx$ has the same solutions as the ordinary eigenvalue problem $B^{-1}Ax = \mu x$. Let $E' = V^{-1}E(V^T)^{-1}$ and $F' = V^{-1}F(V^T)^{-1}$ so that

$$\tilde{A} = V(D_A + E')V^T, \qquad \tilde{B} = V(D_B + F')V^T.$$

Note that $\|F'\| \leq \|V^{-1}\|^2\|F\| < \sigma_{\min}(D_B)$ by assumption, so $\tilde{B}$ is also invertible, hence the generalized eigenvalues $\{\tilde{\mu}_j\}$ of $(\tilde{A}, \tilde{B})$ are just the ordinary eigenvalues of $\tilde{B}^{-1}\tilde{A}$. Since eigenvalues are invariant under change of basis (i.e. similarity transformation), the eigenvalues of $B^{-1}A$ are the same as those of

$$C := V^T B^{-1} A (V^T)^{-1} = D_B^{-1} D_A$$

and the eigenvalues of $\tilde{B}^{-1}\tilde{A}$ are the same as those of

$$\tilde{C} := V^T \tilde{B}^{-1} \tilde{A} (V^T)^{-1} = (D_B + F')^{-1}(D_A + E').$$

Therefore if we let $\mathcal{E} := C - \tilde{C}$, then by Gershgorin's circle theorem[9], the Cauchy-Schwarz inequality, and the equivalence of Frobenius and spectral norms, we find

$$|\mu_j - \tilde{\mu}_j| \le 2\sum_{i,j} |\mathcal{E}_{ij}| \le 2k\|\mathcal{E}\|_F \le 2k^{3/2}\|\mathcal{E}\| \tag{18}$$

for all j.

It remains to bound $\|\mathcal{E}\|$. By the triangle inequality,

$$\|\mathcal{E}\| = \|D_B^{-1}D_A - (D_B + F')^{-1}(D_A + E')\| \le \|D_B^{-1}D_A - (D_B + F')^{-1}D_A\| + \|(D_B + F')^{-1}E'\|.$$

To bound the second term on the right hand side, we observe

$$\|(D_B + F')^{-1}E'\| \le \|(D_B + F')^{-1}\|\|E'\| \le \frac{\|E'\|}{\sigma_{\min}(D_B) - \|F'\|}.$$

To bound the first term, we observe the following useful matrix identity for $\|M\| < 1$:

$$(I - M)^{-1} = \sum_{i=0}^{\infty} M^i = I + M(I - M)^{-1}.$$

This gives

$$(D_B + F')^{-1} = D_B^{-1}(I + F'D_B^{-1})^{-1} = D_B^{-1} - D_B^{-1}F'(D_B + F')^{-1}.$$

which is valid because $\|F'\| \le \|F\|\|V^{-1}\|^2 < \sigma_{\min}(D_B)$ by assumption. Therefore

$$\|D_B^{-1}D_A - (D_B + F')^{-1}D_A\| = \|D_B^{-1}F'(D_B + F')^{-1}D_A\| \le \frac{\|D_B^{-1}\|\|F'\|\|D_A\|}{\sigma_{\min}(D_B) - \|F'\|}.$$

Combining these two parts completes the bound on $\|\mathcal{E}\|$:

$$\begin{aligned}\|\mathcal{E}\| &\le \frac{1}{\sigma_{\min}(D_B) - \|F'\|}(\|D_B^{-1}\|\|F'\|\|D_A\| + \|E'\|)\\ &\le \frac{\|V^{-1}\|^2}{\sigma_{\min}(D_B) - \|V^{-1}\|^2\|F\|}(\|D_B^{-1}\|\|F\|\|D_A\| + \|E\|).\end{aligned}$$

Rewriting the last expression and combining with (18) gives the result. □

C.3 Proof of Theorem 2.3

When we put the results of Appendix C.2 back into the context of learning mixtures of exponentials, we should think about the perturbation errors E and F as essentially being the same, just offset in the row/column indexing from each other. This is because when we are using the Matrix Pencil Method (Section 2.2) to learn an exponential mixture, the entries of A and B are simply noisy versions of $v_0, v_1, v_2, \ldots, v_{2k-1}$.

We remind the reader of the normalization assumption which restricts $\lambda_1, \ldots, \lambda_k \in [0, 1]$. An application of the previous lemmas yields the following result:

[9] Here we use that the connected component made of r Gershgorin discs has exactly r eigenvalues, which follows by a standard continuity argument. Therefore the distance an eigenvalue moves under perturbation is at most the sum of the diameters of the discs.

Lemma C.4. *Consider $\widetilde{v}_0 \dots, \widetilde{v}_{2k-1}$ as inputs to the Matrix Pencil Method, in place of $v_0, \dots, v_{2k-1}$. Let $p_{\min} := \min_i p_i$. If*

$$\sup_{t\geq 0} |v_t - \widetilde{v}_t| = \frac{\beta p_{\min}}{k^2} \left(\frac{2e}{\Delta}\right)^{-2k}$$

for some $\beta \in [0, 1)$, then for all j,

$$|\alpha_j - \widetilde{\alpha}_j| \leq \frac{k^{3/2}\left(\frac{1}{p_{\min}} + 1\right)}{\frac{1}{\beta} - 1}.$$

Proof. The strategy is to apply Lemma C.3. The setup prescribes $D_A = D_p D_\alpha$, $D_B = D_p$, which gives $\sigma_{\min}(D_B) = p_{\min}$ and $\sigma_{\max}(D_A) \leq 1$. Observe that the hypothesized bound on the estimated v_t's implies

$$\max(\|E\|_\infty, \|F\|_\infty) = \max_{t\in\{0,\dots,2k-1\}} v_t \leq \frac{\beta p_{\min}}{k^2}\left(\frac{2e}{\Delta}\right)^{-2k}.$$

This, in turn, gives us bounds on $\|E\|$ (and $\|F\|$), since $\|E\| \leq k\|E\|_\infty$. In particular, by Lemma C.2, $\|F\| \leq \sigma_{\min}(D_B)\sigma_{\min}(V)^2$.

This allows us to directly apply Lemma C.3, by substituting α_j and $\widetilde{\alpha}_j$ as the eigenvalues of (A, B) and $(\widetilde{A}, \widetilde{B})$ respectively. This immediately gives the desired bound on $|\alpha_j - \widetilde{\alpha}_j|$. □

Lemma C.4 provides sufficient conditions for the computed exponents $\widetilde{\lambda}_j$ to be accurate. It remains to analyze the resulting error in the coefficients p_j. To do so, we recall the following result, attributed to Weyl:

Theorem C.5 (Singular Value Stability)**.** *Let A and B be $n \times n$ matrices with entries in $\mathbb{R}$. Then for $j = 1, \dots, n$, we have $|\sigma_j(A + B) - \sigma_j(A)| < \|B\|$.*

Lemma C.6. *Consider step (3) of the Matrix Pencil Method, using $\widetilde{v}_0 \dots, \widetilde{v}_{2k-1}$ and $\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_k$ in place of their true counterparts. Let $\rho := \max_i |\alpha_i - \widetilde{\alpha}_i|$, $\epsilon := \sup_{t\geq 0} |v_t - \widetilde{v}_t|$, and $\widetilde{V} := V_k(\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_n)$. Then if $\sigma_{min}(V) > k\rho$,*

$$\|\widetilde{p} - p\|_\infty \leq \frac{\sqrt{k}\epsilon + k\rho}{\sigma_{\min}(V) - k\rho}. \tag{19}$$

Proof. Let $\eta = \widetilde{v} - v$, and recall that $v = Vp$, where p is the vector $(p_1, \dots, p_n)$. By the setup, we may write $\widetilde{v} = Vp + \eta = \widetilde{V}p + (V - \widetilde{V})p + \eta$. Let $\tilde{p}$ be such that $\widetilde{v} = \widetilde{V}\widetilde{p}$. Equating the two formulae and moving terms around gives

$$\widetilde{p} - p = \widetilde{V}^{-1}\left((V - \widetilde{V})p + \eta\right)$$

which yields

$$\|\widetilde{p} - p\| \leq \|\widetilde{V}^{-1}\|(\|V - \widetilde{V}\|\|p\| + \|\eta\|) = \frac{\|V - \widetilde{V}\|\|p\| + \|\eta\|}{\sigma_{\min}(\widetilde{V})}. \tag{20}$$

Now observe that

1. By Gershgorin's circle theorem, $\|V - \widetilde{V}\| \leq k\|V - \widetilde{V}\|_\infty = k \cdot \max_i |\alpha_i - \widetilde{\alpha}_i| = k\rho$.

2. p is a vector $(p_1, \dots, p_n)$ such that $\sum_i p_i = 1$, so $\|p\| \leq 1$.

3. $\|\eta\| \leq \sqrt{k}\|\eta\|_\infty \leq \sqrt{k}\epsilon$.

4. Using Theorem C.5, we get $\sigma_{\min}(\widetilde{V}) \geq \sigma_{\min}(V) - k\|\widetilde{V} - V\|_\infty$, given the hypothesis $\sigma_{\min}(V) > k\rho$.

These, together with Equation 20 and the bound $\|\widetilde{p} - p\|_\infty < \|\widetilde{p} - p\|$ give the desired result. □

The hypothesis $\sigma_{\min}(V) > k\rho$ of Lemma C.6 is satisfied for sufficiently small β. Qualitatively, this is the regime where β is small enough so that $\sigma_{\min}(\widetilde{V}) \approx \sigma_{\min}(V)$. Indeed, the statement of Theorem 2.3 specifies the case for large N, which will induce such a scenario. To complete the analysis, we recall the Dvoretzky-Kiefer-Wolfowitz concentration inequality which gives uniform control of the error for the empirical CDF:

Theorem C.7 (DKW Inequality [6, 16]). *Suppose $X_1, \ldots, X_n$ are i.i.d. samples from an unknown distribution with CDF F. Let F_n denote the n-sample empirical CDF, i.e.*

$$F_n(x) := \frac{1}{n}\#\{i : X_i \leq x\}.$$

Then for every $\epsilon > 0$,

$$\Pr\left(\sup_{t\in\mathbb{R}} |F(t) - F_n(t)| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

We are now ready to finish the proof of Theorem 2.3.

Proof of Theorem 2.3. Recall that v_t is the tail probability $\Pr[T \geq t] = 1 - F(t)$, while the empirical estimate of v_t is computed as $\widetilde{v}_t = 1 - F_n(t)$; therefore $|F(t) - F_n(t)| = |v_t - \widetilde{v}_t|$. Let $\epsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$, so that by the DKW inequality (Theorem C.7),

$$\sup_t |v_t - \widetilde{v}_t| \leq \epsilon$$

with probability at least $1 - \delta$. Let $\beta = \frac{\epsilon k^2}{p_{\min}}\left(\frac{2e}{\Delta}\right)^{2k}$, as in the hypothesis of Lemma C.4. Observe that both ϵ and β decrease to zero as N increases for any fixed δ. Specifically, we may take N large enough so that $\frac{1}{\beta} - 1 > \frac{1}{2\beta}$.

According to Lemma C.4, the event $|v_t - \widetilde{v}_t| < \epsilon$ implies the desired bound on the α's:

$$|\alpha_j - \widetilde{\alpha}_j| \leq k^{1.5}\frac{\frac{1}{p_{\min}} + 1}{\frac{1}{\beta} - 1} \leq k^{1.5} \cdot \frac{2}{p_{\min}} \cdot 2\beta = \frac{4k^{3.5}\epsilon}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k}.$$

To get the bound on the p's, again we require N sufficiently large, this time so that $\frac{1}{\epsilon\sqrt{k}\left(\frac{2e}{\Delta}\right)^k} - \frac{4k^{4.5}}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k} \geq \frac{1}{2\epsilon\sqrt{k}\left(\frac{2e}{\Delta}\right)^k}$. Such a condition ensures that $\frac{4k^{4.5}\epsilon}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k} < \sigma_{\min}(V)$ by Lemma C.2, which allows us to apply Lemma C.6. This gives us

$$|p_j - \widetilde{p}_j| < \frac{\sqrt{k}\epsilon + \frac{4k^{4.5}\epsilon}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k}}{\sigma_{\min}(V) - \frac{4k^{4.5}\epsilon}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k}} \leq \frac{\frac{5k^{4.5}}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k}}{\frac{1}{\epsilon\sqrt{k}}\left(\frac{2e}{\Delta}\right)^{-k} - \frac{4k^{4.5}}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{2k}} \leq \frac{10k^5\epsilon}{p_{\min}^2}\left(\frac{2e}{\Delta}\right)^{3k}.$$

In particular, both of the stated conditions on N are satisfied for $N \geq N_0$, where

$$N_0 = \frac{32k^{10}}{p_{\min}^4}\left(\frac{2e}{\Delta}\right)^{6k}\log\frac{2}{\delta}.$$

□

D Information-Theoretic Lower Bounds

We will show that an exponential dependence on the number of components is not just a limitation of the matrix pencil method, but is in fact information-theoretically necessary. Furthermore we will show that our dependence on the gap parameter Δ is also information-theoretically necessary, proving that the Matrix pencil method has essentially the correct dependence on all parameters.

Our constructions rely on interpreting polynomials as the difference between two mixtures of exponentials, under the re-parametrization $x \mapsto e^{-x}$. From this point of view, and in light of our upper bound (Theorem 2.3), the hardest examples should come from families of polynomials whose supremum on $[0, 1]$ are exponentially small in the number of terms.

Of particular interest are two families of polynomials – one attributed to Chebyshev and the other to Turań [20, 29]. These polynomials are tightest examples of extremal families, in the sense that they have the largest growth rate outside of the unit interval. In particular, the Chebyshev family are tight examples of the Remez Inequality [24] which is a bound on the supremum norm of a polynomial, while Turań's polynomial family (essentially) serve as tight examples for the closely related Nazarov-Turań lemma (Theorem E.1). Notably, the latter result is specialized to linear combinations of exponential functions – precisely the setting that we are working in, which make Turań's polynomials fitting candidates for constructing matching lower bounds.

D.1 A lower bound construction

D.1.1 Turán's polynomials and their properties

First, we recall the construction by Turán's family of polynomials. As mentioned earlier in this section, a slight multivariate generalization of this construction is central to the proof of Turán's First Main Theorem [29] and to "Turán's proof" of the Nazarov-Turán Lemma [20].

Definition 2 (Turán Polynomials)**.** Fix positive integers m and n such that $m > n$. The (m, n)-th Turań polynomial $Q_{m,n}(z)$ is defined to be $(1 - z)^n \sigma_m(z)$, where $\sigma_m(z) = \sum_{i=0}^{m} a_i z^i$ is the degree-m truncation of the power series expansion of $(1 - z)^{-n}$.

These polynomials have several very useful properties; we recall these facts below, along with proofs for the reader's convenience. First we recall the following basic fact:

Proposition D.1. *For any k and n,*

$$[z^k]\frac{1}{(1-z)^n} = \binom{n+k-1}{n-1}$$

where the left hand side denotes the coefficient of z^k in the power series expansion about 0.

Proof. Consider the power series expansion of $1/(1 - z)$ as $1 + z + z^2 + \cdots$. It follows that the kth coefficient of the expansion of $\left(\frac{1}{1-z}\right)^n$ is exactly the number of ways in which we can select n nonnegative numbers to sum up to k, which is $\binom{n+k-1}{n-1}$. □

From this one can verify that the Turan polynomials has several very interesting and useful properties.

Lemma D.2. *Let n, m be positive integers such that $m > n$, and let $Q = Q_{m,n}$. Then*

1. *$Q(1) = 0$ so the sum of the coefficients equals* 0.

2. *The only nonzero monomials in $Q(z)$ are the constant term z^0 and $z^{m+1}, \ldots, z^{m+n}$.*
3. *The leading coefficient has norm at least $\left(\frac{m+n-1}{n-1}\right)^{n-1}$.*
4. *$Q(z) \in [0,1]$ for $z \in [0,1]$.*
5. *The signs of the coefficients of $Q(z)$ are alternating.*

Proof. The first point is immediate. For Property (2), observe that degree of Q is at most $m+n$ and furthermore that, for $0 < k \leq m$,

$$[z^k]Q(z) = [z^k]\sigma_m(z)(1-z)^n = [z^k]\frac{(1-z)^n}{(1-z)^n} = 0.$$

For Property (3), we use Proposition D.1 to observe that

$$\begin{aligned} |[z^{m+n}]Q(z)| &= [z^m]\frac{1}{(1-z)^n} = \binom{m+n-1}{n-1} \\ &= \left(1+\frac{m}{n-1}\right)\left(1+\frac{m}{n-2}\right)\cdots\left(1+\frac{m}{1}\right) \geq \left(1+\frac{m}{n-1}\right)^{n-1}. \end{aligned}$$

For Property (4), observe that the coefficients of the power series of $\left(\frac{1}{1-z}\right)^n$ are all nonnegative, so for $z \in [0,1)$

$$0 \leq (1-z)^n\sigma_m(z) < (1-z)^n\frac{1}{(1-z)^n} \leq 1.$$

and furthermore that $Q(1) = 0$. Finally, to infer Property (5), observe that Q has exactly $n+1$ monomials and a positive root of multiplicity n at $z = 1$. Apply Descartes' Rule of Signs, which tells us that the number of sign changes is equal to the number of positive roots with multiplicity. □

D.2 Hard exponential mixtures from Turán polynomials

We remind ourselves of some tools from information theory. For a pair of probability measures P and Q corresponding to densities p and q, their χ^2-*divergence* is defined by

$$\chi^2(P,Q) := \int \frac{(p(x)-q(x))^2}{q(x)}dx.$$

We recall the following well-known fact relating Hellinger and χ^2-divergence; we include the proof for the reader's convenience.

Proposition D.3.

$$H^2(P,Q) \leq \frac{1}{2}\chi^2(P,Q)$$

Proof. Directly, we have

$$H^2(P,Q) = \frac{1}{2}\int(\sqrt{q(x)}-\sqrt{p(x)})^2dx = \frac{1}{2}\int\frac{(q(x)-p(x))^2}{(\sqrt{q(x)}+\sqrt{p(x)})^2} \leq \frac{1}{2}\int\frac{(q(x)-p(x))^2}{q(x)}dx.$$

□

This fact will be useful in obtaining the following estimate, which is tight in terms of the exponent of the gap parameter Δ.

Theorem D.4. *For any positive integers m, n such that $m > 2n$, let $\Delta = 1/(m + 2n + 1)$. Fix $\alpha \in (0, 1/2)$. Then there exists two mixtures of exponentials P_1, P_2 with CDFs defined on $[0, \infty)$ by*

$$1 - F_1(t) = (1 - \alpha + a_0)e^{-\nu t} + \sum_{j=1}^{n} a_j e^{-\lambda_j t}$$

and

$$1 - F_2(t) = (1 - \alpha)e^{-\nu t} + \sum_{j=1}^{n} b_j e^{-\mu_j t}$$

such that

1. *(Normalization) $\nu, \lambda_j, \mu_j \in (0, 1]$.*
2. *(Exponents are well-separated) All of the elements of the set $\nu \cup \{\lambda_j\} \cup \{\mu_j\}$ are separated by at least Δ. Furthermore, the sets $\{\nu\}, \{\lambda_j\}$, and $\{\mu_j\}$ are disjoint and interlaced, e.g. $\nu < \lambda_1 < \mu_1 < \lambda_2 < \mu_2 < \cdots < \lambda_n < \mu_n$.*
3. *(Coefficients are bounded) $\sum_j a_j = \alpha$ and $\sum_j b_j = \alpha$.*
4. *(Indistinguishability) $H^2(P_1, P_2) \leq \frac{\alpha^2}{(2n-1)^2}\left[\Delta(2n - 1)\right]^{4n-4}$.*

Proof. Let $Q_{m,2n}$ be the $(m, 2n)$-th Turań polynomial. We start by re-centering it so that its average value over the unit interval is zero, by writing

$$R_{m,2n}(x) := Q_{m,2n}(x) - \int_0^1 Q_{m,2n}(y)dy.$$

Let $C_{m,2n}$ be the sum of the positive entries of $Q_{m,2n}$. By Property 3 of Lemma D.2, we have $C_{m,2n} \geq [(m + 2n - 1)/(2n - 1)]^{2n-1}$. Integrate $R_{m,2n}$ to give the polynomial

$$S_{m,2n}(x) := \int_0^x R_{m,2n}(y)dy$$

and observe that $S_{m,2n}(0) = 0$ and $S_{m,2n}(1) = \int_0^1 Q_{m,2n}(y)dy - \int_0^1 Q_{m,2n}(y)dy = 0$. Define $C'_{m,2n}$ to be the sum of the positive entries of $S_{m,2n}(x)$. Observe, by the power rule for integrating monomials, that

$$C'_{m,2n} \geq C_{m,2n}/(m + 2n + 1) \geq \frac{[(m + 2n - 1)/(2n - 1)]^{2n-2}}{4n - 2}.$$

Define $f(x) = \alpha S_{m,2n}(x)/C'_{m,2n}$, so

$$f(x) = f_1 x + \sum_{k=m+2}^{m+2n+1} f_k x^k$$

and the sum of all of the positive (or all of the negative) coefficients of f is α. The coefficients here will be exactly the a_j's and b_j's, which verifies property (3).

We now split the polynomial into a positive part and a negative part. Define $f_+(x) = (1 - \alpha + f_1)x + \sum_{k=m+2; f_k>0}^{m+2n+1} f_k x^k$ and $f_-(x) = (1-\alpha)x + \sum_{k=m+2; f_k<0}^{m+2n+1}(-f_k)x^k$ so that

$$f(x) = f_+(x) - f_-(x)$$

and define CDFs of mixtures of exponentials F_1, F_2 by

$$F_1(t) := 1 - f_+(e^{-t}), \qquad F_2(t) := 1 - f_-(e^{-t}).$$

Observe by the chain rule that

$$p_1(t) := \frac{d}{dt}F_1(t) = f_+'(e^{-t})e^{-t}, \qquad p_2(t) := \frac{d}{dt}F_2(t) = f_-'(e^{-t})e^{-t}$$

so by linearity of the derivative,

$$p_1(t) - p_2(t) = e^{-t}(f_+'(e^{-t}) - f_-'(e^{-t})) = e^{-t}f'(e^{-t}).$$

Observe that for $x \in [0,1]$

$$|f'(x)| = \frac{\alpha}{C'_{m,2n}}|S'_{m,2n}(x)| = \frac{\alpha}{C'_{m,2n}}|R_{m,2n}(x)| = \frac{\alpha}{C'_{m,2n}}\left|Q_{m,2n}(x) - \int_0^1 Q_{m,2n}(y)dy\right| \le \frac{\alpha}{C'_{m,2n}}.$$

In the last step, we used Property 4 of Lemma D.2; the quantity inside the absolute value is the difference between $Q_{m,2n}$ and its average over $[0,1]$, which implies that the difference is bounded in absolute value by 1.

Finally, notice that $f_-'(x) \ge (1-\alpha) \ge 1/2$ for all $|x| \ge 0$. If P_1 and P_2 are measures having densities p_1 and p_2, then

$$\begin{aligned}\chi^2(P_1, P_2) = \int_t \frac{(p_1(t) - p_2(t))^2}{p_2(t)}dt &= \int_t e^{-t}\frac{(f_+'(e^{-t}) - f_-'(e^{-t}))^2}{f_-'(e^{-t})}dt \\ &\le \int_t e^{-t}\frac{(\alpha/C'_{m,2n})^2}{(1/2)}dt \\ &\le 2\alpha^2(4n-2)^2\left[\frac{2n-1}{m+2n-1}\right]^{4n-4}.\end{aligned}$$

Therefore,

$$H^2(P_1, P_2) \le \frac{1}{2}\chi^2(P_1, P_2) \le \alpha^2(4n-2)^2\left[\frac{2n-1}{m+2n-1}\right]^{4n-4} \le \alpha^2(4n-2)^2\left[\Delta(2n-1)\right]^{4n-4}.$$

This computation verifies condition (4). Finally, we re-scale t by making the transformation $t \mapsto \frac{t}{m+2n+1}$. By Properties 2 and 5 of Lemma D.2, these mixtures satisfy (1) and (2). □

Corollary D.5. *Let P_1 and P_2 be as in Theorem D.4, $\alpha \in (0, 1/2)$ be arbitrary, and let $k = n+1$ be the number of components in each of the two mixtures. Let $P^{\otimes N}$ denote the product measure corresponding to taking N iid samples from probability measure P. Then*

$$\mathbf{TV}(P_1^{\otimes N}, P_2^{\otimes N}) \le \frac{\alpha\sqrt{2N}}{2k-3}\left[\Delta(2k-3)\right]^{2k-4}.$$

Proof. We use the comparison inequality between TV and H^2, together with the tensorization inequality for H^2 to conclude

$$\mathbf{TV}(P_1^{\otimes N}, P_2^{\otimes N}) \leq \sqrt{2H^2(P_1^{\otimes N}, P_2^{\otimes N})} \leq \sqrt{2NH^2(P_1, P_2)}$$

and then apply Theorem D.4. □

Note that if we take m to be much larger than k, Δ is much smaller than $2k$. Informally speaking, Corollary D.5 translates to a sample complexity lower bound of

$$\Omega\left(\frac{k^2}{\alpha^2}\left(\frac{1}{\Delta}\right)^{4k-4}\right)$$

in the hypothesis testing problem of distinguishing between P_1 and P_2 for any choice of α. Compare this to Equation 5 of Theorem 2.6, which gives an upper bound of

$$O\left(\frac{1}{p_{leading}^2}(c_\Delta/\Delta)^{4k-2}\log(2/\alpha)\right)$$

where $c_\Delta = 8e^2/\left(\min(1/\Delta_{leading}, (2k-1))\right)$ for simple-versus-composite hypothesis testing at a constant significance level. This shows the matching the quadratic dependence on the coefficients as well as the exponent $4k$ in the gap parameter.

For comparison, we also juxtapose the result from Theorem 2.3, which gives an upper bound of

$$O\left(\frac{k^7}{p_{\min}^4}\left(\frac{2e}{\Delta}\right)^{4k}\right)$$

for the learning problem, which gives a quartic dependence on the coefficients instead. We suspect this difference is not simply an artifact of the analysis; in practice, the convex programming approach seems to succeed with fewer samples than the MPM.

Finally, we can also give a lower bound on the minimax rate of learning the coefficients of a mixture of exponentials:

Theorem D.6. *Let $k > 3$, $m > k$ and $\Delta = 1/(m+k-1)$. There exists $\lambda_1, \ldots, \lambda_k$ which are Δ-separated such that for any estimator $\hat{p}$ of the coefficients from N samples of the mixture of exponentials with CDF $F(t) = 1 - \sum_j p_j e^{-\lambda_j t}$:*

$$\inf_{\hat{p}} \max_{p} \mathbb{E}_p \|p - \hat{p}\|_1 \geq \frac{1}{4}\min\left(1, \frac{k-3}{\sqrt{2N}}\left(\frac{1}{\Delta(k-3)}\right)^{k-4}\right)$$

Proof. This is derived by the usual reduction to hypothesis testing. Let $\alpha \in (0, 1/2)$ to be fixed later, and let P_1 and P_2 be the mixtures of exponentials from Corollary D.5 with coefficients $\{a_i\}$ and $\{b_j\}$ (supported at disjoint indices) with parameter α. As a reminder, Corollary D.5 gives (where we write k' as a placeholder for k)

$$\mathbf{TV}(P_1^{\otimes N}, P_2^{\otimes N}) \leq \frac{\alpha\sqrt{2N}}{2k'-3}\left(\Delta(2k'-3)\right)^{2k'-4}.$$

Now let

$$\alpha := \min\left(\frac{1}{2}, \frac{2k-3}{2\sqrt{2N}}\left(\frac{1}{\Delta(2k'-3)}\right)^{2k'-4}\right)$$

so $\mathbf{TV}(P_1, P_2) \leq 1/2$. This means we can couple the distributions so that they draw identical outputs with probability at least 1/2. Let $k = 2k'$. Under such a coupling, for at least one of P_1 or P_2, its coefficients p are far from $\hat{p}$:

$$\|\hat{p} - p\|_1 \geq \frac{1}{2}(a_0 + \sum_{j=1}^{k} |a_j - b_j|) = \alpha$$

with probability at least 1/4, so the expected error is at least $\alpha/4$. For the last equality in the above expression, we again keep in mind that the coefficients a_j and b_j are supported on disjoint sets of indices. □

D.3 Equally-spaced exponents via Chebyshev Polynomials

In this section, we briefly discuss a slightly different family of polynomials, which are classically attributed to Chebyshev. It will turn out that they give a similar exponential-type bound for learning the λ_j's when the gap Δ is as large as possible. We remind ourselves of their definition now:

Definition 3. The kth Chebyshev polynomial (of the first kind) T_k is given by the recursive relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

starting with $T_0(x) = 1$ and $T_1(x) = x$.

These polynomials have the property that for any k,

1. $|T_k(x)| \leq 1$ for all $x \in [-1, 1]$.
2. If k is even, the monomials that appear in T_k are all even powers of x up to k. If k is odd, then the monomials are all odd powers of x up to k.
3. The coefficient of x^k is 2^{k-1} for $k \geq 1$.

Note that Property 3 of T_k is comparable to Property 3 of $Q_{m,n}$ in Lemma D.2. Of particular interest here is Property 2 of T_k, which states that the exponents that appear are equally spaced. Indeed, the same technique of normalizing then re-weighting the coefficients by an arbitrary $\alpha \in (0, 1/2)$ used in the proof of Theorem D.4 yields the following lower bound:

Theorem D.7. *Fix any positive integer k, and fix $\alpha \in (0, 1/2)$. There exist two mixtures of exponentials P_1, P_2 with CDFs defined on $[0, \infty]$ by*

$$1 - F_1(t) = (1 - \alpha + a_0)e^{-\nu t} + \sum_{j=1}^{k} a_j e^{-\lambda_j t}$$

and

$$1 - F_2(t) = (1 - \alpha)e^{-\nu t} + \sum_{j=1}^{k} b_j e^{-\mu_j t}$$

such that

1. *(Normalization) $\nu, \lambda_j, \mu_j \in (0, 1]$.*

2. *(Well-separated exponents) All of the elements of the set* $\nu \cup \{\lambda_j\} \cup \{\mu_j\}$ *are separated by at least* $1/(2k)$.

3. *(Bounded coefficients)* $\sum_j \alpha_j = \alpha$ *and* $\sum_j b_j = \alpha$.

4. *(Indistinguishability)* $H^2(P_1, P_2) \leq \frac{\alpha^2(2k+1)^2}{2^{4k-2}}$.

E Hypothesis Testing using the Nazarov-Túran Lemma

In this section we use the Nazarov-Túran lemma, originating in works on analytic number theory and approximation theory, to give guarantees on the number of samples needed for the Kolmogorov-Smirnov test described in Section 2.4 to successfully solve the simple-vs-composite hypothesis testing Problem 1.

As a first step towards proving Theorem 2.6, we now recall our key technical tool, the Nazarov-Turán Lemma[10].

Theorem E.1 (Nazarov-Turán [20])**.** *Suppose* $p(t) = \sum_{j=1}^{n} c_j e^{-\lambda_j t}$. *Then for* $t_1, t_2 > 0$

$$\max_{t \in [-t_2, t_1]} |p(t)| \leq e^{(t_1+t_2)\lambda_1} [4e(t_1 + t_2)/t_1]^{n-1} \max_{t \in [0,t_1]} |p(t)|.$$

We will use the Nazarov-Turán lemma by analytically continuing the mixture of exponential CDFs to negative time; this allows us to derive the following lemma which is the heart of our hypothesis testing result.

Lemma E.2. *Suppose* $p(t) = \sum_{j=1}^{k_0} c_j e^{-\lambda_j t} - \sum_{j=1}^{k_1} d_j e^{-\mu_j t}$ *such that the* $c_j, d_j \geq 0$ *and* $\sum_j c_j, \sum_j d_j \leq 1$. *Also suppose that* $1 \geq \lambda_1 > \cdots > \lambda_{k_0} \geq 0$ *and similarly for the* μ_j. *Then if* $k := \frac{k_0+k_1}{2}$ *and* $\delta := \min(|c_1|, |d_1|)$

$$\frac{\delta}{2(8e^2 \log(2/\delta))^{2k-1}} \cdot \min\left(1, ((2k-1)(\lambda_1 - \mu_1))^{2k-1}\right) \leq \max_{t \in [0, 2k-1]} |p(t)|$$

Proof. Without loss of generality we suppose that $\lambda_1 > \mu_1$. By Theorem E.1, for any t_1, t_2 we have

$$|p(-t_2)| \leq e^{(t_1+t_2)\lambda_1} [4e(t_1 + t_2)/t_1]^{2k-1} \max_{t \in [0,t_1]} |p(t)|.$$

To use this, we will need to choose a t_2 so that we can get an explicit lower bound on $p(-t_2)$. It will suffice if

$$|c_1 e^{\lambda_1 t_2}|/2 \geq |\sum_{j=1}^{k_1} d_j e^{\mu_j t_2}|$$

or, phrased slightly differently:

$$|c_1|/2 \geq |\sum_{j=1}^{k_1} d_j e^{(\mu_j - \lambda_1)t_2}| \tag{21}$$

Observe that all the exponents on the right hand side have negative coefficients, because λ_1 is the largest. In order to bound the right hand side, we use Holder's Inequality:

$$|\sum_{j=1}^{n} d_j e^{(\mu_j - \lambda_1)t_2}| \leq e^{(\mu_1 - \lambda_1)t_2} \sum_{j=1}^{k_1} |d_j| \leq e^{(\mu_1 - \lambda_1)t_2}.$$

[10] Here we have specialized the Nazarov-Turán lemma to real exponents and to intervals instead of general measurable sets.

Therefore, taking $t_2 = \frac{\log 2/|c_1|}{\lambda_1 - \mu_1}$ suffices for (21) to hold. We then have

$$
\begin{aligned}
|c_1 e^{\lambda_1 t_2}|/2 &\le e^{(t_1+t_2)\lambda_1}[4e(t_1+t_2)/t_1]^{2k-1} \max_{t\in[0,t_1]} |p(t)| \\
\implies |c_1|/2 &\le e^{t_1\lambda_1}[4e(t_1 + \frac{\log(2/|c_1|)}{\lambda_1 - \mu_1})/t_1]^{2k-1} \max_{t\in[0,t_1]} |p(t)|.
\end{aligned}
$$

which gives a lower bound on the maximum value of $|p(t)|$:

$$
\frac{|c_1|}{2} e^{-t_1\lambda_1} \left[\frac{t_1}{4e(t_1 + \frac{\log(2/|c_1|)}{\lambda_1-\mu_1})} \right]^{2k-1} \le \max_{t\in[0,t_1]} |p(t)|.
$$

Letting $t_1 = (2k-1)$ and using that $|\lambda_1| \le 1$, we find

$$
\frac{|c_1|}{2(4e^2)^{2k-1}} \left[\frac{2k-1}{(2k-1) + \frac{\log(2/|c_1|)}{\lambda_1-\mu_1}} \right]^{2k-1} \le \max_{t\in[0,2k-1]} |p(t)| \tag{22}
$$

Using the inequality that $x + y \le 2\max(x,y)$, we bound the second term on the LHS as follows:

$$
\begin{aligned}
\left[\frac{2k-1}{(2k-1) + \frac{\log(2/|c_1|)}{\lambda_1-\mu_1}} \right]^{2k-1} &\ge \frac{1}{2^{2k-1}} \left[\frac{2k-1}{\max\left(2k-1, \frac{\log(2/|c_1|)}{\lambda_1-\mu_1}\right)} \right]^{2k-1} \\
&\ge \frac{1}{2^{2k-1}} \min\left(1, \frac{(2k-1)(\lambda_1-\mu_1)}{\log(2/|c_1|)}\right)^{2k-1} \\
&\ge \frac{1}{(2\log(2/|c_1|)^{2k-1}} \min\left(1, ((2k-1)(\lambda_1-\mu_1))^{2k-1}\right)
\end{aligned} \tag{23}
$$

Using Equation 23, we simplify the lhs of (22) to give

$$
\frac{|c_1|}{2(8e^2\log(2/|c_1|))^{2k-1}} \cdot \min\left(1, ((2k-1)(\lambda_1-\mu_1))^{2k-1}\right) \le \max_{t\in[0,2k-1]} |p(t)|.
$$

□

We are now ready to prove the finite sample guarantee (Theorem 2.6) for the Kolmogorov-Smirnov test from Section 2.4. For the reader's convenience, we re-state the test now:

1. Let $\alpha > 0$ be the significance level.
2. Let F_N be the empirical CDF and let F be the CDF under the null hypothesis H_0.
3. Reject H_0 if $\sup_t |F_n(t) - F(t)| > \sqrt{\log(2/\alpha)/2N}$.

Proof of Theorem 2.6. Let F be the CDF under the null hypothesis, that the samples are drawn from P. By the DKW inequality (Theorem C.7), under the null hypothesis

$$
\Pr_P(\sup_t |F_N(t) - F(t)| > s) \le 2e^{-2Ns^2}
$$

Therefore with $s = \sqrt{\log(2/\alpha)/2N}$ then the probability of type I error is bounded by α, as desired. On the other hand, observe by Lemma E.2 that if F_Q is the CDF of $Q \in H_1$, then

$$\frac{\delta}{2(8e^2)^{2k-1}} \cdot \min\left(1, ((2k-1)(\epsilon))^{2k-1}\right) \leq \max_{t\in[0,2k-1]} |F(t) - F_Q(t)|$$

It follows that the test is guaranteed to reject as long as

$$\frac{\delta}{2(8e^2)^{2k-1}} \cdot \min\left(1, ((2k-1)(\epsilon))^{2k-1}\right) - \sup_t |F_N(t) - F_Q(t)| > \sqrt{\log(2/\alpha)/2N}. \tag{24}$$

Under the alternative hypothesis, by the DKW Inequality (Theorem C.7). Equation 24 happens with probability

$$\begin{aligned}&\Pr\left(\sup_t |F_N(t) - F_Q(t)| < \frac{\delta}{2(8e^2)^{2k-1}} \cdot \min(1, ((2k-1)\epsilon)^{2k-1}) - \sqrt{\log(2/\alpha)/2N}\right)\\ \geq& 1 - 2\exp\left(-\frac{N\delta^2 \min(1, ((2k-1)\epsilon)^{4k-2})}{8(8e^2)^{4k-2}}\right)\end{aligned}$$

since $\sqrt{\log(2/\alpha)/2N} < \frac{\delta\cdot\min(1,((2k-1)\epsilon)^{2k-1})}{4(8e^2)^{2k-1}}$ under the assumption $N \geq N_0$. □

Remark 6. The point of the above argument is to derive finite sample guarantees. In the asymptotic regime, one can immediately derive more precise results by combining the guarantee from Lemma E.2, applied to the difference of CDFs, with the classical version of the Kolmogorov-Smirnov test, using the critical values for the Kolmogorov distribution.

F A reconstruction algorithm for population history

In this section, we describe an algorithm that takes as input L independent and identically distributed 2-sample coalescence times $c_1, \ldots c_L$ and outputs a population shape. In particular, we assume that each sample is an accurate measurement of coalescence times, with the interpretation that they are i.i.d. from the distribution of T (Section 1.3.1).

Our strategy is reminiscent of the algorithm provided by [14] for the 1-component case. We will be reconstructing the history iteratively, starting from the most recent event and going as far backwards in time as possible. The hope is that if we have a guarantee that we have accurate constructions of the subpopulation structures in the intervals $I_1, \ldots I_j$, then we should be able to provide a good guess of what happened at the boundaries between I_{j-1} and I_j at time t_{j-1}.

F.1 Preliminary ideas

F.1.1 Inferring events – split, merge or size change?

Assume, for the sake of argument, that we are given the exact population shape $\mathbf{N}$ for the interval I_{j-1}. Since our model only allows for up to one merge or one split at any given time, the only possibilities are $D_j \in \{D_{j-1} - 1, D_{j-1}, D_{j-1} + 1\}$. Therefore, the maximal number of components in the exponential mixture in I_j including the constant term is $D_{j-1} + 2$.

In the noiseless scenario, we can run the matrix pencil method with component size $k = D_{j-1} + 2$ on estimates $v_0, \ldots, v_{2D_{j-1}+3}$ describing CDF values on I_j. In this case, we get some collection of k eigenvalues $\alpha_0 > \cdots > \alpha_{D_{j-1}+1}$. If in reality $D_j < D_{j-1} + 1$, then 0 must be an eigenvalue. Therefore, we can discard α_i that are equal to zero to get the true collection of components.

In the noisy case, by Theorem 2.3, given a sufficiently large collection of L iid samples that lie in I_j and to its right, we can discard α_i that are close to zero: those that are smaller than some error threshold[11], $\eta_1 = \Theta\left(\left(\frac{2e}{\Delta}\right)^{2k}\sqrt{\frac{1}{L}}\right)$. Just as in Theorem 2.3, Δ is a lower bound on the gap between any two distinct exponents in the model. For the reconstruction algorithm, it is a free parameter that determines the number of samples required. Conversely, it is the threshold for learning given the number of samples available.

Once this is done, all that's left is to semantically link the subpopulations in I_{j-1} to those of I_j. While it is tempting to simply index subpopulations in decreasing order of α's and say that corresponding subpopulations are matching, the labeling becomes inconsistent, for instance, under consideration of large size changes or under merges/splits. We address this issue in the upcoming section, Appendix F.1.2.

F.1.2 The recursion

We briefly recall the model derived in Appendix A. If the population history $\mathbf{N}$ is constant in the interval $I = [a, b]$, then T satisfies

$$\Pr(T > a + t \mid T > a) = p_0 + \sum_{\ell=1}^{D} p_\ell e^{-\lambda_\ell t},$$

where for each $\ell \geq 1$, $p_\ell := \Pr(\mathcal{E}_{\ell\ell} \mid T > a)$.

Since our algorithm works iteratively, it would be helpful to relate the exponential mixture parameters of the interval I_j to those of I_{j-1}, assuming $\mathbf{N}$ is constant in $I_j = [b, c]$ and in $I_{j-1} = [a, b]$, for $a < b < c$. *The overall goal here is to describe a "matching" scheme for subpopulations of I_{j-1} to those of I_j.* For the sake of brevity, we introduce some notation.

- Let $n = D_j$ and $m = D_{j-1}$.
- For $i = 1$ to D_j, let $\mathcal{E}_i^j$ denote the event that both lineages trace back to subpopulation indexed i in the interval I_j.
- Let $\{(p_i, \alpha_i)\}_{i=1}^n$ denote the coefficients and exponentials ($\alpha_i = e^{-\lambda_i}$) for the interval I_j.
- Let $\{(q_i, \beta_i)\}_{i=1}^m$ denote the analogous parameters for I_{j-1}.

We only need to consider subpopulations i in I_j which is linked to subpopulation i' in I_{j-1}, in the absence of splits or merges specifically involving i or i'. Indeed, since we only allow one split or merge at a time, either of these events can be inferred from the fact that all but at most two subpopulations from I_j can be "matched" with subpopulations in I_{j-1} if a split/merge occurred.

Observe that $\mathcal{E}_i^j = \mathcal{E}_{i'}^{j-1}$, and that $\{T > b\} \subset \{T > a\}$ as events. We derive the following:

$$\begin{aligned}
p_i &= \Pr(\mathcal{E}_i^j \mid T > b) \\
&= \Pr(\mathcal{E}_{i'}^{j-1} \wedge \{T > a\} \mid T > b) \\
&= \Pr(T > b \mid \mathcal{E}_{i'}^{j-1} \wedge \{T > a\}) \cdot \frac{\Pr(\mathcal{E}_{i'}^{j-1} \wedge \{T > a\})}{\Pr(T > b)} \\
&= \Pr(T > b \mid \mathcal{E}_{i'}^{j-1} \wedge \{T > a\}) \cdot \Pr(\mathcal{E}_{i'}^{j-1} \mid T > a) \cdot \frac{\Pr(T > a)}{\Pr(T > b)} \\
&= \left(\beta_{i'}^{b-a}\right) \cdot (q_{i'}) \cdot \frac{1}{\Pr(T > b \mid T > a)}
\end{aligned}$$

[11] Note that we provide the exact constants in the full proof of Theorem 2.3 found in Appendix C.

In particular, if we can accurately estimate $\hat{p}, \hat{\alpha}, \hat{q}, \hat{\beta}$ from coalescent samples $t_1, \ldots, t_L$, we can infer that population i links to population i' if

$$\hat{p}_i \cdot \frac{\#\{k : t_k > b\}}{\#\{k : t_k > a\}} \approx \hat{q}_{i'} \hat{\beta}_{i'}^{b-a} \tag{25}$$

where $\approx$ denotes an approximation up to some additive error η_2. To determine what η_2 should be, we analyze both sides of Eq. (25) via the bounds from Theorem 2.3.

Let L_{j-1} and L_j respectively denote the number of samples that fall in the intervals $[a, \infty)$ and $[b, \infty)$. For the RHS, note that if the additive errors ϵ_1 and ϵ_2 for $q_{i'}$ and $\beta_{i'}$ respectively are small, then the error of the right-hand side of Eq. (25) is bounded by $\beta\epsilon_1 + q\epsilon_2 + \epsilon_1\epsilon_2 \leq \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2 = O\left(\left(\frac{2e}{\Delta}\right)^{3D_{j-1}} \sqrt{\frac{1}{L_{j-1}}}\right)$. The same analysis goes for the LHS. In particular, the error for $\Pr(T > b \mid T > a)$, by the DKW inequality, is $\sqrt{\frac{1}{L_{j-1}} \log \frac{2}{\delta_0}}$ with probability δ_0, which is dominated by the error $O\left(\left(\frac{2e}{\Delta}\right)^{3D_j} \sqrt{\frac{1}{L_j}}\right)$ in p_i. Therefore, the tolerance should be $\eta_2 = \Theta\left(\left(\frac{2e}{\Delta}\right)^{3D_j} \sqrt{\frac{1}{L_j}}\right)$.

F.2 The algorithm description

Now we implement the proposed strategy, which is built upon the ideas of the previous section (Appendix F.1).

Inputs: $\{c_1, \ldots, c_L\}$, a sample collection of L iid coalescence times.

Parameters:

- D_0, an initial upper bound for the number of subpopulations at $t = 0$.
- K, the number of intervals for reconstruction.
- δ, failure probability budget.
- ϵ, a scaling parameter for the interval sizes.
- N, the present-day ($t = 0$) total population size.
- P, threshold lower bound for component weight.
- Δ, threshold lower bound for exponential gap over all intervals.

Output: A population history $\{(I_j, D_j, \mathbf{N}_j, \mathcal{E}_j)\}_{j=1}^{K}$.

Procedure:

1. Partition time (oriented towards the past) into intervals by setting $t_j = \epsilon j N$, for $j = 0, \ldots, K$. Let $I_1 = [t_0, t_1]$, $\ldots$, $I_K = [t_{K-1}, t_K]$. Initialize $j = 1$.

2. **Collect CDF statistics in I_j:** Abbreviate $D := D_{j-1}$. For $\ell = 0, 1, \ldots, 2D + 3$, compute the statistic

$$\hat{v}_\ell = \frac{\#\{i : c_i \geq t_{j-1} + \frac{\epsilon N}{2D+4}\ell\}}{\#\{i : c_i \geq t_{j-1}\}}$$

3. **Learn the model parameters in I_j**: apply the Matrix Pencil Method on $D+2$ components using inputs $\hat{v}_0, \dots, \hat{v}_{2D+3}$, which outputs eigenvalues $\hat{\alpha}_0 > \cdots > \hat{\alpha}_{D+1}$ and corresponding coefficients $\hat{p}_0, \hat{p}_1, \dots, \hat{p}_{D+1}$.

   Discard all $\hat{\alpha}$'s (and corresponding $\hat{p}$'s) of absolute value at most η, where

$$\eta_1 = \frac{4(D+2)^{3.5}}{P^2} \left(\frac{2e}{\Delta}\right)^{2D+4} \sqrt{\frac{1}{Le^{-\epsilon jN}} \log \frac{K}{\delta}}. \tag{26}$$

   Set D_j equal to the number of leftover $\hat{\alpha}$'s, minus 1 (to account for the constant term), and re-normalize the remaining $\hat{p}$'s so that they sum to 1.

4. **Learn the population event at t_{j-1} & sizes in I_j**: If $j = 1$, let $\hat{\mathbf{N}}_1 = (N\sqrt{\hat{p}_1}, \dots, N\sqrt{\hat{p}_{D_1}})$. (The leading coefficient $\hat{p}_0$ is omitted from this calculation.)

   Otherwise ($j > 1$), set $\hat{\mathbf{N}}_j = (\frac{\epsilon N}{(2D+4)\hat{\mu}_1}, \dots, \frac{\epsilon N}{(2D+4)\hat{\mu}_{D_j}})$, where $\hat{\mu}_i = -\log(\hat{\alpha}_i)$. Let $\{(\hat{q}_i, \hat{\beta}_i)\}_{i=0}^{D_{j-1}}$ be the (coefficient, eigenvalue) pairs recovered from the previous iteration.

   Say that subpopulation i of I_j *matches* subpopulation i' of I_{j-1} if

$$\left| \hat{p}_i - \hat{q}_{i'} \hat{\beta}_{i'}^{2D+4} \cdot \frac{\#\{k : t_k > a\}}{\#\{k : t_k > b\}} \right| < \eta_2 \tag{27}$$

   where

$$\eta_2 = \frac{40(D+2)^5}{P^2} \left(\frac{2e}{\Delta}\right)^{3(D+2)} \sqrt{\frac{1}{Le^{-\epsilon jN}} \log(K/\delta)}. \tag{28}$$

   Do one of the following.

   - If $D_j = D + 1$, infer "Split": Find a valid matching. There should be one $\ell \in [D_{j-1}]$ and two indices $k_1, k_2 \in [D_j]$ left over. Assign the event: $\text{Split}(\ell \to \{k_1, k_2\})$.
   - If $D_j = D - 1$, infer "Merge": Find a valid matching. There should be two $\ell_1, \ell_2 \in [D_{j-1}]$ and one $k \in [D_j]$ left over. Assign the event: $\text{Merge}(\{\ell_1, \ell_2\} \to k)$.
   - If $D_j = D$, infer "Change Size": Identify a bijection from $[D_j]$ to $[D_{j-1}]$ via the matching scheme.
   - If none of the above cases are true, or if there is no such prescribed matching, fail by default and terminate the procedure.

5. If $j = K$, stop. Otherwise, set $j \leftarrow j + 1$, go back to step 2.

F.3 Reconstruction Guarantees

Consider the setting where the intervals are known and are equal in size (which decides ϵ). Given enough samples, the algorithm for population reconstruction presented in Section F succeeds in reconstructing the correct population history with high probability. Indeed, with a large enough number of samples, the conditional tail distribution $F(t) = \Pr(T > t \mid T \geq t_j)$ can be approximated to arbitrarily high precision with high probability. In turn, this means that the empirical (conditional) tail probabilities in I_j that are collected in step (2) of the algorithm are arbitrarily close to the ground truth tail probabilities.

We choose $2(D+2)$ maximally spread out points in the interval I_j, and input the empirical tail probabilities at these points to the Matrix Pencil Method. The *robustness* of the Matrix Pencil

Method (Theorem 2.3) guarantees that the estimated exponents $\hat{\mu}_i$ are sufficiently close to their true values μ_i with high probability. Furthermore, each true exponent μ_i is roughly the reciprocal of N_i, the size of the i^{th} subpopulation in the interval I_j, and so approximating the μ_i's gives us a good approximation of the subpopulation sizes in the interval I_j.

Note the scaling factor $\nu = \frac{\epsilon N}{(2D+4)}$ that appears in the algorithm, which comes from the change of variables $t = \nu\tau$. The rescaling reflects a time units conversion from *generations* to *coalescent units*, which is convenient because the robustness guarantees are provided via interpolation at integer points $\tau = 0, 1, 2$, etc. The precise relation between the μ_is and N_is in our algorithm is

$$\mu_i = \frac{\epsilon N}{(2D+4)N_i}.$$

Steps 3 and 4 contain errors η_1, η_2 with correction terms $Le^{-\epsilon jN}$ and K/δ. The first term $Le^{-\epsilon jN}$ is the expected number of samples (out of L total) found inside and to the right of the interval I_j. By the observations made in Appendix F.1 and Appendix F.1.2, Steps 3 and 4 of the algorithm provide accurate reconstructions of $\mathbf{N}$ for I_j with probability δ/K, given sufficiently large samples. By union bounding over all K intervals, the algorithm accurately reconstructs all K pieces of $\mathbf{N}$ with probability δ.

Remark 7. It is almost always the case, however, that we do not know the interval endpoints, and do not know a small enough value of ϵ for which the algorithm will work. According to the model, there exists a sufficiently small value of ϵ which will allow for correct inference, where "sufficiently small" means $I_1, \ldots, I_K$ in Step 1 of the algorithm is fine-grained enough to capture all intervals. Therefore, a natural strategy is to try a search-based approach for an optimal value, as follows. Choose various values of ϵ, then learn the parameters of the model for each. From these, we may choose the best ϵ via a goodness-of-fit test, by taking the ϵ that outputs a learned distribution $\hat{f}$ that is closest to the empirical distribution f_{emp} in total variation distance. Note that f_{emp} converges to the true f as the number of samples increases, for some appropriate ϵ.

G Simulation Methods

G.1 A convex programming approach to learning

In addition to using the Matrix Pencil Method to learn mixtures of exponentials in each interval, we also implemented a convex program. Here, the goal is to learn a mixture of exponentials, whose support is perhaps restricted to an interval $I = [a, b]$. The idea is as follows: assume that we know the interval $\Lambda = [0, c]$ for which we can assume $\lambda_1, \ldots, \lambda_n \in \Lambda$. We first discretize the space of possible exponents by choosing n equally spaced points $\lambda_1, \ldots, \lambda_n$ inside Λ. Solve the convex program

$$\begin{aligned} \underset{\vec{p}}{\text{minimize}} \quad & \sup_{t\in I} \left| \sum_{i=1}^{n} p_i e^{\lambda_i t} - v_t \right| \\ \text{subject to} \quad & \sum_i p_i = 1 \\ & p_i \geq 0, i = 1, \ldots, n. \end{aligned}$$

In practice, we replace $\sup_{t\in I}$ with the discretization $\max_{t\in\mathcal{S}}$, where $\mathcal{S} \subset I$ is a finite mesh of points in I. Since we are learning from samples, we also substitute v_t with $\widetilde{v}_t$, the empirical estimate of the tail probability $\Pr[T = t \mid T \geq a]$. Since the ℓ_1-norm of the p_i is fixed to be 1, we do not expect to need additional regularization to get sparse output.

For small instances (see Section 3), the convex program is more sample-efficient than the Matrix Pencil Method. In the context of this paper, however, it does not come with robustness guarantees. Results on convex programming approaches for super-resolution are known, due to Candès and Fernandez-Granda [4]; for our (real-exponent) setting, a different analysis will be required and we leave this to future work. If we assume that the program does return a sparse output (which occurs in practice), some guarantees for the accuracy of the output follow automatically from the analysis of Theorem 2.3 and Theorem 2.6, since for sparse mixtures they (implicitly) bound parameter error in terms of the closeness in CDF-distance.

Implementation in simulations: In our experiments, we solved the above convex program using the barrier method of CPLEX version 12.8 with numerical emphasis enabled.

G.2 Simulations: Additional Details

Earthmover's distance between parameters: The Earthmover's (or 1-Wasserstein) distance between P and Q measures the minimum transport cost to move the "mass" corresponding to probability distribution P to that of Q. Rigorously, in one dimension it can be defined by

$$EMD(P, Q) := \min_{\pi : \pi|_X = P, \pi|_Y = Q} \mathbb{E}_{(X,Y)\sim\pi}[|X - Y|]$$

where here π ranges over all possible couplings of marginal distributions P and Q. The following definition makes the notion of Earthmover's distance between the parameters of two mixtures of exponentials precise:

Definition 4. Let P and Q be two mixtures of exponentials $P(T > t) = \sum_i p_i e^{-\lambda_i t}$ and $Q(T > t) = \sum_i q_i e^{-\gamma_i t}$. The *Earthmover's distance in parameter space* between P and Q is the Earthmover's distance between corresponding atomic measures $\mu_P := \sum_i p_i \delta_{\lambda_i}$ and $\mu_Q := \sum_i q_i \delta_{\gamma_i}$ where δ_x represents a Dirac mass at point x.

Derivation of Per-Instance Information-Theoretic Lower bounds: Given the alternative instance, we derived the bound by computing the H^2 (Hellinger squared) distance between the true distribution and the alternative distribution, and then applying standard tensorization and comparison inequalities to bound the TV (as used in the proof of Corollary D.5).

Upper bound simulations: We ran 300 trials for each setting of k and number of samples; in order to run the simulation for very large numbers of samples, we directly generated the corresponding noisy CDF estimates by adding Gaussian noise of order $O(1/\sqrt{N})$ where N is the number of samples. For reasonable size N we also ran the methods using actual sample-estimated CDFs and the results were consistent with the simulated Gaussian-noise CDFs. The lower bound is analytically computed, not simulated, so it is unaffected by this Gaussian-noise approximation.

Plotted Data: Here we provide the data plotted in Figure 3, that was found via simulation as described above.

k	CVX	MPM	LB
1	2.98×10^5	9.28×10^4	1.34×10^4
2	3.25×10^8	3.45×10^{10}	8.18×10^6
3	3.55×10^{11}	3.87×10^{14}	1.44×10^8
4	1.21×10^{14}	1.40×10^{19}	1.13×10^9
5	N/A	4.89×10^{22}	1.43×10^{13}

Table 1: Values plotted on a log (base 10) scale in Figure 3.