On the surprising effectiveness of a simple matrix exponential derivative approximation, with application to global SARS-CoV-2

Gustavo Didier[a], Nathan E. Glatt-Holtz[a], Andrew J. Holbrook[b,1], Andrew F. Magee[b], and Marc A. Suchard[b,c,d]

[a]Department of Mathematics, Tulane University; [b]Department of Biostatistics, University of California, Los Angeles; [c]Department of Biomathematics, University of California, Los Angeles; [d]Department of Human Genetics, University of California, Los Angeles

The continuous-time Markov chain (CTMC) is the mathematical workhorse of evolutionary biology. Learning CTMC model parameters using modern, gradient-based methods requires the derivative of the matrix exponential evaluated at the CTMC's infinitesimal generator (rate) matrix. Motivated by the derivative's extreme computational complexity as a function of state space cardinality, recent work demonstrates the surprising effectiveness of a naive, first-order approximation for a host of problems in computational biology. In response to this empirical success, we obtain rigorous deterministic and probabilistic bounds for the error accrued by the naive approximation and establish a "blessing of dimensionality" result that is universal for a large class of rate matrices with random entries. Finally, we apply the first-order approximation within surrogate-trajectory Hamiltonian Monte Carlo for the analysis of the early spread of SARS-CoV-2 across 44 geographic regions that comprise a state space of unprecedented dimensionality for unstructured (flexible) CTMC models within evolutionary biology.

Continuous-time Markov chains | Hamiltonian Monte Carlo | Matrix exponential | Molecular epidemiology | Random matrix theory

When measuring the size of "big data", one should consider the structures that implicitly give rise to one's data and the computational complexities of modeling schemes that appropriately take such structures into account. Phylogeographic methods (1–4) model large-scale viral transmission between human populations as a function of the shared evolutionary history of the viral population of interest. Data take the form of dates, locations and genome sequences associated to individual viral samples. Spatiotemporal structure interfaces with network structure given by the phylogeny, or family tree, describing the viruses' collective history beginning with the most recent common ancestor. While one cannot directly observe this history, one may statistically reconstruct the phylogenetic tree by positing that changes in the viral genome happen randomly at regular intervals, thereby capturing the intuition that viral samples with more differences between their (aligned) sequences should find themselves further apart on the family tree.

The continuous-time Markov chain (CTMC) (5) represents the gold-standard mathematical model for such evolution of characters (e.g., nucleotides) within a fixed span of evolutionary time. A CTMC defined over a discrete, d-element state space consists of a row vector $\boldsymbol{\pi}_0$ whose individual components describes the probability of inhabiting each of the possible states at time $t = 0$, as well as a $d \times d$ infinitesimal generator (or rate) matrix $\mathbf{Q}$ with non-negative off-diagonal elements q_{ij}, $i \neq j$, and non-positive diagonal elements $q_{ii} = -\sum_j q_{ij}$.

For any lag $t \geq 0$, the matrix exponential (6, 7) provides the Markov chain's transition probability matrix

$$\mathbf{P}_t := e^{t\mathbf{Q}} := \sum_{n=0}^{\infty} \frac{t^n \mathbf{Q}^n}{n!}, \qquad [1]$$

which has elements $[\mathbf{P}_t]_{ij}$ that dictate the probability of the process jumping from state i to state j after time t. It is straightforward to verify that $\mathbf{P}_t$ is a valid transition matrix, having probability vectors for rows: if $\mathbf{1}$ and $\mathbf{0}$ are the column vectors of ones and zeros, respectively, then $\mathbf{Q1} = \mathbf{0}$ and, therefore, $\mathbf{P}_t \mathbf{1} = \mathbf{1}$. The law of total probability then provides the marginal probability of the process at any time $t \geq 0$ as $\boldsymbol{\pi}_t = \boldsymbol{\pi}_0 e^{t\mathbf{Q}}$.

Whether frequentist (8) or Bayesian (9–12), likelihood-based approaches to phylogenetic reconstruction allow phylogenetic tree branch lengths to parameterize time lags within the CTMC framework. We present the exact statement of the phylogenetic CTMC paradigm below (see Section 2). Here, we note that the historical importance of tree-reconstruction from aligned sequences leads to an early emphasis on the sparse specification of $\mathbf{Q}$ based on biologically motivated assumptions (13–15). Classical Markov chain Monte Carlo (MCMC) procedures (16, 17) work well for such low-dimensional models. But the phylogenetic CTMC framework has applications beyond simple nucleotide substitution models. Within, e.g., Bayesian phylogeography, the work in (1) provides a phylogenetic CTMC model for the spread of avian influenza across $d = 20$ global geographic locations but, for computational reasons, favors a low-dimensional $\mathcal{O}(d)$ parameterization of $\mathbf{Q}$. Similarly, (2) model the spread of influenza A H1N1 and H3N2 between as many as $d = 26$ geographic regions but—again

---

Significance Statement

Recent work uses a simplistic approximation to the matrix exponential derivative to apply gold-standard models from evolutionary biology to a collection of challenging data analyses. Whereas one may expect the naive approach to break down with increasing model dimensionality, empirical results show no such failure. Here, we (1) develop rigorous error bounds that improve—in a certain sense—as model dimension grows and (2) demonstrate the scalability of the naive approach to a higher-dimensional analysis of the global spread of the virus responsible for the COVID-19 pandemic.

---

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

PNAS | September 12, 2023 | vol. XXX | no. XX | 1–12

for computational reasons—fit the model with approximation techniques that provide no inferential guarantees.

Recently, (18) demonstrate the feasibility of approximate gradient-based methods for both maximum *a posteriori* and full Bayesian inference of flexible and fully-parameterized rate models and apply these methods to a gold-standard $\mathcal{O}(d^2)$ mixed-effect CTMC model for the spread of A H3N2 influenza between $d = 14$ geographic locations. The usual CTMC log-likelihood gradient calculations feature the matrix exponential derivative (see Eq. (13), Eq. (15) below)

$$\nabla_{\mathbf{J}}e^{t\mathbf{Q}} := \lim_{\epsilon \to 0} \frac{e^{t(\mathbf{Q}+\epsilon\mathbf{J})} - e^{t\mathbf{Q}}}{\epsilon} \qquad [2]$$

$$= e^{t\mathbf{Q}} \sum_{n=0}^{\infty} \frac{t^{n+1}}{(n+1)!} \left(\sum_{\ell=0}^{n} (-1)^{\ell} \binom{n}{\ell} \mathbf{Q}^{\ell} \mathbf{J} \mathbf{Q}^{n-\ell} \right) \qquad [3]$$

computed in the direction $\mathbf{J}$ of each of the d^2 natural basis elements $\mathbf{J}_{ij}$ spanning the space of real-valued, $d \times d$ matrices $\mathcal{M}(d) = \mathcal{M}(d, \mathbb{R})$, thereby requiring at least $\mathcal{O}(d^5)$ floating point operations (19).

Within the phylogenetic CTMC models of Section 2, log-likelihood derivative computations that require $\nabla_{\mathbf{J}}e^{t\mathbf{Q}}$ balloon to $\mathcal{O}(KNd^5)$, for N the number of biological specimens observed and K the number of parameters parameterizing $\mathbf{Q}$. To address this overwhelming computational cost, (18) leverage the simplistic approximation obtained by setting $n = 0$ within Eq. (3):

$$\widetilde{\nabla}_{\mathbf{J}}e^{t\mathbf{Q}} := te^{t\mathbf{Q}}\mathbf{J}. \qquad [4]$$

(18) show that this approximation helps reduce total cost to $\mathcal{O}(Kd^2 + Nd^3)$ and use this speedup within surrogate-trajectory Hamiltonian Monte Carlo (see Appendix D) to obtain a 34-fold improvement in effective sample size per second (ESS/s) over random-walk MCMC within their 14 region phylogeographic example. When trying to explain the remarkable empirical performance of the naive approximation, the authors derive an error upper bound (for an arbitrary matrix norm)

$$\|\widetilde{\nabla}_{\mathbf{J}}e^{t\mathbf{Q}} - \nabla_{\mathbf{J}}e^{t\mathbf{Q}}\| \le \frac{\|\mathbf{J}\|\|\mathbf{Q}\|}{2}(e^{2t} - 2t - 1) \qquad [5]$$

that fails to leverage the specific forms of $\mathbf{Q}$ and $\mathbf{J}$. Notably, this bound explodes as either t or $\|\mathbf{Q}\|$ diverges to ∞. Of course, the latter quantity would be expected to grow large with dimension d without more careful structural assumptions, e.g., that $\mathbf{Q}$ is a rate matrix belonging to the class

$$\mathcal{R}(d) := \left\{ \mathbf{Q} \in \mathcal{M}(d) | \mathbf{Q}_{ij} \ge 0 \text{ for } j \ne i, \sum_{j=1}^{d} \mathbf{Q}_{ij} = 0 \right\}. \qquad [6]$$

In the following, we use the finer structural properties of $\mathbf{Q}$ to obtain more precise bounds on

$$\mathbf{E}(t) := \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}. \qquad [7]$$

In Theorem 1, we provide an affine (in t) correction to the approximation Eq. (4) that yields an exponentially tight $t \to \infty$ asymptotic for the error Eq. (7). Then, in Theorem 4, we establish precise probabilistic bounds in the high-dimensional $d \to \infty$ limit for a large class of randomly drawn rate matrices $\mathbf{Q} \in \mathcal{R}$. Here we show, for any $\mathbf{Q} \in \mathcal{R}$ whose off-diagonal

elements are determined by independently and identically distributed (iid) draws from a positive, sub-exponential distribution F, that all of the non-zero singular values grow as $\sigma_j(\mathbf{Q}) \sim d$ along with asymptotically valid almost sure bounds for these rates as $d \to \infty$.

In regards to this second result, Theorem 4, note that random rate (or 'Laplacian') matrices have attracted a great deal of attention from the probability research community, especially in regard to their high-dimensional properties; see e.g., (20–25). In particular, seminal papers such as (26–28) establish broad characterizations of bulk behavior for the Laplacian eigenspectrum. One contribution of this paper, of independent interest, is a short and self-contained construction of useful bounds for the singular values of $\mathbf{Q} \in \mathcal{R}$.

In Theorem 3 and Corollary 5, we show how Theorem 1 and Theorem 4 combine to provide a more refined analysis of $\mathbf{E}$ for suitable randomly generated $\mathbf{Q} \in \mathcal{R}(d)$. In Theorem 3, we establish that particular terms appearing in the bound Eq. (24) in Theorem 1 decay with a rate on the order of $1/d$ in the operator norm topology for large d for certain classes of symmetric generators $\mathbf{Q}$. This class includes the random elements considered in Theorem 4. Here, although Corollary 5 applies only for symmetric matrices composed of sub-exponential draws, we provide strong supplemental numerical evidence that our bounds remain valid well beyond this special symmetric, sub-exponential special case; see Remark 7 and Figure 1.

One notable practical implication of Theorem 3, Corollary 5 and Remark 7 is the identification of a further correction to Eq. (4). Crucially this correction has the same $\mathcal{O}(d^3)$ computational cost as Eq. (4) while leading to an asymptotically temporally uniformly accurate approximation of $\nabla_{\mathbf{J}}e^{t\mathbf{Q}}$. See Remark 7 and Eq. (46) below.

In Section 2 we follow these theoretical investigations with an application of the naive, first-order gradient approximation Eq. (4) to a challenge in phylogeography requiring Bayesian inference of a rate-matrix of unprecedented dimensionality. Namely, we apply the approximation to a gold-standard mixed-effects, phylogenetic CTMC model that uses 1,897 parameters to describe the spread of SARS-CoV-2 across a $d = 44$ dimensional state space consisting of different global geographic locations. Such an application complements the empirical studies of (18) in a manner that emphasizes the naive approximation's potential for impact.

1. Rigorous Results

This section lays out our rigorous results Theorem 1, Theorem 3, Theorem 4 and Corollary 5, the proofs of which appear in the supplemental material below.

In what follows we adopt the following notational conventions. For any $\mathbf{A} \in \mathcal{M}(d)$, we list the associated (not necessarily distinct) eigenvalues of $\mathbf{A}$ in ascending order according to their real part, namely,

$$\Re\lambda_1(\mathbf{A}) \le \cdots \le \Re\lambda_d(\mathbf{A}). \qquad [8]$$

Similarly, the singular values of $\mathbf{A}$ are written in ascending order

$$\sigma_1(\mathbf{A}) \le \cdots \le \sigma_d(\mathbf{A}). \qquad [9]$$

We let $\mathcal{S}(d) = \mathcal{S}(d, \mathbb{R})$ and $\mathcal{S}(d, \mathbb{C})$ represent the spaces of $d \times d$ symmetric and Hermitian matrices, respectively.

We make use of multiple matrix norms leading to materially different bounds as $d \to \infty$ (29). Take

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j=1}^{d} \mathbf{A}_{ij}^2} = \sqrt{\sum_{j=1}^{d} \sigma_j(\mathbf{A})^2} \qquad [10]$$

for the *Frobenius norm* and

$$\|\mathbf{A}\|_{\mathrm{op}} := \sqrt{\lambda_d(\mathbf{A}^*\mathbf{A})} = \sqrt{\lambda_d(\mathbf{A}\mathbf{A}^*)} = \sigma_d(\mathbf{A}) \qquad [11]$$

for the *operator norm* of $\mathbf{A}$. Finally note that when we simply write $\|\cdot\|$ the statement then holds for any valid matrix norm as in our formulation of Theorem 1.

Deterministic bounds on approximation error in time. We begin by deriving a dynamical equation for the error $\mathbf{E}(t)$, defined in Eq. (7). Recall that $\mathbf{X}(t) := e^{t\mathbf{Q}}$ for any $\mathbf{Q} \in \mathcal{M}(d)$ obeys the (matrix-valued) ordinary differential equation

$$\frac{d\mathbf{X}}{dt} = \mathbf{Q}\mathbf{X}, \quad \mathbf{X}(0) = \mathbf{I}. \qquad [12]$$

Setting $\mathbf{Y}^\epsilon = \epsilon^{-1}(e^{t(\mathbf{Q}+\epsilon\mathbf{J})} - e^{t\mathbf{Q}})$ and taking a limit as $\epsilon \to 0$ we find that $\mathbf{Y} = \nabla_{\mathbf{J}} e^{t\mathbf{Q}}$ obeys

$$\frac{d\mathbf{Y}}{dt} = \mathbf{Q}\mathbf{Y} + \mathbf{J}\mathbf{X} = \mathbf{Q}\mathbf{Y} + \mathbf{J}e^{t\mathbf{Q}}, \quad \mathbf{Y}(0) = \mathbf{0}. \qquad [13]$$

Thus, variation of constants yields that, for any $t \geq 0$,

$$\nabla_{\mathbf{J}} e^{t\mathbf{Q}} = e^{t\mathbf{Q}} \int_0^t e^{-s\mathbf{Q}} \mathbf{J} e^{s\mathbf{Q}} ds \qquad [14]$$

$$= e^{t\mathbf{Q}} \sum_{k,m=0}^{\infty} \int_0^t \frac{(-s)^k \mathbf{Q}^k \mathbf{J} s^m \mathbf{Q}^m}{k!m!} ds$$

$$= e^{t\mathbf{Q}} \sum_{n=0}^{\infty} \frac{t^{n+1}}{(n+1)!} \left(\sum_{\ell=0}^{n} (-1)^\ell \binom{n}{\ell} \mathbf{Q}^\ell \mathbf{J} \mathbf{Q}^{n-\ell} \right). \qquad [15]$$

Taking the first order ($n = 0$) approximation in Eq. (15) produces Eq. (4). Note that, from Eq. (14), this approximation $\widetilde{\nabla} e^{t\mathbf{Q}} \mathbf{J}$ is evidently exact in the special case when $\mathbf{J}$ and $\mathbf{Q}$ commute.

Next notice that, if we differentiate $\widetilde{\mathbf{Y}} := te^{t\mathbf{Q}}\mathbf{J}$ in t, we find that $\widetilde{\mathbf{Y}}$ obeys

$$\frac{d\widetilde{\mathbf{Y}}}{dt} = \mathbf{Q}\widetilde{\mathbf{Y}} + e^{t\mathbf{Q}}\mathbf{J}, \quad \widetilde{\mathbf{Y}}(0) = \mathbf{0}. \qquad [16]$$

Thus, taking the error $\mathbf{E}(t)$ as in Eq. (7) and combining Eq. (13) with Eq. (16) yields

$$\frac{d\mathbf{E}}{dt} = \mathbf{Q}\mathbf{E} + \mathbf{J}e^{t\mathbf{Q}} - e^{t\mathbf{Q}}\mathbf{J}, \quad \mathbf{E}(0) = \mathbf{0}. \qquad [17]$$

Hence, again integrating this expression, we find

$$\mathbf{E}(t) = e^{t\mathbf{Q}} \int_0^t (e^{-s\mathbf{Q}} \mathbf{J} e^{s\mathbf{Q}} - \mathbf{J}) ds \qquad [18]$$

$$= e^{t\mathbf{Q}} \sum_{n=1}^{\infty} \frac{t^{n+1}}{(n+1)!} \left(\sum_{\ell=0}^{n} (-1)^\ell \binom{n}{\ell} \mathbf{Q}^\ell \mathbf{J} \mathbf{Q}^{n-\ell} \right) \qquad [19]$$

as could also be directly deduced from Eq. (14), Eq. (15).

Given a rate matrix $\mathbf{Q}$ in $\mathcal{R}(d)$, recall that $\Re\lambda_{d-1}(\mathbf{Q}) \leq \lambda_d(\mathbf{Q}) = 0$ by the Gershgorin circle theorem (30, Theorem

6.1.1). Imposing a further non-degeneracy assumption (e.g., that $\Re\lambda_{d-1}(\mathbf{Q}) < 0$) we therefore have an exponential decay in $\mathbf{Q}e^{t\mathbf{Q}}$. This starting point suggests that, under fairly general conditions, we may decompose Eq. (18) into a component where $\mathbf{Q}e^{t\mathbf{Q}}$ induces an exponential decay in time and a complementary component taking the form of a time-affine correction term.

These observations lead to the following theorem, the proof of which appears in the supplement as Appendix A.

Theorem 1. *Suppose that $\mathbf{Q}, \mathbf{J} \in \mathcal{M}(d)$ for some $d \geq 1$. We assume that we can find an element $\mathbf{Q}^+ \in \mathcal{M}(d)$ such that $\mathbf{Q}$ is a generalized inverse of $\mathbf{Q}^+$, namely,*

$$\mathbf{Q}^+\mathbf{Q}\mathbf{Q}^+ = \mathbf{Q}^+ \qquad [20]$$

and such that

$$e^{\tau\mathbf{Q}}(\mathbf{I} - \mathbf{Q}^+\mathbf{Q}) = \mathbf{I} - \mathbf{Q}^+\mathbf{Q}, \ (\mathbf{I} - \mathbf{Q}\mathbf{Q}^+)e^{\tau\mathbf{Q}} = \mathbf{I} - \mathbf{Q}\mathbf{Q}^+, \ [21]$$

for any $\tau \in \mathbb{R}$. Furthermore, we suppose that $\mathbf{Q}^+$ and $\mathbf{Q}$ commute

$$\mathbf{Q}\mathbf{Q}^+ = \mathbf{Q}^+\mathbf{Q}. \qquad [22]$$

Finally, taking $\|\cdot\|$ be any matrix norm, we assume that

$$\|\mathbf{Q}e^{\tau\mathbf{Q}}\| \leq C_0 e^{-\kappa\tau}, \quad \textit{for all } \tau \geq 0, \qquad [23]$$

where the constants $C_0 > 0, \kappa > 0$ are independent of τ. Then, under these circumstances,

$$\|\mathbf{Q}^+\mathbf{J}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^+) - (\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}^+ + t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+$$
$$+ \nabla_{\mathbf{J}} e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\| \leq C(1 + t)e^{-\kappa t} \qquad [24]$$

for any $t \geq 0$. Here, $C > 0$ is a t-independent constant which is given explicitly as

$$C_0 \left(\|(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}(\mathbf{Q}^+)^2\| + \|(\mathbf{Q}^+)^2\mathbf{J}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^+)\| + \|\mathbf{Q}^+\mathbf{J}\| \right)$$
$$+ C_0^2 \|\mathbf{Q}^+\mathbf{J}\mathbf{Q}^+\|. \qquad [25]$$

Remark 2. *To illuminate the scope of Theorem 1, we have the following three classes of matrices maintaining the conditions Eq. (20)–Eq. (23) as follows.*

(i) Suppose that $\mathbf{Q} \in \mathcal{M}(d)$ is such that

$$\Re\lambda_{d-1}(\mathbf{Q}) < \Re\lambda_d(\mathbf{Q}) \leq 0, \lambda_d(\mathbf{Q}) \textit{ is simple} \qquad [26]$$

and such that, if $\lambda_d(\mathbf{Q})$ has an imaginary component, then its real part is strictly negative. Under these circumstances, writing $\mathbf{Q}$ in its Jordan canonical form yields, for some $m \geq 1$,

$$\mathbf{Q} = \mathbf{M}\mathrm{diag}(J_1, \ldots, J_{m-1}, \lambda_d(\mathbf{Q}))\mathbf{M}^{-1}. \qquad [27]$$

Here, under Eq. (26) each of these blocks J_j must be invertible and so we may take

$$\mathbf{Q}^+ := \mathbf{M}\,\mathrm{diag}(J_1^{-1}, \ldots, J_{m-1}^{-1}, \lambda_d(\mathbf{Q})^+)\,\mathbf{M}^{-1}, \qquad [28]$$

where

$$\lambda_d(\mathbf{Q})^+ = \begin{cases} 0 & \textit{if } \lambda_d(\mathbf{Q}) = 0, \\ \lambda_d(\mathbf{Q})^{-1} & \textit{otherwise.} \end{cases} \qquad [29]$$

(ii) We next consider the case where $\mathbf{Q} \in \mathcal{M}(d)$ is diagonalizable and its spectrum lies strictly on the left half plane or at the origin. This time we can write

$$\mathbf{Q} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^{-1} \ \text{where} \ \mathbf{\Lambda} = \mathrm{diag}(\lambda_1(\mathbf{Q}), \ldots, \lambda_d(\mathbf{Q})) \quad [30]$$

and we set

$$\mathbf{Q}^+ = \mathbf{M}\mathbf{\Lambda}^+\mathbf{M}^{-1} \ \text{with} \ \mathbf{\Lambda}^+ = \mathrm{diag}(\lambda_1(\mathbf{Q})^+, \ldots, \lambda_d(\mathbf{Q})^+). \quad [31]$$

The complex numbers $\lambda_j(\mathbf{Q})^+$, $j = 1, \ldots, d$ are defined as in Eq. (29).

(iii) Finally, we specialize to the case where $\mathbf{Q} \in \mathcal{S}(d)$ is symmetric. In this case, $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$, where $\mathbf{U}$ is a unitary matrix and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1(\mathbf{Q}), \ldots, \lambda_d(\mathbf{Q}))$, $\lambda_j(\mathbf{Q})$ are its (real) eigenvalues. We suppose that these eigenvalue are all non-positive,

$$\lambda_d(\mathbf{Q}) \le 0, \quad [32]$$

Here we take $\mathbf{Q}^+$ as the Moore-Penrose inverse, namely,

$$\mathbf{Q}^+ = \mathbf{U}\mathbf{\Lambda}^+\mathbf{U}^*, \quad [33]$$

where $\mathbf{\Lambda}^+$ is as in Eq. (31).

In anticipation of Theorem 4 below and our desired application in Section 2 we are preoccupied with the dimensional dependence of the constants in Eq. (23), Eq. (24) and Eq. (25) in our formulation of Theorem 1. We next provide some such desirable bounds in case (iii) of Remark 2. Note that analogous results for generators $\mathbf{Q}$ in the classes (i) or (ii) would seemingly require a delicate analysis of the associated eigenspaces (i.e., of the structure of $\mathbf{M}$ in Eq. (27) or Eq. (30) respectively). However, Remark 6 and Figure 1 provide numerical evidence of a broader scope for dimensionally improving approximations beyond the symmetric case, at least for certain classes of randomly drawn matrices.

Theorem 3. *Let $\mathbf{Q} \in \mathcal{S}(d)$ be non-positive, i.e., suppose that* Eq. (32) *holds. Take $\mathbf{Q}^+$ as in* Eq. (33) *and define*

$$d_- = \max\{1 \le j \le d \mid \Re\lambda_j(\mathbf{Q}) < 0\}. \quad [34]$$

Then, for any $\mathbf{J} \in \mathcal{M}(d)$,

$$\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_F \quad [35]$$
$$\le \Big(\sqrt{d_-}|\lambda_1(\mathbf{Q})| \cdot \Big[2\sqrt{d - d_-}\|\mathbf{Q}^+\|_F^2 + \|\mathbf{Q}^+\|_F$$
$$+ \sqrt{d_-}|\lambda_1(\mathbf{Q})|\|\mathbf{Q}^+\|_F^2\Big](1+t)e^{-t|\lambda_{d-}(\mathbf{Q})|}$$
$$+ 2\sqrt{d - d_-}\|\mathbf{Q}^+\|_F\Big)\|\mathbf{J}\|_F,$$

with $\|\mathbf{Q}^+\|_F^2 := \sum_{k=1}^{d_-} \frac{1}{\lambda_k^2(\mathbf{Q})}$, whereas

$$\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_{\mathrm{op}} \quad [36]$$
$$\le \Big(\frac{|\lambda_1(\mathbf{Q})|(2 + |\lambda_{d-}(\mathbf{Q})| + |\lambda_1(\mathbf{Q})|)}{\lambda_{d-}^2(\mathbf{Q})}(1+t)e^{-t|\lambda_{d-}(\mathbf{Q})|}$$
$$+ \frac{2}{|\lambda_{d-}(\mathbf{Q})|}\Big)\|\mathbf{J}\|_{\mathrm{op}}.$$

Under the further assumption that $\lambda_d(\mathbf{Q}) = 0$ and

$$\mu_1 d \le |\lambda_{d-1}(\mathbf{Q})| \le |\lambda_1(\mathbf{Q})| \le \mu_2 d \quad [37]$$

for some $0 < \mu_1 \le \mu_2$, we have

$$\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_F \quad [38]$$
$$\le \Big(\frac{\mu_2}{\mu_1^2} \cdot \Big[2\sqrt{d} + \mu_1 d + \mu_2 d^2\Big](1+t)^{-t\mu_1 d} + \frac{2}{\mu_1\sqrt{d}}\Big)\|\mathbf{J}\|_F,$$

and that

$$\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_{\mathrm{op}} \quad [39]$$
$$\le \Big(\frac{\mu_2}{\mu_1^2} \cdot \Big[\frac{2}{d} + \mu_1 + \mu_2\Big](1+t)e^{-t\mu_1 d} + \frac{2}{\mu_1 d}\Big)\|\mathbf{J}\|_{\mathrm{op}}.$$

High-dimensional asymptotics via random matrix theory. We turn to our probabilistic bounds on the singular values of randomly generated rate matrices, Theorem 4. Although interesting in its own right, this result leads to consequences for the bounds in Eq. (23) and Eq. (24) when applied in Theorem 3. Before proceeding, we briefly introduce further mathematical preliminaries associated with the so-called subexponential random variables.

Definition 1. A random variable X is called *sub-exponential* if there exists some constant $K > 0$ for which its tails satisfy

$$\mathbb{P}(|X| \ge t) \le 2\exp\{-t/K\}, \quad \forall t \ge 0. \quad [40]$$

In this case, the *sub-exponential norm* of X is defined by

$$\|X\|_{\psi_1} = \inf\big\{s > 0 : \mathbb{E}\exp(|X|/s) \le 2\big\}. \quad [41]$$

The *class of sub-exponential distributions* is denoted by

$$L_{\psi_1} = \big\{F_X(dx) : \|X\|_{\psi_1} < \infty\big\}.$$

Remark 1. In fact, by (31, Proposition 2.7.1, p. 33), condition Eq. (40) is equivalent to the existence of some $s_0 > 0$ such that $\mathbb{E}\exp(|X|/s_0) \le 2$, namely, $\|X\|_{\psi_1} < \infty$ in Eq. (41). As a notable example, if $X \sim \exp(\lambda)$, $\lambda > 0$, then it is easy to see that $X \in L_{\psi_1}$, indeed.

We formulate our second major result as follows.

Theorem 4. *Let $F_X \in L_{\psi_1}$ be a distribution such that $X \ge 0$ a.s., $\mathbb{E}X = \mu > 0$ and $\mathrm{Var}\,X = \sigma^2 > 0$. We consider a sequence of random matrices $\mathbf{Q} \equiv \mathbf{Q}(d) = \{q_{ij}\}_{i,j=1,\ldots,d} \in \mathcal{R}(d)$, $d \in \mathbb{N}$, where either*

$$\{q_{ij}\}_{i,j=1,\ldots,d,\ i\ne j} \overset{\text{iid}}{\sim} F_X \ \text{and}$$
$$-q_{ii} = \sum_{j\in\{1,\ldots,d\}\setminus\{i\}} q_{ij}, i = 1, \ldots, d \quad [42]$$

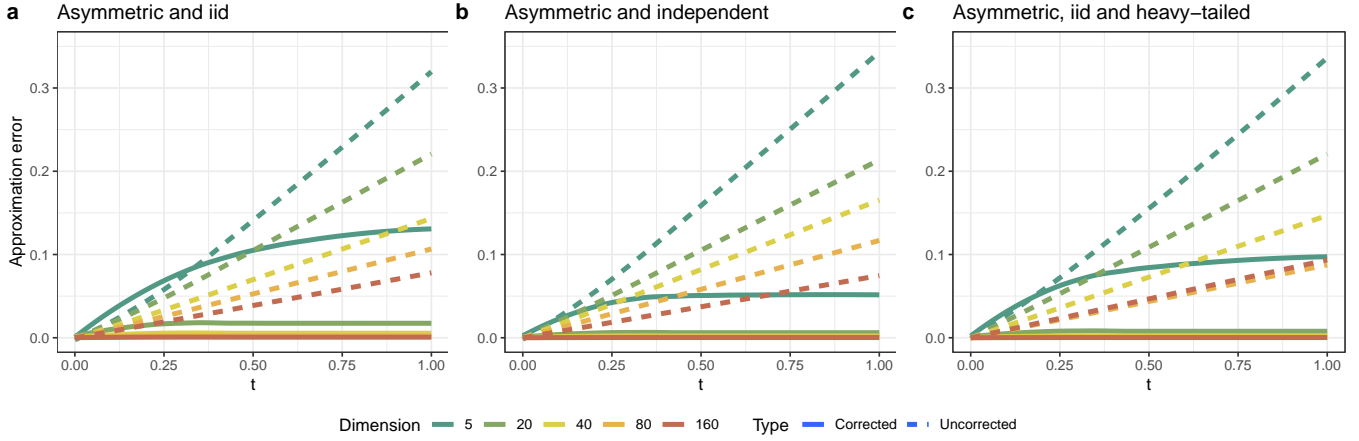*or we impose that $\mathbf{Q} \in \mathcal{R}(d) \cap \mathcal{S}(d, \mathbb{R})$ as*

$$\{q_{ij}\}_{i,j=1,\ldots,d,\ i>j} \overset{\text{iid}}{\sim} F_X, q_{ij} := q_{ji} \ \text{for} \ i < j \ \text{and}$$
$$-q_{ii} = \sum_{j\in\{1,\ldots,d\}\setminus\{i\}} q_{ij}, i = 1, \ldots, d. \quad [43]$$

Then, in either of these cases, for any $d \in \mathbb{N}\setminus\{1\}$, we have

$$\mu + O_{\text{a.s.}}\Big(\sqrt{\frac{\log d}{d}}\Big) \le \frac{\sigma_2(\mathbf{Q})}{d}$$
$$\le \frac{\sigma_d(\mathbf{Q})}{d} \le \mu + O_{\text{a.s.}}\Big(\sqrt{\frac{\log d}{d}}\Big) \quad [44]$$

almost surely. In Eq. (44), *$O_{\text{a.s.}}$ is as in* Eq. (64).

Fig. 1. Frobenius norm errors obtained by first-order approximation $te^{t\mathbf{Q}}\mathbf{J}$ and by affine-corrected first-order approximation $te^{t\mathbf{Q}}\mathbf{J} - t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{JQQ}^+$ (Theorem 3) under increasingly relaxed assumptions. Within each assumption set, we average over 20 independent Monte Carlo simulations of random generator matrices for each dimension. Plot **a** corresponds to asymmetric generators with off-diagonal elements having independent and identically distributed (iid) standard exponential random variables that correspond to the sub-exponential distribution hypothesis. Plot **b** drops the identical distribution assumption by allowing each row and column of the generator matrix to additively contribute its own mean—itself given by a standard exponential—to its corresponding exponentially distributed entries. Plot **c** features rate matrices with iid Cauchy entries truncated to be positive. Empirically, the results of Corollary 5 extend beyond the symmetric, iid and sub-exponential hypotheses, suggesting scope of future work.

Remark 2. The bounds constructed in Theorem 4 are strongly reminiscent of the bounds for eigenvalues provided in Theorem 1.5 of the seminal paper (28) (see also, for instance, Corollary 1.6 in (26) and Corollary 1.1 in (27)).

Finally, let us observe that Theorem 3, Theorem 4 as well as the fact that

$$\sigma_j(\mathbf{Q}) = |\lambda_{d-j+1}(\mathbf{Q})|, \text{ for e.g. any } \mathbf{Q} \in \mathcal{S}(d) \cap \mathcal{R}(d) \quad [45]$$

combine to produce the following immediate corollary.

Corollary 5. *Consider any sequence of random matrices* $\mathbf{Q} \equiv \mathbf{Q}(d) = \{q_{ij}\}_{i,j=1,\ldots,d} \in \mathcal{R}(d),\ d \in \mathbb{N}$, *as in Theorem 4 under the second (symmetric) case* Eq. (43). *Then, taking* $\mu_1 = \mu_1(d)$ *and* $\mu_2 = \mu_2(d)$ *as the resulting lower and upper bounds defined by* Eq. (44), *we have that* $\mathbf{Q}$ *satisfies both* Eq. (38), Eq. (39), *cf.* Eq. (45) *relative to this sequence of* μ_1, μ_2 *for any* $\mathbf{J} \in \mathcal{M}(d)$.

Remark 6. *Our rigorous formulation of Corollary 5 is limited to symmetric random rate matrices whose above diagonal elements are independent and identically distributed draws from a sub-exponential distribution. However, strong numerical evidence suggest that the scope of the approximations* Eq. (38), Eq. (39) *reach far beyond limitations of Corollary 5 in several different ways. Figure 1 explores the consequences of relaxing various assumptions of Corollary 5. In general, there appears to be no significant departure from the idea that the proxy form* $t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{JQQ}^+$ *corresponds increasingly well to the true approximation error (*Eq. (7)*) as the dimension increases.*

Remark 7. *Calculation of* $\widetilde{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}} := te^{t\mathbf{Q}}\mathbf{J}$ *requires* $\mathcal{O}(d^3)$ *operations by, e.g., computing the spectral decomposition* $\mathbf{Q}$ *as in* Eq. (30). *One may then recycle this decomposition to determine the additional term* $t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{JQQ}^+$ *for little extra cost. In view of Corollary 5 and Figure 1,*

$$\widehat{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}} := te^{t\mathbf{Q}}\mathbf{J} - t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{JQQ}^+ \quad [46]$$

provides an accurate approximation of $\nabla_{\mathbf{J}} e^{t\mathbf{Q}}$ *for asymptotically for large* d. *Thus, we anticipate further computational*

improvements when fitting large, gold-standard models using this refined approximation. That said, we leave the efficient and scalable application of $\widehat{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}}$ *to future work.*

2. Application: global spread of SARS-CoV-2

(18) consider phylogenetic models that involve CTMC priors and show that the first-order approximation Eq. (4) of the matrix exponential derivative Eq. (2) performs remarkably well within surrogate HMC (see Appendix D in the Supplement), achieving an over 30-fold efficiency gain compared to random-walk Metropolis for a 14-state model with over 180 model parameters. Here, we demonstrate similar strong performance of the first-order approximation within surrogate HMC for a 44-state model with almost 1,900 model parameters.

Phylogenetic CTMC. Start with a (possibly unknown) rooted and bifurcating phylogenetic tree $\mathcal{T}$ consisting of N leaf nodes that correspond to observed biological specimens and $N-1$ internal nodes that correspond to unobserved ancestors. The tree also contains $2N-2$ branches of length t_v connecting each child node v to its parent u.

Given $\mathcal{T}$, we model the evolution of d characters along each branch of the tree according to a CTMC model with $d \times d$ generator matrix $\mathbf{Q}$ and stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d) = \lim_{t\to\infty} \widetilde{\boldsymbol{\pi}} e^{t\mathbf{Q}}$, for $\widetilde{\boldsymbol{\pi}}$ for any arbitrary probability vector. Examples of characters are the $d = 4$ nucleotides within a set of aligned genome sequences (13) and the set of $d = 15$ geographic regions visited by an influenza subtype (2). We may scale t_v to be raw time (e.g., years) or the expected number of substitutions with respect to $\boldsymbol{\pi}$ depending on the given problem. In the former case, one may augment the model with a rate scalar γ that modulates the expected number of substitutions across all branches, and the finite-time transition probability matrix along branch v becomes $\mathbf{P}_v := e^{\gamma t_v \mathbf{Q}}$. In the following, we further posit $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ a vector of parameters.

Let data $\mathbf{Y}$ be the $d \times N$ matrix with columns $\mathbf{y}_n$, $n \in \{1, \ldots, N\}$, each having a single non-zero entry (set to 1) corresponding to the observed state of the biological specimen.

One may use any node v to express the likelihood

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \mathbf{p}_v^T \mathbf{q}_v \,, \qquad [47]$$

where $\mathbf{p}_v$ and $\mathbf{q}_v$ are the post-order and pre-order partial likelihood vectors, respectively (32). The former describes the probability of the observed states for all observed specimens (i.e., leaf nodes) that descend from node v, conditioned on the state at node v. The latter describes analogous probabilities for all observed specimens not descending from node v. For leaf nodes, $\mathbf{p}_n := \mathbf{y}_n$, $n \in \{1, \ldots, N\}$, and one may specify the root node's pre-order partial likelihood to be any arbitrary probability vector *a priori*. Let '∘' denote the Hadamard or elementwise product between matrices or vectors of equal dimensions. If we suppose that node u gives rise to two child nodes, v and w, then

$$\mathbf{p}_u = \mathbf{P}_v \mathbf{p}_v \circ \mathbf{P}_w \mathbf{p}_w \,, \quad \mathbf{q}_v = \mathbf{P}_v^T \left(\mathbf{q}_u \circ \mathbf{P}_w \mathbf{p}_w \right) \,. \qquad [48]$$

Using the chain rule and the fact that $\mathbf{p}_v$ does not depend on $\mathbf{P}_v$, one may write the likelihood's derivative with respect to a single parameter θ_k thus:

$$\begin{aligned}
\frac{\partial}{\partial \theta_k} p(\mathbf{Y}|\boldsymbol{\theta}) &\propto \sum_{v=1}^{2N-2} \text{tr} \left(\frac{\partial (\mathbf{p}_v^T \mathbf{q}_v)}{\partial \mathbf{P}_v} \frac{\partial \mathbf{P}_v^T}{\partial \theta_k} \right) \qquad [49] \\
&= \sum_{v=1}^{2N-2} \text{tr} \left((\mathbf{q}_u \circ \mathbf{P}_w \mathbf{p}_w) \mathbf{p}_v^T \left(\frac{\partial e^{\gamma t_v \mathbf{Q}}}{\partial \theta_k} \right)^T \right) \\
&= \sum_{v=1}^{2N-2} \mathbf{p}_v^T \left(\sum_{i,j=1}^{d} \frac{\partial e^{\gamma t_v \mathbf{Q}}}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial \theta_k} \right)^T (\mathbf{q}_u \circ \mathbf{P}_w \mathbf{p}_w) \,,
\end{aligned}$$

where we suppress the dependence of u and w on v.

Whereas the recursions of Eq. (48) facilitate fast likelihood computation, inferring $\boldsymbol{\theta}$ using gradient-based approaches such as HMC requires a large number of repeated evaluations of the matrix exponential derivative. These computations become particularly onerous when one opts for a gold-standard mixed-effects model (18) and specifies

$$\log q_{ij} = b_{ij} + \epsilon_{ij} \,, \quad i \neq j \,. \qquad [50]$$

Here, the fixed effects b_{ij} are elements of some non-random matrix $\mathbf{B}$, and the random effects ϵ_{ij} are mutually independent *a priori* and inferred as model parameters. The dimension K of $\boldsymbol{\theta}$ in this model is $\mathcal{O}(d^2)$, so the $\mathcal{O}(KNd^5)$ cost of the log-likelihood derivative Eq. (49) becomes a massive $\mathcal{O}(Nd^7)$. In this context, (18) show that an approximate log-likelihood derivative based on the first-order approximation to the matrix exponential helps achieve a considerable speedup over the exact derivative, requiring only $\mathcal{O}(d^4 + Nd^3)$ floating-point operations yet facilitating high-quality proposals in the context of surrogate HMC.

In the following section, we use this method to analyze the global spread of SARS-CoV-2 and show that the first-order approximation maintains its performance for an even higher-dimensional problem than previously considered.

Bayesian analysis of SARS-CoV-2 contagion. We use the phylogenetic CTMC framework to model the early spread of SARS-CoV-2—the virus responsible for the ongoing COVID-19 pandemic—based on $N = 285$ observed viral samples collected from 31 regions worldwide between December 24, 2019

and March 19, 2020. These regions comprise 13 provinces within China and 18 countries without. Understanding the manner in which viruses travel between human populations is an object of ongoing study, and phylogeographic analyses point to the central role of travel networks including those measured by airline passenger counts (3) or google mobility data (33). Here, we include three such predictors of travel in our CTMC model by expanding the fixed effects in regression model Eq. (50) to take the form

$$\mathbf{B} = \theta_1 \mathbf{X}_1 + \theta_2 \mathbf{X}_2 + \theta_3 \mathbf{X}_3 \qquad [51]$$

for $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ fixed 44×44 matrix predictors and θ_1, θ_2 and θ_3 real-valued regression coefficients. Note we have expanded the number of regions between which viruses may travel to $d = 44$ by including two additional Chinese provinces and 11 additional countries. Figure 2 presents the three predictors of interest: $\mathbf{X}_1$ contains air travel proximities between locations as measured by annualized air passenger counts between airports contained within a region (3); $\mathbf{X}_2$ contains intracontinental proximities arising from physical distances when two regions inhabit the same continent and fixed at the minimum otherwise; finally, $\mathbf{X}_3$ describes the Hubei asymmetry, i.e., the non-existence of human travel out of the Hubei province in early 2020.

In the context of a Bayesian analysis, we specify independent normal priors on θ_1, θ_2 and θ_3 with means of 0 and variances of 2. We also assume that the 1,892 random effects ϵ_{ij} follow sparsity inducing Bayesian bridge priors with global scale parameter τ and exponent $\alpha = 0.25$. Here, we follow (34) and specify a Gamma prior on $\tau^{-\alpha}$ with a shape parameter of 1 and a rate parameter of 2. Finally, we place a flat prior on the rate scalar γ. Inferring the posterior distribution of all 1,897 model parameters requires an advanced MCMC strategy. Namely, we adopt an HMC-within-Gibbs approach, updating the scalars γ and τ independently but updating all 1,895 regression parameters (both fixed and random effects) using surrogate HMC accomplished with the first-order approximation

$$\gamma t_v e^{\gamma t_v \mathbf{Q}} \mathbf{J}_{ij} \approx \frac{\partial e^{\gamma t_v \mathbf{Q}}}{\partial q_{ij}}$$

within Eq. (49).

We generate 8 million MCMC samples in this manner, saving 1 in 10,000 Markov chain states, in order to guarantee a minimum effective sample size (ESS) greater than 100. By far, the worst-mixing parameter is the global scale τ, which obtains an ESS of 185. The three fixed-effects regression coefficients θ_1, θ_2 and θ_3 obtain ESS's of 721, 721 and 644, respectively. The median and minimum ESS for the 1,892 random effects $\epsilon_{ss'}$ are 721 and 190, respectively. We note that, after thinning and removing burn-in, the sample only consists of 721 states, so ESS of 721 implies negligible autocorrelation between samples.

The left plot of Figure 3 presents posterior densities for the fixed-effect coefficients from regression Eq. (50). The air traffic proximity, intracontinental proximity and Hubei asymmetry coefficients have posterior means and 95% credible intervals of 0.76 (0.50, 1.03), 0.04 (-0.04, 0.12) and 0.27 (-0.39, 0.78), respectively. While these results suggest a positive association between each predictor matrix and the infinitesimal generator matrix $\mathbf{Q}$, the only predictor with a statistically significant association is air traffic proximity. This result
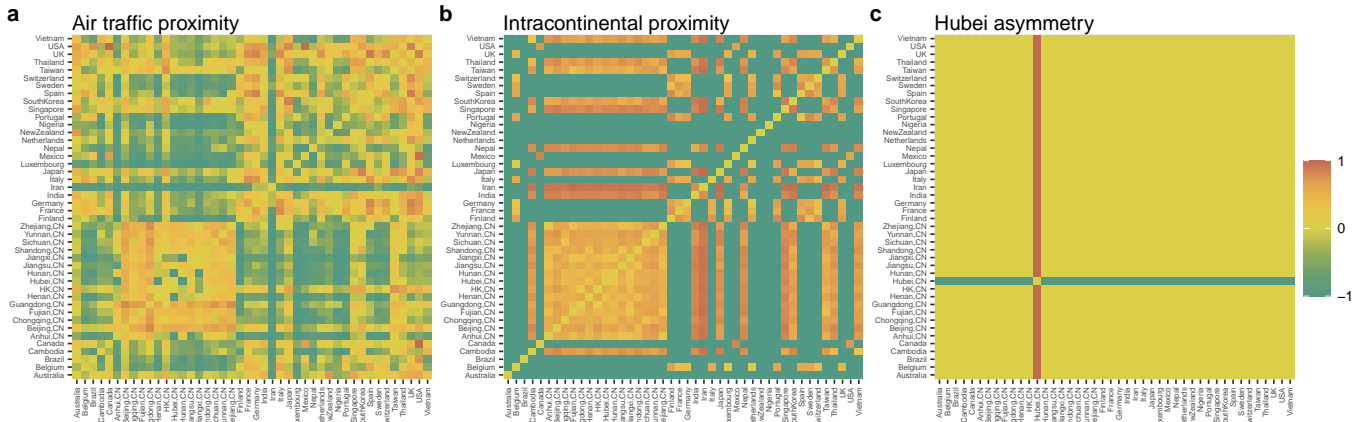
Fig. 2. Matrix predictors described in Eq. (51) combine in a linear manner to form the fixed-effect matrix **B** featured within the mixed-effects regression model Eq. (50). Air traffic proximities *(a)* are proportional the number of air passengers exchanged between airports within respective regions (3). Intracontinental proximities *(b)* take values between -1 for regions on different continents to 1 for adjacent regions. The Hubei asymmetry *(c)* roughly characterizes the Hubei quarantine of early 2020.
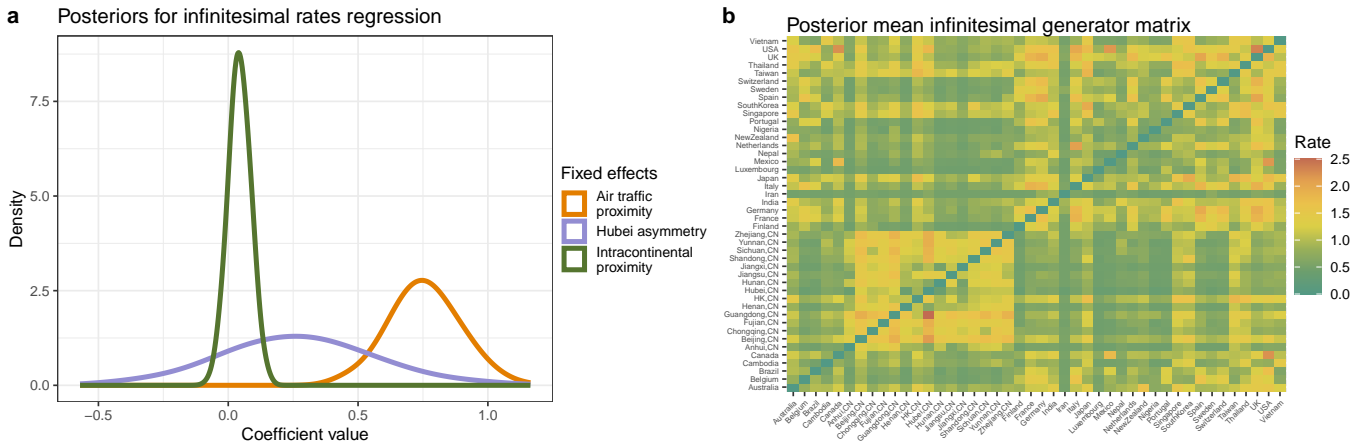
Fig. 3. Posterior inference. Posterior densities *(a)* for the three fixed-effect regression coefficients corresponding to the predictor matrices of Figure 2 reflect largely positive associations between predictors and infinitesimal rate matrix, although air traffic proximity has the only statistically significant coefficient with posterior mean of 0.76 and 95% credible interval of (0.50, 1.03). The posterior mean *(b)* infinitesimal rate matrix closely resembles the air traffic predictor, while reflecting the Hubei asymmetry to a lesser extent.

agrees with a previous phylogeographic analysis of the global spread of influenza (3). The right plot of Figure 3 presents the posterior mean for generator **Q**. As one may expect, the matrix looks similar to that of the air traffic predictor matrix in Figure 2, but one may also see the influence of the Hubei asymmetry in, e.g., the squares corresponding to travel between Guangdong and Hubei provinces. Finally, we randomly generate regions of unobserved ancestors from their posterior predictive distributions every 100-thousandth MCMC iteration. Figure 4 projects the empirical posterior predictive mode of these locations onto the phylogenetic tree $\mathcal{T}$.

In addition to these scientific questions of interest, we are interested in the performance of the first-order approximation as a surrogate gradient for HMC in such a high-dimensional setting. To isolate this performance, we fix the Bayesian bridge global-scale τ at 2.5×10^{-5}. We generate a Markov chain with 80,000 states, saving every tenth state and removing the first 1,000 states as burn-in. Despite the relatively small number of iterations, we observe large ESS that suggest satisfactory accuracy of the first-order approximation within the context of high-dimensional surrogate HMC. The ESS for the three fixed-effect regression coefficients θ_1, θ_2 and θ_3 are 1,053, 1,343

and 498, respectively. The median and minimum ESS for the 1,892 random effects ϵ_{ij} are 1,514 and 1,161, respectively.

Discussion

We develop tight probabilistic bounds on the error associated with a simplistic approximation to the matrix exponential derivative for a large class of CTMC infinitesimal generator matrices with random entries. Our "blessing of dimensionality" result shows that this error improves for higher-dimensional matrices. We apply the numerically naive approach to the analysis of the global spread of SARS-CoV-2 using a mixed-effects model of unprecedented dimensions. The results obtained herein suggest the further study of CTMCs through the lens of random matrix theory. Furthermore, this analysis suggests a refinement of the first-order approximation to the matrix exponential derivative that may be particularly useful within modern, high-dimensional settings.
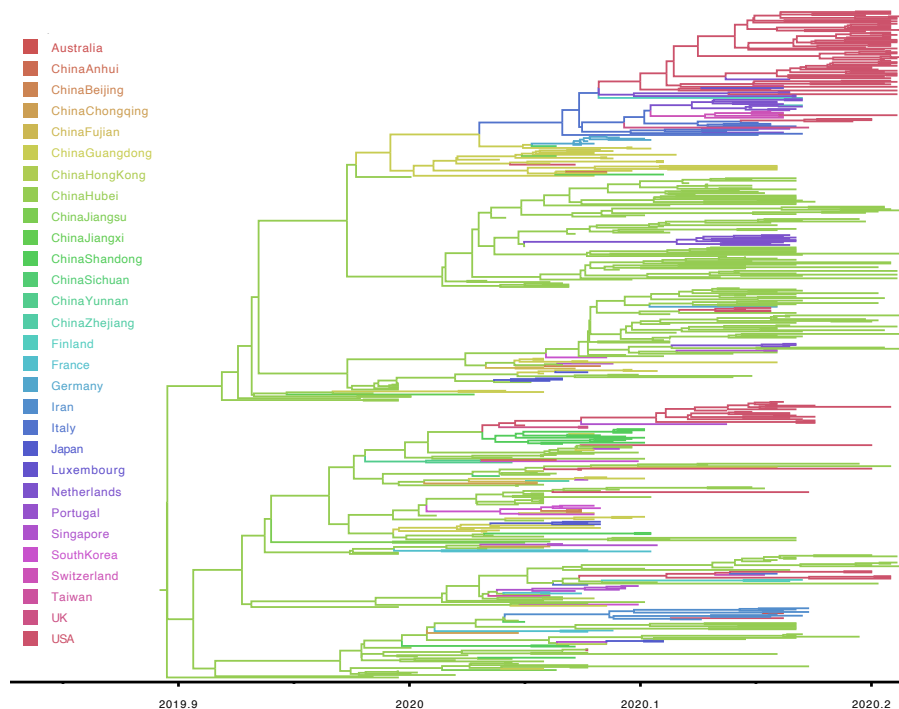
Fig. 4. Posterior predictive modes for (unobserved) ancestral locations color a phylogenetic tree that describes the shared evolutionary history of 285 SARS-CoV-2 samples.

1. P Lemey, A Rambaut, AJ Drummond, MA Suchard, Bayesian phylogeography finds its roots. *PLoS computational biology* **5**, e1000520 (2009).
2. P Lemey, et al., Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS pathogens* **10**, e1003932 (2014).
3. AJ Holbrook, et al., Massive parallelization boosts big bayesian multidimensional scaling. *J. Comput. Graph. Stat.* **30**, 11–24 (2021).
4. AJ Holbrook, X Ji, MA Suchard, From viral evolution to spatial contagion: a biologically modulated hawkes model. *Bioinformatics* **38**, 1846–1856 (2022).
5. JR Norris, *Markov chains.* (Cambridge university press) No. 2, (1998).
6. C Moler, C Van Loan, Nineteen dubious ways to compute the exponential of a matrix. *SIAM review* **20**, 801–836 (1978).
7. C Moler, C Van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review* **45**, 3–49 (2003).
8. J Felsenstein, Evolutionary trees from dna sequences: a maximum likelihood approach. *J. molecular evolution* **17**, 368–376 (1981).
9. JS Sinsheimer, JA Lake, RJ Little, Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* pp. 193–210 (1996).
10. Z Yang, B Rannala, Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Mol. biology evolution* **14**, 717–724 (1997).
11. B Mau, MA Newton, B Larget, Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics* **55**, 1–12 (1999).
12. MA Suchard, RE Weiss, JS Sinsheimer, Bayesian selection of continuous-time markov chain evolutionary models. *Mol. biology evolution* **18**, 1001–1013 (2001).
13. TH Jukes, CR Cantor, , et al., Evolution of protein molecules. *Mammalian protein metabolism* **3**, 21–132 (1969).
14. M Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. molecular evolution* **16**, 111–120 (1980).
15. M Hasegawa, H Kishino, Ta Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. molecular evolution* **22**, 160–174 (1985).
16. N Metropolis, AW Rosenbluth, MN Rosenbluth, AH Teller, E Teller, Equation of state calculations by fast computing machines. *The journal chemical physics* **21**, 1087–1092 (1953).
17. WK Hastings, Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
18. AF Magee, et al., Random-effects substitution models for phylogenetics via scalable gradient approximations. *arXiv preprint arXiv:2303.13642* **1** (2023).
19. I Najfeld, TF Havel, Derivatives of the matrix exponential and their computation. *Adv. applied mathematics* **16**, 321–375 (1995).
20. Y Takahashi, Markov chains with random transition matrices in *Kodai Mathematical Seminar Reports*. Vol. 21, pp. 426–447 (1969).
21. ZD Bai, Methodologies in spectral analysis of large dimensional random matrices, a review. *Stat. Sinica* **9**, 611–677 (1999).
22. D Chafaï, The Dirichlet Markov Ensemble. *J. Multivar. Analysis* **101**, 555–567 (2010).
23. C Bordenave, P Caputo, D Chafaï, Circular law theorem for random markov matrices. *Probab. Theory Relat. Fields* **152**, 751–779 (2012).
24. A Chatterjee, RS Hazra, Spectral properties for the Laplacian of a generalized Wigner matrix. *Random Matrices: Theory Appl.* **11**, 2250026 (2022).
25. G Nakerst, S Denisov, M Haque, Random sparse generators of markovian evolution and their spectral properties. *Phys. Rev. E* **108**, 014102 (2023).
26. W Bryc, A Dembo, T Jiang, Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Annals Probab.* **34**, 1–38 (2006).
27. X Ding, T Jiang, Spectral distributions of adjacency and laplacian matrices of random graphs. *Annals Appl. Probab.* pp. 2086–2117 (2010).
28. C Bordenave, P Caputo, D Chafaï, Spectrum of markov generators on sparse random graphs. *Commun. on Pure Appl. Math.* **67**, 621–669 (2014).
29. LN Trefethen, D Bau, *Numerical linear algebra.* (Siam) Vol. 181, (2022).
30. RA Horn, CR Johnson, *Matrix Analysis.* (Cambridge University Press), (2012).
31. R Vershynin, *High-Dimensional Probability: an Introduction with Applications in Data Science.* (Cambridge University Press) Vol. 47, (2018).
32. X Ji, et al., Gradients do grow on trees: a linear-time o (n)-dimensional gradient for statistical phylogenetics. *Mol. biology evolution* **37**, 3047–3060 (2020).
33. M Worobey, et al., The emergence of sars-cov-2 in europe and north america. *Science* **370**, 564–570 (2020).
34. A Nishimura, MA Suchard, Shrinkage with shrunken shoulders: Gibbs sampling shrinkage model posteriors with guaranteed convergence rates. *Bayesian Analysis* **1**, 1–24 (2022).
35. YQ Yin, ZD Bai, PR Krishnaiah, On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Theory Relat. Fields* **78**, 509–521 (1988).
36. S Duane, AD Kennedy, BJ Pendleton, D Roweth, Hybrid monte carlo. *Phys. letters B* **195**, 216–222 (1987).
37. R Neal, Mcmc using hamiltonian dynamics. *Handb. Markov Chain Monte Carlo* **1**, 113–162 (2011).
38. B Leimkuhler, S Reich, *Simulating hamiltonian dynamics.* (Cambridge university press) No. 14, (2004).
39. AJ Holbrook, X Ji, MA Suchard, Bayesian mitigation of spatial coarsening for a hawkes model applied to gunfire, wildfire and viral contagion. *The Annals Appl. Stat.* **16**, 573–595 (2022).
40. NE Glatt-Holtz, JA Krometis, CF Mondaini, On the accept-reject mechanism for metropolis-hastings algorithms. *Annals Appl. Probab.* (2023 (to appear)).
41. C Zhang, B Shahbaba, H Zhao, Precomputing strategy for hamiltonian monte carlo method based on regularity in parameter space. *Comput. Stat.* **32**, 253–279 (2017).
42. CE Rasmussen, Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals in *Seventh Valencia international meeting, dedicated to Dennis V. Lindley.* (Oxford University Press), pp. 651–659 (2003).
43. S Lan, T Bui-Thanh, M Christie, M Girolami, Emulation of higher-order tensors in manifold monte carlo methods for bayesian inverse problems. *J. Comput. Phys.* **308**, 81–101 (2016).
44. C Zhang, B Shahbaba, H Zhao, Hamiltonian monte carlo acceleration using surrogate functions with random bases. *Stat. computing* **27**, 1473–1490 (2017).
45. L Li, A Holbrook, B Shahbaba, P Baldi, Neural network gradient hamiltonian monte carlo. *Comput. statistics* **34**, 281–299 (2019).

Didier *et al.*

A. Proof of Theorem 1

Recall that $\mathbf{E}(t)$ is defined as in Eq. (7), and consider the representation Eq. (18). We begin by decomposing the first term $e^{-s\mathbf{Q}}\mathbf{J}e^{s\mathbf{Q}}$ inside of the integral in Eq. (18) as

$$
\begin{aligned}
&e^{-s\mathbf{Q}}\mathbf{J}e^{s\mathbf{Q}}\\
&= e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}} + e^{-s\mathbf{Q}}(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}}\\
&\quad + e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})e^{s\mathbf{Q}}\\
&\quad + e^{-s\mathbf{Q}}(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})e^{s\mathbf{Q}}\\
&= e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}} + (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}}\\
&\quad + e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) + (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}),
\end{aligned}
\tag{52}
$$

where we used Eq. (21) for the second identity. Regarding the middle two terms, with our commutativity Eq. (22) assumption, we observe that

$$
\begin{aligned}
&(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}} + e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) \tag{53}\\
&= (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+}\mathbf{Q}e^{s\mathbf{Q}} + e^{-s\mathbf{Q}}\mathbf{Q}\mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\\
&= \frac{d}{ds}\left((\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+}e^{s\mathbf{Q}} - e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\right).
\end{aligned}
$$

Hence, integrating in time, and then using Eq. (20), Eq. (22) for the first and last term terms on the right hand side, we have

$$
\begin{aligned}
&\int_{0}^{t}\left\{(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}} + e^{-s\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\right\}ds\\
&= (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+}e^{t\mathbf{Q}} - e^{-t\mathbf{Q}}\mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\\
&\quad - (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+} + \mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\\
&= (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}(\mathbf{Q}^{+})^{2}\mathbf{Q}e^{t\mathbf{Q}} - e^{-t\mathbf{Q}}\mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\\
&\quad - (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+} + \mathbf{Q}(\mathbf{Q}^{+})^{2}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) \tag{54}
\end{aligned}
$$

Finally let us note, regarding the second term on the right hand side of Eq. (18), using again Eq. (21) and Eq. (20), Eq. (22)

$$
-te^{t\mathbf{Q}}\mathbf{J} = -te^{t\mathbf{Q}}\mathbf{Q}\mathbf{Q}^{+}\mathbf{J} - t(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}. \tag{55}
$$

By combining Eq. (52)–Eq. (55) and comparing with Eq. (18), we now find that

$$
\begin{aligned}
\mathbf{E}(t) = &\int_{0}^{t}e^{(t-s)\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}}ds + (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}(\mathbf{Q}^{+})^{2}\mathbf{Q}e^{t\mathbf{Q}}\\
&+ e^{t\mathbf{Q}}\mathbf{Q}(\mathbf{Q}^{+})^{2}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) - te^{t\mathbf{Q}}\mathbf{Q}\mathbf{Q}^{+}\mathbf{J}\\
&- \mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) - (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+}\\
&- t(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}. \tag{56}
\end{aligned}
$$

Rearranging Eq. (56), we have

$$
\begin{aligned}
&\mathbf{Q}^{+}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) + (\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}^{+} + t(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}\\
&\quad + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J} = \mathbf{T}_{1} + \mathbf{T}_{2},
\end{aligned}
$$

where

$$
\mathbf{T}_{1} := \int_{0}^{t}e^{(t-s)\mathbf{Q}}\mathbf{Q}^{+}\mathbf{Q}\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}e^{s\mathbf{Q}}ds,
$$

$$
\begin{aligned}
\mathbf{T}_{2} := &(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}(\mathbf{Q}^{+})^{2}\mathbf{Q}e^{t\mathbf{Q}}\\
&+ e^{t\mathbf{Q}}\mathbf{Q}(\mathbf{Q}^{+})^{2}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}) - te^{t\mathbf{Q}}\mathbf{Q}\mathbf{Q}^{+}\mathbf{J}.
\end{aligned}
$$

We estimate each of $\mathbf{T}_{1}$ and $\mathbf{T}_{2}$ in turn. Regarding $\mathbf{T}_{1}$, using once more Eq. (22), as well as Eq. (23), we have

$$
\begin{aligned}
\|\mathbf{T}_{1}\| &\leq \int_{0}^{t}\|\mathbf{Q}e^{(t-s)\mathbf{Q}}\|\|\mathbf{Q}^{+}\mathbf{J}\mathbf{Q}^{+}\|\|\mathbf{Q}e^{s\mathbf{Q}}\|ds\\
&\leq C_{0}^{2}\int_{0}^{t}e^{-(t-s)\kappa}\|\mathbf{Q}^{+}\mathbf{J}\mathbf{Q}^{+}\|e^{-s\kappa}ds\\
&= C_{0}^{2}te^{-t\kappa}\|\mathbf{Q}^{+}\mathbf{J}\mathbf{Q}^{+}\|.
\end{aligned}
$$

Turning to $\mathbf{T}_{2}$, by Eq. (23) we have

$$
\begin{aligned}
\|\mathbf{T}_{2}\| &\leq C_{0}(\|(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}(\mathbf{Q}^{+})^{2}\| + \|(\mathbf{Q}^{+})^{2}\mathbf{J}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+})\|\\
&\quad + t\|\mathbf{Q}^{+}\mathbf{J}\|)e^{-t\kappa}.
\end{aligned}
$$

Combining these two bound completes the proof.

B. Proof of Theorem 3

This result is established from Theorem 1 and some basic properties $\mathbf{Q}$ under the given assumptions. First, note that, cf. Eq. (32), (8)

$$
|\lambda_{d}(\mathbf{Q})| \leq \max_{j=1,\dots,d_{-}}|\lambda_{j}(\mathbf{Q})| \leq |\lambda_{1}(\mathbf{Q})|. \tag{57}
$$

Referring back to Eq. (10), by Eq. (57) we have the estimate

$$
\begin{aligned}
\|\mathbf{Q}e^{t\mathbf{Q}}\|_{F} &= \left(\sum_{j=1}^{d_{-}}|\lambda_{j}(\mathbf{Q})e^{t\lambda_{j}(\mathbf{Q})}|^{2}\right)^{1/2}\\
&\leq \sqrt{d_{-}}|\lambda_{1}(\mathbf{Q})|e^{-t|\lambda_{d_{-}}(\mathbf{Q})|}. \tag{58}
\end{aligned}
$$

Similarly, from Eq. (11) and Eq. (57), we have that

$$
\begin{aligned}
\|\mathbf{Q}e^{t\mathbf{Q}}\|_{\mathrm{op}} &\leq \max_{j=1,\dots,d_{-}}|\lambda_{j}(\mathbf{Q})|e^{-t|\lambda_{j}(\mathbf{Q})|}\\
&\leq |\lambda_{1}(\mathbf{Q})|e^{-t|\lambda_{d_{-}}(\mathbf{Q})|}. \tag{59}
\end{aligned}
$$

Likewise, we observe that

$$
\|\mathbf{Q}\|_{F}^{2} = \sum_{k=1}^{d}\lambda_{k}(\mathbf{Q})^{2}, \quad \|\mathbf{Q}^{+}\|_{F}^{2} = \sum_{k=1}^{d_{-}}\frac{1}{\lambda_{k}(\mathbf{Q})^{2}}, \tag{60}
$$

and that

$$
\|\mathbf{Q}\|_{\mathrm{op}}^{2} = \lambda_{1}(\mathbf{Q})^{2}, \quad \|\mathbf{Q}^{+}\|_{\mathrm{op}}^{2} = \frac{1}{\lambda_{d_{-}}(\mathbf{Q})^{2}}. \tag{61}
$$

Let $\mathbf{U}$ be as in Eq. (33). Noting, furthermore, that $\mathbf{I} - \mathbf{Q}\mathbf{Q}^{+} = \mathbf{U}(\mathbf{I} - \mathbf{I}_{d_{-}})\mathbf{U}^{*}$, where $\mathbf{I}_{d_{-}}$ is the matrix with 1's along the first d_{-} diagonal elements and zero otherwise, we have the bounds

$$
\|\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}\|_{F} = \sqrt{d-d_{-}}, \quad \|\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}\|_{\mathrm{op}} = 1. \tag{62}
$$

Now, for any matrix norm $\|\cdot\|$, Eq. (24) implies that

$$
\begin{aligned}
&\|t(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+} + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|\\
&\leq C(1+t)e^{-\kappa t} + 2\|\mathbf{Q}^{+}\|\|\mathbf{J}\|\|\mathbf{I}-\mathbf{Q}\mathbf{Q}^{+}\| \tag{63}
\end{aligned}
$$

for any $t \geq 0$, where $C > 0$ is given by Eq. (25). Thus, for the choice of norm $\|\cdot\| = \|\cdot\|_{F}$, by Eq. (63), Eq. (58), Eq. (61) and Eq. (62), we obtain Eq. (35). Similarly, for the choice of norm $\|\cdot\| = \|\cdot\|_{\mathrm{op}}$, Eq. (63), in conjunction with Eq. (59), Eq. (61) and Eq. (62), yields the bound Eq. (36). Since Eq. (38) follows immediately from Eq. (35), and similarly between Eq. (39) and Eq. (36), the proof is now complete.

C. Proof of Theorem 4

For use in this section, we introduce a probabilistic version of the typical O and o asymptotic notations. Given collections of random variable $\{X_{d}\}_{d\in\mathbb{N}}$, $\{Y_{d}\}_{d\in\mathbb{N}}$, we write

$$
X_{d} = O_{\mathrm{a.s.}}(f(d)), \quad Y_{d} = o_{\mathrm{a.s.}}(f(d)), \tag{64}
$$

for some $f : \mathbb{N} \to \mathbb{R}^{+}$ to mean, respectively, that there exists a random variable $C = C(\omega)$, not dependent on d, such that $|X_{d}|/f(d) \leq C$ for all $d \in \mathbb{N}$ a.s., and also that $\lim_{d\to\infty}Y_{d}/f(d) = 0$ a.s.

In everything that follows we will make use of the so-called *Weyl's inequalities* (see e.g. (30, Theorem 4.3.1, p. 239)). Namely, let $\mathbf{A}, \mathbf{B} \in \mathcal{S}(d, \mathbb{C})$, each with its own eigenvalues listed in increasing order as in our convention Eq. (8). Then, for $i = 1, \dots, d$,

$$
\lambda_{i}(\mathbf{A}+\mathbf{B}) \leq \lambda_{i+j}(\mathbf{A}) + \lambda_{d-j}(\mathbf{B}), \quad j = 0, 1, \dots, d-i, \tag{65}
$$

and

$$
\lambda_{i-j+1}(\mathbf{A}) + \lambda_{j}(\mathbf{B}) \leq \lambda_{i}(\mathbf{A}+\mathbf{B}), \quad j = 1, \dots, i. \tag{66}
$$

In what follows, we also make use of the fact that, for $\mathbf{A} \in \mathcal{S}(d, \mathbb{C})$,

$$\lambda_\ell(-\mathbf{A}) = -\lambda_{d-\ell+1}(\mathbf{A}), \quad \ell = 1, \ldots, d. \qquad [67]$$

With these preliminaries now in hand, we turn to the proof of Theorem 4. This result is established with the aide of four auxiliary results, Lemma 1, Lemma 2, Lemma 3 and Lemma 4, whose precise statements and proof are provided immediately afterward.

Proof. To demonstrate our desired result, Eq. (44), it suffices to establish the lower bound

$$\sigma_2^2(\mathbf{Q}) \geq (d-1)^2 \left\{ \left(\frac{\mu d}{d-1} \right)^2 + O_{\text{a.s.}} \left(\sqrt{\frac{\log d}{d}} \right) \right\} \qquad [68]$$

as well as the upper bound

$$\sigma_d^2(\mathbf{Q}) \leq (d-1)^2 \left\{ \left(\frac{\mu d}{d-1} \right)^2 + O_{\text{a.s.}} \left(\sqrt{\frac{\log d}{d}} \right) \right\}. \qquad [69]$$

With this in mind, we now decompose $\mathbf{Q}$ as follows, starting with the case Eq. (42). Fix $d \in \mathbb{N} \backslash \{1\}$. Taking F_X to be the distribution defining the elements in Eq. (42), draw

$$\{\widetilde{q}_{ii}\}_{i=1,\ldots,d} \overset{\text{iid}}{\sim} F_X \text{ independently of } \{q_{ij}\}_{i,j=1,\ldots,d}. \qquad [70]$$

Here note carefully that $\widetilde{q}_{ii} \overset{d}{=} q_{ij}$ and $\widetilde{q}_{ii} \overset{d}{\neq} q_{ii}$. Now recast $\mathbf{Q} = \{q_{ij}\}_{i,j}$ as

$$\begin{pmatrix} 0 & q_{12}-\mu & \cdots & q_{1d}-\mu \\ q_{21}-\mu & 0 & \cdots & q_{2d}-\mu \\ \vdots & \vdots & \ddots & \vdots \\ q_{d1}-\mu & q_{d2}-\mu & \cdots & 0 \end{pmatrix} + \begin{pmatrix} q_{11} & \mu & \cdots & \mu \\ \mu & q_{22} & \cdots & \mu \\ \vdots & \vdots & \ddots & \vdots \\ \mu & \mu & \cdots & q_{dd} \end{pmatrix}$$

$$= \left\{ \begin{pmatrix} \widetilde{q}_{11}-\mu & q_{12}-\mu & \cdots & q_{1d}-\mu \\ q_{21}-\mu & \widetilde{q}_{22}-\mu & \cdots & q_{2d}-\mu \\ \vdots & \vdots & \ddots & \vdots \\ q_{d1}-\mu & q_{d2}-\mu & \cdots & \widetilde{q}_{dd}-\mu \end{pmatrix} \right. \qquad [71]$$

$$+ \begin{pmatrix} -\widetilde{q}_{11}+\mu & 0 & \cdots & 0 \\ 0 & -\widetilde{q}_{22}+\mu & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\widetilde{q}_{dd}+\mu \end{pmatrix} \right\}$$

$$+ (1-d) \left\{ \begin{pmatrix} \frac{-q_{11}-(d-1)\mu}{d-1} & 0 & \cdots & 0 \\ 0 & \frac{-q_{22}-(d-1)\mu}{d-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{-q_{dd}-(d-1)\mu}{d-1} \end{pmatrix} \right.$$

$$+ \left. \begin{pmatrix} \mu & \frac{-\mu}{d-1} & \cdots & \frac{-\mu}{d-1} \\ \frac{-\mu}{d-1} & \mu & \cdots & \frac{-\mu}{d-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-\mu}{d-1} & \frac{-\mu}{d-1} & \cdots & \mu \end{pmatrix} \right\} \qquad [72]$$

$$=: \mathbf{Q}_1 + \mathbf{Q}_2 + (1-d)\left\{\mathbf{Q}_3 + \mathbf{Q}_4\right\} =: \mathbf{R} + (1-d)\mathbf{Q}_4 \qquad [73]$$

Working from Eq. (73) we start by establishing the first bound Eq. (68) under Eq. (42). Noting from Lemma 4 that $\mathbf{Q}_4$ is symmetric and positive, making use of Eq. (66), first with $i = 2, j = 1$, then with $i = 1, j = 1$, invoking Eq. (67) and finally using that $\mathbf{R}\mathbf{R}^*$ is symmetric and non-negative, we find

$$\sigma_2^2(\mathbf{Q}) = \lambda_2(\mathbf{Q}\mathbf{Q}^*)$$

$$\geq (d-1)^2 \left\{ \lambda_2(\mathbf{Q}_4^2) + \lambda_1 \left(\frac{\mathbf{R}\mathbf{R}^*}{(d-1)^2} - \frac{\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*}{d-1} \right) \right\}$$

$$\geq (d-1)^2 \left\{ \lambda_2(\mathbf{Q}_4^2) + \lambda_1 \left(\frac{\mathbf{R}\mathbf{R}^*}{(d-1)^2} \right) + \lambda_1 \left(\frac{\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*}{1-d} \right) \right\}$$

$$\geq (d-1)^2 \left\{ \lambda_2(\mathbf{Q}_4^2) - \lambda_d \left(\frac{\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*}{d-1} \right) \right\}. \qquad [74]$$

Now observe

$$|\lambda_j(\mathbf{A})| \leq \sigma_d(\mathbf{A}) = \|\mathbf{A}\|_{op} \quad \text{for any } j = 1, \ldots, d, \qquad [75]$$

whenever $\mathbf{A}$ is symmetric. Furthermore,

$$\lambda_2(\mathbf{Q}_4^2) = \lambda_2^2(\mathbf{Q}_4) = \frac{\mu^2 d^2}{(d-1)^2}, \qquad [76]$$

where the equality follows from Eq. (90) in Lemma 4. Combining Eq. (74), Eq. (75) and Eq. (76), we therefore infer

$$\sigma_2^2(\mathbf{Q}) \geq (d-1)^2 \left\{ \left(\frac{\mu d}{d-1} \right)^2 - \frac{1}{d-1} \|\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*\|_{op} \right\}. \qquad [77]$$

However, by Eq. (80) in Lemma 1, Eq. (83) in Lemma 2 and Eq. (85) from Lemma 3 we have

$$\frac{d}{d-1} \frac{\|\mathbf{R}\|_{op}}{d} \leq \frac{d}{d-1} \left(\frac{\|\mathbf{Q}_1\|_{op}}{d} + \frac{\|\mathbf{Q}_2\|_{op}}{d} \right) + \|\mathbf{Q}_3\|_{op}$$

$$= O_{\text{a.s.}} \left(\sqrt{\frac{\log d}{d}} \right). \qquad [78]$$

Hence, by Eq. (78) and Eq. (90) in Lemma 4,

$$\frac{1}{d-1} \|\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*\|_{op} \leq \frac{2d}{d-1} \|\mathbf{Q}_4\|_{op} \left\| \frac{\mathbf{R}}{d} \right\|_{op}$$

$$= O(1) \cdot O_{\text{a.s.}} \left(\sqrt{\frac{\log d}{d}} \right) = O_{\text{a.s.}} \left(\sqrt{\frac{\log d}{d}} \right). \qquad [79]$$

Thus, by Eq. (77), Eq. (78) and Eq. (79), the lower bound Eq. (68) holds.

We now turn to verify Eq. (69). Here, by Eq. (73), Eq. (65) with $j = 0, i = d$, Eq. (76) and finally Eq. (75),

$$\sigma_d^2(\mathbf{Q}) = \lambda_d(\mathbf{Q}\mathbf{Q}^*)$$

$$\leq (d-1)^2 \left\{ \lambda_d(\mathbf{Q}_4^2) + \lambda_d \left(\frac{\mathbf{R}\mathbf{R}^*}{(d-1)^2} \right) + \lambda_d \left(\frac{\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*}{1-d} \right) \right\}$$

$$\leq (d-1)^2 \left\{ \left(\frac{\mu d}{d-1} \right)^2 + \left\| \frac{\mathbf{R}}{d-1} \right\|_{op}^2 + \frac{1}{d-1} \|\mathbf{Q}_4\mathbf{R}^* + \mathbf{R}\mathbf{Q}_4^*\|_{op} \right\}.$$

The upper bound Eq. (69) now follows from Eq. (78) and Eq. (79). Thus, as a consequence of Eq. (68) and Eq. (69), Eq. (44) holds under condition Eq. (42), as claimed. We have established the desired result in the first case Eq. (42).

Now suppose that condition Eq. (43) holds. In order to establish Eq. (44), it suffices to replace q_{ij}, $i < j$, with q_{ji} in expression Eq. (71) and in the definition of $\mathbf{Q}_3$. We then follow the rest of the argument for proving Eq. (44) under condition Eq. (42) noting that Lemma 1, Lemma 3 also hold in the symmetric case. This concludes the proof. □

We turn now to the Lemmata 1, 2, 3 and 4, which are used in the proof of Theorem 4. These results correspond to the bounds we use for each of elements in the decomposition Eq. (73). We start off with Lemma 1 which essentially packages results found in (35), (21).

Lemma 1. *Let* $\mathbf{Q}_1 = \mathbf{Q}_1(d)$, $d \in \mathbb{N} \setminus \{1\}$, *be the sequence of random matrices defined as in* Eq. (71). *Here we suppose that the off diagonal elements* q_{ij} *defining* $\mathbf{Q}_1$ *are either as in* Eq. (42) *or as in* Eq. (43) *and that the diagonal elements* $\widetilde{q}_{ii}$ *are drawn as in* Eq. (70). *Then, in either case, we have*

$$\frac{\|\mathbf{Q}_1\|_{op}}{d} = O_{\text{a.s.}} \left(\frac{1}{\sqrt{d}} \right). \qquad [80]$$

Proof. To establish Eq. (80) in the first case, Eq. (42), we note that $d^{-1}\mathbf{Q}_1\mathbf{Q}_1^*$ forms a sample covariance matrix, where each entry of $\mathbf{Q}_1$ is centered and has infinitely many moments; cf. Remark 1. Thus, by (35, Theorem 3.1, p. 517),

$$\|\mathbf{Q}_1\|_{op} = \sqrt{d} \cdot \sqrt{\lambda_d \left(\frac{\mathbf{Q}_1\mathbf{Q}_1^*}{d} \right)} = \sqrt{d} \cdot \sqrt{4\sigma^2 + o_{\text{a.s.}}(1)}, \qquad [81]$$

which immediately yields Eq. (80).

In the second case, where all the off-diagonal elements are determined starting from Eq. (43), $\mathbf{Q}_1$ is a centered (mean-zero) Wigner matrix in the typical nomenclature followed by (21). Thus, noting Eq. (11) and that, as in the previous case, $\mathbf{Q}_1$ has infinitely many moments Eq. (80) is thus a direct consequence of (21, Theorem 2.12, p. 630). □

We next state and establish Lemma 2. Note that the assumption of independence among the random variables is not needed for this result. Here and below in what follows we adopt the useful notational convention that $o_{\mathrm{a.s.}}(1)$ denotes a random variable that vanishes almost surely, as $d \to \infty$ (cf. Eq. (64)).

Lemma 2. *Let $\{Y_d\}_{d \in \mathbb{N}}$ be an identically distributed sequence of sub-exponential variables with mean μ; cf. Definition 1. Let*

$$Z_d = \max_{k=1,\ldots,d} |Y_k - \mu|, d \in \mathbb{N}. \qquad [82]$$

Then, for any fixed $\varepsilon_0 > 0$,

$$\frac{Z_d}{d^{\varepsilon_0}} = o_{\mathrm{a.s.}}(1). \qquad [83]$$

In particular, for the sequence of diagonal matrices $\mathbf{Q}_2 \equiv \mathbf{Q}_2(d)$ given as in Eq. (73), and any $\varepsilon_0 > 0$,

$$\frac{\|\mathbf{Q}_2\|_{op}}{d^{\varepsilon_0}} = o_{\mathrm{a.s.}}(1). \qquad [84]$$

Proof. For any $M > 0$,

$$\mathbb{P}\left(\frac{Z_d}{d^{\varepsilon_0}} \geq \frac{M}{d^{\varepsilon_0/2}}\right) = \mathbb{P}\left(\bigcup_{k=1}^{d}\left\{\frac{|Y_k - \mu|}{d^{\varepsilon_0}} \geq \frac{M}{d^{\varepsilon_0/2}}\right\}\right)$$

$$\leq d \cdot \mathbb{P}\left(|Y_1 - \mu| \geq d^{\varepsilon_0/2}M\right) \leq 2d \cdot e^{-\frac{d^{\varepsilon_0/2}M}{K}}.$$

In the first and second inequalities, respectively, we use the assumption that the random variables $\{Y_d\}_{d \in \mathbb{N}}$ are identically distributed as well as relation Eq. (40) for some suitable value $K > 0$. Therefore, $\sum_{d=1}^{\infty} \mathbb{P}(Z_d/d^{\varepsilon_0/2} \geq M/d^{\varepsilon_0/2}) < \infty$. Since $\lim_{d \to \infty} M/d^{\varepsilon_0/2} = 0$, Eq. (83) now follows as a consequence of the Borel-Cantelli lemma.

Regarding the second claim observe that, cf. Eq. (70), $\mathbf{Q}_2$ is a diagonal matrix containing identically distributed sub-exponential random variables. This mean that, taking $Y_k := \tilde{q}_{kk}$, $\|\mathbf{Q}_2\|$ is of the form Eq. (82). As such the second claim Eq. (84) now follows from the first Eq. (83), completing the proof. □

Turning to our bounds on $\|\mathbf{Q}_3\|_{op}$, we have the following lemma.

Lemma 3. *Let $\mathbf{Q}_3 \equiv \mathbf{Q}_3(d)$, $d \in \mathbb{N}$, be a sequence of random matrices as in Eq. (73). Here we assume that the q_{ii} are determined either according to Eq. (42) or according to Eq. (43). Then, in either of these cases,*

$$\|\mathbf{Q}_3\|_{op} = O_{\mathrm{a.s.}}\left(\sqrt{\frac{\log d}{d}}\right). \qquad [85]$$

Proof. Under either Eq. (42) or Eq. (43), each of the entries along the main diagonal is a renormalized sums of iid sub-exponential random variables i.e.

$$\frac{-q_{ii} - \mu(d-1)}{d-1} = \frac{1}{d-1}\sum_{j \neq i}^{d}(q_{ij} - \mu) \qquad [86]$$

While these diagonal elements are not independent under Eq. (43) note carefully that we do not use the independence of the rows of $\mathbf{Q}_3$ in the arguments that follow.

Now, by Bernstein's inequality, (31, p. 29, Theorem 2.8.1), there exists a constant $C > 0$ such that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|-q_{ii} - \mu(d-1)\right| > \varepsilon\right)$$

$$\leq 2 \cdot \exp\left\{-C \min\left\{\frac{\varepsilon^2}{(d-1)\|X_\mu\|_{\psi_1}^2}, \frac{\varepsilon}{\|X_\mu\|_{\psi_1}}\right\}\right\}, \qquad [87]$$

for $i = 1, \ldots, d$ where $X_\mu := X - \mu$ for $X \sim F_X$ and the right-hand side of Eq. (87) involves the sub-exponential norm Eq. (41). Fix

$\delta > 0$, and let $\eta := (2 + \delta)/C$ where C is the constant in the upper bound in Eq. (87). Hence, for all $d \in \mathbb{N}\backslash\{1\}$,

$$\mathbb{P}\left(\sqrt{\frac{d-1}{\eta \log d}} \|\mathbf{Q}_3(d)\|_{op} > \|X_\mu\|_{\psi_1}\right)$$

$$= \mathbb{P}\left(\max_{i=1,\ldots,d}\left|\frac{-q_{ii} - \mu(d-1)}{d-1}\right| > \|X_\mu\|_{\psi_1}\sqrt{\frac{\eta \log d}{d-1}}\right)$$

$$= \mathbb{P}\left(\bigcup_{i=1}^{d}\left\{\left|-q_{ii} - \mu(d-1)\right| > \|X_\mu\|_{\psi_1}\sqrt{\eta(d-1)\log d}\right\}\right)$$

$$\leq \sum_{i=1}^{d} \mathbb{P}\left(\left|-q_{ii} - \mu(d-1)\right| > \|X_\mu\|_{\psi_1}\sqrt{\eta(d-1)\log d}\right)$$

$$\leq 2d \cdot \exp\left\{-C \min\left\{\eta \log d, \sqrt{\eta(d-1)\log d}\right\}\right\}. \qquad [88]$$

In the second inequality in Eq. (88), we use Eq. (87) with $\varepsilon := \|X_\mu\|_{\psi_1}\sqrt{\eta(d-1)\log d}$ and the fact that $\{q_{ii}\}_{i=1,\ldots,d}$ are identically (but not necessarily independently) distributed. Therefore, for every d sufficiently large,

$$\mathbb{P}\left(\sqrt{\frac{d-1}{\eta \log d}}\|\mathbf{Q}_3(d)\|_{op} > \|X_\mu\|_{\psi_1}\right) \leq \frac{2}{d^{1+\delta}},$$

and as a consequence,

$$\sum_{d=1}^{\infty} \mathbb{P}\left(\|\mathbf{Q}_3(d)\|_{op} > \|X_\mu\|_{\psi_1}\sqrt{\frac{\eta \log d}{d-1}}\right) < \infty.$$

Hence, by the Borel-Cantelli lemma,

$$\mathbb{P}\left(\|\mathbf{Q}_3(d)\|_{op} > \|X_\mu\|_{\psi_1}\sqrt{\frac{\eta \log d}{d-1}} \text{ i.o.}\right) = 0,$$

so that the relation Eq. (85) holds. The proof is complete. □

Finally we conclude with Lemma 4 as follows.

Lemma 4. *Let $\mathbf{Q}_4 \equiv \mathbf{Q}_4(d)$, $d \in \mathbb{N}$, be the sequence of symmetric matrices defined in Eq. (73). Then, for any $d \in \mathbb{N}\backslash\{1\}$,*

$$\lambda_1(\mathbf{Q}_4) = 0, \qquad [89]$$

and

$$\lambda_\ell(\mathbf{Q}_4) = \frac{\mu d}{d-1}, \quad for \ \ell = 2, \ldots, d. \qquad [90]$$

Proof. The statement Eq. (89) is a consequence of the fact that $\mathbf{Q}_4(1,\ldots,1)^\top = \mathbf{0} \in \mathbb{R}^n$. To establish Eq. (90), it suffices to prove that

$$\frac{d}{d-1} \leq \frac{\lambda_2(\mathbf{Q}_4)}{\mu} \leq \frac{\lambda_d(\mathbf{Q}_4)}{\mu} \leq \frac{d}{d-1}, \quad d \in \mathbb{N}\backslash\{1\}. \qquad [91]$$

Let $\mathbf{1} \in \mathcal{S}(d, \mathbb{R})$ be a matrix of ones and recast

$$\frac{1}{\mu}\mathbf{Q}_4 = \mathbf{I} + \frac{1}{d-1}(\mathbf{I} - \mathbf{1}). \qquad [92]$$

By Eq. (92), Eq. (66) with $i = 2, j = 1$ and Eq. (67),

$$\frac{\lambda_2(\mathbf{Q}_4)}{\mu} \geq \lambda_1(\mathbf{I}) + \lambda_2\left(\frac{1}{d-1}(\mathbf{I} - \mathbf{1})\right)$$

$$= \lambda_1(\mathbf{I}) - \frac{1}{d-1}\lambda_{d-1}(\mathbf{1} - \mathbf{I}). \qquad [93]$$

However, by Eq. (65), in the case $i = d-1, j = 1$,

$$\lambda_{d-1}(\mathbf{1} - \mathbf{I}) \leq \lambda_{d-1}(\mathbf{1}) + \lambda_d(-\mathbf{I}) = -1, \qquad [94]$$

where we used that

$$\lambda_1(\mathbf{1}) = \cdots = \lambda_{d-1}(\mathbf{1}) = 0. \qquad [95]$$

This latter claim follows immediately from the fact that $\mathrm{rank}(\mathbf{1}) = 1$ and $\mathbf{1}(1,\ldots,1)^\top = d(1,\ldots,1)^\top$. The first inequality in Eq. (91) is now a consequence of Eq. (93) and Eq. (94).

On the other hand, again by Eq. (92), Eq. (65), this time with $i = d, j = 0$, and Eq. (67),

$$\frac{\lambda_d(\mathbf{Q}_4)}{\mu} \leq \lambda_d(\mathbf{I}) + \lambda_d\left(\frac{1}{d-1}(\mathbf{I} - \mathbf{1})\right)$$
$$= \lambda_d(\mathbf{I}) - \lambda_1\left(\frac{1}{d-1}(\mathbf{1} - \mathbf{I})\right). \qquad [96]$$

However, by Eq. (66) and Eq. (95),

$$\lambda_1(\mathbf{1} - \mathbf{I}) \geq \lambda_1(\mathbf{1}) + \lambda_1(-\mathbf{I}) = -1. \qquad [97]$$

The third inequality in Eq. (91) is now a consequence of Eq. (96) and Eq. (97). This establishes Eq. (91) and, hence, Eq. (90), completing the proof. □

D. Surrogate-trajectory Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) (36, 37) is an advanced MCMC procedure that uses numerical approximations to Hamiltonian trajectories to generate Metropolis-type (16) proposals far away from the current Markov chain state. On the one hand, this approach to proposal generation helps reduce autocorrelation between chain states and is particularly helpful within higher-dimensional state spaces. On the other hand, numerical integration of Hamilton's equations requires repeated evaluation of the Hamiltonian potential energy's gradient, and these repeated floating-point operations may become computationally burdensome.

Lets briefly recall this HMC approach to resolve a given density $\pi(\cdot)$ of a 'target' probability measure of interest.* We proceed by considering a 'potential energy' of the form $U(\boldsymbol{\theta}) := -\log \pi(\boldsymbol{\theta})$. We then select an associated potential energy $V(\boldsymbol{\xi}) := |\mathbf{G}^{-1/2}\boldsymbol{\xi}|^2$ for an appropriately chosen symmetric-positive-definite mass matrix $\mathbf{G}$ (which is often taken as the identity for simplicity). One then observes that the Gibbs measure, proportional to $e^{-H(\boldsymbol{\theta}, \boldsymbol{\xi})}$ where $H = U + V$, is invariant under the associated Hamiltonian dynamic. Here note that the $\boldsymbol{\theta}$ marginal of $e^{-H(\boldsymbol{\theta}, \boldsymbol{\xi})}$ coincides with π while the $\boldsymbol{\xi}$ marginal is normally distributed as $\mathcal{N}(\mathbf{0}, \mathbf{G})$.

One operationalizes these observations as an algorithmic sampling procedure as follows. At each step, given a current sample $\boldsymbol{\theta}^{(n)}$, one draws $\boldsymbol{\xi}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$. From this initial state $(\boldsymbol{\theta}(0), \boldsymbol{\xi}(0)) := (\boldsymbol{\theta}^{(n)}, \boldsymbol{\xi}^{(n)})$ one then numerically approximates the associated Hamiltonian dynamics using a Störmer-Verlet (velocity Verlet) or leapfrog integrator up to a total integration time $\tau > 0$ and using integration step size $\epsilon > 0$. In this context, note that a single iteration of this integrator takes the form, (38),

$$\boldsymbol{\xi}\left(s + \frac{\epsilon}{2}\right) := \boldsymbol{\xi}(s) + \frac{\epsilon}{2}\nabla \log \pi(\boldsymbol{\theta}(s)),$$
$$\boldsymbol{\theta}(s + \epsilon) := \boldsymbol{\theta}(s) + \epsilon \, \mathbf{G}^{-1}\boldsymbol{\xi}(s + \frac{\epsilon}{2}), \qquad [98]$$
$$\boldsymbol{\xi}(s + \epsilon) := \boldsymbol{\xi}\left(s + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\nabla \log \pi(\boldsymbol{\theta}(s + \epsilon)).$$

In this fashion the proposed new state is given through Eq. (98) by $\boldsymbol{\theta}(\tau)$. To remove bias this procedure can be augmented with an acceptance probability of the form $\alpha^{(n)} := \exp(H(\boldsymbol{\theta}(0), \boldsymbol{\xi}(0)) - H(\boldsymbol{\theta}(\tau), \boldsymbol{\xi}(\tau)) \wedge 1$.

Different strategies aim to speed up the leapfrog integrator's many log-posterior gradient evaluations $\nabla \log \pi(\boldsymbol{\theta})$, as these numerical routines often represent the algorithm's computational bottleneck. In model-specific contexts, (3, 4, 39) yield parallelization strategies, and (32) develops dynamic programming techniques, to accelerate the evaluation of $\nabla \log \pi(\boldsymbol{\theta})$.

A small body of work considers another approach by replacing $\nabla \log \pi(\boldsymbol{\theta})$ with a suitable approximation $\widetilde{\nabla} \log \pi(\boldsymbol{\theta})$ and recognizing that the modified Eq. (98) continues to satisfy path reversibility and volume preservation, two essential ingredients for well-specified HMC. It follows immediately that the resulting 'surrogate trajectory HMC' continues to sample the correct target distribution $\pi(\cdot)$; see further details (37, 40). Nonetheless, the acceptance rates and overall efficiency of such samplers may suffer when approximations are poor.

The majority of surrogate HMC methods first obtain a small sample of exact gradient evaluations and then use some model to interpolate: (41) assumes the approximate gradient follows a piecewise constant form across a grid; (42, 43) construct approximations using Gaussian processes; and (44, 45) do the same using neural networks. But an even simpler approach to surrogate HMC may be appropriate when gradients have series representations as we can leverage here in Eq. (3).

---

*In the Bayesian inference context, we are typically considering the posterior density function $\pi(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\mathbf{Y})$ for the parameter $\boldsymbol{\theta} \in \mathbb{R}^K$ given observed data $\mathbf{Y}$.