REACHABILITY AND ASYMPTOTICS OF GAUSSIAN TRANSFORMER DYNAMICS*

ALBERT ALCALDE†, ZHENGPING JI†, AND ENRIQUE ZUAZUA†‡

Abstract. We formulate data propagation through the Transformer, the machine learning architecture powering large language models, as a nonlinear control system on the space of probability measures. For the mean-field Transformer model with self-attention and affine feed-forward layers, we prove that Gaussian distributions remain exactly Gaussian along the induced flow. This invariance reduces the infinite-dimensional measure dynamics to a finite-dimensional bilinear control system governing the evolution of the mean and covariance, reformulates the expressive capacity of Transformers as a reachability problem for prescribed Gaussian moments, and reveals a novel connection with Riccati-type equations from classical filtering and control.

For time-varying controls, we prove exact finite-time reachability of any target Gaussian distribution whose covariance matrix has the same rank as the initial one, this rank constraint being an intrinsic invariant of the dynamics. For time-invariant parameters, we derive explicit spectral conditions leading either to asymptotic stability toward positive-definite equilibria or to finite-time blow-up of the covariance.

Numerical experiments complement the theory by showing that practical Transformers with Gaussian inputs remain close to moment-matched Gaussian distributions through early and intermediate layers, while Transformers with prescribed attention matrices reproduce the predicted covariance regimes: bounded evolution in stabilizing configurations and blow-up in destabilizing ones.

Key words. deep learning, mean-field transformers, self-attention, Riccati differential equations, covariance control

MSC codes. 68T07, 93B03, 93D20

1. Introduction.

Transformers [10, 44] have become the dominant architecture in modern machine learning, achieving state-of-the-art performance in natural language processing [3, 20], computer vision [16, 32], genomics [2, 30], and scientific machine learning [12, 37]. Despite their empirical success, a rigorous mathematical framework characterizing exactly when and how Transformers reliably represent and propagate information remains elusive, which has motivated a growing theoretical effort to understand them through continuous-depth and mean-field limits, using tools from nonlinear control [13, 15, 17, 25, 39]. When the number of layers and inputs becomes large, the Transformer model can be viewed as a controlled nonlinear flow acting on probability measures [34], in which the controlled state space is the density of the distribution of the inputs, the layer index is interpreted as a continuous time variable, and the layer-varying parameters of the Transformer serve as controls.

A natural and analytically tractable setting for studying this measure flow is

*Submitted to the editors DATE.

Funding: AA was funded by the European Union's Horizon Europe MSCA project ModConFlex (grant number 101073558). EZ was funded by the Alexander von Humboldt-Professorship program, the ERC Advanced Grant CoDeFeL, the Grants PID2020-112617GB-C22 KiLearn and TED2021-131390B-I00-DasEl of MINECO and PID2023-146872OB-I00-DyCMaMod of MICIU (Spain), the European Union's Horizon Europe MSCA project ModConFlex (grant number 101073558), the Transregio 154 Project "Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks" of the DFG, the AFOSR 24IOE027 project, and the SURE-AI Centre grant 357482, Research Council of Norway.

†Chair for Dynamics, Control, Machine Learning & Numerics (Alexander von Humboldt Professorship), Department of Mathematics, Friedrich–Alexander-Universität Erlangen–Nürnberg, 91058 Erlangen, Germany. (albert.alcalde@fau.de, zhengping.ji@fau.de, enrique.zuazua@fau.de)

‡Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

Chair of Computational Mathematics, Fundación Deusto. Av. de las Universidades, 24, 48007 Bilbao, Basque Country, Spain.

the invariant manifold of Gaussian input distributions [17]. This setting is widely adopted in modeling independently sampled data (see Subsection 1.2) and, crucially, the Gaussian invariant manifold reduces the infinite-dimensional controlled evolution characterizing the information-propagation properties of Transformers into a finite-dimensional control system for the mean and covariance (see Subsection 2.1). The Gaussian framework provides a rigorous control-theoretic paradigm to investigate the training dynamics and expressivity of Transformers. Specifically, the approximation capacity of a Transformer translates directly into a reachability problem: does there exist a sequence of time-varying controls (weight parameters) capable of steering an initial Gaussian measure along the nonlinear flow to match an arbitrary target Gaussian? Furthermore, understanding the forward-pass dynamics naturally raises fundamental questions of asymptotic behavior: under what control parameter regimes does the Transformer flow stabilize to a stationary distribution, and when does it diverge? In the language of control theory, these properties correspond precisely to the system's controllability and stability [40].

In this paper, we address the key questions of reachability and asymptotic behavior of the nonlinear control system arising from the mean-field Transformer model in the Gaussian setting. The Transformer architecture poses distinct control-theoretic challenges: (i) due to the strong nonlinear nature of self-attention, the well-posedness of the mean-field evolution is not guaranteed when the initial measure is not of compact support; (ii) the Gaussian reduction leads to coupled nonlinear dynamics of the mean and covariance, complicating the use of classical geometric control techniques [11]; (iii) the non-commutativity of the control matrices arising from the self-attention mechanism presents significant obstacles to establishing the existence of equilibria, and the nonlinearities in the covariance flow complicate global stability analysis. We overcome these difficulties by resolvent techniques and perturbation analysis, explicitly bridging the Transformer dynamics with classical Riccati theory.

1.1. Our contributions. The main contributions of this paper are as follows:

- We study the mean-field Transformer model with affine feed-forward layers and self-attention, and prove that the class of Gaussian measures remains invariant under the resulting system. This yields a Riccati-type ODE system for the evolution of mean and covariance (Proposition 2.1), resembling a classical formulation in optimal filtering and control theory:

$$\begin{cases} \dot{\mu} &= (A+V)\mu + V\Sigma B\mu + b, \\ \dot{\Sigma} &= A\Sigma + \Sigma A^\top + V\Sigma B\Sigma + \Sigma B^\top \Sigma V^\top, \end{cases} \tag{1.1}$$

  where $\mu(t) \in \mathbb{R}^d$ is the mean and $\Sigma(t) \in \mathbb{R}^{d\times d}$ is the covariance of the Gaussian at time t, while $A(t), B(t), V(t) \in \mathbb{R}^{d\times d}$ and $b(t) \in \mathbb{R}^d$ are trainable parameters acting as controls. Further, when the feed-forward is constructed using ReLU activation, we derive quantitative estimates on the discrepancy between measure flows and the Gaussian evolution (1.1) (Proposition 2.3). The argument is not specific to ReLU and indicates how analogous estimates may be obtained for more general Lipschitz activations.
- For time-varying parameter matrices, we show that the rank of the covariance matrix Σ is preserved along the flow of (1.1) (Lemma 3.1). Leveraging matrix congruence transformations, we construct explicit time-varying control paths that achieve exact finite-time reachability of any target Gaussian state sharing this initial covariance rank (Theorem 3.2).

- For time-invariant parameters, we characterize the long-time behavior of the mean/covariance dynamics (1.1). For stabilizing parameter regimes, we show existence of positive-definite equilibria and derive sufficient conditions for local stability (Theorems 4.1 and 4.5). Conversely, under destabilizing parameter configurations, the covariance blows up in finite time (Theorem 4.6). Moreover, one can asymptotically match an arbitrary target mean by static choices of parameters under stabilizing conditions (Theorem 4.8). Interestingly, the resulting stability conditions are consistent with empirical observations in pretrained vision Transformers [43].
- We perform numerical experiments serving two complementary purposes beyond the exact Gaussian theory. First, we show that pretrained Transformers, although outside the assumptions of the affine mean-field model, preserve an approximately Gaussian moment structure over early and intermediate layers when initialized with Gaussian inputs. Second, we test the robustness of the covariance predictions in more realistic discrete architectures with nonlinear feed-forward blocks, showing bounded covariance growth in stabilizing sign configurations and blow-up in destabilizing ones.

Taken together, these results provide a rigorous framework for understanding data propagation in trained Transformers — traditionally studied via static approximation theory — from a dynamical and control-theoretic perspective.

1.2. Motivating the Gaussian setting. The restriction to Gaussian input distributions is not merely an analytical convenience, but also a standard paradigm in theoretical studies of the in-context learning (ICL) capabilities of Transformers [22, 27, 46, 49]. Broadly speaking, ICL investigates how Transformers can solve families of supervised learning tasks directly from contextual examples provided at inference time. In this setting, the input to the Transformer typically consists of a sequence $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m), (x_{\text{query}}, 0)\}$, where each feature $x_i \in \mathbb{R}^d$ (including the query x_{query}) is independently sampled from a Gaussian distribution (often $x_i \sim \mathcal{N}(0, \Sigma)$), and the labels y_i are generated by some unknown task-specific rule (for instance, $y_i = w^\top x_i$ in a linear regression task, with $w \sim \mathcal{N}(0, I_d)$). The Transformer is trained over many such tasks to predict the missing label y_{query} from the contextual examples. An important aspect of this framework is that Gaussian inputs are not intended as realistic models of data distributions. Rather, they provide an analytically tractable ensemble allowing one to isolate and study the mechanisms by which attention layers aggregate and propagate information.

1.3. Related work. Our work builds on recent efforts to understand Transformer architectures through mean-field limits and controlled measure flows. We position our contribution with respect to three closely related directions: asymptotic analyses of attention dynamics, controllability of Transformers, and Riccati-type covariance dynamics.

Asymptotics and mean-field perspective on Transformer dynamics. A recurring theme in theoretical studies is that repeated application of attention layers can drive collapse phenomena: inputs cluster and attention matrices can become effectively low-rank, limiting expressivity. At the particle level, the asymptotic dynamics of Transformers only with self-attention layers have been studied in [6, 8, 24, 35].

These observations are further clarified in the mean-field regime, where the dynamics can be described by partial differential equations on probability measures in $\mathbb{R}^d$. Early contributions identify well-posedness [39], clustering [15, 25], metastability [5, 13, 14, 23], and phase transitions relevant to long-context attention [18]. Closest

to our work is [17], which proves Gaussian invariance for self-attention-only Transformers and derives corresponding mean and covariance equations. Their asymptotic results rely on a strong commutativity condition between the control matrices and the initial covariance. Our contribution is to move beyond the self-attention-only setting by incorporating affine feed-forward layers, and to analyze the resulting Gaussian system relying on weaker sign assumptions on the control matrices.

Reachability and simultaneous controllability of Transformers. The question of controllability and target reachability is central to Transformers. At the discrete level, a first approximate simultaneous controllability result is proved in [48], and extended to the exact setting in [7, 31]. From a mean-field perspective, the simultaneous controllability of Transformers including normalization layers has been established in [26], while [4] takes an optimal control approach to study the training dynamics of Transformers. Although related, our work takes the different perspective of studying how a target Gaussian can be reached with minimal assumptions on the controls.

Bilinear systems and covariance control. As established in Proposition 2.1, (1.1) forms a bilinear system, revealing a surprising link between Transformer models and the linear quadratic regulator theory, as the covariance matrix Σ can be viewed as a "feedback gain" of the mean μ (see Subsection 4.1). It is also related to the optimal estimation problem which aims at designing feedback controllers for the closed-loop system to reach a specified state covariance [29]. Unlike in optimal control where the Riccati equation is a tool to find optimal gains with guaranteed well-posedness, in our analysis the Riccati/Bernoulli-type equation is part of the state dynamics, the stability of which is under question: the coupled structure (1.1) is hence fundamentally new due to the specific structure of Transformers, whose behavior is complicated by the inherent lack of commutativity between control parameters and the covariance. Our controllability analysis of μ and Σ based on congruence transformations avoids the complexity of bracket computation in traditional nonlinear control theory [21].

1.4. Organization of the paper. The remainder of the paper is organized as follows. In Section 2, we introduce the Gaussian Transformer model. Section 3 is devoted to finite-time reachability of the model, while Section 4 addresses its asymptotic behavior for time-invariant parameters. In Section 5, we present our numerical experiments and Section 6 concludes the paper, identifying future perspectives.

2. Transformer dynamics with Gaussian initial conditions. We study a class of nonlinear transport equations arising as mean-field limits of Transformer architectures [44]. Following recent developments at the interface of machine learning, optimal transport, and control theory [17, 26, 39], these models describe the evolution of probability measures driven by parameterized, nonlocal vector fields encoding the architecture of the network. We briefly recall the modeling pathway leading to the equations considered in this work, referring to [25] for detailed derivations from the discrete architecture.

The Transformer is a deep neural network model operating on a finite collection of particles $\{x_i^\ell\}_{i=1}^n \subset \mathbb{R}^d$, called *tokens*, which represent, for instance, words in a sentence or pixels in a picture. In the infinite-depth limit, their evolution is modeled as a continuous-time interacting particle system governed by

$$\dot{x}_i(t) = \sigma(A(t)x_i(t) + b(t)) + \sum_{j=1}^{n} \frac{e^{x_j(t)^\top B(t)x_i(t)}}{\sum_{\ell=1}^{n} e^{x_\ell(t)^\top B(t)x_i(t)}} V(t)x_j(t). \tag{2.1}$$

In this system, $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ denotes a nonlinearity applied componentwise known

as the *activation function*, while $A(t), B(t), V(t) \in \mathbb{R}^{d\times d}$ and $b(t) \in \mathbb{R}^d$ are time-dependent parameters inherited from the trained network. Motivated by the increasing context lengths of modern Transformers [36], we are interested in studying the limit as the number of particles $n \to \infty$. Introducing the empirical measure $\rho_t^n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, the right-hand side of (2.1) can be written as a functional of ρ_t^n. As $n \to \infty$, for compactly supported input measures, this system converges (in the sense made precise in [17]) to a well-posed continuity equation

$$\begin{cases} \partial_t \rho_t + \nabla_x \cdot (\rho_t \Gamma[\rho_t](t,x)) = 0, \\ \rho_{t=0} = \rho_0, \end{cases} \tag{2.2}$$

where $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ and for any $\rho \in \mathcal{P}(\mathbb{R}^d)$, the velocity field $\Gamma[\rho]$ is defined as

$$\Gamma[\rho](t,x) = \sigma\left(A(t)x + b(t)\right) + \frac{\int e^{y^\top B(t)x} V(t) y \,\mathrm{d}\rho(y)}{\int e^{y^\top B(t)x}\,\mathrm{d}\rho(y)}. \tag{2.3}$$

We refer to (2.2) as the *Transformer PDE*. The vector field (2.3) naturally decomposes into a pointwise drift

$$\mathcal{F}(t,x) \coloneqq \sigma(A(t)x + b(t)),$$

corresponding to the so-called *feed-forward* layers, and a nonlocal interaction term

$$\mathcal{A}[\rho](t,x) \coloneqq \frac{\int e^{y^\top B(t)x} V(t) y \,\mathrm{d}\rho(y)}{\int e^{y^\top B(t)x}\,\mathrm{d}\rho(y)}, \tag{2.4}$$

which corresponds to the *self-attention* mechanism. From a control-theoretic viewpoint, (2.2) is a controlled continuity equation with measure-dependent drift, where the time-dependent matrices $A(t), B(t), V(t), b(t)$ act as control variables.

In this work, we focus on the dynamics of (2.2) for Gaussian input measures.

As we will show below in Proposition 2.1, when the initial measure ρ_0 is Gaussian and the activation function σ is the identity, the solution ρ_t remains Gaussian for all t, and the infinite-dimensional dynamics (2.2) reduce to an ODE system governing the evolution of the mean and covariance. We refer to the resulting reduced-order model as the *Gaussian Transformer*. Thus, the question of well-posedness of the model can be studied equivalently in this finite-dimensional setting. Alternatively, for the case of a Gaussian input measure and σ being the ReLU activation function, we will show that the solution remains sub-Gaussian for short times (see Lemma 2.2). This allows us to extend the well-posedness guarantee to this setting.

Throughout the paper, we will denote a (symmetric) positive definite (respectively, positive semi-definite, negative definite and negative semi-definite) matrix $A \in \mathbb{R}^{d\times d}$ by $A \succ 0$ (resp. $A \succeq 0$, $A \prec 0$ and $A \preceq 0$).

2.1. The Gaussian Transformer. For Gaussian measures, the self-attention operator admits an explicit linear representation. If $\rho = \mathcal{N}(\mu, \Sigma)$, then the *attention-only* Transformer PDE (2.2) with $\sigma = 0$ induces the vector field

$$\mathcal{A}[\rho](t,x) = V(t)(\mu + \Sigma B(t)x) \tag{2.5}$$

as shown in [17]. In particular, Gaussian measures are invariant under the self-attention-only flow, and their evolution reduces to a closed system of ODEs for (μ, Σ). We extend this structure to the full Transformer dynamics with an affine feed-forward term, i.e., $\sigma = \mathrm{id}$ in (2.3).

PROPOSITION 2.1 (Gaussian Transformer). *Let ρ_t be the solution of (2.2) with $\sigma = \mathrm{id}$ and initial condition $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$, where $\Sigma_0 \succeq 0$. Then, there exists $T_{\max} > 0$ such that for all $t \in [0, T_{\max})$, ρ_t remains Gaussian, with its mean $\mu(t)$ and covariance $\Sigma(t)$ satisfying*

$$\begin{cases} \dot{\mu} &= (A(t) + V(t) + V(t)\Sigma B(t))\,\mu + b(t), & \mu(0) = \mu_0, \\ \dot{\Sigma} &= A(t)\Sigma + \Sigma A(t)^\top + V(t)\Sigma B(t)\Sigma + \Sigma B(t)^\top \Sigma V(t)^\top, & \Sigma(0) = \Sigma_0. \end{cases} \tag{2.6}$$

Proof. Here and below, unless otherwise stated, A, B, V, b are evaluated at time t. By (2.5), the velocity field is given by

$$\Gamma[\rho_t](t,x) = Ax + b + V\mu + V\Sigma Bx = \big(A + V\Sigma B\big)x + \big(V\mu + b\big),$$

which is affine in x. The pushforward of a Gaussian by an affine map is again Gaussian, so ρ_t stays Gaussian. To derive dynamics for the mean and the covariance, we use standard moment identities: for any smooth $\phi : \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \phi(x)\,\mathrm{d}\rho_t(x) = \int \nabla\phi(x) \cdot \Gamma[\rho_t](t,x)\,\mathrm{d}\rho_t(x). \tag{2.7}$$

Next, we use the definition of $\mu(t) = \int x\,\mathrm{d}\rho_t(x)$ and (2.7) with $\phi(x) = x$ componentwise to obtain

$$\dot{\mu} = \int \Gamma[\rho_t](t,x)\,\mathrm{d}\rho_t(x) = \int \big(Ax + b + V\mu + V\Sigma Bx\big)\,\mathrm{d}\rho_t(x) = A\mu + b + V\mu + V\Sigma B\mu,$$

which gives the desired mean dynamics. For the covariance dynamics, let $M(t) := \mathbb{E}_{\rho_t}[xx^\top]$, so $\Sigma = M - \mu\mu^\top$. Differentiating M gives

$$\dot{M} = \mathbb{E}_{\rho_t}\big[x\Gamma[\rho_t](t,x)^\top + \Gamma[\rho_t](t,x)x^\top\big].$$

Since $\Gamma[\rho_t](t,x) = Ax + b + V\mu + V\Sigma Bx$, linearity and the identity $\mathbb{E}_{\rho_t}[x] = \mu$ yield

$$\begin{aligned} \dot{M} &= \mathbb{E}_{\rho_t}[x(Ax)^\top] + \mathbb{E}_{\rho_t}[x(V\Sigma Bx)^\top] + \mathbb{E}_{\rho_t}[xb^\top] + \mathbb{E}_{\rho_t}[x(V\mu)^\top] \\ &\quad + \mathbb{E}_{\rho_t}[(Ax)x^\top] + \mathbb{E}_{\rho_t}[(V\Sigma Bx)x^\top] + \mathbb{E}_{\rho_t}[bx^\top] + \mathbb{E}_{\rho_t}[(V\mu)x^\top] \\ &= MA^\top + MB^\top\Sigma V^\top + \mu b^\top + \mu\mu^\top V^\top + AM + V\Sigma BM + b\mu^\top + V\mu\mu^\top. \end{aligned}$$

Differentiating $\Sigma = M - \mu\mu^\top$ and substituting $\dot{\mu}$ gives

$$\dot{\Sigma} = \dot{M} - \dot{\mu}\,\mu^\top - \mu\,\dot{\mu}^\top = A\Sigma + \Sigma A^\top + V\Sigma B\Sigma + \Sigma B^\top\Sigma V^\top,$$

which completes the proof. ∎

2.2. Transformer dynamics with ReLU feed-forward layers. Proposition 2.1 shows that Gaussian distributions are invariant under the Transformer PDE (2.2) when the activation function in the feed-forward layers is set to the identity. In this section, we consider the dynamics of the distribution ρ_t in the presence of ReLU activation $\sigma(x) = \max(0, x)$ (componentwise), and quantitatively show that it remains close to the Gaussian evolution (2.6), hence justifying the well-posedness of the Gaussian Transformer under ReLU activations.

First, we prove the following result about the sub-Gaussian behavior of the dynamics (2.2) for short times. Throughout, $\lambda_{\max}(M)$ (resp. $\lambda_{\min}(M)$) denotes the largest (resp. smallest) eigenvalue of a matrix M.

LEMMA 2.2. *Let A, b, B, V be fixed parameters. Denote by ρ_t the solution of* (2.2) *with* ReLU *activation $\sigma = \max(0, x)$ and initial condition $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$. Assume $\kappa_0 \in \mathbb{R}$ satisfies $4\|B\| \le \kappa_0 < \frac{1}{2\lambda_{\max}(\Sigma_0)}$ and define*

$$E_t := \int_{\mathbb{R}^d} e^{\kappa_0 |x|^2} \mathrm{d}\rho_t(x).$$

Then, ρ_t is sub-Gaussian in short time, i.e., there exists $T^* > 0$ such that $E_t \le 2E_0$ for all $t \in [0, T^*]$.

The proof is relegated to Appendix A.1. Thanks to Lemma 2.2, we have the following quantitative estimate on the short-time preservation of Gaussianity in (2.2) in the presence of ReLU activation.

PROPOSITION 2.3. *Let A, B, V and b be fixed parameters. Denote by ρ_t the solution of* (2.2) *with* ReLU *activation $\sigma = \max(0, x)$ and initial condition $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$, where $\Sigma_0 \succeq 0$. If $\lambda_{\max}(\Sigma_0) < \frac{1}{8\|B\|}$, then for $t \in [0, T^*]$,*

$$W_2(\rho_t, \nu_t) \le t \cdot \|\max(0, -(Ax + b))\|_{L^2(\rho_0)} + \mathcal{O}(t^2),$$

where $W_2(\cdot, \cdot)$ is the Wasserstein-2 distance between distributions, ν_t is the Gaussian evolution solving (2.2) *with $\sigma = \mathrm{id}$ and $\nu_0 = \rho_0$.*

The proof is postponed to Appendix A.2. We nonetheless note that the argument for the short-time preservation mechanism in Proposition 2.3 is not specific to ReLU. More generally, for other Lipschitz nonlinear activations σ, under suitable assumptions on the initial covariance matrix, one may expect the existence of a time $T_\sigma > 0$ such that, for $t \le T_\sigma$, $W_2(\rho_t, \nu_t) \le t \cdot \|\sigma(AX_0 + b) - (AX_0 + b)\|_{L^2(\rho_0)} + \mathcal{O}(t^2)$. Thus, if $\sigma(AX_0 + b)$ is close to $AX_0 + b$ under the initial law, the nonlinear evolution remains close to the corresponding Gaussian evolution for short times.

Proposition 2.3 implies that the Gaussian model (2.6) provides a quantifiable short-time approximation of the nonlinear Transformer PDE (2.2). Although Gaussianity is not preserved for nonlinear activations, the result quantifies a tube of validity around the Gaussian dynamics. This justifies focusing the subsequent analysis on the Gaussian setting, where the evolution is finite-dimensional, guaranteeing that the analysis on the Gaussian Transformer (2.6) is informative for real models in practice (see Section 5 for empirical validation with pretrained Transformer models).

3. Finite-time reachability. Our first set of results concerns finite-time behavior and the reachability set of the Gaussian Transformer (2.6) under all possible choices of the time-varying parameters $A(t)$, $V(t)$ and $B(t)$.

The following result establishes that the rank of $\Sigma(t)$ along the flow (2.6) cannot change as a function of t, so it is a natural invariant of the flow (2.6).

LEMMA 3.1 (Rank preservation). *Let $A, B, V\colon [0, \infty) \to \mathbb{R}^{d\times d}$ be time-varying matrices. For any finite time interval $[0, T_{\max})$ on which the solution of* (2.6) *exists, we have* $\mathrm{rank}(\Sigma(t)) = \mathrm{rank}(\Sigma_0)$, $\forall t \in [0, T_{\max})$.

Proof. We define a time-varying matrix $M(t) := A(t) + V(t)\Sigma(t)B(t)$ for $t \in [0, T_{\max})$. By symmetry of the equation we have $\dot{\Sigma}(t) = M(t)\Sigma(t) + \Sigma(t)M(t)^\top$. This is a linear time-varying equation in terms of a congruence transformation. Let $\Psi(t, 0)$ be the state transition matrix from time 0 to time t for the linear system $\dot{x}(t) = M(t)x(t)$, meaning that it is the unique solution of

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi(t, 0) = M(t)\Psi(t, 0), \quad \Psi(0, 0) = I.$$

Then the unique solution for $\Sigma(t)$ is

$$\Sigma(t) = \Psi(t,0)\Sigma(0)\Psi(t,0)^\top. \tag{3.1}$$

The matrix $\Psi(t,0)$ as the state transition matrix of the equation $\dot{x}(t) = M(t)x(t)$ is always invertible for any time $t \in [0, T_{\max})$. This is because its determinant is given by Liouville's formula (see, for example, [41, Lemma 3.11]) as

$$\det(\Psi(t,0)) = \exp\left(\int_0^t \operatorname{tr}(M(\tau))\mathrm{d}\tau\right) \neq 0.$$

Therefore, (3.1) and the non-singularity of $\Psi(t,0)$ imply that the rank of $\Sigma(t)$ in (2.6) is invariant and equal to the rank of $\Sigma(0)$ for all t where the solution exists. ∎

THEOREM 3.2. *Suppose $\Sigma_0 \succeq 0$. For any $T > 0$, $\hat{\mu} \in \mathbb{R}^d$ and $\hat{\Sigma} \succeq 0$ satisfying $\operatorname{rank}(\hat{\Sigma}) = \operatorname{rank}(\Sigma_0)$, there exists continuous A, B, $V : [0,T] \to \mathbb{R}^{d\times d}$ and $b : [0,T] \to \mathbb{R}^d$ such that $\mu(t)$, $\Sigma(t)$ solving (2.6) satisfy $\mu(T) = \hat{\mu}$, $\Sigma(T) = \hat{\Sigma}$.*

Proof. Let $A(t) = -V(t) - V(t)\Sigma(t)B(t)$, and $B(t) = \Sigma(t)^\dagger$ be the Moore–Penrose pseudoinverse of $\Sigma(t)$, acting as smooth feedback controls. Then the dynamics of μ and Σ become

$$\begin{cases} \dot{\mu}(t) = b(t), & \mu(0) = \mu_0, \\ \dot{\Sigma}(t) = -V(t)\Sigma(t) - \Sigma(t)V(t)^\top, & \Sigma(0) = \Sigma_0. \end{cases} \tag{3.2}$$

We shall first prove that there exists $V : [0,T] \to \mathbb{R}^{d\times d}$ such that $\Sigma(t)$ satisfies $\Sigma(0) = \Sigma_0$ and $\Sigma(T) = \hat{\Sigma}$. By the same argument as in Lemma 3.1, the solution of (2.6) becomes $\Sigma(t) = \Psi_V(t)\Sigma_0\Psi_V(t)^\top$ where the transition matrix $\Psi_V(t)$ is the unique solution of

$$\dot{\Psi}_V(t) = -V(t)\Psi_V(t), \quad \Psi_V(0) = I_d. \tag{3.3}$$

Thus, it suffices to show there exists $V(t)$ such that $\hat{\Sigma} = \Sigma(T) = \Psi_V(T)\Sigma_0\Psi_V(T)^\top$.

We start by constructing $\Psi_V(T)$. Since $\Sigma_0 \succeq 0$ and $\hat{\Sigma} \succeq 0$, we can write their spectral decompositions as $\Sigma_0 = U_0\Lambda_0U_0^\top$, $\hat{\Sigma} = U_1\Lambda_1U_1^\top$, where $U_0, U_1 \in \mathrm{SO}(d)$. Without loss of generality, assume that $\operatorname{rank}(\Sigma_0) = \operatorname{rank}(\hat{\Sigma}) = r$, and that the eigenvalues are arranged so that the positive entries occupy the top-left $r \times r$ block of the diagonal matrices Λ_0, Λ_1, while the remaining blocks are zero.

Let $\lambda_{0,i}$ and $\lambda_{1,i}$ be the i-th diagonal elements of Λ_0 and Λ_1, $i = 1, \dots, d$. Construct a nonsingular diagonal matrix D such that $D_{ii} = \sqrt{\frac{\lambda_{1,i}}{\lambda_{0,i}}}$ for $1 \le i \le r$, and $D_{ii} = 1$ for $r < i \le d$.

Let $M = U_1DU_0^\top$. Then $\det(M) > 0$. Since U_0, U_1 are orthogonal matrices, it is easy to verify that $M\Sigma_0M^\top = U_1\Lambda_1U_1^\top = \hat{\Sigma}$, and hence M is the matrix $\Psi_V(T)$ required. Next we construct the time-varying matrix V.

Since $\det(M) > 0$ and the space of d-dimensional nonsingular real matrices with strictly positive determinants is path-connected, there exists a differentiable path $\Phi : [0,T] \to \mathrm{GL}^+(d,\mathbb{R})$ such that $\Phi(0) = I_d$, $\Phi(T) = M$.

Let $V(t) := -\dot{\Phi}(t)\Phi(t)^{-1}$, then $V(t)$ is the time-varying matrix allowing the matrix $\Psi_V(t)$ in (3.3) to satisfy $\hat{\Sigma} = \Sigma(T) = \Psi_V(T)\Sigma_0\Psi_V(T)^\top$.

Meanwhile, by choosing $b(t) = \frac{1}{T}(\hat{\mu} - \mu_0)$, one obtains the finite-time reachability of $\mu(t)$ in (3.2). This proves the result. ∎

Table 1: Asymptotics of (2.6) with $\Sigma(0) = \Sigma_0 \succ 0$. The first row follows from Theorems 4.1 and 4.8, the second from Theorem 4.5, and the third from Theorem 4.6.

V, B **regime**	**Extra condition**	**Covariance** $\Sigma(t)$	**Mean** $\mu(t)$
$V = \eta I_d$, $\eta(B + B^\top) \preceq 0$	$\eta < 0$	$\Sigma(t) \to U \left(\begin{smallmatrix} \Sigma^a_\infty & 0 \\ 0 & 0 \end{smallmatrix}\right) U^\top$	$\mu(t) \to \mu_\infty \in \mathbb{R}^d$
	$\eta > 0$	$\Sigma(t) \to U \left(\begin{smallmatrix} \Sigma^a_\infty & 0 \\ 0 & 0 \end{smallmatrix}\right) U^\top$	$\|\mu(t)\| \to \infty$
$V \prec 0$, $B \succ 0$	$\Re(\mathrm{spec}(A)) > 0$	$\Sigma(t) \to \Sigma_\infty \succ 0$	Depends on $\mathrm{spec}(A+V+V\Sigma_\infty B)$
	$\Re(\mathrm{spec}(A)) \leq 0$	$\Sigma(t) \to 0$	Depends on $\mathrm{spec}(A+V)$
$V \succ 0$, $B \succ 0$	$\lambda_{\min}(A + A^\top) \geq 0$	$\|\Sigma(t)\| \to \infty$	$\|\mu(t)\| \to \infty$
	$\lambda_{\min}(A + A^\top) < 0$	Depends on Σ_0	Depends on Σ_0

Remark 3.3. Theorem 3.2 ensures the existence of time-varying controls (i.e. layer-varying parameters) to match an arbitrary initial Gaussian with rank-compatible target Gaussians, guaranteeing the approximation capacity of Gaussian Transformers. However, from classical control theory one knows that the time-varying control parameters might be physically infeasible when the target lies near the boundary of reachability sets [19], and hence autonomous choice of parameters across layers might be preferred in control applications.

4. Asymptotic dynamics. To characterize the asymptotics of (2.6), we couple the system with suitable initial conditions $\mu(0) = \mu_0 \in \mathbb{R}^d$ and $\Sigma(0) = \Sigma_0 \succeq 0$, as well as restrict the study to constant-in-time parameters $A, V, B \in \mathbb{R}^{d\times d}$ and bias $b \in \mathbb{R}^d$.

The asymptotic analysis of (2.6) is difficult due to the non-commutativity of Σ with the parameter matrices. As the product of positive definite matrices is not necessarily positive definite, it is not easy to derive the existence of the equilibrium of the nonlinear matrix differential equation based on the positive-definiteness of A, B and V. On the other hand, the asymptotics of the mean depend on the spectral properties of the limiting covariance matrix, and can be interpreted as the problem of characterizing stabilizing feedback gains for a linear system with static feedback.

Given these difficulties, the results in this section are split in two. In the first part, we consider the easier case of $V = \eta I_d$, $\eta \in \mathbb{R}$, in which the analysis simplifies thanks to the connections to Riccati theory, and weaker assumptions are needed on the other matrix parameters. In the second part, we tackle the more general yet involved case of definite $V \in \mathbb{R}^{d\times d}$. We provide a summary of results in Table 1.

4.1. Asymptotic dynamics for $V = \eta I_d$. Before introducing our results regarding the asymptotic dynamics of (2.6) for $V = \eta I_d$, we make the connection to Riccati theory rigorous. This observation, although restricted to a certain parameter range, will prove to be useful in illustrating the dynamics of the Gaussian Transformer.

Throughout this subsection, we will fix $V = \eta I_d$, $\eta \in \mathbb{R}$, and assume

$$\eta(B + B^\top) \preceq 0. \tag{4.1}$$

If (4.1) holds, we can write $-H^\top H \coloneqq \eta(B + B^\top)$ for some $H \in \mathbb{R}^{m\times d}$ and $m \leq d$. This yields

$$\dot\Sigma = A\Sigma + \Sigma A^\top - \Sigma H^\top H\Sigma, \quad \Sigma(0) = \Sigma_0, \tag{4.2}$$

which is a particular instance of the differential Riccati equation. For any initial condition $\Sigma_0 \succeq 0$, standard Riccati theory [1, Theorem 4.1.6] proves the existence of a unique solution of (4.2) for all times only for the case when $\eta(B + B^\top) \preceq 0$. Otherwise, finite-time blow-up of the solution may occur. In turn, the mean dynamics becomes

$$\dot{\mu} = (A + \eta I_d)\mu + \Sigma(\eta B)\mu + b = (A - \Sigma H^\top H)\mu + \eta(I_d - \Sigma B^\top)\mu + b.$$

The first term has the same structure as the Kalman–Bucy error dynamics with covariance (4.2), for which exponential decay is known under suitable assumptions [38]. However, the additional term $\eta(I_d - \Sigma B^\top)\mu$ has no direct analogue in the standard Kalman–Bucy setting and prevents us from inferring the asymptotic behavior of μ from Riccati theory alone.

Thus, in the case $V = \eta I_d$, the covariance dynamics admit a precise Riccati interpretation under (4.1). This analogy is useful but incomplete: it does not by itself characterize the mean dynamics, and it does not extend directly to general value matrices. A separate analysis is therefore required.

In what follows, we will denote the spectrum of a matrix $A \in \mathbb{R}^{d\times d}$ by $\mathrm{spec}(A) \subset \mathbb{C}$, and by $\Re(\mathrm{spec}(A))$ the real part of its elements. The next theorem characterizes the asymptotic behavior of $\Sigma(t)$ in terms of the real part of the spectrum of A.

THEOREM 4.1. *Let $\Sigma(t)$ be the solution of (4.2) with $\Sigma_0 \succ 0$. Suppose that A has real Schur decomposition as*

$$A = U\tilde{A}U^\top = U\begin{pmatrix} A_a & A_{12} \\ 0 & A_n \end{pmatrix}U^\top, \tag{4.3}$$

where U is an orthogonal matrix, $\Re(\mathrm{spec}(A_a)) > 0$ and $\Re(\mathrm{spec}(A_n)) \le 0$. Define the negative semidefinite matrix $\tilde{B} \coloneqq \left(\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right) = U^\top\eta(B+B^\top)U$, in which B_{11} has the same dimension as A_a. Then when $t \to \infty$, we have

$$\Sigma(t) \to U\begin{pmatrix} \Sigma^a_\infty & 0 \\ 0 & 0 \end{pmatrix}U^\top$$

where Σ^a_∞ is the unique positive definite solution of the algebraic Riccati equation

$$0 = A_a\Sigma^a_\infty + \Sigma^a_\infty A_a^\top + \Sigma^a_\infty B_{11}\Sigma^a_\infty. \tag{4.4}$$

Proof. As Σ_0 is invertible, we consider the decomposition $\Sigma^{-1}(t) = U\tilde{P}(t)U^\top$ and study the dynamics of $\tilde{P}(t)$ for $t \ge 0$. Using that $U\frac{\mathrm{d}\tilde{P}}{\mathrm{d}t}U^\top = \frac{\mathrm{d}\Sigma^{-1}}{\mathrm{d}t} = -\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}$, from (4.2) we have

$$\dot{\tilde{P}} = -\tilde{A}^\top\tilde{P} - \tilde{P}\tilde{A} - \tilde{B}.$$

Writing $\tilde{P} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$ corresponding to the blocked form of A and $\tilde{B}$, we have

$$\begin{aligned}
\dot{P}_{11} &= -A_a^\top P_{11} - P_{11}A_a - B_{11} \\
\dot{P}_{12} &= -A_a^\top P_{12} - P_{11}A_{12} - P_{12}A_n - B_{12} \\
\dot{P}_{21} &= -A_{12}^\top P_{11} - A_n^\top P_{21} - P_{21}A_a - B_{21} \\
\dot{P}_{22} &= -A_n^\top P_{22} - P_{22}A_n - (A_{12}^\top P_{12} + P_{21}A_{12}) - B_{22}
\end{aligned}$$

Recalling the classical results of Lyapunov differential equations (for example, see [9, Section 6.7, Theorem 7.5]), one sees that P_{11} will converge to the unique positive

definite solution of $0 = -A_a^\top P_{11} - P_{11}A_a - B_{11}$, denoted by P_{11}^∞, due to the fact that $-A_a$ is stabilizing.

As for $P_{12}(t)$, since $P_{11}(t)$ remains bounded for all $t \geq 0$, the limit of $P_{12}(t)$ when $t \to \infty$ depends on the eigenvalues of A_a and A_n, more specifically, its growth rate is

$$\lambda_{12} = \max \Re(\mathrm{spec}(-A_a^\top)) + \max \Re(\mathrm{spec}(-A_n)), \tag{4.5}$$

while for $P_{22}(t)$ we have the growth rate of its homogeneous part as

$$\lambda_{22} = 2 \max \Re(\mathrm{spec}(-A_n)) \tag{4.6}$$

By definition, $\Re(\mathrm{spec}(-A_n)) \geq 0$ and $\Re(\mathrm{spec}(-A_a)) < 0$, hence $\lambda_{22} > \lambda_{12}$ and $\lambda_{22} \geq 0$. As the growth rate of P_{22} is determined by $\lambda_{22} - \lambda_{12}$ and $B_{22} \succ 0$, we have $P_{22}(t) \to \infty$ when $t \to \infty$. Next, we study the limit of $\tilde{P}$ to derive the limit of Σ. Let

$$\Sigma(t) = U \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) \end{pmatrix} U^\top.$$

By definition of the inverse matrix, we have $P_{21}(t)\Sigma_{12}(t) + P_{22}(t)\Sigma_{22}(t) = I$. As $\lim_{t\to\infty} P_{22}(t) = \infty$, it follows that $\lim_{t\to\infty} \Sigma_{22}(t) = 0$; on the other hand, we have

$$P_{11}(t)\Sigma_{12}(t) + P_{12}(t)\Sigma_{22}(t) = 0$$

which implies that $\lim_{t\to\infty} \Sigma_{12}(t) = 0$ (because $P_{11}(t) \succ 0$ for all $t \geq 0$ and $P_{11}^\infty \succ 0$). Finally, we have

$$\Sigma_{11} = (P_{11} - P_{12}P_{22}^{-1}P_{21})^{-1}$$

then, by the estimates of λ_{12} and λ_{22} in (4.5) and (4.6), the growth rate of $P_{12}P_{22}^{-1}P_{21}$ is controlled by $-\min(\Re(\mathrm{spec}(A_a))) < 0$. Hence we have $\lim_{t\to\infty} \Sigma_{11}(t) = (P_{11}^\infty)^{-1} \succ 0$. Defining $\Sigma_\infty^a = (P_{11}^\infty)^{-1}$, one sees that Σ_∞^a solves the equation (4.4). By the symmetry of Σ, we also have $\lim_{t\to\infty} \Sigma_{21}(t) = 0$. The conclusion follows. ∎

Remark 4.2 (Comparison with [17]). We compare our result to that in [17, Proposition 4.8]. In their setting, $A = 0$, thus A_a (and consequently B_{11} and Σ_∞) have no dimension. Then, we conclude that $\Sigma(t) \to 0$. From their result, one also concludes that $\mathrm{rank}(\Sigma_\infty) = 0$ and so $\Sigma_\infty = 0$. Overall, our result allows for non-zero A, while their result allows for indefinite B.

4.2. Asymptotic dynamics for definite V. To extend our results for V beyond scalar multiples of the identity, we study the case when B and V are negative or positive definite, and derive asymptotic results based on their sign.

Our first result concerns the spectrum of a matrix arising in the dynamics of μ in (2.6) at the equilibrium of Σ, which applies to general time-invariant A, B and V. This result will be used later for the asymptotic analysis of the mean in Subsection 4.3.

LEMMA 4.3. *Let $A, B, V \in \mathbb{R}^{d\times d}$ be arbitrary real matrices. Assume $\Sigma_\infty \succ 0$ satisfies the algebraic Bernoulli equation*

$$A\Sigma_\infty + \Sigma_\infty A^\top + V\Sigma_\infty B\Sigma_\infty + \Sigma_\infty B^\top \Sigma_\infty V^\top = 0. \tag{4.7}$$

Then all eigenvalues of $M_\infty \coloneqq A + V\Sigma_\infty B$ are purely imaginary, i.e., $\Re\big(\mathrm{spec}(M_\infty)\big) = 0$.

Proof. By definition, (4.7) implies

$$M_\infty \Sigma_\infty + \Sigma_\infty M_\infty^\top = 0. \tag{4.8}$$

Let $\lambda \in \mathbb{C}$ be an eigenvalue of M_∞ with corresponding left eigenvector $v \in \mathbb{C}^{1\times d} \setminus \{0\}$. Multiplying (4.8) by v and v^* yields $v(M_\infty \Sigma_\infty + \Sigma_\infty M_\infty^\top)v^* = 0$.

Next, we use $vM_\infty = \lambda v$ and $M_\infty^\top v^* = \overline{\lambda} v^*$ to compute $(\lambda + \overline{\lambda})\, v\Sigma_\infty v^* = 0$.

Because $\Sigma_\infty \succ 0$, we have $v\Sigma_\infty v^* > 0$, and hence $\lambda + \overline{\lambda} = 0$, which implies $\Re(\lambda) = 0$. Since λ is an arbitrary eigenvalue of M_∞, all eigenvalues of M_∞ lie on the imaginary axis. ∎

Remark 4.4. Lemma 4.3 implies that if the covariance converges to an equilibrium $\Sigma_\infty \succ 0$, then the linear matrix $M_\infty = A + V\Sigma_\infty B$ contributes only to an oscillatory part of the dynamics of μ.

Turning to the covariance dynamics, the following result characterizes its asymptotic behavior under opposite definiteness conditions on V and B.

THEOREM 4.5. *Let $\Sigma(t)$ be the solution of (2.6) with $\Sigma(0) = \Sigma_0 \succ 0$. Let $V \prec 0$ and $B \succ 0$. Then $\Sigma(t)$ is bounded for all $t \in [0, +\infty)$. Moreover:*

1. *If $\Re(\mathrm{spec}(A)) > 0$, then there exists an equilibrium $\Sigma_\infty \succ 0$ solving the algebraic Bernoulli equation (4.7). Further, if $\Sigma_\infty^{-1} V + V\Sigma_\infty^{-1} \prec 0$, then Σ_∞ is a locally stable equilibrium of $\Sigma(t)$.*
2. *If $\Re(\mathrm{spec}(A)) \le 0$ and $\lambda_{\max}(A + A^\top) \le 0$, then $\lim_{t\to\infty} \Sigma(t) = 0$ for all Σ_0.*
3. *If $\Re(\mathrm{spec}(A)) < 0$ and $\lambda_{\max}(A + A^\top) > 0$, then $\Sigma = 0$ is a locally stable equilibrium of the Σ equation in (2.6).*

The proof is technical and thus relegated to Appendix A.3. Finally, when the signs of definiteness of V and B coincide, we have the following result showing the finite-time blow-up of the evolution of Σ.

THEOREM 4.6. *Let $\Sigma(t)$ be the solution of (2.6) with $\Sigma(0) = \Sigma_0 \succ 0$, where $V \succ 0$ and $B \succ 0$. Then*

1. *If $\lambda_{\min}(A + A^\top) \ge 0$, then $\Sigma(t)$ blows up in finite time T_f, and for $\lambda_{\min}(A + A^\top) > 0$ we have*

$$T_f \le \frac{1}{\lambda_{\min}(A + A^\top)} \log\left(1 + \frac{d\,\lambda_{\min}(A + A^\top)}{2\lambda_{\min}(V)\lambda_{\min}(B)\,\mathrm{tr}(\Sigma_0)}\right). \tag{4.9}$$

2. *If $\lambda_{\min}(A + A^\top) < 0$, there exists $C > 0$ such that when $\|\Sigma_0\| \ge C$, $\Sigma(t)$ blows up in finite time.*

Proof. First, consider the case $\lambda_{\min}(A + A^\top) \ge 0$. Differentiating $\mathrm{tr}(\Sigma)$ yields

$$\begin{aligned} \mathrm{tr}(\dot\Sigma) &= 2\,\mathrm{tr}(A\Sigma) + 2\,\mathrm{tr}(\Sigma V \Sigma B) \\ &\ge \lambda_{\min}(A + A^\top)\,\mathrm{tr}(\Sigma) + 2\lambda_{\min}(V)\lambda_{\min}(B)\,\mathrm{tr}(\Sigma^2) \\ &\ge \lambda_{\min}(A + A^\top)\,\mathrm{tr}(\Sigma) + \frac{2\lambda_{\min}(V)\lambda_{\min}(B)}{d}(\mathrm{tr}(\Sigma))^2, \end{aligned} \tag{4.10}$$

where the inequalities are due to the fact that V, B and Σ are all positive definite, and that $\mathrm{tr}(\Sigma^2) \ge \frac{1}{d}(\mathrm{tr}(\Sigma))^2$. Therefore, as $\mathrm{tr}(\Sigma) > 0$ and $\lambda_{\min}(A + A^\top) \ge 0$, it suffices to consider the following differential equation

$$\dot y = \lambda_{\min}(A + A^\top) y + \frac{2\lambda_{\min}(V)\lambda_{\min}(B)}{d} y^2, \quad y(0) = \mathrm{tr}(\Sigma_0) \tag{4.11}$$

and when $\lambda_{\min}(A + A^\top) > 0$ solving it yields an explicit solution

$$y(t) = \frac{\alpha y_0 e^{\alpha t}}{\alpha - \beta y_0 (e^{\alpha t} - 1)}$$

where $\alpha = \lambda_{\min}(A + A^\top)$, $\beta = \frac{2\lambda_{\min}(V)\lambda_{\min}(B)}{d}$. Therefore the blow-up time of $\operatorname{tr}(\Sigma)$ is shorter than that of y, namely, $T_f \le \frac{1}{\alpha} \log\left(1 + \frac{\alpha}{\beta y_0}\right)$, which is the estimate (4.9). On the other hand, if $\alpha = 0$, then solving (4.11) yields $y(t) = \frac{y_0}{1 - y_0 \beta t}$, confirming the finite-time blow-up of Σ.

Next, consider the case when $\lambda_{\min}(A + A^\top) < 0$. From (4.10)(4.11) one can see that when $\operatorname{tr}(\Sigma) > -\frac{\lambda_{\min}(A+A^\top)d}{2\lambda_{\min}(V)\lambda_{\min}(B)}$, $\operatorname{tr}(\Sigma)$ becomes monotonic with an increasing speed bounded from below by a positive constant, and hence blows up in finite time by the aforementioned analysis of the case $\lambda_{\min}(A + A^\top) > 0$. ∎

Remark 4.7. Note that the boundedness of Σ is determined by the sign of the quadratic term in (2.6): when V and B have opposite signs of definiteness, by defining a Lyapunov function $\operatorname{tr}(\Sigma)$ we obtain the boundedness of Σ. Conversely, Σ becomes unstable or blows up when V and B have the same sign of definiteness. From this argument, one can see that the proofs of Theorem 4.5 and Theorem 4.6 actually do not rely on the signs of V and B, but on whether they coincide or not.

4.3. Asymptotic mean matching. We now study the ability of the Gaussian Transformer (2.6) to asymptotically match a prescribed mean. For $V = \eta I_d$, using the tools developed in Subsection 4.1 we show that, once the covariance converges, the mean can be steered to any desired target by suitable time-invariant parameters.

THEOREM 4.8. *Consider* (2.6) *with* A *having the decomposition* (4.3)*, suppose* $V = \eta I_d$ *and* $B + B^\top \succ 0$.

1. *If* $\eta < 0$*, then* $\lim_{t\to\infty} \mu(t) = -(A + \eta\Sigma_\infty B + \eta I_d)^{-1} b$ *where* Σ_∞ *solves* (4.7).
2. *If* $\eta > 0$ *and* A_a *exists, then* $\|\mu(t)\| \to \infty$.

Proof. Define $M^\eta(t) \coloneqq A + \eta\Sigma(t)B$. Then, the resulting dynamics of μ is

$$\dot{\mu} = (M^\eta(t) + \eta I_d)\mu + b, \tag{4.12}$$

where $\lim_{t\to\infty} \Re(\operatorname{spec}(M^\eta(t))) \le 0$. Indeed, by Theorem 4.1, we have

$$M^\eta_\infty \coloneqq \lim_{t\to\infty} M^\eta(t) = \lim_{t\to\infty} A + \eta\Sigma(t)B = U \begin{pmatrix} A_a + \Sigma^a_\infty B_{11} & A_{12} \\ 0 & A_n \end{pmatrix} U^\top,$$

where U is an orthogonal matrix, $\Re(\operatorname{spec}(A_a)) > 0$ and $\Re(\operatorname{spec}(A_n)) \le 0$. On the other hand, by Lemma 4.3, $\Re(\operatorname{spec}(A_a + \Sigma^a_\infty B_{11})) = 0$ due to the fact that Σ^a_∞ solves (4.4) which is a special case of (4.7). Hence, we have $\Re(\operatorname{spec}(M^\eta_\infty)) \le 0$.

Define $\mu_\infty \coloneqq -(M^\eta_\infty + \eta I_d)^{-1} b$. Calculating the error dynamics yields

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mu(t) - \mu_\infty) = (M^\eta(t) + \eta I_d)(\mu(t) - \mu_\infty) + (M^\eta(t) - M^\eta_\infty)\mu_\infty. \tag{4.13}$$

When $\eta < 0$, choose the Lyapunov function as $\mathcal{V}(x) \coloneqq x^\top P x$ where $P \succ 0$ solves the Lyapunov equation $(M^\eta_\infty + \eta I_d)^\top P + P(M^\eta_\infty + \eta I_d) = -I_d$, then one can use the fact that $M^\eta_\infty = \lim_{t\to\infty} M^\eta(t)$ to show that the driftless dynamics $\dot{x} = (M^\eta(t) + \eta I_d)x$ exponentially decay to zero. Finally by applying Duhamel's principle to (4.13) and using the fact that $\lim_{t\to\infty} M^\eta(t) - M^\eta_\infty = 0$, we have $\lim_{t\to\infty} \mu(t) = \mu_\infty$.

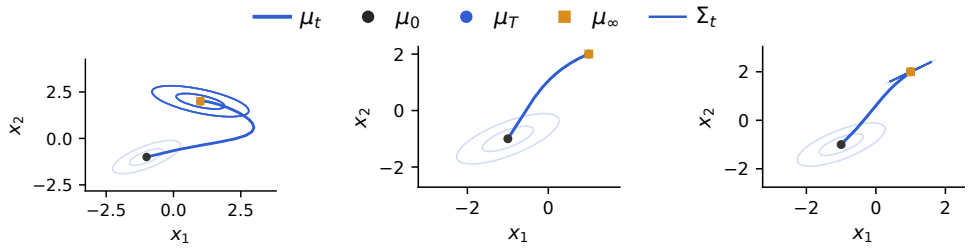
On the other hand, if $\eta > 0$ and there exists an eigenvalue of A with positive real part, then ηI_d plus the purely oscillatory component $A_a + \Sigma_\infty^a B_{11}$ in M_∞^η contributes to an unstable force in the dynamics (4.12), driving $\mu(t)$ to infinity exponentially. ∎

Remark 4.9 (On the role of non-zero bias). In the special case $b = 0$, the mean dynamics reduces to a homogeneous linear system. Thus, by Theorem 4.8, when $V = \eta I_d$, $\mu(t) \to 0$ or $\mu(t) \to \infty$ as $t \to \infty$ for any A and any initial condition μ_0. Hence a nonzero bias term b is essential for asymptotic non-zero mean matching.

Remark 4.10 (On the definite V case). Similarly, one can weaken the assumptions on V by leveraging the results in Subsection 4.2. Indeed, combining the convergence of the covariance (Theorem 4.5) with Lemma 4.3, we obtain that whenever $\Sigma(t) \to \Sigma_\infty$, the long-time behavior of $\mu(t)$ is determined by the spectrum of this limiting matrix, and more specifically, by the stability of V plus a purely oscillatory component $A + V\Sigma_\infty B$.

5. Numerical experiments. In this section, we examine two aspects of the Gaussian Transformer dynamics (2.6). First, we validate the asymptotic mean and covariance behavior predicted in Section 4 in a controlled two-dimensional setting. Second, we test whether an analogous Gaussian moment structure is visible in pretrained Transformer models when their inputs are sampled from a Gaussian distribution. Codes to reproduce our results can be found in https://github.com/DCN-FAU-AvH/gaussianTransformers.

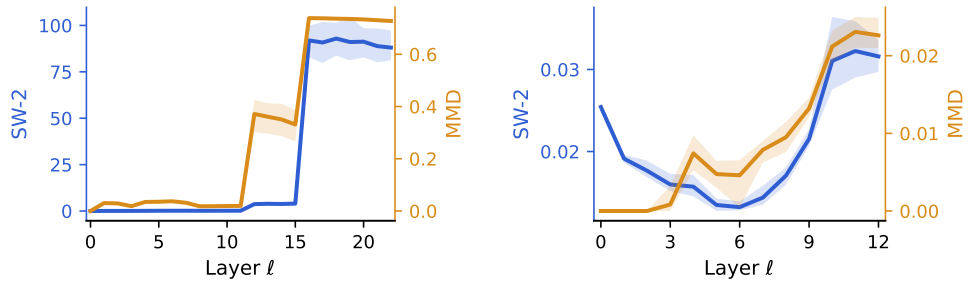
5.1. Validation of asymptotic results. We begin with a two-dimensional example illustrating the asymptotic behavior of Equation (2.6) described in Section 4. We fix $B \succ 0$ and $V = -I_d$, and consider three choices of A: one with $\Re(\mathrm{spec}(A)) > 0$, one with $\Re(\mathrm{spec}(A)) \leq 0$, and one whose eigenvalues have real parts of mixed sign. We set $\mu_0 = (-1, -1)^\top$, $\Sigma_0 = \left(\begin{smallmatrix} 0.4 & 0.2 \\ 0.2 & 0.2 \end{smallmatrix}\right)$, and prescribed limiting mean as $\mu_\infty = (1, 2)^\top$. In each case, the bias term b is chosen according to the mean-matching condition of Theorem 4.8, so that μ_∞ is an equilibrium for the mean dynamics.



(a) $\Re(\mathrm{spec}(A)) > 0$ (b) $\Re(\mathrm{spec}(A)) \leq 0$ (c) mixed sign

Fig. 1: Asymptotic dynamics of Equation (2.6) with $B \succ 0$ and $V = -I_d$. The bias is chosen so that $\mu_\infty = (1, 2)^\top$ is the limiting mean.

The resulting trajectories are shown in Figure 1. In all three regimes, the mean converges to the prescribed target by time $T = 100$. The covariance behavior depends on the spectral regime of A, as predicted by Theorem 4.1. For $\Re(\mathrm{spec}(A)) > 0$, the covariance remains nondegenerate and converges toward a finite limiting shape. When



(a) ModernBERT (b) Tiny-DeiT

Fig. 2: Distances between the empirical distribution $\rho^{(\ell)}$ and its moment-matched Gaussian approximation $\gamma^{(\ell)}$. Curves show the mean across 20 batches of $n = 8192$ tokens, and shaded regions indicate the empirical 0.1–0.9 quantile band.

$\Re(\mathrm{spec}(A)) \leq 0$, the covariance collapses. For the mixed-sign case, the evolution is anisotropic: contraction occurs only along the stable directions.

5.2. Gaussian structure in pretrained Transformers. We next examine whether Gaussian moment structure persists in pretrained Transformers. For a model with embedding dimension d, we draw independent input tokens $x_i^{(0)} \sim \mathcal{N}(0, I_d)$, propagate them through the layers, and denote by $\rho^{(\ell)}$ their empirical distribution at layer ℓ. Further, let $\gamma^{(\ell)} = \mathcal{N}\left(\mu^{(\ell)}, \Sigma^{(\ell)}\right)$ be the Gaussian distribution with the same empirical mean and covariance as $\rho^{(\ell)}$. We compare $\rho^{(\ell)}$ with $\gamma^{(\ell)}$ using the sliced

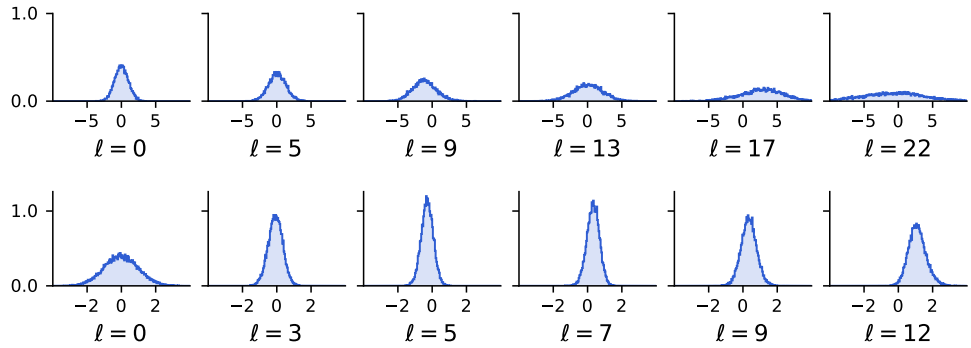
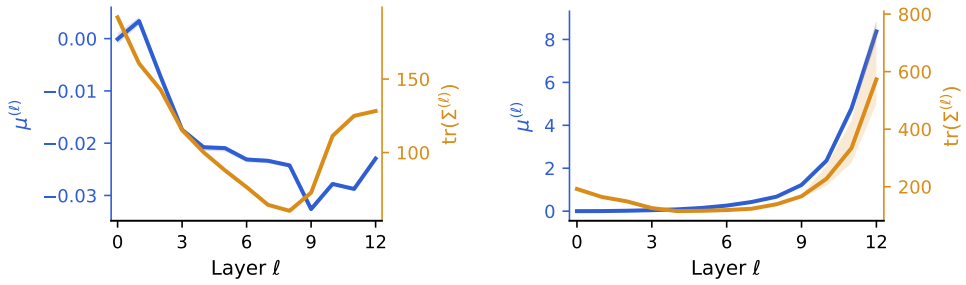


Fig. 3: Marginal distributions across layers of ModernBERT (top row) Tiny-DeiT (bottom row). Each panel shows the coordinate whose variance is closest to the median at that layer, using $n = 8\,192$ Gaussian input tokens.

Wasserstein distance (SW-2):

$$\mathrm{SW}_2(\rho^{(\ell)}, \gamma^{(\ell)}) = \left(\mathbb{E}_\theta W_2^2\left(\langle\theta, x^{(\ell)}\rangle, \langle\theta, z^{(\ell)}\rangle\right)\right)^{\frac{1}{2}},$$



(a) $V \prec 0$, $B \succ 0$ (b) $V \succ 0$, $B \succ 0$

Fig. 4: Evolution of the empirical mean and covariance trace in modified Tiny-DeiT blocks. Each curve reports the average over 20 independent batches, each containing $n = 8192$ tokens initialized from a standard Gaussian distribution. Shaded regions indicate the empirical 0.1–0.9 quantile band.

where $x^{(\ell)} \sim \rho^{(\ell)}$ and $z^{(\ell)} \sim \gamma^{(\ell)}$, and the maximum mean discrepancy (MMD):

$$\mathrm{MMD}^2(\rho^{(\ell)}, \gamma^{(\ell)}) = \mathbb{E}_{x,x' \sim \rho^{(\ell)}}[k(x, x')] + \mathbb{E}_{z,z' \sim \gamma^{(\ell)}}[k(z, z')] - 2\mathbb{E}_{x \sim \rho^{(\ell)},\, z \sim \gamma^{(\ell)}}[k(x, z)],$$

where k is the Gaussian kernel. For our experiments, we use ModernBERT [47], a pretrained language encoder with $L = 22$ layers and $d = 768$, and Tiny-DeiT [42], a compact vision Transformer with $L = 12$ layers and $d = 192$. The resulting distances across layers are shown in Figure 2. For both architectures, the distance to the moment-matched Gaussian remains comparatively small through the early and intermediate layers. The discrepancies increase in deeper layers, indicating that non-Gaussian behavior abruptly emerges at later stages of the network.

We further assess Gaussian preservation in pretrained models by visualizing one-dimensional marginals. At each selected layer, we choose the coordinate whose marginal variance is closest to the median. As shown in Figure 3, the two architectures behave qualitatively differently. In ModernBERT, the marginal broadens with depth and becomes progressively flatter. In contrast, Tiny-DeiT maintains a more Gaussian-like profile across layers, with only modest changes in spread.

Finally, we test whether the covariance regimes predicted by Theorem 4.5 appear in a realistic setting. Starting from Tiny-DeiT, we remove normalization layers and replace self-attention with residual single-head attention maps using prescribed matrices V and B, while leaving the nonlinear feed-forward blocks unchanged. For Gaussian inputs, we track the empirical mean and covariance trace.

Figure 4 shows two distinct regimes: opposing signs of V and B keep the covariance trace bounded, whereas jointly positive signs cause rapid growth. Thus, the sign structure in Theorem 4.5 predicts qualitative covariance behavior even in a discrete residual architecture with nonlinear feed-forward blocks.

6. Conclusions. In this work, we connected modern Transformer architectures with classical control theory by analyzing a mean-field formulation in the Gaussian regime. We proved that, for self-attention with affine feed-forward layers, Gaussian measures are preserved by the induced flow. This invariance reduces the nonlocal transport PDE on probability measures to a finite-dimensional bilinear control system

for the mean and covariance.

Our characterization of mean and covariance dynamics provides a control theoretic interpretation of neural network expressivity. Finite-time reachability yields a minimal interpolation property within the invariant class of Gaussian measures with fixed covariance rank, while the asymptotic analysis identifies parameter regimes leading either to stable covariance dynamics or to finite-time blow-up.

At the model level, covariance instabilities correspond to divergence of token values during forward propagation and therefore suggest possible numerical failure modes. The experiments support this picture: although exact Gaussian invariance is not expected in trained encoder Transformers, Gaussian moment structure remains sufficiently persistent in early and intermediate layers for the reduced dynamics to capture relevant qualitative behavior.

Several questions remain open. First, the long-time analysis relies on structural assumptions on the Transformer parameter matrices, including sign and symmetry conditions on B and V. Extending the theory beyond these regimes is challenging because the resulting Bernoulli-type matrix equations may exhibit non-normal behavior. Second, the present framework treats a simplified architecture: it focuses on single Gaussian input distributions, omits layer normalization, and uses affine feed-forward layers to preserve Gaussian closure. The restriction to a single Gaussian is essential: for Gaussian mixtures, the exponential factors produced by the different components no longer cancel, so the attention field is no longer affine in x, and the closed Bernoulli system for the mean and covariance no longer applies. Consequently, finite Gaussian mixtures do not form an invariant class under the Transformer flow, and no exact finite-dimensional closure analogous to Proposition 2.1 should be expected.

A further perspective is to study the infinite inverse-temperature limit within the Gaussian setting, obtained by replacing B in the attention weights with βB and letting $\beta \to \infty$. For every finite β, the mean and covariance still satisfy a Bernoulli-type system, with the quadratic terms scaled by β. The limit formally corresponds to hardmax attention, but this is not directly well posed on the unbounded support of Gaussians. Whether a meaningful interpretation survives after truncation, renormalization, or projection onto Gaussian moments remains an interesting open problem. More generally, extensions to Gaussian mixtures, multi-head attention, normalization mechanisms, nonlinear feed-forward layers, and singular attention limits would require a more delicate analysis, likely based on projected or approximate moment-closure methods rather than exact Gaussian closure.

Overall, the Gaussian Transformer shows that expressivity and forward-pass stability can be studied jointly through controlled mean–covariance dynamics. This perspective offers a principled route toward the design and analysis of stable, controllable, and mathematically grounded Transformer architectures.

Appendix A. Proofs of technical results. This appendix contains proofs of auxiliary results. Throughout, we omit the domains of integration to ease readability.

A.1. Proof of Lemma 2.2.

Proof. Let $T^* := \sup\{T \geq 0 :\ E_t \leq 2E_0 \text{ for all } t \in [0,T]\}$. We shall show $T^* > 0$.

Define $\Lambda_t(\xi) = \log \int e^{\xi^\top y} \mathrm{d}\rho_t(y)$ as the cumulant generating function of ρ_t. Then, it holds that $\mathcal{A}_{\rho_t}(x) = V\nabla_\xi \Lambda_t(Bx)$. As $E_t \leq 2E_0$, by Young's inequality we have

(A.1)
$$\Lambda_t(\xi) \leq \log\left(e^{\frac{1}{4\kappa_0}|\xi|^2} \int e^{\kappa_0|y|^2} \mathrm{d}\rho_t(y)\right) \leq \log\left(e^{\frac{1}{4\kappa_0}|\xi|^2} \cdot 2E_0\right) \leq \frac{1}{4\kappa_0}|\xi|^2 + \log(2E_0)$$

for all $t \in [0, T^*]$. Note that (A.1) implies that the moment generating function of ρ_t is finite on all of $\mathbb{R}^d$, hence Λ_t is well defined and smooth in ξ, and

$$\nabla \Lambda_t(\xi) = \frac{\int y\, e^{\xi^\top y}\, \mathrm{d}\rho_t(y)}{\int e^{\xi^\top y}\, \mathrm{d}\rho_t(y)}.$$

Since Λ_t is convex and has at most quadratic growth, there exist constants $C_1, C_2 > 0$ such that $\|\nabla \Lambda_t(\xi)\| \le C_1|\xi| + C_2$ for all $t \in [0, T^*]$. It follows that

$$\|\mathcal{A}[\rho_t](t,x)\| = \|V \nabla \Lambda_t(Bx)\| \le \|V\|\big(C_1\|B\||x| + C_2\big).$$

Since $\sigma = \mathrm{ReLU}$ is globally Lipschitz and has linear growth, the velocity field $u_t(x) := \sigma(Ax+b) + \mathcal{A}[\rho_t](t,x)$ is locally Lipschitz in x and satisfies

$$\|u_t(x)\| \le K_1|x| + K_2 \tag{A.2}$$

for all $t \in [0, T^*]$ and some constants $K_1, K_2 > 0$. Therefore, by the Picard–Lindelöf existence-uniqueness theorem for ODEs [41], for each $x_0 \in \mathbb{R}^d$ the characteristic equation

$$\dot{\Phi}_t(x_0) = u_t(\Phi_t(x_0)), \qquad \Phi_0(x_0) = x_0,$$

admits a unique solution on $[0, T^*]$, and the linear-growth bound (A.2) prevents finite-time blow-up. Since ρ_t solves the continuity equation (2.2) with velocity field u_t, the method of characteristics yields $\rho_t = (\Phi_t)_\# \rho_0$ where $(\Phi_t)_\#$ is the pushforward of the flow map Φ_t. Consequently,

$$E_t = \int e^{\kappa_0 |\Phi_t(x_0)|^2}\, \mathrm{d}\rho_0(x_0). \tag{A.3}$$

Thus, for a.e. $t \in [0, T^*]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}|\Phi_t(x_0)| \le \|\sigma(A\Phi_t(x_0)+b) + \mathcal{A}[\rho_t](t, \Phi_t(x_0))\| \le K_1|\Phi_t(x_0)| + K_2. \tag{A.4}$$

Multiplying by $e^{-K_1 t}$ we have $\frac{\mathrm{d}}{\mathrm{d}t}\left(|\Phi_t(x_0)|e^{-K_1 t}\right) \le K_2 e^{-K_1 t}$, and integrating both sides from 0 to t yields $|\Phi_t(x_0)| \le |x_0|e^{K_1 t} + \frac{K_2}{K_1}\left(e^{K_1 t} - 1\right)$, therefore, for any $\epsilon > 0$,

$$|\Phi_t(x_0)|^2 \le (1+\epsilon)e^{2K_1 t}|x_0|^2 + \left(1 + \frac{1}{\epsilon}\right)\left(\frac{K_2}{K_1}(e^{K_1 t} - 1)\right)^2.$$

Then substituting it into (A.3) we have

$$E_t \le \exp\left(\kappa_0\left(1+\tfrac{1}{\epsilon}\right)\left(\tfrac{K_2}{K_1}(e^{K_1 t}-1)\right)^2\right) \int \exp\left(\kappa_0(1+\epsilon)e^{2K_1 t}|x_0|^2\right) \mathrm{d}\rho_0(x_0). \tag{A.5}$$

Since $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $\kappa_0 < \frac{1}{2\lambda_{\max}(\Sigma_0)}$, $\int e^{\alpha|x_0|^2}\, \mathrm{d}\rho_0(x_0)$ is finite for every $\alpha < \frac{1}{2\lambda_{\max}(\Sigma_0)}$, and depends continuously on α. Hence one can choose $\epsilon > 0$ and then $\tau > 0$ sufficiently small so that $\kappa_0(1+\epsilon)e^{2K_1 t} < \frac{1}{2\lambda_{\max}(\Sigma_0)}$ for all $t \in [0, \tau]$, and the right-hand side of (A.5) is bounded by $\frac{3}{2}E_0$ for all $t \in [0, \tau]$. Therefore $E_t \le 2E_0$ on $[0, \tau]$, so $T^* \ge \tau > 0$. ∎

A.2. Proof of Proposition 2.3.

Proof. Consider $X_t \sim \rho_t$, $Y_t \sim \nu_t$ on $\mathbb{R}^d$. Then we have

$$\dot{X}_t = \sigma(AX_t + b) + \mathcal{A}[\rho_t](t, X_t), \qquad \dot{Y}_t = AY_t + b + \mathcal{A}[\nu_t](t, Y_t)$$

Let $e_t = X_t - Y_t$ be the discrepancy and use the shorthand notation $\mathcal{A}[\rho_t](t, X_t) = \mathcal{A}_{\rho_t}(X_t)$ for readability. Differentiation yields

$$\frac{\mathrm{d}}{\mathrm{d}t} e_t = \eta(AX_t + b) + Ae_t + \left(\mathcal{A}_{\rho_t}(X_t) - \mathcal{A}_{\nu_t}(Y_t)\right), \tag{A.6}$$

where $\eta(x) = \sigma(x) - x = \max\{0, -x\}$. We shall bound the term

$$\mathcal{A}_{\rho_t}(X_t) - \mathcal{A}_{\nu_t}(Y_t) = \mathcal{A}_{\rho_t}(X_t) - \mathcal{A}_{\nu_t}(X_t) + \mathcal{A}_{\nu_t}(X_t) - \mathcal{A}_{\nu_t}(Y_t)$$

by calculating the spatial and measure variation, respectively. For the spatial variation we use that ν_t is Gaussian and the expression (2.5), hence $\mathcal{A}_{\nu_t}(X_t) - \mathcal{A}_{\nu_t}(Y_t) = V\Sigma_t B(X_t - Y_t)$. Taking the L^2 norm directly yields the spatial bound:

$$\|\mathcal{A}_{\nu_t}(X_t) - \mathcal{A}_{\nu_t}(Y_t)\|_{L^2} \le \|V\Sigma_t B\| \|X_t - Y_t\|_{L^2}.$$

As for the measure variation, we define for any density ϱ on $\mathbb{R}^d$ two functions $N_\varrho(x) := \int y e^{y^\top Bx} \mathrm{d}\varrho(y)$, $Z_\varrho(x) := \int e^{y^\top Bx} \mathrm{d}\varrho(y)$. Then $\mathcal{A}_\varrho(x) = V \frac{N_\varrho(x)}{Z_\varrho(x)}$. We have

$$\mathcal{A}_{\rho_t}(x) - \mathcal{A}_{\nu_t}(x) = \frac{V}{Z_{\rho_t}(x)} \left(N_{\rho_t}(x) - N_{\nu_t}(x)\right) + \mathcal{A}_{\rho_t}(x) \frac{Z_{\nu_t}(x) - Z_{\rho_t}(x)}{Z_{\rho_t}(x)}. \tag{A.7}$$

Let γ_t be the optimal coupling of ρ_t and ν_t [45], which is a probability distribution on $\mathbb{R}^{2d}$ satisfying $\iint \|y - z\|^2 \mathrm{d}\gamma_t = W_2^2(\rho_t, \nu_t)$. Then we have

$$\begin{aligned}
|Z_{\rho_t}(x) - Z_{\nu_t}(x)| &\le \iint \Big(\|B\|\|x\|\|y - z\|\Big)\Big(e^{y^\top Bx} + e^{z^\top Bx}\Big) \mathrm{d}\gamma_t(y, z) \\
&\le \|B\|\|x\| \left(\iint \|y - z\|^2 \mathrm{d}\gamma_t(y, z)\right)^{\frac{1}{2}} \left(\iint \Big(e^{y^\top Bx} + e^{z^\top Bx}\Big)^2 \mathrm{d}\gamma_t(y, z)\right)^{\frac{1}{2}} \\
&\le \|B\|\|x\| W_2(\rho_t, \nu_t) \left(2 \int e^{2y^\top Bx} \mathrm{d}\rho_t(y) + 2 \int e^{2z^\top Bx} \mathrm{d}\nu_t(z)\right)^{\frac{1}{2}}.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\|N_{\rho_t}(x) - N_{\nu_t}(x)\| &\le \iint \|y - z\| e^{y^\top Bx} \mathrm{d}\gamma_t + \iint \|z\| \left|e^{y^\top Bx} - e^{z^\top Bx}\right| \mathrm{d}\gamma_t \\
&\le \left(\iint \|y - z\|^2 \mathrm{d}\gamma_t\right)^{\frac{1}{2}} \left(\iint e^{2y^\top Bx} \mathrm{d}\gamma_t\right)^{\frac{1}{2}} \\
&\quad + \iint \|z\| \Big(\|B\|\|x\|\|y - z\|\Big)\Big(e^{y^\top Bx} + e^{z^\top Bx}\Big) \mathrm{d}\gamma_t \\
&\le W_2(\rho_t, \nu_t) \left(\int e^{2y^\top Bx} \mathrm{d}\rho_t(y)\right)^{\frac{1}{2}} \\
&\quad + \|B\|\|x\| W_2(\rho_t, \nu_t) \left(\iint 2\|z\|^2 e^{2y^\top Bx} \mathrm{d}\gamma_t + \iint 2\|z\|^2 e^{2z^\top Bx} \mathrm{d}\gamma_t\right)^{\frac{1}{2}}.
\end{aligned}$$

Additionally, by Jensen's inequality, $Z_{\rho_t}(x) \geq \exp\left(\int y^\top Bx \,\mathrm{d}\rho_t(y)\right) \geq \exp\left(\mu_\rho^\top Bx\right)$, where $\mu_\rho = \int y \,\mathrm{d}\rho_t(y)$. Substituting the above estimates of $Z_{\rho_t}(x) - Z_{\nu_t}(x)$, $N_{\rho_t}(x) - N_{\nu_t}(x)$ and $\frac{1}{Z_{\rho_t}(x)} \leq \exp\left(-\mu_\rho^\top Bx\right)$ into (A.7), we obtain

$$\begin{aligned}\|\mathcal{A}_{\rho_t}(x) - \mathcal{A}_{\nu_t}(x)\| \leq & \frac{\|V\|}{Z_{\rho_t}(x)}\left(\|N_{\rho_t}(x) - N_{\nu_t}(x)\| + \frac{\|N_{\nu_t}(x)\|}{Z_{\nu_t}(x)}|Z_{\rho_t}(x) - Z_{\nu_t}(x)|\right)\\ \leq & \|V\|\Big[I_t^1 + I_t^2 + I_t^3\Big]W_2(\rho_t, \nu_t)\end{aligned}$$

where

$$\begin{aligned}I_t^1 &= \left(\int e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\rho_t(y)\right)^{\frac{1}{2}},\\ I_t^2 &= \|B\|\|x\|\left(\iint 2\|z\|^2 e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\gamma_t + \iint 2\|z\|^2 e^{2(z-\mu_\rho)^\top Bx}\mathrm{d}\gamma_t\right)^{\frac{1}{2}},\\ I_t^3 &= \|B\|\|x\|\left(2\int e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\rho_t + 2\int e^{2(z-\mu_\rho)^\top Bx}\mathrm{d}\nu_t\right)^{\frac{1}{2}}(\|\mu_t\| + \|\Sigma_t B\|\|x\|).\end{aligned}$$

We shall show that the $L^2(\rho_t)$ norm of $I_t^1 + I_t^2 + I_t^3$ is finite within $[0, T^*]$. Since $z \sim \nu_t$, terms involving $\int \|z\|^2 e^{c\|z\|}\mathrm{d}\nu_t(z)$, for $c > 0$, are bounded over $t \in [0, T^*]$, and polynomials and exponentials of x integrated against $\rho_t(x)$ are also finite over $t \in [0, T^*]$ due to the sub-Gaussianity proved in Lemma 2.2. Thus, by the algebraic inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, it suffices to consider the worst-case integrals

$$\begin{aligned}J_t^1 &= \int \|x\|^2 \iint \|z\|^2 e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\gamma(y,z)\mathrm{d}\rho_t(x)\\ &= \iiint \|x\|^2\|z\|^2 e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\gamma(y,z)\mathrm{d}\rho_t(x),\\ J_t^2 &= \int \|x\|^4 \int e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\rho_t(y)\mathrm{d}\rho_t(x) = \iint \|x\|^4 e^{2(y-\mu_\rho)^\top Bx}\mathrm{d}\rho_t(y)\mathrm{d}\rho_t(x)\end{aligned}$$

and show that they are both bounded within a short time. By Fubini's theorem and the Cauchy–Schwarz inequality,

$$\begin{aligned}J_t^1 \leq & \left(\int \|x\|^2 e^{\|B\|\|x\|^2}\mathrm{d}\rho_t(x)\right)\left(\iint \|z\|^2 e^{\|B\|\|y-\mu_\rho\|^2}\mathrm{d}\gamma_t(y,z)\right),\\ J_t^2 \leq & \left(\int \|x\|^4 e^{\|B\|\|x\|^2}\mathrm{d}\rho_t(x)\right)\left(\int e^{\|B\|\|y-\mu_\rho\|^2}\mathrm{d}\rho_t(y)\right).\end{aligned}$$

Since $\|B\| < \kappa_0$, by Lemma 2.2, there exists a constant c_0 such that

$$\begin{aligned}\max\Big\{&\left(\int \|x\|^2 e^{\|B\|\|x\|^2}\mathrm{d}\rho_t(x)\right),\\ &\left(\int \|x\|^4 e^{\|B\|\|x\|^2}\mathrm{d}\rho_t(x)\right), \left(\int e^{\|B\|\|y-\mu_\rho\|^2}\mathrm{d}\rho_t(y)\right)\Big\} \leq c_0\end{aligned}$$

for all $t \in [0, T^*]$. As for the second part of J_t^1, we have

$$\begin{aligned}\iint \|z\|^2 e^{\|B\|\|y-\mu_\rho\|^2}\mathrm{d}\gamma_t(y,z) &\leq \left(\int \|z\|^4\mathrm{d}\nu_t(z)\right)^{\frac{1}{2}}\left(\int e^{2\|B\|\|y-\mu_\rho\|^2}\mathrm{d}\rho_t(y)\right)^{\frac{1}{2}}\\ &\leq \left(\int \|z\|^4\mathrm{d}\nu_t(z)\right)^{\frac{1}{2}}\left(e^{4\|B\|\|\mu_\rho\|^2}\int e^{4\|B\|\|y\|^2}\mathrm{d}\rho_t(y)\right)^{\frac{1}{2}},\end{aligned}$$

which is finite over $t \in [0, T^*]$ thanks to Lemma 2.2. Summarizing the above argument, we have $\|\mathcal{A}_{\rho_t}(x) - \mathcal{A}_{\nu_t}(x)\| \le L_\rho(t) W_2(\rho_t, \nu_t)$ where $L_\rho(t)$ is bounded over $t \in [0, T^*]$. Hence by (A.6), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|e_t\|_{L^2} \le \Big(\|A\| + \|V\Sigma_t B\| + L_\rho(t)\Big)\|e_t\|_{L^2} + \|\eta(AX_t + b)\|_{L^2(\rho_t)}.$$

Since $L_\rho(t)$ and $\|V\Sigma_t B\|$ are uniformly bounded over $t \in [0, T^*]$, by Gröwall's lemma, there exists $K > 0$ such that

$$W_2(\rho_t, \nu_t) \le \|e_t\|_{L^2} \le \int_0^t \|\eta(AX_s + b)\|_{L^2(\rho_s)} e^{K(t-s)} \mathrm{d}s. \tag{A.8}$$

By the definition of the pushforward measure, we have

$$\begin{aligned}\|\eta(AX_s + b)\|_{L^2(\rho_s)} =& \|\eta(A\Phi_s(X_0) + b)\|_{L^2(\rho_0)} \\ \le& \|\eta(AX_0 + b)\|_{L^2(\rho_0)} + \|A\|\|\Phi_s(X_0) - X_0\|_{L^2(\rho_0)}.\end{aligned}$$

As we have established in (A.4) that $\frac{\mathrm{d}}{\mathrm{d}t}|\Phi_t(x_0)| \le K_1|\Phi_t(x_0)| + K_2$, it follows that

$$|\Phi_s(X_0) - X_0| \le (K_1|X_0| + K_2)\int_0^s e^{K_1 r}\mathrm{d}r \le (K_1|X_0| + K_2) s e^{K_1 s}.$$

Therefore we have $\|\eta(AX_s + b)\|_{L^2(\rho_s)} \le \|\eta(AX_0 + b)\|_{L^2(\rho_0)} + C_0 s e^{K_1 s}$, where $C_0 = \|A\|(K_1\sqrt{\|\mu_0\|^2 + \mathrm{tr}(\Sigma_0)} + K_2)$. Substituting into (A.8), yields

$$\begin{aligned}W_2(\rho_t, \nu_t) \le& \int_0^t \Big[\|\eta(AX_0 + b)\|_{L^2(\rho_0)} + C_0 s e^{K_1 s}\Big] e^{K(t-s)}\mathrm{d}s \\ \le& \|\eta(AX_0 + b)\|_{L^2(\rho_0)}\left(\frac{e^{Kt} - 1}{K}\right) + C_0 e^{Kt}\int_0^t s e^{(K_1 - K)s}\mathrm{d}s \\ \le& \, t \cdot \|\eta(AX_0 + b)\|_{L^2(\rho_0)} + \mathcal{O}(t^2)\end{aligned}$$

for all $t \in [0, T^*]$, which completes the proof. ∎

A.3. Proof of Theorem 4.5.

Proof. We split the proof in two cases, depending on the sign of $\Re(\mathrm{spec}(A))$.

Case $\Re(\mathrm{spec}(A)) > 0$. We first show the existence of an equilibrium of the Σ equation in (2.6) (i.e., the solution of (4.7)) using the Poincaré-Hopf theorem (see [33, Chapter 7]). More precisely, we construct a domain $\mathcal{K} \subset \mathbb{R}^{d\times d}$ making sure that the vector field defining the Σ equation in (2.6) is transversal to $\partial\mathcal{K}$, and then show the existence of equilibria inside $\mathcal{K}$ by a topological argument.

We begin by constructing the domain $\mathcal{K}$ using hyperplanes defined via Lyapunov functions. Calculating the trace of the derivative of $\Sigma(t)$ yields $\mathrm{tr}(\dot\Sigma) = \mathrm{tr}(A\Sigma + \Sigma A^\top) + \mathrm{tr}(V\Sigma B\Sigma + \Sigma B^\top \Sigma V^\top) = 2\,\mathrm{tr}(A\Sigma) + 2\,\mathrm{tr}(\Sigma V \Sigma B)$. We shall compare the growth of the two terms in this equation to give a sign to $\mathrm{tr}(\dot\Sigma)$ when the norm of Σ is large. On the one hand, there exists $C_1 \ge 0$ such that $\mathrm{tr}(A\Sigma) \le C_1\|\Sigma\|_F$ where $\|\cdot\|_F$ is the Frobenius norm of square matrices. Meanwhile, since $\Sigma V \Sigma \prec 0$ (by Sylvester's law of inertia) and $B \succ 0$, we have $\mathrm{tr}(\Sigma V \Sigma B) < 0$, and further by applying von Neumann's trace inequality (refer to, for example, [28, Theorem 4.3.53]), we have

$$\mathrm{tr}(\Sigma V \Sigma B) \le \lambda_{\min}(B)\,\mathrm{tr}(\Sigma V \Sigma) \le \lambda_{\min}(B)\lambda_{\max}(V)\,\mathrm{tr}(\Sigma^2) \le -C_2\|\Sigma\|_F^2$$

where $C_2 > 0$. Therefore, since $\Sigma \succ 0$ by Lemma 3.1, there exists a constant $R > 0$ such that if $\mathrm{tr}(\Sigma) = R$, then $\mathrm{tr}(\dot{\Sigma}) < 0$. Next, denote $x := \mathrm{vec}(\Sigma)$. Consider a boundary function $U(\Sigma) = x^\top P x = \mathrm{vec}(\Sigma)^\top P \mathrm{vec}(\Sigma)$ where $P \succ 0$ solves the Lyapunov equation

$$P(I_d \otimes A + A \otimes I_d) + (I_d \otimes A + A \otimes I_d)^\top P = I_{d^2}.$$

The existence of P is guaranteed by the fact that $\Re(\mathrm{spec}(A)) > 0$. Then, differentiating the boundary function yields

$$\dot{U} = x^\top P(I_d \otimes A + A \otimes I_d) + (I_d \otimes A + A \otimes I_d)^\top P x \\ + 2x^\top P \mathrm{vec}(\mathcal{N}(\Sigma)) = \|x\|^2 + 2x^\top P \mathrm{vec}(\mathcal{N}(\Sigma)),$$

where $\mathcal{N}(\Sigma) := V\Sigma B\Sigma + \Sigma B^\top \Sigma V^\top$ is the quadratic nonlinearity. Therefore, there exists $\varepsilon > 0$ such that when $U(\Sigma) = \varepsilon$, we have $\dot{U} = \langle \nabla U, \mathrm{vec}(\dot{\Sigma}) \rangle > 0$. Summarizing the above arguments, we construct the set $\mathcal{K} := \{\Sigma \succ 0 \,|\, U(\Sigma) \geq \varepsilon, \mathrm{tr}(\Sigma) \leq R\}$. Note that $\mathcal{K}$ is a compact convex set in the set of $d \times d$ positive definite matrices (and hence in $\mathbb{R}^{d(d+1)/2}$) with nonempty interior. Moreover, on the boundary of $\mathcal{K}$, the Σ vector field $\mathcal{V}(\Sigma) := A\Sigma + \Sigma A^\top + V\Sigma B\Sigma + \Sigma B^\top \Sigma V^\top$ is transversal to the boundary $\partial\mathcal{K}$ and points towards the interior of $\mathcal{K}$ at every point on the boundary, due to the above calculation of the derivatives of $U(\Sigma)$ and $\mathrm{tr}(\Sigma)$. Therefore, by the Poincaré-Hopf theorem, we have

$$\sum_{\{\Sigma^* \in \mathcal{K} \,|\, \mathcal{V}(\Sigma^*) = 0\}} \mathrm{ind}(\Sigma^*) = (-1)^{\frac{d(d+1)}{2}} \chi(\mathcal{K}) = (-1)^{\frac{d(d+1)}{2}}, \tag{A.9}$$

where the last equality is due to the fact that any compact, convex subset of $\mathbb{R}^N$ ($N > 0$) with non-empty interior is homeomorphic to the closed unit ball $\mathbb{D}^N$, whose Euler characteristic χ is 1.

The non-zeroness of (A.9) implies that inside $\mathcal{K}$ there exists at least one equilibrium of the vector field $\mathcal{V}$, guaranteeing that (4.7) has at least one positive definite solution.

Let $\Sigma_\infty \succ 0$ be one solution of (4.7). We shall prove the local stability of Σ_∞ under the condition that $\Sigma_\infty^{-1} V + V \Sigma_\infty^{-1} \prec 0$, by showing that locally the flow of equation of Σ (2.6) converges to the solution of (4.7). Define $\Delta(t) := \Sigma(t) - \Sigma_\infty$, then

$$\dot{\Delta} = M_\infty \Delta + \Delta M_\infty^\top + V\Delta B\Sigma_\infty + \Sigma_\infty B^\top \Delta V^\top + V\Delta B\Delta + \Delta B^\top \Delta V^\top, \tag{A.10}$$

where $M_\infty := A + V\Sigma_\infty B$. Write $P := \Sigma_\infty^{-1} \succ 0$, and define the quadratic functional $\Phi(\Delta) := \frac{1}{2}\mathrm{tr}(\Delta P \Delta P) \equiv \frac{1}{2}\|P^{\frac{1}{2}} \Delta P^{\frac{1}{2}}\|_F^2$. It holds that

$$\frac{1}{2}\lambda_{\min}(P)^2 \|\Delta\|_F^2 \;\leq\; \Phi(\Delta) \;\leq\; \frac{1}{2}\lambda_{\max}(P)^2 \|\Delta\|_F^2. \tag{A.11}$$

We compute the derivative of Φ along solutions of (A.10). Using the cyclic property of the trace, we obtain $\dot{\Phi}(\Delta) = \mathrm{tr}(\Delta P \dot{\Delta} P)$. Substituting (A.10), we decompose $\dot{\Phi} = \mathrm{I} + \mathrm{II} + \mathrm{III}$, where

$$\mathrm{I} := \mathrm{tr}\big(\Delta P(M_\infty \Delta + \Delta M_\infty^\top) P\big),$$
$$\mathrm{II} := \mathrm{tr}\big(\Delta P(V\Delta B\Sigma_\infty + \Sigma_\infty B\Delta V) P\big),$$
$$\mathrm{III} := \mathrm{tr}\big(\Delta P(V\Delta B\Delta + \Delta B\Delta V) P\big).$$

Using cyclicity of the trace, we have $\mathrm{I} = \operatorname{tr}\big(\Delta P \Delta (P M_\infty + M_\infty^\top P)\big)$. Multiplying (4.7) on the left and right by P yields $PM_\infty + M_\infty^\top P = 0$, hence $\mathrm{I} = 0$. Further, using $\Sigma_\infty P = P\Sigma_\infty = I$, we obtain

$$\mathrm{II} = \operatorname{tr}(\Delta P V \Delta B) + \operatorname{tr}(\Delta B \Delta V P) = \operatorname{tr}\big(B\Delta(PV + VP)\Delta\big).$$

By assumption, we have that $PV + VP \prec 0$. Set $S := -(PV + VP) \succ 0$. Then $\mathrm{II} = -\operatorname{tr}(B\Delta S\Delta)$. Since $B \succ 0$ and $S \succ 0$, we have

$$\operatorname{tr}(B\Delta S\Delta) = \|B^{\frac{1}{2}}\Delta S^{\frac{1}{2}}\|_F^2 \;\geq\; \lambda_{\min}(B)\lambda_{\min}(S)\|\Delta\|_F^2.$$

Therefore by (A.11), there exists $c > 0$ such that $\mathrm{II} \leq -c\,\Phi(\Delta)$. As for the cubic term III, by (A.11), there exists $C_1 > 0$ such that $|\mathrm{III}| \leq C_1\Phi(\Delta)^{\frac{3}{2}}$. Combining the above estimates, we obtain $\dot\Phi(\Delta) \leq -c\Phi(\Delta) + C_1\Phi(\Delta)^{\frac{3}{2}}$. Hence, there exists $r > 0$ such that if $\Phi(\Delta) \leq r$, then $\dot\Phi(\Delta) \leq -\frac{c}{2}\Phi(\Delta)$. This implies exponential decay of $\Phi(\Delta(t))$ for all initial data sufficiently close to Σ_∞. Using (A.11) yields $\|\Sigma(t) - \Sigma_\infty\|_F \leq C_2 e^{-\gamma t}\|\Sigma(0) - \Sigma_\infty\|_F$ for some $C_2, \gamma > 0$, proving this case.

Case $\Re(\mathrm{spec}(A)) \leq 0$. We note that $\Sigma_\infty = 0$ is always a solution of (4.7). Next, we show its global or local stability depending on the sign of $\lambda_{\max}(A + A^\top)$. Calculating the derivative of $\operatorname{tr}(\Sigma)$ yields

$$\operatorname{tr}(\dot\Sigma) \leq \lambda_{\max}(A + A^\top)\operatorname{tr}(\Sigma) + 2\lambda_{\max}(V)\lambda_{\max}(B)\operatorname{tr}(\Sigma^2). \tag{A.12}$$

Hence if $\lambda_{\max}(A + A^\top) < 0$ and $V \prec 0$, $B \succ 0$, then $\operatorname{tr}(\dot\Sigma) \leq \lambda_{\max}(A + A^\top)\operatorname{tr}(\Sigma)$, and by Grönwall's Lemma, $\Sigma(t)$ converges to 0 exponentially. If $\lambda_{\max}(A + A^\top) = 0$, then (A.12) yields $\operatorname{tr}(\dot\Sigma) \leq 2\lambda_{\max}(V)\lambda_{\max}(B)\operatorname{tr}(\Sigma^2)$, which is strictly negative as long as $\Sigma \neq 0$, hence assuring that $\lim_{t\to\infty}\Sigma(t) = 0$.

If, instead, $\Re(\mathrm{spec}(A)) < 0$ and $\lambda_{\max}(A + A^\top) > 0$, by first-order linearization of the dynamics of Σ in (2.6), 0 is a locally stable equilibrium. Further, Σ remains uniformly bounded over $t \in (0, +\infty)$ as we have the estimate

$$\operatorname{tr}(\dot\Sigma) \leq \lambda_{\max}(A + A^\top)\operatorname{tr}(\Sigma) + \frac{2}{d}\lambda_{\max}(V)\lambda_{\max}(B)\operatorname{tr}(\Sigma)^2 < 0$$

when $\operatorname{tr}(\Sigma) > -\frac{\lambda_{\max}(A+A^\top)}{\frac{2}{d}\lambda_{\max}(V)\lambda_{\max}(B)} > 0$. Hence completes the proof. ∎

Acknowledgments. The authors thank the Speinshart Scientific Center for AI and SuperTech for its hospitality, where part of this research was conducted.

REFERENCES

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati equations*, Systems & Control: Foundations & Applications, Birkhäuser Verlag, Basel, 2003, https://doi.org/10.1007/978-3-0348-8081-7.

[2] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al., *Accurate structure prediction of biomolecular interactions with alphafold 3*, Nature, 630 (2024), pp. 493–500.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., *Gpt-4 technical report*, arXiv preprint arXiv:2303.08774, (2023).

[4] K. Akman, N. Saldi, and S. Yüksel, *An optimal control approach to transformer training*, arXiv preprint arXiv:2603.09571, (2026).

[5] A. Alcalde, L. Bungert, K. Riedl, and T. Roith, *Quantifying concentration phenomena of mean-field transformers in the low-temperature regime*, arXiv preprint arXiv:2605.10931, (2026).

[6] A. Alcalde, G. Fantuzzi, and E. Zuazua, *Clustering in pure-attention hardmax transformers and its role in sentiment analysis*, SIAM Journal on Mathematics of Data Science, 7 (2025), pp. 1367–1393.

[7] A. Alcalde, G. Fantuzzi, and E. Zuazua, *Exact sequence interpolation with transformers*, Mathematical Foundations of Machine Learning, 2 (2026), p. 2, https://doi.org/10.1007/s44439-026-00005-y.

[8] C. Altafini, *Multistability of self-attention dynamics in transformers*, IEEE Transactions on Automatic Control, (2026), pp. 1–16, https://doi.org/10.1109/TAC.2026.3687461.

[9] P. J. Antsaklis and A. N. Michel, *Linear systems*, Birkhäuser Boston, Inc., Boston, MA, 2006.

[10] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473, (2014).

[11] R.-M. Bianchini and M. Kawski, *Needle variations that cannot be summed*, SIAM Journal on Control and Optimization, 42 (2003), pp. 218–238, https://doi.org/10.1137/S0363012902402876.

[12] C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, A. Allen, J. Brandstetter, P. Garvan, M. Riechert, J. A. Weyn, H. Dong, et al., *A foundation model for the earth system*, Nature, 641 (2025), pp. 1180–1187.

[13] G. Bruno, F. Pasqualotto, and A. Agazzi, *Emergence of meta-stable clustering in mean-field transformer models*, in The Thirteenth International Conference on Learning Representations, 2025.

[14] G. Bruno, F. Pasqualotto, and A. Agazzi, *A multiscale analysis of mean-field transformers in the moderate interaction regime*, in The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2026.

[15] M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand, *Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization*, Philos. Trans. Roy. Soc. A, 383 (2025), pp. Paper No. 20240233, 48, https://doi.org/10.1098/rsta.2024.0233.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-end object detection with transformers*, in European conference on computer vision, Springer, 2020, pp. 213–229.

[17] V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré, *A unified perspective on the dynamics of deep transformers*, arXiv preprint arXiv:2501.18322, (2025).

[18] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet, *Critical attention scaling in long-context transformers*, in The Fourteenth International Conference on Learning Representations, 2026.

[19] Y. Chen, T. T. Georgiou, and M. Pavon, *Optimal steering of a linear stochastic system to a final probability distribution, Part I*, IEEE Trans. Automat. Control, 61 (2016), pp. 1158–1169, https://doi.org/10.1109/TAC.2015.2457784.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[21] D. Elliott, *Bilinear Control Systems: Matrices in Action*, Springer Publishing Company, Incorporated, 1st ed., 2009.

[22] S. Garg, D. Tsipras, P. Liang, and G. Valiant, *What can transformers learn in-context? a case study of simple function classes*, in Advances in Neural Information Processing Systems, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds., 2022.

[23] B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet, *Dynamic metastability in the self-attention model*, arXiv preprint arXiv:2410.06833, (2024).

[24] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *The emergence of clusters in self-attention dynamics*, Advances in Neural Information Processing Systems, 36 (2023), pp. 57026–57037.

[25] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, Bulletin of the American Mathematical Society, 62 (2025), pp. 427–479.

[26] B. Geshkovski, P. Rigollet, and D. Ruiz-Balet, *Measure-to-measure interpolation using transformers*, arXiv preprint arXiv:2411.04551, (2024).

[27] G. Goel, M. Soltanolkotabi, and P. Bartlett, *Training dynamics of softmax self-attention: Fast global convergence via preconditioning*, arXiv preprint arXiv:2603.01514, (2026).

[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[29] A. Hotz and R. E. Skelton, *Covariance control theory*, Internat. J. Control, 46 (1987), pp. 13–32, https://doi.org/10.1080/00207178708933880.
[30] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., *Highly accurate protein structure prediction with alphafold*, nature, 596 (2021), pp. 583–589.
[31] J. Kim, M. Kim, and B. Mozafari, *Provable memorization capacity of transformers*, in The Eleventh International Conference on Learning Representations, 2023.
[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, *Swin transformer: Hierarchical vision transformer using shifted windows*, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[33] J. W. Milnor and D. W. Weaver, *Topology from the differentiable viewpoint*, Princeton University Press, 1997.
[34] G. Peyré, *Optimal and diffusion transports in machine learning*, arXiv preprint arXiv:2512.06797, (2025).
[35] D.-T. Pham, A. N. The, V.-H. Tran, N.-P. Chung, X. T. Tong, T. M. Nguyen, and T. Vo, *Dynamical properties of tokens in self-attention and effects of positional encoding*, in The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2026.
[36] O. Press, N. Smith, and M. Lewis, *Train short, test long: Attention with linear biases enables input length extrapolation*, in International Conference on Learning Representations, 2022.
[37] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, et al., *Probabilistic weather forecasting with machine learning*, Nature, 637 (2025), pp. 84–90.
[38] P. A. Ruymgaart and T. T. Soong, *Mathematics of Kalman-Bucy filtering*, vol. 14 of Springer Series in Information Sciences, Springer-Verlag, Berlin, 1985, https://doi.org/10.1007/978-3-642-96842-6.
[39] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, *Sinkformers: Transformers with doubly stochastic attention*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3515–3530.
[40] P. Tabuada and B. Gharesifard, *Universal approximation power of deep residual neural networks through the lens of control*, IEEE Trans. Automat. Control, 68 (2023), pp. 2715–2728, https://doi.org/10.1109/tac.2022.3190051.
[41] G. Teschl, *Ordinary differential equations and dynamical systems*, vol. 140 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2012, https://doi.org/10.1090/gsm/140.
[42] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, *Training data-efficient image transformers & distillation through attention*, in Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, PMLR, 18–24 Jul 2021, pp. 10347–10357.
[43] A. Trockman and J. Z. Kolter, *Mimetic initialization of self-attention layers*, in International Conference on Machine Learning, PMLR, 2023, pp. 34456–34468.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, *Attention is all you need*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[45] C. Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, Berlin, Heidelberg, 2009, https://doi.org/10.1007/978-3-540-71050-9.
[46] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov, *Transformers learn in-context by gradient descent*, in International Conference on Machine Learning, PMLR, 2023, pp. 35151–35174.
[47] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, et al., *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*, in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 2526–2547.
[48] C. Yun, S. Bhojanapalli, A. S. Rawat, S. Reddi, and S. Kumar, *Are transformers universal approximators of sequence-to-sequence functions?*, in International Conference on Learning Representations, 2020.
[49] R. Zhang, S. Frei, and P. L. Bartlett, *Trained transformers learn linear models in-context*, Journal of Machine Learning Research, 25 (2024), pp. 1–55.