2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Heat Death of Generative Models in Closed-Loop Learning

Matteo Marchi, Stefano Soatto, Pratik Chaudhari, Paulo Tabuada

Abstract— **Improvement and adoption of generative machine learning models is rapidly accelerating, as exemplified by the popularity of LLMs (Large Language Models) for text, and diffusion models for image generation. As generative models become widespread, data they generate is incorporated into shared content through the public web. This opens the question of what happens when data generated by a model is fed back to the model in subsequent training campaigns. This is a question about the stability of the training process, whether the distribution of publicly accessible content, which we refer to as "knowledge", remains stable or collapses.**

Small scale empirical experiments reported in the literature show that this closed-loop training process is prone to degenerating. Models may start producing gibberish data, or sample from only a small subset of the desired data distribution (a phenomenon referred to as mode collapse). So far there has been only limited theoretical understanding of this process, in part due to the complexity of the deep networks underlying these generative models.

The aim of this paper is to provide insights into this process (that we refer to as "generative closed-loop learning") by studying the learning dynamics of generative models that are fed back their own produced content in addition to their original training dataset. The sampling of many of these models can be controlled via a "temperature" parameter. Using dynamical systems tools, we show that, unless a sufficient amount of external data is introduced at each iteration, any non-trivial temperature leads the model to asymptotically degenerate. In fact, either the generative distribution collapses to a small set of outputs, or becomes uniform over a large set of outputs.

I. Introduction

Generative models have exploded in popularity in recent years, primarily driven by the adoption of diffusion models [5] for image generation, and so-called LLMs (Large Language Models) [15] for textual generation. With this explosion, came renewed concerns about AI, especially tied to the *generative* nature of these models. Large scale neural networks underlie most of these models, including, for example, Llama 2 which is trained on 2 trillion tokens [13]. As these models generate data that is published on the internet, they pollute their own training datasets with synthetic data, possibly leading to a spiraling decay of the quality of these models and of the internet by extension.

We are concerned with the setting where a generative model is iteratively trained, and the outcome of each iteration is dependent on the current data distribution encoded by the model (typically by including samples generated by the model in the training set). Serious concerns about decay of such a training process arose first in GANs (Generative Adversarial Networks), where the problem of "mode collapse" [12] was identified. Analogous issues seem to be a general feature of violating distributional assumptions about the training dataset, even for non-GAN models. One way of framing such violations is as "data poisoning" [3], a problem that is likely to become more common, as models trained from public domain internet data are especially susceptible to data poisoning attacks [6]. This is also related to the notion of "distribution shift" [7], although most existing work focuses on the distribution shift occurring at test time and coming from an external source. In our setting the shift occurs at *training time* and has an *internal* origin.

As this is such a new development, there is still only partial understanding of the phenomenon, and much published work is empirical in nature. In [9] and [10], the authors train image diffusion models, iteratively including synthetic samples, and show significant degradation of the quality of the produced images. In [11], it is shown, both theoretically and experimentally, that generative Gaussian models undergo degenerative collapse. A case of closed-loop learning when the sampling of the model is biased (samples may be taken closer to the mean) under a variety of synthetic data policies is studied in [1]. In their results, non-degeneration could be ensured only by introducing a sufficient fraction of fresh data at each training iteration. This aligns with the results in [2], where the authors establish (theoretically and experimentally) that maintaining a high enough fraction of fresh data is a sufficient condition to prevent degeneration.

Most generative models include a way to modulate their sampling probabilities through "temperature", typically as a way to make the outputs more or less random. In this work, we focus on the effect of temperature on the closed-loop learning dynamics of generative models, a perspective that received little attention so far. In particular: 1) We define a class of "generative closed-loop learning models with temperature" that captures many real-world scenarios. 2) We perform a theoretical analysis of the resulting closed-loop learning dynamics, and establish that modulating sampling with temperature leads to degeneration of the learning process. 3) We characterize the type of degeneration depending on one of three possible temperature regimes. As the models degenerate (for any amount of temperature modulation), so do their datasets, consequently losing any knowledge they originally contained, if not explicitly preserved and re-introduced. When applied to the internet, this predicts that unless a copy of the pre-generative-models Internet is preserved, eventually no model will be able to be trained effectively using the internet as a data source. Our results share some similarities with [1], [2], and are compatible with their conclusions, but in contrast to those papers we use tools and techniques from dynamical and control systems for the analysis.

979-8-3503-1632-2/24/$31.00 ©2024 IEEE

II. Notation

- We denote by e_i the i-th element of the standard basis of $\mathbb{R}^n$, i.e., the vector of all zeroes except for a one in its i-th entry.
- The symbol $\mathbf{1}$ denotes the vector $x \in \mathbb{R}^n$ with all elements equal to one.
- We define Δ^n as the n-dimensional probability simplex $\Delta^n = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1, x_i \geq 0\}$, and its restriction to strictly positive probabilities as $\Delta^n_{>0}$. An element of Δ^n is called a "probability vector". The boundary of Δ^n is denoted by $\partial\Delta^n$.
- Given some $X \in \Delta^n$, we say that the random variable Y is sampled according to X, or $Y \sim X$, to mean that for all $i \in \{1, 2, \ldots, n\}$ we have $\mathrm{P}(Y = e_i) = X_i$.
- If $X(k)$ is a stochastic process, $\mathcal{F}_k$ denotes the filtration adapted to the stochastic process up to time k. We say an event happens a.s. to mean "almost surely", i.e., with probability 1 (w.p. 1).
- Unless otherwise noted, $\|\cdot\|$ denotes the usual vector 2-norm over $\mathbb{R}^n$, and $d(x, y) = \|x - y\|$ with $x, y \in \mathbb{R}^n$ is the distance between x and y. If one of the arguments is a set $\Omega \subseteq \mathbb{R}^n$, it denotes the distance from a point to that set $d(x, \Omega) = \inf_{y \in \Omega} d(x, y)$.
- The notation $f(x) \xrightarrow[x \to a]{} \Omega$, with Ω a set means $\lim_{x \to a} d(f(x), \Omega) = 0$.
- We normally use capitalized letters to denote random variables, and lower-case when they are deterministic, or when the randomness is not relevant (i.e., X vs x).
- A continuous function $\alpha : [0, a) \to \mathbb{R}_{\geq 0}$, with $a \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$, is said to be of class kappa ($\alpha \in \mathcal{K}$) if it is strictly increasing and $\alpha(0) = 0$.

III. Generative Closed-Loop Learning

We describe a generative model as a parameterized family of probability distributions over a finite set of $n \in \mathbb{N}$ possible elements $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_n\}$[1]. We denote such family as $\phi : \mathbb{R}^p \to \Delta^n$, a map from a parameter vector $w \in \mathbb{R}^p$ to a probability vector $\phi(w) \in \Delta^n$ for the elements $\mathcal{Y}$. These are the outputs that the model can generate when sampled. Without loss of generality, we identify each $\mathcal{Y}_i$ with the i-th vector of the standard basis of $\mathbb{R}^n$. These elements can be interpreted differently depending on the specific generative model, e.g., for a language model each $\mathcal{Y}_i$ could be a word, token, sentence, or sentence class from a large but finite set.

A. Model sampling with temperature control

For a trained generative model, letting Y be the output of the model when sampled, we denote by $\Theta = \phi(w)$ the "nominal" probability of generating each of the possible elements of $\mathcal{Y}$. Specifically, the probability of generating the i-th element corresponds to the i-th entry of the vector Θ. However, when sampled, the actual generation probabilities are filtered through a *temperature* function $\tau : \Delta^n \to \Delta^n$. Therefore, for $i \in \{1, 2, \ldots, n\}$, the sampled output $Y \in \mathcal{Y}$ satisfies:

$$P(Y = \mathcal{Y}_i) = \tau(\Theta)_i, \tag{1}$$

where a subscripted index i denotes the i-th vector element. For our results to hold, we require the temperature function to satisfy some assumptions (see Sec. IV). We show that these assumptions hold for the temperature function induced by the *softmax* operation, typically used in deep learning.

B. Learning process

We use the term "generative closed-loop learning", or just "closed-loop learning", to refer to a generative model trained on data that includes its own output from prior runs. When a generative model learns from its own output, the probability vector Θ becomes a stochastic process $\Theta(k)$ evolving over (discrete) time $k \in \mathbb{Z}_{\geq 0}$. We assume that a model is initially trained on some externally provided dataset of some size $\ell \in \mathbb{N}$:

$$D(\ell) = \{Y(1), Y(2), \ldots, Y(\ell)\},$$

where we use $Y(k)$ with $k \leq \ell$ to denote the externally provided data, i.e., training only starts at time $k = \ell$. Similarly, for each time $k \geq \ell$ we have a parameter vector $w(k)$ and its associated probability vector $\Theta(k) = \phi(w(k))$. Finally, let the training be represented by a (in general stochastic) function[2] f that maps a parameter vector w and training data D to a "retrained" parameter $f(w, D)$. Then, for each time $k \in \mathbb{Z}_{\geq 0}$, $k \geq \ell$, the closed-loop learning stochastic process unfolds as follows:

$$\begin{aligned} Y(k+1) &\sim \tau(\Theta(k)) \\ D(k+1) &= D(k) \cup \{Y(k+1)\} \\ w(k+1) &= f(w(k), D(k+1)) \\ \Theta(k+1) &= \phi(w(k+1)), \end{aligned} \tag{2}$$

where $D(k) \cup \{Y(k+1)\}$ models "adding" the generated output sample to the current set of training data[3]. For some initial dataset of ℓ samples, the process has the initial conditions $w(\ell) = f(w_0, D(k))$, and $\Theta(\ell) = \phi(w(\ell))$ for some $w_0 \in \mathbb{R}^p$. The recursive process (2) induces a probability distribution for each $\Theta(k)$. Like for the temperature function τ, in Sec. IV we will require that the process (2) satisfies some general properties.

C. Problem statement

Given the closed-loop learning process (2), we want to know what are the long term properties of the probability vector $\Theta(k)$, i.e., what is the asymptotic behavior of $\Theta(k)$ as time increases? Since $\Theta(k)$ describes the probability of generated data, the asymptotic behavior of $\Theta(k)$ determines the ultimate composition of the dataset $D(k)$ as well. For example, if $\Theta(k)$ were to converge to a point independent

[1] This family is a subset of the set of categorical distributions over n categories. In general we do not require that the family of distributions expressible by the model is the full set of categorical distributions, which is unrealistic for very high n (for example, if the outcomes $\mathcal{Y}$ are rgb-images).

[2] While the retraining function here takes the current parameters as an argument, it can also represent a form of retraining where the model is "reset" and trained from scratch over a new dataset by ignoring w.

[3] For notational simplicity, in (2), the process retrains the model after each generated sample, however our results hold even in the case where some variable but bounded number of samples $N \geq 1$ is generated and added to the dataset before retraining.

of the initial dataset $D(\ell)$, any initial knowledge encoded by the dataset is eventually lost.

IV. A COMMON CLASS OF MODELS

In the previous section we presented an abstracted notion of closed-loop learning. We now give specific conditions on the temperature function τ and the behavior of the training algorithm represented by f and ϕ in (2), and show that they are realistic for common closed-loop learning models.

A. Temperature

We assume that the class of temperature functions τ as defined in Sec. III satisfies a few properties:

Assumption 1: The temperature function $\tau : \Delta^n \to \Delta^n$ in (2) is assumed to satisfy the following properties:

1) It is continuous and strictly element-wise order preserving, i.e., for any $\theta \in \Delta^n$ and $i, j \in \{1, 2, \dots, n\}$:
$$\begin{aligned} \theta_i < \theta_j &\implies \tau(\theta_i) < \tau(\theta)_j \\ \theta_i = \theta_j &\implies \tau(\theta_i) = \tau(\theta_j). \end{aligned} \tag{3}$$
2) Given an index set $I \subseteq \{1, 2, \dots, n\}$, let $V_I : \Delta^n \to \mathbb{R}_{\geq 0}$ be defined as[4]:
$$V_I(\theta) = \max_{i \in I} \left\{ \frac{\theta_i}{\sum_{j \in I} \theta_j} \right\} - \min_{i \in I} \left\{ \frac{\theta_i}{\sum_{j \in I} \theta_j} \right\}. \tag{4}$$
Then, τ satisfies exactly one of the properties:
   a) It is the identity: $\tau(\theta) = \theta$.
   b) For any index set $I \subseteq \{1, 2, \dots, n\}$ where $\min_{i \in I} \theta_i > 0$ and $\max_{i \in I} \theta_i > \min_{i \in I} \theta_i$:
   $$V_I(\tau(\theta)) - V_I(\theta) < 0.$$
   c) For any index set $I \subseteq \{1, 2, \dots, n\}$ where $\min_{i \in I} \theta_i > 0$ and $\max_{i \in I} \theta_i > \min_{i \in I} \theta_i$:
   $$V_I(\tau(\theta)) - V_I(\theta) > 0.$$

Intuitively, case 2.b represents a "contracting" τ, and case 2.c an "expanding" τ. The function V_I quantifies how close a probability vector $\theta \in \Delta^n$ is to uniform when conditioned to a specific subset of variables. Note that when an index set I includes all non-zero elements of θ, the renormalizing term $\sum_{j \in I} \theta_j$ in (4) is equal to one, and (4) reduces to $V_I(\theta) = \max_{i \in I} \theta_i - \min_{i \in I} \theta_i$.

The notion of temperature typically used in generative models satisfies the requirements listed above. In fact, this is the case for the *softmax* temperature, as we now show. Many machine learning models do not directly output a set of probabilities, but a vector of so-called *logits* $z \in \mathbb{R}^n$ that is converted into a probability vector via the $\mathrm{softmax}$ function and a positive temperature parameter $T > 0$ as follows:

$$\begin{aligned} \theta_T &= \mathrm{softmax}\left(zT^{-1}\right) \\ &= \frac{1}{\sum_{i=1}^n \exp\left(\frac{z_i}{T}\right)} \begin{bmatrix} \exp\left(\frac{z_1}{T}\right) & \dots & \exp\left(\frac{z_n}{T}\right) \end{bmatrix}^\top. \end{aligned} \tag{5}$$

Note that τ, as defined in Sec. III, is a map between probability vectors, but (5) maps logits to probabilities. Consider a logit vector $z \in \mathbb{R}^n$. While (5) is not a valid τ, it induces a unique map transforming $\theta = \mathrm{softmax}(z)$ (the "nominal" probabilities associated to z) to $\theta_T = \mathrm{softmax}\left(T^{-1}z\right)$ (the "temperature filtered" probabilities associated to the same z). Defining $Z = \sum_{i=1}^n \exp(z_i)$ we have:

$$\begin{aligned} \theta_T &= \mathrm{softmax}\left(zT^{-1}\right) \\ &= \mathrm{softmax}\left(T^{-1}\left(\begin{bmatrix} \ln(\theta_1) & \dots & \ln(\theta_n) \end{bmatrix}^\top + \mathbf{1}\ln(Z)\right)\right) \\ &= \mathrm{softmax}\left(T^{-1}\begin{bmatrix} \ln(\theta_1) & \dots & \ln(\theta_n) \end{bmatrix}^\top\right) \\ &= \frac{1}{\sum_{i=1}^n \theta^{\frac{1}{T}}} \begin{bmatrix} \theta_1^{\frac{1}{T}} & \dots & \theta_n^{\frac{1}{T}} \end{bmatrix}^\top, \end{aligned} \tag{6}$$

where the third equality holds because $\mathrm{softmax}$ is invariant to addition of the same constant ($T^{-1}\ln(Z)$) to all input entries. This induced map $\tau(\theta) = \theta_T$ satisfies exactly one of the three requirements previously introduced. We formalize this in the following Lemma:

Lemma 1: Consider the function $\tau : \Delta^n \to \Delta^n$ defined by $\tau(\theta) = \mathrm{softmax}\left(T^{-1}\ln(\theta)\right)$, with $T \in \mathbb{R}_{>0}$. The function τ satisfies Assumption 1. In particular, it satisfies properties 2.a, 2.b, and 2.c for $T = 1$, $T > 1$, and $T < 1$ respectively.

Proof: Since τ takes the form (6), and $x \mapsto x^{\frac{1}{T}}$ is a continuous strictly monotone increasing function, it is immediate that τ is also continuous (the denominator in (6) is always bounded away from zero) and preserves the order of the elements of θ. Further, note that if $\theta_i = 0$, then $\tau(\theta)_i = 0$.

For the case $T = 1$, it is immediate to see that τ becomes the identity function (Note that $\sum_{i=1}^n \theta_i = 1$).

For the cases where $T \neq 1$, let $I \subseteq \{1, 2, \dots, n\}$ be as in Assumption 1, then, $V_I(\tau(\theta)) < V_I(\theta)$. To see that this is true, let $M, m \in I$ be respectively the (not necessarily unique) indices of the greatest and smallest non-zero elements of θ, and consider the derivative with respect to the temperature parameter T of $V_I(\tau(\theta))$:

$$\begin{aligned} \frac{\partial}{\partial T} V_I(\tau(\theta)) &= \frac{\partial}{\partial T}\left\{ \frac{\theta_M^{\frac{1}{T}}}{\sum_{j \in I} \theta_j^{\frac{1}{T}}} - \frac{\theta_m^{\frac{1}{T}}}{\sum_{j \in I} \theta_j^{\frac{1}{T}}} \right\} \\ &= -T^{-2}\left[\frac{\theta_M^{\frac{1}{T}}}{\sum_{j \in I} \theta_j^{\frac{1}{T}}} \sum_{i \in I} \left(\frac{\theta_i^{\frac{1}{T}}}{\sum_{j \in I} \theta_j^{\frac{1}{T}}} \left(\log(\theta_M) - \log(\theta_i)\right) \right) \right. \\ &\qquad \left. - \frac{\theta_m^{\frac{1}{T}}}{\sum_{j \in I} \theta_j^{\frac{1}{T}}} \sum_{i \in I} \left(\frac{\theta_i^{\frac{1}{T}}}{\sum_{j \in I} \theta_j^{\frac{1}{T}}} \left(\log(\theta_m) - \log(\theta_i)\right) \right) \right]. \end{aligned}$$

The sums are convex combinations of non-negative terms for the first and non-positive for the second summation, as $\log(\theta_M) \geq \log(\theta_i)$ and $\log(\theta_m) \leq \log(\theta_i)$ for all $i \in I$. Further, by Assumption 1 $\max_{i \in I} \theta_i > \min_{i \in I} \theta_i$, therefore the sums are strictly positive and strictly negative respectively, and $\frac{\partial}{\partial T} V_I(\tau(\theta)) < 0$. Then, fixing some specific temperature $\overline{T}$:

$$V_I(\tau(\theta)) - V_I(\theta) = \int_{T=1}^{\overline{T}} \frac{\partial}{\partial T} V_I(\tau(\theta))\, \mathrm{d}T, \tag{7}$$

where (7) is negative for $\overline{T} > 1$, and positive for $\overline{T} < 1$. ∎

[4] V_I will be used as a Lyapunov function later in the analysis.

B. Closed-loop learning

For a training dataset $D(\ell) = \{Y(1), Y(2), \ldots, Y(\ell)\}$, "learning" a generative model is usually framed in a maximum-likelihood sense, i.e., we want to find the set of parameters $w \in \mathbb{R}^p$ whose associated probability vector $\Theta = \phi(w)$ maximizes the log-probability of the observed data:

$$\begin{aligned} w^* &= \arg\max_{w\in\mathbb{R}^p} \ \frac{1}{\ell}\sum_{i=1}^{\ell}\sum_{j=1}^{n} \mathbb{1}_{Y(i)=\mathcal{Y}_j} \log\left(\phi(w)_j\right) \\ &= \arg\min_{w\in\mathbb{R}^p} \ -\sum_{j=1}^{n}\left(\frac{1}{\ell}\sum_{i=1}^{\ell}\mathbb{1}_{Y(i)=\mathcal{Y}_j}\right)\log\left(\phi(w)_j\right) \\ &= \arg\min_{w\in\mathbb{R}^p} \ -\sum_{j=1}^{n}\Theta^*_j \log\left(\phi(w)_j\right) \\ &= \arg\min_{w\in\mathbb{R}^p} \ H\left(\Theta^*, \phi(w)\right), \end{aligned} \tag{8}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function (takes value 1 if the subscript expression is true and 0 otherwise), H is the cross-entropy between probability vectors, and $\Theta^* = \frac{1}{\ell}\sum_{i=1}^{\ell} Y(i)$ is the "empirical" probability vector associated to the data D (remember that with no loss of generality we take $\mathcal{Y}_j$ to be the j-th standard basis element of $\mathbb{R}^n$). Therefore, although usually not expressed in this way, learning the model is also equivalent to minimizing the cross entropy with respect to Θ^*. Note that if ϕ is surjective (where its codomain is Δ^n), there always exists a w^* such that $\phi(w^*) = \Theta^*$ and is thus the optimal solution of (8).

When ϕ is surjective, $\Theta = \Theta^*$, thus we study the behavior of Θ^* under the closed-loop learning dynamics. Consider the process (2), and let $\Theta^*(k) = \frac{1}{k}\sum_{i=1}^{k} Y(i)$ be the empirical probability vector corresponding to the dataset at time k, $D(k)$. When a new sample $Y(k+1)$ is generated by the model, Θ^* evolves as:

$$\begin{aligned} \Theta^*(k+1) &= \frac{1}{k+1}\sum_{i=1}^{k+1} Y(i) \\ &= \frac{k}{k+1}\Theta^*(k) + \frac{1}{k+1}Y(k+1) \\ &= \Theta^*(k) + \frac{1}{k+1}\left(Y(k+1) - \Theta^*(k)\right). \end{aligned} \tag{9}$$

Since $Y(k+1)$ is a random variable, we perform a Martingale decomposition [8]:

$$\begin{aligned} \Theta^*(k+1) &= \Theta^*(k) + \frac{1}{k+1}\big(\mathbb{E}\left[Y(k+1)|\mathcal{F}_k\right] - \Theta^*(k) \\ &\quad + Y(k+1) - \mathbb{E}\left[Y(k+1)|\mathcal{F}_k\right]\big) \\ &= \Theta^*(k) + \frac{1}{k+1}\big(\tau(\Theta(k)) - \Theta^*(k) + U(k+1)\big), \end{aligned}$$

where $U(k+1) = Y(k+1) - \mathbb{E}\left[Y(k+1)|\mathcal{F}_k\right]$ is a bounded Martingale difference sequence, and the second equality holds because $\mathbb{E}\left[Y(k+1)|\mathcal{F}_k\right] = \tau(\Theta)$.

Then, if the update function f in (2) is the maximum-likelihood optimization (8):

$$\begin{aligned} w(k+1) &= f(w(k), D(k+1)) \\ &= \arg\max_{w\in\mathbb{R}^p} H\left(\Theta^*(k+1), \phi(w)\right), \end{aligned} \tag{10}$$

and ϕ is surjective on Δ^n, at each step $\Theta(k) = \phi(w(k)) = \Theta^*(k)$, leading to the following dynamics:

$$\Theta(k+1) = \Theta(k) + \frac{1}{k+1}\left(\tau(\Theta(k)) - \Theta(k) + U(k+1)\right).$$

An actual model is unlikely to have enough expressivity as to represent *any* element of Δ^n, especially for very high dimension n. However, we assume the model is able to approximate a probability vector with some small finite accuracy:

Assumption 2: There exists some $\delta \in \mathbb{R}_{\geq 0}$ such that for all $k \in \mathbb{Z}_{\geq 0}$ the process (2) satisfies:

$$\|\Theta^*(k) - \Theta(k)\| = \|\Theta^*(k) - \phi(w^*(k))\| \leq \delta. \tag{11}$$

Under Assumption 2, the dynamics of $\Theta^*(k)$ obey:

$$\Theta^*(k+1) = \Theta^*(k) + \frac{1}{k+1}\left(\tau(\Theta(k)) - \Theta^*(k) + U(k+1)\right). \tag{12}$$

If we now define the perturbation $\varepsilon(k) = \tau(\Theta) - \tau(\Theta^*)$, recalling that a continuous function on a compact set is uniformly continuous, we have the following inequality, where η is the modulus of continuity of τ:

$$\|\varepsilon(k)\| = \|\tau(\Theta) - \tau(\Theta^*)\| \leq \eta(\delta). \tag{13}$$

Then, the dynamics of $\Theta^*(k)$ become a stochastic approximation (see [4]) of the form:

$$\begin{aligned} \Theta^*(k+1) = \Theta^*(k) + \frac{1}{k+1}(\tau\left(\Theta^*(k)\right) - \Theta^*(k) + \\ + \varepsilon(k) + U(k+1)). \end{aligned} \tag{14}$$

If we understand the behavior of (14), we automatically understand the behaviour of $\Theta(k)$, since by (11) $\Theta(k)$ is always within a distance δ from $\Theta^*(k)$.

V. Main Results

We now present the main results describing the asymptotic behavior of (14) (and hence of $\Theta(k)$ up to error δ). This asymptotic behavior is important, as it determines the long-term composition of the dataset $D(k)$, and how much it may diverge from its initial distribution. The results are different depending on which of the three conditions (2.a, 2.b, 2.c) in Assumption 1 is satisfied by τ, therefore we split the analysis in three different cases.

A. Identity temperature leads to Martingale-like behavior

This case is the most straightforward and does not require any machinery beyond standard analysis tools of stochastic processes. If condition 2.a of Assumption 1 is satisfied, the stochastic process (14) reduces to:

$$\Theta^*(k+1) = \Theta^*(k) + \frac{1}{k+1}\left(U(k+1) + \varepsilon(k)\right), \tag{15}$$

and we state the following formal result.

Theorem 1: Consider the closed-loop learning stochastic process (2), where τ satisfies Assumption 1 with property 2.a (τ is the identity), and Assumption 2. Then it holds that:

$$\mathbb{E}\left[\Theta(k+1) \mid \mathcal{F}_\ell\right] = \Theta(\ell) + \sum_{i=\ell}^{k} \frac{1}{i+1} \mathbb{E}\left[\varepsilon(i) \mid \mathcal{F}_\ell\right] + e_\delta(k+1),$$

where $e_\delta : \mathbb{Z}_{\geq 0} \to \mathbb{R}^n$ is such that $\|e_\delta(k)\| \leq \delta$. In addition, if ε is a Martingale difference sequence, there is a constant $b \in \mathbb{R}_{\geq 0}$ such that the asymptotic variance is bounded as:

$$\lim_{k\to\infty} \operatorname{var}\left(\Theta(k) - \Theta(\ell)\right) \leq b \left(\sum_{i=\ell}^{\infty} \left(\frac{1}{i+1} \right)^2 + \delta^2 \right). \quad (16)$$

Proof: If τ is the identity function, $\Theta^* : \mathbb{Z}_{\geq 0} \to \Delta^n$ satisfies (15). Then, because U is a Martingale difference sequence:

$$\begin{aligned} &\mathbb{E}\left[\Theta^*(k+1) \mid \mathcal{F}_k\right] = \\ &\quad = \mathbb{E}\left[\Theta^*(k) + \frac{1}{k+1} \left(U(k+1) + \varepsilon(k) \right) \mid \mathcal{F}_k \right] \\ &\quad = \Theta^*(k) + \frac{1}{k+1} \varepsilon(k), \end{aligned}$$

and by the tower property of expectation:

$$\begin{aligned} &\mathbb{E}\left[\Theta^*(k+1) \mid \mathcal{F}_{k-1}\right] = \\ &\quad = \mathbb{E}\left[\mathbb{E}\left[\Theta^*(k+1) \mid \mathcal{F}_k\right] \mid \mathcal{F}_{k-1}\right] \\ &\quad = \mathbb{E}\left[\Theta^*(k) + \frac{1}{k+1} \varepsilon(k) \mid \mathcal{F}_{k-1} \right] \\ &\quad = \Theta^*(k-1) + \frac{1}{k} \varepsilon(k-1) + \frac{1}{k+1} \mathbb{E}\left[\varepsilon(k) \mid \mathcal{F}_{k-1}\right]. \end{aligned}$$

Finally, by recursion we arrive at:

$$\mathbb{E}\left[\Theta^*(k+1) \mid \mathcal{F}_\ell\right] = \Theta^*(\ell) + \sum_{i=\ell}^{k} \frac{1}{i+1} \mathbb{E}\left[\varepsilon(i) \mid \mathcal{F}_\ell\right],$$

and the statement is obtained once we take into account that $\Theta(k)$ is always within a distance δ from $\Theta^*(k)$.

In addition, if ε is a Martingale difference sequence, the whole $\Theta^*(k)$ process reduces to a sum of bounded Martingale differences, and it immediately follows that its variance is bounded by a term of the order of the converging sum $\sum_{i=\ell}^{\infty} (i+1)^{-2}$. ∎

Theorem 1 states that in this case $\Theta(k)$ is essentially a Martingale biased by the perturbation ε. In general the asymptotic behavior can be arbitrary as it is dominated by the behavior of $\varepsilon(k)$. However, if ε is also a Martingale difference sequence, with probability one $\Theta(k)$ will only drift a finite amount from its initial value. The magnitude of this drift depends on the converging sum $\sum_{i=\ell}^{\infty} (i+1)^{-2}$, which is smaller the greater the initial dataset size ℓ is. This shows that in this case the initial data distribution of the dataset is not necessarily lost. However, this condition requires hard to verify assumptions on the training behavior (captured by $\varepsilon(k)$) and, if the training dataset is shared by multiple generative models, that no model is biasing their own sampling via temperature. We consider this especially unlikely for data on the public web. From a control perspective, the behavior with identity temperature is similar to that of a marginally stable system, and any arbitrarily small perturbation $\varepsilon(k)$ can destabilize it.

B. High temperature leads to uniformly generated data

While the analysis of the previous case is relatively straightforward, we need to introduce additional machinery for the remaining two. Let us define a vector field $F : \Delta^n \to T\Delta^n$ over the probability simplex as $F(\theta) = \tau(\theta) - \theta$. The behavior of stochastic approximations in the long-term approaches that of a continuous-time ODE (ordinary differential equation) or differential inclusion (see [4]). Thus, under our assumptions, the limit sets of $\Theta^*(k)$ in (14) are determined by the attractors of:

$$\dot{\theta}(t) = F(\theta(t)) + \varepsilon(t). \quad (17)$$

Then, we introduce two families of sets, parameterized by $a \in \mathbb{R}_{\geq 0}$ and an index set $I \subseteq \{1, 2, \ldots, n\}$, that will be used to characterize the attractors and basins of attraction of (17):

$$\begin{aligned} \underline{\Omega}_I(a) &= \left\{ \theta \in \Delta^n \;\middle|\; V_I(\theta) \leq a \right\} \\ \overline{\Omega}_I(a) &= \left\{ \theta \in \Delta^n \mid \min_{i \in I} \theta_i \leq a \right\}. \end{aligned} \quad (18)$$

With respect to the subset of variables indexed by I, the set $\underline{\Omega}_I(a)$ is a compact neighborhood of the uniform probability vector (all θ_i are equal), while $\overline{\Omega}_I(a)$ is a compact neighborhood of the boundary of the probability simplex (at least one θ_i is zero). We can now state the following lemma:

Lemma 2: Consider the continuous-time ODE:

$$\dot{\theta}(t) = F(\theta(t)) + \varepsilon(t), \quad (19)$$

where $\theta \in \Delta^n$, $\|\varepsilon\| \leq \eta \in \mathbb{R}_{\geq 0}$, and $F(\theta) = \tau(\theta) - \theta$. Let $\tau : \Delta^n \to \Delta^n$ satisfy Assumption 1 and property 2.b (contractive τ). Then, for any index set $I \subseteq \{1, 2, \ldots, n\}$ there exists $\kappa \in \mathcal{K}$ such that for any $t_0 \in \mathbb{R}$:

$$\theta(t_0) \notin \overline{\Omega}_I(\kappa(\eta)) \quad (20)$$

implies:

$$\theta(t) \xrightarrow[t\to\infty]{} \underline{\Omega}_I(\kappa(\eta)). \quad (21)$$

Proof: Consider a point $\theta \in \Delta^n$, and an index set $I \subseteq \{1, 2, \ldots, n\}$ where $\min_{i \in I} \theta_i > 0$ (note that the lemma only makes a claim for index sets such that $\theta(t_0) \notin \overline{\Omega}_I(\kappa(\eta))$, which satisfy this condition). Let $M, m \in I$ be the (not necessarily unique) indices of the greatest and smallest elements of $\{\theta_i \mid i \in I\}$. The time derivative of V_I at some

point θ is:

$$\begin{aligned}\dot{V}_I(\theta) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{\left(\sum_{i\in I}\theta_i\right)^{-1}\left(\max_{i\in I}\theta_i - \min_{i\in I}\theta_i\right)\right\}\\ &= -\left(\sum_{i\in I}\theta_i\right)^{-2}\left(\sum_{i\in I}\dot{\theta}_i\right)(\theta_M-\theta_m)\\ &\qquad + \left(\sum_{i\in I}\theta_i\right)^{-1}\left(\dot{\theta}_M - \dot{\theta}_m\right)\\ &\leq \frac{\sum_{i\in I}\tau(\theta)_i}{\sum_{i\in I}\theta_i}\left(\frac{\tau(\theta)_M-\tau(\theta)_m}{\sum_{i\in I}\tau(\theta)_i} - \frac{\theta_M-\theta_m}{\sum_{i\in I}\theta_i}\right) + b\eta\\ &\leq c\left(V_I(\tau(\theta)) - V_I(\theta)\right) + b\eta,\end{aligned}$$

where the second equality holds because τ preserves the indices of the maximum and minimum elements of θ by its order preserving property, guaranteeing that the derivative of V_I is well-defined. In the last two inequalities $b, c \in \mathbb{R}_{>0}$ are finite positive constants, as the sums that appear are always positive and bounded from above and below.

We now seek to establish that for η small enough, there is some $a \in \mathbb{R}_{\geq 0}$ such that the set difference $\mathcal{B}(a) = \Delta^n \setminus (\underline{\Omega}_I(a) \cup \overline{\Omega}_I(a))$ is a basin of attraction for $\underline{\Omega}_I(a)$. Remember that by property 2.b in Assumption 1, $V_I(\tau(\theta)) - V_I(\theta) < 0$ over points that lie in $\mathcal{B}(0)$ (i.e., points that are not the uniform probability vector over I and with no zero elements). Consider the closure of the set $\mathcal{B}(a)$ for some $a > 0$, and let us define the "worst-case" decrease of V_I for points in $\mathrm{clo}(\mathcal{B}(a))$ as:

$$\beta(a) = \sup_{\theta\in\mathrm{clo}(\mathcal{B}(a))}\left\{V_I(\tau(\theta)) - V_I(\theta)\right\}. \tag{22}$$

By continuity, and because for $a > 0$, the set $\mathcal{B}(a)$ always excludes an open neighborhood of the region where $V_I(\tau(\theta)) - V_I(\theta) = 0$, we have that $\beta(a) < 0$. Then, for any $\mathcal{B}(a) \neq \varnothing$, as long as $\eta < -cb^{-1}\beta(a)$, the term $\dot{V}_I(\theta)$ is strictly negative for all $\theta \in \mathcal{B}(a)$.

Observe that $\beta(\cdot)$ is decreasing, since $\mathcal{B}(a_2) \subseteq \mathcal{B}(a_1)$ for $a_2 > a_1$, and $\beta(0) = 0$. Then, we can define $\kappa \in \mathcal{K}$ as any strictly increasing continuous function such that:

$$\kappa(\eta) > \inf\left\{a \in \mathbb{R}_{\geq 0} \;\middle|\; \eta < -cb^{-1}\beta(a)\right\}, \tag{23}$$

for any η where this $\inf$ is finite, which is guaranteed for any η smaller than some finite threshold. Then, any initial condition in the basin $\theta(t_0) \in \mathcal{B}(\kappa(\eta))$ guarantees that $\dot{V}_I(\theta) < 0$, and $\theta(t)$ will converge to $\underline{\Omega}(\kappa(\eta))$.

■

This lemma essentially claims that the solutions of the limiting ODE (17) over a subset of indices whose elements are sufficiently away from zero converge to a neighborhood of the uniform probability vector conditioned over that subset of indices. This key result allows us to then claim the following theorem:

Theorem 2: Consider the closed-loop learning stochastic process (2), where τ satisfies Assumption 1 with property 2.b (τ is contractive), and Assumption 2. Then, for any index set $I \subseteq \{1, 2, \ldots, n\}$ there exists $\sigma \in \mathcal{K}$ such that:

$$\Theta(k) \xrightarrow[k\to\infty]{} \underline{\Omega}_I(\sigma(\delta)) \cup \overline{\Omega}_I(\sigma(\delta)) \qquad \text{w.p. 1.} \tag{24}$$

Further, for a given time $k_0 \in \mathbb{Z}_{\geq 0}$:

$$\mathrm{P}\left(\Theta(k) \xrightarrow[k\to\infty]{} \underline{\Omega}_I(\sigma(\delta)) \;\middle|\; \Theta(k_0) \notin \overline{\Omega}_I(\sigma(\delta))\right) \xrightarrow[k_0\to\infty]{} 1.$$

Proof: Under these conditions, $\Theta^*(k)$ satisfies the stochastic approximation (12). Then, by Corollary 4 (Chapter 5) of [4], the iterates of $\Theta^*(k)$ converge a.s. to a closed connected internally chain transitive invariant set of the continuous-time differential inclusion:

$$\dot{\theta}(t) \in \left\{x \in \Delta^n \mid \|x - F(\theta(t))\| \leq \eta(\delta)\right\}, \tag{25}$$

where $F(\theta) = \tau(\theta) - \theta$.

Lemma 2 characterizes the solutions of (25), thus, letting $a = \kappa(\eta(\delta))$, any solution that enters $\mathcal{B}(a) = \Delta^n \setminus (\underline{\Omega}_I(a) \cup \overline{\Omega}_I(a))$ will converge to $\underline{\Omega}_I(a)$. Then, the set $\mathcal{L}$ of limit points of (25) is contained in:

$$\mathcal{L} \subseteq \underline{\Omega}_I(a) \cup \overline{\Omega}_I(a), \tag{26}$$

In turn, any internally chain transitive set $\mathcal{A}$ of (25) is contained in the closure of the limit set $\mathcal{L}$:

$$\mathcal{A} \subseteq \mathrm{clo}(\mathcal{L}) \subseteq \mathrm{clo}\left(\underline{\Omega}_I(a) \cup \overline{\Omega}_I(a)\right) = \underline{\Omega}_I(a) \cup \overline{\Omega}_I(a),$$

so that w.p. 1: $\Theta^*(k) \xrightarrow[k\to\infty]{} \mathcal{A} \subseteq \underline{\Omega}_I(a) \cup \overline{\Omega}_I(a)$. Finally, because $\|\Theta(k) - \Theta^*(k)\| \leq \delta$, we know that

$$\Theta(k) \xrightarrow[k\to\infty]{} \underline{\Omega}_I(\sigma(\delta)) \cup \overline{\Omega}_I(\sigma(\delta)),$$

with $\sigma(\delta) = a + 2\delta = \kappa(\eta(\delta)) + 2\delta$.

Because any point $\theta \notin \overline{\Omega}_I(a)$ either is in $\underline{\Omega}_I(a)$, or is in $\mathcal{B}(a)$, which is an open basin of attraction for $\underline{\Omega}_I(a)$, the last result holds by [14, Theorem III.2], noting that the stochastic recursion (12) satisfies assumptions A1-3. In fact, the theorem gives explicit bounds for this probability. ■

The result essentially states that as long as the learning model is sufficiently powerful, and the training sufficiently good (low δ), every set of elements of Θ will either converge to a neighborhood of the uniform probability vector conditioned over that subset, or at least one of its elements remains trapped close to zero. The reason this second possibility can occur, is that the vector field induced by the temperature function may vanish at the boundary $\partial\Delta^n$, so that some small perturbation ε over the process can keep it trapped. However, the greater the size of the initial dataset, the higher the probability of the process converging towards the uniform probability.

Either way, regardless of the size of the initial dataset (for small δ, and iterating the result over all sets I), any information it originally contained is lost as $k \to \infty$. Some subset of output probabilities will approach zero, and the rest will approach their (conditioned) uniform distribution. In summary, in the limit, as k increases, the set of possible outputs is partitioned into the outcomes that will (almost)

never be generated, and the outcomes that will be (almost) uniformly generated.

Remark 1: While we are considering the setting where there is only a fixed amount of external initial data $D(\ell)$, our results hold even when some limited amount of external data is introduced at each training iteration.[5] To see this, let $\lambda \in [0,1]$ be the fraction of external data we introduce at each time step. Then (14) becomes:

$$\Theta^*(k+1) = \Theta^*(k) + \frac{1}{k+1}\Big(\tau\left(\Theta^*(k)\right) - \Theta^*(k) + \\ + \lambda\left(\widetilde{Y}(k) - \tau\left(\Theta^*(k)\right) - \Theta^*(k)\right) + \varepsilon(k) + U(k+1)\Big),$$

where $\widetilde{Y}(k)$ is the external data point at time k. Because $\widetilde{Y}$ is bounded, the term $\lambda(\widetilde{Y}(k) - \tau\left(\Theta^*(k)\right) - \Theta^*(k))$ is bounded and can be absorbed into $\varepsilon(k)$.

Remark 2: It may be the case that for very high dimensional outputs ($n >> 1$), the assumption that δ is sufficiently small for every output probability is unrealistic. However, the assumption may still hold over a "coarse-grained" model, where we group outputs $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_n\}$ into a set of $m < n$ categories $\left\{\widehat{\mathcal{Y}}_1, \ldots, \widehat{\mathcal{Y}}_m\right\}$. In this case the result would reduce to some categories disappearing, and others appearing uniformly randomly as $k \to \infty$.

C. Low temperature leads to mode collapse

The low temperature case is identical to the high temperature one, but with the roles of $\underline{\Omega}$ and $\overline{\Omega}$ swapped, so we only state the corresponding theorem. In the proof, the direction of the Lyapunov inequalities is swapped and the sign inverted.

Theorem 3: Consider the closed-loop learning stochastic process (2), where τ satisfies Assumption 1 with property 2.c (τ is expanding), and Assumption 2. Then, for any index set $I \subseteq \{1, 2, \ldots, n\}$ there exists $\sigma \in \mathcal{K}$ such that:

$$\Theta(k) \xrightarrow[k\to\infty]{} \underline{\Omega}_I(\sigma(\delta)) \cup \overline{\Omega}_I(\sigma(\delta)) \qquad \text{w.p. 1.} \qquad (27)$$

Further, for a given time $k_0 \in \mathbb{Z}_{\geq 0}$:

$$\mathrm{P}\left(\Theta(k) \xrightarrow[k\to\infty]{} \overline{\Omega}_I(\sigma(\delta)) \;\middle|\; \Theta(k_0) \notin \underline{\Omega}_I(\sigma(\delta))\right) \xrightarrow[k_0\to\infty]{} 1.$$

Just like for the high temperature case, any information in the original dataset is lost, with data generated by the asymptotic behavior of $\Theta(k)$ dominating the dataset. Unlike the high temperature case, with high probability Θ will converge to a region where most outputs have very low probability mass, and only a few outputs are likely to be sampled. In the limit of $\delta \to 0$, for almost every initial condition the generative probabilities of every output approach zero except for a single output element, that will completely dominate the dataset.

[5]These two scenarios are analogous to the "synthetic augmentation loop" and "fresh data loop" in [1]. In the first one, the amount of external data is fixed at the start, so that over time the proportion of synthetic data dominates the dataset. In the second one, some proportion λ of external data is introduced at each step (this may be additional copies of samples from the initial dataset), guaranteeing that the proportion of synthetic data is always less than $1 - \lambda$.

VI. Conclusion

We have shown that when a generative model is trained on the data it generates, and this generation is biased by temperature (no matter how small the biasing), there is a dichotomy between the accuracy of the learning model and preserving the initial distribution of the dataset unless that initial dataset is preserved and re-injected purposefully. A model capable of accurately reproducing the distribution of a training dataset (low δ in (11)) will inevitably degenerate into never producing some outputs and producing the rest uniformly randomly.

Our theoretical analysis adds to the increasing concern about data self-ingestion, especially in the current age where large scale deep networks are trained on data scraped from the internet, and data generated by these models inevitably finds its way back to their training processes.

References

[1] S. Alemohammad et al. Self-consuming generative models go mad. *International Conference on Learning Representations (ICLR)*, 2024.

[2] Q. Bertrand et al. On the stability of iterative retraining of generative models on their own data. *International Conference on Learning Representations (ICLR)*, 2024.

[3] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, page 1467–1474, 2012.

[4] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 of *Texts and Readings in Mathematics*. Springer, 2023.

[5] H. Cao et al. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[6] N. Carlini et al. Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy (SP)*, 2024.

[7] P. W. Koh et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'21, pages 5637–5664, 2021.

[8] R. Liptser and A. N. Shiryayev. *Theory of Martingales*, volume 49 of *Mathematics and its Applications*. Springer, 1989.

[9] G. Martínez et al. Combining generative artificial intelligence (ai) and the internet: heading towards evolution or degradation? *arXiv preprint arXiv:2303.01255*, 2023.

[10] G. Martínez et al. Towards understanding the interplay of generative artificial intelligence and the internet. *arXiv preprint arXiv:2306.06130*, 2023.

[11] I. Shumailov et al. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[12] H. Thanh-Tung and T. Tran. Catastrophic forgetting and mode collapse in gans. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[13] H. Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[14] V. G. Yaji and S. Bhatnagar. Analysis of stochastic approximation schemes with set-valued maps in the absence of a stability guarantee and their stabilization. *IEEE Transactions on Automatic Control*, 65(3):1100–1115, 2019.

[15] W. X. Zhao et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.