Stable Alignment Research Agenda

Marvin Koß
Heidelberg University

June 20, 2025

Contents

1 Introduction

This is a preliminary outline of the project's resarch agenda, mainly for internal use and providing collaborators with a high-level overview of the project's aims and tools.

The various parts of the problem are likely interlinked in intricate ways. To make analyses tractable it is, as is always the case, still recommendable to start with simple (mathematical) models and propositions, and only connect to the other mathematical parts afterwards. Still, since most of the mathematical tools mentioned below are expected to play a part in the architecture and analyses of the ultimate federated model, it is of utmost importance to always keep the high level picture in mind while phrasing and proving isolated propositions, lest we digress into modelling for the sake of modelling - we are doing **applied** mathematics. For this reason, we should make mathematical choices that likely play well together, such as starting with a base Transformer model that is amenable to an Optimal Transport Analysis, which in turn lets us more easily connect it to Game Theory. At the same time, it is important to also take on an open-ended, investigative approach, and it is possible and desirable that we stumble upon insights that are useful only outside of the context of the project's vision, such as new interpretability methods or improved understanding of how language models do causal inference. We may aim to write multiple papers within this project.

This Research Agenda is continously updated, with a recent version hosted online. It does not yet contain sufficient references to literature for claims made, should be treated with care and is not an official document or a funding proposal. It makes sense at this exploratory point of the project to not yet commit to specific mathematical models and notation but instead handwave at several options and later on figure out which fits best.

The document continues with an outline of the project's vision, followed by an aside on terminology and a brief introduction to LLMs and, furthermore, their variant we work with, the Sinkformer, and then outlines the various mathematical aspects of the research agenda. Each agenda piece is not developed rigorously; rather, questions, problems and potential answers, tools and solutions are briefly sketched.

1.1 Vision

The project's framing is loosely based on four intertwined premises:

Government: Many nation states will, in the near future, be governed by LLMs. There are degrees to which this can become true, such as more and more (in number and degree) important decisions being made by AI, or AI actually being legally instantiated as economic and political actors.

Alignment: AI Alignment has enough efforts going into it that the development of sufficiently benign AI systems can be expected.[1] Since achieving peace politically is a very difficult problem, it is therefore a reasonable approach to try to use aligned AI systems to align humans with one another.

Values: LLMs can find good values (contextual decision heuristics) for the progressively more complex social world in which we cohabit with agentic computational systems. Good values for the near future could for example be:

(a) One should strive to use any general intelligence one has access to - one's own or that of other entities - in a preferrably rather **generative** than **discriminative** manner.

(b) Society should avoid the design and instantiation of **zero-sum games**.

It is unclear whether these two values are in some sense dual to one another. It would be nice to live in a world where the main thesis of Cicero's fifth Tusculan Disputation holds true, namely one wherein egoism necessitates altruism, a game theoretic situation type toward which societal incentives should always be guiding us.

On a related note, reaching goals is becoming more and more trivial. General Intelligence, or goal reaching competence, will soon be worth at best 20 Euros a month. The more pertinent problem now becomes: What are the right goals?

Incentives: Left to themselves, the currently (mainly) economic incentives driving the development of powerful AI systems may lead to malign dynamics, such as influential states, companies - and soon, autonomous AI systems - striving to convert game theoretic environments that are currently still iterated (relatively safe) into non-iterated ones (extremely dangerous).

These considerations, one more contentious than the next, do not need to all hold or be agreed with for one to recognise at least some value in this project.

[1] Is is wrong to ever think of AI Alignment as trivial or solved. Alignment efforts must always be made, and from as many mathematical, algorithmic and humanitarian perspectives as possible.

The items above lead us to attempt to design a communication protocol between governing LLMs that leads to provably stable equilibria. We make various reasonable assumptions, such as for the model's latent spaces to be semantically identical or isomorphic, and try to design an optimal control for the models' processing of their latent distribution during deployment that leads to consensus being the "preferred" option by each individual model.

This is a highly ambitious vision; and even if ultimately getting a working federated model through to policymakers is highly unlikely, there is still much value in designing a vision for language model cooperation that is provably benign.

1.2 Terminology

It is important to use precise terminology. Two terms used in the literature that may prove problematic and need to be differentiated polysemantically in our context are:

Token : In analyses of what we here call the **latent distribution** $\rho = (\rho_t)_{t\in[0,T]}$ of evolving hidden vectors, where

$$\rho_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_t^{(i)}} \in \mathcal{P}_2(S^{d-1})^{2},$$

researchers often use the term *Token* for $x_t^{(i)} \in \mathbb{R}^d$. Although it is reasonable to try to follow established terminology, this particular term is highly misleading: In the context of natural language processing, a token is a discrete element in a finite set called *Vocabulary*. Indeed, the embedding layer of an LLM does map each of these *Tokens* to $x_0^{(i)} \in \mathbb{R}^d$. Further, $x_t^{(i)}$ is always associated to this initial token embedding $x_0^{(i)}$. Importantly, however, $x_t^{(i)}, t > 0$, may point in *any* latent direction and it is therefore usually not sensible anymore to map it to an element of the vocabulary[3]. It is for this reason that it is advisable to avoid the term *token* for arbitary elements of the latent space $x \in \mathbb{R}^d$. Since directions[4] in the latent space ostensibly carry semantic information, I propose the use of the term **Concept** for $x \in \mathbb{R}^d$. Alternative suggestions are appreciated.

Model : There are three main meanings of this term that will occur in this project. I attempt to fix alternative terms for each of them:

LLM: The objects of our study are large language *models*. We shall use the term LLM instead of model for them, even and especially when it is a mathematical object such as $p_\theta(w_{t+1}|w_{\leq t})$ or $(\mathbf{x}, \mathbf{w}) \mapsto \Phi(\mathbf{x}, \mathbf{w})$.

Mathematical: We make mathematical *models*, particularly of LLMs. For instance, we may ultimately decide on a "vanilla" Transformer implementation with run-of-the-mill automatic differentiation, which highlights the fact that our choice to do the analysis using Neural Ordinary Differential Equations (NODEs) is one in *idealisation*, which we call **Mathematical Model**.

Representation: An LLM's latent space, the geometries and logic associated with it *model* the world that produced the input data. We call *models* made by the LLMs **representations**.

1.3 A Primer on LLMs

Essentially no linguistic knowledge is required in this project - one of the major successes of the deep learning revolution within the natural language processing domain.

[2] Depending on context, to avoid thinking of distributions, we may also write $\mathbf{x} \in \mathbb{R}^{n\times d}$ for the latent distribution at time t.

[3] Notwithstanding the applicability of the logit lens view of [14] described in Section 1.3.

[4] Due to Layernorms, the evolution of vectors happens on the unit sphere, and magnitude does not play a role, reducing the dimensionality of the representational space from the latent dimensionality d to $d - 1$.

This section contains a brief introduction to how LLMs process their *latent distribution* ρ. For the sake of this introduction, consider a *vanilla* Transformer, the neural architecture currently the backbone of essentially all LLMs. It consists of L **Blocks** $B_l : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ that, one after another, process the n vectors x_i of dimension d in ρ_t.[5] Each Block contains two main parameterized (learned) maps, firstly the **attention** layer, and secondly the Multi-Layer-Perceptron (**MLP**). Each of the two is followed by[6] a **layernorm**. The functionality of these three components is as follows:

Attention: The attention A_l allows for information flow between the concepts through computation of relevancy. The **attention matrix** $K \in \mathbb{R}^{n\times n}$ is of the form

$$K_{ij} \propto \exp(\langle W_Q x_i, W_K x_j \rangle),$$

meaning that each **query** concept i is projected by the learned $W_Q \in \mathbb{R}^{a\times d}, a << d$ and "compared" by the inner product with the likewise projected **key** concept j, with again $W_K \in \mathbb{R}^{a\times d}$. One then enforces *Gibbs* distributions - called **Softmax** in recent literature - over the columns of K, such that each *row* sums to 1, i.e. each query concept now has a probability distribution over relevancy of all n concepts. The attention map is then given by $A : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$, with

$$A(\mathbf{x}) = [A.(x_1), \ldots, A.(x_n)], \tag{1}$$

$$\text{where } A.(x_i) = \sum_{j=1}^{n} K_{ij} W_V x_j, \quad W_V \in \mathbb{R}^{d\times d}. \tag{2}$$

This setup can be seen as a soft *contextual* dictionary, with each conceptual **key** x_j having a linearly associated conceptual **value** $W_V x_j$.[7]

For LLM training, the definition of A as given thus far is problematic, however: The goal is to predict the next word, and K permits the concept associated with the current token i to "look at" the concept associated with *future* tokens $j > i$, making the learning problem a trivial copying act. To avoid this, a causal mask $m \in \mathbb{R}^{n\times n}$, namely a lower triangular matrix, is elementwise multiplied ($K \leftarrow K \odot m$) before the application of the exponential function. This has, to my knowledge, not been considered in NODE-based analyses of transformers, and the dynamics induced by this is the subject of the sub-agenda outlined in Section 2

Finally, an introduction of attention as it is used in LLMs would not be complete without the mention of **Multihead** attention: Typically, one allows for H heads to each calculate $A = A_h$ in parallel, with unique W_Q^h, W_K^h and a shared W_V.

Attention scales quadratically in time n and is therefore the main bottleneck of inference speed - "inference" here pointing at running the model generatively during deployment, as opposed to the supervised training described below - particularly considering current LLMs allow for n to be in the tens of thousands.

MLP: The MLP M_l is typically a 1-hidden layer Neural Network with a *ReLU*-like Activation function and respective **input** and **output** matrices $W_i \in \mathbb{R}^{f\times d}, W_o \in \mathbb{R}^{d\times f}, f = 4d$. The MLP is often called **position-wise feedforward layer** to emphasize the fact that it processes each token in parallel, with no interaction between them.

[5] t comes from the continuous interpretation of depth, see Section 1.4. Whether the layers are indexed by $[0, T]$ or $\{0, 1, \ldots, L, L+1\}$ is not of importance here, but keep the discrete case in mind for the moment.

[6] Depending on implementation, instead preceded by.

[7] I still use the term *conceptual* here, even though the latent space here is $\mathbb{R}^a$ and not $\mathbb{R}^d$, with the latent space $\mathbb{R}^a$ even possibly differing in its semantic directions among attentions in different layers - which is an interesting research question! A more precise term for elements of $\mathbb{R}^a$ is needed, such as perhaps **downcast**.

A similarity it does have to the attention mechanism described above is that it too can be thought of as a soft dictionary, however not contextual, but fixed: The input vector x_i does a dot product with each row of W_i, and if the dot product is positive, the corresponding neuron **activates**, leading to the forwarding of the associated column in W_o. Thus the rows of W_i can be thought of as learned fixed conceptual **keys**, while the columns of W_o are the corresponding conceptual **values**.

In this way, LLMs can store factual associations within the MLPs. Effectively, this leads to a highly discontinuous *jump* on the unit sphere when considering $\delta_{x_t^{(i)}}$ before and after the application of M_l.

With a parameter count of $8d^2$, the MLPs are the main contributor to the overall parameter count of the Transformer, but their application scales linearly in time n.

Layernorm : The layernorms project vectors back to the unit sphere S^{d-1}. They may have learned weights.

Each attention and each MLP, as well the block itself, is bypassed by a sum with the identity map, i.e. for each of these **modules** $f : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$, the latent distribution $\mathbf{x}$ is updated as

$$\mathbf{x} \leftarrow \mathbf{x} + f(\mathbf{x}), \quad f \in \{A_l, M_l, B_l \mid l \in [[L]]\}. \tag{3}$$

This summing of the layer output with its input is called **residual (connection)**, and has been shown empirically to not only speed up network training since its popularisation with the introduction of ResNets for Computer Vision, but also boost their performance, likely because of stabler and more informative gradients.

More recent work has shown that it plays a key role in the empirical observation that the latent spaces between layers share the same semantic geometry, leading researchers to say each layer **writes to the residual stream**. With no claim of the following being a proof of the prior assertion, interestingly, one can apply the output layer $W_O \in \mathbb{R}^{v\times d}$ to any intermediate latent distribution $\mathbf{x}$ and observe semantically related output tokens [14].

As hinted at in the above paragraph, the final distribution $\mathbf{x} + B_L(\mathbf{x})$ is projected by the output matrix $W_O \in \mathbb{R}^{v\times d}$ to the vocabulary dimension. Afterwards, a categorical distribution over tokens is obtained through an application of the Softmax function, such that we may write the LLM as $p_\theta(w_t|w_{<t}) \in \mathcal{P}(\mathbb{R}^v)$. p_θ is typically (**pre-**)trained using the **Crossentropy** loss (or, equivalently, **Kullback-Leibler Divergence**) from the data distribution p_{data}, which can be somewhat roughly denoted as

$$\mathcal{L}_{CE}(p_\theta(w_{t+1}|w_{<t})|p_{data}(\hat{w}_{t+1}|w_{<t})) = -\log p_\theta(\hat{w}_t),$$

where $\hat{w}_{t+1}$ is the next token in the training data, sometimes called **gold token**. One typically does some version of (accelerated) stochastic gradient descent on $\mathcal{L}_{CE}$ and lets automatic differentiation cheaply obtain layer Jacobians using the chain rule [15, Prop. 10.16].

It should be mentioned that LLMs are not only pretrained in this fashion, but also adapted to specific application domains - most notoriously, chatbots - in a subsequent, shorter stage called **finetuning**, recently most often using methods from **Reinforcement Learning (RL)**, such as **Reinforcement Learning From Human Feedback (RLHF)** with particular methods such as **PPO, DPO, REINFORCE$_+$, GRPO** - indeed, an evidentially highly useful interpretation of $p_\theta(w_t|w_{<t})$ is as an **agent** with a discrete **action space** of cardinality v; from both the perspective of RL and of Game Theory, the connection and interplay of which is highly interesting in the context of this project!

1.4 Sinkformers

As stated above, it is reasonable to pick both mathematical models that firstly allow for faithful practical implementations of them and secondly interlock well mathematically with the other pieces. We pick a model that can be implemented and analysed as a NODE to gain access to differential equations, and pick a model with doubly stochastic attention matrix K to be able to reason in the language of optimal transport.

1.4.1 NODEs and Normalising Flows

Equation 3 can be seen as one step of explicit Euler for the ordinary differential equation (ODE)

$$\dot{x} = f(x), \tag{4}$$

with the associated mass-preserving[8] continuity equation

$$\partial_t \rho + \nabla \cdot (\rho f) = 0. \tag{5}$$

Equation 4 motivates the mathematical model of **Neural Ordinary Differential Equations (NODEs)** introduced by [4] (initially for vision models), which interprets layers like f as a velocity field and goes from discrete time $l \in [[L]]$ to continuous time $t \in [0, T]$. One may take this as more than a mathematical model, but also an implementation recommendation, and replace the vanilla autograd backward with a discretized ODE integration solve, possibly using finegrained discretization methods such as explicit or implicit Runge Kutta methods. Even though this is then more faithful to the idealised mathematical model, and one even gets the further advantage of letting the discretisation step size and method be adapted automatically to the desired accuracy, this merely lets one choose between a slow and a very slow model: Faithful NODE implementations are still drastically slower than vanilla transformers, partly due to a heretofore lack of specialised CUDA kernels and Runge-Kutta specific ASICs.

We might want to formulate the model as a conditional normalising flow, using perhaps the framework of [17] and schedule of [18].

1.4.2 Sinkformer Analysis

The Sinkformer [16] simply enforces the attention matrix K to be doubly stochastic using Sinkhorn's algorithm, such that each column sums to one and each target key also has a probability distribution over which initial key is relevant. This can be seen as an **inductive bias** as the authors demonstrate in Figure 2 that the attention matrices of learned are usually close to doubly anyways, and in Figure 3 that this indeed leads to faster convergence.

Mathematically, this also turns out to be highly insightful: The authors make some (drastic) simplifying assumptions on the model interpreted as a NODE, namely they omit the MLPs altogether, make W_Q, W_K, W_V time-independent and impose a symmetry assumption on them. After doing so, they obtain two interesting results, understanding which is ***paramount*** to making initial progress in our analysis:

WGF: Proposition 2 shows this idealised model leads to a **Wasserstein Gradient Flow (WGF)** on the functional $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$

$$\mathcal{F}(\mu) := -\frac{1}{2} \int k^\infty \log(\frac{k^\infty}{k^0}) d(\mu \otimes \mu), \tag{6}$$

where k^0 is the attention matrix before any normalisation, and k^∞ is K after applying Sinkhorn's Algorithm. This looks suspiciously close to a KL-Divergence (were it not for the fact that k^0 is unnormalised). If one could obtain a WGF on an actual divergence, we may obtain a **connection** to **Information Geometry**: A Divergence on the **Probability Manifold** $\mathcal{P}(S^{d-1})$ uniquely determines the **Christoffel-Symbol**[9] - in the case of KL $1/-1$ - mapping the landscape ρ_t evolves on.

[8] We always have the same number n of vectors.

[9] The CS then consists in the negative third mixed derivatives of the divergence and determines how vectors must be rotated as they are transported along curves on the probability manifold.

Heat Eqn.: In Theorem 1, the authors show that an idealisation of their idealisation, namely its mean field limit of $n \to \infty$ further regularised by a temperature-like parameter $\epsilon \to 0$ leads to the continuity equation 4 resolving, with $f = -\nabla\rho/\rho$, to

$$\partial_t \rho = \Delta\rho, \tag{7}$$

namely the **heat** equation. This immediately tells us that ρ does **natural gradient ascent** according to **Otto Calculus** [19, Example 15.4, p.438]:

$$\begin{align}
-\Delta\rho &= -\nabla \cdot (\nabla\rho) \tag{8}\\
&= \nabla \cdot (\rho f) \tag{9}\\
&= -\nabla \cdot (\rho\nabla \log \rho) \tag{10}\\
&=: \nabla^{nat} H(\rho), \tag{11}
\end{align}$$

where one can see the similarity of the natural gradient of the functional H to its actual gradient by replacing the divergence with an integral and using linearity. [10]

A subsequent work shone light on some questions left open by this paper:

Sphere: Firstly, note that the analysis conducted by [16] considers only that $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ and not on the unit sphere.

Clustering: Secondly, the fact that their mathematical model predicts that the latent distribution should follow the heat equation, and thereby increase in entropy, contradicts empirical observations [1] which show the entropy of the latent distribution rises until some intermediate layer, from whereon it drops.

The mentioned subsequent work is [9], which conducted the analyses for a similar idealisation of a transformer, without the doubly stochastic attention matrix, and with the layernorm projections added in, more realistically considering the evolution $\rho_t \in \mathcal{P}(S^{d-1})$. They show both empirically (Figure 1) as well as analytically (Theorem 6.1) that the distribution in this idealisation clusters to a dirac mass, namely exponentially in time (Theorem 6.3).[11]

It is pertinent to try and make the idealised mathematical models employed in [16, 9] closer to reality by adding in the MLPs[12], making weights time dependent as they are in real LLMs, removing symmetry assumptions and adding in the causal mask. These added complexities likely lead to PDEs that are much less tractable than a simple heat equation, so concessions and assumptions will still have to be made. One may even rightly ask what the advantage of the continuous NODE perspective is anymore with time dependence and a hybrid ODE with $2L$ jumps. We shall figure out what the right approaches are to make headway here. One promising approach may be that outlined below in Section 3.1.

2 Causality

Though this section is probably the least important for the mathematical success of the project, it is nevertheless interesting to try to understand the way in which Transformers do **causal inference** by finding concepts and **Directed Acyclic Graphs (DAGs)** - enforced by the causal mask - between them. Empirically, consider for instance the groundbreaking performance of the

[10] I am unclear on whether the **Riemannian Manifold** here is S^{d-1} equipped with L_2 or $\mathcal{P}(S^{d-1})$ equipped with the **Fisher Metric**, or the **Probability Simplex** Δ_n of categorical distributions over keys/queries.

[11] This exponential rate is interesting and may have connections to the accuracy-independent $O(\log D)$ depth in the construction of [10], who obtain a **generalisation bound** for transformers trained with MSE on data with low intrinsic dimension D.

[12] Leading to a Hybrid ODE, which ought to be well-posed given that we have finite switches, $2L$ in number.

recent Transformer-based *TabPFN* [12], which is trained to do causal inference on arbitrarily generated synthetic DAGs and looks to be the SOTA in data imputation.

Going back to language though, high level questions to guide our research may be the following:

What does a good ontology for modelling the world that generates language data structurally look like - and can we extract an ontological tree structure from the spherically arranged latent geometry[13]? Intuitively, and perhaps compatible with the structures computable by the attention mechanism, it seems to look like DAGs organised in hierarchies, where nodes in more finegrained (**micro**) conceptual levels may be grouped (*suboptimally*) by *concepts* (binary classifiers) of a higher (**macro**) level into coarser dags that fail to perfectly encapsulate the causality at the lower level.[14] Since I just called a binary classifier a *concept*, lets call these **aggregators** instead, to avoid confusion with the vectors in an LLM's latent space. Appropriate notions of optimality of this aggregation process may be that of **Effective Information** [11] or **τ-abstraction** and **(τ, ω)-transformation** [3]. Since this aggregator framing uses binary classifiers, one may also look towards **Statistical Learning Theory** for **shattering numbers** and **VC-Dimension** [15, Sec. 14.5], which unfortunately suffer from the curse of dimensionality.

Ultimately, it would be great to get to a point where we may properly talk about what "good" **values** are, i.e. contextual decision heuristics, which are ostensibly encoded by LLMs through logic between concepts - comparison, causality and lookup. How does a model find good aggregators between which it can do the logic that lets it act well in the world[15]? Perhaps the framework of [8] could be used to try and discover a latent **Bellman Equation** on concepts.

In the analysis of our causally masked Sinkformer variant, which we may call **Causeformer**, we will likely replace the **Wasserstein Distance** W_2 with one where couplings are constrained to be DAGs [5]. We furthermore must distinguish between permutations in S_n of our empirical measure, TODO.

2.1 The Physics of Metaphysics

It may be interesting to relate the ontological structure found by LLMs (and perhaps empirically extracted by us or others from within them) to the **thermodynamic** considerations in the next section, 3. Perspectives to consider in this light are:

Microstates: The aggregation from micro to macro is reminiscent of the tracking of high level properties with the logarithm in thermodynamics, a function which mathematically represents how many questions one has to ask to know a microstate.

Renormalisation: The ontological aggregation from level to level could perhaps be viewed as a renormalisation group from QFT, perhaps obeying something structurally similar to the form of the renormalisation group equation

$$\frac{\partial g}{\partial \log \mu} = \beta(g) := G\ d/(\partial G/\partial g), \tag{12}$$

$$\text{where} \quad g(\mu) := G^{-1}\big((\frac{\mu}{M})^d G(g(M))\big) \tag{13}$$

is the coupling parameter of energy μ, with reference scale M and scaling function G. g could be the statistics of the layer parameters that vary with time, μ is time t, the update

[13] Clustering with **Gromov-Wasserstein** and bounded transport cost on clustering matrices?

[14] A concrete example of this framing is the hierarchy of the natural sciences - particle physics, chemistry, biology, psychology. It may be the case that particular attention layers implement the logic found in a particular level of such a hierarchy of causal models. It is entirely unclear whether this aggregator framing is appropriate for LLMs at all; a hint at such a hierarchical structure within them may however be the logarithmic depth in the construction of [10].

[15] Which "*simply*" means reducing crossentropy, for now.

rule for layer statistics G could be estimated empirically (finite differences?), and M is the scale estimated for the vocabulary-bound embedding of the input sentence at $t = 0$.

Onto. Entropy: TODO Sketch Tree Entropy

3 Entropy

The main considerations for our theoretical investigation of the behavior of $H(\rho_t)$ over time have been recounted in Section 1.4.2. We may empirically estimate the entropy of ρ_t using the Kozachenko-Leonenko estimator [7]. The below subsections contain approaches to understanding the entropy and its relation to the other components of the project

3.1 Free Energy

We may try to introduce a potential U, representing perhaps the action of the MLP[16], and hope to obtain in lieu of the deterministic continuity equation 5 a **Fokker-Planck Equation (FPE)** of the form

$$\partial_t \rho = \nu \Delta \rho + \nabla \cdot (\rho \nabla U), \tag{14}$$

leading to a WGF on the **Helmholtz free energy** [19, Sec. 23, p.700]

$$\mathcal{F}[\rho] = \int U d\rho + \nu H(\rho), \tag{15}$$

discretised in optimal transport canon through the scheme of **Jordan-Kinderlehrer-Otto (JKO)** [13]. If $\nu \geq 0$ is sufficiently small and $\|\nabla U\|$ is big enough, we should expect $\partial_t H(\rho) < 0$, leading to the observed clustering of concepts. One could then verify the **diffusive** (ν) vs **contractive** ($\|\nabla U\|$) behavior over time empirically by fitting $\nu = \nu_t$ through regressing on $\|\mathbf{x}_{l+1} - \mathbf{x}_l\|_2^2$ and fit ∇U with ridge regression on the latents. At this point, one may hope to apply Bakry-Émery [2][17] or borrow from the analysis of [9] to show an exponential convergence in t to the minimiser of $\mathcal{F}$.

3.2 Morse Theory

The latent distribution ρ_t may find itself, particularly around the intermediate layer where it switches from entropic ascent to descent, near saddle points in the entropy landscape where $\nabla H(\rho_t) \approx 0$. Near such saddle points the topology of the landscape is such that there are multiple basins in which to go, which all lead to a descent in entropy. One could let things take their natural gradient descent way and let the model pursue the *steepest* of the descent directions, or one could analyse the curvature of the loss landscape, approximate the Hessian and identify basin directions if its entries happen to have mixed signs. TODO

4 Optimal Transport

The pertinence of optimal transport has been argued for above in Section 1.4. A couple of further considerations follow.

[16] This potential is a scalar field representing everything that is *not* pure heat-equationesque diffusion, i.e. the drift induced by the learned weights and nonlinearities; in mathematical practice it would be the symmetric part of an affine map, with the nonlinearities incorporated by the fact that $ReLU(x) = \nabla \frac{1}{2} \max(0, x)^2$ is the gradient of a convex function.

[17] In addition to the above, we simply require the curvature of our riem. manifold to satisfy $Ric + \Delta U \geq Cg$, where $C > 0$ is scalar and g is the metric.

4.1 Barycenter Drift

For the game theoretic part, we want to ensure consensus of the LLMs' decisions somehow. One way to get there may be to incorporate a drift towards the **Wasserstein Barycenter** [6] of their latent distributions ρ^i. For N measures $\{\mu^i\}_{k=1}^N \subset \mathcal{P}(\Omega)$, it is defined as $\overline{\mu} :\in \arg\min f$ in $\mathcal{P}(\Omega)$, where

$$f(\mu) := \frac{1}{N} \sum_{k=1}^{N} W_2^2(\mu, \mu^i). \tag{16}$$

We then may want a utility summand u_t^i in the cost functional in Section 5 of the form

$$u_t^i := \lambda \nabla_1 W_2(\rho_t^i, \overline{\rho}_t). \tag{17}$$

Although [6] provides a fast way to compute this barycenter, it may not be necessary to enforce this drift throughout all of the layers t. Taking into account our considerations in Sections **??**, it may be sufficient to enforce barycentre drift only at the initial diffusive layers of the LLM, or indeed just at saddle points of the entropy landscape, i.e. points where decisions can be made about which entropic basin to pursue.

4.2 Generalisation Bounds for Minibatch Sinkhorn

It seems the approximation error and generalisation error of minibatched Sinkhorn [17] are open problems, with empirically good performance. TODO Write down generalisation upper & lower bound approach, hopefully $\Theta(b^{-\frac{1}{2}})$.

5 Game Theory

TODO fill in this section

A Hydrodynamics

Here be dragons (TODO *fill* this in)

References

[1] Riccardo Ali et al. "Entropy-lens: The information signature of transformer computations". In: *arXiv preprint arXiv:2502.16570* (2025).

[2] Dominique Bakry and Michel Émery. "Diffusions hypercontractives". In: *Séminaire de Probabilités XIX 1983/84: Proceedings.* Springer, 2006, pp. 177–206.

[3] Sander Beckers and Joseph Y Halpern. "Abstracting causal models". In: *Proceedings of the aaai conference on artificial intelligence.* Vol. 33. 01. 2019, pp. 2678–2685.

[4] Ricky TQ Chen et al. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).

[5] Patrick Cheridito and Stephan Eckstein. "Optimal transport and Wasserstein distances for causal models". In: *Bernoulli* 31.2 (2025), pp. 1351–1376.

[6] Marco Cuturi and Arnaud Doucet. "Fast computation of Wasserstein barycenters". In: *International conference on machine learning.* PMLR. 2014, pp. 685–693.

[7] Sylvain Delattre and Nicolas Fournier. "On the Kozachenko–Leonenko entropy estimator". In: *Journal of Statistical Planning and Inference* 185 (2017), pp. 69–93.

[8] Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. “Provably efficient reinforcement learning with aggregated states”. In: *arXiv preprint arXiv:1912.06366* (2019).

[9] Borjan Geshkovski et al. “A mathematical perspective on transformers”. In: *arXiv preprint arXiv:2312.10794* (2023).

[10] Alexander Havrilla and Wenjing Liao. “Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 42162–42210.

[11] Erik P Hoel. “When the map is better than the territory”. In: *Entropy* 19.5 (2017), p. 188.

[12] Noah Hollmann et al. “Accurate predictions on small data with a tabular foundation model”. In: *Nature* 637.8045 (2025), pp. 319–326.

[13] Richard Jordan, David Kinderlehrer, and Felix Otto. “The variational formulation of the Fokker–Planck equation”. In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17.

[14] nostalgebraist. *interpreting gpt: the logit lens*. LessWrong. 2020. URL: https://www.lesswrong.%20com/posts/AcKRB8wDpdaN6v6ru/%20interpreting-gpt-the-logit-lens (visited on 05/14/2025).

[15] Philipp Petersen and Jakob Zech. “Mathematical theory of deep learning”. In: *arXiv preprint arXiv:2407.18384* (2024).

[16] Michael E Sander et al. “Sinkformers: Transformers with doubly stochastic attention”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 3515–3530.

[17] Alexander Tong et al. “Improving and generalizing flow-based generative models with minibatch optimal transport”. In: *arXiv preprint arXiv:2302.00482* (2023).

[18] Panos Tsimpos et al. “Optimal Scheduling of Dynamic Transport”. In: *arXiv preprint arXiv:2504.14425* (2025).

[19] Cédric Villani et al. *Optimal transport: old and new*. Vol. 338. Springer, 2008.