2025 IEEE 64th Conference on Decision and Control (CDC)
December 10-12, 2025. Rio de Janeiro, Brazil

Genealogical Searching

Bernard Chazelle Kritkorn Karntikoon

***Abstract*— *Genealogical searching* refers to the algorithmic use of history as an error-correcting device. The adjective "genealogical" points to the availability of a lineage of causally related records. The issue arises in countless fields, such as biology, economics, linguistics, and control engineering. We establish minimal quantitative requirements for genealogical searching to be effective. Our work reveals a fundamental tradeoff between robustness and uncertainty, which we explore in several applications from microeconomics, linguistics, and machine learning.**

I. INTRODUCTION

Given a sequence of functions $(g_k)_{k\geq 0}$ from $\mathbb{R}^d$ to $\mathbb{R}^d$, we define the sequence $(a_k)_{k\geq 0}$ inductively by the rule $a_k = g_k(a_{k+1})$, for all $k \geq 0$. The goal of a *genealogical search (GS)* is to evaluate a_0 with high precision, given noisy access to a_k and g_k. In other words, the objective is to use information about the past as an error-correcting device.[1] Specifically, we assume that each a_k (resp. g_k) is accessible via a proxy $\bar{a}_k$ (resp. $\bar{g}_k$). For any $k \geq 0$, we estimate a_k by forming the convex combination

$$\hat{a}_k \leftarrow q_k \bar{g}_k(\hat{a}_{k+1}) + (1 - q_k)\bar{a}_k, \tag{1}$$

for weights $q_k \in [0, 1]$ of our own choosing. The recursion begins at the smallest k such that $q_k = 0$; we denote this index by k. (Note that the recursion index runs backwards.) The sequence $(\bar{a}_k, \bar{g}_k)_{k\geq 0}$ is called a *lineage*. We interpret $\bar{g}_k$ as a model for the iteration map g_k and $\bar{a}_k$ as a noisy observation of the signal a_k. Our task is to choose t and then set the weights q_k so as to minimize the distance between the prediction $\hat{a}_0$ and the true signal a_0. A *GS* comes with a threshold $\rho > 0$ and the goal is to find a sequence $q_0, \ldots, q_t$ ($q_t = 0$) that keeps that distance (i.e., the error) within ρ. The smallest value of t is called the *depth* $\mathbb{D}_\rho$ of the *GS* . Rather than introducing *GS* with a single, unified formal model, we discuss the main ideas behind it through a diverse range of applications, all of which form instances of control systems:

1) Economics: In markets subject to dynamic elasticities, prices can fluctuate without ever reaching equilibrium [5], [20], [29]. A *GS* approach lets us assess how far back in time we need to go in the exploration of supply/demand functions in order to predict future prices (Section II).
2) Linguistics: Iterated learning (IL) has emerged as a promising alternative to the "innate theory" of language evolution [6], [9], [14]. The idea is to model speakers as rational agents who update their linguistic priors by incorporating speech fragments from other sources. This establishes chains of teaching/learning pairs that trace the chronology of a language's evolution. The effective length of these chains has been a topic of great interest in the literature [2], [8], [16], [23], [24], [40]. We interpret the length as the depth of a *GS* lineage and we derive bounds that are nearly tight under common assumptions (Section III).
3) Machine learning: In the world of generative AI, diffusion probabilistic models have impressed lately by their uncanny ability to produce high-quality images from text [8,13,14,20]. Interpreting these models as *GS* lineages provides new insights into their analysis. In particular, we derive sufficient conditions under which the success of guided image synthesis [27] can be explained mathematically (Section IV).
4) A Bayesian perspective: Sometimes, the proxy $\bar{a}_k$ cannot be observed directly and only correlates y_k of the signal can be measured. In that case, we treat $\bar{g}_k(\hat{a}_{k+1})$ from (1) as a prior, which we update in light of the available data. Viewing *GS* through a Bayesian lens requires the use of matrices for the weights q_k, instead of scalars. For concreteness, we treat the Gaussian case and derive bounds on the depth of the corresponding lineages. We also discuss the issue of *infeasibility*, and we give examples of lineages with infinite depth (Section V).

Discussion

The concept of genealogical search is new (to our knowledge), but many of the underlying ideas are not. For example, using convex combinations of predictions and measurements is known as *filtering* in control theory and machine learning [1], [37]. The issue of physical infeasibility arises in quantum mechanics and classical dynamics [10], [26]. It is also well known that modeling from data can be intractable [13], [21], [22].

The error-correcting feature of a *GS* derives from three sources: (i) the law of large numbers, which permits every step back in time to finetune the final prediction; (ii) the multiplicative effect of the filtering process, which can be viewed as a form of *boosting* [38]; and (iii) the (optional) contractivity of the iteration maps.

Our work is self-contained and does not require any background in dynamical systems. For the experts in the subject, however, we add a few words here to dispel any possible confusion. For example, the low depth of a lineage may suggest a form of *ergodicity* but the analogy is flawed. Such

This work was supported in part by NSF grant CCF-2006125.

Princeton University, chazelle@cs.princeton.edu.
Princeton University, kritkorn@cs.princeton.edu.

[1] Observe that a_{k+1} *predates* a_k; this time-reversed notation simplifies the presentation.

lineages indeed forget about the distant past, but they usually lack stationary distributions. Typically, *GS* are not systems at equilibrium. Likewise, as we discuss in the examples below, low depth is not synonymous with high attraction rate. A better analogy is with the idea of *shadowing* in hyperbolic systems [31], [32]. There the goal is to produce real orbits tracking pseudo-orbits in chaotic systems. In *GS*, our objectives are, in some sense, the exact opposite: we seek pseudo-orbits tracking real (non-chaotic) orbits.

This work is the tip of the iceberg. We hope it opens a fruitful line of research: *GS* infeasibility, for example, is not merely of technical interest. It suggests a formal mechanism for deciding whether a given scientific objective might be unachievable, given the range of tools (theories and experiments) at one's disposal. In that context, the depth of a *GS* lineage can be interpreted as an "experimental" counterpart to the notion of circuit complexity.

II. Economics

For a simple illustration of *GS* in action, consider the classical *cobweb model* from microeconomics [5], [20], [29]. The goal is to predict prices from a supply-and-demand time-series. In this model, we have a single-commodity market with a demand of $d(k)$ at time $k \geq 0$. The demand function is linear: $d(k) = d^* - \alpha_k p(k)$, where $p(k)$ is the price at time k and $\alpha_k > 0$. Since production takes time, the supply's dependency on price lags behind. Up to scaling, we can assume the delay to be one unit of time; therefore $s(k) = s^* + \beta_k p(k-1)$, for some $\beta_k > 0$. The market-clearing price satisfies $d(k) = s(k)$; hence $p(k) = \tilde{c}_k + \tilde{d}_k p(k-1)$, where $\tilde{c}_k = (d^* - s^*)/\alpha_k$ and $\tilde{d}_k = -\beta_k/\alpha_k$ is the elasticity ratio at equilibrium ($|\tilde{d}_k| \leq 1$). The goal is to estimate the price $p(t)$ at time t.

A. The Model

Reframing the model in the language of *GS* , we fix t, the time of interest, and we reverse the chronology by setting $a_k := p(t-k)$. The map g_k satisfies $a_k = g_k(a_{k+1})$, where

$$g_k : x \mapsto c_k + d_k x,$$

for $c_k = \tilde{c}_{t-k}$ and $d_k = \tilde{d}_{t-k}$. The prediction for a_k, denoted by $\hat{a}_k$, is computed recursively: for $q_0, \ldots, q_{t-1} \in (0, 1]$ and $q_t = 0$,

$$\hat{a}_k = \hat{g}_k(\hat{a}_{k+1}) \text{ and } \hat{g}_k : x \mapsto q_k g_k(x) + (1-q_k)\bar{a}_k. \tag{2}$$

To bound the prediction error, we assume that $a_k = \bar{a}_k + \nu_k$, where $\nu_0, \nu_1, \ldots$ are unbiased, independent random variables with variances at most 1 (any uniform bound would work). We define the depth $\mathbb{D}_\rho$ of the *GS* lineage as the smallest t such that $\operatorname{Var}\delta_0 \leq \rho$, where $\delta_k := a_k - \hat{a}_k$. As we shall see below, the optimal setting of the weights is:

$$q_k = \frac{\operatorname{Var}\nu_k}{d_k^2 \operatorname{Var}\delta_{k+1} + \operatorname{Var}\nu_k}.$$

B. Bounding the Lineage Depth

Our objective is to approximate the depth $\mathbb{D}_\rho$. Let $\lambda = \max_{n>0} \frac{1}{n}\sum_{k=0}^{n-1} \ln|d_k|$ be the (maximum) Lyapunov exponent of the system. We state our main result.

Theorem 2.1. *Given any threshold ρ $(0 < \rho \leq 1)$, the depth of the* GS *satisfies*

$$\mathbb{D}_\rho \leq \begin{cases} \frac{1}{2|\lambda|}\ln\frac{1}{\rho} & \textit{if } \lambda < 0 \\ O(1/\rho) & \textit{if } \lambda = 0. \end{cases}$$

We begin the proof of Theorem 2.1 with a general identity:

Lemma 2.2. *If* Prec *denotes the precision of a random variable, i.e., the reciprocal of its variance, then*

$$\operatorname{Prec}\delta_0 = \sum_{k=0}^{t} \Big(\prod_{j=0}^{k-1} d_j^{-2}\Big) \operatorname{Prec}\nu_k\,.$$

Proof. By (2),

$$\begin{cases} \hat{g}_k(a_{k+1}) - \hat{g}_k(\hat{a}_{k+1}) = q_k\big(g_k(a_{k+1}) - g_k(\hat{a}_{k+1})\big) = q_k d_k \delta_{k+1} \\ g_k(a_{k+1}) - \hat{g}_k(a_{k+1}) = (1-q_k)\big(g_k(a_{k+1}) - \bar{a}_k\big) = (1-q_k)\nu_k. \end{cases}$$

Adding these two equations, we find that $\delta_k = g_k(a_{k+1}) - \hat{g}_k(\hat{a}_{k+1}) = q_k d_k \delta_{k+1} + (1-q_k)\nu_k$; hence, for $t > s \geq 0$,

$$\delta_s = \Big(\prod_{j=s}^{t-1} q_j d_j\Big)\delta_t + \sum_{k=s}^{t-1}\Big(\prod_{j=s}^{k-1} q_j d_j\Big)(1-q_k)\nu_k\,, \tag{3}$$

with the convention[2] that $\prod_a^b = 1$ if $a > b$. Because $q_t = 0$, we have $\delta_t = \nu_t$; hence $\mathbb{E}\,\delta_t = 0$ and $\operatorname{Var}\delta_t = \operatorname{Var}\nu_t$. By independence, it follows from (3) that δ_s is unbiased and

$$\operatorname{Var}\delta_s = (q_s d_s)^2 \operatorname{Var}\delta_{s+1} + (1-q_s)^2 \operatorname{Var}\nu_s.$$

We minimize $\operatorname{Var}\delta_s$ by setting $q_s = (\operatorname{Var}\nu_s)/(d_s^2 \operatorname{Var}\delta_{s+1} + \operatorname{Var}\nu_s)$, which gives us

$$\operatorname{Var}\delta_s = \frac{d_s^2\,(\operatorname{Var}\nu_s)\operatorname{Var}\delta_{s+1}}{d_s^2 \operatorname{Var}\delta_{s+1} + \operatorname{Var}\nu_s}\,;$$

hence the lemma. □

It follows from $|d_k| \leq 1$ that $\lambda \leq 0$. If $\lambda < 0$ then, by Lemma 2.2, $\operatorname{Var}\delta_0 \leq e^{2\lambda t}\operatorname{Var}\nu_t \leq e^{2\lambda t}$. If, on the other hand, $\lambda = 0$, then $\operatorname{Prec}\delta_0 \geq \sum_{k=0}^{t}\operatorname{Prec}\nu_k = \Omega(t)$. This completes the proof of Theorem 2.1. □

III. Linguistics

Iterated learning has emerged as a fundamental paradigm in the study of language evolution arose [2], [8], [16], [23], [24], [40]. Rather than appealing to the genetic fitting of innate constraints [6], [9], [14], the new approach situates the rise of universals in the information bottlenecks constraining the cross-generational transmission of language. More generally, the evolution of structure in human languages is now widely viewed as an adaptive response to the demands of iterating learning [2]–[4], [17], [23]–[25]. A key question in

[2] We follow that convention throughout the paper.

the field is how quickly an older language loses its "influence" on current ones [16], [34], [35]. A similar question can be asked about slang, memes, or trends.[3]

By framing the question "genealogically" and making a few standard assumptions, we are able to show that the loss of influence is surprisingly quick. A *lineage* is a sequence of speakers $0, 1, \dots, t$, each one equipped with their own language a_k. Beginning with speaker $k = t$, linguistic transmission from speaker k to speaker $k-1$ creates a chain of influence which results in language a_t leaving its mark on a_0. The minimum t for which this mark is negligible is called the *depth* of the lineage. We show that the depth is logarithmic in the language size. It is worth noting that, in general, such systems do *not* converge to fixed-point attractors.

A. The Model

We define the model formally. Following Chomsky and Lasnik [7], a language is specified as a probabilistic mixture of hypotheses (so-called proto-languages) forming a set $\mathcal{H} = \{h_1, \dots, h_n\}$. A hypothesis h_i is itself a probability distribution over a set $\mathcal{D} = \{d_1, \dots, d_N\}$ of sentences. Initially, each speaker $k = 0, 1, \dots$ is equipped with a prior distribution $\mathfrak{p}_k$ over $\mathcal{H}$. The learning process unfolds as follows: for $k = t-1, \dots, 0$, speaker k picks a random hypothesis from a_{k+1} and samples from it m times, thus producing $d_k \in \mathcal{D}^m$. With this evidence in hand, the speaker updates its prior $\mathfrak{p}_k$:[4]

$$\begin{cases} a_k : \ \mathbb{P}[h|k, d_k] = \mathbb{P}[d_k|h] \cdot \mathbb{P}_{\mathfrak{p}_k}[h] / \mathbb{P}[d_k|k] \\ \mathbb{P}[d_k|k] = \sum_{h \in \mathcal{H}} \mathbb{P}[d_k|h]\, \mathbb{P}_{\mathfrak{p}_k}[h], \end{cases} \tag{4}$$

where $d_k \in \mathcal{D}^m$ and $\mathbb{P}_{\mathfrak{p}_k}[h]$ is given by the prior $\mathfrak{p}_k$. The number m is fixed once and for all: it indicates how many times speaker k samples (without replacement) from a_{k+1}. This implies a causal chain of learning processes, beginning at $k = t$:

$$a_t \xrightarrow{\text{sample}} d_{t-1} \xrightarrow{\text{update}} a_{t-1} \quad \cdots \quad a_1 \xrightarrow{\text{sample}} d_0 \xrightarrow{\text{update}} a_0$$

We interpret the approximation of this process as a trimmed-down form of GS : $g_k(a_{k+1}) = a_k$; $q_t = 0$ and $q_k = 1$ for $0 \le k < t$; $\hat{g}_k := \bar{g}_k := g_k$; and $\bar{a}_t$ is an arbitrary probability distribution over $\mathcal{H}$. We define the depth $\mathbb{D}_\rho$ of the GS lineage as the smallest t such that $\max_{\bar{a}_t} \|a_0 - \hat{a}_0\|_1 \le \rho$.

B. Bounding the Lineage Depth

Let Δ denote the maximum total variation between any two hypotheses in $\mathcal{H}$, i.e., $\Delta = \frac{1}{2} \max_{i,j} \|h_i - h_j\|_1$, where $\|h_i - h_j\|_1 := \sum_{d \in \mathcal{D}^m} \big|\mathbb{P}[d|h_i] - \mathbb{P}[d|h_j]\big|$. For technical convenience, we assume that no prior excludes any hypothesis entirely, i.e., $\mathbb{P}_{\mathfrak{p}_k}[h_i] > 0$ for all k, h_i. We state our main result:

Theorem 3.1. *Given any threshold* ρ $(0 < \rho \le 1)$*, the depth of the* GS *is at most* $\log_{1/\Delta}(4/\rho)$.

Proof. The case $\Delta = 1$ holds trivially, so we assume that $\Delta < 1$. We define P_k as the n-by-n column-stochastic matrix given by

$$(P_k)_{ij} = \mathbb{P}_{\mathfrak{p}_k}[h_i] \sum_{d \in \mathcal{D}^m} \frac{\mathbb{P}[d|h_i]\mathbb{P}[d|h_j]}{\sum_l \mathbb{P}[d|h_l]\mathbb{P}_{\mathfrak{p}_k}[h_l]}\,. \tag{5}$$

By (4), $(P_k)_{ij} = \sum_{d \in \mathcal{D}^m} \mathbb{P}[h_i|k, d]\mathbb{P}[d|h_j]$ is the probability of picking hypothesis h_i from speaker k, conditioned on that speaker picking h_j when sampling the posterior a_{k+1} of speaker $k+1$. Thus, for $0 \le k < t$,

$$a_k = P_k\, a_{k+1}. \tag{6}$$

Note that $\mathbb{P}[d|h_i]\mathbb{P}[d|h_j] = 0$ implies that $|\mathbb{P}[d|h_i] - \mathbb{P}[d|h_j]| = \mathbb{P}[d|h_i] + \mathbb{P}[d|h_j]$. Thus, if the sum over d in (5) is null, $\|h_i - h_j]\|_1 = 2$; hence $\Delta = 1$. This case having been ruled out, it follows from $\mathbb{P}_{\mathfrak{p}_k}[h_i] > 0$ that the matrix P_k is positive.

Lemma 3.2. *If* $\delta_t \in \mathbb{R}^n$ *is a vector orthogonal to* $\mathbf{1}$*, then* $\|\delta_0\|_1 \le 2\Delta^t \|\delta_t\|_1$*, where* $\delta_k := P_k \delta_{k+1}$.

Proof. We define the *coefficient of ergodicity* $\tau(M)$ of the column-stochastic matrix M as $1 - \min_{i,j} \sum_k \min\{M_{ki}, M_{kj}\}$ or, equivalently, as half the maximum ℓ_1-distance between any two of its columns [39]: $\tau(M) = \frac{1}{2} \max_{i,j} \sum_{s=1}^n |M_{si} - M_{sj}|$. By (5),

$$\begin{aligned} \sum_{s=1}^n \big|(P_k)_{si} - (P_k)_{sj}\big| &= \sum_{s=1}^n \Big| \sum_{d \in \mathcal{D}^m} \tfrac{\mathbb{P}[d|h_s]\mathbb{P}_{\mathfrak{p}_k}[h_s]}{\sum_l \mathbb{P}[d|h_l]\mathbb{P}_{\mathfrak{p}_k}[h_l]} \big(\mathbb{P}[d|h_i] - \mathbb{P}[d|h_j]\big) \Big| \\ &\le \sum_{s=1}^n \sum_{d \in \mathcal{D}^m} \tfrac{\mathbb{P}[d|h_s]\mathbb{P}_{\mathfrak{p}_k}[h_s]}{\sum_l \mathbb{P}[d|h_l]\mathbb{P}_{\mathfrak{p}_k}[h_l]} \big|\mathbb{P}[d|h_i] - \mathbb{P}[d|h_j]\big| \\ &\le \sum_{d \in \mathcal{D}^m} \big|\mathbb{P}[d|h_i] - \mathbb{P}[d|h_j]\big| = \|h_i - h_j\|_1 \\ &\le 2\Delta\,. \end{aligned}$$

This proves that $\tau(P_k) \le \Delta$, for all k $(0 \le k < t)$. By the submultiplicativity of the coefficient of ergodicity, it then follows that

$$\tau(M) \le \prod_{k=0}^{t-1} \tau(P_k) \le \Delta^t, \tag{7}$$

where $M = \prod_{k=0}^{t-1} P_k$. Since each P_k is positive, M is primitive; by Perron-Frobenius, therefore, it has a unique stationary distribution π. Let M_j denote the j^{th} column of M, we observe that

$$M_j - \pi = M_j - M\pi = \sum_{i=1}^n \pi_i (M_j - M_i);$$

hence,

$$\|M_j - \pi\|_1 \le \sum_{i=1}^n \pi_i \|M_j - M_i\|_1 \le 2\tau(M).$$

Writing $Q = M - \pi \mathbf{1}^\top$, it then follows that $\|Q\|_1 = \max_j \|M_j - \pi\|_1 \le 2\tau(M)$. Since $\delta_0 = M\delta_t$ and $\mathbf{1}^\top \delta_t = 0$, we have $\delta_0 =$

[3] Informally, a source has lost influence (also called sensitivity) on its product if heavy perturbations do not alter the product significantly.

[4] Conditioning on k is an abuse of notation since k is not a random variable, but it is a useful specification of the relevant speaker.

$Q\delta_t$; hence,

$$\begin{aligned}\|\delta_0\|_1 = \|Q\delta_t\|_1 &\le \sum_{j=1}^{n} \left|(\delta_t)_j\right| \, \|Q_j\|_1 \\ &\le \sum_{j=1}^{n} \left|(\delta_t)_j\right| \, \|Q\|_1 \\ &\le 2\tau(M) \, \|\delta_t\|_1 .\end{aligned}$$

Combining this with (7) proves the lemma. □

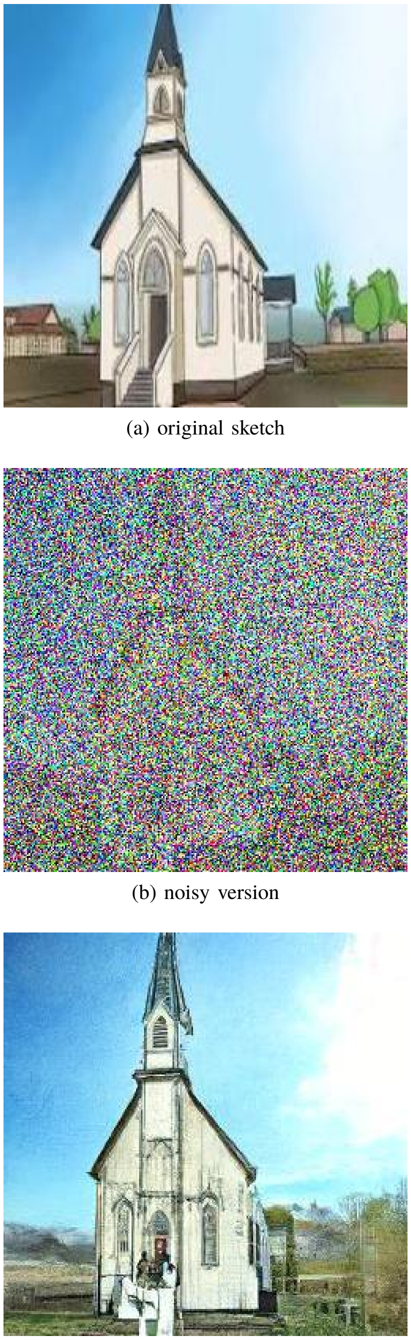
To establish Theorem 3.1, we observe that both a_k and $\hat{a}_k$ satisfy (6): each one is a probability distribution, so $\delta_k = a_k - \hat{a}_k$ is orthogonal to $\mathbf{1}$ and $\|\delta_k\|_1 \le 2$. By Lemma 3.2, therefore, $\|\delta_0\|_1 \le 4\Delta^t$, which completes the proof. □

IV. Machine Learning

Diffusion probabilistic models have been used with great success to generate high-quality images from text [19], [30], [33], [41]. We show how interpreting these models as *GS* lineages provides new insights into their analysis. A diffusion model consists of two parts:

1) The *encoder* receives an image $a_0 \in \mathbb{R}^d$ from a training set collected from an unknown distribution $\mathcal{D}$ and computes a sequence of intermediate latent variables $a_1, \ldots, a_t$. Each image a_k is derived from a_{k-1} by adding a small amount of noise. The process forms a Markov chain whose effect is to blur a_0 by small increments until all of its information content is lost. The operation is repeated over all the images in the training set.
2) The *decoder* attempts to reverse the encoder's sequence, starting with pure noise $\hat{a}_t$ and producing $\hat{a}_{t-1}, \ldots, \hat{a}_0$ in that order via denoising. This is achieved by using a neural net (typically, a *U-net*) in order to guide the sequence toward the unknown distribution $\mathcal{D}$ from which the training set has been drawn. The goal is to produce an image $\hat{a}_0$ that approximates a sampled image from $\mathcal{D}$.

In *guided image synthesis* [27], the goal is to input a rough sketch of, say, a horse and produce a photorealistic enhancement of the sketch. We are *not allowed* to retrain the model. One popular approach is to feed the sketch to a diffusion model and interrupt the encoder before the noising process destroys all the information. The idea then is to run the decoder until it yields an image $\hat{a}_0$ that closely resembles an actual horse yet retains the salient features of our drawing. This process is a good example of genealogical searching. We provide theoretical guarantees for the approach to succeed. We tried it successfully on a public available diffusion model (Fig. 1).



(a) original sketch

(b) noisy version

(c) denoised image

Fig. 1. Adding and removing noise from a hand-drawn picture to make it more realistic. Produced by using von Platen et al's *diffuser* [43] (https://github.com/huggingface/diffusers).

A. The Model

For the encoder, the latent variables are derived by setting the recurrence: for $k > 0$,

$$a_k = \sqrt{1 - \beta_k}\, a_{k-1} + \sqrt{\beta_k}\, \epsilon_k, \tag{8}$$

where $\epsilon_k \sim \mathcal{N}(0, \mathbb{I})$ and $\mathbb{I}$ is the d-by-d identity matrix; the β_k form a sequence of positive reals called the *noise schedule*. (See [19], [33] for a detailed explanation of the formulas.) By standard properties of the normal distribution, unraveling the recurrence gives us the closed form:

$$a_k = \sqrt{\alpha_k}\, a_0 + \sqrt{1-\alpha_k}\, \epsilon'_k, \tag{9}$$

where $\alpha_k = \prod_{i=1}^{k}(1-\beta_i)$ and $\epsilon'_k \sim \mathcal{N}(0,\mathbb{I})$. To set up the decoder, we could reverse the process by applying Bayes' rule, but this would crash against the intractability of the needed marginals. Instead, we use a Gaussian approximation: for $k > 1$,

$$a_{k-1} \sim \mu_\theta(a_k, k) + \sigma_k\,, \tag{10}$$

where $\sigma_k \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}_k^2\,\mathbb{I})$, with untrained time-dependent variances $\tilde{\sigma}_k^2$, and

$$\mu_\theta(y,k) := \frac{1}{\sqrt{1-\beta_k}}\Big(y - \frac{\beta_k}{\sqrt{1-\alpha_k}}\,\epsilon_\theta(y,k)\Big). \tag{11}$$

The model $\epsilon_\theta(a_k, k)$ is sketched below: it is a neural net trained to predict the noise ϵ'_k in (9), for a random $a_0 \sim \mathcal{D}$.

Algorithm 1 Training

```
Input: Training data set $S \subseteq \mathcal{D}$
Output: Model $\epsilon_\theta$
  repeat
     Pick $a_0 \sim S$
     $k \sim \mathrm{Uniform}[1,\ldots,t]$
     $\epsilon \sim \mathcal{N}(0,\mathbb{I})$
     Take gradient descent step on $\nabla_\theta \big\|\epsilon - \epsilon_\theta\big(\sqrt{\alpha_k}\, a_0 + \sqrt{1-\alpha_k}\,\epsilon, k\big)\big\|^2$
  until converged
```

Algorithm 2 Sampling

```
Input: Model $\epsilon_\theta$
Output: New sample $\hat{a}$
  $\hat{a}_t \sim \mathcal{N}(0,\mathbb{I})$
  for $k = t,\ldots,1$ do
     $\sigma_k \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}_k^2\,\mathbb{I})$ if $k > 1$; else $\sigma_k = \mathbf{0}$
     $\hat{a}_{k-1} = \mu_\theta(\hat{a}_k, k) + \sigma_k$
  end for
  return $\hat{a}_0$
```

Since $\epsilon_\theta(\hat{a}_k, k)$ predicts ϵ'_k and the error is Gaussian, it follows from (9) that

$$\epsilon_\theta(\hat{a}_k, k) = \sqrt{\frac{\alpha_k}{1-\alpha_k}}\left(\frac{\hat{a}_k}{\sqrt{\alpha_k}} - a_0\right) + \gamma_k \tag{12}$$

where $\gamma_k \sim \mathcal{N}(\mathbf{0}, \tilde{\gamma}_k^2\,\mathbb{I}_d)$, for real $\tilde{\gamma}_k$.

B. How Much Noise Can Reconstruction Tolerate?

For the purposes of our analysis, we may assume that $d = 1$, the generalization to higher dimension being straightforward. It is common to pick a constant or linearly increasing noise schedule [19]: We assume that $\tilde{\gamma}_k \le \beta_k = \beta$, for all k, for fixed $0 < \beta \le 1/2$. In practice, $\tilde{\sigma}_k^2$ is set to β_k [19]. Fix an arbitrary $a_0 \in \mathcal{D}$, and let $a_1, \ldots, a_t$ (resp. $\hat{a}_t, \ldots, \hat{a}_0$) be its successive iterations through the encoding (resp. decoding) steps. We define the reconstruction error $\mathcal{E}_t = \mathbb{E}_{\hat{a}_t \sim \mathcal{N}(a_t, 1-\alpha_t)}\delta_1^2$, where $\delta_k := a_k - \hat{a}_k$.[5]

Theorem 4.1. *Given any threshold $\rho \ge 18\beta$, the reconstruction error satisfies $\mathcal{E}_t \le \rho$, for any $t \le 1/\beta$.*

Proof. Given $a_0 \in \mathcal{D}$, let $a_k = \sqrt{\alpha_k}\, a_0$, for $k > 0$. This sequence follows the same pattern as defined by (8) but with all noise terms $\epsilon_k = 0$. By (10, 11, 12),

$$\begin{aligned}\frac{\hat{a}_{k-1}}{\sqrt{\alpha_{k-1}}} &= \frac{\hat{a}_k}{\sqrt{\alpha_k}} - \frac{\beta_k}{\sqrt{\alpha_k(1-\alpha_k)}}\,\epsilon_\theta(\hat{a}_k,k) + \frac{\sigma_k}{\sqrt{\alpha_{k-1}}}\\ &= \left(1 - \frac{\beta_k}{1-\alpha_k}\right)\frac{\hat{a}_k}{\sqrt{\alpha_k}} + \left(\frac{\beta_k}{1-\alpha_k}\right)a_0\\ &\quad - \frac{\beta_k\gamma_k}{\sqrt{\alpha_k(1-\alpha_k)}} + \frac{\sigma_k}{\sqrt{\alpha_{k-1}}}\,.\end{aligned} \tag{13}$$

By $a_k = \sqrt{\alpha_k}\, a_0$,

$$\frac{a_{k-1}}{\sqrt{\alpha_{k-1}}} = \left(1 - \frac{\beta_k}{1-\alpha_k}\right)\frac{a_k}{\sqrt{\alpha_k}} + \frac{\beta_k}{1-\alpha_k}a_0. \tag{14}$$

Combining (13, 14) yields:

$$\frac{\delta_{k-1}}{\sqrt{\alpha_{k-1}}} = \left(1 - \frac{\beta_k}{1-\alpha_k}\right)\frac{\delta_k}{\sqrt{\alpha_k}} + \frac{\beta_k\gamma_k}{\sqrt{\alpha_k(1-\alpha_k)}} - \frac{\sigma_k}{\sqrt{\alpha_{k-1}}}\,.$$

By induction on k, we find that $\mathbb{E}\,\delta_k = 0$. Using the independence of the random variables, we derive

$$\mathrm{Var}\,\delta_{k-1} = \lambda_k \mathrm{Var}\,\delta_k + \nu_k, \tag{15}$$

where

$$\lambda_k := (1-\beta_k)\left(\frac{1-\alpha_{k-1}}{1-\alpha_k}\right)^2 \quad\text{and}\quad \nu_k := \frac{\beta_k^2\,\tilde{\gamma}_k^2}{(1-\alpha_k)(1-\beta_k)} + \tilde{\sigma}_k^2\,.$$

It follows that

$$\mathrm{Var}\,\delta_1 = \Big(\prod_{k=2}^{t}\lambda_k\Big)\mathrm{Var}\,\delta_t + \sum_{k=2}^{t}\Big(\prod_{j=2}^{k-1}\lambda_j\Big)\nu_k\,. \tag{16}$$

Since $(1-x)^k \le 1 - kx/2$, for $0 \le xk \le 1$, the inequalities $\beta \le 1/2$ and $t \le 1/\beta$ ensure that

$$\prod_{k=2}^{\ell}\lambda_k \le (1-\beta)^{\ell-1}\Big(\frac{1-\alpha_1}{1-\alpha_\ell}\Big)^2 \le \Big(\frac{\beta}{1-(1-\beta)^\ell}\Big)^2 \le \frac{4}{\ell^2}\,.$$

It follows from our assumption that $\tilde{\gamma}_k \le \beta$ that $\nu_k \le 2\beta^3 + \beta \le 2\beta$. Since a_0 is fixed, so is a_k; the prediction $\hat{a}_t$ is sampled from $\mathcal{N}(a_t, 1-\alpha_t)$; therefore $\mathrm{Var}\,\delta_t = 1-\alpha_t$. By (16) and the inequality $(1-x)^k \ge 1 - kx$ $(0 \le x \le 1)$,

$$\begin{aligned}\mathrm{Var}\,\delta_1 &= \Big(\prod_{k=2}^{t}\lambda_k\Big)(1-\alpha_t) + \sum_{k=2}^{t}\Big(\prod_{j=2}^{k-1}\lambda_j\Big)\nu_k\\ &\le \frac{4\beta}{t} + 8\beta\sum_{k=1}^{\infty}\frac{1}{k^2}\\ &\le 4\beta + 4\beta\pi^2/3 < 18\beta\,.\end{aligned} \tag{17}$$

□

[5] We use δ_1 instead of δ_0 to avoid a minor technical complication caused by the specification of σ_1 in Algorithm 2.

V. A Bayesian Perspective

We consider the common situation where no direct empirical evaluation of the signal $a_k \in \mathbb{R}^d$ is possible. Instead, we have access to measurements encoded in a vector y_k. For nonlinear (smooth) maps g_k, we consider a first-order approximation $g_k : x \mapsto \mathbb{J}_k x + c_k$ (for a nonsingular d-by-d Jacobian matrix $\mathbb{J}_k$) and a model map $\bar{g}_k = g_k$. The predictor is of the form

$$\hat{a}_k = Q_k \bar{g}_k(\hat{a}_{k+1}) + (\mathbb{I} - Q_k)\bar{a}_k. \tag{18}$$

We upgrade the weight Q_k from a scalar to a d-by-d matrix because the variable $\bar{a}_k$ is now hidden. The measurement is of the form $y_k = H_k a_k + v_k$, where H_k is a d'-by-d matrix and $v_k \sim \mathcal{N}(\mathbf{0}, B_k)$, where B_k is a d'-by-d' positive definite covariance matrix.[6] Accordingly, we ensure that $\mathbb{I} - Q_k$ can be written as $M_k H_k$, so that recurrence (18) becomes:

$$\hat{a}_k = Q_k \bar{g}_k(\hat{a}_{k+1}) + M_k y_k. \tag{19}$$

Fix $t > 0$. Inferring a_k from the available sequence of measurements $y_k, \ldots, y_t$ (denoted by $y_{k:t}$) is done by inverting the inference via Bayes' rule. The closure of the normal distribution under conditioning implies that $\mathbb{P}[a_k \mid y_{k:t}]$ is of the form $\mathcal{N}(\hat{a}_k, \Sigma_k)$, for some positive definite matrix Σ_k. We set Q_t to zero and assume that $a_t \sim \mathcal{N}(\bar{a}_t, c\mathbb{I})$, for any $\bar{a}_t$ and a constant c large enough to make the signal arbitrarily noisy; for the purposes of the calculations, we can assume that $c = 1$; hence $\Sigma_t = \mathbb{I}$. The depth $\mathbb{D}_\rho$ of the *GS* lineage is the smallest t such that the spectral radius of Σ_0 is at most ρ.[7]

A. Bounding the Lineage Depth

Theorem 5.1. *If the spectral norm of each transition matrix $\mathbb{J}_k$ is at most $\sigma < 1$; then, given any threshold ρ $(0 < \rho \leq 1)$, the depth of the* GS *satisfies*

$$\mathbb{D}_\rho \leq \tfrac{1}{2} \log \tfrac{1}{\rho} \Big/ \log \tfrac{1}{\sigma}\,.$$

We begin the proof with a recurrence on the covariance matrices of the predictions. (See [28] for an introduction to Bayesian filtering.)

Lemma 5.2. *Using Bayes' rule to integrate the measurements $y_{0:t}$ yields the weight matrix $Q_k = \Sigma_k \big(\mathbb{J}_k \Sigma_{k+1} \mathbb{J}_k^\top\big)^{-1} = \mathbb{I} - M_k H_k$, where $M_k = \Sigma_k H_k^\top B_k^{-1}$. The covariance matrix Σ_k is defined by the recurrence:*

$$\Sigma_k^{-1} = H_k^\top B_k^{-1} H_k + \big(\mathbb{J}_k \Sigma_{k+1} \mathbb{J}_k^\top\big)^{-1}.$$

Proof. By Bayes' rule, $\mathbb{P}[a_k \mid y_{k:t}] = \frac{\mathbb{P}[y_k \mid a_k]\mathbb{P}[a_k \mid y_{k+1:t}]}{\mathbb{P}[y_k \mid y_{k+1:t}]}$; hence,

$$\ln \mathbb{P}[a_k \mid y_{k:t}] = \ln \mathbb{P}[y_k \mid a_k] + \ln \mathbb{P}[a_k \mid y_{k+1:t}] + \mathit{const}, \tag{20}$$

where *const* does not depend on a_k. The distribution of y_k conditioned on a_k is normal with expectation $H_k a_k$; hence it is of the form $\mathcal{N}(H_k a_k, B_k)$; likewise, a_k conditioned on $y_{k+1:t}$ is drawn from $\mathcal{N}(\mathbb{J}_k \hat{a}_{k+1} + c_k, \widehat{\Sigma}_k)$, for some positive definite matrix $\widehat{\Sigma}_k$; thus,

$$\begin{aligned}
-2 \ln \mathbb{P}[y_k \mid a_k] &= (y_k - H_k a_k)^\top B_k^{-1} (y_k - H_k a_k) + \mathit{const} \\
&= a_k^\top H_k^\top B_k^{-1} H_k a_k - 2 a_k^\top H_k^\top B_k^{-1} y_k \\
&\quad + y_k^\top B_k^{-1} y_k + \mathit{const}
\end{aligned}$$

and

$$\begin{aligned}
& -2 \ln \mathbb{P}[a_k \mid y_{k+1:t}] \\
&= (a_k - \mathbb{J}_k \hat{a}_{k+1} - c_k)^\top \widehat{\Sigma}_k^{-1} (a_k - \mathbb{J}_k \hat{a}_{k+1} - c_k) + \mathit{const} \\
&= a_k^\top \widehat{\Sigma}_k^{-1} a_k - 2 a_k^\top \widehat{\Sigma}_k^{-1} (\mathbb{J}_k \hat{a}_{k+1} + c_k) \\
&\quad + (\mathbb{J}_k \hat{a}_{k+1} + c_k)^\top \widehat{\Sigma}_k^{-1} (\mathbb{J}_k \hat{a}_{k+1} + c_k) + \mathit{const}
\end{aligned}$$

Combining both identities with (20), we find that

$$\begin{aligned}
-2 \ln \mathbb{P}[a_k \mid y_{k:t}] &= -2 a_k^\top \big(H_k^\top B_k^{-1} y_k + \widehat{\Sigma}_k^{-1} (\mathbb{J}_k \hat{a}_{k+1} + c_k)\big) \\
&\quad + a_k^\top \big(H_k^\top B_k^{-1} H_k + \widehat{\Sigma}_k^{-1}\big) a_k + \mathit{const}
\end{aligned}$$

By pairwise identification with $-2 \ln \mathbb{P}[a_k \mid y_{k:t}] = (a_k - \hat{a}_k)^\top \Sigma_k^{-1} (a_k - \hat{a}_k) + \mathit{const}$, we derive:

$$\begin{aligned}
\Sigma_k^{-1} &= H_k^\top B_k^{-1} H_k + \widehat{\Sigma}_k^{-1}\ \text{, and} \\
\hat{a}_k &= \Sigma_k \big(H_k^\top B_k^{-1} y_k + \widehat{\Sigma}_k^{-1} (\mathbb{J}_k \hat{a}_{k+1} + c_k)\big),
\end{aligned} \tag{21}$$

where

$$\widehat{\Sigma}_k = \operatorname{Cov}\left[\mathbb{J}_k a_{k+1} + c_k \mid y_{k+1:t}\right] = \mathbb{J}_k \Sigma_{k+1} \mathbb{J}_k^\top. \tag{22}$$

By (19), $\hat{a}_k = Q_k(\mathbb{J}_k \hat{a}_{k+1} + c_k) + M_k y_k$; hence, by (21), $M_k = \Sigma_k H_k^\top B_k^{-1}$ and $Q_k = \Sigma_k \widehat{\Sigma}_k^{-1}$. We verify that $\mathbb{I} - Q_k = M_k H_k$ and that the lemma follows from (22). □

It follows from the lemma and $\Sigma_t = \mathbb{I}$ that

Corollary 5.3..

$$\begin{aligned}
\Sigma_0^{-1} = \sum_{k=0}^{t-1} &\left\{ \prod_{j=0}^{k-1} \big(\mathbb{J}_j^\top\big)^{-1} \right\} H_k^\top B_k^{-1} H_k \left\{ \prod_{j=0}^{k-1} \mathbb{J}_j \right\}^{-1} \\
&+ \left\{ \prod_{j=0}^{t-1} \big(\mathbb{J}_j^\top\big)^{-1} \right\} \left\{ \prod_{j=0}^{t-1} \mathbb{J}_j \right\}^{-1}.
\end{aligned}$$

Proof of Theorem 5.1. We denote by $\Lambda(M)$ the spectrum of a square matrix M and $\mathbb{S}(M)$ its set of singular values. By the submultiplicativity of the matrix 2-norm and the positive

[6] The reason for choosing a normal distribution is the same as with Kalman filters [15]: it is often a sound choice in practice and it allows for the closed-form expression of the relevant distributions.

[7] The spectrum of a square matrix M is the set of its eigenvalues: their largest modulus is called the spectral radius. The spectral norm of M is its largest singular value: it is the square root of the spectral radius of $M^\top M$ (or, equivalently, $M M^\top$) and it is equal to $\max_{\|x\|_2 = 1} \|Mx\|_2$.

semidefiniteness of the summands in Corollary 5.3,

$$\begin{aligned}\frac{1}{\max \Lambda(\Sigma_0)} &= \min \Lambda\big(\Sigma_0^{-1}\big) = \min_{\|x\|_2=1} x^T \Sigma_0^{-1} x \\ &\geq \min_{\|x\|_2=1} \Big\| \Big(\prod_{j=0}^{t-1} \mathbb{J}_j \Big)^{-1} x \Big\|_2^2 \\ &\geq \min \Lambda \Big\{ \Big(\prod_{j=0}^{t-1} (\mathbb{J}_j^\top)^{-1} \Big) \Big(\prod_{j=0}^{t-1} \mathbb{J}_j \Big)^{-1} \Big\} \\ &\geq 1 \Big/ \max \Lambda \Big\{ \Big(\prod_{j=0}^{t-1} \mathbb{J}_j \Big) \Big(\prod_{j=0}^{t-1} \mathbb{J}_j \Big)^\top \Big\} \\ &\geq 1 \Big/ \max_{\|x\|_2=1} \Big\| \Big(\prod_{j=0}^{t-1} \mathbb{J}_j \Big) x \Big\|_2^2 = 1 \Big/ \Big\| \prod_{j=0}^{t-1} \mathbb{J}_j \Big\|_2^2 \\ &\geq 1 \Big/ \prod_{j=0}^{t-1} \|\mathbb{J}_j\|_2^2 \geq \frac{1}{\sigma^{2t}} \,.\end{aligned}$$

□

To draw a parallel with the economics application discussed in Section II, we set the dimension $d = 1$ for a_k and, accordingly, use lower-case letters for the main variables. By Corollary 5.3, we have

$$\operatorname{Prec} a_0 = \sum_{k=0}^{t-1} \frac{h_k^2}{b_k} \Big(\prod_{j=0}^{k-1} \frac{1}{q_j} \Big) + \prod_{j=0}^{t-1} \frac{1}{q_j}, \tag{23}$$

where $q_k = |\mathbb{J}_k|^2$. Note the similarity with Lemma 2.2: in particular, the tension between the noise (which is allowed to grow with time) and the twin denoising effect of (i) contracting maps ($q_k < 1$) together with (ii) the law of large numbers (addition of the precisions). Of course, relation (23) is less general than Lemma 2.2 because of its Gaussian assumptions.

B. When Is Genealogical Searching Feasible?

We show how, regardless of depth, the presence of too much noise in the transition map can make a genealogical search infeasible. Using the notation of Section V, we set

$$g_k : x \mapsto \mathbb{J}_k x + \mu_k,$$

where $\mu_k \sim \mathcal{N}(c_k, A_k)$. The proof of Lemma 5.2 remains unchanged, except for (22), which becomes: $\widehat{\Sigma}_k = \mathbb{J}_k \Sigma_{k+1} \mathbb{J}_k^\top + A_k$. This implies that $Q_k = \Sigma_k \big(\mathbb{J}_k \Sigma_{k+1} \mathbb{J}_k^\top + A_k\big)^{-1}$ and $\Sigma_k^{-1} = H_k^\top B_k^{-1} H_k + \big(\mathbb{J}_k \Sigma_{k+1} \mathbb{J}_k^\top + A_k\big)^{-1}$. The addition of A_k to the equations complicates their resolution. Borrowing a technique used for Riccati equations [15], we write Σ_k as a matrix fraction $U_k V_k^{-1}$ and derive:

$$\begin{aligned}\Sigma_k^{-1} &= V_k U_k^{-1} = H_k^\top B_k^{-1} H_k + \big(\mathbb{J}_k U_{k+1} V_{k+1}^{-1} \mathbb{J}_k^\top + A_k\big)^{-1} \\ &= H_k^\top B_k^{-1} H_k + \big(\big(\mathbb{J}_k U_{k+1} + A_k (\mathbb{J}_k^\top)^{-1} V_{k+1}\big) V_{k+1}^{-1} \mathbb{J}_k^\top\big)^{-1} \\ &= H_k^\top B_k^{-1} H_k + (\mathbb{J}_k^\top)^{-1} V_{k+1} \big(\mathbb{J}_k U_{k+1} + A_k (\mathbb{J}_k^\top)^{-1} V_{k+1}\big)^{-1} \\ &= \big\{H_k^\top B_k^{-1} H_k \mathbb{J}_k U_{k+1} + \big(H_k^\top B_k^{-1} H_k A_k (\mathbb{J}_k^\top)^{-1} + (\mathbb{J}_k^\top)^{-1}\big) V_{k+1}\big\} \\ &\quad \times \big(\mathbb{J}_k U_{k+1} + A_k (\mathbb{J}_k^\top)^{-1} V_{k+1}\big)^{-1};\end{aligned} \tag{24}$$

hence

$$\begin{pmatrix} U_k \\ V_k \end{pmatrix} = \begin{pmatrix} \mathbb{J}_k & A_k (\mathbb{J}_k^\top)^{-1} \\ H_k^\top B_k^{-1} H_k \mathbb{J}_k & H_k^\top B_k^{-1} H_k A_k (\mathbb{J}_k^\top)^{-1} + (\mathbb{J}_k^\top)^{-1} \end{pmatrix} \begin{pmatrix} U_{k+1} \\ V_{k+1} \end{pmatrix}. \tag{25}$$

This provides a linear recurrence relation to compute the covariance matrix Σ_0, as opposed to the recurrence in the first line of (24), which is not linear.

1) The Time-Invariant Scalar Case: For concreteness, we focus on a time-invariant GS system on the line ($d = 1$); For this reason, we drop the subscript k and use lower-case letters. This gives us $\sigma_k = u_k / v_k$, where $u_t = \sigma_t$, $v_t = 1$. By (25),

$$\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \Big(\frac{M}{j}\Big)^t \begin{pmatrix} \sigma_t \\ 1 \end{pmatrix}, \text{ with } M = \begin{pmatrix} q & a \\ qh^2/b & ah^2/b + 1 \end{pmatrix},$$

for $a, b, h > 0$ and $q := j^2 = |\mathbb{J}_k|^2 \in (0, 1)$. Write $\alpha = \sqrt{\big(ah^2 + b(q+1)\big)^2 - 4b^2 q} > 0$.[8] The two eigenvalues λ_1, λ_2 of M are real and

$$\lambda_2 = \frac{ah^2 + b(q+1) + \alpha}{2b} > \lambda_1 = \frac{ah^2 + b(q+1) - \alpha}{2b} > 0,$$

so that $\xi := \lambda_1 / \lambda_2 \in (0, 1)$; this follows from the identity $\alpha = \sqrt{\big(ah^2 + b(1-q)\big)^2 + 4abqh^2}$. Writing $\beta = \alpha + ah^2 + b(1-q)$ and $\gamma = \alpha - ah^2 - b(1-q)$, we diagonalize M to find that

$$M^t = \frac{1}{4q\alpha h^2} \begin{pmatrix} -\beta & \gamma \\ 2qh^2 & 2qh^2 \end{pmatrix} \begin{pmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{pmatrix} \begin{pmatrix} -2qh^2 & \gamma \\ 2qh^2 & \beta \end{pmatrix}.$$

Put $p = 2qh^2$; then

$$\sigma_0 = \frac{1}{p} \cdot \frac{(p\sigma_t - \gamma)\beta\xi^t + (p\sigma_t + \beta)\gamma}{(\gamma - p\sigma_t)\xi^t + p\sigma_t + \beta} \,.$$

As t goes to infinity, σ_0 converges to

$$\begin{aligned}\sigma_0^* := \frac{\gamma}{p} &= \frac{\sqrt{\big(ah^2 + b(1-q)\big)^2 + 4abqh^2} - ah^2 - b(1-q)}{2qh^2} \\ &\leq \frac{ab}{ah^2 + b(1-q)}.\end{aligned} \tag{26}$$

The depth $\mathbb{D}_\rho$ is bounded if $\sigma_0^* \leq \rho$; or, equivalently, if $\operatorname{Prec} a_0 \geq 1/\rho$, for some finite t. In view of (26), we conclude with a sufficient feasibility condition:

Theorem 5.4. *In the scalar, time-invariant version of Bayesian-based GS, the depth $\mathbb{D}_\rho$ is bounded if:*

$$h^2 \operatorname{Prec} \nu + (1 - q) \operatorname{Prec} \mu \geq \frac{1}{\rho} \,.$$

[8] Note that $(ah^2 + b(q+1))^2 - 4b^2 q = a^2h^4 + 2abh^2(q+1) + b^2(q-1)^2 \geq a^2h^4 > 0$.

VI. Conclusions

There are many more applications of *GS* left unmentioned in this work. These include troubleshooting, fault diagnosis, forensic analysis, and control engineering. Perhaps the most pressing application is biology. The physiology of a living cell is mediated by its gene products and a largely inaccessible soup of weakly interacting molecules called "dark matter" [36]. The latter forms a major obstacle in cell biology that data collection alone has not been able to overcome: it corresponds to the $\bar{a}_k$ terms in (1). The alternative is to "recreate" (in vitro or in silico) the production of the missing dark matter (the $\bar{g}_t$ terms). This is an exciting new area of research in theoretical biology [11], [18], [42].

References

[1] Bain, A., Crisan, D, Fundamentals of Stochastic Filtering, (Stoch. Modelling Appl. Prob. 60), Springer-Verlag, New York, 2009.

[2] Brighton, H. *Compositional syntax from cultural transmission*, Artificial life 8, 1 (2002), 25–54.

[3] Briscoe, T. *Linguistic evolution through language acquisition*, Cambridge University Press, 2002.

[4] Chazelle, B., Wang, C. *Iterated Learning in Dynamic Social Networks*, Journal of Machine Learning Research 20, 29 (2019), 1–28.

[5] Chiang, A.C., Wainwright, K. Fundamental Methods of Mathematical Economics, McGraw-Hill, New York, 2005.

[6] Chomsky, N. Aspects of the Theory of Syntax, MIT Press, 50th ed., 1965.

[7] Chomsky, N., Lasnik, H. Principles and Parameters Theory, in Syntax: An International Handbook of Contemporary Research (1993), Berlin: de Gruyter.

[8] Christiansen, M.H., Kirby, S. *Language evolution: Consensus and controversies*, Trends in cognitive sciences 7, 7 (2003), 300–307.

[9] Comrie, B. Language universals and linguistic typology: Syntax and morphology, University of Chicago press, 1989.

[10] Cubitt, T.S., Pérez-García, D., Wolf. M.M. *Undecidability of the spectral gap*, Nature 528 (2015), 207–211.

[11] De Capitani, J., Mutschler, H. *The long road to a synthetic self-replicating central dogma*, Biochemistry 2023, 62, 7 (2023), 1221–1232.

[12] Gagnon, Y.L., Templin, R.M., How, M.J., Marshall, N.J. *Circularly polarized light as a communication signal in mantis shrimps*, Current Biology 25, 23 (2015), 3074–3078.

[13] Goldreich, O., Goldwasser, S., Micali, S. *How to construct random functions*, J. ACM 3, 43 (1986), 792–807.

[14] Greenberg, J.H. Universals of language, MIT Press, 1963.

[15] Grewal, M.S., Andrews, A.P. Kalman Filtering, Wiley-IEEE Press; 4th edition (December 31, 2014).

[16] Griffiths, T.L., Kalish, M.L. *A Bayesian view of language evolution by iterated learning*, Proceedings of the annual meeting of the Cognitive Science Society 27 (2005).

[17] Griffiths, T.L., Kalish, M.L. *Language evolution by iterated learning with Bayesian agents*, Cognitive Science 31, 3 (2007), 441–480.

[18] Guindani, C., Caire da Silva, L., Cao, S., Ivanov, T., Landfester, K. *Synthetic cells: from simple bio-inspired modules to sophisticated integrated systems*, Angew. Chem. International Ed. 61, 16 (2022).

[19] Ho, J., Jain, A., Abbeel, P. *Denoising diffusion probabilistic models*, Advances in Neural Information Processing Systems 33 (2020), 6840–6851.

[20] Hommes, C.H. *Cobwebs, chaos and bifurcations*, Annals of Operations Research 37 (1992), 97–100.

[21] Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapire, R.E., Sellie, L. *On the learnability of discrete distributions* Proc. 26th Annual ACM Symp. Theory of Computing (1994), 273–282.

[22] Kearns, M.J., Valiant, L.G. *Cryptographic limitations on learning Boolean formulae and finite automata*, In: Hanson, S.J., Remmele, W., Rivest, R.L. (eds) Machine Learning: From Theory to Applications. Lecture Notes in Computer Science 661 (1993), Springer, Berlin, Heidelberg.

[23] Kirby, S. *Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity*, IEEE Trans. Evolutionary Computation 5, 2 (2001), 102–110.

[24] Kirby, S., Griffiths, T., Smith, K. *Iterated learning and the evolution of language*, Current opinion in neurobiology 28 (2014), 108–114.

[25] Kirby, S., Hurford, J.R. *The emergence of linguistic structure: An overview of the iterated learning model*, Simulating the evolution of language, 2002, 121–147.

[26] Klingler, A., van der Eyden, M., Stengele, S., Reinhart, T., De las Cuevas, G. *Many bounded versions of undecidable problems are NP-hard*, SciPost Phys. 14, 173 (2023), 1–29.

[27] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., Ermon, S. *SDEdit: Guided image synthesis and editing with stochastic differential equations*, Proc. International Conference on Learning Representations, 2022.

[28] Murphy, K.P. Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

[29] Muth, J. F. *Rational expectations and the theory of price movements*, Econometrica 29, 3 (1961), 315–35.

[30] Nichol, A.Q., Dhariwal, P. *Improved denoising diffusion probabilistic models*, Proc. 38th International Conference on Machine Learning 139 (2021), 8162–8171.

[31] Palmer, K. Shadowing in Dynamical Systems: Theory and applications, Mathematics and its Applications 501 (2000), Kluwer Academic Publishers.

[32] Pilyugin, S. Shadowing in Dynamical Systems, Lecture Notes in Mathematics1706 (1999), Springer-Verlag, Berlin.

[33] Prince, S.J.D. Understanding Deep Learning, MIT Press, 2023.

[34] Rafferty, A.N., Griffiths, T.L., Klein, D. *Convergence bounds for language evolution by iterated learning*, Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society, 2009.

[35] Rafferty, A.N., Griffiths, T.L., Klein, D. *Analyzing the rate at which languages lose the influence of a common ancestor*, Cognitive Science 38, 7 (2014), 1406–1431.

[36] Ross J.L. *The dark matter of biology*, Biophys J. 111, 5 (2016) , 909–16.

[37] Särkkä, S. Bayesian Filtering and Smoothing, Cambridge University Press, 2013.

[38] Schapire, R.E. *The strength of weak learnability*, Machine Learning 5 (1990), 197–227.

[39] Seneta, E. Non-Negative Matrices and Markov Chains, Springer, 2nd ed., 2006.

[40] Smith, K., Kirby, S., Brighton, H. *Iterated learning: A framework for the emergence of language*, Artificial Life 9 (2003), 371–386.

[41] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S. *Deep unsupervised learning using nonequilibrium thermodynamics*, Proc. 32nd International Conference on Machine Learning (2015), 2256–2265.

[42] Szostak, J.W., Bartel, D.P., Luisi, P.L. *Synthesizing life*, Nature 409, 6818 (2001), 387–390.

[43] von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S. Wolf, T., *Diffusers: State-of-the-art diffusion models*, GitHub repository, 2022.