Statistical Physics of Transformers

Honkoop, Beerend

Statistical Physics of Transformers

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCES
in
PHYSICS

Author :	Beer Honkoop
Student ID :	S3979725
Supervisor :	Koenraad Schalm
Second corrector :	Subodh Patil

Leiden, The Netherlands, July 10, 2025

Statistical Physics of Transformers

Beer Honkoop

Huygens-Kamerlingh Onnes Laboratory, Leiden University
P.O. Box 9500, 2300 RA Leiden, The Netherlands

July 10, 2025

Abstract

Following the recent boom of generative Artificial Intelligence, Large Language Models (LLMs) have been used extensively across myriad disciplines. Despite our success in creating and improving these models, they are still seen as a black box with little understanding of the hidden mechanisms. A physics perspective on this model is still lacking, which may provide valuable insights into these hidden mechanisms. The spin-transformer establishes a close connection between the building blocks of LLMs, transformers, and a statistical physics vector-spin glass model. Found and provisionally solved by Bal, this spin-transformer provides high-potential avenues to understanding LLMs. Bal's solution takes advantage of expansions around interaction strength, yielding the Thouless-Anderson-Palmer (TAP) solutions in the second order. To review this correspondence, we use Markov Chain Monte Carlo simulations to sample numerical solutions to the vector-spin glass model. The limits of the TAP solutions are addressed, confirming their failure at low temperatures and revealing an unaddressed paramagnetic phase at high temperatures. We further display replica symmetry breaking of the vector-spin glass model, showing a phase diagram resembling that of the Sherrington-Kirkpatrick model. By closing the gap between spin glass systems and transformers, this work aspires to provide a foundation for understanding LLMs through the principles of physics. The code is accessible at `https://github.com/beerch/ST_MCMC`.

Contents

Version of July 10, 2025– Created July 10, 2025 - 10:07

Chapter 1

Introduction

ChatGPT and similar revolutionary artificial intelligence (AI) are built on a concept called Large Language models (LLMs). At the core of LLMs stand transformer modules [1], which introduced a novel method of measuring relations between tokens in a sequence simultaneously [2]. Despite our now-common interaction with LLMs as part of our daily lives, they still exist as a black-box which we do not thoroughly understand [3]. For example, generative AI like LLMs are capable of displaying emergent behavior, which describe how the model's capacity to not only generate solutions but more importantly to reason, increases to unknown extents as the model grows in terms of depth, parameter space, or training size [4]. Similar to how Hopfield networks showed the power of a physics-based interpretations of artificial neural networks [5], a statistical physics description of LLMs may shed light on these complex systems.

Whereas earlier AI models (DNN, RNN) have close, well-studied connections to the Ising model, LLMs lack a thorough statistical physics interpretation. Many attempts have been made in an effort to understand LLMs and their transformers, with advances in information geometric approaches [6, 7], detailed probing [8] , and linguistics analysis [9]. There have been attempts at establishing a correspondence between statistical physics and transformers, but they are few and scattered. Geshkovski [10] delves into a mathematical framework for transformers, making their connection to spin systems, and uses this framework to analyze clustering behavior. He describes transformers as a variant of residual neural networks in order to incorporate time-dynamics as seen in spin systems. Li [11] connects LLMs to spin glasses, focusing on the in-context learning of language models. However, these methods fail to incorporate an equilibrium framework under which statistical physics methods normally excel.

Version of July 10, 2025– Created July 10, 2025 - 10:07

A series of blog posts by Mathias Bal from Ghent University proposes and solves a spin-transformer (ST) model [12]. This model establishes a novel relation between spin systems and transformer modules. He specifically builds on top of Deep Equilibrium Transformers, which are an emergent form of deep learning models that replace deep stacks of layers with single equilibrium layer [13]. The correspondence found by Bal has a closer connection to statistical spin systems than aforementioned works due to its equilibrating nature, and furthermore is the only literature that acknowledges the importance of asymmetric couplings in Transformers. He finds an analytic Thouless-Anderson-Palmer (TAP) [14] solution and uses it to build a ST.

The detailed correspondence found by Bal has great potential, as it suggests physical explanations for transformer mechanisms beyond that of previous work. For example, he proposes that residuals acts as an applied magnetic field, that the feed-forward layer acts as an Onsager reaction term, and that multi-head attention acts as multi-Replica Symmetry Breaking. Closely connecting transformer modules to equilibrium spin systems, Bal forms a great starting point for studying the statistical physics of transformers.

In this paper, we aim to elucidate and expand upon the ST correspondence by comparing his analytic solutions to our numerical solutions. We restrict our focus to the single transformer module forward pass, and devise a Markov Chain Monte Carlo (MCMC) scheme of a vector-spin glass emulating the ST. Using our numerical results, we show that they closely relate with Bal's analytic solutions but present interesting differences at varying temperatures. The thesis is structured as follows; in Ch. 2 we briefly explain transformers and spin glasses, and elaborate on the ST correspondence and its solution found by Bal. In Ch. 3 we outline our MCMC simulation, and then compare its numerical solution with Bal's analytic solution in Ch. 4. The last chapter Ch. 5 concludes the results, discusses standing problems, and proposes new research directions.

Chapter 2

Chapter 2. Background Theory

We will first give context to the statistical physics used in spin transformers, namely spin glass models. Then we cover the vanilla transformer module, which is the standard building block of any transformer and is our access point into LLMs. Lastly, we draw Bal's proposed correspondence between the two resulting in the spin transformer module, along with his solution.

2.1 Spin Glasses

'Spin models' in this context refer to statistical physics models that describe systems of interacting parameters. By evaluating their interactions using a simple energy function or Hamiltonian, these models allow one to predict and describe complex macroscopic behaviors such as phase transitions. Originally, the Ising model was used to study magnets in which the interacting parameters are electron spins, hence the name spin-model. A simple Ising model takes the form

$$\mathcal{H} = -J \sum_{ij}^{N} s_i s_j - h \sum_{i}^{N} s_i \tag{2.1}$$

where $\{s_1, ..., s_N\}$ denote the electron spins, J the interaction strength, and h the applied field strength. These spin models have since been applied to similar complex interactive systems in many different disciplines, such as climate modeling [15], protein folding [16], and price volatility [17].

2.1.1 Sherrington-Kirkpatrick model

$$\mathcal{H} = -\sum_{i<j}^{N} J_{ij} s_i s_j \tag{2.2}$$

Spin glasses are a category of spin models, with a classic example being the Sherrington-Kirkpatrick (SK) model described above [18]. In this model, N spins denoted as $\{s_i\}$ with $i = 1, ..., N$, which can take a value of spin up or spin down $s_i \in \{-1, 1\}$. Between each pair of spins is a coupling strength defined by J_{ij} drawn from Gaussian distribution with zero mean and variance J^2/N. Given a specific instance of such a Hamiltonian, we can calculate the internal energy of any specific configuration or state $\mathbf{s_k} = \{s_1, ..., s_N\}$. Such a state lives in a state space $\mathbf{s_k} \in \Omega$ that contains all possible configurations $\Omega = \{\mathbf{s_1}, ..., \mathbf{s_M}\}$, which for this SK model has $M = 2^N$ possibilities.

When studying statistics, we prefer to describe a system using a distribution $P(\mathbf{s_k})$ of all possible configurations. This distribution is dependent on the control parameters of the system such as temperature and pressure, as this directly affects the interaction between spins. For a fixed set of control parameters, the distribution will evolve over time from some initial state until the distribution becomes stationary, indicating the system has reached thermal equilibrium. If the energy is conserved, the resulting distribution can be described using the Boltzmann Distribution:

$$P(\mathbf{s_k}) = \frac{e^{-\beta H(\mathbf{s_k})}}{Z} \tag{2.3}$$

where $\beta = 1/k_B T$ captures the inverse temperature and $Z = \sum_k^{\Omega} e^{-\beta H(\mathbf{s}_k)}$ describes the partition function. This partition function acts to normalize the weighted probabilities $e^{-\beta H}$ by summing over all configurations in the state space Ω. A core feature of this Boltzmann distribution is that higher energy states have an exponentially lower probability of occurring, and as temperature decreases this becomes more accentuated while for high temperatures this becomes less relevant.

Measuring the internal energy of this distribution is equivalent to finding its expected energy;

$$U = \langle H \rangle = \sum_{k}^{\Omega} P_k H_k = \frac{1}{Z} \sum_{k}^{\Omega} H_k e^{-\beta H_k} \tag{2.4}$$

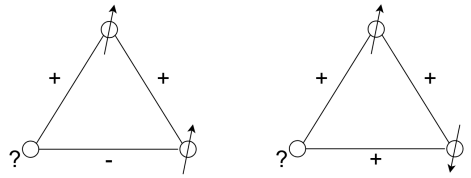

Figure 2.1: *Two sets of triangular spin systems, with arrows indicating the variable spin values, (+/-) indicating the fixed interaction type, and (?) indicating the chosen spin. (Left) Frustrated set of spins, unable to satisfy every connection regardless of the chosen spin. (Right) Not frustrated spins, as the chosen spin will influence the other spins to fully align.*

while the Gibbs entropy of the distribution defined as

$$\begin{aligned} S &= -k_B \sum_k^{\Omega} P_k \ln P_k = -k_B \sum_k^{\Omega} P_k \left(-\beta H_k - \ln Z\right) \\ &= k_B \beta \sum_k^{\Omega} P_k H_k + k_B \ln Z \sum_k^{\Omega} P_k \\ &= \frac{U}{T} + k_B \ln Z \end{aligned} \tag{2.5}$$

where we have derived the free energy $F = -k_B T \ln Z$ which can also be described in terms of internal energy and entropy as $F = U - TS$.

2.1.2 Frustration

The SK model extends upon the Ising model by incorporating fully connected spins $\sum_{i<j}$ and individual couplings J_{ij}, the latter which is characteristic of spin glasses. Strictly speaking, spin glasses require these individual couplings to be distributed over both positive and negative values such as with our Gaussian distribution centered around $\mu = 0$. Doing so will lead to collections of spins that are frustrated with one another, as depicted in Figure 2.1. 'Frustration' is the term given to spins that are unable to satisfy all the energy minimization constraints between their neighboring interactions; once the system cools down, these spins cannot favor certain interactions without frustrating other interactions.

Frustrated interactions are energetically unfavorable as they contribute to increasing $\mathcal{H}$ such as in Eq. (2.2). It makes sense that systems at low temperatures occupy states that are minimally frustrated. However, because

of the size and complexity of the spin system, there will be many equivalently frustrated configurations that have simply favored different sets of spins. These nearly energetically equivalent states often differ by large regions of spins, such that transitioning between two such states using single-spin flips requires one to pass through highly frustrated configurations.

This mechanism can be described by the system's free energy $F = U - TS$. At low T, the influence of entropy is minimal and the free energy is dominated by the internal energy. In this case, minimizing free energy corresponds to minimizing frustrations to find the least energetic ground states. The difficulty in transitioning between these ground states mean that the free energy as function of the parameter space is very rugged, with many minima separated by high walls.

On the other hand, high temperatures cause the entropy term to dominate the free energy. These high temperatures induce stochasticity in the spin parameters, making the system unable to stably hold any specific configuration. This means that the single optimal state starts to transition between very similar states, turning into a slightly broader and more energetic macrostate. On a free energy diagram, this looks like a gradual flattening of many sharp minima. Optimizing the free energy then refers to maximizing entropy rather than minimizing internal energy, which entails finding macrostates that maximize the number of encompassed unique states.

A simplified picture of the free energy of spin glasses can be found in figure 2.2 [19]. This figure also refers to the critical temperature T_c, signifying when the system undergoes a phase transition from ergodic to non-ergodic behavior. Ergodicity here refers to the system's ability to travel across states: an ergodic system features a smooth free energy landscape while a non-ergodic system features minima separated by high walls. That is, a system in a non-ergodic phase would only mutate between configurations within the same macrostate.

2.1.3 Phase behavior

The phase transition from stable states to unstable states in the SK model is depicted in figure 2.3 [19]. The two control parameters in this diagram are temperature T and the mean J_0 around which the couplings J_{ij} are distributed. In the figure, J refers to the variance of couplings $\sigma^2 = J^2/N$. When J_0 increases from 0, the distribution of couplings shift from being equally split negatively and positively to favoring positive couplings, de-

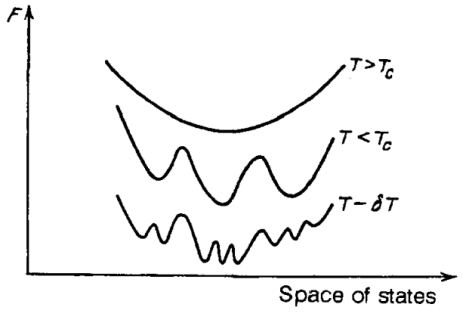

Figure 2.2: *Simple depiction of free energy landscape at different temperatures. T_c indicates critical temperature. Reproduced from de Almeida, Thouless (1978).*

creasing the amount of frustration in the model. For large offsets $J_0 > J$ and low temperatures $T < J_0/k_B$, the spins align with in the favored spin direction and the system thus behaves ferromagnetically. Increasing the temperature such that $k_B T > J_0$ will lead to a second order phase transition to a paramagnetic phase, in which the thermal fluctuations prevent the spins from fixing ferromagnetically.

The unstable region refers to the spin-glass phase, where the system's distribution of spins is non-ergodic. In this case, the thermal fluctuations from temperature aren't large enough $k_B T < J$ to overcome free energy barriers due to frustrated interactions, while the offset isn't large enough $J_0 < J$ to sufficiently imbalance the frustrations. This region is defined as unstable since its many ground states are configurationally disconnected, and non-ergodicity prevents transitions between such ground states. For the same set of parameters at low T, one would find that identical systems converge to different ground states randomly (induced by temperature).

Another control parameter that can be introduced to the system is an applied external field $h_{i \in N}$, which turns the Hamiltonian into:

$$\mathcal{H} = -\sum_{i<j}^{N} J_{ij} s_i s_j - \sum_{i}^{N} h_i s_i \tag{2.6}$$

This field can be incorporated into the stability condition, leading to a new field strength h vs T phase diagram depicted in Figure 2.4 [19]. The non-ergodic unstable phase is suppressed at high temperatures as they provide energy to cross the high free energy walls. This phase is also suppressed as field strength increases as its influence outweighs interactions.

The non-ergodic phase constitutes the spin glass phase, and the transition between unstable and stable phases marks the spin glass phase tran-

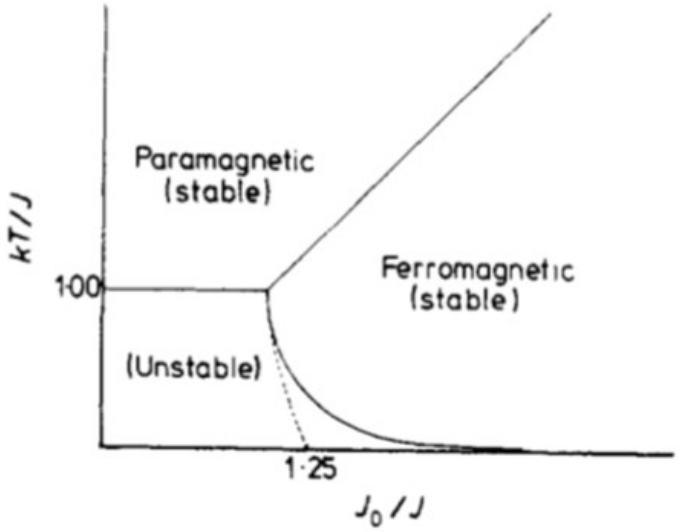

Figure 2.3: *Phase diagram showing stability of SK model solution in absence of a magnetic field. Reproduced from de Almeida, Thouless (1978).*

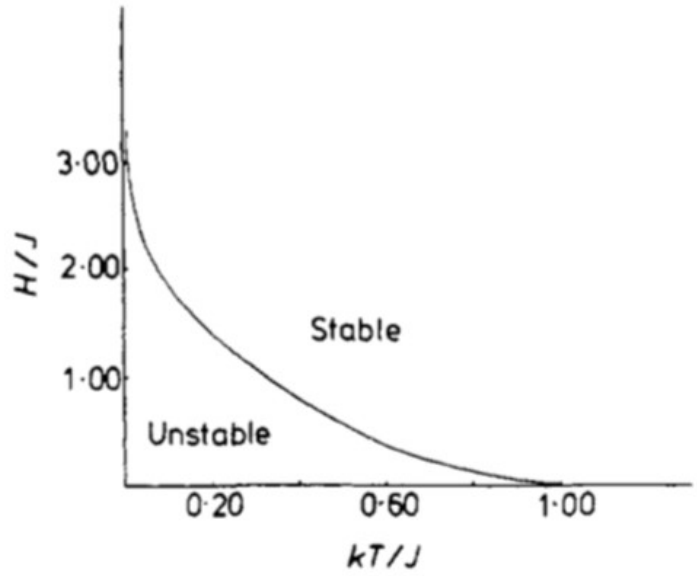

Figure 2.4: *Phase diagram for the SK model for control parameters external field H and temperature T. The graph shows the de Almeida-Thouless instability line, indicating the spin-glass phase transition below which ergodicity breaking occurs. Reproduced from de Almeida, Thouless (1978).*

sition. This line of transition is described by the de Almeida-Thouless inequality

$$\left(\frac{kT}{J}\right)^2 > \frac{1}{\sqrt{2\pi}} \int dz\, e^{-\frac{1}{2}z^2} \operatorname{sech}^4\left(\frac{Jz\sqrt{q}}{kT} + \frac{H}{kT}\right) \tag{2.7}$$

where q is the order parameter, H is the field strength, and $\frac{1}{\sqrt{2\pi}} \int dz\, e^{-\frac{1}{2}z^2}$ is a standard Guassian measure. The method works by inducing a small fluctuation, and seeing whether it gets suppressed (when LHS is greater) or grows. This inequality is essentially a stability analysis of the system's saddle point, where in the RS phase this point is a pure minimum in free energy while in the RSB phase it is a true saddle point [20].

2.1.4 Replica Symmetry Breaking

Quenched disorder refers to systems that feature constant couplings J_{ij} that feature the disordered spin glass phase. Replicas of such a system adopt the same quenched disorder J_{ij} values but develop independently. Due to stochasticity from temperature and transition probabilities, each replica will explore different states. In the ergodic phase, a single ground state exists such that all replicas converge around it. In the non-ergodic phase however, replicas will tend towards and become trapped in different frustrated configurations.

In the absence of a magnetic field, using normally distributed couplings J_{ij}, the mean magnetization $M = \frac{1}{N}\langle \sum_i \vec{s}_i \rangle$ of the system converges to zero for all temperatures. To then distinguish between the ergodic phase and the spin-glass phase, an alternative order parameter known as the *overlap* can be used, which measures the similarity between the configuration of replica a and that of replica b;

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N} \langle \vec{s}_i^{(a)} \cdot \vec{s}_i^{(b)} \rangle \tag{2.8}$$

For a large number of replicas, one can construct an overlap matrix which applies Eq. (2.8) to all pairs of replicas, as seen in Fig. 2.5. In the ergodic phase, all replicas converge to the same ground state and the overlap matrix is uniform. However, in the non-ergodic phase the replicas converge to differing ground states breaking the uniformity of the overlap matrix. Breaking this uniformity is equivalent to breaking the symmetry between replicas, such that ergodicity breaking is synonymous to replica symmetry breaking. The extent of asymmetry can be quantified as well,

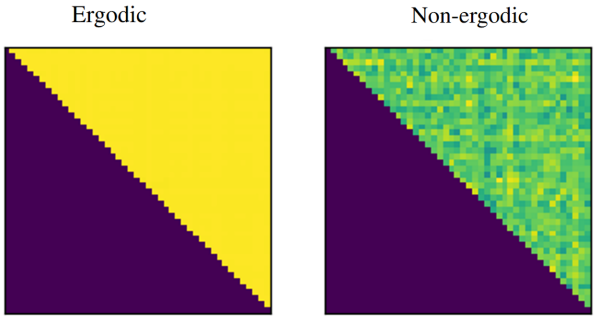

Figure 2.5: *Overlap matrices for a set of 50 replicas, showing replica symmetry phase (left) and full-replica symmetry breaking phase (right).*

where 1-RSB refers to having two quenched disorder ground states and full-RSB referring to every replica ground state being unique [21].

Replicas are also used to approximate the partition function in the RSB phase. Though quenched partition functions can be found for each replica in their respective ground state, the object of interest is the free energy $\overline{F} = -T\overline{\log Z}$ averaged over all replicas. The calculation for $\overline{\log Z}$ is challenging, but can be circumvented by using the replica trick:

$$\overline{\log Z} = \lim_{n \to 0} \frac{\overline{Z^n} - 1}{n} \tag{2.9}$$

with

$$\overline{Z^n} = \int \prod_{ab} \frac{dQ_{ab}}{2\pi} e^{N\mathcal{A}[Q_{ab}]} \tag{2.10}$$

for n replicas with overlap matrix Q_{ab}, which then allows one to estimate $\langle \log Z \rangle$ for free energy calculation.

2.2 Vanilla Transformer

A transformer module takes as input a sequence of tokens each embedded in a latent space. It outputs the same sequence of embedded tokens, but with the latent space incorporating how much each token attends to each other token by using an *attention mechanism*. The complete vanilla transformer architecture shown in Figure 2.6 uses 3 of these modules in an encoder-decoder architecture to enable training [2]. This transformer is a simple model that can be generalized to any sequence, whereas large language models (LLMs) are specifically designed to handle long sequences

Version of July 10, 2025– Created July 10, 2025 - 10:07

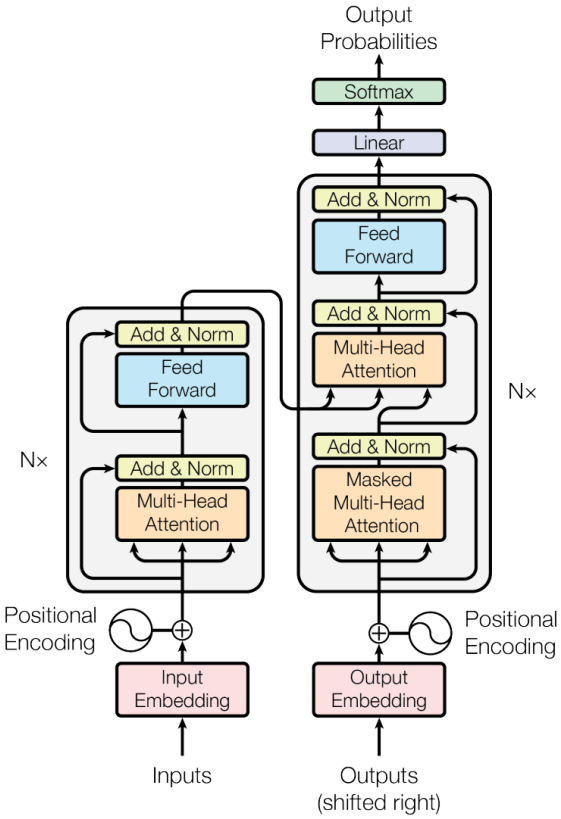

Figure 2.6: *The original transformer architecture, consisting of 3 transformer modules set up in an encoder-decoder manner. Each module features a Multi-Head Attention layer, accompanied by residuals and layer normalizations, and a final feed-forward layer to refine the output. Reproduced from Vaswani, et. al. (2017)*

.

of words. In each of these architectures, the transformer module stands central to their success [1].

Attention mechanism Let the input matrix $\mathbf{x} = \{x_1, ..., x_N\}$ with $x_{i \in N} \in \mathbb{R}^D$ denote a collection of N tokens from a sequence embedded in D latent dimensions. The scaled dot-product attention as shown in Eq. (2.11) uses query Q, key K, and value V matrices and mixes them through matrix multiplication. In the complete transformer architecture, self-attention and cross-attention are distinguished by where these Q, K, V weighted matrices are from. In this report we only use self-attention, thus describing each of these from the same input $\mathbf{x}$ but weighted separately, i.e. $Q =$

Version of July 10, 2025– Created July 10, 2025 - 10:07

$\mathbf{x}W_Q \in \mathbb{R}^{N\times D}$. The attention function starts by multiplying the queries Q and the keys K to form an $N \times N$ matrix. This matrix quantifies the similarity between every pair of tokens in the sequence, called their attention. It is processed through a softmax function $\xi(z_i) = e^{z_i} / \sum_j e^{z_j}$ such that for each token $x_i \in \mathbb{R}^D$, the hidden dimensions are distributed exponentially and normalized to sum up to 1. Finally, this is multiplied by the values V, applying the learned attention to each token x_i in $V = \mathbf{x}W_V$. This process can also be performed separately for different groups or 'heads' of dimensions in D before being recombined later.

$$\text{Attention}(Q, K, V) = \xi(\frac{QK^T}{\sqrt{D}})V \tag{2.11}$$

Full transformer equation In Einstein summation notation, we represent the input as x_i^α with $i = 1, ..., N$ and $\alpha = 1, ..., D$. The processing of the transformer module can be faithfully described as in Eq. (2.12). The multiheaded self attention is first calculated using weights W_Q, W_K, W_V projected along h 'heads'. A residual of the input is added and the matrix is layer-normalized. Layer normalization refers to normalizing each token $x_i \in \mathbb{R}^D$ across its hidden dimensions to have zero mean and fixed deviation. The attended output is then sent through a feed-forward layer for non-linearity [22], before once again adding a residual and layer normalizing. A deeper look at this equation and variations thereof is taken in Appendix A.1.

$$\begin{aligned} g(x_i^\alpha) &= LN\left[\sum_{h=1}^{H} \xi\left(x_i^\alpha W_\alpha^{(\gamma,h)} W^\beta_{(\gamma,h)} x_\beta^j\right) \cdot x_j^\nu W_\nu^{(\alpha,h)} + x_i^\alpha\right] \\ y_i^\alpha &= LN\left[FF(g(x_i^\alpha)) + g(x_i^\alpha)\right] \end{aligned} \tag{2.12}$$

Simple transformer equation For the sake of drawing a correspondence to spin glasses in the next section, we strip this equation to focus on its core mechanism. Layer normalization serves to prevent exploding and vanishing gradients, which is characteristic of any learning model. This is implemented in our simulations and can be ignored in the equation. According to Bal, the nonlinearity of the feed-forward layer can be captured by computing second-order fluctuations in the spin glass [12]. We therefore consider the feed-forward layer to be a correction term, which we ignore for our linear comparison. Lastly, we consider only single headed attention, as the multiple heads primarily contribute to training stability

Version of July 10, 2025– Created July 10, 2025 - 10:07

[23]. This leaves us with a simplified function for the forward pass of a transformer module consisting of pure attention and a residual, taking as input x and parameters $W_{Q,K,V} = \{W_Q, W_K, W_V\}$:

$$f_{attention}(x, W_{Q,K,V}) = x + \xi\left(xW_Q W_K^T x^T\right) \cdot xW_V \tag{2.13}$$

2.3 Spin glass - Transformer correspondence

On one hand, we have an equilibrating Sherrington-Kirkpatrick spin glass model. It minimizes the Hamiltonian described in Eq. (2.6) by maximizing the alignment of the spins s_i with both the field x_i and every other spin s_j. On the other hand, we have the simplified transformer forward pass described by Eq. (2.13). During training, its output $f(x, W_{Q,K,V})$ is compared to desired output y through the loss function $\mathcal{L} = \left[\, y - f(x, W_{Q,K,V})\right]^2$. The cross terms hereof are negative dot products between them, such that minimizing the loss function revolves around maximizing the alignment of y with $f(x, W_{Q,K,V})$. They are therefore both effectively alignment machines.

More importantly, they use similar alignment constraints as shown in Table 2.1. Both external fields in spin glasses and residuals in transformers act as an applied bias, forcing the system to remember this bias across many iterations of their forward pass. The token-to-token relation takes the form of an $N \times N$ matrix, quantifying the relation between every pair of tokens whether in the form of couplings or self-attention. The readout describes the $N \times D$ token embedding that the relation matrix is applied to, their product which incorporates the token-token relations back into the latent space.

Spin Glass		Transformer
$s_i \Leftrightarrow x_i + J_{ij}s_j$	Alignment task	$y_i \Leftrightarrow x_i + \xi\left(x_i W_Q W_K^T x_j^T\right) \cdot x_j W_V$
x_i	applied bias	x_i
J_{ij}	token-token relation	$\xi\left(x_i W_Q W_K^T x_j^T\right)$
s_j	readout	$x_j W_V$

Table 2.1: *Summarized correspondence between generic spin glass model from Section 2.1.3 and the simplified transformer forward pass from Section 2.2.*

Version of July 10, 2025– Created July 10, 2025 - 10:07

Vector spins To complete the correspondence, we can tweak the SK spin glass model to better resemble this simplified transformer model. This involves extending the spins in the spin glass from their binary description to a vector description $s_i \in \{-1, 1\} \Rightarrow \vec{s_i} \in \mathbb{R}^D$, such that the entire system of N spins at time t is described as $\mathbf{s_t} \in \mathbb{R}^{N \times D}$. This thereby matches the embedded-token input of the transformer, where each word (corresponding to spin) is represented as a vector in $\mathbb{R}^D$. This high-dimensional vector spin glass has not yet been comprehensively studied. The Hamiltonian describing a vector spin glass corresponding to a transformer module is thus given by:

$$H = -\sum_{ij} J_{ij}\vec{s}_i\vec{s}_j - \sum_i \vec{x}_i\vec{s}_i = -\mathrm{Tr}\left[\mathbf{s}(\mathbf{x} + J_{ij}\mathbf{s})^T\right] \tag{2.14}$$

Equilibrium depth The remaining difference these two models is that the spin glass remains as an equilibrating system that iterates repeatedly until the alignment is accomplished, while the transformer performs a single-layer direct calculation. In LLMs, these transformer layers are stacked to some finite depth L with each layer having its own set of weights. Under the same light, the spin transformer can be seen as having an infinite depth $L \to \infty$ with a consistent set of weights. This equilibrium framework brings the transformer closer to existing statistical physics models, allowing for better correspondence between the two. This in turn makes the physics of the model more accessible, therefore making the spin-transformer a good starting point for studying the statistical physics of transformers. However, despite the fact that the two models correspond well to one another, we make a point that results found for this equilibrium-based transformer may not apply directly to practical transformers.

Regardless, studying the physics of these equilibrium transformers does have potential in usable language models. Recently developed Deep Equilibrium (DEQ) models [13] show that deep neural networks, including deep transformers, can replace their stack of L layers each with distinct parameters with a single repeating equilibrium layer using the same parameters. They show that similar accuracy to deep networks is attained with significant decreases in memory usage. This implies that besides using spin-transformers for studying the statistical physics of transformers, they also have practical potential within the framework of DEQ models.

Asymmetric Couplings Another important realization in Bal's work is the asymmetric nature of the couplings $J_{ij} \neq J_{ji}$. This is inherent in trans-

formers, where the attention paid by one token to the next is also asymmetric. For example, predicting the word 'thesis' after 'my' is more likely than predicting 'my' after 'thesis'. These asymmetric couplings lead to new, interesting dynamics, which can be captured by using a 'kinetic' Ising model framework [24]. This kinetic framework involves using the readout state s_j from the previous time step, i.e. $H = -\sum_{ij} J_{ij} s_{i,t} s_{j,t-1}$.

2.4 Spin-Transformer solution

Vector Spin Glass model The blog posts provide mean field solutions to the vector-spin glass model. For a system of N interacting vector spins $\vec{s}_{i,t} = \{\vec{s}^1_{i,t}, ..., \vec{s}^D_{i,t}\} \in \mathbb{R}^D$ in D dimensions that evolve in discrete time t, he adapts the kinetic Ising model's dynamics in which all spins get updated at the same time in parallel. The system state at time $t-1$ is described as $\mathbf{s}_{t-1} = \{\vec{s}_{1,t-1}, ..., \vec{s}_{N,t-1}\} \in \mathbb{R}^{N \times D}$. The probability distribution of states at time t is then described by the Markov-chain transition probability:

$$\begin{aligned} P(\mathbf{s}_t|\mathbf{s}_{t-1}) &= \prod_{i=1}^{N} \frac{e^{\beta \vec{s}_{i,t} \cdot \vec{h}_{i,t}}}{\int_{S_{D-1}} d^D \vec{s}_{i,t}\, e^{\beta \vec{s}_{i,t} \cdot \vec{h}_{i,t}}}, \\ \vec{h}_{i,t} &= \vec{x}_{i,t} + \sum_{j=1}^{N} J_{ij} \vec{s}_{j,t-1}, \end{aligned} \tag{2.15}$$

where $\vec{h}_{i,t}$ describes the effective external field, β the inverse temperature, and $S_{D-1}(R) = \{z \in \mathbb{R}^D : ||z|| = R\}$ the latent space $(D-1)$-dimensional hypersphere with radius R. The mean magnetization of the system at time t is then

$$\vec{m}_{i,t} = \sum_{\mathbf{s}_t} \vec{s}_{i,t} P(\mathbf{s}_t) = \sum_{\mathbf{s}_t} \vec{s}_{i,t} \sum_{\mathbf{s}_{t-1}} P(\mathbf{s}_t|\mathbf{s}_{t-1}) ... \sum_{\mathbf{s}_0} P(\mathbf{s}_1|\mathbf{s}_0) P(\mathbf{s}_0), \tag{2.16}$$

which involve the transition probabilities applied iteratively from the initial system distribution. This involves summing over exponentially many spin systems as time passes, making the direct calculation unfeasible. To approximate this mean magnetization $\mathbf{m}_t = \{\vec{m}_{1,t}, ..., \vec{m}_{N,t}\} \in \mathbb{R}^{N \times D}$, Bal uses Plefka expansions. Note that $\mathbf{s}_t$ indicate specific configurations, while $\mathbf{m}_t$ indicate the mean magnetization of probability distribution $P(\mathbf{s}_t)$. This method is effective for handling asymmetric couplings and can extract interactions up to a desired order, including the naive mean-field (nMF) solution and the more accurate Thouless-Anderson-Palmer (TAP) solution [14].

Version of July 10, 2025– Created July 10, 2025 - 10:07

Plefka expansions The Plefka expansions were originally derived using power series expansion of the free energy around a non-interacting spin model [25], and were later formalized in an information geometric perspective [14]. This non-interacting spin model replaces the effective external field $\mathbf{h}_t$ from Eq. (2.18) with independent spins $\boldsymbol{\theta}_t \in \mathbb{R}^{N\times D}$, such that its Markov-chain transition probability is;

$$Q(\mathbf{s}_t|\boldsymbol{\theta}_t) = \prod_{i=1}^{N} \frac{\mathrm{e}^{\beta\,\vec{s}_{i,t}\cdot\vec{\theta}_{i,t}}}{\int_{S_{D-1}} \mathrm{d}^D\vec{s}_{i,t}\; \mathrm{e}^{\beta\,\vec{s}_{i,t}\cdot\vec{\theta}_{i,t}}} \tag{2.17}$$

The method then introduces an interaction strength parameter α, which acts as an interpolator between the fully interacting model and the non-interacting model. The encompassing transition probability is then described as:

$$\begin{aligned} P_\alpha(\mathbf{s}_t|\mathbf{s}_{t-1}) &= \prod_{i=1}^{N} \frac{\mathrm{e}^{\beta\,\vec{s}_{i,t}\cdot\vec{h}_{i,t}(\alpha)}}{\int_{S_{D-1}} \mathrm{d}^D\vec{s}_{i,t}\; \mathrm{e}^{\beta\,\vec{s}_{i,t}\cdot\vec{h}_{i,t}(\alpha)}} \\ \vec{h}_{i,t}(\alpha) &= (1-\alpha)\vec{\theta}_{i,t} + \alpha\left(\vec{x}_{i,t} + \sum_{j=1}^{N} J_{ij}\vec{s}_{j,t-1}\right) \end{aligned} \tag{2.18}$$

Such that $P_{\alpha=0}(\mathbf{s}_t|\mathbf{s}_{t-1}) = Q(\mathbf{s}_t|\boldsymbol{\theta}_t)$ while $P_{\alpha=1}(\mathbf{s}_t|\mathbf{s}_{t-1}) = P(\mathbf{s}_t|\mathbf{s}_{t-1})$. For expectation values hereof, such as mean magnetization in Eq. (2.16), the Plefka expansions are defined as Taylor series expansions of this function around $\alpha = 0$. This allows us to expand the interaction strength of the mean-field to the n^{th} order:

$$\mathbf{m}_t(\alpha) = \mathbf{m}_t(\alpha=0) + \sum_{k=1}^{n} \frac{\alpha^k}{k!} \frac{\partial^k \mathbf{m}_t(\alpha=0)}{\partial\alpha^k} + \mathcal{O}(\alpha^{n+1}) \tag{2.19}$$

Finding the optimal parameters at which $\mathbf{m}_t(\alpha=1) = \mathbf{m}_t(\alpha=0)$ then revolves around setting the expansion term equal to 0 and solving for $\boldsymbol{\theta}_t^*$ that thereby approximates the mean-field values. To evaluate the expectation values, we construct a marginal distribution $P_\alpha(\mathbf{s}_t)$ up to a choice of $\mathbf{s}_{t-\tau}$, and assume that derivatives further back than $t-\tau$ are independent of the expected magnetization $\int d\mathbf{s}_t \vec{s}_{i,t} P_\alpha(\mathbf{s}_t|\mathbf{s}_{t-1})$.

This expansion around interaction strength α allows the TAP solution to consider fluctuations due to higher-order interactions. In the ferromagnetic Ising model, the nMF solution is sufficient as $J_{ij} \sim 1/N$, for which the second-order interactions are trivial. In this spin glass model, coupling strength is of order $1/\sqrt{N}$, making the second-order fluctuations relevant

Version of July 10, 2025– Created July 10, 2025 - 10:07

and the TAP solutions better. The optimal parameters $\boldsymbol{\theta}_t^*$ in the first order nMF expansion is simply the linear effective field $\vec{h}_{i,t}$ as seen in Eq. (2.18). In the TAP solution, the optimal parameters add onto the nMF solution with second order terms as calculated by Bal:

$$\vec{\theta}_{i,t} = \vec{x}_{i,t} + \sum_j J_{ij}\vec{m}_{j,t-1} + \frac{1+\gamma_{i,t}(0)}{2\beta}\left(\frac{\partial^2\vec{m}_{i,t}(\alpha=0)}{\partial\alpha^2} + \frac{\vec{m}_{i,t}\cdot\frac{\partial^2\vec{m}_{i,t}(\alpha=0)}{\partial\alpha^2}}{\frac{R^2\gamma_{i,t}(0)}{1+\gamma_{i,t}(0)} - \vec{m}_{i,t}^2}\vec{m}_{i,t}\right) \tag{2.20}$$

with $\gamma_{i,t}$ stemming from the mean magnetization readout described hereafter.

Mean magnetization The Plefka expansions find the optimal parameters $\boldsymbol{\theta}_t^*$ satisfying Eq. (2.19). The readout function of these parameters, equivalent to *tanh* in the ferromagnetic mean field equations, has to be adapted to the new partition function covering the S_{D-1} hypersphere. Bal uses modified Bessel function recurrence relations to arrive at the mean magnetization per spin of:

$$\vec{m}_{i,t} = \frac{I_{D/2}(\beta R\|\vec{\theta}_{i,t}\|)}{I_{D/2-1}(\beta R\|\vec{\theta}_{i,t}\|)}\frac{R\vec{\theta}_{i,t}}{\|\vec{\theta}_{i,t}\|} \equiv \boldsymbol{\varphi}\left(\vec{\theta}_{i,t}\right), \tag{2.21}$$

He then uses algorithms for efficient computation of the Bessel functions which enforces the radius $R = \sqrt{D/2-1}$ and takes the large dimension limit $D \to \infty$ to allow for the following mean magnetization per spin given parameters $\boldsymbol{\theta}_t$;

$$\vec{m}_{i,t}^{D\to\infty} \approx \frac{\beta}{1+\gamma(\|\vec{\theta}_{i,t}\|)}\vec{\theta}_{i,t} \equiv \boldsymbol{\varphi}^{D\to\infty}\left(\vec{\theta}_{i,t}\right). \tag{2.22}$$

where the γ function is described as

$$\gamma_{i,t}(\alpha) \equiv \gamma\left(\|\vec{h}_{i,t}(\alpha)\|\right) = \sqrt{1+\beta^2\|\vec{h}_{i,t}(\alpha)\|^2/R^2} \tag{2.23}$$

Note that as $\beta \to 0$, $\mathbf{m}_{i,t}^{D\to\infty} \to 0$, meaning that in the high temperature limit the mean magnetization vanishes.

Version of July 10, 2025– Created July 10, 2025 - 10:07

Fixed point equation The readout for mean magnetization as a function of the optimal parameters lead to fixed-point equations. These can be used iteratively until the fixed point is reached, indicating the optimal approximation for $\mathbf{m}_t$. For example, the nMF solution would be computed by starting from some $\mathbf{m}_0$ and repeating the fixed point equation in Eq. (2.24) until convergence.

$$\vec{m}_{i,t}^{nMF} = \frac{\beta}{1+\sqrt{1+\beta^2||\vec{x}_{i,t}+\sum_j J_{ij}\vec{m}_{i,t-1}||^2/R^2}}\left(\vec{x}_{i,t}+\sum_j J_{ij}\vec{m}_{i,t-1}\right) \tag{2.24}$$

The TAP fixed point equation is slightly more complicated, due to the dependence of $\boldsymbol{\theta}_t^*$ on $\mathbf{m}_t$. The author circumvents this apparent bi-level optimization problem by inverting Eq. (2.22) and plugging that into the second-order term in Eq. (2.20). Then the TAP fixed point equation is similar to the nMF fixed point equation, but includes the second-order term in the parameters.

Asymmetry and NESS The final point to which the analytic solutions converge is the equilibrium state. However, this equilibrium state is nuanced due to the inclusion of asymmetric couplings. They directly lead to asymmetry in the transition probability between two states; $P(\mathbf{s}_b|\mathbf{s}_a) \neq P(\mathbf{s}_a|\mathbf{s}_b)$, preventing the system from reaching a stationary distribution. The asymmetric components of J_{ij} thus prevent time-reversal symmetry, leading to positive entropy production [26]. The entropy of the system is a macroscopic observable, and having it change over time means the system is in nonequilibrium. The analytic solutions used by Bal were designed to account for asymmetric couplings, and converge to a near-equilibrium steady state (NESS) that lives in a cycle of near-equilibrium stationary distributions. The analytic solutions approximate an average of these cycling NESS states. For example, using the analytic fixed-point update for nMF in Eq. (2.24), the NESS magnetization for nMF is $\mathbf{m}_{NESS}^{nMF} = \lim_{t\to\infty} \mathbf{m}_t^{nMF}$.

Vector-spin glass & spin transformer In the blog, Bal makes a distinction between this vector-spin glass model and the spin transformer (ST) model, as for the latter he incorporates more transformer-based structures (i.e. field-induced attention, multi-heads) and compiles them in an encoder-decoder structure to form a complete ST. However, this paper does not delve into Bal's complete ST and only addresses the vector-spin glass that corresponds to a transformer module. To maintain the connection between LLMs and spin glasses we refer to this spin glass as a ST, such that

the terms 'ST model' and 'vector-spin glass model' are interchangeable throughout this paper.

Chapter 3

Method

We can verify the analytic solutions by comparing them to the Markov Chain Monte Carlo (MCMC) computational solutions. These Monte Carlo methods are powerful because they can not only accurately converge to a desired probability distribution, but they can do so without access to its partition function [27]. Specifically, we will use the Metropolis-Hastings (MH) algorithm and use a loss-based Hamiltonian to simulate the equilibrium distributions of the spin glasses. We start by describing the general approach to the MCMC-MH algorithm before detailing our MCMC adaptation.

3.1 MH Formulation

Detailed balance MCMC allows for sampling from complex probability distributions, from which direct sampling is unfeasible. Markov chains are sequences of sampled states that forget their history and are based on conditional probabilities that sample using only the previous state, i.e. $P_t(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{s}_{t-2}, ..., \mathbf{s}_0) = P_t(\mathbf{s}_t|\mathbf{s}_{t-1})$.. Considering two states $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{N\times D}$, we can define a transition probability density describing the chance of going from state $\mathbf{x}$ to $\mathbf{y}$ as $P(\mathbf{y}|\mathbf{x})$ such that $\int_{S_{D-1}} d^D\mathbf{y}\, P(\mathbf{y}|\mathbf{x}) = 1$ and marginal probability density $P_t(\mathbf{y}) = \int_{S_{D-1}} d^D\mathbf{x}\, P_{t-1}(\mathbf{x})P(\mathbf{y}|\mathbf{x})$. The goal is to reach a stationary distribution, defined as

$$\pi(\mathbf{y}) = \int_{S_{D-1}} d^D\mathbf{x}\, \pi(\mathbf{x})P(\mathbf{y}|\mathbf{x}) \tag{3.1}$$

We can design our Markov chain to approach this stationary distribution by enforcing the detailed balance condition (DBC) $\pi(\mathbf{x})P(\mathbf{y}|\mathbf{x}) =$

$\pi(\mathbf{y})P(\mathbf{x}|\mathbf{y})$, which defines a balance in the flow of states between all states in state space $\mathbf{x}, \mathbf{y} \in S_{D-1}$.

Metropolis Hastings The MH algorithm allows one to construct Markov chains that approach the above DBC. It splits the transition probability into a proposition probability and acceptance probability $P(\mathbf{y}|\mathbf{x}) = Q(\mathbf{y}|\mathbf{x})A(\mathbf{y}|\mathbf{x})$. The proposition distribution is chosen and fixed throughout the sampling, while the acceptance probability is optimized to approach the stationary distribution. Plugging the MH transition probability into DBC, we can rearrange the terms to arrive at the MH acceptance probability:

$$A(\mathbf{y}|\mathbf{x}) = \min\left(1, \frac{\pi(\mathbf{y})Q(\mathbf{x}|\mathbf{y})}{\pi(\mathbf{x})Q(\mathbf{y}|\mathbf{x})}\right) \tag{3.2}$$

The stationary distribution is commonly based on the Boltzmann distribution that distributes the probabilities of states at equilibrium based on their energy $\pi(\mathbf{y}) = \frac{1}{Z}e^{-\beta H(\mathbf{y})}$. The ratio of the two stationary distributions thus allow the equivalent partition functions to cancel out, such that the MH method does not require calculation of Z. Propositions are usually chosen such that $\frac{Q(\mathbf{x}|\mathbf{y})}{Q(\mathbf{y}|\mathbf{x})} = 1$.

Aperiodicity and Ergodicity To allow for proper sampling, the proposition probabilities must be chosen such that they are aperiodic and ergodic. Aperiodicity prevents the sampling from being in a periodic cycle, which never converges to a static distribution. It can easily be ensured by having $P(\mathbf{x}|\mathbf{x}) > 0$. Ergodicity implies that every state in state space must eventually be reachable from any other state, i.e. $P(\mathbf{s}_{t+\tau} = \mathbf{x}|\mathbf{s}_t = \mathbf{y}) > 0 \; \forall \; \mathbf{x}, \mathbf{y}$ at some time separation τ. This ensures that for any observable $f : S_{D-1} \to \mathbb{R}$, the sample mean $\overline{f(\mathbf{s}_n)} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{s}_i)$ converges to $\langle f(\mathbf{s})\rangle$. Satisfying both these requirements ensures that the sampled distribution converges to the stationary distribution $\lim_{t\to\infty} P(\mathbf{s}_t = \mathbf{x}) = \pi(\mathbf{x})$ irrespective of $\mathbf{s}_0$.

3.2 MCMC algorithm

We now look at our specific choice of probabilities used for the MH algorithm.

Effective Stationary distribution Initially, we used the spin transformer Hamiltonian described in Eq.(2.14) plugged into the Boltzmann distribution for our stationary distribution. However, larger vectors directly led

to greater alignment and thus lower energy, such that the implementation focused more on enlarging vectors and less on aligning them. Even when including normalization constants, at high temperatures the system continued to explode in size.

Thus, we tapped into the known correspondence between Hamiltonians and Loss functions [28]. The Hamiltonian in Eq.(2.14) can be recognized as the cross terms in the loss function:

$$\begin{aligned} H &= \frac{1}{2}\sum_i \left[\vec{s}_{i,t} - \omega_{i,t-1}\left(\vec{x}_i + \sum_j J_{ij}\vec{s}_{j,t-1}\right)\right]^2 \\ &= \frac{1}{2}\sum_i \left[||\vec{s}_{i,t}||^2 - 2\omega_{i,t-1}\vec{s}_{i,t}\left(\vec{x}_i + \sum_j J_{ij}\vec{s}_{j,t-1}\right)^T + R^2\right] \end{aligned} \tag{3.3}$$

where $\omega_{i,t} = R/||\vec{x}_i + J_{ij}\vec{s}_{j,t}||$ normalizes the constraints and $||\vec{s}_{i,t}|| = \sqrt{\sum_\alpha (s_i^\alpha)^2}$ is the vector norm. At this point, for the purpose of our simulations which we will come back to in Sec. 5, we leave behind the kinetic Ising formulation of $H(\mathbf{s_t}|\mathbf{s}_{t-1})$ and define the Hamiltonian in terms of only $H(\mathbf{s}_t)$.

The stationary distribution is thus described as

$$\begin{aligned} \pi(\mathbf{s}_y) &= \frac{1}{Z}\exp\left(-\beta\sum_i^N \left[\vec{s}_{i,y} - \omega_{i,y}\left(\vec{x}_i + \sum_j J_{ij}\vec{s}_{j,y}\right)\right]^2\right) \\ &= \frac{1}{Z}\exp\left(-\beta\ \mathrm{Tr}\left[\mathbf{s}_y \times \Omega_y(\mathbf{x} + \mathbf{J}\mathbf{s}_y)^T\right]\right) \end{aligned} \tag{3.4}$$

where $\Omega_y(\cdot)$ denotes normalizing each attentioned vector $(\vec{x}_i + \sum_j J_{ij}\vec{s}_{j,y})$ by $\omega_{i,y}$.

Effective Proposition The proposition distribution is fixed throughout our sampling and is based on our choice. It first picks the spin-vector to be changed, which is uniformly picked from the N spin-vectors. Then we choose how to propose a new vector, remembering that this proposition has to fulfill both the aperiodicity and ergodicity requirement.

Because our vectors live in $\mathbb{R}^D$, a global proposition drawn as $\vec{x} \sim \mathcal{N}(0, R_D) \in \mathbb{R}^D$ has an exponentially small chance in D to land on the perfect vector, and thus would take impractically long to converge for all N spins. We therefore use local propositions, for which the proposed vector is a slight perturbation across all dimensions of the original vector

$\vec{y}_i = \vec{x}_i + \mathcal{N}(0, \sigma^2)$. Specifically, we used $\sigma = \frac{1}{N\sqrt{D}}$ to keep the perturbations in the range of their vector components. The proposition distribution we used is thus:

$$Q(\vec{y}_i|\vec{x}_i) = \frac{1}{N\sqrt{2\pi D}} \exp\left[-\frac{1}{2}\left(\frac{\vec{y}_i - \vec{x}_i}{N\sqrt{D}}\right)^2\right] \tag{3.5}$$

which satisfies the aperiodic requirement, and though the probability of reaching any other state is very small it is never non-zero, such that it also satisfies ergodicity. This proposition also satisfies $\frac{Q(\mathbf{x}|\mathbf{y})}{Q(\mathbf{y}|\mathbf{x})} = 1$.

Effective Acceptance probability The acceptance probability for our MH algorithm, using the above described stationary and proposition distributions, therefore becomes:

$$\begin{aligned} A(\mathbf{x} \to \mathbf{y}) &= \min\left(1, \frac{\pi(\mathbf{y})Q(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})Q(\mathbf{y},\mathbf{x})}\right) \\ &= \min\left(1, \exp\left[-\beta\left(\mathrm{Tr}[\mathbf{s}_y \times \Omega_y(\mathbf{x} + \mathbf{J}\mathbf{s}_y)^T] - \mathrm{Tr}\left[\mathbf{s}_x \times \Omega_x(\mathbf{x} + \mathbf{J}\mathbf{s}_x)^T\right]\right)\right]\right) \end{aligned}$$

Alignment As explained in Section 4.2, an appropriate order parameter is the replica overlap q_{ab} [29]. In the blog posts, Bal adopts a similar but altered order parameter he calls *cosine similarity*, stemming from the relationship between cosine and dot products $\cos(\theta) = \vec{v} \cdot \vec{u}/(|\vec{v}||\vec{u}|)$. The cosine similarity between two states $\mathbf{s}_x, \mathbf{s}_y \in \mathbb{R}^{N\times D}$ is defined as;

$$\mathcal{A}(\mathbf{s}_x, \mathbf{s}_y) \equiv \frac{1}{N}\sum_i^N \frac{\vec{s}_{i,x} \cdot \vec{s}_{i,y}}{|\vec{s}_{i,x}||\vec{s}_{i,y}|} \tag{3.6}$$

which is essentially the average scaled dot product between the two matrices; tending to 1 when the two states are similar, and to 0 when the two states are dissimilar. This cosine similarity essentially measures the alignment between states $\mathbf{s}_x$ and $\mathbf{s}_y$. For example, Bal plots 3 alignments as function of time; $\mathcal{A}(\mathbf{s}_t|\mathbf{s}_0)$ to track similarity to the initial configuration, $\mathcal{A}(\mathbf{s}_t|\mathbf{s}_{t-1})$ to track convergence, and $\mathcal{A}(\mathbf{s}_t|\mathbf{x})$ to see its relation with the external field.

Version of July 10, 2025– Created July 10, 2025 - 10:07

Chapter 4

Results

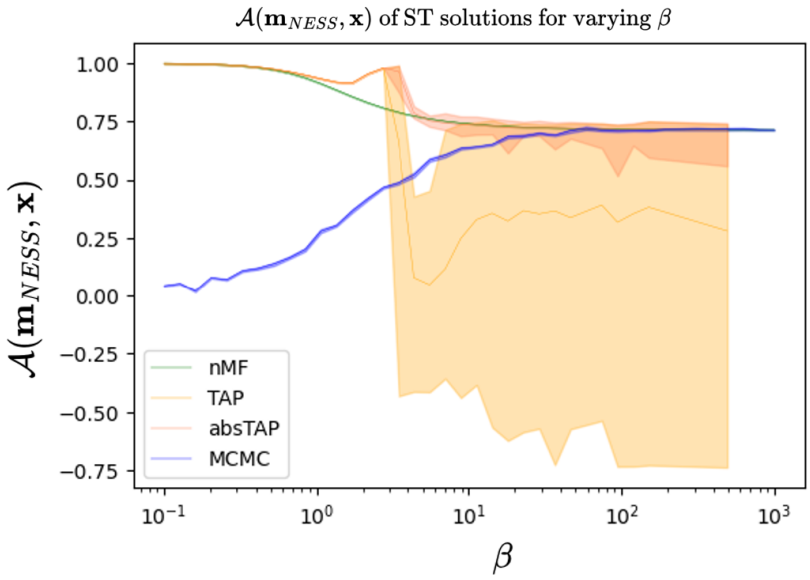

Figure 4.1: *Alignment $\mathcal{A}$ (Eq. 3.6) of equilibrium state $\mathbf{m}_{NESS}$ with applied field $\mathbf{x}$ at varying inverse temperatures β. Absolute alignment of TAP solutions $|\mathcal{A}|$ are shown as absTAP (dark orange). At high temperatures, our numerical MCMC solution becomes paramagnetic with random alignment $\mathcal{A} \to 0$ to the field while analytic nMF & TAP solutions align completely with applied field $\mathcal{A} \to 1$. At low temperatures, TAP becomes unstable while the first-order nMF and MCMC solutions converge to the ground state.*

4.1 MCMC vs TAP

We are only interested in the final $\mathbf{m}_{NESS}$ rather than the exploration towards it. We thus extract the MCMC NESS by sampling many states at convergence and taking $\mathbf{m}_{NESS}^{\text{MCMC}} = \frac{1}{\tau}\sum_t^{\tau} \mathbf{s}_t$ for large τ. We then measure the alignment $\mathcal{A}(\mathbf{m}_{NESS}, \mathbf{x})$ comparing NESS to the random field $\mathbf{x}$ rather than attention $f_{attn}(\mathbf{s_t})$ for two reasons; 1) to show what the simulations are converging to relative to one another (as opposed to their absolute convergence) and 2) to show an interesting interaction between the analytic solutions and the magnetic field.

This alignment of $\mathbf{m}_{NESS}$ with $\mathbf{x}$ is shown for varying inverse temperatures for MCMC, nMF, TAP, and absTAP in Fig. (4.1). We included absTAP $= |\mathcal{A}(\mathbf{m}^{TAP}, \mathbf{x})|$ as we noticed that although $\mathbf{m}^{TAP}$ fluctuates highly in the low temperature regime, much of this fluctuation comes from spin-flips that turn $\mathcal{A}(\mathbf{m}^{TAP}, \mathbf{x})$ negative. If we ignore spin flips and consider them to be aligned, we find that absTAP stays in the neighborhood of alignment with MCMC, though still with considerable variance.

We notice that at high temperatures, MCMC becomes random and paramagnetic while analytic solutions align with the magnetic field. At low temperatures all the solutions converge, with TAP exhibiting considerable fluctuations. In the intermediate temperature regime, all three solutions behave differently. We continue to explain these behaviors in these 3 separate temperature regimes.

High T regime For $\beta < 1$, we see that MCMC becomes paramagnetic while TAP converges with the external field. The analytic solution is contradictory, as one would expect that at extreme T the solution would always be paramagnetic, especially considering this is the regime where TAP excels.

The reason for this alignment with the magnetic field comes down to the calculation for mean magnetization in Eq.(2.22). In the $\beta \to 0$ limit, this expected magnetization goes to 0 which agrees with what one would expect in the paramagnetic phase, and brings the magnitudes of the spin-vectors in the analytic solutions close to 0. The limit of the readout in Eq. (2.22) is;

$$\lim_{\beta\to 0} \frac{\beta}{1+\sqrt{1+\beta^2||\theta||^2/R^2}}\theta = \left[\frac{1}{2}\beta - \frac{1}{8}\beta^2 + ...\right]\theta \approx \frac{1}{2}\beta\theta \tag{4.1}$$

Then in the nMF case using $\theta = (\mathbf{x} + \mathbf{J}\mathbf{m}_{t-1})$ we have

$$\mathbf{m}_t = \frac{1}{2}\beta(\mathbf{x} + \mathbf{J}\mathbf{m}_{t-1}) = \frac{1}{2}\beta\mathbf{x} + \frac{1}{2}\beta\mathbf{J}\mathbf{m}_{t-1} \tag{4.2}$$

Version of July 10, 2025– Created July 10, 2025 - 10:07

As this converges, the magnetization $\lim_{\beta\to 0} \mathbf{m}_{t-1} \to 0$. Therefore, the term $\frac{1}{2}\beta\mathbf{x}$ goes to 0 an order slower than the term $\frac{1}{2}\beta\mathbf{J}\mathbf{m}_{t-1}$ and we see that $\lim_{\beta\to 0} \mathbf{m}_{NESS} \approx \lim_{\beta\to 0}(\frac{1}{2}\beta\mathbf{x})$, which still reduces the vectors to 0 but aligns them with the external field. The TAP solutions have an additional term in the effective field parameters θ, but these similarly reduce to 0 at higher orders than the magnetic field. This approximation is not incorrect, but effectively implies an induced magnetic field that scales with increased temperature.

Intermediate T regime For $1 < \beta < 10$, each of the solutions behaves differently. However, knowing the behavior of analytic solutions at high T, we can see that MCMC and nMF similarly transition smoothly and directly from $\lim_{\beta\to\infty} \mathbf{m}_{NESS}$ to their respective $\lim_{\beta\to 0} \mathbf{m}_{NESS}$. The convergence of TAP however shows a small 'bump', which is entirely induced by the second-order Onsager term. This term captures fluctuations emerging from a spin's effect on itself through interactions around it, reflecting a better approximation than nMF. Furthermore, according to Aguilera [14], this term enables TAP to handle asymmetric couplings and thereby nonequilibrium behavior. Hypothesizing that this bump stems from these nonequilibrium dynamics, we simulated TAP with symmetric $J_{ij} = J_{ji}$ to see whether the bump persists. As shown in Appendix B.3, this bump persists, meaning that this bump doesn't form due to asymmetry but already exists without it. The bump thus likely corresponds to pure self-induced fluctuations as described by the Onsager term. These fluctuations are relevant (as compared to the ferromagnetic case where they are irrelevant) due to the scale of our couplings being of order $1/\sqrt{N}$. It is still possible that asymmetric couplings contribute to the shape or size of this bump, but this is not explored in this thesis.

Low T regime For $\beta > 10$, we find that all methods converge to the same solution with different precision. The simulation and the naive mean-field quickly and precisely find this solution with minimal error. The stability of both lines also indicates that neither seem to capture the spin glass phase. This could be due to the system living outside replica symmetry breaking (RSB) regime, or could be a limitation of the methods.

The TAP solution on the other hand fluctuates significantly, even when eliminating spin-flips. It has been shown that TAP is a replica-symmetric solution [21], such that it fails when crossing the dAT line (Fig. 2.4). Thus, the instability of TAP in this temperature regime may be due to the emergence of the spin glass phase. On the other hand, as explained in appendix

2.4, TAP is a high-temperature approximation and thereby doesn't apply well to low-temperatures. To see whether these reasons apply, we must analyze the replica overlap matrix of the system. Thus to complete our analysis on the low temperature regime, we delve deeper into RSB.

4.2 Replica Symmetry Breaking

We note that the external field vectors, as chosen by Bal and adopted by us, are sampled from a Gaussian with zero mean and variance R^2/D while the couplings are sampled from a Gaussian with zero mean and variance J_v^2/N. For our $N = 64, D = 32$ simulations, the radius was taken to be $R = \sqrt{D/2 - 1} \approx 3.9$ while $J = 1$, such that our external field was significantly stronger than our couplings. As explained in Section 2.1.3, increasing field strengths lead to ergodic solutions. Due to this imbalance, our system likely exists in a replica symmetric phase even for very low temperatures. Therefore, to check whether our MCMC algorithm is capable of exhibiting a spin-glass RSB phase, we lower the external field in an attempt to cross the de Almeida-Thouless line. To this effect a field strength parameter h is implemented, multiplying with each vector in applied field $\mathbf{x}$. Replicas are generated using identical sets of randomly distributed couplings $\mathbf{J}$ and external field $\mathbf{x}$, but starting from different initial states $\mathbf{s_0}$.

dAT recreation We found that after decreasing the magnitude of $\mathbf{x}$, the overlap matrix exhibited clear RSB. Furthermore, we believed that the phase diagram of our ST exhibits a dAT line that resembles that of the SK model, as shown in Fig. 2.4 from Sec. 2.1.3. Thus, we simulated 5-replica overlap matrices for different combinations of h and T as shown in Fig. 4.2. The spin glass transition from the SK model is indicated with a red line, allowing for a rough comparison with the ST model. Though our simulation uses local propositions, we show this causes no issue with ergodicity in Appendix B.3.

The miniature overlap matrices in Fig. 4.2 show 3 different extremes. For high h and low T, the matrix becomes uniform as each replica aligns closely to the applied field and therefore with each other. The opposite happens for low h high T, for which the matrix has no alignment between any pair of different replicas, corresponding to the paramagnetic phase. Finally, we see evidence of the spin glass phase in low h low T, for which the overlap matrix shows RSB.

This field-aligned state at high h low T does not refer to complete alignment with the field, but refers to the ground state balancing the influ-

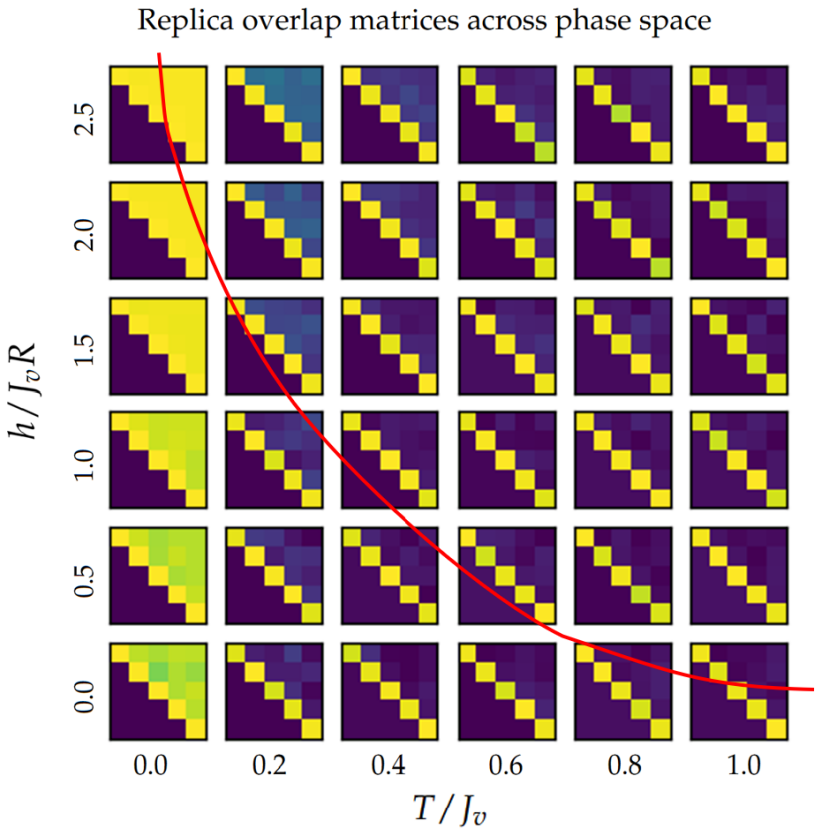

Figure 4.2: *Miniature overlap matrices of MCMC replicas for varying temperatures and field strength h. The overlap matrices show the spin glass transition from stable states (i.e. field-aligned as in top left or paramagnetic as in bottom right) to unstable states (RSB in bottom left). The red line marks the SK model spin glass transition line from Fig. 2.4.*

ences of applied field and spin interactions. Looking at the column for $T/J_v = 0.2, 0.4$, the increasing field strength brings the state from paramagnetic to a partially field-aligned phase. From Fig. 4.1 we see that the analytic solutions transition from these partially field-aligned states at low temperatures to fully field-aligned states at high temperatures. We explained that this is due to an effective scaling of **x** in the analytic solutions, as increasing the field strength brings the state to perfect field alignment. The analytic solutions can thus be interpreted as traveling along a steep $h >> T$ line or curve on the phase diagram from Fig. 4.2.

Our recreated phase diagram can be compared to the two SK model phase diagrams shown in Section 2.1.3. Similar to the TvJ_0 phase diagram from Fig. 2.3, a phase transition between paramagnetic and ferromagnetic

phases is apparent in the stable states, where the ferromagnetic state in the ST model corresponds to the partially field-aligned ground state. Fig 2.4 similarly depicts a hvT phase diagram, offering a direct comparison by plotting the SK model phase transition as a red line in Fig. 4.2. It is clear that the same stability line does not hold exactly for the ST model; the temperature appears to contribute more noise such that the paramagnetic phase persists at temperatures below the dAT line for low h. This might be due to the continuous nature of the vector-spins, which compared to binary spins are susceptible to small temperature fluctuations. However, the general RSB structure as function of field strength h and temperature T implies that a similar spin-glass phase transition to the SK model persists in the ST model.

Chapter 5

Discussion & Conclusion

5.1 Conclusion

Based on a spin-transformer (ST) correspondence found and solved analytically by Bal, we have used an MCMC simulation with the Metropolis-Hastings algorithm to sample states near equilibrium. The behavior of spin systems differ depending on the temperature, leading to a comparison in 3 temperature regimes. We show that this numerical solution converges similarly to the naive mean-field solution in low-temperature regimes $\beta > 10$ under strong applied fields, whereas the Thouless-Anderson-Palmer equations fail as expected due to being high-temperature approximations. Lowering the strength of the field, we simulate many replicas and construct overlap matrices, from which we observe replica-symmetry-breaking. Constructing overlap matrices for varying field strengths and temperatures, we visualize phase transitions between spin-glass, paramagnetic, and field-aligned phases that resemble the de Almeida-Thouless line from the SK model phase diagram. Furthermore, we show that the analytic solutions imply a convergence to the applied field at extreme high temperatures $\beta < 1$, preventing these solutions from entering the paramagnetic phase shown by MCMC. We found that the MCMC numerical solutions do not capture second-order fluctuations at intermediate temperatures $1 < \beta < 10$ as found in TAP, and that under the MH algorithm they are unable to capture nonequilibrium dynamics that follow asymmetric couplings. We conclude that Bal's analytic solutions to the ST model excel in the intermediate temperature regime, outside of which they either fail or are incomplete. Through this research, we reinforce that the spin-transformer model offers a beautiful correspondence between transformer modules and spin glass physics, opening high-potential avenues into ex-

ploring and understanding LLMs.

5.2 Discussion

5.2.1 Asymmetric Couplings

Using asymmetric couplings $\mathbf{J}$ is necessary for Transformers, and furthermore has a deep impact on interactive systems. In spin models, they lead to time reversal asymmetry as the probability of flowing between two states is not equal. This is equivalent to entropy production, leading to path-dependency and complex time dynamics [26].

It is intriguing that the equilibrium MCMC solution settles to the same approximate point as the analytic solutions at low temperatures despite the latter's asymmetric considerations (Fig B.3.2 shows symmetric TAP solutions also fluctuate at low T). If the analytic solutions correctly capture asymmetric couplings, then the results imply that asymmetric couplings have little effect on their NESS in the ergodic phase. In our study, it was difficult to controllably measure varied asymmetric odd components $J_{odd} = (\mathbf{J} - \mathbf{J}^T)/2$ as it directly affects the magnitude of $\mathbf{J}$, thus affecting the sensitive $\mathbf{m}_{NESS}$. It would thus be a possible research direction to explore this nonequilibrium behavior, and answer questions such as; How does asymmetry affect a system's NESS? Does this NESS state develop over time? Do transformers exhibit nonequilibrium behavior?

5.2.2 ST Correspondence

The main purpose behind confirming the ST solution was to add robustness to the sprouting idea and expand its reach to those who read this. It is therefore very interesting to keep building up on this correspondence, and close the gap between ST and deep-equilibrium transformers further. There are multiple interesting avenues to explore in this correspondence.

Multi-head attention as RSB One of the most interesting possibilities was the correspondence between n-head attention and n-RSB overlap matrices. Both methods project the relations between states to block-diagonal matrices. Multi-head attention increases the number of dimensions and forms token-token matrices for each group of dimensions, while RSB organizes replica relations into block-diagonal form to summarize the depth of replica symmetry breaking. Though these two block-diagonal matrices

Version of July 10, 2025– Created July 10, 2025 - 10:07

describe different relations, it could be possible that multi-heads exist to deconstruct asymmetry as one does in n-RSB.

Feed-forward as Onsager term The feed forward layer in transformer modules is elusive, but is vaguely seen as adding nonlinearity in order to make the attention matrix more interpretable [22]. This is intriguing, as adding nonlinearity is also done in the ST solution by considering the second-order interaction term. Bal postulates that the feed-forward layer in transformers acts as to incorporate the second-order Onsager reaction term. It would be interesting to see how similarly the feed-forward layers in language models function as Onsager reaction terms.

Couplings In this study, we've used normally distributed couplings to emulate the attention. To be more precise, these couplings have to be a function of the field $J = xQK^Tx^T/\sqrt{D}$, which would be compelling to see the effects of. Studying the self-attention matrices in true transformers would reveal their degree of frustration and asymmetry, which is relevant to our spin glass model.

5.2.3 Vector Spin glasses

We have found numerical solutions for vector spin glasses using MCMC to compare them to analytical solutions found by Bal. While there is literature on spin glasses in higher dimensions ([30] for a review), large D-dimensional spin glasses haven't been thoroughly studied. Our results suggest that they follow the spin glass transition of SK models, but it is not enough to conclude its universality class.

Deeper understanding of vector spin glasses would provide deeper understanding to the ST, which could extend understanding to language models. This is especially so for nonequilibrium behavior of spin glasses, for which the understanding is still new and developing.

Version of July 10, 2025– Created July 10, 2025 - 10:07

Chapter 6

Appendix

Appendix *A* elaborates on discussions with my supervisor and problems faced during the research process. Appendix *B* provides additional theory and results supporting the paper.

A.1 Transformer heads elaboration

In the computer science approach to transformers, the attention heads are freely manipulated in the sense of separating a matrix into heads and reattaching them after each head has processed its attention. As this separation is done across the token's embedding, the number of heads H are typically a divisor of token dimension D. Writing this formally in an equation is a bit more difficult, as there are varying interpretations of the 'separating' and the 'reattaching'. This appendix summarizes our discussion on this formulation.

In a paper delving into how information runs through different attention heads across many layers [23], the equation for the transformer module forward pass is given by:

$$x_i^{(\ell+1)} = \frac{1}{\sqrt{FH}} \sum_{h=1}^{H} \sum_{j=1}^{N} W_V^{(\ell)h} \cdot x_j^{(\ell)} \zeta \left(\frac{1}{D\sqrt{G}} x_j^{\top} \cdot W_K^{(\ell)h\top} \cdot W_Q^{(\ell)h} \cdot x_i \right) \tag{A.1.1}$$

with the input sequence $x \in \mathbb{R}^{N \times D}$ as usual, layers l, heads $h = \{1, ..., H\}$, softmax function ζ, and where they have used adjustable weight dimensions in $W_V^{(l)h} \in \mathbb{R}^{F \times F}$ and $W_K^{(l)h}, W_Q^{(l)h} \in \mathbb{R}^{G \times D}$. They have excluded aspects of completes transformers such as residuals and layer normalization, but our focus lies on their handling of attention heads. A mathemati-

Version of July 10, 2025– Created July 10, 2025 - 10:07

cal perspective on transformers describes the transformer module forward pass as:

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\sum_{h=1}^{H} \sum_{j=1}^{N} \frac{e^{\beta \langle W_Q^h(t) x_i(t), W_K^h(t) x_j(t) \rangle}}{Z_{\beta,i,h}(t)} W_V^h(t) x_j(t) \right) \tag{A.1.2}$$

where the softmax function is written explicitly, $\mathbf{P}^{\perp}_x y = y - \langle x, y \rangle x$ projects the token $y \in \mathbb{R}^D$ onto tangent space $T_x \mathbb{S}^{D-1}$, and the layer index is interpreted as a time variable. In this formulation, the adjustable weight dimension are chosen such that matrices are square $W^h_{Q,K,V} \in \mathbb{R}^{D \times D}$.

The first important note is that in the summation $\sum_{h=1}^{H}$, each resultant head corresponds to a $N \times d$ column as part of a whole $N \times D$ output matrix. Thus the summation does not entail summing these heads together and taking an average, but rather a concatenation of columns by adding to empty placeholders. We initially imagined this as weight matrices being block-diagonal with each block corresponding to a head. However, this imposes the dimension groupings (heads) onto the input $\mathbf{x}$ as opposed to each head encompassing all dimensions as in applied transformers. This further results in a head block-diagonal output matrix, whereas in practice we want a filled $N \times D$ self attention matrix.

In order to make the equation faithful to a true transformer module, we thus have to use projections in the weight matrices. For example, the full $W_{Q,K,V} \in \mathbb{R}^{D \times D} = W^{\beta}_{\alpha}$ in Einstein notation would become $W^{(\beta,h)}_{\alpha}$ where the columns noted by index β would be divided into heads h. Using projections allows the $N \times d$ columns to calculate attention without mixing, and summing over the heads allows them to concatenate back into a full $N \times D$ output attention matrix. Altogether, this brings us to our complete transformer module equation from Section 2.2:

$$\begin{aligned} g(x_i^{\alpha}) &= LN \left[\sum_{h=1}^{H} \xi \left(x_i^{\alpha} W_{\alpha}^{(\gamma,h)} W^{\beta}_{(\gamma,h)} x^j_{\beta} \right) \cdot x_j^{\nu} W_{\nu}^{(\alpha,h)} + x_i^{\alpha} \right] \\ y_i^{\alpha} &= LN \left[FF(g(x_i^{\alpha})) + g(x_i^{\alpha}) \right] \end{aligned}$$

A.2 Asymmetry elaboration

Asymmetric couplings have an important role in spin transformers (ST), which we hoped to explore at the start of this project. In this appendix we give a complete overview on why they matter, how we tried to handle them, and why it did not work as planned.

Version of July 10, 2025– Created July 10, 2025 - 10:07

Importance of asymmetry In statistical physics, asymmetric couplings are a generic feature of nonequilibrium systems. Compared to thermodynamically equilibrium systems which feature balanced time-independent states, nonequilibrium systems prevent a balance from being reached and endlessly change over time. In equilibrium systems, free energy can clearly be formulated as a balance between internal energy and entropy $F = U - TS$. However, this is a formulation specific to an equilibrium state constant in time, which is able to maximize its entropy to minimize its free energy. This description of free energy breaks down for nonequilibrium systems which change in time, featuring time reversal asymmetry and entropy production [26]. In short, using asymmetric couplings has the significant consequence of breaking equilibrium and introducing new dynamics to the system.

In contrast to multi-layer perceptrons which typically feature symmetric connections between neurons, transformers applied to language models inherently train asymmetric connections between tokens. The interactions between tokens of a sequence are mapped in an attention matrix, quantifying how much each token attends to every other token. However, the connection between two tokens is typically asymmetric. For example, in a generative attention matrix, the probability of predicting the token ”thesis” after ”my” is greater than that of predicting ”my” after ”thesis”. Thus, the coupings in the ST model must similarly be asymmetric.

Even vs. Odd An asymmetric matrix $J_{ij} \neq J_{ji}$ can be split into an even and odd part;

$$J^e_{ij} = \frac{J_{ij} + J_{ji}}{2}, \qquad J^o_{ij} = \frac{J_{ij} - J_{ji}}{2}, \tag{A.2.1}$$

The energy can then be shown to only depend on J^e [31]:

$$E = -\sum_{ij} J_{ij} s_i s_j = -\sum_{i<j} (J_{ij} s_i s_j + J_{ji} s_j s_i) = -\sum_{ij} \frac{J_{ij} + J_{ji}}{2} s_i s_j = -\sum_{ij} J^e s_i s_j \tag{A.2.2}$$

meaning that the internal energy only accounts for the symmetric aspect of the couplings. This proof holds as long as the spins s_i, s_j are drawn from the same state $\mathbf{s} = \{s_1, ..., s_N\}$.

Kinetic Ising framework In our MCMC simulations, we attempted to bypass this strictly symmetric formula by drawing s_i, s_j from different

Version of July 10, 2025– Created July 10, 2025 - 10:07

state configurations. Specifically, we implemented the Kinetic Ising framework that draws $s_{i,t}$ from time t and $s_{j,t-1}$ from the previous step [24]. This not only allows for consideration of J^o, but also incorporates time dynamics into the system. This Kinetic Ising framework is commonly used to address the dynamics of nonequilibrium systems featuring nonzero J^o [12, 14, 24]. This leads to the internal energy of some configuration $\mathbf{s}_t$ at time t:

$$H(\mathbf{s}_t, \mathbf{s}_{t-1}) = -\sum_{ij} J_{ij} s_{i,t} s_{j,t-1} - \sum_i x_i s_{i,t} \tag{A.2.3}$$

Stationary distribution However, a problem arose once we used this framework together inside the Metropolis-Hastings (MH) algorithm. In this algorithm, stationary distributions are defined as $\pi(\mathbf{s}_t) = e^{-\beta E(\mathbf{s}_t)}/Z$ with $E(\mathbf{s}_t)$ being the internal energy of configuration $\mathbf{s}_t$. Then, the acceptance probability of transitioning from state $\mathbf{s}_t$ to proposed state $\mathbf{s}_y$ depends on the ratio of stationary distributions $\frac{\pi(\mathbf{s}_y)}{\pi(\mathbf{s}_t)} = e^{-\beta[E(\mathbf{s}_y)-E(\mathbf{s}_t)]}$. Using our Kinetic Ising framework, this internal energy is redefined as $E(\mathbf{s}_t, \mathbf{s}_{t-1})$. The complication then lies in how we set up the ratio:

$$\frac{\pi(s_y, s_t)}{\pi(s_t, s_{t-1})}, \quad \text{using } s_y \sim s_{t+1} \tag{A.2.4}$$

$$\frac{\pi(s_y, s_{t-1})}{\pi(s_t, s_{t-1})}, \quad \text{using } s_y \text{ as } s_t \text{ alternative} \tag{A.2.5}$$

Using these ratios turned out to work worse than the non-kinetic simple formulation $\pi(\mathbf{s}_y)/\pi(\mathbf{s}_t)$, in that they took longer to approach a worse convergence. Here 'worse convergence' refers to both worse self-attention alignment $\mathcal{A}(\mathbf{s}_t, f_{attn}(\mathbf{s}_t))$ (as shown in Fig. A.2.1) and worse alignment with analytical solutions $\mathcal{A}(\mathbf{s}_t, \mathbf{m}_{NESS}^{nMF,TAP})$. We thus continued our simulations with the energy

$$H(\mathbf{s}_t) = -\sum_{ij} J_{ij} s_{i,t} s_{j,t} - \sum_i x_i s_{i,t} \tag{A.2.6}$$

which means that we were only able to study the symmetric aspects of the vector-spin glass as proven by Eq. A.2.2.

Metropolis Hastings The reason for failing to incorporate asymmetries in our MCMC simulations is likely due to the MH algorithm being equilibrium centric. We acknowledged early on that MH enforces detailed balance, but we took this as necessary in order for the sampling to progress

Version of July 10, 2025– Created July 10, 2025 - 10:07

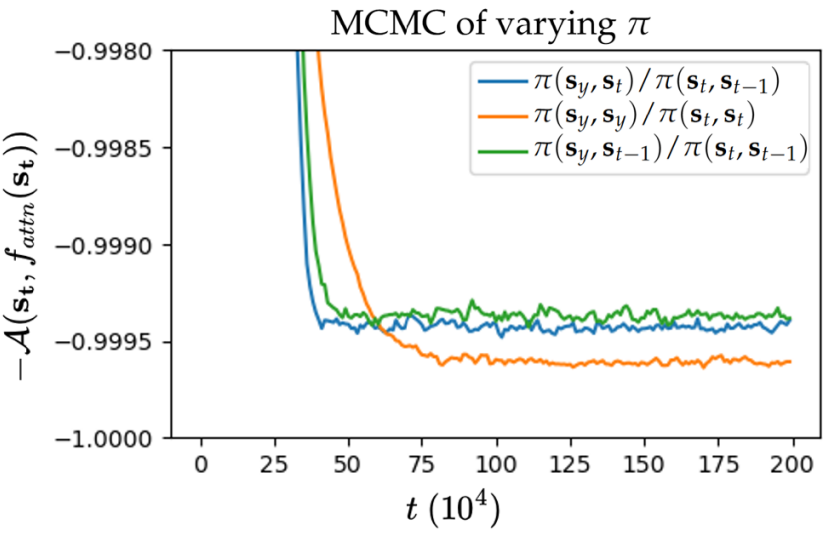

Figure A.2.1: *Attention alignment for varying stationary distributions at* $B = 1000$.

towards the NESS. However, the reliance on equilibrium stands central to MH and is not just a loose constraint to help find nonequilibrium. Specifically, the use of the Boltzmann distribution is also an artifact of equilibrium thermodynamics. Though a nonequilibrium stationary state may also be Boltzmann-like, this is not necessarily true and it would be a mistake to strictly define it as such. It is therefore not a surprise that using a symmetric, equilibrium-friendly energy function works best. What is however surprising is how the analytic solutions show a similar minimal response to asymmetry, as discussed in Section 5.2.1.

Further remarks In general, it was difficult to single out the effects of the odd components J^o as even in the symmetric formulation, J^o affected the convergence due to it influencing the normalization constant $\alpha = ||\mathbf{x} + \mathbf{J}\mathbf{s}_t||^{-1}$. We also briefly attempted to rewrite the energy function by using only $\mathbf{s}_t$ and incorporating time dynamics using a Legendre transform;

$$s_j(t-\epsilon) \to s_j(t) - \epsilon J \dot{s}_j(t) + \frac{\epsilon^2}{2} J^2 \ddot{s}_j(t) \tag{A.2.7}$$

but this was not further explored as we found the problem to lie in our Monte Carlo algorithm.

B.3 Other MCMC Results

For detailed results, please refer to `https://github.com/beerch/ST_MCMC`.

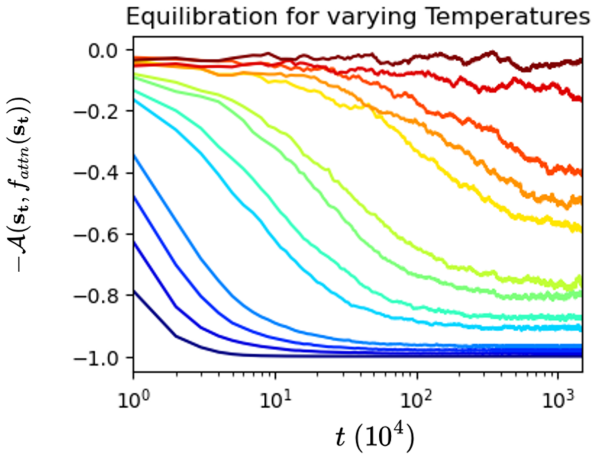

Figure B.3.1: *Convergence of MCMC ST showing expected equilibrating behavior for varying temperatures ranging from $\beta = 1000$ to $\beta = 0.001$. $-\mathcal{A}$ measures negative alignment of state $\mathbf{s}_t$ with its attention $f_{attn}(\mathbf{s}_t)$, where $\mathcal{A} \to 0$ is random/orthogonal alignment, and $\mathcal{A} \to 1$ is full vector alignment.*

Convergence of MCMC Figure B.3.1 shows the convergence of MCMC simulation for different temperatures. At low temperatures, $\mathcal{A} \to 1$ as they converge to the ground state. At high temperatures. $\mathcal{A} \to 0$ as they become random and paramagnetic. The higher the temperature, the longer it takes to converge to its respective equilibrium state.

Local vs. Global proposals Using local propositions in our MCMC method has consequences in our ergodicity analysis from Section 4.2. Even though the probability of leaving a deep RSB exists, it might be unfeasibly small that the system is virtually non-ergodic. Global, fully ergodic propositions on the other hand take too long to converge even if the solution has a clear minimum. Thus we tried many different combinations of local and global propositions, alternating between the two at a variety of intervals. They showed convergence to the same states as when one uses pure local propositions. We also tried using local propositions, but including a chance of spin-flips and perpendicular adjustments. Again, they converged to the

same states as pure local proposals. Thus, we concluded that using pure local proposals are sufficiently ergodic for our system parameters.

Observables Typical to statistical physics is analysis of measurements such as autocorrelation time, specific heat capacity, and free energy diagrams, which help identify phase transitions. A similar analysis was done for the MCMC simulations, for which we found a peak in autocorrelation times at $\beta \approx 1$ and specific heat capacities at $\beta \approx 10$. However, these results were left out due to them not being very meaningful in understanding spin transformers. They were simulated using our default field strength stronger than coupling strength (Bal's parameters), they marked the phase transition between the paramagnetic and 'ferromagnetic' states. It could be interesting to analyze them across the spin glass phase transition. However, if we extend the system to incorporate asymmetric couplings, the phase behavior becomes time-dependent and an alternative approach to analyzing phase transitions is required.

Symmetric Analytic Solutions Figure (B.3.2) shows the convergence of $\mathbf{m}_{NESS}$ with field $\mathbf{x}$ varying β of TAP and absTAP, this time using symmetric couplings $J_{ij} = J_{ji}$. Though J^0 influences the shape of the bump and final $\mathbf{m}_{NESS}$, the bump itself is a manifestation of the second-order interactions.

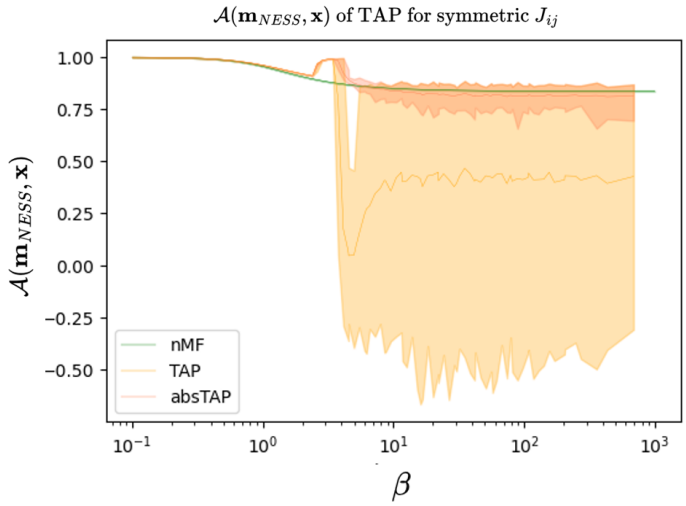

Figure B.3.2: *Alignment of $\mathbf{m}_{NESS}$ with external field for varying β, for a system using symmetric couplings J_{ij}.*

Version of July 10, 2025– Created July 10, 2025 - 10:07

Bibliography

[1] Minghao Shao, Abdul Basit, Ramesh Karri, and Muhammad Shafique. Survey of different large language model architectures: Trends, benchmarks, and challenges. *IEEE Access*, 12:188664â188706, 2024.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[3] Omkar Komera and Rahul Manche. Black-box behavior in large language models: Challenges and implications, 12 2023.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[5] William Bialek. Moving boundaries: An appreciation of john hopfield, 2024.

[6] Zhiquan Tan, Chenghai Li, and Weiran Huang. The information of large language model geometry, 2024.

[7] Emily Cheng, Corentin Kervadec, and Marco Baroni. Bridging information-theoretic and geometric compression in language models, 2023.

[8] Kento Nishi, Rahul Ramesh, Maya Okawa, Mikail Khona, Hidenori Tanaka, and Ekdeep Singh Lubana. Representation shattering in transformers: A synthetic study with knowledge editing, 2025.

[9] Jiali Cheng and Hadi Amiri. Linguistic blind spots of large language models, 2025.

[10] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024.

[11] Yuhao Li, Ruoran Bai, and Haiping Huang. Spin glass model of in-context learning, 2025.

[12] Matthias Bal. Spin-model transformers. December 2023.

[13] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models, 2019.

[14] Miguel Aguilera, S. Amin Moosavi, and Hideaki Shimazaki. A unifying framework for mean-field theories of asymmetric kinetic ising systems. *Nature Communications*, 12(1), February 2021.

[15] Yi-Ping Ma, Ivan Sudakov, Courtenay Strong, and Kenneth M. Golden. Ising model for melt ponds on arctic sea ice, 2017.

[16] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich. Cooperativity of protein folding and the random-field ising model, 1996.

[17] Haochen Li, Yi Cao, Maria Polukarov, and Carmine Ventre. An empirical analysis on financial markets: Insights from the application of statistical physics, 2024.

[18] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, Dec 1975.

[19] J R L de Almeida and D J Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, may 1978.

[20] Ioannis A. Hadjiagapiou. The sherringtonâkirkpatrick spin glass model in the presence of a random field with a joint gaussian probability density function for the exchange interactions and random fields. *Physica A: Statistical Mechanics and its Applications*, 397:1â16, March 2014.

[21] Ada Altieri and Marco Baity-Jesi. *An introduction to the theory of spin glasses*, page 361â370. Elsevier, 2024.

[22] Shashank Sonkar and Richard G. Baraniuk. Investigating the role of feed-forward networks in transformers using parallel attention and feed-forward net design, 2023.

[23] Lorenzo Tiberi, Francesca Mignacco, Kazuki Irie, and Haim Sompolinsky. Dissecting the interplay of attention paths in a statistical mechanics theory of transformers, 2024.

[24] Roy J. Glauber. Timeâdependent statistics of the ising model. *Journal of Mathematical Physics*, 4(2):294–307, 02 1963.

[25] T Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, jun 1982.

[26] Giovanni Gallavotti. Entropy production and thermodynamics of nonequilibrium stationary states: A point of view. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 14(3):680â690, September 2004.

[27] J.-C. Walter and G.T. Barkema. An introduction to monte carlo methods. *Physica A: Statistical Mechanics and its Applications*, 418:78â87, January 2015.

[28] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1â124, May 2019.

[29] Giorgio Parisi. The physical meaning of replica symmetry breaking, 2002.

[30] Valentina Ros and Yan V. Fyodorov. The high-d landscapes paradigm: spin-glasses, and beyond, 2023.

[31] Gabriel Artur Weiderpass, Mayur Sharma, and Savdeep Sethi. Solving the kinetic ising model with nonreciprocity. *Physical Review E*, 111(2), February 2025.