Stochastic Processes and the Neural Tangent Kernel

Teo Benarous

Winter 2024

Abstract

This document presents a study conducted as part of the McGill Directed Reading Program (DRP), which pairs undergraduate students with graduate mentors to explore research-level topics in mathematics and statistics. Over the semester, our focus was on three topics: martingale theory, stochastic processes on trees and and the neural tangent kernel (NTK). These notes aim to elucidate key concepts associated with each topic. The section on the NTK culminates in an experimental project that investigates the convergence behaviour of the empirical NTK to a fixed limit.

Contents

1 Martingale Theory

The notes in this section are compiled from these sources: [18], [6], [13] and [5].

Definition 1.1 (Stochastic Process)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, (S, Σ) be a measurable space and $(\mathbb{T}, \leqslant)$ be an index set equipped with a total order then a stochastic process Y is a collection of S-valued random variables, i.e.,

$$Y = \{X : \mathbb{T} \times \Omega \to S\}$$

where S is called the state space of Y. A sample function is a single outcome of a stochastic process, i.e., $\forall\, \omega \in \Omega$, the map

$$X_\omega \coloneqq X(\cdot, \omega) : \mathbb{T} \to S$$

is called a sample function, or a realization, or particularly a sample path if $\mathbb{T}$ is interpreted as time. Let $t_1, t_2 \in \mathbb{T} : t_1 \leqslant t_2$ then $X_{t_2} - X_{t_1}$ is called an increment of Y.

Remark. An increment can be interpreted as how much the stochastic process changes over a certain time period.

Definition 1.2 (Filtration)**.** Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, $(\mathbb{T}, \leqslant)$ be an index set equipped with a total order and $\forall\, t \in \mathbb{T} : \mathcal{F}_t \subseteq \Sigma$ be a σ-algebra. If

$$\forall\, t_1, t_2 \in \mathbb{T} : t_1 \leqslant t_2 \implies \mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$$

then

$$\mathbb{F} \coloneqq (\mathcal{F}_i : t \in \mathbb{T})$$

is called a filtration. In this case, $(\Omega, \Sigma, \mathbb{F}, \mathbb{P})$ is called a filtered probability space. Moreover, given a sequence of real-valued random variables $(X_t : t \in \mathbb{T})$, if

$$\forall\, t \in \mathbb{T} : \mathcal{F}_t = \sigma\big(X_{\tilde{t}} \,\big|\, \tilde{t} \leqslant t\big)$$

then $\mathbb{F}$ is called a natural filtration.

Definition 1.3 (Adapted Process)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathbb{T}, \leqslant)$ be an index set equipped with a total order, $\mathbb{F} = (\mathcal{F}_t : t \in \mathbb{T})$ be a filtration on $\mathcal{F}$, (S, Σ) be a measurable space then the stochastic process $Y = (X_t : \Omega \to S \,|\, t \in \mathbb{T})$ is said to be adapted to the filtration $\mathbb{F}$ if $\forall\, t \in \mathbb{T} : X_t$ is an $(\mathcal{F}_t, \Sigma)$-measurable map.

Definition 1.4 (Martingale)**.** Let $(\Omega, \Sigma, \mathbb{F}, \mathbb{P})$ be a filtered probability space, and S be a Banach space. A stochastic process $Y = (X_t : \Omega \to S \,|\, t \in \mathbb{T})$ is said to be a martingale with respect to $\mathbb{F}$ and $\mathbb{P}$ if

- $\forall\, t \in \mathbb{T} : X_t$ is a $\mathcal{F}_t$-measurable map.
- $\forall\, t \in \mathbb{T} : X_t \in L^1(\Omega, \mathcal{F}_t, \mathbb{P}, S)$, i.e., $\mathbb{E}[\,\|Y_t\|_S\,]$ is finite.
- $\forall\, t_1, t_2 \in \mathbb{T} : t_1 < t_2 \implies X_{t_1} = \mathbb{E}[X_{t_2} \,|\, \mathcal{F}_{t_1}]$

Moreover, if

$$\forall\, t_1, t_2 \in \mathbb{T} : t_1 < t_2 \implies X_{t_1} \leqslant \mathbb{E}[X_{t_2} \,|\, \mathcal{F}_{t_1}]$$

then Y is called a super-martingale, and if

$$\forall\, t_1, t_2 \in \mathbb{T} : t_1 < t_2 \implies X_{t_1} \geqslant \mathbb{E}[X_{t_2} \,|\, \mathcal{F}_{t_1}]$$

then Y is called a sub-martingale.

Definition 1.5 (Stopping Time)**.** Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space and $\tau \in \Omega \times T$ be a random variable. Then τ is called a stopping time if

$$\forall\, t \in T : \{\tau \leqslant t\} \coloneqq \{\tau(\omega) \leqslant t \,|\, \omega \in \Omega\} \in \mathcal{F}_t$$

Remark. The definition above can be also be interpreted as an adapted process, i.e., $\tau \in \Omega \times T$ is called a stopping time if the stochastic process $X = (X_t : t \in \mathbb{T})$ defined by

$$X_t \coloneqq \begin{cases} 1 & t < \tau \\ 0 & t \geqslant \tau \end{cases}$$

is adapted to the filtration $\mathbb{F} = (\mathcal{F}_t : t \in \mathbb{T})$.

Remark. Stochastic approximation looks like Euler's method

$$x_{k+1} = x_k + \frac{1}{k} f(x_k)$$

with step size $\frac{1}{k}$. Informally, we might expect stochastic approximation to behave like the o.d.e. $x_t' = f(x_t)$. Moreover, we might expect that not only does stochastic approximation find a zero of f, but is should (almost surely) find a stable zero of f.

Theorem 1.6. *The maximum and minimum of two stopping times is also a stopping time.*

Proof. Let $\tau,\ \sigma$ be stopping times then

$$\{\min\{\tau,\ \sigma\} \leqslant j\} = \{\tau \leqslant j\} \cup \{\sigma \leqslant j\} \in F_j$$
$$\{\max\{\tau,\ \sigma\} \leqslant j\} = \{\tau \leqslant j\} \cap \{\sigma \leqslant j\} \in F_j$$

□

Theorem 1.7. *Let (X_n) be a Markov chain with transition probability p and let $f(x,n)$ be a function of the state x and the time n such that*

$$f(x,n) = \sum_y p(x,y)\, f(y,n+1)$$

then $(M_n) = f(X_n, n)$ is a martingale. In particular, if

$$h(x) = \sum_y p(x,y)\, h(y)$$

then $h(X_n)$ is a martingale.

Proof. By the Markov property and the assumption on f,

$$\mathbb{E}[f(X_{n+1}, n+1) \,|\, \mathcal{F}_n] = \sum_{Y \in \mathcal{F}_n} p(X_n, Y)\, f(Y, n+1) = f(X_n, n)$$

hence X_n is martingale. □

Example 1.8 (Gambler's ruin)**.** Consider a gambler who starts with an initial fortune of \$$k$ and then on each successive gamble either wins \$1 or loses \$1 independent of the past with probabilities p and $q = 1-p$. The gambler's objective is to reach a total fortune of \$$N$, without first getting ruined (running out of money).

1. Calculate by brute force, the probability of ruin given the initial state $0 \leqslant k \leqslant N$.
2. Compare the above with the martingale version.

$$\begin{aligned}
\forall\, 1 \leqslant k \leqslant n-1 : p_k = p \cdot p_{k+1} + (1-p) p_{k-1} \implies p_{k+1} &= \frac{1}{p} p_k - \frac{1-p}{p} p_{k-1} \\
\implies p_{k+1} - p_k &= \left(\frac{1}{p} p_k - \frac{1-p}{p} p_{k-1}\right) - p_k \\
&= \left(\frac{1}{p} - 1\right) p_k - \frac{1-p}{p} p_{k-1} \\
&= \frac{1-p}{p}(p_k - p_{k-1}) \\
&= \left(\frac{1-p}{p}\right)^2 (p_{k-1} - p_{k-2}) \\
&\vdots \\
&= \left(\frac{1-p}{p}\right)^k (p_1 - p_0) \\
\implies p_k &= \sum_{k=0}^{i-1} \left(\frac{1-p}{p}\right)^k p_1
\end{aligned}$$

If $p = \frac{1}{2}$ then $\frac{1-p}{p} = 1$ therefore

$$p_N = N \cdot p_1 \implies p_1 = \frac{1}{N} \implies p_k = \frac{k}{N}$$

When $p \neq \frac{1}{2}$ then

$$p_N = p_1 \frac{1 - \left(\frac{1-p}{p}\right)^N}{1 - \frac{1-p}{p}}$$
$$\implies p_1 = \frac{1 - \frac{1-p}{p}}{1 - \left(\frac{1-p}{p}\right)^N}$$

therefore

$$p_k = \frac{1 - \left(\frac{1-p}{p}\right)^k}{1 - \left(\frac{1-p}{p}\right)^N}$$

It follows that the probability of ruin is

$$\tilde{p}_k = \begin{cases} 1 - \frac{k}{N} & p = \frac{1}{2} \\ 1 - \frac{1-\left(\frac{1-p}{p}\right)^k}{1-\left(\frac{1-p}{p}\right)^N} & p \neq \frac{1}{2} \end{cases}$$

For the second part, consider $\left(U_n \sim \mathcal{U}_{\{-1,1\}} : n \in \mathbb{N}\right)$ and let $\mathbb{F} = (\mathcal{F}_n : n \in \mathbb{N}) = \sigma(\{U_i \mid 1 \leqslant i \leqslant n\})$ and S_n be the total amount of dollars gained so far at timestep $n \in \mathbb{N}_0$. Define

$$S_0 \coloneqq k \qquad S_n \coloneqq \sum_{i=1}^{n} U_i$$

Claim. $(S_n : n \in \mathbb{N}_0)$ *is a martingale.*

Proof.

$$\begin{aligned} \mathbb{E}[S_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[S_n + U_{n+1} \mid \mathcal{F}_n] \\ &= S_n + \mathbb{E}[U_{n+1}] \\ &= S_n \end{aligned}$$

□

Let $\tau = \inf\{n \in \mathbb{N} : S_n \in \{0, n\}\}$ then

$$\begin{aligned} \mathbb{E}[S_\tau] &= \mathbb{E}[S_0] = k \\ &= 0 \cdot \mathbb{P}(S_\tau = 0) + N \cdot \mathbb{P}(S_\tau = N) \\ &= N \cdot p_k \end{aligned}$$

therefore $p_k = \frac{k}{N}$. It follows that the probability of ruin is $1 - \frac{k}{N}$.

Let $(U_n : n \in \mathbb{N})$ i.i.d. such that $\mathbb{P}(U_n = 1) = p$ and $\mathbb{P}(U_n = 1) = 1-p =: q$. Set $\mathbb{F} = (\mathcal{F}_n : n \in \mathbb{N}) = \sigma(\{U_i \mid 1 \leqslant i \leqslant n\})$.

Claim. $\left(\frac{q}{p}\right)^{S_n}$ *is a martingale*

Proof.

$$\begin{aligned} \mathbb{E}\left[\left(\frac{q}{p}\right)^{S_{n+1}} \,\middle|\, \mathcal{F}_n\right] &= p\left(\frac{q}{p}\right)^{S_{n+1}} + q\left(\frac{q}{p}\right)^{S_{n-1}} \\ &= \left(\frac{q}{p}\right)^{S_n}\left(p \cdot \frac{q}{p} + q \cdot \frac{p}{q}\right) \\ &= \left(\frac{q}{p}\right)^{S_n} \end{aligned}$$

□

It thus follows that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{q}{p}\right)^{S_\tau}\right] = \mathbb{E}\left[\left(\frac{q}{p}\right)^{S_0}\right] &= \left(\frac{q}{p}\right)^{k} \\ &= p_k\left(\frac{q}{p}\right)^{N} + (1-p_k)\left(\frac{q}{p}\right)^{0} \\ &= 1 - p_k\left(1-\frac{q}{p}\right)^{N}\end{aligned}$$

therefore

$$p_k = \frac{1-\left(\frac{q}{p}\right)^{k}}{1-\left(\frac{q}{p}\right)^{N}}$$

and the probability of ruin is

$$1 - \frac{1-\left(\frac{q}{p}\right)^{k}}{1-\left(\frac{q}{p}\right)^{n}}$$

Theorem 1.9 (Optional Stopping)**.** *Let τ be a stopping time and $(M_n : n \in \mathbb{N}_0)$ be a martingale adapted to $(\mathcal{F}_n : n \in \mathbb{N}_0)$ such that one of the following three conditions holds:*

1. *τ is a.s. bounded*
2. *$\mathbb{E}[\tau] < \infty$ and $\exists\, c \in \mathbb{R}^+ : \mathbb{E}[|M_{t+1} - M_t| \,|\, \mathcal{F}_t] \leqslant c$ a.s. on the event $\{\tau > t\}$, for all $t \in \mathbb{N}$.*
3. *$\exists\, c \in \mathbb{R}^+ : \left|M_{\min\{t,\,\tau\}}\right| \leqslant c$ a.s. for all $t \in \mathbb{N}$.*

Then M_τ is a.s. well defined and

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$$

Proof.

$$M_{\min\{t,\tau\}} = M_0 + \sum_{s=0}^{\min\{\tau-1,\,t-1\}} (M_{s+1} - M_s)$$

yields

$$\left|M_{\min\{t,\tau\}}\right| \leqslant M \coloneqq |M_0| + \sum_{s=0}^{\tau-1} |M_{s+1} - M_s| = |M_0| + \sum_{s=0}^{\infty} |M_{s+1} - M_s| \cdot \mathbf{1}_{\{\tau > s\}}$$

Hence by the monotone convergence theorem,

$$\mathbb{E}[M] = \mathbb{E}[|X_0|] + \sum_{s=0}^{\infty} \mathbb{E}\left[|M_{s+1} - M_s| \cdot \mathbf{1}_{\{\tau > s\}}\right]$$

If (1) holds then the series above only has a finite number of non-zero terms, hence M is integrable. If (2) holds, then

$$\begin{aligned}\mathbb{E}[M] &= \mathbb{E}[|X_0|] + \sum_{s=0}^{\infty} \mathbb{E}\left[\underbrace{\mathbb{E}[|M_{s+1} - M_s| \,|\, \mathcal{F}_s] \cdot \mathbf{1}_{\{\tau > s\}}}_{\leqslant c\mathbf{1}_{\{\tau>s\}} \text{ a.s. by (2)}}\right] \\ &\leqslant \mathbb{E}[|X_0|] + c\sum_{s=0}^{\infty} \mathbb{P}(\tau > s) \\ &= \mathbb{E}[|X_0|] + c\,\mathbb{E}[\tau] < \infty\end{aligned}$$

and if (3) holds then $M \coloneqq c$. Therefore if any of the three conditions hold, then the stopped process is dominated by an integrable M, and converges a.s. to M_τ, the dominated convergence theorem implies that

$$\mathbb{E}[M_\tau] = \lim_{t\to\infty} \mathbb{E}\left[M_{\min\{t,\,\tau\}}\right]$$

By the martingale property of the stopped process,

$$\mathbb{E}\left[M_{\min\{t,\,\tau\}}\right] = \mathbb{E}[M_0]$$

hence

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$$

□

Remark. Under the third condition, $\mathbb{P}(\{\tau = \infty\})$ may be positive. On this event M_τ is defined as the a.s. pointwise limit of M_t as $t \to \infty$.

Corollary 1.10. *Similarly, if $(M_n : n \in \mathbb{N}_0)$ is a submartingale and one of the conditions of the previous theorem above holds, then it follows directly that*

$$\mathbb{E}[M_\tau] \geqslant \mathbb{E}[M_0]$$

and if the process is a supermartingale then

$$\mathbb{E}[M_\tau] \leqslant \mathbb{E}[M_0]$$

Example 1.11. We aim to use optional stopping to compute the expected amount of time Gambler's ruin runs for starting at \$K. Let $(U_n : n \in \mathbb{N})$ i.i.d. such that $U_n \sim \mathcal{U}_{\{-1,1\}}$, $\mathbb{F} = (\mathcal{F}_n : n \in \mathbb{N}) = \sigma(\{U_i \mid 1 \leqslant i \leqslant n\})$ and S_n be the total amount of dollars gained so far at timestep $n \in \mathbb{N}_0$. Define

$$S_0 \coloneqq k \qquad S_n \coloneqq \sum_{i=1}^{n} U_i$$

Assume that $p = q$ and consider $\left(S_n^2 - n : n \in \mathbb{N}\right)$ then

$$\begin{aligned}
\mathbb{E}\left[S_{n+1}^2 - (n+1) \,\middle|\, \mathcal{F}_n\right] &= \mathbb{E}\left[(S_n + U_{n+1})^2 - (n+1) \,\middle|\, \mathcal{F}_n\right] \\
&= \mathbb{E}\left[S_n^2 + 2S_n U_{n+1} + X_{n+1}^2 - n - 1 \,\middle|\, \mathcal{F}_n\right] \\
&= S_n^2 + 2S_n \mathbb{E}[U_{n+1}] - n \\
&= S_n^2 - n
\end{aligned}$$

Hence $\left(S_n^2 - n : n \in \mathbb{N}\right)$ is a martingale. Let $\tau = \inf\{n \in \mathbb{N} : S_n \in \{0, N\}\}$ then

$$\begin{aligned}
\mathbb{E}\left[S_\tau^2 - \tau\right] &= \mathbb{E}\left[S_0^2 - 0\right] = k^2 \\
&= \mathbb{E}\left[S_\tau^2\right] - \mathbb{E}[\tau] \\
&= N^2 \left(\frac{k}{N}\right) - \mathbb{E}[\tau] \\
&= Nk - \mathbb{E}[\tau]
\end{aligned}$$

Therefore $\mathbb{E}[T] = k(N-k)$. Assume that $p \neq q$. Let $(U_n : n \in \mathbb{N})$ i.i.d. such that $\mathbb{P}(U_n = 1) = p$ and $\mathbb{P}(U_n = 1) = 1-p =: q$. Set $\mathbb{F} = (\mathcal{F}_n : n \in \mathbb{N}) = \sigma(\{U_i \mid 1 \leqslant i \leqslant n\})$ and consider $(S_n - n(p-q) : n \in \mathbb{N})$ then

$$\begin{aligned}
\mathbb{E}[S_{n+1} - (n+1)(p-q) \mid \mathcal{F}_n] &= \mathbb{E}[S_n + U_{n+1} - (n+1)(p-q) \mid \mathcal{F}_n] \\
&= S_n + \mathbb{E}[U_{n+1}] - (n+1)(p-q) \\
&= S_n + p - q - (n+1)(p-q) \\
&= S_n - n(p-q)
\end{aligned}$$

Hence $(S_n - n(p-q) : n \in \mathbb{N})$ is martingale.

$$\begin{aligned}
\mathbb{E}[S_T - T(p-q)] &= \mathbb{E}[S_0 - 0(p-q)] = \mathbb{E}[S_0] = k \\
&= \mathbb{E}[S_T] - \mathbb{E}[T](p-q) \\
&= N \cdot p_k - \mathbb{E}[T](p-q) \\
&= N \cdot \frac{1 - \left(\frac{q}{p}\right)^k}{1 - \frac{q}{p}} - \mathbb{E}[T](p-q)
\end{aligned}$$

therefore

$$\mathbb{E}[T] = \frac{k}{q-p} - \left(\frac{n}{q-p}\right)\left(\frac{1 - \left(\frac{q}{p}\right)^k}{1 - \frac{q}{p}}\right)$$

Definition 1.12 (Predictable)**.** A stochastic process is predictable if X_0 is fixed and X_n is F_{n-1} measurable.

Remark. This is a strictly stronger condition than being adapted to a filtration.

Theorem 1.13 (Doob's Decomposition Theorem)**.** *Let $(X_n : n \in \mathbb{N}_0)$ be a process in L^1 adapted to $(\mathcal{F}_n : n \in \mathbb{N}_0)$. Then it can be uniquely decomposed as $X_n = M_n + A_n$ where $(M_n : n \in \mathbb{N}_0)$ is a martingale and $(A_n : n \in \mathbb{N}_0)$ is predictable such that $A_0 \coloneqq 0$. Furthermore,*

$$A_n = \sum_{k=1} \mathbb{E}[X_k - X_{k-1} \,|\, \mathcal{F}_{k-1}] = \sum_{k=1}^{n} \mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}] - X_{k-1}$$

and $(A_n : n \in \mathbb{N}_0)$ is called the compensator of $(X_n : n \in \mathbb{N}_0)$.

Proof. Existence: Let $(A_n : n \in \mathbb{N}_0)$ as above and $\forall\, n \in \mathbb{N}_0$:

$$M_n = X_0 + \sum_{k=1}^{n} X_k - \mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}]$$

First note that the sums for $n = 0$ are empty, and defined to be zero. Moreover, note that A adds up the expected increments of X, and M adds the part for every X_k that is not known one time step before. By definition, A_{n+1} and M_n are $\mathcal{F}_n$ measurable because the process X is adapted. Moreover, $\mathbb{E}[|A_n|] < \infty$ and $\mathbb{E}[|M_n|] < \infty$ since the process X is integrable. Furthermore, the decomposition $X_n = M_n + A_n$ holds for all $n \in \mathbb{N}_0$. Finally,

$$\begin{aligned}
\mathbb{E}[M_{n+1}|F_n] &= \mathbb{E}[X_{n+1} - A_{n+1}|F_n] \\
&= \mathbb{E}[X_{n+1}|F_n] - A_{n+1} \\
&= \mathbb{E}[X_{n+1}|F_n] - \mathbb{E}[X_{n+1}|F_n] + X_n - A_n \\
&= X_n - A_n \\
&= M_n
\end{aligned}$$

Hence $(M_n : n \in \mathbb{N}_0)$ is a martingale.

Uniqueness: Let $X = M' + A'$ be an additional decomposition, then $Y \coloneqq M - M' = A' - A$ is a martingale. Thus

$$\forall\, n \in \mathbb{N} : \mathbb{E}[Y_n \,|\, \mathcal{F}_{n-1}] = Y_{n-1}$$

hence Y is also predictable, thus

$$\forall\, n \in \mathbb{N} : \mathbb{E}[Y_n \,|\, \mathcal{F}_{n-1}] = Y_n$$

Since $Y_0 = A'_0 - A_0 = 0$, then $Y_n = 0$ almost surely, for all $n \in \mathbb{N}$. Therefore the decomposition is almost surely unique. □

Corollary 1.14. *An adapted process X in L^1 is a sub-martingale if and only if it has a Doop decomposition into a martingale M and an integrable predictable A that is almost surely non-decreasing. Similarly, X is a super-martingale if and only if A is almost surely non-increasing.*

Proof. If X is a sub-martingale then

$$\begin{aligned}
X \text{ is a sub-martingale} &\iff \mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}] \geqslant X_{k-1} \\
&\iff \sum_{k=1}^{n} (\mathbb{E}[X_k \,|\, \mathcal{F}_{k-1}] - X_{k-1}) \geqslant 0 \\
&\iff A \text{ is almost surely non-decreasing}
\end{aligned}$$

The equivalence for super-martingales is proved similarly. □

Definition 1.15. For a martingale $(M_n : n \in \mathbb{N})$ in L^2, the bracket process $([M_n] : n \in \mathbb{N}_0)$ is defined as the compensator of $\left(M_n^2 : n \in \mathbb{N}_0\right)$, i.e.,

$$\begin{aligned}
[M_n] &= \mathbb{E}\left[M_0^2\right] + \sum_{k=1}^{n} \mathbb{E}\left[M_k^2 - M_{k-1}^2 \,\middle|\, \mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[M_0^2\right] + \sum_{k=1}^{n} \mathbb{E}\left[(M_k - M_{k-1})^2 \,\middle|\, \mathcal{F}_{k-1}\right]
\end{aligned}$$

Definition 1.16. $(M_n : n \in \mathbb{N})$ is bounded in L^p if

$$\sup_{n \in \mathbb{N}} \|X_n\|_p < \infty$$

Definition 1.17. Let $(X_n : n \in \mathbb{N})$ be a supermartingale. Fix $a < b$, define $T_0 = 0$ and let

$$T_{2k+1} = \inf\{n \geqslant T_{2k} : X_n \leqslant a\} \quad T_{2k+2} = \inf\{n \geqslant T_{2k+1} : X_n \leqslant b\}$$

for all $k \in \mathbb{N}$. The number of upcrossings is then defined as

$$u(a, b) \coloneqq \sup\{k \in \mathbb{N} \mid T_{2k} < \infty\}$$

Remark. In other words, the number of upcrossings is the number of times the supermartingale goes from below a to above b. Due to stochasticity, there may be some accidents, i.e., the presence of upcrossings, but since a supermartingale is expected to decrease, these upcrossings have a cost and it can be proved that the number of upcrossings is almost surely finite. The fact that the supermartingale is bounded in L^1 also prevents the sample paths from drifting off to minus infinity; therefore, almost all the sample paths must converge. We now turn this heuristic argument into a proof, starting with the upcrossing inequality.

Lemma 1.18 (Upcrossing Inequality)**.** *In the context of the above,*

$$\forall k \in \mathbb{N} : \mathbb{P}(u(a, b) > k) \leqslant \mathbb{P}\frac{1}{b - a}\mathbb{E}\left[(X_\infty - a)^- \mathbf{1}_{\{u(a,b)=k\}}\right]$$

Remark. X_n^- denotes the negative part of X_n, i.e., $X_n^- = -\min\{X_n, 0\}$. X_n^+ is defined analogously.

Proof. Since the process $(X_n - a : n \in \mathbb{N})$ is also a supermartingale, it suffices to show the result holds when $a = 0$. Let $(\sigma_1, \sigma_2) = (T_{2k+1}, T_{2k+2})$, then

$$\begin{aligned}
&\{u(a, b) > k\} = \{\sigma_2 < \infty\} \subseteq \{\sigma_1 < \infty\} \cap \{X_{\sigma_2} \geqslant b\} \\
\implies &\mathbb{P}(u(0, b) > k) = \mathbb{E}\left[\mathbf{1}_{\{u(0,b)>k\}}\right] \\
&\qquad \leqslant \frac{1}{b}\,\mathbb{E}\left[X_{\sigma_2}\mathbf{1}_{\{u(0,b)>k\}}\right] \\
&\qquad \leqslant \frac{1}{b}\,\mathbb{E}\left[X_{\sigma_2}^+\mathbf{1}_{\{u(0,b)>k\}}\right] \\
&\qquad \leqslant \frac{1}{b}\,\mathbb{E}\left[X_{\sigma_2}^+\mathbf{1}_{\{\sigma_1<\infty\}}\right]
\end{aligned}$$

Since $(X_n : n \in \mathbb{N})$ is a supermartingale and $X_{\sigma_1} \leqslant 0$ on the event $\{\sigma_1 < \infty\}$, it follows from the optional stopping theorem that

$$\begin{aligned}
\mathbb{E}\left[X_{\sigma_2}^+\mathbf{1}_{\{\sigma_1<\infty\}}\right] - \mathbb{E}\left[X_{\sigma_2}^-\mathbf{1}_{\{\sigma_1<\infty\}}\right] &= \mathbb{E}\left[X_{\sigma_2}\mathbf{1}_{\{\sigma_1<\infty\}}\right] \\
&\leqslant \mathbb{E}\left[X_{\sigma_1}\mathbf{1}_{\{\sigma_1<\infty\}}\right] \\
&\leqslant 0
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}(u(0, b) > k) &\leqslant \frac{1}{b}\,\mathbb{E}\left[X_{\sigma_2}^-\mathbf{1}_{\{\sigma_1<\infty\}}\right] \\
&\leqslant \frac{1}{b}\,\mathbb{E}\left[X_{\sigma_2}^-\mathbf{1}_{\{\sigma_1<\infty,\, X_{\sigma_2}\leqslant 0\}}\right] \\
&\leqslant \frac{1}{b}\,\mathbb{E}\left[X_{\sigma_2}^-\mathbf{1}_{\{\sigma_1<\infty,\, \sigma_2=\infty\}}\right] \\
&\leqslant \frac{1}{b}\,\mathbb{E}\left[X_\infty^-\mathbf{1}_{\{u(0,b)=k\}}\right]
\end{aligned}$$

□

Theorem 1.19 (Martingale Convergence Theorem)**.** *Let* $(X_n : n \in \mathbb{N})$ *be a supermartingale bounded in* L^1*, i.e.,*

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|] < \infty$$

then

$$\begin{aligned}
X_n \xrightarrow[\text{a.s.}]{n\to\infty} X_\infty &< \infty \\
&\in L^1
\end{aligned}$$

Proof. Set $n \in \mathbb{N}$ and denote by $u_n(a,b)$ the number of upcrossings that occur by time n. Applying the upcrossing inequality to the process stopped at time n, we obtain

$$\begin{aligned}\mathbb{E}[u_n(a,b)] &= \sum_{k\in\mathbb{N}} \mathbb{P}(u_n(a,b) > k) \\ &\leqslant \frac{1}{b-a} \sum_{k\in\mathbb{N}} \mathbb{E}\big[(X_n - a)^- \mathbf{1}_{\{u_n(a,b)=k\}}\big] = \frac{1}{b-a}\mathbb{E}\big[(X_n-a)^-\big]\end{aligned}$$

From the monotone convergence theorem, it follows that

$$\begin{aligned}\mathbb{E}[u(a,b)] &= \lim_{n\to\infty} \mathbb{E}[u_n(a,b)] \\ &\leqslant \frac{1}{b-a} \lim_{n\to\infty} \mathbb{E}\big[(X_n-a)^-\big] \\ &\leqslant \frac{1}{b-a} \sup\{\mathbb{E}[(X_n + |a|)] \,|\, n\in\mathbb{N}\} \\ &< \infty\end{aligned}$$

hence $\mathbb{P}(u(a,b) < \infty) = 1$ for all $a < b$. Since $\mathbb{Q}$ is countable, we also have

$$\begin{aligned}\mathbb{P}\left(\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\right) &= \mathbb{P}\left(\left\{\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n \;\middle|\; a,b\in\mathbb{Q}\right\}\right) \\ &= \mathbb{P}(\{u(a,b) = \infty \,|\, a,b\in\mathbb{Q} : a < b\})\end{aligned}$$

which proves almost sure convergence

$$X_n \xrightarrow[\text{a.s.}]{n\to\infty} X_\infty$$

where

$$\mathbb{E}[|X_\infty|] \leqslant \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|] < \infty$$

□

Lemma 1.20. *Let $(M_n : n\in\mathbb{N}_0)$ be a martingale in L^2, and let $s < t \leqslant u < v$ then*

$$\mathbb{E}[(M_t - M_s)(M_v - M_u)] = 0$$

Proof. Since $M_u = \mathbb{E}[M_v \,|\, \mathcal{F}_u]$ and $M_t - M_s \in \mathcal{F}_u$ then

$$\begin{aligned}\mathbb{E}[(M_t - M_s)(M_v - M_u)] &= \mathbb{E}[\mathbb{E}[(M_t - M_s)(M_v - M_u) \,|\, \mathcal{F}_u]] \\ &= \mathbb{E}\left[(M_t - M_s)\underbrace{(\mathbb{E}[M_v \,|\, \mathcal{F}_u] - M_u)}_{=0}\right] \\ &= 0\end{aligned}$$

□

Theorem 1.21. *An L^2 martingale (M_n) is bounded in L^2 if and only if*

$$\sum_{k=1}^{\infty} \mathbb{E}\Big[(M_k - M_{k-1})^2\Big] < \infty$$

In this case, $M_n \xrightarrow[\text{a.s.}]{n\to\infty} M_\infty \in L^2$.

Proof. Due to the orthogonal increments as above, all cross terms of the square below have zero mean, and thus the following holds:

$$\begin{aligned}\mathbb{E}\big[M_n^2\big] &= \mathbb{E}\left[\left(M_0 + \sum_{k=1}^{n}(M_k - M_{k-1})\right)^2\right] \\ &= \mathbb{E}\big[M_0^2\big] + \sum_{k=1}^{n} \mathbb{E}\Big[(M_k - M_{k-1})^2\Big] \xrightarrow{n\to\infty} \mathbb{E}\big[M_0^2\big] + \sum_{k=1}^{\infty} \mathbb{E}\Big[(M_k - M_{k-1})^2\Big]\end{aligned}$$

Hence M_n is bounded in L^2 if and only if $\sum_{k=1}^{\infty} \mathbb{E}\left[(M_k - M_{k-1})^2\right] < \infty$. Moreover, since M_n is bounded in L^2 it follows that it is also bounded in L^1. Thus by the martingale convergence theorem in L^1,

$$M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty \in L^1$$

To see the L^2 convergence, use Fatou's lemma as

$$\begin{aligned}
\mathbb{E}\left[(M_\infty - M_n)^2\right] &= \mathbb{E}\left[\lim_{r\to\infty} (M_{n+r} - M_n)^2\right] \\
&\leqslant \liminf_{r\to\infty} \mathbb{E}\left[(M_{n+r} - M_n)^2\right] \\
&= \liminf_{r\to\infty} \sum_{k=n+1}^{n+r} (M_k - M_{k-1})^2 \\
&= \sum_{k=n+1}^{\infty} \mathbb{E}\left[(M_k - M_{k-1})^2\right] \xrightarrow{n\to\infty} 0
\end{aligned}$$

due to finiteness of the infinite sum. □

Corollary 1.22. *For a martingale $(M_n : n \in \mathbb{N})$ in L^2, consider $[M_n]$. Since $([M_n] : n \in \mathbb{N})$ is a.s. non-decreasing, $[M_\infty] \coloneqq \lim_{n\to\infty} [M_n]$ exists a.s. and may be infinite. On the event, $\{[M_\infty] < \infty\}$, we have that*

$$M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty < \infty$$

Proof. Let $k \geqslant 0$ and

$$\tau_k = \inf\{n \in \mathbb{N}_0 \mid [M]_n \geqslant k\}$$

Note that since $[M]$ is predictable then τ_k is a stopping time. Thus

$$\left[M_{\min\{n,\,\tau_k\}}\right] = [M]_{\min\{n,\,\tau_k\}}$$

is bounded by k so

$$\mathbb{E}\left[\left(M_{\min\{n,\,\tau_k\}}\right)^2\right] \leqslant k$$

hence $\left(M_{\min\{n,\,\tau_k\}} : n \in \mathbb{N}_0\right)$ is bounded in L^2, thus it bounded in L^1, and therefore it converges almost surely as $n \to \infty$ in L^1. In particular, on

$$\{[M]_\infty < k\} \subseteq \{\tau_k = \infty\}$$

we have

$$\begin{aligned}
M_{\min\{n,\,\tau_k\}} \xrightarrow[n\to\infty]{\text{a.s.}} M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty < \infty \\
\in L^2
\end{aligned}$$

Therefore M_∞ exists almost surely on

$$\bigcup_{k=0}^{\infty} \{[M]_\infty < k\} = \{[M]_\infty < \infty\}$$

□

Definition 1.23 (Stochastic Approximation)**.** Let $(X_n : n \in \mathbb{N})$ be a stochastic process in the euclidean space $\mathbb{R}^n$ adapted to a filtration $(\mathcal{F}_n : n \in \mathbb{N})$. Suppose that X_n satisfies

$$X_{n+1} - X_n = \frac{1}{n}(F(X_n) + \xi_{n+1} + R_n)$$

where

- $F : \mathbb{R}^n \to \mathbb{R}^n$
- $\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0$
- and the remainder terms $\mathcal{F}_n \ni R_n \xrightarrow{n\to\infty} 0$ and satisfy $\sum_{n=1}^{\infty} \frac{|R_n|}{n} < \infty$ almost surely.

then such a process is known as a stochastic approximation process.

Remark. Such processes are commonly used to approximate the root of an unknown function in the setting where evaluation queries may be made but the answers are noisy.

2 Galton-Watson Branching Processes

These notes are compiled from [16], [11], [19], [9] and [2].

Percolation on a tree breaks up the tree into random subtrees. Historically, the first random trees to be considered were a model of genealogical (family) trees. Since such trees will be an important source of examples and an important tool in later work, we too will consider their basic theory before turning to percolation. They are also beautiful processes in themselves.

Galton-Watson branching processes are most often defined as Markov chains $(Z_n : n \in \mathbb{N}_0)$, where Z_n represents the size of the n^{th} generation of a family, but we will be interested as well in the underlying family trees. Given numbers $p_k \in [0,1]$ with $\sum_{k\geqslant 0} p_k = 1$, the process is defined as follows. We start with one particle $Z_0 \equiv 1$, unless specified otherwise. It has k children with probability p_k. Then each of these children (should there be any) also has children with the same progeny (or "offspring") distribution $(p_k : k \in \mathbb{N}_0)$, independently of the others and of its parent. This continues forever or until there are no more children. To be formal, let L be a random variable with $\mathbb{P}(L = k) = p_k$, and let $\left(L_i^{(n)} : n,\, i \in \mathbb{N}\right)$ be independent copies of L. The generation sizes of the branching i process are then defined inductively by

$$Z_{n+1} \coloneqq \sum_{i=1}^{Z_n} L_i^{(n+1)}$$

The probability generating function (p.g.f.) of L is defined as

$$f(s) \coloneqq \mathbb{E}\left[s^L\right] = \sum_{k\in\mathbb{N}_0} p_k s^k$$

Note that unless specified otherwise f is defined on $[0,1]$. Note that we interpret $0^0 = 1$ such that $f(0) = \mathbb{P}(L = 0) = p_0$. We call the event $\{\exists\, n \in \mathbb{N}_0 : Z_n = 0\}$ the extinction, which of course is the same as the event $\left\{Z_n \xrightarrow{n\to\infty} 0\right\}$. We will often omit the superscripts on L when not needed. The family (or genealogical) tree associated to a branching process is obtained simply by having one vertex for each particle ever produced and joining two by an edge if one is the parent of the other. We will give a formal definition later (in section 3) of trees and the associated probability measures on them. The first basic result on Galton-Watson processes is that on the event of non-extinction, the population size explodes, except in the trivial case that $p_1 = 1$,

Proposition 2.1. *On the event of non-extinction,* $Z_n \xrightarrow[\text{a.s.}]{n\to\infty} \infty$ *provided that* $p_1 = 1$.

Proof. We want to see that 0 is the only non-transient state of the Markov chain $(Z_n : n \in \mathbb{N}_0)$. If $p_0 = 0$, it is clear, whereas if $p_0 > 0$, then from any state $k \geqslant 1$, eventually returning to k requires not immediately becoming extinct, whence it has probability $\leqslant 1 - p_0^k < 1$. □

What is $q \coloneqq \mathbb{P}(\text{extinction})$? To find out, we use the following property of the p.g.f.

Proposition 2.2.

$$\forall\, s \in [0,1] : \mathbb{E}\left[s^{Z_n}\right] = \underbrace{(f \circ \cdots \circ f)}_{n \text{ times}}(s) =: f^{(n)}(s)$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[s^{Z_n}\right] &= \mathbb{E}\left[\mathbb{E}\left[s^{\sum_{i=1}^{Z_{n-1}} L_i}\right] \middle| Z_{n-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{Z_{n-1}} s^{L_i} \middle| Z_{n-1}\right]\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{Z_{n-1}} \mathbb{E}[\exp(L_i)]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[s^L\right]^{Z_{n-1}}\right] \\
&= \mathbb{E}\left[f(s)^{Z_{n-1}}\right]
\end{aligned}$$

where the random variables $L_i \coloneqq L_i^{(n)}$ are independent of each other and of Z_{n-1} and have the same distribution as L. Iterate this equation n times. □

Remark. Note that within this proof is the identity

$$\mathbb{E}\left[s^{Z_n} \,\middle|\, Z_{0:n-1}\right] = f(s)^{Z_{n-1}}$$

Corollary 2.3 (Extinction Probability)**.** *The extinction probability is* $q = \lim_{n\to\infty} f^{(n)}(0)$.

Proof. Since extinction is the increasing union of the events $\{\exists\, n \in \mathbb{N}_0 : Z_n = 0\}$, it follows that

$$q = \lim_{n\to\infty} P(Z_n = 0) = \lim_{n\to\infty} f^{(n)}(0)$$

□

We finally discover the most used result in the field and value of q,

Proposition 2.4 (Extinction Criterion [16] (Proposition 5.4))**.** *Provided* $p_1 \neq 1$*, we have*

- $q = 1 \iff f'(1) \leqslant 1$*, and*
- q *is the smallest root of* $f(s) = s \in [0, 1]$*, the only other possible root being 1.*

Remark. When we differentiate f at 1, we mean the left-hand derivative. Note that

$$f'(1) = \mathbb{E}[L] =: m = \sum_{k\in\mathbb{N}_0} k\, p_k$$

is the mean number of offspring. We call m simply the mean of the branching process.

By the proposition above, a branching process is called subcritical if $m < 1$, critical if $m = 1$, and supercriticial otherwise.

How quickly does $Z_n \xrightarrow{n\to\infty} \infty$ on the event of non-extinction? The most naive guess would be that it grows approximately like m^n. This is essentially correct. Our first result is that a martingale appears when we divide Z_n by m^n.

Proposition 2.5. *If* m *is finite then* $\left(\frac{Z_n}{m^n} : n \in \mathbb{N}_0\right)$ *is a martingale.*

Proof.

$$\mathbb{E}\left[\frac{Z_{n+1}}{m^{n+1}} \,\middle|\, Z_n\right] = \mathbb{E}\left[\frac{1}{m^{n+1}} \sum_{i=1}^{Z_n} L_i \,\middle|\, Z_n\right] = \frac{1}{m^{n+1}} \sum_{i=1}^{Z_n} \mathbb{E}[L_i \mid Z_n] = \frac{1}{m^{n+1}} \sum_{i=1}^{Z_n} m = \frac{Z_n}{m^n}$$

□

Remark. Actually, we have not verified that we are computing conditional expectations of integrable random variables. One way to avoid calculating (in a similar manner) the unconditional expectation first is to note that all random variables are non-negative. Another way is to use the fact that Z_n takes only countably many values, so that we may work with expectations conditioned on events, rather than on a random variable.

Since the martingale above is non-negative, it has a finite limit a.s. denoted W. Thus, when $W > 0$ the generation sizes Z_n grow as expected, i.e., like m^n up to a random factor. Otherwise, they grow more slowly. Our attention is thus focused on the following two questions

1. When is $W > 0$?
2. When $W = 0$ and the process does not become extinct, what is the rate at which $Z_n \xrightarrow{n\to\infty} \infty$?

To answer these questions, we first note a general zero-one property of Galton-Watson branching processes. Call a property of trees inherited if every finite tree has this property and if whenever a tree has this property, so do all the descendant trees of the children of the root.

Proposition 2.6. *Every inherited property has conditional probability either 0 or 1 given non-extinction.*

Proof. Let A be the set of trees possessing a given inherited property. For a tree T with k children from the root, denote $T^{(1)}, \ldots, T^{(k)}$ as the descendant trees of these children. Then

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{P}(T \in A \mid Z_1)] \leqslant \mathbb{E}\left[\mathbb{P}\left(T^{(1)}, \ldots, T^{(Z_1)} \in A \,\middle|\, Z_1\right)\right]$$

by definition of inheritance. Since $T^{(1)}, \ldots, T^{(Z_1)}$ are i.i.d. given Z_1, the last quantity in the display is equal to

$$\mathbb{E}\left[\mathbb{P}(A)^{Z_1}\right] = f(\mathbb{P}(A))$$

Thus, $\mathbb{P}(A) \leqslant f(\mathbb{P}(A))$. On the other hand, $\mathbb{P}(A) \geqslant q$, since every finite tree is in A. Hence $\mathbb{P}(A) \in \{q, 1\}$, from from which the desired conclusion follows. □

Corollary 2.7. *Suppose m is finite, then $W = 0$ or $W > 0$ a.s. on non-extinction. In other words, $\mathbb{P}(W = 0) \in \{q, 1\}$.*

Proof. The property that $W = 0$ is clearly inherited, whence this is an immediate consequence of the previous proposition. □

In answer to the preceding two questions, we have the following two theorems.

Theorem 2.8 ([11] (Kesten-Stigum Theorem, 1966))**.** *The following are equivalent provided* $1 \leqslant m < \infty$

1. $\mathbb{P}(W = 0) = q$
2. $\mathbb{E}[W] = 1$
3. $\mathbb{E}\left[L \log^+ L\right] < \infty$

Remark. Since (3) requires barely more than the existence of a mean, generation sizes *typically* do grow as expected. However, when (3) fails the means m^n overestimate the rate of growth. Yet there is still an essentially deterministic rate of growth, as shown by Seneta (1968) [19] and Heyde (1970) [9], which is only slightly less than m^n.

Theorem 2.9 (Seneta-Heyde Theorem)**.** *If* $1 \leqslant m < \infty$ *then* $\forall n \in \mathbb{N}_0\ \exists c_n \in \mathbb{R}$ *such that*

1. $\lim_{n\to\infty} \frac{Z_n}{c_n}$ *exists a.s. in* $[0, \infty)$
2. $\mathbb{P}\left(\lim_{n\to\infty} \frac{Z_n}{c_n} = 0\right) = q$
3. $\frac{c_{n+1}}{c_n} = m$

Proof. We will find another martingale to do our work. Choose $s_0 \in (q, 1)$ and set $s_{n+1} \coloneqq f^{-1}(s_n)$ for $n \geqslant 0$. Then $s_n \uparrow 1$. By proposition 4.2, we have that $\left(s^{Z_n} : n \in \mathbb{N}_0\right)$ is a martingale. Being positive and bounded, it converges a.s. and in L^1 to a limit $Y \in [0, 1]$ such that $\mathbb{E}[Y] = \mathbb{E}\left[s_0^{Z_0}\right] = s_0$. Now we can re-formulate these exponentials. Set $c_n \coloneqq -\frac{1}{\log s_n}$ then $s^{Z_n} = \exp\left(-\frac{Z_n}{c_n}\right)$, so that $\lim_{n\to\infty} \frac{Z_n}{c_n}$ exists a.s. in $[0, \infty)$. By l'Hopital's Rule and the fact that $\lim_{s\uparrow 1} f'(s) = m$,

$$\lim_{s\uparrow 1} \frac{-\log f(s)}{-\log s} = \lim_{s\uparrow 1} \frac{f'(s)s}{s} = m$$

Considering this limit along the sequence $(s_n : n \in \mathbb{N}_0)$ we get (3). It follows form (3) that the property that $\frac{Z_n}{c_n} = 0$ is inherited, whence by proposition 4.5 and the fact that $\mathbb{E}[Y] = s_0 < 1$, we deduce (ii). Likewise, the property that $\frac{Z_n}{c_n} < \infty$ is inherited and has probability 1 since $\mathbb{E}[Y] > q$, which implies (1). □

Remark. The proof of the Seneta-Heyde theorem gives a prescription for calculating the constants c_n but does not immediately provide estimates for them. Another approach gives a different prescription that leads sometimes to an explicit estimate (see Asmussen and Hering (1983) [2], pages 45 to 49).

We will often want to consider random trees produced by a Galton-Watson branching process. Up to now, we have avoided that by giving theorems just about the random variables Z_n (except for proposition 4.5, but that was used so far only for studying the limiting behavior of Z_n). One approach to formalize tree-valued random variables is as follows. A rooted labeled tree T is a nonempty collection of finite sequences of positive integers such that if $(i_1, \ldots, i_n) \in T$ then

1. $\forall k \in [0, n]$, also the initial segment $(i_1, \ldots, i_k) \in T$, here the case $k = 0$ is interpreted as the empty sequence, and
2. $\forall j \in [1, i_n]$ the sequence $(i_1, \ldots, i_{n-1}, j) \in T$.

The root of the tree is the empty sequence, $\emptyset$. Thus if $(i_1, \ldots, i_n)$ is the i_n^{th} child of the i_{n-1} of $\ldots$ of the i_1^{th} child of the root. If $x = (i_1, \ldots, i_n) \in T$ then we define $T^x \coloneqq \{(j_1, \ldots, j_k) \mid (i_1, \ldots, i_n, j_1, \ldots, j_k) \in T\}$ to be the descendant tree of the vertex x in T. The height of a tree is the supremum of the lengths of the sequences in the tree. If T is a tree and $n \in \mathbb{N}$, denote the truncation of T to its first n levels as $T^{(n)} \coloneqq \{(i_1, \ldots, i_k) \in T \mid k \leqslant n\}$. This is a tree of height of at most n. A tree is called locally finite if its truncation to every finite level is finite. Let $\mathcal{T}$ be the space of rooted labeled locally finite trees. Finally, we define a metric on $\mathcal{T}$ by setting

$$d\left(T, \tilde{T}\right) \coloneqq \left(1 + \sup\left\{n \in \mathbb{N}_0 \,\middle|\, T^{(n)} = \tilde{T}^{(n)}\right\}\right)^{-1}$$

3 Stochastic Processes on Trees

These notes are compiled from [12], unless specified otherwise.

3.1 Motivation: Broadcasting on Trees

Let $T_n = (V_N, E_N)$ be a binary tree of depth N, rooted at 0. Note that $\forall x, y \in V_n$ there exists a unique self-avoiding path $x = x_0 \sim x_1 \sim \cdots \sim x_n = y$ of neighboring vertices $x_1, \ldots, x_n \in V$, where $\forall a, b \in V_N : a \sim b \iff (a, b) \in E_N$. The path length $n =: d(x, y)$ defines the tree metric. Define the finite-volume state of the model to be $\Omega_N \coloneqq \{-1, 1\}^{|V_N|}$. Together with the product sigma algebra (which is simply the power set on our finite set) this defines our probability space. Denote by $\sigma \in \Omega_n$ a configuration. We define a probability measure μ on Ω_N in two steps. First, the distribution of the spin at the origin is chosen to be a symmetric Bernoulli, i.e.,

$$\mu(\sigma_0 = 1) = \mu(\sigma_0 = -1) = \frac{1}{2}$$

Next, we pass on information from the root to the outside of the tree by putting for all pairs of neighboring vertices $v \to w$, meaning v is the parent of w, i.e., v is closer to the root than w,

$$\mu(\sigma_w = -1 \mid \sigma_v = 1) = \mu(\sigma_w = 1 \mid \sigma_v = -1) = \varepsilon \in \left[0, \frac{1}{2}\right]$$

where ε is an error parameter, and the full probability distribution is obtained by applying this rule from the root to the outside of the tree. In this way, we have the following probability distribution on our finite-volume state space.

Definition 3.1. The probability measure defined by

$$\begin{aligned}\mu(\sigma) &= \frac{1}{2} \prod_{v, w \,:\, v \to w} (1-\varepsilon)^{\mathbf{1}_{\sigma_v = \sigma_w}} \cdot \varepsilon^{\mathbf{1}_{\sigma_v \neq \sigma_w}} \\ &= \frac{1}{2} \prod_{v, w \,:\, v \to w} P_{\sigma_v, \sigma_w}\end{aligned}$$

with

$$P = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$$

is called the symmetric channel on the binary tree.

Remark. This is a specific example of a tree-indexed Markov chain.

We can imagine to replace P by another transition matrix to obtain a different distribution, and we can generalize the local state space. Note that $\varepsilon = \frac{1}{2} \iff$ the σ_v's are independent. Using simple calculations with ± 1-valued variables we can put our probability measure in the exponential form

$$\begin{aligned}\mu(\sigma) &= \frac{1}{2} \prod_{v, w \,:\, v \to w} (1-\varepsilon)^{\mathbf{1}_{\sigma_v = \sigma_w}} \cdot \varepsilon^{\mathbf{1}_{\sigma_v \neq \sigma_w}} \\ &= \frac{1}{Z_N(\beta)} \exp\left(\beta \sum_{v, w \,:\, v \to w} \sigma_v \sigma_w\right)\end{aligned}$$

with $\beta \coloneqq \frac{1}{2} \log \frac{1-\varepsilon}{\varepsilon}$ called the inverse temperature, or equivalently $\varepsilon = \frac{1}{\exp(2\beta)+1}$, and

$$Z_N(\beta) = \sum_{\sigma \in \Omega_N} \exp\left(\beta \sum_{v, w \,:\, v \sim w} \sigma_v \sigma_w\right)$$

is a normalizing constant, called the partition function. We have recovered here the finite-volume Gibbs measure for the Ising model on a tree (with open boundary conditions). We would like to understand this measure. In which way is possibly information persevered over long distances? Such distances will set the tone for subsequent investigations. For $v \in V_N$, $|v|$ is defined to be the distance to the root. For $|w| = N$ define

$$F_N \coloneqq \sigma = (\sigma_0 \sigma_w = -1)$$

This is a meaningful quantity for all N, so we may take a limit.

Proposition 3.2.

$$F_N \xrightarrow{N\to\infty} \frac{1}{2}$$

i.e., an observation of a single spin at the boundary at distance N *does not allow us to deduce anything about the state at the root when* N *tends to infinity.*

Proof. The problem is reduced to the study of a Markov chain along the path which connects the root 0 to the vertex w. Such a problem is elementary and can be treated by diagonalization. With the transition matrix

$$P = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$$

we get

$$F(N) = \sum_{\sigma_1,\dots,\sigma_{N-1}} P_{1\,\sigma_1} P_{\sigma_1\,\sigma_2} \dots P_{\sigma_{N-1}-1}$$

P has eigenvalues 1 and $1-2\varepsilon$, with eigenvectors $(1,1)^\top$ and $(1,-1)^\top$ respectively. Hence

$$O^\top P\, O = \begin{pmatrix} 1 & 0 \\ 0 & 1-2\varepsilon \end{pmatrix}$$

with

$$O = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = O^\top$$

which yields

$$P^N = O \begin{pmatrix} 1 & 0 \\ 0 & (1-2\varepsilon)^N \end{pmatrix} O = \frac{1}{2} \begin{pmatrix} 1+(1-2\varepsilon)^N & 1-(1-2\varepsilon)^N \\ 1-(1-2\varepsilon)^N & 1+(1-2\varepsilon)^N \end{pmatrix}$$

$$\implies F(N) = \frac{1}{2}\Big(1-(1-2\varepsilon)^N\Big) \xrightarrow{N\to\infty} \frac{1}{2}$$

□

Remark. In general one dimensional models have no long range order, unless the interactions are long-range. Markov chains on finite state spaces loose their memory exponentially fast.

A more interesting question now is the following, and this is a typical tree question. When does the information at all of the boundary sites allow us to deduce the state at the origin? The chances are much better now, as there are exponentially many sites in N, and the boundary sites constitute a non-vanishing fraction of all sites of a tree of depth N. Define

$$\partial T_N \coloneqq \{v \in V \mid |v| = N\}$$

for the boundary of the tree of depth N. Consider the conditional probability that the variable at the origin is 1, if we condition on any configuration at distance N from the origin, that is

$$\pi_N(\xi) = \mu(\sigma(0) = 1 \mid \sigma_{\partial T_N} = \xi)$$

Claim (Question 1)**.** *Is it always true that a conditioning of the boundary spins to take their maximal value has no predictive power for the spin at the origin, for large volumes? That is, do we have*

$$\pi_N(1_{\partial T_N}) \xrightarrow{N\to\infty} \frac{1}{2}\,?$$

Claim (Question 2)**.** *Is it true that*

$$\pi_N(1_{\partial T_N}) \xrightarrow{N\to\infty} \frac{1}{2}$$

for typical realizations of the boundary spins ξ*? For which value of* ε*?*

Remark. What do we mean by typicality? More precisely, let us consider the variance of the random variable π_N obtained feeding it random boundary spins ξ distributed according to μ. Then the question above reformulated reads then: When do we have

$$\mathbb{V}\big(\pi^N\big) \xrightarrow{N\to\infty} 0 \xleftarrow{N\to\infty} \mathbb{E}_\mu\left[\left(\pi^N - \frac{1}{2}\right)^2\right]$$

Theorem 3.3. *Let* T *be a regular tree where every vertex has precisely* d *children then Question 1 holds if and only* $\mathrm{d}(\tanh\beta) \leqslant 1$.

Theorem 3.4. *Let* T *be a regular tree where every vertex has precisely* d *children then Question 2 holds if and only* $\mathrm{d}(\tanh\beta)^2 \leqslant 1$.

Remark. Note that β is the second largest eigenvalue of the transition matrix. The above is in accordance with the intuition that the value of the parameter β (which can be considered as a coupling strength) needs to be bigger to ensure propagation of a typical boundary condition. The questions above have been formulated in a pedestrian way, in the sense that we made statements in terms of limits of finite-volume quantities. We did not need any measure theory. However, the appropriate setting to discuss them is the formalism of infinite-volume Gibbs measures to which will come now.

3.2 Gibbsian theory on countable vertex sets

3.2.1 Gibbsian specifications

Let V be a countably infinite set, and let Ω_0 be a Polish space with sigma algebra $\mathcal{F}_0$. We call Ω_0 the local state space, the simplest non-trivial example we previously discussed is $\Omega = \{-1, 1\}$. V could appear as the vertex set of some graph, e.g., $V = \mathbb{Z}^{\mathrm{d}}$ with $\mathrm{d} \in \mathbb{N}$. For any sub-volume $\Lambda \subseteq V$ (possibly infinite) define

$$\Omega_\Lambda = \Omega_0^\Lambda = \left\{(\omega_x)_{x\in\Lambda} \mid \forall x \in \Lambda : \omega_x \in \Omega_0\right\}$$

When $\Lambda = V$, denote $\Omega = \Omega_V$. The measurable structure on Ω_Λ is given by the product sigma algebra

$$\mathcal{B}_\Lambda = \bigotimes_{i\in\Lambda} \mathcal{F}_0 =: \mathcal{F}_0^\Lambda$$

For any $x \in V$, the projection onto the x^{th} coordinate is denoted by

$$\sigma_x : \Omega \to \Omega_0 \quad \omega \mapsto \omega_x$$

The restriction of a configuration in the infinite volume, $\omega \in \Omega$, to sub-volume $\Lambda \subseteq V$, can be given by using the projection $\sigma_\Lambda : \Omega \to \Omega_\Lambda$ with $\sigma_\Lambda(\omega) = \omega_\Lambda$. Similarly, if $\Lambda \subseteq \Delta \subseteq V$, we will use the same notation σ_Λ for the projection from Ω_0^Δ to Ω_0^Λ. The concatenation of two configurations $\omega \in \Omega_0^\Lambda$ and $\rho \in \Omega_0^{\Delta\setminus\Lambda}$ is denoted as $\omega\rho \in \Omega_0^\Delta$ and is defined by having the properties $\sigma_\Lambda(\omega\rho) = \omega$ and $\sigma_{\Delta\setminus\Lambda}(\omega\rho) = \rho$. Denote $\Lambda \Subset V$ is Λ is a finite subset of V. For any sub-volume $\Lambda \Subset V$ define the sigma-algebra of *cylinders with base in* Λ as

$$\mathcal{C}(\Lambda) \coloneqq \sigma_\Lambda^{-1}(\mathcal{B}_\Lambda)$$

For any (possibly infinite) $\Delta \subseteq V$, consider the algebra of cylinders with base in Δ, i.e.,

$$\mathcal{C}_\Delta \coloneqq \bigcup_{\Lambda\Subset\Delta} \mathcal{C}(\Lambda)$$

For each $\Delta \subseteq V$, the sigma algebra $\mathcal{F}_\Delta$, *of all events occurring in* Δ, is then by definition generated by $\mathcal{C}(\Lambda)$, i.e.,

$$\mathcal{F}_\Delta \coloneqq \sigma(\mathcal{C}_\Delta)$$

When $\Delta = V$, we simply denote $\mathcal{F} = \mathcal{F}_V$, and we have by the previous definition that $\mathcal{F}$ is the smallest sigma-algebra on Ω containing the cylinder events, i.e.,

$$\mathcal{F} = \sigma(\mathcal{C}_V) = \bigotimes_{i\in V} \mathcal{F}_0$$

A *spin model* is then simply a probability measure on the product space $(\Omega, \mathcal{F})$. We call Ω_0 the *local state space* and Ω the *configuration space* of the spin model. In the following sections, we will work towards the introduction of a spin model in the infinite volume.

Definition 3.5. Let $\Lambda \Subset V$. A probability kernel from $\mathcal{F}_{\Lambda^c}$ to $\mathcal{F}$ is a map

$$\pi_\Lambda : \mathcal{F} \times \Omega \to [0, 1]$$

such that

- $\pi_\Lambda(\cdot \mid \omega)$ is a probability measure on $(\Omega, \mathcal{F})$ for any $\omega \in \Omega$.

- $\pi_\Lambda(A \mid \cdot)$ is $\mathcal{F}_{\Lambda^c}$ measurable for any $A \in \mathcal{F}$.

Moreover, if

$$\forall A \in \mathcal{F}_{\Lambda^c}\ \forall \omega \in \Omega : \pi_\Lambda(A \mid \omega) = \mathbf{1}_A(\omega)$$

then π_Λ is called *proper*.

A probability kernel pulls functions back and pushes measures forward in the following sense: If μ is a probability measure on the measurable space $(\Omega, \mathcal{F}_{\Lambda^c})$ and π_Λ is a probability kernel from $\mathcal{F}_{\Lambda^c}$ to $\mathcal{F}$ then

$$\forall A \in \mathcal{F} : \mu\pi_\Lambda(A) = \int \pi_\Lambda(A \mid \omega)\mu(\mathrm{d}\omega)$$

defines a probability measure on $(\Omega, \mathcal{F})$. Also, if $f : \Omega \to \mathbb{R}$ is $\mathcal{F}$ measurable then the function $\pi_\Lambda f : \Omega \to \mathbb{R}$ given by

$$\forall \omega \in \Omega : \pi_\Lambda(f \mid \omega) = \int \pi_\Lambda(\mathrm{d}\xi \mid \omega) f(\xi)$$

is measurable with respect to $\mathcal{F}_{\Lambda^c}$. The *composition of kernels* π_Λ and π_Δ is defined as

$$\forall A \in \mathcal{F}\ \forall \omega \in \Omega : \pi_\Lambda\pi_\Delta(A \mid \omega) \coloneqq \int \pi_\Delta(A \mid \omega)\pi_\Lambda(\mathrm{d}\rho \mid \omega)$$

and is itself a kernel from $\mathcal{F}_{\Lambda^c}$ to $\mathcal{F}$.

Assume that we have a proper probability kernel $\pi_\Lambda : \mathcal{F}_{\Lambda^c} \to \mathcal{F}$, then the probability measure $\pi_\Lambda(\cdot \mid \omega)$ is supported on the set $\Omega_\Lambda^\omega \coloneqq \sigma_{\Lambda^c}^{-1}(\omega)$ for any $\omega \in \Omega$ as

$$\pi_\Lambda(\Omega_\Lambda^\omega \mid \omega) = \mathbf{1}_{\Omega_\Lambda^\omega}(\omega) = 1$$

since $\Omega_\Lambda^\omega \in \mathcal{F}_{\Lambda^c}$. Therefore one can interpret the configuration $\omega \in \Omega$ as the *boundary condition* of the measure $\pi_\Lambda(\cdot \mid \omega)$. In the following all kernels π_Λ to be considered will be proper and therefore they will be entirely determined by all the numbers

$$\pi_\Lambda(\eta_\Lambda \omega_{\Lambda^c} \mid \omega) = \pi_\Lambda(\eta_\Lambda \mid \omega_{\Lambda^c})$$

As it turns out, it will be necessary to use an infinite family of probability kernels, $\{\pi_\Lambda \mid \Lambda \Subset V\}$ to describe Gibbs measures in the infinite volume directly. The key concept in that regard is that of a (local) specification.

Definition 3.6 (Specification)**.** A specification is a family of proper probability kernels $\gamma = \{\gamma_\Lambda : \mathcal{F}_{\Lambda^c} \to \mathcal{F} \mid \Lambda \Subset V\}$ which satisfies the consistency relation, i.e.,

$$\forall \Lambda, \Delta \Subset V : \Lambda \subseteq \Delta \implies \gamma_\Delta\gamma_\Lambda = \gamma_\Delta$$

A measure $\mu \in \mathcal{M}_1(\Omega)$ is said to be compatible (or specified by) γ if

$$\forall \Lambda \Subset V : \mu = \mu\gamma_\Lambda$$

The set of measures which are compatible with γ is denoted by $\mathcal{G}(\gamma)$.

A first natural question that arises with regard to this definition is if there is a way to construct specifications. Before we answer this question, consider the following lemma.

Lemma 3.7. *Suppose that π_Λ is a proper probability kernel from $\mathcal{F}_{\Lambda^c}$ to $\mathcal{F}$.*

1. *We have that*

$$\pi_\Lambda(A \cap B \mid \cdot) = \pi_\Lambda(A \mid \cdot)\mathbf{1}_B(\cdot)$$

   for all $A \in \mathcal{F}$ and all $B \in \mathcal{F}_{\Lambda^c}$.

2. *Let $\mu \in \mathcal{M}_1(\Omega)$ then*

$$\mu\pi_\Lambda = \pi_\Lambda \iff \mu(A \mid \mathcal{F}_{\Lambda^c}) \overset{\text{a.s.}}{=} \pi_\Lambda(A \mid \cdot)$$

   for all $A \in \mathcal{F}$.

Proof. (1) First, assume that $\omega \notin B$ then

$$\pi_\Lambda(A \cap B) \leqslant \pi_\Lambda(B \mid \omega) = \mathbf{1}_B(\omega) = 0$$

Now suppose $\omega \in B$ then

$$\pi_\Lambda(A \cap B) = \pi_\Lambda(A \mid \omega) - \pi_\Lambda(A \cap B^c \mid \omega) = \pi_\Lambda(A \mid \omega)$$

(2) If it holds then for all $\Lambda \Subset V$ and for all $A \in \mathcal{F}$, we have

$$\mu\pi_\Lambda(A) = \int \pi_\Lambda(A \mid \omega)\mu(d\omega) = \int \mu(A \mid \mathcal{F}_{\Lambda^c})(\omega)\mu(d\omega) = \mu(A)$$

Now suppose that $\mu\pi_\Lambda = \mu$ then

$$\mu(A \cap B) = \mu\pi_\Lambda(A \cap B) = \int \pi_\Lambda(A \cap B \mid \omega)\mu(d\omega) = \int_B \pi_\Lambda(A \mid \omega)\mu(d\omega)$$

for all $A \in \mathcal{F}$ and for all $B \in \mathcal{F}_{\Lambda^c}$. By conditional probability it follows that

$$\mu(A \cap B) = \int_B \mu(A \mid \mathcal{F}_{\Lambda^c})(\omega)\mu(d\omega)$$

for all $B \in \mathcal{F}_{\Lambda^c}$. And by the almost surely uniqueness of the conditional expectation we see that

$$\mu(A \mid \mathcal{F}_{\Lambda^c})(\cdot) = \pi_\Lambda(A \mid \cdot)$$

μ - a.s. for all $A \in \mathcal{F}$. □

Remark. The second part of the lemma tells us that for a given specification $(\gamma_\Lambda : \Lambda \Subset V)$ the measures $\mu \in \mathcal{G}(\gamma)$ are characterized by having a regular conditional distribution provided by γ_Λ, when conditioning with respect to $\mathcal{F}_{\Lambda^c}$. The most important class of specifications are the so-called *Gibbsian specifications* which we will introduce in the following definition.

Definition 3.8. Let $\Phi = \{\Phi_\Lambda\}_{\Lambda \in V}$ be a family of real-valued functions on the configuration space Ω. We call Φ an interaction potential if it has the following properties:

1. The functions Φ_Λ are $\mathcal{F}_\Lambda$ measurable for any $\Lambda \Subset V$.
2. For all $\Lambda \Subset V$ and $\omega \in \Omega$ the series
$$H_\Lambda^\Phi(\omega) = \sum_{\substack{A \Subset V \\ A \cap \Lambda \neq \emptyset}} \Phi_A(\omega)$$
exists.

We call H_Λ^Φ the Hamiltonian in the finite sub-volume Λ associated to the potential Φ.

Remark. By existence, we mean that for any increasing sequence of volumes Δ_n which converges to V, we have that

$$\lim_{n \uparrow \infty} \sum_{\substack{A \subseteq \Delta_n \\ A \cap \Lambda \neq \emptyset}} \Phi_A(\omega)$$

exists and does not depend on the volume sequence.

Since the sum above contains possibly infinitely many terms there is no guarantee that it converges. However, for an important class of interaction potentials this is not an issue. Let d_G denote the graph distance on V, which is the number of edges in the shortest path connecting two vertices. We define the diameter of a finite set Λ by

$$\operatorname{diam}(\Lambda) \coloneqq \sup_{x, y \in \Lambda} d_G(x, y)$$

Let

$$r(\Phi) \coloneqq \inf\{R > 0 \mid \forall \Lambda \Subset V : \operatorname{diam}(\Lambda) > R \implies \Phi_\Lambda \equiv 0\}$$

If $r(\Phi)$ is finite, the interaction potential Φ is said to be of *finite range* and clearly the Hamiltonian H_Λ^Φ is well defined for any finite sub-volume Λ in this case.

In the following we will assume that the local state space is equipped with a so-called a *prior measure* $\lambda \in \mathcal{M}_1(\Omega_0)$ and denote for any $\Lambda \Subset V$ the product measure on $\left(\Omega_0^\Lambda, \mathcal{F}_0^\Lambda\right)$ by λ^Λ. The *(conditional) partition function* is then defined as

$$Z_\Lambda^\Phi(\omega) = \int \exp\left(-H_\Lambda^\Phi(\zeta_\Lambda \omega_{\Lambda^c})\right)\lambda^\Lambda(d\zeta_\Lambda)$$

A potential Φ is said to be λ-admissible if the partition function $Z_\Lambda^\Phi(\omega)$ is a finite number in the open interval $\mathbb{R}^+$, for all $\Lambda \Subset V$ and all $\omega \in \Omega$.

Proposition 3.9 ([12] (Proposition 2.1.5))**.** *Suppose that Φ is an λ-admissible interaction potential. Then the family of probability kernels*

$$\gamma^\Phi = \left\{\gamma_\Lambda^\Phi : \mathcal{F}_{\Lambda^c} \to \mathcal{F} \mid \Lambda \Subset V\right\}$$

defined by

$$\gamma_\Lambda^\Phi(A \mid \omega) = \frac{1}{Z_\Lambda^\Phi(\omega)} \int \exp\left(-H_\Lambda^\Phi(\zeta_\Lambda \omega_{\Lambda^c})\right)\mathbf{1}_A(\zeta_\Lambda \omega_{\Lambda^c})\lambda^\Lambda(d\zeta_\Lambda)$$

constitutes a specification and it is called the Gibbs specification for Φ. A probability measure $\mu \in \mathcal{G}\left(\gamma^\Phi\right)$ is called an infinite-volume Gibbs measure (or simply a Gibbs measure) associated to the potential Φ.

To verify the specification properties note that the measurability properties are evident, while the consistency is obtained by a rearrangement of sums, see [8] (Proposition 2.5). The measures $\gamma^\Phi(\cdot \mid \omega) \in \mathcal{M}_1(\Omega)$ are also called finite- volume Gibbs measures under boundary condition ω. The way we have defined them they actually are measures on the infinite volume. However, recall that they are supported on the set Ω_Λ^ω which consists only of configurations that are equal to ω outside the finite volume Λ.

3.2.2 Extremal Gibbs measures

One basic observation is that as the DLR equation is linear, $\mathcal{G}(\gamma)$ is a convex set: if $\mu_1, \ldots, \mu_N \in \mathcal{G}(\gamma)$ then so does any convex combination of them. This makes the extremal elements of this set, which we denote by $\operatorname{ex}\mathcal{G}(\gamma)$, especially interesting. The following questions arise naturally:

1. What properties, if any, distinguish the elements of $\operatorname{ex}\mathcal{G}(\gamma)$ from the non-extremal ones?
2. What is the physical interpretation of these extremal points of $\mathcal{G}(\gamma)$?

Before we answer these questions we will give a condition under which the set of extremal Gibbs measures is non-empty. Let $\mathcal{C}_b(\Omega)$ be the set of bounded real-valued functions on Ω that are continuous w.r.t. the product topology obtained from the topology on the Polish local state space Ω_0. A particular class of specifications is given in the following definition.

Definition 3.10. A specification $\gamma = (\gamma_\Lambda)_{\Gamma \Subset V}$ is said to be Feller-continuous if for all $\Lambda \Subset V$, $f \in \mathcal{C}_b(\Omega)$ implies $\gamma_\Lambda f \in \mathcal{C}_b(\Omega)$.

An important example of Feller-continuous specifications is provided by the Gibbsian specifications γ^Φ where the interaction potential Φ is continuous and uniformly convergent (and λ-admissible) [20]. An interaction is by definition uniformly convergent if for every $\Lambda \Subset V$ the sum in (2.1.2) converges uniformly in ω. Note that this is always the case if the interaction is of finite range which will be the case for all models considered in these notes.

Let $(\Lambda_n)_{n \in \mathbb{N}}$ be any sequence of finite subsets of V. We say that $(\Lambda_n)_{n \in \mathbb{N}}$ exhausts V if

$$\forall v \in V \; \exists N \in \mathbb{N} : n \geqslant N \implies v \in \Lambda_n$$

Proposition 3.11. *[20] (Proposition 2.22) Suppose γ is a Feller-continuous specification and let $(\Lambda_n)_{n \in \mathbb{N}}$ be any sequence of finite subsets of V that exhausts V. If $(\nu_n)_{n \in \mathbb{N}}$, a sequence of measures in $\mathcal{M}_1(\Omega)$, converges weakly to some $\mu \in \mathcal{M}_1(\Omega)$ then $\mu \in \mathcal{G}(\gamma)$.*

Note that if Ω_0 is compact, so is $\Omega = \Omega_0^V$ w.r.t. the product topology. Also, Ω is Polish since it is the countable product of Polish spaces. Hence $\mathcal{M}_1(\Omega)$ is weakly compact. Therefore, in the case of a Feller-continuous specification every sequence $(\nu_n \gamma_{\Lambda_n})$ has a convergence subsequence, and hence $\mathcal{G}(\gamma)$ is not empty. In general this might not be true; the question of whether or not $|\mathcal{G}(\gamma)| = 0$ is a non-trivial one. There indeed exist physically reasonable models for which there are no infinite-volume Gibbs measures. Examples are the massless discrete Gaussian free field on the lattice $\mathbb{Z}^d$ in dimensions $d \leqslant 2$ and the solid-on-solid in $d = 1$ [8]. In both cases, the local state space equals the set of all integers.

One nice property of Feller-continuous specifications is that they allow the identification of Gibbs measures as weak limits, at least the extremals. To be more specific, we have the following statement [20] (Proposition 2.23):

Proposition 3.12. *Let Ω_0 be a compact metric space and let $(\gamma_\Lambda : \Lambda \Subset V)$ be a Feller-continuous specification. Furthermore, let $\mu \in \operatorname{ex} \mathcal{G}(\gamma)$. Then for μ-a.s. w,*

$$\gamma_{\Lambda_n}(\cdot \mid \omega) \xrightarrow{n\to\infty} \mu$$

in the weak limit for any sequence of finite sub-volumes $(\Lambda_n : n \in \mathbb{N})$ that exhausts V.

Let us assume a Feller-continuous specification is given. Then the previous two propositions show the connection between the DLR-approach to the Gibbs theory in infinite-volume and the classical approach using the thermodynamic limit of finite-volume Gibbs measures under boundary condition (see Chapter 3 of [7] for a detailed exhibition of this ansatz). Recall that any weak limit of finite-volume Gibbs measures is in fact an infinite-volume Gibbs measure. Conversely, the proposition above states that if we have an extremal Gibbs measure μ and sample any typical configuration from μ and use it as a boundary condition, in the infinite-volume limit we will recover μ itself. The following theorem follows immediately from the proposition, and gives a condition for which there is a unique Gibbs measure.

Theorem 3.13 ([12] (Theorem 2.2.4))**.** *Let Ω_0 be a compact metric space and let $(\gamma_\Lambda : \Lambda \Subset V)$ be a Feller-continuous specification. Suppose that for all sequences of finite sub-volumes $(\Lambda_n : n \in \mathbb{N})$ exhausting V, and every $\omega \in \Omega$, all the possible weak limits of $\gamma_{\Lambda_n}(\cdot \mid \omega)$ are identical. There there exists exactly one Gibbs measure.*

Recall that for any $\Lambda \Subset V$, we defined $\mathcal{F}_{\Lambda^c}$ as the σ-algebra which consists of all the events that only depend on the spins outside the finite set Λ. Now the tail σ-algebra (or tail field) $\mathcal{T}$ is defined as the σ-algebra which only depends on the spins outside any finite region Λ, i.e.,

$$\mathcal{T} \coloneqq \bigcap_{\Lambda \Subset V} \mathcal{F}_{\Lambda^c}$$

The extremal elements of $\mathcal{G}(\gamma)$ are characterized by the following properties [20] (Proposition 2.20):

Proposition 3.14. *Let $\mu \in \mathcal{G}(\gamma)$ then the following are equivalent.*

- *The measure μ is an extremal element of $\mathcal{G}(\gamma)$.*
- *The measure μ is trivial on $\mathcal{T}$, i.e.,*
$$\forall A \in \mathcal{T} : \mu(A) \in \{0, 1\}$$
- *The measure μ has short-range correlations, i.e., for each $A \in \mathcal{F}$ we have*

$$\lim_{\substack{\Lambda \uparrow V \\ \Lambda \Subset V}} \sup\{\mu(A \cap B) - \mu(A)\mu(B) \mid B \in \mathcal{F}_{\Lambda^c}\} = 0$$

Physical systems can in general have one or more possible *macrostates*, depending on the values of some internal free parameters of the system. For example water can be in a gaseous, liquid or solid macrostate depending on the temperature and pressure. While the microscopic quantities change rapidly, the macroscopic quantities remain constant. To turn this into a mathematical exact statement we define the macroscopic quantities or macroscopic observables as the functions on Ω that are measurable w.r.t the tail field $\mathcal{T}$, i.e., the function that do not depend on spins in any finite volume $\Lambda \Subset V$. The physical relevance of the preceding theory presented in this chapter lies in the assumption that the statistical mechanical information of the physical system can be obtained from a suitable specification γ, that is, the space of measures $\mathcal{G}(\gamma)$ describes the macrostates of the system. Proposition 2.2.5 tells us that these macrostates are given by the extremal elements of $\mathcal{G}(\gamma)$.

What is then the interpretation of the non-extremal elements of $\mathcal{G}(\gamma)$? Suppose that the local state space $(\Omega_0, \mathcal{F}_0)$ is Polish. Then every non-extremal measure $\mu \in \mathcal{G}(\gamma)$ in an (integral) convex combination of extremal ones. This decomposition is even unique, that is, $\mathcal{G}(\gamma)$ is a simplex [8] (Theorem 7.26).

This means that a non-extremal Gibbs measure corresponds simply to the preparation of randomly chosen extremal Gibbs measures. The probabilities for this choice are given by the *coefficients* of the convex combination. This extra randomness can be interpreted as the uncertainty in the experiment regarding the true nature of the systems macrostate (for a more detailed discussion, see Chapter 6 of [7]).

Therefore the non-extremal Gibbs measures do not lead to new physics: Everything that we can observe under such a measure is typical for one of the extremal ones that appear in its (unique) decomposition. Hence the extremal Gibbs measures are the physically important ones, which is why they are also called the *pure* states. This is the reason why we say that a physical system exhibits a phase transition when there exist multiple extremal Gibbs measures for the model.

Finally, we want to point out that the extremal Gibbs measures are suitable to describe the different phases of the system as it is possible to distinguish those measures by looking at macroscopic observables only. This is important since we should be able to tell macrostates apart by looking at macroscopic measurements:

Theorem 3.15 ([8] (Theorem 7.7))**.** *Let μ_1, μ_2 be two distinct extremal Gibbs measures w.r.t a specification $(\gamma_\Lambda : \Lambda \Subset V)$. Then there exists some tail-measurable event $A \in \mathcal{T}$ such that $\mu_1(A) = 1$ and $\mu_2(A) = 0$, that is, μ_1 and μ_2 are mutually singular.*

3.2.3 Uniqueness

Please refer to [12] (section 2.3), on a criterion for the uniqueness of Gibbs measures on any graph.

3.3 Gibbs measures on trees

We specialize the index set to be the vertex set of a countably infinite tree. We discuss several Markov properties. There is the notion of a (spatially) Markov specification which means that the finite-volume Gibbs measures depend on their boundary condition only via a boundary layer of thickness one. This notion is meaningful on any graph. Similarly, an infinite-volume measure is called a (spatially) Markov field if its finite-volume conditional probabilities depend on the boundary condition only via the boundary layer of thickness one.

To be distinguished from the above notion, there is the notion of a tree- indexed Markov chain. This is meaningful only on trees. It relies on the definition of past and future vertices relative to a given oriented edge. While each tree-indexed Markov chain is a (spatially) Markov field, the converse statement is ensured only for extremal Gibbs measures. Indeed, the non-trivial part is that any extremal Gibbs measure for a Markov specification is a tree-indexed Markov chain. We will explain in detail why this is true, using conditioning arguments involving *future-tail triviality*.

Then we come to describe the one-to-one correspondence between boundary laws and tree-indexed Markov chains. Boundary laws are families of positive measures on the local state space, indexed by the set of oriented edges which satisfy a consistency equation (tree-recursion). There is also a one-to-one correspondence between boundary laws and transition matrices of the tree-indexed Markov chain, given the specification. We conclude with a discussion of all homogeneous boundary laws on the Cayley tree for concrete examples of the Ising model and the Potts model in zero magnetic field.

3.3.1 Construction of Gibbs measures via boundary laws

One of the most important classes of stochastic processes are Markov chains. A Markov chain in its most elementary form is a sequence of random variables indexed by $\mathbb{N}_0$ (which is usually interpreted as time) which has the property that future events are independent of the past given the information about its present state, i.e.,

$$\forall\, n \in \mathbb{N}_0\ \forall\, x_{n+1}, \ldots, x_0 \in \Omega_0 : \mu(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_0 = x_0) = \mu(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

There is a natural way to generalize this definition to the situation where the stochastic process is no longer indexed by $\mathbb{N}_0$ but by the vertices V of a tree. To formulate this we need some more notation. For any vertex $w \in V$ the set of the directed edges pointing away from w is given by

$$E_w = \{(x, y) \in E \mid d(w, y) = d(w, x) + 1\}$$

This is an orientation of the set of edges induced by the vertex w. Furthermore we define the *past* of any oriented edge $(x, y) \in E$ by

$$(-\infty, xy) = \{w \in V \mid (x, y) \in E_w\}$$

This is the set of sites w from which the oriented edge (x, y) is pointing away. The definition of the future of an oriented edge is analogous. Note that the tree property, i.e., the absence of loops, is clearly needed to give a meaningful definition of the *past* and *future* of an oriented edge. In the following we will always restrict ourselves to the case where the local state space Ω_0 is finite. This simplifies the analysis but still allows the occurrence of phase transitions on trees.

Definition 3.16. Let Ω_0 be the local state space and $\Omega = \Omega_0^V$. A measure $\mu \in \mathcal{M}_1(\Omega)$ is called a tree-index Markov chain if

$$\mu\big(\sigma_y = \omega_y \,\big|\, \mathcal{F}_{(-\infty, xy)}\big) = \mu\big(\sigma_y = \omega_y \,\big|\, \mathcal{F}_{\{x\}}\big)$$

μ-a.s. for all $(x, y) \in E$ and any $\omega_y \in \Omega_0$. Any stochastic matrix P on Ω_0 with

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}} = P_{ij}(\sigma_i, y)\big)$$

μ-a.s. for all $y \in \Omega_0$ is then called a transition matrix from i to j for μ. Moreover, a Markov chain is said to be completely homogeneous with transition matrix P if

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}} = P(\sigma_i, y)\big)$$

μ-a.s. for $y \in \Omega_0$ and $(i, j) \in E$.

Remark. Every tree-indexed Markov chain μ with transition matrices $\big(P_{ij} : (i, j) \in E\big)$ and marginal distribution α_k at vertex $k \in V$ has the following representation

$$\mu(\sigma_\Lambda = \zeta) = \alpha_k(\zeta_k) = \prod_{\substack{(i,j) \in E \\ i,j \in \Lambda}} P_{ij}\big(\zeta_i, \zeta_j\big)$$

for all finite connected sets $\Lambda \Subset V$, $\zeta \in \Omega_0^V$ and $k \in \Lambda$. The above can be proved by induction on the number of vertices in Λ. If μ is completely homogeneous it follows from the equation above that μ is invariant under the group $I(E)$, the group of graph automorphisms of V.

Besides the one-sided Markov property there is also the notion of a spatial Markov property:

Definition 3.17. A a specification γ for Ω_0 and V is said to be a Markov specification if $\gamma_\Lambda(\sigma_\Lambda = \zeta | \cdot)$ is $\mathcal{F}_{\partial\Lambda}$-measurable for all $\zeta \in \Omega_0^V$ and $\Lambda \Subset V$.

Note that $\partial\Lambda = \{i \in V \,|\, d(i, \Lambda) = 1\}$ is the outer boundary layer of thickness one. If γ is a Markov specification, then every $\mu \in \mathcal{G}(\gamma)$ is a Markov field, i.e., μ satisfies the spatial Markov property

$$\mu(\sigma_\Lambda = \zeta \,|\, \mathcal{F}_{\Lambda^c}) = \mu(\sigma_\Lambda = \zeta \,|\, \mathcal{F}_{\partial\Lambda})$$

μ-a.s. for all $\zeta \in \Omega_0^V$ and $\Lambda \Subset V$. Note that every Gibbsian specification which is defined by a nearest-neighbor potential is Markov.

Theorem 3.18. *Every tree-index Markov chain is a Markov field.*

Proof. Assume that μ is a Markov chain. For $\Lambda \Subset V$ let $\Delta \Subset V$ be some finite connected set such that $\Lambda \cup \partial\Lambda \subseteq \Delta$. The explicit form of the finite volume marginals, applied in the bigger volume Δ, shows that

$$\mu(\sigma_\Delta = \zeta\, \omega\, \eta)\mu\Big(\sigma_\Delta = \tilde{\zeta}\, \omega\, \tilde{\eta}\Big) = \mu\Big(\sigma_\Delta = \tilde{\zeta}\, \omega\, \eta\Big)\mu(\sigma_\Delta = \zeta\, \omega\, \tilde{\eta})$$

for all $\zeta, \tilde{\zeta} \in \Omega_0^V$, $\omega \in \Omega_0^{\partial\Lambda}$, $\eta, \tilde{\eta} \in \Omega_0^{\Delta\setminus(\Lambda\cup\partial\Lambda)}$. Summing over $\tilde{\zeta}$ and $\tilde{\eta}$, we obtain

$$\mu(\sigma_\Delta = \zeta\, \omega\, \eta)\mu(\sigma_{\partial\Lambda} = \omega) = \mu\big(\sigma_{\Delta\setminus\Lambda} = \omega\, \eta\big)\mu(\sigma_{\Lambda\cup\partial\Lambda} = \zeta\, \omega)$$

If $\mu\big(\sigma_{\Delta\setminus\Lambda} = \omega\, \eta\big) > 0$, we have

$$\mu\big(\sigma_\Lambda = \zeta \,\big|\, \sigma_{\Delta\setminus\Lambda} = \omega\, \eta\big) = \mu(\sigma_\Lambda = \zeta \,|\, \sigma_{\partial\Lambda} = \omega)$$

which means that

$$\mu\big(\sigma_\Lambda = \zeta \,\big|\, \mathcal{F}_{\Delta\setminus\Lambda}\big) \overset{\text{a.s.}}{=} \mu(\sigma_\Lambda \,|\, \sigma_{\partial\Lambda} = \omega)$$

Since $\mathcal{F}_{\Lambda^c}$ is generated by the union of all $\mathcal{F}_{\Delta\setminus\Lambda}$, we conclude that

$$\mu(\sigma_\zeta \,|\, \mathcal{F}_{\Lambda^c}) \overset{\text{a.s.}}{=} \mu(\sigma_\zeta \,|\, \mathcal{F}_{\partial\Lambda})$$

Hence μ is a Markov field. □

Theorem 3.19. *Let γ be a Markov specification, then each $\mu \in \operatorname{ex} \mathcal{G}(\gamma)$ is a tree-indexed Markov chain.*

Proof. Set an oriented edge $(i,j) \in E$ and let $\Delta(n)$ be the ball of radius n around j and $\Lambda(n) = \Delta(n) \cap (ij,\infty)$ be the future in this ball relative to the oriented edge. As μ is assumed to be extremal, we know that μ is trivial on the tail-σ-algebra $T = \bigcap_{n\in\mathbb{N}} \mathcal{F}_{\Delta(n)^c}$ (see proposition 2.2.5), Hence μ is also trivial on the smaller σ-algebra

$$\bigcap_{n\in\mathbb{N}} \mathcal{F}_{(ij,\infty)\setminus\Lambda(n)}$$

The above is the future tail σ-algebra relative to the oriented edge. This future-tail triviality implies that

$$\mathcal{F}_{\{i\}} = \bigcap_{n\in\mathbb{N}} \mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))} \quad \mu-\text{a.s.}$$

Indeed, the σ-algebra on the left is clearly contained in that on the right. Conversely, if $f:\Omega\to\mathbb{R}$ is bounded and measurable with respect to the latter σ-algebra then $f\big(x\,\sigma_{V\setminus\{i\}}\big)$ is measurable w.r.t. $\bigcap_{n\in\mathbb{N}} \mathcal{F}_{(ij,\infty)\setminus\Lambda(n)}$ and hence

$$f\big(x\,\sigma_{V\setminus\{i\}}\big) = \int f\big(x\,\omega_{V\setminus\{i\}}\big)\mu(d\omega) \quad \mu-\text{a.s.}$$

Therefore f is also measurable w.r.t. $\mathcal{F}_{\{i\}}$. As $\big(\mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))} : n\in\mathbb{N}\big)$ is a decreasing sequence of σ-algebras we can apply the backward martingale convergence theorem, which yield

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))}\big) \xrightarrow[\text{a.s.}]{n\to\infty} \mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}}\big)$$

By the tower property of conditional expectation, the term under the limit on the l.h.s. equals

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))}\big) = \mu\big(\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\Lambda(n)^c}\big) \,\big|\, \mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))}\big)$$

Since μ is a Gibbs measure, we have inside the conditional expectation on the r.h.s.

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\Lambda(n)^c}\big) \overset{\text{a.s.}}{=} \gamma_{\Lambda(n)}\big(\sigma_j = y \,\big|\, \cdot\big)$$

Note that $\{i\}\cup((ij,\infty)\setminus\Lambda(n)) \supseteq \partial\Lambda(n)$. Hence, by the Markov specification property for μ we may pull this out of the conditional expectation and arrive at the μ-a.s. equality

$$\begin{aligned}
&\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))}\big) = \gamma_{\Lambda(n)}\big(\sigma_j = y \,\big|\, \cdot\big) \\
\implies &\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}\cup((ij,\infty)\setminus\Lambda(n))}\big) \xrightarrow{n\to\infty} \lim_{n\to\infty} \gamma_{\Lambda(n)}\big(\sigma_j = y \,\big|\, \cdot\big) \\
&\qquad\qquad \xrightarrow{n\to\infty} \lim_{n\to\infty} \mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\Lambda(n)^c}\big) \\
&\qquad\qquad = \mu\left(\sigma_j = y \,\middle|\, \bigcap_{n\in\mathbb{N}} \mathcal{F}_{\Lambda(n)^c}\right)
\end{aligned}$$

where the second equality follows from the DLR-equation and where the last equation follows again from the backward martingale theorem. Hence,

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}}\big) = \mu\left(\sigma_j = y \,\middle|\, \bigcap_{n\in\mathbb{N}} \mathcal{F}_{\Lambda(n)^c}\right)$$

and by the inclusion

$$\bigcap_{n\in\mathbb{N}} \mathcal{F}_{\Lambda(n)^c} \supseteq \mathcal{F}_{(ij,\infty)} \supseteq \mathcal{F}_{\{i\}}$$

it follows by the tower property that

$$\mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{(ij,\infty)}\big) = \mu\big(\sigma_j = y \,\big|\, \mathcal{F}_{\{i\}}\big)$$

Therefore μ is a Markov chain. □

Let Φ be some nearest-neighbor interaction potential which may contain also single-site terms. Recall that the corresponding Gibbsian specification (the specification kernel) γ^Φ is then given by

$$\gamma^\Phi_\Lambda(\sigma_\Lambda = \omega_\Lambda \,|\, \omega) = Z_\Lambda(\omega)^{-1} \exp(-H_\Lambda(\omega)) = Z_\Lambda(\omega)^{-1} \exp\left(-\sum_{b\,:\,b\cap\Lambda\neq\emptyset} \Phi_b(\omega_b)\right)$$

where the sums runs over all non-oriented edges b touching the finite volume Λ. When we define transfer operators (or transfer matrices) by

$$Q_b(\omega_b) = \exp\Big(-\Phi_b(\omega_b) - |\partial i|^{-1}\Phi_{\{i\}}(\omega_i) - |\partial j|^{-1}\Phi_{\{j\}}(\omega_j)\Big)$$

where $b = \{i, j\} \in E$ and $\omega_b \in \Omega_0^b$, we can rewrite the specification kernel as

$$\gamma_\Lambda^\Phi(\sigma_\Lambda = \omega_\Lambda \,|\, \omega) = Z_\Lambda(\omega)^{-1} \prod_{b\,:\,b\cap\Lambda\neq\emptyset} Q_b(\omega_b)$$

Note that by definition the transfer matrices are symmetric, i.e.,

$$Q_{ij}(x, y) = Q_{ji}(y, x)$$

for all $\{i, j\} \in E$ and $x, y \in \Omega_0$.

In the following we will work towards a representation of tree-indexed Gibbs measures via the notion of so-called boundary laws [4], [8], [22].

Definition 3.20. A family of vectors $\Big\{l_{ij} \,\Big|\, (i,j) \in E,\ l_{ij} \in (0,\infty)^{\Omega_0}\Big\}$ is called a boundary law for the transfer operators $\{Q_b \,|\, b \in E\}$ if for each $(i,j) \in E$ there exists a constant $c_{ij} > 0$ such that the consistency relation

$$l_{ij} = c_{ij} \prod_{k\,\in\,\partial i\setminus\{j\}} \sum_{\omega_k\,\in\,\Omega_0} Q_{ki}(\omega_i, \omega_k) l_{ki}(\omega_k)$$

holds for every $\omega_i n \in \Omega_0$.

Boundary laws are maps from the oriented edges (k, i) to the positive measures on the single-site spin space at the site k.

For any boundary law, the family $\big\{\alpha_{ij}\, l_{ij} \,\big|\, (i,j) \in E\big\}$ for any fixed choice of strictly positive numbers α_{ij} is trivially also a boundary law.

We will now give the main theorem of this section, which shows the equivalence of boundary laws and tree-indexed Markov chains, which are Gibbs measures for the given set of transfer operators.

Theorem 3.21. *Consider a Markov specification γ_Λ^Φ of the form*

$$\gamma_\Lambda^\Phi(\sigma_\Lambda = \omega_\Lambda \,|\, \omega) = Z_\Lambda(\omega)^{-1} \prod_{b\,:\,b\cap\Lambda\neq\emptyset} Q_b(\omega_b)$$

and let $\{Q_b \,|\, b \in E\}$ be its associated family of transfer matrices.

1. *Each boundary law $\big\{l_{ij} \,\big|\, (i,j) \in E\big\}$ for a given family of transfer matrices defines a unique tree-indexed Markov chain $\mu \in \mathcal{G}(\gamma)$ via the equation*

$$\mu(\sigma_{\Lambda\cup\partial\Lambda} = \omega_{\Lambda\cup\partial\Lambda}) = Z_\Lambda^{-1} \prod_{y\,\in\,\partial\Lambda} l_{yy_\Lambda}(\omega_y) \prod_{b\,:\,b\cap\Lambda\neq\emptyset} Q_b(\omega_b)$$

   where $\Lambda \Subset V$ is a finite connected set, $\omega_{\Lambda\cup\partial\Lambda} \in \Omega_0^{\Lambda\cup\partial\Lambda}$ and Z_Λ is a suitable normalizing constant. y_Λ is the unique nearest neighbor of y which lies inside Λ.

2. *Conversely, every tree-indexed Markov chain $\mu \in \mathcal{G}(\gamma)$ admits a representation of the form above in (1) in terms of a boundary law. This representation is unique in the sense that every boundary law is unique up to a positive factor.*

Proof. (1) In the first step we will use Kolmogorov's extension theorem to show that the expressions on the r.h.s. describe the marginals of a unique measure $\mu \in \mathcal{M}(\Omega)$. This holds if the expressions are consistent, i.e.,

$$\sum_{w_V\,\in\,\Omega_0^V} Z_\Delta^{-1} \prod_{k\,\in\,\partial\Delta} l_{kk_\Delta}(\omega_k) \prod_{b\,:\,b\cap\Delta\neq\emptyset} Q_b(\omega_b) = Z_\Lambda^{-1} \prod_{k\,\in\,\partial\Delta} l_{kk_\Delta}(\omega_k) \prod_{b\,:\,b\cap\Delta\neq\emptyset} Q_b(\omega_b)$$

whenever $\Lambda, \Delta \Subset V$ are connected sets with $\Lambda \subseteq \Delta$, $V = (\Delta \cup \partial\Delta) \setminus (\Lambda \cup \partial\Lambda)$ and $\omega_{\Lambda\cup\partial\Lambda} \in \Omega_0^{\Lambda\cup\partial\Lambda}$ (note that in this proof, we deviate from the previous use of the symbol V to denote the infinite vertex set of the tree). By induction, it is enough to check the above when $\Delta = \Lambda = \{i\}$ for any $i \in \partial\Lambda$. In this case, we have $V = \partial i \setminus \{i_\Lambda\}$ and we find that

$$\begin{aligned}
&\sum_{w_V \in \Omega_0^V} Z_\Delta^{-1} \prod_{k\in\partial\Delta} l_{kk_\Delta}(\omega_k) \prod_{b\,:\,b\cap\Delta\neq\emptyset} Q_b(\omega_b) \\
&= Z_\Lambda^{-1} \prod_{k\in\partial\Lambda\setminus\{i\}} l_{kk_\Delta}(\omega_k) \prod_{b\,:\,b\cap\Delta\neq\emptyset} Q_b(\omega_b) \times \sum_{\omega_V\in\Omega_0^V} \left(\prod_{k\in V} l_{ki}(\omega_k) Q_{ki}(\omega_k,\omega_i)\right)
\end{aligned}$$

where

$$\sum_{\omega_V\in\Omega_0^V} \left(\prod_{k\in V} l_{ki}(\omega_k) Q_{ki}(\omega_k,\omega_i)\right) = \prod_{k\in V}\left(\sum_{\omega_k\in\Omega_0} l_{ki}(\omega_k) Q_{ki}(\omega_k,\omega_i)\right)$$

Thus by the boundary law property,

$$\sum_{w_V\in\Omega_0^V} Z_\Delta^{-1} \prod_{k\in\partial\Delta} l_{kk_\Delta}(\omega_k) \prod_{b\,:\,b\cap\Delta\neq\emptyset} Q_b(\omega_b) = Z^{-1}_{\Delta\, c_{ii_\Lambda}} \prod_{k\in\partial\Lambda} l_{kk_\Delta}(\omega_k) \prod_{b\,:\,b\cap\Delta\neq\emptyset} Q_b(\omega_b)$$

Summing over $\omega_{\Lambda\cup\partial\Lambda}$ shows that $Z_{\Delta\, c_{ii_\Lambda}} = Z_\Lambda$. This establishes the consistency. In the next step, we will show that every measure constructed in this way is a tree-indexed Markov chain. Let $(i,j) \in E$ and $\Lambda \Subset (-\infty, ij)$ be any finite connected set in the past of this edge with $j \in \partial\Lambda$. Furthermore, let $x, y \in \Omega_0$ and $\omega_{(\Lambda\cup\partial\Lambda)\setminus\{j\}} \in \Omega_0^{(\Lambda\cup\partial\Lambda)\setminus\{j\}}$. Substituting the finite-volume representation formula in terms of the boundary law, we obtain

$$\frac{\mu\big(\sigma_j = y \,\big|\, \sigma_{(\Lambda\cup\partial\Lambda)\setminus\{j\}} = \omega_{(\Lambda\cup\partial\Lambda)\setminus\{j\}}\big)}{\mu\big(\sigma_j = x \,\big|\, \sigma_{(\Lambda\cup\partial\Lambda)\setminus\{j\}} = \omega_{(\Lambda\cup\partial\Lambda)\setminus\{j\}}\big)} = \frac{l_{ji}(y) Q_{ji}(y,\omega_i)}{l_{ji}(x) Q_{ji}(x,\omega_i)}$$

Summing over $y \in \Omega_0$ yields

$$\mu\big(\sigma_j = x \,\big|\, \sigma_{(\Lambda\cup\partial\Lambda)\setminus\{j\}} = \omega_{(\Lambda\cup\partial\Lambda)\setminus\{j\}}\big) = \frac{l_{ji}(x) Q_{ji}(x,\omega_i)}{\sum_{y\in\Omega_0} l_{ji}(y) Q_{ji}(y,\omega_i)}$$

The expression on the r.h.s. of the above depends on ω via ω_i only. Taking a limit of a sequence of finite sets $\Lambda_n \uparrow V$ yields

$$\mu\big(\sigma_j = x \,\big|\, \mathcal{F}_{(-\infty,ij)}\big) \overset{\text{a.s.}}{=} \mu\big(\sigma_j = x \,\big|\, \mathcal{F}_{\{i\}}\big)$$

and therefore μ is indeed a Markov chain. In the third step, we show that μ is a Gibbs measure. Let $\Lambda \Subset V$ be any finite subset of the infinite-volume vertex set V and take any two configurations $\zeta, \omega \in \Omega$ such that $\zeta_{V\setminus\Lambda} = \omega_{V\setminus\Lambda}$. Let $\Delta \Subset V$ be any connected set such that $\Lambda \subseteq \Delta$, then

$$\begin{aligned}
\frac{\mu\big(\sigma_\Lambda = \zeta_\Lambda \,\big|\, \sigma_{(\Delta\cup\partial\Delta)\setminus\Lambda} = \omega_{(\Delta\cup\partial\Delta)\setminus\Lambda}\big)}{\mu\big(\sigma_\Lambda = \omega_\Lambda \,\big|\, \sigma_{(\Delta\cup\partial\Delta)\setminus\Lambda} = \omega_{(\Delta\cup\partial\Delta)\setminus\Lambda}\big)} &= \frac{\mu(\sigma_{\Delta\cup\partial\Delta} = \zeta_{\Delta\cup\partial\Delta})}{\mu(\sigma_{\Delta\cup\partial\Delta} = \omega_{\Delta\cup\partial\Delta})} \\
&= \prod_{b\,:\,b\cap\Delta\neq\emptyset} \frac{Q_b(\zeta_b)}{Q_b(\omega_b)} \\
&= \prod_{b\,:\,b\cap\Lambda\neq\emptyset} \frac{Q_b(\zeta_b)}{Q_b(\omega_b)} \\
&= \frac{\gamma_\Lambda(\sigma_\Lambda = \zeta_\Lambda \,|\, \omega)}{\gamma_\Lambda(\sigma_\Lambda = \omega_\Lambda \,|\, \omega)}
\end{aligned}$$

Finally we can sum over $\zeta_\Lambda \in \Omega_0^\Lambda$ and take the limit $\Delta \uparrow V$. This way we get $\mu \in \mathcal{G}(\gamma)$.

(2) Now we assume that some Markov chain $\mu \in \mathcal{G}(\gamma)$ is given. On the one hand, we can condition from the inside to the outside using the Markov property. On the other hand we can also condition from the outside to the inside using the Gibbs property. For any $(i,j) \in E$ we define transition probabilities in the usual way by $P_{ij}(x,y) = \mu\big(\sigma_j = y \,\big|\, \sigma_i = x\big)$. Let $\Lambda \Subset V$ by any finite connected set, $\zeta \in \Omega$ and $a \in \Omega_0$ be some fixed reference state. Then

$$\mu(\sigma_{\Lambda\cup\partial\Lambda} = \zeta_{\Lambda\cup\partial\Lambda}) = \mu(A)\,\mu(B\,|\,A)\,\mu(C\,|\,B) \,/\, \mu(A\,|\,B)$$

where

$$A = \{\sigma_\Lambda = a\} \quad B = \{\sigma_{\partial\Lambda} = \zeta_{\partial\Lambda}\} \quad C = \{\sigma_\Lambda = \zeta_\Lambda\}$$

By the Markov property it follows that

$$\mu(B \mid A) = \prod_{k \in \partial\Lambda} P_{k_\Lambda k}(a, \zeta_k)$$

On the other hand, we get from the Gibbs property that

$$\frac{\mu(C \mid B)}{\mu(A \mid B)} = \frac{\mu(C \mid \zeta)}{\mu(A \mid \zeta)} = \frac{\prod_{b\,:\,b \cap \Lambda \neq \emptyset} Q_b(\zeta_b)}{\prod_{b\,:\,b \subseteq \Lambda} Q_b(a, a) \prod_{k \in \partial\Lambda} Q_{k_\Lambda k}(a, \zeta_k)}$$

Hence

$$\mu(\sigma_{\Lambda \cup \partial\Lambda} = \zeta_{\Lambda \cup \partial\Lambda}) = \frac{\mu(\sigma_\Lambda = a)}{\prod_{b\,:\,b \subseteq \Lambda} Q_b(a, a)} \prod_{k \in \partial\Lambda} \frac{P_{k_\Lambda k}(a, \zeta_k)}{Q_{k_\Lambda k}(a, \zeta_k)} \prod_{b\,:\,b \cap \Lambda \neq \emptyset} Q_b(\zeta_b)$$

Therefore the finite-volume representation of the marginals as in the original equation in part (1) of the theorem holds with

$$Z_\Lambda^{-1} = \frac{\mu(\sigma_\Lambda = a)}{\prod_{b\,:\,b \subseteq \Lambda} Q_b(a, a)}$$

and the candidate for a boundary law

$$l_{ij}(x) = \frac{P_{ji}(a, x)}{Q_{ji}(a, x)}$$

for all $(i, j) \in E$ and $x \in \Omega_0$. If we set $\Delta = \Lambda \cup \{i\}$ with $i \in \partial\Lambda$ we can see the defining equation of the boundary by the consistency of μ, if we consider the steps of the proof for part (1) in the opposite direction. To prove the uniqueness of the boundary law we assume that μ admits a second representation for the form as in part (1) of the theorem with a boundary law $\left\{\tilde{l}_{ij} \,\middle|\, (i, j) \in E\right\}$ and normalizing constants $\tilde{z}_\Lambda > 0$. Apply part (1) to the singleton $\Lambda = \{i\}$ and a configuration ω with $\omega_j = x$ for some $j \in \partial i$ and $\omega_k = a$ for all $k \in i \cup (\partial i \setminus \{j\})$. We obtain

$$\frac{\tilde{Z}_i}{Z_i} = \frac{\tilde{l}_{ji}(x)}{l_{ji}(x)} \prod_{k \in \partial i \setminus \{j\}} \frac{\tilde{l}_{ik}(a)}{l_{ik}(a)}$$

and hence $\tilde{l} = l$ up to a positive pre-factor (in general depending on the directed edge). This completes the proof of the theorem. □

Remark. If $l = 1$ is a solution to the boundary law equation we find that this representation is the marginal distribution distribution of a Markov chain $\mu^{\text{free}} \in \mathcal{G}(\gamma)$, called the free Gibbs measure For a boundary law $l \neq 1$, we get a Gibbs measure that is different from this free solution. As every extremal Gibbs measure is a Markov chain theorem 2.7 gives us that $|\mathcal{G}(\gamma)| = 1$ if and only if there exists a unique solution to the boundary law equation.

3.3.2 Completely homogeneous tree-indexed Markov chains on Cayley trees: the Ising and Potts models

In the following we will take a closer look at completely homogeneous Markov chains $\mu \in \mathcal{G}(\gamma)$ on Cayley trees. A Cayley tree of order $k \in \mathbb{N}$, denoted by $\mathcal{CT}(k)$, is an infinite tree where each vertex has $k + 1$ nearest neighbors. The same object is equivalently called a $k + 1$-regular tree. In the case $k = 2$, one commonly speaks of a binary tree. A Markov specification γ on $\mathcal{CT}(k)$ is said to be completely homogeneous with transfer matrix Q if γ can be expressed as this representation with $Q_b = Q$ for all $b \in E$. Recall that in the proof of part (2) of theorem 2.7, we have not only shown that every $\mu \in \mathcal{G}(\gamma)$ admits a representation in the form of part (1) of theorem 2.7, but also that the boundary law is given by $l_{ij} = \frac{P_{ji}(a,x)}{Q_{ji}(a,x)}$ (up to an (i, j)-dependent constant).

Therefore μ is completely homogeneous if and only if $l_{ij} = l$ (up to edge-dependent multiplicative constants) for all $(i, j) \in E$ and some $l \in (0, \infty)^{\Omega_0}$.

As every boundary law is only unique up to a factor we may normalize at a reference state $a \in \Omega_0$. We say that a boundary law $\{l_{ij}\}_{(i,j)\in E}$ is normalized at a if $l_{ij}(a) = 1$ for all $(i, j) \in E$. If l corresponds to a completely homogeneous Markov chain it has to meet

$$l(x) = \left(\sum_{y \in \Omega_0} \frac{Q(x, y)}{Q(a, y)} \right)^k$$

The Ising model in zero magnetic field

In the Ising model the local state space is $\Omega_0 = \{-1, 1\}$. We have some interaction strength $J > 0$, which is fixed and a nearest-neighbor interaction potential Φ such that

$$\Phi_{i,j}(\omega_i, \omega_j) = -J\,\omega_i\,\omega_j$$

The corresponding Markov specification γ is completely homogeneous and the transfer matrix Q is given by

$$Q(-,-) = Q(+,+) = \exp(J) \quad Q(-,+) = Q(+,-) = \exp(-J)$$

According to our previous discussion there is a one-to-one correspondence be- tween the completely homogeneous Markov chains $\mu \in \mathcal{G}(\gamma)$ and the positive solutions $s > 0$ of

$$\left(\frac{Q(-,+) + sQ(+,+)}{Q(-,-) + sQ(+,-)}\right)^k = \left(\frac{s\exp(J) + \exp(-J)}{\exp(J) + s\exp(-J)}\right)^k$$

Above, we normalize at $a = -1$ and hence may look for boundary laws of the form $l = (1, s)$. Introducing a new variable $t = \frac{1}{2}\log s$, the equation above is equivalent to

$$t = \frac{k}{2}\log\frac{\cosh(J+t)}{\cosh(J-t)} =: f_J(t)$$

The r.h.s. of the above is an odd function in t which is concave for $t > 0$ and convex for $t < 0$. Hence the equation has only the trivial solution $s = 1$ if and only if $f_J'(0) = k\tanh(J) \leqslant 1$. If $f_J'(0) > 1$ then we find two additional solutions $\pm s_*$ to the trivial one. Hence, there is a phase transition in this case as every solution s corresponds to a completely homogeneous Markov chain $\mu_s \in \mathcal{G}(\gamma)$. We only know that there is one completely homogeneous Markov chain $\mu \in \mathcal{G}(\gamma)$ if $f_J'(0) = k\tanh(J) \leqslant 1$. It can be shown that there actually is only one Gibbs measure overall in this case, which means that $J = a\tanh(1/k)$ is the sharp threshold for phrase transition in this model [8] (Theorem 12.31). This provides a proof of theorem 2.1.

The Potts model

In the Potts model the local state space is given by $\Omega_0 = \{1, \ldots, q\} \simeq \mathbb{Z}_q$ and the nearest-neighbor potential is $\Phi_{i,j}(\omega_i, \omega_j) = \beta\,\mathbf{1}_{\{\omega_i = \omega_j\}}$, which gives us for the transfer matrix,

$$Q(\omega_i, \omega_j) = \exp\left(\beta\,\mathbf{1}_{\{\omega_i=\omega_j\}}\right) = \theta^{\mathbf{1}_{\{\omega_i=\omega_j\}}}$$

with $\theta = \exp(\beta)$. The homogeneous boundary law equation is

$$l(s) = c\left(\sum_{\tilde{s}} l(\tilde{s})Q(\tilde{s}, s)\right)^k$$

and hence for all $s \in \{1, \ldots, q-1\}$

$$\frac{l(s)}{l(q)} = \left(\frac{l(s)(\theta-1) + \sum_{\tilde{s}=1}^{q-1} l(\tilde{s}) + l(q)}{l(q)\,\theta + \sum_{\tilde{s}=1}^{q-1} l(\tilde{s})}\right)^k$$

For $z_s := \frac{l(s)}{l(q)} \in (0, \infty)$, the above yields

$$z_s = \left(\frac{z_s(\theta-1) + \sum_{\tilde{s}=1}^{q-1} z_{\tilde{s}} + 1}{\theta + \sum_{\tilde{s}=1}^{q-1} z_{\tilde{s}}}\right)^k$$

The solutions to this $(q-1)$-dimensional fixed-point equation are in a one-to-one correspondence with the completely tree-indexed Markov chains $\mu \in \mathcal{G}(\gamma)$.

Proposition 3.22. *For any solution $z = (z_1, \ldots, z_{q-1})$ of the equation above, there exists a set $M \subseteq \{1, \ldots, q-1\}$ and some $z^* > 0$ such that*

$$z_s = \begin{cases} z^* & s \in M \\ 1 & s \notin M \end{cases}$$

Proof. Assume that we have a solution of the boundary law equation. Define the set M to be the set of indices for which the entry of the boundary is different from 1. We will show the boundary law entries will have to be the same for all indices in M. Indeed, take $\theta \neq 1$ and assume w.l.o.g. that $|M| = m$ with $M = \{1, \dots, m\}$. Define $x_s := z_s^{1/k}$ then

$$x_s = \frac{(\theta-1)x_s^k + \left(\sum_{j=1}^m x_j^k + q - m\right)}{\sum_{j=1}^m x_j^k + (q-m-1) + \theta}$$

where $z_s = 1$ for $s \notin M$ and $z_s \neq 1$ if $s \in M$. When we set $R := \sum_{j=1}^m x_j^k + q - m$ we get

$$\begin{aligned} x_s = \frac{(\theta-1)x_s^k + R}{R + \theta - 1} &\iff \left(x_s^k - x_s\right)(\theta-1) = (x_s - 1)R \\ &\iff x_s\left(x_s^{k-2} + x_s^{k-3} + \cdots + 1\right)(\theta-1) = R \end{aligned}$$

The polynomial on the l.h.s. has positive coefficients and is monotone increasing in x_s, hence injective. Therefore $x_s = x_{\tilde{s}}$ for all $s, \tilde{s} \in \{1, \dots, m\}$. □

Corollary 3.23. *Any completely homogeneous tree-indexed Markov chain $\mu \in \mathcal{G}(\gamma)$ corresponds to a solution of*

$$z = f_m(z) := \left(\frac{z(\theta + m - 1) + q - m}{mz + q - m - 1 + \theta}\right)^k$$

for some $m \in \{1, \dots, q-1\}$.

Proof. (Sketch) We focus on the binary case. For $x = \sqrt{z}$, we have

$$x = \frac{x^2(\theta + m - 1) + q - m}{mx^2 + q - m + \theta}$$

Divide out the root $x = 1$ and solve the resulting quadratic equation. Set $\theta_m = 1 + 2\sqrt{m(q-m)}$ for all $1 \leqslant m \leqslant q-1$, and note that $\theta_m = \theta_{q-m}$. We have that $\theta_1 < \theta_2 < \cdots < \theta_{\lfloor q/2 \rfloor - 1} < \theta_{\lfloor q/2 \rfloor} \leqslant q+1$ and the boundary law solutions are given by

$$x_{1,2}(m, \theta) = \frac{\theta - 1 \pm \sqrt{(\theta-1)^2 - 4m(q-m)}}{2m}$$

which exists for $\theta \geqslant \theta_m$. □

4 Neural Tangent Kernel

The notes from section 5.1 to 5.4 are compiled from [21], unless specified otherwise.

Neural networks are well known to be over-parameterized and can often easily fit data with near-zero training loss with decent generalization performance on a test dataset. Although all these parameters are initialized at random, the optimization process can consistently lead to similarly good outcomes. And this is true even when the number of model parameters exceeds the number of training data points.

4.1 Background

4.1.1 Kernel & Kernel Method

Consider a dataset $\mathcal{X} = \{x_i \in \mathbb{R}^p \mid 1 \leqslant i \leqslant n\}$. A kernel is a positive-semidefinite symmetric function of two data points, $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. It describes how sensitive the prediction for one data sample is to the prediction for the other; or in other words, how similar two data points are.

Depending on the problem structure, some kernels can be decomposed into the inner product of the features of the two data points:

$$\mathcal{K}(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$

where $\phi : \mathbb{R}^p \to \mathbb{R}^{\tilde{p}}$ is a feature map (note that $\tilde{p}$ is not necessarily less than p).

Kernel methods are a type of non-parametric, instance-based machine learning algorithms. For example, consider this dataset,

$$\mathcal{X} \times \mathcal{Y} = \{(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\} \mid 1 \leqslant i \leqslant n\}$$

then a kernelized binary classifier typically computes the label for a new input $x \in \mathbb{R}^p$ by a weighted sum of similarities,

$$\hat{y} = \operatorname{sgn}\left(\sum_{i=1}^{n} w_i \, y_i \, \mathcal{K}(x_i, x)\right)$$

where $\{w_i \mid 1 \leqslant i \leqslant n\}$ are weights determined by the learning algorithm.

4.1.2 Gaussian Processes

A Gaussian process (GP) is a non-parametric method by modeling a multivariate Gaussian probability distribution over a collection of random variables. We assume a prior over functions and then updates the posterior over functions based on what data points are observed.

Given a dataset, we assume that the data points follow a joint multivariate Gaussian distribution, defined by a mean μ, and a covariance matrix Σ, such that $\Sigma_{ij} = \mathcal{K}(x_i, x_j)$, where $\mathcal{K}$ is known as a covariance function. The core idea is that if two data points are deemed similar by the kernel, the function outputs should be close, too. Making predictions with a GP for unknown data points is equivalent to drawing samples from this distribution, via a conditional distribution of unknown data points given observed ones.

4.1.3 Notation

Consider a fully-connected neural network with parameter θ, $f(\cdot \mid \theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$. Layers are indexed from 0 (input) to L (output), each containing $n_0, \ldots, n_L$ neurons, including the input of size n_0 and the output of size n_L. There are $P = \sum_{l=0}^{L-1} (n_l + 1) n_{l+1}$ parameters in total and thus we have $\theta \in \mathbb{R}^P$.

Define the training dataset is defined as $\mathcal{D} = \mathcal{X} \times \mathcal{Y} = \{(\mathbf{x}_i, y_i) \mid 1 \leqslant i \leqslant n\}$.

Now consider the forward pass in every layer. For all $0 \leqslant l \leqslant L - 1$, each layer l defines an affine transformation $A^{(l)}$ with a weight matrix $\mathbf{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l+1}}$ and a bias term $\mathbf{b}^{(l)} \in \mathbb{R}^{n_{l+1}}$, as well as a pointwise non-linearity function σ, which is Lipschitz continuous.

$$\begin{aligned}
A^{(0)} &= \mathbf{x} \\
\tilde{A}^{(l+1)}(\mathbf{x}) &= \frac{1}{\sqrt{n_l}} \mathbf{W}^{(l)\top} A^{(l)} + \beta \, \mathbf{b}^{(l)} \in \mathbb{R}^{n_{l+1}} && \text{pre-activations} \\
A^{(l+1)}(\mathbf{x}) &= \sigma\left(\tilde{A}^{(l+1)}(\mathbf{x})\right) \in \mathbb{R}^{n_{l+1}} && \text{post-activations}
\end{aligned}$$

Note that the NTK parameterization applies a rescale weight $\frac{1}{\sqrt{n_l}}$ on the transformation to avoid divergence with infinite-width networks. The constant scalar $\beta \geqslant 0$ controls how much effort the bias terms have.

All the network parameters are initialized as i.i.d standard Gaussians in the following analysis.

4.2 Basics

The neural tangent kernel (NTK) [10], is a kernel to explain the evolution of neural networks during training via gradient descent. It leads to great insights into why neural networks with enough width can consistently converge to a global minimum when trained to minimize an empirical loss. Below, we will do a deep dive into the motivation and definition of the NTK, as well as the proof of a deterministic convergence at different initializations of neural networks with infinite width by characterizing the NTK in such a setting.

Let's start with the intuition behind NTK..

The empirical loss function $\mathscr{L} : \mathbb{R}^P \to \mathbb{R}^+$ to minimize during training is defined as follows, using a per-sample cost function $\ell : \mathbb{R}^{n_0} \times \mathbb{R}^{n_L} \to \mathbb{R}^+$,

$$\mathscr{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i \mid \theta), y_i)$$

and according to the chain rule, the gradient of the loss is

$$\nabla_\theta \mathscr{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\nabla_\theta f(\mathbf{x}_i \mid \theta)}_{p \times n_L} \cdot \underbrace{\frac{\partial \ell(f(\mathbf{x}_i \mid \theta), y_i)}{\partial f(\mathbf{x}_i \mid \theta)}}_{n_L \times 1}$$

When tracking how the network parameter θ evolves over time, each gradient descent update introduces a small incremental change of an infinitesimal step size. Since the update step is small enough, it can be approximately viewed as a derivative on the time dimension:

$$\frac{\partial \theta}{\partial t} = -\nabla_\theta \mathscr{L}(\theta)$$

Again, by the chain rule, the network output evolves as

$$\frac{\partial f(\mathbf{x} \mid \theta)}{\partial t} = \frac{\partial f(\mathbf{x} \mid \theta)}{\partial \theta}\frac{\partial \theta}{\partial t} = -\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta f(\mathbf{x} \mid \theta)^\top \nabla_\theta f(\mathbf{x} \mid \theta) \frac{\partial \ell(f(\mathbf{x}_i \mid \theta), y_i)}{\partial f(\mathbf{x}_i \mid \theta)}$$

The NTK is thus defined as

$$\mathcal{K} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R} \quad \mathcal{K}(\mathbf{x}, \tilde{\mathbf{x}} \mid \theta) = \nabla_\theta f(\mathbf{x} \mid \theta)^\top \nabla_\theta f(\tilde{\mathbf{x}} \mid \theta)$$

where each entry in the Gram matrix induced from the dataset is

$$\mathcal{K}_{mn}(\mathbf{x}, \tilde{\mathbf{x}} \mid \theta) = \sum_{p=1}^{P} \frac{\partial f_m(\mathbf{x} \mid \theta)}{\partial \theta_p}\frac{\partial f_n(\tilde{\mathbf{x}} \mid \theta)}{\partial \theta_p}$$

Note that the feature map is given by $\phi(\mathbf{x}) = \nabla_\theta f(\mathbf{x} \mid \theta)$,

4.3 Infinite Width Networks

To understand why the effect of one gradient descent is so similar for different initializations of network parameters, several pioneering theoretical work starts with infinite width networks. We will consider how the NTK guarantees that infinite width networks can converge to a global minimum when trained to minimize an empirical loss.

4.3.1 Connection with Gaussian Processes

Deep neural networks have deep connection with gaussian processes [17]. The output functions of an L-layer network, $f_i(\mathbf{x} \mid \theta)$ for $i = 1, \ldots, n_L$ are i.i.d. centered Gaussian processes of covariance $\Sigma^{(L)}$, defined recursively as

$$\begin{aligned}
\Sigma^{(1)}(\mathbf{x}, \tilde{\mathbf{x}}) &= \frac{1}{n_0}\mathbf{x}^\top \tilde{\mathbf{x}} + \beta^2 \\
\lambda^{(l+1)}(\mathbf{x}, \tilde{\mathbf{x}}) &= \begin{bmatrix} \Sigma^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}) & \Sigma^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}) \\ \Sigma^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}) & \Sigma^{(l)}(\mathbf{x}, \tilde{\mathbf{x}}) \end{bmatrix} \\
\Sigma^{(l+1)}(\mathbf{x}, \tilde{\mathbf{x}}) &= \mathop{\mathbb{E}}_{(X, \tilde{X}) \sim \mathcal{N}(0, \lambda^{(l)})} \Big[\sigma(f(X)) \cdot \sigma(f(\tilde{X}))\Big] + \beta^2
\end{aligned}$$

We proceed by induction [14]:

(1) Let's start with $L = 1$, when there is no non-linearity function and the input is only processed by a simple affine transformation

$$f(\mathbf{x} \mid \theta) = \tilde{A}^{(1)}(\mathbf{x}) = \frac{1}{\sqrt{n_0}}\mathbf{W}^{(0)\top}\mathbf{x} + \beta \mathbf{b}^{(0)}$$

where $\forall\, 1 \leqslant m \leqslant n_1$

$$\tilde{A}^{(1)}_m(\mathbf{x}) = \frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0} \mathbf{W}^{(0)}_{im} x_i + \beta \mathbf{b}^{(0)}_m$$

Since the weights and biases are initialized i.i.d., all the output dimensions of this network $\tilde{A}^{(1)}_m(\mathbf{x}), \ldots, \tilde{A}^{(1)}_{n_1}(\mathbf{x})$ are also i.i.d. Given different inputs, the m^{th} network outputs $\tilde{A}^{(1)}_m(\cdot)$ have a joint multivariate Gaussian distribution, equivalent to

a Gaussian process with covariance function (with mean $\mu_w = \mu_b = 0$ and variance $\sigma_w^2 = \sigma_b^2 = 1$).

$$\begin{aligned}
\Sigma^{(1)}(\mathbf{x}, \tilde{\mathbf{x}}) &= \mathbb{E}\left[\tilde{A}_m^{(1)}(\mathbf{x})\tilde{A}_m^{(1)}(\tilde{\mathbf{x}})\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\mathbf{W}_{im}^{(0)}x_i + \beta\mathbf{b}_m^{(0)}\right)\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\mathbf{W}_{im}^{(0)}\tilde{x}_i + \beta\mathbf{b}_m^{(0)}\right)\right] \\
&= \frac{1}{n_0}\sigma_w^2\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}x_i\tilde{x}_j + \frac{\beta\mu_b}{\sqrt{n_0}}\sum_{i=1}^{n_0}\mathbf{W}_{im}(x_i + \tilde{x}_i) + \sigma_b^2\beta^2 \\
&= \frac{1}{n_0}\mathbf{x}^\top\tilde{\mathbf{x}} + \beta^2
\end{aligned}$$

(2) We first assume the proposition holds for $L = l$, an l-layer network, and thus $\tilde{A}_m^{(l)}(\cdot)$ is a Gaussian process with covariance $\Sigma^{(l)}$ and $\left\{\tilde{A}_i^{(l)} \,\middle|\, 1 \leqslant I \leqslant n_l\right\}$ are i.i.d.

Then we need to prove the proposition also holds for $L = l + 1$. We compute the outputs by

$$f(\mathbf{x}\,|\,\theta) = \tilde{A}^{(l+1)}(\mathbf{x}) = \frac{1}{\sqrt{n_l}}\mathbf{W}^{(l)\top}\sigma\left(\tilde{A}^{(l)}(\mathbf{x})\right) + \beta\mathbf{b}^{(l)}$$

where $\forall\, 1 \leqslant m \leqslant n_{l+1}$

$$\tilde{A}_m^{(l+1)}(\mathbf{x}) = \frac{1}{\sqrt{n_l}}\sum_{i=1}^{n_l}\mathbf{W}_{im}^{(l)}\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right) + \beta\mathbf{b}_m^{(l)}$$

We can infer that the expectation of the sum of contributions of the previous hidden layers is zero:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{W}_{im}^{(l)}\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)\right] &= \mathbb{E}\left[\mathbf{W}_{im}^{(l)}\right]\mathbb{E}\left[\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)\right] = \mu_w\mathbb{E}\left[\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)\right] = 0 \\
\mathbb{E}\left[\left(\mathbf{W}_{im}^{(l)}\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)\right)^2\right] &= \mathbb{E}\left[\left(\mathbf{W}_{im}^{(l)}\right)^2\right]\mathbb{E}\left[\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)^2\right] = \sigma_w^2\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) = \Sigma^{(l)}(\mathbf{x}, \mathbf{x})
\end{aligned}$$

Since $\left\{\tilde{A}_i^{(l)} \,\middle|\, 1 \leqslant I \leqslant n_l\right\}$ are i.i.d., according to the CLT, when the hidden layer gets infinitely wide, i.e., $n_l \to \infty$, it follows that $\tilde{A}_m^{(l+1)}(\mathbf{x})$ is Gaussian distributed with variance $\beta^2 + \mathbb{V}\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)$. Note that $\tilde{A}_1^{(l+1)}(\mathbf{x}), \ldots, \tilde{A}_{n_{l+1}}^{(l+1)}(\mathbf{x})$ are still i.i.d.

$\tilde{A}_m^{(l+1)}(\cdot)$ is equivalent to a Gaussian process with covariance function,

$$\begin{aligned}
\Sigma^{(l+1)}(\mathbf{x}, \tilde{\mathbf{x}}) &= \mathbb{E}\left[\tilde{A}_m^{(l+1)}(\mathbf{x}) \cdot \tilde{A}_m^{(l+1)}(\tilde{\mathbf{x}})\right] \\
&= \frac{1}{n_l}\sigma\left(\tilde{A}_i^{(l)}(\mathbf{x})\right)^\top\sigma\left(\tilde{A}_i^{(l)}(\tilde{\mathbf{x}})\right) + \beta^2
\end{aligned}$$

When $n_l \to \infty$, according to the CLT,

$$\Sigma^{(l+1)}(\mathbf{x}, \tilde{\mathbf{x}}) \to \mathop{\mathbb{E}}_{(X,\tilde{X})\sim\mathcal{N}\left(0,\lambda^{(l)}\right)}\left[\sigma(f(X))^\top\sigma(f(\tilde{X}))\right] + \beta^2$$

The form of Gaussian processes in the above process is referred to as the *Neural Network Gaussian Process* (NNGP) [14]

4.3.2 Deterministic Neural Tangent Kernel

Finally we are now prepared enough to look into the most critical proposition from the NTK paper:

When $\forall\, 1 \leqslant l \leqslant L : n_l \to \infty$ (i.e., network with infinite width), the NTK converges to be

1. deterministic at initialization, meaning that the kernel is irrelevant to the initialization values and only determined by the model architecture, and
2. stays constant during training.

The proof relies on mathematical induction as well:

(1) First, note that $K^{(0)} = 0$. When $L = 1$, we can get the representation of the NTK directly. It is deterministic and does not depend on the network initialization. There is no hidden layer, so there is nothing to take to the infinite width.

$$
\begin{aligned}
f(\mathbf{x}\,|\,\theta) = \tilde{A}(\mathbf{x}) &= \frac{1}{\sqrt{n_0}}\mathbf{W}^{(0)}\mathbf{x} + \beta\,\mathbf{b}^{(0)} \\
\mathcal{K}^{(1)}(\mathbf{x},\,\tilde{\mathbf{x}}\,|\,\theta) &= \left(\frac{\partial f(\tilde{\mathbf{x}}\,|\,\theta)}{\partial \mathbf{W}^{(0)}}\right)^{\top}\frac{\partial f(\mathbf{x}\,|\,\theta)}{\partial \mathbf{W}^{(0)}} + \left(\frac{\partial f(\tilde{\mathbf{x}}\,|\,\theta)}{\partial \mathbf{b}^{(0)}}\right)^{\top}\frac{\partial f(\mathbf{x}\,|\,\theta)}{\partial \mathbf{b}^{(0)}} \\
&= \frac{1}{n_0}\mathbf{x}^{\top}\tilde{\mathbf{x}} + \beta^2 \\
&= \Sigma^{(1)}(\mathbf{x},\,\tilde{\mathbf{x}})
\end{aligned}
$$

(2) When $L = l$, we assume that an l-layer network with $\tilde{P}$ parameters, $\tilde{\theta} = \left(\mathbf{W}^{(0)},\,\ldots,\,\mathbf{W}^{(l-1)},\,\mathbf{b}^{(0)},\,\ldots,\,\mathbf{b}^{(l-1)}\right) \in \mathbb{R}^{\tilde{P}}$, has a NTK converging to a deterministic limit when $n_1,\,\ldots,\,n_{l-1} \to \infty$.

$$
\mathcal{K}^{(l)}(\mathbf{x},\,\tilde{\mathbf{x}}\,|\,\theta) = \nabla_{\tilde{\theta}}\tilde{A}^{(l)}(\mathbf{x})^{\top}\nabla_{\tilde{\theta}}\tilde{A}^{(l)}(\tilde{\mathbf{x}}) \longrightarrow \mathcal{K}_{\infty}^{(l)}(\mathbf{x},\,\tilde{\mathbf{x}})
$$

Note that $\mathcal{K}_{\infty}^{(l)}$ has no dependency on θ.

Now let's consider the case where $L = l + 1$. Compared to an l-layer network, an $(l+1)$-layer network has additional weight matrix $\mathbf{W}^{(l)}$ and bias $\mathbf{b}^{(l)}$, and thus the total parameters contain $\theta = \left(\tilde{\theta},\,\mathbf{W}^{(l)},\,\mathbf{b}^{(l)}\right)$. The output of this $(l+1)$-layer network is

$$
f(\mathbf{x}\,|\,\theta) = \tilde{A}^{(l+1)}(\mathbf{x}\,|\,\theta) = \frac{1}{\sqrt{n_l}}\mathbf{W}^{(l)\top}\sigma\left(\tilde{A}^{(l)}(\mathbf{x})\right) + \beta\,\mathbf{b}^{(l)}
$$

And we know its derivatives with respect to a different set of parameters. Denote $\tilde{A}^{(l)} = \tilde{A}^{(l)}(\mathbf{x})$ for brevity in the following.

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(l)}} f(\mathbf{x}\,|\,\theta) &= \frac{1}{\sqrt{n_l}}\sigma\left(\tilde{A}^{(l)}\right)^{\top} \in \mathbb{R}^{1\times n_l} \\
\nabla_{\mathbf{b}^{(l)}} f(\mathbf{x}\,|\,\theta) &= \beta \\
\nabla_{\tilde{\theta}} f(\mathbf{x}\,|\,\theta) &= \frac{1}{\sqrt{n_l}}\nabla_{\tilde{\theta}}\sigma\left(\tilde{A}^{(l)}\right)\mathbf{W}^{(l)} \\
&= \frac{1}{\sqrt{n_l}}\begin{bmatrix} \dot{\sigma}\left(\tilde{A}_1^{(l)}\right)\frac{\partial \tilde{A}_1^{(l)}}{\partial \tilde{\theta}_1} & \cdots & \dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}\right)\frac{\partial \tilde{A}_{n_l}^{(l)}}{\partial \tilde{\theta}_1} \\ \vdots & & \vdots \\ \dot{\sigma}\left(\tilde{A}_1^{(l)}\right)\frac{\partial \tilde{A}_1^{(l)}}{\partial \tilde{\theta}_{\tilde{P}}} & \cdots & \dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}\right)\frac{\partial \tilde{A}_{n_l}^{(l)}}{\partial \tilde{\theta}_{\tilde{P}}} \end{bmatrix} \in \mathbb{R}^{\tilde{P}\times n_{l+1}}
\end{aligned}
$$

where $\dot{\sigma}$ is the derivative of σ and for all $1 \leqslant p \leqslant \tilde{P}$, $1 \leqslant m \leqslant n_{l+1}$,

$$
\frac{\partial f_m(\mathbf{x}\,|\,\theta)}{\partial \tilde{\theta}_p} = \sum_{i=1}^{n_l}\mathbf{W}_{im}^{(l)}\dot{\sigma}\left(\tilde{A}_i^{(l)}\right)\nabla_{\tilde{\theta}_p}\tilde{A}_i^{(l)}
$$

The NTK for this $(l+1)$-layer network can be defined accordingly

$$
\begin{aligned}
&\mathcal{K}^{(l+1)}(\mathbf{x},\,\tilde{\mathbf{x}}\,|\,\theta) \\
=&\nabla_{\theta} f(\mathbf{x}\,|\,\theta)^{\top}\nabla_{\theta} f(\tilde{\mathbf{x}}\,|\,\theta) \\
=&\nabla_{\mathbf{W}^{(l)}} f(\tilde{\mathbf{x}}\,|\,\theta)^{\top}\nabla_{\mathbf{W}^{(l)}} f(\mathbf{x}\,|\,\theta) + \nabla_{\mathbf{b}^{(l)}} f(\tilde{\mathbf{x}}\,|\,\theta)^{\top}\nabla_{\mathbf{b}^{(l)}} f(\mathbf{x}\,|\,\theta) + \nabla_{\tilde{\theta}^{(l)}} f(\tilde{\mathbf{x}}\,|\,\theta)^{\top}\nabla_{\tilde{\theta}^{(l)}} f(\mathbf{x}\,|\,\theta) \\
=&\frac{1}{n_l}\left(\sigma\left(\tilde{A}^{(l)}(\mathbf{x})\right)\sigma\left(\tilde{A}^{(l)}(\tilde{\mathbf{x}})\right)^{\top}\right) + \beta^2 + \\
&\frac{1}{n_l}\left(\mathbf{W}^{(l)\top}\begin{bmatrix} \dot{\sigma}\left(\tilde{A}_1^{(l)}(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}_1^{(l)}(\tilde{\mathbf{x}})\right)\sum_{p=1}^{\tilde{P}}\frac{\partial \tilde{A}_1^{(l)}(\mathbf{x})}{\partial \tilde{\theta}_p}\frac{\partial \tilde{A}_1^{(l)}(\tilde{\mathbf{x}})}{\partial \tilde{\theta}_p} & \cdots & \dot{\sigma}\left(\tilde{A}_1^{(l)}(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}(\tilde{\mathbf{x}})\right)\sum_{p=1}^{\tilde{P}}\frac{\partial \tilde{A}_1^{(l)}(\mathbf{x})}{\partial \tilde{\theta}_p}\frac{\partial \tilde{A}_{n_l}^{(l)}(\tilde{\mathbf{x}})}{\partial \tilde{\theta}_p} \\ \vdots & & \vdots \\ \dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}\right)\dot{\sigma}\left(\tilde{A}_1^{(l)}\right)\sum_{p=1}^{\tilde{P}}\frac{\partial \tilde{A}_{n_l}^{(l)}(\mathbf{x})}{\partial \tilde{\theta}_p}\frac{\partial \tilde{A}_1^{(l)}(\tilde{\mathbf{x}})}{\partial \tilde{\theta}_p} & \cdots & \dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}\right)\dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}\right)\sum_{p=1}^{\tilde{P}}\frac{\partial \tilde{A}_{n_l}^{(l)}(\mathbf{x})}{\partial \tilde{\theta}_p}\frac{\partial \tilde{A}_{n_l}^{(l)}(\tilde{\mathbf{x}})}{\partial \tilde{\theta}_p} \end{bmatrix}\mathbf{W}^{(l)}\right) \\
=&\frac{1}{n_l}\left(\sigma\left(\tilde{A}^{(l)}(\mathbf{x})\right)\sigma\left(\tilde{A}^{(l)}(\tilde{\mathbf{x}})\right)^{\top}\right) + \beta^2 + \\
&\frac{1}{n_l}\left(\mathbf{W}^{(l)\top}\begin{bmatrix} \dot{\sigma}\left(\tilde{A}_1^{(l)}(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}_1^{(l)}(\tilde{\mathbf{x}})\right)\mathcal{K}_{11}^{(l)} & \cdots & \dot{\sigma}\left(\tilde{A}_1^{(l)}(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}(\tilde{\mathbf{x}})\right)\mathcal{K}_{1n_l}^{(l)} \\ \vdots & & \vdots \\ \dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}_1^{(l)}(\tilde{\mathbf{x}})\right)\mathcal{K}_{n_l1}^{(l)} & \cdots & \dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}_{n_l}^{(l)}(\tilde{\mathbf{x}})\right)\mathcal{K}_{n_ln_l}^{(l)} \end{bmatrix}\mathbf{W}^{(l)}\right)
\end{aligned}
$$

It follows by the above that $\forall\, 1 \leqslant m, n \leqslant n_{l+1}$,

$$\mathcal{K}^{(l+1)}_{mn} = \frac{1}{n_l}\left(\sigma\left(\tilde{A}^{(l)}_m(\mathbf{x})\right)\sigma\left(\tilde{A}^{(l)}_n(\tilde{\mathbf{x}})\right)\right) + \beta^2 + \frac{1}{n_l}\left(\sum_{i=1}^{n_l}\sum_{j=1}^{n_l} \mathbf{W}^{(l)}_{im}\mathbf{W}^{(l)}_{in}\dot{\sigma}\left(\tilde{A}^{(l)}_i(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}^{(l)}_j(\tilde{\mathbf{x}})\right)\mathcal{K}^{(l)}_{ij}\right)$$

When $n_l \to \infty$, by the previous section, the parts in blue and green converges to $\Sigma^{(l+1)}$, while the red part converges to

$$\sum_{i=1}^{n_l}\sum_{j=1}^{n_l} \mathbf{W}^{(l)}_{im}\mathbf{W}^{(l)}_{in}\dot{\sigma}\left(\tilde{A}^{(l)}_i(\mathbf{x})\right)\dot{\sigma}\left(\tilde{A}^{(l)}_j(\tilde{\mathbf{x}})\right)\mathcal{K}^{(l)}_{\infty,ij}$$

Later, Arora et al. (2019) [1] provided a proof with a weaker limit, that does not require all the hidden layers to be infinitely wide, but only requires the minimum width to be sufficiently large.

4.3.3 Linearized Models

From the previous section, according to the derivative chain rule, we have known that the gradient update on the output of an infinite width network is as follows. For brevity, we omit the inputs in the following analysis

$$\begin{aligned}\frac{\partial f(\theta)}{\partial t} &= -\eta\, \nabla_\theta f(\theta)^\top \nabla_\theta f(\theta)\, \nabla_f \mathscr{L} \\ &= -\eta\, \mathcal{K}(\theta)\, \nabla_f \mathscr{L} \overset{*}{(*)} - \eta\, \mathcal{K}_\infty\, \nabla_f \mathscr{L}\end{aligned}$$

$(*)$: for infinite width networks. To track the evolution of θ over time, let's consider it as a function of a time step t. With Taylor expansion, the network learning dynamics can be simplified as

$$f(\theta(t)) \approx f^{\text{lin}}(\theta(t)) = f(\theta(0)) + \underbrace{\nabla_\theta f(\theta(0))}_{(*)}(\theta(t) - \theta(0))$$

$(*)$: formally, $\nabla_\theta f(\mathbf{x}\,|\,\theta)|_{\theta=\theta(0)}$. Such formation is commonly referred to as the *linearized model*, given $\theta(0)$, $f(\theta(0))$ and $\nabla_\theta f(\theta(0))$ are constants (it is simply a linear approximation of f centered at $\theta(0)$). Assuming that the incremental time step t is small and the parameter is updated by gradient descent,

$$\begin{aligned}\theta(t) - \theta(0) &= \eta\, \nabla_\theta \mathscr{L}(\theta) = -\eta\, \nabla_\theta f(\theta)^\top \nabla_f \mathscr{L} \\ f^{\text{lin}}(\theta(t)) - f(\theta(0)) &= -\eta\, \nabla_\theta f(\theta(0))^\top \nabla_\theta f(\mathcal{X}\,|\,\theta(0))\, \nabla_f \mathscr{L} \\ \frac{\partial (f(\theta(t)))}{\partial t} &= -\eta\, \mathcal{K}(\theta(0))\, \nabla_f \mathscr{L} \overset{(*)}{=} -\eta\, \mathcal{K}_\infty\, \nabla_f \mathscr{L}\end{aligned}$$

$(*)$: for infinite width networks. Eventually we get the same learning dynamics, which implies that a neural network with infinite width can be considerably simplified as governed by the above linearized model [15].

In a simple case when the empirical loss is an MSE loss, $\nabla_\theta \mathscr{L}(\theta) = f(\mathcal{X}\,|\,\theta) - \mathcal{Y}$, the dynamics of the network becomes a simple linear ODE and it can be solved in a closed form

$$\frac{\partial f(\theta)}{\partial t} = -\eta\, \mathcal{K}_\infty\, (f(\theta) - \mathcal{Y})$$

Let $g(\theta) = f(\theta) - \mathcal{Y}$ then

$$\frac{\partial g(\theta)}{\partial t} = -\eta\, \mathcal{K}_\infty\, g(\theta) \implies g(\theta) = C\, \exp(-\eta\, \mathcal{K}_\infty\, t)$$

When $t = 0$, we have $C = f(\theta(0)) - \mathcal{Y}$ and therefore

$$f(\theta) = (f(\theta(0)) - \mathcal{Y})\, \exp(-\eta\, \mathcal{K}_\infty\, t) + \mathcal{Y} = f(\theta(0))\, \exp(-\mathcal{K}_\infty\, t) + (I - \exp(-\eta\, \mathcal{K}_\infty\, t))\mathcal{Y}$$

4.3.4 Lazy Training

When a neural network is heavily over-parameterized, the model is able to learn with the training loss quickly converging to zero, but the network parameters hardly change. Lazy training refers to this phenomenon.

Let $\theta(0)$ be the initial network parameters and $\theta(T)$ be the final network parameters when the loss has been minimized to zero. The change in parameter space can be approximated with a first-order Taylor expansion,

$$\hat{y} = f(\theta(T)) \approx f(\theta(0)) + \nabla_\theta f(\theta(0))(\theta(T) - \theta(0))$$

thus

$$\Delta\theta = \theta(T) - \theta(0) \approx \frac{\|\hat{y} - f(\theta(0))\|}{\|\nabla_\theta f(\theta(0))\|}$$

Still following the first-order Taylor expansion, we can track the change in the differential of f,

$$\begin{aligned}\nabla_\theta f(\theta(T)) &\approx \nabla_\theta f(\theta(0)) + \nabla_\theta^2 f(\theta(0))\Delta\theta \\ &= \nabla_\theta f(\theta(0)) + \nabla_\theta^2 f(\theta(0)) \frac{\|\hat{y} - f(\theta(0))\|}{\|\nabla_\theta f(\theta(0))\|}\end{aligned}$$

Thus

$$\Delta(\nabla_\theta f) = \nabla_\theta f(\theta(T)) - \nabla_\theta f(\theta(0)) = \|\hat{y} - f(\mathbf{x}\,|\,\theta(0))\| \frac{\nabla_\theta^2 f(\theta(0))}{\|\nabla_\theta f(\theta(0))\|^2}$$

Lenaic Chizat, Edouard Oyallon and Francis Bach, proved that for a two-layer neural network, $\mathbb{E}[\kappa(\theta_0)] \to 0$ when the number of hidden neurons tends to infinity [3], that is the network transitions into the lazy regime.

4.4 Project

4.4.1 Model Definition

$$f : \mathbb{R}^d \to \mathbb{R} \quad f(\mathbf{x}) \coloneqq \frac{1}{\sqrt{n}} \mathbf{w}_2^\top \sigma\left(\frac{1}{\sqrt{d}} \mathbf{W}_1^\top \mathbf{x} + \beta \mathbf{b}_1\right) + \beta b_2$$

where

- $\mathbf{W}_1 \in \mathbb{R}^{d \times n}$ and $\mathbf{w}_2 \in \mathbb{R}^n$ are the weights, applied with a rescale weight to avoid divergence with infinite-width networks. They are initialized as standard Gaussians.
- $\mathbf{b}_1 \in \mathbb{R}^n$ and $b_2 \in \mathbb{R}$ are the bias terms, and the constant scalar $\beta \geqslant 0$ controls much effect the bias terms have.
- σ is a pointwise non-linear function which is Lipschitz continuous and twice differentiable.

4.4.2 Inputs

Let $\mathbf{U} \in \mathrm{O}(d)$, the group of $d \times d$ orthogonal matrices. The matrix $\mathbf{U}$ is sampled from the group using the Haar measure, which ensures a uniform distribution over $\mathrm{O}(d)$. Define $\mathbf{\Sigma}$ as a diagonal matrix, where its diagonal elements σ_i are linearly spaced between 0.01 and 1, i.e.,

$$\forall\, 1 \leqslant i \leqslant d : \sigma_i = 0.01 + \frac{0.99 \cdot (i-1)}{d-1}$$

Define $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, which by construction is a symmetric positive definite matrix. Then each input $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\, \mathbf{K})$.

4.4.3 Targets

Let $\mathbf{W}_1^* \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}_{d \times n})$ and $\mathbf{w}_2^* \sim \mathcal{N}(\mathbf{0},\, \mathbf{I}_n)$. Then the target of $\mathbf{x}$ is defined as

$$y = \frac{1}{\sqrt{n}} \mathbf{w}_2^\top \sigma\left(\frac{1}{\sqrt{d}} \mathbf{W}_1^\top \mathbf{x}\right) + \varepsilon$$

where ε is Gaussian noise with standard deviation γ.

4.4.4 NTK

Recall that the (empirical) neural tangent kernel is defined as

$$\mathcal{K} : \mathbb{R}^{\mathrm{d}} \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R} \quad \mathcal{K}(\mathbf{x}, \tilde{\mathbf{x}}) \coloneqq \sum_{\theta \in \Theta} \nabla_\theta \mathrm{f}(\mathbf{x})^\top \nabla_\theta \mathrm{f}(\tilde{\mathbf{x}})$$

where $\Theta = \{\mathbf{W}_1,\, \mathbf{b}_1,\, \mathbf{w}_2,\, \mathrm{b}_2\}$ in our model. Let

$$\begin{aligned}
\mathbf{z}_1 &= \frac{1}{\sqrt{\mathrm{d}}} \mathbf{W}_1^\top \mathbf{x} + \beta \mathbf{b}_1 \\
\mathbf{h} &= \sigma(\mathbf{z}_1) \\
z_2 &= \frac{1}{\sqrt{\mathrm{n}}} \mathbf{w}_2^\top \mathbf{h} + \beta \mathrm{b}_2 \\
\hat{\mathrm{y}} &= \sigma(z_2)
\end{aligned}$$

then

$$\begin{aligned}
\frac{\partial \mathrm{f}}{\partial \mathrm{b}_2} &= \beta \\
\frac{\partial \mathrm{f}}{\partial \mathbf{w}_2} &= \frac{1}{\sqrt{\mathrm{n}}} \mathbf{h} \\
\frac{\partial \mathrm{f}}{\partial \mathbf{b}_1} &= \frac{1}{\sqrt{\mathrm{n}}} \mathbf{w}_2 \odot \sigma'(\mathbf{z}_1) \cdot \beta \\
\frac{\partial \mathrm{f}}{\partial \mathbf{W}_1} &= \frac{1}{\sqrt{\mathrm{n}}} \mathbf{w}_2 \odot \sigma'(\mathbf{z}_1) \cdot \mathbf{x}^\top
\end{aligned}$$

where $\odot$ denotes the Hadamard product.

4.4.5 SGD updates

Consider the MSE loss,

$$\mathscr{L}(\hat{\mathrm{y}}, \mathrm{y}) = \frac{1}{2}(\hat{\mathrm{y}} - \mathrm{y})^2$$

then

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \mathrm{b}_2} &= (\hat{\mathrm{y}} - \mathrm{y}) \cdot \beta \\
\frac{\partial \mathscr{L}}{\partial \mathbf{w}_2} &= (\hat{\mathrm{y}} - \mathrm{y}) \cdot \frac{1}{\sqrt{\mathrm{n}}} \mathbf{h} \\
\frac{\partial \mathscr{L}}{\partial \mathbf{b}_1} &= (\hat{\mathrm{y}} - \mathrm{y}) \cdot \frac{1}{\sqrt{\mathrm{n}}} \mathbf{w}_2 \odot \sigma'(\mathbf{z}_1) \cdot \beta \\
\frac{\partial \mathscr{L}}{\partial \mathbf{W}_1} &= (\hat{\mathrm{y}} - \mathrm{y}) \cdot \frac{1}{\sqrt{\mathrm{n}}} \mathbf{w}_2 \odot \sigma'(\mathbf{z}_1) \cdot \mathbf{x}^\top
\end{aligned}$$

4.4.6 Experiments

At each timestep, we generate a dataset as above and run SGD with constant learning rate 0.001. We fix the input dimension as 10, the number of hidden neuron as 100, the noise standard deviation as 0.5, the number of samples as 1000, the bias hyperparameter as 0.01 and the number of timesteps as 500. We also set the vectors we evaluate the NTK at, and subsequently generate a Gram matrix, as the canonical basis.

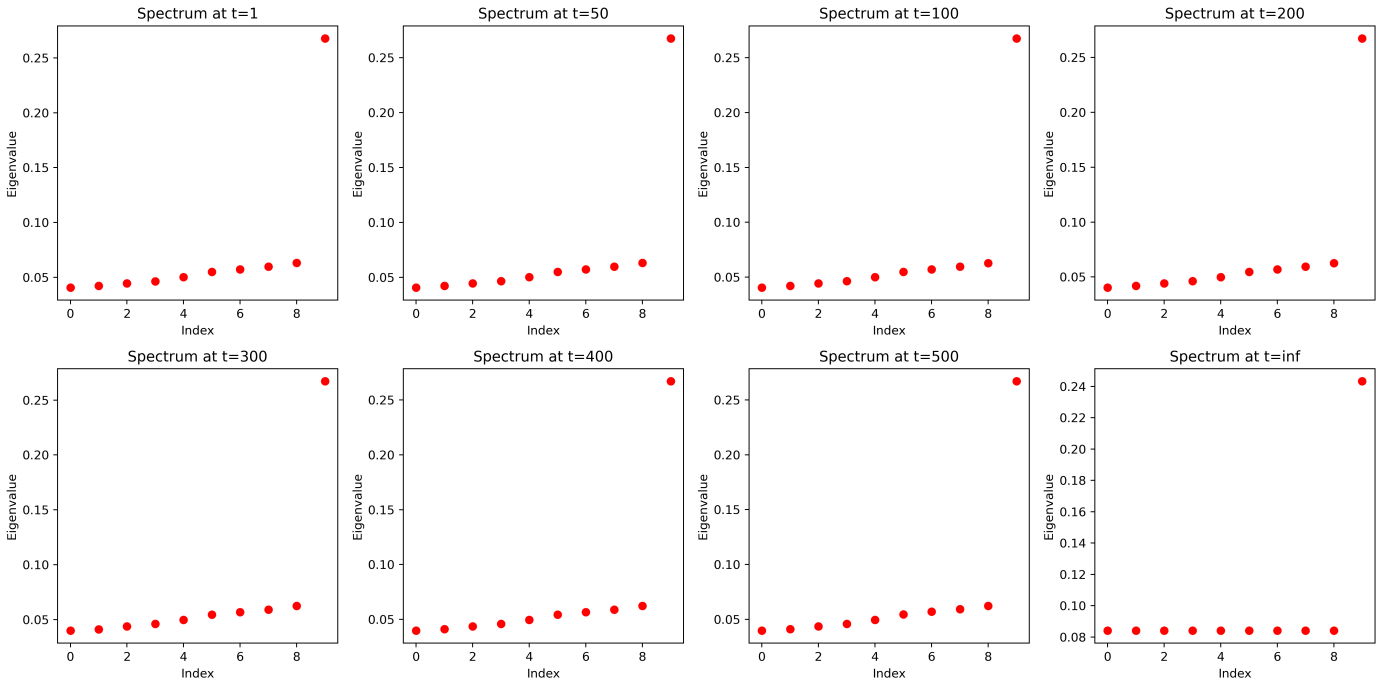

Figure 1: Evolution of the spectrum of the Gram matrix

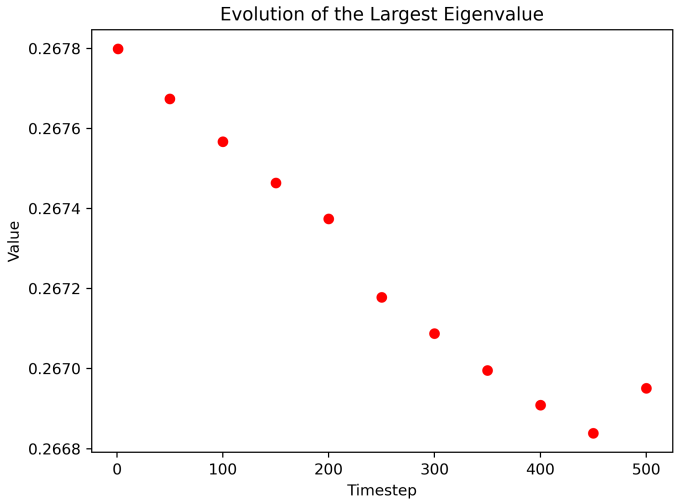

Figure 2: Evolution of the largest eigenvalue of the Gram matrix

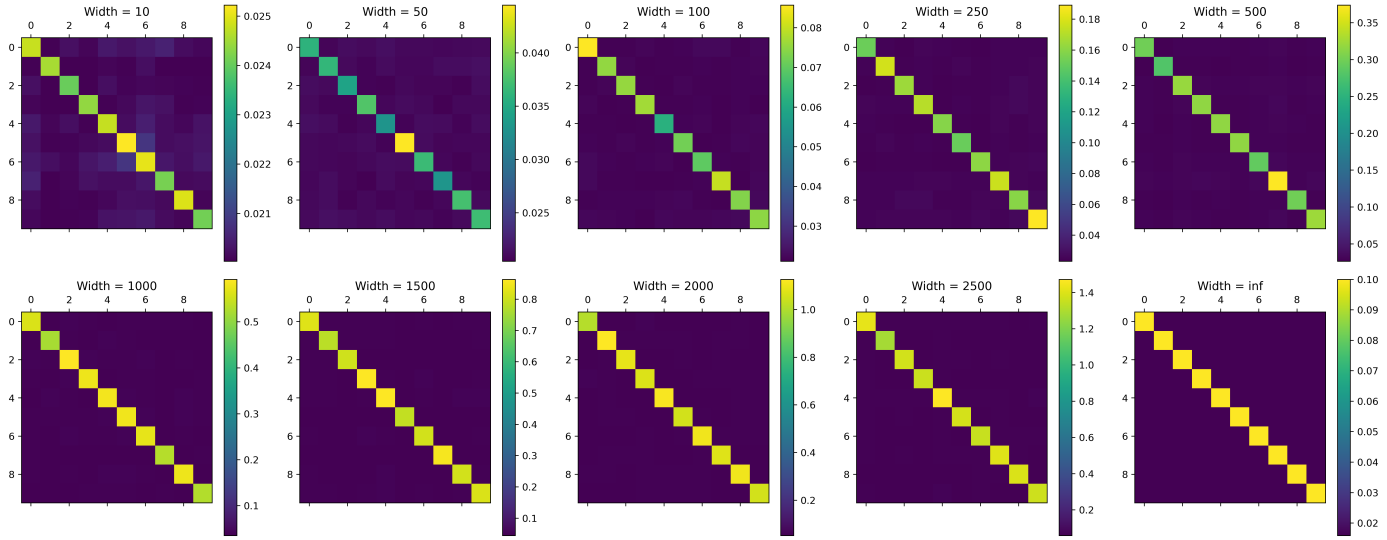

Figure 3: Evolution of the Gram matrix

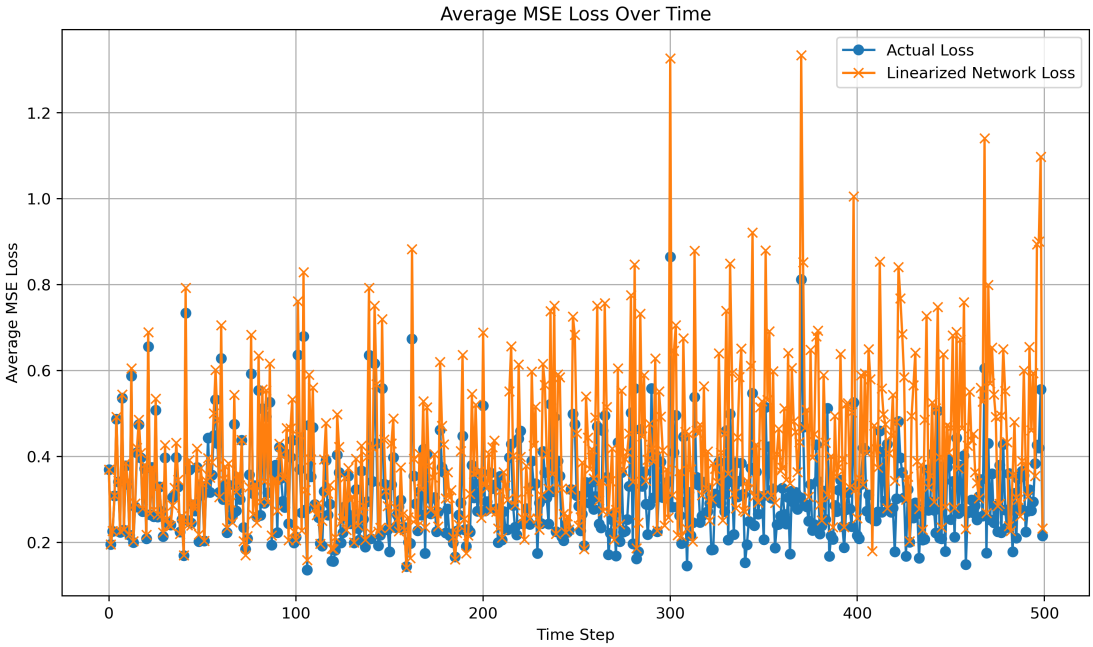

Figure 4: Evolution of the MSE loss over time

4.4.7 Expectation of the Weight Updates

For simplicity, we now consider this network:

$$f : \mathbb{R}^d \to \mathbb{R} \quad x \mapsto \sigma(\langle \theta, x \rangle)$$

where θ is initialized as a standard multivariate Gaussian, scaled by $\frac{1}{\sqrt{d}}$. Note that at each timestep $t \in \mathbb{N}$, we generate a dataset $\mathcal{D}_t = \mathcal{X}_t \times \mathcal{Y}_t$, where the inputs are generated as previously and the targets analogously. Set θ^* as a standard multivariate Gaussian, scaled by $\frac{1}{\sqrt{d}}$, then for each input $x \in \mathcal{X}_t$, define its target $y \in \mathcal{Y}_t$ as

$$y \coloneqq \langle \theta^*, x \rangle + \varepsilon$$

where ε is Gaussian noise with standard deviation γ. We now consider the SGD weight update.

$$\begin{aligned} \theta_{k+1} &= \theta_k + \eta \nabla_\theta \mathscr{L}(x, y, \theta_k) \\ \implies \mathbb{E}[\theta_{k+1} \mid \theta_k] &= \theta_k + \eta \, \mathbb{E}_\theta \left[\nabla_\theta \mathscr{L}(x, y, \theta_k) \right] \\ &= \theta_k + \eta \, \mathbb{E}_\theta \left[\nabla_\theta \left(\frac{1}{2} (\langle \theta_k, x \rangle - y)^2 \right) \right] \end{aligned}$$

Note that

$$\nabla_\theta \left(\frac{1}{2} (\langle \theta_k, x \rangle - y)^2 \right) = (\langle \theta_k, x \rangle - y)x$$

Hence

$$\nabla_\theta \left(\frac{1}{2} (\langle \theta_k, x \rangle - (\langle \theta^*, x \rangle + \varepsilon))^2 \right) = (\langle \theta_k, x \rangle - \langle \theta^*, x \rangle - \varepsilon)x = (\langle \theta_k - \theta^*, x \rangle - \varepsilon)x$$

Since ε is independent of x and has zero mean, it follows that

$$\mathbb{E}_\theta \left[\nabla_\theta \left(\frac{1}{2} (\langle \theta_k, x \rangle - y)^2 \right) \right] = \mathbb{E}_\theta [\langle \theta_k - \theta^*, x \rangle x]$$

Let $\Delta\theta = \theta_k - \theta^*$ then

$$\mathbb{E}_\theta [\langle \Delta\theta, x \rangle x] = xx^\top \Delta\theta$$

Where $xx^\top$ is the covariance matrix of x, which we'll denote as Σ. Therefore

$$\mathbb{E}[\theta_{k+1} \mid \theta_k] = \theta_k + \eta \Sigma \Delta\theta$$

It follows that averaged at timestep t,

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t + \eta \Sigma \Delta\tilde{\theta}$$

where $\Delta\tilde{\theta} = \tilde{\theta}_t - \theta^*$.

Below are visualizations which illustrates the above.

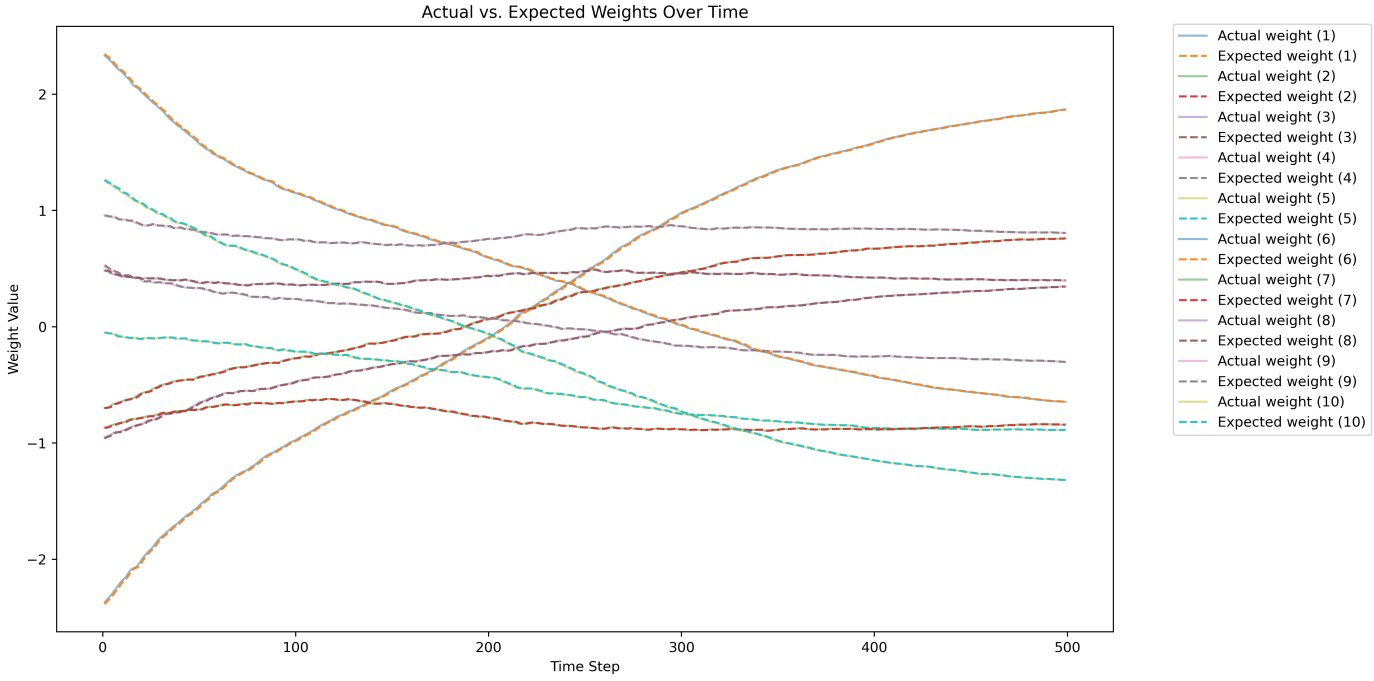

Figure 5: Actual vs. Prediction Weights over Time

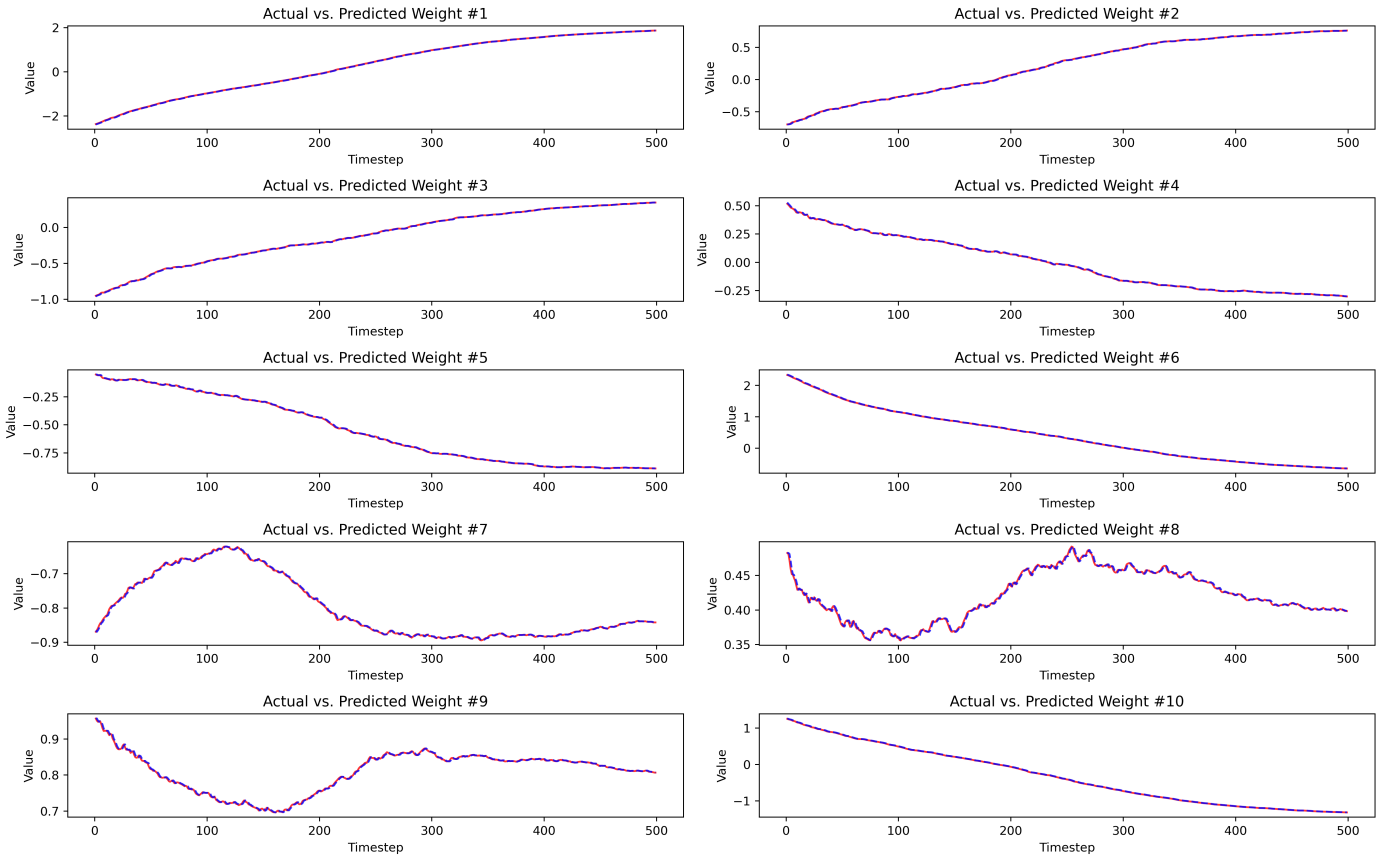

Figure 6: Actual vs. Prediction Weight over Time

Acknowledgements

I would like to express my deepest gratitude to my mentor, Noah Marshall, for his invaluable guidance, support, and encouragement throughout this project, which extended well into the summer term. His insightful feedback and expertise have been instrumental in shaping my understanding of the topics, and I thoroughly enjoyed our discussions. I am also thankful to the committee for providing this enriching opportunity and fostering an environment of academic growth and collaboration.

References

[1] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

[2] Søren Asmussen and Heinrich Hering. *Branching Processes*. Birkhäuser Boston, 1983.

[3] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

[4] J Theodore Cox. Entrance laws for markov chains. *The Annals of Probability*, pages 533–549, 1977.

[5] Richard Durrett and R Durrett. *Essentials of stochastic processes*, volume 1. Springer, 1999.

[6] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[7] Sacha Friedli and Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017.

[8] Hans-Otto Georgii. *Gibbs measures and phase transitions*. Walter de Gruyter GmbH & Co. KG, Berlin, 2011.

[9] Christopher C Heyde. Extension of a result of seneta for the super-critical galton–watson process. *Selected Works of CC Heyde*, pages 115–118, 2010.

[10] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[11] Harry Kesten and Bernt P Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.

[12] Christof Külske. Stochastic processes on trees. *Preprint, Ruhr-University, Bochum*, 2017.

[13] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

[14] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

[15] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[16] Russell Lyons and Yuval Peres. *Probability on trees and networks*, volume 42. Cambridge University Press, 2017.

[17] Radford M Neal and Radford M Neal. Priors for infinite networks. *Bayesian learning for neural networks*, pages 29–53, 1996.

[18] Elliot Paquette. High-dimensional limits of stochastic gradient descent. 2023.

[19] Eugene Seneta. On recent theorems concerning the supercritical galton-watson process. *The Annals of Mathematical Statistics*, 39(6):2098–2102, 1968.

[20] Aernout CD Van Enter, Roberto Fernández, and Alan D Sokal. Regularity properties and pathologies of position-space renormalization-group transformations: Scope and limitations of gibbsian theory. *Journal of Statistical Physics*, 72:879–1167, 1993.

[21] Lilian Weng. Understanding the neural tangent kernel, 2022.

[22] Stan Zachary. Countable state space markov random fields and markov chains on trees. *The Annals of Probability*, pages 894–903, 1983.